

UC Berkeley

UC Berkeley Electronic Theses and Dissertations

Title

Algorithmic Pursuits of Structure: Monotone Operators, Robust Estimation, and Automated Discovery

Permalink

<https://escholarship.org/uc/item/0fd270jz>

ISBN

9798293893379

Author

Chen, Zihao

Publication Date

2025-08-01

Peer reviewed|Thesis/dissertation

Algorithmic Pursuits of Structure: Monotone Operators, Robust Estimation, and
Automated Discovery

By

Zihao Chen

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Laurent El Ghaoui, Chair

Professor Nikita Zhivotovskiy

Professor Christian Borgs

Summer 2025

Algorithmic Pursuits of Structure: Monotone Operators, Robust Estimation, and
Automated Discovery

Copyright 2025
by
Zihao Chen

Abstract

Algorithmic Pursuits of Structure: Monotone Operators, Robust Estimation, and
Automated Discovery

by

Zihao Chen

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor Laurent El Ghaoui, Chair

This dissertation investigates the design and analysis of methods for a few foundational problems in optimization, statistics, and automated discovery. We begin with a theoretical exploration of finding zeros of monotone operators—which generalize convex minimization and convex-concave saddle point problems—providing a geometric characterization of continuous-time oscillatory behavior and introducing a regularization framework for stabilization and acceleration of discrete methods. We then study robust mean estimation under affine equivariance constraints, proposing a high-dimensional median-based estimator and establishing that this natural stability property incurs a dimension-dependent statistical cost. Finally, we present a decision-making methodology that combines machine learning with adaptive experimentation to guide materials discovery, demonstrating improved efficiency in exploring chemical spaces.

To my loved ones.

Contents

Contents	ii
List of Figures	iv
List of Tables	viii
1 Overview	1
2 Obstacles in Monotone Flows	4
2.1 Introduction	4
2.2 Background	6
2.3 Operator Structure on Convex Hulls	8
2.4 The Convex Hull of Orbits in Skew-Symmetric Systems	10
2.5 Synthesis and Concluding Remarks	14
3 A Unified Framework for Monotone Equation	16
3.1 Introduction	16
3.2 Background and Preliminaries	18
3.3 The Regularization Framework	22
3.4 Stabilizing and Accelerating Standard Methods	28
3.5 Conclusion	43
4 Affine-equivariant Mean Estimation	45
4.1 Introduction	45
4.2 Related Work	49
4.3 Intuition	52
4.4 Algorithm	55
4.5 Theoretical Guarantees	58
4.6 Lower Bounds	61
4.7 Summary	66
5 Algorithmic Iterative Reticular Synthesis	68
5.1 Introduction	68

5.2	Results	69
5.3	Methods	77
5.4	Conclusion and Outlook	87
Bibliography		89
A Appendix for Chapter 3		101
A.1	A technical lemma	101
B Appendix for Chapter 4		103
B.1	Auxiliary Results	103
B.2	Relationship between Corruption Models	106
B.3	Comparison to Classical Breakdown Point	109
B.4	Recent Statistical Results	113
C Appendix for Chapter 5		115
C.1	Extended data figures and tables	115
C.2	Determination of reaction condition space	115
C.3	Crystals in the initial dataset	118
C.4	Automated synthesis workflow	120
C.5	Statistical analysis of success rates of the initial dataset	122
C.6	Reaction encoding details	125
C.7	Description of ZIFs discovered	128
C.8	Synthetic experiments for the main methods	131
C.9	Investigation of RFC augmentation factor	135
C.10	Discovery trajectory analysis	136
C.11	Impact of batch selections on ZIF discovery	137

List of Figures

2.1	A single trajectory $\mathbf{z}(t)$ of the flow $\dot{\mathbf{z}} = -F(\mathbf{z})$ in a 3D space. The velocity field drives $\mathbf{z}(t)$ toward the tilted plane, yielding the planar limit cycle. The center $\mathbf{z}^*$ and initial state $\mathbf{z}_0$ are marked. The result of this paper implies that F coincides with a skew-symmetric linear operator over the (dark grey) disk.	5
4.1	An illustration of the affine-equivariance requirement. Here, the result of an estimator $\hat{\mu}$ run on the transformation of the square to the tilted parallelogram is required to coincide with the transformation of the estimate obtained when run on the square itself.	50
4.2	Illustration of hard distribution. The red dot on e_1 denotes higher probability. .	53
4.3	One dimensional projection onto e_1	54
4.4	One dimensional projections onto e_i for $i \neq 1$ and $\mathbf{1}$	55
5.1	Schematic illustration of AIRES cycle and its components. a , Automated synthesis setup showing a robotic liquid handler (left) and heating unit (right). b , Microscope imaging station (top) and optical image collected (bottom) with computer vision classification results (blue: crystal probability, green: non-crystal probability). c , Single crystal X-ray diffraction setup. d , A two-dimensional demonstration of binary reaction outcome maps for different linker systems, with successful reactions in dark blue and failed ones in white. e , A two-dimensional demonstration of ML-predicted crystallization probability maps for different linker systems, with higher predicted probabilities in darker red. The decision-making algorithm, aiming to maximize the number of distinct discoveries, suggests candidates based on predictions for subsequent experimental validation.	70

5.2	Initial ZIF synthesis dataset. a , Chemical structures and abbreviations of eight starting linkers. Abbreviations denote neutral linkers when referring to reactants, but represent their deprotonated forms in ZIF chemical formulae unless specified otherwise. b , Representative optical images (yellow scale bar 100 μm) showing different reaction outcomes: clear solution, precipitate, non-ZIF crystals (red frames, failures), and ZIF single crystals (blue frame, success). Crystal structures represented by ball-and-stick models (right) show $\text{Zn}_5(\text{bIM})_6(\text{HCOO})_4(\text{DMF})_2$ (<i>b</i> -axis view) and $\text{Zn}(\text{bIM})_2$ (<i>c</i> -axis view) with Zn polyhedra in pink. c , Success rate landscapes for each starting linker as a function of temperature and log linker-to-metal ratio, averaged over total concentration. Increasing shades of red indicate higher success rates.	72
5.3	AIRES performance in accelerating ZIF discovery. a , Linker features combining functional group counts (blue, 5 features), linker dimensions (magenta, 4 features), and quantum mechanical properties of neutral/deprotonated linkers (yellow, 5 features) and pre-nucleation species (green, 7 features). b , Two-dimensional PCA projection of the chemical space showing the distribution of 8 starting linkers (blue circles), 24 new imidazole-based (grey triangles), and 24 new benzimidazole-based linkers (red diamonds). The first two principal components capture 50.8% of the total variance. c , Cumulative reward of AIRES equipped with Random Forest Classification (RFC, red) and Gaussian Process Classification (GPC, magenta), and Random Search baseline (RS, grey). Shaded areas represent one standard deviation across 16 repetitions for RFC and GPC, and 100 repetitions for RS. The red dashed line indicates the maximum achievable reward. d , Cumulative reward of AIRES(RFC) (red) and RS (grey) under batch experimentation with 24 experiments per iteration. Shaded areas represent one standard deviation across 16 repetitions for AIRES(RFC) and 100 repetitions for RS. e , Kernel density estimation showing the distribution of the number of experiments required to discover new ZIF crystals for 10 successful linkers using AIRES(RFC) (red) versus RS (grey). Horizontal bars beneath each distribution indicate individual discovery events across multiple repetitions.	75
5.4	Comparison of greedy and Thompson Sampling strategies over 1000 trials. The solid lines show mean cumulative rewards (number of arms with observed successes), while the shaded regions represent one standard deviation. The greedy strategy significantly outperforms Thompson Sampling by focusing on arms with higher posterior means rather than being diverted by uncertainty in the lower-mean arms.	87
C.1	Chemical structures and abbreviations of 48 new linkers.	115
C.2	Distribution of primary synthesis parameters. Dashed vertical lines indicate the boundaries of selected ranges.	116

C.3	Representative optical image of ZIF-2 crystallized from IM. This image was obtained from a reaction whose conditions were: a total concentration of 0.50 mol/L, a log linker-to-metal ratio of 1.00, and a temperature of 100 °C. Yellow scale bar 100 μm	119
C.4	Success rate maps of individual linkers as functions of total concentration and temperature. The success rate is shown as surface plots for each starting linker. The color intensity corresponds to the success rate, with darker red indicating higher success rates.	122
C.5	Success rate maps of individual linkers as functions of total concentration and log linker-to-metal ratio. The success rate is shown as surface plots for each starting linker. The color intensity corresponds to the success rate, with darker red indicating higher success rates.	123
C.6	Success rates of individual linkers as functions of total concentration. Bar plots show the success rates for each linker, with darker blue indicating higher concentrations.	123
C.7	Success rates of individual linkers as functions of log linker-to-metal ratio. Bar plots show the success rates for each linker, with darker blue indicating higher ratios.	124
C.8	Success rates of individual linkers as functions of temperature. Bar plots show the success rates for each linker, with darker blue indicating higher temperatures.	124
C.9	Cumulative explained variance ratio as a function of principal components for decomposed linker features.	127
C.10	Topological classification and crystal structures of ZIFs and LZIFs discovered by AIRES from new linkers. a , Crystal structures of ZIFs (ZIF-A1 through ZIF-A11) are shown and classified based on network topologies. ZIF topologies are visualized using natural tiling models and labeled with their respective three-letter topological codes, where uppercase and lowercase respectively indicate topologies known in zeolites and only in ZIFs. Disordered positions of functional groups in crystal structure are preserved in visualization. b , Crystal structures of LZIFs (LZIF-A1 through LZIF-A4) are shown and classified based on network topologies. LZIF topologies are visualized using 2D tiling models.	128
C.11	Performance of sequential discovery algorithm across 50 trials	133
C.12	Performance of batch discovery algorithm across 50 trials (batch size $N_B = 6$).	134
C.13	Impact of data augmentation on AIRES discovery performance. The plot compares discovery rates between Random Search (RS, grey) and AIRES using RFC with different augmentation factors (5, 10, and 15). Shaded areas represent one standard deviation. The cumulative reward (y-axis) indicates the number of successful discoveries, with a maximum possible reward of 10 (dashed red line). AIRES consistently outperforms RS.	135
C.14	Linker suggestions from A single sequential trajectory of RFC. Each linker suggestion is paired with specific reaction conditions (not shown in this plot).	136

C.15 **Cumulative reward of Batch AIRES(RFC) and Batch RS under different batch sizes.** Shaded areas represent one standard deviation across 16 repetitions for Batch AIRES(RFC) and 100 repetitions for Batch RS. Batch sizes: 24 reactions per iteration (left), 48 reactions per iteration (middle), and 96 reactions per iteration (right). 137

C.16 **Cumulative reward of Batch AIRES(RFC), Batch Top-k(RFC) and Batch RS under batch size of 24.** Shaded areas represent one standard deviation across 16 repetitions for Batch AIRES(RFC) and Batch Top-k(RFC) and 100 repetitions for Batch RS. The procedures of data augmentation and hyperparameter tuning are the same for Batch AIRES(RFC) and Batch Top-k(RFC). 138

List of Tables

C.1	Synthesis conditions of reported Zn-ZIFs used for determining the reaction condition space.	117
C.2	Summary of the crystals formed from the 8 starting linkers. Linker abbreviations represent neutral form in the first column, but represent deprotonated forms in the second column.	118

Acknowledgments

Acknowledgements seem to be read more often than the main body of dissertations, given the current role of dissertations in this field. Though I read some acknowledgements as a junior graduate student, in retrospect I realize such experiences cannot really be glimpsed beforehand.

Thanks to my advisor, Laurent El Ghaoui, for his support and encouragement, and for providing enough space for me to think and grow. It is my honor to be Laurent's last PhD student at Berkeley. Thanks to Nikita Zhivotovskiy, Christian Borgs, and Aditi Krishnapriyan for serving on my dissertation and qualification committees. Thanks to Bin Yu for the mentorship, support, perspectives, and advice.

Thanks to the department staff, Carissa Caloud, Judy Smithson, and especially Shirley Salanio and Jean Nguyen. Also thanks to La Shana Porlaris from the Statistics department. I was fortunate to have benefited from their help at various stages of my graduate school journey, much of which was very important.

My time in Berkeley has overlapped with many graduate students, postdocs, and faculty. Thanks to Yeshwanth Cherapanamjeri and Zichao Rong for being my main collaborators for the work in Chapters 4 and 5, and to Xiang Cheng for the discussions on minimax games. Thanks to the broader Berkeley ML community in EECS and Statistics for countless research discussions, fun chats, lunches, courses, seminars, and reading groups.

Thanks to my good friends over the past years. Surely we are friends in most of the many worlds, and in the ones we are not, the graduate school experience would be far more dull. To Armin Askari, Yeshwanth Cherapanamjeri, Yu Sun, Chandan Singh, Ashish Kumar, Allan Jabri, and Ashvin Nair, for the gatherings at the pool table in Cory 307, almost-weekly China Village dinners, and being with me through the ups and downs. To Efe Aras for always being extremely energetic, loud, and sharing countless laughs. To Forest Yang for always spreading his happiness through daily coffee in our office. To my roommate Junhao Xiong for the hundreds of different dishes together, conversations, soccer, and nature trips.

There are certainly numerous important experiences with people, places, and things outside graduate school during these years—from old friends to strangers with long conversations, from local neighborhoods to distant nature. Thanks to all.

Finally, I would like to thank my family for always being the sun and moon for me.

Chapter 1

Overview

As machine learning continues to mature, its widespread adoption reveals both the power and limitations of our foundational tools. While these methods have proven remarkably effective, their deployment at scale exposes gaps in our theoretical understanding and highlights the need to revisit classical problems with fresh perspectives. This dissertation examines a few foundational challenges in optimization, statistics, and applications in automated material discovery:

Beyond minimization: monotone operators. First-order optimization methods have become the workhorses of modern computation, but many important problems extend beyond simple minimization. Finding equilibria—whether in games, markets, or multi-agent systems—requires different mathematical tools. Applications like generative adversarial networks [47] exemplify this shift, where we seek stable points in competitive dynamics rather than optimal values of a single function. This motivates a revisit and deeper study of monotone operators, a classical mathematical framework that naturally captures these equilibrium-seeking problems and provides the foundation for understanding their convergence properties [97, 42].

Robustness. The problem of robust estimation—recovering signals despite corrupted observations—has been central to statistics for over sixty years [62]. Robustness extends beyond improving accuracy, it imposes structural constraints that shape the design of estimators and more generally, machine learning procedures. Recent work has brought renewed attention to this classical challenge, particularly in obtaining estimation with optimal non-asymptotic guarantees [90, 30].

Non-asymptotic analysis. A defining characteristic of modern theoretical work is the emphasis on finite-sample, non-asymptotic guarantees. In optimization, we need concrete bounds on algorithm performance after a fixed number of iterations—not just convergence “in the limit”—to make informed choices about methods, stopping criteria, and computational trade-offs. Similarly, in statistics, the traditional reliance on asymptotic arguments can be misleading when dimensions are high or data is limited. This shift toward finite-sample guarantees, typically stated with high probability bounds, yields practical insights aligned with the realities of modern computation and data availability.

Machine learning for scientific discovery. Machine learning’s ability to process vast datasets and identify patterns beyond human perception is reshaping discovery pipelines across the sciences. This creates novel problem formulations driven by objectives distinct from traditional optimization. For instance, in materials discovery, the goal is not to find the single best material but to efficiently explore chemical space and maximize the rate of new discoveries [134, 45, 80]. Such applications require rethinking our objectives and developing algorithms tailored to the unique constraints of experimental science.

Organization

Chapters 2 and 3: Monotone Operator Methods. These chapters present complementary investigations into finding zeros of monotone operators—a problem central to modern optimization and equilibrium computation. Chapter 2 provides a complete geometric characterization of the oscillatory behavior exhibited by monotone flows, revealing that the flow fails to converge if and only if the trajectory enters a region where the operator becomes locally indistinguishable from a conservative, skew-symmetric linear system. Chapter 3 takes a more algorithmic perspective, showing that Tikhonov regularization provides a unified lens for understanding, stabilizing, and accelerating discrete-time first-order methods.

Chapter 4: Affine-Equivariant Mean Estimation. This chapter studies a classical problem in robust statistics: estimating the mean of a multivariate heavy-tailed distribution with adversarially corrupted outliers. We focus on the fundamental tension between robustness and affine equivariance—the property that an estimator’s behavior is consistent under linear transformations of the data. We propose a novel estimator based on a new high-dimensional median from a geometric perspective. Additionally, we provide matching upper and lower bounds, giving the first sharp characterization of the trade-off showing that affine equivariance necessarily incurs a dimension-dependent statistical cost.

Chapter 5: Automated Material Discovery. This final chapter applies machine learning to accelerate scientific discovery, specifically in synthesizing new crystalline structures. We present Algorithmic Iterative Reticular Synthesis (AIRES), a framework that combines automated experimentation with adaptive learning to efficiently explore vast chemical spaces. The key insight is that discovery requires different algorithmic strategies than (property) optimization—we show that greedy selection based on success probability better aligns with the problem structure than the exploration-heavy approaches typically used in Bayesian optimization. By learning patterns across diverse chemical systems and dynamically allocating experimental resources, AIRES achieves a two-fold acceleration in the discovery rate of a class of reticular materials and provides a template for untargeted automated discoveries in chemistry and material science.

We conclude this overview with a remark. The theory-practice gap in modern machine learning has perhaps never been larger, with engineering efforts driving most of the field's recent remarkable successes. Yet the advancement of science has always carried an implicit thread of seeking explanations for observed regularities, which come in hierarchies with different levels of abstraction. Theoretical formulation and investigation, as presented in this dissertation, offers one tool for obtaining understanding and explanations—typically at a higher level of abstraction. The first two chapters exemplify this approach: investigating oscillatory behavior and mechanisms for stabilization and acceleration in monotone operators can be viewed as understanding the obstacles and useful principles of minimax training in an idealized setting. The intention is not to provide direct utility to complex practical problems such as training of generative adversarial nets—we should not claim so. Rather, these studies seek explanations and understanding that, when done fundamentally, may have a broad reach to downstream problems. However, it is important to acknowledge the limitations of this approach. As a tool to pursue understanding, theoretical investigation is inherently constrained by the scope of the adopted mathematical formulations and the reach of proofs. Hence, modern machine learning, with its extraordinary complexity and empirical richness, should embrace multiple modes of explanation beyond existing formal mathematical frameworks.

Chapter 2

Obstacles in Monotone Flows

2.1 Introduction

The problem of finding a zero of a (single-valued) monotone operator, $\mathbf{z}^* \in \text{zer}(F)$, is a cornerstone of modern optimization and equilibrium-finding. It provides a unifying framework for problems such as convex optimization, computing a Nash equilibrium in a convex-concave game, and solving variational inequalities, etc [97, 42]. A natural and fundamental approach to this problem is to follow the trajectories of the associated monotone flow, an ordinary differential equation (ODE) given by $\dot{\mathbf{z}}(t) = -F(\mathbf{z}(t))$ (see Eq. (2.2)). This system represents the most direct continuous-time analogue of gradient-based methods. However, its behavior exhibits a well-known and curious dichotomy: while the time-average of the trajectory, $\frac{1}{T} \int_0^T \mathbf{z}(t) dt$, is known to converge to a solution [5], the trajectory $\mathbf{z}(t)$ itself—the “last iterate”—often fails to converge to a point, instead settling into persistent, oscillatory motion. A canonical counterexample is the bilinear saddle-point problem $\min_{x \in \mathbb{R}} \max_{y \in \mathbb{R}} xy$, whose associated monotone flow, governed by the linear skew-symmetric operator $F(x, y) = (y, -x) \in \mathbb{R}^2$, generates trajectories that are perfect circles instead of converging. This paper provides a complete geometric explanation for this dichotomy between asymptotic oscillations and ergodic convergence for the continuously differentiable monotone flow in Euclidean space.

Standard analysis of the monotone flow relies on a Lyapunov argument. By considering the squared Euclidean distance to a solution $\mathbf{z}^*$, one can show that this distance is non-increasing and therefore converges, guaranteeing that the trajectory remains bounded (see Eq. (2.4)). An application of LaSalle’s Invariance Principle (e.g, [67]) further reveals that the trajectory’s limit set, $\omega(\mathbf{z}_0)$, must be contained within the set of points where the Lyapunov function’s derivative vanishes. While this analysis confirms the stability of the system, it is insufficient to explain the lack of asymptotic convergence. It identifies the region where the asymptotic dynamics unfold but does not characterize the structure of the flow within that region, leaving the origin of the oscillations unexplained.

Our central contribution is to provide a complete geometric characterization of these

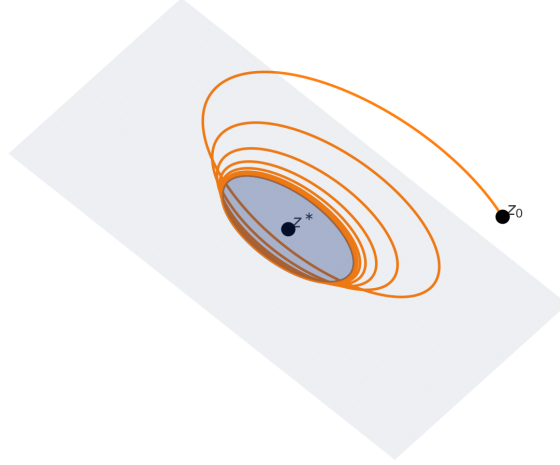


Figure 2.1: A single trajectory $\mathbf{z}(t)$ of the flow $\dot{\mathbf{z}} = -F(\mathbf{z})$ in a 3D space. The velocity field drives $\mathbf{z}(t)$ toward the tilted plane, yielding the planar limit cycle. The center $\mathbf{z}^*$ and initial state $\mathbf{z}_0$ are marked. The result of this paper implies that F coincides with a skew-symmetric linear operator over the (dark grey) disk.

non-convergent dynamics by synthesizing a series of structural results. It is known for specific classes of problems, such as convex-concave saddle-point dynamics, that the flow on the limit set is governed by a skew-symmetric linear system [55]. We first generalize this finding, showing that it holds for any continuously differentiable monotone operator. We then prove that this linear structure is not confined to the limit set's orbit but propagates to its entire convex hull. By analyzing the structure properties of orbits in skew-symmetric linear systems, we demonstrate that this convex hull contains a neighborhood around the trajectory's ergodic limit. This synthesis reveals the fundamental obstruction: asymptotic convergence fails if and only if the trajectory is drawn into a region of the state space where the operator F becomes locally indistinguishable from a conservative, skew-symmetric linear system, thereby permitting persistent, non-dissipative orbits [Figure 2.1](#).

The remainder of this paper is organized to build this argument systematically. [Section 2.2](#) reviews the classical stability analysis and establishes that the dynamics on the invariant limit set are governed by the skew-symmetric part of the operator's Jacobian. [Section 2.3](#) presents a general theorem on how a linear operator structure on a finite set of points extends to their convex hull. [Section 2.4](#) analyzes the geometry of orbits generated by skew-symmetric systems, proving that their convex hull contains an open ball in the topology of the subspace they span. Finally, [Section 2.5](#) synthesizes these results to provide a cohesive

geometric picture of non-convergent trajectories.

Notations. We denote vectors by lowercase bold letters (e.g., $\mathbf{z} \in \mathbb{R}^n$) and matrices by uppercase bold letters (e.g., $\mathbf{A} \in \mathbb{R}^{n \times n}$). The kernel and range of a matrix $\mathbf{A}$ are $\ker(\mathbf{A})$ and $\text{range}(\mathbf{A})$. The Euclidean inner product is $\langle \cdot, \cdot \rangle$ and its induced norm is $\|\cdot\|$. We consider a continuous operator $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$. The set of its zeros is $\text{zer}(F)$, which we assume is non-empty, and an element of this set is denoted by $\mathbf{z}^*$. The Jacobian of F at $\mathbf{z}$ is $DF(\mathbf{z})$; its symmetric and skew-symmetric parts are denoted by $\mathbf{S}(\mathbf{z}^*)$ and $\mathbf{A}(\mathbf{z}^*)$, respectively. For a set $\mathcal{S}$, we denote its convex hull by $\text{conv}(\mathcal{S})$, its interior by $\text{int}(\mathcal{S})$, its span by $\text{span}(\mathcal{S})$, and its closure by $\overline{\mathcal{S}}$. The normal cone to a convex set $\mathcal{K}$ at a point $\mathbf{z} \in \mathcal{K}$ is $N_{\mathcal{K}}(\mathbf{z})$. The trajectory of a dynamical system starting at $\mathbf{z}_0$ is denoted by $\mathbf{z}(t)$, its orbit by $\mathcal{O}(\mathbf{z}_0) = \{\mathbf{z}(t) \mid t \geq 0\}$, and its positive limit set by $\omega(\mathbf{z}_0)$.

2.2 Background

This section reviews the standard framework for continuous-time dynamics driven by a monotone operator. We then provide the characterization of their asymptotic behavior via LaSalle's Invariance Principle and an analysis of the invariant set.

Monotone Operators and Associated Flows

We begin with fundamental definitions. Throughout, we consider an operator $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ with domain $\text{dom}(F) = \mathbb{R}^n$.

Definition 2.2.1 (Monotone Operator). A continuous operator $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is monotone if

$$\langle F(\mathbf{u}) - F(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle \geq 0, \quad \forall \mathbf{u}, \mathbf{v} \in \mathbb{R}^n. \quad (2.1)$$

If F is continuously differentiable, this condition is equivalent to its Jacobian matrix, $DF(\mathbf{z})$, having a positive semidefinite symmetric part for all $\mathbf{z} \in \mathbb{R}^n$.

Our central problem is to find a zero of a monotone operator, i.e., a point $\mathbf{z}^* \in \mathbb{R}^n$ such that $F(\mathbf{z}^*) = \mathbf{0}$. We denote the set of such points by $\text{zer}(F)$ and assume it is non-empty. To find a point in this set, we study the trajectories of the associated *monotone flow*, defined by the ordinary differential equation (ODE):

$$\dot{\mathbf{z}}(t) = -F(\mathbf{z}(t)), \quad \mathbf{z}(0) = \mathbf{z}_0. \quad (2.2)$$

This system generalizes the concept of a gradient flow to operators that may not be the gradient of any potential function. It is known that this system has an ergodic convergence property that $\frac{1}{T} \int_0^T \mathbf{z}(t) dt \rightarrow \mathbf{z}^*$ for some $\mathbf{z}^* \in \text{zer}(F)$ (first shown in [5]).

Lyapunov Stability and Convergence of Distance

A foundational property of the flow (2.2) is its stability with respect to any solution $\mathbf{z}^* \in \text{zer}(F)$. Consider the squared Euclidean distance to $\mathbf{z}^*$ as a Lyapunov function candidate:

$$V(\mathbf{z}) := \frac{1}{2} \|\mathbf{z} - \mathbf{z}^*\|^2. \quad (2.3)$$

The time derivative of $V(\mathbf{z}(t))$ along the trajectories of the flow is non-positive:

$$\dot{V}(\mathbf{z}(t)) = \langle \mathbf{z}(t) - \mathbf{z}^*, \dot{\mathbf{z}}(t) \rangle = -\langle \mathbf{z}(t) - \mathbf{z}^*, F(\mathbf{z}(t)) - F(\mathbf{z}^*) \rangle \leq 0, \quad (2.4)$$

where the final inequality follows from the monotonicity of F .

Since $V(\mathbf{z}(t))$ is non-increasing and bounded below by zero, it must converge to a finite limit. This implies that the distance $\|\mathbf{z}(t) - \mathbf{z}^*\|$ converges. While this guarantees the trajectory is bounded and stable, it does not, by itself, imply that the trajectory $\mathbf{z}(t)$ converges to a point.

Structure of the Limit Set

To characterize the asymptotic behavior of $\mathbf{z}(t)$, we apply LaSalle's Invariance Principle (see, e.g., [67]). This principle states that for a bounded trajectory, its positive limit set, $\omega(\mathbf{z}_0)$, is non-empty, compact, and invariant. Furthermore, $\omega(\mathbf{z}_0)$ is contained in the largest invariant subset of $\mathcal{E} = \{\mathbf{z} \in \mathbb{R}^n \mid \dot{V}(\mathbf{z}) = 0\}$. A set $\mathcal{S}$ is an *invariant set* for the dynamics if any trajectory starting in $\mathcal{S}$ remains in $\mathcal{S}$ for all time.

In our context, the boundedness of $\mathbf{z}(t)$ is guaranteed. The set where the Lyapunov function's derivative vanishes is

$$\mathcal{E}_{\mathbf{z}^*} = \{\mathbf{z} \in \mathbb{R}^n \mid \langle \mathbf{z} - \mathbf{z}^*, F(\mathbf{z}) \rangle = 0\}. \quad (2.5)$$

LaSalle's principle thus asserts that $\omega(\mathbf{z}_0)$ is contained within $\mathcal{M}_{\mathbf{z}^*}$, the largest invariant subset of $\mathcal{E}_{\mathbf{z}^*}$. This constraint imposes a powerful structure on the operator F within this set, which was shown for differentiable convex-concave saddle point dynamics in [55]. Here, we establish a more general result for continuously differentiable monotone operators with a more direct analysis.

Proposition 2.2.2 (Operator Structure on the Invariant Set). *Let F be a continuously differentiable monotone operator. For any $\mathbf{z}^* \in \text{zer}(F)$, let $\mathcal{M}_{\mathbf{z}^*}$ be the largest invariant subset of $\mathcal{E}_{\mathbf{z}^*}$. For any point $\mathbf{z} \in \mathcal{M}_{\mathbf{z}^*}$, the operator satisfies*

$$F(\mathbf{z}) = DF(\mathbf{z}^*)(\mathbf{z} - \mathbf{z}^*) = \mathbf{A}(\mathbf{z}^*)(\mathbf{z} - \mathbf{z}^*), \quad (2.6)$$

where $\mathbf{A}(\mathbf{z}^*) := \frac{1}{2} (DF(\mathbf{z}^*) - DF(\mathbf{z}^*)^\top)$ denotes the skew-symmetric part of $DF(\mathbf{z}^*)$.

Proof. Let $\mathbf{z}(t)$ be a trajectory starting at an arbitrary $\mathbf{z}_0 \in M_{\mathbf{z}^*}$. By invariance, $\mathbf{z}(t) \in M_{\mathbf{z}^*}$ for all $t \geq 0$. This implies two key properties: (i) $\|\mathbf{z}(t) - \mathbf{z}^*\|$ is constant, and (ii) $\langle \mathbf{z}(t) - \mathbf{z}^*, F(\mathbf{z}(t)) \rangle = 0$ for all $t \geq 0$.

Now, consider a second trajectory $\mathbf{w}(t)$ starting from $\mathbf{w}_0 = \lambda \mathbf{z}_0 + (1 - \lambda) \mathbf{z}^*$ for some $\lambda \in (0, 1)$. Due to monotonicity, the distance between any two trajectories is non-increasing: $\frac{d}{dt} \frac{1}{2} \|\mathbf{z}(t) - \mathbf{w}(t)\|^2 \leq 0$. The triangle inequality gives $\|\mathbf{z}(t) - \mathbf{z}^*\| \leq \|\mathbf{z}(t) - \mathbf{w}(t)\| + \|\mathbf{w}(t) - \mathbf{z}^*\|$. Since $\|\mathbf{z}(t) - \mathbf{z}^*\|$ is constant while the other two terms are non-increasing, all three distances must be constant for all $t \geq 0$.

Equality in the triangle inequality implies that $\mathbf{w}(t)$ must lie on the line segment between $\mathbf{z}^*$ and $\mathbf{z}(t)$. Thus, we can write $\mathbf{w}(t) = \lambda(t) \mathbf{z}(t) + (1 - \lambda(t)) \mathbf{z}^*$. Since $\|\mathbf{w}(t) - \mathbf{z}^*\|$ and $\|\mathbf{z}(t) - \mathbf{z}^*\|$ are both constant, the coefficient $\lambda(t)$ must also be constant, so $\lambda(t) = \lambda$.

Differentiating $\mathbf{w}(t) = \lambda \mathbf{z}(t) + (1 - \lambda) \mathbf{z}^*$ with respect to time gives $\dot{\mathbf{w}}(t) = \lambda \dot{\mathbf{z}}(t)$. From the flow definition (2.2), this means $-F(\mathbf{w}(t)) = \lambda(-F(\mathbf{z}(t)))$, which simplifies to $F(\lambda \mathbf{z}(t) + (1 - \lambda) \mathbf{z}^*) = \lambda F(\mathbf{z}(t))$. Since $\mathbf{z}_0 \in M_{\mathbf{z}^*}$ was arbitrary, this relation holds for any point in $M_{\mathbf{z}^*}$. Let $\mathbf{z}$ be such a point. Differentiating $F(\lambda \mathbf{z} + (1 - \lambda) \mathbf{z}^*) = \lambda F(\mathbf{z})$ with respect to λ at $\lambda = 0$ yields the desired result: $DF(\mathbf{z}^*)(\mathbf{z} - \mathbf{z}^*) = F(\mathbf{z})$.

Finally, substituting this back into the orthogonality condition for any $\mathbf{z} \in M_{\mathbf{z}^*}$ gives: $0 = \langle \mathbf{z} - \mathbf{z}^*, F(\mathbf{z}) \rangle = \langle \mathbf{z} - \mathbf{z}^*, DF(\mathbf{z}^*)(\mathbf{z} - \mathbf{z}^*) \rangle$. Since, the symmetric part of $DF(\mathbf{z}^*)$ is PSD, we conclude that $F(\mathbf{z}) = DF(\mathbf{z}^*)(\mathbf{z} - \mathbf{z}^*) = \mathbf{A}(\mathbf{z}^*)(\mathbf{z} - \mathbf{z}^*)$. $\square$

Remark 2.2.3. This characterization reveals a sufficient condition for convergence. If the Jacobian $DF(\mathbf{z}^*)$ is symmetric (i.e., its skew-symmetric part $\mathbf{A}(\mathbf{z}^*) = \mathbf{0}$), then for any point $\mathbf{z} \in M_{\mathbf{z}^*}$, we must have $F(\mathbf{z}) = \mathbf{0}$. This forces the invariant set $M_{\mathbf{z}^*}$ to be a subset of $\text{zer}(F)$.

Remark 2.2.4. Note that the ergodic limit of the trajectory is some $\mathbf{z}^* \in \text{zer}(F)$. Let $DF(\mathbf{z}^*) = \mathbf{S}(\mathbf{z}^*) + \mathbf{A}(\mathbf{z}^*)$ be the decomposition of the Jacobian into its symmetric (PSD) and skew-symmetric part, Proposition 2.2.2 implies that the limit set lies in the affine subspace $\mathbf{z}^* + \text{range}(\mathbf{A}(\mathbf{z}^*))$. Furthermore, since $\langle F(\mathbf{z}), \mathbf{z} - \mathbf{z}^* \rangle = 0$, we have $\mathbf{z} - \mathbf{z}^* \in \ker \mathbf{S}(\mathbf{z}^*)$. Hence, we have that the trajectory must lie in the affine subspace $\mathbf{z}^* + \text{range}(\mathbf{A}(\mathbf{z}^*)) \cap \ker \mathbf{S}(\mathbf{z}^*)$.¹

In summary, any trajectory of a monotone flow either converges to a zero or its limit set is contained in a (non-degenerate) invariant set $M_{\mathbf{z}^*}$, on which the dynamics are governed by a skew-symmetric linear system. This explains the apparent dichotomy that the flow is known to have last-iterate convergence but has ergodic convergence, since the time average of the limiting skew-symmetric linear dynamical system converges to a $\mathbf{z}^* \in \text{zer}(F)$.

2.3 Operator Structure on Convex Hulls

The previous section established that non-convergent dynamics are governed by a skew-symmetric linear system on the invariant set. This section presents a key theorem demon-

¹As we shall see later, the trajectory of the skew-symmetric linear system is a superposition of circular motions in orthogonal 2D planes. Hence, this affine space must allow such motions and otherwise is trivial, $\{\mathbf{z}^*\}$.

strating how this linear structure propagates from a finite set of points to their convex hull.

Theorem 2.3.1. *Let $\mathcal{K}$ be a closed convex subset of a real Hilbert space $\mathcal{H}$, and assume $\mathcal{K}$ has a non-empty interior. Let $F : \mathcal{K} \rightarrow \mathcal{H}$ be a continuous, single-valued, monotone operator, and let $A : \mathcal{H} \rightarrow \mathcal{H}$ be a bounded, skew-adjoint linear operator. Suppose there exists a finite set of points $\{\mathbf{z}_1, \dots, \mathbf{z}_m\} \subseteq \mathcal{K}$ such that $F(\mathbf{z}_i) = A\mathbf{z}_i$ for all $i = 1, \dots, m$.*

Then for any point $\mathbf{z} \in \text{conv}\{\mathbf{z}_1, \dots, \mathbf{z}_m\}$, the deviation of $F(\mathbf{z})$ from $A\mathbf{z}$ is an element of the normal cone to $\mathcal{K}$ at $\mathbf{z}$. That is,

$$A\mathbf{z} - F(\mathbf{z}) \in N_{\mathcal{K}}(\mathbf{z}), \quad (2.7)$$

where $N_{\mathcal{K}}(\mathbf{z}) := \{\mathbf{w} \in \mathcal{H} \mid \langle \mathbf{w}, \mathbf{y} - \mathbf{z} \rangle \leq 0 \text{ for all } \mathbf{y} \in \mathcal{K}\}$. In particular, for any point $\mathbf{z} \in \text{conv}\{\mathbf{z}_1, \dots, \mathbf{z}_m\} \cap \text{int}(\mathcal{K})$, we have $F(\mathbf{z}) = A\mathbf{z}$.

Proof. Let the operator $G : \mathcal{K} \rightarrow \mathcal{H}$ be defined as $G(\mathbf{x}) = F(\mathbf{x}) - A\mathbf{x}$. Since F is monotone and A is skew-adjoint (which implies the operator $\mathbf{x} \mapsto -A\mathbf{x}$ is monotone), their sum G is also a continuous monotone operator. By hypothesis, $G(\mathbf{z}_i) = \mathbf{0}$ for each $i \in \{1, \dots, m\}$.

The monotonicity of G means that for any $\mathbf{y} \in \mathcal{K}$ and any $\mathbf{z}_i$, we have $\langle G(\mathbf{y}) - G(\mathbf{z}_i), \mathbf{y} - \mathbf{z}_i \rangle \geq 0$. As $G(\mathbf{z}_i) = \mathbf{0}$, this simplifies to $\langle G(\mathbf{y}), \mathbf{y} - \mathbf{z}_i \rangle \geq 0$.

Now, let $\mathbf{z} \in \text{conv}\{\mathbf{z}_1, \dots, \mathbf{z}_m\}$. Then $\mathbf{z} = \sum_{i=1}^m \alpha_i \mathbf{z}_i$ for some coefficients $\alpha_i \geq 0$ with $\sum_{i=1}^m \alpha_i = 1$. We take a convex combination of the m inequalities above:

$$\sum_{i=1}^m \alpha_i \langle G(\mathbf{y}), \mathbf{y} - \mathbf{z}_i \rangle = \langle G(\mathbf{y}), \mathbf{y} - \sum_{i=1}^m \alpha_i \mathbf{z}_i \rangle = \langle G(\mathbf{y}), \mathbf{y} - \mathbf{z} \rangle \geq 0.$$

This inequality, $\langle G(\mathbf{y}), \mathbf{y} - \mathbf{z} \rangle \geq 0$, holds for all $\mathbf{y} \in \mathcal{K}$. This is a standard variational inequality. To infer properties of $G(\mathbf{z})$, we use a classic technique (the Minty trick [97]). For any fixed $\mathbf{y} \in \mathcal{K}$, let $\mathbf{y}_t = \mathbf{z} + t(\mathbf{y} - \mathbf{z})$ for $t \in (0, 1]$. Since $\mathcal{K}$ is convex, $\mathbf{y}_t \in \mathcal{K}$. Substituting $\mathbf{y}_t$ for $\mathbf{y}$ in the inequality gives:

$$\langle G(\mathbf{y}_t), \mathbf{y}_t - \mathbf{z} \rangle = \langle G(\mathbf{z} + t(\mathbf{y} - \mathbf{z})), t(\mathbf{y} - \mathbf{z}) \rangle \geq 0.$$

For $t > 0$, we can divide by t to get $\langle G(\mathbf{z} + t(\mathbf{y} - \mathbf{z})), \mathbf{y} - \mathbf{z} \rangle \geq 0$. Taking the limit as $t \rightarrow 0^+$ and invoking the continuity of G yields the Minty Variational Inequality [97]:

$$\langle G(\mathbf{z}), \mathbf{y} - \mathbf{z} \rangle \geq 0 \quad \text{for all } \mathbf{y} \in \mathcal{K}. \quad (2.8)$$

Rearranging this as $\langle -G(\mathbf{z}), \mathbf{y} - \mathbf{z} \rangle \leq 0$ and comparing it with the definition of the normal cone $N_{\mathcal{K}}(\mathbf{z})$ immediately shows that $-G(\mathbf{z}) \in N_{\mathcal{K}}(\mathbf{z})$. Substituting back $G(\mathbf{z}) = F(\mathbf{z}) - A\mathbf{z}$ gives the main result:

$$A\mathbf{z} - F(\mathbf{z}) \in N_{\mathcal{K}}(\mathbf{z}).$$

For the second claim, if $\mathbf{z} \in \text{int}(\mathcal{K})$, the normal cone is trivial, i.e., $N_{\mathcal{K}}(\mathbf{z}) = \{\mathbf{0}\}$. This forces $A\mathbf{z} - F(\mathbf{z}) = \mathbf{0}$, which completes the proof. $\square$

Remark 2.3.2 (Connection to Maximal Monotonicity). The proof relies on the continuity of F . An alternative approach arises if F is assumed to be *maximally monotone*. In that case, $G(\mathbf{x}) = F(\mathbf{x}) - A\mathbf{x}$ is also maximally monotone.² A key property of maximal monotone operators is that their zero set is closed and convex. Since $\{\mathbf{z}_1, \dots, \mathbf{z}_m\} \subseteq \text{zer}(G)$, it follows that $\text{conv}\{\mathbf{z}_1, \dots, \mathbf{z}_m\} \subseteq \text{zer}(G)$, implying $F(\mathbf{z}) = A\mathbf{z}$ for all $\mathbf{z}$ in the convex hull. Our theorem is more general as it does not require maximal monotonicity, only continuity on a closed convex subset of $\mathcal{H}$. See [42] for a comprehensive background on (maximal) monotone operators.

A related result was shown in [9], where if a monotone operator $F : S \rightarrow \mathbb{R}^n$ on a set $S \subseteq \mathbb{R}^n$ satisfies the orthogonality condition $\langle F(\mathbf{z}), \mathbf{z} - \mathbf{z}^* \rangle = 0$ for all $\mathbf{z} \in S$, and if S contains a ball around $\mathbf{z}^*$, then F must be a skew-symmetric linear operator on S . Our result differs by analyzing how a known skew-linear structure on a finite set propagates to its convex hull.

2.4 The Convex Hull of Orbits in Skew-Symmetric Systems

This section analyzes the geometric and topological properties of trajectories generated by linear skew-symmetric systems. We establish the precise conditions under which the convex hull of such a trajectory is “rich” enough to contain a ball of its center point. We consider dynamics governed by the following (shifted) linear system:

$$\dot{\mathbf{z}}(t) = \mathbf{A}\mathbf{z}(t), \quad \text{with initial condition } \mathbf{z}(0) = \mathbf{z}_0, \quad (2.9)$$

where $\mathbf{A} \in \mathbb{R}^{n \times n}$ is a skew-symmetric matrix ($\mathbf{A}^T = -\mathbf{A}$).

Trajectory Decomposition

We start from some elementary analysis of the trajectory (see also [52]). The skew-symmetry of $\mathbf{A}$ implies that its flow, $e^{\mathbf{A}t}$, is orthogonal, conserving the Euclidean norm $\|\mathbf{z}(t)\| = \|\mathbf{z}_0\|$ for all $t \geq 0$. The geometric structure of the orbit, $\mathcal{O}(\mathbf{z}_0) := \{\mathbf{z}(t) \mid t \geq 0\}$, is determined by the spectral properties of $\mathbf{A}$. The real spectral theorem for skew-symmetric matrices guarantees the existence of an orthogonal matrix $\mathbf{Q}$ that block-diagonalizes $\mathbf{A}$:

$$\mathbf{Q}^T \mathbf{A} \mathbf{Q} = \text{diag} \left(\begin{pmatrix} 0 & -\theta_1 \\ \theta_1 & 0 \end{pmatrix}, \dots, \begin{pmatrix} 0 & -\theta_k \\ \theta_k & 0 \end{pmatrix}, \mathbf{0}_\ell \right),$$

where $\theta_j > 0$ are the non-zero eigen-frequencies and $\ell = \dim \ker(\mathbf{A})$. This decomposition induces an orthogonal splitting of the state space $\mathbb{R}^n = \ker(\mathbf{A}) \oplus \text{range}(\mathbf{A})$, where $\text{range}(\mathbf{A}) = \bigoplus_{j=1}^k V_j$ and each V_j is an invariant 2D plane of rotation corresponding to a frequency θ_j .

²A monotone operator is *maximally monotone* if its graph is not properly contained in the graph of any other (set-valued) monotone operator on $\mathcal{H}$.

The solution $\mathbf{z}(t) = e^{\mathbf{A}t}\mathbf{z}_0$ can be decomposed according to the projection of $\mathbf{z}_0$ onto these subspaces. Let $\mathbf{z}_{\ker} = \text{proj}_{\ker(\mathbf{A})}(\mathbf{z}_0)$ and let $\mathbf{z}_j(t) = e^{\mathbf{A}t} \text{proj}_{V_j}(\mathbf{z}_0)$. The trajectory is a superposition of these components:

$$\mathbf{z}(t) = \mathbf{z}_{\ker} + \sum_{j=1}^k \mathbf{z}_j(t). \quad (2.10)$$

The component in the kernel of $\mathbf{A}$ remains fixed, while each component $\mathbf{z}_j(t)$ rotates in its plane V_j with frequency θ_j . The full trajectory $\{\mathbf{z}(t)\}_{t \geq 0}$ is thus confined to the affine subspace $\mathbf{z}_{\ker} + \text{range}(\mathbf{A})$.

Topological Structure and Convex Hull

The character of the trajectory is determined by the components with non-zero initial projections.

Definition 2.4.1 (Active Frequencies and Active Subspace). The set of active frequencies for an initial condition $\mathbf{z}_0$ is $\Theta_{\text{active}} = \{\theta_j \mid \text{proj}_{V_j}(\mathbf{z}_0) \neq \mathbf{0}\}$. Let $m = |\Theta_{\text{active}}|$ be the number of such frequencies. The active subspace is the direct sum of the corresponding planes, $H := \bigoplus_{j=1}^m V_j \subseteq \text{range}(\mathbf{A})$.

The rotating part of the solution, $\mathbf{z}_{\text{rot}}(t)$, evolves on a set formed by the sum of m orthogonal circular motions. This set is diffeomorphic to an m -dimensional torus ($\mathcal{T}^m$), the Cartesian product of m circles ($S^1 \times \cdots \times S^1$), embedded in the active subspace H . The closure of the trajectory within this torus is governed by the arithmetic relations among the active frequencies. A set of frequencies $\{\theta_j\}_{j=1}^m$ is *rationally independent* if the only integer solution $(c_1, \dots, c_m) \in \mathbb{Z}^m$ to the equation $\sum_{j=1}^m c_j \theta_j = 0$ is the trivial one, $c_j = 0$ for all j . If a non-trivial integer solution exists, the frequencies are *rationally dependent* (a condition known as resonance). A classical result in dynamical systems, the Kronecker-Weyl theorem, dictates that if the frequencies are rationally independent, the trajectory is dense on the entire torus $\mathcal{T}^m$. If they are rationally dependent, the trajectory is confined to a *proper subtorus*, which is a closed subgroup of $\mathcal{T}^m$ of lower dimension. See [52] for an in-depth introduction of these concepts.

We are interested in the convex hull of the orbit, $\text{conv}(\mathcal{O}(\mathbf{z}_0))$. Specifically, we seek the condition under which its center point, $\mathbf{z}_{\ker}$, is an interior point. This is equivalent to determining when the convex hull of the rotating part, $\text{conv}(\{\mathbf{z}_{\text{rot}}(t)\})$, contains an open ball centered at the origin in its ambient space, H .

Theorem 2.4.2 (Structure of the Orbit's Convex Hull). *Let $\mathbf{z}_{\text{rot}}(t)$ be the rotating part of the trajectory, and let d be the number of distinct active frequencies. Let $S_{\text{orb}} := \text{span}(\{\mathbf{z}_{\text{rot}}(t) \mid t \geq 0\})$ be the linear subspace spanned by the trajectory. Then:*

- (a) *The convex hull of the trajectory, $\text{conv}(\{\mathbf{z}_{\text{rot}}(t)\})$, contains an open ball centered at the origin relative to the topology of this subspace. That is, $\mathbf{0} \in \text{int}_{S_{\text{orb}}}(\text{conv}(\{\mathbf{z}_{\text{rot}}(t)\}))$.*

- (b) *The dimension of the subspace spanned by the trajectory is twice the number of distinct active frequencies:* $\dim(S_{\text{orb}}) = 2d$.

Proof. Let the m active frequencies be $\{\theta_1, \dots, \theta_m\}$, and let the d distinct frequencies among them be $\{\lambda_1, \dots, \lambda_d\}$. The rotating part of the trajectory is

$$\mathbf{z}_{\text{rot}}(t) = \sum_{j=1}^m r_j (\cos(\theta_j t) \mathbf{u}_j + \sin(\theta_j t) \mathbf{w}_j),$$

where $\{\mathbf{u}_j, \mathbf{w}_j\}_{j=1}^m$ are orthonormal vectors forming a basis for the respective planes of rotation V_j , and $r_j > 0$.

Part (a): Interior of the Convex Hull. We prove that $\mathbf{0}$ is an interior point of $\text{conv}(\{\mathbf{z}_{\text{rot}}(t)\})$ with respect to the subspace S_{orb} . Suppose for contradiction that it is not. By the Supporting Hyperplane Theorem applied within S_{orb} , there exists a non-zero vector $\mathbf{v} \in S_{\text{orb}}$ such that $\langle \mathbf{v}, \mathbf{x} \rangle \geq 0$ for all $\mathbf{x} \in \text{conv}(\{\mathbf{z}_{\text{rot}}(t)\})$. This implies $\langle \mathbf{v}, \mathbf{z}_{\text{rot}}(t) \rangle \geq 0$ for all $t \geq 0$.

Let $g(t) = \langle \mathbf{v}, \mathbf{z}_{\text{rot}}(t) \rangle$. The function $g(t)$ is a finite sum of sinusoids and is therefore almost periodic (see [8]). Since $\theta_j \neq 0$, the time-average of each sinusoidal component is zero, making the mean of $g(t)$ zero, i.e., $M(g) := \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T g(t) dt = 0$. The function $g(t)$ is non-negative with a zero mean, so by the forthcoming Lemma 2.4.4, it must be identically zero: $g(t) = \langle \mathbf{v}, \mathbf{z}_{\text{rot}}(t) \rangle = 0$ for all $t \geq 0$.

This means $\mathbf{v}$ is orthogonal to every vector in the generating set of S_{orb} . Since $\mathbf{v} \in S_{\text{orb}}$, it can be written as a finite linear combination of these generators. Therefore, $\mathbf{v}$ must be orthogonal to itself, i.e., $\langle \mathbf{v}, \mathbf{v} \rangle = 0$, which implies $\mathbf{v} = \mathbf{0}$. This contradicts our choice of $\mathbf{v}$ as a non-zero vector.

Part (b): Dimension of the Span. We first regroup the terms in $\mathbf{z}_{\text{rot}}(t)$ according to the d distinct frequencies $\{\lambda_k\}$:

$$\mathbf{z}_{\text{rot}}(t) = \sum_{k=1}^d \left(\cos(\lambda_k t) \left(\sum_{j|\theta_j=\lambda_k} r_j \mathbf{u}_j \right) + \sin(\lambda_k t) \left(\sum_{j|\theta_j=\lambda_k} r_j \mathbf{w}_j \right) \right).$$

Let us define a set of $2d$ vectors: $\tilde{\mathbf{u}}_k := \sum_{j|\theta_j=\lambda_k} r_j \mathbf{u}_j$ and $\tilde{\mathbf{w}}_k := \sum_{j|\theta_j=\lambda_k} r_j \mathbf{w}_j$ for $k = 1, \dots, d$. The trajectory can now be written compactly as

$$\mathbf{z}_{\text{rot}}(t) = \sum_{k=1}^d (\cos(\lambda_k t) \tilde{\mathbf{u}}_k + \sin(\lambda_k t) \tilde{\mathbf{w}}_k).$$

This expression makes it clear that $S_{\text{orb}} = \text{span}\{\mathbf{z}_{\text{rot}}(t)\} \subset \tilde{H} := \text{span}\{\tilde{\mathbf{u}}_1, \tilde{\mathbf{w}}_1, \dots, \tilde{\mathbf{u}}_d, \tilde{\mathbf{w}}_d\}$.

The set of vectors $\{\tilde{\mathbf{u}}_k, \tilde{\mathbf{w}}_k\}_{k=1}^d$ is orthogonal because the constituent vectors $\{\mathbf{u}_j, \mathbf{w}_j\}$ are orthonormal and the sums are over disjoint index sets for each distinct frequency λ_k .

Furthermore, since all $r_j > 0$, each $\tilde{\mathbf{u}}_k$ and $\tilde{\mathbf{w}}_k$ is a sum of orthogonal, non-zero vectors and is therefore non-zero. An orthogonal set of non-zero vectors is linearly independent, so $\dim(\tilde{H}) = 2d$.

To complete the proof, we show that $S_{\text{orb}} = \tilde{H}$. Let $\mathbf{v} \in \tilde{H}$ be a vector that is orthogonal to S_{orb} , meaning $\langle \mathbf{v}, \mathbf{z}_{\text{rot}}(t) \rangle = 0$ for all $t \geq 0$. Since $\mathbf{v} \in \tilde{H}$, it can be written as $\mathbf{v} = \sum_{k=1}^d (a_k \tilde{\mathbf{u}}_k + b_k \tilde{\mathbf{w}}_k)$. The condition $\langle \mathbf{v}, \mathbf{z}_{\text{rot}}(t) \rangle = 0$ becomes:

$$\left\langle \sum_{i=1}^d (a_i \tilde{\mathbf{u}}_i + b_i \tilde{\mathbf{w}}_i), \sum_{k=1}^d (\cos(\lambda_k t) \tilde{\mathbf{u}}_k + \sin(\lambda_k t) \tilde{\mathbf{w}}_k) \right\rangle = 0.$$

By the orthogonality of $\{\tilde{\mathbf{u}}_k, \tilde{\mathbf{w}}_k\}$, this simplifies to:

$$\sum_{k=1}^d (a_k \|\tilde{\mathbf{u}}_k\|^2 \cos(\lambda_k t) + b_k \|\tilde{\mathbf{w}}_k\|^2 \sin(\lambda_k t)) = 0.$$

Since the frequencies $\{\lambda_k\}$ are distinct, the functions $\{\cos(\lambda_k t), \sin(\lambda_k t)\}$ are linearly independent. This forces all coefficients to be zero. As $\|\tilde{\mathbf{u}}_k\|^2 > 0$ and $\|\tilde{\mathbf{w}}_k\|^2 > 0$, we must have $a_k = 0$ and $b_k = 0$ for all k . This implies $\mathbf{v} = \mathbf{0}$. Thus, the only vector in $\tilde{H}$ orthogonal to S_{orb} is the zero vector. So $S_{\text{orb}} = \tilde{H}$ and $\dim(S_{\text{orb}}) = 2d$. $\square$

The proof above relies on the notion of almost periodic functions and a fundamental property of them, first defined and established in [8].

Definition 2.4.3 (Bohr Almost Periodic Function). A continuous function $g : \mathbb{R} \rightarrow \mathbb{R}$ is almost periodic if for every $\varepsilon > 0$, the set of ε -almost periods $\{\tau \in \mathbb{R} \mid \sup_{t \in \mathbb{R}} |g(t+\tau) - g(t)| < \varepsilon\}$ is relatively dense in $\mathbb{R}$.

The following result is a consequence of Bohr Uniqueness Theorem for the mean function of an almost periodic function.

Lemma 2.4.4 (Uniqueness Lemma). Let $g : \mathbb{R} \rightarrow \mathbb{R}$ be a continuous, almost periodic function. If $g(t) \geq 0$ for all $t \geq 0$ and its mean value $M(g) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T g(t) dt = 0$, then $g(t) = 0$ for all $t \geq 0$.

Distinctness versus Rational Independence It is worth distinguishing *distinct frequencies* in Theorem 2.4.2 from the stronger condition of *rational independence*. The former is an algebraic condition ensuring the hull is full-dimensional, while the latter is an arithmetic condition determining its specific shape and volume.

The rotating part of the trajectory, $\mathbf{z}_{\text{rot}}(t)$, is a sum of m individual circular motions, $\mathbf{z}_j(t)$, in orthogonal 2D subspaces V_j . The convex hull closure of each individual motion,

$\overline{\text{conv}(\{\mathbf{z}_j(t)\}_{t \geq 0})}$, is the closed disk $D_j \subseteq V_j$ that it traces. The convex hull of the total trajectory is contained within the Minkowski sum of these individual disks:

$$\text{conv}(\{\mathbf{y}(t)\}) \subseteq \sum_{j=1}^m \overline{\text{conv}(\{\mathbf{y}_j(t)\})} = \sum_{j=1}^m D_j.$$

The set on the right, $\sum D_j$, is a “filled torus” and represents the largest possible convex hull the trajectory can have. This set always contains an open ball centered at the origin with radius $\min_{j \in [m]} r_j$.

The role of rational independence is to determine whether the trajectory’s convex hull achieves this maximal volume. If the distinct frequencies $\{\theta_j\}$ are also *rationally independent*, the trajectory is dense on the boundary of this filled torus. As a result, its convex hull is the entire filled torus itself, $\overline{\text{conv}(\{\mathbf{y}(t)\})} = \sum D_j$. Conversely, if the frequencies are *rationally dependent* (resonant), the trajectory is restricted to a lower-dimensional subtorus. Its convex hull is then a proper subset of the filled torus.

2.5 Synthesis and Concluding Remarks

We conclude by synthesizing the results from the preceding sections to provide a cohesive picture of the behavior of monotone flows. This synthesis reveals a fundamental dichotomy: trajectories either converge to a solution or enter a region where the dynamics are locally equivalent to a conservative, skew-symmetric linear system—the sole barrier to convergence.

The Geometric Structure of Non-Convergent Trajectories

Our analysis culminates in a detailed geometric characterization of any trajectory of the monotone flow (2.2) that fails to converge to a point in $\text{zer}(F)$. Let $\{\mathbf{z}(t)\}$ be such a trajectory.

First, as established in Section 2.2, its limit set $\omega(\mathbf{z}_0)$ is a non-empty, compact, invariant set where the dynamics become linear and conservative. For any point $\mathbf{z}$ on this limit set, the operator satisfies $F(\mathbf{z}) = \mathbf{A}(\mathbf{z} - \mathbf{z}^*)$, where $\mathbf{z}^*$ is the ergodic limit of the trajectory and $\mathbf{A}$ is the skew-symmetric part of the Jacobian $DF(\mathbf{z}^*)$. The results of Section 2.4 apply directly to this emergent linear system. Theorem 2.4.2 shows that this limit-set trajectory, though potentially a thin subset of its affine span, is nevertheless “rich”. Its convex hull, $\text{conv}(\omega(\mathbf{z}_0))$, contains an open ball centered at $\mathbf{z}^*$ (relative to the topology of the trajectory’s span, S_{orb}). Finally, applying Theorem 2.3.1 from Section 2.3, we can conclude that the operator F must behave as the linear map $\mathbf{z} \mapsto \mathbf{A}(\mathbf{z} - \mathbf{z}^*)$ not just on the trajectory itself, but on this entire neighborhood within the convex hull.

This synthesis reveals the fundamental obstruction: convergence fails if and only if the trajectory is drawn into a region of the state space where the operator F becomes locally equivalent to a conservative, skew-symmetric linear system, permitting persistent, non-dissipative orbits.

Convergent Algorithms

This understanding clarifies that the design of convergent algorithms is fundamentally about breaking or avoiding these emergent skew-symmetric dynamics. Two prominent strategies achieve this.

Implicit Discretization Methods based on implicit updates, such as the *Proximal Point Method (PPM)* [94, 109], have guaranteed last-iterate convergence. The PPM update, $\mathbf{z}_{k+1} = (\mathbf{I} + \eta_k F)^{-1}(\mathbf{z}_k)$, involves the resolvent operator, which is firmly non-expansive and inherently dissipative [42]. By requiring the operator to be evaluated at the next iterate, $\mathbf{z}_{k+1}$, the method prevents the system from getting trapped in the torus as in the continuous flow. Prominent successful algorithms, including the *Extragradient* [69] and *Optimistic Gradient* [106] methods, can be interpreted as effective approximations to the PPM, thereby inheriting its robustness to oscillatory behavior.

Regularization A second approach is to add an explicit regularization term that breaks the underlying symmetry. For instance, *Tikhonov regularization* [118, 10] modifies the flow to $\dot{\mathbf{z}} = -F(\mathbf{z}) - \beta(t)(\mathbf{z} - \mathbf{z}_0)$, where $\beta(t) > 0$ is a decaying coefficient. The regularization term $-\beta(t)(\mathbf{z} - \mathbf{z}_0)$ introduces a dissipative, anchoring force that pulls the trajectory toward $\mathbf{z}_0$, breaking any conservative dynamics from the skew-symmetric part of F , and ensuring convergence.

Finally, note that these two ideas can be combined, as seen in modern algorithms including the *Extra Anchored Gradient* method [139] and the anchored Popov's scheme [124], to yield a faster convergence rate in terms of $\|F(z)\|$.

Chapter 3

A Unified Framework for Monotone Equation

3.1 Introduction

A fundamental task in optimization and applied mathematics is to find a zero of a monotone operator F , i.e., an $\mathbf{x}^*$ such that $F(\mathbf{x}^*) = \mathbf{0}$, or a fixed point of a non-expansive map T , where $T(\mathbf{x}^*) = \mathbf{x}^*$. For first-order iterative methods that generate a sequence $(\mathbf{x}_k)_{k \geq 0}$, a primary benchmark of efficiency is the non-asymptotic convergence rate of the residual norm— $\|F(\mathbf{x}_k)\|$ or $\|T(\mathbf{x}_k) - \mathbf{x}_k\|$ —which measures how closely the iterate $\mathbf{x}_k$ satisfies the solution condition. While many algorithms are known to converge, achieving optimal rates often requires sophisticated and distinct analytical approaches for different problem classes.

The landscape of algorithms reveals these distinct approaches. For finding fixed points, Halpern’s iteration is a classic method that introduces an “anchoring” term to ensure convergence to a particular solution (the one closest to the anchor) [51]. While its asymptotic convergence properties are well-understood [15, 129], its non-asymptotic rate analysis for the residual norm is a more recent development [76, 68, 115, 77]. For finding zeros of monotone operators, the challenge is more pronounced. The most basic algorithm, the Forward Step method ($\mathbf{x}_{k+1} = \mathbf{x}_k - \eta F(\mathbf{x}_k)$), without additional assumptions such as the cocoercivity or strong monotonicity, is known to diverge for merely monotone operators. Convergent methods like the Extragradient (EG) [69] and Optimistic Gradient (OG) [106] methods guarantee a suboptimal $O(1/\sqrt{k})$ rate for the residual norm $\|F(\mathbf{x}_k)\|$ [49, 48]. Attaining the optimal $O(1/k)$ rate has motivated the development of specialized algorithms that modify the update steps in intricate ways to achieve acceleration [33, 138, 74, 124].

In this work, we show that these varied goals—analyzing classic methods, stabilizing divergent ones, and accelerating convergent ones—can be achieved through a single, unified framework grounded in Tikhonov regularization. The core idea is conceptually simple: instead of designing new algorithmic updates, we analyze the performance of standard, well-

known iterative methods when applied to a sequence of regularized operators of the form

$$T_k(\cdot) = (1 - \delta_k)T(\cdot) + \delta_k x_0, \quad F_k(\cdot) := F(\cdot) + \delta_k(\cdot - \mathbf{x}_0),$$

for a nonexpansive mapping T and a monotone operator F respectively, where $\mathbf{x}_0$ is a fixed anchor point and $\delta_k \rightarrow 0$ is a diminishing regularization parameter. This operator-centric, continuation-style approach provides a coherent analytical template that yields new insights and high-performance algorithms.

Our framework delivers the following concrete contributions:

- **A new, intuitive analysis of Halpern Iteration.** We focus on interpreting Halpern’s iteration as applying a single Picard step to a sequence of regularized, contractive maps. While this perspective has historical roots [15], our work provides a new, direct analysis based on this view that recovers the optimal $O(1/k)$ convergence rate for the fixed-point residual $\|T(\mathbf{x}_k) - \mathbf{x}_k\|$ in a general normed space [115, 25].
- **Stabilization of the forward step method.** We apply our framework to the Forward Step method (gradient descent in convex optimization) for monotone operators. The regularization transforms the operator at each step into a strongly monotone one, which is sufficient to ensure convergence. This yields a new algorithm, the Regularized Forward Step method, and the first discrete-time proof that such a scheme converges, achieving an $O(1/\sqrt{k})$ rate for $\|F(\mathbf{x}_k)\|$. This forward method matches the convergence rate of methods including the Proximal Point method and its variants including the Extragradient and Optimistic Gradient methods [49, 48].
- **A unified path to acceleration.** We demonstrate that our framework provides a simple and natural path to acceleration. By applying the standard Extragradient and Optimistic Gradient methods to the sequence of regularized operators $\{F_k\}$, we derive new algorithms (Reg-EG and Reg-OG) that achieve the optimal $O(1/k)$ rate for $\|F(\mathbf{x}_k)\|$. These methods are distinct from existing accelerated algorithms [138, 74, 124] as the acceleration arises from regularizing the operator rather than modifying the algorithm’s structure.

Organization The remainder of this paper is organized as follows. Section 2 establishes our notation, reviews the necessary background on the problems and algorithms we study, and situates our contributions within the context of related literature. In Section 3, we introduce our unified Tikhonov regularization framework, first presenting the core operator-centric principle and then providing a motivating example by reinterpreting the classical Halpern iteration through this lens. Section 4 presents our main results, demonstrating the framework’s power by applying it to stabilize the Forward Step method and to accelerate both the Extragradient and Optimistic Gradient methods to the optimal rate. Finally, we conclude in Section 5 with a summary of our findings and a discussion of their implications

3.2 Background and Preliminaries

Notation and Preliminaries

We consider problems set in either a real Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ with its induced norm, or a general normed space $(\mathcal{X}, \|\cdot\|)$. Our analysis concerns two primary objects: a monotone operator $F : \mathcal{H} \rightarrow \mathcal{H}$, which we typically assume to be L -Lipschitz, and a non-expansive mapping $T : \mathcal{X} \rightarrow \mathcal{X}$. The goal of an iterative method is to generate a sequence of iterates $(\mathbf{x}_k)_{k \geq 0}$ that finds a solution $\mathbf{x}^*$ in the solution set—either the set of zeros, $\text{Zer}(F)$, or the set of fixed points, $\text{Fix}(T)$. Our regularization framework is defined by an anchor point $\mathbf{x}_0$ and a diminishing regularization parameter $\delta_k > 0$. The iterative algorithms we study use a step size denoted by η or η_k .

Fixed-Point Problems. In a normed space $\mathcal{X}$, a mapping $T : \mathcal{X} \rightarrow \mathcal{X}$ is *non-expansive* if it does not increase distances, i.e., for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$,

$$\|T(\mathbf{x}) - T(\mathbf{y})\| \leq \|\mathbf{x} - \mathbf{y}\|.$$

The objective is to find a *fixed point* $\mathbf{x}^* \in \mathcal{X}$ such that $T(\mathbf{x}^*) = \mathbf{x}^*$. We denote the set of all such points by $\text{Fix}(T)$. A mapping is a *contraction* if there exists a factor $\gamma \in [0, 1)$ such that for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, $\|T(\mathbf{x}) - T(\mathbf{y})\| \leq \gamma \|\mathbf{x} - \mathbf{y}\|$.

Monotone Equations. In a Hilbert space $\mathcal{H}$, an operator $F : \mathcal{H} \rightarrow \mathcal{H}$ is *monotone* if for all $\mathbf{x}, \mathbf{y} \in \mathcal{H}$,

$$\langle F(\mathbf{x}) - F(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq 0.$$

The operator is *μ -strongly monotone* for some $\mu > 0$ if it satisfies the stricter condition

$$\langle F(\mathbf{x}) - F(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \mu \|\mathbf{x} - \mathbf{y}\|^2.$$

The goal is to find a *zero* of the operator, i.e., an $\mathbf{x}^* \in \mathcal{H}$ such that $F(\mathbf{x}^*) = \mathbf{0}$. We denote the solution set by $\text{Zer}(F)$. In our analysis, we will typically assume F is also *L -Lipschitz*, satisfying $\|F(\mathbf{x}) - F(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\|$ for some constant $L > 0$. We make the standing assumption that the solution sets $\text{Fix}(T)$ and $\text{Zer}(F)$ are non-empty for the problems under consideration.

A canonical example of a monotone operator arises from finding a saddle point of a convex-concave function $\mathcal{L}(\mathbf{u}, \mathbf{v})$. The operator $F(\mathbf{u}, \mathbf{v}) := (\nabla_{\mathbf{u}} \mathcal{L}(\mathbf{u}, \mathbf{v}), -\nabla_{\mathbf{v}} \mathcal{L}(\mathbf{u}, \mathbf{v}))$ is monotone, and finding its zero corresponds to solving the minimax problem.

Connections and Stronger Conditions. A monotone operator F is *β -cocoercive* for $\beta > 0$ if

$$\langle F(\mathbf{x}) - F(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \beta \|F(\mathbf{x}) - F(\mathbf{y})\|^2.$$

Cocoercivity is a stronger condition than monotonicity. These two problem classes are intimately related. Finding a zero of a monotone operator F is equivalent to finding a fixed point of its *resolvent* operator $J_\eta := (I + \eta F)^{-1}$ for any $\eta > 0$. The resolvent is known to be *firmly non-expansive*, meaning $\|J_\eta(\mathbf{x}) - J_\eta(\mathbf{y})\|^2 \leq \langle \mathbf{x} - \mathbf{y}, J_\eta(\mathbf{x}) - J_\eta(\mathbf{y}) \rangle$, which implies it is non-expansive. Furthermore, if F is β -cocoercive, the operator $T := I - \eta F$ is non-expansive for any step size $\eta \in (0, 2\beta]$. This establishes a direct path from a cocoercive monotone inclusion problem to a non-expansive fixed-point problem.

Standard Iterative Methods The methods we analyze are built upon the following foundational algorithms.

Picard Iteration. For a contraction mapping T , the sequence generated by

$$\mathbf{x}_{k+1} = T(\mathbf{x}_k) \quad (3.1)$$

converges to the unique fixed point of T , as guaranteed by the Banach Fixed-Point Theorem.

Halpern Iteration. For a non-expansive map T , the iteration is given by

$$\mathbf{x}_{k+1} = (1 - \delta_k)T(\mathbf{x}_k) + \delta_k \mathbf{x}_0, \quad (3.2)$$

where $\mathbf{x}_0$ is a fixed anchor point and $\delta_k \in (0, 1]$ is a sequence of parameters.

Forward Step (Gradient) Method. To find a zero of an operator F , the update is

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta_k F(\mathbf{x}_k), \quad (3.3)$$

where $\eta_k > 0$ is a step size. This method is not guaranteed to converge for merely monotone operators.

Extragradient (EG) Method. This method uses a two-step update structure:

$$\mathbf{x}_{k+1/2} = \mathbf{x}_k - \eta F(\mathbf{x}_k), \quad \mathbf{x}_{k+1} = \mathbf{x}_k - \eta F(\mathbf{x}_{k+1/2}). \quad (3.4)$$

Optimistic Gradient (OG) Method. The OG method also involves two steps, often written using an auxiliary sequence $(\tilde{\mathbf{x}}_k)$:

$$\tilde{\mathbf{x}}_k = \mathbf{x}_k - \eta F(\tilde{\mathbf{x}}_{k-1}), \quad \mathbf{x}_{k+1} = \mathbf{x}_k - \eta F(\tilde{\mathbf{x}}_k), \quad (3.5)$$

with $\tilde{\mathbf{x}}_{-1} := \mathbf{x}_0$.

A Key Recurrence Lemma

Our convergence analyses for each method will arrive at a similar recurrence relation. We state the following deterministic lemma, which will be used repeatedly to establish our main results.

Lemma 3.2.1. *Let $c, c_1, R, \alpha, \beta, \gamma > 0$ with $1 < \gamma\alpha \leq \beta$ and define*

$$\delta_n = \frac{\alpha}{n + \beta}, \quad n \in \mathbb{N}.$$

Assume the sequence $(e_n)_{n \geq 1}$ satisfies $e_1 \leq c_1 R$ and

$$e_{n+1} \leq (1 - \gamma\delta_n) \left(e_n + c(\delta_{n-1} - \delta_n) R \right), \quad n \geq 1. \quad (3.6)$$

Then for every $n \geq 1$,

$$e_n \leq \frac{M}{n + \beta - 1}, \quad \text{where } M := \max\left\{\frac{c\alpha}{\gamma\alpha - 1}, c_1\beta\right\} R.$$

Related Work

Our work builds upon three distinct but interconnected lines of research: the development of fixed-point iterations with strong convergence guarantees, the pursuit of accelerated rates for monotone equations, and the overarching principle of regularization in optimization.

Fixed-Point Iterations: From Weak Convergence to Optimal Rates The study of fixed-point iterations begins with the Banach Fixed-point theorem, which establishes the linear convergence of the Picard iteration, $\mathbf{x}_{k+1} = T(\mathbf{x}_k)$, for contractive mappings [6]. For the broader class of non-expansive operators, foundational algorithms like the Krasnosel'skii-Mann (KM) iteration, $\mathbf{x}_{k+1} = \alpha_k \mathbf{x}_k + (1 - \alpha_k)T(\mathbf{x}_k)$, were developed but could only guarantee weak convergence in a general Hilbert space [71, 91, 15].

To achieve strong convergence, Halpern introduced a pivotal modification: anchoring the iteration to a fixed point, $\mathbf{x}_{k+1} = \alpha_k \mathbf{u} + (1 - \alpha_k)T(\mathbf{x}_k)$ [51]. This concept of regularization was explored concurrently by Browder, who studied regularized monotone inclusions and showed that the solution to the regularized problem converges weakly to a solution of the original problem [15]. Wittmann later provided a streamlined proof of Halpern iteration's strong convergence in Hilbert spaces to the fixed point closest to the anchor [129]. This perspective was generalized by Moudafi to the viscosity approximation scheme, which replaces the constant anchor $\mathbf{u}$ with a contractive mapping $f(\mathbf{x}_k)$ [99], a result later extended to uniformly smooth Banach spaces [133].

While this line of work focused on the nature of convergence, more recent efforts have focused on establishing non-asymptotic convergence rates for the residual norm $\|T(\mathbf{x}_k) - \mathbf{x}_k\|$. Initial results for Halpern iteration showed a rate of $O(1/\log k)$ [76], which was later improved to $O(1/\sqrt{k})$ [68] and finally to the optimal $O(1/k)$ rate in general normed spaces [115], with recent work tightening the constants [25]. In Hilbert spaces, an optimal rate of $O(1/k)$ was established by Lieder [77] and shown to be exact by a matching lower bound [104].

Methods for Monotone Equations: The Quest for Acceleration While conceptually linked to fixed-point problems, finding a zero of a monotone operator has its own distinct line of algorithmic development. The simplest approach, the Forward Step method, $\mathbf{x}_{k+1} = \mathbf{x}_k - \eta F(\mathbf{x}_k)$, is known to diverge for merely monotone operators that are not cocoercive. This failure necessitated more sophisticated designs, such as the Extragradient (EG) method [69] and the Optimistic Gradient (OG) method [106], which can be conceptually viewed as practical variants of the Proximal Point Method (PPM) [95, 109]. These methods are guaranteed to converge, and recent analyses have established a last-iterate convergence rate of $O(1/\sqrt{k})$ for the operator norm [49, 48].

Inspired by the success of anchoring in Halpern-type methods, a significant body of recent work has focused on achieving an accelerated $O(1/k)$ rate for $\|F(\mathbf{x}_k)\|$. A near-optimal $\tilde{O}(1/k)$ rate was first achieved by combining Halpern iteration with inexact proximal updates [33]. This was followed by a series of algorithms that attain the optimal $O(1/k)$ rate by integrating an anchoring or Halpern-style term directly into the algorithmic steps. Notable examples include the Extra Anchored Gradient (EAG) method [138], a variant for weakly-monotone settings [74], and an accelerated anchored OG variant [124]. The acceleration phenomenon in these methods is often illuminated by studying their continuous-time ODE analogs [114, 118, 24, 11].

While these state-of-the-art methods modify the *algorithmic updates* to achieve acceleration, our framework offers a conceptually distinct approach. We show that acceleration can be achieved by regularizing the *operator* itself and then applying standard, un-modified methods like EG and OG to this sequence of regularized operators.

The Broader Principle of Regularization and Continuation The strategy of solving a sequence of subproblems with progressively weaker regularization is a historically significant principle in numerical analysis and optimization, broadly known as the continuation or homotopy method [1]. The core idea is to solve a difficult problem by first solving an easier, more regularized version, and then using its solution to “warm-start” a sequence of slightly less regularized problems until the original problem is recovered.

A canonical example in optimization is the Proximal Point Method, introduced by [95] and analyzed for weak convergence by [109], which iteratively solves a sequence of regularized monotone inclusion subproblems. This conceptual framework has guided the design of practical algorithms, such as the Catalyst acceleration scheme [79]. The general principle is widespread: penalty and barrier methods in constrained optimization gradually increase a parameter to enforce feasibility [101, 130]; homotopy algorithms for sparse regression trace the entire piecewise-linear solution path of the Lasso as the ℓ_1 -penalty is decreased [103, 41]; and graduated optimization in machine learning avoids poor local minima by starting with a heavily smoothed objective and gradually reducing the smoothing [7, 53]. The doubling trick in online learning follows a similar pattern of solving a successive sequence of problems to adapt to the known parameters of an environment [18]. Our work applies this powerful, general principle to develop a unified analysis and new algorithms for fixed-point and

monotone inclusion problems.

3.3 The Regularization Framework

An Operator-Centric View of Regularization

The central principle of our work is a shift in perspective: rather than designing novel, complex algorithms, we analyze the performance of standard, well-known iterative methods when applied to a sequence of regularized operators. This operator-centric view provides a unified analytical framework that can stabilize divergent methods and accelerate convergent ones without altering the underlying structure of the base algorithm.

This approach, a form of continuation, works by solving a sequence of progressively harder subproblems. We begin with a heavily regularized, and thus easier, problem and use its solution to warm-start the next, slightly less regularized problem. For a non-expansive mapping T or a monotone operator F , this corresponds to generating a sequence of iterates by applying a standard method to the operators:

$$T_k(\cdot) := (1 - \delta_k)T(\cdot) + \delta_k \mathbf{x}_0, \quad \text{or} \quad F_k(\cdot) := F(\cdot) + \delta_k(\cdot - \mathbf{x}_0),$$

where $\mathbf{x}_0$ is a fixed anchor point and $\delta_k \rightarrow 0$ is a diminishing regularization parameter. The subsequent sections will demonstrate the power and versatility of this single, coherent framework.

A Motivating Example: A Continuation View of Halpern Iteration

To make our abstract framework concrete, we begin with its simplest application: finding a fixed point for a non-expansive mapping T . We will demonstrate that the classical Halpern iteration can be elegantly reinterpreted through our operator-centric lens. Specifically, we view the algorithm as applying a single Picard iteration to the sequence of contractive mappings $\{T_k\}$ defined in our framework. This perspective yields a new, intuitive analysis that recovers the optimal convergence rate and serves as a blueprint for the more complex problems studied later in this paper.

We begin by recalling the problem setup and notation. Let $\mathcal{X}$ be a normed space and let $T : \mathcal{X} \rightarrow \mathcal{X}$ be a non-expansive mapping, meaning $\|T(\mathbf{x}) - T(\mathbf{y})\| \leq \|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$. We assume the set of fixed points, $\text{Fix}(T)$, is non-empty and let $\mathbf{x}^* \in \text{Fix}(T)$ be one such fixed point.

The regularized operator at step k is the contractive mapping:

$$T_k(\mathbf{x}) := (1 - \delta_k)T(\mathbf{x}) + \delta_k \mathbf{x}_0, \tag{3.7}$$

where the regularization parameter $\delta_k \in (0, 1]$ ensures T_k is a contraction with factor $(1 - \delta_k)$. The Halpern iteration is simply one step of the Picard iteration applied to this operator:

$$\mathbf{x}_{k+1} = T_k(\mathbf{x}_k). \tag{3.8}$$

Our analysis hinges on a simple idea: each step makes progress on the current regularized problem T_k , while the shift from T_{k-1} to T_k introduces a controllable perturbation that diminishes as $k \rightarrow \infty$. We now state our main result on the convergence rate of the residual $\|\mathbf{x}_k - T(\mathbf{x}_k)\|$.

Theorem 3.3.1 (Convergence of Halpern Iteration). *Let T be a non-expansive mapping with $\text{Fix}(T) \neq \emptyset$. Let the sequence $(\mathbf{x}_k)_{k \geq 1}$ be generated by the Halpern iteration (3.8) with from $\mathbf{x}_0$ and $\delta_k = \frac{2}{k+2}$. For any $\mathbf{x}^* \in \text{Fix}(T)$, the residual norm satisfies, for all $k \geq 1$, we have*

$$\|\mathbf{x}_k - T(\mathbf{x}_k)\| \leq \frac{8R}{k+1},$$

where $R := \|\mathbf{x}_0 - \mathbf{x}^*\|$.

Proof. Our proof strategy is to first show that the iterates $(\mathbf{x}_k)$ remain bounded. We then define $E_k := \|\mathbf{x}_k - T_{k-1}(\mathbf{x}_k)\|$ representing the residual norm for the regularized mapping T_{k-1} , and establish a recurrence relation for it. Applying a technical lemma on such recurrences (Lemma 3.2.1 in the Appendix) yields a rate for E_k , which we finally translate into the desired bound on $\|\mathbf{x}_k - T(\mathbf{x}_k)\|$.

Step 1: Boundedness of the Iterates. We first establish that the sequence $(\mathbf{x}_k)$ remains in a ball of radius R around the fixed point $\mathbf{x}^*$. This property is crucial for bounding key terms in the analysis.

Lemma 3.3.2 (Boundedness of the Trajectory). *Let $\mathbf{x}^*$ be a fixed point of T and $R := \|\mathbf{x}_0 - \mathbf{x}^*\|$. For any $\delta \in [0, 1]$ and any $\mathbf{x}$ such that $\|\mathbf{x} - \mathbf{x}^*\| \leq R$, the mapping $T_\delta(\mathbf{x}) := (1 - \delta)T(\mathbf{x}) + \delta\mathbf{x}_0$ satisfies*

$$\|T_\delta(\mathbf{x}) - \mathbf{x}^*\| \leq R.$$

Consequently, if we initialize with $\mathbf{x}_0$, the Halpern iterates satisfy $\|\mathbf{x}_k - \mathbf{x}^\| \leq R$ for all $k \geq 1$.*

Proof. The proof follows directly from the definitions. Using the triangle inequality and the non-expansive property of T :

$$\begin{aligned} \|T_\delta(\mathbf{x}) - \mathbf{x}^*\| &= \|(1 - \delta)(T(\mathbf{x}) - \mathbf{x}^*) + \delta(\mathbf{x}_0 - \mathbf{x}^*)\| \\ &= \|(1 - \delta)(T(\mathbf{x}) - T(\mathbf{x}^*)) + \delta(\mathbf{x}_0 - \mathbf{x}^*)\| \\ &\leq (1 - \delta)\|T(\mathbf{x}) - T(\mathbf{x}^*)\| + \delta\|\mathbf{x}_0 - \mathbf{x}^*\| \\ &\leq (1 - \delta)\|\mathbf{x} - \mathbf{x}^*\| + \delta R. \end{aligned}$$

Since we assume $\|\mathbf{x} - \mathbf{x}^*\| \leq R$, the right-hand side is bounded by $(1 - \delta)R + \delta R = R$. The claim about the iterates $(\mathbf{x}_k)$ follows by a simple induction. $\square$

An immediate consequence of this boundedness is that

$$\|T(\mathbf{x}_k) - \mathbf{x}_0\| \leq \|T(\mathbf{x}_k) - T(\mathbf{x}^*)\| + \|\mathbf{x}^* - \mathbf{x}_0\| \leq 2R \quad (3.9)$$

Step 2: The Error Recurrence. Let's define the error term $E_k := \|\mathbf{x}_k - T_{k-1}(\mathbf{x}_k)\|$ for $k \geq 1$. Since $\mathbf{x}_{k+1} = T_k(\mathbf{x}_k)$ and T_k is a $(1 - \delta_k)$ -contraction, we have

$$E_{k+1} = \|\mathbf{x}_{k+1} - T_k(\mathbf{x}_{k+1})\| \leq (1 - \delta_k) \|\mathbf{x}_k - T_k(\mathbf{x}_k)\|.$$

We can bound the term on the right-hand side by introducing and removing $T_{k-1}(\mathbf{x}_k)$:

$$\begin{aligned} \|\mathbf{x}_k - T_k(\mathbf{x}_k)\| &= \|\mathbf{x}_k - T_{k-1}(\mathbf{x}_k) + T_{k-1}(\mathbf{x}_k) - T_k(\mathbf{x}_k)\| \\ &\leq \|\mathbf{x}_k - T_{k-1}(\mathbf{x}_k)\| + \|T_{k-1}(\mathbf{x}_k) - T_k(\mathbf{x}_k)\| \\ &= E_k + \|((1 - \delta_{k-1})T - (1 - \delta_k)T)(\mathbf{x}_k) + (\delta_{k-1} - \delta_k)\mathbf{x}_0\| \\ &= E_k + |\delta_k - \delta_{k-1}| \|T(\mathbf{x}_k) - \mathbf{x}_0\|. \end{aligned}$$

We require the sequence (δ_k) to be non-increasing, so $|\delta_k - \delta_{k-1}| = \delta_{k-1} - \delta_k$. Using the bound $\|T(\mathbf{x}_k) - \mathbf{x}_0\| \leq 2R$, we arrive at the following recurrence relation for the error:

$$E_{k+1} \leq (1 - \delta_k) (E_k + 2(\delta_{k-1} - \delta_k)R), \quad \text{for } k \geq 1. \quad (3.10)$$

Step 3: Solving the Recurrence. Next, since $x_1 = T_0(\mathbf{x}_0) = \mathbf{x}_0$, for the base case $n = 1$, by the contraction of T_0 and triangle inequality we have

$$\begin{aligned} E_1 &= \|x_1 - T_0(x_1)\| \leq (1 - \delta_0) \|\mathbf{x}_0 - T_0(\mathbf{x}_0)\| \\ &\leq (1 - \delta_0)^2 (\|\mathbf{x}_0 - \mathbf{x}^*\| + \|T(\mathbf{x}_0) - \mathbf{x}^*\|) \leq 2(1 - \delta_0)^2 R. \end{aligned}$$

Then by Lemma 3.2.1, choosing $\delta_k = \frac{2}{k+2}$ we have

$$E_k \leq \frac{4}{k+1} R.$$

Step 4: Bounding the Final Residual. Finally, we relate our error term E_k back to the quantity of interest, $\|\mathbf{x}_k - T(\mathbf{x}_k)\|$.

$$\begin{aligned} \|\mathbf{x}_k - T(\mathbf{x}_k)\| &= \|\mathbf{x}_k - T_{k-1}(\mathbf{x}_k) + T_{k-1}(\mathbf{x}_k) - T(\mathbf{x}_k)\| \\ &\leq \|\mathbf{x}_k - T_{k-1}(\mathbf{x}_k)\| + \|T_{k-1}(\mathbf{x}_k) - T(\mathbf{x}_k)\| \\ &= E_k + \|(1 - \delta_{k-1})T(\mathbf{x}_k) + \delta_{k-1}\mathbf{x}_0 - T(\mathbf{x}_k)\| \\ &= E_k + \delta_{k-1} \|\mathbf{x}_0 - T(\mathbf{x}_k)\|. \end{aligned}$$

Using our bound $E_k \leq \frac{4R}{k+1}$ and the fact that $\|\mathbf{x}_0 - T(\mathbf{x}_k)\| \leq 2R$, we obtain:

$$\|\mathbf{x}_k - T(\mathbf{x}_k)\| \leq \frac{4R}{k+1} + \delta_{k-1}(2R).$$

With $\delta_{k-1} = \frac{2}{(k-1)+2} = \frac{2}{k+1}$, the final bound becomes:

$$\|\mathbf{x}_k - T(\mathbf{x}_k)\| \leq \frac{4R}{k+1} + \frac{2}{k+1}(2R) = \frac{8R}{k+1}.$$

This concludes the proof. □

The above analysis accomodates the flexibility of the regularization schedule. While our specific choice of $\delta_k = \frac{2}{k+2}$ leads to a clean bound, any schedule of the form $\delta_k = O(1/k)$ maintains the optimal $O(1/k)$ convergence rate, albeit with different constants. The technical conditions in our recurrence lemma can be relaxed if one is only concerned with the asymptotic rate, allowing for even greater flexibility in the choice of parameters.

An alternative regularization scheme involves applying the anchoring point inside the operator: $\tilde{T}_\delta(\mathbf{x}) = T((1 - \delta_\delta)\mathbf{x} + \delta_k\mathbf{x}_0)$. Since T is non-expansive, $\tilde{T}_k$ is also a $(1 - \delta_k)$ -contraction. This is commonly called the Tikhonov-Mann iteration, and its equivalence with the classical outer Halpern step is established in [22].

Generalization: An Episodic Warm-Restart Scheme

The regularization viewpoint suggests a natural algorithmic extension. While Halpern iteration represents an extreme case of taking only a single step toward solving each subproblem, a more intuitive implementation of the continuation idea is to perform multiple steps. This allows the iterates to get closer to the fixed point of the current contractive map T_k before the map is changed to T_{k+1} . This leads to an episodic, warm-restarted algorithm.

Formally, at each phase k , we apply the operator T_k for t_k iterations, initialized from the output of the previous phase:

$$\mathbf{x}_{k+1} = (T_k)^{t_k}(\mathbf{x}_k) \equiv \underbrace{T_k \circ \cdots \circ T_k}_{t_k \text{ times}}(\mathbf{x}_k). \quad (3.11)$$

The goal is to choose the step sizes (δ_k) and iteration counts (t_k) to achieve a fast, e.g., geometric, reduction in the residual norm. We establish the following proposition:

Proposition 3.3.3 (Episodic Halpern Scheme with Geometric Rate). *Let T be a non-expansive mapping with $\text{Fix}(T) \neq \emptyset$. Consider the episodic scheme where the sequence $(\mathbf{x}_k)_{k \geq 0}$ is generated by*

$$\mathbf{x}_{k+1} = (T_k)^{t_k}(\mathbf{x}_k), \quad \text{where} \quad T_k(\mathbf{x}) := (1 - \delta_k)T(\mathbf{x}) + \delta_k\mathbf{x}_0.$$

With a geometrically decreasing regularization schedule $\delta_k = \delta_0 \cdot 2^{-k}$ for some $\delta_0 \in (0, 1]$ and a number of inner iterations $t_k = \lceil \ln(4)/\delta_k \rceil$, the residual norm after K phases satisfies

$$\|\mathbf{x}_K - T(\mathbf{x}_K)\| = O(2^{-K}R).$$

Furthermore, in terms of the total number of iterations $N_K = \sum_{k=1}^K t_k$, this corresponds to the optimal rate:

$$\|\mathbf{x}_K - T(\mathbf{x}_K)\| = O\left(\frac{R}{N_K}\right).$$

Proof. The analysis follows the same structure as before. Let $E_k := \|\mathbf{x}_k - T_{k-1}(\mathbf{x}_k)\|$. The error recurrence simply gains an exponent t_k :

$$E_{k+1} \leq (1 - \delta_k)^{t_k} (E_k + 2(\delta_{k-1} - \delta_k)R).$$

Let's choose a geometrically decreasing step size, $\delta_k = 2^{-k}$ for $k \geq 0$. This gives $\delta_{k-1} - \delta_k = 2^{-(k-1)} - 2^{-k} = 2^{-k} = \delta_k$ for $k \geq 1$. To simplify the recurrence, we can choose the inner loop iterations t_k to ensure a constant factor reduction. For instance, setting $t_k = \lceil \ln(4)/\delta_k \rceil$ implies $(1 - \delta_k)^{t_k} \leq e^{-\delta_k t_k} \leq 1/4$. The recurrence becomes:

$$E_{k+1} \leq \frac{1}{4}(E_k + 2\delta_k R).$$

One can show by induction that $E_k \leq O(R\delta_{k-1})$. For the base case $k = 0$, we have $E_1 \leq 2\delta_0 R$ as before. For the inductive step, if $E_k \leq 4R\delta_{k-1}$, then

$$E_{k+1} \leq \frac{1}{4}(4R\delta_k + 2\delta_k R) = \frac{1}{4}(8R\delta_k + 2\delta_k R) = 2.5R\delta_k \leq 4R\delta_k.$$

Hence, the claims holds by induction and this implies the residual norm bound:

$$\|\mathbf{x}_k - T(\mathbf{x}_k)\| \leq E_k + \delta_{k-1}\|\mathbf{x}_0 - T(\mathbf{x}_k)\| \leq 4R\delta_{k-1} + 2R\delta_{k-1} = 6R\delta_{k-1} = O(2^{-k}R).$$

To express this in terms of total iterations $N_K = \sum_{k=1}^K t_k$, we note that $N_K \approx \frac{\ln 4}{\delta_1} \sum_{j=0}^{K-1} 2^j = O(2^K)$. Therefore, the residual norm after K phases is $\|x_K - T(x_K)\| = O(2^{-K}) = O(R/N_K)$, which recovers the optimal rate. $\square$

Generalization: The Viscosity Approximation Method

Our framework can also be extended to the viscosity approximation method, where the anchor term $\delta_k \mathbf{x}_0$ is replaced by a more general contraction mapping $\delta_k G(\mathbf{x})$. Let $G : \mathcal{X} \rightarrow \mathcal{X}$ be a contraction with factor $(1 - \rho)$ for some $\rho \in (0, 1]$.

The regularized operator is now

$$T_k(\mathbf{x}) := (1 - \delta_k)T(\mathbf{x}) + \delta_k G(\mathbf{x}), \quad (3.12)$$

which is a $(1 - \delta_k \rho)$ -contraction. The algorithm remains $\mathbf{x}_{k+1} = T_k(\mathbf{x}_k)$. The analysis proceeds almost identically to the Halpern case, with ρ modifying the contraction factor. Note that this scheme includes the special case of regularizing based on a fixed anchor (whether being $\mathbf{x}_0$ or not), and was shown to converge to the projection of the unique fixed point of G onto $\text{Fix}(T)$. Here we present the formal convergence rate guarantee:

Proposition 3.3.4. *Let T be a non-expansive mapping with $\text{Fix}(T) \neq \emptyset$. Let the sequence $(\mathbf{x}_k)_{k \geq 1}$ be generated by the Halpern iteration (3.12) with from $\mathbf{x}_0$ and $\delta_k = \frac{1 + 1/\rho^2}{k + \rho + 1/\rho}$. For any $\mathbf{x}^* \in \text{Fix}(T)$, the residual norm satisfies, for all $k \geq 1$, we have*

$$\|\mathbf{x}_k - T(\mathbf{x}_k)\| \leq \frac{4(\rho^2 + 1)}{\rho k + \rho^2 - \rho + 1} R.,$$

where $R := \max \left\{ \|\mathbf{x}^* - \mathbf{x}_0\|, \frac{\|G(\mathbf{x}^*) - \mathbf{x}^*\|}{\rho} \right\}$.

Proof. First, we establish a new boundedness radius,

$$R := \max \left\{ \|\mathbf{x}^* - \mathbf{x}_0\|, \frac{\|G(\mathbf{x}^*) - \mathbf{x}^*\|}{\rho} \right\}. \quad (3.13)$$

As before, suppose we have $\|x - \mathbf{x}^*\| \leq R$, then

$$\begin{aligned} \|T_\delta(x) - \mathbf{x}^*\| &\leq (1 - \delta) \|T(x) - T(\mathbf{x}^*)\| + \delta \|G(x) - \mathbf{x}^*\| \\ &\leq (1 - \delta) \|x - \mathbf{x}^*\| + \delta \|G(x) - G(\mathbf{x}^*)\| + \delta \|G(\mathbf{x}^*) - \mathbf{x}^*\| \\ &\leq (1 - \rho\delta)R + \rho\delta R = R. \end{aligned}$$

Hence, through a similar induction argument as before, we have $\|\mathbf{x}_k - \mathbf{x}^*\| \leq R$ for all k .

Next, denote $E_k := \|\mathbf{x}_k - T_{k-1}(\mathbf{x}_k)\|$, by the same contraction argument, we have

$$E_{k+1} \leq (1 - \delta_k \rho) (E_k + \|T_{k-1}(\mathbf{x}_{k-1}) - T_k(\mathbf{x}_{k-1})\|)$$

and

$$\begin{aligned} \|T_{k-1}(\mathbf{x}_{k-1}) - T_k(\mathbf{x}_{k-1})\| &\leq |\delta_{k-1} - \delta_k| \|T(\mathbf{x}_{k-1}) - G(\mathbf{x}_{k-1})\| \\ &\leq |\delta_{k-1} - \delta_k| (\|T(\mathbf{x}_{k-1}) - \mathbf{x}^*\| + \|G(\mathbf{x}_{k-1}) - G(\mathbf{x}^*)\| + \|G(\mathbf{x}^*) - \mathbf{x}^*\|) \\ &\leq |\delta_{k-1} - \delta_k| (R + (1 - \rho)R + \rho R) = 2|\delta_{k-1} - \delta_k|R, \end{aligned}$$

where the last inequality is due to the boundedness of the trajectory above. Moreover, as before, we have

$$E_1 \leq (1 - \rho\delta_0) \|\mathbf{x}_0 - T_0(\mathbf{x}_0)\| \leq (1 - \rho\delta_0) (\|\mathbf{x}_0 - \mathbf{x}^*\| + \|T_0(\mathbf{x}_0) - \mathbf{x}^*\|) \leq 2(1 - \rho\delta_0) R.$$

By choosing $\delta_n = \frac{\alpha}{n+\beta}$ with $\alpha = \frac{\rho^2+1}{\rho^2}$ and $\beta = \frac{\rho^2+1}{\rho}$ and using [Lemma 3.2.1](#) we obtain that

$$E_k \leq \frac{2(\rho^2 + 1)}{\rho k + \rho^2 - \rho + 1} R. \quad (3.14)$$

Hence,

$$\|T(\mathbf{x}_k) - \mathbf{x}_k\| \leq E_k + \delta_{k-1} \|T(\mathbf{x}_k) - G(\mathbf{x}_k)\| \leq E_k + 2\delta_{k-1} R \leq \frac{4(\rho^2 + 1)}{\rho k + \rho^2 - \rho + 1} R.$$

Note that this recovers the Halpern iteration with $G(x) \equiv \mathbf{x}_0$ and $\rho = 1$ shown before. $\square$

This demonstrates that our regularization-based analysis directly extends from the classic Halpern iteration to the more general viscosity approximation method, recovering the optimal convergence rate in each case.

3.4 Stabilizing and Accelerating Standard Methods

Having established our regularization framework and demonstrated its utility on the Halpern iteration, we now apply it to the more general monotone equation problem. This section showcases the full power of the operator-centric approach to achieve two distinct goals. First, we demonstrate its ability to *stabilize* a classically divergent algorithm, the Forward Step method, guaranteeing convergence. Second, we show how the same framework can be used to *accelerate* already-convergent algorithms—the Extragradient and Optimistic Gradient methods—to the optimal $O(1/k)$ rate.

A Regularized Forward Step Method

Our first goal is to address a foundational algorithm known for its instability: the Forward Step method. For a general monotone and L -Lipschitz operator $F : \mathcal{H} \rightarrow \mathcal{H}$, the standard update

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta_k F(\mathbf{x}_k), \quad \eta_k > 0, \quad (3.15)$$

is well-known to fail to converge for operators that are merely monotone and not cocoercive. For instance, if $F(\mathbf{x}) = \mathbf{A}\mathbf{x}$ for a full-rank skew-symmetric matrix $\mathbf{A}$ in an Euclidean space, the norm of the operator, $\|F(\mathbf{x}_k)\|$, can strictly increase at each step for any choice of $\eta_k > 0$.

By applying the Forward Step update not to the original operator F , but to our sequence of regularized operators $\{F_k\}$, we can guarantee convergence. The regularization transforms the operator at each step into a strongly monotone one, which is sufficient to ensure the update is contractive. The resulting algorithm, which we call the Regularized Forward Step method, is defined as:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - F_k(\mathbf{x}_k), \quad \text{where} \quad F_k(\mathbf{x}) := \eta_k F(\mathbf{x}) + \delta_k(\mathbf{x} - \mathbf{x}_0). \quad (3.16)$$

As we will show, coordinating the step size η_k with the regularization strength δ_k yields the first discrete-time proof that such a scheme converges, achieving an $O(1/\sqrt{k})$ rate for the residual norm $\|F(\mathbf{x}_k)\|$.

Theorem 3.4.1 (Convergence of the Anchored Forward Step Method). *Let F be a monotone, L -Lipschitz operator with a solution $\mathbf{x}^*$ such that $F(\mathbf{x}^*) = \mathbf{0}$. Consider the anchored forward step method (3.16) with parameters chosen such that $2\eta_k L = \sqrt{\delta_k}$ and $\delta_k = \frac{4}{k+4}$ for $k \geq 0$. Let $R := \|\mathbf{x}_0 - \mathbf{x}^*\|$. Then the iterates satisfy:*

$$\|F(\mathbf{x}_k)\| = O\left(\frac{RL}{\sqrt{k}}\right).$$

Proof. Our proof proceeds in three main steps. First, we show that the update map $\mathbf{x} \mapsto \mathbf{x} - F_k(\mathbf{x})$ is a contraction. Second, we use this property to establish that the sequence of iterates $(\mathbf{x}_k)$ is bounded. Finally, we derive a recurrence relation for an error term and apply Lemma 3.2.1 to obtain the final convergence rate.

Step 1: The regularized update is a contraction. We begin by analyzing the properties of the mapping $T_k(\mathbf{x}) := \mathbf{x} - F_k(\mathbf{x})$. For any $\mathbf{x}, \mathbf{y} \in \mathcal{H}$, we have

$$\begin{aligned} T_k(\mathbf{x}) - T_k(\mathbf{y}) &= (\mathbf{x} - \mathbf{y}) - (F_k(\mathbf{x}) - F_k(\mathbf{y})) \\ &= (\mathbf{x} - \mathbf{y}) - \eta_k(F(\mathbf{x}) - F(\mathbf{y})) - \delta_k(\mathbf{x} - \mathbf{y}) \\ &= (1 - \delta_k)(\mathbf{x} - \mathbf{y}) - \eta_k(F(\mathbf{x}) - F(\mathbf{y})). \end{aligned}$$

Taking the squared norm and using the monotonicity of F (i.e., $\langle \mathbf{x} - \mathbf{y}, F(\mathbf{x}) - F(\mathbf{y}) \rangle \geq 0$), we get:

$$\begin{aligned} \|T_k(\mathbf{x}) - T_k(\mathbf{y})\|^2 &= (1 - \delta_k)^2 \|\mathbf{x} - \mathbf{y}\|^2 - 2\eta_k(1 - \delta_k) \langle \mathbf{x} - \mathbf{y}, F(\mathbf{x}) - F(\mathbf{y}) \rangle + \eta_k^2 \|F(\mathbf{x}) - F(\mathbf{y})\|^2 \\ &\leq (1 - \delta_k)^2 \|\mathbf{x} - \mathbf{y}\|^2 + (\eta_k L)^2 \|\mathbf{x} - \mathbf{y}\|^2 \\ &= ((1 - \delta_k)^2 + \eta_k^2 L^2) \|\mathbf{x} - \mathbf{y}\|^2. \end{aligned}$$

By our choice of parameters $2\eta_k L = \sqrt{\delta_k}$, and for $\delta_k \in [0, 1]$, the contraction factor is bounded by:

$$(1 - \delta_k)^2 + \eta_k^2 L^2 = (1 - \delta_k)^2 + \frac{\delta_k}{4} = 1 - 2\delta_k + \delta_k^2 + \frac{\delta_k}{4} \leq (1 - \delta_k/2)^2.$$

Thus, T_k is a contraction with factor $(1 - \delta_k/2)$.

Step 2: The trajectory is bounded. With the contractive property established, we now show that the sequence $(\mathbf{x}_k)$ remains in a bounded region. Since $F(\mathbf{x}^*) = \mathbf{0}$, we have $F_k(\mathbf{x}^*) = \eta_k F(\mathbf{x}^*) + \delta_k(\mathbf{x}^* - \mathbf{x}_0) = \delta_k(\mathbf{x}^* - \mathbf{x}_0)$. The update rule $\mathbf{x}_{k+1} = T_k(\mathbf{x}_k)$ gives:

$$\|\mathbf{x}_{k+1} - (\mathbf{x}^* - F_k(\mathbf{x}^*))\| = \|T_k(\mathbf{x}_k) - T_k(\mathbf{x}^*)\| \leq \left(1 - \frac{\delta_k}{2}\right) \|\mathbf{x}_k - \mathbf{x}^*\|.$$

Using the triangle inequality, we isolate $\|\mathbf{x}_{k+1} - \mathbf{x}^*\|$:

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\| \leq \left(1 - \frac{\delta_k}{2}\right) \|\mathbf{x}_k - \mathbf{x}^*\| + \|F_k(\mathbf{x}^*)\| = \left(1 - \frac{\delta_k}{2}\right) \|\mathbf{x}_k - \mathbf{x}^*\| + \delta_k \|\mathbf{x}^* - \mathbf{x}_0\|.$$

Next, we show by induction that $\|\mathbf{x}_k - \mathbf{x}^*\| \leq 2R$. The base case $\|\mathbf{x}_0 - \mathbf{x}^*\| \leq 2R$ is trivial. Assuming $\|\mathbf{x}_k - \mathbf{x}^*\| \leq 2R$, we have:

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\| \leq \left(1 - \frac{\delta_k}{2}\right) (2R) + \delta_k R = 2R.$$

The claim holds by induction. This also implies

$$\|\mathbf{x}_k - \mathbf{x}_0\| \leq \|\mathbf{x}_k - \mathbf{x}^*\| + \|\mathbf{x}^* - \mathbf{x}_0\| \leq 2R + R = 3R$$

Step 3: Error recurrence and final bound. We define the error term $E_k := \|F_{k-1}(\mathbf{x}_k)\|$. Since $\mathbf{x}_{k+1} = \mathbf{x}_k - F_k(\mathbf{x}_k)$ and $F_k(\mathbf{x}_{k+1})$ is the residual at the next point, we can relate them via the contraction of the map $Id - F_k$:

$$\begin{aligned} \|F_k(\mathbf{x}_{k+1})\| &= \|(\mathbf{x}_{k+1} - F_k(\mathbf{x}_{k+1})) - (\mathbf{x}_k - F_k(\mathbf{x}_k))\| \\ &\leq \left(1 - \frac{\delta_k}{2}\right) \|\mathbf{x}_{k+1} - \mathbf{x}_k\| = \left(1 - \frac{\delta_k}{2}\right) \|F_k(\mathbf{x}_k)\|. \end{aligned}$$

Next, we relate $\|F_k(\mathbf{x}_k)\|$ to E_k . From the definition of F_k , we have:

$$F_k(\mathbf{x}_k) = \frac{\eta_k}{\eta_{k-1}} F_{k-1}(\mathbf{x}_k) + \left(\delta_k - \frac{\eta_k}{\eta_{k-1}} \delta_{k-1}\right) (\mathbf{x}_k - \mathbf{x}_0).$$

Given that δ_k is non-increasing and $\eta_k/\eta_{k-1} = \sqrt{\delta_k/\delta_{k-1}} \leq 1$, we bound the norm:

$$\|F_k(\mathbf{x}_k)\| \leq \|F_{k-1}(\mathbf{x}_k)\| + (\delta_{k-1} - \delta_k) \|\mathbf{x}_k - \mathbf{x}_0\| \leq E_k + 3(\delta_{k-1} - \delta_k)R.$$

This yields the recurrence:

$$E_{k+1} \leq \left(1 - \frac{\delta_k}{2}\right) (E_k + 3(\delta_{k-1} - \delta_k)R).$$

For the base case, with $\mathbf{x}_1 = \mathbf{x}_0 - F_0(\mathbf{x}_0)$:

$$E_1 = \|F_0(\mathbf{x}_1)\| \leq \left(1 - \frac{\delta_0}{2}\right) \|F_0(\mathbf{x}_0)\| = \left(1 - \frac{\delta_0}{2}\right) \|\eta_0 F(\mathbf{x}_0)\| \leq \eta_0 LR = \frac{\sqrt{\delta_0}}{2} R.$$

where the first inequality is due to the contraction of $I - F_0$ and the last inequality is by Lipschitzness. Choosing $\delta_n = \frac{4}{n+4}$ and use [Lemma 3.2.1](#) we obtain that

$$E_k \leq \frac{12}{k+3} R$$

and hence

$$\|F(\mathbf{x}_k)\| \leq \frac{E_k + 3\delta_{k-1}R}{\eta_{k-1}} = O\left(\frac{RL}{\sqrt{k}}\right).$$

□

The $O(1/\sqrt{k})$ rate is a consequence of the operator only being Lipschitz. If F is also $1/L$ -cocoercive (a stronger condition than monotonicity), the analysis changes significantly. The operator $T(\mathbf{x}) = \mathbf{x} - \eta F(\mathbf{x})$ becomes non-expansive for $\eta \in (0, 2/L]$. One could then apply the Halpern iteration from the previous section directly to the fixed-point problem $\mathbf{x} = T(\mathbf{x})$, which is equivalent to $F(\mathbf{x}) = 0$. This approach allows a constant step size $\eta = O(1)$ instead of $\eta_k = O(1/\sqrt{k})$ and yields a faster $O(1/k)$ convergence rate for $\|F(\mathbf{x}_k)\|$, matching the

convergence of vanilla gradient descent over smooth convex functions (a condition further stronger than cocoercivity).

Additionally, note that the anchoring scheme $F_k(\mathbf{x}) = \eta_k F(\mathbf{x}) + \delta_k(\mathbf{x} - \mathbf{x}_0)$ can be readily generalized to the viscosity approximation $F_k(\mathbf{x}) = \eta_k F(\mathbf{x}) + \delta_k(\mathbf{x} - G(\mathbf{x}))$, where G is any contractive mapping. We can directly plug this into the forward step method, and the analysis follows the same structure: the update remains a contraction, the trajectory remains unbounded and recurrence is the same with slightly different definitions of E_k and constants.

Acceleration: A Regularized Extragradient Method

Having shown that our framework can impose convergence on a divergent method, we now demonstrate its second key capability: acceleration. The standard Extragradient (EG) method, unlike Forward Step, is guaranteed to converge for monotone operators, but its last-iterate residual norm, $\|F(\mathbf{x}_k)\|$, converges at the suboptimal $O(1/\sqrt{k})$ rate [48]. We will show that by applying the standard EG algorithm not to the original operator F , but to our sequence of regularized operators $\{F_k\}$, we accelerate this rate to the optimal $O(1/k)$.

The resulting algorithm, the Regularized Extragradient (Reg-EG) method, does not alter the structure of the EG update. It simply applies the two-step iteration to the strongly-monotone operator F_k at each phase k :

$$\mathbf{x}_{k+1/2} = \mathbf{x}_k - F_k(\mathbf{x}_k), \quad \mathbf{x}_{k+1} = \mathbf{x}_k - F_k(\mathbf{x}_{k+1/2}), \quad (3.17)$$

where $F_k(\mathbf{x}) := \eta F(\mathbf{x}) + \delta_k(\mathbf{x} - \mathbf{x}_0)$. This simple modification, driven entirely by our operator-centric framework, is sufficient to achieve an accelerated rate.

Theorem 3.4.2 (Convergence of the Regularized Extragradient Method). *Let F be a monotone, L -Lipschitz operator with a solution $\mathbf{x}^*$ such that $F(\mathbf{x}^*) = \mathbf{0}$. Consider the regularized extragradient method (3.17) with parameters $\eta = \frac{1}{8L}$ and a non-increasing sequence of regularization parameters (δ_k) given by $\delta_k = \frac{4}{k+32}$ for $k \geq 0$. Let $R := \|\mathbf{x}_0 - \mathbf{x}^*\|$. Then the iterates satisfy:*

$$\|F(\mathbf{x}_k)\| = O\left(\frac{LR}{k}\right).$$

Our Reg-EG method should be distinguished from the Extra Anchored Gradient (EAG) algorithm proposed by [138], which was the first EG variant to achieve an accelerated rate. The distinction lies in how the anchoring is applied. The EAG algorithm has the general form:

$$\mathbf{x}_{k+1/2} = \mathbf{x}_k + \beta_k(\mathbf{x}_0 - \mathbf{x}_k) - \alpha_k F(\mathbf{x}_k), \quad \mathbf{x}_{k+1} = \mathbf{x}_k + \beta_k(\mathbf{x}_0 - \mathbf{x}_k) - \alpha_k F(\mathbf{x}_{k+1/2}).$$

Note that the primary difference between our Reg-EG and EAG lies in the second step for the update $\mathbf{x}_{k+1}$. Reg-EG directly regularizes the operator while EAGs shrinks the update of $\mathbf{x}_{k+1}$ to the anchor $\mathbf{x}_0$. While both methods achieve the same optimal rate, the Reg-EG

approach leaves the base algorithm unchanged. This operator-centric design is conceptually clean, intuitive, and directly extensible to other contexts, such as the viscosity approximation schemes.

Proof Overview of Theorem 3.4.2

The proof proceeds in two stages. First, we establish a series of preparatory lemmas concerning the properties of the regularized operator F_k , its unique zero $\mathbf{x}_k^*$, and the behavior of standard extragradient on strongly monotone operators. Second, we use these lemmas to prove that the trajectory of the iterates $(\mathbf{x}_k)$ is bounded. With this boundedness guarantee, we derive a master recurrence relation for the norm of the residual, $\|F_{k-1}(\mathbf{x}_k)\|$, and solve it using Lemma 3.2.1 to obtain the final result.

Preparatory Lemmas

We begin with four lemmas that form the foundation of our analysis.

Lemma 3.4.3 (Boundedness of the Regularized Zero). *Let $F : \mathcal{H} \rightarrow \mathcal{H}$ be monotone, and let $\mathbf{x}^*$ satisfy $F(\mathbf{x}^*) = 0$. Fix $\mathbf{x}_0 \in \mathcal{H}$ and, for $\delta > 0$, define*

$$F_\delta(\mathbf{x}) := F(\mathbf{x}) + \delta(\mathbf{x} - \mathbf{x}_0), \quad \mathbf{x}_\delta^* \text{ as the unique zero of } F_\delta.$$

Then we have,

$$\|\mathbf{x}_\delta^* - \mathbf{x}^*\| \leq \|\mathbf{x}_0 - \mathbf{x}^*\| \quad \text{and} \quad \|\mathbf{x}_\delta^* - \mathbf{x}_0\| \leq \|\mathbf{x}_0 - \mathbf{x}^*\|.$$

Proof. Since $F(\mathbf{x}^*) = 0$, we have $F_\delta(\mathbf{x}^*) = F(\mathbf{x}^*) + \delta(\mathbf{x}^* - \mathbf{x}_0) = \delta(\mathbf{x}^* - \mathbf{x}_0)$. By the monotonicity of F_δ ,

$$0 \leq \langle F_\delta(\mathbf{x}_\delta^*) - F_\delta(\mathbf{x}^*), \mathbf{x}_\delta^* - \mathbf{x}^* \rangle = \langle -\delta(\mathbf{x}^* - \mathbf{x}_0), \mathbf{x}_\delta^* - \mathbf{x}^* \rangle = -\delta \langle \mathbf{x}^* - \mathbf{x}_0, \mathbf{x}_\delta^* - \mathbf{x}^* \rangle.$$

This implies $\langle \mathbf{x}_\delta^* - \mathbf{x}^*, \mathbf{x}_\delta^* - \mathbf{x}_0 \rangle \leq 0$. Using the identity $2\langle \mathbf{a}, \mathbf{b} \rangle = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - \|\mathbf{a} - \mathbf{b}\|^2$ with $\mathbf{a} = \mathbf{x}_\delta^* - \mathbf{x}^*$ and $\mathbf{b} = \mathbf{x}_\delta^* - \mathbf{x}_0$, we get

$$\|\mathbf{x}_\delta^* - \mathbf{x}^*\|^2 + \|\mathbf{x}_\delta^* - \mathbf{x}_0\|^2 - \|\mathbf{x}^* - \mathbf{x}_0\|^2 \leq 0.$$

The claimed inequalities follow directly. $\square$

Lemma 3.4.4 (Stability of the Regularized Zero). *Let $F : \mathcal{H} \rightarrow \mathcal{H}$ be monotone, and let $\mathbf{x}^*$ satisfy $F(\mathbf{x}^*) = 0$. Fix $\mathbf{x}_0 \in \mathcal{H}$ and, for $\delta_1, \delta_2 > 0$, $i = 1, 2$, define*

$$F_{\delta_i}(\mathbf{x}) := F(\mathbf{x}) + \delta_i(\mathbf{x} - \mathbf{x}_0), \quad \mathbf{x}_{\delta_i}^* \text{ the unique zero of } F_{\delta_i}.$$

Then we have

$$\|\mathbf{x}_{\delta_1}^* - \mathbf{x}_{\delta_2}^*\| \leq \frac{|\delta_1 - \delta_2|}{\max\{\delta_1, \delta_2\}} \|\mathbf{x}^* - \mathbf{x}_0\|$$

Proof. Since $F_{\delta_i}(\mathbf{x}_{\delta_i}^*) = 0$ for $i = 1, 2$, we have

$$\langle F_{\delta_1}(\mathbf{x}_{\delta_1}^*) - F_{\delta_2}(\mathbf{x}_{\delta_2}^*), \mathbf{x}_{\delta_1}^* - \mathbf{x}_{\delta_2}^* \rangle = 0.$$

Next, note that $F_{\delta_i}(\mathbf{x}^*) = \delta_i(\mathbf{x}^* - \mathbf{x}_0)$, plugging this and the definition of F_δ we obtain that:

$$\langle F(\mathbf{x}_{\delta_1}^*) - F(\mathbf{x}_{\delta_2}^*), \mathbf{x}_{\delta_1}^* - \mathbf{x}_{\delta_2}^* \rangle + \delta_1 \|\mathbf{x}_{\delta_1}^* - \mathbf{x}_{\delta_2}^*\|^2 + (\delta_1 - \delta_2) \langle \mathbf{x}_{\delta_2}^* - \mathbf{x}_0, \mathbf{x}_{\delta_1}^* - \mathbf{x}_{\delta_2}^* \rangle = 0.$$

Monotonicity makes the first term non-negative, so

$$\delta_1 \|\mathbf{x}_{\delta_1}^* - \mathbf{x}_{\delta_2}^*\|^2 \leq |\delta_1 - \delta_2| \|\mathbf{x}_{\delta_2}^* - \mathbf{x}_0\| \|\mathbf{x}_{\delta_1}^* - \mathbf{x}_{\delta_2}^*\|.$$

Applying Lemma 3.4.3 ($\|\mathbf{x}_{\delta_2}^* - \mathbf{x}_0\| \leq \|\mathbf{x}^* - \mathbf{x}_0\|$), we obtain:

$$\delta_1 \|\mathbf{x}_{\delta_1}^* - \mathbf{x}_{\delta_2}^*\| \leq |\delta_1 - \delta_2| \|\mathbf{x}^* - \mathbf{x}_0\|.$$

The result follows by replacing δ_1 on the left with $\max\{\delta_1, \delta_2\}$ on the right, which weakens the bound but provides the stated symmetric form. $\square$

Lemma 3.4.5 (Linear Convergence of Extragradient). *Let $G : \mathcal{H} \rightarrow \mathcal{H}$ be L_G -Lipschitz and μ_G -strongly monotone with condition number $\kappa = L_G/\mu_G$. The extragradient iteration for finding the zero of G , given by $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta G(\mathbf{x}_t - \eta G(\mathbf{x}_t))$, with step size $\eta = \frac{1}{4L_G}$, converges linearly:*

$$\|\mathbf{x}_{t+1} - \mathbf{x}_G^*\| \leq \left(1 - \frac{1}{8\kappa}\right) \|\mathbf{x}_t - \mathbf{x}_G^*\|,$$

where $\mathbf{x}_G^*$ is the unique solution to $G(\mathbf{x}_G^*) = 0$.

This is a standard result for the extragradient method. The proof relies on expanding the norms $\|\mathbf{x}_{t+1/2} - \mathbf{x}_G^*\|^2$ and $\|\mathbf{x}_{t+1} - \mathbf{x}_G^*\|^2$ and carefully applying the Lipschitz and strong monotonicity properties of G . Here we provide a self-contained proof for completeness.

Proof. Let $\mathbf{x}_{t+1/2} := \mathbf{x}_t - \eta G(\mathbf{x}_t)$. We first establish a lower bound for $\|\mathbf{x}_{t+1/2} - \mathbf{x}_G^*\|^2$:

$$\begin{aligned} \|\mathbf{x}_{t+1/2} - \mathbf{x}_G^*\|^2 &= \|\mathbf{x}_t - \mathbf{x}_G^* - \eta(G(\mathbf{x}_t) - G(\mathbf{x}_G^*))\|^2 \\ &= \|\mathbf{x}_t - \mathbf{x}_G^*\|^2 - 2\eta \langle \mathbf{x}_t - \mathbf{x}_G^*, G(\mathbf{x}_t) - G(\mathbf{x}_G^*) \rangle + \eta^2 \|G(\mathbf{x}_t) - G(\mathbf{x}_G^*)\|^2 \\ &\geq \|\mathbf{x}_t - \mathbf{x}_G^*\|^2 - 2\eta L_G \|\mathbf{x}_t - \mathbf{x}_G^*\|^2 \\ &= (1 - 2\eta L_G) \|\mathbf{x}_t - \mathbf{x}_G^*\|^2 \end{aligned}$$

where in the inequality we used the Cauchy-Schwarz inequality and the L_G -Lipschitz continuity of G .

Next, we bound $\|\mathbf{x}_{t+1} - \mathbf{x}_G^*\|^2$:

$$\begin{aligned} \|\mathbf{x}_{t+1} - \mathbf{x}_G^*\|^2 &= \|\mathbf{x}_t - \mathbf{x}_G^* - \eta G(\mathbf{x}_{t+1/2})\|^2 \\ &= \|\mathbf{x}_t - \mathbf{x}_G^*\|^2 - 2\eta \langle \mathbf{x}_t - \mathbf{x}_G^*, G(\mathbf{x}_{t+1/2}) \rangle + \eta^2 \|G(\mathbf{x}_{t+1/2})\|^2 \end{aligned} \quad (3.18)$$

We decompose the inner product term as follows:

$$\langle \mathbf{x}_t - \mathbf{x}_G^*, G(\mathbf{x}_{t+1/2}) \rangle = \langle \mathbf{x}_{t+1/2} - \mathbf{x}_G^*, G(\mathbf{x}_{t+1/2}) \rangle + \eta \langle G(\mathbf{x}_t), G(\mathbf{x}_{t+1/2}) \rangle.$$

By μ_G -strong monotonicity, the first term on the right-hand side is bounded below by $\mu_G \|\mathbf{x}_{t+1/2} - \mathbf{x}_G^*\|^2$. For the second term, we use the identity $2\langle \mathbf{a}, \mathbf{b} \rangle = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - \|\mathbf{a} - \mathbf{b}\|^2$ and the L_G -Lipschitz property to get:

$$\begin{aligned} 2\eta \langle G(\mathbf{x}_t), G(\mathbf{x}_{t+1/2}) \rangle &\geq \eta (\|G(\mathbf{x}_t)\|^2 + \|G(\mathbf{x}_{t+1/2})\|^2 - L_G^2 \|\mathbf{x}_t - \mathbf{x}_{t+1/2}\|^2) \\ &= \eta (\|G(\mathbf{x}_t)\|^2 + \|G(\mathbf{x}_{t+1/2})\|^2 - \eta^2 L_G^2 \|G(\mathbf{x}_t)\|^2) \\ &= \eta (1 - \eta^2 L_G^2) \|G(\mathbf{x}_t)\|^2 + \eta \|G(\mathbf{x}_{t+1/2})\|^2. \end{aligned}$$

Substituting these bounds into (3.18):

$$\|\mathbf{x}_{t+1} - \mathbf{x}_G^*\|^2 \leq \|\mathbf{x}_t - \mathbf{x}_G^*\|^2 - 2\eta \mu_G \|\mathbf{x}_{t+1/2} - \mathbf{x}_G^*\|^2 - \eta^2 (1 - \eta^2 L_G^2) \|G(\mathbf{x}_t)\|^2.$$

Using the bounds $\|G(\mathbf{x}_t)\|^2 \geq \mu_G^2 \|\mathbf{x}_t - \mathbf{x}_G^*\|^2$ (from strong monotonicity) and our initial lower bound $\|\mathbf{x}_{t+1/2} - \mathbf{x}_G^*\|^2 \geq (1 - 2\eta L_G) \|\mathbf{x}_t - \mathbf{x}_G^*\|^2$, we get:

$$\|\mathbf{x}_{t+1} - \mathbf{x}_G^*\|^2 \leq (1 - 2\eta \mu_G (1 - 2\eta L_G) - \eta^2 \mu_G^2 (1 - \eta^2 L_G^2)) \|\mathbf{x}_t - \mathbf{x}_G^*\|^2.$$

For the step size $\eta = \frac{1}{4L_G}$, the contraction factor for the squared norm is:

$$1 - 2\frac{\mu_G}{4L_G} \left(1 - \frac{1}{2}\right) - \left(\frac{\mu_G}{4L_G}\right)^2 \left(1 - \frac{1}{16}\right) = 1 - \frac{1}{4\kappa} - \frac{15}{256\kappa^2}.$$

We can thus bound the factor by a simpler term:

$$\|\mathbf{x}_{t+1} - \mathbf{x}_G^*\|^2 \leq \left(1 - \frac{1}{4\kappa}\right) \|\mathbf{x}_t - \mathbf{x}_G^*\|^2.$$

Taking the square root and applying the inequality $\sqrt{1-z} \leq 1 - z/2$ for $z = \frac{1}{4\kappa} \in [0, 1)$, we conclude that

$$\|\mathbf{x}_{t+1} - \mathbf{x}_G^*\| \leq \left(1 - \frac{1}{8\kappa}\right) \|\mathbf{x}_t - \mathbf{x}_G^*\|.$$

□

Lemma 3.4.6 (Boundedness of the Trajectory). *Let $(\mathbf{x}_k)_{k \geq 0}$ be generated by the Reg-EG scheme (3.17) with parameters $\eta = \frac{1}{8L}$ and δ_k chosen such that $\delta_k \leq \delta_{k-1} \leq \dots \leq \delta_0 \leq \frac{1}{8L}$. Denote by $\mathbf{x}_k^*$ the unique zero of F_k . If $\delta_k = \frac{c}{k+8c}$ for some $c \geq 2$, then for every $k \geq 0$, we have $\|\mathbf{x}_k - \mathbf{x}_0\| \leq 2R$.*

Proof. At phase k , the algorithm performs one EG step on the operator F_k . This operator is δ_k -strongly monotone and has a Lipschitz constant $L_k = \eta L + \delta_k$. With our choice of $\eta = 1/(8L)$ and $\delta_k \leq 1/(8L)$, we have $L_k \leq 1/(4L)$. We can apply Lemma 3.4.5 with $G = F_k$, $\mu_G = \delta_k$, $L_G = 1/4$, and a step size of 1. This gives the local contraction around the zero $\mathbf{x}_k^*$:

$$\|\mathbf{x}_{k+1} - \mathbf{x}_k^*\| \leq \left(1 - \frac{\delta_k}{8 \cdot \frac{1}{4}}\right) \|\mathbf{x}_k - \mathbf{x}_k^*\| = \left(1 - \frac{\delta_k}{2}\right) \|\mathbf{x}_k - \mathbf{x}_k^*\|.$$

Using the triangle inequality and Lemma 3.4.4 to handle the shift from $\mathbf{x}_k^*$ to $\mathbf{x}_{k-1}^*$:

$$\begin{aligned} \|\mathbf{x}_k - \mathbf{x}_k^*\| &\leq \|\mathbf{x}_k - \mathbf{x}_{k-1}^*\| + \|\mathbf{x}_{k-1}^* - \mathbf{x}_k^*\| \\ &\leq \|\mathbf{x}_k - \mathbf{x}_{k-1}^*\| + \frac{\delta_{k-1} - \delta_k}{\delta_{k-1}} \|\mathbf{x}^* - \mathbf{x}_0\|. \end{aligned}$$

Combining these yields the recurrence:

$$\|\mathbf{x}_{k+1} - \mathbf{x}_k^*\| \leq \left(1 - \frac{\delta_k}{2}\right) \left(\|\mathbf{x}_k - \mathbf{x}_{k-1}^*\| + \frac{\delta_{k-1} - \delta_k}{\delta_{k-1}} R\right).$$

We show $\|\mathbf{x}_k - \mathbf{x}_{k-1}^*\| \leq R$ by induction. The base case $k = 1$ holds by contraction from $\mathbf{x}_0$ and Lemma 3.4.3. For the inductive step, assuming $\|\mathbf{x}_k - \mathbf{x}_{k-1}^*\| \leq R$, and with $\delta_k = \frac{c}{k+8c}$, the recurrence gives $\|\mathbf{x}_{k+1} - \mathbf{x}_k^*\| < R$. Finally, by Lemma 3.4.3, $\|\mathbf{x}_k^* - \mathbf{x}_0\| \leq R$. The triangle inequality gives the result:

$$\|\mathbf{x}_{k+1} - \mathbf{x}_0\| \leq \|\mathbf{x}_{k+1} - \mathbf{x}_k^*\| + \|\mathbf{x}_k^* - \mathbf{x}_0\| < R + R = 2R.$$

□

Main Proof Argument

We now proceed with the proof of the main theorem. The main idea of the proof is similar to before. We track the residuals for the current monotone equation problem associated with F_k and show that a shift from F_{k-1} to F_k is a controllable perturbation. The mapping $\mathbf{x} \mapsto \mathbf{x} - F_\delta(\mathbf{x} - F_\delta(\mathbf{x}))$ is not a contraction and thus arguments from the previous sections are not directly applicable. Nonetheless, we can still obtain a contraction of $\|F_k(\mathbf{x}_{k+1})\|$ under suitable step size choices. This is analogous to a known fact about EG: the mapping $\mathbf{x} \mapsto \mathbf{x} - \eta F(\mathbf{x} - \eta F(\mathbf{x}))$ is not co-coercive and yet is non-increasing under suitable step size choices [48].

To begin, let us define the following shorthand notations for the residuals at step k :

$$\mathbf{v}_1 = F_k(\mathbf{x}_k), \quad \mathbf{v}_2 = F_k(\mathbf{x}_{k+1/2}), \quad \mathbf{v}_3 = F_k(\mathbf{x}_{k+1}).$$

A direct expansion of the difference of squared norms yields:

$$\|\mathbf{v}_3\|^2 - \|\mathbf{v}_1\|^2 = \|(\mathbf{v}_3 - \mathbf{v}_2) + \mathbf{v}_2\|^2 - \|(\mathbf{v}_1 - \mathbf{v}_2) + \mathbf{v}_2\|^2$$

$$\begin{aligned}
&= \|\mathbf{v}_3 - \mathbf{v}_2\|^2 - \|\mathbf{v}_1 - \mathbf{v}_2\|^2 + 2\langle \mathbf{v}_3 - \mathbf{v}_1, \mathbf{v}_2 \rangle \\
&\leq ((\eta L + \delta_k)^2 - 1) \|\mathbf{v}_1 - \mathbf{v}_2\|^2 - 2\delta_k \|\mathbf{v}_2\|^2,
\end{aligned}$$

where the final inequality is obtained by applying two key properties. First, we use the Lipschitz continuity of F_k to bound the term $\|\mathbf{v}_3 - \mathbf{v}_2\|$:

$$\|\mathbf{v}_3 - \mathbf{v}_2\| \leq (\eta L + \delta_k) \|\mathbf{x}_{k+1} - \mathbf{x}_{k+1/2}\| = (\eta L + \delta_k) \|\mathbf{v}_1 - \mathbf{v}_2\|.$$

Second, the strong-monotonicity of F_k provides a bound for the inner product term: $\langle \mathbf{v}_3 - \mathbf{v}_1, \mathbf{v}_2 \rangle \leq -\delta_k \|\mathbf{v}_2\|^2$.

We can further simplify this expression. Given our choice of parameters, we have:

$$1 - (\eta L + \delta_k) \geq 1 - (\eta L + \delta_0) = \frac{3}{4}.$$

Furthermore, applying the Lipschitz continuity of F_k between $\mathbf{x}_k$ and $\mathbf{x}_{k+1/2}$ gives:

$$\|\mathbf{v}_1 - \mathbf{v}_2\| \leq (\eta L + \delta_k) \|\mathbf{v}_1\|,$$

and then, by the reverse-triangle inequality,

$$\|\mathbf{v}_2\| \geq \|\mathbf{v}_1\| - \|\mathbf{v}_1 - \mathbf{v}_2\| \geq (1 - (\eta L + \delta_k)) \|\mathbf{v}_1\| \geq \frac{3}{4} \|\mathbf{v}_1\|,$$

which in turn implies a lower bound on the squared norm of $\mathbf{v}_2$: $\|\mathbf{v}_2\|^2 \geq \frac{1}{2} \|\mathbf{v}_1\|^2$. Substituting this bound back into our main inequality, we get:

$$\|\mathbf{v}_3\|^2 - \|\mathbf{v}_1\|^2 \leq -2\delta_k \frac{1}{2} \|\mathbf{v}_1\|^2 = -\delta_k \|\mathbf{v}_1\|^2,$$

This result leads to a direct contraction for the norm:

$$\|\mathbf{v}_3\| \leq \sqrt{1 - \delta_k} \|\mathbf{v}_1\| \leq (1 - \frac{\delta_k}{2}) \|\mathbf{v}_1\|.$$

Let us define the error at step $k + 1$ as the residual:

$$E_{k+1} := \|F_k(\mathbf{x}_{k+1})\| = \|\mathbf{v}_3\|.$$

To establish a recurrence relation for E_k , we first relate $F_k(\mathbf{x}_k)$ to $F_{k-1}(\mathbf{x}_k)$:

$$F_k(\mathbf{x}_k) - F_{k-1}(\mathbf{x}_k) = (\delta_k - \delta_{k-1})(\mathbf{x}_k - \mathbf{x}_0).$$

Combining this with the previously derived contraction on $\|\mathbf{v}_3\|$ yields the following master recurrence:

$$E_{k+1} \leq (1 - \frac{\delta_k}{2}) \left(E_k + (\delta_{k-1} - \delta_k) \|\mathbf{x}_k - \mathbf{x}_0\| \right). \quad (3.19)$$

For the base case, we examine the initialization condition:

$$E_1 = \|F_0(\mathbf{x}_1)\| \leq (1 - \frac{\delta_0}{2}) \|F_0(\mathbf{x}_0)\| < \eta L \|\mathbf{x}_0 - \mathbf{x}^*\| = \frac{R}{8}.$$

Using the boundedness from Lemma 3.4.6 ($\|\mathbf{x}_k - \mathbf{x}_0\| \leq 2R$) and the base case $E_1 < R/8$, we apply Lemma 3.2.1 with $\delta_k = \frac{4}{k+32}$. This yields the bound:

$$E_{k+1} \leq \frac{8}{k+32}R.$$

Finally, using the definition of F_k and the bound on E_k , we can establish the convergence rate for the original problem:

$$\|F(\mathbf{x}_k)\| \leq \frac{E_k + 2\delta_{k-1}R}{\eta} \leq \frac{128LR}{k+31} = O\left(\frac{LR}{k}\right).$$

□

Acceleration: A Regularized Optimistic Gradient Method

To further demonstrate the versatility of our framework, we now apply it to another cornerstone algorithm for monotone operators: the Optimistic Gradient (OG) method. Showing that our regularization strategy also accelerates OG provides compelling evidence that the acceleration is a robust outcome of our operator-centric approach, rather than an artifact of the specific structure of the Extragradient method.

The standard OG method, which uses past gradient information, can be written in a two-step form with an auxiliary sequence ($\tilde{\mathbf{x}}_k$):

$$\begin{aligned}\tilde{\mathbf{x}}_k &= \mathbf{x}_k - \eta F(\tilde{\mathbf{x}}_{k-1}), \\ \mathbf{x}_{k+1} &= \mathbf{x}_k - \eta F(\tilde{\mathbf{x}}_k).\end{aligned}$$

Consistent with our framework, we create the Regularized Optimistic Gradient (Reg-OG) method by simply replacing the operator F in the update with its regularized version, F_k . This yields the algorithm:

$$\tilde{\mathbf{x}}_k = \mathbf{x}_k - F_k(\tilde{\mathbf{x}}_{k-1}), \tag{3.20}$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k - F_k(\tilde{\mathbf{x}}_k), \tag{3.21}$$

where $F_k(\mathbf{x}) := \eta F(\mathbf{x}) + \delta_k(\mathbf{x} - \mathbf{x}_0)$ and we initialize $\tilde{\mathbf{x}}_{-1} = \mathbf{x}_0$. As with Reg-EG, this method achieves the optimal $O(1/k)$ convergence rate, providing a final, powerful confirmation of our framework's utility.

Theorem 3.4.7 (Convergence of the Regularized Optimistic Gradient Method). *Let F be a monotone, L -Lipschitz operator with a solution $\mathbf{x}^*$ such that $F(\mathbf{x}^*) = \mathbf{0}$. Consider the Regularized Optimistic Gradient method defined by (3.20)-(3.21) with parameters $\eta = \frac{1}{8L}$ and a non-increasing sequence of regularization parameters (δ_k) given by $\delta_k = \frac{8}{k+64}$ for $k \geq 0$. Let $R := \|\mathbf{x}_0 - \mathbf{x}^*\|$. Then the iterates satisfy:*

$$\|F(\mathbf{x}_k)\| = O\left(\frac{LR}{k}\right).$$

Our Reg-OG method should be distinguished from the *Anchored Popov* (A-Popov) algorithm proposed by [124], the first Popov-type scheme to achieve an $O(1/k)$ rate. The distinction, once again, lies in the way anchoring is incorporated. A-Popov applies the anchor inside the iterates:

$$y_k = \beta_k(\mathbf{x}_0 - \mathbf{x}_k) + \mathbf{x}_k - \eta_k F(y_{k-1}), \quad \mathbf{x}_{k+1} = \beta_k(\mathbf{x}_0 - \mathbf{x}_k) + \mathbf{x}_k - \eta_k F(y_k),$$

whereas Reg-OG directly replaces the operator with the regularized one.

Proof overview of Theorem 3.4.7

The proof is more intricate than for the previous methods and relies on a specialized potential function. Our strategy is as follows:

1. We present the preparatory lemmas that establish the core contraction property of the OMD update and a key inequality for norms.
2. We conduct a detailed analysis to prove that the iterates $(\mathbf{x}_k)$ of the anchored algorithm remain within a bounded region. This proof uses its own Lyapunov-like potential function, Φ_k .
3. With boundedness established, we introduce a second potential function, E_k , tied to the operator residual. We show that E_k follows a recurrence that, when combined with the boundedness result, yields the final convergence rate via Lemma 3.2.1.

Preparatory Lemmas

Lemma 3.4.8 (One-Step Contraction). *Let $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d} \in \mathbb{R}^n$ and suppose $0 < \delta \leq \frac{1}{3}$. We assume the following two conditions hold:*

$$\langle \mathbf{b} - \mathbf{a}, \mathbf{c} \rangle \geq \delta \|\mathbf{c}\|^2, \quad (\text{strong monotonicity}) \quad (3.22)$$

$$\|\mathbf{a} - \mathbf{c}\| \leq \frac{1}{3} \|\mathbf{c} - \mathbf{d}\|, \quad (\text{Lipschitz bound}) \quad (3.23)$$

Let us define a potential function $\Phi(\mathbf{x}, \mathbf{y}) := \|\mathbf{x}\|^2 + 2\|\mathbf{x} - \mathbf{y}\|^2$. Then, we have the following contraction property:

$$\Phi(\mathbf{a}, \mathbf{c}) \leq \left(1 - \frac{\delta}{2}\right) \Phi(\mathbf{b}, \mathbf{d}). \quad (3.24)$$

Proof. The proof proceeds by establishing a lower bound on the difference $\Phi(\mathbf{b}, \mathbf{d}) - \Phi(\mathbf{a}, \mathbf{c})$ and showing that this is greater than or equal to $\frac{\delta}{2} \Phi(\mathbf{b}, \mathbf{d})$.

Let's begin by expanding the difference between the potentials:

$$\Phi(\mathbf{b}, \mathbf{d}) - \Phi(\mathbf{a}, \mathbf{c}) = (\|\mathbf{b}\|^2 - \|\mathbf{a}\|^2) + 2(\|\mathbf{b} - \mathbf{d}\|^2 - \|\mathbf{a} - \mathbf{c}\|^2).$$

We first find a lower bound for the term $\|\mathbf{b}\|^2 - \|\mathbf{a}\|^2$. By decomposing $\mathbf{a}$ as $(\mathbf{a} - \mathbf{c}) + \mathbf{c}$ and invoking the strong monotonicity condition from Eq. (3.22), we obtain:

$$\begin{aligned}\|\mathbf{b}\|^2 - \|\mathbf{a}\|^2 &= \|\mathbf{b} - \mathbf{a}\|^2 + 2\langle \mathbf{a}, \mathbf{b} - \mathbf{a} \rangle \\ &= \|\mathbf{b} - \mathbf{a}\|^2 + 2\langle \mathbf{a} - \mathbf{c}, \mathbf{b} - \mathbf{a} \rangle + 2\langle \mathbf{c}, \mathbf{b} - \mathbf{a} \rangle \\ &\geq \|\mathbf{b} - \mathbf{a}\|^2 + 2\langle \mathbf{a} - \mathbf{c}, \mathbf{b} - \mathbf{a} \rangle + 2\delta\|\mathbf{c}\|^2.\end{aligned}$$

Using the identity $\|\mathbf{x}\|^2 + 2\langle \mathbf{y}, \mathbf{x} \rangle = \|\mathbf{x} + \mathbf{y}\|^2 - \|\mathbf{y}\|^2$ with $\mathbf{x} = \mathbf{b} - \mathbf{a}$ and $\mathbf{y} = \mathbf{a} - \mathbf{c}$, the first two terms simplify to $\|\mathbf{b} - \mathbf{c}\|^2 - \|\mathbf{a} - \mathbf{c}\|^2$. This gives:

$$\|\mathbf{b}\|^2 - \|\mathbf{a}\|^2 \geq \|\mathbf{b} - \mathbf{c}\|^2 - \|\mathbf{a} - \mathbf{c}\|^2 + 2\delta\|\mathbf{c}\|^2.$$

Substituting this result back into the expression for the potential difference yields:

$$\Phi(\mathbf{b}, \mathbf{d}) - \Phi(\mathbf{a}, \mathbf{c}) \geq \|\mathbf{b} - \mathbf{c}\|^2 + 2\|\mathbf{b} - \mathbf{d}\|^2 - 3\|\mathbf{a} - \mathbf{c}\|^2 + 2\delta\|\mathbf{c}\|^2. \quad (3.25)$$

Next, we apply the Lipschitz bound from Eq. (3.23) to handle the negative term involving $\mathbf{a} - \mathbf{c}$:

$$\Phi(\mathbf{b}, \mathbf{d}) - \Phi(\mathbf{a}, \mathbf{c}) \geq \|\mathbf{b} - \mathbf{c}\|^2 + 2\|\mathbf{b} - \mathbf{d}\|^2 - \frac{1}{3}\|\mathbf{d} - \mathbf{c}\|^2 + 2\delta\|\mathbf{c}\|^2. \quad (3.26)$$

Our next step is to relate this lower bound to $\Phi(\mathbf{b}, \mathbf{d})$. We establish an upper bound for $\frac{\delta}{2}\Phi(\mathbf{b}, \mathbf{d})$ by applying the triangle inequality in the form $\|\mathbf{x} + \mathbf{y}\|^2 \leq 2(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2)$, which gives $\|\mathbf{b}\|^2 = \|(\mathbf{b} - \mathbf{c}) + \mathbf{c}\|^2 \leq 2(\|\mathbf{b} - \mathbf{c}\|^2 + \|\mathbf{c}\|^2)$. Therefore,

$$\frac{\delta}{2}\Phi(\mathbf{b}, \mathbf{d}) = \frac{\delta}{2}(\|\mathbf{b}\|^2 + 2\|\mathbf{b} - \mathbf{d}\|^2) \leq \delta(\|\mathbf{b} - \mathbf{c}\|^2 + \|\mathbf{c}\|^2 + \|\mathbf{b} - \mathbf{d}\|^2). \quad (3.27)$$

Finally, we combine our results. Subtracting the bound in Eq. (3.27) from the inequality in Eq. (3.26) yields:

$$\Phi(\mathbf{b}, \mathbf{d}) - \Phi(\mathbf{a}, \mathbf{c}) - \frac{\delta}{2}\Phi(\mathbf{b}, \mathbf{d}) \geq (1 - \delta)\|\mathbf{b} - \mathbf{c}\|^2 + (2 - \delta)\|\mathbf{b} - \mathbf{d}\|^2 - \frac{1}{3}\|\mathbf{d} - \mathbf{c}\|^2 + \delta\|\mathbf{c}\|^2.$$

To simplify the right-hand side, we again use the triangle inequality: $\|\mathbf{d} - \mathbf{c}\|^2 = \|(\mathbf{d} - \mathbf{b}) + (\mathbf{b} - \mathbf{c})\|^2 \leq 2\|\mathbf{b} - \mathbf{d}\|^2 + 2\|\mathbf{b} - \mathbf{c}\|^2$. This leads to:

$$\begin{aligned}(1 - \frac{\delta}{2})\Phi(\mathbf{b}, \mathbf{d}) - \Phi(\mathbf{a}, \mathbf{c}) &\geq (1 - \delta - \frac{2}{3})\|\mathbf{b} - \mathbf{c}\|^2 + (2 - \delta - \frac{2}{3})\|\mathbf{b} - \mathbf{d}\|^2 + \delta\|\mathbf{c}\|^2 \\ &= (\frac{1}{3} - \delta)\|\mathbf{b} - \mathbf{c}\|^2 + (\frac{4}{3} - \delta)\|\mathbf{b} - \mathbf{d}\|^2 + \delta\|\mathbf{c}\|^2.\end{aligned}$$

Since we assumed $0 < \delta \leq \frac{1}{3}$, all the coefficients on the right-hand side are non-negative, which implies the entire expression is non-negative. Thus,

$$(1 - \frac{\delta}{2})\Phi(\mathbf{b}, \mathbf{d}) - \Phi(\mathbf{a}, \mathbf{c}) \geq 0,$$

which rearranges to $\Phi(\mathbf{a}, \mathbf{c}) \leq (1 - \delta/2)\Phi(\mathbf{b}, \mathbf{d})$, the desired result. $\square$

Lemma 3.4.9 (Square-Root Perturbation Inequality). *Let $(\mathcal{X}, \|\cdot\|)$ be a normed space. For any $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d} \in \mathcal{X}$ with $\|\mathbf{b}\| \leq \|\mathbf{d}\|$,*

$$\sqrt{\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2} \leq \sqrt{\|\mathbf{c}\|^2 + \|\mathbf{d}\|^2} + \|\mathbf{a} - \mathbf{c}\|.$$

Proof. Since $\|\mathbf{b}\| \leq \|\mathbf{d}\|$,

$$\sqrt{\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2} \leq \sqrt{\|\mathbf{a}\|^2 + \|\mathbf{d}\|^2}.$$

We can view the pair of scalar norms $(\|\mathbf{a}\|, \|\mathbf{d}\|)$ as a vector in $\mathbb{R}^2$ and write

$$(\|\mathbf{a}\|, \|\mathbf{d}\|) = (\|\mathbf{c}\|, \|\mathbf{d}\|) + (\|\mathbf{a}\| - \|\mathbf{c}\|, 0).$$

Applying the standard triangle inequality for vectors in $\mathbb{R}^2$ gives

$$\sqrt{\|\mathbf{a}\|^2 + \|\mathbf{d}\|^2} \leq \sqrt{\|\mathbf{c}\|^2 + \|\mathbf{d}\|^2} + \|\|\mathbf{a}\| - \|\mathbf{c}\|\| \leq \sqrt{\|\mathbf{c}\|^2 + \|\mathbf{d}\|^2} + \|\mathbf{a} - \mathbf{c}\|,$$

where the last step follows from the reverse triangle inequality. Combining the two displayed inequalities yields the claim. $\square$

Proof of Trajectory Boundedness

First, we establish that the iterates generated by the algorithm remain within a bounded region. We set the step-size to $\eta = \frac{1}{8L}$ and let $\delta_k \leq \delta_0 \leq \frac{1}{8}$. This choice ensures that F_k is $1/4$ -Lipschitz for any $k \geq 0$. Our immediate goal is to prove the following inequality, which indicates that a specific Lyapunov function is contracting:

$$\|\mathbf{x}_{k+1} - \mathbf{x}_k^*\|^2 + \frac{1}{16} \|\tilde{\mathbf{x}}_k - \tilde{\mathbf{x}}_{k-1}\|^2 \leq (1 - \delta_k) \left(\|\mathbf{x}_k - \mathbf{x}_k^*\|^2 + \frac{1}{16} \|\tilde{\mathbf{x}}_{k-1} - \tilde{\mathbf{x}}_{k-2}\|^2 \right).$$

The proof of this statement relies on the following key lemma.

Lemma 3.4.10. *By combining the δ_k -strong monotonicity and L -Lipschitz properties of F_k with the algorithm's update rules, we obtain the following one-step inequality:*

$$\|\mathbf{x}_{k+1} - \mathbf{x}_k^*\|^2 \leq (1 - \delta_k) \|\mathbf{x}_k - \mathbf{x}_k^*\|^2 - (1 - 2\delta_k) \|\mathbf{x}_k - \tilde{\mathbf{x}}_k\|^2 + \frac{1}{16} \|\tilde{\mathbf{x}}_k - \tilde{\mathbf{x}}_{k-1}\|^2$$

Proof. We begin by establishing a lower bound on the term $2 \langle F_k(\tilde{\mathbf{x}}_k), \tilde{\mathbf{x}}_k - \mathbf{x}_k^* \rangle$. Using the δ_k -strong monotonicity of F_k and the inequality $\|\mathbf{a} + \mathbf{b}\|^2 \leq 2\|\mathbf{a}\|^2 + 2\|\mathbf{b}\|^2$, we have:

$$\begin{aligned} 2 \langle F_k(\tilde{\mathbf{x}}_k), \tilde{\mathbf{x}}_k - \mathbf{x}_k^* \rangle &\geq 2\delta_k \|\tilde{\mathbf{x}}_k - \mathbf{x}_k^*\|^2 \\ &\geq \delta_k (\|\mathbf{x}_k - \mathbf{x}_k^*\|^2 - 2\|\mathbf{x}_k - \tilde{\mathbf{x}}_k\|^2). \end{aligned}$$

Next, we derive an upper bound for the same term. From the update rule, we know that $F_k(\tilde{\mathbf{x}}_k) = \mathbf{x}_k - \mathbf{x}_{k+1}$. Substituting this into the expression gives:

$$2 \langle F_k(\tilde{\mathbf{x}}_k), \tilde{\mathbf{x}}_k - \mathbf{x}_k^* \rangle = 2 \langle \mathbf{x}_k - \mathbf{x}_{k+1}, \tilde{\mathbf{x}}_k - \mathbf{x}_k^* \rangle$$

$$= \|\mathbf{x}_k - \mathbf{x}_k^*\|^2 - \|\mathbf{x}_{k+1} - \mathbf{x}_k^*\|^2 - \|\mathbf{x}_k - \tilde{\mathbf{x}}_k\|^2 + \|\tilde{\mathbf{x}}_k - \mathbf{x}_{k+1}\|^2.$$

The final term in this expression can be bounded using the Lipschitz property of F_k :

$$\|\tilde{\mathbf{x}}_k - \mathbf{x}_{k+1}\|^2 = \|F_k(\tilde{\mathbf{x}}_k) - F_k(\tilde{\mathbf{x}}_{k-1})\|^2 \leq \frac{1}{16} \|\tilde{\mathbf{x}}_k - \tilde{\mathbf{x}}_{k-1}\|^2.$$

Combining the lower and upper bounds and rearranging the terms yields the result stated in the lemma. $\square$

With this lemma, we can now proceed to prove the main contractive inequality. First, we relate $\|\mathbf{x}_k - \tilde{\mathbf{x}}_k\|^2$ to quantities from the previous iteration. By the triangle inequality and the update rule:

$$\begin{aligned} \|\tilde{\mathbf{x}}_k - \tilde{\mathbf{x}}_{k-1}\|^2 &\leq 2\|\mathbf{x}_k - \tilde{\mathbf{x}}_k\|^2 + 2\|\mathbf{x}_k - \tilde{\mathbf{x}}_{k-1}\|^2 \\ &\leq 2\|\mathbf{x}_k - \tilde{\mathbf{x}}_k\|^2 + 2\|F_{k-1}(\tilde{\mathbf{x}}_{k-2}) - F_{k-1}(\tilde{\mathbf{x}}_{k-1})\|^2 \\ &\leq 2\|\mathbf{x}_k - \tilde{\mathbf{x}}_k\|^2 + \frac{1}{8} \|\tilde{\mathbf{x}}_{k-1} - \tilde{\mathbf{x}}_{k-2}\|^2. \end{aligned}$$

Rearranging this gives a lower bound on $\|\mathbf{x}_k - \tilde{\mathbf{x}}_k\|^2$, which we substitute into the result of [Lemma 3.4.10](#):

$$\begin{aligned} \|\mathbf{x}_{k+1} - \mathbf{x}_k^*\|^2 &\leq (1 - \delta_k) \|\mathbf{x}_k - \mathbf{x}_k^*\|^2 + \frac{1}{16} \|\tilde{\mathbf{x}}_k - \tilde{\mathbf{x}}_{k-1}\|^2 \\ &\quad - (1 - 2\delta_k) \left(\frac{1}{2} \|\tilde{\mathbf{x}}_k - \tilde{\mathbf{x}}_{k-1}\|^2 - \frac{1}{16} \|\tilde{\mathbf{x}}_{k-1} - \tilde{\mathbf{x}}_{k-2}\|^2 \right) \\ &= (1 - \delta_k) \|\mathbf{x}_k - \mathbf{x}_k^*\|^2 + \left[\frac{1}{16} - \frac{1}{2} (1 - 2\delta_k) \right] \|\tilde{\mathbf{x}}_k - \tilde{\mathbf{x}}_{k-1}\|^2 \\ &\quad + \frac{1}{16} (1 - 2\delta_k) \|\tilde{\mathbf{x}}_{k-1} - \tilde{\mathbf{x}}_{k-2}\|^2. \end{aligned}$$

After rearranging, we get:

$$\|\mathbf{x}_{k+1} - \mathbf{x}_k^*\|^2 + \left(\frac{7}{16} - \delta_k \right) \|\tilde{\mathbf{x}}_k - \tilde{\mathbf{x}}_{k-1}\|^2 \leq (1 - \delta_k) \|\mathbf{x}_k - \mathbf{x}_k^*\|^2 + \frac{1}{16} (1 - 2\delta_k) \|\tilde{\mathbf{x}}_{k-1} - \tilde{\mathbf{x}}_{k-2}\|^2.$$

Given our choice of $\delta_k \leq 1/8$, we have $(\frac{7}{16} - \delta_k) \geq \frac{1}{16}(1 - 2\delta_k)$. This allows us to define the Lyapunov function Φ_k and show its contraction. Let:

$$\begin{aligned} \Phi_{k+1} &:= \sqrt{\|\mathbf{x}_{k+1} - \mathbf{x}_k^*\|^2 + \frac{1}{16} \|\tilde{\mathbf{x}}_k - \tilde{\mathbf{x}}_{k-1}\|^2} \\ &\leq \sqrt{(1 - \delta_k) \left(\|\mathbf{x}_k - \mathbf{x}_k^*\|^2 + \frac{1}{16} \|\tilde{\mathbf{x}}_{k-1} - \tilde{\mathbf{x}}_{k-2}\|^2 \right)} \\ &\leq \left(1 - \frac{\delta_k}{2} \right) \sqrt{\|\mathbf{x}_k - \mathbf{x}_k^*\|^2 + \frac{1}{16} \|\tilde{\mathbf{x}}_{k-1} - \tilde{\mathbf{x}}_{k-2}\|^2} \end{aligned}$$

$$\leq \left(1 - \frac{\delta_k}{2}\right) \left(\sqrt{\|\mathbf{x}_k - \mathbf{x}_{k-1}^*\|^2 + \frac{1}{16} \|\tilde{\mathbf{x}}_{k-1} - \tilde{\mathbf{x}}_{k-2}\|^2} + \|\mathbf{x}_{k-1}^* - \mathbf{x}_k^*\| \right).$$

Here, the second inequality uses $\sqrt{1-x} \leq 1-x/2$ for $x \in [0, 1]$, and the third inequality uses the triangle inequality property $\sqrt{\|\mathbf{a} + \mathbf{b}\|^2 + \|\mathbf{c}\|^2} \leq \sqrt{\|\mathbf{a}\|^2 + \|\mathbf{c}\|^2} + \|\mathbf{b}\|$. Applying the bound on the stability of the regularized solution from [Lemma 3.4.4](#), we have:

$$\Phi_{k+1} \leq \left(1 - \frac{\delta_k}{2}\right) \left(\Phi_k + \frac{\delta_{k-1} - \delta_k}{\delta_{k-1}} R \right). \quad (3.28)$$

We now use induction to show that the iterates are bounded, i.e., $\Phi_k \leq R$. For the base case $k = 1$, using [Lemma 3.4.10](#) and [Lemma 3.4.3](#), we find:

$$\begin{aligned} \Phi_1^2 &= \|\mathbf{x}_1 - \mathbf{x}_0^*\|^2 + \frac{1}{16} \|\tilde{\mathbf{x}}_0 - \tilde{\mathbf{x}}_{-1}\|^2 \\ &\leq (1 - \delta_0) \|\mathbf{x}_0 - \mathbf{x}_0^*\|^2 - (1 - 2\delta_0) \|\mathbf{x}_0 - \tilde{\mathbf{x}}_0\|^2 + \frac{1}{8} \|\mathbf{x}_0 - \tilde{\mathbf{x}}_0\|^2 \\ &\leq \|\mathbf{x}_0 - \mathbf{x}_0^*\|^2 \leq R^2. \end{aligned}$$

Assuming $\Phi_k \leq R$ holds, and choosing $\delta_k = \frac{c}{k+8c}$ with $c \geq 2$, the inductive step follows:

$$\Phi_{k+1} \leq \left(1 - \frac{1}{k+8c}\right) \left(R + \frac{1}{k+8c} R \right) \leq R.$$

This completes the induction. We can now conclude that $\|\mathbf{x}_k - \mathbf{x}_{k-1}^*\| \leq \Phi_k \leq R$, which in turn implies that the trajectory is bounded:

$$\|\mathbf{x}_{k+1} - \mathbf{x}_0\| \leq \|\mathbf{x}_{k+1} - \mathbf{x}_k^*\| + \|\mathbf{x}_k^* - \mathbf{x}_0\| \leq R + R = 2R. \quad (3.29)$$

Main Proof Argument

Our primary goal is to establish a recurrence relation for the potential function E_k , defined as

$$E_k := \sqrt{\|F_k(\mathbf{x}_k)\|^2 + 2\|F_k(\mathbf{x}_k) - F_k(\tilde{\mathbf{x}}_{k-1})\|^2}.$$

To do this, we need to bound the difference between E_{k+1} and E_k . The key challenge lies in bounding the expression:

$$\sqrt{\|F_k(\mathbf{x}_k)\|^2 + 2\|F_k(\mathbf{x}_k) - F_k(\tilde{\mathbf{x}}_{k-1})\|^2} - \sqrt{\|F_{k-1}(\mathbf{x}_k)\|^2 + 2\|F_{k-1}(\mathbf{x}_k) - F_{k-1}(\tilde{\mathbf{x}}_{k-1})\|^2}.$$

Let's first analyze the difference in the second parts of the two norms. By the definition of F_k , we have:

$$\begin{aligned} &2\|F_k(\mathbf{x}_k) - F_k(\tilde{\mathbf{x}}_{k-1})\|^2 - 2\|F_{k-1}(\mathbf{x}_k) - F_{k-1}(\tilde{\mathbf{x}}_{k-1})\|^2 \\ &= 4\eta(\delta_k - \delta_{k-1}) \langle F(\mathbf{x}_k) - F(\tilde{\mathbf{x}}_{k-1}), \mathbf{x}_k - \tilde{\mathbf{x}}_{k-1} \rangle + 4(\delta_k^2 - \delta_{k-1}^2) \|\mathbf{x}_k - \tilde{\mathbf{x}}_{k-1}\|^2 \leq 0, \end{aligned}$$

where the inequality follows from the monotonicity of F and the fact that $\delta_{k-1} \geq \delta_k$. Furthermore, the difference between the operators themselves can be bounded:

$$\|F_k(\mathbf{x}_k) - F_{k-1}(\mathbf{x}_k)\| = (\delta_{k-1} - \delta_k) \|\mathbf{x}_k - \mathbf{x}_0\|.$$

Using these observations and applying [Lemma 3.4.9](#), we can establish an initial recurrence for E_k :

$$E_{k+1} \leq \left(1 - \frac{\delta}{4}\right) (E_k + (\delta_{k-1} - \delta_k) \|\mathbf{x}_k - \mathbf{x}_0\|).$$

Now, we are ready to incorporate the trajectory boundedness result from the previous section. We have shown that $\|\mathbf{x}_{k+1} - \mathbf{x}_0\| \leq 2R$. Substituting this into the recurrence gives:

$$E_{k+1} \leq \left(1 - \frac{\delta}{4}\right) (E_k + 2(\delta_{k-1} - \delta_k) R). \quad (3.30)$$

The initial value E_0 can be bounded as:

$$E_0 = \sqrt{\|F_0(\mathbf{x}_0)\|^2 + 2\|F_0(\mathbf{x}_0) - F_0(\tilde{\mathbf{x}}_{-1})\|^2} \leq \eta LR = \frac{1}{8}R.$$

By setting the regularization parameter schedule to $\delta_k = \frac{8}{k+64}$ and applying [Lemma 3.2.1](#), we solve the recurrence to get:

$$E_{k+1} \leq \frac{16}{k+64}R.$$

Finally, from the definition of F_k , we can unwind the bound on E_k to get a bound on the norm of the true operator $F(\mathbf{x}_k)$:

$$\|F(\mathbf{x}_k)\| \leq \frac{E_k + 2\delta_{k-1}R}{\eta} \leq \frac{256LR}{k+63} = O\left(\frac{LR}{k}\right). \quad (3.31)$$

This concludes the proof.

3.5 Conclusion

We have demonstrated that Tikhonov regularization provides a unified framework for analyzing and designing first-order methods for both fixed-point and monotone inclusion problems. By applying standard algorithms to a sequence of regularized operators $F_k(\cdot) = F(\cdot) + \delta_k(\cdot - \mathbf{x}_0)$ with diminishing δ_k , we recover optimal convergence rates without modifying the algorithmic structure itself. This operator-centric approach yields three concrete contributions: a simplified analysis of Halpern iteration achieving $O(1/k)$ rate for fixed points, the first discrete-time convergent forward-step method for monotone operators with $O(1/\sqrt{k})$ rate, and accelerated variants of EG and OG achieving the optimal $O(1/k)$ rate.

Our framework reveals that acceleration can emerge from regularization alone rather than from algorithmic modifications—a conceptual shift from recent approaches that embed anchoring directly into update rules. The modularity of our approach enables natural extensions: the viscosity approximation ($G(\cdot)$ instead of $\mathbf{x}_0$) and episodic schemes with multiple iterations per regularization level. A natural follow-up direction is extending this framework to stochastic settings where operator regularization may offer similar benefits.

Chapter 4

Affine-equivariant Mean Estimation

4.1 Introduction

The design of estimators, robust to gross outliers in the data, has remained a mainstay of statistical research for the past 60 years [62, 87, 30]. A natural consideration of such a procedure is its stability under natural transformations of the data. Focusing on *affine* transformations in particular and the fundamental estimation problem of mean estimation, several early results constructed estimators *equivariant* to affine transformations of the data and studied the degree of their resilience to outliers [125, 93, 117, 37, 102, 111, 112, 60, 81, 38]. Unfortunately, while these estimators guarantee *finite* error, a sharp *quantitative* understanding of their performance remains elusive. On the other hand, recent estimators [90, 89] with quantitatively *optimal* performance do not satisfy this natural stability property. As a consequence, their use pre-supposes an appropriate *standardization* of the input data which is often challenging to estimate in heavy-tailed settings. In contrast, affine-equivariant estimators, by nature, offer stable performance even when such a standardization is not available.

Formally, we study the robust mean estimation problem where we are given n independent and identically distributed (i.i.d.) data points $\widetilde{\mathbf{X}} = \{\widetilde{X}_i\}_{i=1}^n \subset \mathbb{R}^d$ drawn from a distribution, D , with mean μ and variance Σ along with a target failure probability δ . Furthermore, an arbitrarily chosen η fraction of the data points may be modified in a possibly adversarial way to obtain a corrupted dataset, $\mathbf{X} = \{X_i\}_{i=1}^n$ ¹, which is then presented to the estimator. The goal, now, is to design an estimator $\widehat{\mu}$ with the smallest $r_{\delta,\eta}$ satisfying:

$$\mathbb{P} \{ \|\widehat{\mu}(\mathbf{X}) - \mu\|_* \leq r_{\delta,\eta} \} \geq 1 - \delta$$

where $\|\cdot\|_*$ is a target norm of interest which will be specified subsequently.

We start by overviewing the Euclidean setting where the error is measured in the *Euclidean* norm and which has been the subject of much recent attention leading to optimal

¹See [Appendix B.2](#) for a formal definition of the corruption model.

quantitative characterizations. Here, one aims to design an estimator minimizing $r_{\delta,\eta}$ satisfying:

$$\mathbb{P} \{ \|\hat{\mu}(\mathbf{X}) - \mu\| \leq r_{\delta,\eta} \} \geq 1 - \delta \text{ where } \|x\| = \sqrt{x^\top x}.$$

That is, one aims to find an estimator with the smallest possible *Euclidean* confidence interval as a function of n , d , and δ . Surprisingly, this fundamental question was only resolved recently starting with the pioneering work of Lugosi and Mendelson [90], who established a sharp characterization when $\eta = 0$, and subsequent work by Depersin and Lecué [29] which showed that a variant of this estimator also applies when $\eta \neq 0$ yielding the following rate:

$$r_{\delta,\eta} = O \left(\sqrt{\|\Sigma\|} \cdot \left(\sqrt{\frac{r(\Sigma) + \log(1/\delta)}{n}} + \sqrt{\eta} \right) \right) \text{ where } r(\Sigma) := \frac{\text{Tr } \Sigma}{\|\Sigma\|}.$$

When $\eta = 0$, this rate, referred to as the *sub-Gaussian* rate, is known to be *optimal* for Gaussians and hence, cannot be improved in general. This striking result holds despite making no other assumptions on the data distribution. Note that when $\Sigma \preceq I$, the rate simplifies to:

$$r_{\delta,\eta} = O \left(\sqrt{\frac{d + \log(1/\delta)}{n}} + \sqrt{\eta} \right). \quad (4.1)$$

However, as remarked previously, the output of these estimators is sensitive to the precise scaling of the data, which is often challenging to estimate. This sensitivity motivates the main setting of the paper, where the error of an estimator is measured in the *Mahalanobis* norm, a natural affine-invariant metric, defined below:

$$\mathbb{P} \{ \|\hat{\mu}(\mathbf{X}) - \mu\|_\Sigma \leq r_{\delta,\eta} \} \geq 1 - \delta \text{ where } \|x\|_\Sigma = \sqrt{x^\top \Sigma^{-1} x}. \quad (\text{MAH})$$

Equivalently, the Mahalanobis distance may be viewed as measuring the *Euclidean* distance under the linear transformation that renders the distribution isotropic². Hence, for affine-equivariant estimators, our results characterize the optimal achievable *Euclidean* error on distributions with $\Sigma \preceq I$ ³. Lastly, in addition to the aforementioned benefits, a bound on the Mahalanobis norm straightforwardly yields a bound on the *Euclidean* error as:

$$\forall x \in \mathbb{R}^d : \|x\| \leq \sqrt{\|\Sigma\|} \cdot \|x\|_\Sigma.$$

Driven by these motivations, one may ask whether bounds analogous to Eq. (4.1)⁴ are possible for the Mahalanobis norm. Under more stringent assumptions on the data distribution, often allowing the estimation of a multiplicative approximation of Σ , a series of recent estimators [88, 17, 27, 40, 16] indeed achieve the rates in Eq. (4.1). In this context, we ask

²Our results also apply to non-invertible Σ with the *psuedo-inverse* $\Sigma^\dagger$ taking the place of Σ^{-1} and further restricting the estimator to achieve *zero* error along the null space of Σ . See Section 4.5.

³Note that without any restriction on Σ , no finite error bound is possible.

⁴Note that dimension independent error bounds are not possible in this setting as any distribution with non-denerate covariance may be transformed to one with identity covariance by an isotropic transformation.

whether such rates are obtainable *even when* Σ is not estimable from data. In other words, what is the optimal estimation error in this setting?

Unfortunately, we show that [Eq. \(4.1\)](#) is *not* obtainable in the general setting, and that a strict degradation by $\sqrt{d}$ is required. Our first contribution is a novel estimator which makes no assumptions on the distribution beyond the existence of a variance and obtains the following guarantees which we subsequently prove are near-optimal.

Theorem 4.1.1. *There exist absolute constants $C_1, C_2 > 0$ such that the following hold. Let $n, d \in \mathbb{N}, \delta \in (0, 1)$ and $\eta \in [0, 1/(6d)]$. Suppose $\widetilde{\mathbf{X}} = \{\tilde{X}_i\}_{i=1}^n$ are generated i.i.d. from a distribution D with mean μ and covariance Σ . Then, there exists an affine-equivariant estimator, $\widehat{\mu}$, which when given any η -corrupted version of $\widetilde{\mathbf{X}}$, $\mathbf{X}$, satisfies:*

$$\|\widehat{\mu}(\mathbf{X}) - \mu\|_{\Sigma} \leq C_1 \left(\sqrt{\frac{d \log(1/\delta)}{n}} + \sqrt{d\eta} \right)$$

with probability at least $1 - \delta$ over $\widetilde{\mathbf{X}}$ when $n \geq C_2 d \log(1/\delta)$.

Furthermore, we exhibit lower bounds establishing that the above rate and restrictions on η are essentially tight. Our lower bounds are proved separately for the heavy-tailed and adversarial settings and hence, our upper bound which holds for *both* settings simultaneously is optimal. We start with the heavy-tailed setting.

Theorem 4.1.2. *There exist absolute constants, $C_1, C_2, c > 0$ such that the following holds. Let $n, d \in \mathbb{N}$ and $\delta \in (0, 1)$ be such that $n \geq C_1 d \log(1/\delta)$ and $\log(1/\delta) \geq C_2 \log(2d)$. Then, there exists a family of distributions $\mathcal{D}$ such that for any estimator $\widehat{\mu}$:*

$$\max_{D \in \mathcal{D}} \mathbb{P}_{\mathbf{X} \sim D^n} \left\{ \|\widehat{\mu}(\mathbf{X}) - \mu(D)\|_{\Sigma(D)} \geq c \sqrt{\frac{d \log(1/\delta)}{n \log(d)}} \right\} \geq \delta.$$

Next, we present our lower bounds for the adversarial contamination model. Furthermore, our lower bounds hold for the weaker *Huber* contamination model where an adversary is only allowed to *add* corrupted points to the dataset as opposed to corrupting existing points⁵. In the first theorem, we show that the error is unbounded if the corruption fraction exceeds $1/(d+1)$ for estimators that are eventually even *approximately* consistent.

Theorem 4.1.3. *For any $d > 3$ and $r > 0$, there exists a family of distributions, $\mathcal{D}$, and a distribution $D_0 \in \mathcal{D}$ such that for any $D \in \mathcal{D}$, there exists distribution P satisfying:*

$$D_0 = \frac{d}{d+1} D + \frac{1}{d+1} P.$$

Furthermore, we have for any estimator, $\widehat{\mu}$, and any $n \in \mathbb{N}$:

$$\max_{D \in \mathcal{D}} \mathbb{P}_{\mathbf{X} \sim D_0^n} \left\{ \|\widehat{\mu}(\mathbf{X}) - \mu(D)\|_{\Sigma(D)} \geq r \right\} \geq \frac{1}{d+1}.$$

⁵A formal comparison to the stronger corruption model of [Theorem 4.1.1](#) is included in [Appendix B.2](#).

In the above theorem and the one following it, the distribution P is to be interpreted as the distribution over corruptions that an adversary adds to the sample. Next, we show that the dependence on the corruption fraction when $\eta < 1/(d+1)$ in [Theorem 4.1.1](#) is tight.

Theorem 4.1.4. *For any $d > 3$ and $\eta < 1/(d+1)$, there exists a family of distributions, $\mathcal{D}$, and a distribution $D_0 \in \mathcal{D}$ such that for any $D \in \mathcal{D}$, there exists distribution P satisfying:*

$$D_0 = (1 - \eta)D + \eta P.$$

Furthermore, we have for any estimator, $\hat{\mu}$, and any $n \in \mathbb{N}$:

$$\max_{D \in \mathcal{D}} \mathbb{P}_{\mathbf{X} \sim D_0^n} \left\{ \|\hat{\mu}(\mathbf{X}) - \mu(D)\|_{\Sigma(D)} \geq \frac{1}{2} \sqrt{\frac{d\eta}{1-d\eta}} \right\} \geq \frac{1}{d+1}.$$

Taken together, our bounds imply a marked departure from the Euclidean setting. The breakdown point and the dependence of the recovery guarantees on the failure probability and corruption factor *all* decay by a factor of d . Consider the following concrete instantiation of the two bounds for the particular setting of $\delta = e^{-d}$:

$$\text{Affine-equivariant bound: } \frac{d}{\sqrt{n}} \qquad \text{Euclidean Bound: } \sqrt{\frac{d}{n}}.$$

Intuitively, in the Euclidean setting, the optimal recovery guarantees essentially match the *Gaussian* rate. However, in the affine-equivariant setting, a significant cost is incurred when weaker assumptions are placed on the data distribution.

Our estimator utilizes the classical median-of-means framework for building robust estimators [100]. The key ingredient of our approach is a novel high-dimensional median, inspired by the well-known Tukey median [125] and the Stahel-Donoho estimator [117, 37] and may be of independent interest. We aim to find a point whose distance to the mean along any direction is small with respect to the variance along that direction. However, the main difficulty in analyzing the estimator is establishing that such a point always exists. We define an appropriate proxy for the variance which guarantees the existence of such a median while allowing for optimal recovery guarantees. Interestingly, our analysis, similar to that of the Tukey Median, relies strongly on Helly's Theorem, a central result in convex geometry.

The key challenge in proving our lower bounds is establishing the correct dependence on the failure probability in the heavy-tailed setting. Our lower bound construction uses a family of distributions with different covariances in the standard Bayesian estimation-to-testing framework for proving minimax lower bounds (see, for example, [127]). In the typical heavy-tailed setting, where the terms depending on the failure probability and the dimension are decoupled, two lower bounds are established separately for each setting and are subsequently combined to obtain the final bound. However, in our case, these two elements are intimately coupled making the application of standard techniques challenging. To overcome this, we perform an explicit analysis of the posterior over the set of candidate

distributions once the data points have been generated, but only for a carefully chosen set of observations. We show that when such samples are obtained, the posterior is well-spread and that any proposed estimate performs poorly on at least some distributions in the support of the posterior. Here, the differences in the covariance matrices across the distributions in the family play a critical role and the sensitivity of the Mahalanobis norm to such differences yields our lower bound.

Notation

We use $\mathbb{N}$ and $\mathbb{R}$ to denote the set of natural and real numbers respectively and for $n \in \mathbb{N}$, $[n] := \{1, 2, \dots, n\}$. For $d \in \mathbb{N}$, $\mathbb{R}^d$ refers to the standard d -dimensional Euclidean vector space with $\mathbf{1}$, $\mathbf{0}$, and $\{e_i\}_{i=1}^d$ denoting the all 1s vector, the all 0s vector, and the standard basis vectors respectively. $\mathbb{S}^{d-1}$ denotes the unit sphere in d dimensions. For $M_1, M_2 \in \mathbb{R}^{d \times d}$, $M_1 \succcurlyeq M_2$ (and $M_1 \succ M_2$) indicates that M_1 is larger (and strictly larger) than M_2 in the standard PSD ordering. For $x \in \mathbb{R}^d$ and $M \in \mathbb{R}^{d \times d}$, $\|x\|$ and $\|M\|$ refer to the Euclidean and spectral norm of x and M respectively, and $\text{Tr } M$ denotes the trace of M . Furthermore, for $\Sigma \succ 0$, $\|x\|_\Sigma := \sqrt{x^\top \Sigma^{-1} x}$. For a distribution D over $\mathbb{R}^d$, we use $\mu(D)$ and $\Sigma(D)$ to denote its mean and covariance, and for $n \in \mathbb{N}$, D^n denotes the product distribution formed by taking n i.i.d. draws from D . We will typically use $\mathbf{X}$ and $\mathbf{Y}$ to denote multisets of points in $\mathbb{R}^d$ and with a slight abuse of notation write $\mathbf{X}, \mathbf{Y} \subset \mathbb{R}^d$. For a point set $\mathbf{Y} = \{y_1, \dots, y_n\} \subset \mathbb{R}$, we define:

$$\mu(\mathbf{Y}) := \frac{1}{n} \sum_{i=1}^n y_i, \quad \sigma_1(\mathbf{Y}) := \frac{1}{n} \sum_{i=1}^n |y_i - \mu(\mathbf{Y})|.$$

Furthermore, $\text{Med}(\mathbf{Y})$ denotes its median and $|\mathbf{Y}|$ denotes its cardinality. Additionally, for a (possibly infinite) subset $T \subset \mathbb{R}^d$, we use $\mu(T)$ to denote the average of the uniform distribution over T and $\text{conv}(T)$ to denote its convex hull. Additionally, for a function f , we use $f(T)$ to denote the set T with f applied to each element. For a proposition A , $\mathbf{1}\{A\}$ is the indicator function which is 1 if the proposition is true and 0 otherwise. Throughout this paper, we adopt the convention $\frac{0}{0} := 0$ and $\frac{p}{0} := \infty$ for any $p > 0$. The use of this convention will be clear from context.

In the remainder of the paper, we first overview related work in [Section 4.2](#). We then provide some insight into the failure modes of prior work in [Section 4.3](#). Next, we formally present our estimator along with a proof of its existence in [Section 4.4](#). Finally, we establish upper bounds on its performance in [Section 4.5](#) and present nearly-matching lower bounds in [Section 4.6](#).

4.2 Related Work

Here, we review prior work closely related to our own.

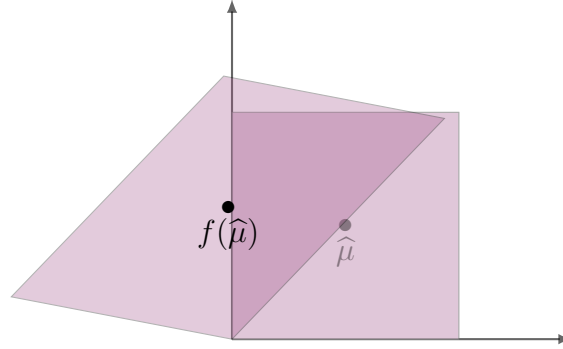


Figure 4.1: An illustration of the affine-equivariance requirement. Here, the result of an estimator $\hat{\mu}$ run on the transformation of the square to the tilted parallelogram is required to coincide with the transformation of the estimate obtained when run on the square itself.

Affine-equivariant Estimators

There is a rich line of work in theoretical statistics analyzing the stability and robustness of affine-equivariant estimators. Formally, these estimators commute with affine transformations; i.e., for any non-singular affine map $f : \mathbb{R}^d \rightarrow \mathbb{R}^d$, an affine-equivariant estimator $\hat{\mu}$ satisfies for any $\mathbf{X} \subset \mathbb{R}^d$, $\hat{\mu}(f(\mathbf{X})) = f(\hat{\mu}(\mathbf{X}))$ (see Figure 4.1). This intuitive property ensures that the outcomes of an inferential procedure utilizing the estimator do not depend on the particular choice of scaling transformations applied to the data. The empirical mean is a prominent example of such an estimator but its performance is severely degraded when faced with heavy-tailed or adversarial data. A significant advance toward the construction of more robust estimators was made by Tukey [125], who proposed the Tukey Median as an alternative. The starting point of this estimator is the concept of a one-dimensional *depth function*. Such a function measures how *close to the center* a point is with respect to a set of points. For instance, the depth function corresponding to the Tukey median is defined as follows for $\mathbf{Y} = \{y_i\}_{i=1}^n \subset \mathbb{R}$:

$$D_{\tau}^1(y; \mathbf{Y}) = \frac{1}{n} \min(|*| \{i : y_i \geq y\}, |*| \{i : y_i \leq y\}).$$

The Tukey Median of a set of points $\mathbf{X} = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$ is a point that maximizes its minimum depth over all one-dimensional projections:

$$\hat{\mu}_{\tau}(\mathbf{X}) = \arg \max_{x \in \mathbb{R}^d} D_{\tau}^d(x; \mathbf{X}) \text{ where } D_{\tau}^d(x; \mathbf{X}) = \min_{u \in \mathbb{S}^{d-1}} D_{\tau}^1(\langle u, x \rangle; \{\langle u, x_i \rangle\}_{i=1}^n).$$

When $d = 1$, this reduces to the standard Median which is known to be robust (and scale-equivariant). The Tukey Median naturally generalizes this simple estimator to higher dimensions. Since then, the breakdown properties of the Tukey Median and other affine-equivariant estimators have been closely investigated by Maronna [93] and Huber [61]. These works concluded that the *breakdown point* [36] of these well-known affine-equivariant estimators is at

most $1/(d+1)$ ⁶. This somewhat disappointing discovery led to the search for estimators with improved breakdown properties. One of the first such approaches was the Stahel-Donoho estimator, independently discovered by Stahel [117] and Donoho [37]. Here, one utilizes an alternative notion of *outlyingness*:

$$D_{\text{SD}}^1(y; \mathbf{Y}) = \frac{|y - \text{Med}(\mathbf{Y})|}{\text{MAD}(\mathbf{Y})} \text{ where } \text{MAD}(\mathbf{Y}) = \text{Med}(\{|y_i - \text{Med}(\mathbf{Y})|\}_{i=1}^n).$$

Now, the Stahel-Donoho estimate is a point with *minimum* outlyingness over all one-dimensional projections of the dataset:

$$\hat{\mu}_{\text{SD}}(\mathbf{X}) = \arg \min_{x \in \mathbb{R}^d} D_{\text{SD}}^d(x; \mathbf{X}) \text{ where } D_{\text{SD}}^d(x; \mathbf{X}) = \max_{u \in \mathbb{S}^{d-1}} D_{\text{SD}}^1(\langle u, x \rangle; \{\langle u, x_i \rangle\}_{i=1}^n).$$

This estimator is known to have a breakdown point approaching $1/2$. However, all these approaches have significant drawbacks:

1. There exist no quantitative bounds on the performance of these estimators without additional distributional assumptions allowing the estimation of Σ (see [Appendix B.4](#)).
2. Furthermore, the breakdown point depends on the non-degeneracy of the dataset with error bounds growing arbitrarily large as the dataset approaches degeneracy ([Proposition B.3.5](#)).

Since the Stahel-Donoho estimator, numerous alternative approaches with differing notions of depth have been proposed: these include estimators based on the simplicial volume [102], S-estimators [111], the minimum volume ellipsoid [112], and the simplicial depth [82]. In addition, the robustness [93, 137, 26, 39, 86, 38, 81, 84, 83] and consistency properties [142, 143] of these estimators have been studied. However, despite this interest, there exist no *quantitative* accuracy guarantees in terms of the number of data points n , dimension d , failure probability δ , and corruption fraction η for these estimators in the general setting.

Computationally Efficient Adversarially Robust Estimation

The first *computationally* efficient estimator (in *Euclidean* norm) with near-optimal guarantees was proposed in a breakthrough result of Diakonikolas, Kamath, Kane, Li, Moitra, and Stewart [32]. This estimator is computable in polynomial time with polynomial sample complexity. The statistical and computational complexity has been substantially improved in follow-up works [20, 35] resulting in estimators with near-optimal statistical and computational performance. We direct the interested reader to the excellent survey [30] for more applications of these ideas.

⁶See [Appendix B.3](#) for a discussion on the classical breakdown point of affine-equivariant estimators.

Heavy-tailed Estimation

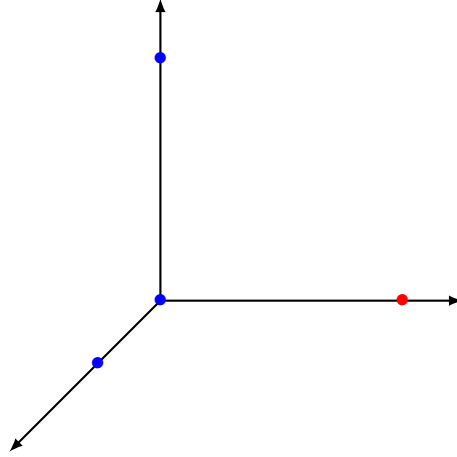
An alternative statistical model for outliers is the heavy-tailed corruption model. In this setting, minimal assumptions are made about the data generating distribution (for instance, the covariance of the data-generating distribution exists as opposed to stronger ones such as Gaussianity) and hence, outliers occur naturally as part of the data. This is in contrast to an adversary maliciously corrupting the datapoints in the adversarial setting. Here, estimators such as the empirical mean remain *consistent* but suffer from poor *statistical* performance and the emphasis is on designing estimators which avoid this degradation. In one dimension, optimal estimators based on the median-of-means framework have been known (and independently discovered) in a series of classical works [100, 65, 2]. Finding a corresponding *high-dimensional* estimator remained open till the pioneering work of Lugosi and Mendelson [90] whose estimator achieves the optimal sub-Gaussian rate in the Euclidean setting. A computationally efficient estimator by Hopkins [56] with the same guarantees followed shortly after along with alternative approaches [89]. Since then, these ideas have been improved and extended to numerous other settings leading to estimators with strong statistical and computational performance [21, 28, 75]. Some recent works have also focused on the strong connections between the heavy-tailed and adversarially robust settings yielding estimators simultaneously robust to both corruption models [28, 89, 58, 31]. Another line of work has also incorporated privacy guarantees into these estimators [85, 57, 70, 59].

In this context, we compare our results to three recently developed methods: the work of Depersin and Lecue [27], the setting considered by Duchi, Haque and Kuditipudi [40] and Brown, Hopkins, and Smith [16] which in turn build upon approaches by Brown, Gaboardi, Smith, Ullman, and Zakynthinou [17], and the recent result of Lugosi and Mendelson [88]. Depersin and Lecue [27] consider the Stahel-Donoho estimator and show that it achieves sub-Gaussian statistical performance. However, their approach requires strong assumptions on the data-generating process where a multiplicative approximation to Σ may be estimated from the data samples. On the other hand, in [17], the authors construct affine-equivariant estimators with sub-Gaussian error and strong privacy guarantees with subsequent work [40, 16] achieving computational efficiency. However, these works also require the existence of higher order-moments (beyond merely, the existence of the covariance matrix) of the distribution. Finally, sub-Gaussian estimators with direction-dependent accuracy are developed in [88]. However, these bounds scale with the expected *Euclidean* deviation of a sample from its mean which when evaluated in the *Mahalanobis* norm may be arbitrarily large. In addition, this estimator also requires additional assumptions for provable guarantees. A more comprehensive overview of these recent works is provided in [Appendix B.4](#).

4.3 Intuition

In this section, we provide some intuition for our estimator. We analyze the performance of two prominent affine-equivariant estimators: the Tukey Median and the Stahel-Donoho

Figure 4.2: Illustration of hard distribution. The red dot on e_1 denotes higher probability.



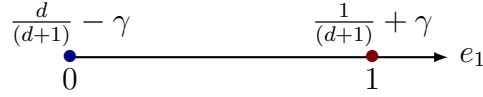
estimator. We consider a simple setting where both these estimators perform poorly. While these medians may be used with a median-of-means pre-processing step, we analyze the extreme setting where no bucketing takes place and the estimator is evaluated directly on the distribution. As we will see, the Tukey median incurs error $\sqrt{d}$ while our median when run in this setting incurs constant error.

Our hard example will essentially be the simple uniform distribution over the standard basis vectors and the origin $\{e_i\}_{i=1}^d \cup \{\mathbf{0}\}$. However, we will assume one of the standard basis vectors (say e_1) is mildly more likely to be observed. Formally, the distribution is defined for parameter γ as follows:

$$\mathbb{P}_{X \sim D_\gamma} \{X = x\} = \begin{cases} \frac{1}{d+1} + \gamma & \text{if } x = e_1 \\ \frac{1}{d+1} - \frac{\gamma}{d} & \text{otherwise} \end{cases}.$$

The example is illustrated in 3 dimensions in Figure 4.2.

We also assume for the sake of simplicity that the estimators are run directly on the distribution itself with distributional analogues of the corresponding depth and outlyingness functions as opposed to samples from the distribution. While these estimators are typically biased even when the data remains *uncorrupted*, the standard Median-of-Means pre-processing step can be used to trade-off bias for larger failure probability. This example concerns the extreme setting of the failure probability where no bucketing is implemented. We start with the Tukey median and establish that e_1 is the unique point with largest Tukey depth. Notice that the depth of e_1 is $1/(d+1) + \gamma$. Let the support of the distribution be S . For any point not in the convex hull of S , the separating hyperplane theorem ensures that they have Tukey depth 0. Now, consider the case where x belongs to the convex hull and $x \neq e_1$. We consider the two possibilities $x \in T$ where $T = S \setminus \{e_1\}$ and $x \notin T$ separately. First, let $x \in T$ and consider the vector $v = x - \mathbf{1}\{x = \mathbf{0}\}$. We have $\langle x, v \rangle > \langle y, v \rangle$ for all

Figure 4.3: One dimensional projection onto e_1 .


$y \in S \setminus x$. Hence, the depth of x is at most $1/(d+1) - \frac{\gamma}{d}$. Secondly, consider the alternative case when $x \notin T$ (but lies in its convex hull). We must have:

$$x = \sum_{y \in S} w_y y \text{ where } \sum_{y \in S} w_y = 1 \text{ and } w_y \geq 0.$$

Furthermore, since $x \notin S$ and $x \neq e_1$, there exists $y \in T$ with $0 < w_y < 1$. Consider such a vector y and the vector $v = y - \mathbf{1}\{y = \mathbf{0}\}\mathbf{1}$. We now get:

$$\forall z \in S \setminus y : \langle v, y \rangle > \langle v, x \rangle > \langle v, z \rangle.$$

Since $y \neq e_1$, the depth of x is also at most $1/(d+1) - \frac{\gamma}{d}$. The previous two cases establish that e_1 is the unique point of maximum Tukey depth. Unfortunately, the error of e_1 is rather large. Consider the one-dimensional projection of the distribution, denoted D_γ^1 , onto e_1 :

$$\mathbb{P}_{Y \sim D_\gamma^1} \{Y = y\} = \begin{cases} \frac{1}{d+1} + \gamma & \text{if } y = 1 \\ \frac{d}{d+1} - \gamma & \text{if } y = 0 \end{cases}.$$

By considering the error along e_1 using [Lemma B.1.5](#), we get for $\mu_\gamma = \mu(D_\gamma)$ for $\gamma \leq 1/(10d)$:

$$\|e_1 - \mu_\gamma\|_{\Sigma(D_\gamma)} \geq \frac{d/(d+1) - \gamma}{\sqrt{(1/(d+1) + \gamma)(d/(d+1) - \gamma)}} = \sqrt{\frac{d/(d+1) - \gamma}{1/(d+1) + \gamma}} \geq \frac{\sqrt{d}}{2}.$$

As we will see later, this error is larger than optimal by a $\sqrt{d}$ factor. However, notably, the Tukey median can be shown to exist for *any* set of data points. The main drawback of the Tukey median is that it remains insensitive to the *variance* along different directions. As illustrated in [Figure 4.3](#), the true mean (along e_1) lies at $\frac{1}{d+1} + \gamma$ while the Tukey estimate projects to 1. The variance along e_1 is also at most $\mathbb{E}[\langle e_1, X \rangle^2] = \frac{1}{d+1} + \gamma$. Therefore, incorporating variance information into the estimator can help mitigate some of this degradation.

The Stahel-Donoho estimator attempts to incorporate such variance information. However, analyzing the estimator requires non-degeneracy assumptions on the data and even after doing so, these do not provide any *quantifiable* bounds on its performance ([Proposition B.3.5](#)). For our example, the Stahel-Donoho estimator is not even *defined*. Consider the projection of the distribution onto the standard basis vectors e_i and the all-ones $\mathbf{1}$ direction. From the one-dimensional projections in [Figure 4.4](#), we have the following straightforward

Figure 4.4: One dimensional projections onto e_i for $i \neq 1$ and $\mathbf{1}$.


observations where D_v^1 denotes the projection of D_γ onto v :

$$\text{Med}(D_v^1) = \begin{cases} 0 & \text{if } v = e_i \\ 1 & \text{if } v = \mathbf{1} \end{cases} \text{ and } \forall v \in \{e_i\}_{i=1}^d \cup \{\mathbf{1}\} : \text{MAD}(D_v^1) = 0.$$

Consequently, for an estimate x to have finite Stahel-Donoho outlyingness, it must satisfy $\langle x, e_i \rangle = 0$ for all i and $\langle x, \mathbf{1} \rangle = 1$ which is a contradiction.

Our previous discussion shows that the Tukey median and the Stahel-Donoho estimator fail in two complementary ways. The Tukey median is always defined for any set of data points but its failure to incorporate directional variances into its estimation procedure leads to large error. On the other hand, the variance estimates used in the Stahel-Donoho estimator may not allow for a well-defined estimate in certain settings and even when it is defined, existing analyses do not yield quantitative bounds on its performance. As we will see in [Sections 4.4](#) and [4.5](#), our estimator simultaneously addresses the shortcomings of both the Tukey median and the Stahel-Donoho estimator. Our median estimator accounts for directional variances like the Stahel-Donoho median but at the same time, is defined for *any* collection of data points like the Tukey median.

4.4 Algorithm

Both the Tukey median and Stahel-Donoho median aim to find the estimate with the optimal worst-direction outlyingness. For each direction, the Tukey median employs the empirical quantile to measure outlyingness, while the Stahel-Donoho median uses the ratio of deviation from the median to the median absolute deviation. In this section, we propose a similar formulation for a high-dimensional median, but with a novel notion of outlyingness.

Consider any one-dimensional multi-set of points $\mathbf{Y}$. For a subset $S \subset [n]$ we define:

$$\mu_S(\mathbf{Y}) := \frac{1}{|S|} \sum_{i \in S} y_i \text{ and } \sigma_{1,S}(\mathbf{Y}) := \frac{1}{|S|} \sum_{i \in S} |y_i - \mu_S(\mathbf{Y})|,$$

We propose the following measure of outlyingness:

$$D_{\text{Ours}}^1(y; \mathbf{Y}) := \max_{S \subset [n], |S| \geq (1-\nu)n} \frac{|y - \mu_S(\mathbf{Y})|}{\sigma_{1,S}(\mathbf{Y})},$$

where $\nu = c/d$ for some $c > 0$. Analogous to the Tukey and Stahel-Donoho medians, our median can be defined as:

$$\hat{\mu}_{\text{Ours}}(\mathbf{X}) = \arg \min_{x \in \mathbb{R}^d} D_{\text{Ours}}^d(x; \mathbf{X}) \text{ where } D_{\text{Ours}}^d(x; \mathbf{X}) := \sup_{v \in \mathbb{S}^{d-1}} D_{\text{Ours}}^1(\langle x, v \rangle; \{\langle x_i, v \rangle\}_{i=1}^n)$$

We demonstrate that with these definitions, our estimate *always* exists for *any* point set and exhibits a *constant* outlyingness. This median is defined in [Algorithm 1](#) and the proof of its existence is provided in [Theorem 4.4.1](#). The proof relies on Helly's Theorem ([Theorem B.1.1](#)), a fundamental result in convex geometry, also utilized to analyze the Tukey Median.

Algorithm 1 High-dimensional Median

- 1: **Input:** Point set $\mathbf{X} = \{x_i\}_{i=1}^k \subset \mathbb{R}^d$
- 2: Let $\nu = 1/(3d)$ and $\mathcal{S} = \{S \subset [k] : |S| \geq (1 - \nu)k\}$
- 3: Define for all $v \in \mathbb{S}^{d-1}, S \in \mathcal{S}$:

$$\mu_{v,S} = \mu_S \left(\{\langle v, x_i \rangle\}_{i=1}^k \right) \quad \sigma_{v,S} = \sigma_{1,S} \left(\{\langle v, x_i \rangle\}_{i=1}^k \right)$$

- 4: Define convex compact sets for all $v \in \mathbb{S}^{d-1}, S \in \mathcal{S}$:

$$T_{v,S} = \{x \in \mathbb{R}^d : |\langle x, v \rangle - \mu_{v,S}| \leq 2\sigma_{v,S}\} \cap \text{conv}(\mathbf{X})$$

- 5: Let $T = \bigcap_{v \in \mathbb{S}^{d-1}, S \in \mathcal{S}} T_{v,S}$
 - 6: **Return:** $\mu(T)$
-

For a point set $\mathbf{X} = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$, let $\hat{\mu}(\mathbf{X})$ denote the output of [Algorithm 1](#). The main result of this section establishes the existence of this estimator.

Theorem 4.4.1. *For any $k \in \mathbb{N}$ and $\mathbf{X} = \{x_i\}_{i=1}^k \subset \mathbb{R}^d$, $\hat{\mu}(\mathbf{X})$ is well defined and affine-equivariant.*

Proof. We tackle the two claims of the theorem in turn.

Existence of $\hat{\mu}$: We first show that T is non-empty, convex, and compact implying the first claim. Note that T is the intersection of compact convex sets and is hence, convex and compact. To establish the non-emptiness of T , an application of Helly's Theorem ([Theorem B.1.1](#)) allows us to restrict to intersections of at most $d + 1$ of the sets $T_{v,S}$. Consider any $d + 1$ sized collection $H = \{v_j, S_j\}_{j \in [d+1]}$ with $v_j \in \mathbb{S}^{d-1}$ and $S_j \in \mathcal{S}$ ([Algorithm 1](#)). We have:

$$R := \bigcap_{j \in [d+1]} S_j, |R| \geq (1 - (d + 1)\nu)k \geq \frac{k}{2}.$$

Defining:

$$\mu_R = \frac{1}{|R|} \cdot \sum_{i \in R} x_i,$$

we will show that μ_R lies in $\cap_{(v,S) \in H} T_{v,S}$. For any $(v, S) \in H$, we have:

$$|\langle \mu_R, v \rangle - \mu_{v,S}| = |*| \frac{1}{|R|} \sum_{i \in R} (\langle x_i, v \rangle - \mu_{v,S}) \leq \frac{1}{|R|} \cdot \sum_{i \in R} |*| \langle x_i, v \rangle - \mu_{v,S} \leq \frac{|S|}{|R|} \sigma_{v,S} \leq 2\sigma_{v,S}.$$

An application of Helly's theorem now establishes that T is non-empty proving the claim.

Affine-equivariance of $\hat{\mu}$: Next, we establish that $\hat{\mu}$ is affine-equivariant. Let $\mathbf{X} = \{x_i\}_{i=1}^k \subset \mathbb{R}^d$ and $f(x) = Ax + b$ with $A \in \mathbb{R}^{d \times d}$, $b \in \mathbb{R}^d$ and A be non-singular. Hence, f is an invertible affine transformation. Furthermore, let $\mathbf{X}' = f(\mathbf{X}) = \{x'_i = f(x_i)\}_{i=1}^k$ and T be the set obtained in [Algorithm 1](#) on input $\mathbf{X}$ and T' be the corresponding set on $\mathbf{X}'$. We will show $T' = f(T)$ proving the second claim.

First, let $x \in T$ and we prove $f(x) \in T'$. Observe for any $v \in \mathbb{S}^{d-1}$, $i \in [k]$:

$$\begin{aligned} \langle v, x'_i \rangle &= \langle v, f(x_i) \rangle = \langle v, Ax_i \rangle + \langle v, b \rangle \\ &= \langle A^\top v, x_i \rangle + \langle v, b \rangle = \|A^\top v\| \left\langle \frac{A^\top v}{\|A^\top v\|}, x_i \right\rangle + \langle v, b \rangle. \end{aligned}$$

We have by defining $v' = \frac{A^\top v}{\|A^\top v\|}$ for any $S \subset [k]$ with $|S| \geq (1 - \nu)k$:

$$\begin{aligned} \mu(\{\langle v, x'_i \rangle\}_{i \in S}) &= \|A^\top v\| \cdot \mu(\{\langle v', x_i \rangle\}_{i \in S}) + \langle v, b \rangle \\ \sigma_1(\{\langle v, x'_i \rangle\}_{i \in S}) &= \|A^\top v\| \cdot \sigma_1(\{\langle v', x_i \rangle\}_{i \in S}). \end{aligned}$$

As a consequence, we get:

$$\begin{aligned} |\langle v, Ax + b \rangle - \mu(\{\langle v, x'_i \rangle\}_{i \in S})| &= \|A^\top v\| \cdot |*| \langle v', x \rangle - \mu(\{\langle v', x_i \rangle\}_{i \in S}) \\ &\leq 2 \cdot \|A^\top v\| \cdot \sigma_1(\{\langle v', x_i \rangle\}_{i \in S}) = 2\sigma_1(\{\langle v, x'_i \rangle\}_{i \in S}) \end{aligned}$$

where the inequality follows from $x \in T$. Since, the above holds for all $x \in T$, $v \in \mathbb{S}^{d-1}$, $S \subset [k]$ with $|S| \geq (1 - \nu)k$, we get that $f(T) \subseteq T'$. By repeating the above argument for $f^{-1}(z) = A^{-1}z - A^{-1}b$, we get that $f^{-1}(T') \subset T$ which implies $T' \subset f(T)$ concluding the proof of the claim and hence, the theorem. $\square$

Next, we have a simple lemma analyzing the setting where the inliers have singularities.

Lemma 4.4.2. *Suppose $\mathbf{X} = \{x_i\}_{i=1}^k \subset \mathbb{R}^d$, $v \in \mathbb{S}^{d-1}$, $t \in \mathbb{R}$, $S \subset [k]$ satisfy:*

$$|S| \geq (1 - \nu)k \quad \text{and} \quad \forall i \in S : \langle v, x_i \rangle = t.$$

Then, the output of [Algorithm 1](#) on input $\mathbf{X}$, $\hat{\mu}(\mathbf{X})$, satisfies:

$$\langle \hat{\mu}(\mathbf{X}), v \rangle = t.$$

Proof. Observe for the direction v and set S (Algorithm 1):

$$\mu_{v,S} = t \text{ and } \sigma_{v,S} = 0.$$

Hence, we get from Theorem 4.4.1 and Algorithm 1:

$$|\langle \hat{\mu}(\mathbf{X}), v \rangle - \mu_{v,S}| \leq 2\sigma_{v,S} \implies \langle \hat{\mu}(\mathbf{X}), v \rangle = t$$

concluding the proof. $\square$

4.5 Theoretical Guarantees

Here, we prove Theorem 4.1.1 using the high-dimensional median described in the previous section. To account for scenarios where Σ is singular, we extend the definition of the Mahalanobis norm below:

Definition 4.5.1. For general $\Sigma \succcurlyeq 0$, define the Mahalanobis norm as follows:

$$\|x\|_{\Sigma} = \sup_{v \in \mathbb{S}^{d-1}} \frac{\langle v, x \rangle}{\sqrt{v^{\top} \Sigma v}}.$$

Note that when Σ is invertible, the above definition corresponds to the definition in MAH (Lemma B.1.5). Here, we use the convention that $\frac{0}{0} := 0$ and $\frac{p}{0} := \infty$ for any $p > 0$. Hence, for an estimator to achieve finite Mahalanobis error, it may not output an estimate incurring error along the directions with *zero* variance (the null space of Σ).

Our estimator achieving the guarantees of Theorem 4.1.1 is defined in Algorithm 2. Note that since our high-dimensional median is affine-equivariant (Theorem 4.4.1), so is Algorithm 2. Hence, it suffices to establish Theorem 4.1.1 in the setting $\mu = 0$ and

$$\Sigma = \begin{bmatrix} I_{r \times r} & \mathbf{0}_{r \times (d-r)} \\ \mathbf{0}_{(d-r) \times r} & \mathbf{0}_{(d-r) \times (d-r)} \end{bmatrix}$$

where $I_{r \times r}$ denotes the $r \times r$ identity matrix and $\mathbf{0}_{l \times m}$ the $l \times m$ matrix all of whose entries are 0. Note that the *uncorrupted* samples, $\tilde{\mathbf{X}} = \{\tilde{X}_i\}_{i=1}^n$, satisfy $\langle \tilde{X}_i, e_j \rangle = 0$ for all $j > r$. Hence, this also holds for the *uncorrupted* bucketed means $\tilde{\mu}_i$. Since there are at most ηn corrupted points and the number of bucketed means $k \geq 6\eta dn$ (Algorithm 2), we have from Lemma 4.4.2 that the output $\hat{\mu}$ of Algorithm 2 satisfies:

$$\forall i > r : \langle \hat{\mu}, e_i \rangle = 0.$$

As a consequence, it suffices to now bound the *Euclidean* error of the estimator. Next, we prove the following technical lemma which establishes the required concentration properties on the bucketed means, $\hat{\mu}_i$'s. Before we proceed, we define the thresholding operator for a threshold $\tau \geq 0$ as follows where $\text{sgn}(\cdot)$ denotes the signum function:

$$\psi_{\tau}(x) = \begin{cases} x & \text{if } |x| \leq \tau \\ \text{sgn}(x)\tau & \text{otherwise} \end{cases}.$$

Algorithm 2 Affine-equivariant Estimator

-
- 1: **Input:** Point set $\mathbf{X} = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$, Confidence Parameter δ , Corruption Fraction η
 - 2: $k \leftarrow \max(6\eta dn, Cd \log(1/\delta))$
 - 3: Partition $\mathbf{X}$ into k equally sized buckets $\{\mathcal{B}_i\}_{i \in [k]}$
 - 4: Compute $\hat{\mu}_i = \mu(\mathcal{B}_i)$
 - 5: $\hat{\mu} = \text{High-dimensional Median}(\{\hat{\mu}_i\}_{i \in [k]})$
 - 6: **Return:** $\hat{\mu}$
-

Lemma 4.5.2. *There exists an absolute constant $C > 0$ such that the following holds. Let $Y_1, \dots, Y_k$ be k i.i.d. random vectors drawn from a distribution D with mean 0 and variance $\mathbb{E}[Y_k Y_k^\top] \preceq \sigma^2 I$. Then, we have for $\tau = 24\sigma d$:*

$$\max_{v \in \mathbb{S}^{d-1}} \frac{1}{k} \sum_{i=1}^k |\psi_\tau(\langle v, Y_i \rangle)| \leq 2\sigma$$

with probability at least $1 - \delta$ when $k \geq Cd \log(1/\delta)$.

Proof. Define the random variable Z as follows with Y drawn from D :

$$Z = \max_{v \in \mathbb{S}^{d-1}} \frac{1}{k} \sum_{i=1}^k |\psi_\tau(\langle Y_i, v \rangle)| - \mathbb{E}[|\psi_\tau(\langle Y, v \rangle)|].$$

We first bound the expectation of Z with Y'_i and γ_i denoting i.i.d. vectors from D and Rademacher random variables respectively. The fourth inequality follows from the Ledoux-Talagrand contraction principle (Corollary B.1.4) and the observation that $|\psi_\tau(\cdot)|$ is 1-Lipschitz:

$$\begin{aligned} \mathbb{E}[Z] &\leq \mathbb{E} \left[\max_{v \in \mathbb{S}^{d-1}} \left| \frac{1}{k} \sum_{i=1}^k |\psi_\tau(\langle Y_i, v \rangle)| - \mathbb{E}[|\psi_\tau(\langle Y, v \rangle)|] \right| \right] \\ &\leq \mathbb{E} \left[\max_{v \in \mathbb{S}^{d-1}} \left| \frac{1}{k} \sum_{i=1}^k |\psi_\tau(\langle Y_i, v \rangle)| - |\psi_\tau(\langle Y'_i, v \rangle)| \right| \right] \\ &= \mathbb{E} \left[\max_{v \in \mathbb{S}^{d-1}} \left| \frac{1}{k} \sum_{i=1}^k \gamma_i (|\psi_\tau(\langle Y_i, v \rangle)| - |\psi_\tau(\langle Y'_i, v \rangle)|) \right| \right] \\ &\leq 2\mathbb{E} \left[\max_{v \in \mathbb{S}^{d-1}} \left| \frac{1}{k} \sum_{i=1}^k \gamma_i |\psi_\tau(\langle Y_i, v \rangle)| \right| \right] \leq 4\mathbb{E} \left[\max_{v \in \mathbb{S}^{d-1}} \left| \frac{1}{k} \sum_{i=1}^k \gamma_i \langle Y_i, v \rangle \right| \right] \\ &= \frac{4}{k} \mathbb{E} \left[\left\| \sum_{i=1}^k \gamma_i Y_i \right\| \right] \leq \frac{4}{k} \sqrt{\mathbb{E} \left[\left\| \sum_{i=1}^k \gamma_i Y_i \right\|^2 \right]} \leq 4\sqrt{\frac{d}{k}} \sigma. \end{aligned}$$

Additionally, noting that $\psi_\tau(x) \leq \tau$ for all $x \in \mathbb{R}$:

$$Y_{i,v} := \frac{1}{\tau} (|\psi_\tau(\langle Y_i, v \rangle)| - \mathbb{E}[|\psi_\tau(\langle Y, v \rangle)|]) \leq 1.$$

Furthermore, we have for all $v \in \mathbb{S}^{d-1}$:

$$\sum_{i=1}^k \mathbb{E}[Y_{i,v}^2] \leq \frac{k}{\tau^2} \mathbb{E}[\psi_\tau(\langle Y, v \rangle)^2] \leq \frac{k}{\tau^2} \mathbb{E}[\langle Y, v \rangle^2] \leq \frac{k\sigma^2}{\tau^2}.$$

Hence, we get by an application of Bousquet's inequality ([Theorem B.1.2](#)):

$$\mathbb{P}\{Z \geq \mathbb{E}[Z] + t\} \leq \exp\left(-\left(\frac{k}{\tau}\right)^2 \cdot \frac{t^2}{2(v + kt/(3\tau))}\right) \text{ where } v = \frac{8\sigma\sqrt{kd}}{\tau} + \frac{k\sigma^2}{\tau^2}.$$

Setting $t = \frac{\sigma}{2}$ and from our setting of τ and k , we get:

$$Z \leq \sigma$$

with probability at least $1 - \delta$. The lemma now follows as:

$$\mathbb{E}[|\psi_\tau(\langle Y, v \rangle)|] \leq \sqrt{\mathbb{E}[\psi_\tau(\langle Y, v \rangle)^2]} \leq \sqrt{\mathbb{E}[\langle Y, v \rangle^2]} \leq \sigma.$$

□

We now proceed to the proof of [Theorem 4.1.1](#). For the sake of analysis let $\tilde{\mu}_i$ denote the *uncorrupted* versions of the bucketed means $\hat{\mu}_i$. For these, we have:

$$\mathbb{E}[\tilde{\mu}_i] = 0 \quad \text{and} \quad \mathbb{E}[\tilde{\mu}_i \tilde{\mu}_i^\top] \preceq \frac{k}{n} I.$$

Hence, we get by [Lemma 4.5.2](#) and the setting of k in [Algorithm 2](#):

$$\forall v \in \mathbb{S}^{d-1} : \frac{1}{k} \sum_{i=1}^k |\psi_\tau(\langle v, \tilde{\mu}_i \rangle)| \leq 2\tilde{\sigma} \text{ where } \tilde{\sigma} = \sqrt{\frac{k}{n}} \text{ and } \tau = 24\tilde{\sigma}d$$

with probability at least $1 - \delta$. We condition on this event in the remainder of the proof. Note, furthermore, that there are at most ηn many corrupted points in $\mathbf{X}$. Therefore, we have for $|\{i : \tilde{\mu}_i = \hat{\mu}_i\}| \geq (1 - 1/(6d))k$ again from the setting of k in [Algorithm 2](#) and that:

$$\forall v \in \mathbb{S}^{d-1} : \frac{1}{k} \sum_{i=1}^k \mathbf{1}\{|\langle v, \tilde{\mu}_i \rangle| \geq \tau\} \leq \frac{1}{\tau} \cdot \frac{1}{k} \sum_{i=1}^k |\psi_\tau(\langle v, \tilde{\mu}_i \rangle)| \leq \frac{1}{12d}.$$

Therefore, we get from the previous two observations that:

$$\forall v \in \mathbb{S}^{d-1} : |\mathcal{G}_v| \geq \left(1 - \frac{1}{4d}\right) k \text{ where } \mathcal{G}_v = \{i : \tilde{\mu}_i = \hat{\mu}_i \text{ and } \psi_\tau(\langle v, \tilde{\mu}_i \rangle) = \langle v, \tilde{\mu}_i \rangle\}.$$

Now, let $v \in \mathbb{S}^{d-1}$. We have for $\mathcal{G}_v$ from [Algorithm 1](#) and [Theorem 4.4.1](#):

$$|\langle v, \hat{\mu} \rangle - \mu(\{\langle v, \hat{\mu}_i \rangle\}_{i \in \mathcal{G}_v})| \leq 2\sigma_1(\{\langle v, \hat{\mu}_i \rangle\}_{i \in \mathcal{G}_v}).$$

For the mean term, we get:

$$|\mu(\{\langle v, \hat{\mu}_i \rangle\}_{i \in \mathcal{G}_v})| \leq \mu(\{|\langle v, \hat{\mu}_i \rangle|\}_{i \in \mathcal{G}_v}) \leq \frac{1}{(1 - 1/(4d))} \mu(\{|\psi_\tau(\langle v, \hat{\mu}_i \rangle)|\}_{i \in [k]}) \leq 3\tilde{\sigma}.$$

For the deviation term, we get:

$$\sigma_1(\{\langle v, \hat{\mu}_i \rangle\}_{i \in \mathcal{G}_v}) = \mu(\{|\langle v, \hat{\mu}_i \rangle - \mu(\{\langle v, \hat{\mu}_i \rangle\}_{i \in \mathcal{G}_v})|\}_{i \in \mathcal{G}_v}) \leq \mu(\{|\langle v, \hat{\mu}_i \rangle|\}_{i \in \mathcal{G}_v}) + 3\tilde{\sigma} \leq 6\tilde{\sigma}.$$

The above two bounds imply:

$$\forall v \in \mathbb{S}^{d-1} : |\langle v, \hat{\mu} \rangle| \leq 15\tilde{\sigma} = 15\sqrt{\frac{k}{n}}$$

establishing the theorem. $\square$

We conclude the section with some additional remarks. For clarity and ease of exposition, we impose a minor constraint that $\eta \in [0, 1/(6d)]$. The analysis reveals that the results are applicable for any $\eta \leq c/(d+1)$ with any constant $c < 1$. Moreover, in the absence of adversarial contamination or bucketing, our analysis yields an $O(1)$ upper bound in the distributional setting. This stands in contrast with the $\Omega(\sqrt{d})$ lower bound of Tukey median, as elucidated in [Section 4.3](#).

4.6 Lower Bounds

Here, we present the proofs of [Theorems 4.1.2](#) to [4.1.4](#) which show that the guarantees of [Theorem 4.1.1](#) are nearly tight. For the heavy-tailed setting with *no* adversarial contamination (i.e., $\eta = 0$), [Theorem 4.1.2](#) shows that the recovery error of our estimator is optimal up to a $\sqrt{\log(d)}$ factor and for the adversarial contamination model, [Theorems 4.1.3](#) and [4.1.4](#) establish that *no* affine-equivariant estimator can achieve breakdown point greater than $1/(d+1)$ and that $O(\sqrt{d\eta})$ is the best achievable recovery error for *any* affine-equivariant estimator.

Heavy-tailed Lower Bound ([Theorem 4.1.2](#))

To define our class of distributions, let:

$$\varepsilon = \frac{1}{4} \sqrt{\frac{d \log(1/(d\delta))}{n \log(d)}}.$$

Our hard class will contain d distributions with support over the standard basis vectors and the origin, i.e., $\{e_i\}_{i=1}^d \cup \{\mathbf{0}\}$ such that each distribution puts a smaller mass at one of the standard basis vectors. More formally, we have $\mathcal{D} = \{D_i\}_{i=1}^d$ with:

$$\mathbb{P}_{X \sim D_i} \{X = e_j\} = \begin{cases} \frac{\varepsilon^2}{d} & \text{if } i \neq j \\ \frac{\varepsilon^2}{d^2} & \text{if } i = j \end{cases},$$

and

$$\mathbb{P}_{X \sim D_i} \{X = \mathbf{0}\} = 1 - \frac{d-1}{d}\varepsilon^2 - \frac{\varepsilon^2}{d^2}.$$

By a straightforward calculation, we have:

$$\Sigma(D_i) \preceq M^i \text{ where } M_{jk}^i = \begin{cases} 0 & \text{if } j \neq k \\ \frac{\varepsilon^2}{d} & \text{if } j = k \text{ and } j \neq i \\ \frac{\varepsilon^2}{d^2} & \text{if } j = k = i \end{cases}.$$

Now consider the following procedure of generating the data X :

1. Sample a random integer I from the index set $\{1, 2, \dots, d\}$.
2. Given $I = i$, draw n i.i.d. samples $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$ from D_i .

Next, it suffices to show that for any estimator $\hat{\mu}(\mathbf{X})$, we have

$$\mathbb{P} \left\{ \|\hat{\mu}(\mathbf{X}) - \mu(D_I)\|_{\Sigma(D_I)} \geq \frac{1}{4}\varepsilon \right\} \geq \delta.$$

For each distribution D_i , consider a set of instances

$$S_i = \left\{ \mathbf{X} = (x_1, \dots, x_n) : m_i(\mathbf{X}) \geq \frac{4\varepsilon^2 n}{d} \text{ and } \sum_{j=1}^d \mathbf{1} \left\{ m_j(\mathbf{X}) < \frac{4\varepsilon^2 n}{d} \right\} \geq \frac{d}{2} \right\}$$

$$\text{where } m_j(\mathbf{X}) := \sum_{k=1}^n \mathbf{1} \{x_k = e_j\}.$$

Next, consider $S := \cup S_i$. Now for any $\mathbf{X} \in S$, let

$$\mathcal{J} := \left\{ j : m_j(\mathbf{X}) < \frac{4\varepsilon^2 n}{d} \right\} \text{ and } z := \hat{\mu}(\mathbf{X}).$$

Suppose $z_j \leq \frac{\varepsilon^2}{2d}$ for all $j \in \mathcal{J}$. Then, by considering the cumulative error on $\mathcal{J}$ we have

$$\|\hat{\mu}(\mathbf{X}) - \mu(D_k)\|_{\Sigma(D_k)}^2 \geq \left(\frac{d}{2} - 1 \right) \frac{(\varepsilon^2/2d)^2}{\varepsilon^2/d} \geq \frac{1}{16}\varepsilon^2$$

where we use [Lemma B.1.6](#) with M^k and $\Sigma(D_k)$ for any $D_k \in \mathcal{D}$.

On the other hand, suppose there exists $j \in \mathcal{J}$ such that $z_j > \frac{\varepsilon^2}{2d}$. If $I = j$ is the sampled index, then by considering the error on e_j we have by [Lemma B.1.5](#)

$$\|\hat{\mu}(\mathbf{X}) - \mu(D_j)\|_{\Sigma(D_j)} \geq \frac{|\varepsilon^2/2d - \varepsilon^2/d^2|}{\sqrt{\varepsilon^2/d^2}} \geq \frac{1}{4}\varepsilon.$$

Since $\mathbf{X} \in S$, let i be an index such that $\mathbf{X} \in S_i$. Then, we have the posterior probability of $I = j$ is at least that of $I = i$. From the previous two inequalities, we obtain

$$\mathbb{P}\left\{\|\hat{\mu}(\mathbf{X}) - \mu(D_I)\|_{\Sigma(D_I)} \geq \frac{1}{4}\varepsilon \middle| \mathbf{X}\right\} \geq \mathbb{P}\{I = j | \mathbf{X}\} \geq \mathbb{P}\{I = i | \mathbf{X}\}.$$

From the above, we have

$$\begin{aligned} \mathbb{P}\left\{\|\hat{\mu}(\mathbf{X}) - \mu(D_I)\|_{\Sigma(D_I)} \geq \frac{1}{4}\varepsilon\right\} &\geq \sum_{\mathbf{X} \in S_i} \mathbb{P}\left\{\|\hat{\mu}(\mathbf{X}) - \mu(D_I)\|_{\Sigma(D_I)} \geq \frac{1}{4}\varepsilon \middle| \mathbf{X}\right\} \mathbb{P}\{\mathbf{X}\} \\ &\geq \sum_{\mathbf{X} \in S_i} \mathbb{P}\{I = i | \mathbf{X}\} \mathbb{P}\{\mathbf{X}\} = \sum_{\mathbf{X} \in S_i} \mathbb{P}\{I = i, \mathbf{X}\} \\ &= \sum_{\mathbf{X} \in S_i} \mathbb{P}\{\mathbf{X} | I = i\} \mathbb{P}\{I = i\}. \end{aligned}$$

It remains to prove that $\mathbb{P}\{S_i | I = i\} \geq d\delta$. For simplicity of notation, denote $\mathbb{P}_i$ as the conditional distribution of $\mathbf{X}$ under $I = i$. Define events:

$$\begin{aligned} A &= \left\{\mathbf{X} = (x_1, \dots, x_n) : m_i(\mathbf{X}) \geq \frac{4\varepsilon^2 n}{d}\right\} \\ B &= \left\{\mathbf{X} = (x_1, \dots, x_n) : \sum_{j=1, j \neq i}^d \mathbf{1}\left\{m_j(\mathbf{X}) < \frac{4\varepsilon^2 n}{d}\right\} \geq \frac{d}{2}\right\} \\ C &= \left\{\mathbf{X} = (x_1, \dots, x_n) : m_0(\mathbf{X}) \geq (1 - 2\varepsilon^2)n\right\} \text{ where } m_0(\mathbf{X}) := \sum_{k=1}^n \mathbf{1}\{x_k = \mathbf{0}\}. \end{aligned}$$

Note that $\mathbb{P}_i\{S_i\} = \mathbb{P}_i\{A \cap B\}$ and $C \subseteq B$. So we have

$$\mathbb{P}_i(A \cap B) = \mathbb{P}_i(B|A)\mathbb{P}_i(A) \geq \mathbb{P}_i(B)\mathbb{P}_i(A) \geq \mathbb{P}_i(C)\mathbb{P}_i(A).$$

where the first inequality follows from [Lemma B.1.7](#).

We first use a Binomial tail lower bound to bound $\mathbb{P}_i(A)$ (see e.g., [3]):

$$\mathbb{P}_i\{B(n, p) \geq k\} \geq \frac{1}{\sqrt{8n\frac{k}{n}(1 - \frac{k}{n})}} \exp\left(-nD\left(\frac{k}{n} \parallel p\right)\right),$$

where $B(n, p)$ denotes a Binomial random variable and $D(a \parallel p) = a \log \frac{a}{p} + (1 - a) \log \frac{1-a}{1-p}$ denotes the KL divergence.

Plugging in $k = 4dnp$ and $p = \varepsilon^2/d^2$ we obtain that

$$\mathbb{P}_i(A) \geq \frac{1}{\sqrt{32ndp}} \exp(-4ndp \log(4d)) \geq 2d\delta.$$

Also, note that $\mathbb{P}(C) \geq 1/2$ since $m_0(x)$ follows a Binomial distribution with $\mathbb{E}[m_0(x)] \geq (1 - \varepsilon^2)n$. Therefore, we have $\mathbb{P}_i(S_i) \geq d\delta$ concluding the proof. $\square$

Adversarial Contamination ([Theorems 4.1.3](#) and [4.1.4](#))

We start with [Theorem 4.1.3](#).

Proof of [Theorem 4.1.3](#)

Let $r > 0$ and $S = \{e_i\}_{i=1}^d \cup \{\mathbf{0}\}$. First define the family $\tilde{\mathcal{D}} = \{\tilde{D}_i\}_{i=0}^{d+1}$ with $\tilde{D}_0$ and $\tilde{D}_{d+1}$ denoting the uniform distributions over S and $S \setminus \{\mathbf{0}\}$ respectively and for $i \in [d]$, $\tilde{D}_i$ is defined as follows:

$$\mathbb{P}_{X \sim \tilde{D}_i} \{X = x\} = \begin{cases} 0 & \text{if } x = e_i \\ \frac{1}{d} & \text{if } x \in S \setminus \{e_i\} \end{cases}.$$

Defining $\sigma = 1/(2dr)$, our hard family of distributions $\mathcal{D} = \{D_i\}_{i=0}^{d+1}$ is defined in the following way:

1. First, generate $\tilde{X} \sim \tilde{D}_i$.
2. Independently, generate $Z \sim \text{Unif}(\{\pm 1\}^d)$.
3. Observe $X = \tilde{X} + \sigma Z$.

Note that $\Sigma(D_i)$ is non-singular for each i and $\tilde{D}_0$ satisfies:

$$\tilde{D}_0 = \frac{d}{d+1} \tilde{D}_i + \frac{1}{d+1} \delta_{e_i}$$

where δ_{e_i} denotes a Dirac-delta distribution on e_i . Suppose $\hat{\mu}$ satisfies for some $n \in \mathbb{N}$:

$$\forall D \in \mathcal{D} : \mathbb{P}_{\mathbf{X} \sim D_0^n} \{ \|\hat{\mu}(\mathbf{X}) - \mu(D)\|_{\Sigma(D)} \geq r \} < \frac{1}{d+1}.$$

Then, by the union bound, there must exist a sample $\mathbf{X}$ in the support of D_0^n and an outcome $\hat{\mu} = \hat{\mu}(\mathbf{X})$ over the randomness of $\hat{\mu}$ such that:

$$\forall D \in \mathcal{D} \setminus \{D_0\} : \|\hat{\mu} - \mu(D)\|_{\Sigma(D)} \leq r.$$

Then, considering the direction e_i for any $i \in [d]$, we have by [Lemma B.1.5](#):

$$r \geq \|\hat{\mu} - \mu(D_i)\|_{\Sigma(D_i)} \geq 2dr |\hat{\mu}_i|.$$

This, implies:

$$|\hat{\mu}_i| \leq \frac{1}{2d} \implies \sum_{i=1}^d |\hat{\mu}_i| \leq \frac{1}{2}.$$

However, note that we have for the direction $\mathbf{1}$ and distribution D_{d+1} by [Lemma B.1.5](#):

$$r \geq \|\hat{\mu} - \mu(D_{d+1})\|_{\Sigma(D_{d+1})} \geq \left| \frac{1 - \sum_{i=1}^d \hat{\mu}_i}{\sqrt{d}/(2dr)} \right| \geq \sqrt{d}r,$$

which is a contradiction thus establishing the theorem. $\square$

We now move on to [Theorem 4.1.4](#).

Proof of [Theorem 4.1.4](#)

As before, we will construct a hard family of distributions. For support set $S = \{e_i\}_{i=1}^d \cup \{\mathbf{1}/d\}$, define the set of distributions $\tilde{\mathcal{D}} = \{\tilde{D}_i\}_{i=0}^d$ as follows:

$$\begin{aligned} \forall i \in [d] : \mathbb{P}_{X \sim \tilde{D}_i}(X = x) &= \begin{cases} 0 & \text{if } x = e_i \\ \frac{\eta}{1-\eta} & \text{if } x = e_j \text{ for } j \neq i \text{ and} \\ \frac{1-d\eta}{1-\eta} & \text{if } x = \frac{\mathbf{1}}{d} \end{cases} \\ \mathbb{P}_{X \sim \tilde{D}_0}(X = x) &= \begin{cases} \eta & \text{if } x = e_j \text{ for any } j \in [d] \\ 1-d\eta & \text{if } x = \frac{\mathbf{1}}{d} \end{cases}. \end{aligned}$$

Let

$$r := \frac{1}{2} \sqrt{\frac{d\eta}{1-d\eta}} \text{ and } \sigma := \frac{\eta}{8r}.$$

The hard family of distributions, $\mathcal{D} = \{D_i\}_{i=0}^d$ is defined as follows:

1. First, generate $\tilde{X} \sim \tilde{D}_i$.
2. Independently, generate $Z \sim \text{Unif}(\{\pm 1\}^d)$.
3. Observe $X = \tilde{X} + \sigma Z$.

As before, for each i , $\Sigma(D_i)$ is non-singular and D_0 may be written as a mixture of D_i and the distribution with all its mass on e_i . Now, suppose $\hat{\mu}$ is an estimator that satisfies for some $n \in \mathbb{N}$:

$$\forall D \in \mathcal{D} : \mathbb{P}_{\mathbf{X} \sim D_0^n} \{ \|\hat{\mu}(\mathbf{X}) - \mu(D)\|_{\Sigma(D)} \geq r \} < \frac{1}{d+1}.$$

Then, by the union bound, there must exist a sample $\mathbf{X}$ in the support of D_0^n and an outcome $\hat{\mu} = \hat{\mu}(\mathbf{X})$ over the randomness of $\hat{\mu}$ such that:

$$\forall D \in \mathcal{D} : \|\hat{\mu} - \mu(D)\|_{\Sigma(D)} \leq r.$$

We have for the direction $\mathbf{1}$ by Lemma B.1.5:

$$r \geq \|\hat{\mu} - \mu(D_0)\|_{\Sigma(D_0)} \geq \frac{1}{\sqrt{d}\sigma} \left| \sum_{i=1}^d \hat{\mu}_i - 1 \right|.$$

Therefore, there exists $i \in [d]$ with:

$$\hat{\mu}_i \geq \frac{1 - \sqrt{d}\sigma r}{d}.$$

For this i , by considering the direction e_i we have by Lemma B.1.5:

$$\begin{aligned} \|\hat{\mu} - \mu(D_i)\|_{\Sigma(D_i)} &\geq \frac{(1-\eta)d}{\sqrt{\eta(d-1)(1-d\eta)} + (1-\eta)d\sigma} \cdot \left(\frac{1 - \sqrt{d}\sigma r}{d} - \frac{1-d\eta}{1-\eta} \cdot \frac{1}{d} \right) \\ &= \frac{1}{\sqrt{\eta(d-1)(1-d\eta)} + (1-\eta)d\sigma} \cdot \left((d-1)\eta - (1-\eta)\sqrt{d}\sigma r \right) \\ &> \frac{1}{\sqrt{\eta d(1-d\eta)} + d\sigma} \cdot \left(\frac{3d}{4}\eta - \frac{d}{2}\sigma r \right) \geq \frac{3}{4} \frac{1}{\sqrt{\eta d(1-d\eta)}} \cdot \frac{2}{3} d\eta = r, \end{aligned}$$

where the second inequality follows from the fact that $d > 3$ and the third from our setting of σ . This leads to a contradiction concluding the proof of the theorem. $\square$

4.7 Summary

In this work, we studied the problem of affine-equivariant robust mean estimation when the data is drawn from a bounded covariance distribution with no additional assumptions. We obtain nearly-tight characterizations of the optimal achievable error. In sharp contrast to the Euclidean setting, the sub-Gaussian rate is no longer achievable here. Furthermore, we established traditional affine-equivariant estimators are *sub-optimal* even within this class. Finally, we derived a novel estimator which addresses these deficiencies.

There remain several open questions from our work. Firstly, similar to previous affine-equivariant estimators such as the Tukey median, our estimator is not computationally tractable. In fact, the best-known computational results for this problem only guarantee a point with Tukey depth $\Omega(1/d^2)$ [23]; a factor of d worse than that achievable in the computationally unbounded setting. Consequently, this only guarantees robustness to a $\Omega(1/d^2)$ fraction of outliers. Hence, addressing this shortcoming would constitute a significant step towards building more practical estimators. Secondly, our results hold in the general setting

of covariance bounded distributions. However, when the right standardization of the data is available, sub-Gaussian rates are achievable for the Mahalanobis norm. Whether there exist natural settings where such rates are obtainable even when these standardization are *not* estimable is another direction for future research.

Chapter 5

Algorithmic Iterative Reticular Synthesis

5.1 Introduction

The discovery of novel crystalline materials presents a fundamental combinatorial optimization problem with significant real-world applications. These materials, including metal-organic frameworks (MOFs) and zeolitic imidazolate frameworks (ZIFs), form ordered porous structures essential for environmental applications like water harvesting [45, 43] and CO₂ capture [80, 110]. The challenge lies in determining which combinations of molecular building blocks will successfully crystallize from an exponentially large search space. Despite decades of research guided by geometric principles of reticular synthesis [134], this process remains largely empirical—we lack predictive models to determine a priori which molecular combinations will yield stable crystal structures. This inefficiency has left vast regions of chemical space unexplored, creating both a practical bottleneck in materials development and a rich domain for machine learning applications. The problem differs fundamentally from traditional synthesis optimization: rather than refining parameters for a single target, we must efficiently allocate limited experimental resources across many candidates to maximize the total number of distinct discoveries.

Given the scale and complexity of this challenge, computational and ML-driven methods offer great opportunity because they can process vast amounts of data and identify patterns beyond human capability. However, current implementations of these approaches in chemistry and materials science, while powerful for their intended purposes, are mismatched to this discovery objective. Bayesian Optimization and genetic algorithms excel at finding optimal conditions but concentrates resources on single targets [96, 126, 132, 98, 34]. Experimental screening of computationally predicted and ML-generated structures treats each candidate independently, missing cross-system learning opportunities and failing to adapt when predictions prove incorrect [119, 131, 140]. The discovery challenge demands a fundamentally different approach: one that simultaneously learns crystallization patterns across

diverse chemical systems and strategically allocates experimental resources based on this evolving understanding.

Here we present AIRES (Algorithmic Iterative Reticular Synthesis), an integrated machine learning pipeline that transforms crystal discovery from empirical exploration to systematic decision-making focused on maximizing the number of distinct new structures found. AIRES combines automated synthesis, computer vision for crystal identification, and structure determination, with the critical addition of an algorithmic framework that decides which experiments to perform next (Fig. 5.1). The system employs probabilistic machine learning models trained on carefully engineered chemical descriptors to predict crystallization outcomes across diverse systems, enabling cross-system transfer learning where patterns from one chemical family accelerate discovery in others. These predictions feed a theoretically grounded selection strategy: we prove that for discovery objectives, greedy selection of highest-probability candidates outperforms the exploration-heavy strategies optimal for property optimization. AIRES dynamically allocates resources, immediately excluding successful building blocks to focus on undiscovered ones, while continuously learning from both successes and failures to refine predictions across the chemical space. To implement this strategy in high-throughput experimentation, we developed a principled batch selection algorithm that identifies experiment sets maximizing expected discoveries while accounting for conditional dependencies—preventing the severe redundancy that would arise from simply selecting the top-k candidates.

We demonstrated AIRES on ZIFs, discovering crystals from 10 out of 48 new molecular building blocks with twice the efficiency of random exploration (700 versus 1,400 experiments). This expanded the known single-component Zn-ZIF library by one-third, breaking a decade-long discovery drought. From a computational perspective, ZIFs present unique challenges: (1) while sharing common geometric constraints, minor variations in molecular features can lead to entirely different crystal structures, creating a complex, non-convex optimization landscape where multiple arrangements can be thermodynamically stable [4]; (2) the synthesis involves competing reaction pathways with multiple potential products, making the prediction task inherently multi-modal. This combination—high structural diversity from subtle feature variations coupled with complex characterization requirements—establishes ZIFs as an ideal yet rigorous testbed for machine learning methods. The fact that AIRES uncovered new discoveries in this mature field, where over a hundred known structures suggested exhausted possibilities [141], demonstrates its ability to identify patterns in high-dimensional spaces that elude human intuition.

5.2 Results

Curation of initial ZIF synthesis dataset

The first step in our data-driven approach was to create a comprehensive initial dataset. This involved a pipeline of automated chemical synthesis, optical imaging for preliminary

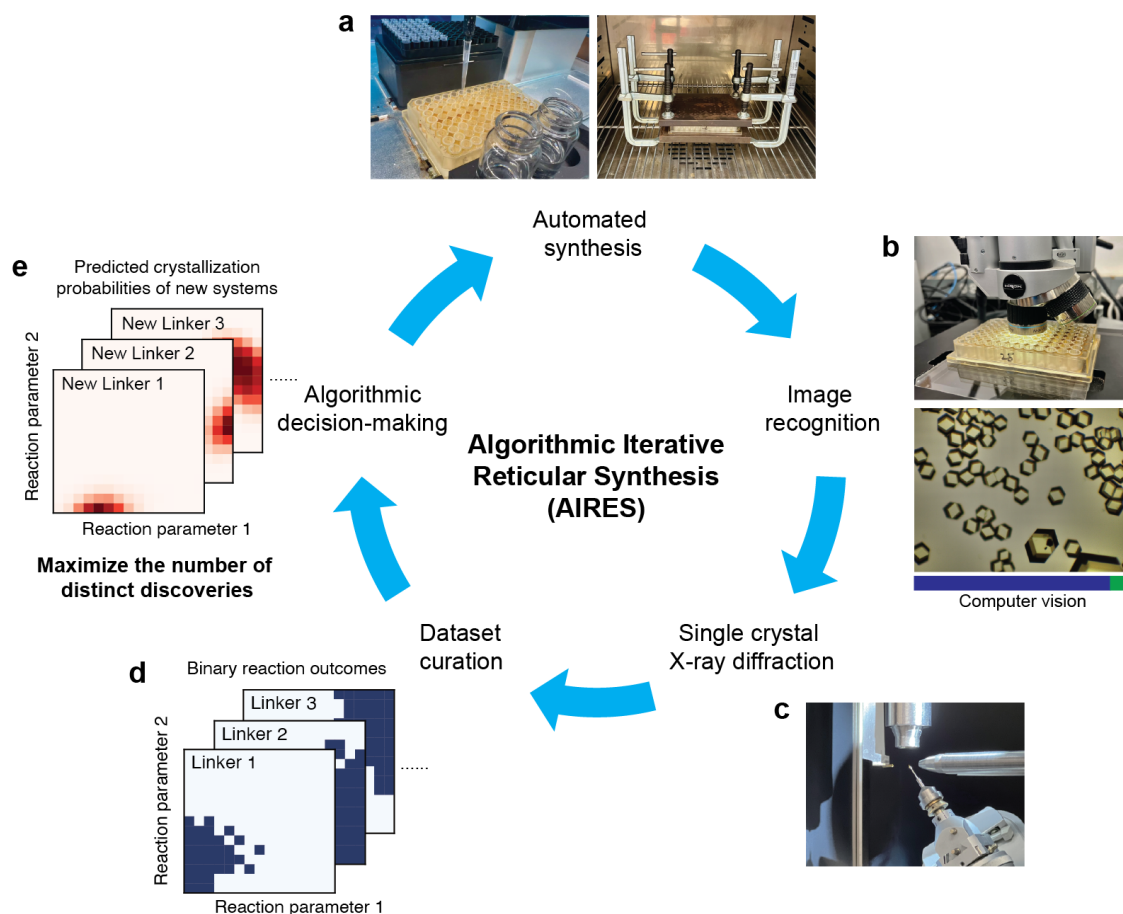


Figure 5.1: **Schematic illustration of AIRES cycle and its components.** **a**, Automated synthesis setup showing a robotic liquid handler (left) and heating unit (right). **b**, Microscope imaging station (top) and optical image collected (bottom) with computer vision classification results (blue: crystal probability, green: non-crystal probability). **c**, Single crystal X-ray diffraction setup. **d**, A two-dimensional demonstration of binary reaction outcome maps for different linker systems, with successful reactions in dark blue and failed ones in white. **e**, A two-dimensional demonstration of ML-predicted crystallization probability maps for different linker systems, with higher predicted probabilities in darker red. The decision-making algorithm, aiming to maximize the number of distinct discoveries, suggests candidates based on predictions for subsequent experimental validation.

classification, and definitive analysis to produce labeled training data (Fig. 5.1a-d).

A significant challenge in applying ML to materials synthesis is the biased nature of existing public data. Scientific literature typically reports only successful experiments, creating a dataset composed almost entirely of positive examples [107, 64]. To train a robust classifier, a dataset containing both successes and failures is essential. Therefore, we generated our own initial dataset by systematically exploring the reaction outcomes for a set of known ZIF-forming building blocks.

For this study, we focused on a specific class of materials (zinc-based ZIFs) built from a single type of organic molecule, or "linker". We selected 8 well-established linkers known to form ZIFs, each with distinct structural features (Fig. 5.2a). For each linker, we performed experiments across a grid of reaction conditions. We varied three key experimental parameters (features): total reactant concentration, the ratio of linker to metal, and temperature. By systematically sampling these parameters, we generated 336 unique experiments for each of the 8 linkers, resulting in an initial dataset of 2,688 labeled reactions (Appendix C.2).

Our high-throughput experimentation platform operates in two stages: automated synthesis and automated characterization. First, a robotic liquid handling system prepares the reaction mixtures based on the 2,688 predefined experimental conditions. After the reaction period, initial screening is performed using automated optical microscopy. To manage the high volume of images, we trained an EfficientNetV2-S convolutional neural network [122] to act as a preliminary classifier. The model was trained on images manually labeled by experts to distinguish between outcomes containing potential crystals and those without (e.g., clear solutions, precipitates).

Human experts then validated the model's selections to identify candidates for the final characterization step: single crystal X-ray diffraction (SCXRD), which provides the definitive atomic structure. This step is crucial, as optical appearance alone can be misleading. For instance, one linker produced two visually distinct crystal types (Fig. 5.2b), which SCXRD confirmed were two completely different materials—one a known ZIF and the other a new, unexpected structure.

To make the final labeling process efficient, we applied two heuristics. First, we assumed that visually similar crystals produced from the same linker have the same underlying atomic structure. Second, we classified a crystal as a "known ZIF" if its fundamental structural parameters (unit cell) matched a known entry in our database. This allowed for rapid confirmation without requiring a full, time-consuming structural analysis for every crystal. Ultimately, each of the 2,688 experiments was assigned a binary label: '1' for the successful formation of a desired ZIF crystal and '0' for all other outcomes (including failures, non-ZIF crystals, and other byproducts).

Visualizing this initial dataset reveals a complex and challenging problem landscape (Fig. 5.2c). The plots show the "success rate" as a function of the experimental parameters (features). These "crystallization landscapes" are highly non-linear and difficult to predict. The underlying physics and chemistry, involving concepts like nucleation theory [66] and coordination chemistry [78], create a complex interplay between the experimental inputs and the final outcome. It is this complexity that makes a purely physics-based prediction model

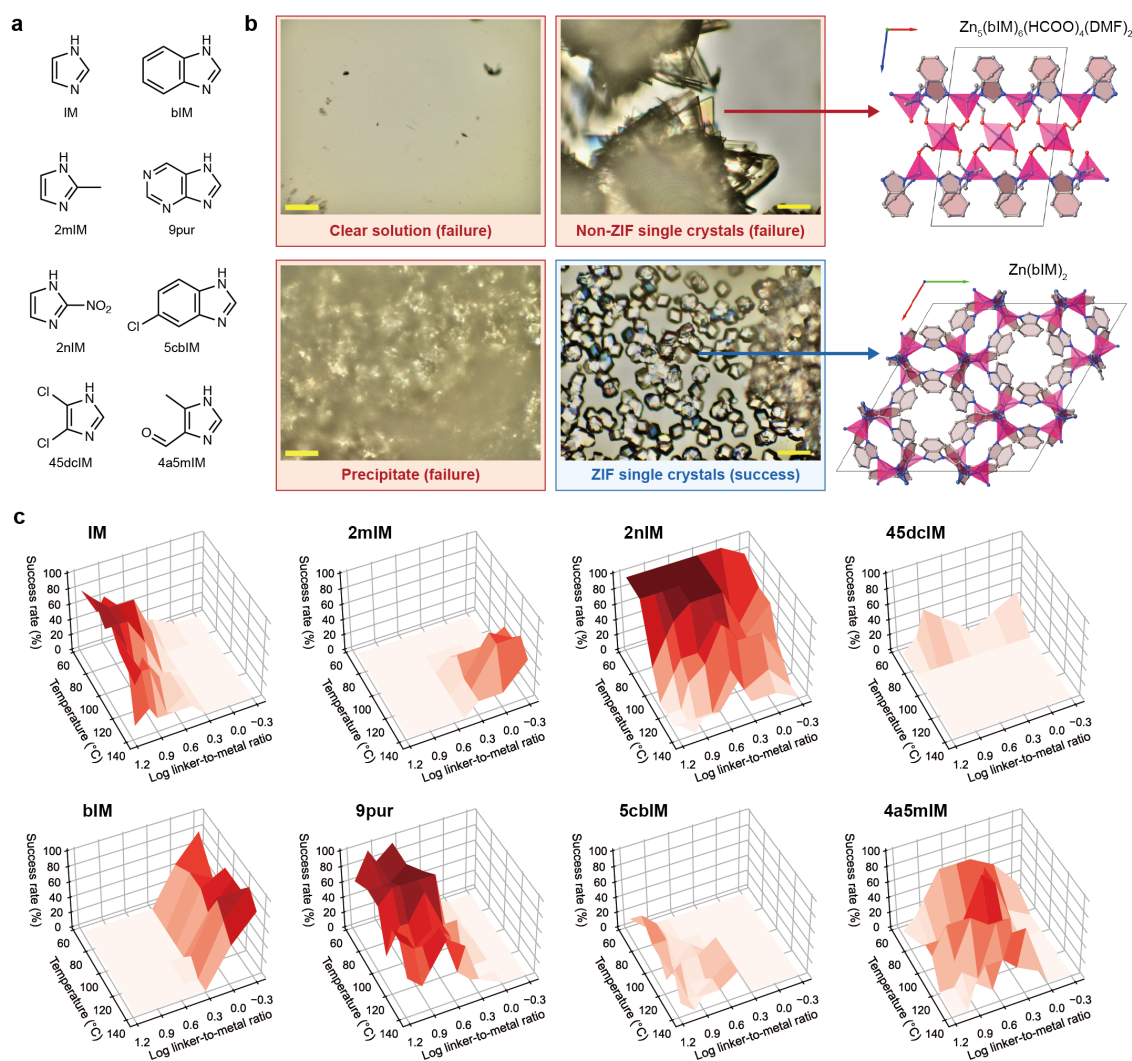


Figure 5.2: **Initial ZIF synthesis dataset.** **a**, Chemical structures and abbreviations of eight starting linkers. Abbreviations denote neutral linkers when referring to reactants, but represent their deprotonated forms in ZIF chemical formulae unless specified otherwise. **b**, Representative optical images (yellow scale bar 100 μm) showing different reaction outcomes: clear solution, precipitate, non-ZIF crystals (red frames, failures), and ZIF single crystals (blue frame, success). Crystal structures represented by ball-and-stick models (right) show $\text{Zn}_5(\text{bIM})_6(\text{HCOO})_4(\text{DMF})_2$ (*b*-axis view) and $\text{Zn}(\text{bIM})_2$ (*c*-axis view) with Zn polyhedra in pink. **c**, Success rate landscapes for each starting linker as a function of temperature and log linker-to-metal ratio, averaged over total concentration. Increasing shades of red indicate higher success rates.

intractable and motivates a data-driven ML approach.

The data highlights two key insights. First, the overall success rate for finding a crystal is low (averaging $25\% \pm 16\%$ across the 8 linkers), confirming that a random or brute-force search strategy is highly inefficient. Second, each linker displays a unique and complex "fingerprint" in the parameter space. For example, some linkers succeed only in very narrow, isolated regions of the space, while others are successful across a broader range of conditions. There are no simple, universal rules that govern success. This heterogeneity underscores the need for a sophisticated model that can learn the specific behavior of each linker while also generalizing patterns across the entire chemical space.

ML-guided discovery methodology

Our discovery methodology is a closed-loop system that integrates ML-based decision-making with automated experimentation (Fig. 5.1). Each iteration of the AIRES loop consists of three steps: 1) an ML model predicts promising experiments, 2) the robotic platform executes these experiments, and 3) the new experimental outcomes are used to update the model for the next iteration. This active learning cycle continues until the experimental budget is exhausted.

To validate our methodology, we defined a search space of 48 new, unexplored linker molecules. These were selected based on practical constraints of our automated platform (e.g., solubility) and chemical principles designed to avoid known failure modes. Of these 48 candidates, 45 had never been successfully used to form the target class of crystals, providing a challenging and realistic testbed for discovery.

A critical component of any ML approach is feature engineering. We developed a comprehensive feature vector to represent each potential experiment, combining experimental parameters with molecular descriptors for each linker. The linker features were designed to capture multiple aspects of its chemistry and physics, including: counts of key functional groups, molecular dimensions (to represent size and shape), and quantum mechanical properties derived from density functional theory (DFT) simulations (Fig. 5.3a). A principal component analysis (PCA) of this feature space shows some separation between the major classes of linkers but also reveals significant complexity that cannot be captured by simple linear projections, justifying the use of non-linear models (Fig. 5.3b).

Our methodology treats crystallization as a binary outcome; namely, success or failure drawn from an underlying probability distribution that reflects both inherent chemical variability and experimental precision. This methodology combines two key components: a prediction model to estimate crystallization probabilities, and a greedy selection strategy that always chooses the experiment with the highest predicted probability of yielding new ZIF crystals. This strategy is model-agnostic, in that it can work with any probabilistic prediction model. We explored two specific models: Random Forest Classification (RFC) [14] and Gaussian Process Classification (GPC) [128, 108]. The RFC captures complex, non-linear relationships through an ensemble of decision trees. However, a key challenge emerged during the discovery process: each initial training linker has approximately 300 data points,

while new linkers are explored one experiment at a time. This data imbalance, combined with inherent differences between training and new linkers, could cause the model to overlook crucial patterns in new data. We addressed this through strategic data augmentation and dynamic hyperparameter optimization that gradually increased the weight of new experimental results. Separately, the GPC provides naturally calibrated probability estimates and we adopt a composite product kernel that captures different patterns of variation in linker features versus reaction conditions. While principally neural network-based prediction models could also be incorporated into our decision-making strategy, our dataset characteristics made simpler models more appropriate: with only 8 training linkers (hence 8 data points for the linker representation), we had insufficient data to effectively train neural networks and hence selected RFC and GPC for their uncertainty quantification properties and ability to perform well with limited data.

Building on these predictive models, we developed a greedy selection strategy that prioritizes experiments with the highest predicted probability of success. Critically, our objective is explicitly defined as maximizing the number of unique successful crystallizations rather than identifying the optimal reaction conditions. Once a linker yields a ZIF crystal, the algorithm excludes further experiments with that linker, focusing resources on unexplored candidates. This objective fundamentally differs from traditional optimization formulations like Bayesian Optimization [116] or Multi-Armed Bandits [72], where they aim to maximize success probability, and cumulative (non-unique) successes, respectively. While such traditional methodologies often favor exploration over exploitation, we demonstrated through both theoretical analysis (Appendix 5.3) and empirical validation (Appendix C.8) that our greedy strategy of selecting the highest-probability candidates is particularly well-suited for crystal discovery, where each success represents a unique scientific advance. Furthermore, to fully utilize high-throughput capabilities, we extended this methodology to batch experimentation, developing a novel principled batch selection algorithm. In each iteration, the batch selection strategy works in conjunction with the trained ML model to identify the set of experiments that collectively maximize the expected number of discoveries; accounting for potential interactions between experiments targeting the same linker through conditional probability estimation (Methods 5.3).

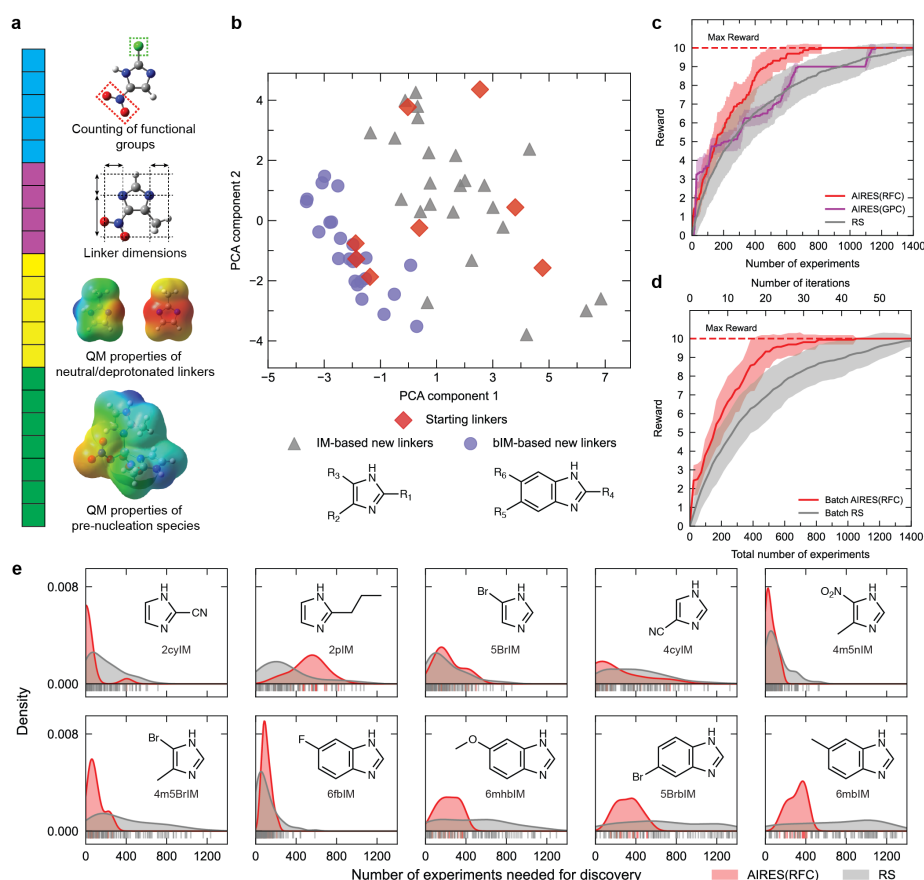


Figure 5.3: **AIRES performance in accelerating ZIF discovery.** **a**, Linker features combining functional group counts (blue, 5 features), linker dimensions (magenta, 4 features), and quantum mechanical properties of neutral/deprotonated linkers (yellow, 5 features) and pre-nucleation species (green, 7 features). **b**, Two-dimensional PCA projection of the chemical space showing the distribution of 8 starting linkers (blue circles), 24 new imidazole-based (grey triangles), and 24 new benzimidazole-based linkers (red diamonds). The first two principal components capture 50.8% of the total variance. **c**, Cumulative reward of AIRES equipped with Random Forest Classification (RFC, red) and Gaussian Process Classification (GPC, magenta), and Random Search baseline (RS, grey). Shaded areas represent one standard deviation across 16 repetitions for RFC and GPC, and 100 repetitions for RS. The red dashed line indicates the maximum achievable reward. **d**, Cumulative reward of AIRES(RFC) (red) and RS (grey) under batch experimentation with 24 experiments per iteration. Shaded areas represent one standard deviation across 16 repetitions for AIRES(RFC) and 100 repetitions for RS. **e**, Kernel density estimation showing the distribution of the number of experiments required to discover new ZIF crystals for 10 successful linkers using AIRES(RFC) (red) versus RS (grey). Horizontal bars beneath each distribution indicate individual discovery events across multiple repetitions.

To rigorously compare our ML methodology with random exploration, multiple trials are necessary to account for the inherent randomness in these approaches. However, conducting repeated reactions would be resource-intensive. Instead, we adopted an equivalent validation strategy: we first experimentally executed a comprehensive grid of reaction conditions across all 48 new linkers, generating a dataset of 1,728 individual reactions. This allowed us to simulate both ML-guided and random exploration by having each method iteratively suggest experiments, then retrieving the corresponding reaction outcomes from this new dataset, rather than repeatedly conducting experiments. Different approaches revealed distinct discovery patterns on this dataset (Fig. 5.3c). Random search (RS) required approximately 1,400 experiments to discover all new ZIF crystals from the ten successful new linkers, showing a steadily diminishing discovery rate characteristic of exhaustive exploration. The GPC approach demonstrated promising early performance with several rapid discoveries, but its effectiveness ultimately plateaued at levels comparable to RS. This behavior likely reflects the challenge of extrapolating from our initial dataset to substantially different linkers. The RFC emerged as the most effective approach, discovering all possible structures in approximately 700 experiments, a two-fold acceleration compared to RS. To balance the influence between initial and new data in RFC, we used data augmentation with 10 copies per new experiment, a choice informed by the relative sample sizes between initial and new linkers. While other augmentation factors also showed decent acceleration over RS, this balanced approach proved particularly effective (Supplementary Fig. C.13). The batch implementation of our methodology maintained the efficiency of the best performing model RFC, consistently requiring approximately half the RS experiments to achieve the same number of discoveries (Fig. 5.3d). In addition, the advantage over RS remained robust across various batch sizes (Supplementary Fig. C.15). This is a nontrivial achievement, as simply selecting the top-k most promising candidates would introduce severe redundancy. Such candidates often share high similarity and provide limited new information, resulting in discovery performance substantially worse than Batch RS (Supplementary Fig. C.16). This demonstrates that our batch selection strategy, which carefully accounts for dependencies between related experiments through conditional probability estimation, successfully extends the advantages of ML-guided discovery to high-throughput settings.

Analysis of distributions of experiments needed for discovery provides insights into both the chemical space and our algorithmic approach (Fig. 5.3e). For linkers such as 2cyIM, both RS and RFC achieved early discovery of the corresponding ZIFs, though the latter still provided modest acceleration. This suggests robust crystallization conditions that are relatively insensitive to precise reaction parameters. In contrast, more challenging targets like 6mbIM showed dramatically different trajectories between methods. While RS discovered these structures at a roughly uniform rate through blind exploration, RFC significantly accelerated their discovery through pattern recognition. Interestingly, 2pIM proved an exception where RS outperformed ML. Despite having moderate success rates overall (explaining RS’s earlier discovery), its unique structural features and resulting ZIF crystal topology differed substantially from other positive examples in our training data. Such unique characteristics caused our greedy algorithm to prioritize other candidates with more familiar patterns evi-

dent in the training data, rather than selecting based on intrinsic success rates that RS relies upon. Nevertheless, overall AIRES equipped with RFC clearly outperformed RS leading to a global acceleration of new ZIF discovery.

Further investigations into what linkers are actually suggested at every iteration reveal the model behavior. The discovery trajectory (Supplementary Figure C.14) shows the model is able to evaluate the tendency of different linkers, therefore some linkers are suggested more frequently while some others are less frequently suggested. This is in contrast to the behavior of RS where each linker will have the same frequency of being suggested. This indicates AIRES(RFC) learns what structural features help and therefore prioritizes promising linkers to try. One can find that when there are substituents both on 2- and 4/5- positions of the imidazole backbone, those linkers are less likely to form single-linker Zn-ZIFs as there will be too many atoms around a single metal coordination center. Moreover, even for successful linkers, we noticed AIRES can find the discovery from some of them very quickly, while for some others it requires more iterations.

5.3 Methods

Reaction encoding

Reaction data were encoded by concatenating linker features with experimental features into comprehensive feature vectors. The experimental features included three primary reaction parameters plus metal and linker concentrations, with the latter two derived from total concentration and log linker-to-metal ratio. The linker features were computed through two sets of density functional theory (DFT) calculations in Gaussian 16 [44]. For individual linkers, geometry optimization was performed for both neutral and deprotonated forms using the B3LYP exchange-correlation functional with 6-31g* basis set, accounting for DMF solvent effects through the Polarizable Continuum Model [123]. From these calculations, we extracted linker dimensions (using imidazolate nitrogen coordinates as reference points) and electronic properties, including the energy difference between neutral and deprotonated forms and atomic electrostatic potentials. To represent the pre-nucleation species, additional DFT calculations were performed using the M06-L functional with def2SVP basis set. These calculations examined clusters containing one Zn center coordinated by two neutral linkers and two nitrates. For each linker type, up to 16 different clusters with varying linker orientations were optimized. Single-point energy calculations using SMD implicit continuum solvent model [92] identified the most thermodynamically stable configurations, from which we extracted formation energies, structural parameters, and APT charges on Zn and N atoms. A complete list of features and their meanings is provided in Appendix C.6.

Decision-making methodology

Our discovery methodology implements a greedy active learning framework optimized for maximizing unique discoveries rather than minimizing prediction error. Unlike traditional active learning that seeks to reduce model uncertainty, our objective function counts distinct successful molecular systems, making each discovery equally valuable regardless of the number of successful conditions found for that system. We developed algorithms for both sequential (single experiment) and parallel (batch) selection, compatible with any probabilistic classification model.

Sequential algorithm

In the sequential discovery process, candidates were selected individually, with each subsequent choice informed by previous outcomes. Our algorithm employed a greedy strategy compatible with any probabilistic prediction model. At each iteration, the algorithm selected the candidate with the highest predicted probability of success. Upon successful crystallization, all remaining candidates for that linker were removed from consideration to focus resources on unexplored candidates. The model was retrained after each iteration to incorporate new reaction data.

The sequential greedy algorithm operates as follows:

Algorithm 3 Sequential Greedy Discovery Algorithm

Require: Initial training data $\mathcal{D}_{\text{train}}$

- 1: Train model $\mathcal{M}$ on $\mathcal{D}_{\text{train}}$
 - 2: Initialize experiment space $\mathcal{X}$
 - 3: **while** budget remains **do**
 - 4: Select $x_t \leftarrow \arg \max_{x \in \mathcal{X}} \mathcal{M}(x)$
 - 5: Observe outcome y_t
 - 6: Update $\mathcal{D}_{\text{train}} \leftarrow \mathcal{D}_{\text{train}} \cup \{(x_t, y_t)\}$
 - 7: Update $\mathcal{X}$ based on y_t
 - 8: Retrain model $\mathcal{M}$ on updated $\mathcal{D}_{\text{train}}$
 - 9: **end while**
-

The greedy approach emphasizes exploitation (selecting promising experiments) over exploration (selecting uncertain experiments), departing from conventional strategies in Bayesian Optimization and Multi-Armed Bandits that typically balance both aspects. While greedy methods can potentially become fixated on initially promising candidates in traditional optimization settings, our distinct objective – maximizing the number of unique discoveries – naturally avoids this limitation. Upon discovering a successful crystallization for a linker, the algorithm automatically moves on to explore other candidates. Conversely, each negative result decreases the predicted success probability for the corresponding linker, naturally shifting focus to more promising alternatives. Indeed, theoretical analysis demonstrated the

optimality of the greedy method under conditions where the posterior knowledge of each linker is independent (Section 5.3).

Though the sequential greedy algorithm required model retraining at each iteration, the computational costs remained manageable in our implementation. While online variants of the prediction models could offer a trade-off between computational efficiency and statistical performance, we maintained the full retraining approach.

The greedy approach emphasizes exploitation (selecting promising experiments) over exploration (selecting uncertain experiments), departing from conventional strategies in Bayesian Optimization and Multi-Armed Bandits that typically balance both aspects. While greedy methods can potentially become fixated on initially promising candidates in traditional optimization settings, our distinct objective – maximizing the number of unique discoveries – naturally avoids this limitation. Upon discovering a successful crystallization for a linker, the algorithm automatically moves on to explore other candidates. Conversely, each negative result decreases the predicted success probability for the corresponding linker, naturally shifting focus to more promising alternatives. Indeed, theoretical analysis demonstrated the optimality of the greedy method under conditions where the posterior knowledge of each linker is independent (Section 5.3).

Though the sequential greedy algorithm required model retraining at each iteration, the computational costs remained manageable in our implementation. While online variants of the prediction models could offer a trade-off between computational efficiency and statistical performance, we maintained the full retraining approach.

Batch algorithm

The batch algorithm extended our greedy methodology to select groups of size N_B that collectively maximize the expected number of new discoveries. While finding the globally optimal batch would be computationally intractable due to combinatorial complexity, we developed an efficient sequential batch construction method that preserved the greedy principle.

For each batch, candidates were selected iteratively by computing their marginal contribution to the batch’s expected reward. This approach accounted for dependencies between candidates within the same batch, particularly when multiple candidates targeted the same linker. The algorithm is of the following.

The key challenge in batch selection is accurately estimating the marginal reward of adding a new reaction outcome to the current batch. For any candidate reaction $x_{\text{new}} = (x_l, x_e)$, where x_l denotes the linker feature and x_e the experimental feature, we compute the marginal reward $R(x_{\text{new}})$ as:

$$\begin{aligned}
 R(x_{\text{new}}) &= \mathbb{P}[\text{success with } x_{\text{new}}] - \mathbb{P}[\text{success without } x_{\text{new}}] \\
 &= \mathbb{P}[\exists x \in B_t^l \cup \{x_{\text{new}}\} : Y(x) = 1] - \mathbb{P}[\exists x \in B_t^l : Y(x) = 1] \\
 &= \mathbb{P}[Y(x_{\text{new}}) = 1 \mid \forall x \in B_t^l : Y(x) = 0] \cdot \mathbb{P}[\forall x \in B_t^l : Y(x) = 0]
 \end{aligned} \tag{5.1}$$

Algorithm 4 Batch Greedy Discovery Algorithm**Require:** Initial training data $\mathcal{D}_{\text{train}}$, batch size N_B

- 1: Train model $\mathcal{M}$ on $\mathcal{D}_{\text{train}}$
- 2: Initialize experiment space $\mathcal{X}$
- 3: **while** budget remains **do**
- 4: Initialize batch $\mathcal{B}_t \leftarrow \emptyset$
- 5: **for** $b \leftarrow 1$ **to** N_B **do**
- 6: Estimate expected reward contributions $R(x)$ for all $x \in \mathcal{X}$
- 7: Select $x_t^b \leftarrow \arg \max_{x \in \mathcal{X}} R(x)$
- 8: Update $\mathcal{B}_t \leftarrow \mathcal{B}_t \cup \{x_t^b\}$
- 9: Update R according to Eq. (5.1)
- 10: Remove x_t^b from $\mathcal{X}$
- 11: **end for**
- 12: Run experiments on $\mathcal{B}_t$ and observe outcomes $\{y_t^b\}$
- 13: Update $\mathcal{D}_{\text{train}} \leftarrow \mathcal{D}_{\text{train}} \cup \{(x_t^b, y_t^b)\}_{b=1}^{N_B}$
- 14: Retrain model $\mathcal{M}$ on updated $\mathcal{D}_{\text{train}}$
- 15: **end while**

where $\mathcal{B}_t^l$ represents the set of already-selected candidates in the current batch using the same linker l , and $Y(x)$ denotes the random outcome of candidate x . Hence, estimating the expected reward reduces to estimating two probabilities: a conditional probability and a joint probability. The joint probability can be computed using the chain rule, decomposing it into a product of conditional probabilities that can be estimated individually:

$$\mathbb{P}[\forall x \in \mathcal{B}_t^l : Y(x) = 0] = \prod_{k=1}^l \mathbb{P}[Y(x_t^{b_k}) = 0 \mid Y(x_t^{b_{k-1}}) = \dots = Y(x_t^{b_1}) = 0]. \quad (5.2)$$

To estimate the conditional probabilities, we can temporarily augment the training set with virtual observations of failures for previously selected candidates $\mathcal{B}_t^l$ and then retrain the model. While such estimation is exact for pure probabilistic models like Gaussian Process Classification, it serves as a practical heuristic for other prediction models like Random Forest Classification.

The computational cost remained similar to the sequential approach. Within each batch, selecting a candidate required updating the expected reward contributions for other candidates under the same linker. This update involved refitting the model with the training set augmented by virtual labels to estimate conditional probabilities. By storing these conditional probabilities, we efficiently computed the joint probabilities. The total number of model fits equaled the total suggested experiments, matching the sequential case.

Our batch algorithm naturally accommodated practical experimental constraints, such as the requirement for all reactions in a batch to share the same temperature due to equipment limitations. While the underlying greedy principle remained unchanged, we adapted the

algorithm by restricting the search space during batch construction. For example, to handle temperature constraints imposed by our experimental setup (where multiple reactions in a well plate shared an oven temperature), we evaluated the algorithm’s selected batch across a discrete set of temperatures and selected the configuration with the largest expected reward.

Prediction models

Our framework requires probabilistic classifiers that provide calibrated confidence estimates for the greedy selection strategy. We implemented two complementary approaches, each offering different trade-offs between computational efficiency, uncertainty quantification, and handling of sparse data regions. The models were selected based on their ability to handle high-dimensional sparse data (26 features, 8 training systems) and provide reliable probability estimates for out-of-distribution test cases.

Random Forest Classification

RFC combines predictions from multiple decision trees, each trained on bootstrap samples of the training data. We used the proportion of trees predicting success as our probability estimate. The RFC implementation in the scikit-learn library [105] was used in this study. Two key modifications were necessary to adapt RFC for our discovery process:

First, we addressed the substantial data imbalance between initial and newly collected data. While each starting linker had 336 samples, new linker data accumulated incrementally. We implemented a data augmentation procedure that injected multiple copies of newly collected data during iterative experimentation. The optimal degree of data augmentation depends on the chemical similarity between new and training linkers. When a new linker closely resembles a training linker, each new experimental result should carry similar weight to the training data, requiring minimal augmentation. Conversely, if a new linker’s synthesis mechanism differs substantially from all training linkers, the initial dataset is merely a distraction, necessitating stronger augmentation to prioritize learning from new experimental outcomes. We selected an augmentation number of 10, matching the ratio between the number of experiments for each initial (336) and new linker. This choice provided a balance between leveraging prior knowledge and adapting to novel chemical space. Additional results for other augmentation numbers can be found in Supplementary Fig. C.13.

Second, we implemented dynamic hyperparameter optimization to accommodate the evolving training set. Every 100 iterations we performed five-fold cross-validation on newly collected data: for each fold, we trained on all data (including the initial training set) except the validation fold, selecting the maximum allowed tree depth that minimized classification cross-entropy loss. We kept other hyperparameters as default in the scikit-learn library.

Gaussian Process Classification

GPC models the probability of success through a latent Gaussian process, providing naturally calibrated probability estimates with uncertainty quantification. Our initial dataset revealed distinct patterns: outcomes varied smoothly across reaction conditions for fixed linkers, but showed less continuity across linker features at fixed conditions, reflecting the underlying chemical complexity. To capture these characteristics, we designed a composite kernel function:

$$k(x, x') = k_{\text{RBF}}(x_e, x'_e) \cdot k_{\text{Mat52}}(x_l, x'_l), \quad (5.3)$$

where k_{RBF} represents a Radial Basis Function (RBF) kernel operating on experimental features x_e , capturing smooth variations across experimental features, and k_{Mat52} denotes a Matérn-5/2 kernel operating on linker features x_l , allowing less smooth transitions across molecular structures. We computed posterior probabilities and determined the hyperparameters using the variational inference approximation implemented in the GPyTorch library [46].

Theoretical results for the greedy method

In this section, we create a theoretical and mathematical sandbox to illustrate that the greedy objective is the right decision-making principle for this problem.

Setup and preliminaries

To distill the essence of greedy and optimistic methods, we consider a classical Bayesian bandit setting, which provides a fundamental framework for sequential decision-making under uncertainty. The name “bandit” originates from the analogy of a casino with multiple slot machines (one-armed bandits), where a gambler must decide which machine to play at each turn. Such settings arise naturally in scientific discovery, where a sequence of experiments must be conducted across multiple possible directions.

In our setup, each arm $i \in [k] = \{1, 2, \dots, k\}$ follows a Bernoulli distribution $P_i = \text{Bern}(\nu_i)$ with an unknown mean parameter $\nu_i \in [0, 1]$. A Bernoulli distribution models binary outcomes: when played, arm i yields either a success (1) with probability ν_i or a failure (0) with probability $1 - \nu_i$. In the Bayesian formulation, these mean parameters are themselves random variables, drawn independently from known prior distributions Q_i before the game begins, forming a product prior $Q = Q_1 \times \dots \times Q_k$. The Bayesian aspect of our setting comes from maintaining and updating beliefs about these unknown parameters through the prior distributions.

The classical multi-armed bandit literature typically considers two main objectives [72, 116]:

1. **Cumulative reward maximization:** The traditional objective is to maximize the sum of expected rewards over T rounds, $\mathbb{E} \left[\sum_{t=1}^T X_t \right]$, where X_t is the reward obtained

at time t . This objective models scenarios where each reward contributes to the final performance, such as in online advertising where each click generates value.

2. **Simple regret minimization:** Another common objective is to identify the arm with the highest mean, with the objective $\max_{i \in [k]} \nu_i$. This objective is relevant in property maximization or A/B testing, where the goal is to find the best option for future use rather than maximize immediate returns during the exploration. This objective is typically used in Bayesian Optimization.

In contrast to these classical objectives, our pure discovery reward is formulated as the following: the goal is to observe at least one success (outcome of 1) from as many arms as possible within a fixed horizon T . Formally, let τ_i denote the first time we observe a success on arm i , with $\tau_i = \infty$ if no success is observed. The reward after T rounds is then defined as $R_T = \sum_{i=1}^k \mathbf{1}\{\tau_i \leq T\}$, counting the number of arms from which we have observed at least one success. This objective naturally models scientific discovery scenarios where the goal is to make discoveries across multiple system.

At each round t , the agent selects an arm A_t based on the history of previous actions and observations, denoted by the filtration $\mathcal{F}_{t-1} = \sigma(A_1, X_1, \dots, A_{t-1}, X_{t-1})$. The outcome $X_t \sim P_{A_t}$ is then observed, and the process continues. A policy π is a mapping from histories to actions, and our goal is to find the Bayes-optimal policy π^* that maximizes $\mathbb{E}[R_T]$.

This setup creates a natural tension between two fundamental principles in sequential decision-making:

- **The Greedy principle:** Select the arm that appears most promising based on current information, maximizing immediate expected reward. In our setting, this means choosing the arm with the highest posterior mean among those that haven't yet yielded a success.
- **The Optimistic principle:** Select arms that have high potential but high uncertainty, explicitly accounting for the value of exploration. This principle underlies many algorithms in bandits and reinforcement learning, such as Thompson Sampling (TS) [113] and Upper Confidence Bound (UCB) methods.

As we show in the following section, the greedy strategy is surprisingly optimal in this setting, despite the apparent need for exploration. This result challenges the conventional wisdom that effective learning algorithms must explicitly account for the exploration-exploitation trade-off.

Main results

We establish our central theoretical result: the optimality of a greedy policy for the pure discovery objective in the Bayesian MAB framework. This result demonstrates that, perhaps surprisingly, a purely exploitative approach achieves Bayes-optimality for our reward

structure, contrasting with the conventional wisdom that effective bandit algorithms require explicit exploration.

We begin with an intuitive lemma about the behavior of posterior means after observing failures.

Lemma 5.3.1 (Decrease of Posterior Mean). *Let ν be a random variable with prior distribution Q supported over $[0, 1]$. Let $X \mid \nu \sim \text{Bern}(\nu)$ be a Bernoulli random variable. Then,*

$$\mathbb{E}[\nu \mid X = 0] \leq \mathbb{E}[\nu].$$

Proof. The proof follows from the law of total expectation and Jensen's inequality. By the definition of conditional expectation, we have:

$$\mathbb{E}[\nu \mid X = 0] = \frac{\mathbb{E}[\nu \cdot \mathbf{1}\{X = 0\}]}{\mathbb{P}(X = 0)}.$$

Expanding the numerator, we get:

$$\mathbb{E}[\nu \cdot \mathbf{1}\{X = 0\}] = \mathbb{E}[\mathbb{E}[\nu \cdot \mathbf{1}\{X = 0\} \mid \nu]] = \int \nu(1 - \nu)dQ(\nu) = \mathbb{E}[\nu] - \mathbb{E}[\nu^2].$$

Similarly, for the denominator, we have:

$$\mathbb{P}(X = 0) = \int \mathbb{P}(X = 0 \mid \nu)dQ(\nu) = \int (1 - \nu)dQ(\nu) = 1 - \mathbb{E}[\nu].$$

Combining these results and applying Jensen's inequality, we obtain:

$$\mathbb{E}[\nu \cdot \mathbf{1}\{X = 0\}] \leq \mathbb{P}(X = 0) \cdot \mathbb{E}[\nu],$$

which completes the proof. $\square$

This lemma establishes a key property: observing a failure can only decrease our estimate of an arm's success probability. We now present our main theorem:

Theorem 5.3.2 (Optimality of greedy policy). *Consider a k -armed Bayesian bandit problem where the prior factorizes into independent priors $Q = Q_1 \times \cdots \times Q_k$, with $\nu_i \sim Q_i$ for each arm i . Let $P_i = \text{Bern}(\nu_i)$, and define the reward as*

$$R_T = \sum_{i=1}^k \mathbf{1}\{\tau_i \leq T\},$$

where $\tau_i = \min\{t : A_t = i, X_t = 1\}$ is the first time a success is observed on arm i , or $\tau_i = \infty$ if no success is observed.

At each round t , let $\mathcal{I}_t = \{i : \tau_i > t\}$ denote the set of arms that have not yet succeeded, and let $\mathcal{F}_{t-1}$ denote the history of actions and observations up to time $t - 1$. Then, for any finite horizon T , the Bayes-optimal policy is greedy: at each round t , it selects the arm with the highest posterior mean among the unsucceeded arms:

$$A_t \in \arg \max_{i \in \mathcal{I}_t} \mathbb{E}[\nu_i \mid \mathcal{F}_{t-1}].$$

Proof. We prove the theorem by backward induction on t .

Base case ($t = T$). At the final round T , the policy has a single opportunity to gain a success from the remaining arms $\mathcal{I}_T$. The optimal choice is to select the arm with the highest posterior mean $\mathbb{E}[\nu_i \mid \mathcal{F}_{T-1}]$ to maximize the probability of success. This is trivially a greedy choice.

Inductive hypothesis. Assume that at some time $t + 1$, the greedy policy is optimal for all future rounds $t + 1, t + 2, \dots, T$. That is, at each round $t' \geq t + 1$, the policy selects the arm with the highest posterior mean among the unsucceeded arms.

Inductive step. Consider round t and let $(Q(t), \mathcal{I}_t)$ represent the current state, where $Q(t) = (Q_1(t), \dots, Q_k(t))$ are the posterior distributions of $\nu_1, \dots, \nu_k$ given the history $\mathcal{F}_{t-1}$, and $\mathcal{I}_t$ is the set of arms that have not yet succeeded. The value function at time t is given by:

$$V_t(Q(t), \mathcal{I}_t) = \max_{i \in \mathcal{I}_t} \left\{ \mathbb{E}_{Q_i(t)}[\nu_i] (1 + V_{t+1}(Q(t+1), \mathcal{I}_t \setminus \{i\})) + (1 - \mathbb{E}_{Q_i(t)}[\nu_i]) V_{t+1}(Q(t+1), \mathcal{I}_t) \right\},$$

where $V_{t+1}(Q(t+1), \cdot)$ represents the optimal expected reward from time $t + 1$ onward. The posterior $Q(t + 1)$ is updated based on the outcome of pulling arm i at time t .

Suppose, for the sake of contradiction, that the greedy policy is not optimal at time t . Let $i^* \in \arg \max_{j \in \mathcal{I}_t} \mathbb{E}[\nu_j \mid \mathcal{F}_{t-1}]$ be the greedy action, and suppose some other arm $i \neq i^*$ achieves strictly higher V_t . Note that, in this case, the optimal action at time $t + 1$ is picking i^* . If $X_t = 0$, then according to Lemma 5.3.1, the posterior mean of i decreases, and i^* remains the optimal arm. Conversely, if $X_t = 1$, then arm i is removed from the active arms and the posterior mean of all other arms remain unchanged.

Now, define a new policy that is identical to π in all respects except that it swaps the order of the actions at times t and $t + 1$: at time t , it picks i^* (the greedy arm), and at time $t + 1$, it picks arm i . Note that $\tilde{\pi}$ achieves the same expected reward as π . To be more detailed, denote p_i and p_{i^*} as the probability of success choosing i and i^* respectively. Since, two policies are identical starting from time $t + 2$, it suffices to compare the rewards obtained during times t and $t + 1$. The expected reward under policy π over these two time steps is:

$$V_\pi = p_i(1 + p_{i^*}) + (1 - p_i)p_{i^*} = p_i + p_{i^*},$$

which is identical to the reward of policy $\tilde{\pi}$:

$$V_{\tilde{\pi}} = p_{i^*}(1 + p_i) + (1 - p_{i^*})p_i = p_{i^*} + p_i.$$

According to the induction hypothesis, we know that the best action in $t + 1$ is greedy, hence the greedy policy which picks i^* at time t has an expected reward no smaller than $\tilde{\pi}$ and thus π , contradicting our assumption that π achieves a strictly higher value. Therefore, the greedy policy must be optimal at time t , completing the proof. $\square$

Sub-optimality of optimistic methods

In this section, we perform synthetic experiments to show that the optimistic principle can be highly suboptimal under our reward definition. We consider one of the most prominent

optimistic methods – Thompson Sampling (TS). TS is a classical Bayesian approach to the exploration-exploitation dilemma that maintains posterior distributions over the reward parameters of each arm. At each round, TS draws random samples from these posteriors and selects the arm with the highest sampled value. This randomized strategy naturally balances exploration and exploitation: arms with higher uncertainty have more variable samples, giving them opportunities to be selected even if their posterior means are not the highest. While TS is known to be asymptotically optimal for many traditional bandit objectives, we demonstrate that it can be substantially suboptimal for our pure discovery reward. We construct a synthetic experiment with $k = 300$ arms divided into two groups with carefully chosen Beta priors:

- 200 arms with Beta(2,4) priors, representing arms with moderate success probabilities ($\mathbb{E}[\nu] \approx 0.33$)
- 100 arms with Beta(601,301) priors, representing arms with higher expected success probabilities ($\mathbb{E}[\nu] \approx 0.67$) but much higher concentration

This setup creates a challenging scenario for TS. The second group of arms has higher means but much tighter distributions around these means. TS’s tendency to sample from the full posterior leads it to frequently explore the first group of arms due to their higher variance, even though the greedy strategy’s focus on posterior means would correctly prioritize the second group. Our experimental results, shown in Supplementary Fig 5.4, confirm that TS performs substantially worse than the greedy strategy in this setting. The key insight is that TS’s exploration, which is typically beneficial for cumulative reward maximization, becomes detrimental for the pure discovery objective. While TS spends time exploring arms from the first group due to their uncertainty, the greedy strategy efficiently accumulates successes by focusing on the higher-mean arms. This demonstrates that traditional exploration strategies, even sophisticated ones like TS, can be fundamentally misaligned with the pure discovery objective, validating our theoretical result about the optimality of greedy approaches in this setting.

Discussion

One limitation of the analysis above is assuming the independence of different arms. This often does not hold in practice, for instance, machine learning models leveraging the structural similarity of different linkers tend to lead to dependent posteriors. Nonetheless, it suffices to consider this simple setting to obtain the key insight: optimistic exploration strategies, often considered essential for bandit problems, can be substantially suboptimal for the discovery objective. A natural follow-up question would be extending the analysis to correlated arms, particularly in Gaussian Process bandits [116] where observations inform our beliefs about similar arms.

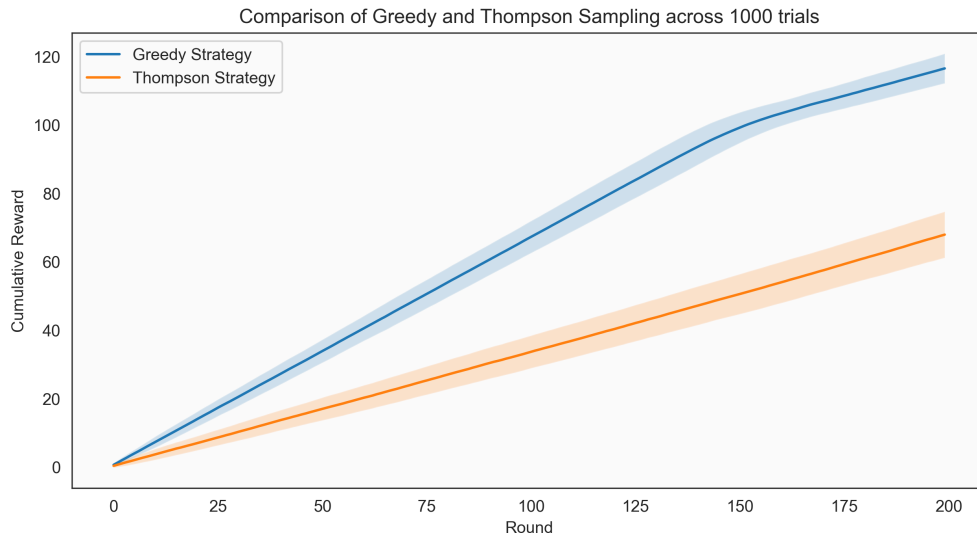


Figure 5.4: **Comparison of greedy and Thompson Sampling strategies over 1000 trials.** The solid lines show mean cumulative rewards (number of arms with observed successes), while the shaded regions represent one standard deviation. The greedy strategy significantly outperforms Thompson Sampling by focusing on arms with higher posterior means rather than being diverted by uncertainty in the lower-mean arms.

5.4 Conclusion and Outlook

As demonstrated in the preceding section, AIRES successfully navigated the complex ZIF chemical space to expand the known materials library by over 30%. This effort yielded not only a significant number of new 3D frameworks, including some with previously inaccessible or entirely novel topologies, but also unexpected layered structures (LZIFs), highlighting the platform’s efficacy in exploring syntheses with diverse and non-linear outcomes.

The AIRES platform demonstrates significant potential beyond the specific material class explored in this work. The core methodology is designed to be generalizable. The algorithmic framework, particularly the greedy discovery strategy and the batch selection algorithm, is readily transferable to other discovery problems. Future work could involve extending the system to more complex chemical spaces, such as syntheses involving multiple building blocks, where the spectrum of possible outcomes is far richer. This would require more sophisticated data labeling to classify a wider range of products, potentially integrating multi-modal characterization data (e.g., from techniques like powder X-ray diffraction and infrared spectroscopy) to provide a more complete picture of the reaction outcomes. While the feature engineering would need to be adapted for new chemical domains, the fundamental closed-loop approach remains the same.

AIRES is positioned to be a foundational tool in a larger materials development ecosystem. The validated structures it discovers serve as high-quality data points for training the

next generation of ML models, such as those that predict material properties from structure or learn direct synthesis-to-structure correlations. Equally important, the successful "hits" from AIRES provide the starting point for process optimization. The initial successful synthesis conditions discovered by the platform can seed a subsequent Bayesian Optimization loop aimed at scaling up production, maximizing yield, and improving material quality. In this way, AIRES bridges the gap between high-throughput discovery and targeted materials engineering, creating a more efficient pipeline from initial idea to final application.

Bibliography

- [1] Eugene L Allgower and Kurt Georg. *Numerical continuation methods: an introduction*. Vol. 13. Springer Science & Business Media, 2012.
- [2] Noga Alon, Yossi Matias, and Mario Szegedy. “The space complexity of approximating the frequency moments”. In: vol. 58. 1, part 2. Twenty-eighth Annual ACM Symposium on the Theory of Computing (Philadelphia, PA, 1996). 1999, pp. 137–147. DOI: [10.1006/jcss.1997.1545](https://doi.org/10.1006/jcss.1997.1545). URL: <https://doi.org/10.1006/jcss.1997.1545>.
- [3] Robert B Ash. *Information theory*. Courier Corporation, 2012.
- [4] IA Baburin, Stefano Leoni, and G Seifert. “Enumeration of not-yet-synthesized zeolitic zinc imidazolate MOF networks: a topological and DFT approach”. In: *The Journal of Physical Chemistry B* 112.31 (2008), pp. 9437–9443.
- [5] Jean-Bernard Baillon and Haïm Brézis. “Une remarque sur le comportement asymptotique des semigroupes non linéaires”. In: *Houston Journal of Mathematics* 2.1 (1976), pp. 5–7.
- [6] Stefan Banach. “Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales”. In: *Fundamenta mathematicae* 3.1 (1922), pp. 133–181.
- [7] Andrew Blake and Andrew Zisserman. *Visual reconstruction*. MIT press, 1987.
- [8] Harald Bohr. *Almost Periodic Functions*. Trans. by Harvey Cohn. Vol. 27. AMS Chelsea Publishing. New York: Chelsea Publishing Company, 1947, p. 113. ISBN: 082840027X, 9780828400275.
- [9] Jonathan M. Borwein and Herre Wiersma. “Asplund Decomposition of Monotone Operators”. In: *SIAM Journal on Optimization* 18.3 (2007), pp. 946–960. DOI: [10.1137/060658357](https://doi.org/10.1137/060658357).
- [10] Radu Ioan Bot and Dang Khoa Nguyen. “Tikhonov regularization of monotone operator flows not only ensures strong convergence of the trajectories but also speeds up the vanishing of the residuals”. In: *arXiv preprint arXiv:2406.00852* (2025). DOI: [10.48550/arXiv.2406.00852](https://doi.org/10.48550/arXiv.2406.00852). arXiv: [2406.00852](https://arxiv.org/abs/2406.00852) [math.OC].
- [11] Radu Ioan Bot and Dang-Khoa Nguyen. “Tikhonov regularization of monotone operator flows not only ensures strong convergence of the trajectories but also speeds up the vanishing of the residuals”. In: *arXiv preprint arXiv:2406.00852* (2024).

- [12] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities*. A nonasymptotic theory of independence, With a foreword by Michel Ledoux. Oxford University Press, Oxford, 2013, pp. x+481. ISBN: 978-0-19-953525-5. DOI: [10.1093/acprof:oso/9780199535255.001.0001](https://doi.org/10.1093/acprof:oso/9780199535255.001.0001). URL: <https://doi.org/10.1093/acprof:oso/9780199535255.001.0001>.
- [13] Olivier Bousquet. “Concentration Inequalities and Empirical Processes Theory Applied to the Analysis of Learning Algorithms”. PhD thesis. Ecole Polytechnique, 2002.
- [14] Leo Breiman. “Random forests”. In: *Machine learning* 45 (2001), pp. 5–32.
- [15] Felix E Browder. “Convergence of approximants to fixed points of nonexpansive nonlinear mappings in Banach spaces”. In: *Archive for Rational Mechanics and Analysis* 24.1 (1967), pp. 82–90.
- [16] Gavin Brown, Samuel B. Hopkins, and Adam D. Smith. “Fast, Sample-Efficient, Affine-Invariant Private Mean and Covariance Estimation for Subgaussian Distributions”. In: *CoRR* abs/2301.12250 (2023). DOI: [10.48550/arXiv.2301.12250](https://doi.org/10.48550/arXiv.2301.12250). arXiv: [2301.12250](https://doi.org/10.48550/arXiv.2301.12250). URL: <https://doi.org/10.48550/arXiv.2301.12250>.
- [17] Gavin Brown et al. “Covariance-Aware Private Mean Estimation Without Private Covariance Estimation”. In: *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*. Ed. by Marc’Aurelio Ranzato et al. 2021, pp. 7950–7964.
- [18] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.
- [19] Mengjie Chen, Chao Gao, and Zhao Ren. “Robust covariance and scatter matrix estimation under Huber’s contamination model”. In: *Ann. Statist.* 46.5 (2018), pp. 1932–1960. ISSN: 0090-5364. DOI: [10.1214/17-AOS1607](https://doi.org/10.1214/17-AOS1607). URL: <https://doi.org/10.1214/17-AOS1607>.
- [20] Yu Cheng, Ilias Diakonikolas, and Rong Ge. “High-Dimensional Robust Mean Estimation in Nearly-Linear Time”. In: *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019*. Ed. by Timothy M. Chan. SIAM, 2019, pp. 2755–2771. DOI: [10.1137/1.9781611975482.171](https://doi.org/10.1137/1.9781611975482.171). URL: <https://doi.org/10.1137/1.9781611975482.171>.
- [21] Yeshwanth Cherapanamjeri, Nicolas Flammarion, and Peter L. Bartlett. “Fast Mean Estimation with Sub-Gaussian Rates”. In: *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*. Ed. by Alina Beygelzimer and Daniel Hsu. Vol. 99. Proceedings of Machine Learning Research. PMLR, 2019, pp. 786–806. URL: <http://proceedings.mlr.press/v99/cherapanamjeri19b.html>.
- [22] Horațiu Cheval, Ulrich Kohlenbach, and Laurențiu Leuştean. “On modified halpern and tikhonov–mann iterations”. In: *Journal of Optimization Theory and Applications* 197.1 (2023), pp. 233–251.

- [23] Kenneth L. Clarkson et al. “Approximating Center Points with Iterated Radon Points”. In: *Proceedings of the Ninth Annual Symposium on Computational Geometry San Diego, CA, USA, May 19-21, 1993*. Ed. by Chee Yap. ACM, 1993, pp. 91–98. DOI: [10.1145/160985.161004](https://doi.org/10.1145/160985.161004). URL: <https://doi.org/10.1145/160985.161004>.
- [24] Roberto Cominetti, Juan Peypouquet, and Sylvain Sorin. “Strong asymptotic convergence of evolution equations governed by maximal monotone operators with Tikhonov regularization”. In: *Journal of Differential Equations* 245.12 (2008), pp. 3753–3763.
- [25] Juan Pablo Contreras and Roberto Cominetti. “Optimal error bounds for non-expansive fixed-point iterations in normed spaces”. In: *Mathematical Programming* 199.1 (2023), pp. 343–374.
- [26] P. L. Davies. “Asymptotic behaviour of S -estimates of multivariate location parameters and dispersion matrices”. In: *Ann. Statist.* 15.3 (1987), pp. 1269–1292. ISSN: 0090-5364. DOI: [10.1214/aos/1176350505](https://doi.org/10.1214/aos/1176350505). URL: <https://doi.org/10.1214/aos/1176350505>.
- [27] Jules Depersin and Guillaume Lecué. “On the robustness to adversarial corruption and to heavy-tailed data of the Stahel–Donoho median of means”. In: *Information and Inference: A Journal of the IMA* 12.2 (2022), pp. 814–850. DOI: [10.1093/imaiai/iaac026](https://doi.org/10.1093/imaiai/iaac026).
- [28] Jules Depersin and Guillaume Lecué. “Robust sub-Gaussian estimation of a mean vector in nearly linear time”. In: *Ann. Statist.* 50.1 (2022), pp. 511–536. ISSN: 0090-5364. DOI: [10.1214/21-aos2118](https://doi.org/10.1214/21-aos2118). URL: <https://doi.org/10.1214/21-aos2118>.
- [29] Jules Depersin and Guillaume Lecué. *Robust subgaussian estimation of a mean vector in nearly linear time*. 2019. arXiv: [1906.03058](https://arxiv.org/abs/1906.03058) [math.ST].
- [30] Ilias Diakonikolas and Daniel M Kane. “Recent advances in algorithmic high-dimensional robust statistics”. In: *arXiv preprint arXiv:1911.05911* (2019).
- [31] Ilias Diakonikolas, Daniel M. Kane, and Ankit Pensia. “Outlier Robust Mean Estimation with Subgaussian Rates via Stability”. In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. Ed. by Hugo Larochelle et al. 2020.
- [32] Ilias Diakonikolas et al. “Robust estimators in high dimensions without the computational intractability”. In: *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*. Ed. by Irit Dinur. IEEE Computer Society, 2016, pp. 655–664. ISBN: 978-1-5090-3933-3. DOI: [10.1109/FOCS.2016.85](https://doi.org/10.1109/FOCS.2016.85). URL: <https://doi.org/10.1109/FOCS.2016.85>.
- [33] Jelena Diakonikolas. “Halpern iteration for near-optimal and parameter-free monotone inclusion and strong solutions to variational inequalities”. In: *Conference on Learning Theory*. PMLR. 2020, pp. 1428–1451.

- [34] Nency P Domingues et al. “Using genetic algorithms to systematically improve the synthesis conditions of Al-PMOF”. In: *Communications Chemistry* 5.1 (2022), p. 170.
- [35] Yihe Dong, Samuel B. Hopkins, and Jerry Li. “Quantum Entropy Scoring for Fast Robust Mean Estimation and Improved Outlier Detection”. In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. Ed. by Hanna M. Wallach et al. 2019, pp. 6065–6075. URL: <https://proceedings.neurips.cc/paper/2019/hash/a4d92e2cd541fca87e4620aba658316d-Abstract.html>.
- [36] David Donoho and Peter J. Huber. “The notion of breakdown point”. In: *A Festschrift for Erich L. Lehmann*. Wadsworth Statist./Probab. Ser. Wadsworth, Belmont, CA, 1983, pp. 157–184.
- [37] David L Donoho. *Breakdown properties of multivariate location estimators*. Tech. rep. Technical report, Harvard University, Boston., 1982.
- [38] David L. Donoho and Miriam Gasko. “Breakdown properties of location estimates based on halfspace depth and projected outlyingness”. In: *Ann. Statist.* 20.4 (1992), pp. 1803–1827. ISSN: 0090-5364. DOI: [10.1214/aos/1176348890](https://doi.org/10.1214/aos/1176348890). URL: <https://doi.org/10.1214/aos/1176348890>.
- [39] David L. Donoho and Richard C. Liu. “The “automatic” robustness of minimum distance functionals”. In: *Ann. Statist.* 16.2 (1988), pp. 552–586. ISSN: 0090-5364. DOI: [10.1214/aos/1176350820](https://doi.org/10.1214/aos/1176350820). URL: <https://doi.org/10.1214/aos/1176350820>.
- [40] John C. Duchi, Saminul Haque, and Rohith Kudritipudi. “A Fast Algorithm for Adaptive Private Mean Estimation”. In: *CoRR* abs/2301.07078 (2023). DOI: [10.48550/arXiv.2301.07078](https://doi.org/10.48550/arXiv.2301.07078). arXiv: [2301.07078](https://arxiv.org/abs/2301.07078). URL: <https://doi.org/10.48550/arXiv.2301.07078>.
- [41] Bradley Efron et al. “Least Angle Regression”. In: *The Annals of Statistics* 32.2 (2004), pp. 407–499.
- [42] Francisco Facchinei and Jong-Shi Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer, 2003.
- [43] Farhad Fathieh et al. “Practical water production from desert air”. In: *Science advances* 4.6 (2018), eaat3198.
- [44] M. J. Frisch et al. *Gaussian~16 Revision A.03*. Gaussian Inc. Wallingford CT. 2016.
- [45] Hiroyasu Furukawa et al. “Water adsorption in porous metal–organic frameworks and related materials”. In: *Journal of the American Chemical Society* 136.11 (2014), pp. 4369–4381.
- [46] Jacob Gardner et al. “Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration”. In: *Advances in neural information processing systems* 31 (2018).

- [47] Ian Goodfellow et al. “Generative adversarial networks”. In: *Communications of the ACM* 63.11 (2020), pp. 139–144.
- [48] Eduard Gorbunov, Nicolas Loizou, and Gauthier Gidel. “Extragradient method: $O(1/k)$ last-iterate convergence for monotone variational inequalities and connections with cocoercivity”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pp. 366–402.
- [49] Eduard Gorbunov, Adrien Taylor, and Gauthier Gidel. “Last-iterate convergence of optimistic gradient method for monotone variational inequalities”. In: *Advances in neural information processing systems* 35 (2022), pp. 21858–21870.
- [50] P. M. Gruber and J. M. Wills, eds. *Handbook of convex geometry. Vol. A, B*. North-Holland Publishing Co., Amsterdam, 1993, Vol. A: lxvi+735 pp., Vol. B: pp. i–lxvi and 737–1438. ISBN: 0-444-89598-1.
- [51] Benjamin Halpern. “Fixed points of nonexpanding maps”. In: *Bulletin of the American Mathematical Society* (1967).
- [52] Boris Hasselblatt and Anatole Katok. *A first course in dynamics: with a panorama of recent developments*. Cambridge University Press, 2003.
- [53] Elad Hazan, Kfir Yehuda Levy, and Shai Shalev-Shwartz. “On graduated optimization for stochastic non-convex problems”. In: *International conference on machine learning*. PMLR. 2016, pp. 1833–1841.
- [54] Eduard Helly. “Über Mengen konvexer Körper mit gemeinschaftlichen Punkte.” In: *Jahresbericht der Deutschen Mathematiker-Vereinigung* 32 (1923), pp. 175–176. URL: <http://eudml.org/doc/145659>.
- [55] Thomas Holding and Ioannis Lestas. “Stability and Instability in Saddle Point Dynamics—Part I”. In: *IEEE Transactions on Automatic Control* 66.7 (2021), pp. 2933–2944. DOI: [10.1109/TAC.2020.3019375](https://doi.org/10.1109/TAC.2020.3019375).
- [56] Samuel B. Hopkins. “Mean estimation with sub-Gaussian rates in polynomial time”. In: *Ann. Statist.* 48.2 (2020), pp. 1193–1213. ISSN: 0090-5364. DOI: [10.1214/19-AOS1843](https://doi.org/10.1214/19-AOS1843). URL: <https://doi.org/10.1214/19-AOS1843>.
- [57] Samuel B. Hopkins, Gautam Kamath, and Mahbod Majid. “Efficient mean estimation with pure differential privacy via a sum-of-squares exponential mechanism”. In: *STOC ’22: 54th Annual ACM SIGACT Symposium on Theory of Computing, Rome, Italy, June 20 - 24, 2022*. Ed. by Stefano Leonardi and Anupam Gupta. ACM, 2022, pp. 1406–1417. DOI: [10.1145/3519935.3519947](https://doi.org/10.1145/3519935.3519947). URL: <https://doi.org/10.1145/3519935.3519947>.

- [58] Samuel B. Hopkins, Jerry Li, and Fred Zhang. “Robust and Heavy-Tailed Mean Estimation Made Simple, via Regret Minimization”. In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. Ed. by Hugo Larochelle et al. 2020. URL: <https://proceedings.neurips.cc/paper/2020/hash/8a1276c25f5efe85f0fc4020fbf5b4f8-Abstract.html>.
- [59] Samuel B. Hopkins et al. “Robustness Implies Privacy in Statistical Estimation”. In: *Proceedings of the 55th Annual ACM Symposium on Theory of Computing, STOC 2023, Orlando, FL, USA, June 20-23, 2023*. Ed. by Barna Saha and Rocco A. Servedio. ACM, 2023, pp. 497–506. DOI: [10.1145/3564246.3585115](https://doi.org/10.1145/3564246.3585115). URL: <https://doi.org/10.1145/3564246.3585115>.
- [60] Peter J. Huber. “Projection pursuit”. In: *Ann. Statist.* 13.2 (1985). With discussion, pp. 435–525. ISSN: 0090-5364. DOI: [10.1214/aos/1176349519](https://doi.org/10.1214/aos/1176349519). URL: <https://doi.org/10.1214/aos/1176349519>.
- [61] Peter J. Huber. “Robust covariances”. In: *Statistical decision theory and related topics, II (Proc. Sympos., Purdue Univ., Lafayette, Ind., 1976)*. Academic Press, New York, 1977, pp. 165–191.
- [62] Peter J. Huber. “Robust estimation of a location parameter”. In: *Ann. Math. Statist.* 35 (1964), pp. 73–101. ISSN: 0003-4851. DOI: [10.1214/aoms/1177703732](https://doi.org/10.1214/aoms/1177703732). URL: <https://doi.org/10.1214/aoms/1177703732>.
- [63] Peter J. Huber and Elvezio M. Ronchetti. *Robust statistics*. Second. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, 2009, xvi+354 pp. + loose erratum. ISBN: 978-0-470-12990-6. DOI: [10.1002/9780470434697](https://doi.org/10.1002/9780470434697). URL: <https://doi.org/10.1002/9780470434697>.
- [64] Haoyan Huo et al. “Machine-learning rationalization and prediction of solid-state synthesis conditions”. In: *Chemistry of Materials* 34.16 (2022), pp. 7323–7336.
- [65] Mark R. Jerrum, Leslie G. Valiant, and Vijay V. Vazirani. “Random generation of combinatorial structures from a uniform distribution”. In: *Theoret. Comput. Sci.* 43.2-3 (1986), pp. 169–188. ISSN: 0304-3975. DOI: [10.1016/0304-3975\(86\)90174-X](https://doi.org/10.1016/0304-3975(86)90174-X). URL: [https://doi.org/10.1016/0304-3975\(86\)90174-X](https://doi.org/10.1016/0304-3975(86)90174-X).
- [66] Dimo Kashchiev. *Nucleation*. Elsevier, 2000.
- [67] Hassan K Khalil and Jessy W Grizzle. *Nonlinear systems*. Vol. 3. Prentice hall Upper Saddle River, NJ, 2002.
- [68] Ulrich Kohlenbach. “On quantitative versions of theorems due to FE Browder and R. Wittmann”. In: *Advances in Mathematics* 226.3 (2011), pp. 2764–2795.
- [69] Galina M Korpelevich. “The extragradient method for finding saddle points and other problems”. In: *Matecon* 12 (1976), pp. 747–756.

- [70] Pravesh Kothari, Pasin Manurangsi, and Ameya Velingker. “Private Robust Estimation by Stabilizing Convex Relaxations”. In: *Conference on Learning Theory, 2-5 July 2022, London, UK*. Ed. by Po-Ling Loh and Maxim Raginsky. Vol. 178. Proceedings of Machine Learning Research. PMLR, 2022, pp. 723–777. URL: <https://proceedings.mlr.press/v178/kothari22a.html>.
- [71] Mark Aleksandrovich Krasnosel’skii. “Two remarks on the method of successive approximations”. In: *Uspekhi matematicheskikh nauk* 10.1 (1955), pp. 123–127.
- [72] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.
- [73] Michel Ledoux and Michel Talagrand. *Probability in Banach spaces*. Classics in Mathematics. Isoperimetry and processes, Reprint of the 1991 edition. Springer-Verlag, Berlin, 2011, pp. xii+480. ISBN: 978-3-642-20211-7.
- [74] Suchool Lee and Donghwan Kim. “Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 22588–22600.
- [75] Zhixian Lei et al. “A Fast Spectral Algorithm for Mean Estimation with Sub-Gaussian Rates”. In: *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*. Ed. by Jacob D. Abernethy and Shivani Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, 2020, pp. 2598–2612. URL: <http://proceedings.mlr.press/v125/lei20a.html>.
- [76] Laurențiu Leuştean. “Rates of asymptotic regularity for Halpern iterations of nonexpansive mappings”. In: *Journal of Universal Computer Science* 13.11 (2007), pp. 1680–1691.
- [77] Felix Lieder. “On the convergence rate of the Halpern-iteration”. In: *Optimization letters* 15.2 (2021), pp. 405–418.
- [78] Ivy H Lim, Wolfgang Schrader, and Ferdi Schüth. “Insights into the molecular assembly of zeolitic imidazolate frameworks by ESI-MS”. In: *Chemistry of Materials* 27.8 (2015), pp. 3088–3095.
- [79] Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. “A universal catalyst for first-order optimization”. In: *Advances in neural information processing systems* 28 (2015).
- [80] Jian-Bin Lin et al. “A scalable metal-organic framework as a durable physisorbent for carbon dioxide capture”. In: *Science* 374.6574 (2021), pp. 1464–1469.
- [81] Regina Y. Liu. “Data depth and multivariate rank tests”. In: *L₁-statistical analysis and related methods (Neuchâtel, 1992)*. North-Holland, Amsterdam, 1992, pp. 279–294.
- [82] Regina Y. Liu. “On a notion of data depth based on random simplices”. In: *Ann. Statist.* 18.1 (1990), pp. 405–414. ISSN: 0090-5364. DOI: [10.1214/aos/1176347507](https://doi.org/10.1214/aos/1176347507). URL: <https://doi.org/10.1214/aos/1176347507>.

- [83] Regina Y. Liu, Jesse M. Parelius, and Kesar Singh. “Multivariate analysis by data depth: descriptive statistics, graphics and inference”. In: *Ann. Statist.* 27.3 (1999). With discussion and a rejoinder by Liu and Singh, pp. 783–858. ISSN: 0090-5364. DOI: [10.1214/aos/1018031260](https://doi.org/10.1214/aos/1018031260). URL: <https://doi.org/10.1214/aos/1018031260>.
- [84] Regina Y. Liu and Kesar Singh. “Ordering directional data: concepts of data depth on circles and spheres”. In: *Ann. Statist.* 20.3 (1992), pp. 1468–1484. ISSN: 0090-5364. DOI: [10.1214/aos/1176348779](https://doi.org/10.1214/aos/1176348779). URL: <https://doi.org/10.1214/aos/1176348779>.
- [85] Xiyang Liu et al. “Robust and differentially private mean estimation”. In: *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*. Ed. by Marc’Aurelio Ranzato et al. 2021, pp. 3887–3901. URL: <https://proceedings.neurips.cc/paper/2021/hash/1fc5309ccc651bf6b5d22470f67561ea-Abstract.html>.
- [86] Hendrik P. Lopuhaä and Peter J. Rousseeuw. “Breakdown points of affine equivariant estimators of multivariate location and covariance matrices”. In: *Ann. Statist.* 19.1 (1991), pp. 229–248. ISSN: 0090-5364. DOI: [10.1214/aos/1176347978](https://doi.org/10.1214/aos/1176347978). URL: <https://doi.org/10.1214/aos/1176347978>.
- [87] Gábor Lugosi and Shahar Mendelson. “Mean estimation and regression under heavy-tailed distributions: A survey”. In: *Found. Comput. Math.* 19.5 (2019), pp. 1145–1190. DOI: [10.1007/s10208-019-09427-x](https://doi.org/10.1007/s10208-019-09427-x). URL: <https://doi.org/10.1007/s10208-019-09427-x>.
- [88] Gábor Lugosi and Shahar Mendelson. *Multivariate mean estimation with direction-dependent accuracy*. 2020. arXiv: [2010.11921](https://arxiv.org/abs/2010.11921) [math.ST].
- [89] Gábor Lugosi and Shahar Mendelson. “Robust multivariate mean estimation: the optimality of trimmed mean”. In: *Ann. Statist.* 49.1 (2021), pp. 393–410. ISSN: 0090-5364. DOI: [10.1214/20-AOS1961](https://doi.org/10.1214/20-AOS1961). URL: <https://doi.org/10.1214/20-AOS1961>.
- [90] Gábor Lugosi and Shahar Mendelson. “Sub-Gaussian estimators of the mean of a random vector”. In: *Ann. Statist.* 47.2 (2019), pp. 783–794. ISSN: 0090-5364. DOI: [10.1214/17-AOS1639](https://doi.org/10.1214/17-AOS1639). URL: <https://doi.org/10.1214/17-AOS1639>.
- [91] W Robert Mann. “Mean value methods in iteration”. In: *Proceedings of the American Mathematical Society* 4.3 (1953), pp. 506–510.
- [92] Aleksandr V Marenich, Christopher J Cramer, and Donald G Truhlar. “Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions”. In: *The Journal of Physical Chemistry B* 113.18 (2009), pp. 6378–6396.
- [93] Ricardo Antonio Maronna. “Robust M -estimators of multivariate location and scatter”. In: *Ann. Statist.* 4.1 (1976), pp. 51–67. ISSN: 0090-5364. URL: [http://links.jstor.org/sici?sici=0090-5364\(197601\)4:1%3C51:ROMLAS%3E2.0.CO;2-I&origin=MSN](http://links.jstor.org/sici?sici=0090-5364(197601)4:1%3C51:ROMLAS%3E2.0.CO;2-I&origin=MSN).

- [94] Bernard Martinet. “Brève communication. Régularisation d’inéquations variationnelles par approximations successives”. In: *Revue française d’informatique et de recherche opérationnelle. Série rouge* 4.R3 (1970), pp. 154–158.
- [95] Bernard Martinet. “Régularisation d’inéquations variationnelles par approximations successives. Rev. Française Informat”. In: *Recherche Opérationnelle* 4 (1970), pp. 154–158.
- [96] Flore Mekki-Berrada et al. “Two-step machine learning enables optimized nanoparticle synthesis”. In: *npj Computational Materials* 7.1 (2021), p. 55.
- [97] George J Minty. “Monotone (nonlinear) operators in Hilbert space”. In: (1962).
- [98] Seyed Mohamad Moosavi et al. “Capturing chemical intuition in synthesis of metal-organic frameworks”. In: *Nature communications* 10.1 (2019), p. 539.
- [99] Abdellatif Moudafi. “Viscosity approximation methods for fixed-points problems”. In: *Journal of mathematical analysis and applications* 241.1 (2000), pp. 46–55.
- [100] Arkadi S. Nemirovsky and David B. Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience Series in Discrete Mathematics. Translated from the Russian and with a preface by E. R. Dawson. John Wiley & Sons, Inc., New York, 1983, pp. xv+388. ISBN: 0-471-10345-4.
- [101] Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*. SIAM, 1994.
- [102] Hannu Oja. “Descriptive statistics for multivariate distributions”. In: *Statist. Probab. Lett.* 1.6 (1983), pp. 327–332. ISSN: 0167-7152. DOI: [10.1016/0167-7152\(83\)90054-8](https://doi.org/10.1016/0167-7152(83)90054-8). URL: [https://doi.org/10.1016/0167-7152\(83\)90054-8](https://doi.org/10.1016/0167-7152(83)90054-8).
- [103] Michael R. Osborne, Brett Presnell, and Berwin A. Turlach. “A New Approach to Variable Selection in Least Squares Problems”. In: *IMA Journal of Numerical Analysis* 20.3 (2000), pp. 389–403.
- [104] Jisun Park and Ernest K Ryu. “Exact optimal accelerated complexity for fixed-point iterations”. In: *International Conference on Machine Learning*. PMLR. 2022, pp. 17420–17457.
- [105] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.
- [106] Leonid Denisovich Popov. “A modification of the Arrow-Hurwicz method for search of saddle points”. In: *Mathematical notes of the Academy of Sciences of the USSR* 28.5 (1980), pp. 845–848.
- [107] Paul Raccuglia et al. “Machine-learning-assisted materials discovery using failed experiments”. In: *Nature* 533.7601 (2016), pp. 73–76.
- [108] CE Rasmussen and CKI Williams. *Gaussian processes for machine learning (Vol. 2, No. 3, p. 4)*. 2006.

- [109] R Tyrrell Rockafellar. “Monotone operators and the proximal point algorithm”. In: *SIAM journal on control and optimization* 14.5 (1976), pp. 877–898.
- [110] Rachel C Rohde et al. “High-temperature carbon dioxide capture in a porous material with terminal zinc hydride sites”. In: *Science* 386.6723 (2024), pp. 814–819.
- [111] P. Rousseeuw and V. Yohai. “Robust regression by means of S-estimators”. In: *Robust and nonlinear time series analysis (Heidelberg, 1983)*. Vol. 26. Lect. Notes Stat. Springer, New York, 1984, pp. 256–272. DOI: [10.1007/978-1-4615-7821-5_15](https://doi.org/10.1007/978-1-4615-7821-5_15). URL: https://doi.org/10.1007/978-1-4615-7821-5_15.
- [112] Peter Rousseeuw. “Multivariate estimation with high breakdown point”. In: *Mathematical statistics and applications, Vol. B (Bad Tatzmannsdorf, 1983)*. Reidel, Dordrecht, 1985, pp. 283–297.
- [113] Daniel J Russo et al. “A tutorial on thompson sampling”. In: *Foundations and Trends® in Machine Learning* 11.1 (2018), pp. 1–96.
- [114] Ernest K Ryu, Kun Yuan, and Wotao Yin. “Ode analysis of stochastic gradient methods with optimism and anchoring for minimax problems”. In: *arXiv preprint arXiv:1905.10899* (2019).
- [115] Shoham Sabach and Shimrit Shtern. “A first order method for solving convex bilevel optimization problems”. In: *SIAM Journal on Optimization* 27.2 (2017), pp. 640–660.
- [116] Bobak Shahriari et al. “Taking the human out of the loop: A review of Bayesian optimization”. In: *Proceedings of the IEEE* 104.1 (2015), pp. 148–175.
- [117] Werner A. Stahel. “Robuste Schätzungen. infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen”. de. Doctoral Thesis. Zürich: ETH Zurich, 1981. DOI: [10.3929/ethz-a-000231580](https://doi.org/10.3929/ethz-a-000231580).
- [118] Jaewook Suh, Jisun Park, and Ernest Ryu. “Continuous-time analysis of anchor acceleration”. In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 32782–32866.
- [119] Nathan J Szymanski et al. “An autonomous laboratory for the accelerated synthesis of novel materials”. In: *Nature* 624.7990 (2023), pp. 86–91.
- [120] Michel Talagrand. “New concentration inequalities in product spaces”. In: *Invent. Math.* 126.3 (1996), pp. 505–563. ISSN: 0020-9910. DOI: [10.1007/s002220050108](https://doi.org/10.1007/s002220050108). URL: <https://doi.org/10.1007/s002220050108>.
- [121] Michel Talagrand. “Sharper bounds for Gaussian and empirical processes”. In: *Ann. Probab.* 22.1 (1994), pp. 28–76. ISSN: 0091-1798. URL: [http://links.jstor.org/sici?sici=0091-1798\(199401\)22:1%3C28:SBFGAE%3E2.0.CO;2-W&origin=MSN](http://links.jstor.org/sici?sici=0091-1798(199401)22:1%3C28:SBFGAE%3E2.0.CO;2-W&origin=MSN).
- [122] Mingxing Tan and Quoc Le. “Efficientnetv2: Smaller models and faster training”. In: *International conference on machine learning* 139 (2021), pp. 10096–10106.
- [123] Jacopo Tomasi, Benedetta Mennucci, and Roberto Cammi. “Quantum mechanical continuum solvation models”. In: *Chemical reviews* 105.8 (2005), pp. 2999–3094.

- [124] Quoc Tran-Dinh and Yang Luo. “Halpern-type accelerated and splitting algorithms for monotone inclusions”. In: *arXiv preprint arXiv:2110.08150* (2021).
- [125] John W. Tukey. “Mathematics and the picturing of data”. In: *Proceedings of the International Congress of Mathematicians (Vancouver, B.C., 1974), Vol. 2*. Canad. Math. Congress, Montreal, Que., 1975, pp. 523–531.
- [126] Oleksandr Voznyy et al. “Machine learning accelerates discovery of optimal colloidal quantum dot synthesis”. In: *Acs Nano* 13.10 (2019), pp. 11122–11128.
- [127] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, Cambridge, 2019. DOI: [10.1017/9781108627771](https://doi.org/10.1017/9781108627771). URL: <https://doi.org/10.1017/9781108627771>.
- [128] Christopher KI Williams and David Barber. “Bayesian classification with Gaussian processes”. In: *IEEE Transactions on pattern analysis and machine intelligence* 20.12 (1998), pp. 1342–1351.
- [129] Rainer Wittmann. “Approximation of fixed points of nonexpansive mappings”. In: *Archiv der mathematik* 58.5 (1992), pp. 486–491.
- [130] Stephen J Wright. *Primal-dual interior-point methods*. SIAM, 1997.
- [131] Yilei Wu et al. “Universal machine learning aided synthesis approach of two-dimensional perovskites in a typical laboratory”. In: *Nature Communications* 15.1 (2024), p. 138.
- [132] Yunchao Xie et al. “Accelerate synthesis of metal–organic frameworks by a robotic platform and bayesian optimization”. In: *ACS Applied Materials & Interfaces* 13.45 (2021), pp. 53485–53491.
- [133] Hong-Kun Xu. “Viscosity approximation methods for nonexpansive mappings”. In: *Journal of Mathematical Analysis and Applications* 298.1 (2004), pp. 279–291.
- [134] Omar M Yaghi et al. “Reticular synthesis and the design of new materials”. In: *Nature* 423.6941 (2003), pp. 705–714.
- [135] Jingjing Yang et al. “Principles of designing extra-large pore openings and cages in zeolitic imidazolate frameworks”. In: *Journal of the American Chemical Society* 139.18 (2017), pp. 6448–6455.
- [136] Qing-Feng Yang et al. “A series of metal–organic complexes constructed from in situ generated organic amines”. In: *CrystEngComm* 10.11 (2008), pp. 1534–1541.
- [137] Víctor J. Yohai. “High breakdown-point and high efficiency robust estimates for regression”. In: *Ann. Statist.* 15.2 (1987), pp. 642–656. ISSN: 0090-5364. DOI: [10.1214/aos/1176350366](https://doi.org/10.1214/aos/1176350366). URL: <https://doi.org/10.1214/aos/1176350366>.
- [138] TaeHo Yoon and Ernest K Ryu. “Accelerated algorithms for smooth convex-concave minimax problems with $O(1/k^2)$ rate on squared gradient norm”. In: *International Conference on Machine Learning*. PMLR, 2021, pp. 12098–12109.

- [139] Taeho Yoon and Ernest K. Ryu. “Accelerated Algorithms for Smooth Convex-Concave Minimax Problems with $\mathcal{O}(1/k^2)$ Rate on Squared Gradient Norm”. In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 18–24 Jul 2021, pp. 12098–12109.
- [140] Zhiling Zheng et al. “Shaping the water-harvesting behavior of metal–organic frameworks aided by fine-tuned GPT models”. In: *Journal of the American Chemical Society* 145.51 (2023), pp. 28284–28295.
- [141] Zhiling Zheng et al. “Structural chemistry of zeolitic imidazolate frameworks”. In: *Inorganic Chemistry* 62.51 (2023), pp. 20861–20873.
- [142] Yijun Zuo. “Projection-based depth functions and associated medians”. In: *Ann. Statist.* 31.5 (2003), pp. 1460–1490. ISSN: 0090-5364. DOI: [10.1214/aos/1065705115](https://doi.org/10.1214/aos/1065705115). URL: <https://doi.org/10.1214/aos/1065705115>.
- [143] Yijun Zuo, Hengjian Cui, and Xuming He. “On the Stahel-Donoho estimator and depth-weighted means of multivariate data”. In: *Ann. Statist.* 32.1 (2004), pp. 167–188. ISSN: 0090-5364. DOI: [10.1214/aos/1079120132](https://doi.org/10.1214/aos/1079120132). URL: <https://doi.org/10.1214/aos/1079120132>.

Appendix A

Appendix for Chapter 3

A.1 A technical lemma

We first state a lemma that we will use throughout the manuscript:

Lemma (Restatement of [Lemma 3.2.1](#)). *Let $c, c_1, R, \alpha, \beta, \gamma > 0$ with $1 < \gamma\alpha \leq \beta$ and define*

$$\delta_n = \frac{\alpha}{n + \beta}, \quad n \in \mathbb{N}.$$

Assume the sequence $(e_n)_{n \geq 1}$ satisfies $e_1 \leq c_1 R$ and

$$e_{n+1} \leq (1 - \gamma\delta_n) \left(e_n + c(\delta_{n-1} - \delta_n) R \right), \quad n \geq 1. \quad (\text{A.1})$$

Then for every $n \geq 1$,

$$e_n \leq \frac{M}{n + \beta - 1}, \quad \text{where } M := \max\left\{\frac{c\alpha}{\gamma\alpha - 1}, c_1\beta\right\} R.$$

Proof. We prove by induction that $e_n \leq M/(n + \beta)$ for all $n \geq 0$.

Base case ($n = 1$). By the definition of M we have

$$e_1 \leq c_1 R \leq \frac{M}{\beta}.$$

Inductive step. Assume $e_n \leq M/(n - 1 + \beta)$ for some $n \geq 1$. Since $\delta_{n-1} - \delta_n = \frac{\alpha}{(n - 1 + \beta)(n + \beta)}$, recurrence (A.1) yields

$$\begin{aligned} e_{n+1} &\leq \left(1 - \frac{\gamma\alpha}{n + \beta}\right) \left(\frac{M}{n - 1 + \beta} + \frac{c\alpha R}{(n - 1 + \beta)(n + \beta)} \right) \\ &= \frac{M(1 - \gamma\alpha) + c\alpha R - \frac{\gamma\alpha c\alpha R}{n + \beta}}{(n - 1 + \beta)(n + \beta)} + \frac{M}{n + \beta}. \end{aligned}$$

By the definition of M ,

$$M \geq \frac{c\alpha R}{\gamma\alpha - 1} \implies M(1 - \gamma\alpha) + c\alpha R \leq 0.$$

Hence the first fraction is non-positive, giving $e_{n+1} \leq M/(n + \beta)$. Thus the inductive step closes and the claimed bound holds for all $n \geq 1$, implying $e_n = O(1/n)$. $\square$

Appendix B

Appendix for Chapter 4

B.1 Auxiliary Results

Here we list auxiliary results important to our analysis. The first is Helly's celebrated theorem [54] on convex intersections as stated in [50, Theorem 1.1, Chapter 2.1].

Theorem B.1.1 ([54, 50]). *Let $\mathcal{K}$ be a family of convex sets in $\mathbb{R}^d$, and suppose $\mathcal{K}$ is finite or each member of $\mathcal{K}$ is compact. If every $d + 1$ or fewer members of $\mathcal{K}$ have a common point, then there is a point common to all members of $\mathcal{K}$.*

We also require Bousquet's famous inequality on the suprema of empirical processes [13] which builds on prior results by Talagrand [121, 120].

Theorem B.1.2 ([13, 12]). *Let $X_1, \dots, X_n$ be independent identically distributed random vectors indexed by an index set $\mathcal{T}$. Assume that $\mathbb{E}[X_{i,s}] = 0$, and $X_{i,s} \leq 1$ for all $s \in \mathcal{T}$. Let $Z = \sup_{s \in \mathcal{T}} \sum_{i=1}^n X_{i,s}$, $\nu = 2\mathbb{E}Z + \sigma^2$ where $\sigma^2 = \sup_{s \in \mathcal{T}} \sum_{i=1}^n \mathbb{E}X_{i,s}^2$ is the wimpy variance. Let $\phi(u) = e^u - u - 1$ and $h(u) = (1 + u) \log(1 + u) - u$, for $u \geq -1$. Then for all $\lambda \geq 0$,*

$$\log \mathbb{E} e^{\lambda(Z - \mathbb{E}Z)} \leq \nu \phi(\lambda).$$

Also, for all $t \geq 0$,

$$\mathbb{P}\{Z \geq \mathbb{E}Z + t\} \leq e^{-\nu h(t/\nu)} \leq \exp\left(-\frac{t^2}{2(\nu + t/3)}\right).$$

We present the well-known Ledoux-Talagrand contraction inequality [73].

Theorem B.1.3 ([73, 12]). *Let $x_1, \dots, x_n$ be vectors whose real-valued components are indexed by $\mathcal{T}$, that is, $x_i = (x_{i,s})_{s \in \mathcal{T}}$. For each $i = 1, \dots, n$, let $\phi_i : \mathbb{R} \rightarrow \mathbb{R}$ be a 1-Lipschitz function such that $\phi_i(0) = 0$. Let $\varepsilon_1, \dots, \varepsilon_n$ be independent Rademacher random variables, and let $\Psi : [0, \infty) \rightarrow \mathbb{R}$ be a non-decreasing convex function. Then,*

$$\mathbb{E} \left[\Psi \left(\sup_{s \in \mathcal{T}} \sum_{i=1}^n \varepsilon_i \phi_i(x_{i,s}) \right) \right] \leq \mathbb{E} \left[\Psi \left(\sup_{s \in \mathcal{T}} \sum_{i=1}^n \varepsilon_i x_{i,s} \right) \right]$$

and

$$\mathbb{E} \left[\Psi \left(\frac{1}{2} \sup_{s \in \mathcal{T}} |*| \sum_{i=1}^n \varepsilon_i \phi_i(x_{i,s}) \right) \right] \leq \mathbb{E} \left[\Psi \left(\sup_{s \in \mathcal{T}} |*| \sum_{i=1}^n \varepsilon_i x_{i,s} \right) \right].$$

We will use the following simple corollary of the second conclusion in our proof.

Corollary B.1.4. *Assume the setting of Theorem B.1.3. Then,*

$$\mathbb{E} \left[\sup_{s \in \mathcal{T}} |*| \sum_{i=1}^n \varepsilon_i \phi_i(x_{i,s}) \right] \leq 2 \mathbb{E} \left[\sup_{s \in \mathcal{T}} |*| \sum_{i=1}^n \varepsilon_i x_{i,s} \right].$$

We also require the following simple lemma frequently in our analysis.

Lemma B.1.5. *Let $\Sigma \in \mathbb{R}^{d \times d}$ and $\Sigma \succ 0$. Then, we have:*

$$\forall x, \mathbb{R}^d : \|x\|_{\Sigma} = \max_{v \in \mathbb{S}^{d-1}} \frac{|\langle x, v \rangle|}{\sqrt{v^{\top} \Sigma v}}.$$

Proof. We have:

$$\|\Sigma^{-1/2}x\| = \max_{u \in \mathbb{S}^{d-1}} \langle u, \Sigma^{-1/2}x \rangle = \max_{v \in \mathbb{S}^{d-1}} |*| \left\langle \frac{\Sigma^{1/2}v}{\|\Sigma^{1/2}v\|}, \Sigma^{-1/2}x \right\rangle = \max_{v \in \mathbb{S}^{d-1}} \frac{|\langle x, v \rangle|}{\sqrt{v^{\top} \Sigma v}}$$

concluding the proof of the lemma. $\square$

Next, we relate the PSD ordering of two matrices to their induced Mahalanobis norms.

Lemma B.1.6. *Let $\Sigma_1, \Sigma_2 \in \mathbb{R}^{d \times d}$ and $\Sigma_1 \succcurlyeq \Sigma_2 \succ 0$. Then,*

$$\forall x \in \mathbb{R}^d : \|x\|_{\Sigma_1} \leq \|x\|_{\Sigma_2}.$$

Proof. Observe

$$\Sigma_1 \succcurlyeq \Sigma_2 \implies \Sigma_2^{-1/2} \Sigma_1 \Sigma_2^{-1/2} \succcurlyeq I \implies \Sigma_2^{1/2} \Sigma_1^{-1} \Sigma_2^{1/2} \preccurlyeq I \implies \Sigma_1^{-1} \preccurlyeq \Sigma_2^{-1}$$

yielding the lemma by the definition of $\|\cdot\|_{\Sigma}$. $\square$

Lemma B.1.7. *Let $k \in \mathbb{N}$ and D be a distribution over $[k]$ with probabilities $\{p_i\}_{i=1}^k$ with $p_i \geq 0$ and $\sum_{i=1}^k p_i = 1$. Furthermore, let $n \in \mathbb{N}$ and $\mathbf{X} = \{X_1, \dots, X_n\}$ be generated i.i.d. from D . For any $j \in [k]$ and thresholds $\{\tau_i\}_{i=1}^n$ and τ with $\tau_i, \tau > 0$, define the events*

$$\begin{aligned} A &:= \{\mathbf{X} \in [k]^n : m_j(\mathbf{X}) \geq \tau_j\} \\ B &:= \left\{ \mathbf{X} \in [k]^n : \sum_{i=1, i \neq j}^k \mathbf{1}\{m_i(\mathbf{X}) \leq \tau_i\} \geq \tau \right\} \\ &\text{where } m_i(\mathbf{X}) = \sum_{l=1}^n \mathbf{1}\{X_l = i\}. \end{aligned}$$

Then, we must have:

$$\mathbb{P}(B \mid A) \geq \mathbb{P}(B).$$

Proof. Note that

$$\mathbb{P}(B \mid A) - \mathbb{P}(B) = \mathbb{P}(A^c) \cdot (\mathbb{P}(B \mid A) - \mathbb{P}(B \mid A^c)).$$

Furthermore, we have:

$$\begin{aligned} \mathbb{P}(B \mid A) &= \frac{1}{\sum_{t \geq \tau_j} \mathbb{P}(m_j(\mathbf{X}) = t)} \sum_{t \geq \tau_j} \mathbb{P}(m_j(\mathbf{X}) = t) \mathbb{P}(B \mid m_j(\mathbf{X}) = t) \\ \mathbb{P}(B \mid A^c) &= \frac{1}{\sum_{t < \tau_j} \mathbb{P}(m_j(\mathbf{X}) = t)} \sum_{t < \tau_j} \mathbb{P}(m_j(\mathbf{X}) = t) \mathbb{P}(B \mid m_j(\mathbf{X}) = t). \end{aligned} \quad (\text{B.1})$$

Moreover, observe by symmetry:

$$\begin{aligned} \mathbb{P}(B \mid m_j(\mathbf{X}) = t) &= \mathbb{P}(B \mid T_t) \\ \text{where } T_t &:= \{\mathbf{X} : X_n, X_{n-1}, \dots, X_{n-t+1} = j \text{ and } \forall l \leq n-t : X_l \neq j\} \end{aligned} \quad (\text{B.2})$$

and that conditioned on T_t , X_l for $l \leq n-t$ are drawn i.i.d. from D conditioned on $[k] \setminus \{j\}$. Next, we show for all $t_1 \geq t_2$

$$\mathbb{P}(B \mid T_{t_1}) \geq \mathbb{P}(B \mid T_{t_2}). \quad (\text{B.3})$$

Let D_j denote the conditional distribution of D conditioned on $[k] \setminus \{j\}$ and $Y_1, \dots, Y_n$ be i.i.d. draws from D_j . Then define

$$\mathbf{X}^t = \{X_1^t, \dots, X_n^t\} \text{ with } X_i^t = \begin{cases} Y_i & \text{if } i \leq n-t \\ j & \text{otherwise} \end{cases}.$$

Note that $\mathbf{X}^t$ is identical in distribution to $\mathbf{X}$ conditioned on T_t . Furthermore, note that by definition $\mathbf{X}^{t_2} \in B$ implies $\mathbf{X}^{t_1} \in B$. Hence, Eq. (B.3) is established. This, by Eq. (B.2), also establishes

$$\forall t_1 \geq t_2 : \mathbb{P}(B \mid m_j(\mathbf{X}) = t_1) \geq \mathbb{P}(B \mid m_j(\mathbf{X}) = t_2)$$

which yields the lemma by Eq. (B.1). □

B.2 Relationship between Corruption Models

In this section, we compare the corruption models used in the upper bounds of [Theorem 4.1.1](#), and the lower bounds of [Theorems 4.1.3](#) and [4.1.4](#). In the former, we used the strong contamination model where an adversary is allowed to inspect the generated sample and arbitrarily tamper with an η -fraction of them while in the latter results established under the Huber contamination model, the adversary supplies a contaminating distribution, P , such that each data point is drawn from a mixture distribution $(1 - \eta)D + \eta P$. Informally, the strong contamination model allows an adversary to explicitly inspect samples generated from D and carefully choose adversarial points tailored to the specific sample while in the Huber setting, the adversarial points are chosen independently of the clean samples. While the randomness of the mixture model might lead to the drawing of more than ηn samples in the Huber setting, this is not significant. This intuition is formalized in [Proposition B.2.6](#) where it is shown that a strong contamination adversary can effectively simulate a Huber contamination adversary with negligibly larger corruption factor.

We start with the definition of the strong contamination model which is the contamination model used in [Theorem 4.1.1](#) and as we prove in [Proposition B.2.6](#), subsumes several classical contamination models. This model is the main focus of most recent work on robust estimation (see, for example, [\[28, 89, 58, 31\]](#)).

Definition B.2.1. For $\eta \in [0, 1]$, and sample set, $\widetilde{\mathbf{X}} = \{\widetilde{X}_1, \dots, \widetilde{X}_n\}$, we say $\mathbf{X} = \{X_1, \dots, X_n\}$ is η -strongly contaminated version of $\widetilde{\mathbf{X}}$ if:

$$|\ast| \left\{ i : X_i \neq \widetilde{X}_i \right\} \leq \eta n.$$

Furthermore, we say an adversary, $\mathcal{A}$, is an η -strong contamination adversary if $\mathcal{A}(\widetilde{\mathbf{X}})$ is a η -strongly contaminated version of $\widetilde{\mathbf{X}}$.

We reproduce the definition of the Huber contamination model below [\[63\]](#).

Definition B.2.2. For $\eta \in [0, 1]$, and a distribution, $\widetilde{D}$, we say D is an η -Huber contaminated version of $\widetilde{D}$ if:

$$D = (1 - \eta)\widetilde{D} + \eta P$$

for some distribution P .

We also recall the notion of *strong* Huber contamination where the adversary is allowed to replace an η -fraction of the distribution with an arbitrary adversarial distribution, as opposed to simply mixing in an adversarial distribution with the original one [\[63\]](#).

Definition B.2.3. For $\eta \in [0, 1]$, and a distribution, $\widetilde{D}$, we say D is an η -Huber strongly contaminated version of $\widetilde{D}$ if:

$$\text{TV}(\widetilde{D}, D) \leq \eta.$$

for some distribution P .

Remark B.2.4. The strong Huber contamination model of [Definition B.2.3](#) subsumes the weaker notion of corruption in [Definition B.2.2](#).

We first show that the adversarial contamination model assumed in [Theorem 4.1.1](#) subsumes the Huber contamination model assumed in [Theorems 4.1.3](#) and [4.1.4](#). Hence, the algorithm from [Theorem 4.1.1](#) is applicable to the weaker Huber contamination model while our lower bounds also apply to algorithms only resilient to the Huber model. We will utilize Bernstein's inequality in the proof, stated here for convenience.

Theorem B.2.5. [[12](#), [Theorem 6.12](#)] *Let $X_1, \dots, X_n$ be n i.i.d. Bernoulli random variables with probability p ; i.e we have $\mathbb{P}\{X_i = 1\} = p$, $\mathbb{P}\{X_i = 0\} = 1 - p$. Then, for all $t \geq 0$:*

$$\mathbb{P}\left\{\sum_{i=1}^n X_i - np \geq t\right\} \leq \exp\left(-\frac{t^2}{2np + 2t/3}\right).$$

In the following proposition, we show that the Huber contamination model with corruption fraction η can be simulated by a strong contamination adversary with corruption factor negligibly more than η .

Proposition B.2.6. *Let $\delta \in (0, 1/2)$, $\tilde{D}$ be any distribution over $\mathbb{R}^d$, and D be any η -Huber strongly contaminated version of D . That is:*

$$\text{TV}(\tilde{D}, D) \leq \eta.$$

Then there exists an η' -strong contamination adversary $\mathcal{A}$ such that given $\tilde{\mathbf{X}} = \{\tilde{X}_i\}_{i=1}^n$ i.i.d. samples from $\tilde{D}$ satisfies for $\mathbf{X} = \{X_i\}_{i=1}^n$ being i.i.d. samples from D :

$$\text{TV}(\mathcal{A}(\tilde{\mathbf{X}}), \mathbf{X}) \leq \delta \text{ with } \eta' = \left(1 + 4 \max\left(\frac{\log(1/\delta)}{n\eta}, \sqrt{\frac{\log(1/\delta)}{n\eta}}\right)\right) \eta.$$

Proof. Note that we have for some distributions D', P, Q :

$$\begin{aligned}\tilde{D} &= (1 - \eta)D' + \eta P \\ D &= (1 - \eta)D' + \eta Q.\end{aligned}$$

Consider the following coupling between the distributions $\tilde{D}$ and D . A sample $(\tilde{X}, X)$ with $\tilde{X}$ and X distributed according to $\tilde{D}$ and D respectively as follows:

1. First draw $W \in \{0, 1\}$ i.i.d. with $\mathbb{P}\{W = 1\} = \eta$.
2. If $W = 0$, draw $X \sim D'$ and generate $(\tilde{X} = X, X)$.
3. Otherwise, draw $\tilde{X}, X$ independently from distributions P and Q respectively.

Consider now a sequence of n samples generated $(\widetilde{\mathbf{X}} = \{\widetilde{X}_i\}_{i=1}^n, \mathbf{X} = \{X_i\}_{i=1}^n)$ with $(\widetilde{X}_i, X_i)$ drawn i.i.d. from the coupling just described. Denote the distribution over $(\widetilde{\mathbf{X}}, \mathbf{X})$ as C . Note that C is a joint distribution over pairs of i.i.d. datasets. Consider, now, the following adversary where $C(\mathbf{X} \mid \widetilde{\mathbf{X}})$ denotes the conditional distribution of $\mathbf{X}$ given $\widetilde{\mathbf{X}}$:

1. Given $\widetilde{\mathbf{X}}$, draw $\mathbf{X}$ from $C(\mathbf{X} \mid \widetilde{\mathbf{X}})$.

2. Output $\mathbf{X}$ if

$$\sum_{i=1}^n \mathbf{1}\{X_i \neq \widetilde{X}_i\} \leq \eta' n.$$

3. Otherwise, output $\widetilde{\mathbf{X}}$.

It is clear from the definition that $\mathcal{A}$ is an η' -strong contamination adversary. We now show that they indeed satisfy the conclusions of the lemma. Observe first that the pair $(\widetilde{\mathbf{X}}, \mathbf{X})$ are distributed according to C . Let $\widetilde{\mathbf{X}} = \{\widetilde{X}_i\}_{i=1}^n$ and $\mathbf{X} = \{X_i\}_{i=1}^n$ with $(\widetilde{X}_i, X_i)$ distributed according to the coupling. Furthermore, let W_i denote the Bernoulli random variables drawn during their generation. We now have:

$$\sum_{i=1}^n \mathbf{1}\{X_i \neq \widetilde{X}_i\} \leq \sum_{i=1}^n W_i.$$

Finally, we have from our setting of η' and [Theorem B.2.5](#) that

$$\mathbb{P}\left\{\mathcal{A}(\widetilde{\mathbf{X}}) \neq \mathbf{X}\right\} = \mathbb{P}\left\{\sum_{i=1}^n \mathbf{1}\{X_i \neq \widetilde{X}_i\} > \eta' n\right\} \leq \mathbb{P}\left\{\sum_{i=1}^n W_i \geq \eta' n\right\} \leq \delta$$

concluding the proof of the lemma. □

B.3 Comparison to Classical Breakdown Point

Here, we compare classical notions of the breakdown point to the notions used in this paper. In the classical setting, the breakdown point of an estimator is defined for *fixed* pointsets, whereas our paper is concerned with the distributional setting where the pointset is generated from a distribution. The main goal of classical work was to characterize the robustness properties of various robust estimators *independently* of their statistical properties. Hence, they do not feature the statistical setting where the points are sampled from a population distribution. However, recent work in robust statistics has aimed to characterize the statistical rates achieved by various estimators in addition to their robustness properties. Therefore, they feature a setting where the data is generated from a distribution as opposed to a fixed point set that is fully revealed to an estimator. The (distributional) Huber contamination model (Definition B.2.2) generalization was proposed by Chen, Gao, and Ren [19]. However, most recent work (see, for example, [29, 89, 31, 58]) has focused on the stronger contamination model (Definition B.2.1) which subsumes this model as shown in Proposition B.2.6.

In this section, we show that even under the classical setting with *no* statistical error and the point set is fully revealed to the estimator, no affine equivariant estimator can tolerate more than $1/d$ fraction of corruptions. The first result shows that with no restriction on the pointset, no affine-equivariant estimator can tolerate more than $1/d$ fraction of outliers (Proposition B.3.3) on datasets where the inliers are *bounded*. Our next result shows that even when the dataset (again, with bounded inliers) is in general position, a condition under which estimators such as the Stahel-Donoho median achieve breakdown point $1/2$, their error in estimating the mean of the inlier pointset may be arbitrarily large (Proposition B.3.5).

Below, we recall the classical definition of the breakdown point from [63] where we use the weakest corruption model where the adversary is only allowed to *add* points to the dataset.

Definition B.3.1. [63, Definition 11.1] The (contamination) breakdown point of an estimator $\hat{\mu}$ for a dataset $\mathbf{X} = \{x_1, \dots, x_n\}$ is defined as follows:

$$\sup \left\{ \eta : \sup_{\mathbf{Y}: |\mathbf{Y}| \leq \eta n} \|\hat{\mu}(\mathbf{X} \cup \mathbf{Y}) - \hat{\mu}(\mathbf{X})\| < \infty \right\}$$

We will denote the breakdown point of $\hat{\mu}$ on $\mathbf{X}$ as $\text{Break}(\hat{\mu}; \mathbf{X})$.

As previously remarked, the primary distinction between the classical notion above and the distributional ones in Definitions B.2.1 to B.2.3 is that the classical version features a *fixed* pointset entirely revealed to the estimator whereas in the distributional setting, they only access samples. The distributional setting further incurs statistical error due to the sampling of the points in addition to the error caused by the adversary. Note, that any estimator for the distributional setting straightforwardly generalizes to the classical setting.

Remark B.3.2. Suppose $\mathbf{Z} = \mathbf{X} \cup \mathbf{Y}$ featuring the inliers $\mathbf{X} = \{X_i\}_{i=1}^n$ and $\mathbf{Y} = \{Y_i\}_{i=1}^{\eta n}$. Then, any estimator $\hat{\mu}$ that is robust to a η -Huber contamination adversary (Definition B.2.2)

may be used for the classical setting by running it on a pointset $\mathbf{Z}' = \{Z_1, \dots, Z_{n'}\}$ generated i.i.d. from $\mathbf{Z}$.

Note, furthermore, that the breakdown point is established for the *Euclidean* norm in line with classical results. However, this implies a corresponding bound on the breakdown point for the Mahalanobis norm as:

$$\|x\| \leq \|\Sigma\| \|x\|_\Sigma.$$

Since our inliers are bounded, $\|\Sigma\| \leq 1$ and hence, our results apply to the Mahalanobis norm.

In what follows, we show that for any affine-equivariant estimator, there exists a *bounded* dataset with breakdown point at most $1/d$. In fact, the example will comprise of a set of d points with their (Euclidean) norm bounded by 1. This can be scaled to a arbitrary number of points simply by repeating the points in the construction.

Proposition B.3.3. *For any $d \in \mathbb{N}$ and affine-equivariant estimator $\hat{\mu}$, there exists a dataset $\mathbf{X} = \{x_i\}_{i=1}^d$ with $\|x_i\| \leq 1$ for all $i \in [d]$ such that:*

$$\text{Break}(\hat{\mu}; \mathbf{X}) \leq \frac{1}{d}.$$

Proof. Consider the dataset $\mathbf{Z} = \{\mathbf{0}, e_1, \dots, e_d\}$. We first consider the case where $\hat{\mu}(\mathbf{Z}) \neq \mathbf{0}$. Without loss of generality we assume $\langle \hat{\mu}(\mathbf{Z}), e_1 \rangle = \gamma \neq 0$ and choose $\mathbf{X} = \{\mathbf{0}, e_2, \dots, e_d\}$. Now, for any r , consider the affine map f_r defined as follows:

$$f_r(x)_i := \begin{cases} x_i & \text{if } i > 1 \\ \frac{r}{\gamma} x_i & \text{if } i = 1 \end{cases}.$$

Then, we have for the dataset $\mathbf{X} \cup \mathbf{Y} := \{\mathbf{0}, \frac{r}{\gamma} e_1, e_2, e_3, \dots, e_d\}$ where $\mathbf{Y} := \{\frac{r}{\gamma} e_1\}$:

$$\hat{\mu}(\mathbf{X} \cup \mathbf{Y}) = \hat{\mu}(f_r(\mathbf{Z})) = f_r(\hat{\mu}(\mathbf{Z})).$$

Furthermore, by considering the change of estimate along e_1 , we obtain that for any r :

$$\|\hat{\mu}(\mathbf{X} \cup \mathbf{Y}) - \hat{\mu}(\mathbf{X})\| \geq |r - \hat{\mu}(\mathbf{X})_1|.$$

Hence, since r is unbounded, we obtain from [Definition B.3.1](#):

$$\text{Break}(\hat{\mu}; \mathbf{X}) \leq \frac{1}{d}.$$

In the alternative case where $\hat{\mu}(\mathbf{Z}) = \mathbf{0}$, we set $\mathbf{X} = \{e_1, e_2, \dots, e_d\}$ and consider the affine map:

$$g_r(x) := r(\langle \mathbf{1}, x \rangle - 1)\mathbf{1} + x.$$

Then, we have for the dataset $\mathbf{X} \cup \mathbf{Y} := \{-r\mathbf{1}, e_1, e_2, \dots, e_d\}$ where $\mathbf{Y} := \{-r\mathbf{1}\}$:

$$\hat{\mu}(\mathbf{X} \cup \mathbf{Y}) = \hat{\mu}(g_r(\mathbf{Z})) = g_r(\hat{\mu}(\mathbf{Z})) = -r\mathbf{1},$$

which concludes the proof since r can be chosen arbitrarily. $\square$

Classical results on the Stahel-Donoho median [37] establish that it achieves *classical* breakdown points approaching $1/2$ when the data is in *general position*. This notion is defined below following the exposition in [60]:

Definition B.3.4 ([60]). For $n, d \in \mathbb{N}$ with $n > d$, we say that a dataset $\mathbf{X} = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ is general position if no subset of $d + 1$ points from $\mathbf{X}$ lie in a $(d - 1)$ -dimensional affine subspace.

However, this notion does not provide quantitative *finite* error bounds on the Stahel-Donoho estimator even for pointsets in general position. In the next result, we show that the *quantitative* error of the Stahel-Donoho (and any affine equivariant estimator) may be arbitrarily large simply by allowing an adversary to *add* n/d additional corrupted points even in the restricted setting with bounded inliers and the points in general position.

Proposition B.3.5. For any $r > 1$, $n, d \in \mathbb{N}$ with $n > d$, and affine-equivariant estimator $\hat{\mu}$, there exists a datasets $\mathbf{X} = \{x_i\}_{i=1}^l, \mathbf{Y} = \{y_i\}_{i=1}^m \subset \mathbb{R}^d$ with $m \leq n/(d + 1)$, $l + m = n$, and for all $i \in [l]$, $\|x_i\| \leq 1$, such that:

$$\|\hat{\mu}(\mathbf{X} \cup \mathbf{Y}) - \mu(\mathbf{X})\| \geq r.$$

Furthermore, $\mathbf{X}$ is in general position.

Proof. Let $p = n/(d + 1)$ and generate $\mathbf{X}' = \{X_i\}_{i=1}^n$ as follows with $\text{Unif}(\mathbb{B}(c, r))$ denoting the uniform distribution over the d -dimensional ball centered at c with radius r :

$$X_i = \begin{cases} \mathbf{0} + u_i & \text{if } j = 0 \\ \frac{1}{2}e_j + u_i & \text{if } j \geq 1 \end{cases} \text{ where } j = \left\lceil \frac{i}{p} \right\rceil - 1 \text{ and } u_i \sim \text{Unif}(\mathbb{B}(\mathbf{0}, 1/(32rd))).$$

Note that with probability 1, $\mathbf{X}'$ are in general position.

We first consider the case where there exists j such that $\hat{\mu}(\mathbf{X}')_j > \frac{1}{4d}$. Without loss of generality assume $j = 1$, and consider the affine mapping for any $r > 0$:

$$\forall x \in \mathbb{R}^d, i \in [d] : f_r(x)_i = \begin{cases} x_i & \text{if } i > 1 \\ 8rd \cdot x_i & \text{if } i = 1 \end{cases}.$$

Then, by the affine-equivariance of $\hat{\mu}$ for $f_r(\mathbf{X}') = \{f_r(X_i)\}_{i=1}^n$ and observing

$$\hat{\mu}(f_r(\mathbf{X}')) = f_r(\hat{\mu}(\mathbf{X}')),$$

we have that $\hat{\mu}(f_r(\mathbf{X}'))_1 \geq 2r$. Now, consider the point sets $\mathbf{X} = \{f_r(X_i)\}_{i \leq p} \cup \{f_r(X_i)\}_{i \geq 2p+1}$ and $\mathbf{Y} = \{f_r(X_i)\}_{p < i \leq 2p}$ so $\mathbf{X} \cup \mathbf{Y} = f_r(\mathbf{X}')$. We have for any $X_i \in \mathbf{X}'$ with $i \leq p$ or $i \geq 2p + 1$ that for some $j \neq 1$ and $u_i \in \mathbb{B}(\mathbf{0}, 1/(32rd))$ with e_0 denoting $\mathbf{0}$:

$$X_i = \frac{1}{2}e_j + u_i \implies \|f_r(X_i)\| \leq \left\| f_r\left(\frac{1}{2}e_j\right) \right\| + \|f_r(u_i)\| \leq 1.$$

Furthermore, we have :

$$\|\hat{\mu}(\mathbf{X} \cup \mathbf{Y}) - \mu(\mathbf{X})\| \geq |\hat{\mu}(\mathbf{X} \cup \mathbf{Y})_1 - \mu(\mathbf{X})_1| \geq |2r - 1/(32rd)| \geq r.$$

Now, suppose $\hat{\mu}(\mathbf{X}')_j \leq \frac{1}{4d}$ for all j , we have $\langle \hat{\mu}(\mathbf{X}'), \mathbf{1} \rangle \leq \frac{1}{4}$. Consider the affine mapping

$$g_r(x) := 8r \left(\langle \mathbf{1}, x \rangle - \frac{1}{2} \right) \mathbf{1} + x.$$

and the point sets $\mathbf{X} = \{g_r(X_i)\}_{i \geq p+1}$ and $\mathbf{Y} = \{g_r(X_i)\}_{i \leq p}$, we have that for any $i \geq p+1$,

$$X_i = \frac{1}{2}e_j + u_i \implies \|g_r(X_i)\| = \left\| 8r \langle \mathbf{1}, u_i \rangle \mathbf{1} + \frac{1}{2}e_j + u_i \right\| \leq \frac{8rd}{32rd} + \frac{1}{2} + \frac{1}{32rd} \leq 1.$$

By the affine-equivariance of $\hat{\mu}$ we have:

$$\hat{\mu}(\mathbf{X} \cup \mathbf{Y}) = \hat{\mu}(g_r(\mathbf{X}')) = g_r(\hat{\mu}(\mathbf{X}')).$$

Hence by the definition of g_r and the fact that $\langle \hat{\mu}(\mathbf{X}'), \mathbf{1} \rangle \leq \frac{1}{4}$,

$$\begin{aligned} \|\hat{\mu}(\mathbf{X} \cup \mathbf{Y}) - \mu(\mathbf{X})\| &\geq \frac{1}{\sqrt{d}} \langle \mu(\mathbf{X}) - g_r(\hat{\mu}(\mathbf{X}')), \mathbf{1} \rangle \\ &\geq \frac{1}{\sqrt{d}} \left(\frac{1}{2} - \frac{\sqrt{d}}{32rd} \right) - \frac{1}{\sqrt{d}} \left(\langle \hat{\mu}(\mathbf{X}'), \mathbf{1} \rangle (8rd + 1) - \frac{8rd}{2} \right) \\ &\geq 0 + 4r\sqrt{d} - \frac{(8rd + 1)}{4\sqrt{d}} \geq r. \end{aligned}$$

Observing that r is arbitrary and that $|\mathbf{Y}| = p = n/(d+1)$ concludes the proof. $\square$

B.4 Recent Statistical Results

Here, we overview recent results that aim to build estimators with direction-dependent accuracy guarantees, expanding on their exposition in [Section 4.2](#). All these results impose additional assumptions on the distribution beyond the existence of a covariance; typically, these take the form of contractivity assumptions on the higher-order moments of the distribution. Notably, these assumptions enable (approximate) estimation of the covariance matrix Σ allowing for standardization of the data. The utility of such an estimate is explained subsequently. Their quantitative guarantees fall into two classes:

1. The first show that there exist estimators achieving sub-Gaussian rates in the Mahalanobis norm with a constant fraction of corruptions. These include the works of Depersin, and Lecué [27], Duchi, Haque and Kuditipudi [40], and Brown, Hopkins, and Smith [16] which in turn build upon approaches by Brown, Gaboardi, Smith, Ullman, and Zakynthinou [17]. The additional assumptions in these works allow for (approximate) estimation of the covariance matrix Σ . The utility of such an estimate is subsequently explained. However, the key contribution of our work concerns the scenario where Σ is *not* estimable.
2. The second concerns the recent results of Lugosi and Mendelson [88] who show that under an additional contractivity assumption, it is possible to obtain the following:

$$\forall \|v\| = 1 : \langle \hat{\mu}_{\text{LM}}(\mathbf{X}) - \mu, v \rangle \leq C \left(\sigma(v) \sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{\sum_{j \geq c \log(1/\delta)} \sigma_j}{n}} \right)$$

where $\sigma(v) = \sqrt{v^\top \Sigma v}$ and σ_i denote the eigenvalues of Σ . This result avoids an explicit dependence on the dimension which is inherent to the Mahalanobis norm while also, achieving better accuracy guarantees for directions with small variance. On the other hand, these results only apply under additional contractivity assumptions under which the sub-Gaussian rate ([Eq. \(4.1\)](#)) is achievable in the Mahalanobis norm. Furthermore, the error of this estimator in Mahalanobis norm may be arbitrarily large.

We now briefly explain the utility of estimating Σ when estimating in the Mahalanobis norm. For the sake of simplicity, consider the setting where Σ is known *exactly*. In this case, one may compute $\mathbf{X}' = \{X'_i = \Sigma^{-1/2} X_i\}_{i=1}^n$. Notice, now that the covariance of the (uncorrupted) X'_i now become I . As a consequence, applying an estimator in the *Euclidean norm* (say, the near linear-time algorithm from [\[29\]](#)), we obtain an estimate $\hat{\mu}'$ satisfying:

$$\|\hat{\mu}' - \Sigma^{-1/2} \mu\| \leq C \left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \right).$$

Hence, we get for the estimate $\hat{\mu} = \Sigma^{1/2} \hat{\mu}'$, we get:

$$\|\hat{\mu} - \mu\|_{\Sigma} = \|\Sigma^{-1/2}(\hat{\mu} - \mu)\| = \|\hat{\mu}' - \Sigma^{-1/2} \mu\| \leq C \left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \right).$$

The main drawback of the above approach is that Σ is assumed to be known while in practice, it is usually estimated from the same samples used to estimate μ . However, in several cases of interest, it suffices obtain an approximation of Σ , and the analysis can be adapted to account for the dependencies for particular estimators such as the Stahel-Donoho median [27]. As remarked previously, our estimator concerns scenarios where such an approximation is *not* obtainable from the data.

Appendix C

Appendix for Chapter 5

C.1 Extended data figures and tables

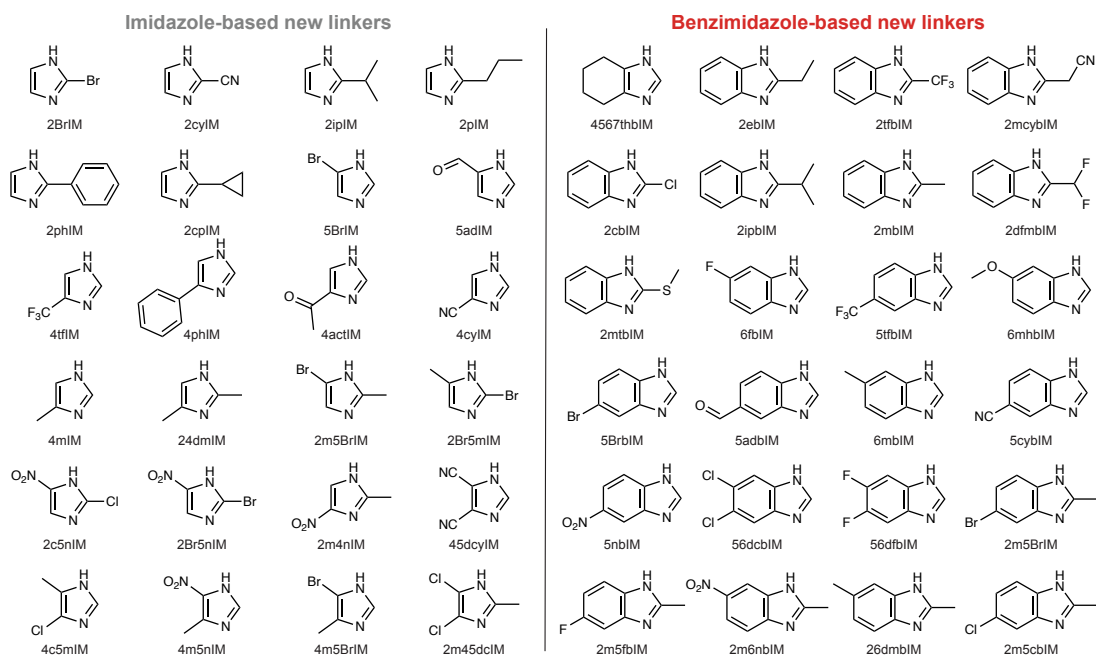


Figure C.1: Chemical structures and abbreviations of 48 new linkers.

C.2 Determination of reaction condition space

The reaction condition space in this study is defined by three primary parameters: total concentration, logarithmic linker-to-metal ratio, and temperature. We established the parameter ranges by first extracting synthesis conditions of some reported Zn-ZIF single crystal

structures (Supplementary Table C.1) and then plotting the distributions of these parameters (Supplementary Fig. C.2). The ranges are determined so as to cover these distributions: total concentration (0.05 to 0.5 mol/L), log linker-to-metal ratio (-0.3 to 1.3), and temperature (65 to 140 °C).

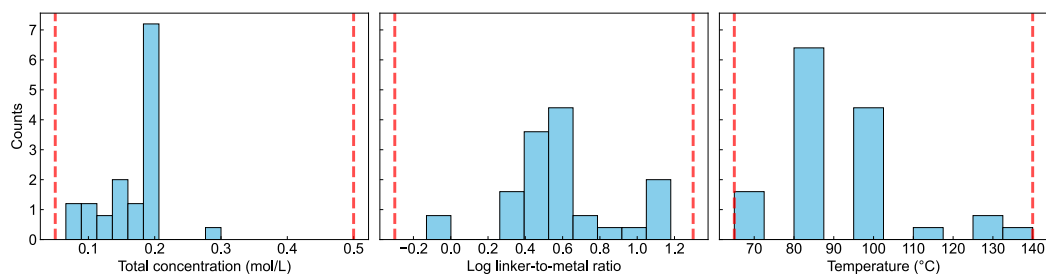


Figure C.2: **Distribution of primary synthesis parameters.** Dashed vertical lines indicate the boundaries of selected ranges.

Table C.1: Synthesis conditions of reported Zn-ZIFs used for determining the reaction condition space.

ZIF	DOI	Total concentration (mol/L)	Log linker-to-metal ratio	Temperature (°C)
ZIF-1	10.1073/pnas.0602439103	0.141	0.81	85
ZIF-2	10.1073/pnas.0602439103	0.141	1.18	85
ZIF-3	10.1073/pnas.0602439103	0.160	1.06	85
ZIF-4	10.1073/pnas.0602439103	0.198	0.46	130
ZIF-6	10.1073/pnas.0602439103	0.139	1.08	85
ZIF-7	10.1073/pnas.0602439103	0.066	-0.13	130
ZIF-8	10.1073/pnas.0602439103	0.085	-0.04	140
ZIF-10	10.1073/pnas.0602439103	0.160	1.06	85
ZIF-11	10.1073/pnas.0602439103	0.104	1.18	100
ZIF-11	10.1126/science.1152516	0.200	0.43	100
ZIF-14	10.1126/science.1152516	0.200	0.43	85
ZIF-20	10.1126/science.1152516	0.200	0.40	65
ZIF-20	10.1038/nmat1927	0.300	0.70	65
ZIF-60	10.1126/science.1152516	0.120	0.48	85
ZIF-61	10.1126/science.1152516	0.090	0.30	100
ZIF-62	10.1126/science.1152516	0.150	0.30	100
ZIF-64	10.1126/science.1152516	0.200	0.95	100
ZIF-68	10.1126/science.1152516	0.160	0.60	100
ZIF-68	10.1021/ja809459e	0.200	0.30	100
ZIF-69	10.1126/science.1152516	0.200	0.60	100
ZIF-69	10.1021/ja809459e	0.200	0.30	85
ZIF-70	10.1126/science.1152516	0.200	0.60	100
ZIF-70	10.1021/ja809459e	0.200	0.40	110
ZIF-71	10.1126/science.1152516	0.075	0.60	85
ZIF-71	10.1021/ja104035j	0.200	0.48	85
ZIF-72	10.1126/science.1152516	0.200	0.60	65
ZIF-73	10.1126/science.1152516	0.110	0.43	85
ZIF-74	10.1126/science.1152516	0.190	0.73	85
ZIF-76	10.1126/science.1152516	0.150	0.48	65
ZIF-77	10.1126/science.1152516	0.200	0.64	100
ZIF-78	10.1021/ja809459e	0.120	0.60	100
ZIF-79	10.1021/ja809459e	0.200	0.60	85
ZIF-80	10.1021/ja809459e	0.200	0.60	85
ZIF-81	10.1021/ja809459e	0.200	0.60	85
ZIF-82	10.1021/ja809459e	0.200	0.60	85

C.3 Crystals in the initial dataset

All crystals containing Zn and linker in the initial dataset are summarized below in Supplementary Table C.2. Only reactions yielding ZIF crystals are labeled as ‘1’ in the initial dataset. Representative optical images of these crystals are shown in Supplementary Fig. C.3.

Table C.2: **Summary of the crystals formed from the 8 starting linkers.** Linker abbreviations represent neutral form in the first column, but represent deprotonated forms in the second column.

Linker	Crystal composition	ZIF identity
IM	Zn(IM)_2	ZIF-2 (Yes)
	Zn(IM)_2	ZIF-3 (Yes)
2mIM	Zn(2mIM)_2	ZIF-8 (Yes)
2nIM	Zn(2nIM)_2	ZIF-65(Zn) (Yes)
	Zn(2nIM)(HCOO)	No
45dcIM	Zn(45dcIM)_2	ZIF-dcIM-sod (Yes)
bIM	Zn(bIM)_2	ZIF-7
	$\text{Zn}_5(\text{bIM})_6(\text{HCOO})_4(\text{DMF})_2$	No
9pur	Zn(9pur)_2	ZIF-20 (Yes)
	$\text{Zn}_2(\text{9pur})_2(\text{HCOO})_2(\text{DMF})$	No
5cbIM	Zn(5cbIM)_2	ZIF-95 (Yes)
	$\text{Zn}_{20}(\text{5cbIM})_{39}(\text{OH})$	ZIF-100 (Yes)
4a5mIM	Zn(4a5mIM)_2	ZIF-93 (Yes)
	Zn(4a5mIM)_2	ZIF-94 (Yes)

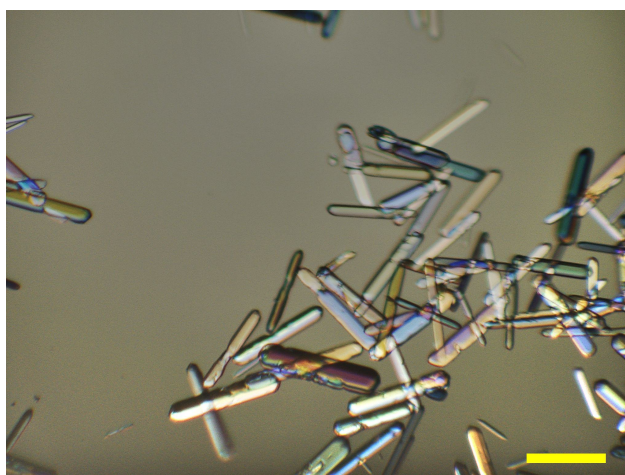


Figure C.3: **Representative optical image of ZIF-2 crystallized from IM.** This image was obtained from a reaction whose conditions were: a total concentration of 0.50 mol/L, a log linker-to-metal ratio of 1.00, and a temperature of 100 °C. Yellow scale bar 100 μm .

C.4 Automated synthesis workflow

The synthesis was automated using an Opentrons OT-2 robot equipped with dual pipettes: a P20 (1-20 μL) and a P300 (20-300 μL). The robot’s 11-slot platform was configured with two tip racks, a 96-well plate, a waste collection tube rack, and three custom 3D-printed holders, each accommodating six 20 mL vials containing stock solutions. Stock solutions of metal salts and linkers were prepared at two concentrations (0.15 mol/L and 0.5 mol/L), with pure DMF available for dilution. The system was controlled through the Opentrons App by Jupyter notebook, using custom Python scripts and JSON files defining labware specifications.

For the 8 starting linkers, we sampled a reaction condition space comprehensively using six values of total concentration (0.05, 0.10, 0.15, 0.25, 0.35, and 0.50 mol/L), eight values of log linker-to-metal ratio (-0.30, 0.00, 0.30, 0.48, 0.70, 0.90, 1.00, and 1.18), and seven values of temperature (65, 85, 100, 110, 120, 130, and 140 $^{\circ}\text{C}$). The boundary of this reaction condition space was determined as described in Appendix C.2. For the 48 new linkers, considering their higher cost and our understanding of parameter sensitivity gained from initial dataset, we sampled fewer conditions in the same reaction condition space: two values of total concentration (0.15 and 0.40 mol/L), six values of log linker-to-metal ratio (0.00, 0.25, 0.50, 0.75, 1.00, and 1.25), and three values of temperature (70, 100, and 130 $^{\circ}\text{C}$). This sampling strategy resulted in 336 reactions per starting linker and 36 candidate reactions per new linker.

To execute these reactions using the robotic system, the sampled reaction parameters need to be converted into specific volumes of solutions. As mentioned, the robotic system used three components: a stock solution of metal salt (concentration $c_{\text{metal_stock}}$), a stock solution of linker (concentration $c_{\text{linker_stock}}$), and pure solvent. Given target parameters of total concentration (c_{total}) and log linker-to-metal ratio ($\log LMR$), and a fixed total volume (V_{total}), the required volumes of each component can be calculated using Equations (C.1)-(C.3).

$$V_{\text{metal_stock}} = V_{\text{total}} \cdot \frac{1}{1 + 10^{\log LMR}} \cdot \frac{c_{\text{total}}}{c_{\text{metal_stock}}} \quad (\text{C.1})$$

$$V_{\text{linker_stock}} = V_{\text{total}} \cdot \frac{10^{\log LMR}}{1 + 10^{\log LMR}} \cdot \frac{c_{\text{total}}}{c_{\text{linker_stock}}} \quad (\text{C.2})$$

$$V_{\text{solvent}} = V_{\text{total}} - V_{\text{metal_stock}} - V_{\text{linker_stock}} \quad (\text{C.3})$$

The reactions were conducted in a TOPAS polymer 96-well plate with 0.5 mL borosilicate glass inserts purchased from Chemglass Life Sciences LLC. A CSV file read by the python script guided dispensing volumes for each well, combining metal salt solution, linker solution, and DMF as needed. While maintaining a constant 300 μL total volume, we varied total concentrations and linker-to-metal ratios across the reactions on a plate. To prevent cross-contamination, pipettes discarded and reloaded fresh tips between different stock solutions. The complete plate preparation took approximately 30 minutes. Following automated dispensing, the plate was sealed with PTFE/Silicone stoppers and secured between steel bars

using aluminum clamps before heating in isothermal ovens at the sampled temperature for that plate.

Post-reaction, the cooled plate was unsealed and returned to the robot for supernatant removal. A separate Python script controlled the pipette to aspirate solution 3 mm above each well's bottom, leaving only a thin liquid layer and solid products for imaging.

C.5 Statistical analysis of success rates of the initial dataset

The dataset has a structure of $y(L, C_1, C_2, C_3)$ where the observed reaction outcome is a binary label (y) dependent on the linker (L) and the three primary reaction parameters (C_1, C_2, C_3). To analyze reaction success patterns, we define a success rate (SR) by fixing certain parameters while randomly sampling others within their predetermined ranges. For example, when analyzing the dependence on linker type and the first parameter, the success rate is calculated as: $SR(L, C_1) = \langle y(L, C_1, C_2, C_3) \rangle_{C_2, C_3}$, where $\langle \rangle_{C_2, C_3}$ denotes averaging over samples of C_2 and C_3 . The success rates for each starting linker are visualized as functions of two parameters in Supplementary Figs C.4 and C.5, and as functions of individual parameters in Supplementary Figs C.6, C.7, and C.8.

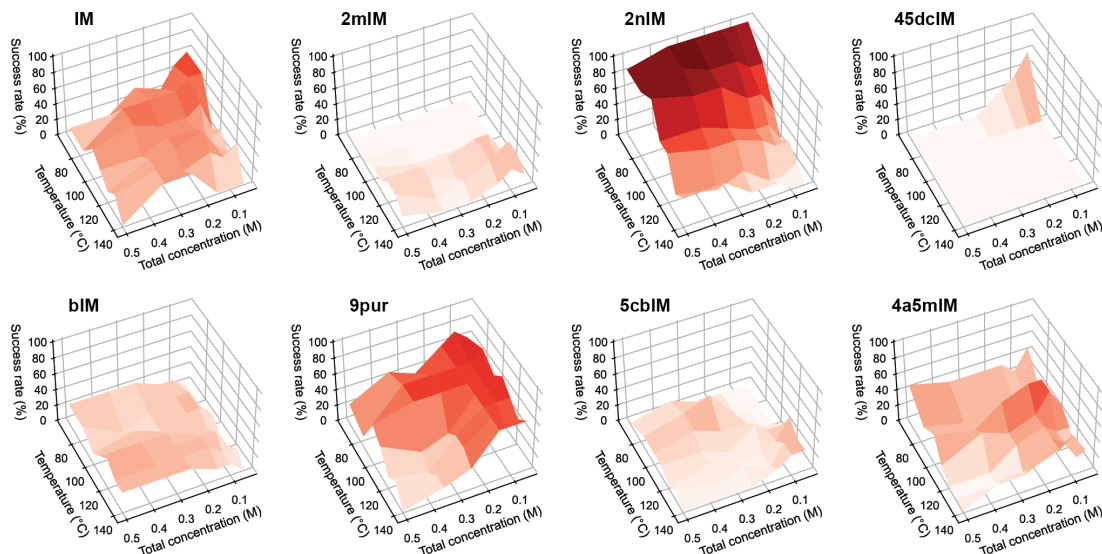


Figure C.4: **Success rate maps of individual linkers as functions of total concentration and temperature.** The success rate is shown as surface plots for each starting linker. The color intensity corresponds to the success rate, with darker red indicating higher success rates.

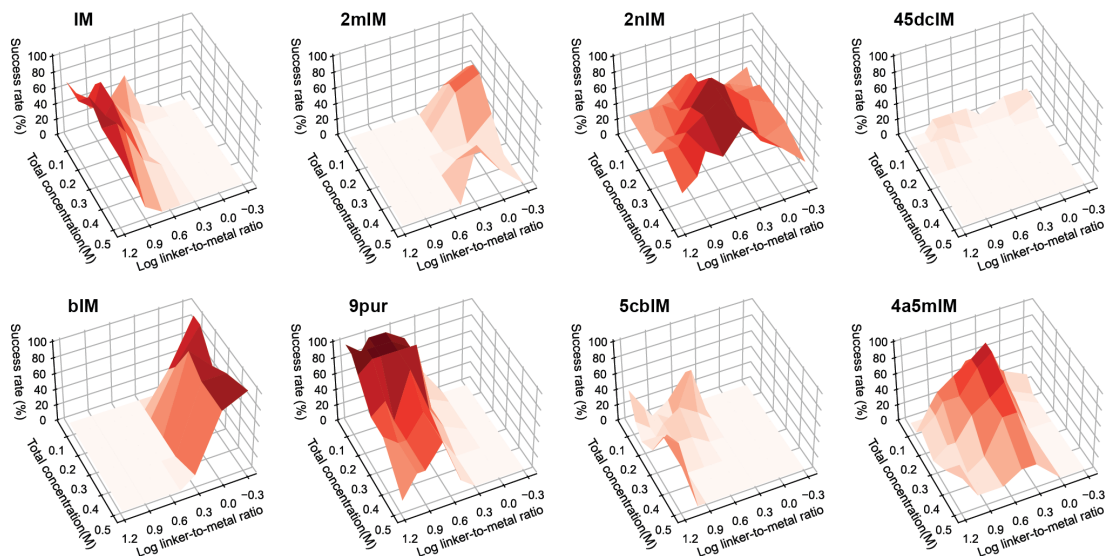


Figure C.5: **Success rate maps of individual linkers as functions of total concentration and log linker-to-metal ratio.** The success rate is shown as surface plots for each starting linker. The color intensity corresponds to the success rate, with darker red indicating higher success rates.

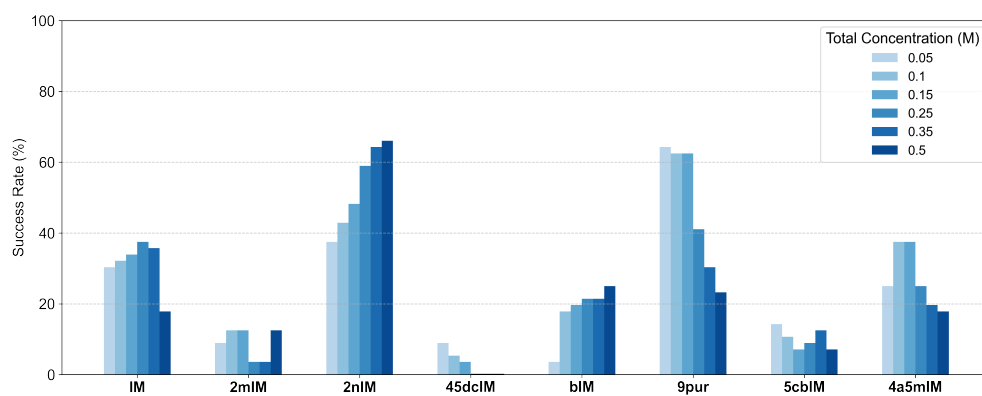


Figure C.6: **Success rates of individual linkers as functions of total concentration.** Bar plots show the success rates for each linker, with darker blue indicating higher concentrations.

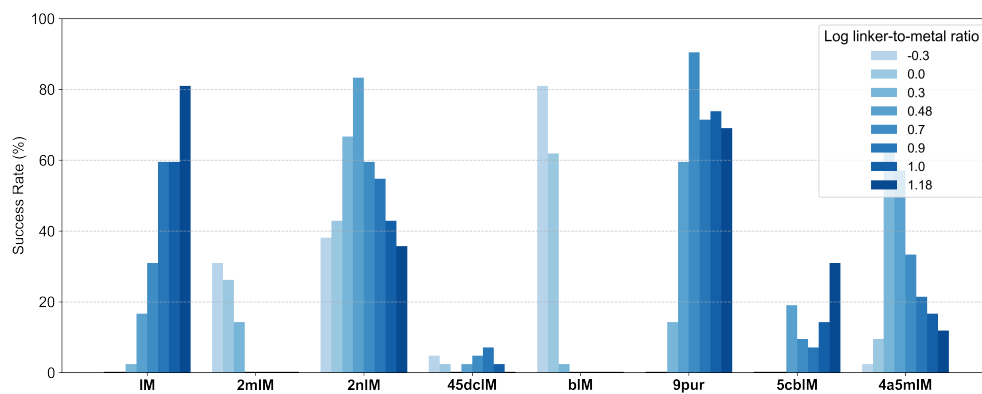


Figure C.7: **Success rates of individual linkers as functions of log linker-to-metal ratio.** Bar plots show the success rates for each linker, with darker blue indicating higher ratios.

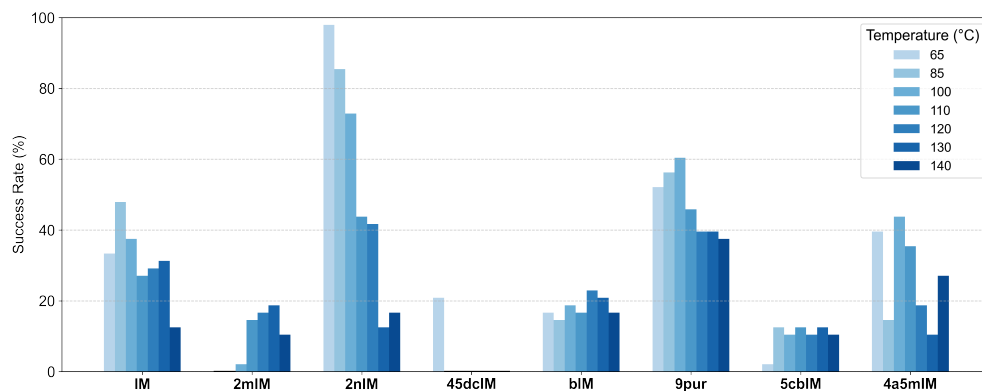


Figure C.8: **Success rates of individual linkers as functions of temperature.** Bar plots show the success rates for each linker, with darker blue indicating higher temperatures.

C.6 Reaction encoding details

Linker features:

- **feat_HbondAcceptorCount**: Total number of hydrogen bond acceptors present on the neutral linker.
- **feat_HalogenOnC2**: Binary indicator (0 or 1) for the presence of a halogen atom at the C2 position of the linker.
- **feat_HalogenOnC4orC5_IM-based**: Count of halogen substituents at C4/C5 positions, applicable only to imidazole-based linkers (set to 0 for benzimidazole-based linkers)
- **feat_NitroCount**: Total number of nitro (-NO₂) groups present on the linker.
- **feat_CyanoCount**: Total number of cyano (-CN) groups present on the linker.
- **feat_TopDistance(Å)**: Distance from the furthest C2 substituent atom to the line segment connecting the two imidazole nitrogen atoms, measured using the optimized geometry of the deprotonated linker.
- **feat_BottomDistance(Å)**: Distance from the furthest C4/C5 substituent atom to the line segment connecting the two imidazole nitrogen atoms, measured using the optimized geometry of the deprotonated linker.
- **feat_LeftRightDistanceMax(Å)**: The larger of the horizontal distances measured between vertical planes that contain the outermost atoms at either C2 or C4/C5 positions to their nearest imidazole nitrogen planes in the optimized deprotonated structure.
- **feat_LeftRightDistanceMin(Å)**: The smaller of the horizontal distances measured between vertical planes that contain the outermost atoms at either C2 or C4/C5 positions to their nearest imidazole nitrogen planes in the optimized deprotonated structure.
- **feat_N1ESPChargeNeutralLowEnergy**: Electrostatic potential charge on N1 from the lower-energy neutral tautomer, determined by comparing electronic energies of two possible neutral structures formed by placing the hydrogen on either N atoms.
- **feat_H1ESPChargeNeutralLowEnergy**: Electrostatic potential charge on the H1 (attached to nitrogen) from the lower-energy neutral tautomer.
- **feat_ElectronicEnergyDifferenceNeutralLowEnergyCharged(kJ/mol)**: Electronic energy difference between the lower-energy neutral tautomer and the deprotonated (charged) form of the linker.
- **feat_N1N3ESPChargeChargedMax**: The more positive electrostatic potential charge between the two N atoms in the deprotonated (charged) form.

- **feat_N1N3ESPChargeChargedMin**: The more negative electrostatic potential charge between the two N atoms in the deprotonated (charged) form.
- **feat_dHinDMF@298K**: Enthalpy of formation of the Zn-linker cluster relative to isolated Zn ions and neutral linkers at 298 K.
- **feat_dGinDMF@298K**: Free energy of formation of the Zn-linker cluster relative to isolated Zn ions and neutral linkers at 298 K.
- **feat_d(Zn-N)Avg**: Bond length of the Zn center and coordinated N atoms, averaged across all linkers in the cluster.
- **feat_A(N-Zn-N)**: Angle between N atoms of adjacent linkers and bridging the Zn center.
- **feat_qZn(APT)**: APT charge of the Zn center.
- **feat_q[N]-Zn(APT)Avg**: APT charge of the N atoms bonded to the Zn center, averaged across all linkers in the cluster.
- **feat_q[N]-H(APT)Avg**: APT charge of the N atoms away from the Zn center and terminated with H atoms, averaged across all linkers in the cluster.

Experimental features:

- **rxn_TotalConcentration(mol/L)**: Combined concentration of metal ions and linkers in the reaction mixture.
- **rxn_Log10LinkerMetalRatio**: $\log_{10}$ of the linker-to-metal concentration ratio in the reaction mixture.
- **rxn_Temperature(°C)**: Reaction temperature during the heating process.
- **rxn_MetalConcentration(mol/L)**: Concentration of metal ions in the reaction mixture.
- **rxn_LinkConcentration(mol/L)**: Concentration of linkers in the reaction mixture.

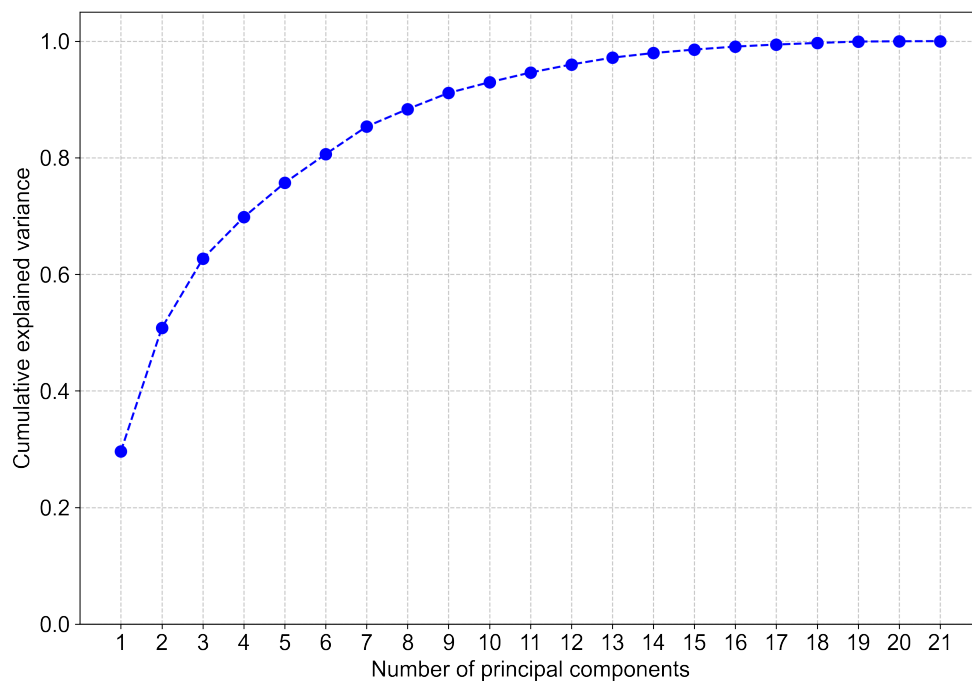


Figure C.9: Cumulative explained variance ratio as a function of principal components for decomposed linker features.

C.7 Description of ZIFs discovered

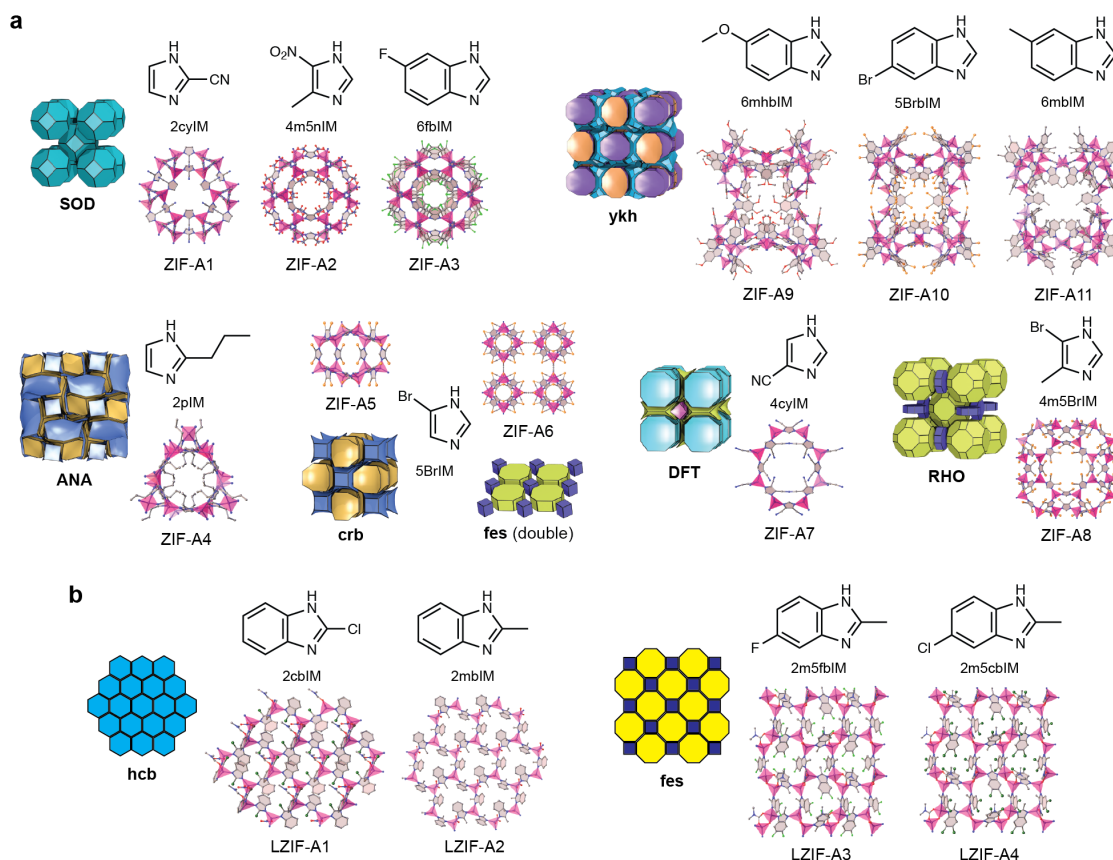


Figure C.10: **Topological classification and crystal structures of ZIFs and LZIFs discovered by AIRES from new linkers.** **a**, Crystal structures of ZIFs (ZIF-A1 through ZIF-A11) are shown and classified based on network topologies. ZIF topologies are visualized using natural tiling models and labeled with their respective three-letter topological codes, where uppercase and lowercase respectively indicate topologies known in zeolites and only in ZIFs. Disordered positions of functional groups in crystal structure are preserved in visualization. **b**, Crystal structures of LZIFs (LZIF-A1 through LZIF-A4) are shown and classified based on network topologies. LZIF topologies are visualized using 2D tiling models.

The discovery process yielded 11 new crystal structures from 10 successful molecular systems, representing a 33% expansion of the known single-component Zn-ZIF library. From a computational perspective, this demonstrates the algorithm’s ability to navigate a complex search space where minor variations in molecular features (single atom substitutions) can lead to entirely different crystal topologies—a challenging test case for transfer learning where the mapping from features to outcomes is highly non-linear and multi-modal.

Despite their common geometry of tetrahedral building blocks, ZIFs display diverse network topologies, patterns of metal connectivity denoted by three-letter codes. Through AIRES, ten successful linkers yielded eleven new crystals belonging to seven distinct topologies (Fig. C.10a). Crystal structures exhibiting SOD and ANA topologies (ZIF-A1 and ZIF-A4) were obtained from 2cyIM and 2pIM respectively, matching structures originally discovered in alkyl-alcohol solvents, in contrast to DMF used in this study. Additional SOD topology structures (ZIF-A2 and ZIF-A3) emerged from 4m5nIM and 6fbIM. A structure with DFT topology (ZIF-A7) was generated from 4cyIM. 4m5BrIM formed the only RHO topology structure (ZIF-A8) among all the new linkers. Structures of the ykh topology (ZIF-A9, ZIF-A10, and ZIF-A11) were produced by 6mhbIM, 5BrbIM, and 6mbIM, which were previously accessible only through mixed linker approaches that incorporate multiple linker types [135]. Structural versatility was observed with 5BrIM, where it produced a crb topology structure (ZIF-A5) and additionally an unprecedented double fes topology structure (ZIF-A6). During the AIRES process with RFC, ZIF-A6 was discovered earlier than ZIF-A5 in one sixth of attempts, in accordance with their relative success rates. While there was a reported layered ZIF featuring metal nodes arranged in a single atomic layer [136], ZIF-A6 exhibits a new double-thick layer arrangement where two layers of metal nodes are covalently bridged by 5BrIM, representing a new structural motif in ZIF chemistry. Moreover, as mentioned earlier, constructing isorecticular structures for ZIFs with specific topologies presents significant challenges, as minor functional group modifications can trigger alternative frameworks by altering linker-linker interactions (exemplified by ZIF-7 made of bIM showing SOD topology versus ZIF-20 made of 9pur displaying LTA topology) or by introducing critical steric constraints (as seen with ZIF-8 made of 2mIM forming SOD topology versus ZIF-A4 made of 2pIM adopting ANA topology). Therefore, we highlight the significance of ZIF-A5 made of 5BrIM exhibiting crb topology and ZIF-A7 made of 4cyIM exhibiting DFT topology, as these topologies were previously achieved only by unsubstituted IM linker (ZIF-1, ZIF-2 and ZIF-3). The strategic use of linkers with single functionality at the 4/5 position of the imidazole backbone creates new perspectives in ZIF structure design. Overall, AIRES’s systematic exploration expanded the known single-linker Zn-ZIF library developed over the last twenty years by one-third, with several providing new structural insights. The achievement of AIRES not only breaks through a decade-long discovery plateau but also demonstrates that significant untapped potential still exists in a seemingly well-explored chemical space.

The rich coordination chemistry of ZIF synthesis led to unexpected discoveries beyond the target structure motif. RFC also identified four new linkers (2cbIM, 2mbIM, 2m5fbIM, and 2m5cbIM) that produced other types of crystals before discovering all ZIF crystals. While these structures maintain tetrahedral Zn coordination, they exhibit competitive binding between linkers and auxiliary ligands at the metal centers. This competition results in layered structures rather than three-dimensional frameworks, leading us to classify them as layered-ZIFs (LZIFs) (Fig. C.10b). In LZIF-A1 (from 2cbIM) and LZIF-A2 (from 2mbIM), DMF and nitrate compete for coordination sites, while in LZIF-A3 (from 2m5fbIM) and LZIF-A4 (from 2m5cbIM), the competition comes from dimethylcarbamate, which forms through DMF oxidation. Although the metal centers deviate from coordination characteristic of

ZIFs, the metal-linker coordination bonds are still preserved, implying that ML captured fundamental coordination principles.

C.8 Synthetic experiments for the main methods

To systematically evaluate our methodology, we constructed an artificial synthetic experimental pipeline that mirrors the key characteristics of real-world crystallization discovery. Experiments showed a superior performance of our discovery methodology over Random Search.

Data-generating process

Our synthetic data generation process carefully preserves the essential structure of the crystallization problem while providing a controlled environment for algorithmic validation.

The process consists of the following steps:

1. Generate a set of n_l synthetic linkers by sampling uniformly from $[-1, 1]^{d_l}$, where d_l represents the dimensionality of the linker feature space. Each point in this space represents a unique molecular structure through its abstract features, analogous to how real molecular descriptors (such as molecular weight, number of rotatable bonds, etc.) characterize actual chemical compounds.
2. Create an experimental condition space by constructing a regular grid of $n_e = k \times k$ points in $[-1, 1]^2$, representing standardized reaction parameters such as temperature and concentration. This grid structure reflects the systematic exploration of experimental conditions typically employed in crystallization studies. The standardization to $[-1, 1]$ follows common practice in machine learning, where input variables are normalized to a consistent scale.
3. Form the complete experimental space through a Cartesian product of the linker set and reaction condition grid, yielding $n_l \times n_e$ potential experiments. This construction captures the combinatorial nature of the crystallization discovery process, where each linker must be tested under various experimental conditions.
4. Randomly select a subset of linkers as the initial training set, simulating the historical data available at the start of a discovery campaign. The remaining linkers constitute the exploration space for evaluating discovery algorithms.
5. Generate crystallization outcomes through a two-stage probabilistic process:
 - First, sample an underlying function f from a Gaussian Process with a Radial Basis Function (RBF) kernel over the experimental space. This function represents the true but unknown relationship between experimental parameters and crystallization success probability.
 - Then, for each experimental point x , sample the binary outcome y from a Bernoulli distribution with probability $\sigma(c \cdot f(x))$, where σ is the sigmoid function and $c > 0$ controls the signal-to-noise ratio. The signal-to-noise ratio parameter c determines

how clearly the true pattern emerges from random variation – higher values of c create sharper distinctions between successful and unsuccessful regions, while lower values introduce more uncertainty.

The RBF kernel in our Gaussian Process ensures smooth transitions in crystallization probability across the experimental space, reflecting the chemical intuition that similar conditions and molecular structures should yield similar outcomes. The kernel computes similarity between points x and x' as:

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2l^2}\right) \quad (\text{C.4})$$

where l is the length scale parameter controlling how quickly correlations decay with distance. Our chosen length scale produces realistic spatial patterns in crystallization success probability, neither too smooth (which would be unrealistic) nor too rough (which would make learning impossible).

Implementation details and results

We implemented this framework with the following specifications:

- Number of linkers n_l : 20
- Linker feature dimensionality d_l : 8, chosen to represent key molecular descriptors while maintaining computational tractability
- Experimental grid size: 6×6 ($n_e = 36$), providing sufficient resolution to capture meaningful patterns in the experimental space
- Signal-to-noise ratio parameter c : 2.0, selected to create realistic levels of experimental uncertainty while maintaining learnable patterns
- Mean function offset: -1.0 , inducing a natural imbalance between successful and failed experiments

The resulting synthetic dataset exhibits several key characteristics that align with real-world crystallization discovery:

1. **Reaction spatial correlation:** The RBF kernel induces smooth variations in crystallization probability across experimental conditions, this reflects the physical reality that similar reaction conditions tend to produce similar outcomes.
2. **Molecular similarity effects:** The kernel structure captures both local smoothness within each linker’s experimental space and discontinuities across different molecular structures, mirroring the complex relationship between molecular structure and crystallization propensity.

3. **Success rate imbalance:** The negative mean function produces a realistic success-failure ratio, with successful crystallization being relatively rare (success rate 29%), matching typical experimental environments.

We validated our sequential and batch discovery algorithms described in previous sections over the synthetic dataset. The prediction model we used is a Gaussian Process Classification with RBF kernel and a 0 mean function, which was a mildly mis-specified model. We chose the batch size of 6 and randomly selected the 8 initial linkers for each trial and ran experiments over 50 trials. Indeed, the sequential and batch methods equipped with an ML prediction model substantially outperformed Random Search:

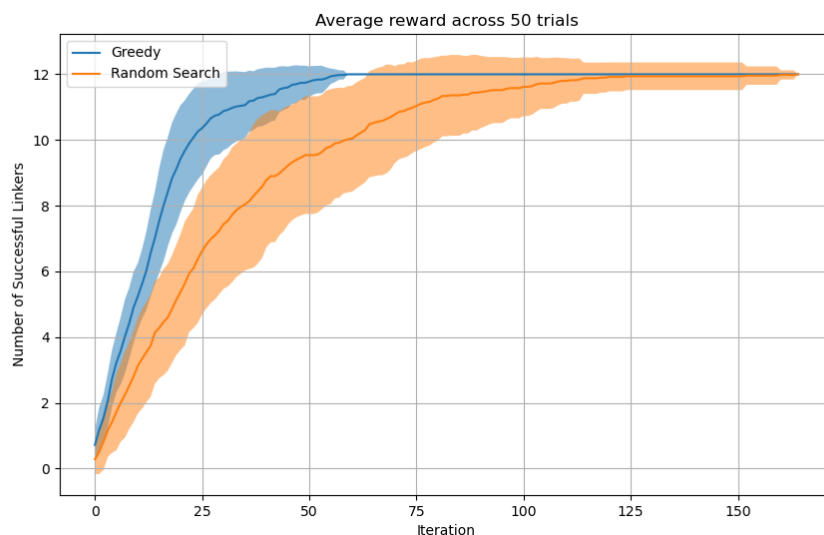


Figure C.11: Performance of sequential discovery algorithm across 50 trials

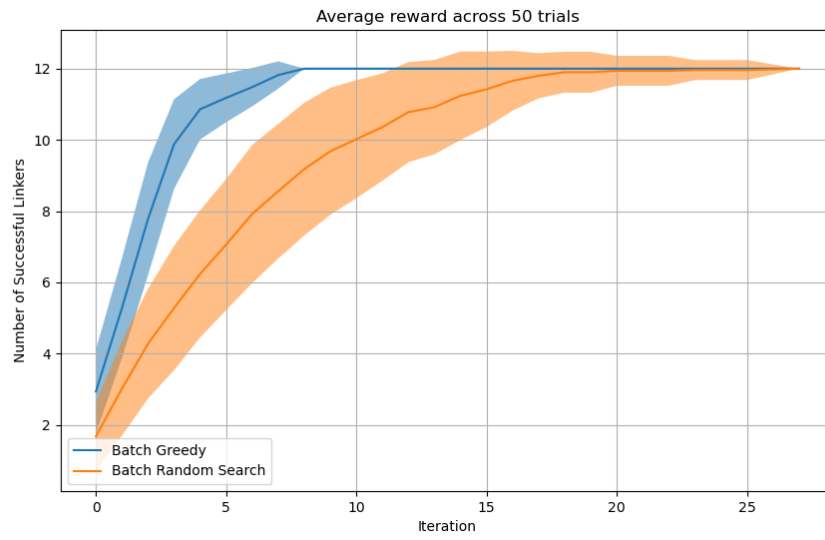


Figure C.12: Performance of batch discovery algorithm across 50 trials (batch size $N_B = 6$).

C.9 Investigation of RFC augmentation factor

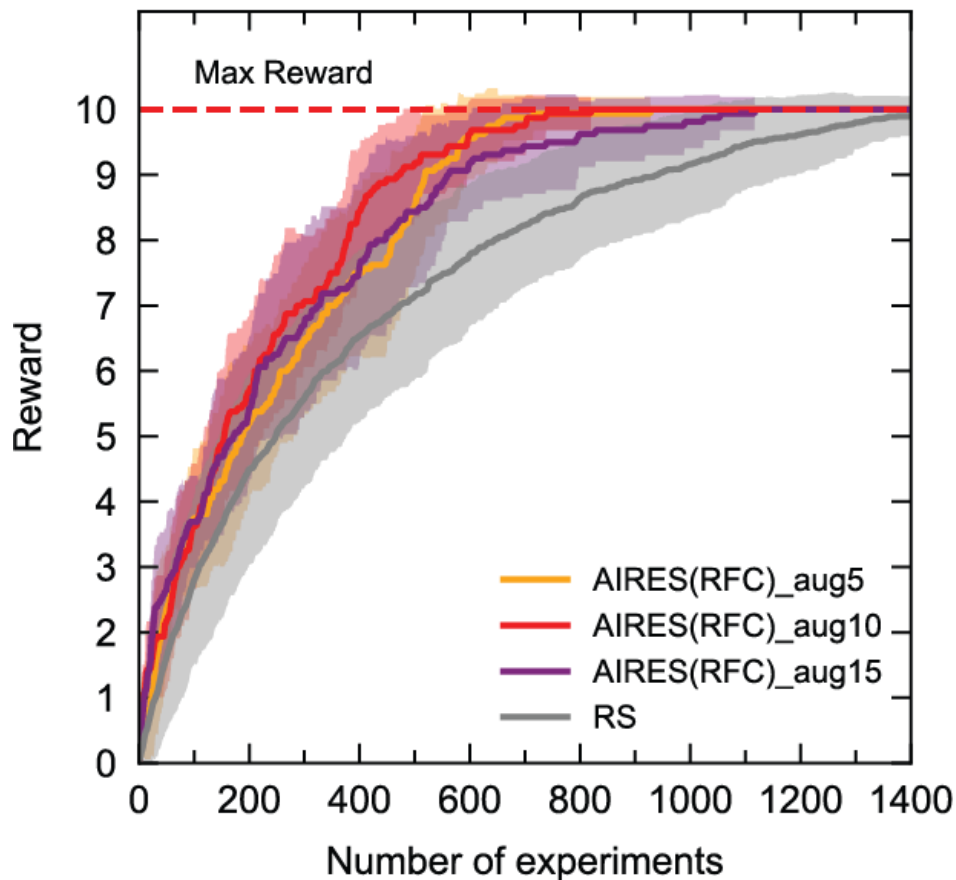


Figure C.13: **Impact of data augmentation on AIRES discovery performance.** The plot compares discovery rates between Random Search (RS, grey) and AIRES using RFC with different augmentation factors (5, 10, and 15). Shaded areas represent one standard deviation. The cumulative reward (y-axis) indicates the number of successful discoveries, with a maximum possible reward of 10 (dashed red line). AIRES consistently outperforms RS.

C.10 Discovery trajectory analysis

Here we visualize which linker is suggested by AIRES(RFC) at every iteration in the sequential setting. We take the first trajectory from AIRES(RFC) output for this demonstration. The higher frequency of experiments suggested for certain linkers (shown by denser clustering of markers) reveals RFC’s ability to identify and prioritize linkers with higher potential for ZIF crystal formation.

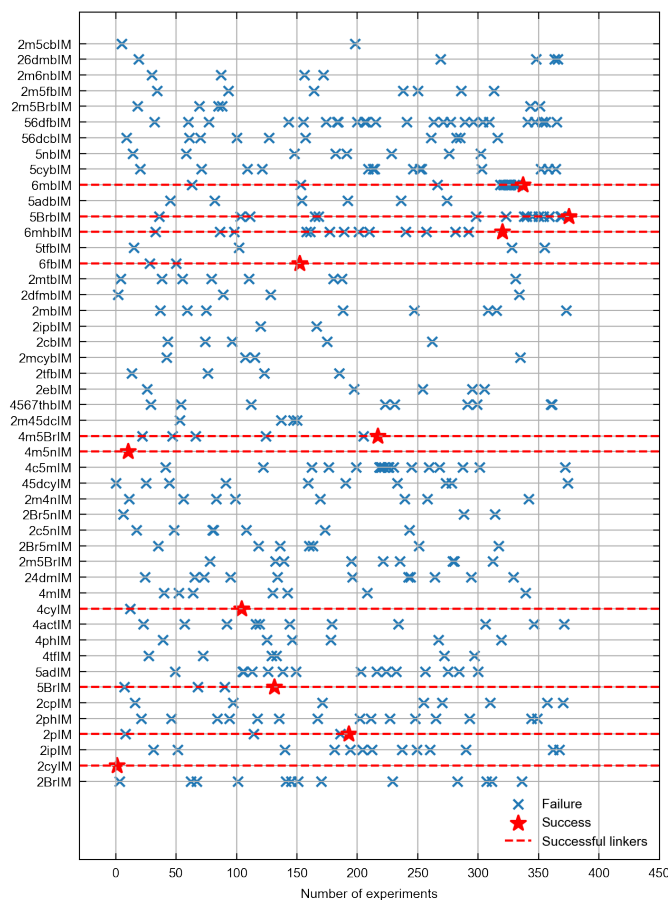


Figure C.14: **Linker suggestions from A single sequential trajectory of RFC.** Each linker suggestion is paired with specific reaction conditions (not shown in this plot).

C.11 Impact of batch selections on ZIF discovery

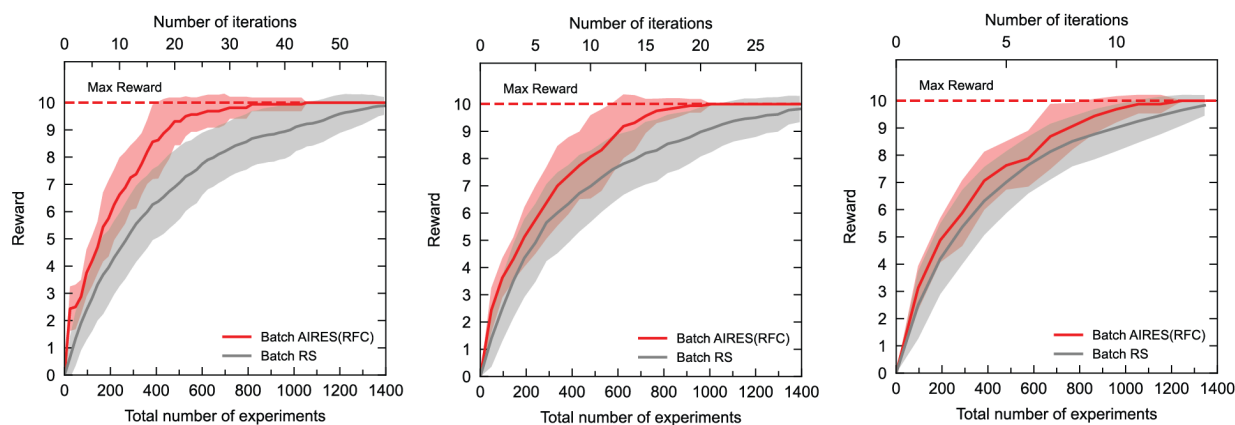


Figure C.15: **Cumulative reward of Batch AIRES(RFC) and Batch RS under different batch sizes.** Shaded areas represent one standard deviation across 16 repetitions for Batch AIRES(RFC) and 100 repetitions for Batch RS. Batch sizes: 24 reactions per iteration (left), 48 reactions per iteration (middle), and 96 reactions per iteration (right).

Another batch strategy is to naively select the top-k candidates with the highest predicted success probabilities. However, it often results in choosing candidates that are too similar to each other. This lack of diversity limits both the collective information gain and the overall discovery potential of the batch. A comparison of this top-k batch strategy and our Batch AIRES is presented below.

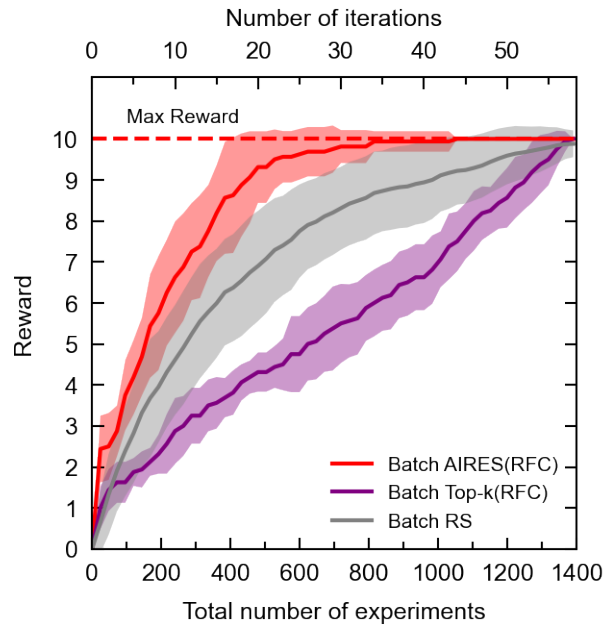


Figure C.16: **Cumulative reward of Batch AIRES(RFC), Batch Top-k(RFC) and Batch RS under batch size of 24.** Shaded areas represent one standard deviation across 16 repetitions for Batch AIRES(RFC) and Batch Top-k(RFC) and 100 repetitions for Batch RS. The procedures of data augmentation and hyperparameter tuning are the same for Batch AIRES(RFC) and Batch Top-k(RFC).