8-16-2023 10:30 AM

On Computing Optimal Repairs for Conditional Independence

Alireza Pirhadi, *Western University*

Supervisor: Milani, Mostafa, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Computer Science

Abstract

This thesis focuses on the concept of Conditional Independence (CI) and its testing, which holds immense significance across various fields, including economics, social sciences, and biomedical research. Notably, within computer science, CI has become an integral part of building probabilistic and causal models. It aids efficient inference and plays a key role in uncovering causal relationships.

The primary aim of this thesis is to broaden the scope of CI beyond its testing aspect. We introduce the pioneering problem of data repair, designed to adhere to particular CI constraints. The value and pertinence of this problem are highlighted through two contrasting applications. The first application is debiasing data and developing fair machine learning (ML) models. As fairness becomes a central issue in machine learning, exploring techniques for debiasing data to construct more equitable models is crucial. The proposed data repair methodology supports this, assisting in creating fairer models. The second application is about improving data quality and cleaning processes. Maintaining data quality is a continuous challenge across various fields, and our repair methods present a novel way to address this, enhancing the overall quality and reliability of the data.

The proposed repairs use optimal transport (OT) and Earth Mover's distance as dissimilarity measures. This approach ensures the preservation of the underlying probability distribution's geometry. In the context of fairness, this contributes to increased downstream model accuracy. In the realm of data cleaning, it offers a robust method for error detection. To facilitate the efficient generation of the repairs, we present novel techniques, including relaxed OT and block coordinate descent methods. The effectiveness of the repair methodologies is validated through experiments conducted on synthetic and real-world datasets. This comprehensive exploration highlights the potential of data repair in addressing critical issues in machine learning and data quality, offering a new perspective on using CI in these fields.

Keywords: Conditional independence, optimal transport, Wasserstein distance, fairness, data-cleaning

Lay Summary

This thesis delves into conditional independence (CI), a principle commonly used in economics, social sciences, and healthcare research. In simpler terms, CI is a way to understand how two random variables in a probabilistic model are connected or influence each other when a third random variable is considered. In computer science, this principle is fundamental to building models that can predict or figure out cause-effect relationships.

The main aim of this research is to extend the use of this concept to fix issues in data so it aligns with certain rules of CI. This idea is explored through two main applications. The first application involves making data more fair and unbiased, which in turn helps create machine learning models that treat all information fairly. In essence, this means creating a model that does not favor one set of data over another based on biased or unequal information. The second application focuses on improving the overall quality of data and making it error-free. This is a critical step because high-quality, clean data is essential for accurate predictions and decision-making. To fix or repair the data, the study uses methods that consider the difference between how data is distributed in its original and repaired states. This approach ensures the essence of the original data is maintained while enhancing the accuracy of the models built from it.

The study also introduces new techniques to make these repairs more efficient and tests the effectiveness of the repair methods using both made-up and real-world datasets. Overall, this research shines a light on how the principle of CI can be used innovatively to address critical issues in machine learning and data quality.

Acknowledgements

I would like to thank Dr. Mostafa Milani, my exceptional supervisor, for playing a crucial role in making this work possible. His guidance and support have been incredibly valuable. Additionally, my deepest gratitude to Dr. Babak Salimi for his invaluable contributions and advice, which have significantly enriched the quality of this thesis.

I also wish to acknowledge the esteemed members of my committee: Dr. Grace Li, Dr. Katarina Grolinger, and Dr. Boyu Wang. Their expertise, feedback, and constructive critiques have been indispensable in shaping this work.

I am grateful to my friends and fellow lab members for their generous feedback and for accompanying me on this journey. A special thanks to Yasaman Jafari for sharing her precious insights into this project.

Finally, A big shout-out goes to my family for always being there for me and believing in me. I profoundly appreciate all the favors they've provided to make this achievement possible.

Contents

List of Abbreviations

AUC Area Under Curve

BCD Block Coordinate Descent

CI Conditional Independence

CMI Conditional Mutual Information

DP Demographic Parity

EO Equalized Odds

FD Functional Dependency

FPR False Positive Rate

KL Kullback-Leibler

LP Linear Programming

ML Machine Learning

MVD Multi-Valued Dependency

NMF Non-negative Matrix Factorization

OT Optimal Transport

QCLP Quadratically Constrained Linear Programs

ROC Receiver Operating Characteristic

ROD Ratio of Observation Discrimination

TPR True Positive Rate

List of Figures

Chapter 1

Introduction

Conditional Independence (CI) is a central concept that permeates numerous disciplines, providing a critical framework for understanding and modeling the interplay between variables. This core statistical concept is represented as $X \perp\!\!\!\perp Y \mid Z$, where X, Y, Z are random variables. The relationship translates that the value of variable X imparts no additional information about variable Y (and vice versa), given the values of Z. In other words, when fixing the value of Z, changing the value of variable X doesn't change the probability of possible values of variable Y. This foundational principle is pivotal in fields such as statistics, machine learning [16, 47, 50], bioinformatics [58], and genetics [28], where it underpins key concepts such as sufficiency, ancillary, causal discovery, and inference [41].

In the domain of computer science, the role of CI is accentuated in the construction and interpretation of graphical models [33]. These models, including Bayesian networks and Markov Random Fields, use CI assumptions to simplify intricate probability distributions, rendering them more manageable for computation. This aspect is particularly vital when probing causal relationships among variables, where CI tests serve as a fundamental tool for delineating the true causal structure, especially in scenarios where experimental interventions are infeasible. Additionally, the arena of feature selection and engineering in machine learning applications reaps significant benefits from CI testing. Feature selection is integral to the identification of relevant input variables, thereby enhancing the predictive prowess of machine learning models, mitigating overfitting, and optimizing computational efficiency. In this regard, CI facilitates the selection of a subset of features that exert the most predictive power for the target variable, considering the other features. The influence of CI further permeates the specialized areas of machine learning such as domain adaptation and transfer learning [40, 15]. In these subfields, where the objective is to transfer knowledge derived from one task to another related task, CI enables researchers to model alterations in underlying data-generating processes across domains or tasks. This assists in identifying which components of the model should be transferred and which require adaptation. This expansive influence of CI underscores its importance across diverse domains, attesting to its role as a cornerstone in statistical understanding and application.

The application of CI extends to the discipline of genetics and bioinformatics, where CI testing forms the bedrock for identifying genomic mutations linked directly to diseases, thereby propelling the design of personalized therapies. CI helps isolate the direct correlations between genetic variations and diseases, even within a high-dimensional milieu of potential confound-

ing variables.

1.1 Motivation and Knowledge Gap

As previously outlined, CI is vital in establishing and validating the relationships between variables in various applications. It allows us to assess if two variables retain their independence when influenced by another variable, forming the backbone of decision-making strategies in many domains. However, there are scenarios where the data or its underlying distribution fail to comply with the anticipated CI conditions. In such instances, an intriguing yet relatively uncharted avenue is to "repair" the data or its distribution to satisfy the CI constraints. This approach to data repair draws parallels to the concept of database repair, a well-established practice in the realm of databases.

Database repair, specifically in the context of relational databases, is a fundamental operation that focuses on modifying a database to conform to specified integrity constraints, such as functional dependencies and inclusion dependencies [2, 5, 3]. Functional dependencies, for instance, denote a relationship where the value of one set of attributes (the determinant) determines the value of another set of attributes. Inclusion dependencies, on the other hand, represent a constraint between two sets of attributes in a relational database, typically used to specify a foreign key constraint. Although the satisfaction of the integrity constraints are usually ensured by DBMSs, in some applications a given database might violate a constraint; e.g., when a new constraint is added to the system. In those cases, repair operations are executed to ensure compliance. The repair should ideally be minimal, implying that the least possible modifications are made to the database to satisfy the constraints. Such operations become critical in ensuring the robustness of databases, preserving data integrity, and improving query processing efficiency. Classic applications of database repair include resolving data inconsistencies caused by data integration, dealing with imperfect data inputs, or correcting errors that arise due to updates or deletions [37, 56, 11]. These repair operations typically involve correcting or eliminating data that violates the given constraints, which can be challenging if the constraints are complex or if there are dependencies between them.

Similarly, in our context, when the CI constraints in the data or its distribution are not met, one can adapt the concept of database repair to "repair" the data. Existing work already draws a connection between testing CI and multivalued dependencies in databases [54, 17]. Repairing with respect to CI can be executed by minimally modifying the data or distribution to satisfy the CI conditions. Like database repair, this process can potentially maintain data integrity, enhance data usability, and facilitate more efficient data analysis. In the following sections, we will delve deeper into this concept, discussing specific examples where such data repair operations were necessitated due to initial CI testing failures. These examples illustrate the potential of data repair in statistical and machine learning applications, introducing an exciting new avenue for research.

1.1.1 Interventional Fairness

Our application of fairness utilizes the principle of "interventional fairness" outlined in Salimi et al. [48], deploying a causal pre-processing technique to promote fairness. In the realm of

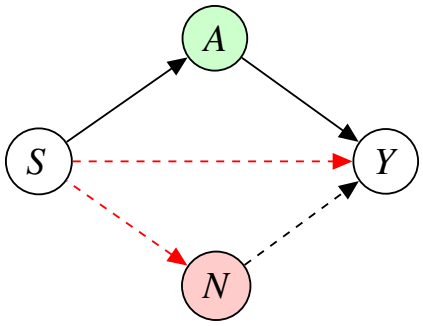


Figure 1.1: Graphical model representing the variables' relationships in the admission example.

fairness, pre-processing techniques address bias before data enters the machine learning model, modifying the training data to diminish bias. This makes such techniques universally adaptable, unlike in-processing fairness methods that infuse fairness into model training, adjusting the learning algorithm or introducing constraints to balance fairness and accuracy. Causal fairness strategies alleviate bias by eliminating spurious causal connections between protected attributes and outcomes, resulting in unbiased decisions.

To illustrate interventional fairness, we refer to a university admission example from Salimi et al. [48]. The goal is to develop a prediction model for admissions, considering applicant information such as qualifications, hobbies, and the department applied to, and providing an admission or rejection outcome. In constructing a fair model, we aim to sever any causal links between the protected attribute, i.e., gender, and the outcome, i.e., the admission result. The straightforward approach of excluding protected attributes before inputting data into the machine learning model referred to as "fairness through unawareness", has proven ineffective [25, 34]. The remaining attributes often correlate with the protected attributes, maintaining spurious connections between the protected and outcome attributes. For instance, hobbies are closely related to gender and thus carry the unintended impact of gender on the outcome. Another approach might suggest eliminating any attribute correlating with the protected attributes, but this is clearly impractical as it would lead to the exclusion of numerous attributes and a significant loss of information for the model.

Interventional fairness goes beyond "fairness through unawareness", removing all spurious causal connections between the protected attribute and the outcome. Consider the graphical model in Figure 1.1 to elaborate on this concept. Here, attributes are divided into four categories: protected attributes S, e.g., gender; the outcome attribute Y, e.g., admission; admissible attributes A, e.g., department; and inadmissible attributes N, e.g., hobby. Both admissible and inadmissible attributes correlate with protected attributes, influencing the outcome. The impact of admissible attributes is acceptable; for example, the choice of the department can affect the chance of admission and may, therefore, carry the impact of gender on the outcome – an impact considered acceptable. In contrast, inadmissible attributes like hobbies carry an unacceptable, spurious influence of gender on the outcome. These spurious links are emphasized by dashed red lines in Figure 1.1.

Interventional fairness aims to remove these spurious links between the protected attribute and the outcome. This is done in two steps: first, enforcing a CI $\sigma : N \perp\!\!\!\perp S \mid A$, and second, removing the protected attribute S. The CI in the first step ensures that when admissible

attributes A are held constant, there is no correlation between the protected attribute S and the inadmissible attribute N. Thus, the unwanted impact of the protected attribute on the outcome is removed, as inadmissible attributes no longer carry information from the protected attribute once the admissible attributes are fixed. This is preferred over removing inadmissible attributes, as their correlation with the outcome, which can contribute to a better prediction, is still preserved. The second step is intended to remove the direct impact of the protected attribute as in fairness through unawareness.

1.1.2 Cleaning through Data Repair with respect to CIs

Our data quality application draws upon the concept of "statistical constraints" as outlined in Yan et al. [57], representing instances of probabilistic dependence or independence. Consider, for instance, the task of creating a regression model to predict car prices. In this scenario, certain expectations of dependence and independence should exist between the input variables (such as color, model, and fuel type) and the target variable (in this case, price). For example, the car model and price are generally not independent, while price and color often are, particularly when the model and make of a car are specified. If these statistical constraints are not satisfied in the training data, the cause could be errors that decrease accuracy when the model is used on unseen test data that don't exhibit the same errors.

The work by Yan et al. [57] presents statistical constraints as conditional dependences or independences expected to be found in a dataset. They also draw comparisons and make connections between statistical constraints and integrity constraints, such as functional dependencies and multi-valued dependencies in databases. In the realm of data quality applications, they showcase how these constraints can aid in error detection. However, they overlook the potential use of these constraints in data cleaning through data repairs, which alter data to satisfy the constraints. In our work, we expand beyond mere error detection and leverage statistical constraints in the form of CI to rectify and clean errors. We note that CI constraints represent a specific form of statistical constraints wherein independence is assumed as opposed to dependence.

1.2 Optimal Transport for Data Repair

This thesis delves into the unique territory of data repair, specifically focusing on CI constraints. However, our approach deviates from conventional techniques that apply repair mechanisms directly to the dataset. Instead, we operate within a probabilistic framework, inferring a probability distribution from the given dataset and modifying this distribution to comply with the CI constraints. This methodology aligns with the fundamental premise of data repair: achieving an optimal repair strategy that satisfies a specific set of constraints—in this case, the CI constraints—while preserving the integrity of the original data as much as possible.

In data repair, the notion of "distance" is crucial. It measures the deviation from the initial data representation, effectively quantifying the extent of the repair. While numerous metrics, such as f-divergences like the Kullback-Leibler (KL) divergence, could serve as this distance measure, this study opts for the Optimal Transport (OT) theory and its associated Wasserstein distance [53]. The theory of OT, rooted in mathematical economics, offers a robust and intuitive

framework for transforming one distribution into another cost-effectively. In this context, the term "cost" is quantified by the Wasserstein distance, a well-defined metric that calculates the least "work" required to reshape one distribution into another. The notion of work here corresponds to the product of the amount of distribution mass moved and the distance it is moved.

In recent years, OT and the Wasserstein distance have gained significant traction due to their potential to handle discrepancies in complex, high-dimensional data distributions. They have found applications in fields as diverse as computer graphics [6], image recognition [19], and machine learning [13, 4, 21], where they are often used for tasks such as domain adaptation, generative modeling, and clustering. The robustness, intuition, and theoretical soundness of OT and the Wasserstein distance make them ideal for quantifying the distance between probability distributions. Hence, they're well-suited for our probabilistic approach to data repair.

The Wasserstein distance choice over measures like the KL-divergence offers two key advantages in our setting. First, it incorporates the mass and spatial location of the distributions in the feature space, resulting in a more natural and intuitive measure of distance. It considers the amount of probability mass and the effort in moving this mass from its original location to a new one. This enables the Wasserstein distance to capture differences in shape, spread, and location between two distributions more holistically than many other metrics. Second, the Wasserstein distance remains finite and well-defined even when the support of the distributions does not overlap. This attribute is particularly beneficial when dealing with distributions with disparate or disjoint supports, a scenario frequently encountered in complex data-driven applications.

1.3 Thesis Contributions and Structure

In this thesis, the principal objective is to tackle the problem of data repair concerning CI constraints, leveraging the powerful concept of OT. We make the following contributions to this thesis:

- We formalize the problem of data repair with respect to CI constraints where we use OT as a principled way to compare probability distributions and measure the deviation of repair from its initial distribution.

- We cast the repair problem to a Quadratically Constrained Linear Program (QCLp). This representation provides an analytical pathway to identify the OT plan. The plan aims to minimize the distance between the original and repaired data distribution while ensuring adherence to the imposed CI constraints. Existing optimization techniques can be employed to solve the QCLP model and yield a precise solution.

- We present a novel algorithm based on relaxed OT and block coordinate descent (BCD) to address the scalability issues inherent to the QCLP approach in data repair. Utilizing entropic regularization and a relaxed version of the OT problem, this method transforms the original non-convex problem into a regularized optimization problem, providing a scalable and efficient solution for large-scale datasets. The developed algorithm ensures

adherence to CI constraints and effectively maintains data fidelity, showcasing its superiority over conventional methods through comprehensive experiments.

- We conduct extensive experiments to showcase the effectiveness of our solutions in minimizing OT for optimal repair and demonstrate the usefulness of repair with respect to CI in the two applications of fairness and data cleaning.

This thesis is structured as follows. In Chapter 4, we delve into the formal definition of our repair problem and unveil our algorithms, built upon QCLP and the relaxed OT. Before this, we provide some necessary background information in Chapter 3, covering aspects such as OT, its entropic and relaxed versions, BCD, and QCLP. Our experimental results, which demonstrate the effectiveness and practicality of our proposed methodologies, are then thoroughly discussed in Chapter 5. Chapter 2 offers a comprehensive review of related concepts, exploring the realms of algorithmic fairness and data quality and cleaning. Finally, in Chapter 6, we provide a summary of our work, its implications, and potential directions for future research.

Chapter 2

Related Work

2.1 Fairness in Machine Learning

The concept of fairness in Machine Learning (ML) has evolved over the years, gaining substantial attention due to its far-reaching implications in ML systems' outcomes.

Measures of fairness are divided into two groups: Group fairness and individual fairness. Satisfying a measure from one of these groups does not necessarily make an improvement for measures from another group. Group fairness measures partition records into some groups usually based on their value of protected attributes (like race, gender, etc.), and checks how similar the outcomes for different partitions are. A widely recognized measure of group fairness is demographic parity [9], which necessitates that decisions are independent of protected attributes. In other words, all demographic groups should have equal probabilities of receiving a particular decision outcome, irrespective of their proportions in the population. An extension of demographic parity is conditional statistical parity [12], which allows for fairness adjustment based on specific non-sensitive attributes. The idea is to achieve a balance within subgroups defined by such attributes, ensuring the decision system does not discriminate against a particular demographic when these attributes are controlled. Furthermore, measures like equalized odds and equal opportunity focus on fairness at the outcome level [26]. Equalized odds require that the True Positive Rate (TPR) and False Positive Rate (FPR) should be the same for different demographic groups. Similarly, equal opportunity requires equal true positive rates across groups, thereby emphasizing the need for fairness among individuals who should receive the same outcome.

Individual fairness, on the other hand, posits that similar individuals should be treated similarly. This involves defining a similarity metric based on the task, allowing for nuanced interpretations of fairness. Since the choice of similarity metric is arbitrary, it needs careful consideration and domain knowledge, making it less applicable in cases where this information is unavailable. For counterfactual fairness as a member of this group, it is shown in the literature that its estimation is not possible from data. [46, 44, 45]

Fairness measures can also be categorized from a different point of view to be either associational or causal measures. Associational fairness is a fundamental group of measures that primarily revolves around statistical dependencies between sensitive attributes and decisions. It doesn't take into consideration how attributes impact each other and what their relationship

is. It basically computes some statistical values (like TPR, FPR, etc.) from data and uses them for reporting fairness level. Demographic parity and equalized odds are some examples. Many methods are proposed in the literature for enforcing associational fairness measures by pre/post-processing input and output data [18, 10, 26] and also modifying decision-making algorithm [9].

While associational measures only consider the statistical constraints, causal fairness, also known as fairness in causality, represents another paradigm in fairness based on causal reasoning rather than purely statistical correlations [29, 22]. The premise is to consider the underlying causes of decisions to ensure fairness, not just the associations observed in the data. For finding discriminatory impacts on outcome attribute, many methodologies can be used, such as causal inference techniques. Then it will be addressed by debiasing data used for training ML models [48] or mitigating the models themselves [29, 39]. So there are no universal solutions for measuring and enforcing these measures, as the sources of discrimination may vary in different cases.

Counterfactual fairness [34] is a prime measure in the domain of causal fairness and belongs to the family of individual fairness measures, which proposes that a decision is fair towards an individual if the same decision would have been made in a "counterfactual" world where the individual belonged to a different demographic group. A challenging part about counterfactual fairness is that its computation can not be done by having only data itself since when we flip the value of the protected attribute, there might not be any record in the dataset with that value telling us the outcome.

Interventional fairness [49] is another example of causal fairness measures that introduces the concept of interventions in causal models. It postulates that a decision is fair if it remains unchanged when an intervention is made on the protected attribute. By ignoring protected and outcome attributes, it considers all subsets of other attributes and fixes their values by removing all edges going to them in the causal graph. Then the distribution of outcome must be equal when trying all possible values of the protected attribute.

2.1.1 Bias in Machine Learning and Approaches to De-Biasing

While fairness measures aim to ensure equitable decisions, biases in ML can undermine these efforts, leading to fairness issues. Bias can infiltrate ML systems at various stages, including data collection, preprocessing, model training, and deployment. One common form of bias is sampling bias, which arises when the training data is not representative of the population. Another type is measurement bias, which occurs when certain demographic groups are systematically misclassified due to limitations in measurement processes. Further, latent bias can emerge from seemingly innocuous features heavily correlated with protected attributes, leading to indirect discrimination. Prejudices in society can also manifest as historical bias in ML systems, propagating and amplifying societal inequalities.

To counteract these biases, researchers have proposed three general approaches to debiasing machine learning models: preprocessing, in-processing, and post-processing. Preprocessing methods aim to modify the training data to reduce or eliminate biases before the learning process begins. Techniques like reweighing, oversampling, or undersampling can adjust the representation of different demographic groups. Other preprocessing methods may involve modifying the features used by the model to prevent correlations with sensitive attributes. In-

processing, or in-training, techniques aim to incorporate fairness constraints directly into the learning process. This often involves modifying the model's objective function to include a term that penalizes unfair predictions. This approach enables the model to balance the trade-off between accuracy and fairness during training. Post-processing techniques, on the other hand, adjust the model's predictions after training to meet specific fairness criteria. For example, they may recalibrate the model's output probabilities or adjust the decision threshold for different demographic groups to ensure fair treatment. Each method offers unique advantages and trade-offs and may be more or less suited to different contexts or applications. The understanding of these de-biasing techniques informs the methodologies used in this thesis, as the pursuit of fairness and the mitigation of bias are both integral to our work of designing an algorithm for optimal data repair that respects fairness constraints.

2.2 Data Cleaning and Data Repair

The fundamental importance of error detection and data cleaning in data management is highlighted by the development of a wide range of strategies aimed at improving data quality. Our extensive survey of the literature reveals that these data-cleaning solutions can be grouped into four broad categories: rule-based methods, probabilistic and statistical approaches, machine learning techniques, and automated machine learning (AutoML) systems.

Rule-based Methods: These methods are often considered the most traditional form of data cleaning. They involve the application of pre-defined rules or constraints derived from domain knowledge or database schema to identify and correct inconsistencies. The concept here is to transform the data cleaning task into a data repair process, where the goal is to modify the data in a way that respects the set rules, thereby preserving the integrity of the original data [24, 56, 11]. For instance, one might employ a rule-based approach to ensure that no employee in a dataset is listed with a salary below the minimum wage. The major limitation of rule-based methods is that they rely heavily on the quality of the defined rules: if the rules are incomplete or incorrectly specified, the cleaning will be suboptimal. Despite this limitation, rule-based methods remain popular due to their transparency and simplicity.

Probabilistic and Statistical Approaches: With the rise of big data, probabilistic and statistical techniques have been introduced to tackle the uncertainty and complexity inherent in modern datasets. These methods typically build probabilistic models that represent what "clean" data should look like and then apply these models to identify and correct errors in "dirty" data. For example, HoloClean, developed by Ilyas et al., uses probabilistic soft logic to infer the likelihood of various data values being errors and predicts the correct values to replace them [42]. By taking into account the inherent uncertainty of data, these methods can handle complex error patterns and dependencies among data attributes, thereby providing more robust data cleaning solutions.

Machine Learning Techniques: Machine learning techniques, especially deep learning, offer a data-driven approach to data cleaning. These methods require labeled training data, which

they use to learn patterns of errors and generate models to identify and correct similar errors in new data. For example, Baran, a system proposed by Heidari et al., employs active learning to iteratively improve its understanding of data quality rules [27]. Due to their ability to learn high-level patterns and generalize from training data, machine learning techniques can detect and correct complex and subtle data errors that may be missed by other methods.

Automated Machine Learning (AutoML) Systems: The latest trend in data cleaning involves integrating the above methods into automated end-to-end systems, commonly referred to as AutoML systems. These systems, such as Raha and HoloClean, provide comprehensive solutions for data cleaning, encompassing error detection, data repair, and subsequent data analysis tasks [42, 38]. By automating the data cleaning pipeline, these systems minimize the amount of manual work and domain expertise required, thereby enabling non-expert users to clean their data efficiently and effectively.

The current field of data cleaning is highly diverse, with a variety of methods available, each with its own strengths and trade-offs. Notably, the work presented in this thesis on a repair with respect to CI constraints could be considered a fusion of the first two approaches: rule-based methods and probabilistic/statistical techniques. It combines the use of rule-based logic, in this case, CI constraints, with probabilistic and statistical modeling to handle uncertainty and complexity in the data. This represents a promising direction for future data cleaning strategies, providing a balance between the interpretability and domain-specificity of rule-based methods and the robustness of probabilistic and statistical approaches. The ongoing challenge for researchers and practitioners is to continue refining these methods, innovating new strategies, and integrating different approaches to effectively handle the increasing complexity of modern datasets while ensuring data privacy and maintaining the high quality of the cleaned data.

Chapter 3

Background

This chapter introduces the fundamental concepts that are vital to the understanding of the rest of the thesis. Starting with probability theory (Section 3.1), it guides the reader through essential domains such as optimal transport (Section 3.2), evaluation metrics for both accuracy and fairness (Section 3.3), optimization programs including Quadratically Constrained Linear Programs (Section 3.4), block coordinate descent (Section 3.5), and non-negative matrix factorization (Section 3.6).

3.1 Basic Concepts from Probability Theory

A probability space $(\Omega, \mathcal{F}, \mu)$ is a mathematical construct that comprises three elements; the sample space Ω that represents the set of all possible outcomes of a random experiment, the event space $\mathcal{F}$ that is a collection of events, where each event is a subset of the sample space, and the probability measure μ that assigns a probability to each event within the event space. For $(\Omega, \mathcal{F}, \mu)$ to form a valid probability space, it has to satisfy certain properties. $\mathcal{F}$ must be non-empty, contain Ω, and be closed under complement and countable union. Also, μ must be a valid probability measure, which means the probability of the union of any disjoint events must equal the sum of the probabilities of each event individually, and the probability of the entire sample space occurring must be one. In essence, a probability space models a random procedure by delineating the possible outcomes and assigning probabilities to certain subsets of those outcomes. The pair $(\Omega, \mathcal{F})$, without the probability measure μ, is often called a measurable space.

Other essential concepts in probability theory include random variables and probability distributions. Given a sample space Ω, a random variable X is a real-valued function $X : \Omega \mapsto \mathbb{R}$ that maps the samples in Ω to real numbers. Ω is often called the support of X, which we denote by $\mathcal{X}$. In this work, we consider discrete random variables where the range of X in $\mathbb{R}$ is a countable set. A probability distribution P with the random variable X is a function that specifies the likelihood of the values or subsets of values in the range of X, e.g., $P(X \leq x)$ for $x \in \mathbb{R}$ is the likelihood that X returns a value less than x. Also, $P(X = x)$, which is often shortened to $P(x)$, is the likelihood that X returns exactly x. Note that the probabilities returned from P correspond to the probabilities of the samples mapped to those values returned from X occur. Thus, a probability distribution P with random variable X, specifies a probability space

$(\Omega, \mathcal{F}_X, \mu_X)$ where $|\mathcal{F}_X| = 2^\Omega$ and μ_X is $\mu_X(A) = \sum_{x \in A} P(X = x)$ for any $A \in \mathcal{F}_X$.

Given two random variables X and Y with respective sample spaces Ω_X and Ω_Y, the joint probability distribution of X and Y is a function $P_{X,Y}$ that specifies the likelihood of any subset of $\Omega_X \times \Omega_Y$, e.g., $P_{X,Y}(X = x, Y = y) = P_{X,Y}(x, y)$ specifies the likelihood that samples from Ω_X and Ω_Y happen that are mapped to x and y. The marginal probability of X is obtained by summing the joint probability over all possible values of Y, and similarly for Y: $P_X(x) = \sum_y P_{X,Y}(x, y)$ and $P_Y(y) = \sum_x P_{X,Y}(x, y)$. For the joint probability distribution $P_{X,Y}$, the conditional probability of X given Y is defined as $P_{X|Y}(x \mid y) = P_{X,Y}(x, y)/P_Y(y)$ where $P_Y(y) > 0$.

Let us assume a set $\mathbf{V}$ of random variables. We use bold capital letters to refer to sets of random variables. Let $\mathbf{V}$ include three random variables $X, Y, Z \in \mathbf{V}$. We denote X as being *conditionally independent* (CI) of Y in P, given Z, represented as $\sigma : Y \perp\!\!\!\perp X \mid Z$, if the conditional joint probability $P_{X,Y|Z}(x, y|z)$ equals the product of the conditional probabilities $P_{X|Z}(x|z)$ and $P_{Y|Z}(y|z)$. We define $\sigma : Y \perp\!\!\!\perp X \mid Z$ as a CI constraint. A probability distribution P is said to be consistent with σ or to satisfy σ, represented as $P \models \sigma$ if the conditional independence $Y \perp\!\!\!\perp X \mid Z$ holds in P. A CI constraint is termed *saturated* if $\mathbf{V} = \{Y, X, Z\}$. A set of CI constraints, represented as Σ, is satisfied by P (i.e., $P \models \Sigma$) if P satisfies every CI constraint present in Σ. A CI constraint naturally extends to sets of random variables $\mathbf{X}, \mathbf{Y}$, and $\mathbf{Z}$ in $\mathbf{V}$.

When $\sigma : Y \perp\!\!\!\perp X \mid Z$ is not satisfied by P, we quantify the violation using *conditional mutual information (CMI)*, denoted as $I(X; Y \mid Z)$. This measure captures the information gained about Y by knowing X, given Z. For CI, $P_{X,Y|Z}$ should equal $P_{X|Z} \otimes P_{Y|Z}$, where $\otimes$ is the product of two functions, so it is natural to measure their divergence. CMI uses Kullback-Leibler (KL) divergence for this purpose for every fixed value of Z:

$$\begin{aligned} I(X; Y \mid Z) &= \mathbb{E}_Z[KL(P_{X,Y|Z} \parallel P_{X|Z} \otimes P_{Y|Z})] \\ &= \sum_{c \in \mathcal{Z}} P_Z(c) \sum_{a \in \mathcal{X}} \sum_{b \in \mathcal{Y}} P_{X,Y|Z}(a, b \mid c) \log \left(\frac{P_{X,Y|Z}(a, b \mid c)}{P_{X|Z}(a \mid c) \cdot P_{Y|Z}(b \mid c)} \right) \end{aligned} \tag{3.1}$$

As a result, it returns zero when CI holds.

Consider a dataset $D = \{r_1, ..., r_n\}$ with attributes $A_1, ..., A_m$. We use $\text{DOM}(A_j)$ to refer to the domain of attribute A_j and $r_i[A_j]$ to refer to the value of r_i for attribute A_j. The empirical distribution of D, denoted by P_D, is a joint probability distribution with m random variables $X_1, ..., X_m$ with supports $\mathcal{X}_j = \text{DOM}(A_j)$. The sample space Ω of P_D is the set of possible records $\text{DOM}(A_1) \times .... \times \text{DOM}(A_m)$ and P_D is defined as $P_D(r) = f_D(r)/n$ where $f_D(r)$ is a function that returns the number of times record r appears in D. In this work, when given a dataset D, we estimate the probability distribution from which the dataset is sampled using P_D. However, our problem formulation and algorithms are amendable to other ways of estimating a distribution from data.

3.2 Optimal Transport

The Optimal Transport (OT) problem seeks to determine the most efficient way of transferring mass from one probability distribution to another while preserving the total mass. The

OT problem's classical formulation is *the Monge problem* where the objective is to identify a *transport map* $\mathcal{T}$ that pushes a measure from a source measurable space forward to a measure in a target measurable space. Here, we explain it using probability distribution functions as a special case. $\mathcal{T}$ pushes a probability function P with random variables $\mathbf{X}$ forward to a distribution function Q with random variables $\mathbf{Y}$, while minimizing the total cost of transporting mass. Formally, Q, known as *the pushforward* of P under the transport map $\mathcal{T}$, is a new distribution function defined as $Q(A) = P(\mathcal{T}^{-1}(A))$ for any set $A \subseteq \mathcal{Y}$. In other words, the pushforward Q characterizes the distribution of the images of P under the map $\mathcal{T}$ from $\mathcal{X}$ to $\mathcal{Y}$.

With this definition, the Monge problem can be formally defined as follows: Given two distributions P and Q supported on finite discrete domains $\mathcal{X}$ and $\mathcal{Y}$, respectively, and a cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$, the goal is to find a transport map $\mathcal{T} : \mathcal{X} \to \mathcal{Y}$ that pushes forward P to Q, such that the total cost of transporting mass is minimized:

$$\mathrm{OT}_{Monge}(P, Q) = \underset{\mathcal{T}:\mathcal{X}\to\mathcal{Y}}{\operatorname{argmin}} \sum_{\bar{a}\in\mathcal{X}} c(\bar{a}, \mathcal{T}(\bar{a})), \tag{3.2}$$

where $\mathcal{T}$ is a transport map and $\mathcal{T}_{\#}P = Q$.

The Monge problem does not always have a solution, meaning a pushforward may not exist between two distributions. This can be shown by a counterexample where the support of P is one and the support of Q is greater than one. Since mass can not be split, no transport map can result in the final measure, regardless of the choice of the cost function. To address this limitation, *the Kantorovich relaxation* relaxes the Monge problem to allow for *couplings* (or *transport plans*). A coupling is a joint distribution M over a discrete space $\mathcal{X} \times \mathcal{Y}$ that has marginal probabilities P and Q over $\mathcal{X}$ and $\mathcal{Y}$: $P = M_{\mathbf{X}}$ and $Q = M_{\mathbf{Y}}$. We use $\Pi(P, Q)$ to refer to the set of all such couplings. The Kantorovich relaxation of the Monge problem is referred to as *the primal Kantorovich problem* and is defined as follows:

$$\mathrm{OT}(P, Q) = \underset{M\in\Pi(P,Q)}{\operatorname{argmin}} \sum_{\bar{a}\in\mathcal{X}} \sum_{\bar{b}\in\mathcal{Y}} c(\bar{a}, \bar{b}) M(\bar{a}, \bar{b}). \tag{3.3}$$

The optimal coupling M minimizes the transport cost in Equation 3.3 which is referred to as *the Wasserstein distance* between P and Q, denoted by $W(P, Q)$, when c is the Euclidean distance.

3.2.1 Entropic Optimal Transport

Finding the optimal coupling in Equation 3.3 is costly, and entropic regularization is used to optimize it. The entropic OT is defined as follows:

$$\underset{M\in\Pi(P,Q)}{\operatorname{argmin}} \sum_{\bar{a}\in\mathcal{X}} \sum_{\bar{b}\in\mathcal{Y}} c(\bar{a}, \bar{b}) M(\bar{a}, \bar{b}) - \frac{1}{\rho} H(M), \tag{3.4}$$

where $H(M)$ is the entropic regularizer:

$$H(M) = \sum_{\bar{a}\in\mathcal{X}} \sum_{\bar{b}\in\mathcal{Y}} M(\bar{a}, \bar{b}) \log(M(\bar{a}, \bar{b})),$$

with the entropic regularization parameter $1/\rho$. The coupling M in Equation 3.4 has marginals P and Q, similar to the formulation in Equation 3.3. We denote the distance obtained from the entropic OT by $W_\rho(P, Q)$.

The entropic OT formulation allows an efficient iterative algorithm, called *the Sinkhorn algorithm*, for finding M, which consists of the following steps (see [14] for more detail):

1. Initialize a vector of variables $\bar{v}$ with $\bar{1}_{d_\mathbf{Y}}$, the 1-vector of size $d_\mathbf{Y} = |\text{DOM}(Y)|$.
2. Compute the matrix $K = e^{-\rho C - 1}$ where C is a matrix of size $d_\mathbf{X} \times d_\mathbf{Y}$ representing the cost function c.
3. Iteratively update two vectors of variables $\bar{u}$ and $\bar{v}$ as follows until convergence:
$$\bar{u} = \bar{a} \oslash (K \cdot \bar{v}) \qquad \text{and} \qquad \bar{v} = \bar{b} \oslash (K^\top \cdot \bar{u})$$
where $\bar{a}$ and $\bar{b}$ are vectors of size $d_\mathbf{X}$ and $d_\mathbf{Y}$, respectively, and represent the probability values in the probability functions P and Q, and $\oslash$ is the element-wise vector division.
4. Compute the matrix $diag(\bar{u}) \cdot K \cdot diag(\bar{v}))$ of size $d_\mathbf{X} \times d_\mathbf{Y}$ where $diag(\bar{u})$ and $diag(\bar{v})$ are the diagonal matrices of sizes $d_\mathbf{X}^2$ and $d_\mathbf{Y}^2$ with entries from the vectors $\bar{u}$ and $\bar{v}$, respectively. Then use the matrix to generate a coupling M and return it.

The parameter $1/\rho \in [0, +\infty)$ controls the amount of smoothing in the solution; a smaller value of $1/\rho$ (close to 0) results in a more precise but possibly spiky solution, while a larger value of $1/\rho$ results in a smoother but possibly less precise solution. The algorithm converges linearly[32], resulting in a coupling M that provides an approximate solution to the optimal transport problem. This means $W_\rho(P, Q)$ is an estimation of $W(P, Q)$ and reduces to W if ρ is large enough.

Relaxed Optimal Transport Relaxed OT was first introduced in [21] for measuring training loss. It is also used in [1] for sparse matrix factorization. The main idea is to replace the hard constraints for marginal distributions with soft penalties with respect to KL divergence to be added to Equation 3.4:

$$\underset{M:\mathcal{X}\times\mathcal{Y}\to\mathbb{R}_{\geq 0}}{\text{argmin}} \sum_{\bar{a}\in\mathcal{X}} \sum_{\bar{b}\in\mathcal{Y}} c(\bar{a}, \bar{b}) M(\bar{a}, \bar{b}) - \frac{1}{\rho} H(M) \; + \lambda(KL(M_\mathbf{Y}, Q) + KL(M_\mathbf{X}, P)). \tag{3.5}$$

Here, $M_\mathbf{Y}$ and $M_\mathbf{X}(x)$ are the marginal probability functions obtained from the coupling M, λ is the relaxation regularization coefficient, and KL refers to the KL divergence between two probability distributions. Compared with the entropic OT in Equation 3.4, The space of M is relaxed to be any function with domain $\mathcal{X} \times \mathcal{Y}$ and range $\mathbb{R}_{\geq 0}$, and the new regularizer $\lambda(KL(M_\mathbf{Y}, Q) + KL(M_\mathbf{X}, P))$ represents a soft constraint that requires M to be a probability distribution with marginals equal to P and Q. We denote the distance obtained from this relaxed entropic OT by $W_{Rx}(P, Q)$.

The Sinkhorn algorithm also works for the relaxed version of the entropic OT in Equation 4.4 but with different update rules for $\bar{u}$ and $\bar{v}$ that considers the regularization parameters ρ and λ (see [21, Proposition 4.2] for more detail):

$$\bar{u} = (\bar{a} \oslash (K \cdot \bar{v}))^{\frac{\rho\lambda}{\rho\lambda+1}} \qquad \text{and} \qquad \bar{v} = (\bar{b} \oslash (K^\top \cdot \bar{u}))^{\frac{\rho\lambda}{\rho\lambda+1}} \tag{3.6}$$

We use relaxed OT and its soft constraint to represent the CI constraints in our repair.

3.3 Evaluation Measures

In our experimental chapter, we use several measures to evaluate the accuracy and fairness of the machine learning models in this work. We briefly review these measures in this section.

3.3.1 Accuracy Measures

Machine learning models are typically evaluated using a variety of measures that capture different aspects of the model's performance. This thesis uses the following measures: accuracy, precision, recall, F1 score, and AUC.

- *Accuracy* is the most straightforward metric, defined as the ratio of correctly predicted observations to the total observations. It generally measures how well the model performs across all classes.
- *Precision* is the ratio of correctly predicted positive observations to the total predicted positives. High precision indicates a low false positive rate, meaning the model is reliable when it predicts a positive class.
- *Recall (Sensitivity)* is the ratio of correctly predicted positive observations to all observations in the actual class. It captures the ability of the model to find all the positive instances.
- *F1 Score* is the harmonic mean of precision and recall, balancing the two metrics. It is beneficial in situations where the data has imbalanced classes.
- *AUC* stands for Area Under the Receiver Operating Characteristics curve or ROC. It illustrates the performance of a binary classifier as its discrimination threshold is varied. The AUC measures the entire two-dimensional area underneath the ROC curve, providing an aggregate performance measure across all possible classification thresholds. A model whose predictions are 100% wrong has an AUC of 0.0, while a model whose predictions are 100% correct has an AUC of 1.0.

Each of these metrics provides different insights, and their use depends on your machine learning task's specific objectives and constraints. The best measure will depend on these factors, as well as the specific characteristics of your data.

3.3.2 Fairness and Bias Measures

The widespread use of machine learning models in critical decisions that significantly impact individuals' lives, such as in credit scoring, hiring, or criminal justice, has raised concerns about fairness. Fairness in machine learning refers to the absence of systematic bias or discrimination in the decisions made by machine learning models. A fair model does not disproportionately harm or benefit any particular group of individuals based on their protected attributes, such as race, gender, age, or religion. However, achieving fairness is challenging due to biases in training data that reflect historical or societal inequalities. Furthermore, fairness is subjective

and context-dependent, with different stakeholders holding varied perspectives on what fairness entails in a given context. Multiple fairness measures have been proposed, though no single measure can capture all fairness aspects, and appropriateness depends on the context. A comprehensive evaluation may involve multiple metrics, considering their strengths and limitations.

Fairness assessment in machine learning can be approached from individual or group perspectives, resulting in two primary fairness categories: Individual and group fairness. Individual fairness pertains to treating similar individuals, while group fairness ensures fair outcomes for groups concerning a protected attribute. This project primarily focuses on improving group fairness, using various metrics, including:

- *Demographic parity (DP)* measures whether the positive outcome probability is the same across different groups concerning a protected attribute.
- *Equal Opportunity (EO)* measures if the true positive rate is the same across different groups concerning a protected attribute.

Beyond DP and EO, several other standard fairness measures, such as Equalized Odds (EOD), Treatment Equality (TE), Predictive Parity (PP), False Positive Rate Equality (FPR Equality), False Negative Rate Equality (FNR Equality), Overall Accuracy Equality (OAE), and Disparate Impact (DI), contribute to different facets of fairness. A single measure cannot encapsulate all fairness aspects; hence, the context determines the suitability of each measure. A thorough fairness assessment may necessitate multiple metrics, carefully considering their relative strengths and application-specific limitations. To evaluate the fairness of machine learning models in this work, we employ DP and EO because of their widespread use, simplicity, and straightforward implementation. We also use a less common measure called *ratio of observational discrimination (ROD)* that is specifically useful to evaluate interventional fairness [48]. We explain ROD in detail in Section 5.1.

3.4 Optimization Programs

This section provides an overview of linear optimization programs and quadratically constrained quadratic programs, which will be used to solve the repair problem in this thesis. *Linear Optimization Programs (LOPs)* are optimization problems where the objective function and constraints are linear functions of a set of variables. More precisely, the objective function and the constraints are often written in the following general forms:

$$f(\bar{x}) = \bar{c}^T \bar{x} \tag{3.7}$$

$$A\bar{x} \leq \bar{b} \tag{3.8}$$

where $\bar{x}$ is the vector of decision variables and has size n, $\bar{c} \in \mathbb{R}^n$ is the coefficient vector that defines the objective function $f(\bar{x})$ in Equation 3.7. The constraints in Equation 3.8 are characterized by the constraint matrix $A \in \mathbb{R}^{m \times n}$ and the vector $b \in \mathbb{R}^m$ that specify constraint bounds. The goal in LOPs is to find the values of the decision variables $\bar{x}$ that optimize the objective function in Equation 3.7 while satisfying the linear constraints Equation 3.8. LOPs have

been extensively studied and can be efficiently solved using algorithms such as the Simplex or interior-point methods.

Quadratically Constrained Linear Programs (QCLPs) are a type of optimization problem involving linear objective functions and quadratic constraints. They generalize LOPs and are a particular case of *Quadratically Constrained Quadratic Programs (QCQP)*, where both the objective and constraints are the decision variables' quadratic functions. The general formulation of a QCQP is as follows:

$$f(\bar{x}) = \frac{1}{2}\bar{x}^T F_0 \bar{x} + \bar{c}^T \bar{x} \tag{3.9}$$

$$\frac{1}{2}\bar{x}^T F_i \bar{x} + \bar{a}_i^T \bar{x} \leq b_i, \quad i \in \{1, 2, \ldots, m\} \tag{3.10}$$

where $\bar{x}$ is the vector of variables, $F_0, F_i \in \mathbb{R}^{n \times n}$ are symmetric matrices, $\bar{c}, \bar{a}_i \in \mathbb{R}^n$ are coefficient vectors, $b_i \in \mathbb{R}$ are scalar constants bounds, and m is the number of quadratic constraints.

The objective function in Equation 3.9 and the constraints in Equation 3.10 are quadratic. QCQPs are more general and flexible than linear optimization problems. QCQP and QCLP are non-convex optimization problems that cannot be solved efficiently. Solving QCQPs is generally NP-hard, but some relaxations make it feasible in practice. Interior-point, Semi-Definite Programming (SDP), Second-Order Cone Programming (SOCP), and Linear Programming (LP) relaxations are some of them which can guarantee the exact solution in different scenarios. If diagonal elements of F_is are all zero, then the optimal value returned by SDP, SOCP, and LP would be equal [31]. It is shown that when solving non-convex QCQPs with non-positive off-diagonal elements in F_ks, SDP and SOCP relaxations are equivalent to solving the original problem [30]. Also, if all coefficient matrices are positive-definite, the problem becomes convex and interior-point and SDP methods can be used.

The problem in this thesis is QCLP, a particular case of QCQP, where the objective function is linear as in Equation 3.7, and the constraints are quadratic as in Equation 3.10. These programs are solved using the same techniques for QCQP and are computationally expensive.

3.5 Block Coordinate Descent

The concept of local search is the backbone of a fundamental optimization algorithm known as gradient descent. Given a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, the procedure begins with an arbitrary starting point $\bar{x}_0 \in \mathbb{R}^n$ and improves this guess iteratively by advancing in the direction of the steepest descent. This direction is determined by the negative gradient of the function at the current point, $-\nabla f(\bar{x}_k)$. The gradient descent update rule can formally be stated as follows:

$$\bar{x}_{k+1} = \bar{x}_k - \alpha \nabla f(\bar{x}_k), \tag{3.11}$$

where $\nabla f(\bar{x}_k)$ is the gradient of f evaluated at the point $\bar{x}_k$, and α is the step size. The choice of α is crucial as it influences the convergence properties of the algorithm and the speed at which it converges. An excessively large α can cause the algorithm to oscillate or diverge, while a very small α may result in very slow convergence.

However, the standard gradient descent method may prove inefficient when it comes to non-convex optimization problems, such as QCLP. Non-convex functions can have several local

minima, complicating the search for the global minimum. As a local optimization method, gradient descent is more likely to converge to a local minimum, the selection of which is heavily influenced by the choice of the initial starting point.

One way to navigate the complex landscape of non-convex optimization problems is through the *Block Coordinate Descent (BCD)* method. In BCD, the parameter vector $\bar{x}$ is partitioned into m subsets, referred to as "blocks" or "coordinates." These blocks may correspond to single parameters or groups of parameters. The optimization process then cyclically minimizes the objective function for one block at a time while holding all other blocks constant.

Formally, the BCD update rule can be written as follows:

$$\bar{x}^i_{k+1} = \bar{x}^i_k - \alpha \nabla f_i(\bar{x}^i_k), \tag{3.12}$$

where $\bar{x}^i_k$ denotes the i-th block at the k-th iteration, and $\nabla f_i(\bar{x}^i_k)$ is the gradient of f with respect to the i-th block evaluated at $\bar{x}^i_k$. The main advantage of BCD is its ability to significantly cut down the computational cost of each iteration, particularly when the problem can be easily minimized for each block.

Despite its efficiency and simplicity, BCD, like its parent gradient descent, can also converge to non-optimal stationary points in the case of non-convex problems, including local minima or saddle points. However, it has gained wide acceptance in many practical applications due to its often satisfactory performance and because, under certain conditions, it can converge to the global minimum. One general condition for ensuring convergence is called the cyclic rule that says i-th block should be optimized in iterations $i, i + m, i + 2m, ...$ to ensure each block is optimized at least once in each consecutive m iterations. [51] (regardless of the order of optimizing them) We will show the satisfaction of this rule in our solution.

3.6 Non-negative Matrix Factorization

Non-negative Matrix Factorization (NMF) refers to the factorization of a non-negative matrix V into two matrices, W and H, such that all three matrices consist of non-negative elements. The non-negativity condition facilitates a more intuitive interpretation of the resulting matrices. In mathematical terms, the NMF of a non-negative matrix V is represented as $V \approx WH$, where W and H are non-negative matrices.

NMF has found successful application in various fields, including image analysis, text mining, and bioinformatics, attributed to its dimensionality reduction capabilities and its potential to generate parts-based representations of data [36, 55, 8, 23]. For instance, NMF is employed in image analysis for face recognition by learning a parts-based representation of human faces. In text mining, NMF assists in extracting topics from large text corpora by interpreting the factor matrices W and H as document-to-topic and word-to-topic mappings, respectively. In the field of bioinformatics, NMF helps identify patterns or clusters in gene expression data, thereby extracting biologically significant information from these patterns.

The task of data repairing concerning conditional independence involves rectifying data to ensure it satisfies a set of Conditional Independence (CI) constraints. These constraints are often derived from domain expertise or a learned statistical model. The common approach to data repairing involves the use of probabilistic models, but this often encounters computational complexity issues, particularly in high dimensions.

In this context, NMF, as a low-rank approximation method, can be instrumental in reducing data dimensionality and making the problem more tractable. More specifically, NMF can be utilized to generate a reduced-dimensionality representation of the original data matrix while preserving its non-negative characteristics. This representation can then be used as an input to CI testing methods, effectively minimizing the computational complexity associated with the repairing process. Additionally, the parts-based representation of NMF allows for the intuitive interpretation of the factorized matrices, thereby facilitating the diagnosis and repair of CI constraint violations. For example, if the CI constraints correspond to missing links in a graphical model representing the data, the matrices W and H derived from the NMF can provide insights into the parts of the data most significantly contributing to these missing links. Despite NMF not directly enforcing CI constraints during the factorization process, its capability for dimensionality reduction, interpretability of its results, and computational efficiency establish it as a valuable tool in data repairing tasks concerning conditional independence.

In this work, we employ NMF as a baseline for our repair algorithms. Specifically, our repair algorithm commences with a repair generated through NMF and refines it to derive an optimal repair.

Chapter 4

Methodology

This chapter provides a formal definition of our repair problem in Section 4.1, followed by the proposition of two distinct solutions for the problem in Section 4.2. Both solutions are designed for saturated cases where all attributes appear in the constraint but we will show in Section 4.2.3 that our solution can also be used for unsaturated cases. Through the discussion in this chapter, we provide an efficient and robust solution to the data repair problem at hand.

4.1 Problem Definition

Data repair primarily involves modifying a dataset to align with a pre-established set of constraints. As explored in Chapter 1, some applications that involve a data distribution necessitate that this distribution adheres to specific CI constraints. Therefore, it is essential to "repair" the distribution itself, regardless of its representation, to satisfy these CI requirements. In this section, we clarify and formalize our understanding of the repair concept utilized in this study.

Definition 4.1.1. Consider a probability distribution P with a set of random variables $\mathbf{V} = \{X, Y, Z\}$. We denote the set of all possible probability distributions with the same random variables over the same sample space as $\Delta(\mathbf{V})$. Let $\sigma : X \perp\!\!\!\perp Y \mid Z$, be a saturated CI constraint that is not satisfied by P, i.e., $P \not\models \sigma$. An *optimal repair* of P with respect to σ is a probability distribution $Q \in \Delta(\mathbf{V})$ that satisfies the following conditions:

1. Q is a repair, i.e., it is consistent with $\sigma : \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$ ($Q \models \sigma$)
2. Q is optimal, i.e., for every repair $Q' \in \Delta(\mathbf{V})$, $\mathit{Dist}(P, Q) \leq \mathit{Dist}(P, Q')$, where *Dist* is a distance between probability distributions.

We denote the set of all optimal repairs of P with respect to σ by $\mathit{repairs}(P, \sigma)$. For every repair $Q \in \mathit{repairs}(P, \sigma)$, we use $\mathit{couplings}(P, Q)$ to denote the couplings that convert P to Q. ■

Regarding the repair problem as outlined in Definition 4.1.1, a few points warrant further discussion. Initially, we utilize the Wasserstein distance for *Dist*, as justified in Chapter 1. In Chapter 3, we discussed that the sample space for $\mathbf{V}$ is formed by the Cartesian product of the supports of the random variables $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$. This implies that certain values in P that originally had zero probability could assume positive probabilities in Q. This observation

underscores that our methods do more than merely re-weighting existing data points; they can also introduce new ones. Concerning the number of potential optimal repairs, the set of feasible repairs $repairs(P, \sigma)$ is never empty. This is because at least one probability function always complies with σ, which can invariably be derived from P through a particular coupling. This is justified by considering that there is no restriction on the repaired measure, so for example, a constant probability function (representing a uniform distribution) is always a repair. Moreover, the set of possible repairs may encompass more than a single repair. This scenario arises when the identical mass from P can be transported to various positions in Q to fulfill σ, thereby leading to different distributions Q. In the context of an optimal repair Q, there could be multiple possible couplings M in $couplings(P, Q)$, as masses from diverse positions of P can be transported to generate the same optimal repair Q.

The application of OT_{Monge} and a transport map to define the distance W in Definition 4.1.1 alters the repair problem, as the set $repairs(P, \sigma)$ could potentially be empty. However, when a repair Q exists, one or more couplings may still be derived from the possible Monge's maps.

Another point of note pertains to the distribution P. In practical scenarios, P is typically unknown, and only a dataset D sampled from this distribution is accessible. Consequently, we operate under the assumption that P can be effectively estimated from D using a variety of existing estimation techniques, such as Maximum Likelihood Estimation (MLE), Kernel Density Estimation (KDE), and Generative Adversarial Networks (GANs). In our experiments, we determine the empirical distribution of a dataset and defer the exploration of other techniques to future work.

The final discussion point relates to the resulting distribution (optimal repair) Q. Starting with dataset D and subsequently estimating P, a repaired dataset can be produced by sampling from Q, as illustrated in our experimental section. When utilizing OT_{Monge}, it is possible to apply the Monge mapping to transform each data point in the input dataset into a point in the repaired dataset. This implies that repair can be implemented at the record level. Depending on the specific application, the distribution Q may be employed in various other ways.

4.2 Computing Optimal Repairs

Obtaining an optimal repair as defined in Definition 4.1.1 can be computationally expensive. This is because there are numerous possible candidate repairs, and computing the Wasserstein distance for each candidate requires solving the optimization problem in Equation 3.3. To address this issue, we propose two solutions that, given a probability distribution P and a CI constraint σ, compute an optimal repair Q.

The first solution involves minimizing the Wasserstein distance between the initial and reconstructed distribution while ensuring that the CI constraint is satisfied through a QCLP detailed in Section 4.2.1. This is an exact solution as it finds an optimal repair. The second approach uses a relaxed version of the OT and satisfies the CI constraint through the structure of the reconstructed distribution. This approach is referred to as the approximate solution.

4.2.1 QCLP Formulation

We propose a QCLP to find an optimal repair that minimizes the optimal transport distance between the initial and reconstructed distribution while satisfying the given CI constraint $\sigma : X \perp\!\!\!\perp Y \mid Z$. The inputs for this program are the probability distribution P, the CI constraint σ, and the cost function c. We assume that the given CI constraint σ is saturated (i.e., $\mathbf{V} = \{X, Y, Z\}$) and X, Y, Z are individual random variables rather than sets of random variables for ease of explanation. The discussions can be trivially extended to sets of random variables. In Chapter 6, we explain an extension to unsaturated CI constraints.

The QCLP consists of two types of decision variables that are organized in a vector $\tilde{q}_i$ and a matrix $\tilde{M}_{i,j}$:

- The *repair variables* $\tilde{q}_i$ with $i \in [1, d_V]$ that represent the probabilities in the optimal repair Q and form a stochastic vector with values summing to 1.
- The *coupling variables* $\tilde{M}_{i,j}$ with $i, j \in [1, d_V]$ that represent the coupling which gives the optimal repair Q. This is a *doubly stochastic matrix*, a square matrix of probability values with each row and column summing to 1.

The variables in the QCLP formulation carry probabilities; hence, they are restricted to the range $[0, 1]$. We assume that the sets of values $\text{DOM}(X)$, $\text{DOM}(Y)$, and $\text{DOM}(Z)$ are ordered, and we define an order for the elements in $\text{DOM}(V)$. This order on $\text{DOM}(V)$ determines the order of the repair variables in $\tilde{q}$, as P is represented with three random variables $\bar{V} = \{X, Y, Z\}$. The coupling variables $\tilde{M}$ represent a joint probability distribution between the input and the repair distributions for $\bar{V}$. Hence, the order on $\text{DOM}(V)$ also determines the order of the variables in the matrix $\tilde{M}$. This particular order of the variables in $\tilde{q}$ and $\tilde{M}$ is useful in expressing the constraints in the QCLP formulation.

To facilitate writing the constraints in the QCLP, we introduce an *index function* that allows us to access elements in $\tilde{q}$:

$$idx(i, j, k) = (k - 1) \times (d_X \times d_Y) + (j - 1) \times d_X + i,$$

where d_X, d_Y, d_Z are the domain sizes of X, Y, Z, respectively. The index function maps a triple of indices (i, j, k) to a unique index in the vector $\tilde{q}$. Using this index function, we can easily access variables in $\tilde{q}$ corresponding to the random variables X, Y, and Z. For instance, the variable $\tilde{q}_{idx(i,j,k)}$ corresponds to the probability for the i-th value of X, the j-th value of Y, and the k-th value of Z.

We can now formulate the QCLP with the following objective function:

$$\min_{\tilde{M}, \tilde{q}} \sum_{i=1}^{d_V} \sum_{j=1}^{d_V} c(\bar{e}_i, \bar{e}_j) \times \tilde{M}_{i,j} \tag{4.1}$$

where we assume $\text{DOM}(V) = \{\bar{e}_1, \bar{e}_2, ..., \bar{e}_{d_V}\}$. The program includes three linear constraints that the first two ones ensure the coupling specified by $\tilde{M}$ is consistent with its marginal probabilities specified by P and Q, and the last one guarantees the non-negativity of the coupling

variables $\tilde{M}_{i,j}$s:

$$\sum_{i=1}^{d_V} \tilde{M}_{i,j} = \tilde{q}_j \qquad \sum_{j=1}^{d_V} \tilde{M}_{i,j} = P(\bar{e}_i) \qquad \tilde{M}_{i,j} \geq 0 \tag{4.2}$$

where $P(\bar{e}_i)$ represents constant probabilities for $\bar{e}_i \in \text{Dom}(V)$. Furthermore, the following quadratic constraints ensure that the CI constraint σ is satisfied by the distribution specified by $\tilde{Q}$:

$$\sum_{t=1}^{d_Y} \tilde{q}_{idx(i,t,k)} \times \sum_{t=1}^{d_X} \tilde{q}_{idx(t,j,k)} = \tilde{q}_{idx(i,j,k)} \times \sum_{t=1}^{d_X} \sum_{t'=1}^{d_Y} \tilde{q}_{idx(t,t',k)}. \tag{4.3}$$

There is one constraint for every $i \in [1, d_X], j \in [1, d_Y], k \in [1, d_Z]$. Equation 4.3 represents $Q_{X,Z}(x, z) \times Q_{Y,Z}(y, z) = Q(x, y, z) \times Q_Z(z)$, which is equivalent to $Q_{X|Z}(x, z) \times Q_{Y|Z}(y, z) = Q_{X,Y|Z}(x, y \mid z)$ and expresses CI, as we explained in Chapter 3. In addition to these constraints, the program includes constraints that ensure the variables are probability values (in the range $[0, 1]$).

The program specified by the objective function in Equation 4.1 and the constraints in Equations 4.2 and 4.3 is a QCLP [52]. The objective function is a linear function of the coupling variables, and the constraints in Equation4.2 are also linear concerning the repair variables. However, the constraints in Equation 4.3 are non-linear and quadratic, as the left side of each constraint is a product of the sum of some repair variables. The QCLP falls in the category of problems known as QCQPs or Second-Order Cone Programs (SOCPs) with diverse applications in finance, control systems, signal processing, and other fields. QCLP is NP-hard and is considered a non-convex optimization problem [52, 7]. Various efficient methods such as sequential quadratic programming, augmented Lagrangian, interior-point, and active set have been used to find sub-optimal solutions for these programs [7]. To obtain an optimal repair by solving the QCLP program, we implemented an alternating algorithm that linearizes the quadratic constraints iteratively.

Analysis of the QCLP solution: The existing efficient algorithms for solving QCLP guarantee convergence. However, they have two issues when applied to our application. They generally return only a sub-optimal solution and do not scale in our QCLP solution. The latter is because the number of variables in our solution grows exponentially with the number of attributes. The program in our QCLP solution comprises d_V^2 coupling variables, d_V repair variables, and $d_V^2 + 2d_V$ constraints specified in Equation 4.2 and d_V constraints specified in Equation 4.3. The domain size d_V includes exponential values in the number of attributes in $\mathbf{V}$. Therefore, solving the QCLP is computationally infeasible for such scenarios. We propose a second solution based on relaxed OT to address this issue.

4.2.2 BCD with Relaxed OT

The limitations of the initial approach lead us to explore alternative solutions. While these alternatives might trade off some degree of accuracy, they offer significantly increased computational efficiency. One such solution is built upon an approximate algorithm for computing the relaxed OT, using the Sinkhorn algorithm as a foundational component.

BCD is a common approach to solving non-convex optimization problems, where the variables are partitioned into blocks. The objective function is iteratively optimized with respect to the variables in one block while treating the variables in the remaining blocks as constants [51]. In our method, we employ the relaxed OT to efficiently find an optimal repair through BCD, thereby achieving end-to-end optimal repair and coupling learning. This is accomplished by integrating the search for an optimal repair Q with the Sinkhorn algorithm, used for finding the regularized OT between P and Q. This process is described as follows:

$$\underset{\substack{Q\in\Delta(\Omega,\mathcal{F})\\ M:\mathcal{V}\times\mathcal{V}\to\mathbb{R}_{\geq 0}}}{\operatorname{argmin}} \sum_{\bar{a}\in\mathcal{V}}\sum_{\bar{b}\in\mathcal{V}} c(\bar{a},\bar{b})M(\bar{a},\bar{b}) - \frac{1}{\rho}H(M) + \lambda(KL(M_Y,Q) + KL(M_X,P)) + \mu\ \delta_\sigma(Q), \tag{4.4}$$

In this formulation, M_Q and M_P are the marginals of the coupling M with respect to the variables of Q and P, respectively. Furthermore, $\delta_\sigma(Q)$ represents the measure of dissatisfaction with the conditional independence constraint σ for Q, as explained in Chapter 3 (see Equation 3.1).

There are two crucial observations to be made about Equation 4.4. First, μ is a coefficient that modulates the influence of σ and its associated dissatisfaction; for $\mu \geq 0$, σ acts as a soft constraint, becoming a hard constraint as μ tends towards infinity. Second, Equation 4.4 deviates from the relaxed OT formulation by minimizing the sum over the coupling M and the repair probability distribution Q rather than only over M.

Given the restriction to distributions Q in $\Delta(\Omega,\mathcal{F})$ that satisfy σ, the Sinkhorn algorithm (using the update rules in Equation 3.6) cannot directly solve the optimization problem posed in Equation 4.4. To overcome this, we employ BCD rather than standard gradient descent. The variables in the optimization problem are partitioned into $d_Z + 1$ blocks: one block represents the variables for the coupling M, while the remaining d_Z blocks depict slices of variables from the distribution Q with fixed values of the random variable Z. Algorithm 2 delineates our algorithm, implementing BCD alongside the relaxed OT.

The algorithm begins with the *Initialize* procedure, which sets Q to a probability function in compliance with σ (Line 1). Multiple initial choices can satisfy the CI constraint, such as the marginal probabilities with respect to random variables in σ. We use NMF as our starting point, which experimentally proved to make faster convergence of the algorithm. A while loop follows, executing the Sinkhorn algorithm within each iteration. Within this loop, the algorithm initializes vector $\bar{v}$ and matrix K before executing the Sinkhorn algorithm (Line 3). It then computes the distance between P and the current repair candidate Q by iteratively updating $\bar{u}$ and $\bar{V}$ until convergence (Line 5). The coupling M is then computed as a joint probability distribution in Line 3.

The for loop from Lines 4 to 5 updates the repair Q via BCD. A naive BCD implementation for Q uses the gradient of W_{Rx} as computed in [21]. However, this approach tends to be slow. We accelerate the gradient computation for each block by implementing matrix factorization in the *Update* procedure (Line 5). Here, we first extract a matrix representing the conditional probability $Q_{X,Y|Z}$ when $Z = c_k$ (a slice of the conditional distribution where Z is fixed to value c_k). This matrix is then factorized into two matrices, representing $Q_{X|Z}$ and $Q_{Y|Z}$, which are

used to update the values in Q corresponding to $Z = c_k$. The matrix factorization applies the multiplicative update rule within a rapid iterative algorithm [35].

Algorithm 1: *RelaxedRepair*

Input: A probability function P, a cost function c, and CI constraint $\sigma : X \perp\!\!\!\perp Y \mid Z$
Output: An optimal repair Q of P w.r.t. σ and coupling M that generates Q

1 $Q \leftarrow Initialize(P)$; ▷ Initial guess using NMF
2 **while** Q *is not converged* **do**
3 $\quad \bar{v} \leftarrow \mathbb{1}_{d_Y}; K \leftarrow e^{-\rho C - 1}$; ▷ Sinkhorn Initialization
4 $\quad$ **while** $\bar{u}$ *and* $\bar{v}$ *are not converged* **do** ▷ Sinkhorn iterations
5 $\quad\quad \bar{u} \leftarrow (\bar{p} \oslash (K \cdot \bar{v}))^{\frac{\rho\lambda}{\rho\lambda+1}}, \bar{v} \leftarrow (Q \oslash (K \cdot \bar{u}))^{\frac{\rho\lambda}{\rho\lambda+1}}$;
6 $\quad M = diag(\bar{u}) \cdot K \cdot diag(\bar{v})$;
7 $\quad$ **for** $c_k \in \text{Dom}(Z)$ **do** ▷ BCD for updating Q
8 $\quad\quad Q \leftarrow Update(Q, M_Q, c_k)$; ▷ Updating Q using NMF
9 **return** Q, M;

Analysis of the RelaxedRepair Algorithm: Let us first address the correctness of Algorithm 2 in creating a distribution repair. Our central claim is that the algorithm generates a distribution in compliance with the ICs while minimizing the relaxed OT outlined in Equation 4.4. This implies a sub-optimal solution for the repair problem, given that the relaxed OT offers an estimation of OT and the Wasserstein distance in the repair problem. The proof for this assertion is two-pronged. Firstly, M, as defined by the equation in Line 6 of the algorithm, minimizes the relaxed OT in Equation 4.4. The substantiation for this segment echoes the proof of Proposition 4.1 in [21]. Secondly, the BCD iterations (Lines 2-8) are always convergent since Q invariably gravitates towards a stationary point. This phenomenon occurs because the algorithm iteratively solves NMF to update the marginal Q for $c_k \in \text{Dom}(Z)$ while the remainder of Q is held constant. Note that since all blocks of variables are optimized in each iteration of the while loop, our solution meets the cyclic rule, which is important for ensuring convergence. The subsequent theorem encapsulates these claims.

Theorem 4.2.1. Given a probability distribution $P \in \Delta(\mathbf{V})$ with random variables $\mathbf{V}$ and a CI constraint $\sigma : \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$ with $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$, *RelaxedRepair* (Algorithm 2) converges and returns a probability distribution $Q \in \Delta(\mathbf{V})$ that minimizes the relaxed OT as per Equation 4.4. ■

The runtime of Algorithm 2 is subject to several determinants, including the initial approximation for the distribution Q and the initialization of vectors $\bar{u}$ and $\bar{v}$ during Sinkhorn iterations. These elements dictate the iteration count necessary for algorithm convergence and final Q acquisition. The major computational cost within each iteration arises from two pivotal operations: the Sinkhorn iterations (Line 5) and the NMF computation (Line 8), each of which is executed d_Z times. Owing to optimized matrix multiplication techniques, these operations maintain efficiency, with a computational complexity approximating $O(d_V \times d_V)$. To further enhance the performance of the algorithm, we have integrated several optimization strategies.

These involve refined initialization tactics for Q, $\bar{u}$, and $\bar{v}$, alongside strategies to lessen the computational toll of large domain size (d_V) on the overall runtime. Together, these optimizations significantly bolster the overall efficiency of the algorithm.

4.2.3 Unsaturated Constraints

In the solutions offered in the previous section, we operated under the assumption that σ is a saturated CI constraint, meaning the set of all random variables $\mathbf{V}$ is the union of $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$. However, in real-world applications, especially those involving high-dimensional datasets, CI constraints can be unsaturated. For example, in our fairness application setting, the attribute Y does not form part of the CI constraint $S \perp\!\!\!\perp N \mid A$, yet it's needed for the training of a classifier using the repaired distribution. This observation prompted us to devise a method to address this situation.

Two strategies can be adopted to repair with respect to unsaturated ICs. The first, and simpler, approach is to consider the additional variables $\mathbf{W} = \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z})$ in the random variables in P and Q and find the coupling M that includes $\mathbf{W}$. However, this method is expensive, as an increase in the number of variables in $\mathbf{W}$ and their domains directly impacts the size of M and the computation cost of finding an optimal repair.

The second and more practical approach, which we utilize in our experiments, involves converting P to a marginal distribution P' containing only $\mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z}$ and finding a much smaller coupling M' that leads to a marginal distribution Q' of Q with $\mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z}$. Any Q that satisfies $Q = Q' \times Q_{W|X,Y,Z}$ will be an acceptable repair. We apply the following technique to generate Q from Q'. For every $\bar{g} \in \text{DOM}(W)$ and $\bar{a} \in \text{DOM}(V)$, we compute Q as follows:

$$Q_{\mathbf{W}|\mathbf{X},\mathbf{Y},\mathbf{Z}}(\bar{g} \mid \bar{a}) = \sum_{\bar{a}' \in \text{DOM}(V)} M(\bar{a}', \bar{a}) \times P_{\mathbf{W}|\mathbf{X},\mathbf{Y},\mathbf{Z}}(\bar{g} \mid \bar{a}') \tag{4.5}$$

Here, we set Q to be a linear combination of $P_{\mathbf{W}|\mathbf{X},\mathbf{Y},\mathbf{Z}}(\bar{g} \mid \bar{a}')$s. We employ this approach in Chapter 5 for the fairness application.

Algorithm 2: *BCD*

Input: A probability function P, a cost function c, and CI constraint $\sigma : X \perp\!\!\!\perp Y \mid Z$
Output: An optimal repair Q of P w.r.t. σ and coupling M that generates Q

```
Q ← Initialize(P);
while Q is not converged do
    M ← Sinkhorn(P, Q, c);
    for z ∈ Dom(Z) do
        Q_{X,Z}(X, z), Q_{Y|Z}(Y|z) ← NMF(M_Q(X, Y, z));
return Q, M;
```

Chapter 5

Experimental Evaluations

In the experimental chapter of this thesis, we utilize both real-world and synthetic datasets with the aim of accomplishing the following objectives:

1. To fine-tune the BCD algorithm by identifying the optimal hyperparameters (ρ and λ) for each dataset.

2. To demonstrate the efficacy of the BCD algorithm in minimizing the relaxed OT and to confirm its convergence to a repair solution.

3. To assess the quality of estimation in the relaxed OT by comparing it with the actual Wasserstein distance.

4. To establish the superiority of our repair solution over baseline algorithms in the context of debiasing in our fairness application.

5. To illustrate the application of our repairs in data error detection.

The chapter is structured as follows: We first present our experimental setup in Section 5.1, which includes an overview of the datasets used for both applications in Subsection 5.1.1, and a discussion on the baseline models in Subsection 5.1.2. Following this, we outline our experimental results in Section 5.2 and conclude with a comprehensive discussion and summary of key findings in Section 5.3.

5.1 Experimental Setup

Our methodologies were implemented using Python version 3.10.10, which is highly respected for its robust support and extensive library ecosystem for scientific computation and data analysis. Our experiments were performed on a high-performance server with 64GB of CPU and 24GB of GPU, balancing substantial computational power and efficient memory utilization.

For computations involving variations of the Wasserstein distance, we utilized the PyTorch library[1], a widely utilized open-source machine learning library for Python. PyTorch provides

[1] https://pytorch.org

an extensive set of tools and libraries designed explicitly for tasks in computer vision and natural language processing. It also offers robust support for various operations, such as matrix factorization and multiplication, which are vital for our implementation.

Alongside PyTorch, we made use of the Python Optimal Transport (POT) library[2], a comprehensive toolbox facilitating OT computations. Despite the POT library providing efficient OT implementations and being compatible with PyTorch, we encountered challenges with implementing the relaxed OT regarding memory efficiency. One such issue arose during an experiment wherein the cost matrix was expected to fit in GPU memory but did not. The multiplication of the cost matrix by $-\rho$ was not an in-place operation, leading to consuming twice the necessary memory. We refined the original implementation to overcome these issues, integrating it into our code for improved memory efficiency. This custom-optimized version of the relaxed OT played a pivotal role in successfully executing our experiments, emphasizing the importance of memory-efficient implementations for high-dimensional computational tasks.

In addition to DP and EO, widely accepted group fairness measures, we also employed a measure known as Ratio of Observation Discrimination (ROD), introduced in [48]. ROD is part of the interventional fairness family and is defined for scenarios where covariates are divided into sensitive, inadmissible, and admissible attributes akin to our setting. By intervening on admissible attributes, all incoming edges to them are removed in the causal graph. Thus, there will be no path from sensitive to label passing from admissible attributes. Then, it will check whether any path is from the sensitive attribute to the label attribute. It is formally defined as

$$ROD(S, \hat{Y} \mid A) = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \frac{P(\hat{Y} = 1 \mid S = 0, a) P(\hat{Y} = 0 \mid S = 1, a)}{P(\hat{Y} = 0 \mid S = 0, a) P(\hat{Y} = 1 \mid S = 1, a)}.$$

In this equation, $\hat{Y}$ is the value predicted by the classifier for the label. ROD effectively captures the discriminatory impact of the sensitive attribute on the label, gauging how closely $P(\hat{Y} = 1 \mid S = 0, a)$ and $P(\hat{Y} = 1 \mid S = 1, a)$ align. These probabilities are expected to be identical in conditions that satisfy full fairness. Conditioned on admissible attributes, it only encapsulates the direct and indirect discriminatory impacts via inadmissible attributes. Consequently, when considering samples with identical admissible attribute values, the probability of a successful outcome should be equal for both privileged and unprivileged groups. When ROD is equal to one, the CI constraint $\hat{Y} \perp\!\!\!\perp S \mid A$ is satisfied [48]. This observation establishes a direct link between CI constraints and fairness measures. It suggests that fairness can be achieved by debiasing in accordance with a CI constraint. The specific constraint to enforce is contingent upon the fairness measure we intend to satisfy. For instance, the constraints corresponding to DP and EO are $\hat{Y} \perp\!\!\!\perp S$ and $\hat{Y} \perp\!\!\!\perp S \mid Y$, respectively.

5.1.1 Datasets

Understanding the statistics and characteristics of the datasets employed in an experiment is crucial. It clarifies the rationale behind their selection for the specific experiment in question. In this context, the following datasets, which have been used in this study, are introduced in detail:

[2] https://pythonot.github.io

UCI Adult Dataset The Adult dataset enjoys popularity in fairness literature. [3] This dataset comprises 32,561 records with 15 attributes. We have utilized seven for our experiments: gender, income, marital status, age, education-num, hours-per-week, and occupation. The dataset has been preprocessed similarly to the approach in Calmon et al. [10] to make the domain discrete and also reducing domain size; age is quantized into decades, and ages above 70 are capped at 70. Education-num values below five and above 13 are excluded, and the value of hours-per-week is replaced with the nearest lower multiple of 10, reducing domain size from 23,497 to 9,281. The level of discretization of attributes in addition to specifying the semantic of constraint, controls the information loss as well.

Each record in this dataset delineates various attributes of an individual, with the binary income attribute indicating if their income exceeds 50k per year. This has been chosen as the label attribute. Extensive research indicates a dependency between gender (the protected attribute) and income, which introduces a bias in machine learning models trained on this dataset. The remaining attributes are divided into two groups: Admissibles and Inadmissibles. Admissible attributes are those through which the protected attribute can indirectly influence the label. Conversely, any indirect impact of the protected attribute through inadmissible attributes is deemed discriminatory and must be removed. In this context, marital status is classified as inadmissible, while age, education-num, hours per week, and occupation are treated as admissible attributes. The selection of admissible and inadmissible attributes mirrors the work of Salimi et al. [48], who first proposed this fairness application setting. Our focus is solely on demonstrating the performance of our methods under a similar setting, so we did not delve into the rationale behind their choice.

Despite the wide usage of this dataset in the literature, it has some limitations. The number of records is not enough for presenting its large domain size, and as a result, discretization is needed to reduce it, which causes information loss. There is an imbalance in sensitive and label attributes; around 75.9% and 66.9 % of records refer to individuals with less than 50K income and males, respectively. In addition, the choice of 50K as a threshold for classifying incomes is debatable and might not be generalized to current incomes.

Synthetic Datasets: The synthetic dataset crafted for the fairness application emulates a similar graphical model as depicted in Figure 1.1. However, it omits the direct impact of S on Y and construes N as a function of A. Additionally, a hidden attribute Z is considered, which impacts both N and Y. It is not considered post-data generation. This endows N with prediction power independent of S, thereby rendering it non-discriminatory. All attributes are binary, and the dataset contains 5,000 samples.

The synthetic dataset curated for error detection applications comprises just two main attributes, with domain sizes of 25 and 32. The first attribute (X) is sampled from a uniform distribution (ranging from 1 to 20), with the addition of a normally distributed noise (standard deviation of 1). The second attribute (Y) is formed by the summation of a Poisson distribution (parameter λ=10) and a normally distributed noise (standard deviation of 3). The CMI of the generated dataset, which consists of 10,000 samples, is zero, making it entirely suitable for error detection applications.

[3] https://archive.ics.uci.edu/dataset/2/adult

5.1.2 Baselines

For the evaluation of our proposed approach, comparisons with established baseline methods are necessary. While our fairness application setting is inspired by Salimi et al., [48], a direct comparison with their method is not viable. This is due to the label attribute not being included in the CI constraint and thus not being present in the dataset returned by their approach. Instead, we compare our method with two other baseline approaches, detailed below.

NMF: The output of this method forms the initial input for our proposed method. Given the CI constraint $\sigma : X \perp\!\!\!\perp Y \mid Z$, this method factorizes the joint probability distribution of X and Y for all possible values of Z. For each value of $c \in \text{DOM}(Z)$, NMF factorizes $P_{X,Y|Z}(X, Y, Z = c)$ by minimizing the KL divergence between the original and repaired distribution using multiplicative update solver [20, 35]. Despite its advantage in terms of runtime efficiency, this method has a key limitation: it cannot perform the repair by altering the value of Z, as the probability distribution of Z is fixed.

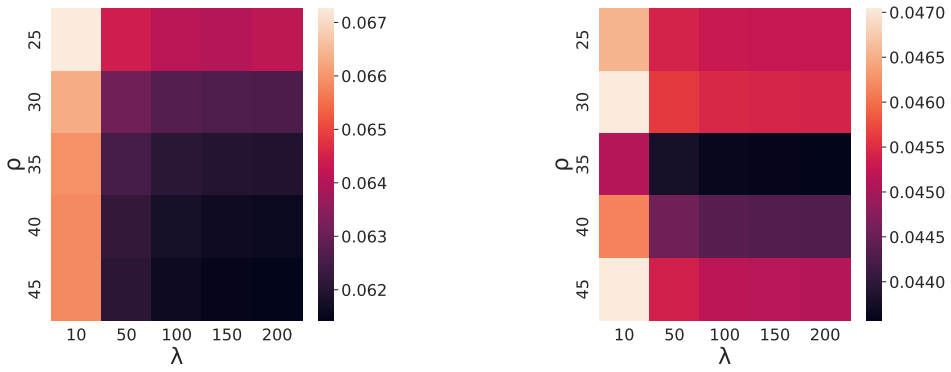
Dropped: As discussed earlier, a protected attribute can exert indirect discriminatory influence on the label through inadmissible attributes in addition to its direct impact. The direct impact can be mitigated by removing the protected attribute from the training features given to the classifier. The 'Dropped' approach is a simple method that removes the inadmissible attributes, leaving only the admissible ones for classifier training. While this can deliver optimal results regarding fairness measures, since all sources of discriminatory impact are eliminated, it compromises accuracy as the predictive power of the inadmissible attributes is disregarded.

5.2 Experiments

Our experimental setup is segmented into four parts. The first two parts are aimed at evaluating the efficacy of our proposed approach in a general context, while the latter two specifically assess our method's performance in two distinct applications.

5.2.1 Tuning Hyper-parameters

The effectiveness of our solution in minimizing the Wasserstein distance largely depends on two hyper-parameters, namely λ and ρ. The former dictates the proximity of the source and target of the mapping to the original and repaired distributions, while the latter restricts the entropy of the mapping. We employed a grid search strategy to determine the optimal values for these hyper-parameters. For each value attempted during the grid search, we reported the $W(P, Q)$, and plotted the results in a heatmap, which is illustrated in Figure 5.1. The enforced constraint is analogous to that of the fairness application (protected attributes are independent of inadmissible attributes given the admissible ones), and the convergence threshold for both datasets is set at 10^{-7}. The cost function permits alterations only to the values of the inadmissible attribute, a choice which will be justified in Section 5.2.3. (Note: For this section, we only used education-num from the admissible attributes in the adult dataset to accelerate the runtime.)



a) UCI Adult dataset b) The synthetic dataset

Figure 5.1: The minimum normalized distance for the hyper-parameters ρ and λ

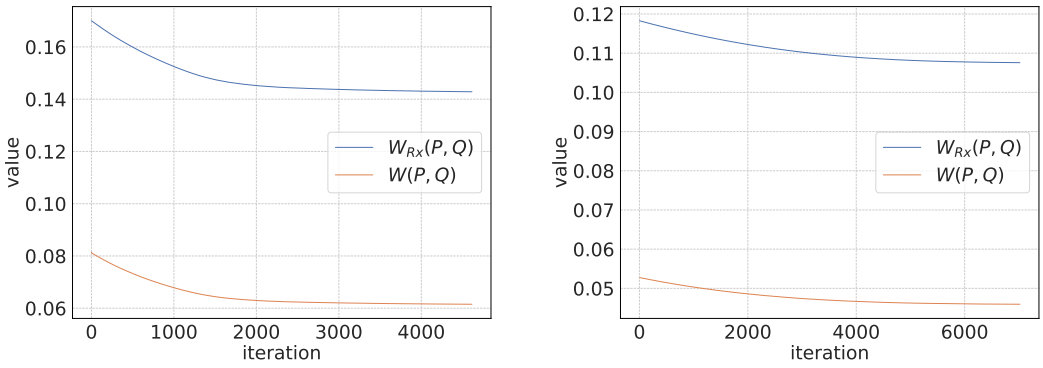
As can be observed from the results, for each ρ value, the performance improves with an increase in λ. This improvement is due to the fact that assigning a higher weight to the two KL divergence terms in the objective function brings the relaxed Wasserstein distance ($W_{Rx}(P, Q)$) closer to the entropic Wasserstein distance ($W_\rho(P, Q)$), leading to a more accurate estimation. However, this enhanced outcome comes with the trade-off of increased iterations required for convergence. For instance, in the case of the adult dataset, with $\rho = 45$, an increase in the λ value from 10 to 200 results in a monotonic rise in the iteration count, from approximately 400 to around 4,500. Given this trade-off between the number of iterations and the answer quality achieved, we set $\lambda = 200$ for both datasets.

As a general rule, an increase in the ρ value causes $W_{Rx}(P, Q)$ to draw closer to $W(P, Q)$, as the weight of the entropy term decreases. However, a maximum ρ value exists, beyond which the Sinkhorn algorithm fails due to the expanded search space. For the experiments conducted here, this maximum value is 50. The optimal values of ρ that yield the best performance are 45 and 35 for the Adult and synthetic datasets, respectively.

To further clarify, let's consider λ. As this parameter controls the weight of the KL divergence terms in the objective function, a higher λ results in a solution that aligns more closely with the entropic Wasserstein distance. However, it also leads to an increased number of iterations, resulting in higher computational costs. On the other hand, the ρ parameter, which controls the weight of the entropy term in the entropic Wasserstein distance, affects the accuracy of the approximation. As ρ increases, the entropy term's weight decreases, bringing the relaxed and true Wasserstein distances closer together. Nevertheless, the algorithm struggles beyond a certain threshold (50 in this case) due to the enlarged search space. Therefore, the best performance is achieved by finding the optimal balance between accuracy and computational complexity, which, in these experiments, resulted in ρ values of 45 for the Adult dataset and 35 for the synthetic dataset.

5.2.2 Minimizing OT and the Wasserstein Distance

After determining suitable hyper-parameter values in the previous section, we can now evaluate the effectiveness of our solution from two perspectives. Firstly, the solution should be



a) UCI Adult dataset

b) The synthetic dataset

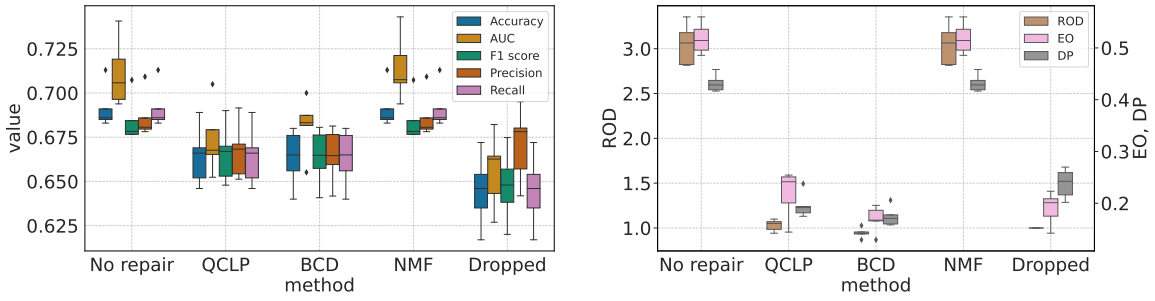
Figure 5.2: Effectiveness of our solution in minimizing Wasserstein distance

capable of reducing the value of the objective function with each iteration, i.e., the objective function should be a monotonically decreasing function of the iteration number. Secondly, since the primary aim of our method was to minimize the Wasserstein distance, we need to demonstrate that by minimizing the objective function, we are indeed reducing the Wasserstein distance of the repair in every iteration. To this end, we plotted the values of the objective function ($W_{Rx}(P,Q)$) and the actual Wasserstein distance between the initial (P) and repaired (Q) distributions for all iterations in Figure 5.2.

As illustrated in both Figures 5.2a and 5.2b, W_{Rx} is a strictly decreasing function of the iteration number, indicating that our solution can successfully solve our optimization problem. Another key observation from these figures is the distinct relationship between minimizing W_{Rx} and W. This suggests that the objective function we selected aligns well with the purpose of our method. By comparing Figures 5.2a and 5.2b, we notice that the reduction in the Wasserstein distance in the synthetic dataset is not as significant as in the Adult dataset. This discrepancy is due to the difference in the domain size of the inadmissible attribute in the two datasets, which is 7 and 2 for the Adult and synthetic datasets, respectively. With the larger dimension in the Adult dataset, our method has a more extensive search space and greater freedom to diverge from the starting point. If we permit the alteration of other attributes, we would observe a drastic reduction in the Wasserstein distance with our method.

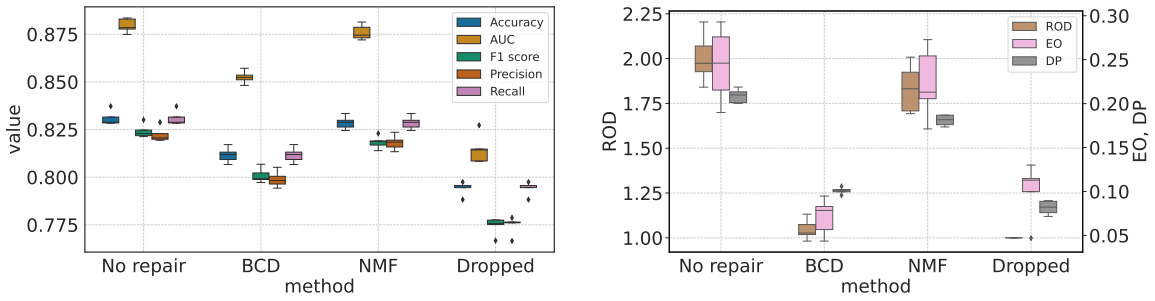
5.2.3 Fairness Application

The following presents the experimental results of our proposed method in a fairness application setting. In this scenario, we start with a given dataset. It is divided into training and test sets using 5-fold cross-validation. For each pairing of training and test sets, we apply our repair process to the training data, after which a logistic regression model is trained using the repaired distribution. We report a variety of metrics based on the testing of the trained model with the test data: accuracy, AUC, F1 score, precision, recall, DP, EO, and ROD. These results are compared with those of a model trained on the original, unrepaired training data. Additionally, we contrast our results with two baseline approaches: NMF and Dropped. It's important to note that our LP method can only be applied to the synthetic dataset due to its size. The



a) Accuracy measures b) Fairness measures

Figure 5.3: Performance of logistic regression classifier when trained on training data repaired by each method (Synthetic dataset)



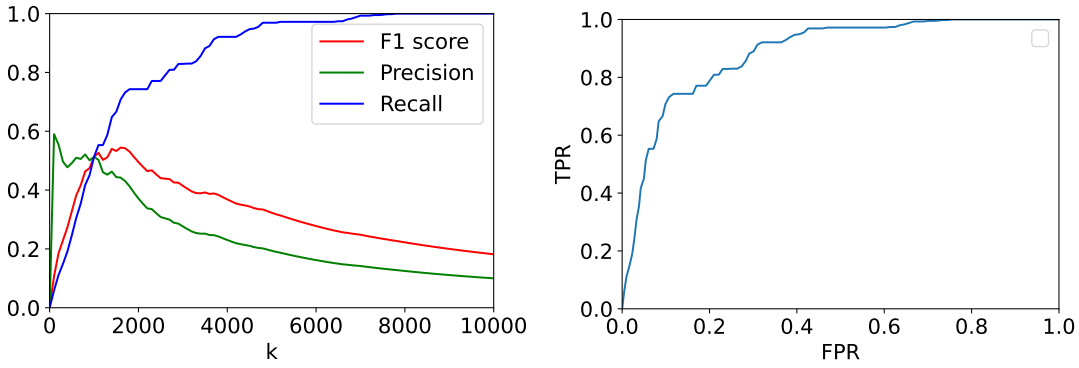
a) Accuracy measures b) Fairness measures

Figure 5.4: Performance of logistic regression classifier when trained on training data repaired by each method (UCI Adult dataset)

Adult dataset's larger domain size makes the LP method inapplicable.

As outlined in the introduction, the CI constraint enforced here is $S \perp\!\!\!\perp N \mid A$ to remove the unwanted influence of S on Y through N. The direct impact of S on Y is eliminated by not including the protected attribute in the classifier. The hyper-parameter values utilized in this section are the same as those in Section 5.2.2. In our methods (LP and BCD), we preprocess test data before feeding it to the classifier. (The original value of the test data will be used to report fairness measures.) We adjust each sample according to the coupling returned by our methods. For instance, if in the returned coupling, the value of (s, n, a) is changed to (s, n', a) with a probability of 25%, this modification will be applied to the test sample in the same way. The results for the Adult and synthetic datasets are provided in Figures 5.3 and 5.4.

Both figures demonstrate that our methods can achieve a level of fairness similar to the Dropped approach but with higher accuracy measures. This outcome underscores the value of applying a repair rather than simply dropping sources of discrimination. By retaining the inadmissible attribute, we utilize its predictive power, which is non-discriminatory. For the synthetic data, where the LP method is applicable, we observe that the accuracy measures of LP and BCD are comparable, making it challenging to determine which one outperforms the other. This comparison suggests that BCD can achieve results almost equivalent to LP and can



a) change of performance by varying k

b) ROC curve

Figure 5.5: Effectiveness of our method for detecting sorting error

also be applied when repairing datasets with large domain sizes.

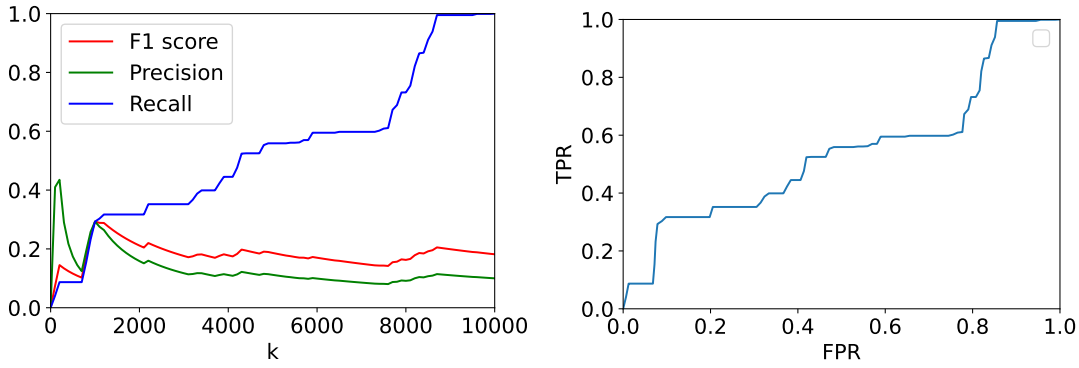
On the other hand, NMF fails to significantly alter the fairness measures, mostly due to its inability to preprocess test data. When we repair the training data, we apply a covariate shift to the data distribution, and the classifier learns this shifted distribution. Hence, when making predictions on test data, we must consider the difference between the training and test data distributions. However, since NMF doesn't generate a coupling, we can't apply the same preprocessing approach as we did for our methods. As NMF relies on discriminatory information for making predictions, it can maintain higher accuracy measures than our methods.

Upon examining the fairness measures more closely, we find that BCD, LP, and Dropped can nearly reduce the ROD value to its optimal value, which is one. However, their DP and EO values are not very close to zero. To understand this, we must remember the distinction between the definition of ROD and that of DP and EO. In the case of ROD, the influence of the protected attribute on the label through admissible is not considered discriminatory; thus, accurately capturing the unwanted impact that we aim to eliminate. However, DP and EO capture any impact of the protected attribute on the label. Thus, it's clear that we cannot bring these measures to zero as we don't intend to remove the indirect impact of the protected attribute through admissible attributes.

5.2.4 Application in Error Detection and Data Cleaning

Data cleaning, a broad area of research, primarily involves two steps: the detection of data errors and the subsequent rectification of these errors within the dataset. The first step identifies records that violate a constraint, while the second replaces the current values of "dirty" records with values that closely resemble the original, clean data. The selection of a constraint can range from any database integrity constraints, such as Functional Dependencies (FDs) and Multi-Valued Dependencies (MVDs), to uniquely designed constraints for outlier detection. In this section, we discuss the application of our method in the context of error detection.

The use of CI constraints for error detection is a novel area, first discussed in [57]. Intriguingly, it has been shown that there exists a relationship between independence and integrity constraints, indicating that repairing based on one type of constraint is akin to applying repair



a) change of performance by varying k

b) ROC curve

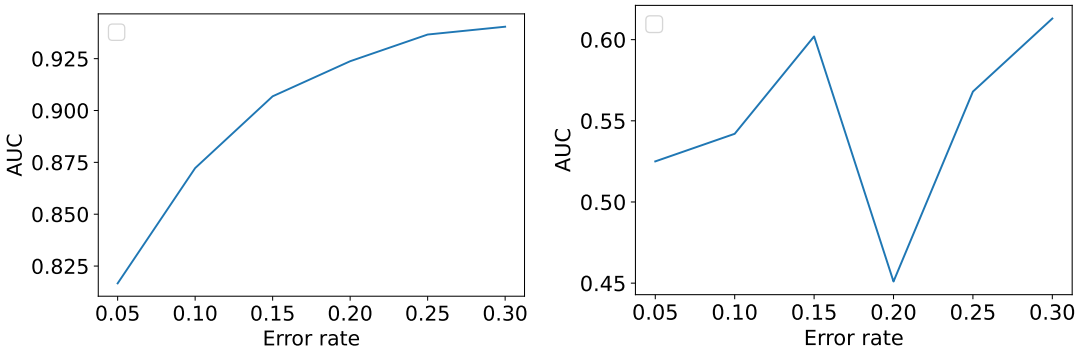
Figure 5.6: Effectiveness of our method for detecting imputation error

for both constraints. Our algorithms return a coupling and a repaired distribution, but we need a systematic approach to assign a score to records, indicating their likelihood of being dirty. This score, defined over the domain values $\text{DOM}(V)$, is calculated as the mass transported from one point to the other points. The higher the score, the more likely a value requires repair.

Our experiments use a synthetic dataset, carefully constructed to satisfy almost all constraints. Then, we randomly select $\alpha\%$ of records and add noise to one of their columns. Our algorithm is subsequently applied, and the generated coupling is used to detect records requiring more repair. We test the detection of two simulated error types: sorting error and imputation error. Sorting error is introduced by organizing the values in one column relative to values in another, making them more dependent. This kind of error is inspired by the KDD-Cup 2008, where a team exploited the dependency between patient ID and target label in their prediction model, eventually winning the competition [43]. Imputation error, on the other hand, is added when enforcing the constraint $\sigma : X \perp\!\!\!\perp Y$ by replacing the values of column X with the mean of that column in the selected records [57].

To enhance results, domain knowledge is leveraged to specify the cost matrix. Assuming that the error has been added to only one column and knowing that dirty column, we set the cost of altering the value of another column exceedingly high. This strategy ensures that mass only moves between points differing in the value of the dirty column in the returned coupling. Performance evaluation is done by reporting AUC. By calculating a score for "dirtiness" and trialing different thresholds, the number of returned records (k) can be adjusted. As we vary k, the False Positive Rate (FPR) and True Positive Rate (TPR) also change, enabling us to plot the ROC curve and calculate the AUC by computing the area under this curve. The change in f1 score, precision, and recall for different k values is also examined.

Sorting and imputation errors were added to 10% of records in the first two experiments of this section, and the results are depicted in Figures 5.5 and 5.6. The AUC when detecting sorting and imputation errors was 0.884 and 0.542, respectively, indicating a significant difference. This is due to the increase in CMI from zero to 0.035 when adding the sorting error, whereas the change in CMI when introducing imputation error is negligible. As a result, the contribution of imputed records in violating the independence constraint is akin to other records, causing our method to fail in detecting them.



a) sorting error

b) imputation error

Figure 5.7: trend of changing AUC when increasing error rate

In the final part of this section, we plot the AUC change when increasing the error rate, as seen in Figure 5.7. A comparison of Figures 5.7a and 5.7b reveals that with sorting error, AUC increases with the error rate, while with imputation error, there is no clear relationship. The difference lies in the relationship between CMI and error rate in these cases. In the case of sorting error, CMI increases from 0.01 to 0.32 as the error rate rises from 5 to 30 percent. But for imputation error, CMI remains close to zero. Hence, we can conclude that when CMI is high post-error addition, our coupling is better at distinguishing between clean and dirty records.

5.3 Discussion and Takeaways

Our research involves extensive experimentation in both the domains of data fairness and data cleaning, offering interesting insights into the efficacy of our proposed method in both areas.

The experiments related to data fairness have demonstrated our methods' ability to achieve fairness levels comparable to the baseline model, which simply drops the discriminatory sources while maintaining better accuracy. This outcome highlights the benefit of repairing the dataset over completely eliminating potential sources of discrimination. By retaining the predictive power of inadmissible attributes that aren't intrinsically discriminatory, our methods provide an effective means of achieving fair data handling. It was further observed that both our methods, LP and BCD, performed similarly when applied to synthetic datasets. This result indicates that BCD is capable of delivering comparable results to LP, with the added advantage of being applicable in scenarios involving datasets with large domain sizes. However, the NMF method did not significantly alter the fairness measures, which was mainly due to its inability to preprocess test data. Analyzing the fairness measures, it was evident that while BCD, LP, and Dropped methods reduced the value of ROD effectively, their DP and EO values were not as close to zero. This discrepancy reminds us of the difference in definitions of ROD, DP, and EO, and reinforces the challenge of achieving absolute fairness in data handling.

In terms of data cleaning, our method successfully demonstrated its utility in detecting and rectifying both sorting and imputation errors in datasets. The versatility of our method to rectify data based on one type of constraint and extend its capability to other types of constraints

adds to its robustness. Nevertheless, the effectiveness of our approach is highly contingent on the type of errors present in the dataset, as observed in the experiment involving sorting and imputation errors. The necessity of domain knowledge and the strategic selection of cost matrices was underscored during our experimentation. The ability to understand the error structure in the dataset before applying our method is critical for its successful implementation. Finally, our experiments indicate that our method excels in scenarios with a high CMI post-error addition. However, its efficiency drops in cases where the CMI remains low, such as in the case of imputation errors.

In conclusion, our methods provide a promising approach to address fairness and data cleaning challenges. Yet, they require careful application and further refinement to enhance their universal effectiveness in different data scenarios.

Chapter 6

Conclusion and Future Work

This chapter aims to review the challenges solved in the thesis and summarize all of the discussions we have had so far, and also highlight potential areas for future research in the domain of repairing with respect to CI constraints. A significant amount of work has been conducted on testing these constraints, but their utility in repairs has been less explored. We emphasized the choice of an appropriate distance function and proposed the use of the Wasserstein distance, which we believe offers improved results.

At first, we came up with a linear program solution that could achieve the objective but had bad memory efficiency. This moved us to our second approach, which utilized relaxed Wasserstein distance and BCD for solving the problem. The efficacy of our proposed methods has been examined in two applications in the preceding chapter. We showed we can outperform both baselines in the fairness application. We also showed promising performance in the error detection task but couldn't compare our results with the existing method as we didn't have access to their code. Now, we wish to outline some potential future research directions:

- *Cost matrix design for Wasserstein distance:* The selection of the cost matrix is a challenging but vital aspect of our method, as it incorporates domain knowledge about our data. Although we manually chose suitable cost functions for our data cleaning and fairness applications, there's no guarantee of optimality. Future work could focus on developing systematic approaches for cost matrix design.

- *Handling labels in CI constraints:* In the context of fairness applications, our current approach does not allow for the preprocessing of test data if labels are included within the CI constraint attributes. A naive solution involves the marginalization of the coupling, however, this is not accurate. Future research could focus on modifying our method to learn a coupling that only comprises a subset of all involved attributes.

- *Multiple CI constraints:* Our current method is limited to supporting a single CI constraint. Extending the method to enforce multiple constraints simultaneously, as opposed to sequentially, presents an exciting avenue for future research.

- *Approximate CI enforcement:* We currently enforce full CI satisfaction, potentially leading to considerable information loss. An interesting extension of this work could involve exploring the trade-off between the level of constraint violation and information loss by enforcing approximate CI instead of exact constraints.

- *Continuous random variables:* The current approach is designed to repair discrete distributions, premised on the assumption that each column in the dataset represents a discrete random variable. Future work could look at the repair of continuous distributions, necessitating a wholly different approach.

- *Application in other fields:* The evaluation of our method has been limited to applications in fairness and data error detection. However, CI constraints have a broad range of uses in areas like bioinformatics, statistics, and genetics. Future research could extend the evaluation of our method's effectiveness to these domains.

Bibliography

[1] Ardavan Afshar, Kejing Yin, Sherry Yan, Cheng Qian, Joyce C Ho, Haesun Park, and Jimeng Sun. Swift: Scalable wasserstein factorization for sparse nonnegative tensors. In *Proceedings of the AAAI Conference*, 2021.

[2] Marcelo Arenas, Leopoldo Bertossi, and Jan Chomicki. Consistent query answers in inconsistent databases. In *Proceedings of the eighteenth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 68–79, 1999.

[3] Marcelo Arenas, Leopoldo Bertossi, and Jan Chomicki. Answer sets for consistent query answering in inconsistent databases. *Theory and practice of logic programming*, 3(4-5):393–424, 2003.

[4] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[5] Leopoldo Bertossi. *Database Repairs and Consistent Query Answering*. Morgan & Claypool Publishers, 2011.

[6] Nicolas Bonneel, Michiel Van De Panne, Sylvain Paris, and Wolfgang Heidrich. Displacement interpolation using lagrangian mass transport. In *Proceedings of the 2011 SIGGRAPH Asia conference*, pages 1–12, 2011.

[7] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[8] Jean-Philippe Brunet, Pablo Tamayo, Todd R Golub, and Jill P Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the national academy of sciences*, 101(12):4164–4169, 2004.

[9] Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data mining and knowledge discovery*, 21:277–292, 2010.

[10] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pages 3992–4001, 2017.

[11] Fei Chiang and Renée J Miller. A unified model for data and constraint repair. In *2011 IEEE 27th International Conference on Data Engineering*, pages 446–457. IEEE, 2011.

[12] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806, 2017.

[13] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. *Advances in neural information processing systems*, 30, 2017.

[14] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[15] Hal Daumé III. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, 2007.

[16] Gary Doran, Krikamol Muandet, Kun Zhang, and Bernhard Schölkopf. A permutation-based kernel conditional independence test. In *UAI*, pages 132–141, 2014.

[17] Ronald Fagin. Multivalued dependencies and a new normal form for relational databases. *ACM Transactions on Database Systems (TODS)*, 2(3):262–278, 1977.

[18] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

[19] Sira Ferradans, Nicolas Papadakis, Gabriel Peyré, and Jean-François Aujol. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882, 2014.

[20] Cédric Févotte and Jérôme Idier. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural computation*, 23(9):2421–2456, 2011.

[21] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a wasserstein loss. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[22] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. Fairness testing: testing software for discrimination. In *Proceedings of the 2017 11th Joint meeting on foundations of software engineering*, pages 498–510, 2017.

[23] Yuan Gao and George Church. Improving molecular cancer class discovery through sparse non-negative matrix factorization. *Bioinformatics*, 21(21):3970–3975, 2005.

[24] Floris Geerts, Giansalvatore Mecca, Paolo Papotti, and Donatello Santoro. The llunatic data-cleaning framework. *Proceedings of the VLDB Endowment*, 6(9):625–636, 2013.

[25] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS symposium on machine learning and the law*, volume 1, page 11. Barcelona, Spain, 2016.

[26] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[27] Hoda Heidari and Andreas Krause. Preventing disparate treatment in sequential decision making. In *IJCAI*, pages 2248–2254, 2018.

[28] Amit V Khera and Sekar Kathiresan. Genetics of coronary artery disease: discovery, biology and clinical translation. *Nature Reviews Genetics*, 18(6):331–344, 2017.

[29] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. *Advances in neural information processing systems*, 30, 2017.

[30] Sunyoung Kim and Masakazu Kojima. Exact solutions of some nonconvex quadratic optimization problems via sdp and socp relaxations. *Computational optimization and applications*, 26:143–154, 2003.

[31] Masaki Kimizuka, Sunyoung Kim, and Makoto Yamashita. Solving pooling problems with time discretization by lp and socp relaxations and rescheduling methods. *Journal of Global Optimization*, 75:631–654, 2019.

[32] Philip A Knight. The sinkhorn–knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008.

[33] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[34] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

[35] Daniel Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13, 2000.

[36] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[37] Maurizio Lenzerini. Data integration: A theoretical perspective. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 233–246, 2002.

[38] Mohammad Mahdavi, Ziawasch Abedjan, Raul Castro Fernandez, Samuel Madden, Mourad Ouzzani, Michael Stonebraker, and Nan Tang. Raha: A configuration-free error detection system. In *Proceedings of the 2019 International Conference on Management of Data*, pages 865–882, 2019.

[39] Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[40] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[41] Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19, 2000.

[42] Theodoros Rekatsinas, Xu Chu, Ihab F Ilyas, and Christopher Ré. Holoclean: Holistic data repairs with probabilistic inference. *arXiv preprint arXiv:1702.00820*, 2017.

[43] Saharon Rosset, Claudia Perlich, Grzergorz Świrszcz, Prem Melville, and Yan Liu. Medical data mining: insights from winning two competitions. *Data Mining and Knowledge Discovery*, 20:439–468, 2010.

[44] Donald B Rubin. Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, 81(396):961–962, 1986.

[45] Donald B Rubin. Comment: The design and analysis of gold standard randomized experiments. *Journal of the American Statistical Association*, 103(484):1350–1353, 2008.

[46] Donald Bruce Rubin. *The use of matched sampling and regression adjustment in observational studies*. PhD thesis, Harvard University, 1971.

[47] Jakob Runge. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In *International Conference on Artificial Intelligence and Statistics*, pages 938–947. PMLR, 2018.

[48] Babak Salimi, Bill Howe, and Dan Suciu. Database repair meets algorithmic fairness. *ACM SIGMOD Record*, 49(1):34–41, 2020.

[49] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. Interventional fairness: Causal database repair for algorithmic fairness. In *Proceedings of the 2019 International Conference on Management of Data*, pages 793–810. ACM, 2019.

[50] Rajat Sen, Ananda Theertha Suresh, Karthikeyan Shanmugam, Alexandros G Dimakis, and Sanjay Shakkottai. Model-powered conditional independence test. *Advances in neural information processing systems*, 30, 2017.

[51] Paul Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109:475–494, 2001.

[52] C Van de Panne. Programming with a quadratic constraint. *Management Science*, 12(11):798–815, 1966.

[53] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

[54] SK Michael Wong, Cory J Butz, and Dan Wu. On the implication problem for probabilistic conditional independency. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 30(6):785–805, 2000.

[55] Wei Xu, Xin Liu, and Yihong Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 267–273, 2003.

[56] Mohamed Yakout, Ahmed K Elmagarmid, Jennifer Neville, Mourad Ouzzani, and Ihab F Ilyas. Guided data repair. *arXiv preprint arXiv:1103.3103*, 2011.

[57] Jing Nathan Yan, Oliver Schulte, MoHan Zhang, Jiannan Wang, and Reynold Cheng. SCODED: Statistical Constraint Oriented Data Error Detection. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pages 845–860, 2020.

[58] Zhihong Zhu, Zhili Zheng, Futao Zhang, Yang Wu, Maciej Trzaskowski, Robert Maier, Matthew R Robinson, John J McGrath, Peter M Visscher, Naomi R Wray, et al. Causal associations between risk factors and common diseases inferred from gwas summary data. *Nature communications*, 9(1):1–12, 2018.

Curriculum Vitae

Name:	Alireza Pirhadi
Post-Secondary Education and Degrees:	Amirkabir University of Technology Tehran 2017 - 2021 B.Sc. University of Western Ontario London, ON 2022 - 2023 M.Sc.
Related Work Experience:	Teaching Assistant The University of Western Ontario 2022 - 2023 Teaching Assistant Amirkabir University of Technology 2019 - 2020