Fairness–Accuracy Trade-offs in Machine Learning: γ-Controlled LEACE and Pareto Front Analysis

Hyungho (Daniel) Na

Student ID: 12918652

Supervisor: Floris Holstege

June 26, 2025

Abstract

Fairness-accuracy trade-offs have become a central concern in machine learning, as sensitive attributes (e.g., gender, nationality) may introduce biases into model predictions and affect the performance and reliability of the model. In this paper, we propose γ-controlled LEACE, an extension of LEACE method by Belrose et al. (2023). Our method introduces a covariance ratio γ to control the strength of projection and fairness constraints. By adjusting the parameter, we can analyze how predictive accuracy changes over varying levels of fairness. We conduct experiments on synthetic datasets to study the theoretical fairness-accuracy trade-off on γ-controlled LEACE approach and identify Pareto optimal points to offer practical guidelines for selecting appropriate projection strength.

1 Statement of Originality

This document is written by Student Hyungho (Daniel) Na who declares to take full responsibility for the contents of this document. I declare that the text and the work presented in this document are original and that no sources other than those mentioned in the text and its references have been used in creating it. I have not used generative AI (such as ChatGPT) to generate or rewrite text. UvA Economics and Business is responsible solely for the supervision of completion of the work and submission, not for the contents.

2 Introduction

Machine Learning (ML) refers to computational systems that have the ability to learn from experience and improve their performance without being explicitly programmed. It has emerged as a core technology leading the fourth industrial revolution in a wide range of fields (Sarker (2021)).

Machine learning is increasingly replacing human in automated decision-making processes including credit scoring and job hiring. The growing influence and complexity of ML algorithms used in high-stakes decision-making have raised concerns about how they process the data and identify patterns (Liu and Vicente (2022)). More specifically, because ML algorithms are not entirely objective and may reflect human biases, the issue of **algorithmic fairness** has attracted increasing academic and public attention (Verma and Rubin (2018)). In algorithmic contexts, an ML predictor is generally considered fair if it treats individuals equally regardless of sensitive attributes such as gender, race, or nationality (Liu and Vicente (2022)).

Thus, in recent years, researchers have proposed several projection-based techniques to mitigate the influence of sensitive attributes from the machine learning process in order to improve algorithmic fairness. A prominent example is the **LEAst Squares Concept Erasure** (LEACE) method proposed by Belrose et al. (2023), which erases correlations between the input data and sensitive information by projecting the input onto a orthogonal subspace of correlation between the input and the sensitive attribute. While this method enhances fairness, it has been observed to degrade in

predictive performance, as the linearly available information of the sensitive attribute is removed after the projection (Belrose et al. (2023)). Prediction accuracy here refers to the model's ability to perform the main classification task of the target variable after the removal of information through projection (Belrose et al. (2023)). Motivated by this observation, this research investigates the trade-off between fairness and accuracy in the LEACE method by varying the strength of projection. The central research question is: "How does the varying level of strength of projection in LEACE affect the predictive performance of a machine learning model, and what does the resulting Pareto front look like?"

This research conducts a Pareto-front analysis of the LEACE projection to empirically examine the trade-off between fairness and prediction accuracy. By preserving the core structure of LEACE and adjusting the level of covariance between the projected input and the sensitive attribute, it is possible to derive a set of Pareto-optimal solutions. This allows us to trace the shape of the Pareto front to characterize the nature of the trade-off between fairness and accuracy in the LEACE method. Furthermore, we analyze how this trade-off behaves across datasets with different data characteristics, particularly in terms of correlation between the target variable and the sensitive attribute. The following section outlines the relevant projection-based methods, with a particular focus on the LEACE method, to establish the baseline for our study.

3 Literature Review

Recent breakthroughs in Artificial Intelligence (AI) and Machine Learning (ML) have led to the use of algorithms in several human decision-making processes (Verma and Rubin (2018)). This includes decisions such as whether a person gets a loan or not (Olson (2011)) or who will be hired (Miller (2015)). Researchers have increasingly focused on the concept of 'fairness' in ML algorithms (Verma and Rubin (2018)) since the algorithm may discriminate against people with sensitive attributes (Liu and Vicente (2022)). For example, according to a Reuters report, an AI recruitment algorithm used by Amazon penalized female applicants, failing to rate them in a gender-neutral manner. This bias arose because the ML algorithm was trained on historical data, primarily

consisting of CVs submitted by male applicants (BBC 2018).

While there is yet no clear agreement on the definition of fairness in ML algorithms across different contexts, Verma and Rubin (2018) provide a comprehensive overview of the definitions in classification tasks. One of the simplest and most intuitive definitions is demographic parity (Verma and Rubin (2018)), also known as statistical parity, which is widely used in research, including the study by Belrose et al. (2023). Belrose et al. (2023) define statistical parity as the condition where the input $X \in \mathbb{R}^d$ exhibits statistical parity with respect to sensitive attribute Z a random vector taking values in $\{0,1\}^k$, meaning that for all $z \in Z$ and function f with domain $\mathbb{R}^d$, $\mathbb{E}[f(X) \mid Z = z] = \mathbb{E}[f(X)]$. This definition focuses only on the first moment condition, thus less restrictive than the classical definition by Verma and Rubin (2018), which requires the full distribution of input X to be independent of attribute Z.

To achieve fairness in ML algorithms, several methods aim to reduce bias directly through projection in the embedding space—the vector-based representation of input data. For instance, Bolukbasi et al. (2016) propose removing gender bias from word vector embeddings by identifying a specific "gender direction" (such as the difference between the vectors for the words "he" and "she"). By projecting all word vectors onto an orthogonal subspace of this identified direction, gender-related information can be effectively removed. However, Gonen and Goldberg (2019) point out that the systematic gender bias remains after the projection, as gendered words tend to cluster together. To address this, Ravfogel et al. (2020) introduce the Iterative Nullspace Projection (INLP) method. This method find patterns in the data that are linked to a sensitive attribute by training a linear classifier to predict that attribute. By iteratively projecting the input into the null space of classifier's weight vector for the prediction, INLP learns and removes the these patterns more efficiently and effectively.

Despite its effectiveness in achieving fairness, INLP may damage the useful parts of the data as it deletes several dimensions throughout the iterative projection process (Belrose et al. (2023)). To address this limitation, Belrose et al. (2023) suggest *Least Squares Concept Erasure* (LEACE) as an alternative. LEACE projects the vector only once onto a subspace that is orthogonal to the direction of correlation between input and the sensitive attribute. In doing so, it subtracts

the component of the input that aligns with the sensitive attribute's direction, thereby precisely removing information in the input that are strongly correlated with the sensitive attribute in linear way (Belrose et al. (2023)).

According to Belrose et al. (2023), although the LEACE projection enforces $\text{Cov}(\mathbf{P}X, Z) = 0$ thereby achieves fairness in terms of demographic parity, it still leads to a decrease in the prediction accuracy, indicating a trade-off between accuracy and fairness. To analyze this trade-off in ML algorithms, Liu and Vicente (2022) propose a Pareto front analysis. The Pareto front represents the set of non-dominated solutions in multi-objective optimization, where improving one objective necessarily worsens at least one another objective (Wei and Niethammer (2020)).

Based on this theoretical foundation, this research conducts Pareto-front analysis of the LEACE projection to empirically examine the trade-off between fairness and prediction accuracy. Furthermore, we investigate how the shape of the Pareto-front varies depending data characteristics, by adjusting the degree of correlation between the target variable and the sensitive attribute.

4 Methodology

4.1 Preliminaries and Notation

Symbol	Meaning
$X \in \mathbb{R}^d$	d-dimensional random input vector
$X^* \in \mathbb{R}^d$	Realization of the input vector
$Z \in \mathbb{R}^k$	Sensitive attribute vector
$z \in \mathbb{R}$	Sensitive attribute
$y \in \mathbb{R}$	Continuous target variable
$\mathbf{P}_{\text{LEACE}} \in \mathbb{R}^{d\times d}$	Projection matrix in LEACE
$\mathbf{P}_\gamma \in \mathbb{R}^{d\times d}$	Projection matrix in γ controlled LEACE
$\mathbf{P}_{\mathbf{W}\mathbf{\Sigma}_{Xz}} \in \mathbb{R}^{d\times d}$	Projection matrix onto column space of $\mathbf{W}\mathbf{\Sigma}_{Xz}$
$\gamma \in [0,1]$	Covariance ratio
$\mathbf{W} = (\Sigma_{XX}^{\frac{1}{2}})^+$	Whitening transformation matrix
$V = \mathbf{W}\mathbf{\Sigma}_{Xz}$	Whitened covariance vector between X and z
$\mathbf{A}^+$	Moore-Penrose pseudoinverse of matrix A
$\|\cdot\|_2$	Euclidean norm
$\|\cdot\|_M^2$	Mahalanobis norm
β_y	Target coefficient vector, generating $y = X\beta_y + \epsilon_y$
β_z	Sensitive attribute coefficient vector, generating $z = X\beta_z + \epsilon_z$
θ	Angle between β_y and β_z, controlling the correlation between y and z

Table 1: Summary of symbols and notation used in the paper

In this paper, we adopt a consistent notation: bold-uppercase letters denote matrices, uppercase letters represent high-dimensional vectors, and lowercase letters are used for scalar values or low-dimensional vectors. In particular, we distinguish between random input vector X and observed input vector X^*, which is the realization of X. This distinction is necessary to separate theoret-

ical derivation based on expectations and computation performed on observed data. In addition, the superscript $+$ denotes the Moore-Penrose pseudoinverse (Belrose et al. (2023)), which is the generalization of the matrix inverse when the matrix is not a full rank (Barata and Hussein (2011)).

4.2 Baseline Projection Matrix from Belrose et al. (2023)

The objective of LEACE method proposed by Belrose et al. (2023) is to minimize the distance between input vector X and projected input vector $\mathbf{P}X$, subject to the constraint that the projected input is uncorrelated with the sensitive attribute Z.

$$\arg\min_{\mathbf{P}\in\mathbb{R}^{d\times d}} \mathbb{E}[\|\mathbf{P}X - X\|_M^2] \quad \text{subject to} \quad \text{Cov}(\mathbf{P}X, Z) = 0$$

According to Belrose et al. (2023), the unique solution to this objective is the following:

$$\mathbf{P}_{\text{LEACE}} = \mathbf{I} - \mathbf{W}^{+}\mathbf{P}_{\mathbf{W}\boldsymbol{\Sigma}_{XZ}}\mathbf{W}$$

Based on this projection matrix, LEACE performs projection in three main steps. First, the data is demeaned(centered) and whitened to equalize variance in all distances. Next, the whitened data is projected onto the column space of $\mathbf{W}\boldsymbol{\Sigma}_{XZ}$, which is the whitened covariance space between X and Z, responsible for the correlation between X and Z. Lastly, unwhitened data is subtracted from the original input, thereby removing all linearly available information about Z while preserving the structure. Putting together we derive the LEACE formula for input X^* (Belrose et al. (2023)):

$$r_{LEACE}(X^*) = x - \mathbf{W}^{+}\mathbf{P}_{\mathbf{W}\boldsymbol{\Sigma}_{XZ}}\mathbf{W}(X^* - \mathbb{E}[X])$$

4.3 γ-Controlled LEACE and Projection Matrix

To analyze the trade-off between fairness and prediction accuracy, we can reformulate the objective of LEACE projection from paper by Belrose et al. (2023). The following objective aims to find projection matrix $\mathbf{P}_\gamma$ such that it will achieve certain level of non-zero covariance and minimize

the information loss.

$$\arg\min_{\mathbf{P}\in\mathbb{R}^{d\times d}} \mathbb{E}[\|\mathbf{P}X - X\|_M^2] \quad \text{subject to} \quad \|\text{Cov}(\mathbf{P}X, z)\|_2^2 = \gamma\|\text{Cov}(X, z)\|_2^2$$

The sensitive attribute Z is simplified to a scalar vector z in this research for simplification. As cross-covariance $\text{Cov}(X, z)$ reduces to a vector, we use squared Euclidean norm (l_2 norm) to quantify the covariance between input matrix and sensitive attribute. Euclidean norm is defined for a vector $X \in \mathbb{R}^d$ as $\|X\|_2 = (\sum_{i=1}^d X_i^2)^{\frac{1}{2}} = \sqrt{X^\top X}$ (Horn and Johnson (2013)).

The main difference between objective of original LEACE projection and the new formulation lies in the covariance condition. While Belrose et al. (2023) project the input vector to achieve zero covariance between $\mathbf{P}X$ and Z, our research aims to analyze how prediction accuracy changes as we gradually reduce the information of z in $\mathbf{P}X$—that is, under varying levels of covariance between $\mathbf{P}X$ and z.

While preserving the basic structure of LEACE projection, this research introduces a scalar multiplier into the LEACE formula to construct different projection matrices for relaxed fairness constraint. Instead of completely removing the sensitive component by projecting the input vector onto the null space, the multiplier method allows 'partial' removal of information, enabling the analysis of accuracy—fairness tradeoff. The following is the new LEACE projection matrix in this paper, incorporating the derived multiplier that satisfies the new objective and covariance constraint. As $\mathbf{W} = (\Sigma_{XX}^{\frac{1}{2}})^+$ is full-rank in our setting, we use inverse instead of the Moore-Penrose pseudoinverse for simplicity.

$$\mathbf{P}_\gamma = I - (1 - \sqrt{\gamma})\mathbf{W}^{-1}\mathbf{P}_{\mathbf{W}\Sigma_{Xz}}\mathbf{W}$$

Consequently, different projection matrices corresponding to different γ values share the same structure, differing solely in scalar multiplier $1 - \sqrt{\gamma}$ value. Thus, we can derive the closed form of multiplier as a function of the desired covariance ratio γ.

As $\mathbf{W}\mathbf{\Sigma}_{Xz}$ is a rank-one, the projection matrix $\mathbf{P}_{\mathbf{W}\mathbf{\Sigma}_{Xz}}$ can be written in form as

$$\mathbf{P}_{\mathbf{W}\mathbf{\Sigma}_{Xz}} = \frac{VV^\top}{\|V\|^2} \text{ where } V = \mathbf{W}\Sigma_{Xz}$$

Thus, our final projection matrix for this research is the following:

$$\mathbf{P}_\gamma = I - (1-\sqrt{\gamma})\mathbf{W}^{-1}\left(\frac{VV^\top}{\|V\|^2}\right)\mathbf{W}, \text{ where } V = \mathbf{W}\Sigma_{Xz}, \mathbf{W} = (\mathbf{\Sigma}_{XX}^{\frac{1}{2}})^{-1} \tag{1}$$

For simplicity, we refer to our variant of LEACE utilizing the γ covariance constraint as γ-controlled LEACE throughout this paper.

4.4 Verification of γ-controlled LEACE projection matrix

In this section, we verify that the new matrix $\mathbf{P}_\gamma$ achieves the target covariance $\gamma\|\text{Cov}(X,z)\|_2^2$ and whether it recovers the original LEACE matrix $\mathbf{P}_{\text{LEACE}}$ when $\gamma = 0$.

Using the fact that $\mathbf{\Sigma}_{Xz}$ can be expressed as ηS where $S = \frac{\mathbf{\Sigma}_{Xz}}{\|\mathbf{\Sigma}_{Xz}\|}$ and η is a scalar multiplier:

$$\begin{aligned}
\mathbf{P}_\gamma\mathbf{\Sigma}_{Xz} &= \mathbf{\Sigma}_{Xz} - (1-\sqrt{\gamma})\mathbf{W}^{-1}\frac{\mathbf{W}\mathbf{\Sigma}_{Xz}}{\|\mathbf{W}\mathbf{\Sigma}_{Xz}\|}\left(\frac{\mathbf{W}\mathbf{\Sigma}_{Xz}}{\|\mathbf{W}\mathbf{\Sigma}_{Xz}\|}\right)^\top \mathbf{W}\mathbf{\Sigma}_{Xz} \\
&= \eta\left(S - (1-\sqrt{\gamma})\mathbf{W}^{-1}\frac{\mathbf{W}S}{\|\mathbf{W}S\|}\cdot\frac{\mathbf{W}S}{\|\mathbf{W}S\|}^\top \mathbf{W}S\right) \\
&= \eta\left(S - (1-\sqrt{\gamma})\mathbf{W}^{-1}\mathbf{W}S\right) = \eta\sqrt{\gamma}v \\
\therefore \|\mathbf{P}_\gamma\mathbf{\Sigma}_{Xz}\|_2^2 &= \eta^2\cdot\gamma\cdot S^\top S = \gamma\|\mathbf{\Sigma}_{Xz}\|_2^2
\end{aligned}$$

Furthermore, as we set $\gamma = 0$, the γ-controlled LEACE matrix reduces to $I - \mathbf{W}^{-1}\left(\frac{VV^\top}{\|V\|^2}\right)\mathbf{W}$. Noting that $= \frac{VV^\top}{\|V\|^2}$ corresponds to projection matrix $\mathbf{P}_{\mathbf{W}\mathbf{\Sigma}_{Xz}}$, this expression is equivalent to the original projection matrix $\mathbf{P}_{\text{LEACE}}$ introduced in Belrose et al. (2023).

While a formal proof that proposed γ-controlled LEACE matrix $\mathbf{P}_\gamma$ is the optimal solution to the objective remains an open question, the preceding verification suggest that it is a strong candidate for optimal solution and a suitable choice for this research.

4.5 Dataset Generation

To empirically evaluate the effect and trade-off from γ-controlled LEACE projection, we utilize the simulated dataset with a controllable correlation between target variable y and the sensitive attribute z. We first generate each random input vector $X \in \mathbb{R}^d$ with independent and identically distributed standard normal entries. The target variable and sensitive attribute are then defined as $y = X\beta_y + \epsilon_y$ and $z = X\beta_z + \epsilon_z$, where $\epsilon_y, \epsilon_z \sim \mathcal{N}(0, 1)$. All noise terms ϵ_y and ϵ_z are independent of each other and of the input vector X. The coefficient vectors β_y and β_z are constructed such that angle between two coefficients is fixed to a predetermined value θ.

Specifically, first, β_y is sampled from a standard normal distribution and normalized. Then, β_z is generated by rotating the random unit vector v in the subspace orthogonal to β_y ($\beta_z = \cos\theta\beta_y + \sin\theta v$), ensuring that angle between β_y and β_z corresponds exactly to θ. Both coefficient vectors are normalized ($\|\beta_y\|_2, \|\beta_z\|_2 = 1$), so that angle θ fully controls the correlation. This setup allows precise control over the correlation between y and z, thereby enabling an evaluation of trade-off under varying levels of dependence.

We simulate a dataset with input dimension $d = 10$, a sample size of $n = 5000$, and a fixed angle $\theta = 60^\circ$ between β_y and β_z. This results in final dataset consisting of an input matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, sensitive attribute vector $Z \in \mathbb{R}^n$, and target variable vector $Y \in \mathbb{R}^n$. To ensure the reproducibility, we also set the random seed to 42. The generated dataset is then split into training and test sets using an 80/20 ratio using random partitioning. All training of prediction model and projection matrix computations are performed using the training dataset, while evaluation metrics for prediction accuracy are computed on the test dataset for generalization.

4.6 Classifier Selection and Trade-off Analysis

To evaluate the prediction accuracy in the accuracy-fairness trade-off, we employ a **Random Forest Regressor** (RFR) as the predictive model. Unlike the original LEACE research setup by Belrose et al. (2023), which focuses on the classification task with a linear model, our setting incor-

porates the continuous target variable y, requesting a regression approach. Furthermore, this was motivated by its robustness to overfitting and ability to capture non-linear relationships (Lin et al. (2023)).

For each covariance ratio γ, a corresponding projection matrix is generated using the training dataset. The projected training features are then used to train the Random Forest Regressor (RFR). Subsequently, the same projection is applied to the test dataset, and the trained RFR is used on projected test features to generate predictions of the true target value y. The Mean Squared Error (MSE) is then calculated between the predicted outputs from RFR and true continuous target values. To account for randomness in the dataset splitting and model training, and to ensure robustness, this procedure is repeated 10 times, and the average MSE is computed for each value of γ. The average MSE is used as the final predictive accuracy for each value of γ. By computing the MSE across varying levels of desired covariance constraint, we quantify changes in predictive accuracy resulting from the projection.

The level of fairness will be quantified using the squared Euclidean norm of original covariance level between X and z, denoted as $\|\text{Cov}(X, z)\|_2^2$, and the covariance ratio γ introduced in derivation of the γ-controlled projection matrix. We will select a series of γ values between 0 and 1—e.g."$\gamma \in \{1.0, 0.9, 0.8, 0.7, \dots\}$"—each corresponding to 90%, 80%, 70% of the original $\|\text{Cov}(X, z)\|_2^2$. A smaller value of γ corresponds to a stricter covariance condition, as it enforces a lower covariance between the projected features and the sensitive attribute z. Thus, it will result in more 'fair' projection by removing more sensitive information from the input. While we use the covariance between the sensitive attribute and the (projected) input to quantify 'fairness', its theoretical connection to demographic parity—particularly when $\gamma \neq 0$, where we enforce $\text{Cov}(\mathbf{P}X, z) = 0$—remains a limitation (see Discussion and Conclusion). For clarity, we refer to multiplier γ as **covariance ratio** in this paper.

For each γ value we derive projection matrix that satisfies our objective, then record the model's average prediction accuracy for different projection matrices. Plotting the prediction accuracy against γ will show us how covariance constraint affects the prediction performance. This allows us to analyze the trade-off between fairness and accuracy in our γ-controlled LEACE method.

4.7 Constructing the Pareto Front

This paper constructs the Pareto front to analyze the trade-off boundary and identify optimal points between fairness and accuracy. The Pareto front consists of non-dominated points (x_i, y_i), where a point (x_j, y_j) is said to dominate (x_i, y_i) if $x_j \leq x_i$, $y_j \leq y_i$, and at least of the inequalities is strict ($x_j < x_i$ or $y_j < y_i$) (Liu and Vicente (2022)). This condition ensures that each point on the Pareto front represents a solution for which no other configuration simultaneously achieves both lower covariance and lower MSE. The resulting front provides a visualization and interpretation of the fairness-accuracy trade-off.

To conclude, our research consists of the following steps:

1. Estimate the covariance vector $\|\text{Cov}(X, z)\|_2^2$ from the training dataset and construct the γ-controlled LEACE projection matrix using the training dataset for different values of γ (representing desired level of fairness).

2. Train a Random Forest Regressor (RFR) based on the projected training features.

3. Evaluate prediction accuracy by applying the trained RFR to the projected test dataset features and computing Mean Squared Error (MSE).

4. Repeat this process 10 times to obtain average MSE and trace the resulting Pareto front.

5. Repeat the entire process on different datasets with varying level of correlation between y and z to analyze any changes in the Pareto front.

While our study focuses on analyzing the Pareto-front of γ-controlled LEACE without a formal conventional hypothesis, it is expected that reduction of prediction accuracy will vary depending on the strength of projection, creating a measurable trade-off between fairness and accuracy. Specifically, we expect that achieving higher level of fairness is expected to come at the cost of predictive accuracy, as shown from the observation in Belrose et.al (2023). We aim to empirically characterize and confirm this trade-off by tracing Pareto front across different values of γ, corresponding to varying degrees of fairness.

5 Results

5.1 Accuracy-Fairness Trade-off Analysis in γ-Controlled LEACE

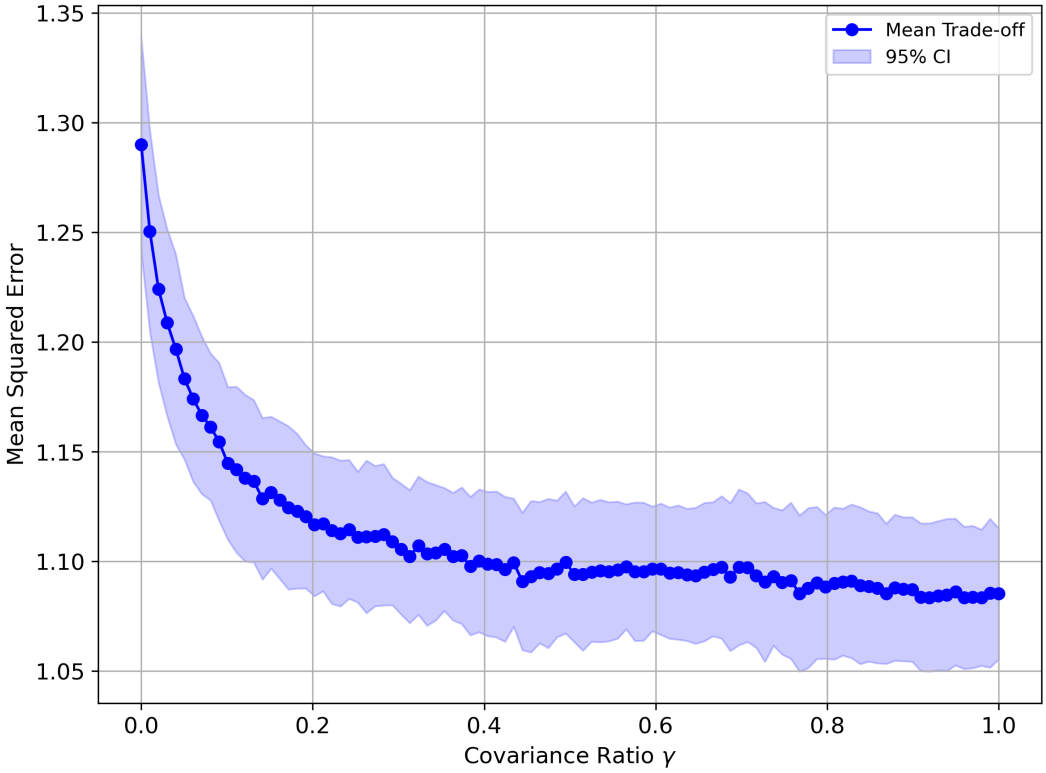

Figure 1: Fairness–Accuracy Trade-off Curve from γ-Controlled LEACE projection. The x-axis represents the covariance ratio γ, and the y-axis shows the Mean Squared Error (MSE) of the prediction model using a Random Forest Regressor (RFR). The sky-blue area represents the 95% confidence interval.

Figure 1 illustrates the fairness-accuracy trade-off curve obtained by evaluating γ-controlled LEACE across different covariance ratio γ. The trade-off curve exhibits a clear inverse relationship between fairness and predictive accuracy. As the fairness condition is relaxed by taking higher γ value (the post-projection covariance increases), the MSE continuously decreases. Conversely, using a stronger fairness condition with lower γ value (the covariance after projection decreases), results in higher MSE values, indicating a loss in predictive accuracy.

This demonstrates that achieving a higher level of fairness comes at the cost of model performance. This trade-off pattern strongly aligns with the theoretical expectations and empirical findings of Belrose et al. (2023), who also observed that enforcing a zero-covariance constraint through original LEACE projection resulted in a decrease in predictive accuracy.

We note that the 95% confidence interval closely follow the overall shape of our trade-off curve. This stability indicates that the observed fairness-accuracy trade-off is consistent across repeated runs, and not a result of stochastic variation. Furthermore, we observe that the confidence intervals are narrower at lower γ values, where stronger fairness constraint was enforced. This suggest that as more information is removed via projection, the projected features become simpler and the model exhibits more stable behavior.

The differing slopes of the trade-off curve in the high-covariance and low-covariance regions suggest that the marginal loss of predictability varies depending on the strength of the projection. Figure 1 exhibits a steep slope in the low-γ region and a much flatter slope in the high-γ region. This indicates that under strong covariance constraints, even a small reduction in the desired covariance level can lead to sharp increase in MSE, resulting in a rapid decline in predictive accuracy. Moreover, while the increase in MSE is gradual at high-γ region, it accelerates rapidly beyond a certain covariance threshold. This suggests the presence of a critical point after which further fairness enforcement causes significant loss in accuracy. This observation is consistent with the paper by Tang et al. (2023), who show that over-pursuing the fairness can result in a sharp decrease in model accuracy.

5.2 Pareto Front Analysis on γ-Controlled LEACE Projection

Next, we construct the Pareto front for the γ-controlled LEACE projection by identifying and marking Pareto optimal points from previous trade-off graph.

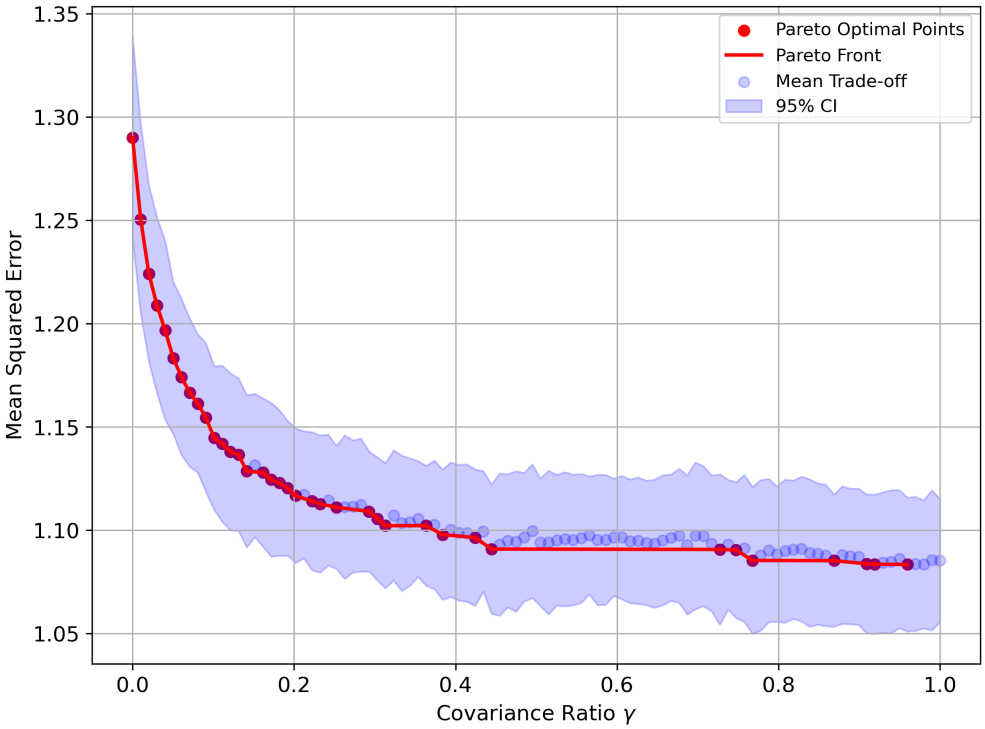

Figure 2: Pareto Front from γ-Controlled LEACE Projection. The full set of trade-off points is shown in light blue, while the Pareto-optimal points and Pareto front are marked in red.

Figure 2 exhibits the Pareto-optimal points and the resulting Pareto front. It is clearly shown that the Pareto-optimal points are distributed across the trade-off curve. This supports the presence of trade-off between fairness and accuracy, as even when considering only the optimal points, improving one objective necessarily comes at the cost of the other. It also suggests the potential for selecting Pareto-optimal points based on the desired level of fairness or prediction accuracy in future research.

Although spread across the curve, Pareto-optimal points tend to cluster around the low-γ region. This indicates that stronger projections and stricter fairness constraints offer more optimal and efficient choices compared to relaxed fairness constraints. Furthermore, in the region roughly between

$\gamma = 0.45$ and $\gamma = 0.75$ no Pareto-optimal points are identified. This suggests the inefficiency of selecting γ values within this range of 0.45 and 0.75 for γ-controlled LEACE projection.

5.3 Pareto Front Analysis Across Different Data Conditions

We next generalize our analysis by constructing the Pareto front under four distinct simulation settings, corresponding to angles $\theta \in \{0°, 30°, 60°, 90°\}$. As the angle θ controls the linear dependence between y and z, each resulting Pareto front illustrates the fairness-accuracy tradeoff at a different level of correlation between the target variable and the sensitive attribute.

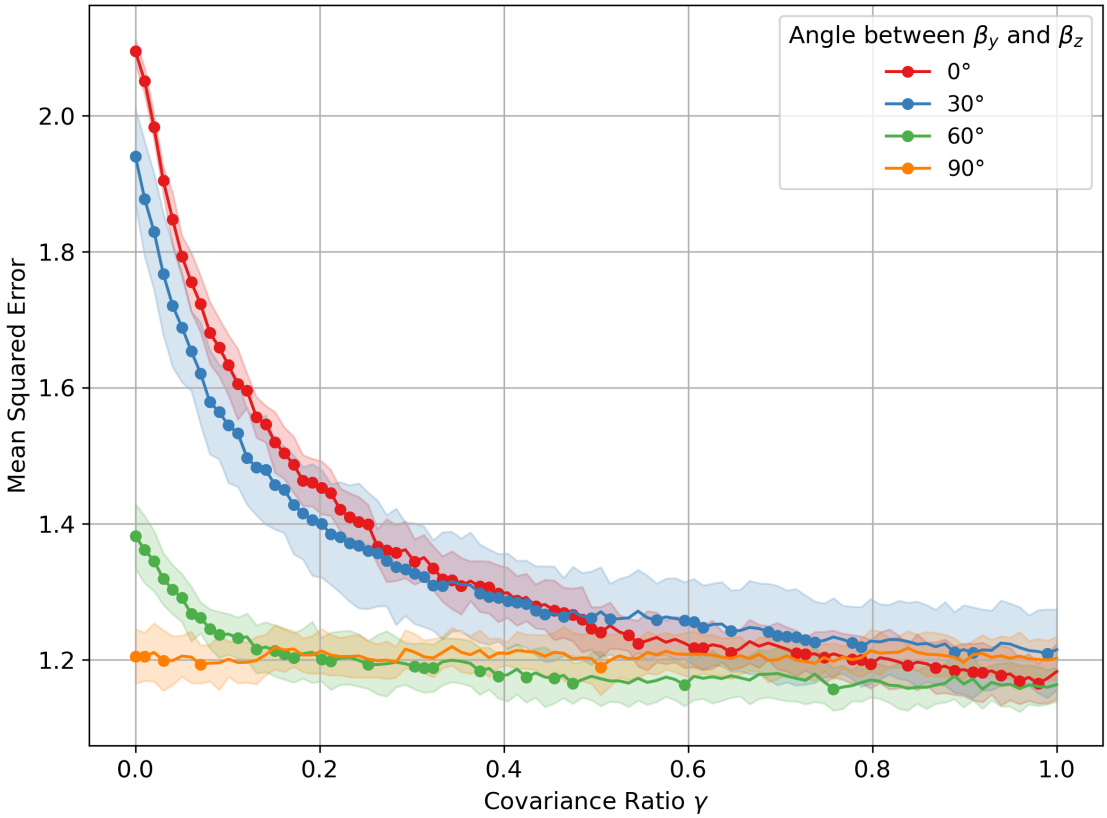

Figure 3: Different Pareto Fronts from γ-Controlled LEACE Projection. Each colored curve corresponds to a trade-off curve resulting from different θ values, with markers indicating the Pareto-optimal points. The horizontal axis shows the covariance ratio γ, and the vertical axis shows the mean squared error (MSE). The color-shaded region denotes 95% confidence interval over 10 runs.

Figure 3 shows the Pareto front of four different simulated datasets. Despite the variation in the

data generation settings, resulting Pareto fronts exhibit similar shapes and trade-off aspect. This resulting plot suggest that trade-off structure between prediction accuracy and fairness remains stable, regardless of the data characteristics. This consistency implies a level of robustness in γ-controlled LEACE and the possibility of generalization across different datasets.

When $\theta = 0°$, as β_z and β_y point in the same direction, y and z are perfectly aligned. Consequently, imposing a strict covariance constraint with low value γ will remove almost all predictive information for y. This is reflected in the Pareto front at $\theta = 0°$, which shows a steep increase in the MSE as γ decreases. Similarly,at $\theta = 90°$, y and z are statistically independent, so projection does not remove predictive information for y, and the trade-off curve remains essentially flat as covariance ratio γ varies.

For intermediate angles ($\theta = 30°$ and $\theta = 60°$), the fairness-accuracy trade-off is still visible but with reduced severity. At $\theta = 30°$, enforcing stricter covariance constraints still results in a steep rise in MSE - less abrupt than at $\theta = 0°$. At $\theta = 60°$, the increase in MSE is clearly milder, as the curve remains flat until around $\gamma = 0.2$ and there is a gradual increase in MSE for $\gamma < 0.2$.

Moreover, for all angles except $\theta = 0°$, the MSE shows a gradual increase in the high-γ region but accelerates rapidly beyond a certain covariance ratio threshold. Although the exact value of corresponding covariance ratio γ and MSE varies for each dataset, this also confirms the existence of 'critical point' in fairness enforcement as described by the paper by Tang et al. (2023). As noted in the previous sections, once this critical threshold is exceeded, further pursuing fairness results in a sharp decrease in prediction accuracy.

6 Discussion and Conclusion

Across all experiments, we observe a consistent trade-off pattern between fairness and prediction accuracy. Despite varying data generation parameters and settings, the resulting Pareto fronts exhibit similar shapes. This aligns with the theoretical trade-off between fairness and accuracy in projection method, and observation of information loss from Belrose et al. (2023).

The ideal balance and desired level of accuracy and fairness may vary depending on the user's objective and purpose of projection. A key strength of our γ-controlled LEACE approach lies in its flexibility: by adjusting the value of γ, users can effectively adjust the projection to achieve desired level of fairness or prediction accuracy, based on the dataset features and their preferences.

While our experimental results demonstrate the theoretical trade-off and suggest robustness of the γ-controlled LEACE method, several limitations remain. Our research is conducted on simulated data, which may not fully reflect the complexity and biases of real-world data. We assume a linear dependence between the sensitive attribute z and y via coefficient β_y, β_z, and input X, which may not be generalized in practical applications and complicated correlation structures. Lastly, it remains an open question whether the γ-controlled LEACE method and trade-off pattern holds under more complex, non-linear models, for example.

These limitations lead to several directions for future research. This research assumes sensitive attribute z as a scalar value. To further reflect the complexity in real-world data, we can extend the γ-controlled LEACE further to high-dimensional sensitive attribute vector Z or apply the method on non-linear datasets. Furthermore, the idea of controlling the projection strength may be applicable to classification problem with binary target variable y.

Moreover, one important limitation of this research is the quantification of 'unfairness'. Belrose et al. (2023) prove that reaching $\text{Cov}(X, Z) = 0$ ensures perfect demographic parity of X over sensitive attribute Z. However, the case where $\text{Cov}(X, Z) = c \neq 0$ remains unexplored. The clear relationship between the magnitude of covariance and the level of deviation from demographic parity has not been characterized. In other words, we do not yet offer a closed-form expression or any theoretical bound for the degree of unfairness across varying levels of non-zero covariance. Ultimately, quantifying 'unfairness' in future research will allow our projections to satisfy a more rigorous definition of fairness and to analyze the corresponding trade-offs.

To conclude, we propose the γ-controlled LEACE, a projection method that enables the control over the desired level of fairness and accuracy. By adjusting the projection ratio γ, users can explore different projection strengths, analyze the fairness-accuracy trade-offs, and identify possible optimal points.

7 References

Barata, João C. A., and Hussein, M.S. (2011): "The Moore-Penrose Pseudoinverse: A Tutorial Review of the Theory," *Brazilian Journal of Physics* 42(1): 146-165.

BBC. 2018. "Amazon Scrapped 'Sexist AI' Tool," *BBC News*, October 10, Accessed May 12, 2025. https://www.bbc.com/news/technology-45809919.

Belrose, N., Schneider-Joseph, D., Ravfogel, S., Cotterell, R., Raff, E., and Biderman, S. (2023): "LEACE: Perfect linear concept erasure in closed form," *Advances in Neural Information Processing Systems*, 36, 66044–66063.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016): "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings," *Advances in neural information processing systems*, 29.

Gonen, H., and Goldberg, Y. (2019): "Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 609–614.

Horn, A. R., and Johnson, R. C. (2013): *Matrix Analysis*, 2nd ed., Cambridge University Press.

Lin, J., Zhuang, Y., Zhao, Y., Li, H., He, X., and Lu, S. (2023): "Measuring the Non-Linear Relationship between Three-Dimensional Built Environment and Urban Vitality Based on a Random Forest Model," *International Journal of Environmental Research and Public Health*, 20(1), 734.

Liu, S., and Vicente, L. N. (2022): "Accuracy and fairness trade-offs in machine learning: a stochastic multi-objective approach," *Computational Management Science*, 19, 513-537.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. (2018): "Learning Adversarially Fair and Transferable Representations," *Proceedings of the 35th International Conference on Machine Learning*, 80: 3384-3393.

Miller, C. C. (2015): "Can an Algorithm Hire Better Than a Human?" *The New York Times*, June 25, 2015. Available at: https://www.nytimes.com/2015/06/26/upshot/can-an-algorithm-hire-better-than-a-human.html.

Olson, P. (2011): "The Algorithm That Beats Your Bank Manager," *Forbes*, March 15, 2011. Available at: https://www.forbes.com/sites/parmyolson/2011/03/15/the-algorithm-that-beats-your-bank-manager/.

Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., and Goldberg, Y. (2020): "Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7237-7256.

Sarker, I., H. (2021): "Machine Learning: Algorithms, Real-World Applications and Research Directions," *2021 SN Computer Science,* 2:160

Satopää, V., Albrecht, J., Irwin D., and Raghavan, B. (2011): "Finding a Kneedle in a Haystack: Detecting Knee Points in System Behavior," *Proceedings of the 31st International Conference on Distributed Computing Systems Workshops(ICDCSW),* 161-171.

Tang, H., Cheng, L., Liu, N., Du, M. (2023): "*A Theoretical Approach to Characterize the Accuracy-Fairness Trade-off Pareto Frontier*," arXiv, arXiv:2310.12785

Verma, S. and Rubin, J. (2018). "Fairness definitions explained." *Proceedings of the International Workshop on Software Fairness*, 1-7.

Wei, S., and Niethammer, M. (2021): "The Fairness-Accuracy Pareto Front," *Statistical Analysis and Data Mining: The ASA Data Science Journal 15* (1): 1-13.

Zhang, B.H., Lemoine, B., and Mitchell, M. (2018): "Mitigating Unwanted Biases with Adversarial Learning," *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335-340

Zhang, J., Moon A., Zhuo, X., Son, S.W. (2019): "Towards Improving Rate-Distortion Performance of Transform-Based Lossy Compression for HPC," *Proceedings of the IEEE High Performance Extreme Computing Conference(HPEC)*, Waltham, MA, 1-7.

8 Appendix

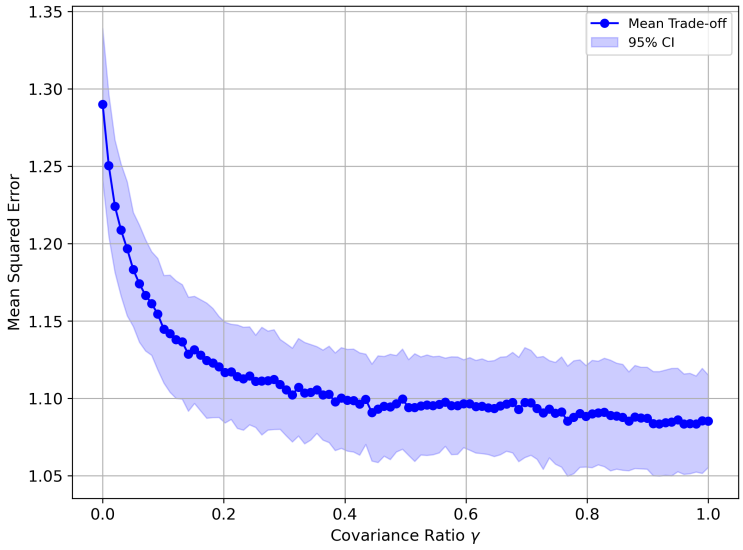

Figure A.1: Fairness–Accuracy Trade-off Curve from γ-Controlled LEACE projection. The x-axis represents the covariance ratio γ, and the y-axis shows the Mean Squared Error (MSE) of the prediction model using a Random Forest Regressor (RFR). The sky-blue area represents the 95% confidence interval.

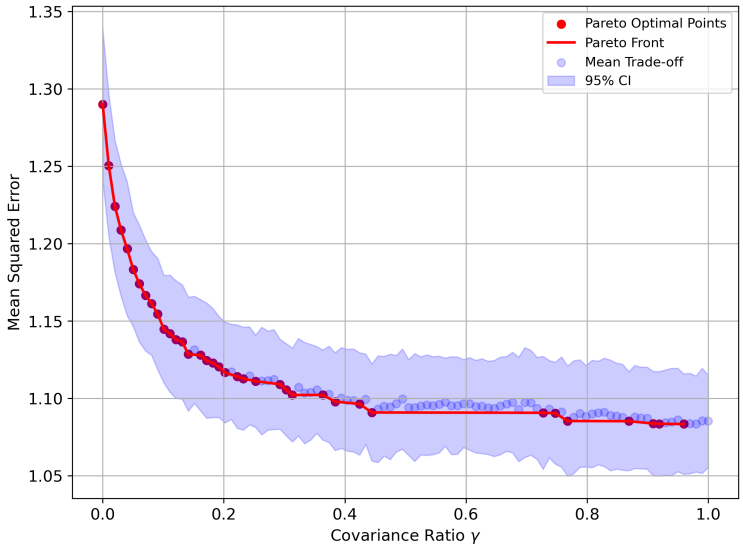

Figure A.2: Pareto Front from γ-Controlled LEACE Projection. The full set of trade-off points is shown in light blue, while the Pareto-optimal points are marked in red.

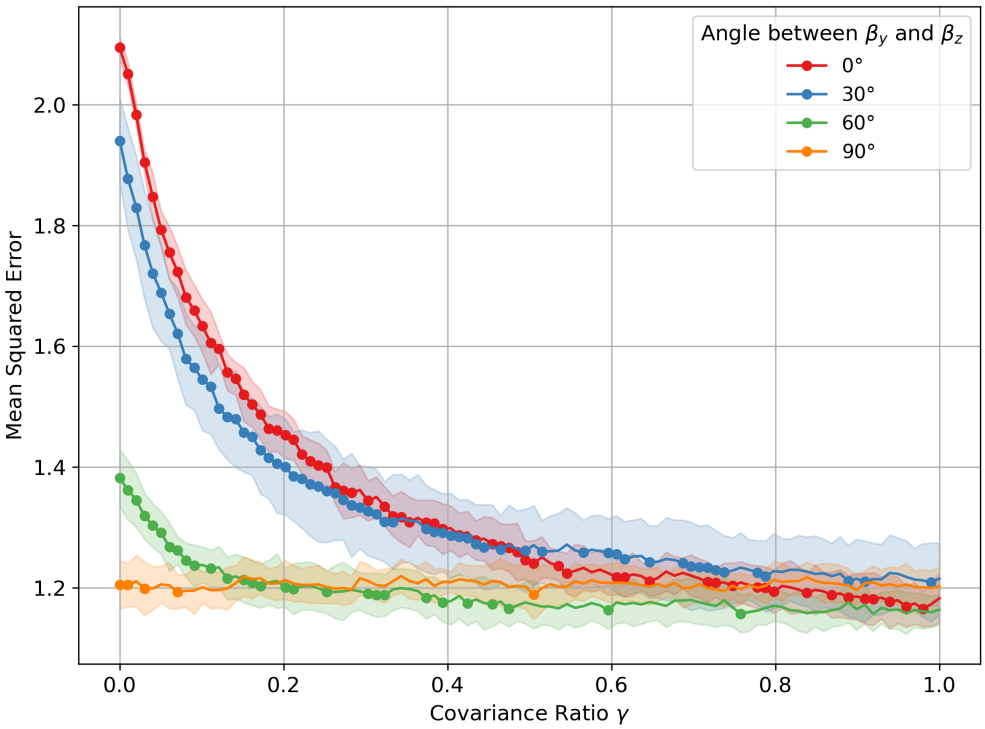

Figure A.3: Different Pareto Fronts from γ-Controlled LEACE Projection. Each colored curve corresponds to a trade-off curve resulting from different θ values, with markers indicating the Pareto-optimal points. The horizontal axis shows the covariance ratio γ, and the vertical axis shows the mean squared error (MSE). The color-shaded region denotes 95% confidence interval over 10 runs.