

Data Science in Finance: Robustness, Fairness, and Strategic Modeling

Mike Li

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2024

© 2024

Mike Li

All Rights Reserved

Abstract

Data Science in Finance: Robustness, Fairness, and Strategic Modeling

Mike Li

In the multifaceted landscape of financial markets, the understanding and application of data science methods are crucial for achieving robustness, fairness, and strategic advancement. This dissertation addresses these critical areas through three interconnected studies.

The first study investigates the problem of data imbalance, with particular emphasis on financial applications such as credit risk assessment, where the prevalence of non-defaulting entities overshadows defaulting ones. Traditional classification models often falter under such imbalances, leading to biased predictions. By analyzing linear discriminant functions under conditions where one class's sample size grows indefinitely while the other remains fixed, this study reveals that certain parameters stabilize, providing robust predictions. This robustness ensures model reliability even in skewed data environments.

The second study explores anomalies in option pricing, specifically the total positivity of order 2 (TP_2) in call options and the reverse sign rule of order 2 (RR_2) in put options within the S&P 500 index. By examining the empirical significance and occurrence patterns of these violations, the research identifies potential trading opportunities. The findings demonstrate that while these conditions are mostly satisfied, violations can be strategically exploited for consistent positive returns, providing practical insights into profitable trading strategies.

The third study addresses the fairness of regulatory stress tests, which are crucial for assessing the capital adequacy of banks. The uniform application of stress test models across diverse banks

raises concerns about fairness and accuracy. This study proposes a method to aggregate individual models into a common framework, balancing forecast accuracy and equitable treatment. The research demonstrates that estimating and discarding centered bank fixed effects leads to more reliable and fair stress test outcomes.

The conclusions of these studies highlight the importance of understanding the behavior of commonly used models in handling imbalanced data, the strategic exploitation of option pricing anomalies for profitable trading, and the need for fair regulatory practices to ensure financial stability. Together, these findings contribute to a deeper understanding of data science in finance, offering practical insights for regulators, financial institutions, and traders.

Table of Contents

Acknowledgments	xii
Dedication	xiv
Preface	1
Chapter 1: Robustness of Linear Classifiers under Infinite Imbalance	2
1.1 Introduction	2
1.2 Discriminant Functions	6
1.2.1 Logistically Consistent Objectives	6
1.2.2 Examples of Objective Functions	8
1.2.3 Empirical Loss	10
1.3 Existence of a Minimizer	11
1.4 Convergence Under Infinite Imbalance	13
1.5 Robustness Interpretation of β_*	17
1.5.1 Asymptotically Subexponential Weight Functions	17
1.5.2 Infinite Upsampling	20
1.5.3 The Gaussian Case	21
1.5.4 Asymptotically Exponential Weight Functions	24
1.6 Numerical Examples	26

1.6.1	Convergence Simulations	27
1.6.2	High-Sensitivity and High-Specificity Regions	28
1.7	A Credit Risk Application	31
1.7.1	Freddie Mac Data	31
1.7.2	High-Sensitivity Classifiers	34
1.7.3	pAUC plots	36
1.7.4	Choice of λ	36
1.8	Concluding Remarks	38
Chapter 2: Trading TP ₂ /RR ₂ Violations in Options		40
2.1	Introduction	40
2.2	Background	42
2.3	Options on S&P 500 Index	44
2.3.1	Evolution of S&P Options	44
2.3.2	Data	46
2.4	TP ₂ -Violating Option Pairs	47
2.4.1	Determining TP ₂ -Violating Option Pairs	47
2.4.2	Evolution of Violations	49
2.4.3	Violation Rates and Market Conditions	51
2.4.4	Maturities in Violation Pairs	52
2.4.5	Option Deltas in Violation Pairs	54
2.4.6	Violation Correction Rates	55
2.5	Trading TP ₂ and RR ₂ Violations	56

2.5.1	From TP_2 and RR_2 to Trading Strategies	57
2.5.2	Per trade profits	58
2.5.3	Dynamic Trading strategy	64
2.5.4	Cashflow Analysis	72
2.5.5	Trading Costs Considerations	77
2.6	Concluding Remarks	79
Chapter 3: Fairness in Regulatory Stress Tests		80
3.1	Introduction	80
3.2	Background	84
3.2.1	Regulatory Bank Stress Tests	85
3.2.2	Bank Heterogeneity	86
3.2.3	Heterogeneity and Fairness	88
3.3	Pooling: Fairness Through Unawareness?	90
3.3.1	Basic Model	90
3.3.2	Average Treatment Effects	95
3.4	Fair Regressions	96
3.4.1	Projection to Fairness	97
3.4.2	Formal Equality of Opportunity	100
3.4.3	Conditional Expectation Model	104
3.4.4	Substantive Equality of Opportunity	105
3.4.5	A Unified Perspective: Legitimate Information	107
3.4.6	Extension of FEO for Interaction Effects	109

3.5	Nonlinear Models	111
3.6	Concluding Remarks	113
	Epilogue	115
	References	117
	Appendix A: Robustness of Linear Classifiers under Infinite Imbalance	122
A.1	Proofs for Section 1.3	122
A.2	Proofs for Section 1.4	127
A.2.1	Proof of Lemma 1.4.1	127
A.2.2	Proof of Proposition 1.4.1	128
A.2.3	A Convergence Result	129
A.2.4	Proof of Theorem 1.4.2	137
A.2.5	Boundedness of $V(u)/e^u w(u)$	144
A.3	Proofs for Section 5	145
A.3.1	Proof of Proposition 1.5.1	145
A.4	Connection with Nonlinear Classifiers	146
A.5	Supplementary Tables	148
A.5.1	Freddie Mac Dataset AUC	148
A.5.2	Freddie Mac Testing TPR	149
A.6	Delta Function Weight	149
	Appendix B: Trading TP_2 and RR_2 Violations	152
B.1	Data Sources	152

B.2	AM vs PM-Settled Options	153
B.3	Robustness Checks	154
B.3.1	Next-available Trades	154
B.3.2	Two-Strike Approximation	155
B.3.3	Short-only Strategies	157
Appendix C: Fairness in Regulatory Stress Tests		161
C.1	Proofs	161
C.2	Cross-Bank Parameter Externalities	169
C.3	Sensitivity Analysis	173
C.3.1	Forecast Bias	173
C.3.2	Improvement in Intercept α_s	175
C.3.3	Improvement in Loan Quality	176
C.3.4	Improvement in Loan Management	177
C.4	Convex Combinations of Coefficients	178
C.5	Empirical Evidence	179
C.5.1	Data	179
C.5.2	Heterogeneity in Slopes and Intercepts	184
C.5.3	Robustness Checks	188
C.6	Nonlinear Models	192
C.7	Additional Information on Empirical Analysis	195
C.8	Revenue Models	197
Appendix References		199

List of Figures

1.1	Examples of weight functions w and penalty functions U and V	10
1.2	F surrounds point x^* if it assigns mass at least $\delta > 0$ to the shaded half-space, for every direction ω , for some $\epsilon > 0$	12
1.3	Classification boundaries with unequal covariance matrices. Dashed lines show the effect of varying λ	23
1.4	Two-Dimensional Example	28
1.5	Illustration of a specificity-oriented (left) and sensitivity-oriented (right) pAUC. . .	30
1.6	Comparison of pAUC values for logistic and exponential ($\lambda = 0.1, 0.5, 0.9$) classifiers	31
1.7	As N increases, the logistic pAUC values move closer to the exponential pAUC values with small λ in both the high-sensitivity and high-specificity regions	32
1.8	Freddie Mac Summary Data	34
1.9	Comparison of pAUC values in test data for logistic and exponential ($\lambda = 0.1, 0.5, 0.9$) classifiers using Freddie Mac loan data. SMOTE upsampling with logistic regression is also included for comparison	37
1.10	Test data pAUC plots with 90% confidence bands for $\lambda = 0.1$ and $\lambda = 0.9$	37
2.1	Total yearly trading volume separated to AM (orange), PM-settled and non-ODTE (light blue), and PM-settled ODTE (dark blue).	46
2.2	TP ₂ violations count by settlement type.	49
2.3	RR ₂ violations count by settlement type.	50
2.4	TP ₂ and RR ₂ violations (rolling 30-day average).	51

2.5	(a) TP ₂ violation counts; (b) TP ₂ violation rates. T_1 in the horizontal axis and T_2 vertical axis. DTEs are grouped into 7-day blocks.	53
2.6	(a) RR ₂ violation counts; (b) RR ₂ violation rates. T_1 in the horizontal axis and T_2 vertical axis. DTEs are grouped into 7-day blocks.	54
2.7	(a) Heatmap of count of TP ₂ -violating option deltas (b) Heatmap of count of RR ₂ -violating option deltas. x-axis: delta for (a) $C(K_1, T_1)$ or (b) $P(K_1, T_1)$; y-axis: delta for (a) $C(K_2, T_2)$ or (b) $P(K_2, T_2)$	55
2.8	Cumulative profits in log-scale. Left: T_1 -denominated TP ₂ trades. Right: K_2 -denominated TP ₂ trades. The black lines are the S&P 500 index. The solid and dashed red lines are cumulative total profits and cash premiums, respectively. . . .	68
2.9	Cumulative profits in log-scale. The black line is the S&P 500 index. The solid and dashed red lines are cumulative total profits and cash premiums, respectively. . .	72
3.1	Heterogeneity among large banks. Left: Distribution of 2022 stress test banks by GIC sub-industry. Right: The percentage of loans in each of four categories for each of the U.S. G-SIBs, based on Y-9C reports for Q4 2021.	87
B.1	(a) TP ₂ violation rate of AM-settled options. (b) TP ₂ violation rate of PM-settled options.	153
B.2	(a) RR ₂ violation rate of AM-settled options. (b) RR ₂ violation rate of PM-settled options.	154
B.3	Cumulative returns for T_1 -denominated, next-available trades.	155
B.4	Slope of S&P 500 futures price (30-day vs 7-day).	156
B.5	TP ₂ violations count (two-strike approximation).	156
B.6	RR ₂ violations count (two-strike approximation).	157
B.7	TP ₂ and RR ₂ violation rates (two-strike approximation).	157
B.8	Cumulative returns. Left: T_1 -denominated TP ₂ trades. Right: K_2 -denominated TP ₂ trades. (two-strike approximation).	158
B.9	Cumulative returns of T_1 -denominated RR ₂ trades. (two-strike approximation). . .	159
B.10	Cumulative returns for short-only TP ₂ trades.	160

C.1	First principal component (PC1) of macro variables from 1990 Q2 to 2021 Q4. The dashed lines correspond to the 5th and 95th percentiles of PC1.	181
C.2	Past due rates (winsorized) by bank and loan category. The dots show mean values and each horizontal bar corresponds to ± 1.96 standard errors.	184
C.3	Pooled and FEO predicted loss rates for Citigroup's first lien loans.	189
C.4	Banks' generalized fixed effects. Y-axis is in %.	194
C.5	Pooled and FEO predicted loss rates for Citigroup's first lien loans.	195

List of Tables

1.1	Convergence of coefficients as the sample size N of the majority class grows. Numbers in parentheses are standard errors.	27
1.2	True negative rates (in percent) in test data for classifiers trained at a true positive rate of 99%	35
1.3	Average default loan amount of classified positive instances when TPR is 99% . . .	38
2.1	Regressing i) log of call violation rates; ii) log of put violation rates; and iii) differences between ii) and i) on log of S&P 500 index daily returns, CBOE's VIX index, and CBOE's SKEW index.	52
2.2	Statistics for TP ₂ and RR ₂ -violating option deltas.	54
2.3	Violation survival rates (in %) for TP ₂ and RR ₂ violating option pairs with $T_1 = 3, 7, 14$ days and $T_2 \leq 60$ days.	56
2.4	Strategies and positions.	57
2.5	Cashflow for a TP ₂ -violating trade entered at time t at time $t + T_1$ and $t + T_2$. In addition, at t , a positive cash of $\text{premium}_{C,t}$ is received by all four denominations.	59
2.6	Cashflow for a RR ₂ -violating trade entered at time t at time $t + T_1$ and $t + T_2$. In addition, at t , a positive cash of $\text{premium}_{P,t}$ is received by all four strategies.	60
2.7	Mean profit from TP ₂ -violating trades of different denominations on \$1 cashness. Numbers in parentheses are the percentage of trades yield positive returns.	62
2.8	Mean profit from RR ₂ -violating trades of different denominations on \$1 cashness. Numbers in parentheses are the percentage of trades yield positive returns.	63
2.9	Yearly returns of TP ₂ strategies of different denominations (numbers are in percent).	67

2.10	Annualized Sharpe Ratio for monthly T_1 -denominated returns and S&P 500 index returns.	69
2.11	Regress monthly T_1 -denominated returns on (i) monthly S&P returns only, and (ii) monthly S&P returns and average CBOE's VIX closing levels in the month.	70
2.12	Yearly returns of RR_2 strategies of different denominations (numbers are in percent).	71
2.13	Yearly returns for T_1 -denominated RR_2 trades and S&P 500 index and annualized Sharpe Ratio for monthly T_1 -denominated returns and S&P index returns.	73
2.14	Regress monthly RR_2 -violating T_1 -denominated returns on (i) monthly S&P returns only, and (ii) monthly S&P returns and average VIX closing levels in the month.	73
2.15	Per-trade profit of T_1 -denominated trades for TP_2 and RR_2 violations with trading costs. Numbers in brackets are the percentage of trades receiving positive profits.	78
3.1	Summary of forecast model forms and constraints.	108
A.1	AUC, Logistic Regression	148
A.2	AUC, $\lambda = 0.1$	148
A.3	AUC, $\lambda = 0.5$	148
A.4	AUC, $\lambda = 0.9$	148
A.5	TPR (in percent) in test data using classification thresholds that achieve TPR=99% in training data.	149
B.1	Differences between the probability of an average call option expiring ITM and that of a comparable TP_2 -violating option. Numbers are in percent.	159
B.2	Differences between the probability of an average put option expiring ITM and that of a comparable RR_2 -violating option. Numbers are in percent.	160
C.1	Sensitivity of results for bank l in response to a decrease in parameter μ_s , α_s , or β_s for bank s . Sensitivities shown are for predicted loss $\hat{Y}_l(x)$ (top), mean predicted loss $E[\hat{Y}(X_l)]$ (middle), and the bias $E[\hat{Y}(X_l) - Y_l]$	172

C.2	Loadings of first principal component on macro variables.	180
C.3	Descriptive statistics in percent. Columns 2–4 are calculated from banks’ time averages, and columns 5–8 are calculated from all observations, with mean and standard deviation are stressed time and loan balance weighted.	183
C.4	<i>P</i> -values for heterogeneity tests. In each loan category, the first two rows are for a model with <i>PastDueRate</i> only, and the last two rows are for a model with <i>PastDueRate</i> and <i>MacroPC</i> . The two rows for each model show results under alternative assumptions on the error covariance matrix.	188
C.5	Comparison of coefficients for one-year forecasts. We regress <i>LossRate</i> on (i) <i>PastDueRate</i> and (ii) <i>PastDueRate</i> and <i>MacroPC</i> . Difference is calculated as $\beta_{Pool} - \beta_F$. <i>p</i> -values test $H_0 : \beta_{Pool} = \beta_F$ for <i>PastDueRate</i> or $H_0 : \gamma_{Pool} = \gamma_F$ for <i>MacroPC</i> . . .	188
C.6	Mean and median of relative prediction differences between the pooled and the FEO estimates (in %).	189
C.7	Heterogeneity tests using pre-COVID data with allowance rate as an additional proxy for banks’ portfolio risks.	190
C.8	Comparison of pooled and FEO coefficients using pre-COVID data with allowance rate as an additional proxy for banks’ portfolio risks.	191
C.9	Mean and median of relative prediction differences between the pooled and the FEO estimates (in %) for GAMs.	194
C.10	Symbols and names of included bank holding companies.	196
C.11	Loan variables and FR Y-9C form correspondence.	197
C.12	Coefficient estimates for the AR models. The first two columns use a one-quarter lag, and the last two use a one-year average lag. Columns 1 and 3 correspond to AR models with bank fixed effects, and columns 2 and 4 are pooled AR models. . .	200

Acknowledgements

The Covid-19 pandemic hit in the first year of my doctoral studies, bringing with it a period of significant changes and uncertainties. Born and raised in Wuhan, the epicenter of the pandemic, I experienced its impact firsthand. Despite these circumstances, I'm incredibly grateful for the immense support I received from many people, which has been invaluable in navigating through these times and completing this thesis.

First and foremost, I would like to extend my deepest gratitude to my advisor, Professor Paul Glasserman. His guidance, wisdom, and unwavering support have been instrumental in shaping my academic journey. His support during the Covid-19 pandemic were particularly impactful, as he consistently made time for our meetings whenever I needed them. Under his mentorship, I have not only gained invaluable knowledge in research methodology but have also become a better researcher, inspired by his approach to scholarship. His dedication and passion for the field have motivated me to strive for excellence in my work.

I would also like to express my heartfelt thanks to my fellow PhD students and friends who have been an integral part of this journey. In particular, I am grateful to Shangzhou (Shawn), Wenxin, Wen, and Yibo. Their friendship has been a source of great joy and motivation. Whether it was exploring new places, enjoying meals together, or simply spending time discussing our research and aspirations, their presence has made this journey enjoyable and rewarding.

Additionally, I would like to thank my long-time childhood friend Anran for the countless meals we shared and the various places in New York and around the world we explored together. From the warm and sunny beaches of Cancun and Jamaica to the cold, snowy landscapes of Quebec

and Maine, these invaluable memories and emotional support have been a cornerstone throughout this journey.

Finally, I am deeply thankful to my family. My mother worked on the front lines of Covid-19, from day one of the lockdown in January to the discharge of the last ICU unit in Wuhan in mid-June. Her short hair had grown long by the time she could return home. Throughout this period, she never complained to me about the hardships she faced. Witnessing her unwavering commitment has served as a profound source of inspiration, instilling within me the true meaning of perseverance, dedication, and devotion. I am also grateful to my life partner, whose unwavering patience, understanding, and insightful perspective have proven invaluable, particularly during challenging times. My family's unconditional love, support, and encouragement have been my bedrock of this entire academic journey. Their belief in me has empowered me to persevere and achieve my goals.

To my family.

Preface

The financial markets have always intrigued me with their inherent complexities and rapid evolution. This fascination has guided my academic and professional journey, leading to the research presented in this thesis. The core focus of this work is on three pivotal areas in financial data science: the robustness of linear classifiers under extreme data imbalance, the empirical analysis of option pricing anomalies, and the fairness of regulatory stress tests.

Each chapter in this thesis reflects a step in tackling some of the pressing challenges in finance using data science, exploring different dimensions of financial modeling and strategic decision-making. The insights gained from studying linear classification models under extreme data imbalance provide a theoretical foundation for commonly used models. The investigation into TP_2 and RR_2 violations in options trading opens new avenues for strategic trading, while the analysis of regulatory stress tests underscores the importance of fairness and precision in financial regulations.

Throughout this research, I have been driven by the goal of bridging the gap between theoretical models and practical applications. The dynamic nature of financial markets requires data science approaches that are robust, fair, and strategic. I believe that the findings presented in this thesis contribute to this objective, offering insights for both academics and practitioners in the field.

The journey of completing this thesis has been challenging yet rewarding. This process has not only deepened my understanding of financial data science but has also reinforced my commitment to advancing this field. It is my sincere wish that these studies will inspire further exploration and innovation in financial modeling, ultimately contributing to the stability and efficiency of financial markets.

Chapter 1: Robustness of Linear Classifiers under Infinite Imbalance

In financial applications, particularly credit risk assessment, data imbalance poses a significant challenge. Defaulting entities are considerably fewer than non-defaulting ones, leading traditional classification models to produce biased predictions. This chapter delves into the robustness of linear discriminant functions in such highly imbalanced datasets. By investigating the behavior of classifier coefficients when one class's sample size grows indefinitely while the other remains fixed, we uncover that with appropriate weight functions, the intercept term diverges but the slope vector converges to some robust fixed point. This robustness is crucial for developing models that retain accuracy even in skewed data environments, offering a convenient and reliable tool for risk assessment in finance.

1.1 Introduction

Binary classification tasks often face a severe problem of imbalanced data: observations from one class are plentiful but observations from the other class are scarce. In detecting rare diseases, for example, one may have access to nearly unlimited measurements from healthy patients but only a few from sick patients; lenders often have rich data on low-risk borrowers but fewer observations of borrowers who default.

Under extreme imbalance, a simple rule that predicts that all observations are from the majority class achieves near-perfect accuracy on training data. Such a rule is clearly useless in detecting observations from the minority class, which is often the main objective of the classification.

A common approach to binary classification evaluates a scoring rule at each observation and then applies a threshold to the scores to assign each observation to one class or the other. In analyzing performance under imbalance we would like to separate the choice of threshold (which

can be adjusted to correct for imbalance) from the effect of the features used in the scoring rule.

In the case of logistic regression, this separation is provided by the intercept and the rest of the coefficient vector. Owen [65] analyzed the behavior of logistic regression in the infinitely imbalanced limit, where the size of the majority class becomes infinite while the size of the minority class remains fixed. He showed that the intercept tends to negative infinity, as a consequence of the growing imbalance, but the rest of the coefficient vector approaches a finite limit.

We extend Owen's [65] result to a wide class of linear discriminant functions exhibiting a variety of behaviors. These classifiers assign a score to each observation using a linear combination of features, with coefficients chosen to minimize a loss function; the score is compared to a threshold to classify the observation. In the framework of Eguchi and Copas [24], the loss is determined by a weight function that penalizes high scores for one class and low scores for the other class. We prove infinite-imbalance limits for the coefficient vectors of a broad family of such classifiers, with explicit expressions for the limits. Even in the case of logistic regression, our results extend Owen's [65] because we directly analyze the empirical loss rather than the approximation used in Owen [65].

We distinguish two broad categories of classifiers we call *asymptotically subexponential* and *asymptotically exponential*, based on the left-tail growth rates of their weight functions. The first category includes bounded weight functions, which further include logistic regression as a special case. We show that all classifiers in this category have the same limit under infinite imbalance and are therefore equivalent to logistic regression in this limiting regime.

The asymptotically exponential category includes the loss function in the AdaBoost method (as formulated in Freund and Schapire [33], and Friedman, Hastie, and Tibshirani [34]) and asymmetric extensions of this method. For this category we show that the limiting coefficient vector depends on the exponent in the left-tail growth rate of the weight function. From this perspective, the limit for logistic regression can be seen as a very special boundary case, corresponding to an exponent of zero.

The asymptotically exponential case allows a richer set of limits, and varying the exponent in

this family of methods provides useful flexibility in controlling the performance of a classifier with highly imbalanced data. By varying the exponent we can put more weight on specificity (the true negative rate) or sensitivity (the true positive rate) in the classification task.

To support this interpretation, we study the form of the limiting coefficient vectors to understand what the infinite-imbalance limit says about the classification rules. We show that the limits reflect robustness properties, in the sense that they are optimized against certain worst-case alternatives. Different types of robustness properties can also be seen as different types of conservatism in selecting which errors to emphasize under extreme imbalance.

This robustness or conservatism is easiest to appreciate when the weight function is asymptotically subexponential, which includes the case of logistic regression. We know from the Neyman-Pearson lemma that the optimal rule for classifying an observation as coming from one probability distribution or another uses the likelihood ratio between the two distributions. The limiting coefficient vector with a subexponential weight function is the log likelihood ratio between the distribution of the majority class and a “worst-case” alternative. Among the set of distributions having the mean of the minority class, this alternative is the distribution closest to that of the majority class, with closeness measured through relative entropy or Kullback-Leibler divergence. This is the worst case because distributions that are closer are harder to separate. Thus, we show that the limiting coefficient vector provides the best (Neyman-Pearson) classifier for the worst alternative to the majority distribution among all distributions with the mean of the minority class.

We also prove a version of this result when the left tail of the weight functions grows exponentially. The subexponential case implicitly emphasizes conservatism with respect to false positives in identifying draws from the minority class. The asymptotically exponential case balances concerns about false negatives and false positives, with the relative weight determined by the choice of exponent.

Imbalance is often addressed through downsampling (discarding observations from the majority class) or upsampling (reproducing or creating synthetic observations from the minority class); see, for example, the methods in Chawla et al. [15], Drummond and Holte [21], and Kubat and

Matwin [53], and the comparison of methods in Liu, Wu, and Zhou [60]. The infinite imbalance limits we study can also be understood from this perspective in the asymptotically subexponential case. Linear discriminant rules in this case (including logistic regression) become equivalent, in the infinite-imbalance limit, to an implicit choice of upsampling distribution. This implicit rule upsamples the minority class using the worst-case alternative to the majority class.

We illustrate these ideas through numerical examples and an empirical application. As is customary, we examine classification performance through the receiver operating characteristic (ROC) curve. We use partial area-under-the-curve (pAUC) measures, as introduced in McClish [62], to focus attention on the high-specificity and high-sensitivity endpoints. We argue that these regions are where the choice of weights for the discriminant function matters most. Using exponential weight functions, we find a consistent ranking of performance according to the size of the exponent, but the ordering flips between the high-sensitivity and high-specificity regions. Consistent with our limiting results, the behavior of logistic regression becomes similar to that of a classifier from an exponential weight function with a small exponent, as the degree of imbalance grows.

We apply these ideas in a credit risk setting using mortgage data from Freddie Mac. We consider the problem of predicting default in the first two years of a loan, using features available at the time the loan was made. Defaults are rare, making the data highly imbalanced. We take the view that in an initial screening, a lender would want to achieve a high level of sensitivity in detecting likely defaulters. We calibrate logistic and exponential classifiers to achieve high true positive rates in training data and then compare true positive and true negative rates in test data. The relative performance of the exponential classifiers and logistic regression are consistent with the predictions from our theoretical analysis.

Section 1.2 discusses the class of linear discriminant functions we study based on minimizing expected loss measures. Section 1.3 establishes the existence of unique minimizers for empirical loss measures. Section 1.4 presents our main theoretical results, the coefficient limits under infinite imbalance. Section 1.5 discusses the interpretation of the limits. Section 1.6 presents numerical results, and Section 1.7 develops the credit risk application. Proofs appear in appendices.

1.2 Discriminant Functions

1.2.1 Logistically Consistent Objectives

We consider data in which each observation takes the form of a pair $(x, y) \in \mathbb{R}^d \times \{0, 1\}$, where x is a vector of features or attributes, and y is a binary class label. When we introduce imbalance, 0 will label the majority class, and 1 will label the minority class. A discriminant function $\eta(x)$ assigns a score to each feature vector x , with the intention that points from class 1 will tend to have higher scores than points from class 0, so that the score can be used for classifying unlabeled observations: an observation x is predicted to be from class 1 if and only if $\eta(x) > t$, for some threshold t . A linear discriminant function takes the form $\eta(x) = \alpha + \beta^\top x$, for some $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^d$. We are primarily interested in the vector β ; if our rule for predicting class 1 based on features x is $\alpha + \beta^\top x > t$, then the effect of α can be absorbed into the threshold t .

We select (α, β) by minimizing a loss function. To formulate the objective, it is useful to introduce distributions F_0 and F_1 on \mathbb{R}^d , describing the distributions of features in the two classes, and marginal probabilities π_0 and $\pi_1 = 1 - \pi_0$ for the two labels. Write \mathbb{E}_i for expectation with respect to F_i , $i = 0, 1$, and write \mathbb{P}_i for the corresponding probability measures. The loss function is defined by two increasing functions U, V on \mathbb{R} , which yield the objective

$$C(\alpha, \beta) = -\pi_1 \mathbb{E}_1[U(\alpha + \beta^\top X)] + \pi_0 \mathbb{E}_0[V(\alpha + \beta^\top X)]. \quad (1.1)$$

The first term on the right penalizes small scores in class 1, and the second term penalizes large scores in class 0. We can also write this loss function as

$$C(\alpha, \beta) = \mathbb{E}[-YU(\alpha + \beta^\top X) + (1 - Y)V(\alpha + \beta^\top X)], \quad (1.2)$$

by taking the expectation with respect to the unconditional distribution of (X, Y) , under which $\mathbb{P}(Y = i) = \pi_i$ and X has distribution F_i , given $Y = i$, $i = 0, 1$.

Let η be the log likelihood ratio of F_1 with respect to F_0 , which is well-defined on the intersec-

tion of the support of F_0 and F_1 , and which is not in general linear. Let $p(x) = \mathbb{P}(Y = 1|X = x)$, and define the log-odds

$$\eta_o(x) = \log \frac{p(x)}{1-p(x)} = \log \frac{\pi_1}{\pi_0} + \log \frac{dF_1}{dF_0}(x) \equiv \log \frac{\pi_1}{\pi_0} + \eta(x). \quad (1.3)$$

By the Neyman-Pearson lemma, (1.3) provides the optimal discriminant function in the sense that it minimizes the error probability $\mathbb{P}_1(\eta(X) \leq t)$ for any value of the error probability $\mathbb{P}_0(\eta(X) > t)$, as t varies. (See, e.g., Theorem 3.2.1 of Lehmann and Romano [57] for a precise statement.) The log-odds η_o provides an equivalent classifier because it differs from η by a constant that can be absorbed in the threshold t .

The log-odds need not be linear and thus need not be achievable by minimizing (1.2). Eguchi and Copas [24] proposed the following *logistically consistent* restriction on U and V : if the log-odds function is linear, $\eta_o(x) = \alpha_o + \beta_o^\top x$, then (1.2) should be minimized at (α_o, β_o) . In other words, U and V should deliver the optimal classifier if the optimal classifier is linear.

Eguchi and Copas [24] show that this consistency condition holds if the penalty functions U and V satisfy

$$\frac{\partial V(u)}{\partial u} = e^u \frac{\partial U(u)}{\partial u};$$

equivalently,

$$U(u) = C_U - \int_u^\infty w(s)ds, \quad V(u) = C_V + \int_{-\infty}^u e^s w(s)ds, \quad (1.4)$$

for some positive function w and some constants C_U and C_V . Our analysis applies to linear discriminant functions obtained by minimizing an empirical counterpart of (1.2) with U and V of this form. The constants C_U and C_V have no effect in minimizing (1.2).

Before proceeding further, we briefly review some common evaluation metrics in binary classification that we will use later. The *sensitivity* of a classifier refers to the true positive rate, or the probability that a positive instance will be classified as positive. For a discriminant function η and threshold t , the sensitivity is $\mathbb{P}_1(\eta(X) > t)$, if we interpret a positive instance to be an observation from class 1. The *specificity* of a classifier refers to the true negative rate, or the probability

$\mathbb{P}_0(\eta(X) \leq t)$ that a negative instance will be classified as negative. Setting the threshold t to positive infinity classifies all instances to the negative class (class 0) and achieves 100% specificity but suffers from 0% sensitivity. As one decreases the decision threshold, the classifier's sensitivity improves at the expense of specificity, until it reaches the other extreme point where all instances are classified as positive, achieving 100% sensitivity and 0% specificity.

The intrinsic trade-off between sensitivity and specificity is often illustrated through the *receiver operating characteristic curve* (ROC curve). The ROC curve is the set of points traced by coordinates $(\mathbb{P}_0(\eta(X) > t), \mathbb{P}_1(\eta(X) > t))$, $-\infty \leq t \leq \infty$, connecting point $(0, 0)$ when $t \rightarrow \infty$ and $(1, 1)$ when $t \rightarrow -\infty$. The second coordinate is the sensitivity of the classifier, and the first is 1 minus the specificity, so higher levels of the ROC curve indicate better performance. We will see examples of ROC curves in Section 1.6.

The Neyman-Pearson lemma implies (see Lehmann and Romano [57], p.62) that the ROC curve for the log-odds classifier lies above the ROC curve for any other discriminant function. As discussed in Section 2.3 of Eguchi and Copas [24], the loss $C(\alpha, \beta)$ can be interpreted as the weighted area between the log-odds ROC curve and the ROC curve for the linear classifier determined by (α, β) ; the weight assigned to the gap between the curves at a score of u is $w(u)$. By minimizing $C(\alpha, \beta)$, we find the linear score $\alpha + \beta^\top x$ that is closest to the true log-odds function, in the sense of this weighted area. Different weight functions balance the sensitivity-specificity trade-offs differently. The weight $w(u)$ at large positive u emphasizes the high-specificity/low-sensitivity region of the ROC curve, and the weight $w(u)$ at large negative u emphasizes the high-sensitivity/low-specificity region of the ROC curve. We return to these ideas in Section 1.5.4 and in the application of Section 1.7.

1.2.2 Examples of Objective Functions

With $w(u) = w_0(u) = 1/(1 + e^u)$, and $C_U = C_V = 0$, we get

$$U(u) = \log \frac{e^u}{1 + e^u}, \quad V(u) = -\log \frac{1}{1 + e^u}, \quad (1.5)$$

and the loss (1.2) becomes

$$C(\alpha, \beta) = -\mathbf{E} \left[Y \log \frac{e^{\alpha + \beta^\top X}}{1 + e^{\alpha + \beta^\top X}} + (1 - Y) \log \frac{1}{1 + e^{\alpha + \beta^\top X}} \right].$$

Minimizing this expression (or more precisely its empirical counterpart) is equivalent to maximizing the likelihood function in ordinary logistic regression. That is, ordinary logistic regression is a special case of this family of objectives. The discriminant functions $x \mapsto \alpha + \beta^\top x$ and $x \mapsto \exp(\alpha + \beta^\top x)/(1 + \exp(\alpha + \beta^\top x))$ yield equivalent classification rules because each is a monotone transformation of the other.

Among other examples, we will also consider exponential weight functions,

$$w(u) = \lambda(1 - \lambda)e^{-\lambda u}, \quad U(u) = -(1 - \lambda)e^{-\lambda u}, \quad V(u) = \lambda e^{(1-\lambda)u},$$

with $\lambda \in (0, 1)$, for which the loss function becomes

$$C(\alpha, \beta) = \mathbf{E}[Y(1 - \lambda)e^{-\lambda(\alpha + \beta^\top X)} + (1 - Y)\lambda e^{(1-\lambda)(\alpha + \beta^\top X)}]. \quad (1.6)$$

As illustrated in Figure 1.1, the logistic weight w_0 is bounded whereas the exponential w is not. A larger $\lambda \in (0, 1)$ attaches a greater penalty to large negative values of $\alpha + \beta^\top x$ when $y = 1$, and a smaller $\lambda \in (0, 1)$ attaches a greater penalty to large positive values of $\alpha + \beta^\top x$ when $y = 0$. Informally, λ balances a trade-off between false negative and false positive probabilities. (Recall that we take draws from class 1 to be positive cases.) We will see that a larger λ is more sensitive to the distribution of the minority class. The symmetric case $\lambda = 1/2$ corresponds to the loss function behind the AdaBoost method of Freund and Schapire [33], as discussed in Section 4.1 of Friedman, Hastie, and Tibshirani [34] and Section 2.4 of Eguchi and Copas [24].

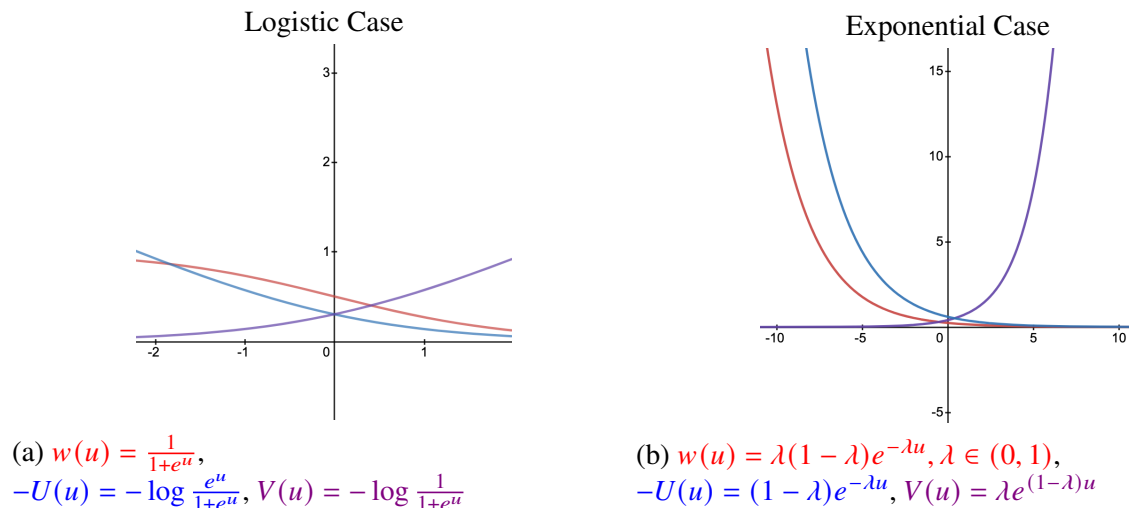


Figure 1.1: Examples of weight functions w and penalty functions U and V .

Our analysis also allows weight functions like

$$w(u) = \begin{cases} 1 - 2u, & u \leq 0; \\ (1 + u)^{-2}, & u > 0, \end{cases}$$

that are in between an exponential weight function and the bounded weight function underlying logistic regression in (1.5), in the sense that in this example $w(u)$ grows linearly as $u \rightarrow -\infty$. We provide precise conditions on w in Section 1.4.

1.2.3 Empirical Loss

In estimating (α, β) , we minimize an empirical version of the loss (1.2). Let x_1, \dots, x_n denote n observations from class 1, and let X_1, \dots, X_N denote N independently and identically distributed (i.i.d.) samples from class 0, with underlying distribution F_0 . Define

$$\bar{C}_N(\alpha, \beta) = \sum_{i=1}^n -U(\alpha + \beta^\top x_i) + \sum_{j=1}^N V(\alpha + \beta^\top X_j). \quad (1.7)$$

To go from (1.1) to (1.7), replace the expectations with sample means, replace π_1 with $n/(n+N)$ and π_0 with $N/(n+N)$, and multiply by $n+N$. For fixed N , let (α_N, β_N) minimize (1.7). We study

the behavior of (α_N, β_N) as $N \rightarrow \infty$, with n fixed.

1.3 Existence of a Minimizer

In this section, we provide conditions ensuring the existence of a unique, finite minimizer of (1.7) for all sufficiently large N , a.s. We first state some basic assumptions:

Condition 1 (Basic properties). *The weight function w is strictly positive on \mathbb{R} . The penalty functions U and V in (1.4) are well-defined and finite on all of \mathbb{R} .*

The conditions on U and V imply that $w(s) \rightarrow 0$ and $e^{-s}w(-s) \rightarrow 0$, as $s \rightarrow \infty$. For the loss to be convex, we want $w(u)$ to be decreasing and $e^u w(u)$ to be increasing, so we assume

Condition 2 (Convexity). *For all $u \in \mathbb{R}$, $w'(u) \leq 0$ and $w(u) + w'(u) > 0$.*

Although weak inequalities would suffice for convexity, we make the second inequality strict to ensure strict convexity.

For ordinary logistic regression, Silvapulle [73] provides necessary and sufficient conditions for the existence of maximum likelihood estimates. These conditions include a requirement of overlap between the two classes. Owen [65] proposes a stronger, but broadly applicable condition to prevent degeneracy. The key idea is that the empirical distribution of the minority class and the true distribution F_0 for the majority class should overlap at least to the extent that there is probability mass (with respect to F_0) along every possible direction away from the empirical mean of the minority class. We will use this condition as well. It relies on the following definition from Owen [65].

Definition 1.3.1 (Surrounding property). *The distribution F on \mathbb{R}^d surrounds the point x^* if for some $\epsilon > 0$, for some $\delta > 0$ and for all $\omega \in \Omega = \{\omega \in \mathbb{R}^d | \omega^\top \omega = 1\}$*

$$\int_{(x-x^*)^\top \omega > \epsilon} dF(x) > \delta. \tag{1.8}$$

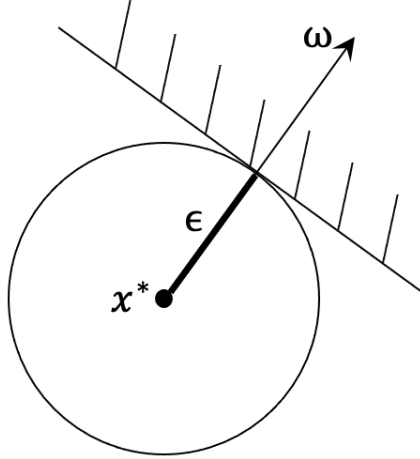


Figure 1.2: F surrounds point x^* if it assigns mass at least $\delta > 0$ to the shaded half-space, for every direction ω , for some $\epsilon > 0$.

Figure 1.2 illustrates the surrounding condition: F surrounds point $x^* \in \mathbb{R}$ if F has at least δ mass in *every* half-space that is ϵ away from x^* . This holds, for example, if F has a density that is bounded away from zero in a neighborhood of x^* .

Lemma 1.3.1 (Convex Objective). *Under Conditions 1–2, \bar{C}_N is a.s. convex. If in addition F_0 surrounds at least one point in \mathbb{R}^d , \bar{C}_N is strictly convex for all sufficiently large N , a.s.*

It turns out that if F_0 surrounds the minority class mean \bar{x} , then the loss function $\bar{C}_N(\alpha, \beta)$ in (1.7) has a unique finite minimizer, as established below.

Condition 3 (Surrounding minority mean). *F_0 surrounds \bar{x} with some parameters $\epsilon, \delta > 0$, where $\bar{x} = (x_1 + \dots + x_n)/n$ is the mean of the minority class.*

Lemma 1.3.2 (Existence). *Let $n \geq 1$ and $x_1, \dots, x_n \in \mathbb{R}^d$ be given. If Conditions 1–3 hold then the loss function $\bar{C}_N(\alpha, \beta)$ in (1.7) has a unique finite minimizer (α_N, β_N) for all sufficiently large N , a.s.*

Although the existence result Lemma 1.3.2 only requires that F_0 surround the minority class mean \bar{x} , for our main results we will assume that F_0 surrounds every minority class observation x_1, \dots, x_n , as stated in the following condition:

Condition 4 (Surrounding all minority observations). *With some parameters $\epsilon^o, \delta^o > 0$, F_0 surrounds every minority class observation x_1, \dots, x_n .*

1.4 Convergence Under Infinite Imbalance

This section states our main result, describing the behavior of the optimal (α_N, β_N) in the limit of infinite imbalance. Our analysis considers weight functions $w(u)$ based primarily on their properties for large negative values of u . We are particularly interested in distinguishing weight functions that grow exponentially or subexponentially as $u \rightarrow -\infty$, and to make this distinction precise we introduce additional conditions. After stating these conditions, we will show that they are satisfied by simple and easily interpreted examples, including all the examples in Section 1.2.2. We use the relation \sim to indicate that the ratio of two functions converges to 1.

Definition 1.4.1. *For a weight function w we define the following conditions.*

- *Left-tail condition: $w(u) \sim e^{-\lambda u} h(u)$, as $u \rightarrow -\infty$, for some $\lambda \in [0, 1)$ and some differentiable $h : \mathbb{R} \mapsto \mathbb{R}_+$ that satisfies*

(i) $h(u) > 0$, for all $u \in \mathbb{R}$;

(ii) $-(1 - \lambda)h(u) < h'(u) \leq \lambda h(u)$, for all $u \in \mathbb{R}$;

(iii) $\liminf_{u \rightarrow -\infty} h'(u)/h(u) \geq 0$;

(iv) *there exists some $C > 0$ and $\xi > 0$ such that for any $\epsilon > 0$, there exists some $u_0 < 0$ such that for any $u, u + s \leq u_0$, $s \in \mathbb{R}$, we have*

$$\left| \frac{h(u+s)}{h(u)} - 1 \right| \leq \epsilon \max\{C, e^{\xi|s|}\}. \quad (1.9)$$

- *Right-tail condition: If w is unbounded then $\limsup_{u \rightarrow \infty} \frac{V(u)}{e^{u w(u)}} < \infty$.*

The first condition on h is natural since we require $w(u) > 0$; the second condition is needed for Condition 2. The third condition implies that h grows subexponentially in the following sense:

for any $\epsilon_h > 0$, there exists some $u_h < 0$ such that for all $u \leq u_h$, $h(u) \leq C e^{-\epsilon_h u}$. An immediate consequence of condition (iv) is that for any $s \in \mathbb{R}$,

$$\left| \frac{h(u+s)}{h(u)} - 1 \right| \rightarrow 0$$

as $u \rightarrow -\infty$. Condition (iv) controls the speed of this convergence as s varies.

Our focus in Definition 1.4.1 is on the left-tail behavior of the weight function w because under infinite imbalance we expect $\alpha_N \rightarrow -\infty$; this limit is suggested by letting $\pi_1 \rightarrow 0$ in (1.3). The additional condition on the right tail for unbounded w helps ensure that α_N indeed diverges and β_N remains bounded. The right-tail condition is satisfied if $e^u w(u)$ is log-convex on some interval $[u_0, \infty)$, and this condition is satisfied if $w(u) = C e^{-\lambda u}$, $\lambda \in (0, 1)$, or $w(u) = C u^{-k}$, $k > 0$, for large u ; see Section A.2.5 of the appendix for a more general result.

The left-tail condition in Definition 1.4.1 is satisfied by the following examples.

Lemma 1.4.1. *The following weight functions satisfy the left-tail condition in Definition 1.4.1:*

- $w(u) \sim C$ as $u \rightarrow -\infty$, for some $C > 0$;
- $w(u) \sim C |u|^k$ as $u \rightarrow -\infty$, for some $C > 0$, $k \geq 0$;
- $w(u) \sim C e^{-\lambda u} |u|^k$ as $u \rightarrow -\infty$, for some $C > 0$, $\lambda \in (0, 1)$, and $k \geq 0$.

The first two examples in Lemma 1.4.1 correspond to taking $\lambda = 0$ with $h(u) = C$ and $h(u) = |u|^k$, respectively, in Definition 1.4.1; the last example corresponds to taking $\lambda \in (0, 1)$ and $h(u) = |u|^k$. For logistic regression (1.5), the weight function is bounded, so the first case in Lemma 1.4.1 applies, and the right-tail condition in Definition 1.4.1 is not needed.

In our main result, Theorem 1.4.2, the exponent λ in Definition 1.4.1 determines the behavior of linear classifiers under infinite imbalance. It will be useful to distinguish the following two cases:

- **asymptotically subexponential case:** weight functions with $\lambda = 0$.

- **asymptotically exponential case:** weight functions with $\lambda \in (0, 1)$.

The first two examples in Lemma 1.4.1 are in the subexponential category, and the last example is in the exponential category. For these categories to be meaningful, we need to ensure that λ is well-defined, which we do through the following result.

Proposition 1.4.1. *Suppose w satisfies Definition 1.4.1 with $w(u) \sim e^{-\lambda u} h(u)$ as $u \rightarrow -\infty$. Suppose w also satisfies Definition 1.4.1 with $w(u) \sim e^{-\tilde{\lambda} u} \tilde{h}(u)$ as $u \rightarrow -\infty$. Then $\lambda = \tilde{\lambda}$ and $h(u) \sim \tilde{h}(u)$ as $u \rightarrow -\infty$.*

For our main result, we require the following tail condition on the majority class distribution F_0 :

Condition 5 (Tail condition). *For some $r > \max\{1, 1 - \lambda + \xi\}/\gamma$,*

$$\int e^{r\|x\|} dF_0(x) < \infty, \quad (1.10)$$

where $\gamma = (1 - \lambda)\epsilon\delta$, with $\epsilon, \delta > 0$ the surrounding parameters in Condition 3, and where λ and ξ are the parameters from Definition 1.4.1.

This condition is satisfied by distributions F_0 with bounded support, with Gaussian tails, or with tails that decay exponentially at a rate faster than r . A larger ξ imposes a weaker condition in (1.9) but a stronger condition in Condition 5. For the second and third cases in Lemma 1.4.1 we can take $\xi > 0$ arbitrarily small, and for bounded w we can take $\xi = 0$. We can now show that β_N converges a.s. under infinite imbalance and we can identify its limit.

Theorem 1.4.2. *Suppose Conditions 1–5 hold and w satisfies Definition 1.4.1 with $w(u) \sim e^{-\lambda u} h(u)$ as $u \rightarrow -\infty$. Then the minimizer (α_N, β_N) of \bar{C}_N in (1.7) satisfies $\alpha_N \rightarrow -\infty$ and $\beta_N \rightarrow \beta_*$, a.s., where β_* is the unique solution to*

$$\frac{\int x e^{(1-\lambda)\beta_*^\top x} dF_0(x)}{\int e^{(1-\lambda)\beta_*^\top x} dF_0(x)} = \frac{\sum_{i=1}^n x_i e^{-\lambda\beta_*^\top x_i}}{\sum_{i=1}^n e^{-\lambda\beta_*^\top x_i}}. \quad (1.11)$$

In particular, when $\lambda = 0$, β_* is the unique solution to

$$\frac{\int x e^{\beta_*^\top x} dF_0(x)}{\int e^{\beta_*^\top x} dF_0(x)} = \bar{x}, \quad (1.12)$$

where $\bar{x} = n^{-1} \sum_{i=1}^n x_i$ is the minority class mean.

The second part of Theorem 1.4.2 extends Owen's [65] result by showing that a broad class of linear discriminant functions are asymptotically equivalent to logistic regression under infinite imbalance: all subexponential ($\lambda = 0$) weight functions share the same limiting coefficient vector β_* in (1.12). In contrast, asymptotically exponential weight functions allow a wider range of limits (1.11), dependent on the exponent λ . The limiting β_N in (1.12) depends on the minority class observations only through their mean \bar{x} . With $\lambda \in (0, 1)$, (1.11) shows that the limiting β_N will depend on the full empirical distribution of x_1, \dots, x_n .

We can interpret the left side of (1.12) as the mean of the distribution obtained after applying an exponential tilt or reweighting to F_0 . Equation (1.12) then says that β_* tilts F_0 to the mean of the minority class. Equation (1.11) tilts both F_0 and the empirical distribution of x_1, \dots, x_n to a common mean. We build on this interpretation in the next section.

Even in the case of logistic regression, Theorem 1.4.2 extends Owen's [65] result because our (α_N, β_N) minimize the empirical loss function (1.7) rather than an approximate loss function that replaces the empirical distribution over X_1, \dots, X_N with the population distribution F_0 .

Although our focus is on the behavior of β_N , as a byproduct of our analysis we show that α_N diverges at least as fast as $-\log N$. See Corollary A.2.1 in the appendix.

To see that not all choices of w lead to similar limits, we briefly consider a case suggested in Section 2.4 of Eguchi and Copas [24], in which the weight function is replaced by a measure. Taking w to be a delta function with unit mass at a point u_0 , they arrive at the objective

$$C(\alpha, \beta) = \pi_1 \mathbf{P}_1(\alpha + \beta^\top X \leq u_0) + \pi_0 e^{u_0} \mathbf{P}_0(\alpha + \beta^\top X > u_0). \quad (1.13)$$

Through appropriate choice of u_0 , $C(\alpha, \beta)$ can be interpreted as balancing two types of misclassi-

fication costs. However, we show in Appendix A.6 that the resulting discriminant function degenerates under imbalance, in the sense that for all sufficiently large N , $\beta_N = 0$ a.s. and α_N can be any value less than or equal to u_0 .

Li, Bellotti, and Adams [58] consider regularized logistic regression with an L_1 or L_2 penalty on β . They show that the optimal β_N converges to zero under infinite imbalance. In this sense, the regularized discriminant function degenerates under imbalance.

1.5 Robustness Interpretation of β_*

We now turn to the interpretation of the limits β_* defined by (1.12) and (1.11). We will show that these limits reflect robustness properties, in the sense that the coefficients are optimized against certain worst-case errors or combinations of errors. These robustness properties reflect implicit choices of conservatism towards different types of errors.

1.5.1 Asymptotically Subexponential Weight Functions

The robustness interpretation is easiest to formulate in the limit (1.12) for the subexponential case, which includes logistic regression. For $\beta \in \mathbb{R}^d$, define the cumulant generating function of F_0 by setting

$$\psi(\beta) = \log \int e^{\beta^\top x} dF_0(x), \quad (1.14)$$

and let $B_\psi = \{\beta : \psi(\beta) < \infty\}$. Define the exponential family of distributions F_β , $\beta \in B_\psi$, on \mathbb{R}^d by setting

$$dF_\beta(x) = e^{\beta^\top x - \psi(\beta)} dF_0(x).$$

Part of the content of Theorem 1.4.2 is that $\beta_* \in B_\psi$; the normalizing factor $e^{\psi(\beta_*)}$ is the denominator in (1.12). Write F_* for the special case F_{β_*} . Equation (1.12) tells us that

$$\int x dF_*(x) = \int x e^{\beta_*^\top x - \psi(\beta_*)} dF_0(x) = \bar{x}, \quad (1.15)$$

so the mean of F_* is \bar{x} . We can also write (1.15) as $\nabla\psi(\beta_*) = \bar{x}$. The robustness interpretation comes from identifying F_* as a worst-case alternative and identifying β_* as the optimal classifier for this worst-case alternative.

For distributions G and F on \mathbb{R}^d , define the relative entropy (or Kullback-Leibler divergence),

$$D(G\|F) = \int \log \frac{dG}{dF} dG, \quad (1.16)$$

with $D(G\|F) = \infty$ if the support of G is not contained within the support of F . Relative entropy is always non-negative, and it is zero if and only if G and F coincide. It is not symmetric, but $D(G\|F)$ can be interpreted as a measure of the “distance” of G from F . If $D(G\|F)$ is small then G is close to F , making the problem of discriminating between the two distributions difficult. Kullback and Leibler [54] interpret $D(G\|F)$ as the mean information for discriminating between G and F per observation from G .

Let $\mathcal{M}_{\bar{x}}$ be the set of probability distributions on \mathbb{R}^d with mean \bar{x} . We know from (1.15) that $F_* \in \mathcal{M}_{\bar{x}}$. In fact, of all elements of $\mathcal{M}_{\bar{x}}$, the one “closest” to F_0 with respect to relative entropy is F_* , as the following result shows. This result is a special case of Corollary 3.1 of Csiszar [19].

Lemma 1.5.1. *The problem*

$$\min_G D(G\|F_0) \text{ subject to } \int x dG(x) = \bar{x}, \quad (1.17)$$

where G is a probability distribution on \mathbb{R}^d , is solved by $G = F_*$.

Lemma 1.5.1 leads to a robustness property of β_* . Let G be any distribution on \mathbb{R}^d with the same support as F_0 . As discussed at the end of Section 1.2.1, the Neyman-Pearson lemma implies that the ROC curve defined by the log likelihood ratio $\eta(x) = \log dG(x)/dF_0(x)$ lies above the ROC curve for any other discriminant function.

In the case of the distribution F_* defined by β_* , the log likelihood ratio is given by

$$\eta_*(x) = \log \frac{dF_*}{dF_0}(x) = \beta_*^\top x - \psi(\beta_*). \quad (1.18)$$

The linear discriminant function $\beta_*^\top x$ coincides with the optimal classifier η_* . The constant $\psi(\beta_*)$ shifts the threshold t , but the two functions trace the same ROC curve. That is, the limit in (1.12) picks the optimal classifier for discriminating between F_0 and F_* .

Combining this lemma with (1.18), we arrive at the following conclusion: The limiting coefficient β_* in (1.12) provides the optimal classifier for the worst-case alternative to the majority class distribution F_0 , among all distributions with the same mean \bar{x} as the observations from the minority class. The distribution F_* presents the worst case because it is hardest to distinguish from F_0 , in the sense of Lemma 1.5.1. This robustness property does not necessarily translate to better performance; optimizing against a worst case can be overly conservative if the true distribution of the minority class distribution is very different from F_* .

The definition in (1.16) suggests an interpretation of the conservatism implicit in the focus on β_* and F_* . In making $D(G||F_0)$ smaller, we are, roughly speaking, making the optimal discriminant function (the log likelihood ratio $\log(dG/dF_0)$) smaller at observations that have higher probability under G . In other words, in focusing on F_* we are focusing on a distribution whose optimal classifier (relative to F_0) will have low sensitivity. This perspective suggests that the limiting coefficient β_* in (1.12) — the optimal classifier for F_* — is implicitly conservative in classifying to the minority class and should perform better at low-sensitivity (high-specificity) thresholds than at high-sensitivity thresholds. This interpretation will be supported by our discussion of exponential weight functions in Section 1.5.4 and the numerical results of Sections 1.6 and 1.7, because the subexponential case considers only $\lambda = 0$ whereas the exponential case considers all $\lambda \in (0, 1)$.

1.5.2 Infinite Upsampling

The problem of imbalance is sometimes dealt with in practice through upsampling — creating artificial data from the minority class. We now show that for the class of linear discriminant functions defined by minimizing (1.1) with asymptotically subexponential w (including logistic regression), the estimate of β obtained in the infinite imbalance limit coincides with the estimate obtained from a specific choice of upsampling distribution, namely F_* .

Suppose, then, that we have infinitely upsampled the minority class so that a fraction π_1 of our data is drawn from F_* . Suppose we also have infinitely many observations, a fraction $\pi_0 = 1 - \pi_1$ of the total, from F_0 . The loss function then takes the form (1.1). Differentiating (1.1) with respect to β and recalling that $U'(s) = w(s)$, $V'(s) = e^s w(s)$, we get the first-order condition

$$-\pi_1 \mathbb{E}_1 [w(\alpha + \beta^\top X)X] + \pi_0 \mathbb{E}_0 [e^{\alpha + \beta^\top X} w(\alpha + \beta^\top X)X] = 0,$$

where \mathbb{E}_1 is now expectation with respect to F_* . Using the likelihood ratio dF_*/dF_0 , as in (1.18), we can write the first term as an expectation with respect to F_0 to get

$$-\pi_1 \mathbb{E}_0 [e^{\beta_*^\top X - \psi(\beta_*)} w(\alpha + \beta^\top X)X] + \pi_0 \mathbb{E}_0 [e^{\alpha + \beta^\top X} w(\alpha + \beta^\top X)X] = 0.$$

This equation is solved by taking

$$\beta = \beta_*, \quad e^\alpha = (\pi_1/\pi_0) e^{-\psi(\beta_*)}.$$

The limiting coefficient vector β_* in (1.12) is thus precisely what one would obtain through infinite upsampling of the minority class using F_* when we have infinitely many observations from the majority class. The relative degree of upsampling, as reflected in π_1/π_0 , affects α but not β .

The choice of F_* as an upsampling distribution is not arbitrary. We know from Section 1.5.1 that F_* is the distribution “closest” to F_0 among all distributions with the mean \bar{x} in the data from the minority class. We have argued that the optimal classifier β_* corresponding to F_* is implicitly

conservative in classifying into the minority class. In this sense, upsampling according to F_* is conservative if one is particularly concerned about classifying with high specificity.

1.5.3 The Gaussian Case

The Gaussian case provides some convenient simplifications and helps illustrate the transition from the subexponential limit in (1.12) to the exponential limit in (1.11). We begin by reviewing some properties of the Gaussian setting.

Suppose that F_0 is the multivariate normal distribution $N(\mu_0, \Sigma_0)$ on \mathbb{R}^d , with mean μ_0 and full-rank covariance matrix Σ_0 . The cumulant generating function (1.14) becomes

$$\psi(\beta) = \log \int e^{\beta^\top x} dF_0(x) = \beta^\top \mu_0 + \frac{1}{2} \beta^\top \Sigma_0 \beta,$$

for all $\beta \in \mathbb{R}^d$; this is a special case of Brown [14], Example 1.14. The tilted distribution defined by

$$e^{\beta^\top x - \psi(\beta)} dF_0(x) \tag{1.19}$$

is multivariate normal with mean $\mu_0 + \Sigma_0 \beta$. In other words, the ratio of the $N(\mu_0 + \Sigma_0 \beta, \Sigma_0)$ density to the $N(\mu_0, \Sigma_0)$ density is the exponential factor in (1.19). Every point in \mathbb{R}^d can be expressed as $\mu_0 + \Sigma_0 \beta$, for some $\beta \in \mathbb{R}^d$, so as β ranges over \mathbb{R}^d , (1.19) ranges over all multivariate normal distributions with covariance matrix Σ_0 . That is, the multivariate normal distributions with covariance matrix Σ_0 and arbitrary mean form an exponential family with parameter β .

Setting aside the issue of imbalance for a moment, let F_1 be the multivariate normal distribution $N(\mu_1, \Sigma_0)$, for some mean vector μ_1 . Let β_* be as in (1.12), but with \bar{x} replaced by μ_1 . Then β_* is the parameter that makes the mean of (1.19) equal to μ_1 , or $\mu_1 = \mu_0 + \Sigma_0 \beta_*$, and therefore

$$\beta_* = \Sigma_0^{-1}(\mu_1 - \mu_0). \tag{1.20}$$

We recognize (1.20) as the coefficient vector in classical linear discriminant analysis. (See, e.g.,

Theorem 6.4.1 of Anderson [3].) Linear discriminant analysis classifies an observation x to class 1 or 0 depending on whether $\beta_*^\top x$ is larger or smaller than some threshold. We ignore the intercept in the discriminant function because it can be absorbed into the choice of threshold. Thus, the coefficient vector defined by (1.12) when F_0 is multivariate normal coincides with the coefficient vector in linear discriminant analysis that applies when F_0 and F_1 are multivariate normal with a common covariance matrix. That is, when F_0 is multivariate normal, the infinite imbalance limit (1.12) chooses β_* as if F_1 were multivariate normal with the same covariance matrix as F_0 . This is a special case of the upsampling interpretation in Section 1.5.2 because the β_* we get from (1.12) is the same coefficient vector we would get if we upsampled the minority class to the distribution $N(\bar{x}, \Sigma_0)$ and then applied logistic regression.

In the left panel of Figure 1.3, the ellipses show probability contours for two bivariate normal distributions $N(\mu_i, \Sigma_i)$, $i = 0, 1$,

$$\mu_0 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \quad \Sigma_0 = \begin{pmatrix} 1.4^2 & 1.4 \cdot 2.2 \cdot 0.6 \\ 1.4 \cdot 2.2 \cdot 0.6 & 2.2^2 \end{pmatrix}, \quad \mu_1 = \begin{pmatrix} 6 \\ 8 \end{pmatrix}, \quad \Sigma_1 = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix},$$

Class 0 is indicated in blue, and class 1 in red. The blue line illustrates a set of points x with a constant value of $\beta_*^\top x$, where the vector $\beta_* = \Sigma_0^{-1}(\mu_1 - \mu_0)$ is computed from the perspective of F_0 , using the covariance matrix Σ_0 . The intercept of the line is arbitrary; the classification rule defined by β_* should be thought of as the set of lines parallel to the blue line, with different intercepts corresponding to different classification thresholds. The only property of F_1 used in calculating β_* is the mean μ_1 , just as in (1.12). With different covariance matrices for the two classes, linear discriminant analysis is still useful but it loses optimality properties (Anderson [3], Section 6.10.2); the choice of β_* in the infinitely imbalanced limit (1.12) fails to capture the difference in covariance matrices and is akin to assuming that F_1 has the same covariance as F_0 . The red line similarly illustrates a classification boundary, but now using $\beta_* = \Sigma_1^{-1}(\mu_0 - \mu_1)$ calculated with the roles of F_0 and F_1 reversed.

Now consider the analog of (1.11) in which the distributions on the two sides (F_0 on the left,

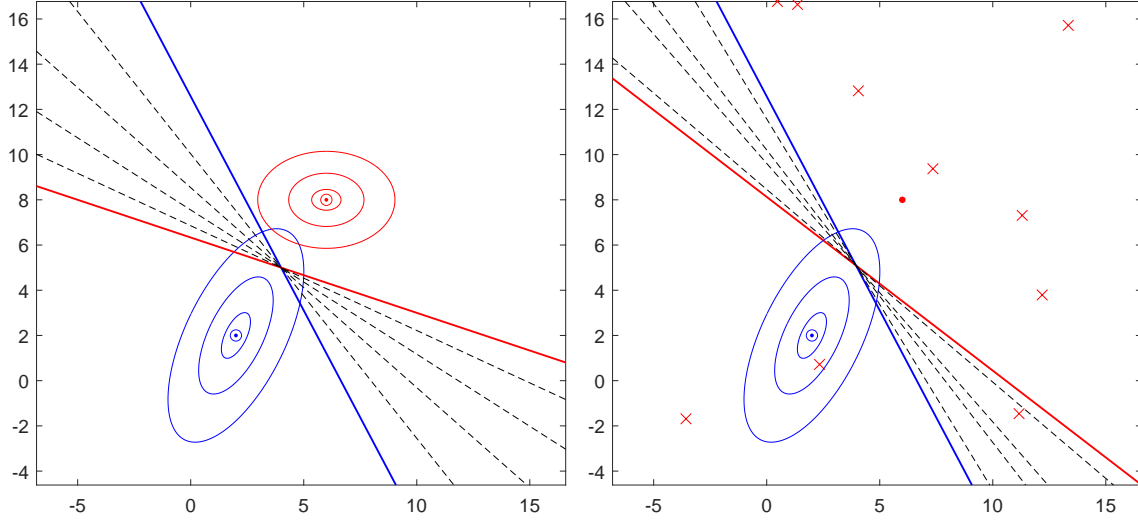


Figure 1.3: Classification boundaries with unequal covariance matrices. Dashed lines show the effect of varying λ .

the empirical distribution on x_1, \dots, x_n on the right) are replaced with the normal distributions F_0 and F_1 . In this case (1.11) reduces to

$$\mu_0 + (1 - \lambda)\Sigma_0\beta_* = \mu_1 - \lambda\Sigma_1\beta_*;$$

i.e.,

$$\beta_* = (\lambda\Sigma_1 + (1 - \lambda)\Sigma_0)^{-1}(\mu_1 - \mu_0). \quad (1.21)$$

The exponent λ in the asymptotically exponential weight function w balances the relative influence of the two distributions in setting the slope of the classifier. The dashed lines in the left panel of Figure 1.3 show the effect of varying λ between 0 and 1.

In classical linear discriminant analysis, Anderson [3], p.247, shows that a coefficient vector of the form (1.21) solves a minimax problem of balancing the two types of misclassification errors that can occur in binary classification. Based on the discussion in Section 1.5.1, we interpret the extremes $\lambda = 0$ and $\lambda = 1$ as two forms of conservatism, from the perspectives F_0 and F_1 , respectively.

Now replace the normal distribution for F_1 with the empirical distribution on x_1, \dots, x_n , while

keeping F_0 normal. In this case (1.11) becomes

$$\mu_0 + (1 - \lambda)\Sigma_0\beta_* = \frac{\sum_{i=1}^n x_i e^{-\lambda\beta_*^\top x_i}}{\sum_{i=1}^n e^{-\lambda\beta_*^\top x_i}},$$

which can be solved numerically. The crosses in the right panel of Figure 1.3 represent x_1, \dots, x_n , $n = 10$. Their mean is at $(6, 8)$, just as it is for the red normal distribution in the left panel. The slopes of the solid blue lines are therefore the same in the two panels: the discriminant computed from the blue distribution ($\lambda = 0$) depends on the red distribution only through its mean. At $\lambda = 0$, the infinite imbalance limit chooses β_* as if F_1 were the multivariate normal distribution $N(\bar{x}, \Sigma_0)$. The solid red line ($\lambda = 1$) similarly depends on F_0 only through μ_0 .

The dashed black lines correspond to intermediate values of λ . These make use of the distributions of both classes, not just their means, indicating a potential advantage of the exponential objective over logistic regression and other asymptotically subexponential cases. Although it is not evident from the figure, the slopes of the lines need not change monotonically with λ , nor is the slope at an intermediate λ necessarily between the slopes of the blue and red lines.

The explicit expressions in (1.20) and (1.21) are potentially useful in non-Gaussian settings as starting values for numerical calculation of optimal coefficient vectors using estimated means and covariance matrices. Related suggestions are made in Owen [65] and Deo and Juneja [20].

1.5.4 Asymptotically Exponential Weight Functions

We have seen that in the limit (1.12) for asymptotically subexponential weight functions, the linear classifier under infinite imbalance is implicitly optimized for low sensitivity and high specificity. In this section, we extend this interpretation to the asymptotically exponential case (1.11) with $\lambda \in (0, 1)$, and we argue that the exponent λ balances the classifier's emphasis along the sensitivity-specificity trade-off.

For distributions F_0, F_1 on \mathbb{R}^d , define the cumulant generating functions ψ_i , $i = 0, 1$, as in

(1.14), with domains B_i , $i = 0, 1$, and define the exponential families of distributions

$$dF_{i,\beta}(x) = e^{\beta^\top x - \psi_i(\beta)} dF_i(x), \quad \beta \in B_i.$$

Proposition 1.5.1. *For $\lambda \in (0, 1)$, suppose there is a $\beta_* \in B_0 \cap B_1$ for which*

$$\nabla\psi_0((1 - \lambda)\beta_*) = \nabla\psi_1(-\lambda\beta_*). \quad (1.22)$$

Then the problem

$$\min_{G_0, G_1} \lambda D(G_0 \| F_0) + (1 - \lambda) D(G_1 \| F_1) \text{ subject to } \int x dG_0(x) = \int x dG_1(x), \quad (1.23)$$

where G_0 and G_1 are distributions on \mathbb{R}^d , is solved by $G_0 = F_{0, (1-\lambda)\beta_}$ and $G_1 = F_{1, -\lambda\beta_*}$.*

Equation (1.22) generalizes (1.11); it reduces to (1.11) when F_1 is the empirical distribution on x_1, \dots, x_n (and the existence of β_* solving (1.11) is proved as part of Theorem 1.4.2). The limiting coefficient vector in (1.11) can therefore be interpreted as the result of minimizing the objective in (1.23), with this substitution for F_1 . Proposition 1.5.1 then generalizes Lemma 1.5.1. Indeed, as λ approaches 0, $D(G_1 \| F_1)$ dominates the objective function, so the minimizer G_1 approaches F_1 . In particular, the mean of G_1 approaches the mean of F_1 . Then in solving for G_0 , we are solving $\min_{G_0} D(G_0 \| F_0)$ subject to $\int x dG_0(x) = \int x dF_1(x)$, reducing to the minimization problem to Lemma 1.5.1. Whereas the limit of the asymptotically subexponential case implicitly focuses on the worst-case distribution from the perspective of F_0 , the objective in (1.23) balances the worst case as seen from both F_0 and F_1 . In doing so, it balances the focus on the high-specificity and high-sensitivity regions.

To make this balance more explicit, recall from the discussion at the end of Section 1.2.1 that the weight function $w(u)$ can be interpreted as a penalty on the difference between the ROC curves for the optimal discriminant function and an approximating linear discriminant function. With $w(u)$ proportional to $e^{-\lambda u}$, we thus expect a better approximation at large negative u (the

high-sensitivity region of the ROC curve) when λ is close to 1, and a better approximation at large positive u (the high-specificity region of the ROC curve) when λ is close to 0:

$$\lambda \approx 0 \Rightarrow \text{emphasizes high-specificity region;}$$

$$\lambda \approx 1 \Rightarrow \text{emphasizes high-sensitivity region.}$$

This pattern is what we find in the experiments of Sections 1.6 and 1.7. In Appendix A.4, we provide further discussion on the connection between (1.23) and the original classification problem by generalizing the minimization of (1.6) over possibly nonlinear discriminant functions.

The symmetric case $\lambda = 1/2$ admits a further interpretation. For any discriminant function, the area under the curve measure AUC, discussed further in the next section, equals the probability that a draw X_1 from F_1 scores higher than an independent draw X_0 from F_0 . Thus, for a linear discriminant function $x \mapsto \beta^\top x$, Markov's inequality yields

$$\text{AUC} = 1 - \mathbf{P}(\beta^\top X_0 \geq \beta^\top X_1) \geq 1 - \mathbf{E}[e^{\beta^\top (X_0 - X_1)/2}] \geq 1 - e^{\psi_0(\frac{1}{2}\beta) + \psi_1(-\frac{1}{2}\beta)}.$$

A cumulant generating function is convex on its domain. Maximizing the lower bound over β therefore leads to the first-order condition $\nabla\psi_0(\beta/2) = \nabla\psi_1(-\beta/2)$, which is the condition in (1.22) with $\lambda = 1/2$. This observation is consistent with the idea that taking $\lambda = 1/2$ balances overall performance without emphasizing either specificity or sensitivity over the other.

1.6 Numerical Examples

We use simulations to examine the convergence of β_N and to illustrate properties of the classifiers derived using various choices of the penalty function w .

1.6.1 Convergence Simulations

For simplicity, we examine convergence in a one-dimensional example. We have just two observations from the minority class, $x_1 = 0$ and $x_2 = 1$. For the majority class, we use N i.i.d. samples from the standard normal distribution, $N(0, 1)$. We compare results at several values of N .

Table 1.1 reports the mean and the standard error of the coefficients (α_N, β_N) for ordinary logistic regression, an exponential objective, and an asymptotically linear objective, averaging over 1,000 independent runs. For the exponential objective we consider $\lambda = 0.5$, and for the asymptotically linear objective we use $w(u) = -2u + 1$, for $u \leq 0$, and $w(u) = (u + 1)^{-2}$ for $u > 0$, as in Section 1.2.2.

We solve for the optimal coefficients using the *minimize* function in the *scipy.optimize* package with the *Newton-CG* optimization method. The last row of Table 1.1 reports the limiting value β_* determined by Theorem 1.4.2, calculated by solving (1.11).

N	Logistic w		$\lambda = 0.5$		Linear w	
	α_N	β_N	α_N	β_N	α_N	β_N
10	-1.75 (0.047)	0.86 (0.272)	-1.85 (0.052)	1.11 (0.372)	-1.77 (0.062)	1.17 (0.432)
100	-4.04 (0.002)	0.53 (0.014)	-4.17 (0.004)	0.82 (0.020)	-4.08 (0.003)	0.63 (0.016)
1,000	-6.34 (0.000)	0.50 (0.001)	-6.48 (0.000)	0.80 (0.002)	-6.37 (0.000)	0.57 (0.001)
10,000	-8.64 (0.000)	0.50 (0.000)	-8.78 (0.000)	0.80 (0.000)	-8.66 (0.000)	0.55 (0.000)
100,000	-10.95 (0.000)	0.50 (0.000)	-11.08 (0.000)	0.80 (0.000)	-10.96 (0.000)	0.54 (0.000)
True β_*		0.50		0.80		0.50

Table 1.1: Convergence of coefficients as the sample size N of the majority class grows. Numbers in parentheses are standard errors.

In all cases, we see that $\alpha_N \rightarrow -\infty$ at rate $\log N$, consistent with the findings in Corollary A.2.1 of the appendix. For logistic regression and the exponential objective, β_N is close to β_* at $N = 1,000$, when the data is only 0.2% imbalanced, and we have observed a similar convergence rate for other values of λ . When the left tail of the weight function diverges linearly, we know from Theorem 1.4.2 that β_N approaches the same limit β_* as in the case of logistic regression, but the results in the table indicate that the convergence is much slower. We have observed the same behavior with other weight functions whose left tail diverges at a polynomial rate.

1.6.2 High-Sensitivity and High-Specificity Regions

We turn next to a comparison of logistic regression with exponential objectives at various values of the exponent λ . As we discussed in Sections 1.2.2 and 1.5.4, we expect the value of λ to control the relative performance of a classifier as measured by sensitivity or specificity.

We consider a two-dimensional example in which F_0 is the bivariate standard normal distribution. Samples from the minority class are drawn from a mixture of two normals: we have a sample of $n = 500$, of which 10% are drawn from $N(\mu_{1,1}, \Sigma_{1,1})$ and 90% from $N(\mu_{1,2}, \Sigma_{1,2})$, with

$$\mu_{1,1} = \begin{pmatrix} 0 \\ 2 \end{pmatrix}, \Sigma_{1,1} = \begin{pmatrix} 0.3 & 0 \\ 0 & 0.3 \end{pmatrix}, \quad \mu_{1,2} = \begin{pmatrix} 2.3 \\ 2.3 \end{pmatrix}, \Sigma_{1,2} = \begin{pmatrix} 0.2 & 0 \\ 0 & 0.2 \end{pmatrix}.$$

We will compare results at various values of the sample size N for the majority class.

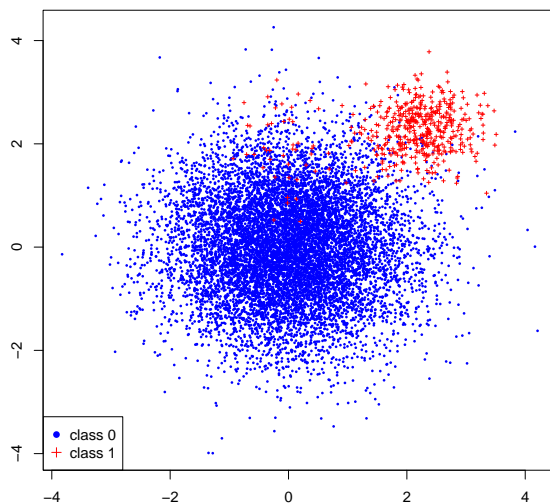


Figure 1.4: Two-Dimensional Example

Figure 1.4 shows points drawn from the two classes, with $N = 10,000$ and $n = 500$. The example is designed so that 90% of the minority class comes from a distribution that is easily distinguishable from the majority class, but the remaining 10% makes the classification task challenging.

Recall from Section 1.2.1 that the ROC curve for a linear discriminant function $\beta^\top x$ is a plot of the true positive rate $P_1(\beta^\top X > t)$ against the false positive rate $P_0(\beta^\top X > t)$, for all $t \in \mathbb{R}$. The area under the ROC, abbreviated AUC, provides an overall summary of the performance of the classifier, but we are more interested in comparing performance at high levels of sensitivity (high true positive rates) and high levels of specificity (low false positive rates). We therefore make comparisons based on partial AUC (pAUC) measures, as introduced in McClish [62], for the regions of interest.

The calculation of a specificity-oriented pAUC measure is illustrated in the left panel of Figure 1.5. In this example, we focus on the area under the curve between 0 and FP_1 (call that “Area”) and then normalize it to fall between 1/2 and 1 through the transformation

$$pAUC = \frac{1}{2} \left(1 + \frac{\text{Area} - \min}{\max - \min} \right) = \frac{1}{2} \left(1 + \frac{\text{Area} - FP_1^2/2}{FP_1 - FP_1^2/2} \right).$$

Here, \max is the area FP_1 of the shaded rectangle, and \min is the area $FP_1^2/2$ of the triangular portion of the rectangle below the diagonal. An ideal classifier over the interval from 0 to FP_1 would have a pAUC of 1, whereas a random assignment of observations to classes would have a pAUC of 1/2. To focus on high specificity, we consider values of FP_1 decreasing from 0.10 to 0, which corresponds to a true negative rate (TNR) increasing from 0.90 to 1.

The calculation of a sensitivity-oriented pAUC measure on the right side of Figure 1.5 works similarly. To focus on high sensitivity, we take a rectangle along the top end of the unit square. The lower boundary of that rectangle is defined by a true positive rate (TPR) that we initially set equal to 0.90 and then increase toward 1. In the normalization of the area for this case, \min is the area to the right of the diagonal. We use the R package *pROC* (Robin et al. [70]) to facilitate the calculation and plotting of pAUC values.

Figure 1.6 compares pAUC values for logistic and exponential classifiers with a sample size of $N = 10,000$ for the majority class. Panel (a) plots pAUCs in the high-sensitivity region, with the true positive rate $P_1(\beta^\top X > t)$ increasing from 0.90 toward 1. Among the exponential classifiers,

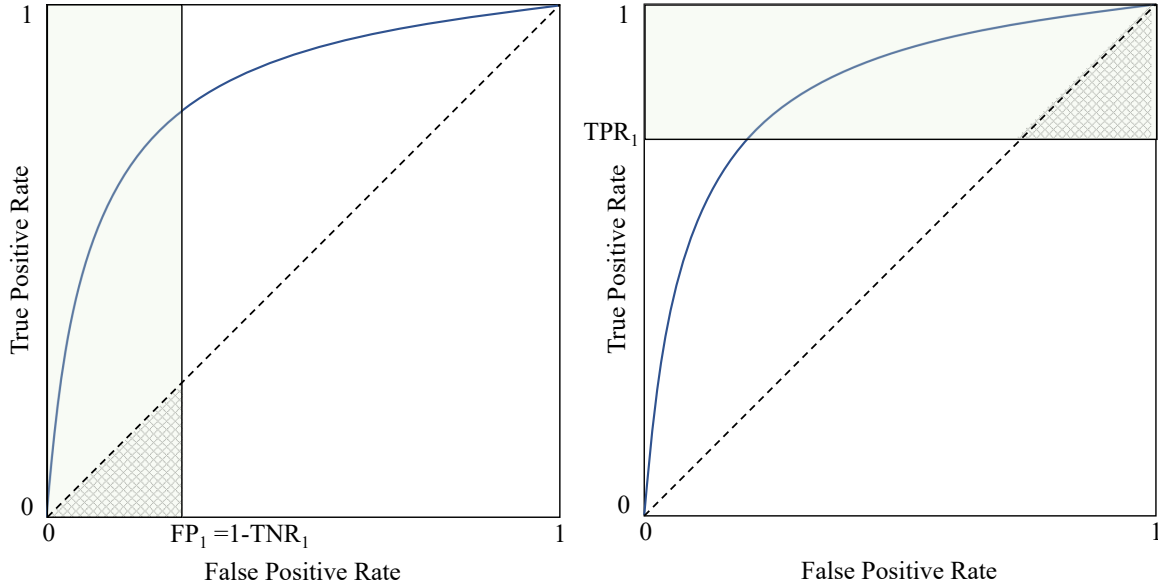


Figure 1.5: Illustration of a specificity-oriented (left) and sensitivity-oriented (right) pAUC.

we see that, at high levels of sensitivity, the classifier with $\lambda = 0.9$ outperforms the classifier with $\lambda = 0.5$, which outperforms the classifier with $\lambda = 0.1$. Panel (b) focuses on the region of high specificity, where the true negative rate $\mathbf{P}_0(\beta^\top X \leq t)$ increases from 0.90 toward 1. Here we see the ordering of the exponential classifiers reversed. This pattern is consistent with our interpretation of the exponential objective in Section 1.5.4: higher λ puts more weight on sensitivity, and lower λ puts more weight on specificity.

We see this pattern as the key consideration in choosing λ . In applications such as disease testing or screening for default risk, where a false negative may be much more costly than a false positive, a larger λ should be preferred; but if the goal is to maintain high specificity while optimizing for sensitivity, then a smaller λ is more appropriate. The choice of λ does not solve the problem of imbalanced data, but it helps control the consequences of the imbalance.

At both extremes, Figure 1.6 indicates that the performance of the logistic classifier falls between the exponential classifiers with $\lambda = 0.1$ and $\lambda = 0.5$. We investigate this pattern further in Figure 1.7, where we consider the effect of a smaller ($N = 1,000$) or larger ($N = 50,000$) sample size. Comparing these results with those in Figure 1.6 reveals a consistent pattern: as N increases, the performance of the logistic classifier becomes indistinguishable from that of an exponential

classifier with small λ . This pattern is consistent with Theorem 1.4.2: if we think of $\beta_*(\lambda)$ as a function of the exponent λ , then (1.11) suggests that $\lim_{\lambda \downarrow 0} \beta_*(\lambda) = \beta_*$, where β_* is the limiting coefficient vector for ordinary logistic regression.

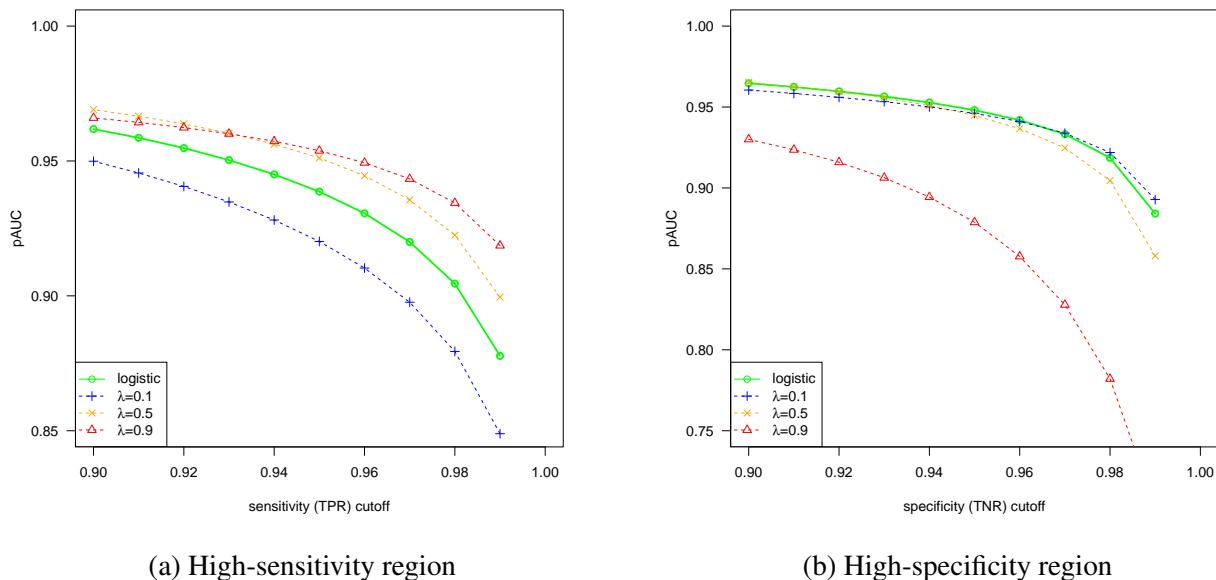


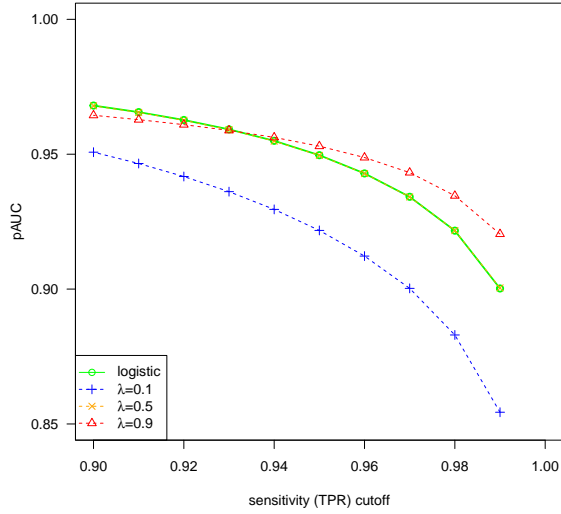
Figure 1.6: Comparison of pAUC values for logistic and exponential ($\lambda = 0.1, 0.5, 0.9$) classifiers

1.7 A Credit Risk Application

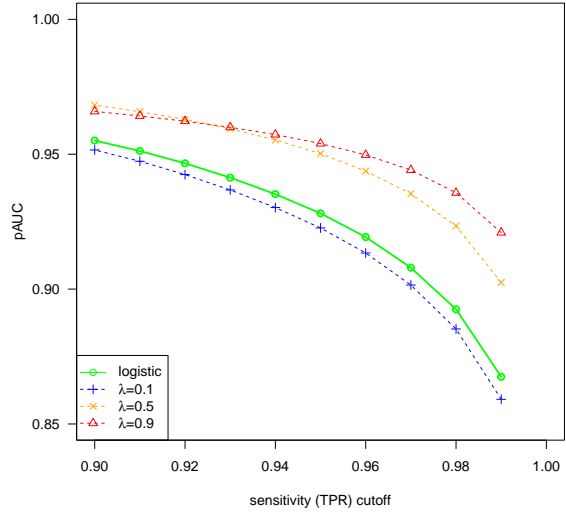
1.7.1 Freddie Mac Data

The task of classifying borrowers by their credit risk is challenged by imbalanced data in settings where defaults are rare. In this section, we apply ideas from previous sections to quarterly data from the Freddie Mac Single Family Loan-Level Dataset, from 2003 to 2016. The dataset can be accessed from http://www.freddiemac.com/research/datasets/sf_loanlevel_dataset.page. The dataset covers mortgages purchased or guaranteed by Freddie Mac.

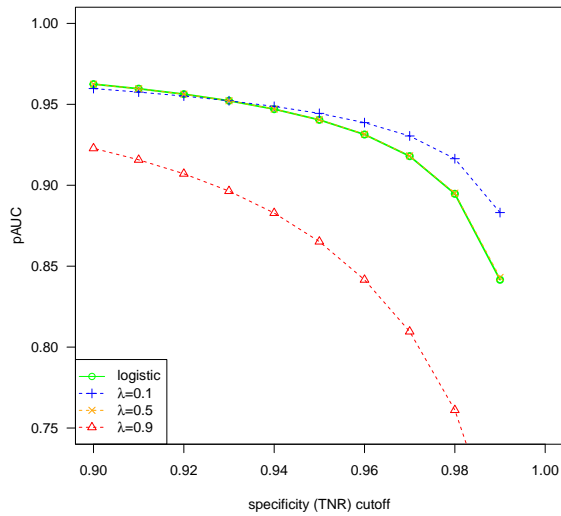
Our outcome of interest — the binary label we attach to each loan — is whether the loan defaults within two years of origination. We define a loan to be in default if it is 180 days or more



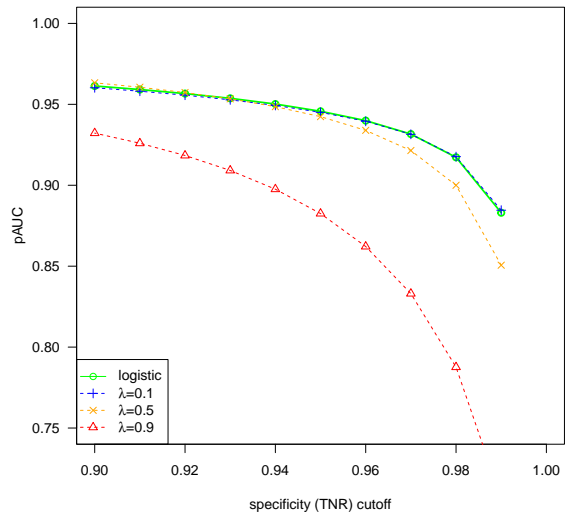
(a) High-sensitivity region, $N = 1,000$



(b) High-sensitivity region, $N = 50,000$



(c) High-specificity region, $N = 1,000$



(d) High-specificity region, $N = 50,000$

Figure 1.7: As N increases, the logistic pAUC values move closer to the exponential pAUC values with small λ in both the high-sensitivity and high-specificity regions

past due. Our goal is to predict this outcome based on loan and borrower features available at origination. This setup is consistent with Li et al. [58], although our sample is much larger.

Figure 1.8a plots the number of loans originated in each quarter, and Figure 1.8b plots the default rate over time from 2003 to 2019. We exclude from our analysis all loans that were repurchased within two years of origination. The default rate is almost always less than 0.03%, except around the financial crisis of 2008 when it climbs near 3.5%. We are thus dealing with extremely imbalanced data and considerable variation in the degree of imbalance.

In predicting outcomes, we use a combination of numerical and categorical attributes. The numerical variables are credit score, original debt-to-income ratio, log of original unpaid principal balance, original loan-to-value ratio, and original interest rate; the categorical variables are number of borrowers (one or more than one), first time homebuyer flag, number of units, occupancy status, loan origination channel, prepayment penalty mortgage flag, property type, and loan purpose. Precise definitions of these variables can be found in the Freddie Mac [32] user guide.

We estimate linear classifiers over a rolling window, for $t = 2003, \dots, 2013$. For $t = 2003$, the process works as follows. We use 80% of loans originated in any of the four quarters of 2003 and their default status in the corresponding quarter of 2005 ($t + 2$) to estimate a model, reserving the other 20% of the data for later validation. This is our training data for $t = 2003$. We then apply the trained model to the attributes of loans originated in the first quarter of 2004 to predict default status as of the first quarter of 2006 ($t + 3$). This is our test data for $t = 2003$. We apply the same process, retraining the model with $t = 2004$, to predict default status in the first quarter of 2007 for loans originated in the first quarter of 2005. Our last forecast is for defaults in the first quarter of 2016, for loans originated in the first quarter of 2014, trained based on loans originated in $t = 2013$.

We remove loans that are missing values for any numerical variables. For each categorical variable, we interpret missing values as a separate category. At each t , we check each variable to ensure that we have at least two distinct values of the variable in the data to avoid degeneracy. We omit the variable for that t if the variable fails this check, which happens in fewer than 1% of cases.

Using this process, we estimate four classifiers at each t , using logistic regression and exponen-

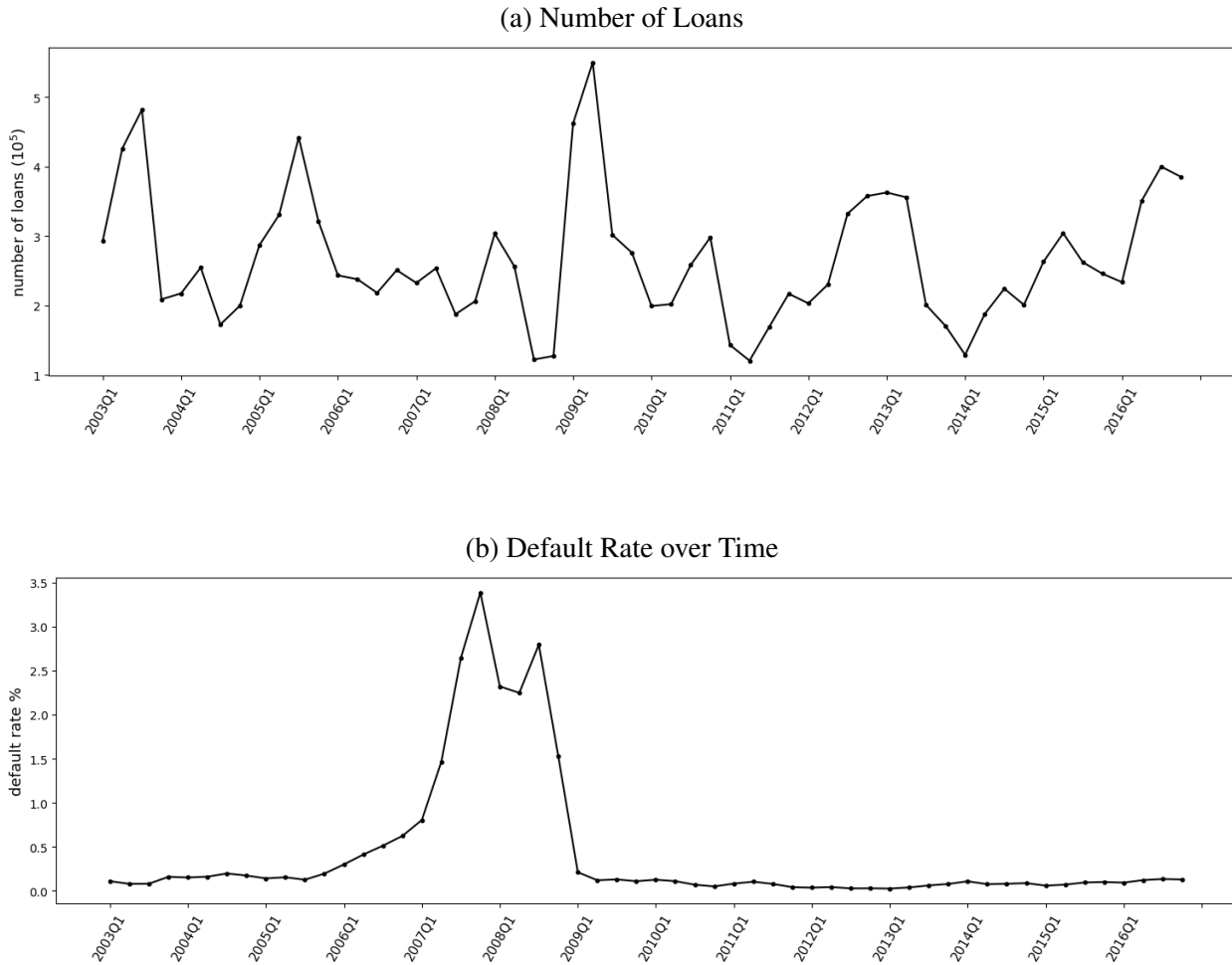


Figure 1.8: Freddie Mac Summary Data

tial objectives with $\lambda = 0.1, 0.5, 0.9$. In Appendix A.5.1, we report AUCs for training, validation, and test data for each classifier, for each t . All AUCs are above 0.8, indicating that linear classifiers perform reasonably well in this task. The validation AUCs and testing AUCs are all very close to the training AUCs, allaying any concerns about overfitting.

1.7.2 High-Sensitivity Classifiers

We consider a lender that would like to apply a simple first-pass classifier that correctly identifies at least 99% of customers who would default as high risk. Those classified as high risk would then undergo a costlier in-depth review. The lender thus wants the first-pass classifier to have a high TPR to make it highly sensitive to likely defaulters. We have seen that, in highly imbalanced

settings, logistic regression becomes similar to an exponential classifier with λ close to zero; but we have also seen that in the high-sensitivity region we should prefer to take λ close to one. We investigate this comparison using the Freddie Mac data.

For each classifier and each year t , we set a classification threshold to achieve a TPR of 99% in the training data. We then evaluate the TPR and TNR in the test set for each classifier and each year.

Appendix A.5.2 reports the test TPRs for all methods and all years. In all cases, the test TPR is close to 99%, indicating that the threshold set in the training data works well in the test data. However, we see clear differences in the test TNRs reported in Table 1.2. In all years, the exponential classifier with $\lambda = 0.9$ achieves the best or near the best performance, as we expected in this high-sensitivity region. The classifier with $\lambda = 0.5$ consistently outperforms the logistic classifier and the case $\lambda = 0.1$, and the last two are difficult to distinguish. All of these findings are consistent with our interpretation of the effect of the parameter λ and the relationship between the logistic and exponential objectives.

Year	Logistic	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 0.9$
2003	29.57	29.06	33.11	36.34
2004	24.34	24.11	26.68	26.11
2005	23.51	23.59	24.45	24.14
2006	22.93	23.62	26.17	28.06
2007	21.97	21.84	23.09	24.07
2008	23.57	23.13	24.33	24.50
2009	29.65	25.91	29.29	32.14
2010	17.14	17.80	22.26	26.33
2011	31.68	30.59	29.52	33.67
2012	33.08	32.98	35.63	42.17
2013	26.74	26.25	23.72	30.25

Table 1.2: True negative rates (in percent) in test data for classifiers trained at a true positive rate of 99%

1.7.3 pAUC plots

To gain further insight into the comparison of the classifiers, we examine pAUC plots like those introduced in Section 1.6, but now using the Freddie Mac data. Figure 1.9 shows results for 2007, but we find the same pattern in all years: as expected, a higher λ gives better results in the high-sensitivity region, and a lower λ works better in the high-specificity region. The performance of logistic regression is similar to that of $\lambda = 0.1$ in the first case but closer to that of $\lambda = 0.5$ in the second case due to the finite imbalance in the data.

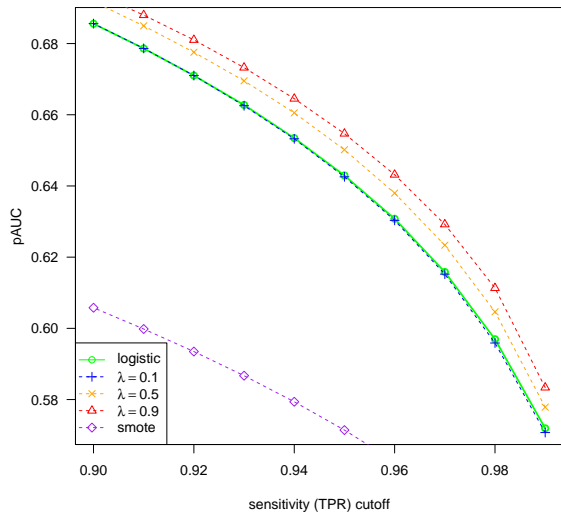
To compare these linear classifiers with upsampling methods often used in practice to address data imbalance, we apply the SMOTE method of Chawla et al. [15]. We apply it using 5 nearest neighbors, and we upsample the default class to match the sample size of the non-default class. We then plot the pAUC curve of the logistic regression classifier trained on the transformed data in Figure 1.9. The results in the figure indicate that SMOTE does not improve performance in either the high-sensitivity or high-specificity regions. We have also found that it results in a lower overall AUC than the other methods.

To gauge the statistical significance of differences across λ values, Figure 1.10 includes 90% bootstrap confidence intervals around the pAUC curves. For clarity, we compare just two cases, $\lambda = 0.9$ in red and $\lambda = 0.1$ in blue. The confidence bands barely overlap, indicating that the ordering of the two curves is reliable.

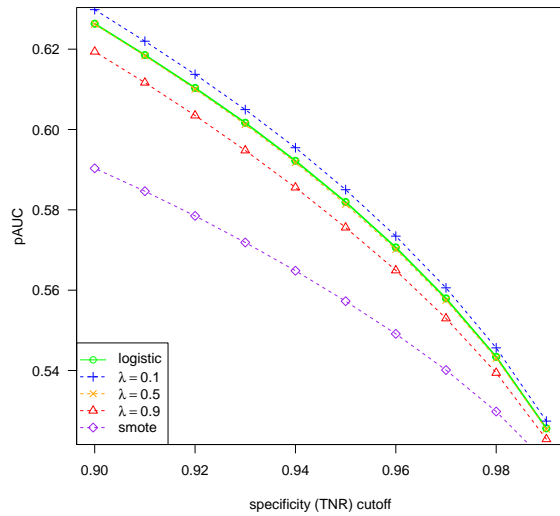
1.7.4 Choice of λ

In the Freddie Mac example, where false negative instances are costly, we have focused on the high sensitivity region and established that a larger λ is preferable. We have centered our discussion on accuracy or the true negative rate. We now give an example to show how λ can be chosen to optimize some other objective or utility function.

For each $\lambda \in [0, 1)$, we can find the corresponding β_λ by minimizing the empirical version of (1.1). Let R_λ^t denote the linear classification rule parameterized by β_λ with threshold t . That is, given feature X_i of instance i , $R_\lambda^t(X_i) = \mathbf{1}\{\beta_\lambda^\top X_i \geq t\}$. We classify instance i to be positive (or risky

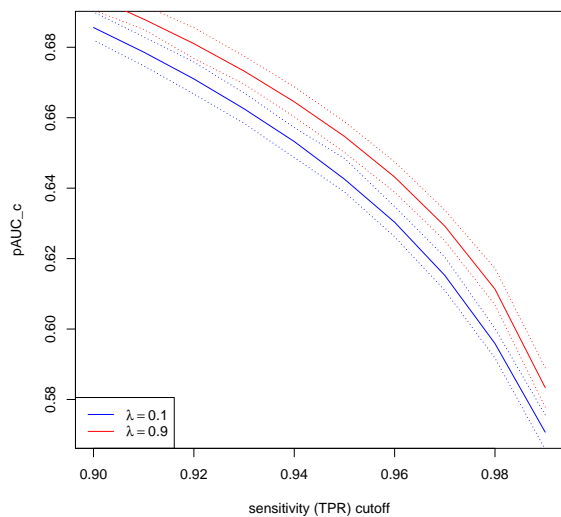


(a) High-sensitivity region

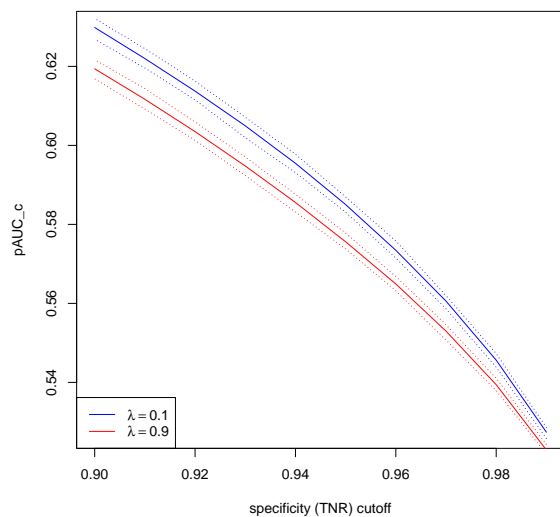


(b) High-specificity region

Figure 1.9: Comparison of pAUC values in test data for logistic and exponential ($\lambda = 0.1, 0.5, 0.9$) classifiers using Freddie Mac loan data. SMOTE upsampling with logistic regression is also included for comparison



(a) High-sensitivity region



(b) High-specificity region

Figure 1.10: Test data pAUC plots with 90% confidence bands for $\lambda = 0.1$ and $\lambda = 0.9$.

in the credit risk example) if $R_\lambda^t(X_i) = 1$. In choosing λ , we may be concerned about the size of loans to risky borrowers as well as the error rates in classifications. This leads us to maximize the average value of true positive loans — loans that are classified to be positive that actually default — averaging over all loans classified as positive, subject to the constrained that TPR meets some threshold l :

$$\max_{\lambda, t} \frac{\sum_i V_i \mathbf{1}\{R_\lambda^t(X_i) = 1\} \mathbf{1}\{Y_i = 1\}}{\sum_i \mathbf{1}\{R_\lambda^t(X_i) = 1\}} \quad s.t. \quad TPR(R_\lambda^t) = l. \quad (1.24)$$

As before, we desire a high TPR and therefore take l to be close to 1. Table 1.3 reports the value of (1.24) for different values of λ across the years when t is chosen such that $TPR(R_\lambda^t) = 0.99$. We observe that except for the years 2004 and 2005, $\lambda = 0.9$ consistently yields the best classifier in the sense that the loans that are classified as risky lead to the largest average loss.

Year	Logistic	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 0.9$
2003	175.65	174.82	185.53	194.69
2004	307.75	306.32	316.54	313.92
2005	303.25	309.50	312.96	311.59
2006	1135.06	1125.45	1162.96	1191.07
2007	5355.27	5353.43	5432.49	5494.36
2008	6406.06	6383.23	6476.76	6489.64
2009	426.72	414.86	433.74	451.58
2010	187.01	188.86	199.73	210.68
2011	200.77	200.91	198.29	210.69
2012	90.66	92.58	96.07	106.96
2013	119.23	118.19	114.62	124.52

Table 1.3: Average default loan amount of classified positive instances when TPR is 99%

1.8 Concluding Remarks

We have shown that a broad family of linear discriminant functions have explicit limits as the sample size of one class grows while the sample size for the other remains fixed. Linear discriminant functions defined by asymptotically subexponential weight functions share a common limit with logistic regression. A wider range of limits applies using asymptotically exponential weights. The limits of these classifiers reflect different types of robustness or conservatism to-

wards worst-case false-positive and false-negative errors. Our analysis does not solve the problem of imbalanced data, but it does provide the user some control over the consequences of imbalance through the exponent λ , favoring performance in either the high-specificity or high-sensitivity regions. We illustrated these ideas through numerical examples and an application to credit risk in predicting mortgage defaults.

Our analysis is limited to linear discriminant functions. Linear classifiers are widely used in practice, at least in part because they are easy to interpret. Note, also, that the features used for classification could include scores computed from nonlinear models. In other words, a linear discriminant function can be used to aggregate results from multiple models. It would be interesting to know if similar limits hold for scoring rules derived from regression trees or neural networks. The loss functions used to optimize these types of rules over parameters are typically nonconvex, which significantly complicates any analysis of their limiting behavior.

Chapter 2: Trading TP_2/RR_2 Violations in Options

The options market presents unique opportunities for identifying and exploiting pricing anomalies. This chapter investigates the empirical significance of the total positivity of order 2 (TP_2) property in call options and the reverse sign rule of order 2 (RR_2) in put options within the S&P 500 index. By analyzing the conditions under which these properties are violated, we assess the potential trading opportunities that arise from these violations. Our findings indicate that while these conditions are mostly satisfied, violations, when they occur, can be strategically exploited for consistent positive returns. This chapter demonstrates how theoretical insights into option pricing anomalies can be translated into practical trading strategies, providing a strategic approach to navigating the options market.

2.1 Introduction

The concept of *total positivity* (TP) is crucial in understanding the behavior of certain financial models. Consider a square matrix that is formed by function values. We say that the function is *totally positive* (TP) if any square submatrix of arbitrary order in this matrix has nonnegative determinants. A related property is the *reverse sign rule* (RR), where the determinants of the submatrices are nonpositive. Previous research demonstrated that under the Black-Scholes model, out-of-the-money (OTM) call options necessarily satisfy the TP condition in strike and days-to-maturity (DTE), while OTM put options satisfy the RR property [38]. More recently, it has been established that a particular order of the general TP and RR condition – total positivity of order 2 (TP_2) and reverse sign rule of order 2 (RR_2) – apply more broadly for all call and put options, not just OTM strikes [39].

We examine the empirical significance of the TP_2 and RR_2 conditions by analyzing options on

the Standard and Poor 500 (S&P 500) index (SPX options) from 2000 to 2022. We compute the violation rates of TP_2 and RR_2 conditions daily and evaluate the extent of these violations. Our findings reveal that TP_2 and RR_2 conditions are satisfied by the majority of liquid options, with 30-day moving averages of violation rates rarely exceeding 6%. However, violation rates show significant variation over different periods, with lower rates observed post-pandemic.

Interestingly, the peaks in call and put violation rates do not coincide, with higher violation rates observed for put options. To explore these differences, we regress the violation rates against market conditions such as the CBOE's market volatility index (VIX), the CBOE's skew index (SKEW), and market returns. Our results indicate that perceived risk, as measured by VIX levels and the SKEW index, is more effective in explaining the violation rates than market returns. Put violations are particularly influenced by market sentiment, aligning with the practice of purchasing OTM puts for downside protection, whereas OTM calls are used for different investment intentions.

We also analyze how violations are distributed across different maturities and strikes. While options near maturity and near the money are the most traded, they do not necessarily exhibit a higher likelihood of forming a violation pair. We observe similar violation rates across a range of maturities up to two months for call options. However, for put options, soon-to-expire pairs have a noticeably higher probability of forming a violation pair.

As previously reasoned in [39], TP_2 and RR_2 conditions can be viewed as slightly stronger conditions than the no static arbitrage condition. When a violation occurs, it presents potential trading opportunities. The challenge lies in the fact that TP_2 and RR_2 conditions involve products of two options on each side of the equation, complicating the trading based on these conditions.

We propose a long-short strategy to exploit these violations. For a violating pair, the direction of the inequality in TP_2 or RR_2 is reversed, making one side "undervalued" and the other "overvalued." Thus, we long the undervalued side and short the overvalued side, treating one option value as the coefficient and the other as the actual option contract to be traded.

Our findings indicate that this long-short strategy consistently generates positive returns, both per trade and when applied in a dynamic trading strategy. Testing our strategy with data from 2000

to 2022, we find that it outperforms the S&P 500 index by a substantial margin for both calls and puts, with only half the volatility of the S&P 500 daily returns. Notably, the S&P 500 index itself yields nearly a 300% return over the same period.

Our analysis uses end-of-day data from *OptionMetrics*. The core analysis assumes trading at mid prices, but similar patterns hold when adopting a more conservative approach, assuming purchases at the best offer price and sales at the best bid price, or accounting for the full bid-ask spread in trading costs.

The rest of the paper is organized as follows: Section 2.2 reviews total positivity and provide its connections with Black-Scholes calls. Section 2.3 overviews the history and evolution of SPX options and the data source for our analysis. Section 2.4 analyzes the behavior of the violations, and Section 2.5 examines trading these violations. Additional experiment details are given in the Appendices.

2.2 Background

In this section, we briefly review the total positivity condition and its applications in mathematical finance. We also revisit our earlier results on TP (RR) conditions for Black-Scholes call (put) options.

Total Positivity Condition: A function $K(\cdot, \cdot) : I_x \times I_y \rightarrow \mathbb{R}$ is said to be *totally positive of order r* (TP_r) if for all $n \in [r]$, all ordered $x_1 < \dots < x_n \in I_x$ and ordered $y_1 < \dots < y_n \in I_y$, the matrix formed by evaluating K at (x_i, y_j) satisfies:

$$\det[K(x_i, y_j)]_{i \in [n], j \in [n]} \geq 0. \tag{2.1}$$

If K is totally positive of all orders $r = 1, 2, \dots$, then it is totally positive (TP).

Relatedly, reversing the order of one of x or y in (2.1) yields *reverse rule of order r* (RR_r)

condition. A function K is RR_r if

$$\det[K(x_i, y_j)]_{i \in [n], j \in [n]} \leq 0 \quad (2.2)$$

for all $n \in [r]$, all $x_1 > \dots > x_n$ and all $y_1 < \dots < y_n$.

For further details on totally positive functions and their properties, readers can refer to Chapter 2 of Karlin [50]. Total positivity has seen various applications in mathematical finance. For example, Zipkin used total positivity conditions to study bond prices under stochastic interest rate settings [75]. More recently, Keller-Ressel studied yield curve shapes under Vasicek Model using total positivity condition [51].

TP Conditions and Black-Scholes Calls: Consider an underlying asset $\{S_t, t \geq 0\}$ and the undiscounted call and put options with strike K and expiry T

$$\bar{C}(K, T) = \mathbf{E}[(S_T - K)_+], \quad \bar{P}(K, T) = \mathbf{E}[(K - S_T)_+]$$

where $(\cdot)_+ = \max(\cdot, 0)$. [38] shows that under the Black-Scholes model (and among other models), if $\mathbf{E}[S_t] = S_0 = 1$, then $\bar{C}(K, T)$ and $\bar{P}(K, T)$ are TP and RR, respectively, for all out-of-the-money (OTM) $K > 1$. It is further shown in [39] that if we restrict our attention to the order of 2, then \bar{C} and \bar{P} are TP_2 and RR_2 for all K , not just OTM strikes.

TP_2 implies that for all $K_1 \leq K_2$,

$$\frac{\bar{C}(K_2, T)}{\bar{C}(K_1, T)}$$

is increasing in T . In particular, TP_2 implies the no strike nor calendar spread static arbitrage under Black-Scholes models. Specifically, proposition 5.1 in [39] shows that if $\bar{C}(K, T)$ is TP_2 , then it is arbitrage-free in the sense that there exists a martingale $\bar{S}_t, t \geq 0$ for which $\bar{C}(K, T) = \mathbf{E}[(\bar{S}_T - K)_+]$ for all $K, T \geq 0$. However, it is possible that some call options in an arbitrage-free setting are not TP_2 .

Extending to Dividend-paying Asset: Let $\{\tilde{S}_t, t \geq 0\}$ denote a dividend-paying asset. The TP_2 conditions can be summarized as follows: for two pairs of strikes K_1, K_2 , and expirations T_1, T_2 , if $T_1 < T_2$ and $K_1/F_{T_1} < K_2/F_{T_2}$ then

$$C(K_1, T_1)C(K_2, T_2) \geq C(K_1 F_{T_2}/F_{T_1}, T_2)C(K_2 F_{T_1}/F_{T_2}, T_1), \quad (2.3)$$

where F_t denotes the futures price at time t , $C(K, T)$ represents the value of a European-style call option with strike K and day-to-maturity (DTE) T , and the option itself we denote (K, T) -Call.

Similarly, put options satisfy the RR_2 property if

$$P(K_1, T_1)P(K_2, T_2) \leq P(K_1 F_{T_2}/F_{T_1}, T_2)P(K_2 F_{T_1}/F_{T_2}, T_1). \quad (2.4)$$

For notation simplicity, we define $\tilde{K}_1 = K_1 F_{T_2}/F_{T_1}$ and $\tilde{K}_2 = K_2 F_{T_1}/F_{T_2}$. We say options (K_1, T_1) -Call and (K_2, T_2) -Call ((K_1, T_1) -Put, (K_2, T_2) -Put) constitute a TP_2 -violating (RR_2 -violating) pair if TP_2 (RR_2) condition above is violated.

Empirical observations indicate that TP_2 or RR_2 violations are relatively rare (Section 6.5 in [38]). In this work, we expand the empirical analysis on TP_2 and RR_2 violations, studying when and how often these violations occur, and examining put and call violations across different times for S&P 500 options. Because TP_2 and RR_2 conditions are shown to be slightly stronger versions than the no-arbitrage condition, we investigate if violations lead to profitable trading strategies.

2.3 Options on S&P 500 Index

We consider S&P 500 index options (SPX) traded on the Chicago Board Options Exchange (CBOE) from January 2000 to December 2022.

2.3.1 Evolution of S&P Options

Since their introduction by the Chicago Board Options Exchange (CBOE) on July 1, 1983, S&P 500 index (SPX) options quickly attracted interests from investors, now standing as the most

heavily traded index options globally. These options offer a wide range of expiration dates and strike prices, catering to a diverse investor base with their robust liquidity and trading volume.

Historically, AM-settled SPX options were the standard, with their settlement prices determined on the morning of the third Friday of each month for monthly options. These options are settled based on the Special Opening Quotation (SOQ), which is derived from the opening prices of the stocks within the S&P 500 index. Different from S&P 500 opening price, the SOQ is calculated only after all constituent stocks have begun trading, which can introduce a delay, as the opening of stocks can vary due to market conditions. These options have a cut-off for trading at the close of the market on the Thursday before expiration, requiring traders to finalize their positions without the ability to react to market movements on the expiration day itself.

PM-settled SPX Options, on the other hand, conclude at the market's close and are settled with the closing S&P 500 index level. The CBOE introduced ¹ these options in 2007, under the SEC's PM Option Expiration Pilot program. This move expanded the choices available to traders, starting with end-of-month expirations and subsequently including weekly expirations in 2010, and monthly expirations on the 3rd Friday of each month starting in 2011. The first weekly options expired only on Fridays, with Wednesday (February 23, 2016)² and Monday (August 15, 2016)³ expiries added over time. On April 18 and May 11, 2022, respectively, Tuesday and Thursday expirations were added to complete the suite, making options expiring five days a week.

Figure 2.1 plots the total yearly trading volume (across all strikes and expiries) for AM and PM-settled SPX options from 2000 to 2022. The bottom orange bar is the AM-settled options, the light blue bar is the PM-settled options whose day-to-maturity (DTE) is at least one, and the dark blue bar is the same-day expiring (0DTE) PM-settled options. The 0DTE options, as the name implies, have a very short lifespan, expiring on the same day they are traded. This characteristic makes

¹More precisely, PM-settled options were *reintroduced* in 2007, as all SPX options were PM-settled initially between 1983 and 1986. However, to facilitate better inventory management among dealers at the time, a shift to AM-settled options was made in June 1987.

²<https://www.prnewswire.com/news-releases/cboe-to-list-spx-wednesday-expiring-weekly-options-300212876.html>

³<https://ir.cboe.com/news/news-details/2016/CBOE-to-List-SPX-Monday-Expiring-Weekly-Options-07-11-2016/default.aspx>

ODTE options a compelling choice for traders looking for opportunities to capitalize on short-term market movements. The overall interest in SPX options is steadily increasing; in particular, we observe exponential growth in SPX volume with the introduction of PM-settled options, although trading volumes for AM-settled options stay roughly flat after the introduction of PM-settled options. The trading volume for non-ODTE PM-settled options stays relatively steady since 2018, but the interest in ODTE options has increased significantly from 2018 to 2022.

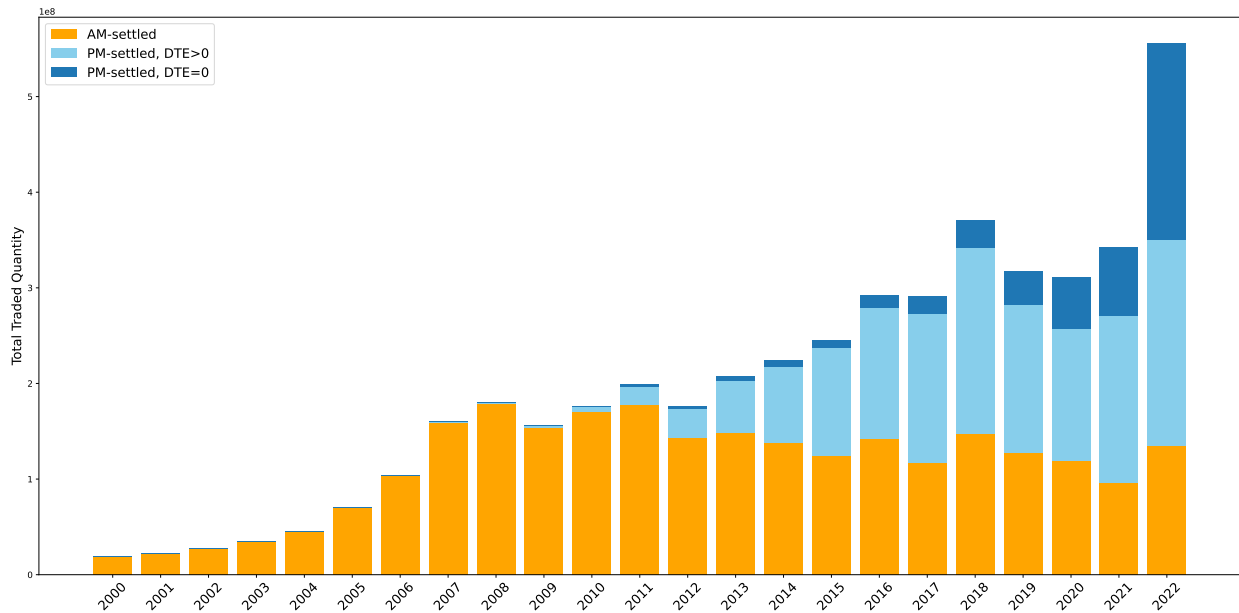


Figure 2.1: Total yearly trading volume separated to AM (orange), PM-settled and non-ODTE (light blue), and PM-settled ODTE (dark blue).

2.3.2 Data

We study SPX options because they are the most traded index product on the market, with large volumes traded across different strikes and expirations, attracting both retail and professional investors. We detail the three data sources in our numerical exercise. A more detailed description of the data can be found in Appendix B.1.

S&P 500 index data: We obtain data associated with the underlying S&P 500 index levels from CRSP. The closing S&P price on option expiry day determines the PM-settled SPX option payoffs.

We also collect the Settlement Price (SET) information from the CBOE website. The SET index tracks the special opening quotation (SOQ) that determines the settlement value of the AM-settled SPX options. Finally, we collect CBOE’s market volatility index (VIX) and skewness (SKEW) time series data.

SPX Option data: We collect historical end-of-day SPX option prices from the Option Price dataset in *OptionMetrics*. This dataset includes information such as option prices (best ask and best bid), quantity traded, greeks, and settlement type (AM-settled or PM-settled) at market close (or more precisely, 3.59 pm) on each trading day. We also collect the S&P 500 forward price information, available as a standalone dataset in *OptionMetrics*. This information is useful for evaluating the right-hand sides of equations (2.3 and 2.4).

Market Data: We consider CBOE’s VIX and SKEW indices that characterize the market conditions, particularly perceived tail risks. We use the 3-month CRSP Risk-Free Rates as the benchmark risk-free rate in our experiments. This time series is useful for calculating the Sharpe Ratio of our portfolio in Section 2.5.

2.4 TP₂-Violating Option Pairs

We call an option 2-tuple $((K_1, T_1)$ -Call, (K_2, T_2) -Call) violating TP₂ condition in (2.3) a TP₂-violating option pair. We similarly define RR₂-violating pairs for puts. We detail how we evaluate the violations before we explore some time-evolving characteristics of the violations.

2.4.1 Determining TP₂-Violating Option Pairs

To avoid additional errors introduced by interpolation, we restrict our attention to the strikes and the maturities available in *OptionMetrics*. For every trading day, we loop through option values of all possible pairs of strikes and maturities (K_1, T_1) and (K_2, T_2) satisfying $0 < T_1 < T_2$ and $K_1/F_{T_1} < K_2/F_{T_2}$, store their value, and check if those, along with the option values of parameters (\tilde{K}_1, T_2) and (\tilde{K}_2, T_1) , violate the TP₂ condition in (2.3) for calls or the RR₂ condition in (2.4) for

puts. A violation pair can be formed between i) two AM-settled options; ii) two PM-settled options; or iii) one AM-settled and one PM-settled option. We do not separate the violations based on these settlement types unless explicitly mentioned otherwise. For now, we use options' mid prices as their option values. Later in Section 2.5.5 we evaluate the effect of accounting for bid-ask spread in violation determination.

Removing illiquid options: We only consider options that have been traded at least 1 lot on the current day. We further remove options whose best bid price is 0, or whose delta is larger than 0.99 or less than 0.01 in absolute terms. These options are deep in-the-money (ITM) or out-of-the-money (OTM) and present limited trading opportunities.

Rounding up the strikes: Because \tilde{K}_1 and \tilde{K}_2 are not necessarily integers or strikes available for trade, we round up \tilde{K}_1 and \tilde{K}_2 . Due to the monotonicity of the options, doing so results in a conservative approach to evaluating both TP_2 and RR_2 violations: for calls in TP_2 violations, we inflate the RHS of (2.3); for puts in RR_2 violations, we decrease the RHS of (2.4), both of which form a more conservative approximation of the respective TP_2 or RR_2 condition.

If the rounded-up strikes are not listed on CBOE, we look for the nearest strike that is *larger* than the rounded-up strike; if no strike is traded within \$50 of the rounded-up strike, we discard this option as it indicates that this particular strike is in a highly illiquid region where few neighboring contracts are traded on the market, leading to inaccurate pricing.

Remark: The "rounding up the strikes" approach may be more conservative than it appears. As we will soon see, the majority of the violations are formed by options expiring in the near term and when T_1 and T_2 are close together. That is, the ratio F_{T_2}/F_{T_1} is close to 1. Imagine a case where $F_{T_2}/F_{T_1} - 1 = 0.05\%$. Then, by the procedure above, we necessarily search for the next strike $\tilde{K}_1 > K_1$, and $\tilde{K}_2 = K_2$ for near-the-money strikes K_1, K_2 . Because most SPX options strikes are at least \$5 apart, looking at the next strike, the option value could change by a material amount. This renders our TP_2 or RR_2 violating pairs a potentially extremely conservative approximation, and the actual violation rates may be considerably larger if fractional strikes were listed and traded.

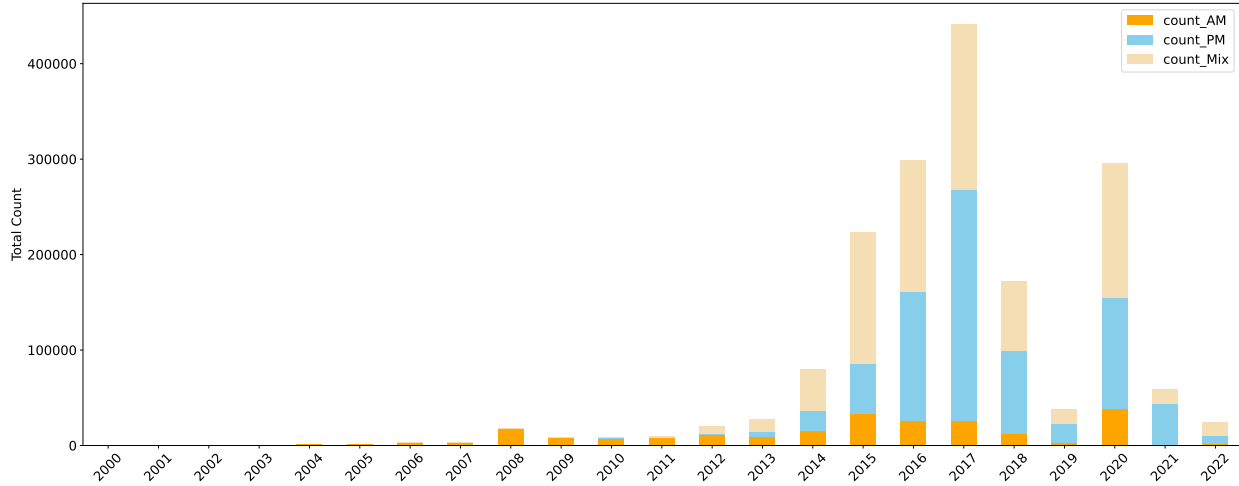


Figure 2.2: TP₂ violations count by settlement type.

We also present additional findings under the assumption $F_{T_2}/F_{T_1} = 1$ and without any “rounding up of strikes” in Appendix B.3.2.

2.4.2 Evolution of Violations

With growing interest in SPX options, the number of TP₂ violations has also increased significantly over the years. Figures 2.2 and 2.3 plot the number of TP₂-violating calls and RR₂-violating puts from 2000 to 2022. The bottom orange bar represents when both options in the violating pair are AM-settled options, the light blue bar represents PM-settled options, and the top light beige bar represents when one option is AM-settled and the other PM-settled. The figures reveal an upward trend in violations, with the total violation count increasing exponentially from the beginning of the period to 2017. The number of total violating pairs decreased in 2018 and further in 2019, before increasing upwards again after the global pandemic in 2020. After 2020, TP₂ violations remain near the 2019 level, but RR₂ violations increase to a record-high in 2021, before dropping significantly in 2022. The number of TP₂-violating or RR₂-violating pairs is mostly driven by the increased popularity in PM-settled options; the number of purely AM-settled violating pairs increases at a much smaller rate.

Figure 2.4 plots the 30-day moving average of violation rates for call options (in blue) and

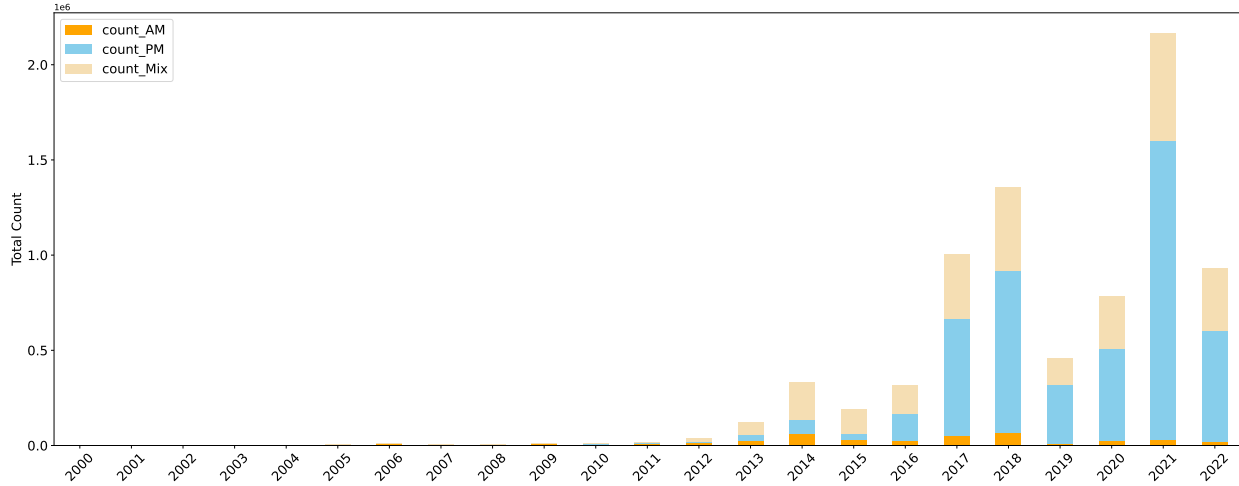


Figure 2.3: RR₂ violations count by settlement type.

put options (in orange). The violation rate is calculated as the daily number of violating pairs divided by the total number of option pairs that could have constituted a TP₂ or RR₂ violating pair. Specifically, we consider an option pair that satisfies: (i) $K_1/F_1 < K_2/F_2$, (ii) $T1 < T2$, and (iii) for both \tilde{K}_1 and \tilde{K}_2 , there is some listed strike traded at least one lot and are not \$50 larger than \tilde{K}_1 or \tilde{K}_2 as an *eligible* candidate pair (denominator of the ratio), and the pairs that violation (2.3) or (2.4) as the violating pair (numerator of the ratio). We observe that the violation rate is generally low: less than 6.5% for RR₂ violations and 4% for TP₂ violations. Additionally, RR₂ violations are more likely than TP₂ violations, and the peak violations do not usually occur simultaneously for puts and calls. Indeed, RR₂ violation rates peaked in 2006 and in 2014-2015, whereas TP₂ violation rates peaked in late 2008 to early 2009 and then near 2016. As a general trend, we observe an increasing trend in both violation rates from the beginning of the period to 2017 or early 2018, after which we observe a sharp decline in the rates, and this decline is more pronounced for call options than for put options.

Earlier we remarked that our approach to finding the next available strike for \tilde{K}_1 and \tilde{K}_2 is conservative. We plot a relaxed approximation of the TP₂ and RR₂ conditions where the futures ratios are assumed to be 1 (or equivalently, $\tilde{K}_1 = K_1$ and $\tilde{K}_2 = K_2$). We call this the “two-strike approximation”, and the resulting violation rate is given in Figure B.7 in Appendix B.3.2. As

expected, we see higher violation rates than in Figure 2.4, but overall violation rates are still low.

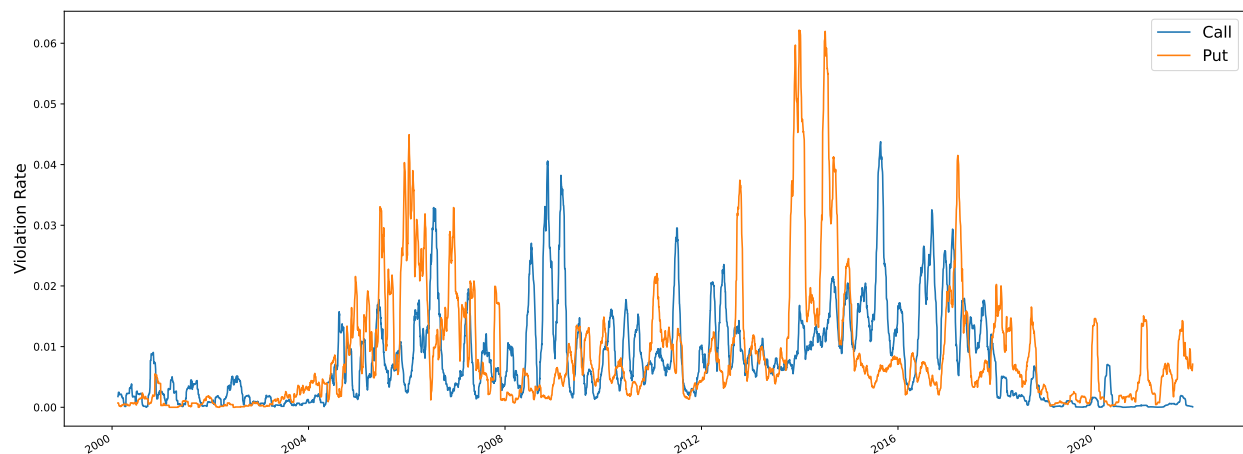


Figure 2.4: TP_2 and RR_2 violations (rolling 30-day average).

2.4.3 Violation Rates and Market Conditions

We regress the log of violation rates on the log of S&P returns (close to close), CBOE's VIX close levels, and CBOE's SKEW index close levels. The results are presented in Table 2.1, where the columns represent the dependent variables: log of call violation rate, log of put violation rate, and their differences. Newey-West robust standard errors are used to capture the time series nature of the data.

The regression results indicate that the change in the S&P 500 index does not contribute to the change in the violation rates for both calls and puts. However, both the VIX and SKEW indices are significant predictors. Higher market volatility (VIX) is associated with fewer violations, while higher perceived tail risk (SKEW) is associated with more violations. This pattern holds for both calls and puts. Moreover, the gap in put and call violation rates increases with higher perceived tail risk and decreases with higher market volatility.

The R-squared values for the regression are 5% for calls, but 21% for puts, indicating that variations in RR_2 violation rates are more closely related to market conditions than TP_2 violations. The near-perfect alignment of the coefficients from the third regression with the differences in the coefficients from the first two regressions suggests a consistent relationship between the predictors

and the dependent variables. This consistency indicates that the error terms in the individual regressions ($\log(\text{call vio})$ and $\log(\text{put vio})$) do not introduce significant bias or variation that would disrupt the linear relationships when taking the differences.

	$\log(\text{call vio})$	$\log(\text{put vio})$	$\log(\text{put vio}) - \log(\text{call vio})$
const	-10.663*** (1.110)	-13.329*** (0.849)	-2.667*** (0.990)
log SP Return	-0.103 (0.067)	0.026 (0.054)	0.128* (0.066)
CBOE's VIX	-0.056*** (0.012)	-0.114*** (0.009)	-0.058*** (0.010)
CBOE's SKEW	0.063*** (0.008)	0.108*** (0.006)	0.045*** (0.007)
R-squared	0.054	0.208	0.035
R-squared Adj.	0.053	0.207	0.034

Table 2.1: Regressing i) log of call violation rates; ii) log of put violation rates; and iii) differences between ii) and i) on log of S&P 500 index daily returns, CBOE's VIX index, and CBOE's SKEW index.

2.4.4 Maturities in Violation Pairs

Figure 2.4 shows that violations are overall rare, but they may not be rare for certain maturity or strike pairs. To study the effect of different maturities on violations, we plot the count of shorter day-to-maturity T_1 (x-axis) and longer day-to-maturity T_2 (y-axis) in violation pairs for calls from 2000 to 2022 in 2.5a. The DTEs are grouped into 7-day blocks. The bottom left area (option pairs with small DTEs) contributes to the majority of the violations. In particular, the majority of the violations occur when T_1 expires in about one month, and T_2 expires within a month from T_1 's expiration, with T_1 and T_2 being one or two weeks apart the most prevalent.

We plot the violation rate heatmap in Figure 2.5b to see which values of T_1 and T_2 lead to the

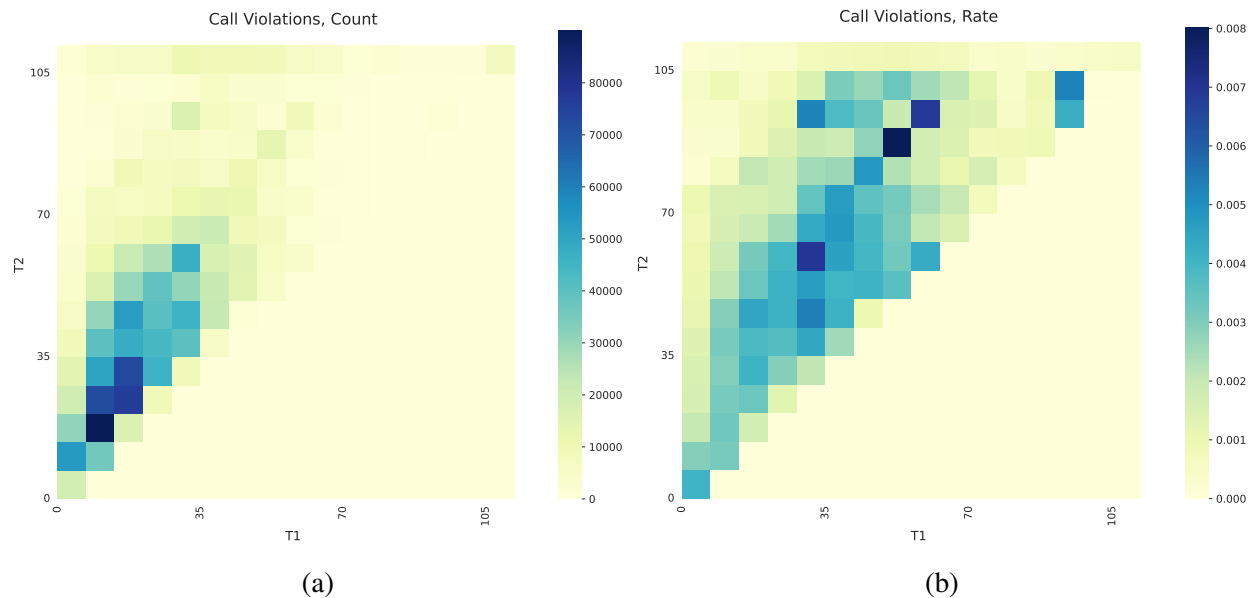


Figure 2.5: (a) TP₂ violation counts; (b) TP₂ violation rates. T_1 in the horizontal axis and T_2 vertical axis. DTEs are grouped into 7-day blocks.

largest chance of forming a violation, as smaller time-to-maturity options are traded more heavily than options that expire longer. The violation rate is calculated similarly to that in Figure 2.4, but is restricted to option pairs expiring in that 7-day block. The result shows a much more uniformly distributed, almost a band-like pattern where for T_1 less than 2-month and T_2 expires within 5–6 weeks after T_1 , the probability of observing a violation is relatively comparable. We also note that the violation rate does not exceed 10% in any cells.

We repeat the analysis for RR₂-violating put option pairs. The results are plotted in Figures 2.6a– 2.6b. The behavior of RR₂ violations differs significantly from that of TP₂ violations: Figure 2.6a shows that RR₂ violations are highly concentrated in extremely small T_1 and T_2 ; in particular, most violations occur for T_1 and T_2 are one week apart, when T_1 expires in 0–1 week and T_2 1 – 2 weeks. However, it is the two near-the-term put options expiring in one week that are the most likely to form a violation pair, as is evident in Figure 2.6b. In contrast to TP₂ violations in Figure 2.5b, we do not observe a higher violation rate for options expiring longer terms, and few violations are observed for options expiring longer than one month.

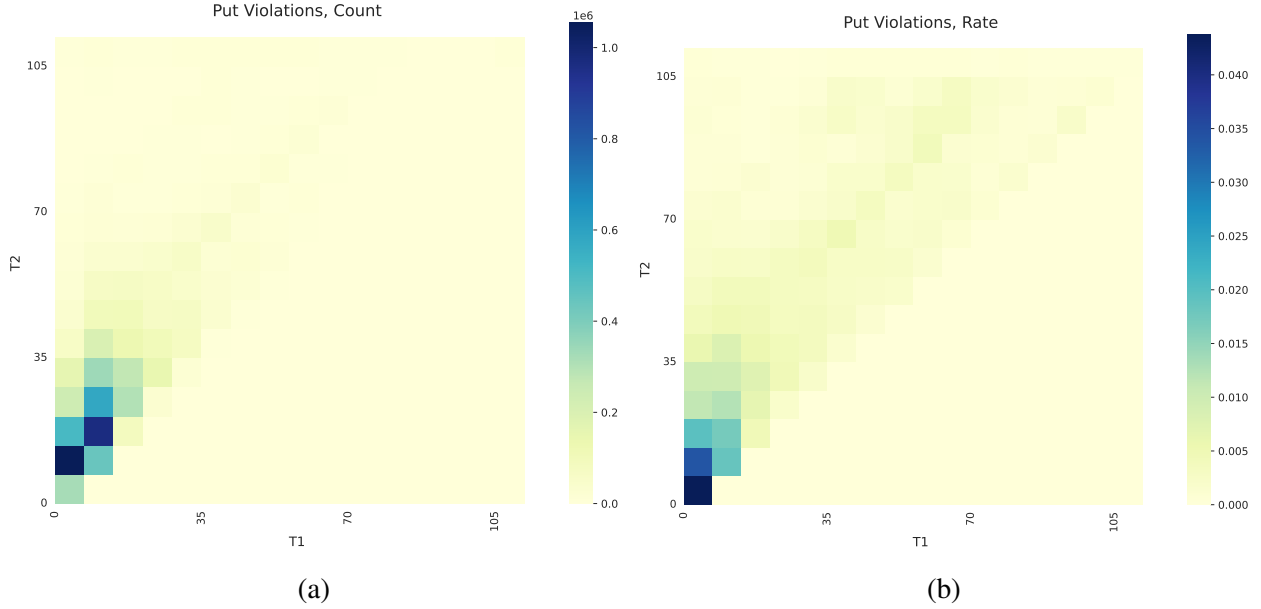


Figure 2.6: (a) RR_2 violation counts; (b) RR_2 violation rates. T_1 in the horizontal axis and T_2 vertical axis. DTEs are grouped into 7-day blocks.

2.4.5 Option Deltas in Violation Pairs

In the option delta space, more than 98% of the TP_2 violations and 99.5% of the RR_2 violations are formed by out-of-the-money (OTM) options. The remaining 2% for TP_2 or 0.5% for RR_2 are formed either by one in-the-money (ITM) option and one OTM option, or two ITM options.

Table 2.2 reports the mean and median of $\delta_i^V, i \in \{1, 2\}, V \in \{C, P\}$, the option delta of option contract $V(K_i, T_i)$ in the violation set. The deltas δ_1 and δ_2 of TP_2 -violating call options are much closer than those of RR_2 -violating put option. Also, all statistics are around or less than 0.1, showing that it is the options with extremely small deltas that constitute the majority of the violations.

	$E[\delta_1^C]$	$\text{Median}(\delta_1^C)$	$E[\delta_2^P]$	$\text{Median}(\delta_2^P)$
Calls	0.074	0.041	0.052	0.038
Puts	-0.020	-0.015	-0.106	-0.091

Table 2.2: Statistics for TP_2 and RR_2 -violating option deltas.

Figure 2.7 further shows the distribution of violation pair counts in the (δ_1^V, δ_2^V) -space. In both cases, the most violations are observed in the tail region of the delta space. However, for call

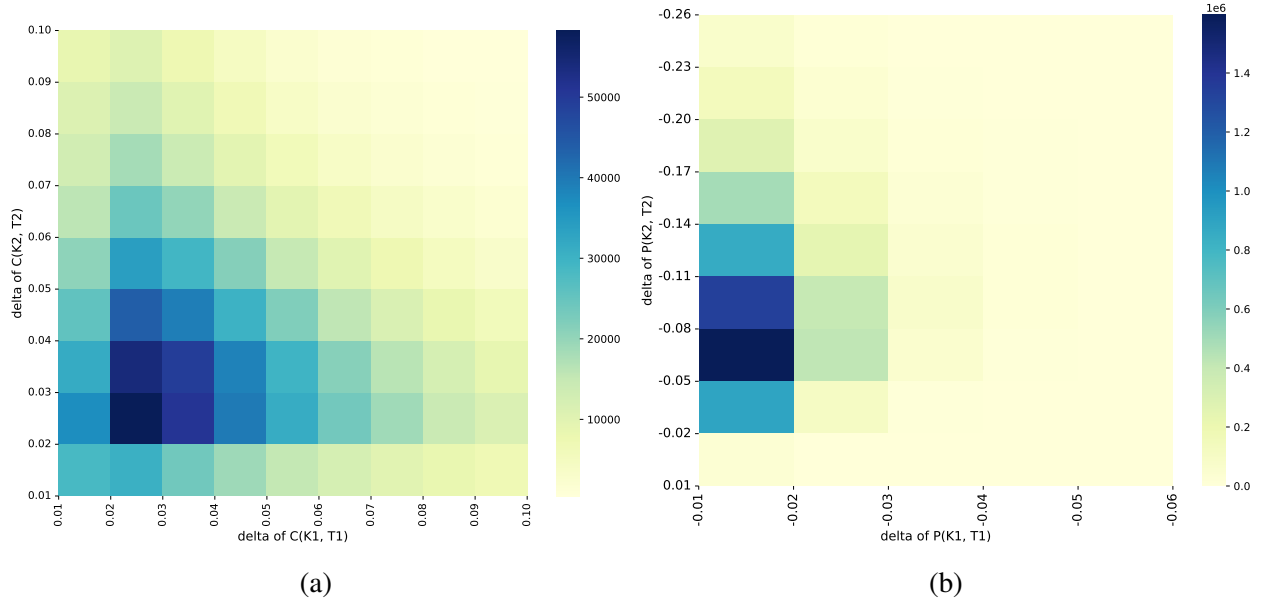


Figure 2.7: (a) Heatmap of count of TP₂-violating option deltas (b) Heatmap of count of RR₂-violating option deltas. x-axis: delta for (a) $C(K_1, T_1)$ or (b) $P(K_1, T_1)$; y-axis: delta for (a) $C(K_2, T_2)$ or (b) $P(K_2, T_2)$.

violations, the plot is almost symmetric, suggesting δ_1 and δ_2 are close to each other when forming the violations. For puts, however, it is the extreme tail $\delta_1 \in (-0.03, -0.01]$ that is paired with a range of different δ_2 values, the majority of which are less OTM than $P(K_1, T_1)$.

2.4.6 Violation Correction Rates

How soon do violation pairs correct? For a TP₂ or RR₂-violating pair, we consider the last violated day as the last trading day on which this option pair violates TP₂ or RR₂ conditions. If a violation pair does not violate in any future days, then the current day is the last violated day.

It is possible that a violating pair violates the TP₂ or RR₂ for a few days, stops violating for a few days, but violates again before maturity. We consider options in the intermediate time as “potentially violating”, as it may simply be the case that we do not have the price information. Instead, we consider the last time this option pair forms a violation before maturity, and after which the violation pair is permanently corrected.

Table 2.3 reports the violation correction results for option pairs with $T_1 = 3, 7, 14$ days and $T_2 \leq 60$ days. The majority of the violations are corrected very quickly – more than 83%, 77%,

and 50% are corrected permanently within one day for call violation pairs with $T_1 = 3, 7, 14$ days, respectively, and 82%, 63%, 55% for puts. By the day before maturity, more than 96% violations would have been corrected in all scenarios, for both calls and puts.

	Call, $T_1 =$			Put, $T_1 =$		
	3d	7d	14d	3d	7d	14d
T+1	16.19	22.86	48.94	17.95	36.59	44.73
T+2	2.73	16.44	42.74	3.01	28.41	37.61
T+3		11.25	40.20		22.71	35.57
T+4		6.72	31.88		10.13	28.08
T+5		3.45	25.84		4.53	21.14
T+6		1.07	19.54		0.58	17.50
T+7			14.62			12.89
T+8			8.12			6.39
T+9			7.31			4.85
T+10			5.83			3.71
T+11			4.48			1.99
T+12			2.64			1.18
T+13			0.89			0.04

Table 2.3: Violation survival rates (in %) for TP_2 and RR_2 violating option pairs with $T_1 = 3, 7, 14$ days and $T_2 \leq 60$ days.

2.5 Trading TP_2 and RR_2 Violations

We have seen that while TP_2 and RR_2 are rare, they do occur in empirical data. [39] established that these conditions are a slightly stronger form of the no static arbitrage conditions. In this section, we explore a long-short strategy that aims to profit from TP_2 and RR_2 violations. We first define our long-short trading strategy, then present trading performance from 2000 to 2022. As

the majority of the violations involve out-of-the money options, we also examine the short-only strategy, with the results included in Appendix B.3.3.

2.5.1 From TP_2 and RR_2 to Trading Strategies

The TP_2 and RR_2 conditions do not directly translate to straightforward trading strategies because each of the conditions in (2.3) and (2.4) involves two products of two options. To trade these violations, for each option value product, we interpret one option value as the number of contracts to be traded (or the coefficient) and the other as the actual option contract to be traded.

This interpretation yields four different trading combinations. For instance, to trade the left-hand side (LHS) of TP_2 violation in (2.3), we can either buy $C(K_1, T_1)$ shares of (K_2, T_2) -Call, or buy $C(K_2, T_2)$ shares of (K_1, T_1) -Call. Similarly, to trade the right-hand side (RHS), we can either buy $C(\tilde{K}_2, T_1)$ shares of (\tilde{K}_1, T_2) -Call, or buy $C(\tilde{K}_1, T_2)$ shares of (\tilde{K}_2, T_1) -Call where $\tilde{K}_1 = K_1 F_{T_2} / F_{T_1}$ and $\tilde{K}_2 = K_2 F_{T_1} / F_{T_2}$. Similarly, we have two possible interpretations for either side of the RR_2 condition for puts.

For a TP_2 (RR_2) violation, the relation is reversed in (2.3) and (2.4). We can think of it as the LHS being undervalued (overvalued) compared to the RHS. If the violation corrects itself, one can long the undervalued side and short the overvalued side.

In total, we have four different trading combinations, which we group based on the actual contract parameters. For TP_2 violations, we define

Strategy	Long Position	Contract	Short Position	Contract
\mathbf{T}_1 -denominated	$+C(K_2, T_2)$	(K_1, \mathbf{T}_1) -Call	$-C(K_1, T_2)$	(K_2, \mathbf{T}_1) -Call
\mathbf{T}_2 -denominated	$+C(K_1, T_1)$	(K_2, \mathbf{T}_2) -Call	$-C(K_2, T_1)$	(K_1, \mathbf{T}_2) -Call
\mathbf{K}_1 -denominated	$+C(K_2, T_2)$	(\mathbf{K}_1, T_1) -Call	$-C(K_2, T_1)$	(\mathbf{K}_1, T_2) -Call
\mathbf{K}_2 -denominated	$+C(K_1, T_1)$	(\mathbf{K}_2, T_2) -Call	$-C(K_1, T_2)$	(\mathbf{K}_2, T_1) -Call

Table 2.4: Strategies and positions.

For example, in the T_1 -denominated strategy, we long $C(K_2, T_2)$ shares of (K_1, T_1) -Call and

short $C(\tilde{K}_1, T_2)$ shares of (\tilde{K}_2, T_1) -Call. The trading strategies for RR_2 violations of the puts are defined analogously.

Each of these positions incurs a net cash premium to be collected when entering the contracts for TP_2 and RR_2 violating option pairs. For calls, the positive premium equals the difference between the cash dispensed from longing the LHS and the cash collected from shorting the RHS; for puts, the positive premium is collected from the short position in the LHS offset by the long position in the RHS.

To evaluate the profitability of a trade, we assume all positions are held until expiry. At expiry, we determine the payoff of the call or put options based on the index level. The above four positions will differ in the payoffs at expiration times T_1 and T_2 based on actual index level and option payoffs.

2.5.2 Per trade profits

If a TP_2 or RR_2 violation were to reverse, we expect to make positive returns from trading on TP_2 or RR_2 -violating pairs. We assume trades are executed at mid prices, defined as the halfway point between the best ask price and the best bid price, and analyze the per-trade profit of these positions. We will consider the full bid-ask spread in Section 2.5.5

Initial cash premium $\text{premium}_{V,t}$: Let $\text{premium}_{V,t}, V \in \{C, P\}$ denote the cash premium of entering the trade for a violation of option type V at time t . For TP_2 violation pairs, we have

$$\text{premium}_{C,t}(K_1, K_2, \tilde{K}_1, \tilde{K}_2, T_1, T_2) = -C_t(K_1, T_1)C_t(K_2, T_2) + C_t(\tilde{K}_1, T_2)C_t(\tilde{K}_2, T_1),$$

and for RR_2 violation pairs, we have

$$\text{premium}_{P,t}(K_1, K_2, \tilde{K}_1, \tilde{K}_2, T_1, T_2) = P_t(K_1, T_1)P_t(K_2, T_2) - P_t(\tilde{K}_1, T_2)P_t(\tilde{K}_2, T_1),$$

where $C_t(K, T)$ (or $P_t(K, T)$) denotes the value of a call (or put) option contract at time t with strike price K and day-to-maturity T (so the option expires at time $t + T$).

We note that $\text{premium}_{V,t}$ is positive because the option pair violates TP_2 or RR_2 conditions. In the mid-price setting, all trade denominations for the same TP_2 or RR_2 -violating pair will have the same $\text{premium}_{V,t}$.

Option payoff at maturity $\text{payoff}_{V,t}^D$: Each trade of the TP_2 (RR_2) violating pair involves a short and a long position in call (put) options. For call options, the payoff at expiry T is simply $[S_T - K]_+$, where S_T is the S&P closing level at time T for PM-settled options or SET price at time T for AM-settled options; for a put option, the payoff at expiry is $[K - S_T]_+$. Then $\text{payoff}_{V,t}^D$ is the option payoffs multiplied by the respective number of contracts traded.

Table 2.5 and Table 2.6 detail the payoffs at time $t + T_1$ and $t + T_2$ for different trading denominations of TP_2 and RR_2 violations, respectively, at time t .

Trade	Payoff at $t + T_1$	Payoff at $t + T_2$
T_1	$C_t(K_2, T_2)[S_{t+T_1} - K_1]_+ - C_t(\tilde{K}_1, T_2)[S_{t+T_1} - \tilde{K}_2]_+$	0
T_2	0	$C_t(K_1, T_1)[S_{t+T_2} - K_2]_+ - C_t(\tilde{K}_2, T_1)[S_{t+T_2} - \tilde{K}_1]_+$
K_1	$C_t(K_2, T_2)[S_{t+T_1} - K_1]_+$	$-C_t(\tilde{K}_2, T_1)[S_{t+T_2} - \tilde{K}_1]_+$
K_2	$-C_t(\tilde{K}_1, T_2)[S_{t+T_1} - \tilde{K}_2]_+$	$C_t(K_1, T_1)[S_{t+T_2} - K_2]_+$

Table 2.5: Cashflow for a TP_2 -violating trade entered at time t at time $t + T_1$ and $t + T_2$. In addition, at t , a positive cash of $\text{premium}_{C,t}$ is received by all four denominations.

Trade	Payoff at $t + T_1$	Payoff at $t + T_2$
T_1	$-P_t(K_2, T_2)[K_1 - S_{t+T_1}]_+ + P_t(\tilde{K}_1, T_2)[\tilde{K}_2 - S_{t+T_1}]_+$	0
T_2	0	$-P_t(K_1, T_1)[K_2 - S_{t+T_2}]_+ + P_t(\tilde{K}_2, T_1)[\tilde{K}_1 - S_{t+T_2}]_+$
K_1	$-P_t(K_2, T_2)[K_1 - S_{t+T_1}]_+$	$P_t(\tilde{K}_2, T_1)[\tilde{K}_1 - S_{t+T_2}]_+$
K_2	$P_t(\tilde{K}_1, T_2)[\tilde{K}_2 - S_{t+T_1}]_+$	$-P_t(K_1, T_1)[K_2 - S_{t+T_2}]_+$

Table 2.6: Cashflow for a RR_2 -violating trade entered at time t at time $t + T_1$ and $t + T_2$. In addition, at t , a positive cash of $\text{premium}_{P,t}$ is received by all four strategies.

In the analysis of per-trade profit, we assume interest is 0 and analyze the trade payoff by summing up the option payoffs at time $t + T_1$ and $t + T_2$. We call it $\text{payoff}_{V,t}^D$, where $D \in \{T_1, T_2, K_1, K_2\}$ represents the trade denomination. We will come back to Tables 2.5 and 2.6 when examining the performance of dynamic trading strategy in Section 2.5.3 and analyzing the trade performance in Section 2.5.4.

Normalization by gross cash exposure: If we consider the total amount of cash involved in forming a TP_2 or RR_2 trade, a trade of an ITM option pair will have a much larger cash exposure than that of an OTM pair. To put violating option pairs on the same scale, we normalize our trading position for each violating pair to \$1 cash exposure by dividing the sum of absolute values on both sides of (2.3) and (2.4). More formally, let $\mathcal{G}_{V,t}(K_1, K_2, \tilde{K}_1, \tilde{K}_2, T_1, T_2), V \in \{C, P\}$ denote the gross cash exposure of violating option pair V at time t . Then

$$\mathcal{G}_{V,t}(K_1, K_2, \tilde{K}_1, \tilde{K}_2, T_1, T_2) = V_t(K_1, T_1)V_t(K_2, T_2) + V_t(\tilde{K}_1, T_2)V_t(\tilde{K}_2, T_1).$$

Similar to $\text{premium}_{V,t}$, in the mid-price setting the gross exposure is the same for different trading denominations.

We compute the profits on each violation pair, assuming that (i) we hold the options to expiry, and (ii) the interest rate is 0, so that we can sum up the profits across different times. The total profits from D-denominated TP_2 or RR_2 -violating trade is thus the sum of the positive premium

received from TP₂ or RR₂ violation and the profits (or losses) from holding the options to maturity, or

$$\text{profit}_{V,t}^D = \frac{1}{G_{V,t}} [\text{premium}_{V,t} + \text{payoff}_{V,t}^D].$$

Table 2.7 reports the yearly average profits for different TP₂ trading denominations on \$1 cash exposure, and the numbers in brackets are the percentage of trades that yield positive returns. The first observation is that T_1 and K_2 -denominated strategies yield consistently large positive profits, and nearly all trades lead to positive gains in the time period, suggesting potential profits to be gained from TP₂ violations. Also, it is evident that T_2 and K_1 denominated strategies exhibit somewhat opposite behavior to those of T_1 and K_2 strategies, with consistently negative yearly returns. It is also worth noting that the best years for T_2 and K_1 strategies are the beginning of the year 2000 to 2002 and around 2008, coinciding with the period of the Great Recession in the early 2000 and the financial crisis in 2008.

Table 2.8 reports the results for RR₂ strategies. T_1 and K_1 denominated strategies yield consistently large positive profits, demonstrating robust performance across different market conditions. Conversely, the T_2 and K_2 strategies suffer severely, particularly in the years of the Great Recession and the 2008 financial crisis.

The differing behaviors of the trading combinations are rooted in the exposure of different denominations. For TP₂ violations, T_1 and K_2 -denominated options are inherently longing the index, whereas T_2 and K_1 denominated options are shorting the index. Neglecting the normalization factor, the T_2 -denominated strategy can be decomposed to (i) a short position of $C(\tilde{K}_2, T_1)$ shares in $(\tilde{K}_1, K_2; T_2)$ -CS, where $(K', K''; T)$ -CS is the Call Spread (CS) with strikes $K' < K''$ and day-to-expiry T^4 , and (ii) a long position of $C(K_1, T_1) - C(\tilde{K}_2, T_1)$ shares in (K_2, T_2) -Call. The Call Spread comprises a long position in the lower strike (K', T) -Call and a short position in the higher strike (K'', T) -Call, and is longing the underlying from K' to K'' . Because most violations are OTM calls, this decomposition shows that the position is negatively correlated with the index levels up to the

⁴To be more precise, we only have $K_1 < K_2$ but not $\tilde{K}_1 < K_2$, but empirical evidence suggests that the latter inequality holds almost always.

	$E[\text{profit}_C^{T_1}]$	$E[\text{profit}_C^{T_2}]$	$E[\text{profit}_C^{K_1}]$	$E[\text{profit}_C^{K_2}]$
2000	0.18 (100.0%)	0.14 (97.7%)	0.17 (98.7%)	0.15 (99.3%)
2001	0.24 (100.0%)	0.09 (95.0%)	0.19 (97.5%)	0.14 (100.0%)
2002	0.14 (100.0%)	0.08 (96.7%)	0.09 (95.9%)	0.13 (100.0%)
2003	0.22 (100.0%)	-0.17 (53.2%)	-0.49 (48.9%)	0.54 (96.4%)
2004	0.33 (100.0%)	-0.26 (57.7%)	-0.32 (57.9%)	0.39 (100.0%)
2005	0.42 (100.0%)	-0.30 (71.8%)	-0.29 (73.8%)	0.41 (100.0%)
2006	0.22 (100.0%)	-0.08 (76.5%)	-0.23 (75.9%)	0.36 (100.0%)
2007	0.50 (100.0%)	-0.45 (72.6%)	-0.63 (72.0%)	0.69 (100.0%)
2008	0.44 (100.0%)	0.20 (98.4%)	0.46 (98.4%)	0.18 (99.9%)
2009	0.28 (100.0%)	-0.14 (75.7%)	-0.62 (74.1%)	0.77 (100.0%)
2010	0.28 (100.0%)	-0.30 (74.7%)	-0.29 (75.1%)	0.27 (100.0%)
2011	0.21 (100.0%)	-0.00 (86.5%)	-0.08 (85.1%)	0.29 (100.0%)
2012	0.29 (100.0%)	-0.17 (74.5%)	-0.15 (77.1%)	0.28 (100.0%)
2013	0.40 (100.0%)	-0.26 (74.9%)	-0.21 (77.1%)	0.35 (99.9%)
2014	0.65 (99.7%)	-0.15 (81.9%)	0.23 (85.4%)	0.27 (99.1%)
2015	0.23 (100.0%)	-0.05 (89.5%)	-0.00 (91.0%)	0.18 (100.0%)
2016	0.43 (100.0%)	-0.30 (76.0%)	-0.11 (79.4%)	0.24 (99.9%)
2017	0.78 (99.9%)	-0.38 (61.8%)	-1.14 (60.1%)	1.53 (99.7%)
2018	0.14 (100.0%)	-0.01 (93.0%)	0.03 (94.5%)	0.10 (100.0%)
2019	0.32 (99.4%)	-0.56 (66.4%)	-0.59 (70.4%)	0.34 (99.7%)
2020	0.92 (98.8%)	-0.63 (64.2%)	-1.21 (60.7%)	1.49 (98.7%)
2021	0.22 (99.9%)	-0.75 (68.3%)	-1.35 (64.7%)	0.81 (99.9%)
2022	0.04 (99.6%)	0.02 (95.2%)	0.03 (96.7%)	0.04 (98.9%)

Table 2.7: Mean profit from TP_2 -violating trades of different denominations on \$1 cashness. Numbers in parentheses are the percentage of trades yield positive returns.

	$E[\text{profit}_P^{T_1}]$	$E[\text{profit}_P^{T_2}]$	$E[\text{profit}_P^{K_1}]$	$E[\text{profit}_P^{K_2}]$
2000	0.13 (100.0%)	-0.09 (64.4%)	0.51 (100.0%)	-0.47 (63.7%)
2001	0.08 (100.0%)	-0.20 (79.4%)	0.78 (100.0%)	-0.90 (72.0%)
2002	0.15 (100.0%)	-0.12 (81.4%)	0.46 (98.6%)	-0.43 (78.6%)
2003	0.10 (100.0%)	0.10 (99.8%)	0.10 (100.0%)	0.10 (99.8%)
2004	0.15 (100.0%)	0.13 (98.5%)	0.14 (100.0%)	0.14 (98.5%)
2005	0.15 (100.0%)	0.10 (96.5%)	0.15 (100.0%)	0.11 (96.5%)
2006	0.15 (100.0%)	0.08 (95.5%)	0.16 (100.0%)	0.08 (95.8%)
2007	0.16 (100.0%)	-0.05 (77.2%)	0.36 (100.0%)	-0.25 (75.7%)
2008	0.16 (99.8%)	0.27 (82.2%)	1.79 (96.7%)	-1.36 (76.7%)
2009	0.06 (100.0%)	0.05 (99.6%)	0.06 (100.0%)	0.05 (99.5%)
2010	0.21 (100.0%)	-0.01 (95.6%)	0.06 (98.8%)	0.14 (95.7%)
2011	0.15 (100.0%)	-0.13 (92.6%)	0.30 (99.4%)	-0.28 (91.5%)
2012	0.08 (100.0%)	0.02 (95.3%)	0.08 (100.0%)	0.02 (95.3%)
2013	0.06 (100.0%)	0.05 (99.4%)	0.06 (100.0%)	0.05 (99.4%)
2014	0.07 (100.0%)	0.04 (98.4%)	0.06 (100.0%)	0.05 (98.5%)
2015	0.10 (100.0%)	-0.01 (96.8%)	0.07 (100.0%)	0.01 (97.0%)
2016	0.05 (100.0%)	0.04 (99.8%)	0.04 (100.0%)	0.05 (99.9%)
2017	0.03 (100.0%)	0.03 (99.8%)	0.03 (100.0%)	0.03 (99.8%)
2018	0.39 (100.0%)	-0.28 (89.2%)	0.06 (99.4%)	0.06 (90.8%)
2019	0.05 (100.0%)	0.02 (99.1%)	0.06 (100.0%)	0.01 (99.0%)
2020	0.02 (99.6%)	0.03 (97.2%)	0.09 (99.4%)	-0.04 (96.9%)
2021	0.05 (100.0%)	-0.01 (98.1%)	0.03 (100.0%)	0.01 (98.1%)
2022	0.09 (99.9%)	-0.27 (84.5%)	0.10 (99.8%)	-0.28 (85.1%)

Table 2.8: Mean profit from RR_2 -violating trades of different denominations on \$1 cashness. Numbers in parentheses are the percentage of trades yield positive returns.

point $S_{T_2} = K_2$.

Similarly, the T_1 -denominated strategy can be decomposed into (i) a long position of $C(K_2, T_2)$ shares in $(K_1, \tilde{K}_2; T_1)$ -CS, and (ii) a short position of $C(\tilde{K}_1, T_2) - C(K_2, T_2)$ shares in (\tilde{K}_2, T_1) -Call. Using a similar argument, the T_1 -denominated strategy is longing the market up to $S_{T_1} = \tilde{K}_2$. This contrast leads to the almost opposite behaviour of T_1 and T_2 -denominated strategies. Because S&P index is mostly steadily increasing, T_1 -denominated strategies tend to lead to positive returns whereas those of T_2 suffer.

The analysis of K_1 and K_2 -denominated strategies is more complex, as now the options on the two sides of the violations expire at different times, potentially with different strikes. However, once time T_1 is reached and one of the options expires, K_1 -denominated strategy is exposed to an unprotected short position in a call option, whereas K_2 -denominated strategy holds a long position in a call option. This difference exposes K_1 -strategy to potential significant downside risk. We defer a more detailed analysis of the strategy performance to Section 2.5.4 where we break down the option positions in greater detail, analyzing the cashflow of different denominations for both TP_2 and RR_2 violations.

2.5.3 Dynamic Trading strategy

In this section, we transition from static per-trade profit to a dynamic version of our strategies in which we measure their performances from the beginning to the end period. We first define the strategy in Section 2.5.3, where further normalization is needed to make the strategy dynamic while maintaining certain risk exposure. We also introduce the notion of return, as our long-short trading strategy has no initial investment. We then report the performances in Sections 2.5.3 and 2.5.3. Enlightened by the per-trade performance in Section 2.5.2, we mainly focus our attention on T_1 and K_2 -denominated strategies for TP_2 -violating trades, and T_1 -denominated strategies for RR_2 -violating trades. We assume we hold the options until expiry. As a robustness check, we also consider exiting the option positions at the first possible time after entering into a TP_2 or RR_2 trade. We report the results in Appendix B.3.1 and note that we find similar patterns to our

hold-to-maturity strategies.

Strategy Description

As before, we assume the option value is the mid-price, and we assume we can trade the options at this price. The $\text{premium}_{V,t}$ from the previous section remains unchanged.

Normalization of the trade: In the previous section where we analyzed per-trade profits, we normalize each trade by the total cash exposure $\mathcal{G}_{V,t}$, or the sum of absolute values of the two sides of the condition. However, because our trading position may lead to unprotected short positions in options, the potential loss may be unbounded. We thus need to introduce additional normalization to control the total risk with our strategies.

Fix a violating option type V and a trading denomination $D \in \{T_1, T_2, K_1, K_2\}$. For simplicity, we omit these two parameters in our notations. Consider days $t = 0, \dots$ and for each day we index the violations by $i \in \{1, \dots, N_t\}$, where N_t denotes the total number of trades on day t . Each violation i comprises option parameter information $\mathcal{I}_{t,i} = (K_{i,1}, K_{i,2}, \tilde{K}_{i,1}, \tilde{K}_{i,2}, T_{i,1}, T_{i,2})$ and we consider normalization factor $\mathcal{F}_{t,i}$ comprised of 3 components:

- Gross cash exposure $\mathcal{G}_{t,i}$: the gross cash exposure for a violation pair on day t given by

$$\mathcal{G}_{t,i}(\mathcal{I}_{t,i}) = V_t(K_{i,1}, T_{i,1})V_t(K_{i,2}, T_{i,2}) + V_t(\tilde{K}_{i,1}, T_{i,2})V_t(\tilde{K}_{i,2}, T_{i,1});$$

- Number of trades on current day t , N_t : if we were to do many trades in a day, we want each trade to be smaller in size compared to those in a day when we do few trades. Because current-day N_t is not known until the end-of-day, we use the previous-day violation counts N_{t-} as a proxy for current-day N_t ; ⁵
- Leverage factor κ ,

⁵Note: we define “previous day” as the previous day with at least one violation. This is mostly a problem early in the period when we do not observe violations every day.

and $\mathcal{F}_{t,i} = 1/\kappa * N_{\zeta} * \mathcal{G}_{t,i}$. We then normalize each trade i by multiplying the traded quantity by $1/\mathcal{F}_{t,i}$. If we sum over all trades in one day, the product of gross exposure \mathcal{G} and the number of trades N normalize the total cash exposure to \$1, and the constant κ serves as a leverage factor, further normalizing our position based on our chosen risk appetite.

Return on day t : To analyze the performance of the trading strategies on each day, we assume we receive the profits from each option position at its expiry. We keep track of the positions we have and book the profits when entering into the trades (initial premium) or when options expire.

That is, we split the payoffs of one long-short trade $\text{payoff}_{t,i}$ in the previous section to $\text{payoff}_{t,i}^{T_1}$ and $\text{payoff}_{t,i}^{T_2}$, where the new superscript T_1 and T_2 denote the option payoff at time $t+T_1$ and $t+T_2$, respectively, as in Table 2.5 or 2.6. For every T_1 or T_2 -denominated trades, cashflow occurs at the inception and at $t + T_1$ or $t + T_2$, respectively, but not both $t + T_1$ and $t + T_2$. For K_1 or K_2 -denominated trades, cashflow may occur at the inception and at both $t + T_1$ and $t + T_2$.

The return (on cashness) of day t , r_t , is then

$$r_t = \kappa \left[\sum_i \frac{1}{N_{\zeta}} \frac{\text{premium}_{t,i}}{\mathcal{G}_{t,i}} + \sum_{t' < t} \frac{1}{N_{\zeta}} \sum_{i', T_1: t'+T_1=t} \frac{\text{payoff}_{t',i'}^{T_1}}{\mathcal{G}_{t',i'}} + \sum_{t' < t} \frac{1}{N_{\zeta}} \sum_{i', T_2: t'+T_2=t} \frac{\text{payoff}_{t',i'}^{T_2}}{\mathcal{G}_{t',i'}} \right], \quad (2.5)$$

where the first term in the bracket is the cash premium received from day t trades, the second term is the payoffs of existing trade whose the earlier expiring option expires today, and the last term is the payoffs of the existing trade whose later expiring option expires today. For T_1 -denominated strategy, the third term is always 0; for T_2 , the second term is always 0.

The total profit and loss (P&L) of data t is then r_t multiplied by the initial cash at day t . We remark that as we are trading options and from our trading strategies, we may be left with an unprotected short position in options, the potential loss may be unbounded. Therefore, the leverage ratio κ needs to be sufficiently small to control our risk. In the result shown in Sections 2.5.3 and 2.5.3, we choose our leverage ratio κ such that the volatility of our trading strategy is half of that of the historical S&P 500 index return volatility.

Performance of TP₂-Violating Strategy

Yearly returns: Table 2.9 shows the yearly returns (in %) of different denomination TP₂ violation trades. The corresponding yearly returns on the S&P 500 index are included in the last column for reference. We observe consistent double-digit returns for T_1 -denominated strategy, outperforming that of the index return. We also observe that T_1 -denominated strategy yields positive returns across all years, even during times of the financial crises.

Year	T_1	T_2	K_1	K_2	S&P
2000	4.63	2.56	1.74	2.73	-10.14
2001	3.26	0.56	1.02	1.09	-13.04
2002	1.80	-0.09	-0.04	1.22	-23.37
2003	0.36	-0.90	-0.56	0.33	26.38
2004	8.46	-2.14	-0.26	4.17	8.99
2005	18.53	-9.15	-1.89	6.41	3.00
2006	10.28	-8.02	-5.30	8.28	13.24
2007	24.65	-7.53	-4.43	16.78	3.53
2008	24.78	9.97	8.70	11.24	-38.49
2009	17.01	-15.33	-19.22	32.95	23.45
2010	23.63	-16.76	-6.02	8.65	12.78
2011	14.25	-1.37	-2.56	13.00	-0.00
2012	14.27	-13.40	-7.94	9.94	13.41
2013	57.32	-21.41	-4.56	18.96	29.60
2014	53.26	-10.77	7.15	5.95	11.39
2015	9.24	-1.08	0.40	4.56	-0.73
2016	24.61	-9.79	-1.75	8.95	9.54
2017	23.00	-18.30	-10.27	14.17	19.42
2018	63.29	9.16	-3.84	61.21	-6.24
2019	47.42	-24.48	-7.61	15.31	28.88
2020	34.58	-53.76	-48.00	55.79	16.26
2021	36.73	-12.22	-15.54	45.10	26.89
2022	2.36	-0.55	-0.18	1.37	-19.44

Table 2.9: Yearly returns of TP₂ strategies of different denominations (numbers are in percent).

Cumulative returns: We also examine the cumulative returns for different trading strategies from 2000 to 2022. The results are plotted in 2.8. The solid black lines represent the corresponding S&P 500 index levels. The solid red lines are the cumulative returns of T_1 -denominated strategy

on the left and K_2 -denominated strategy on the right. The dashed red lines on both plots represent the cumulative returns excluding option payoffs at maturity, or the cumulative returns of the cash premiums received from TP_2 violations assuming all positions expire OTM. The y-axis on both plots is plotted in the log scale. We observe that T_1 -denominated strategy yields significantly higher profits compared to simply longing the index. K_2 -denominated returns also generate substantially better returns than the S&P index, although to a lesser extent than T_1 -denominated returns.

The returns from TP_2 violation cash premiums show a smooth, steady increase, consistent with the constant inflow of cash premiums from TP_2 trades. However, the rate of increase is slower than that of the total returns; this difference comes from our trading positions that expire ITM and make profits – or the jumps in the solid red lines – and is consistent with our analysis in Section 2.5.4 that a steadily increasing S&P 500 index benefits both the T_1 and K_2 -denominated hold-to-maturity strategies.

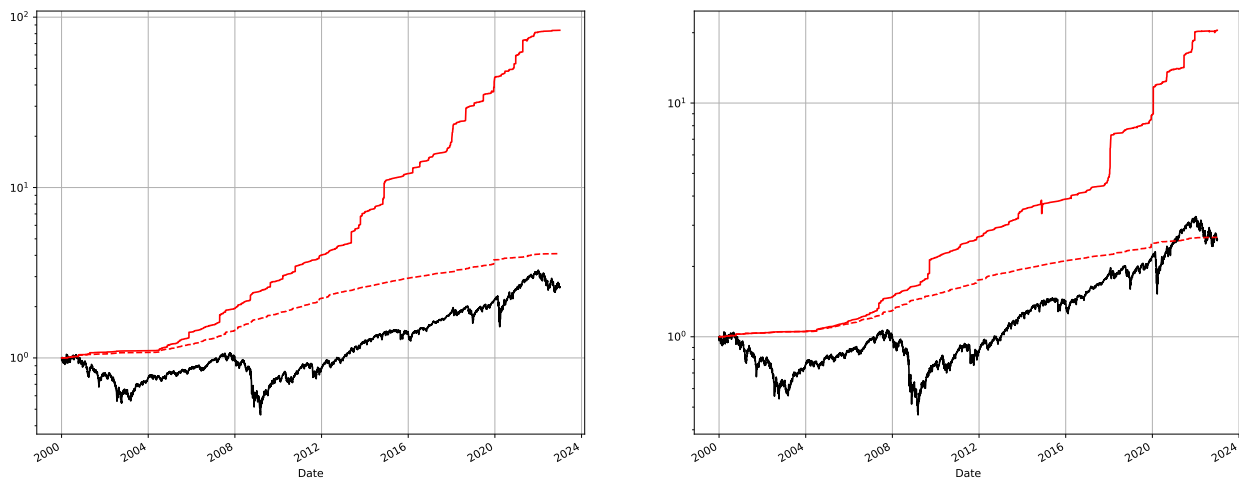


Figure 2.8: Cumulative profits in log-scale. Left: T_1 -denominated TP_2 trades. Right: K_2 -denominated TP_2 trades. The black lines are the S&P 500 index. The solid and dashed red lines are cumulative total profits and cash premiums, respectively.

We further analyze the trading performance of the best-performing T_1 -denominated strategy. We calculate the annualized Sharpe Ratio for monthly T_1 -denominated returns from 2004 to 2022. We discard the results for the years 2000 to 2003 because very few trades occur in this period and most of the gains are from the cash premium as can be seen from Figure 2.8. The results are

reported in Table 2.10, including the Sharpe Ratio of the monthly S&P 500 index return in the same year for comparison. We observe higher Sharpe Ratios for T_1 -denominated returns compared to those of S&P index in most years, including the years when S&P has a negative Sharpe Ratio. In many instances, the Sharpe ratio for T_1 -denominated returns exceeds 2.

Year	Sharpe T_1	Sharpe SP
2004	2.39	1.06
2005	1.88	0.02
2006	3.29	1.33
2007	1.78	-0.04
2008	2.83	-2.23
2009	2.70	1.05
2010	2.71	0.71
2011	3.87	0.07
2012	5.71	1.24
2013	2.58	3.10
2014	1.82	1.37
2015	4.82	0.01
2016	2.42	0.92
2017	3.02	4.38
2018	1.70	-0.47
2019	1.75	1.88
2020	2.33	0.69
2021	1.97	2.22
2022	0.56	-0.91

Table 2.10: Annualized Sharpe Ratio for monthly T_1 -denominated returns and S&P 500 index returns.

Table 2.11 regresses T_1 -denominated monthly trading returns against (i) monthly S&P returns, and (ii) monthly S&P returns and average CBOE's VIX levels in the month. As expected, T_1 -denominated returns are positively correlated with S&P returns, consistent with the payoff structure that it is inherently a long position in the market in a given range of the underlying movement. The level of the VIX index negatively impacts T_1 -denominated returns, indicating that T_1 -denominated strategy tends to do well when the market is forecasted to be calm, although this effect is less pronounced than the returns on the S&P index. Overall, S&P returns and the VIX index explain only a small fraction of T_1 -denominated trading returns.

	(i)	(ii)
const	0.015*** (0.002)	0.024*** (0.006)
SP_Return	0.147*** (0.047)	0.124*** (0.043)
CBOE's VIX		-0.041* (0.024)
R-squared	0.034	0.043
R-squared Adj.	0.031	0.036

Table 2.11: Regress monthly T_1 -denominated returns on (i) monthly S&P returns only, and (ii) monthly S&P returns and average CBOE's VIX closing levels in the month.

Performance of RR_2 -Violating Strategy

In this section, we analyze the performance of the RR_2 -violating strategies. We apply the same methodology used for TP_2 -violating strategies to evaluate the RR_2 -violating trades from 2000 to 2022.

Table 2.12 reports the yearly returns of trades of different trading denominations. We observe T_1 -denominated trades outperform S&P index in the majority of the years with double-digit returns. K_1 -denominated strategy also receives net positive yearly returns except for the year 2020, and in 2008 it achieves 58.6% returns. However, in other years K_1 -denominated trades perform worse than T_1 -denominated trades.

In contrast, T_2 and K_2 -denominated strategies suffer losses in many years, although they achieved large positive gains in the period of the 2008 Financial Crisis where S&P 500 index drops almost 40%. These behaviors are consistent with our cashflow analysis later in Section 2.5.4 that these two strategies benefit from a sharp decline in the underlying index levels.

We plot the total T_1 -denominated cumulative returns (solid red line) and those composed of only cash premiums from RR_2 violations (dashed red line) in Figure 2.9. Similar to TP_2 trades, T_1 -

Year	T_1	T_2	K_1	K_2	S&P
2000	0.60	0.19	0.11	0.07	-10.14
2001	0.29	0.34	2.68	-3.48	-13.04
2002	0.11	-0.07	0.14	-0.19	-23.37
2003	1.91	1.11	0.38	0.35	26.38
2004	5.81	3.44	0.98	1.30	8.99
2005	12.02	0.76	1.60	0.97	3.00
2006	14.41	2.95	1.72	2.14	13.24
2007	11.33	-1.74	1.17	0.51	3.53
2008	22.33	40.93	58.59	-50.13	-38.49
2009	5.66	3.43	1.17	1.00	23.45
2010	29.75	-4.21	1.48	3.70	12.78
2011	18.62	-0.47	5.99	-3.87	-0.00
2012	7.49	2.24	1.49	0.58	13.41
2013	6.67	3.60	1.12	1.37	29.60
2014	9.77	3.14	1.28	1.72	11.39
2015	8.37	-4.43	1.47	-1.56	-0.73
2016	12.10	1.17	1.08	1.86	9.54
2017	3.96	2.45	0.66	0.94	19.42
2018	57.47	-25.66	1.83	-1.31	-6.24
2019	5.02	-0.33	1.21	-0.54	28.88
2020	-1.66	-0.94	-1.75	1.17	16.26
2021	4.57	0.95	0.57	0.68	26.89
2022	37.21	-16.57	1.91	-0.51	-19.44

Table 2.12: Yearly returns of RR_2 strategies of different denominations (numbers are in percent).

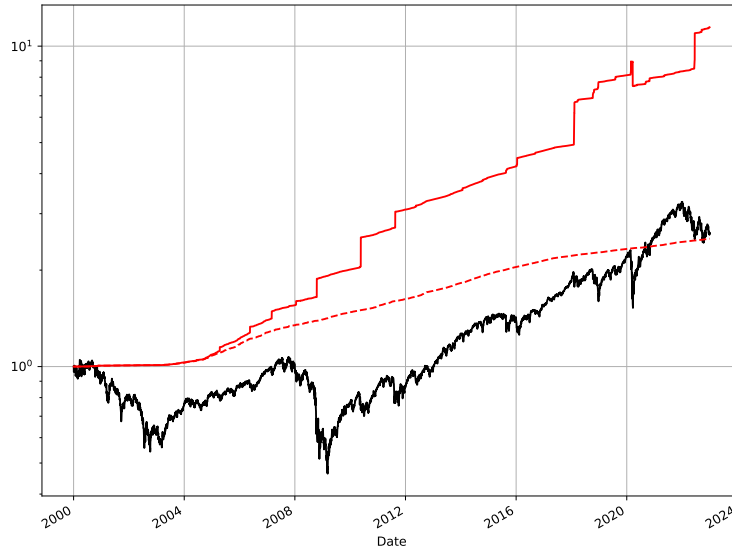


Figure 2.9: Cumulative profits in log-scale. The black line is the S&P 500 index. The solid and dashed red lines are cumulative total profits and cash premiums, respectively.

denominated strategy outperforms S&P, and the option maturity payoffs (the difference between the total profits and RR_2 violation premiums) constitute a significant part of the overall strategy. We also observe large jumps of the total strategy returns during periods of some S&P 500 index declines, but one significant drop in 2020 when S&P crashes. From the analysis in Section 2.5.4. These jumps are from the modest size drops that lead to positive option payoffs; the drop is when S&P 500 index dips beyond the threshold which lead to a loss in the portfolio.

Table 2.13 reports the Sharpe ratio for T_1 -denominated returns for RR_2 -violating puts. Similar to TP_2 violations, the T_1 -denominated shape ratios are consistently higher than those of S&P index returns. We regress T_1 -denominated returns on S&P returns and market VIX, with the results presented in Table 2.14. T_1 -denominated returns are now negatively correlated with S&P returns, but similar to the case of TP_2 violations, S&P returns and VIX levels account for only a small fraction of variability in T_1 -denominated returns.

2.5.4 Cashflow Analysis

We come back to the definition of different trading strategies and analyze the associated positions and P&Ls. Tables 2.5 and 2.6 detail the payoffs of different trading denominations of TP_2 and

Year	T_1 Returns	SP Returns	Sharpe T_1	Sharpe SP
2004	5.81	8.99	4.07	1.06
2005	12.01	3.00	3.23	0.02
2006	14.40	13.24	2.03	1.33
2007	11.32	3.53	1.24	-0.04
2008	22.32	-38.49	1.43	-2.23
2009	5.65	23.45	15.12	1.05
2010	29.73	12.78	1.18	0.71
2011	18.59	0.28	1.48	0.09
2012	7.48	13.41	6.75	1.24
2013	6.66	29.60	12.30	3.10
2014	9.77	11.39	6.74	1.37
2015	8.37	-0.73	3.51	0.01
2016	12.09	9.54	1.92	0.92
2017	3.94	19.55	6.93	4.39
2018	57.42	-7.25	1.39	-0.51
2019	4.99	29.11	1.40	1.89
2020	-1.66	16.26	-0.06	0.69
2021	4.56	26.89	4.78	2.22
2022	37.18	-19.44	1.16	-0.91

Table 2.13: Yearly returns for T_1 -denominated RR_2 trades and S&P 500 index and annualized Sharpe Ratio for monthly T_1 -denominated returns and S&P index returns.

	(i)	(ii)
const	0.010*** (0.002)	0.008 (0.006)
S&P Return	-0.182** (0.075)	-0.178*** (0.063)
CBOE's VIX		0.009 (0.033)
R-squared	0.056	0.057
R-squared Adj.	0.053	0.050

Table 2.14: Regress monthly RR_2 -violating T_1 -denominated returns on (i) monthly S&P returns only, and (ii) monthly S&P returns and average VIX closing levels in the month.

RR₂-violating pairs at option maturing times. At each future trading day t , all strategies receive payoffs from three distinct sets of positions:

- \mathcal{S}_0 : premium from new contracts;
- $\mathcal{S}_{T_1=t}$: cashflow from violation pair whose earlier expiring option expires today;
- $\mathcal{S}_{T_2=t}$: cashflow from violation pair whose later expiring option expires today.

The cashflow from \mathcal{S}_0 is always positive. We analyze the cashflow for the options in the other two sets for each of the trading denominations. For the analysis, we assume options are issued at time 0 and we assume $K_1 < \tilde{K}_2$ and $\tilde{K}_1 < K_2$ ⁶.

TP₂ payoffs

T -denominated strategies: For T_1 -denominated strategy, both options in a TP₂-violating pair expire at the earlier expiration T_1 , so the payoffs from $\mathcal{S}_{T_2=t}$ is 0, and we are left analyzing the cashflow from $\mathcal{S}_{T_1=t}$.

Consider the payoff as a function of varying index levels at T_1 . If $S_{T_1} < K_1$, both options expire worthless, and the payoff from $\mathcal{S}_{T_1=t}$ is 0. The payoff increases when S_{T_1} increases from K_1 to \tilde{K}_2 , reaches its maximum at $S_{T_1} = \tilde{K}_2$, before decreases when S_{T_1} further increases. T_1 -denominated strategy will incur a *negative* cash flow from $\mathcal{S}_{T_1=t}$ if and only if

$$S_{T_1} > \tilde{K}_2 + C(K_2, T_2) \frac{\tilde{K}_2 - K_1}{C(\tilde{K}_1, T_2) - C(K_2, T_2)}.$$

That is, as long as the index level S_{T_1} does not increase beyond \tilde{K}_2 plus some buffer on the RHS, T_1 -denominated violating pair will have positive profits from $\mathcal{S}_{T_1=t}$; even if the index level increases beyond the RHS, the total cash flow today can still be positive because of the positive cash premiums from \mathcal{S}_0 .

⁶Empirically we observe $K_1 < K_2 F_{T_1}/F_{T_2}$ and $K_2 < K_1 F_{T_2}/F_{T_1}$ almost always, as the majority of the violations have short time-to-maturity, and the ratio F_{T_1}/F_{T_2} is close to 1.

We have seen in Figure 2.7 that most violations comprise out-of-the-money options, or $S_0 < K_1$. T_1 -denominated strategy is thus a long position in the S&P index up to the point \tilde{K}_2 . This explains why we observe steady, consistent growth in T_1 -denominated returns in Figure 2.8.

For T_2 -denominated strategy, the payoffs from $\mathcal{S}_{T_1=t}$ is 0 and we only need to focus on $\mathcal{S}_{T_2=t}$. If $S_{T_2} < \tilde{K}_1$, both options expire worthless. As the index level increases from \tilde{K}_1 to K_2 , the payoff at T_2 decreases. The payoff from $\mathcal{S}_{T_2=t}$ only starts to increase again when S_{T_2} increases beyond K_2 . The payoff from $\mathcal{S}_{T_2=t}$ is positive if and only if

$$S_{T_2} > K_2 + C(\tilde{K}_2, T_1) \frac{K_2 - \tilde{K}_1}{C(K_1, T_1) - C(\tilde{K}_2, T_1)},$$

or when there is a jump in the index levels from the inception of the trade to T_2 . We therefore observe that T_2 -denominated strategy generally benefits from a decline in the S&P 500 index as the payoffs from \mathcal{S}_0 is positive, but suffers when the index is increasing and the positive cash premium does not compensate for the losses from $\mathcal{S}_{T_2=t}$.

K -denominated strategies: We first examine K_2 -denominated strategy. For K_2 -denominated strategies, $\mathcal{S}_{T_2=t}$ brings potential profits and $\mathcal{S}_{T_1=t}$ potential loss. A steady, fast increase in the underlying brings the highest profits. To see this, as T_2 is larger than T_1 , K_2 can be deep OTM for options in $\mathcal{S}_{T_2=t}$ as those were issued relatively long ago, bringing in positive payoffs from $\mathcal{S}_{T_2=t}$, but near-the-money or even ITM for options in $\mathcal{S}_{T_1=t}$ and rendering the payoffs from $\mathcal{S}_{T_1=t}$ less negative. Together, the positive cashflows from \mathcal{S}_0 and $\mathcal{S}_{T_2=t}$ dominate the negative cashflows from $\mathcal{S}_{T_1=t}$, yielding positive returns.

Now for K_1 -denominated strategies, $\mathcal{S}_{T_1=t}$ brings potential profits and $\mathcal{S}_{T_2=t}$ potential loss. Now an upward trend in S&P 500 index is detrimental to K_1 -denominated strategy because now options in $\mathcal{S}_{T_2=t}$ bring negative cashflow that cannot be offset by the positive cashflows from \mathcal{S}_0 and $\mathcal{S}_{T_1=t}$.

RR₂ payoffs

T-denominated strategies: For T_1 -denominated strategy, we again analyze the payoffs from $\mathcal{S}_{T_1=t}$ because payoffs from $\mathcal{S}_{T_2=t}$ is 0. When $S_{T_1} > \tilde{K}_2$, both options in the RR₂-violating pair expires worthless, and the payoff is 0. As S_{T_1} decreases from \tilde{K}_2 , the payoff from $\mathcal{S}_{T_1=t}$ increases, reaching its maximum when $S_{T_1} = K_1$, before decreases when S_{T_1} decreases further. Therefore, T_1 -denominated will incur a *negative* cashflow for options in $\mathcal{S}_{T_1=t}$ if and only if

$$S_{T_1} < K_1 - P(\tilde{K}_1, T_2) \frac{\tilde{K}_2 - K_1}{P(K_2, T_2) - P(\tilde{K}_1, T_2)}.$$

Because most violations are OTM, or $S_0 > K_1$, T_1 -denominated strategy makes profits when the index is increasing or flat, and makes the greatest profit when the index is modestly decreasing. Only when the index dip is so steep and below the RHS of the equation that we suffer from a loss, just as in the case of 2020 in Figure 2.9.

For T_2 -denominated strategy, we study the payoff from $\mathcal{S}_{T_2=t}$. If $S_{T_2} > K_2$, both options expire worthless. As the index level decreases from K_2 to \tilde{K}_1 , the payoff from $\mathcal{S}_{T_2=t}$ becomes increasingly negative. The payoff only starts to increase when S_{T_2} decreases further below \tilde{K}_1 . The payoff from $\mathcal{S}_{T_2=t}$ is positive if and only if

$$S_{T_2} < \tilde{K}_1 - P(K_1, T_1) \frac{K_2 - \tilde{K}_1}{P(\tilde{K}_2, T_1) - P(K_1, T_1)},$$

or when there is a sharp decline in the index levels from the inception of the trade to T_2 . We therefore observe that T_2 -denominated strategy brings profits when the S&P 500 index is flat or increases, but generally suffers if is decreasing, unless the dip is so large as is the case in the 2008 Financial Crisis.

K-denominated strategies: A steady, fast downward trend in the S&P index level benefits K_1 -denominated trades but hurt K_2 -denominated trades. This is because K_2 can be OTM for options in $\mathcal{S}_{T_2=t}$, but near-the-money or even ITM for options in $\mathcal{S}_{T_1=t}$. However, violations are much

more likely for OTM options, and the cashflow from $\mathcal{S}_{T_2=t}$ dominates that from $\mathcal{S}_{T_1=t}$. For K_1 -denominated strategy, the cashflow for $\mathcal{S}_{T_2=t}$ is positive, and that for $\mathcal{S}_{T_1=t}$ is negative, so the overall the strategy benefits from the fast steady decline in the index level. For K_2 -denominated strategy, the sign is reversed the the downward trend is harmful for the profits.

2.5.5 Trading Costs Considerations

Options' true values are not observable; we only see the best bid and best offer prices. Previous sections have assumed that mid-price is a fair proxy for option values, and that the contract is sufficiently liquid to trade both sides at the mid-price.

So far we have used mid-price to (i) determine violations when checking TP_2 conditions in (2.3) and RR_2 conditions in (2.4), and (ii) calculate trade profit by assuming we can trade both sides at the mid-price. We have *partly* relaxed the second condition in the trading strategy analysis where we plot the positive initial cash premium in Figures 2.8 and 2.9. We conclude that even without the initial premium, the option payoffs outperform the market. This suggests that if trading costs are smaller than the initial cash premium received from entering TP_2 trades, then T_1 and K_2 -denominated strategies are still profitable.

In this section, we experiment with a more conservative setting where we consider the TP_2 or RR_2 condition violated only when the most conservative interpretation of (2.3) and (2.4) is violated. Specifically, a call option pair violates TP_2 if

$$C^A(K_1, T_1)C^A(K_2, T_2) < C^B(\tilde{K}_1, T_2)C^B(\tilde{K}_2, T_1)$$

and a put option pair violates RR_2 if

$$P^B(K_1, T_1)P^B(K_2, T_2) > P^A(\tilde{K}_1, T_2)P^A(\tilde{K}_2, T_1)$$

where the superscript A, B denote best ask and best bid prices. respectively. By adopting this more restrictive definition of the TP_2 or RR_2 condition, we effectively trade a subset of the original TP_2

violations, averaging a greater degree of violation (differences between the two sides) if mid-price is used, but we can only trade them at a worse price.

Table 2.15 shows the per-trade profits for TP_2 and RR_2 violations from 2014 to 2022. Years before 2014 are discarded due to the limited number of violation pairs observed before the increase in trading volumes in 2014 under the stricter TP_2 and RR_2 conditions. In all years, we consistently observe strong average positive returns per \$1 cashness for both TP_2 and RR_2 trades. For RR_2 -violating pairs, the average profit per \$1 cashness is approximately 10 cents across the years. For TP_2 -violating pairs, the average profit varies more widely but still consistently averages more than 10%. Remarkably, 100% of the trades yield positive profits for both calls and puts, even after accounting for the full bid-ask spread as trading costs. This analysis further demonstrates that violations of TP_2 and RR_2 conditions can lead to significant arbitrage opportunities.

Year	$E[\text{profit}_C^{T_1}]$	$E[\text{profit}_P^{T_1}]$
2014	0.26 (100.0%)	0.08 (100.0%)
2015	0.26 (100.0%)	0.09 (100.0%)
2016	0.30 (100.0%)	0.10 (100.0%)
2017	1.31 (100.0%)	0.05 (100.0%)
2018	0.16 (100.0%)	0.27 (100.0%)
2019	0.46 (100.0%)	0.07 (100.0%)
2020	2.11 (100.0%)	0.07 (100.0%)
2021	0.10 (100.0%)	0.08 (100.0%)
2022	0.09 (100.0%)	0.07 (100.0%)

Table 2.15: Per-trade profit of T_1 -denominated trades for TP_2 and RR_2 violations with trading costs. Numbers in brackets are the percentage of trades receiving positive profits.

2.6 Concluding Remarks

We examine the empirical significance of TP_2 and RR_2 properties in S&P 500 options, using data from 2000 to 2022. Our analysis reveals that these conditions are largely upheld, with violation rates seldom exceeding 6%, and that put options are more prone to violations, particularly during periods of elevated market risk.

A key contribution of this research is the development of a long-short trading strategy based on TP_2 and RR_2 violations. This strategy has consistently yielded positive returns, outperforming the S&P 500 index by a substantial margin with lower volatility for both puts and calls. The per-trade profits are almost always positive, providing empirical evidence that these conditions may be viewed as some slightly stronger no-arbitrage conditions. These findings highlight the practical potential of TP_2 and RR_2 properties for developing profitable trading strategies.

While our focus has been on a long-short strategy to fully replicate TP_2 and RR_2 violations, future research could explore alternative approaches. For instance, as examined in Appendix B.3.3, a strategy centered on solely shorting overvalued options and collecting premiums might offer additional benefits and could be particularly effective in certain market conditions. Additionally, the CBOE provides FLEX Options, allowing option contracts to be customized, including strike prices and expiration dates. This flexibility may create more violations and trading opportunities. Moreover, our use of end-of-day data provides robust insights but suggests another avenue for future work: employing intraday data could allow for quicker adjustments to positions, potentially further enhancing the profitability of the trading strategy.

In summary, our study contributes to the understanding of TP_2 and RR_2 properties in option pricing and demonstrates their practical utility in trading strategies. Future research exploring alternative strategies and using more granular data could further uncover the potential of these properties in financial markets.

Chapter 3: Fairness in Regulatory Stress Tests

Regulatory stress tests are a critical tool for assessing the capital adequacy of banks and ensuring financial stability. However, the application of a uniform stress test models across diverse banks raises questions of fairness and accuracy. This chapter addresses the challenge of fair aggregation of individual models into a common industry model. We explore various notions of regression fairness to balance forecast accuracy and equal treatment, proposing methods to enhance the reliability and fairness of stress tests. By ensuring that stress tests accurately and equitably reflect the risks faced by different banking institutions, this research contributes to more equitable and precise regulatory practices, supporting the stability of the financial system.

3.1 Introduction

In the aftermath of the 2008 financial crisis, U.S. banking regulators adopted stress testing as a primary tool for monitoring the capital adequacy of the largest banks. For each round of annual stress tests, the Federal Reserve announces a “severely adverse stress scenario,” defined by a hypothetical path of economic variables over the next several quarters. A typical path includes an increase in unemployment, a decline in GDP, and projections for the level and volatility of the stock market, among other variables. The largest banks provide the Fed with detailed information about their loan portfolios and other assets. The Fed then applies internally developed models to project revenues and losses for each bank through the stress scenario. Banks are required to have sufficient capital to weather the projected losses.

The Fed does not disclose details of the models it uses to project revenues and losses. The Fed makes clear that to ensure consistent treatment for different banks it uses “industry models,” as opposed to models tailored to individual banks. As a matter of policy, the same models are applied

to all banks. Quoting Board of Governors [11] (p.3), “two firms with the same portfolio receive the same results for that portfolio.” We will refer to this statement as the Fed’s principle of equal treatment.

Banks have countered that the Fed’s models fail to capture bank-specific features that could lower projected losses. They have made these arguments in requests for reconsideration of stress test results. Of course, banks are not objective critics of the Fed’s supervision; but significant heterogeneity among the largest banks is indisputable. The banks subject to annual stress testing include universal banks, investment banks, large regional banks, the U.S. subsidiaries of certain foreign banks, and a variety of more specialized financial firms. It is certainly possible that bank-specific models would produce more accurate forecasts than a single industry model, in which case using a single model entails a trade-off between forecast accuracy and consistency across banks. Indeed, a recent industry article (Baer and Hopper [7]) argues that “banks’ own models are trained on each bank’s specific experience, rather than relying on a one-size-fits-all assumption," and that “[t]here is a strong argument that banks’ own models generate far more accurate results than the Fed’s”.

The heterogeneity among large banks motivates the questions we study: What is the best way to aggregate bank-specific models into an industry model? How should the Fed’s principle of equal treatment be interpreted and implemented? Is simply ignoring bank identity in estimating and applying models the best way to achieve fairness? To what extent is fairness at odds with accuracy? Although the heterogeneity of large banks is widely recognized, we know of no prior work that seeks to address this property within the constraints of the Fed’s policy of equal treatment. We will argue that addressing heterogeneity is preferable to ignoring it.

The question of fairness in algorithms and models has received a great deal of renewed interest in recent years, in some cases reviving earlier debates over fairness in testing and related policies that were not explicitly “algorithmic;” see, for example, the overviews in Barocas, Hardt, and Narayanan [8] and Hutchinson and Mitchell [47]. We draw on this literature, but our setting differs in important ways from most discussions of fairness.

Algorithmic fairness is usually concerned with ensuring that certain protected attributes — race or gender, for example — do not influence outcomes such as hiring decisions or loan approvals. Different methods can be compared based on alternative measures of influence and the degree to which sensitive attributes are indeed protected.

The counterpart of a protected attribute in our setting is a bank’s identity; but this attribute is not so much protected (in the sense that race and gender are) as inadmissible for the Fed’s purpose. In stating that “two firms with the same portfolio receive the same results for that portfolio,” the Fed is stating that bank identity is not a legitimate predictor of losses. Perhaps, then, fairness is achieved as long as the Fed uses the same model for all banks. In other words, perhaps “fairness through unawareness,” paraphrasing Dwork et al. [23], is sufficient in this setting. Moreover, in questioning whether the Fed’s models apply to them, banks are not claiming discrimination; on the contrary, they are asking for discrimination — asking that the Fed change its models to recognize ways in which an individual bank differs from other banks.

To investigate these issues, we focus primarily on a simple setting in which the “true” loss rate for each bank is described by a bank-specific regression on portfolio features and scenario features. The regulator’s goal is to aggregate these bank-specific models into a single model. A natural interpretation of an “industry” model in this setting is a pooled regression based on combining results across banks. The pooled model treats banks equally, but we show that it has at least two significant deficiencies: when applied to heterogeneous banks, it can produce poor measures of the marginal impact of individual features, even resulting in the wrong sign; and it implicitly misdirects legitimate information in portfolio features to infer (or proxy for) bank identity in forecasting losses. The second of these deficiencies works against the spirit of equal treatment of banks, even if bank identity is not explicitly used in the model.

We then investigate the application of ideas from algorithmic fairness in our setting. The fairness literature has mainly focused on classification problems (hiring decisions and credit approvals, for example), with regression problems getting somewhat less attention. Chzhen et al. [16] and Le Gouic et al. [9] developed a method of particular importance for regression that Le Gouic et al. [9]

call “projection to fairness.” This method produces optimal forecasts (in the least-squares sense) subject to a fairness constraint known as *demographic parity*. We examine the application of this approach in our setting and conclude that it goes too far in leveling results across banks.

The pooled method ignores fairness and the projection method goes too far in imposing fairness, so we seek an intermediate solution. Johnson, Foster, and Stine [48] introduce a variety of methods for introducing fairness considerations in regression. These include methods they call “full equality of opportunity” (FEO) and “substantive equality of opportunity” (SEO). We examine these methods in our setting and conclude that the FEO method provides an attractive solution. In particular, we show that it addresses the two deficiencies of the pooled method highlighted above: it removes the distortion in the pooled coefficients that results from bank heterogeneity, and it prevents the misdirection of legitimate information to infer bank identity. Indeed, we show that the only way to achieve lower forecast errors than the FEO method is through such misdirection, a result that sheds light on the trade-off between accuracy and fairness.

Moreover, the method is easy to interpret and implement: fit a pooled model with centered bank fixed effects, and then *discard* the centered fixed effects to forecast losses. Including the fixed effects prevents misdirection of legitimate information; discarding them is necessary to treat banks equally; centering ensures that the overall mean forecast remains unchanged. Although we mainly work with linear models, we show that these ideas can be extended to nonlinear models as well. We also derive an extension of FEO to remove certain interaction effects, as opposed to just fixed effects.

To help position our work, we briefly discuss some other research on bank stress tests. Covas, Rump, and Zakrajsek [18], Kapinos and Mitnik [49], and Kupiec [55] find strong evidence of heterogeneity in banks’ responses to macroeconomic shocks, and Kapinos and Mitnik [49] argue that ignoring heterogeneity can substantially underestimate projected capital requirements. The related models of Hirtle et al. [46] and Guerrieri and Welch [44] forecast aggregate results and are therefore not concerned with differences among banks. Heterogeneity in the accuracy of the Fed’s models for different banks is suggested by the comparisons in Agarwal et al. [1], Bassett and

Berrospide [9], and Flannery, Hirtle, and Kovner [29] between the Fed’s results and results based on the banks’ own models.

A separate line of research considers the design of stress scenarios. Several studies (including Breuer et al. [13], Flood and Korenko [31], Glasserman et al. [37], Pritsker [69], and Schuermann [72]) have advocated the use of multiple scenarios to capture different combinations of risk factors. Cope et al. [17] and Flood et al. [30] recommend designing scenarios to reflect bank heterogeneity. Parlatore and Philippon [66] propose a theoretical framework for scenario design as a problem of optimal information acquisition.

Several studies have investigated the information content of stress test results, either through market responses (as in Fernandes, Igan, and Pinheiro [27], Flannery, Hirtle, and Kovner [29], Georgescu et al. [35], Glasserman and Tangirala [40], Guerrieri and Modugno [43], Morgan, Peristiani, and Savino [63], and Sahin, de Haan, and Neretina [71]) or through subsequent bank performance (as in Kupiec [55] and Philippon, Pessarossi, and Camara [67]). Flannery [28] discusses just how much information the Fed should disclose about stress testing procedures and outcomes. For perspectives on the effectiveness of the Fed’s stress tests, see Kohn and Liang [52] and Schuermann [72].

We provide additional background on the Federal Reserve’s stress tests in Section 3.2. Section 3.3 lays out our modeling framework and analyzes the pooled industry model within this framework. Section 3.4 analyzes various ways to introduce fairness considerations, including the projection-to-fairness and FEO methods. Section 3.5 considers nonlinear models. Proofs of our main results appear in the appendix. Additional supporting theoretical (Sections C.2–C.4) and empirical (Sections C.5–C.8) material is included in the Electronic Companion. Most of our discussion considers loss models, but we consider revenue models in Section C.8.

3.2 Background

This section provides background on the Federal Reserve’s stress testing process and on the heterogeneity of the participating banks.

3.2.1 Regulatory Bank Stress Tests

In early 2009, in the depths of the Global Financial Crisis, the Federal Reserve launched a stress test of the 19 largest U.S. bank holding companies to gauge how much more capital they would need if economic conditions continued to worsen. The results of the stress test were made public, and the transparency and credibility of the process have been credited with restoring public confidence and helping to end the crisis.

The Dodd-Frank Act, the package of reforms that followed the crisis, codified the use of stress testing for bank supervision. The number of banks subject to DFAST (Dodd-Frank Act Stress Tests) has varied over time. The current requirement applies annually to banks with over \$250 billion in assets and every other year to banks with assets between \$100 billion and \$250 billion. The 2022 DFAST covered 34 banks. We refer to the participating firms as “banks,” but they are more precisely holding companies, including the U.S. subsidiaries of some foreign banks.

The inputs to the stress test analysis are the stress scenario, which is common to all banks, and bank-specific balance sheet information. A scenario is specified through a hypothetical path of economic variables over the next 13 quarters. The 2022 DFAST specified paths for 28 variables, including GDP, inflation, unemployment, stock market and real estate indexes, interest rates, exchange rates, and measures of overseas economic activity. Each bank submits detailed information on its loans and other assets.

The Fed uses 21 models to integrate the stress scenarios with bank-level information to make bank-level projections. For example, one model applies to credit cards, one to first lien residential mortgages, one to commercial real estate loans, and another to commercial and industrial loans. These models project losses in each of these portfolios. Some other models project revenues.

The Fed does not disclose details of its models, either to banks or the general public. But it does describe its general modeling approach in public documents. At a high level, a model assigns a loss rate to a set of bank-specific loan portfolio features x and a common set of scenario variables z through a function $f(x, z)$. The function f is estimated from past observations of the macro variables and portfolio features for multiple banks. Thus, f is estimated as an industry-wide

model and then applied individually to each bank.

This approach is described, for example, on p.3 of Board of Governors [11], where we read, “The Federal Reserve generally develops its models under an industry-level approach calibrated using data from many financial institutions... The Federal Reserve models the response of specific portfolios and instruments to variations in macroeconomic and financial scenario variables such that differences across firms are driven by differences in firm-specific input data, as opposed to differences in model parameters and specifications. As a result, two firms with the same portfolio receive the same results for that portfolio in the supervisory stress test, facilitating the comparability of results.”

As noted in the introduction, we refer to the principle that banks with the same portfolio receive the same results as *equal treatment*.

3.2.2 Bank Heterogeneity

The appropriateness of equal treatment seems incontrovertible. But the right notion of consistency across firms becomes less clear when portfolios vary widely, and the largest U.S. banks are a highly heterogeneous group. We may not expect a regional bank to have an investment bank’s skill in the capital markets, nor do we expect the investment bank to have the regional bank’s skill in making single-family residential loans.

Heterogeneity among large banks is illustrated in Figure 3.1. The left panel applies to the banks that participated in the Federal Reserve’s 2022 stress test. It shows the distribution of the banks by their Global Industry Classification Standard sub-industry classifications. This group includes diversified banks (such as JPMorgan Chase and Bank of America); regional banks (like PNC Financial and Citizens Financial); consumer finance companies (including American Express and Discover); custody banks (such as Bank of New York Mellon and State Street); investment banks (including Goldman Sachs and Morgan Stanley); and intermediate holding companies comprising the U.S. subsidiaries of foreign banks (such as TD Group and Credit Suisse USA). The distribution of banks across categories reflects important differences in their areas of specialization.

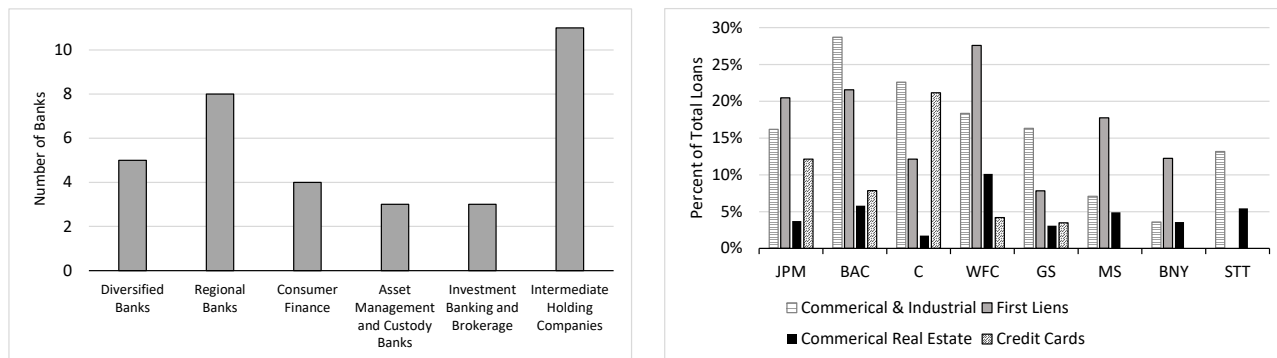


Figure 3.1: Heterogeneity among large banks. Left: Distribution of 2022 stress test banks by GIC sub-industry. Right: The percentage of loans in each of four categories for each of the U.S. G-SIBs, based on Y-9C reports for Q4 2021.

The right panel of Figure 3.1 applies to the U.S. Global Systemically Important Banks (G-SIBs). It shows heterogeneity in the fractions of loans the banks hold in each of the four categories. For example, for Wells Fargo (WFC) first lien mortgages are a relatively large fraction of its loans, whereas for Citigroup (C), credit cards make up a relatively large fraction. The figure suggests different areas of specialization in lending, even among the largest U.S. banks. Heterogeneity across banks is further explored empirically in the Electronic Companion.

Beginning in 2020, the Federal Reserve allowed banks to submit requests for reconsideration of the stress capital buffer set by the Fed through the stress testing process. (The capital buffer is set through the Comprehensive Capital Analysis and Review, or CCAR, process, which accompanies the stress test.) The banks' requests are confidential, but the Fed's responses to these requests are public. The responses show that the banks were arguing for reconsideration at least in part based on claims that the Fed's models do not capture distinctive features of the banks' businesses. For example, Regions Financial claimed that the Fed's models overlook the bank's hedging of interest rate risk. Goldman Sachs claimed that the Fed's models omit information relevant to the credit quality of the bank's mortgage loans. Citizens Financial claimed that the Fed's models overlook the bank's loss-sharing agreements in its retail portfolio.

Five firms requested reconsideration in 2020, and all five requests were rejected. In its response¹ to Goldman Sachs, the Fed wrote, "the Board has determined that it will follow its pub-

¹<https://www.federalreserve.gov/supervisionreg/files/goldman-sachs-group-inc-20200904.pdf>

lished principles for stress testing, including the principle of creating industry-level models, and not modify the existing results of these models. In particular, models used in the supervisory stress test are generally developed according to an industry-level approach, calibrated using data from many institutions.” Similar statements appear in all five rejections. These exchanges point to a debate in which the banks highlight their heterogeneity and the Fed asserts the importance of consistency.

3.2.3 Heterogeneity and Fairness

Read narrowly, the principle of equal treatment — the Fed’s statement that “two firms with the same portfolio receive the same results for that portfolio” — is easy to satisfy. It holds in any model that forecasts losses based only on portfolio features and the stress scenario, without using any other bank-specific information. Even this narrow reading has important implications. For example, the quality of a bank’s IT systems or the strength of its “culture”² may be important factors in determining a bank’s losses under stress, but as they are not features of a loan portfolio, the Fed’s modeling principle would preclude incorporating them into the Fed’s models. Matters like the quality of a bank’s internal governance and controls must be addressed in other parts of the overall bank supervision process, outside of stress testing.³ Within the Basel framework, these considerations are part of the Pillar 2 supervisory process, as described in BCBS [10].

Under a broader interpretation of the principle of equal treatment, a regulatory model should also exclude indirect proxies for bank identity. Suppose, for example, that a bank with outdated IT systems had a particularly large number of loans to the energy sector. Suppose further that because of its weak IT the bank was a poor monitor of its borrowers and suffered abnormally large losses in downturns. With information about IT excluded, a predictive model of losses that uses this history would likely overstate the risk of loans to the energy sector. This outcome is arguably unfair to all banks making energy loans, in that they would be indirectly penalized for one bank’s weak

²For a perspective on the importance of culture, see, for example, “Enhancing Financial Stability by Improving Culture in the Financial Services Industry,” a speech given by then president of the Federal Reserve Bank of New York, William C. Dudley, on October 20, 2014, <https://www.newyorkfed.org/newsevents/speeches/2014/dud141020a.html>.

³The Fed’s stress tests previously included a qualitative component, but this component was dropped in 2019.

IT. Addressing these types of indirect effects drives our investigation. In its narrow sense, equal treatment requires an indifference to which banks hold which portfolios once a model is selected; the broader interpretation seeks to remove the influence of bank identity in the design of the model.

The Fed’s stated principle implicitly responds to concerns for *disparate treatment* of banks. The broader interpretation — precluding proxies for bank identity — aligns with a concern for a particular notion of *disparate impact* used in the literature on algorithmic fairness (see, for example, Chapter 6 of Barocas et al. [8], Section 3 of Lipton et al. [59], and Prince and Schwarcz [68]), sometimes called “proxy discrimination.” The banks’ objections, as reflected in their reconsideration requests, can be seen as concerns for a different type of disparate impact: even if the same model is applied to all banks, and even if the model is free of bank-identity proxies, some banks may claim to be more adversely affected than others by the model’s limitations. Most of the objections raised by banks can be understood as pointing to omitted variables — features omitted from the Fed’s models that a bank believes would result in a more favorable outcome if included in the models. The Fed’s responses suggest a reluctance to incorporate overly narrow features into models, particularly features that might affect only a single bank. Model limitations, of the type claimed by the banks are likely inevitable, given the limited data available on bank performance in scenarios of severe stress. The Fed should strive to continue to improve its models, but our concern is not primarily for the banks’ objections. Our focus is rather on how best to interpret and implement the Fed’s stated principle of equal treatment, particularly under the broader interpretation that addresses elements of both disparate treatment and disparate impact, within the overarching goal of accurately forecasting stressed losses for each bank.

The fairness literature distinguishes notions of individual fairness and group fairness, where the members of a group often share a sensitive or protected attribute. More abstractly, an individual is defined by a fixed set of features, and a group is characterized by a probability distribution over features; see, for example, the characterizations of individuals and groups in Sections 2 and 3 of Dwork et al. [23]. From this perspective, individual fairness is concerned with fairness conditional on a set of features, whereas measures of group fairness incorporate distributions over features.

Individual fairness typically requires that individuals with similar features be treated similarly. For a portfolio loss model, this condition is satisfied if the predicted loss is a suitably smooth function of the features of individual portfolios. But group fairness is more relevant to our setting than individual fairness because we think of each bank, with its particular mix of businesses and areas of focus, not as one portfolio but as a distribution over portfolios the bank might hold at different times. We are interested in accuracy and fairness with respect to these distributions of bank portfolio features. We therefore view each bank as a group of (and probability distribution over) individual portfolios that share the attribute of bank identity. In contrast, individual fairness would be relevant to evaluating accuracy and fairness conditional on a specific portfolio for each bank. We will make this formulation of groups and individuals more explicit in the next section after introducing our basic model.

3.3 Pooling: Fairness Through Unawareness?

3.3.1 Basic Model

To capture bank heterogeneity, we consider a market with multiple banks, indexed by $s = 1, \dots, \bar{S}$. The loss rate (or net charge-off rate) Y_s for bank s is given by

$$Y_s = \alpha_s + \beta_s^\top X_s + \epsilon_s, \quad (3.1)$$

with $\alpha_s \in \mathbb{R}$ and $\beta_s \in \mathbb{R}^d$. Here, X_s is a d -dimensional random vector of predictive variables whose distribution defines bank s ; at this point, we do not distinguish between portfolio characteristics and macro variables. The portfolio characteristics include information about a bank's borrowers and loan terms. We use a linear specification in (3.1) because it offers the simplest setting to explore the interaction of heterogeneity and fairness; we discuss nonlinear extensions in Section 3.5. We take (3.1) to be the true relationship between the loss rate Y_s for bank s over the forecast horizon and characteristics X_s known at the date the forecast is made. Loss rates are normalized by loan balances to make values of Y_s comparable across banks of different sizes.

We think of X_s as a draw from some distribution with

$$\mu_s = \mathbf{E}[X_s] \in \mathbb{R}^d, \quad \Sigma_s = \mathbf{var}[X_s] \in \mathbb{R}^{d \times d}. \quad (3.2)$$

The randomness in X_s can be interpreted as reflecting the variation in the characteristics for bank s (and the macro variables) over time — in particular, times of stress. As we discuss in greater detail in Remark 3.3.1, we think of each bank as a group, in the sense of a probability distribution over portfolios. We assume throughout that each Σ_s (hence μ_s) is finite and each Σ_s is nonsingular. The error ϵ_s in (3.1) is assumed to satisfy, for each s ,

$$\mathbf{E}[\epsilon_s] = 0 \quad \text{and} \quad \mathbf{cov}[X_s, \epsilon_s] = 0. \quad (3.3)$$

The regulator's problem is to choose a model g that forecasts the loss rate $g(x, s)$ for bank s if the bank's portfolio characteristic vector is x . The forecasts should, at a minimum, satisfy the following narrow property, which prohibits the regulator from applying different models to different banks:

Definition 3.3.1 (Equal treatment). *Model $g : \mathbb{R}^d \times \{1, \dots, \bar{S}\} \rightarrow \mathbb{R}$ satisfies equal treatment if $g(x, s) = g(x, s')$, for all $x \in \mathbb{R}^d$, for all $s, s' \in \{1, \dots, \bar{S}\}$.*

As the true relationship for each bank is linear in (3.1), we mainly focus on the case of a linear industry-wide model. The regulator's problem is then to choose a single $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^d$ that it will use to form a forecast

$$\hat{Y}(x) = \alpha + \beta^\top x, \quad (3.4)$$

given portfolio characteristics x . The forecast (3.4) satisfies equal treatment because it has no functional dependence on bank identity s . The parameters of the industry model (3.4) may depend on the bank-specific parameters (α_s, β_s) and on the mean and variance in (3.2), but they should not depend on the realized features X_s .

The regulator would like the forecast loss $\hat{Y}(X_s)$ to be close to the actual loss Y_s in (3.1) for

every bank s . To aggregate errors across banks, we introduce a random variable S that picks a bank according to a distribution

$$P(S = s) = p_s, \quad s = 1, \dots, \bar{S}, \quad (3.5)$$

with the probabilities p_s summing to 1. In the simplest case, all banks get equal weight, and the p_s are all equal; but the p_s could also reflect relative asset sizes or other weighting schemes. When we replace a bank label s with the random variable S , we get a mixture over banks. In particular, we can combine the bank-specific models (3.1) into a mixture or hierarchical model by writing

$$Y_S = \alpha_S + \beta_S^\top X_S + \epsilon_S. \quad (3.6)$$

In choosing parameters α and β in (3.4), the regulator would like to make the forecast errors small for all banks. A natural way to aggregate forecast errors across banks is to consider the average squared error, in which case the regulator's problem becomes choosing α and β in (3.4) to solve

$$\min_{\alpha, \beta} \mathbf{E}[(\hat{Y}(X_S) - Y_S)^2]. \quad (3.7)$$

The objective in (3.7) averages squared forecast errors over banks. It can also be written as $\sum_s p_s \mathbf{E}[(\hat{Y}(X_s) - Y_s)^2]$.

Remark 3.3.1. Before solving (3.7), we make several comments on our problem formulation.

(i) *Targets versus estimators.* The problem posed by (3.7), like the more general problem of choosing industry parameters in (3.4), is one of characterizing ideal coefficients α and β . This is a question of choosing the correct *targets* of estimation, rather than a question of choosing *estimators*. In particular, α and β are population quantities rather than sample quantities. In practice, the regulator may have a panel of time series of observations across banks. Estimation methods for panel data ordinarily focus on coefficients that are common to all units and exploit the panel structure to estimate these shared values. Our concern is precisely with the case of heterogeneous coefficients, where we need to identify suitable targets before we can consider their estimation.

(ii) *Groups versus individuals.* As discussed in Section 3.2.3, individual fairness is concerned with ensuring that if two feature vectors x and x' are close, then the predicted losses $\hat{Y}(x)$ and $\hat{Y}(x')$ are also close. Our concern is for accuracy and fairness with respect to the distributions of (X_s, Y_s) , $s = 1, \dots, \bar{S}$, and not just for individual outcomes; we do not want the choice of industry model to depend on the realization of X_s , $s = 1, \dots, \bar{S}$. Each (X_s, Y_s) reflects a distribution over individual portfolio features and losses — individuals that share the bank identity attribute s . Each bank thus represents a group of potential individual portfolios, and we are interested in accuracy and fairness with respect to the probability distributions that define these groups.

(iii) *Stressed versus unstressed.* For the application to stress testing, it is helpful to think of the portfolio features and scenario variables in X_s and the losses Y_s in (3.1) as having their conditional distributions given stress conditions. The regulator is then interested in forecasting the conditional mean loss for each bank, given stress conditions. By focusing on the conditional mean, this formulation makes the squared error (3.7) a reasonable benchmark for studying accuracy and fairness.

(iv) *Regulator versus banks.* As discussed in Section 3.2.3, some of the objections raised by banks can be understood as pointing to features missing from (3.1) and (3.4), features that are rejected by the Fed as overly narrow. Our investigation assumes the bank-specific models (3.1)–(3.3) are correct; in particular, under (3.3) any relevant omitted features are uncorrelated with included features. We focus on the regulator’s problem of how best to aggregate the bank-specific models (assuming their correctness) into an industry model, considering both accuracy and fairness; we do not address the banks’ claims regarding which features should be included in the models.

For the solution to (3.7), write

$$\bar{\mu} = \mathbf{E}[X_S] = \mathbf{E}[\mu_S] = \sum_s p_s \mu_s \in \mathbb{R}^d, \quad (3.8)$$

and

$$\text{var}[X_S] = \mathbf{E}[(X_S - \bar{\mu})(X_S - \bar{\mu})^\top] = \mathbf{E}[W_S] = \sum_s p_s W_s, \quad (3.9)$$

with

$$W_s = \Sigma_s + \mu_s \mu_s^\top - \bar{\mu} \bar{\mu}^\top \in \mathbb{R}^{d \times d}. \quad (3.10)$$

Similarly,

$$\text{COV}[\alpha_S, \mu_S] = \sum_s p_s \alpha_s (\mu_s - \bar{\mu}) \in \mathbb{R}^d.$$

Proposition 3.3.1. *Problem (3.7) is solved by*

$$\beta_{Pool} = \mathbf{E}[W_S]^{-1} (\text{COV}[\alpha_S, \mu_S] + \mathbf{E}[W_S \beta_S]) \quad (3.11)$$

and

$$\alpha_{Pool} = \mathbf{E}[Y_S] - \beta_{Pool}^\top \bar{\mu}. \quad (3.12)$$

Loss forecasts using α_{Pool} and β_{Pool} in (3.4) provide *fairness through unawareness*, in that they ignore bank identity. They satisfy equal treatment in the narrow sense of Definition 3.3.1. Given our starting point (3.1), problem (3.7) would seem to be the most direct interpretation of the Fed’s policy of developing an “industry-level approach calibrated using data from many financial institutions.”

However, the solution in (3.11) is not a satisfactory target. Indeed, (3.11) shows where heterogeneity is most problematic. If the intercepts α_s covary with the means μ_s , this effect can distort β_{Pool} through what is commonly known as Simpson’s paradox. As an extreme example, consider the case that $\beta_s = 0$ for all s ; in other words, none of the features in X_s is predictive of losses for any of the banks. The regulator’s model (3.4) using β_{Pool} would nevertheless forecast losses based on these features if $\text{COV}[\alpha_S, \mu_S]$ is nonzero. This covariation would create the illusion of predictability. In applying (3.11), we would be forecasting losses based on irrelevant features, purely as a consequence of the way we aggregated the bank-specific models.

Even in a less extreme setting in which the β_s are nonzero, the presence of the $\text{COV}[\alpha_S, \mu_S]$ term in (3.11) reflects an indirect influence of bank identity on loss forecasts. If the bank-level mean

characteristics μ_s positively covary with the bank-level intercepts α_s , then in the pooled model this covariance will lead to a higher loss forecast for a bank with a higher value of X_s . This is arguably unfair, in the sense that the loss forecast is not based on the legitimate influence of the feature X_s . We will formalize the idea that the pooled method misdirects legitimate information in Sections 3.4.2 and 3.4.5.

This effect is reminiscent of the bias incurred in panel regressions when fixed effects are present in the data but omitted from a model. As we emphasized in Remark 3.3.1(i), in our setting the primary objective is to define the appropriate target of estimation, given the heterogeneity in the coefficients. We cannot say the term $\text{COV}[\alpha_s, \mu_s]$ introduces bias until we have decided what we are trying to estimate.

3.3.2 Average Treatment Effects

We can gain additional insight by considering the case of scalar X_s . In this case, the pooled coefficient β_{Pool} in (3.11) becomes

$$\beta_{Pool} = \frac{\text{COV}[\alpha_S, \mu_S] + \sum_s p_s (\sigma_s^2 + \mu_s^2 - \bar{\mu}\mu_s) \beta_s}{\sum_s p_s (\sigma_s^2 + \mu_s^2 - \bar{\mu}\mu_s)}. \quad (3.13)$$

In the special case that $\text{COV}[\alpha_S, \mu_S] = 0$ and $\sigma_s^2 + \mu_s^2 - \bar{\mu}\mu_s \geq 0$, for all s , (3.13) becomes a convex combination of the individual β_s . In Section C.4, we state some simple properties that an aggregation of the individual β_s into a single industry value should satisfy, and we show that only a convex combination satisfies these properties. Equation (3.13) thus shows a further potential problem with the pooled method. Even if $\text{COV}[\alpha_S, \mu_S] = 0$, the coefficient on some β_s could be negative, which would mean that a reduction in β_s would increase β_{Pool} . This could mean that an improvement in risk management by one bank *increases* predicted losses at all banks. We investigate these types of cross-bank effects further in Section C.2.

We will refer to any convex combination of the β_s as a *weighted average treatment effect* or WATE parameter. This terminology is suggested by thinking of a unit increase in a portfolio

characteristic X_s as a treatment, and β_s as the response to that treatment. The (ordinary) average treatment effect is the expected coefficient,

$$\beta_{ATE} = \mathbf{E}[\beta_S] = \sum_s p_s \beta_s, \quad (3.14)$$

but weighting the individual coefficients allows other combinations. In particular, if the μ_s are all equal, the pooled coefficient (3.13) becomes

$$\beta_{Pool} = \frac{\sum_s p_s \sigma_s^2 \beta_s}{\sum_s p_s \sigma_s^2}. \quad (3.15)$$

We will say more about these cases in subsequent sections.

To translate a WATE coefficient into a loss projection \hat{Y} , as in (3.4), we also need to specify an intercept. Setting

$$\alpha_{WATE} = \mathbf{E}[Y_S] - \beta_{WATE}^\top \bar{\mu},$$

ensures that the forecasts

$$\hat{Y}_{WATE}(X_s) = \alpha_{WATE} + \beta_{WATE}^\top X_s, \quad s = 1, \dots, \bar{S},$$

have zero expected error, in the sense that

$$\mathbf{E}[\hat{Y}_{WATE}(X_S) - Y_S] = \sum_s p_s (\alpha_{WATE} + \beta_{WATE}^\top \mu_s) - \mathbf{E}[Y_S] = 0.$$

3.4 Fair Regressions

We have seen that if the regulator's sole objective is to minimize average squared forecast errors subject to equal treatment, then the solution is given by the pooled coefficients in (3.11) and (3.12). However, we have also seen that (3.11) has consequences that are undesirable and even unfair, in the sense that it is indirectly influenced by bank identity. In this section, we turn to meth-

ods that expand the squared loss minimization objective (3.7) to include fairness considerations. Because the pooled method minimizes (3.7), any method that addresses fairness will entail a loss of accuracy as measured by (3.7).

3.4.1 Projection to Fairness

In the literature on fairness in classification methods, *demographic parity* is among the most widely discussed fairness principles; see, for example, Chapter 3 of Barocas et al. [8]. In the simplest classification setting, the counterpart of our forecast is a binary outcome $\hat{Y} \in \{0, 1\}$. For example, $\hat{Y} = 1$ may indicate a hiring decision, a loan approval, or a school admission decision. The decision is to be based on certain features of a candidate that are deemed legitimate. Demographic parity requires that the event $\{\hat{Y} = 1\}$ be statistically independent of a protected attribute, such as race or gender. This objective is difficult to achieve when legitimate features covary with the protected attribute.

Chzhen et al. [16] and Le Gouic et al. [9] extend the notion of demographic parity to the regression setting by requiring that model predictions be independent of a protected attribute. These two articles solve the problem of finding the model that minimizes mean squared prediction errors while achieving demographic parity. We will use the term *projection to fairness* (PTF), coined in Le Gouic et al. [9], for the method in these papers.

Both papers reduce the problem of regression fairness to one of finding the Wasserstein barycenter of a set of distributions, in the sense of Agueh and Carlier [2]. The barycenter is the distribution closest to the set of distributions in an average sense. For a squared error and one-dimensional distributions, the barycenter can be described as the distribution whose quantile function is a weighted average of the individual quantile functions. (The quantile function is the inverse of the cumulative distribution function.)

In the setting of Section 3.3.1, the resulting solution can be interpreted as follows. Let F_s denote the cumulative distribution function of $\hat{Y}_s(X_s)$, the forecast for bank s . Given realized features $X_s = x$, the regulator first forms the forecast $\hat{Y}_s(x) = \alpha_s + \beta_s^\top x$, using the bank-specific

coefficients. The regulator then computes $q_s = F_s(\hat{Y}_s(x))$, meaning that the forecast $\hat{Y}_s(x)$ is at the q_s quantile of the distribution F_s . The PTF forecast is then achieved by taking the weighted average of the corresponding quantile of all banks' forecast distributions, $\mathbb{E}_S[F_S^{-1}(q_s)]$. If, for example, $\hat{Y}_s(x)$ falls at the 80th percentile of the forecast distribution for bank s , then the regulator takes a weighted average of the 80th percentile forecast for all of the bank-specific models. That weighted average becomes the PTF forecast for bank s .

To make this procedure more explicit and to specialize the general framework of Chzhen et al. [16] and Le Gouic et al. [9] to our setting, we consider the case (for this section only) that each feature vector X_s has a multivariate normal distribution $N(\mu_s, \Sigma_s)$. Write $\Sigma_s^{1/2}$ for the symmetric square root of Σ_s , and define the standardized feature vectors

$$Z_s = \Sigma_s^{-1/2}(X_s - \mu_s); \quad (3.16)$$

each Z_s has a multivariate standard normal distribution. Write the basic identity (3.1) using standardized variables as

$$Y_s = \alpha_s^o + \beta_s^{o\top} Z_s + \epsilon_s,$$

with standardized coefficients

$$\beta_s^o = \Sigma_s^{1/2} \beta_s, \quad \alpha_s^o = \alpha_s + \beta_s^\top \mu_s. \quad (3.17)$$

Suppose $\|\beta_s^o\| \neq 0$, for all s , with $\|\cdot\|$ denoting the usual Euclidean norm. Consider the model that assigns, to each bank $s = 1, \dots, \bar{S}$, with features $X_s = x$ the forecast

$$\hat{Y}^o(x, s) = \sum_i p_i \alpha_i^o + \sum_i p_i \|\beta_i^o\| \frac{\beta_i^{o\top} z_s}{\|\beta_i^o\|}, \quad z_s = \Sigma_s^{-1/2}(x - \mu_s). \quad (3.18)$$

If there exists a $\beta \in \mathbb{R}^d$ and scalars $a_s > 0$ for which

$$\beta_s^o = a_s \beta, \quad s = 1, \dots, \bar{S}, \quad (3.19)$$

then we will see that (3.18) simplifies to the weighted average

$$\hat{Y}^o(x, s) = \bar{\alpha}^o + \bar{\beta}^{o\top} z_s, \quad \bar{\alpha}^o = \sum_i p_i \alpha_i^o, \quad \bar{\beta}^o = \sum_i p_i \beta_i^o. \quad (3.20)$$

In the case of scalar X_s , (3.19) holds whenever all β_s have the same sign.

Proposition 3.4.1. *Suppose that the X_s are multivariate normal and $\|\beta_s\| \neq 0$, for all $s = 1, \dots, \bar{S}$. Then (3.18) is the projection-to-fairness of the bank-specific models (3.1), meaning that (3.18) minimizes $\mathbb{E}[(\hat{Y}^o(X_S, S) - Y_S)^2]$ among all models (whether linear or not) that satisfy demographic parity. If (3.19) holds, the projection-to-fairness is given by (3.20).*

We can see from (3.18) that the PTF model does not satisfy equal treatment: to calculate the loss forecast for a bank, we need to know its identity s . We have included the special case of (3.20) because it more nearly parallels the type of model we seek in (3.4). The coefficients in (3.20) are weighted averages of bank-specific coefficients. The model in (3.20) satisfies equal treatment with respect to the standardized features Z_s , rather than the raw features X_s : two banks with the same standardized features will receive the same forecasts. But the means for the two banks could be very different — the standardization is done separately for each bank — indicating that one bank’s portfolio may be much riskier than the other bank’s. In treating standardized characteristics for different banks as comparable, the PTF model implicitly evaluates the riskiness of each bank relative to the distribution for that bank. The suitability of PTF in our setting is therefore questionable.

The root of the problem is that demographic parity is too strong a property for our setting. Ensuring that a hiring decision is independent of race or gender is important; but forcing the distribution of loss projections to be independent of bank identity ignores relevant differences in banks’ portfolios. Whereas the pooled model (3.11)–(3.12) does too little to address heterogeneity across banks, the PTF model goes too far in leveling differences. The next section provides a better balance.

3.4.2 Formal Equality of Opportunity

Johnson, Foster, and Stine [48] introduce the concept of formal equality of opportunity (FEO) in regression, based on the use of the term in political philosophy, for which they cite the review in Arneson [6]. According to Arneson [6], FEO means that “positions and posts that confer superior advantages should be open to all applicants. Applications are assessed on their merits.”

In adapting this idea to our setting, it is helpful to make a contrast with the previous section: whereas demographic parity requires that loss forecasts be independent of bank identity, FEO allows bank-dependence, but only through legitimate portfolio characteristics — through the bank’s “merits.” This notion aligns well with the Fed policy, quoted earlier, that “two firms with the same portfolio receive the same results.” The objective of FEO in regression, as developed by Johnson et al. [48], is to ensure that a protected attribute (for us, bank identity) has no direct or “causal” impact on a model’s predictions. The predictions may be correlated with bank identity if different banks tend to have different levels of exposure to legitimate portfolio features.

To develop this idea in our setting, we introduce the centered dummy variables

$$U_i(s) = \mathbf{1}\{s = i\} - p_i, \quad i = 1, \dots, \bar{S} - 1, \quad s = 1, \dots, \bar{S}. \quad (3.21)$$

We discuss the implications of centering below. For any coefficients $\alpha, \delta_1, \dots, \delta_{\bar{S}-1} \in \mathbb{R}$ and $\beta \in \mathbb{R}^d$, and any $x \in \mathbb{R}^d$, let

$$\hat{Y}(x, s) = \alpha + \sum_i \delta_i U_i(s) + \beta^\top x. \quad (3.22)$$

We have included the bank label s as an argument of \hat{Y} because U_i depends on s . Let $\alpha_F, \{\delta_i, i = 1, \dots, \bar{S} - 1\}$, and β_F solve the error minimization problem

$$\min_{\alpha, \{\delta_i\}, \beta} \mathbb{E}[(\hat{Y}(X_S, S) - Y_S)^2]. \quad (3.23)$$

With the coefficients that minimize (3.7), (3.22) becomes the linear projection of Y_S onto the span of $\{1, U_1(S), \dots, U_{\bar{S}-1}(S), X_S\}$, evaluated at $S = s$ and $X_S = x$. Now drop the centered dummy

variables U_i and define

$$\hat{Y}_F(x) = \alpha_F + \beta_F^\top x. \quad (3.24)$$

The FEO loss forecast for bank s is $\hat{Y}_F(X_s)$.

Steps (3.22)–(3.24) result from applying the definition of an impartial estimate (their Definition 2) in Johnson et al. [48]. (More precisely, steps (3.22)–(3.24) define a population counterpart of the sample formulation in [48].) The procedure in (3.22)–(3.24) can be interpreted as follows: pool losses and portfolio features across banks; regress losses on portfolio features with bank fixed-effects included; throw away the fixed effects in forecasting future losses. The resulting model (3.24) is an equal-treatment model, with no explicit dependence on bank identity. Centering the discarded variables U_i ensures that $\mathbf{E}[\hat{Y}_F(X_S)] = \mathbf{E}[Y_S]$, so dropping the fixed effects does not introduce an overall bias.

We will say more about the implications of this approach, but we first show that our setting allows an explicit expression for the FEO coefficients:

Proposition 3.4.2. (i) *The FEO coefficients are given by*

$$\beta_F = \mathbf{E}[\Sigma_S]^{-1} \mathbf{E}[\Sigma_S \beta_S], \quad (3.25)$$

and

$$\alpha_F = \mathbf{E}[Y_S] - \beta_F^\top \bar{\mu}. \quad (3.26)$$

In particular, in the scalar case,

$$\beta_F = \frac{\sum_s p_s \sigma_s^2 \beta_s}{\sum_s p_s \sigma_s^2}. \quad (3.27)$$

(ii) *We also have*

$$\beta_F = \text{var}[X_S - \mu_S]^{-1} \text{cov}[X_S - \mu_S, Y_S], \quad (3.28)$$

so $\beta_F^\top (X_S - \mu_S)$ is the linear projection of $Y_S - \mathbf{E}[Y_S]$ onto $X_S - \mu_S$.

We encountered (3.27) in (3.15) as a special case of the pooled coefficient when the bank means

μ_s are constant. The general case in (3.25) similarly coincides with the pooled coefficient in (3.11) when the means are constant. In other words, introducing the bank-level fixed effects in (3.22) purges β_F of the effect of different feature means across banks; dropping these fixed effects in (3.24) ensures that the regulator's model has no explicit dependence on bank identity and satisfies equal treatment.

In what sense is this procedure fair? We adapt the interpretation in Johnson et al. [48] to our setting. Write $U = (U_1, \dots, U_{\bar{S}-1})^\top$ for the vector of centered dummy variables. Write $\text{cov}[X_S, U(S)]$ for the $d \times (\bar{S} - 1)$ matrix of covariances between the components of X_S and $U(S)$. Let

$$\Lambda = (\text{var}[X_S])^{-1} \text{cov}[X_S, U(S)]. \quad (3.29)$$

This matrix minimizes $\mathbf{E}[\|U(S) - \Lambda^\top (X_S - \bar{\mu})\|^2]$, so $\Lambda^\top (X_S - \bar{\mu})$ is the linear projection of the bank-identity variables $U(S)$ onto the centered portfolio features $X_S - \bar{\mu}$. The relationship between β_{Pool} and β_F can be expressed as follows.

Proposition 3.4.3. *The coefficients β_{Pool} and β_F satisfy*

$$\beta_{Pool} = \beta_F + \Lambda \delta, \quad (3.30)$$

where $\delta = (\delta_1, \dots, \delta_{\bar{S}-1})^\top$ is the vector of coefficients from (3.22)–(3.23). In particular,

$$\delta_s = (\alpha_s + \beta_s \mu_s) - (\alpha_{\bar{S}} + \beta_{\bar{S}} \mu_{\bar{S}}) - \beta_F^\top (\mu_s - \mu_{\bar{S}}), \quad s = 1, \dots, \bar{S} - 1. \quad (3.31)$$

We can write the forecast in (3.22), using the optimal coefficients from (3.23) as

$$\hat{Y}(x, s) = \mathbf{E}[Y_S] + \delta^\top U(s) + \beta_F^\top (x - \bar{\mu}); \quad (3.32)$$

This is the linear projection of Y_S onto $(1, U(S), X_S)$, evaluated at $S = s$, $X_S = x$. Let $\hat{Y}_P(x) = \alpha_{Pool} + \beta_{Pool}^\top x$ denote the forecast based on the pooled coefficients (3.11) and (3.12). Decomposing $U(S)$ into its projection onto $X_S - \bar{\mu}$ and an orthogonal component leads to the following contrast

of these forecasts:

$$\hat{Y}(x, s) = \mathbf{E}[Y_S] + \delta^\top \Lambda^\top (x - \bar{\mu}) + \delta^\top [U(s) - \Lambda^\top (x - \bar{\mu})] + \beta_F^\top (x - \bar{\mu}) \quad (3.33)$$

$$\hat{Y}_P(x) = \mathbf{E}[Y_S] + \delta^\top \Lambda^\top (x - \bar{\mu}) + \beta_F^\top (x - \bar{\mu}) \quad (3.34)$$

$$\hat{Y}_F(x) = \mathbf{E}[Y_S] + \beta_F^\top (x - \bar{\mu}) \quad (3.35)$$

The term $\delta^\top [U(s) - \Lambda^\top (x - \bar{\mu})]$ in (3.33) affects the forecast through information in bank identity that is orthogonal to the legitimate features x . This would be *disparate treatment*, as in Johnson et al. [48]. Through “unawareness” (meaning that it has no functional dependence on bank identity) the pooled forecast (3.34) drops this term, but it retains $\delta^\top \Lambda^\top (x - \bar{\mu})$, as can be seen from (3.30).

The term $\delta^\top \Lambda^\top (x - \bar{\mu})$ is the problematic component of the pooled method. Although it does not explicitly use bank identity, this term relies on the fact that bank identity is to some extent predictable from portfolio features. Imagine the regulator forming loss forecasts from blinded data — the regulator does not know the identity of the bank. The term $\Lambda^\top (x - \bar{\mu})$ is the least-squares prediction of $U(s)$ from $x - \bar{\mu}$. In the pooled forecast (3.34), the regulator is implicitly “misdirecting” the data in the features $x - \bar{\mu}$ to try to identify the bank and then to adjust the forecast based on the inferred identity. The FEO forecast (3.35) removes this effect and retains only the direct effect of portfolio features on the loss rate.

In the terminology of Section 3.2.3, dropping $\delta^\top [U(s) - \Lambda^\top (x - \bar{\mu})]$ ensures the narrow sense of equal treatment — that loss forecasts not depend explicitly on bank identity. Dropping $\Lambda^\top (x - \bar{\mu})$ ensures a broader sense of equal treatment — that loss forecasts not depend on proxies for bank identity. Johnson et al. [48] refer to their counterpart of $\Lambda^\top (x - \bar{\mu})$ as *disparate impact*, which is consistent with the notion of “proxy discrimination” as a particular type of disparate impact (as in Prince and Schwarcz [68]). In our setting, as noted in Section 3.2.3, the disparate impact of most immediate concern to banks is the omission of features from the Fed’s models that might otherwise benefit individual banks. Omitted features may contribute to $\Lambda^\top (x - \bar{\mu})$, but dropping

this term does not necessarily dispel banks' complaints. The banks' disagreements with the Fed concern the scope of portfolio features that should be modeled. We therefore prefer to associate $\delta^\top [U(s) - \Lambda^\top(x - \bar{\mu})]$ and $\Lambda^\top(x - \bar{\mu})$ with narrow and broad interpretations of the Fed's own principle of equal treatment (as discussed in Section 3.2.3), rather than with separate concerns for disparate treatment and disparate impact by the Fed and the banks. We emphasize the interpretation of $\Lambda^\top(x - \bar{\mu})$ as a misdirection of legitimate information, rather than as a contributor to disparate impact.

The FEO method offers a further advantage over the pooled method. Recall again from Section 3.2.3 (and Remark 3.3.1(iv)) that we interpret the banks' objections as calls for the inclusion of features that are omitted from the Fed's models. Under the condition $\text{cov}[X_s, \epsilon_s] = 0$ in (3.3), the FEO coefficients of included features are unaffected by the omission of other features. The pooled coefficients do not in general have this property.

We will conclude in Section 3.4.5 that the FEO forecast is, in a precise sense, the best way to aggregate the bank-specific models into a single regulatory model. The FEO forecast has no direct dependence on bank identity; but it also removes the indirect dependence that results when bank identity is partly predictable from portfolio features. We discuss other methods for comparison.

3.4.3 Conditional Expectation Model

A similar misdirection of information occurs if we project the bank-specific models to an industry model in the sense of conditional expectation, rather than least squares. Suppose X_s has density g_s , and suppose $\mathbb{E}[\epsilon_s|X_s] = 0$, $s = 1, \dots, \bar{S}$. Then, by Bayes' rule,

$$\hat{Y}_C(x) \equiv \mathbb{E}[Y_S|X_S = x] = \frac{\sum_s p_s g_s(x) (\alpha_s + \beta_s^\top x)}{\sum_s p_s g_s(x)}. \quad (3.36)$$

This model satisfies equal treatment — $\hat{Y}_C(x)$ depends on the portfolio features x but not on a bank's identity. However, the point of the weights $p_s g_s(x)$ is to infer the identity of the bank from the features. Indeed, as discussed in Section 3.5, the conditional expectation $\mathbb{E}[Y_S|X_S = x]$ can be

viewed as a nonlinear generalization of the pooled method, with some of the same shortcomings.

3.4.4 Substantive Equality of Opportunity

As discussed in Arneson [6], a system in which admission decisions are made through a competitive exam open to everyone achieves formal equality of opportunity; but if only the wealthy have access to the preparation required for the exam, the system fails to achieve *substantive* equality of opportunity (SEO). In the regression setting, Johnson et al. [48] interpret SEO to mean that any influence of protected attributes should be removed from other variables included in a regression model. In the analogy with Arneson’s [6] example, SEO would seek to remove the effect of economic status from performance on the exam, whereas FEO would accept exam scores as a legitimate basis for decision-making. (Our use of “SEO” follows Johnson et al. [48]. For a broader interpretation of substantive equality in algorithmic fairness, see Green [41].)

To apply these ideas to our setting, define the $(\bar{S} - 1) \times d$ matrix

$$M = \text{var}[U(S)]^{-1} \text{cov}[U(S), X_S]; \quad (3.37)$$

then M minimizes $\mathbb{E}[\|X_S - \bar{\mu} - M^\top U(S)\|^2]$. In accordance with Definition 2 of Johnson et al. [48], define

$$\hat{Y}_{SEO}(x, s) = \alpha_F + \beta_F^\top (x - M^\top U(s)), \quad (3.38)$$

with α_F and β_F defined by (3.23). The SEO forecast adjusts the portfolio features x to remove the linear projection onto the centered bank dummy variables U . We can write (3.38) somewhat more explicitly as follows:

Proposition 3.4.4. *With M as in (3.37)*

$$M^\top U(s) = \sum_i (\mu_i - \mu_{\bar{S}}) U_i(s) = \mu_s - \bar{\mu}, \quad (3.39)$$

so the SEO forecast (3.38) is given by

$$\hat{Y}_{SEO}(x, s) = \alpha_F + \beta_F^\top(x - \mu_s + \bar{\mu}). \quad (3.40)$$

The SEO forecast is the linear projection of Y_S onto a constant and $X_S - \mu_S$.

Recall from Section 3.4.1 that a model satisfies demographic parity if its forecasts are independent of bank identity. Let us say that a model satisfies *weak* demographic parity if its forecasts are *uncorrelated* with the bank-identity variables $U_i(S)$. The centered features $X_S - \mu_S$ are uncorrelated with the $U_i(S)$. It therefore follows from Proposition 3.4.4 that SEO forecasts are uncorrelated with the $U_i(S)$. In other words, we have the following result:

Corollary 3.4.1. *The SEO forecast satisfies weak demographic parity.*

Under additional conditions, we get a stronger conclusion:

Corollary 3.4.2. *If the covariance matrix Σ_s and the distribution of Z_s in (3.16) are the same for all s , then the SEO model coincides with the standardized model (3.20), and both satisfy demographic parity.*

Under the conditions in the corollary, the mean adjustment in (3.40) is sufficient to give $\hat{Y}_{SEO}(X_s, s)$ the same distribution for all s . Put differently, PTF considers only the quantile of $\alpha_s + \beta_s^\top X_s$, relative to the distribution for bank s , to be legitimate information; SEO considers $X_s - \mu_s$ to be legitimate information. Under the conditions of the corollary, the two concepts coincide.

The mean adjustment in (3.40) requires knowledge of the bank identity s , so (3.38) does not satisfy Definition 3.3.1. The intent of the mean adjustment is to achieve a greater degree of equality. Consider the example with which began this section. If x represents an exam score and $\mu_1 > \mu_0$ are the mean scores among wealthy and non-wealthy exam takers, (3.40) adjusts scores downward for wealthy exam takers and upward for non-wealthy exam takers.

Such an adjustment may be appropriate when the individuals or firms under evaluation are, in some sense, not responsible for their mean characteristic (or the mean in their peer group) and

are therefore evaluated based on deviations from the mean. This type of consideration does not seem applicable to the stress-test setting, but it could arise more generally in settings where capital regulation intersects with other policy objectives.

One such example is suggested by the Paycheck Protection Program Lending Facility (PPPL) launched by the Federal Reserve early in the COVID crisis. The PPPL provided for loans to small businesses to be made by banks and guaranteed by the Small Business Administration. Under normal circumstances, the loans would increase participating banks' balance sheets and thus potentially increase their capital requirements. To promote use of the facility, banking regulators issued a rule excluding PPPL loans from capital requirements, thus “neutralizing the effects of participating in the PPPL Facility on regulatory capital requirements.”⁴ This “neutralizing” action is somewhat analogous to the SEO adjustment in that it removes responsibility for the larger balance sheet from the bank. The adjustments differ in that SEO adjusts for the mean whereas the PPPL adjustment removes the amount lent through the program.

3.4.5 A Unified Perspective: Legitimate Information

All of the methods we have discussed can be seen as ways of choosing forecasts $\hat{Y}_s, s = 1, \dots, \bar{S}$, (of the form $\hat{Y}(X_s)$ or $\hat{Y}(X_s, s)$) to minimize

$$E[(\hat{Y}_s - Y_s)^2], \tag{3.41}$$

subject to additional considerations. Table 3.1 summarizes the cases we have considered. In rows (i), (iv), and (v), we minimize (3.41) over the indicated coefficients. In (ii) and (iii), we allow g to be an arbitrary (suitably measurable) function of the indicated arguments. In (iii) we strengthen the condition (3.3) on the errors ϵ_s .

Proposition 3.4.5. *In each row of Table 3.1, the squared loss (3.41) is minimized over forecasts of the form in the first column, subject to the constraint in the second column, by the model in the last*

⁴Federal Register, Vol. 85, No. 71, p.20389, April 13, 2020.

	Form	Constraint	Forecast
(i)	$\hat{Y}_s = \alpha + \beta^\top X_s$		Pooled (3.11)–(3.12)
(ii)	$\hat{Y}_s = g(X_s, s)$, some g	\hat{Y}_s independent of S	PTF ([16, 9])
(iii)	$\hat{Y}_s = g(X_s)$, some g , $\mathbf{E}[\epsilon_s X_s] = 0$		Cond. exp. (3.36)
(iv)	$\hat{Y}_s = \alpha + \beta^\top X_s$	$\text{cov}[Y_S - \hat{Y}_S, X_S - \mu_S] = 0$	FEO (3.24)
(v)	$\hat{Y}_s = \alpha + \lambda^\top U(s) + \beta^\top X_s$	$\text{cov}[\hat{Y}_S, U(S)] = 0$	SEO (3.38)

Table 3.1: Summary of forecast model forms and constraints.

column.

The constraint in Table 3.1(v) is weak demographic parity. SEO implicitly takes the view that the only legitimate information in forecasting losses for bank s is the deviation $X_s - \mu_s$. In contrast, FEO takes the full set of features X_s as legitimate information. Through the constraint in Table 3.1(iv), it enforces a requirement we call *no misdirection of legitimate information*. FEO uses all of X_s in forecasting losses; but it chooses the coefficient β_F to be the coefficient in a regression of Y_S on $X_S - \mu_S$, which is the part of X_S orthogonal to bank identity. This condition ensures that the information in X_S is not misdirected to infer bank identity.

To make this idea precise, consider any model of the form (3.4). If we assume the intercept is chosen to match the unconditional mean, we may write the model as

$$\hat{Y}_\gamma(x) = \mathbf{E}[Y_S] + (\beta_F + \gamma)^\top (x - \bar{\mu}), \quad (3.42)$$

for some $\gamma \in \mathbb{R}^d$. With $\gamma = 0$, we get the FEO forecast (3.24).

Proposition 3.4.6. *If γ reduces errors in the sense that $\mathbf{E}[(\hat{Y}_\gamma(X_S) - Y_S)^2] < \mathbf{E}[(\hat{Y}_F(X_S) - Y_S)^2]$, then the forecast \hat{Y}_γ misdirects legitimate information in the sense that*

(i) $\text{cov}[\gamma^\top X_S, \delta^\top \Lambda^\top X_S] > 0$, and

(ii) $\text{cov}[\gamma^\top M^\top U(S), \delta^\top U(S)] > 0$.

Recall that $\Lambda^\top (X_S - \bar{\mu})$ is the linear projection of the centered bank identity variables $U(S)$ onto the centered portfolio features $X_S - \bar{\mu}$. The condition in (i) therefore indicates that γ misdirects

some of the legitimate information in X_S toward inferring bank identity. Thus, deviating from β_F in (3.42) either increases errors or misdirects information.

Property (ii) has a similar interpretation. The term $\delta^\top U(S)$ is the direct influence of bank identity on losses Y_S . The proposition states that any deviation γ that reduces forecast errors (relative to $\gamma = 0$) implicitly picks up some of the information in bank identity.

To further illustrate the contrast between FEO and SEO consider a simple example in which some component of X_S measures exposure to community development projects. Suppose for simplicity that this feature is uncorrelated with other features. In the SEO forecast, the only legitimate information from this exposure is a bank's deviation from its own mean. Years in which a bank had above average exposure would lead to higher loss forecasts, but the bank's average exposure to community development would not directly inform the forecasts — it is neutralized. In contrast, FEO treats the bank's total exposure (mean plus deviation) as legitimate information. Like SEO, in evaluating the impact of this exposure — that is, in estimating the coefficient on the exposure — it relies only on the within-bank variation. This ensures that the information in the exposure is not misdirected toward inferring the bank's identity, as could happen in the pooled regression.

3.4.6 Extension of FEO for Interaction Effects

Recall that the FEO forecast controls for bank fixed effects. One might similarly consider controlling for interactions between bank indicators and components of the feature vectors. This leads to a family of extensions of FEO that differ in which interactions they include. We will show that with a full set of interactions, the extended FEO model becomes the ATE model (3.14).

To examine this case, suppose the feature vector for each bank s is partitioned into two components, X_s and V_s . We extend FEO by including interactions with components of V_s but not with components of X_s . (Thus, in our discussion of FEO, V_s was empty.) We assume that for every bank s , the components of X_s are uncorrelated with the components of V_s . This allows a clear delineation between variables with and without interactions. Let $\nu_s = E[V_s]$. The bank-specific

models (3.1) now take the form

$$Y_s = \alpha_s + \beta_s^\top X_s + \gamma_s^\top V_s + \epsilon_s, \quad (3.43)$$

with ϵ_s uncorrelated with X_s and V_s .

We extend FEO to the following procedure:

- 1) Project Y_s linearly onto $1, U_1(S), \dots, U_{\bar{S}-1}(S), X_S - \mu_S, V_S - \nu_S, U_1(S)V_S, \dots, U_{\bar{S}}(S)V_S$.

Let β_F denote the coefficient of $X_S - \mu_S$ and let γ_F denote the coefficient of $V_S - \nu_S$.

- 2) Set $\hat{Y}_F(x, v) = \alpha_F + \beta_F^\top x + \gamma_F^\top v$, with α_F chosen so that $\mathbb{E}[\hat{Y}(X_S, V_S)] = \mathbb{E}[Y_S]$.

If V_S is empty, then we know from (3.28) that these steps do indeed reduce to the original FEO forecast. We have included the interaction $U_{\bar{S}}(S)V_S$ in the first step (even though we omitted $U_{\bar{S}}(S)$) to simplify the derivation of γ_F . Including this term means that the coefficients on the interactions $U_i(S)V_S$ are determined only up to constant, because $U_1(S)V_S + \dots + U_{\bar{S}}(S)V_S = 0$. These coefficients are dropped in the second step, so their value is immaterial.

Proposition 3.4.7. *Suppose $\text{var}[X_s]$ and $\text{var}[V_s]$ have full rank and X_s and V_s are uncorrelated, for each $s = 1, \dots, \bar{S}$. Then β_F is given by (3.25) and (3.28), and $\gamma_F = \bar{\gamma} = \sum_s p_s \gamma_s$. In particular, if interactions with $U(S)$ are included for all features, the FEO vector of coefficients reduces to the average treatment effect (3.14).*

This result allows us to interpret the ATE forecast as a version of the FEO forecast that removes the effects of certain interactions. As a convex combination of the bank-specific coefficients, the ATE coefficient retains some of the advantages of the FEO coefficient, particularly for the cross-bank effects studied in Section C.2.

However, we do not see a compelling case for controlling for interactions between bank identity and portfolio features. When we control for the bank-identity variables in FEO, we are ensuring that the industry β for legitimate features is not affected by heterogeneity in the banks' constants

(the fixed effects). This reasoning does not necessarily extend to removing the influence of heterogeneity in exposures to portfolio features.

3.5 Nonlinear Models

Most of the ideas developed in previous sections for linear regressions extend to generalized linear models through a transformation of the response variable. For example, instead of working with the loss rate Y_s , we could specify a linear model for its logit transformation $\log(Y_s/(1 - Y_s))$.

But we can also extend ideas from previous sections to more fully nonlinear models. Replace the mixture model in (3.6) with a general representation of the form

$$Y_s = g(S, X_s) + \epsilon_s, \quad \mathbf{E}[\epsilon_s | S, X_s] = 0. \quad (3.44)$$

In other words, the loss for bank s is given by $g(s, X_s) + \epsilon_s$. We assume that $g(S, X_s)$ and ϵ_s are square-integrable.

The counterpart of the pooled estimate becomes

$$f_{Pool}(x) \equiv \mathbf{E}[Y_s | X_s = x] = \mathbf{E}[g(S, X_s) | X_s = x].$$

This rule satisfies equal treatment — it has no functional dependence on S — but we argued earlier (in Section 3.4.3) that this forecast implicitly uses the information in the portfolio features x to infer bank identity.

To introduce a nonlinear version of the FEO forecast, we will make the relatively modest assumption that (3.44) admits a decomposition of the form

$$Y_s = f_0 + f_1(S) + f_2(X_s) + \epsilon, \quad \mathbf{E}[\epsilon | S] = \mathbf{E}[\epsilon | X_s] = 0, \quad (3.45)$$

with $f_0 = \mathbf{E}[Y_S]$, $f_1 : \{1, \dots, \bar{S}\} \rightarrow \mathbb{R}$, $f_2 : \mathbb{R}^d \rightarrow \mathbb{R}$, and

$$\mathbf{E}[Y_S - f_0 - f_1(S)|X_S] = f_2(X_S) \quad (3.46)$$

$$\mathbf{E}[Y_S - f_0 - f_2(X_S)|S] = f_1(S), \quad (3.47)$$

$\mathbf{E}[f_1^2(S)] < \infty$, $\mathbf{E}[f_2^2(X_S)] < \infty$, and

$$\mathbf{E}[f_1(S)] = \mathbf{E}[f_2(X_S)] = 0.$$

Equations (3.46)–(3.47) are population versions of the backfitting algorithm in Hastie and Tibshirani [45], which is a special case of the alternating conditional expectations algorithm of Breiman and Friedman [12]. Given an initial choice of f_1 (and known f_0), (3.46) defines an initial choice of f_2 through the regression of the residual $Y_S - f_0 - f_1(S)$ on X_S . Equation (3.47) then defines an updated choice of f_1 . The algorithm iterates over (3.46) and (3.47). In writing (3.45), we are positing that this algorithm has a fixed point. Convergence of the backfitting algorithm is established under widely applicable conditions in Ansley and Kohn [5].

We now introduce

$$\hat{Y}_F(x) = f_0 + f_2(x) \quad (3.48)$$

as a nonlinear counterpart of the FEO forecast. We justify this interpretation by showing that \hat{Y}_F exhibits properties that are nonlinear counterparts of the key properties of the FEO forecast in Section 3.4.2 and 3.4.5. To state the result, consider forecasts of the form

$$\hat{Y}_\gamma(x) = f_0 + f_2(x) + \gamma(x), \quad (3.49)$$

for some $\gamma : \mathbb{R}^d \rightarrow \mathbb{R}$ with $\mathbf{E}[\gamma(X_S)^2] < \infty$.

Proposition 3.5.1. *The nonlinear FEO forecast (3.48) satisfies*

$$\text{cov}[\hat{Y}_F(X_S) - Y_S, X_S - \mathbf{E}[X_S|S]] = 0. \quad (3.50)$$

For \hat{Y}_γ as in (3.49), if γ reduces errors, in the sense that $\mathbf{E}[(\hat{Y}_\gamma(X_S) - Y_S)^2] < \mathbf{E}[(\hat{Y}_F(X_S) - Y_S)^2]$, then it misdirects legitimate information, in the sense that

$$\text{cov}[\gamma(X_S), \mathbf{E}[f_1(S)|X_S]] > 0 \quad (3.51)$$

and

$$\text{cov}[\mathbf{E}[\gamma(X_S)|S], f_1(S)] > 0. \quad (3.52)$$

Property (3.50) parallels the condition in row (iv) of Table 3.1 that characterizes the FEO forecast in the linear setting. It says that the forecast error $\hat{Y}_F(X_S) - Y_S$ is uncorrelated with the legitimate information $X_S - \mathbf{E}[X_S|S]$, which is the component of X_S orthogonal to bank identity S . Properties (3.51)–(3.52) parallel conditions (i) and (ii) in Proposition 3.4.6. In particular, in (3.51), $\mathbf{E}[f_1(S)|X_S]$ is the expected impact of bank identity inferred from portfolio features; the positive covariance with $\gamma(X_S)$ thus indicates that γ misdirects some of the information in X_S to inferring S . The approach of this section is further explored numerically in the Electronic Companion.

We briefly contrast our FEO forecast in (3.48) with an alternative approach to extending fairness concerns to complex, nonlinear models. The alternative seeks to strip X_S of any protected attributes before a model is estimated. Examples of this general approach include Grūnewālder and Khaleghi [42] and Madras et al. [61]. This approach is primarily concerned with ensuring demographic parity: if a model has no access — not even indirect access — to a protected attribute, its forecasts will be independent of the attribute. But we argued previously that demographic parity is too strong a condition for our setting. Our FEO forecast in (3.48) treats all the information in X_S as legitimate information — even elements that could help infer S — but it ensures that the information is not in fact misdirected to infer S .

3.6 Concluding Remarks

The current practice of regulatory stress testing ignores bank heterogeneity in loss models as a matter of policy and principle. We have argued that simply pooling banks can distort coefficients

on legitimate features and is vulnerable to implicit misdirection of legitimate information to infer bank identity. We have examined various ways of incorporating fairness considerations and shown that estimating and discarding centered bank fixed effects addresses the deficiencies of pooling — and it does so in an optimal sense.

Beyond this specific recommendation, the broader conclusion to be drawn from our analysis is that accuracy and equal treatment can more effectively be addressed by accounting for bank heterogeneity rather than ignoring it. Although we have focused on the stress testing application, our analysis applies more generally to settings requiring the fair aggregation of individually tailored models into a single common model.

Epilogue

The landscape of financial data science is both challenging and rewarding. In this thesis, I have explored three areas within this domain: the quest for robust and reliable models in the face of data imbalance, the strategic exploitation of market anomalies, and the pursuit of fairness in regulatory practices. Each of these areas underscores the complexity and importance of designing fair and robust data science models in finance.

The research on linear classifiers under infinite imbalance proved particularly enlightening, revealing the often neglected aspects of the most commonly used classification tools in skewed data environments. It reinforced the notion that robust modeling is not just an empirical exercise but also requires theoretical foundations.

The exploration of TP_2 and RR_2 violations in options pricing opened new avenues for strategic trading. It was fascinating to see how theoretical conditions could translate into real-world trading opportunities, providing consistent positive returns. This study serves as a reminder of the potential that lies in uncovering and understanding market anomalies.

Addressing the fairness of regulatory stress tests highlighted the critical role of equitable practices in maintaining financial stability. The proposed method for aggregating individual models into a common industry model showcased the importance of precision and fairness in regulatory assessments, ensuring that all institutions are evaluated on a level playing field.

Looking ahead, the field of financial data science will continue to evolve. Emerging technologies, such as artificial intelligence and machine learning, promise to bring new insights and capabilities. Nevertheless, the core principles of robustness, fairness, and strategic thinking will

remain central to navigating the complexities of financial markets.

As I conclude this thesis, I am filled with a sense of accomplishment and anticipation for the future. I look forward to seeing how these findings will influence future research and practice, and I am excited about the continued exploration of fair and robust financial models in the years to come.

Bibliography

- [1] Agarwal, S., An, X., Cordell, L., and Roman, R.A. (2020) Bank stress test results and their impact on consumer credit markets. Working paper 20-30, Federal Reserve Bank of Philadelphia.
- [2] Agueh, M., and Carlier, G. (2011) Barycenters in the Wasserstein space, *SIAM Journal on Mathematical Analysis* 43(2), 904–924
- [3] Anderson, T.W. (2003) *An Introduction to Multivariate Statistical Analysis*, Third Edition. Wiley, Hoboken, New Jersey.
- [4] Angrist, J.D. and Pischke, J.S. (2008) *Mostly Harmless Econometrics*, Princeton University Press, Princeton, New Jersey.
- [5] Ansley, C.F., and Kohn, R. (1994) Convergence of the backfitting algorithm for additive models, *Journal of the Australian Mathematical Society (Series A)* 57, 316–329.
- [6] Arneson, R. (2015) Equality of opportunity, *Stanford Encyclopedia of Philosophy* (Summer 2015 edition), Edward N. Zalta, ed.
- [7] Baer, G., and Hopper, G. (2023) The Fed’s stress test models are inaccurate. Something has to change. *Risk*, September 14.
- [8] Barocas, S., Hardt, M., and Narayanan, A. (2019) *Fairness in Machine Learning*, <https://fairmlbook.org/>
- [9] Bassett, W.F., and Berrospide, J.M. (2018) The impact of post stress test capital on bank lending. Working paper 2018-097, Federal Reserve Board, Washington, D.C.
- [10] BCBS (2019) Overview of Pillar 2 supervisory review practices and approaches, Bank for International Settlements, Basel, Switzerland.
- [11] Board of Governors (2021) Dodd-Frank Act Stress Test 2021: Supervisory Stress Test Methodology. Federal Reserve System, Washington, D.C.
- [12] Breiman, L., and Friedman, J.H. (1985) Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association* 80, 580–598.
- [13] Breuer, T., Jandacka, M., Rheinberger, K., and Summer, M. (2009) How to find plausible, severe, and useful stress scenarios. *International Journal of Central Banking* 5, 205–224.
- [14] Brown, L.D. (1986) *Fundamental of Statistical Exponential Families*, Institute of Mathematical Statistics, Hayward, California.
- [15] Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P. (2002) SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research* 16:321–357.

- [16] Chzhen, E., Denis, C., Hebiri, M., Oneto, L., and Pontil, M. (2020) Fair regression with Wasserstein barycenters, *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*.
- [17] Cope, D., Hsu, C., Lively, C., Morgan, J., Schuermann, T., and Sekeris, E. (2022) Stress testing for commercial, investment, and custody banks. *Handbook of Financial Stress Testing*, 247–270, Cambridge University Press.
- [18] Covas, F.B., Rump, B., and Zakrajsek, E. (2014) Stress-testing US bank holding companies: A dynamic panel quantile regression approach. *International Journal of Forecasting* 30, 691–713.
- [19] Csiszar, I. (1975) I-Divergence Geometry of Probability Distributions and Minimization Problems. *Annals of Probability* 3(1):146–158.
- [20] Deo, A., and Juneja, S. (2021) Credit Risk: Simple Closed-Form Approximate Maximum Likelihood Estimator. *Operations Resesarch* 69(2), 361–379.
- [21] Drummond, C. and Holte, R.C. (2003) C4. 5, Class Imbalance, and Cost Sensitivity: Why Under-Sampling Beats Over-Sampling. *Workshop on Learning from Imbalanced Datasets*, Washington, D.C.
- [22] Durrett, R. (2019) *Probability: Theory and Examples*, Fifth Edition. Cambridge University Press.
- [23] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012) Fairness through awareness, pp.214–226, in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*.
- [24] Eguchi, S., and Copas, J. (2002) A Class of Logistic-Type Discriminant Functions. *Biometrika* 89(1):1–22.
- [25] Embrechts P., Klüppelberg C., Mikosch T. (1997) Modelling Extremal Events: for Insurance and Finance. Springer, Berlin, pp 169—170.
- [26] van Erven, T., and Harremoës, P. (2014) Rényi Divergence and Kullback-Leibler Divergence, *IEEE Transactions on Information Theory* 60(7):3793–3820.
- [27] Fernandes, M., Igan, D., and Pinheiro, M. (2020) March madness in Wall Street: (What) does the market learn from stress tests? *Journal of Banking and Finance* 112, 105250.
- [28] Flannery, M.J. (2019) Transparency and model evolution in stress testing, Available at SSRN: <https://ssrn.com/abstract=3431679>.
- [29] Flannery, M., Hirtle, B., and Kovner, A. (2017) Evaluating the information in the Federal Reserve stress tests. *Journal of Financial Intermediation* 29, 1–18.
- [30] Flood, M.D., Jones, J., Pritsker, M., and Siddique, A. (2022) The role of heterogeneity in scenario design for financial stability stress testing. *Handbook of Financial Stress Testing*, 98–127, Cambridge University Press.

- [31] Flood, M.D. and Korenko, G.G. (2015) Systematic scenario selection: stress testing and the nature of uncertainty. *Quantitative Finance* 15, 43–59.
- [32] Freddie Mac (2021) Single Family Loan-Level Dataset General User Guide. Freddie Mac, McLean, Virginia.
- [33] Freund, Y., and Schapire, R.E. (1997) A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences* 55:119–139.
- [34] Friedman, J., Hastie, T., and Tibshirani, R. (2000) Additive Logistic Regression: A Statistical View of Boosting. *Annals of Statistics* 28(2):337–407.
- [35] Georgescu, O.-M., Gross, M., Kapp, D., and Kok, C. (2017) Do stress tests matter? Evidence from the 2014 and 2016 stress test. Working paper 2054, European Central Bank, Frankfurt, Germany.
- [36] Gill, P.M., Pearce, C.E.M., and Pečarić, J. (1997) Hadamard’s Inequality for r -Convex Functions, *Journal of Mathematical Analysis and Applications* 215:461–470.
- [37] Glasserman, P., Kang, C., and Kang, W. (2015) Stress scenario selection by empirical likelihood. *Quantitative Finance* 15, 25–41.
- [38] Glasserman, P. and Pirjol, D. (2022) Total Positivity and Relative Convexity of Option Prices. to appear in *Frontiers in Mathematical Finance*, special issue dedicated to the Peter Carr Gedenkschrift Conference, Nov. 12-13, 2022., Available at SSRN: <https://ssrn.com/abstract=4283971>
- [39] Glasserman, P. and Pirjol, D. (2024) When Are Options TP2?. Working paper.
- [40] Glasserman, P., and Tangirala, G. (2016) Are the Federal Reserve’s stress test results predictable? *Journal of Alternative Investments: Systemic Risk Special Edition* 18, 82–97.
- [41] Green, B. (2022) Escaping the impossibility of fairness: from formal to substantive algorithmic fairness. *Philosophy & Technology* 35:90, 1–32.
- [42] Grūnewālder, S., and Khaleghi, A. (2021) Oblivious data for fairness with kernels, *Journal of Machine Learning Research* 22, 1–36.
- [43] Guerrieri, L., and Modugno, M. (2021) The information content of stress test announcements. Working paper 2012-012, Federal Reserve Board, Washington, D.C.
- [44] Guerrieri, L., and Welch, M. (2012) Can macro variables used in stress test forecast the performance of banks? Working paper 2012-49, Federal Reserve Board, Washington, D.C.
- [45] Hastie, T., and Tibshirani, R. (1986) Generalized additive models, *Statistical Science* 1(3), 297–318.

- [46] Hirtle, B., Kovner, A., Vickery, J., and Bhanot, M. (2016) Assessing financial stability: The Capital and Loss Assessment under Stress Scenarios (CLASS) model. *Journal of Banking and Finance* 69, S35–S55.
- [47] Hutchinson, B., and Mitchell, M. (2019) 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 49–58.
- [48] Johnson, K., Foster, D., and Stine, R. (2020) Impartial predictive modeling: ensuring group fairness in arbitrary models, arXiv:1608.00528.
- [49] Kapinos, P., and Mitnik, O.A. (2016) A top-down approach to stress-testing banks. *Journal of Financial Services Research* 49, 229–264.
- [50] Karlin, S. (1968) *Total Positivity*, Stanford University Press, Stanford, CA, 1968.
- [51] M. Keller-Ressel. (2021) Total positivity and the classification of term structure shapes in the two-factor Vasicek model, *International Journal of Theoretical and Applied Finance*, 24 (2021), 2150027, 27 pp. doi: 10.2139/ssrn.3441116.
- [52] Kohn, D., and Liang, N. (2019) *Understanding the effects of the U.S. stress tests*. Brookings Institution, Washington, D.C.
- [53] Kubat, M., and Matwin, S. (1997) Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. *Proceedings of the 14th International Conference in Machine Learning*. Morgan Kaufmann, San Francisco, 179–186.
- [54] Kullback, S., and Leibler, R.A. (1951) On Information and Sufficiency. *Annals of Mathematical Statistics* 22(1):79–86.
- [55] Kupiec, P. (2020) Policy uncertainty and bank stress testing. *Journal of Financial Stability* 51, 100761.
- [56] Le Gouic, T., Loubes, J.-M., and Rigollet, P. (2020), Projection to fairness in statistical learning, arXiv:2005.11720
- [57] Lehmann, E.L., and Romano, J.P. (2005) *Testing Statistical Hypotheses*, 3rd Ed., Springer.
- [58] Li, Y., Bellotti, T., and Adams, N. (2019) Issues Using Logistic Regression with Class Imbalance, with a Case Study from Credit Risk Modelling. *Foundations of Data Science* 1(4):389–417.
- [59] Lipton, Z., Chouldechova, A., and McAuley, J. (2018) Does mitigating ML’s impact disparity require treatment disparity?, *32nd Conference on Neural Information Processing Systems*, Montréal, Canada.
- [60] Liu, X.-Y., Wu, J., and Zhou, Z.-H. (2008) Exploratory Undersampling for Class-Imbalance Learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39(2):539–550.

- [61] Madras, D., Creager, E., Pitassi, T., and Zemel, R. (2018) Learning adversarially fair and transferable representations, arXiv:1802.06309
- [62] McClish, D. (1989) Analyzing a Portion of the ROC Curve. *Medical Decision Making* 9:190–195.
- [63] Morgan, D.P., Peristiani, S., and Savino, V. (2014) The information value of the stress test. *Journal of Money, Credit and Banking* 46, 1479–1500.
- [64] Newey, W. and West, K. (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, 55(3), 703–708. <https://doi.org/10.2307/1913610>
- [65] Owen, A. (2007) Infinitely Imbalanced Logistic Regression. *Journal of Machine Learning Research* 8:761–773.
- [66] Parlato, C., and Philippon, T. (2022) Designing stress scenarios. Working paper w29901, National Bureau of Economic Research, Cambridge, Mass.
- [67] Philippon, T., Pessarossi, P., and Camara, B. (2017) Backtesting european stress tests. Working paper w23083, National Bureau of Economic Research, Cambridge, Mass.
- [68] Prince, A.E., and Schwarcz, D. (2019) Proxy discrimination in the age of artificial intelligence and big data. *Iowa Law Review* 105, 1257–1358.
- [69] Pritsker, M.G. (2017) Choosing stress scenarios for systemic risk through dimension reduction. Risk and Policy Analysis Unit Paper No. RPA 17-4, Federal Reserve Bank of Boston.
- [70] Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011) pROC: an Open-source Package for R and S+ to Analyze and Compare ROC Curves. *BMC Bioinformatics*, 12:77.
- [71] Sahin, C., de Haan, J., and Neretina, E. (2020) Banking stress test effects on returns and risks. *Journal of Banking and Finance* 117, 105843.
- [72] Schuermann, T. (2020) Capital adequacy pre- and postcrisis and the role of stress testing. *Journal of Money, Credit and Banking* 52, 87–105.
- [73] Silvapulle, M. (1981) On the Existence of Maximum Likelihood Estimates for the Binomial Response Models. *Journal of the Royal Statistical Society, Series B* 43:310–313.
- [74] Wooldridge, J.M. (2010) *Econometric Analysis of Cross Section and Panel Data*, Second Edition, MIT Press.
- [75] Zipkin, P. (1992) The relationship between risk and maturity in a stochastic setting. *Mathematical Finance*, 2 (1992), 33-46. doi: 10.1111/j.1467-9965.1992.tb00024.x.

Appendix A: Robustness of Linear Classifiers under Infinite Imbalance

A.1 Proofs for Section 1.3

Proof of Lemma 1.3.1. Differentiating (1.4) yields

$$\frac{dU}{du} = w(u), \quad \frac{d^2U}{du^2} = w'(u)$$

and

$$\frac{dV}{du} = e^u w(u), \quad \frac{d^2V}{du^2} = e^u (w(u) + w'(u)),$$

so Condition 2 implies that U is concave and V is strictly convex. It follows that each term $-U(\alpha + \beta^\top x_i)$ and $V(\alpha + \beta^\top X_i)$ is convex in (α, β) , and thus that \bar{C}_N in (1.7) is convex.

To establish strict convexity of \bar{C}_N , we first claim that, almost surely, X_1, \dots, X_N do not fall on a hyperplane, for all sufficiently large N . Permuting the order of the X_i does not change whether they fall on a hyperplane, so we may apply the Hewitt-Savage zero-one law (as in, e.g., Durrett [22], p.71) to conclude that the probability that all X_i fall on a hyperplane is zero or one. By the surrounding condition, F_0 is not supported on any hyperplane, so the probability that all X_i fall on a hyperplane is less than one and must therefore be zero.

Suppose, then, that N is sufficiently large that X_1, \dots, X_N do not fall on a hyperplane. Then for any distinct (α_1, β_1) and (α_2, β_2) there is some $i_0 \in \{1, \dots, N\}$ for which $\alpha_1 + \beta_1^\top X_{i_0} \neq \alpha_2 + \beta_2^\top X_{i_0}$.

Then, for any $\nu \in (0, 1)$,

$$\begin{aligned}
& \sum_{i=1}^N V(\nu(\alpha_1 + \beta_1^\top X_i) + (1 - \nu)(\alpha_2 + \beta_2^\top X_i)) \\
&= V(\nu(\alpha_1 + \beta_1^\top X_{i_0}) + (1 - \nu)(\alpha_2 + \beta_2^\top X_{i_0})) + \sum_{i \neq i_0}^N V(\nu(\alpha_1 + \beta_1^\top X_i) + (1 - \nu)(\alpha_2 + \beta_2^\top X_i)) \\
&< \nu V(\alpha_1 + \beta_1^\top X_{i_0}) + (1 - \nu)V(\alpha_2 + \beta_2^\top X_{i_0}) + \sum_{i \neq i_0}^N V(\nu(\alpha_1 + \beta_1^\top X_i) + (1 - \nu)(\alpha_2 + \beta_2^\top X_i)) \\
&\leq \nu \sum_{i=1}^N V(\alpha_1 + \beta_1^\top X_i) + (1 - \nu) \sum_{i=1}^N V(\alpha_2 + \beta_2^\top X_i).
\end{aligned}$$

The strict inequality follows from the strict convexity of V . Strict convexity of \bar{C}_N follows. \square

The following result allows us to translate a surrounding condition on F_0 to a surrounding condition on its empirical counterpart.

Lemma A.1.1. *Let \hat{F}_N be the empirical distribution of independent random variables X_1, \dots, X_N drawn from F . Suppose F surrounds x^* with parameters (ϵ, δ) . Then for any $\epsilon_1 < \epsilon$ and $\delta_1 < \delta$, \hat{F}_N surrounds x^* with parameters (ϵ_1, δ_1) for all sufficiently large N , a.s.*

Proof for Lemma A.1.1. For any constants $M > 0$ and $\lambda \in (0, 1)$, we can choose fixed points $\nu_1, \dots, \nu_K \in \Omega$, with K depending on M and λ , such that, for every $\omega \in \Omega$,

$$\min_{k=1, \dots, K} \|\omega - \nu_k\| \leq \frac{\lambda \epsilon}{M}. \tag{A.1}$$

This follows from the relative compactness of Ω . For any $0 < \delta' < \delta$, we may take M sufficiently large that

$$\mathbf{P}(\|X - x^*\| > M) < \delta',$$

with X having distribution F .

It follows from (A.1) that for any sequence $\omega_N \in \Omega$ we may choose $k_N \in \{1, \dots, K\}$ such that,

for all N ,

$$\|\omega_N - v_{k_N}\| \leq \frac{\lambda\epsilon}{M}.$$

The sequences ω_N and k_N may be stochastic. For any $x \in \mathbb{R}^d$,

$$\begin{aligned} \mathbf{1}\{(x - x^*)^\top v_{k_N} > \epsilon\} &= \mathbf{1}\{(x - x^*)^\top \omega_N + (x - x^*)^\top (v_{k_N} - \omega_N) > \epsilon\} \\ &\leq \mathbf{1}\{(x - x^*)^\top \omega_N + \|x - x^*\| \lambda\epsilon/M > \epsilon\} \\ &\leq \mathbf{1}\{(x - x^*)^\top \omega_N > (1 - \lambda)\epsilon\} + \mathbf{1}\{\|x - x^*\| \lambda\epsilon/M > \lambda\epsilon\} \\ &= \mathbf{1}\{(x - x^*)^\top \omega_N > (1 - \lambda)\epsilon\} + \mathbf{1}\{\|x - x^*\| > M\}. \end{aligned}$$

Thus, for any $i = 1, \dots, N$,

$$\mathbf{1}\{(X_i - x^*)^\top \omega_N > (1 - \lambda)\epsilon\} \geq \mathbf{1}\{(X_i - x^*)^\top v_{k_N} > \epsilon\} - \mathbf{1}\{\|X_i - x^*\| > M\},$$

a.s., and also, a.s.,

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{(X_i - x^*)^\top \omega_N > (1 - \lambda)\epsilon\} &\geq \\ \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{(X_i - x^*)^\top v_{k_N} > \epsilon\} - \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{\|X_i - x^*\| > M\}. \end{aligned} \quad (\text{A.2})$$

For the first term on the right, note that the strong law of large numbers and the surrounding condition for F imply that, for each $k = 1, \dots, K$, we have the almost sure limit

$$\frac{1}{N} \sum_{i=1}^N \mathbf{1}\{(X_i - x^*)^\top v_k > \epsilon\} \rightarrow \mathbf{P}((X - x^*)^\top v_k > \epsilon) > \delta,$$

so

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{(X_i - x^*)^\top v_{k_N} > \epsilon\} > \delta, \quad \text{a.s.}$$

For the second term on the right side of (A.2), we have, a.s.,

$$\frac{1}{N} \sum_{i=1}^N \mathbf{1}\{\|X_i - x^*\| > M\} \rightarrow \mathbf{P}(\|X - x^*\| > M) < \delta',$$

by the strong law of large numbers and our choice of M . Thus, (A.2) yields

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{(X_i - x^*)^\top \omega_N > (1 - \lambda)\epsilon\} > \delta - \delta', \quad \text{a.s.}$$

This implies

$$\liminf_{N \rightarrow \infty} \min_{\omega \in \Omega} \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{(X_i - x^*)^\top \omega > (1 - \lambda)\epsilon\} > \delta - \delta', \quad \text{a.s.}$$

because for each N the sum takes only finitely many values as ω varies, so the minimum over ω is attained, and we may take ω_N to be the minimizing ω . It follows that,

$$\min_{\omega \in \Omega} \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{(X_i - x^*)^\top \omega > (1 - \lambda)\epsilon\} > \delta - \delta',$$

for all sufficiently large N , a.s. As λ and δ' may be arbitrarily close to zero, the result follows. \square

Proof of Lemma 1.3.2. We know from Lemma 1.3.1 that \bar{C}_N is strictly convex in (α, β) for all sufficiently large N , a.s. Strict convexity implies that either \bar{C}_N has a unique minimizer or it strictly decreases along some ray $\{(\lambda\alpha_0, \lambda\beta_0) | 0 \leq \lambda < \infty\}$, with (α_0, β_0) not identically zero. We will show that the latter case is not possible.

We treat separately the cases $\beta_0 = 0$ (with $\alpha_0 \neq 0$) and $\beta_0^\top \beta_0 = 1$. The normalization in the second case is justified by our scaling by λ .

Case 1: $\beta_0 = 0, \alpha_0 \neq 0$. Differentiation yields

$$\begin{aligned} \frac{\partial \bar{C}_N(\lambda\alpha_0, \lambda\beta_0)}{\partial \lambda} &= - \sum_{i=1}^n \frac{\partial}{\partial \lambda} U(\lambda\alpha_0) + \sum_{i=1}^N \frac{\partial}{\partial \lambda} V(\lambda\alpha_0) \\ &= -nw(\lambda\alpha_0)\alpha_0 + Ne^{\lambda\alpha_0}w(\lambda\alpha_0)\alpha_0 \\ &= (-n + Ne^{\lambda\alpha_0})w(\lambda\alpha_0)\alpha_0. \end{aligned}$$

If $\alpha_0 > 0$, then for λ large, $e^{\lambda\alpha_0} > n/N$, and the derivative is strictly positive. If $\alpha_0 < 0$, then for λ large, $e^{\lambda\alpha_0} < n/N$ and the derivative is again strictly positive.

Case 2: $\beta_0^\top \beta_0 = 1$. Differentiation yields

$$\begin{aligned} \frac{\partial \bar{C}_N(\lambda\alpha_0, \lambda\beta_0)}{\partial \lambda} &= \sum_{i=1}^n -w(\lambda\alpha_0 + \lambda\beta_0^\top x_i)(\alpha_0 + \beta_0^\top x_i) \\ &\quad + \sum_{i=1}^N e^{\lambda\alpha_0 + \lambda\beta_0^\top X_i} w(\lambda\alpha_0 + \lambda\beta_0^\top X_i)(\alpha_0 + \beta_0^\top X_i) \\ &= \sum_{i:\alpha_0 + \beta_0^\top x_i < 0} -w(\lambda\alpha_0 + \lambda\beta_0^\top x_i)(\alpha_0 + \beta_0^\top x_i) \end{aligned} \quad (\text{A.3})$$

$$+ \sum_{i:\alpha_0 + \beta_0^\top x_i > 0} -w(\lambda\alpha_0 + \lambda\beta_0^\top x_i)(\alpha_0 + \beta_0^\top x_i) \quad (\text{A.4})$$

$$+ \sum_{i:\alpha_0 + \beta_0^\top X_i < 0} e^{\lambda\alpha_0 + \lambda\beta_0^\top X_i} w(\lambda\alpha_0 + \lambda\beta_0^\top X_i)(\alpha_0 + \beta_0^\top X_i) \quad (\text{A.5})$$

$$+ \sum_{i:\alpha_0 + \beta_0^\top X_i > 0} e^{\lambda\alpha_0 + \lambda\beta_0^\top X_i} w(\lambda\alpha_0 + \lambda\beta_0^\top X_i)(\alpha_0 + \beta_0^\top X_i). \quad (\text{A.6})$$

We will show that as λ increases, the liminf of the derivative on the left is strictly positive. We will prove this by showing that (A.4) and (A.5) approach zero as λ increases, and the sum of (A.3) and (A.6) remains positive and bounded away from zero as λ increases.

Recall from the comments after Condition 1 that $w(s) \rightarrow 0$ and $e^{-s}w(-s) \rightarrow 0$ as $s \rightarrow \infty$. Thus, in (A.4), $\alpha_0 + \beta_0^\top x_i > 0$ implies $\lambda(\alpha_0 + \beta_0^\top x_i) \rightarrow \infty$, and $w(\lambda\alpha_0 + \lambda\beta_0^\top x_i) \rightarrow 0$. In (A.5), $\alpha_0 + \beta_0^\top X_i < 0$ implies $e^{\lambda\alpha_0 + \lambda\beta_0^\top X_i} w(\lambda\alpha_0 + \lambda\beta_0^\top X_i) \rightarrow 0$. Thus, (A.4) and (A.5) approach zero as λ increases.

The terms in (A.3) and (A.6) are nonnegative. We need to show that at least one of them remains bounded away from zero. Using the fact that $w(s) > 0$ is decreasing we get a lower bound

for (A.3),

$$\begin{aligned}
\sum_{i:\alpha_0+\beta_0^\top x_i < 0} -w(\lambda\alpha_0 + \lambda\beta_0^\top x_i)(\alpha_0 + \beta_0^\top x_i) &\geq -w(0) \sum_{i:\alpha_0+\beta_0^\top x_i < 0} (\alpha_0 + \beta_0^\top x_i) \\
&\geq -w(0) \sum_{i=1}^n (\alpha_0 + \beta_0^\top x_i) \\
&= -n \cdot w(0)(\alpha_0 + \beta_0^\top \bar{x}). \tag{A.7}
\end{aligned}$$

Using the fact that $e^s w(s)$ is increasing we get a lower bound for (A.6),

$$\begin{aligned}
\sum_{i:\alpha_0+\beta_0^\top X_i > 0} e^{\lambda\alpha_0+\lambda\beta_0^\top X_i} w(\lambda\alpha_0 + \lambda\beta_0^\top X_i)(\alpha_0 + \beta_0^\top X_i) \\
&\geq w(0) \sum_{i=1}^N \mathbf{1}\{\alpha_0 + \beta_0^\top X_i > 0\}(\alpha_0 + \beta_0^\top X_i) \\
&\geq w(0)\epsilon_1 \sum_{i=1}^N \mathbf{1}\{\alpha_0 + \beta_0^\top X_i > \epsilon_1\} \\
&= w(0)\epsilon_1 \sum_{i=1}^N \mathbf{1}\{\beta_0^\top (X_i - \bar{x}) > \epsilon_1 - (\alpha_0 + \beta_0^\top \bar{x})\}. \tag{A.8}
\end{aligned}$$

In light of Lemma A.1.1, we may suppose N is sufficiently large that the empirical distribution of X_1, \dots, X_N surrounds \bar{x} . If $\alpha_0 + \beta_0^\top \bar{x} < 0$, then (A.7) is strictly positive; if $\alpha_0 + \beta_0^\top \bar{x} \geq 0$, then the surrounding condition implies that (A.8) is strictly positive, for sufficiently small $\epsilon_1 > 0$. \square

A.2 Proofs for Section 1.4

A.2.1 Proof of Lemma 1.4.1

Proof of Lemma 1.4.1. To show that the three examples of functions satisfy the left-tail conditions in Definition 1.4.1, we need to show that for any $k \geq 0$ and $C > 0$, $h(u) = C|u|^k$ satisfies the requirements on h in Definition 1.4.1. Then we can choose $\lambda = k = 0$; or $\lambda = 0$ and $k > 0$; or $\lambda \in (0, 1)$ and $k > 0$ for the three cases of weight functions.

To see why $h(u) = C|u|^k$ satisfies the requirements on h in Definition 1.4.1, we note that for

$u < 0$, $h'(u) = -Ck|u|^{k-1}$, and

$$\liminf_{u \rightarrow -\infty} h'(u)/h(u) = \liminf_{u \rightarrow -\infty} -k/|u| = 0,$$

so left-tail condition (iii) in Definition 1.4.1 is satisfied. Now notice that

$$\frac{h(u+s)}{h(u)} = \left|1 + \frac{s}{u}\right|^k.$$

Let $\epsilon > 0$. We may find $u_1 < 0$ such that for any $u \leq u_1$ and any $|s| \leq 1$,

$$1 - \epsilon \leq \left|1 + \frac{s}{u}\right|^k \leq 1 + \epsilon. \quad (\text{A.9})$$

We may find $u_2 < 0$ such that for any $u \leq u_2$ and any $|s| > 1$,

$$1 - \epsilon|s|^k \leq \left|1 + \frac{s}{u}\right|^k \leq 1 + \epsilon|s|^k. \quad (\text{A.10})$$

Combining (A.9) and (A.10), we have for $u \leq \min\{u_1, u_2\}$ and for any s ,

$$1 - \epsilon \max\{1, |s|^k\} \leq \left|1 + \frac{s}{u}\right|^k \leq 1 + \epsilon \max\{1, |s|^k\}. \quad (\text{A.11})$$

That is,

$$\left| \left|1 + \frac{s}{u}\right|^k - 1 \right| \leq \epsilon \max\{1, |s|^k\},$$

so (1.9) is satisfied by any $C > 1$, $\xi > 0$, $s_0 > 1$ such that $C \geq \max\{1, |s|^k\}$ for $|s| < s_0$, and $e^{\xi|s|} \geq |s|^k$ for $|s| \geq s_0$. \square

A.2.2 Proof of Proposition 1.4.1

Proof of Proposition 1.4.1. Suppose $w(u) \sim e^{-\lambda u} h(u)$, where λ and $h(u)$ satisfy the conditions in Definition 1.4.1. Suppose $\tilde{\lambda}$ and \tilde{h} also satisfy the conditions in Definition 1.4.1 and $w(u) \sim$

$e^{-\tilde{\lambda}u} \tilde{h}(u)$.

Let $\Delta = \tilde{\lambda} - \lambda$. Then $w(u) \sim e^{-(\lambda+\Delta)u} \tilde{h}(u)$. Since $w(u) \sim e^{-\lambda u} h(u)$, we must have $\tilde{h}(u) \sim e^{\Delta u} h(u)$. We will show that \tilde{h} fails left-tail condition (iv) in Definition 1.4.1 unless $\Delta = 0$.

By (1.9), for any $s \in \mathbb{R}$,

$$\frac{h(u+s)}{h(u)} \rightarrow 1,$$

as $u \rightarrow -\infty$. Therefore,

$$\lim_{u \rightarrow -\infty} \frac{\tilde{h}(u+s)}{\tilde{h}(u)} - 1 = \lim_{u \rightarrow -\infty} \frac{h(u+s)}{h(u)} e^{\Delta s} - 1 = e^{\Delta s} - 1,$$

and for $s \neq 0$,

$$\left| \frac{\tilde{h}(u+s)}{\tilde{h}(u)} - 1 \right| \not\rightarrow 0$$

violating (1.9) unless $\Delta = 0$. Thus, (1.9) requires $\tilde{\lambda} = \lambda$ and $\tilde{h}(u) \sim h(u)$ as $u \rightarrow -\infty$. \square

A.2.3 A Convergence Result

The following proposition is key to our main result.

Proposition A.2.1. *Suppose Conditions 1–4 hold, and suppose w satisfies Definition 1.4.1 with $w(u) \sim e^{-\lambda u} h(u)$ as $u \rightarrow -\infty$. Let $\gamma = (1 - \lambda)\epsilon\delta$, where $\epsilon, \delta > 0$ are the surrounding parameters in Condition 3. Then, almost surely,*

$$\alpha_N \rightarrow -\infty \quad \text{and} \quad \limsup_{N \rightarrow \infty} \|\beta_N\| \leq 1/\gamma. \quad (\text{A.12})$$

We separate the proof into two steps, first showing that $\alpha_N + \beta_N^\top \bar{x} \rightarrow -\infty$, a.s., and then showing (A.12).

Lemma A.2.1 (Step 1). *Under the conditions of Proposition A.2.1, $\alpha_N + \beta_N^\top \bar{x} \rightarrow -\infty$, a.s.*

We will prove two cases separately: (i) bounded weight functions with $\lambda = 0$ and $h(u) \equiv C > 0$ in Definition 1.4.1, and (ii) unbounded weight functions with non-constant h or $\lambda > 0$.

Proof of Lemma A.2.1 for bounded weight functions. Recalling that w is decreasing, let $C = \lim_{u \rightarrow -\infty} w(u)$ and let (ϵ, δ) be the parameters in Condition 3. Then at any (α, β) and for any $\delta_1 \in (0, \delta)$,

$$\begin{aligned}
\frac{\partial \bar{C}_N}{\partial \alpha} &= \sum_{i=1}^n -w(\alpha + \beta^\top x_i) + \sum_{i=1}^N e^{\alpha + \beta^\top X_i} w(\alpha + \beta^\top X_i) \\
&= \sum_{i=1}^n -w(\alpha + \beta^\top x_i) + \sum_{i=1}^N e^{\alpha + \beta^\top \bar{x} + \beta^\top (X_i - \bar{x})} w(\alpha + \beta^\top \bar{x} + \beta^\top (X_i - \bar{x})) \\
&\geq -nC + e^{\alpha + \beta^\top \bar{x}} w(\alpha + \beta^\top \bar{x}) \sum_{i=1}^N \mathbf{1}\{\beta^\top (X_i - \bar{x}) \geq 0\} \\
&\geq -nC + Ne^{\alpha + \beta^\top \bar{x}} w(\alpha + \beta^\top \bar{x}) \delta_1,
\end{aligned} \tag{A.13}$$

for all sufficiently large N , a.s., where going from the second to the third line we used the conditions that $w(u)$ is decreasing and $w(u)e^u$ is increasing, and going from the third to the fourth line we applied Lemma A.1.1.

Consider any (α, β) for which $e^{\alpha + \beta^\top \bar{x}} w(\alpha + \beta^\top \bar{x}) > nC/(N\delta_1)$. For any such (α, β) , (A.13) implies that $\partial \bar{C}_N / \partial \alpha > 0$. It follows that no such (α, β) can be optimal; the optimal (α_N, β_N) must satisfy the reverse inequality

$$e^{\alpha_N + \beta_N^\top \bar{x}} w(\alpha_N + \beta_N^\top \bar{x}) \leq nC/(N\delta_1), \tag{A.14}$$

from which we get $\alpha_N + \beta_N^\top \bar{x} \rightarrow -\infty$, a.s., which completes the proof for the bounded case. \square

To prove Lemma A.2.1 for unbounded weight functions, we will need the following result. (The following result also holds for bounded weight functions, which will be useful in Corollary A.2.1.)

Lemma A.2.2. *Suppose Conditions 1, 2, and 4 hold. Then*

$$\min_i \alpha_N + \beta_N^\top x_i \rightarrow -\infty. \tag{A.15}$$

Proof of Lemma A.2.2. Let $(\epsilon^\circ, \delta^\circ)$ be the surrounding parameters in Condition 4. Then for any

$j = 1, \dots, n$ at any (α, β) and for any $\delta_1 \in (0, \delta^o)$,

$$\begin{aligned}
\frac{\partial \bar{C}_N}{\partial \alpha} &= \sum_{i=1}^n -w(\alpha + \beta^\top x_i) + \sum_{i=1}^N e^{\alpha + \beta^\top X_i} w(\alpha + \beta^\top X_i) \\
&= \sum_{i=1}^n -w(\alpha + \beta^\top x_i) + \sum_{i=1}^N e^{\alpha + \beta^\top x_j + \beta^\top (X_i - x_j)} w(\alpha + \beta^\top x_j + \beta^\top (X_i - x_j)) \\
&\geq \sum_{i=1}^n -w(\alpha + \beta^\top x_i) + e^{\alpha + \beta^\top x_j} w(\alpha + \beta^\top x_j) \sum_{i=1}^N \mathbf{1}\{\beta^\top (X_i - x_j) \geq 0\} \\
&\geq -n \max_i w(\alpha + \beta^\top x_i) + N e^{\alpha + \beta^\top x_j} w(\alpha + \beta^\top x_j) \delta_1^o, \tag{A.16}
\end{aligned}$$

for all sufficiently large N , a.s., where going from the second to the third line we used the conditions that $w(u)$ is decreasing and $w(u)e^u$ is increasing, and going from the third to the fourth line we applied Lemma A.1.1.

Let $j(N) \in \operatorname{argmin}_{i=1,2,\dots,n} \{\alpha_N + \beta_N^\top x_i\}$. Because $-w(s)$ and $e^s w(s)$ are monotonically increasing, at the minimizer (α_N, β_N) we have

$$\begin{aligned}
0 = \frac{\partial \bar{C}_N}{\partial \alpha}(\alpha_N, \beta_N) &\geq -n w(\alpha_N + \beta_N^\top x_{j(N)}) + N e^{\alpha_N + \beta_N^\top x_{j(N)}} w(\alpha_N + \beta_N^\top x_{j(N)}) \delta_1^o \\
&= w(\alpha_N + \beta_N^\top x_{j(N)}) (-n + N e^{\alpha_N + \beta_N^\top x_{j(N)}} \delta_1^o), \tag{A.17}
\end{aligned}$$

so, $\alpha_N + \beta_N^\top x_{j(N)} \rightarrow -\infty$ a.s., which is (A.15). \square

We can now prove Lemma A.2.1 for unbounded w . Recall that for unbounded w we require the right-tail condition in Definition 1.4.1.

Proof of Lemma A.2.1 for unbounded weight functions. Let (ϵ^o, δ^o) be the surrounding parameters in Condition 4. We now use (A.15) to show that $\alpha_N + \beta_N^\top \bar{x} \rightarrow -\infty$ almost surely.

For $j = 1, \dots, n$, we introduce the centered loss (centered around x_j)

$$C^j(\alpha, \beta) = \sum_{i=1}^n -U(\alpha + \beta^\top (x_i - x_j)) + \sum_{k=1}^N V(\alpha + \beta^\top (X_k - x_j)). \tag{A.18}$$

With (α_N, β_N) the minimizer of \bar{C}_N , the centered loss C^j is minimized at (α_N^j, β_N^j) , where

$$\alpha_N^j = \alpha_N + \beta_N^\top x_j, \quad \beta_N^j = \beta_N.$$

Consider

$$\begin{aligned} & C^j(\alpha, 0) - C^j(\alpha, \beta) \\ &= \sum_{i=1}^n [-U(\alpha) + U(\alpha + \beta^\top(x_i - x_j))] + NV(\alpha) - \sum_{k=1}^N V(\alpha + \beta^\top(X_k - x_j)). \end{aligned} \quad (\text{A.19})$$

Since U is concave with $dU/du = w(u)$,

$$U(\alpha + \beta^\top(x_i - x_j)) \leq U(\alpha) + w(\alpha)\beta^\top(x_i - x_j)$$

and

$$\sum_{i=1}^n [-U(\alpha) + U(\alpha + \beta^\top(x_i - x_j))] \leq \sum_{i=1}^n w(\alpha)\beta^\top(x_i - x_j) \leq nw(\alpha)\|\beta\|C, \quad (\text{A.20})$$

where $C = \max_{i,j}\|x_i - x_j\|$.

Similarly, V is convex and strictly positive with $dV/du = e^u w(u)$, so, for $x \neq x_j$,

$$V(\alpha + \beta^\top(x - x_j)) \geq [V(\alpha) + e^\alpha w(\alpha)\beta^\top(x - x_j)]_+ \geq e^\alpha w(\alpha)[\beta^\top(x - x_j)]_+. \quad (\text{A.21})$$

By Condition 4 and Lemma A.1.1, the empirical distribution of X_1, \dots, X_N surrounds all x_i , $i = 1, 2, \dots, n$, for any parameters $\epsilon_1^o \in (0, \epsilon^o)$ and $\delta_1^o \in (0, \delta^o)$, for all sufficiently large N a.s. Therefore,

$$\min_i \inf_{\omega \in \Omega} \sum_{k=1}^N [(X_k - x_i)^\top \omega]_+ \geq \min_i \inf_{\omega \in \Omega} \sum_{k: (X_k - x_i)^\top \omega > \epsilon_1^o}^N [(X_k - x_i)^\top \omega]_+ \geq \epsilon_1^o \delta_1^o \equiv \gamma_1^o > 0$$

where $\Omega = \{\omega \in \mathbb{R}^d | \omega^\top \omega = 1\}$. Applying this bound with (A.21) we get

$$-\sum_{k=1}^N V(\alpha + \beta^\top (X_k - x_j)) \leq -e^\alpha w(\alpha) \sum_{k=1}^N [\beta^\top (X_k - x_j)]_+ \leq -N e^\alpha w(\alpha) \|\beta\| \gamma_1^o. \quad (\text{A.22})$$

Applying (A.20) and (A.22) in (A.19), we get

$$\begin{aligned} C^j(\alpha, 0) - C^j(\alpha, \beta) &\leq NV(\alpha) - N e^\alpha w(\alpha) \|\beta\| \gamma_1^o + n w(\alpha) \|\beta\| C \\ &= NV(\alpha) - w(\alpha) (N e^\alpha \gamma_1^o - nC) \|\beta\|. \end{aligned}$$

At the minimizer (α_N^j, β_N) , this becomes

$$0 \leq C^j(\alpha_N^j, 0) - C^j(\alpha_N^j, \beta_N) \leq NV(\alpha_N^j) - w(\alpha_N^j) (N e^{\alpha_N^j} \gamma_1^o - nC) \|\beta_N\|,$$

which implies

$$\|\beta_N\| \left(\gamma_1^o - \frac{nC}{N} e^{-\alpha_N^j} \right) \leq \frac{V(\alpha_N^j)}{e^{\alpha_N^j} w(\alpha_N^j)}. \quad (\text{A.23})$$

The right side is bounded almost surely for large α_N^j , by the right-tail condition in Definition 1.4.1. Through any subsequence N_m through which $\|\beta_{N_m}\|$ grows without bound, this inequality is eventually violated unless $\alpha_{N_m}^j \rightarrow -\infty$. Thus, if $\|\beta_{N_m}\|$ is unbounded, we must have $\alpha_{N_m}^j \equiv \alpha_{N_m} + \beta_{N_m}^\top x_j \rightarrow -\infty$. Suppose $\|\beta_{N_m}\|$ remains bounded. We know from (A.15) that $\min_i \{\alpha_{N_m} + \beta_{N_m}^\top x_i\} \rightarrow -\infty$, so we must have $\alpha_{N_m} \rightarrow -\infty$, and thus we again have $\alpha_{N_m} + \beta_{N_m}^\top x_j \rightarrow -\infty$. We conclude that $\alpha_N + \beta_N^\top x_j \rightarrow -\infty$, for all $j = 1, \dots, n$, and thus $\alpha_N + \beta_N^\top \bar{x} \rightarrow -\infty$ a.s. \square

We complete the proof of Proposition A.2.1 by showing that $\|\beta_N\|$ remains bounded. In light of Lemma A.2.1, boundedness of $\|\beta_N\|$ implies that $\alpha_N \rightarrow -\infty$ a.s., as required for (A.12).

Lemma A.2.3 (Step 2). *Under the conditions of Proposition A.2.1, $\limsup_{N \rightarrow \infty} \|\beta_N\| \leq 1/\gamma$ where $\gamma = (1 - \lambda)\epsilon\delta$, with $\epsilon, \delta > 0$ the surrounding parameters in Condition 3, and λ the exponential parameter in Definition 1.4.1.*

Proof of Lemma A.2.3. We will work with the centered loss, centered around \bar{x} ,

$$\tilde{C}_N(\alpha, \beta) = \sum_{i=1}^n -U(\alpha + \beta^\top(x_i - \bar{x})) + \sum_{i=1}^N V(\alpha + \beta^\top(X_i - \bar{x})), \quad (\text{A.24})$$

which is minimized at $(\tilde{\alpha}_N, \tilde{\beta}_N)$, with

$$\tilde{\alpha}_N = \alpha_N + \beta_N^\top \bar{x}, \quad \tilde{\beta}_N = \beta_N.$$

For any $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^d$,

$$\begin{aligned} & \tilde{C}_N(\alpha, 0) - \tilde{C}_N(\alpha, \beta) \\ &= \sum_{i=1}^n [-U(\alpha) + U(\alpha + \beta^\top(x_i - \bar{x}))] + NV(\alpha) - \sum_{i=1}^N V(\alpha + \beta^\top(X_i - \bar{x})). \end{aligned} \quad (\text{A.25})$$

We will bound the expression on the right.

Using the concavity of U , as in (A.20), we get

$$\sum_{i=1}^n [-U(\alpha) + U(\alpha + \beta^\top(x_i - \bar{x}))] \leq \sum_{i=1}^n w(\alpha) \beta^\top(x_i - \bar{x}) = 0. \quad (\text{A.26})$$

Using the convexity of V , as in (A.21), we get

$$\sum_{i=1}^N V(\alpha + \beta^\top(X_i - \bar{x})) \geq e^\alpha w(\alpha) \sum_{i=1}^N [\beta^\top(X_i - \bar{x})]_+. \quad (\text{A.27})$$

Recalling that ϵ and δ are the surrounding parameters in Condition 3, let $\epsilon_1 \in (0, \epsilon)$ and $\delta_1 \in (0, \delta)$.

By Lemma A.1.1, the empirical distribution of X_1, \dots, X_N surrounds \bar{x} with parameters (ϵ_1, δ_1)

for all sufficiently large N , a.s., and so

$$\inf_{\omega \in \Omega} \frac{1}{N} \sum_{i=1}^N [(X_i - \bar{x})^\top \omega]_+ \geq \inf_{\omega \in \Omega} \frac{1}{N} \sum_{i: (X_i - \bar{x})^\top \omega > \epsilon_1} [(X_i - \bar{x})^\top \omega]_+ \geq \epsilon_1 \delta_1 \equiv \gamma_1 > 0.$$

Applying this bound in (A.27), we get, for sufficiently large N , a.s.,

$$\sum_{i=1}^N V(\alpha + \beta^\top (X_i - \bar{x})) \geq N e^\alpha w(\alpha) \|\beta\| \gamma_1. \quad (\text{A.28})$$

Now, by applying (A.26) and (A.28) in (A.25) we get

$$\tilde{C}_N(\alpha, 0) - \tilde{C}_N(\alpha, \beta) \leq N V(\alpha) - N e^\alpha w(\alpha) \|\beta\| \gamma_1. \quad (\text{A.29})$$

Let $C = \lim_{u \rightarrow -\infty} w(u)/(e^{-\lambda u} h(u))$, with λ and h as in Definition 1.4.1. Take any $\epsilon_h \in (0, (1 - \lambda)/C)$, By property (iii) in Definition 1.4.1, $\liminf_{u \rightarrow -\infty} h'(u)/h(u) \geq 0$, so there exists some $u_h < 0$ such that for any $u \leq u_h$,

$$-\epsilon_h h(u) \leq h'(u). \quad (\text{A.30})$$

For any $\epsilon_0 > 0$, there is a $u_0 \leq u_h$ such that for all $u \leq u_0$,

$$(1 - \epsilon_0) C e^{-\lambda u} h(u) \leq w(u) \leq (1 + \epsilon_0) C e^{-\lambda u} h(u).$$

These bounds yield, for $u \leq u_0 \leq u_h$,

$$\begin{aligned} V(u) &\leq (1 + \epsilon_0) C \int_{-\infty}^u e^{(1-\lambda)s} h(s) ds \\ &= (1 + \epsilon_0) C \left(\frac{e^{(1-\lambda)u}}{1-\lambda} h(u) - \int_{-\infty}^u \frac{e^{(1-\lambda)s}}{1-\lambda} h'(s) ds \right) \\ &\leq (1 + \epsilon_0) C \left(\frac{e^{(1-\lambda)u}}{1-\lambda} h(u) + \int_{-\infty}^u \frac{e^{(1-\lambda)s}}{1-\lambda} \epsilon_h h(s) ds \right) \\ &= (1 + \epsilon_0) C \frac{e^{(1-\lambda)u}}{1-\lambda} h(u) + C \frac{\epsilon_h}{1-\lambda} (1 + \epsilon_0) \int_{-\infty}^u e^{(1-\lambda)s} h(s) ds \\ &\leq (1 + \epsilon_0) C \frac{e^{(1-\lambda)u}}{1-\lambda} h(u) + C \frac{1 + \epsilon_0}{1 - \epsilon_0} \frac{\epsilon_h}{1-\lambda} V(u), \end{aligned}$$

where going from the first line to the second we used integration by parts and from the second to the third we applied (A.30).

This bound holds, in particular, for any $\epsilon_0 \in (0, \frac{1-\lambda-C\epsilon_h}{1-\lambda+C\epsilon_h})$, recalling that we took $C\epsilon_h < 1 - \lambda$.

For any such ϵ_0 , we have $C \frac{1+\epsilon_0}{1-\epsilon_0} \frac{\epsilon_h}{1-\lambda} < 1$, and so

$$V(u) \leq \frac{(1+\epsilon_0)C e^{(1-\lambda)u} h(u)/(1-\lambda)}{1 - C \frac{1+\epsilon_0}{1-\epsilon_0} \frac{\epsilon_h}{1-\lambda}} = \frac{(1+\epsilon_0)C e^{(1-\lambda)u} h(u)}{1-\lambda - C\epsilon_h(1+\epsilon_0)/(1-\epsilon_0)}. \quad (\text{A.31})$$

By the optimality of $\tilde{\alpha}_N$ and $\tilde{\beta}_N$, and noting that $\beta_N = \tilde{\beta}_N$,

$$\tilde{C}_N(\tilde{\alpha}_N, 0) - \tilde{C}_N(\tilde{\alpha}_N, \tilde{\beta}_N) = \tilde{C}_N(\tilde{\alpha}_N, 0) - \tilde{C}_N(\tilde{\alpha}_N, \beta_N) \geq 0. \quad (\text{A.32})$$

From Lemma A.2.1, we have $\tilde{\alpha}_N \rightarrow -\infty$, so we may take N large enough that $\tilde{\alpha}_N \leq u_0$, and then (A.32), (A.31), and (A.29) yield

$$\begin{aligned} 0 \leq \tilde{C}_N(\tilde{\alpha}_N, 0) - \tilde{C}_N(\tilde{\alpha}_N, \beta_N) &\leq NV(\tilde{\alpha}_N) - Ne^{\tilde{\alpha}_N} w(\tilde{\alpha}_N) \|\beta_N\| \gamma_1 \\ &\leq NC \frac{(1+\epsilon_0)e^{(1-\lambda)\tilde{\alpha}_N} h(\tilde{\alpha}_N)}{1-\lambda - C\epsilon_h(1+\epsilon_0)/(1-\epsilon_0)} - NC(1-\epsilon_0)e^{(1-\lambda)\tilde{\alpha}_N} h(\tilde{\alpha}_N) \|\beta_N\| \gamma_1 \\ &= NCe^{(1-\lambda)\tilde{\alpha}_N} h(\tilde{\alpha}_N) \left(\frac{1+\epsilon_0}{1-\lambda - C\epsilon_h(1+\epsilon_0)/(1-\epsilon_0)} - (1-\epsilon_0) \|\beta_N\| \gamma_1 \right). \end{aligned}$$

Therefore, we have

$$\|\beta_N\| \leq \frac{1+\epsilon_0}{(1-\lambda)(1-\epsilon_0) - C(1+\epsilon_0)\epsilon_h} \frac{1}{\gamma_1}$$

for all sufficiently large N , a.s.

Notice that ϵ_h and ϵ_0 may be taken arbitrarily close to 0, and by Lemma A.1.1 ϵ_1 and δ_1 may be taken arbitrarily close to ϵ and δ , respectively, so recalling that $\gamma_1 = \epsilon_1 \delta_1$, we conclude that

$$\limsup_N \|\beta_N\| \leq \frac{1}{(1-\lambda)\epsilon\delta} = \frac{1}{\gamma}, \quad \text{a.s.}$$

□

We can say more about the rate at which α_N diverges:

Corollary A.2.1. *Under the conditions of Proposition A.2.1, there is a constant κ for which*

$$\limsup_{N \rightarrow \infty} (\alpha_N + \log N) < \kappa, \quad \text{a.s.} \quad (\text{A.33})$$

Proof of Corollary A.2.1. Let $K = \max_{i=1,2,\dots,n} \|x_i\|$ and let $r > 1/\gamma$. For N sufficiently large,

$$\max_{i=1,\dots,n} |\beta_N^\top x_i| \leq \max_{i=1,\dots,n} \|\beta_N^\top\| \|x_i\| \leq rK, \quad \text{a.s.} \quad (\text{A.34})$$

Because Lemma A.2.2 holds for any weight function (bounded or unbounded) satisfying Definition 1.4.1, (A.17) implies that for N sufficiently large,

$$N e^{\alpha_N + \beta_N^\top x_j(N)} \delta_1^o \leq n, \quad \text{a.s.}$$

Taking logs on both sides and rearranging yields

$$\alpha_N + \log N \leq \log(n/\delta_1^o) - \beta_N^\top x_j(N), \quad \text{a.s.}$$

Therefore, for any $\kappa > \log(n/\delta_1^o) + rK$ we can apply (A.34) to get (A.33). □

A.2.4 Proof of Theorem 1.4.2

We will need the following lemma in the proof of Theorem 1.4.2.

Lemma A.2.4. *Suppose the conditions in Corollary A.2.1 hold. Then for any $u_0 \in \mathbb{R}$*

$$\max_{i=1,\dots,N_n} \{\alpha_N + \beta_N^\top X_i\} \leq u_0, \quad (\text{A.35})$$

for all sufficiently large N , a.s.

Proof of Lemma A.2.4. Condition 5 implies that for any $\epsilon' > 0$,

$$\sum_{k=1}^{\infty} \mathbf{P}(e^{r\|X\|} > \epsilon'k) < \infty;$$

so for any $v \in \mathbb{R}$,

$$\sum_{k=1}^{\infty} \mathbf{P}(r\|X\| > v + \log k) < \infty.$$

By Theorem 3.5.1 of Embrechts, Kluppelberg, and Mikosch [25], this implies

$$\mathbf{P}(\max_{i=1,\dots,k} r\|X_i\| > v + \log k \text{ i.o.}) = 0,$$

where “i.o.” means infinitely often. It follows that

$$\mathbf{P}(\max_{i=1,\dots,N_n} r\|X_i\| > v + \log N_n \text{ i.o.}) = 0.$$

Now $|\beta_N^\top X_i| \leq \|\beta_N\| \cdot \|X_i\| \leq r\|X_i\|$, for all sufficiently large N , a.s., so

$$\mathbf{P}(\max_{i=1,\dots,N} |\beta_N^\top X_i| > v + \log N \text{ i.o.}) = 0.$$

This means that, a.s., for all sufficiently large N ,

$$\max_{i=1,\dots,N_n} \beta_N^\top X_i \leq v + \log N_n.$$

By Corollary A.2.1, $\log N < \kappa - \alpha_N$, for all sufficiently large N , a.s., so

$$\max_{i=1,\dots,N} \{\alpha_N + \beta_N^\top X_i\} \leq v + \kappa.$$

Choosing $v = u_0 - \kappa$ proves (A.35). □

Proof of Theorem 1.4.2. Taking the partial derivative of \bar{C}_N with respect to α and its gradient with

respect to β , we get

$$\frac{\partial \bar{C}_N}{\partial \alpha} = \sum_{i=1}^n -w(\alpha + \beta^\top x_i) + \sum_{i=1}^N e^{\alpha + \beta^\top X_i} w(\alpha + \beta^\top X_i)$$

$$\frac{\partial \bar{C}_N}{\partial \beta} = \sum_{i=1}^n -w(\alpha + \beta^\top x_i) x_i + \sum_{i=1}^N e^{\alpha + \beta^\top X_i} w(\alpha + \beta^\top X_i) X_i.$$

At the minimizer (α_N, β_N) , these derivatives equal zero, so, a.s., for all sufficiently large N ,

$$\sum_{i=1}^n w(\alpha_N + \beta_N^\top x_i) = \sum_{i=1}^N e^{\alpha_N + \beta_N^\top X_i} w(\alpha_N + \beta_N^\top X_i) \quad (\text{A.36})$$

and

$$\sum_{i=1}^n w(\alpha_N + \beta_N^\top x_i) x_i = \sum_{i=1}^N e^{\alpha_N + \beta_N^\top X_i} w(\alpha_N + \beta_N^\top X_i) X_i. \quad (\text{A.37})$$

Because $\alpha_N \rightarrow -\infty$ and β_N is bounded a.s., for any $x \in \mathbb{R}^d$,

$$\frac{w(\alpha_N + \beta_N^\top x)}{w(\alpha_N)} - e^{-\lambda \beta_N^\top x} \rightarrow 0 \quad \text{a.s.} \quad (\text{A.38})$$

Therefore

$$\frac{\sum_{i=1}^n w(\alpha_N + \beta_N^\top x_i) x_i}{\sum_{i=1}^n w(\alpha_N + \beta_N^\top x_i)} - \frac{\sum_{i=1}^n x_i e^{-\lambda \beta_N^\top x_i}}{\sum_{i=1}^n e^{-\lambda \beta_N^\top x_i}} = \frac{\sum_{i=1}^n w(\alpha_N + \beta_N^\top x_i) x_i / w(\alpha_N)}{\sum_{i=1}^n w(\alpha_N + \beta_N^\top x_i) / w(\alpha_N)} - \frac{\sum_{i=1}^n x_i e^{-\lambda \beta_N^\top x_i}}{\sum_{i=1}^n e^{-\lambda \beta_N^\top x_i}} \rightarrow 0 \quad \text{a.s.}$$

Taking the ratios of the two sides in (A.36)–(A.37), we get

$$\frac{\sum_{i=1}^N e^{\beta_N^\top X_i} w(\alpha_N + \beta_N^\top X_i) X_i}{\sum_{i=1}^N e^{\beta_N^\top X_i} w(\alpha_N + \beta_N^\top X_i)} - \frac{\sum_{i=1}^n x_i e^{-\lambda \beta_N^\top x_i}}{\sum_{i=1}^n e^{-\lambda \beta_N^\top x_i}} \rightarrow 0 \quad \text{a.s.} \quad (\text{A.39})$$

Lemma A.2.5. *Suppose the conditions of Theorem 1.4.2 hold. Suppose there is a possibly random $\beta \in \mathbb{R}^d$ independent of $\{X_i\}_{i \in \mathbb{N}}$ and a possibly stochastic subsequence $N_n \rightarrow \infty$ through which*

$\beta_{N_n} \rightarrow \beta$, a.s. Then

$$\frac{1}{N_n} \sum_{i=1}^{N_n} e^{\beta_{N_n}^\top X_i} \frac{w(\alpha_{N_n} + \beta_{N_n}^\top X_i)}{w(\alpha_{N_n})} \rightarrow \int e^{(1-\lambda)\beta^\top x} dF_0(x), \quad a.s.,$$

and

$$\frac{1}{N_n} \sum_{i=1}^{N_n} e^{\beta_{N_n}^\top X_i} \frac{w(\alpha_{N_n} + \beta_{N_n}^\top X_i)}{w(\alpha_{N_n})} X_i \rightarrow \int e^{(1-\lambda)\beta^\top x} x dF_0(x), \quad a.s.$$

Proof. Condition 5 ensures that the limiting integrals are well-defined and finite because Lemma A.2.3 implies that $\|\beta^\top x\| \leq \|\beta\| \|x\| \leq \|x\|/\gamma < r\|x\|$. We detail the argument for the second limit; the first limit works the same way. Write

$$\begin{aligned} & \left\| \frac{1}{N_n} \sum_{i=1}^{N_n} e^{\beta_{N_n}^\top X_i} \frac{w(\alpha_{N_n} + \beta_{N_n}^\top X_i)}{w(\alpha_{N_n})} X_i - \int e^{(1-\lambda)\beta^\top x} x dF_0(x) \right\| \\ & \leq \left\| \frac{1}{N_n} \sum_{i=1}^{N_n} e^{\beta_{N_n}^\top X_i} \frac{w(\alpha_{N_n} + \beta_{N_n}^\top X_i)}{w(\alpha_{N_n})} X_i - \frac{1}{N_n} \sum_{i=1}^{N_n} e^{(1-\lambda)\beta_{N_n}^\top X_i} X_i \right\| \end{aligned} \quad (\text{A.40})$$

$$+ \left\| \frac{1}{N_n} \sum_{i=1}^{N_n} e^{(1-\lambda)\beta_{N_n}^\top X_i} X_i - \frac{1}{N_n} \sum_{i=1}^{N_n} e^{(1-\lambda)\beta^\top X_i} X_i \right\| \quad (\text{A.41})$$

$$+ \left\| \frac{1}{N_n} \sum_{i=1}^{N_n} e^{(1-\lambda)\beta^\top X_i} X_i - \int e^{(1-\lambda)\beta^\top x} x dF_0(x) \right\|. \quad (\text{A.42})$$

We examine these terms in reverse order. The term in (A.42) vanishes, a.s., by the strong law of large numbers.

Turning next to (A.41), we use Taylor's theorem to write

$$e^{(1-\lambda)\beta_{N_n}^\top X_i} = e^{(1-\lambda)\beta^\top X_i} + e^{(1-\lambda)\tilde{\beta}_{N_n,i}^\top X_i} (1-\lambda)(\beta_{N_n} - \beta)^\top X_i,$$

for some $\tilde{\beta}_{N_n,i}$ on the line segment connecting β_{N_n} and β . So,

$$\left\| \frac{1}{N_n} \sum_{i=1}^{N_n} e^{(1-\lambda)\beta_{N_n}^\top X_i} X_i - \frac{1}{N_n} \sum_{i=1}^{N_n} e^{(1-\lambda)\beta^\top X_i} X_i \right\| \quad (\text{A.43})$$

$$\begin{aligned} &= \left\| \frac{1}{N_n} \sum_{i=1}^{N_n} e^{(1-\lambda)\tilde{\beta}_{N_n,i}^\top X_i} (1-\lambda)(\beta_{N_n} - \beta)^\top X_i \cdot X_i \right\| \\ &\leq (1-\lambda) \|\beta_{N_n} - \beta\| \frac{1}{N_n} \sum_{i=1}^{N_n} e^{(1-\lambda)\|\tilde{\beta}_{N_n,i}\| \|X_i\|} \|X_i\|^2. \end{aligned} \quad (\text{A.44})$$

With r as in Condition 5 and any $r > r' > 1/\gamma$, we know from Lemma A.2.3 that, a.s., $\|\beta_{N_n}\| \leq r'$ for all sufficiently large N_n , and $\|\beta\| \leq 1/\gamma$, so $\|\tilde{\beta}_{N_n,i}\| \leq r', i = 1, \dots, N_n$, a.s., for all sufficiently large N_n because each $\tilde{\beta}_{N_n,i}$ is a convex combination of β_{N_n} and β . Thus,

$$\limsup_{N_n \rightarrow \infty} \frac{1}{N_n} \sum_{i=1}^{N_n} e^{(1-\lambda)\|\tilde{\beta}_{N_n,i}\| \|X_i\|} \|X_i\|^2 \leq \lim_{N_n \rightarrow \infty} \frac{1}{N_n} \sum_{i=1}^{N_n} e^{(1-\lambda)r' \|X_i\|} \|X_i\|^2 < \infty,$$

in light of Condition 5. Thus, (A.44) vanishes as $N_n \rightarrow \infty$, a.s., and then (A.41) vanishes, a.s., as $N_n \rightarrow \infty$.

To bound (A.40), we need to bound

$$\left| \frac{w(\alpha_{N_n} + \beta_{N_n}^\top X_i)}{w(\alpha_{N_n})} - e^{-\lambda \beta_{N_n}^\top X_i} \right|.$$

By Definition 1.4.1, for any $\epsilon \in (0, 1)$ we may choose $u_0 < 0$ such that for any $u \leq u_0$ and $u + s \leq u_0$,

$$(1 - \epsilon)C e^{-\lambda u} h(u) \leq w(u) \leq (1 + \epsilon)C e^{-\lambda u} h(u)$$

and

$$(1 - \epsilon)C e^{-\lambda(u+s)} h(u+s) \leq w(u+s) \leq (1 + \epsilon)C e^{-\lambda(u+s)} h(u+s).$$

We therefore have

$$\frac{1 - \epsilon}{1 + \epsilon} e^{-\lambda s} \frac{h(u+s)}{h(u)} \leq \frac{w(u+s)}{w(u)} \leq \frac{1 + \epsilon}{1 - \epsilon} e^{-\lambda s} \frac{h(u+s)}{h(u)}. \quad (\text{A.45})$$

By condition (1.9) in Definition 1.4.1, we can find some $C_1 > 0$ and some $u_1 \leq u_0 < 0$ such that for any $u \leq u_1$ and $u + s \leq u_1$,

$$\left| \frac{h(u+s)}{h(u)} - 1 \right| \leq \frac{2\epsilon}{1+\epsilon} \max\{C_1, e^{\xi|s|}\}.$$

Let $g(s) = \max\{C_1, e^{\xi|s|}\}$. Expanding the absolute value yields

$$1 - \frac{2\epsilon}{1+\epsilon}g(s) \leq \frac{h(u+s)}{h(u)} \leq 1 + \frac{2\epsilon}{1+\epsilon}g(s).$$

Because $g(s) > 0$, we have

$$1 - \frac{2\epsilon}{1-\epsilon}g(s) \leq 1 - \frac{2\epsilon}{1+\epsilon}g(s) \leq \frac{h(u+s)}{h(u)} \leq 1 + \frac{2\epsilon}{1+\epsilon}g(s).$$

Substituting back into (A.45), we get

$$\frac{1-\epsilon}{1+\epsilon}e^{-\lambda s} \left(1 - \frac{2\epsilon}{1-\epsilon}g(s) \right) \leq \frac{w(u+s)}{w(u)} \leq \frac{1+\epsilon}{1-\epsilon}e^{-\lambda s} \left(1 + \frac{2\epsilon}{1+\epsilon}g(s) \right).$$

Subtracting $e^{-\lambda s}$ from all three parts of the above inequality, we have

$$e^{-\lambda s} \left(\frac{1-\epsilon}{1+\epsilon} - 1 \right) - e^{-\lambda s} \frac{2\epsilon}{1+\epsilon}g(s) \leq \frac{w(u+s)}{w(u)} - e^{-\lambda s} \leq e^{-\lambda s} \left(\frac{1+\epsilon}{1-\epsilon} - 1 \right) + e^{-\lambda s} \frac{2\epsilon}{1-\epsilon}g(s),$$

and so

$$-\frac{2\epsilon}{1+\epsilon}e^{-\lambda s}(1+g(s)) \leq \frac{w(u+s)}{w(u)} - e^{-\lambda s} \leq \frac{2\epsilon}{1-\epsilon}e^{-\lambda s}(1+g(s)).$$

Because

$$-\frac{2\epsilon}{1-\epsilon}e^{-\lambda s}(1+g(s)) \leq -\frac{2\epsilon}{1+\epsilon}e^{-\lambda s}(1+g(s)),$$

letting $\epsilon_0 = 2\epsilon/(1-\epsilon)$ we get

$$\left| \frac{w(u+s)}{w(u)} - e^{-\lambda s} \right| \leq \epsilon_0 e^{-\lambda s} (1+g(s)).$$

This implies, for all sufficiently large N_n ,

$$\left| \frac{w(\alpha_{N_n} + \beta_{N_n}^\top X_i)}{w(\alpha_{N_n})} - e^{-\lambda \beta_{N_n}^\top X_i} \right| \leq \epsilon_0 e^{-\lambda \beta_{N_n}^\top X_i} (1 + g(|\beta_{N_n}^\top X_i|)).$$

Now

$$\begin{aligned} & \left\| \frac{1}{N_n} \sum_{i=1}^{N_n} e^{\beta_{N_n}^\top X_i} \frac{w(\alpha_{N_n} + \beta_{N_n}^\top X_i)}{w(\alpha_{N_n})} X_i - \frac{1}{N_n} \sum_{i=1}^{N_n} e^{(1-\lambda)\beta_{N_n}^\top X_i} X_i \right\| \\ &= \left\| \frac{1}{N_n} \sum_{i=1}^{N_n} e^{\beta_{N_n}^\top X_i} \left(\frac{w(\alpha_{N_n} + \beta_{N_n}^\top X_i)}{w(\alpha_{N_n})} - e^{-\lambda \beta_{N_n}^\top X_i} \right) X_i \right\| \\ &\leq \frac{1}{N_n} \sum_{i=1}^{N_n} e^{\beta_{N_n}^\top X_i} \left| \frac{w(\alpha_{N_n} + \beta_{N_n}^\top X_i)}{w(\alpha_{N_n})} - e^{-\lambda \beta_{N_n}^\top X_i} \right| \|X_i\| \\ &\leq \epsilon_0 \frac{1}{N_n} \sum_{i=1}^{N_n} e^{(1-\lambda)\beta_{N_n}^\top X_i} (1 + g(|\beta_{N_n}^\top X_i|)) \|X_i\| \\ &\leq \epsilon_0 (1 + C_1) \frac{1}{N_n} \sum_{i=1}^{N_n} e^{(1-\lambda)\beta_{N_n}^\top X_i} \|X_i\| + \epsilon_0 \frac{1}{N_n} \sum_{i=1}^{N_n} e^{(1-\lambda)\beta_{N_n}^\top X_i} e^{\xi |\beta_{N_n}^\top X_i|} \|X_i\|, \end{aligned}$$

where the last inequality uses $g(u) = \max\{C_1, e^{\xi|u|}\} \leq C_1 + e^{\xi|u|}$. For any r' such that $\max\{1, 1 - \lambda + \xi\}/\gamma < r' < r$,

$$\limsup_{N_n \rightarrow \infty} \frac{1}{N_n} \sum_{i=1}^{N_n} e^{(1-\lambda)\beta_{N_n}^\top X_i} \|X_i\| \leq \limsup_{N_n \rightarrow \infty} \frac{1}{N_n} \sum_{i=1}^{N_n} e^{(1-\lambda)r' \|X_i\|} \|X_i\| < \infty$$

and

$$\limsup_{N_n \rightarrow \infty} \frac{1}{N_n} \sum_{i=1}^{N_n} e^{(1-\lambda)\beta_{N_n}^\top X_i} e^{\xi |\beta_{N_n}^\top X_i|} \|X_i\| \leq \frac{1}{N_n} \sum_{i=1}^{N_n} e^{(1-\lambda+\xi)\beta_{N_n}^\top X_i} \|X_i\| \leq \frac{1}{N_n} \sum_{i=1}^{N_n} e^{r' \|X_i\|} \|X_i\| < \infty.$$

As $\epsilon_0 > 0$ can be arbitrarily small, we have shown that (A.40) vanishes, a.s., as do (A.41) and (A.42). We have thus proved the second limit in the lemma. The first limit works the same way. \square

We can now complete the proof of Theorem 1.4.2. Our bound on $\|\beta_N\|$ ensures that β_N has at least one limit point; for any limit point $\tilde{\beta}$ of β_N , there is a subsequence β_{N_n} such that $\beta_{N_n} \rightarrow \tilde{\beta}$

a.s. By combining the two limits in Lemma A.2.5 in (A.39) we have

$$\frac{\int e^{(1-\lambda)\tilde{\beta}^\top x} dF_0(x)}{\int e^{(1-\lambda)\tilde{\beta}^\top x} dF_0(x)} = \frac{\sum_{i=1}^n x_i e^{-\lambda\tilde{\beta}^\top x_i}}{\sum_{i=1}^n e^{-\lambda\tilde{\beta}^\top x_i}}, \quad \text{a.s.}$$

We claim that there can be at most one $\beta_* \in \mathbb{R}^d$ satisfying this equation, and so $\tilde{\beta} = \beta_*$ a.s. To see this, define the cumulant generating function $\psi(\beta) = \log \mathbb{E}[e^{\beta^\top W}]$ of $W = (1-\lambda)X_0 - \lambda X_1$, where $X_0 \sim F_0$, X_1 is uniform over of x_1, \dots, x_n , and X_0, X_1 are independent. Equation (1.11) reads $\nabla\psi(\beta_*) = 0$. The surrounding condition on F_0 ensures that the support of F_0 has full dimension. By Theorem 1.13(iv) of Brown [14], this implies that ψ is strictly convex. Strict convexity implies that for $\beta \neq \beta_*$,

$$\nabla\psi(\beta) \cdot (\beta_* - \beta) < \psi(\beta_*) - \psi(\beta) < \nabla\psi(\beta_*) \cdot (\beta_* - \beta),$$

and thus $\nabla\psi(\beta) \neq \nabla\psi(\beta_*)$.

We have thus shown that any β_{N_n} has an almost sure constant limit β_* , and we conclude that $\beta_N \rightarrow \beta_*$ a.s. □

A.2.5 Boundedness of $V(u)/e^u w(u)$

Definition 1.4.1 requires an upper bound on $V(u)/e^u w(u)$ for unbounded weight functions, and this condition is used in (A.23). The following lemma shows that this condition is satisfied by a broad family of weight functions, including $w(u) = Ce^{-\lambda u}$, $\lambda \in (0, 1)$, and $w(u) = Cu^{-k}$, $k > 0$, for large u .

Lemma A.2.6. *Suppose there is an increasing log-convex function g for which $Cg(u) \leq e^u w(u) \leq g(u)$, for all $u \geq u_0$, for some $u_0 \in \mathbb{R}$ with $g'(u_0) \neq 0$, and some $C > 0$. Then $V(u)/e^u w(u)$ is bounded above on $[u_0, \infty)$.*

Proof. Theorem 2.1 of Gill, Pearce, and Pečarić [36] provides an upper bound on the integral of a

log-convex function g , which yields, for any $u > u_0$,

$$V(u) = V(u_0) + \int_{u_0}^u e^s w(s) ds \leq V(u_0) + \int_{u_0}^u g(s) ds \leq V(u_0) + \frac{(u - u_0)(g(u) - g(u_0))}{\log g(u) - \log g(u_0)}.$$

Since $g(u_0) \geq e^{u_0} w(u_0) \geq 0$, the bound remains valid if we remove $g(u_0)$ from the numerator of the last term. Convexity of $\log g$ implies that $\log g(u) - \log g(u_0) \geq (\log g(u_0))'(u - u_0)$, and $(\log g(u_0))' = g'(u_0)/g(u_0)$, so

$$\frac{V(u)}{e^u w(u)} \leq \frac{V(u_0)}{e^{u_0} w(u_0)} + \frac{g(u)}{e^u w(u)} \frac{(u - u_0)}{\log g(u) - \log g(u_0)} \leq \frac{V(u_0)}{e^{u_0} w(u_0)} + \frac{1}{C} \frac{g(u_0)}{g'(u_0)}.$$

□

A.3 Proofs for Section 5

A.3.1 Proof of Proposition 1.5.1

Proof of Proposition 1.5.1. For any value μ of the common mean required by the constraint in (1.23), we know from Lemma 1.5.1 that $D(G_i \| F_i)$, $i = 0, 1$, is minimized by taking G_i to be an exponentially tilted distribution F_{i, β_i} , with $\nabla \psi_i(\beta_i) = \mu$, $i = 0, 1$. We then get

$$D(G_i \| F_i) = \int \beta_i^\top x - \psi_i(\beta_i) dG_i = \beta_i^\top \mu - \psi_i(\beta_i), \quad i = 0, 1,$$

and an objective function value of

$$\lambda[\beta_0^\top \mu - \psi_0(\beta_0)] + (1 - \lambda)[\beta_1^\top \mu - \psi_1(\beta_1)]. \quad (\text{A.46})$$

We can solve (1.23) by minimizing (A.46) over μ , keeping in mind that β_0 and β_1 depend on μ .

Each function $\mu \mapsto \beta_i^\top \mu - \psi_i(\beta_i)$ in (A.46) is the convex conjugate of ψ_i , $i = 0, 1$, at μ , defined by

$$\sup_b \{b^\top \mu - \psi_i(b)\},$$

and is therefore convex in μ . It follows that any point at which the derivative of (A.46) with respect to μ is zero minimizes (A.46).

Differentiating (A.46) with respect to μ , writing $\dot{\beta}_i$ for the derivative matrix of β_i with respect to μ , and setting the derivative equal to zero to get

$$\lambda[\dot{\beta}_0 \cdot \mu + \beta_0 - \dot{\beta}_0 \cdot \nabla\psi_0(\beta_0)] + (1 - \lambda)[\dot{\beta}_1 \cdot \mu + \beta_1 - \dot{\beta}_1 \cdot \nabla\psi_1(\beta_1)] = 0.$$

But $\nabla\psi_i(\beta_i) = \mu$, $i = 0, 1$, so this equation simplifies to

$$\lambda\beta_0 + (1 - \lambda)\beta_1 = 0.$$

The solution is then of the form $\beta_0 = (1 - \lambda)\beta$ and $\beta_1 = -\lambda\beta$, where β solves

$$\nabla\psi_0((1 - \lambda)\beta) = \nabla\psi_1(-\lambda\beta),$$

which is (1.22). □

A.4 Connection with Nonlinear Classifiers

In this section, we provide further insight into the connection between (1.23) and the original classification problem by generalizing (1.6) to the problem of minimizing

$$\bar{C}_\lambda(R) = \mathbb{E} [Y(1 - \lambda)e^{-\lambda R(X)} + (1 - Y)\lambda e^{(1-\lambda)R(X)}] \quad (\text{A.47})$$

over possibly nonlinear discriminant functions $R : \mathbb{R}^d \rightarrow \mathbb{R}$. Suppose for simplicity that F_0 and F_1 have densities f_0 and f_1 . Then arguing as in Lemma 1 of Friedman, Hastie, and Tibshirani [34], (A.47) is minimized at

$$R(x) = \log \frac{\pi_1}{\pi_0} + \log \frac{f_1(x)}{f_2(x)}.$$

Making this substitution in (A.47) and simplifying yields

$$\bar{C}_\lambda(R) = \int (\pi_1 f_1(x))^{1-\lambda} (\pi_0 f_0(x))^\lambda dx. \quad (\text{A.48})$$

The case $\lambda = 1/2$ appears in equation (28) of Eguchi and Copas [24]. We can write (A.48) as

$$\bar{C}_\lambda(R) = \pi_1^{1-\lambda} \pi_0^\lambda e^{(\lambda-1)D_\lambda(F_0\|F_1)}, \quad (\text{A.49})$$

using the Rényi divergence,

$$D_\lambda(H\|F) = \frac{1}{\lambda-1} \log \int dH^\lambda dF^{(1-\lambda)}.$$

The Rényi divergence has a representation in terms of the Kullback-Leibler divergence as

$$(1-\lambda)D_\lambda(H\|F) = \inf_G \{\lambda D(G\|H) + (1-\lambda)D(G\|F)\},$$

(Erven and Harremoës [26], Theorem 30) which allows us to write (A.49) as

$$\bar{C}_\lambda(R) = \pi_1^{1-\lambda} \pi_0^\lambda e^{-\inf_G \{\lambda D(G\|F_0) + (1-\lambda)D(G\|F_1)\}}. \quad (\text{A.50})$$

In other words, when we drop the requirement that the discriminant function be linear, the minimal loss (A.47) is determined by

$$\inf_G \{\lambda D(G\|F_0) + (1-\lambda)D(G\|F_1)\}.$$

The loss in (A.50) is small when there is no G that is “close” to both F_0 and F_1 in the sense of this weighted divergence. When we require the discriminant function $R(\cdot)$ in (A.47) to be linear, the optimal β is determined by (1.23), which relaxes the constraint that $G_0 = G_1$ to the requirement that these two distributions have the same mean. But this discussion shows that the effect of

λ discussed in Section 1.5.4 extends at least in part to nonlinear discriminant functions defined through (A.47).

A.5 Supplementary Tables

A.5.1 Freddie Mac Dataset AUC

Tables A.1–A.4 report AUC values for training, validation, and test sets for four choices of classifiers. Results are indexed by training year, so test results for 2003, for example, are based on defaults predicted for the first quarter of 2006. In all cases, the test and validation AUCs are close to the training AUCs and above 0.8.

Year	Train	Val	Test
2003	0.8899	0.8748	0.8867
2004	0.8514	0.8368	0.8484
2005	0.8456	0.8465	0.8457
2006	0.8314	0.8220	0.8294
2007	0.8258	0.8234	0.8254
2008	0.8487	0.8512	0.8492
2009	0.8763	0.8769	0.8764
2010	0.8421	0.8663	0.8470
2011	0.8730	0.8539	0.8683
2012	0.6573	0.6379	0.6536
2013	0.8589	0.8383	0.8546

Table A.1: AUC, Logistic Regression

Year	Train	Val	Test
2003	0.8911	0.8766	0.8880
2004	0.8530	0.8338	0.8490
2005	0.8479	0.8493	0.8482
2006	0.8322	0.8206	0.8298
2007	0.8277	0.8259	0.8273
2008	0.8503	0.8526	0.8508
2009	0.8783	0.8796	0.8785
2010	0.8440	0.8636	0.8480
2011	0.8748	0.8580	0.8707
2012	0.8759	0.8583	0.8726
2013	0.8600	0.8372	0.8553

Table A.2: AUC, $\lambda = 0.1$

Year	Train	Val	Test
2003	0.8914	0.8771	0.8883
2004	0.8555	0.8379	0.8519
2005	0.8531	0.8549	0.8534
2006	0.8339	0.8235	0.8318
2007	0.8281	0.8264	0.8278
2008	0.8514	0.8534	0.8518
2009	0.8794	0.8793	0.8794
2010	0.8444	0.8654	0.8487
2011	0.8754	0.8626	0.8723
2012	0.8760	0.8617	0.8733
2013	0.8600	0.8367	0.8551

Table A.3: AUC, $\lambda = 0.5$

Year	Train	Val	Test
2003	0.8875	0.8740	0.8847
2004	0.8537	0.8376	0.8504
2005	0.8526	0.8558	0.8532
2006	0.8316	0.8215	0.8295
2007	0.8247	0.8231	0.8244
2008	0.8501	0.8518	0.8505
2009	0.8774	0.8763	0.8771
2010	0.8403	0.8633	0.8450
2011	0.8642	0.8622	0.8637
2012	0.8697	0.8579	0.8674
2013	0.8542	0.8263	0.8484

Table A.4: AUC, $\lambda = 0.9$

A.5.2 Freddie Mac Testing TPR

Table A.5 shows true positive rates in test data for four classifiers. In each case, the classification threshold was set in the training data to achieve a TPR of 99%. The results show that the same thresholds achieve very similar TPRs in the test data. Results are indexed by training year, so test results for 2003, for example, are based on defaults predicted for the first quarter of 2006.

Year	Logistic	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 0.9$
2003	98.76	98.83	98.91	98.76
2004	98.86	98.72	98.65	98.72
2005	99.17	99.17	99.17	99.17
2006	99.00	98.95	98.93	98.91
2007	98.98	99.00	99.02	99.04
2008	99.02	99.01	99.03	99.05
2009	99.09	99.04	99.00	99.00
2010	99.23	99.23	99.10	98.97
2011	98.90	99.12	99.34	99.34
2012	98.70	98.70	98.70	98.70
2013	99.01	99.01	99.21	98.81

Table A.5: TPR (in percent) in test data using classification thresholds that achieve TPR=99% in training data.

A.6 Delta Function Weight

The loss function defined by setting, for some $u_0 \in \mathbb{R}$,

$$U(s) = -\mathbf{1}\{s \leq u_0\}, \quad V(s) = e^{u_0} \mathbf{1}\{s > u_0\},$$

can be interpreted as taking w to be a delta function with unit mass at u_0 . This case leads to the objective in (1.13) and its counterpart

$$\bar{C}_N(\alpha, \beta) = \sum_{i=1}^n \mathbf{1}\{\alpha + \beta^\top x_i \leq u_0\} + e^{u_0} \sum_{i=1}^N \mathbf{1}\{\alpha + \beta^\top X_i > u_0\}. \quad (\text{A.51})$$

As discussed in Example 2 in Section 2.4 of Eguchi and Copas [24], through appropriate choice of u_0 , $C(\alpha, \beta)$ can be interpreted as balancing misclassification costs. However, $\bar{C}_N(\alpha, \beta)$ is not convex, and we will show that the resulting linear discriminant function degenerates under imbalance, in the sense that $\beta_N = 0$ a.s. for all sufficiently large N .

For any α, β , we have $\bar{C}_N(\alpha, \beta) \geq 0$, and for any $\alpha \leq u_0$, we have $\bar{C}_N(\alpha, 0) = n$. In particular, then, if (α_N, β_N) minimizes \bar{C}_N ,

$$0 \leq \bar{C}_N(\alpha_N, \beta_N) \leq \bar{C}_N(\alpha, 0) = n, \quad (\text{A.52})$$

for all $\alpha \leq u_0$.

Lemma A.6.1. *Suppose F_0 surrounds some x_j , $1 \leq j \leq n$ with parameter (ϵ, δ) . Then $\alpha_N + \beta_N^\top x_j \leq u_0$ a.s. for all sufficiently large N .*

Proof. Let $\delta_1 \in (0, \delta)$. At any (α, β) for which $\alpha + \beta^\top x_j > u_0$, we have

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{\alpha + \beta^\top X_i > u_0\} &= \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{\beta^\top (X_i - x_j) > u_0 - (\alpha + \beta^\top x_j)\} \\ &\geq \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{\beta^\top (X_i - x_j) \geq 0\} \\ &\geq \delta_1 \end{aligned}$$

a.s., for all sufficiently large N , where going from the second last step to the last step we use Lemma A.1.1.

For $N > n/e^{u_0} \delta_1$, (A.51) then implies $\bar{C}_N(\alpha, \beta) > n$. But we know $\bar{C}_N(\alpha_N, \beta_N) \leq n$ from (A.52), so we must have $\alpha_N + \beta_N^\top x_j \leq u_0$ a.s., for all sufficiently large N . \square

With this lemma, we have the following result.

Proposition A.6.1. *Suppose Condition 4 holds. Then for all sufficiently large N , every pair (α, β) of the form $\alpha \leq u_0, \beta = 0$ is a global minimizer of \bar{C}_N .*

Proof. If (α_N, β_N) minimizes \bar{C}_N , then we know from the previous lemma that $\alpha_N + \beta_N^\top x_i \leq u_0$ a.s., for all $i = 1, \dots, n$, for all sufficiently large N . It follows from (A.51) that $\bar{C}_N(\alpha_N, \beta_N) \geq n$. At any $\alpha \leq u_0$, (A.51) also shows that $\bar{C}_N(\alpha, 0) = n$. Thus, every $(\alpha, 0)$, $\alpha \leq u_0$, is a global minimizer for sufficiently large N . \square

We conclude from this result that the objective (A.51) degenerates under sufficient imbalance, in the sense that it returns $\beta_N = 0$ a.s., for all sufficiently large N . The linear discriminant function $\alpha_N + \beta_N^\top x$ assigns the same value α_N to every observation, and α_N could be any value less than or equal to u_0 .

Appendix B: Trading TP_2 and RR_2 Violations

B.1 Data Sources

Data sources: from *OptionMetrics*

- `option_price`: This file includes option prices (best ask and best bid) at market close (or more precisely, 3.59pm), quantity traded, greeks, and settlement type (AM or PM).
- `forward`: This file includes forward prices for option contracts at market close. We will use this information to calculate K/F

From CRSP

- `sp`: This is the daily S&P 500 index level. This file is useful to determine PM-settled option payoffs at expiry using the Close price
- `riskfree.csv`: CRSP Risk-Free Rates File. We mainly consider the 1-month rate. Details see https://wrds-www.wharton.upenn.edu/documents/407/CRSP_US_Treasury_Database_Guide_G3ggXhY.pdf and https://wrds-www.wharton.upenn.edu/documents/409/CRSP_Monthly_US_Treasury_Guide.pdf.

From CBOE

- `SET`: This file contains daily SET index price (<https://www.cboe.com/us/indices/dashboard/set/>). SET price determines the price for which AM-settled options are settled.
- `vix`: This is the daily CBOE Volatility Index (VIX Index) file, downloaded from https://www.cboe.com/tradable_products/vix/vix_historical_data/

B.2 AM vs PM-Settled Options

Options on the S&P 500 index can be either AM-settled or PM-settled. In the main analysis, we do not differentiate between the two but treat same-day expiring AM-settled options as expiring earlier than PM-settled options, allowing them to potentially form a violation pair. In this section, we break down violations between AM-settled and PM-settled options.

Figures B.1 and B.2 are analogous to Figure 2.5a, where we plot the violation rates in the T -space, but here we separate violation pairs into AM-settled and PM-settled options. We observe that TP_2 and RR_2 violations are more consistent for AM-settled options, whereas the violation rates for PM-settled options differ significantly. Notably, RR_2 violations are highly concentrated on near-the-term options expiring in 1–2 weeks.

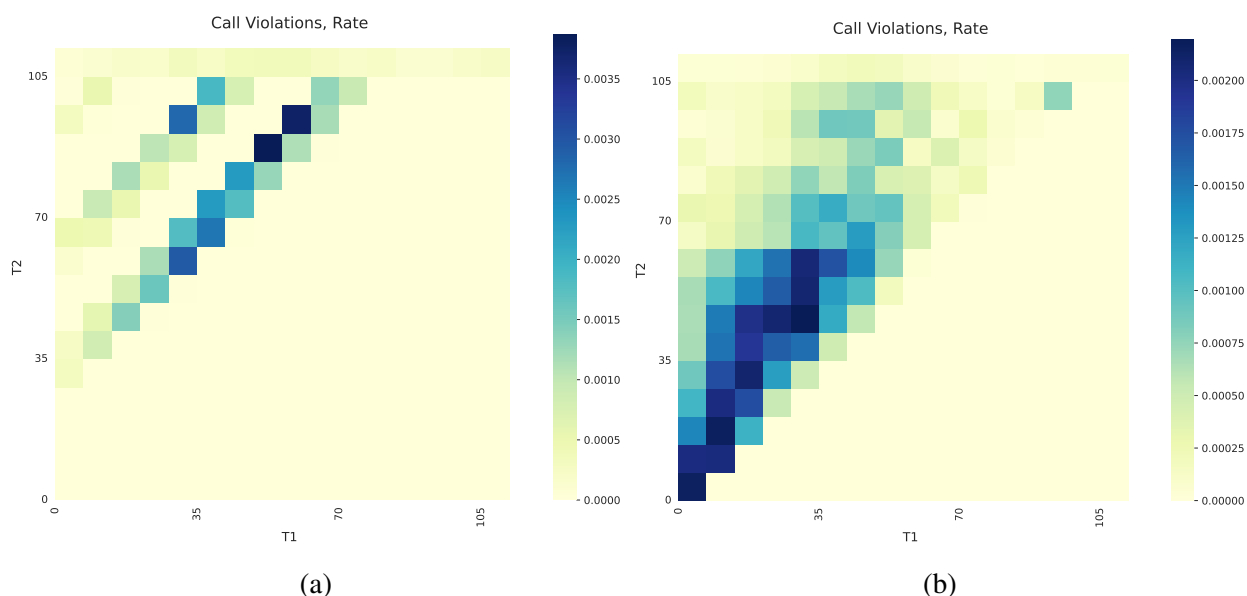


Figure B.1: (a) TP_2 violation rate of AM-settled options. (b) TP_2 violation rate of PM-settled options.

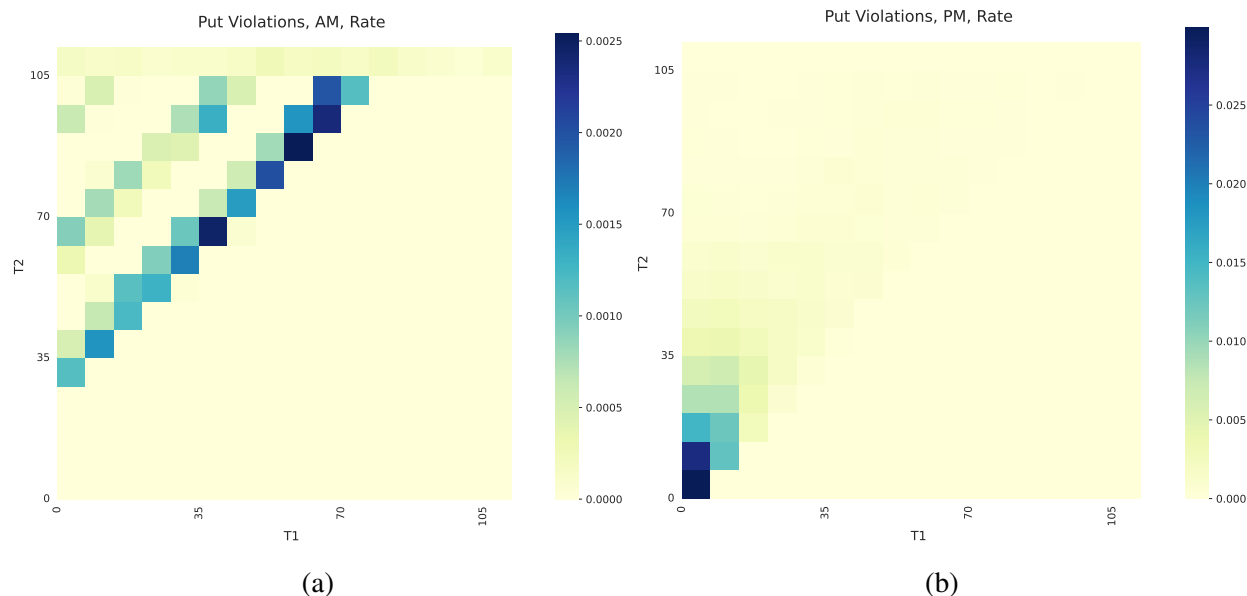


Figure B.2: (a) RR_2 violation rate of AM-settled options. (b) RR_2 violation rate of PM-settled options.

B.3 Robustness Checks

B.3.1 Next-available Trades

Our trading performance relies on the assumption that we hold the options to maturity. A natural question arises: what if we hold the options only until the first available time we can close the position, instead of holding them to maturity? Since many options are deep OTM and not traded every day, the next available times may not be the next trading day. In this section, we focus on T_1 -denominated trades. We assume we close the positions at the first time possible, and we allow the instances when the two T_1 -denominated options are closed at different times.

Figure B.3 plots the cumulative returns for T_1 -denominated TP_2 and RR_2 trades when positions are offloaded at the next available times after the initial trades. The leverage ratio is chosen to match that of the hold-to-maturity strategies. In general, we find a similar overall pattern to the hold-to-maturity (HTM) approach. Our strategies still outperform the S&P index. Notably, the put cumulative return curve is now much smoother than the HTM approach. This is expected, as we are now collecting option payoffs earlier rather than at maturity when large jumps may occur. It would

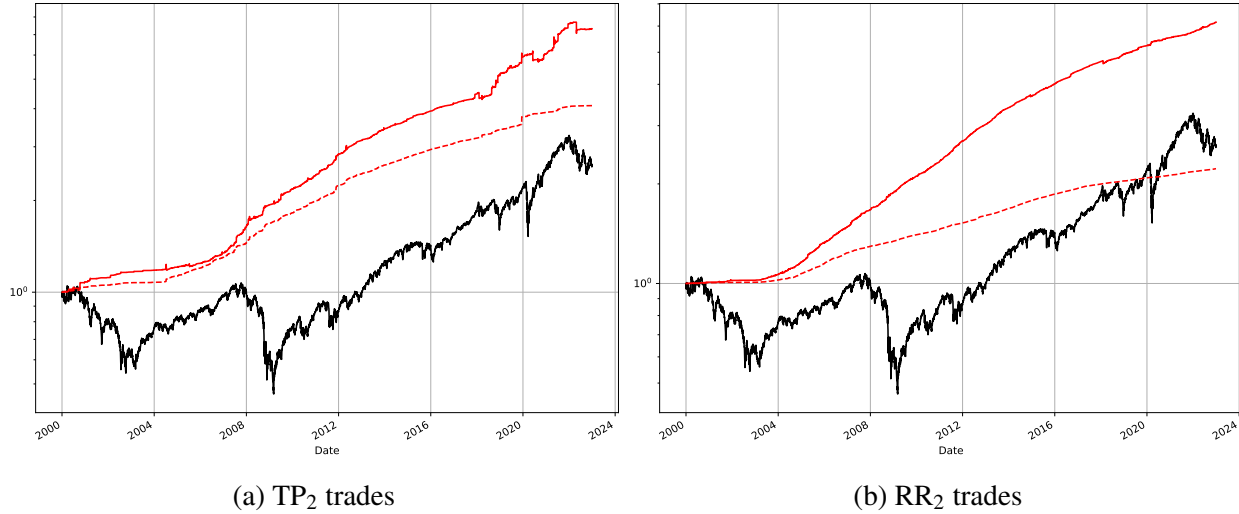


Figure B.3: Cumulative returns for T_1 -denominated, next-available trades.

be interesting to repeat the exercise using intraday data. Offloading the positions at shorter time intervals may show its advantage by more precisely capturing the TP_2 or RR_2 condition violations.

B.3.2 Two-Strike Approximation

The futures ratios on the RHS of (2.3 and (2.4) adjust the strikes K_1 and K_2 and drive our “round up the strikes” approximation. However, as shown in Figures 2.5a and 2.6a most violations occur when T_1 and T_2 are relatively close together.

Figure B.4 plots the historical 30-day S&P futures price relative to the 7-day futures price. Futures prices are influenced by dividends and interest rates; we observe that in the early 2000s, the ratio deviated most significantly from 1. However, in recent years, the ratio has remained close to 1. This suggests that we can potentially approximate \tilde{K}_1 and \tilde{K}_2 in (2.3) and (2.4) with the original K_1 and K_2 . We call this approach “two-strike approximation”, as now only two strike prices are needed in determining violation instead of the original four strikes.

While this approach might lead to a less accurate TP_2 or RR_2 condition, it avoids the conservative “look up the strikes” method in the original violation identification and may uncover more trading opportunities. Indeed, Figures B.5 and B.6 show the decomposition of the total count of TP_2 and RR_2 violations by settlement type. If the two-strike approximation were deployed, the

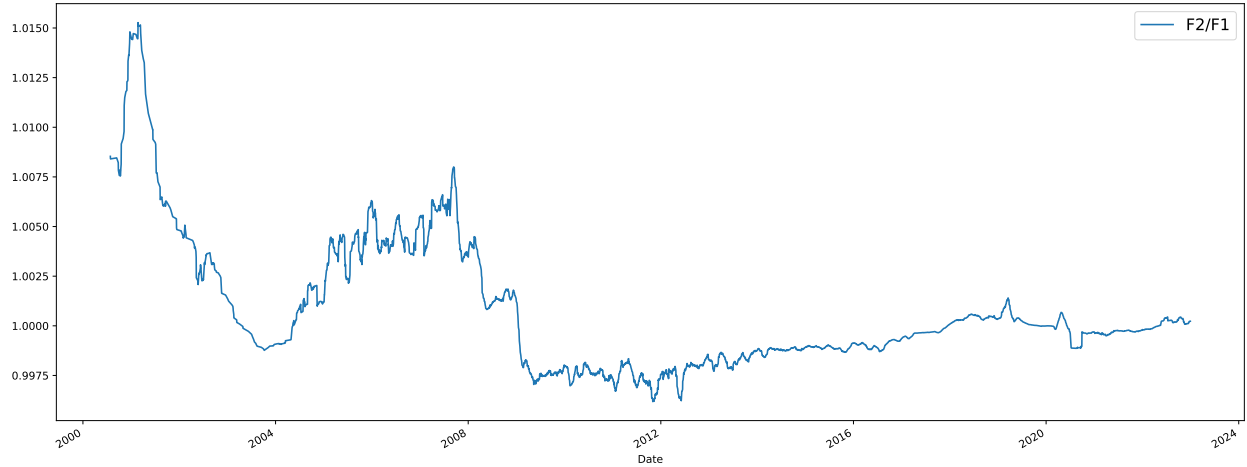


Figure B.4: Slope of S&P 500 futures price (30-day vs 7-day).

total number of violations would increase; this effect is most pronounced post-pandemic when the TP_2 violation count in the original setting decreases but increases significantly in Figure B.5. The violation rates are plotted in Figure B.7. The overall pattern remains consistent, although slightly higher than that in Figure 2.4.

We also repeat our dynamic long-short strategy the two-strike approximation approach. Figures B.8 and B.9 report the cumulative returns for TP_2 and RR_2 trades, respectively. We find similar patterns to the original approach, suggesting the two-strike approximation is useful in uncovering trading opportunities.

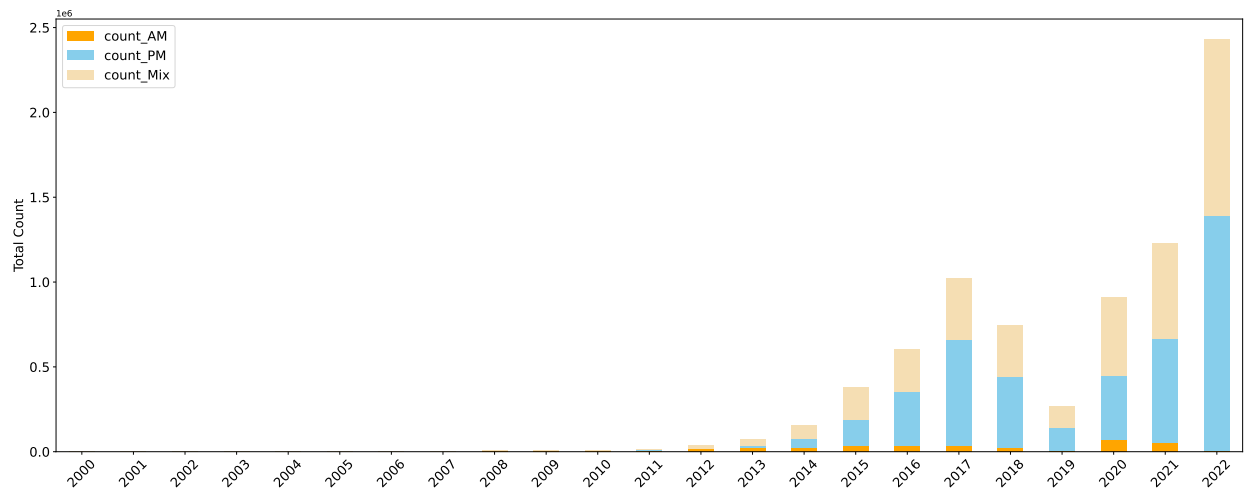


Figure B.5: TP_2 violations count (two-strike approximation).

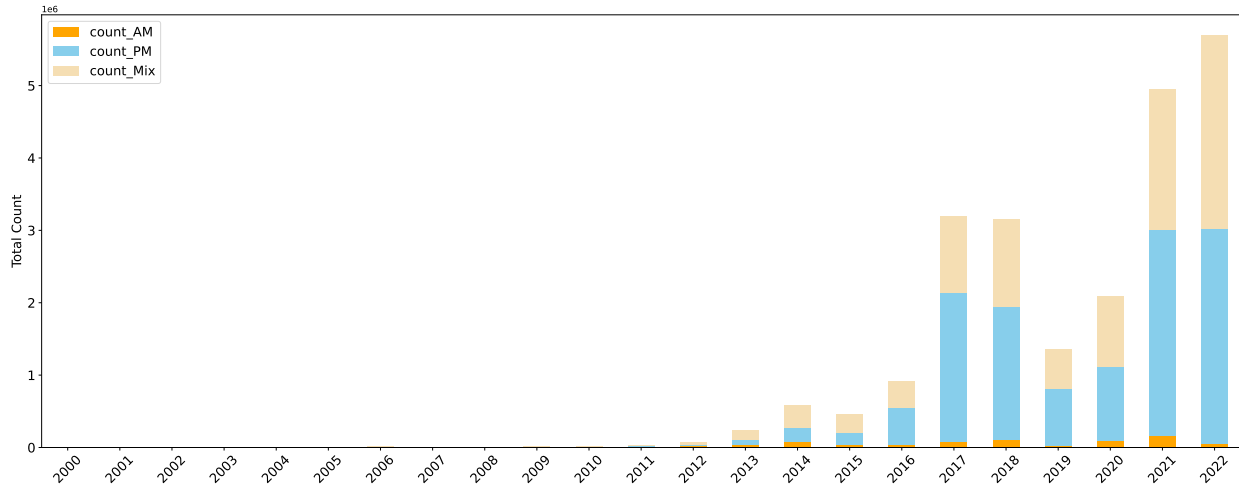


Figure B.6: RR₂ violations count (two-strike approximation).

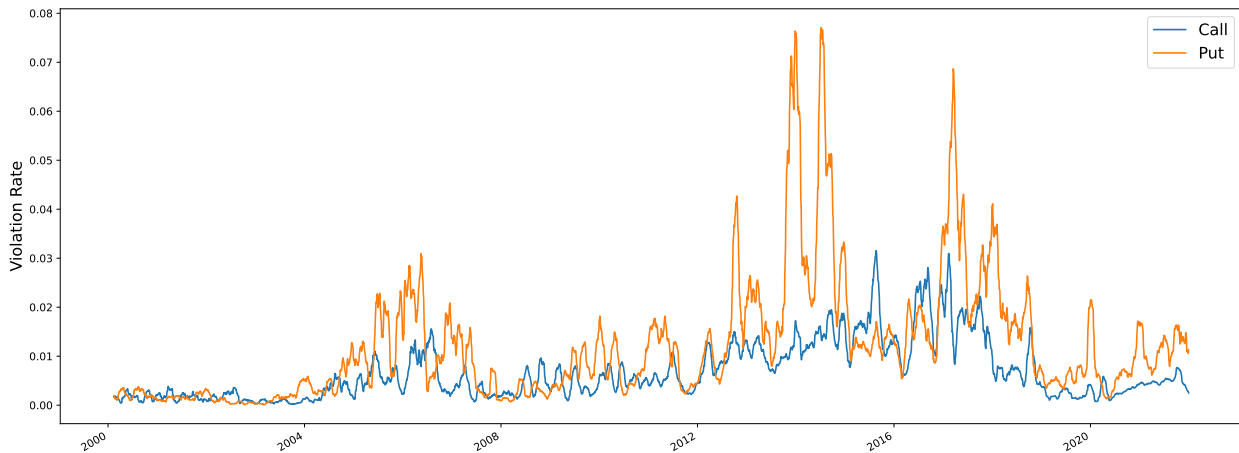


Figure B.7: TP₂ and RR₂ violation rates (two-strike approximation).

B.3.3 Short-only Strategies

We have primarily focused on the long-short strategy, where we trade both sides of the TP₂ or RR₂ condition. As shown in Figures 2.8 and 2.9, the net cash premium received from the long-short trades represent a significant portion of the total profits. These net cash premiums are derived from the cash received from the short positions minus the investment in the long positions. An interesting question arises: what if we only trade the overvalued side by taking short positions and receiving the position premiums?

Figure B.10 plots the cumulative returns of a strategy that only shorts the RHS of (2.3). We

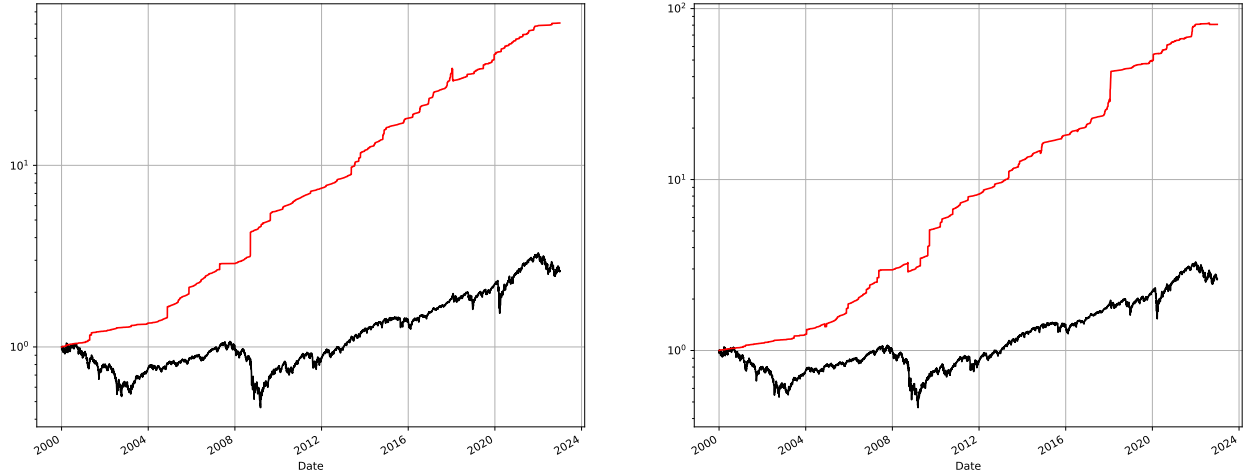


Figure B.8: Cumulative returns. Left: T_1 -denominated TP_2 trades. Right: K_2 -denominated TP_2 trades. (two-strike approximation).

call this the “short-only” strategy and we compare Figure B.10 with the T_1 -denominated long-short returns in Figure 2.8. The short-only strategy modestly outperforms the long-short strategy in total cumulative returns. However, this comes at the expense of experiencing a maximum daily drawdown of 15%, compared to a 1% drawdown in the T_1 -denominated long-short strategy, where the long positions provided protection.

By only shorting the TP_2 -violating options, Figure B.10 shows that this strategy outperforms the index by two orders of magnitude from 2000 to 2022. This raises the question: does TP_2 provide insights on which options to short, or could one achieve similar results by simply shorting deep OTM options? To investigate this, we compare the probability of the options we short in the short-only trade ((\tilde{K}_2, T_1) -Call for TP_2 or (K_1, T_1) -Put for RR_2) expiring in-the-money to that of all options with similar deltas.

More specifically, for each year from 2012 to 2022 we find the $\delta_q, q \in \{0, 10\%, 25\%, 50\%, 75\%, 90\%\}$ quantiles of the option deltas that we short in TP_2 or RR_2 trades. In each delta region, we find the average probability of options we short expiring in-the-money (p_{short}) and compare it to the average probability for all options listed listed on CBOE with deltas in the same region ($p_{average}$). Tables B.1 and B.2 report $p_{all} - p_{short}$ for calls and puts, respectively. The small magnitudes in the numbers are due to the fact that the options considered are generally deep out-of-the-money,

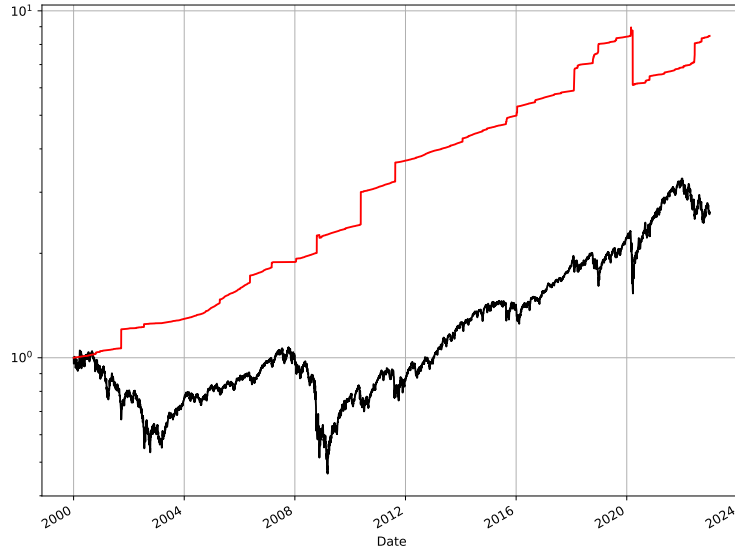


Figure B.9: Cumulative returns of T_1 -denominated RR_2 trades. (two-strike approximation).

with a slim chance of expiring in-the-money (ITM). We observe that in the majority of the years, the difference is positive, suggesting that TP_2 or RR_2 can select options that are less likely to end up ITM than an average comparable option with a similar option delta, and thus shorting a TP_2 or RR_2 -violating option is on average safer than an otherwise comparable option. This provides further evidence that TP_2 and RR_2 violations uncover price anomalies with potential opportunities for profits.

Year	10%	25%	50%	75%	90%
2012	0.00	0.38	0.00	0.39	0.45
2013	0.00	0.00	0.00	0.00	0.82
2014	0.83	0.17	1.22	1.06	2.69
2015	0.00	0.00	0.15	0.25	0.55
2016	-0.00	0.21	0.20	0.30	0.11
2017	0.31	0.83	0.86	1.54	2.23
2018	0.28	0.07	0.26	0.54	0.72
2019	0.00	0.23	0.20	0.32	0.57
2020	0.00	0.08	0.34	0.46	1.45
2021	0.48	0.35	0.49	0.70	1.33
2022	0.08	0.00	0.14	0.41	2.96

Table B.1: Differences between the probability of an average call option expiring ITM and that of a comparable TP_2 -violating option. Numbers are in percent.

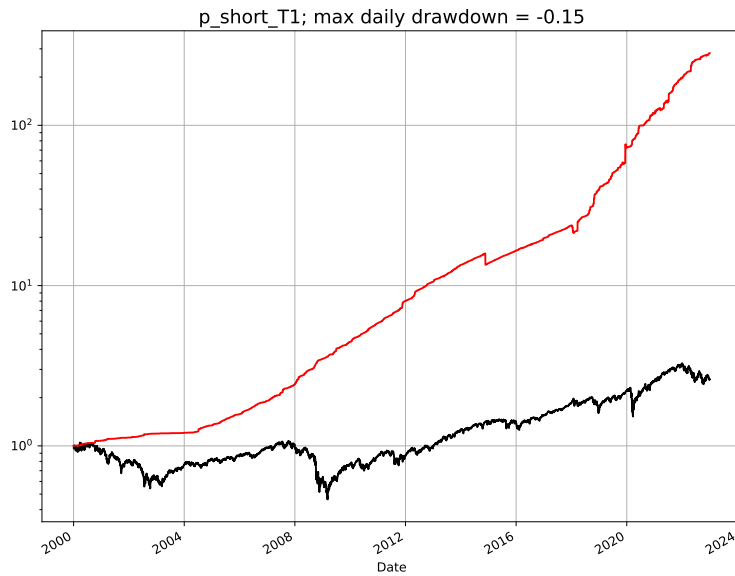


Figure B.10: Cumulative returns for short-only TP₂ trades.

Year	10%	25%	50%	75%	90%
2012	0.00	0.00	0.00	0.00	0.00
2013	0.00	0.00	0.00	0.00	0.00
2014	0.00	0.00	0.00	0.00	0.06
2015	0.00	0.00	0.00	0.00	0.05
2016	0.00	0.00	0.00	0.00	0.00
2017	0.00	0.00	0.00	0.00	0.00
2018	-0.33	-0.16	-0.24	-0.53	0.23
2019	0.09	0.22	0.23	0.12	0.25
2020	1.90	1.73	2.12	3.28	4.37
2021	0.00	0.00	0.00	0.00	0.01
2022	0.21	0.06	0.19	0.30	0.40

Table B.2: Differences between the probability of an average put option expiring ITM and that of a comparable RR₂-violating option. Numbers are in percent.

Appendix C: Fairness in Regulatory Stress Tests

C.1 Proofs

Proposition 3.3.1. Problem (3.7) is solved by the linear projection of Y_S onto the span of 1 and X_S . If $\text{var}[X_S]$ is invertible, then the coefficients of the linear projection are given by (3.12) and

$$\beta_{Pool} = \text{var}[X_S]^{-1} \text{cov}[Y_S, X_S];$$

see, for example, Wooldridge [74], p.25. In (3.9)–(3.10) we can write

$$\text{var}[X_S] = \sum_s p_s W_s = \sum_s p_s \Sigma_s + \text{var}[\mu_S].$$

This matrix is positive definite because we assumed that each Σ_s is positive definite, so $\text{var}[X_S] = \text{E}[W_S]$ is indeed invertible. To evaluate $\text{cov}[Y_S, X_S]$ for Y_S in (3.6), we first note that

$$\text{cov}[X_S, \epsilon_S] = \text{E}[\text{cov}[X_S, \epsilon_S|S]] + \text{cov}[\text{E}[X_S|S], \text{E}[\epsilon_S|S]] = \text{E}[0] + \text{cov}[\mu_S, 0] = 0.$$

It follows that

$$\begin{aligned} \text{cov}[Y_S, X_S] &= \text{E}[\text{cov}[\alpha_S, X_S|S]] + \text{cov}[\text{E}[\alpha_S|S], \text{E}[X_S|S]] + \text{E}[\text{cov}[\beta_S^\top X_S, X_S|S]] + \text{cov}[\text{E}[\beta_S^\top X_S|S], \text{E}[X_S|S]] \\ &= 0 + \text{cov}[\alpha_S, \mu_S] + \text{E}[\Sigma_S \beta_S] + \text{E}[\text{var}[\mu_S] \beta_S] \\ &= \text{cov}[\alpha_S, \mu_S] + \text{E}[W_S \beta_S]. \end{aligned}$$

□

Proposition 3.4.1. By Proposition 3.4 of Chzhen et al. [16] or Theorem 6 of Le Gouic et al. [9], the expected squared error is minimized subject to demographic parity by the rule that assigns to bank s with features x the loss forecast

$$\hat{Y}_{PTF}(x, s) = \sum_i p_i F_i^{-1}(F_s(\alpha_s + \beta_s^\top x)), \quad (\text{C.1})$$

where F_s is the cumulative distribution function of $\alpha_s + \beta_s^\top X_s$. By construction, F_s is then also the cumulative distribution function of $\alpha_s^o + \beta_s^{o\top} Z_s$, which is normal with mean α_s^o and variance $\|\beta_s^o\|^2$. Writing Φ for the standard normal distribution function, we get

$$F_s(y) = \Phi\left(\frac{y - \alpha_s^o}{\|\beta_s^o\|}\right), \quad F_i^{-1}(q) = \alpha_i^o + \|\beta_i^o\| \Phi^{-1}(q).$$

Making these substitutions in (C.1) and writing $\alpha_s^o + \beta_s^{o\top} z_s$ for $\alpha_s + \beta_s^\top x$, with $z_s = \Sigma_s^{-1}(x - \mu_s)$, we get

$$\begin{aligned} \hat{Y}_{PTF}(x, s) &= \sum_i p_i F_i^{-1}(F_s(\alpha_s^o + \beta_s^{o\top} z_s)) \\ &= \sum_i p_i F_i^{-1}(\Phi(\beta_s^{o\top} z_s / \|\beta_s^o\|)) \\ &= \sum_i p_i \{\alpha_i^o + \|\beta_i^o\| \Phi^{-1}(\Phi(\beta_s^{o\top} z_s / \|\beta_s^o\|))\} \\ &= \sum_i p_i \{\alpha_i^o + \|\beta_i^o\| \beta_s^{o\top} z_s / \|\beta_s^o\|\}, \end{aligned}$$

which is (3.18). (Demographic parity holds because the distribution of $\beta_s^{o\top} Z_s / \|\beta_s^o\|$ does not depend on s .) Under (3.19), $\|\beta_i^o\| \|\beta_s^o\| / \|\beta_s^o\| = \|a_i \beta\| \|a_s \beta\| / \|a_s \beta\| = a_i \beta = \beta_i^o$, and we get (3.20). \square

Proposition 3.4.2. We can rewrite $\hat{Y}(x, s)$ in (3.22) as

$$\hat{Y}(x, s) = \sum_{i=1}^{\bar{s}} a_i \mathbf{1}\{s = i\} + \beta^\top x,$$

for suitable a_i . Minimizing (3.23) over the a_i and β yields the same value for β as minimizing

(3.23) using (3.22) because the indicators $\mathbf{1}\{s = i\}$ have the same span as the $U_i(s)$ and a constant. Thus, the β_F defined by (3.23) is the coefficient of X_S in the regression of Y_S on X_S and the indicators $\mathbf{1}\{S = i\}$. By the Frisch-Waugh-Lovell Theorem (as in Angrist and Pischke [4], pp.35–36), we can therefore evaluate β_F as the coefficient in the regression of Y_S on the component of X_S orthogonal to the other variables, which in our case are the indicators. The projection of X_S onto the indicators is given by $\sum_i \mu_i \mathbf{1}\{S = i\} = \mu_S$, so the orthogonal component is $X_S - \mu_S$. We may therefore evaluate β_F as the coefficient in the regression of $Y_S - \mathbf{E}[Y_S]$ on $X_S - \mu_S$, which is (3.28). For the first factor in (3.28), we have

$$\text{var}[X_S - \mu_S] = \mathbf{E}[\text{var}[X_S - \mu_S|S]] + \text{var}[\mathbf{E}[X_S - \mu_S|S]] = \mathbf{E}[\Sigma_S] + 0.$$

For the second factor, we similarly have

$$\text{cov}[X_S - \mu_S, Y_S] = \mathbf{E}[\text{cov}[X_S - \mu_S, Y_S|S]] = \mathbf{E}[\text{cov}[X_S - \mu_S, \beta_S^\top X_S|S]] = \mathbf{E}[\Sigma_S \beta_S],$$

so (3.25) follows. The optimal α_F in (3.23) ensures that $\mathbf{E}[\hat{Y}(X_S, S)] = \mathbf{E}[Y_S]$, which yields (3.26). □

Proposition 3.4.3. The minimization in (3.23) yields coefficients α_F , δ , and β_F , with which we can write

$$Y_S = \alpha_F + \sum_{i=1}^{\bar{S}-1} \delta_i U_i(S) + \beta_F^\top X_S + u, \quad (\text{C.2})$$

where the error u has mean zero and is uncorrelated with $U(S)$ and X_S . We thus have

$$\begin{aligned} \beta_{Pool} &= \text{var}[X_S]^{-1} \text{cov}[X_S, Y_S] \\ &= \text{var}[X_S]^{-1} \{ \text{cov}[X_S, \beta_F^\top X_S] + \text{cov}[X_S, \delta^\top U(S)] \} \\ &= \text{var}[X_S]^{-1} \{ \text{var}[X_S] \beta_F + \text{cov}[X_S, U(S)] \delta \} \\ &= \beta_F + \Lambda \delta, \end{aligned}$$

using the expression for Λ in (3.29) for the last step.

Next, we evaluate δ . Using (C.2), we can derive δ as the vector of coefficients in a regression of $Y_S - \beta_F^\top X_S$ on $U(S)$. Thus,

$$\begin{aligned}\delta &= \text{var}[U(S)]^{-1} \text{cov}[U(S), Y_S - \beta_F^\top X_S] \\ &= \text{var}[U(S)]^{-1} \text{cov}[U(S), Y_S] - \text{var}[U(S)]^{-1} \text{cov}[U(S), X_S] \beta_F.\end{aligned}\quad (\text{C.3})$$

To evaluate $\text{var}[U(S)]^{-1}$, we first note that

$$\text{var}[U(S)] = \begin{pmatrix} p_1 - p_1^2 & -p_1 p_2 & \cdots & -p_1 p_{\bar{S}-1} \\ -p_2 p_1 & p_2 - p_2^2 & \cdots & -p_2 p_{\bar{S}-1} \\ \vdots & \vdots & \ddots & \vdots \\ -p_{\bar{S}-1} p_1 & -p_{\bar{S}-1} p_2 & \cdots & p_{\bar{S}-1} - p_{\bar{S}-1}^2 \end{pmatrix};$$

direct multiplication then verifies that

$$\text{var}[U(S)]^{-1} = \begin{pmatrix} 1/p_1 + 1/p_{\bar{S}} & 1/p_{\bar{S}} & \cdots & 1/p_{\bar{S}} \\ 1/p_{\bar{S}} & 1/p_2 + 1/p_{\bar{S}} & \cdots & 1/p_{\bar{S}} \\ \vdots & \vdots & \ddots & \vdots \\ 1/p_{\bar{S}} & 1/p_{\bar{S}} & \cdots & 1/p_{\bar{S}-1} + 1/p_{\bar{S}} \end{pmatrix}.$$

The vector $\text{cov}[U(S), Y_S]$ has elements

$$[\text{cov}[U(S), Y_S]]_s = p_s(\mathbf{E}[Y_s] - \mathbf{E}[Y_S]), \quad s = 1, \dots, \bar{S} - 1;$$

and row s of the matrix $\text{cov}[U(S), X_S]$ is given by $p_s(\mu_s - \bar{\mu})^\top$. Thus, for $s = 1, \dots, \bar{S} - 1$, we have the vector elements

$$(\text{var}[U(S)]^{-1} \text{cov}[U(S), Y_S])_s = (\mathbf{E}[Y_s] - \mathbf{E}[Y_S]) + \sum_{i=1}^{\bar{S}-1} p_i(\mathbf{E}[Y_i] - \mathbf{E}[Y_S])/p_{\bar{S}} = \mathbf{E}[Y_s] - \mathbf{E}[Y_{\bar{S}}],$$

and similarly row s of the matrix $\text{var}[U(S)]^{-1}\text{cov}[U(S), X_S]$ is given by

$$(\mu_s - \bar{\mu})^\top + \sum_{i=1}^{\bar{s}-1} p_i (\mu_i - \bar{\mu})^\top / p_{\bar{s}} = (\mu_s - \mu_{\bar{s}})^\top. \quad (\text{C.4})$$

Combining these terms in (C.3) yields (3.31). \square

Proposition 3.4.4. We derived an expression for the rows of M in (C.4), and (3.39) follows from that expression. By applying expression (3.28) for β_F in (3.40), we see that \hat{Y}_{SEO} is the claimed projection. \square

Corollary 3.4.2. From (3.25) we know that if $\Sigma_s \equiv \Sigma$ then $\beta_F = \mathbf{E}[\beta_S]$. From (3.20), we get $\bar{\beta}^o = \sum_i p_i \beta_i^o = \sum_i p_i \Sigma^{1/2} \beta_i = \Sigma^{1/2} \mathbf{E}[\beta_S] = \Sigma^{1/2} \beta_F$. Thus, $\beta_F^\top (x - \mu_s) = \bar{\beta}^{o\top} \Sigma^{-1/2} (x - \mu_s) = \bar{\beta}^{o\top} z_s$. It follows that (3.20) and (3.40) coincide because they have the same overall mean. If the distribution of Z_s does not depend on s , then (3.20) satisfies demographic parity. \square

Proposition 3.4.5. The claim for (i) simply restates Proposition 3.3.1. The constraint in (ii) is demographic parity, so the optimizer follows from the definition of projection to fairness. The constraint in (v) requires $\text{cov}[\lambda^\top U(S) + \beta^\top X_S, U(S)] = 0$. Rearranging this equation, we get $\lambda = -(\text{var}[U(S)])^{-1} \text{cov}[U(S), X_S] \beta$; i.e., $\lambda = -M\beta$. Making this substitution in the form of \hat{Y}_S in row (v), (3.41) becomes

$$\mathbf{E}[(Y_S - \alpha - \lambda^\top U(S) - \beta^\top X_S)^2] = \mathbf{E}[(Y_S - \alpha - \beta^\top [X_S - M^\top U(S)])^2]. \quad (\text{C.5})$$

Minimizing this expression over α and β yields the coefficients in a linear regression of Y_S on a constant $X_S - M^\top U(S)$. In light of Proposition 3.4.4, the optimal β in (C.5) is then the coefficient on $X_S - \mu_S$ in a regression of Y_S on a constant $X_S - \mu_S$. It follows from (3.28) that the optimal β in (C.5) is therefore β_F . Because $\mathbf{E}[U(S)] = 0$, the minimizing α in (C.5) is the α_F defined by (3.23). We have thus shown that the optimal forecast in row (v) is

$$\hat{Y}_s = \alpha_F + \beta_F^\top (X_s - M^\top U) = \alpha_F + \lambda^\top U(s) + \beta_F^\top X_s.$$

In case (iv), by applying (3.39) we see that the constraint requires $\text{cov}[Y_S - \beta^\top X_S, X_S - M^\top U(S)] = 0$, so $\beta = (\text{cov}[X_S, X_S - M^\top U(S)])^{-1} \text{cov}[Y_S, X_S - M^\top U(S)]$. Using the fact that $X_S - M^\top U(S)$ is orthogonal to $U(S)$, we get

$$\begin{aligned} \text{cov}[X_S, X_S - M^\top U(S)] &= \text{cov}[X_S - M^\top U(S), X_S - M^\top U(S)] + \text{cov}[M^\top U(S), X_S - M^\top U(S)] \\ &= \text{var}[X_S - M^\top U(S)], \end{aligned}$$

and therefore $\beta = (\text{var}[X_S - M^\top U(S)])^{-1} \text{cov}[Y_S, X_S - M^\top U(S)]$. In other words, the optimal β in (iv) is the coefficient in a linear regression of Y_S on $X_S - M^\top U(S)$. As noted in the discussion of (v), this is β_F , and it follows from $\mathbf{E}[U(S)] = 0$ that the optimal α in (iv) is α_F . \square

Proposition 3.4.6. By construction, the least-squares projection of Y_S onto a constant and X_S is given by the pooled forecast, so

$$Y_S = \mathbf{E}[Y_S] + \beta_{\text{pool}}^\top (X_S - \bar{\mu}) + \epsilon_P,$$

for some orthogonal error ϵ_P with a variance σ_P^2 that does not depend on γ . We therefore have

$$\begin{aligned} \mathbf{E}[(\hat{Y}_\gamma(X_S) - Y_S)^2] &= \mathbf{E}[\{\hat{Y}_\gamma(X_S) - \mathbf{E}[Y_S] - \beta_{\text{pool}}^\top (X_S - \bar{\mu})\}^2] + \sigma_P^2 \\ &= \mathbf{E}[\{(\gamma - \Lambda\delta)^\top (X_S - \bar{\mu})\}^2] + \sigma_P^2, \end{aligned}$$

from which (i) follows.

Using the linear projection of Y_S onto $(1, U(S), X_S)$ in (3.32), we can write

$$Y_S = \mathbf{E}[Y_S] + \delta^\top U(S) + \beta_F^\top (X_S - \bar{\mu}) + \epsilon,$$

for some orthogonal error ϵ with a variance σ_ϵ^2 that does not depend on γ . We therefore have

$$\begin{aligned}
\mathbf{E}[(\hat{Y}_\gamma(X_S) - Y_S)^2] &= \mathbf{E}[\{\gamma^\top(X_S - \bar{\mu}) - \delta^\top U(S)\}^2] + \sigma_\epsilon^2 \\
&= \mathbf{E}[\{\gamma^\top(X_S - \mu_S) + \gamma^\top(\mu_S - \bar{\mu}) - \delta^\top U(S)\}^2] + \sigma_\epsilon^2 \\
&= \mathbf{E}[\{\gamma^\top(X_S - \mu_S) + (\gamma^\top M^\top - \delta^\top)U(S)\}^2] + \sigma_\epsilon^2 \\
&= \mathbf{E}[\{\gamma^\top(X_S - \mu_S)\}^2] + \mathbf{E}[\{(\gamma^\top M^\top - \delta^\top)U(S)\}^2] + \sigma_\epsilon^2,
\end{aligned}$$

where the third equality uses (3.39), and the last equality uses the orthogonality of $X_S - \mu_S$ and $U(S)$. If this expression is smaller than the corresponding value with $\gamma = 0$, then (ii) must hold. \square

Proposition 3.4.7. We saw in the proof of Proposition 3.4.2 that $X_S - \mu_S$ is uncorrelated with the centered indicators $U_i(S)$. It is also uncorrelated with $V_S - \nu_S$ because

$$\mathbf{E}[(X_S - \mu_S)(V_S - \nu_S)] = \sum_s p_s \mathbf{E}[(X_s - \mu_s)(V_s - \nu_s)] = 0,$$

under our assumption that X_s and V_s are uncorrelated. Similarly,

$$\mathbf{E}[(X_S - \mu_S)U_i(S)V_S] = p_i \mathbf{E}[(X_i - \mu_i)V_i] - p_i \mathbf{E}[(X_S - \mu_S)V_S] = 0,$$

so $X_S - \mu_S$ is uncorrelated with the interaction terms. Thus, $X_S - \mu_S$ is uncorrelated with all the elements of $\mathcal{O} = \{1, U(S), V_S - \nu_S, U_1(S)V_S, \dots, U_{\bar{s}}(S)V_S\}$.

Starting from the representation of (3.43) as

$$Y_S = \sum_{i=1}^{\bar{s}} \mathbf{1}\{S = i\} \{\alpha_i + \beta_i^\top X_S + \gamma_i^\top V_S + \epsilon_i\},$$

we may write

$$\begin{aligned}
Y_S &= \beta_S^\top(X_S - \mu_S) + \sum_{i=1}^{\bar{s}} \mathbf{1}\{S = i\}(\alpha_i + \beta_i^\top \mu_i) + \sum_{i=1}^{\bar{s}} \mathbf{1}\{S = i\} \gamma_i^\top V_S + \epsilon_S \\
&\equiv \beta_S^\top(X_S - \mu_S) + \tilde{Y} + \epsilon_S,
\end{aligned}$$

which expresses Y_S as the sum of three mutually orthogonal terms. As $X_S - \mu_S$ is uncorrelated with O , and \tilde{Y} is uncorrelated with $X_S - \mu_S$, we may calculate the projection of Y_S onto the span of $X_S - \mu_S$ and O by projecting $\beta_S^\top (X_S - \mu_S)$ onto $X_S - \mu_S$ and projecting \tilde{Y} onto O .

We know from (3.28) that the projection of $\beta_S^\top (X_S - \mu_S)$ onto $X_S - \mu_S$ is $\beta_F^\top (X_S - \mu_S)$; in other words, including V_S and the interaction terms does not change β_F .

For the projection of \tilde{Y} onto O , let $a_i = \alpha_i + \beta_i^\top \mu_i + \bar{\gamma}^\top \nu_i$ and $\bar{a} = \sum_i p_i a_i$. Then,

$$\begin{aligned}
\tilde{Y} &= \sum_{i=1}^{\bar{s}} \mathbf{1}\{S=i\}(\alpha_i + \beta_i^\top \mu_i) + \sum_{i=1}^{\bar{s}} \mathbf{1}\{S=i\} \gamma_i^\top V_S \\
&= \sum_{i=1}^{\bar{s}} \mathbf{1}\{S=i\}(\alpha_i + \beta_i^\top \mu_i) + \sum_{i=1}^{\bar{s}} U_i(S) \gamma_i^\top V_S + \sum_{i=1}^{\bar{s}} p_i \gamma_i^\top V_S \\
&= \sum_{i=1}^{\bar{s}} \mathbf{1}\{S=i\}(\alpha_i + \beta_i^\top \mu_i + \bar{\gamma}^\top \nu_i) + \sum_{i=1}^{\bar{s}} U_i(S) \gamma_i^\top V_S + \sum_{i=1}^{\bar{s}} p_i \gamma_i^\top (V_S - \nu_S) \\
&= \bar{a} + \sum_{i=1}^{\bar{s}-1} U_i(S)(a_i - a_{\bar{s}}) + \sum_{i=1}^{\bar{s}} U_i(S) \gamma_i^\top V_S + \bar{\gamma}^\top (V_S - \nu_S).
\end{aligned}$$

Thus, \tilde{Y} is in the span of O , and its coefficient on $V_S - \nu_S$ is $\bar{\gamma}$. With all $\text{var}[V_S]$ having full rank, $V_S - \nu_S$ is not spanned by the other elements of O , so its coefficient $\bar{\gamma}$ is uniquely determined. \square

Proposition 3.5.1. For the first claim, we have

$$\begin{aligned}
\text{cov}[\hat{Y}_F(X_S) - Y_S, X_S - \mathbf{E}[X_S|S]] &= -\mathbf{E}[(f_1(S) + \epsilon)(X_S - \mathbf{E}[X_S|S])] \\
&= -\mathbf{E}[f_1(S)(X_S - \mathbf{E}[X_S|S])] - \mathbf{E}[\epsilon X_S] + \mathbf{E}[\epsilon \mathbf{E}[X_S|S]] \\
&= 0 + \mathbf{E}[\mathbf{E}[\epsilon|S] \mathbf{E}[X_S|S]] - \mathbf{E}[\mathbf{E}[\epsilon|X_S] X_S] = 0.
\end{aligned}$$

For the second claim, we have

$$\begin{aligned}
\mathbf{E}[(\hat{Y}_\gamma(X_S) - Y_S)^2] &= \mathbf{E}[(\gamma(X_S) - f_1(S) - \epsilon)^2] \\
&= \mathbf{E}[(\gamma(X_S) - f_1(S))^2] + \mathbf{E}[\epsilon^2].
\end{aligned}$$

The last step uses

$$\mathbf{E}[(\gamma(X_S) - f_1(S))\epsilon] = \mathbf{E}[(\gamma(X_S) - f_1(S))\mathbf{E}[\epsilon|S]] = 0.$$

It now follows that if γ reduces the expected squared forecast error then $\mathbf{E}[\gamma(X_S)f_1(S)] > 0$, which implies (3.51) and (3.52). \square

C.2 Cross-Bank Parameter Externalities

As a consequence of aggregating bank-specific results into a single industry model, changes at one bank can affect loss forecasts at other banks, and the results are sometimes counterintuitive. In this section, we argue that these cross-bank externalities are generally more reasonable under FEO forecasts than under the pooled method.

For simplicity, we consider a setting with a single scalar feature x . More generally, we can think of this as a feature that is uncorrelated with all other features. We adopt the convention that this feature is nonnegative, and that higher values of x are associated with higher losses. Thus, for each bank s we assume $\mu_s \geq 0$ and $\beta_s \geq 0$. In reducing μ_s , a bank improves its portfolio quality; in reducing β_s , a bank improves its ability to manage portfolio risk; and in reducing α_s , a bank improves unobserved features to reduce its losses. We examine how these improvements — reductions in μ_s , α_s , and β_s — affect stress test results for bank s and other banks l .

We can write the FEO loss forecast (3.24) for bank l evaluated at $X_l = x$ as

$$\hat{Y}_{F,l}(x) = \hat{Y}_F(x) = \sum_s p_s (\alpha_s + \beta_s \mu_s) + \beta_F (x - \bar{\mu}), \quad (\text{C.6})$$

with $\beta_F = \sum_i p_i \sigma_i^2 \beta_i / \sum_i p_i \sigma_i^2$, as in (3.27). The forecast is the same for all banks l because FEO satisfies equal treatment. It is now easy to see that

$$\frac{\partial \hat{Y}_F(x)}{\partial \mu_s} = p_s \beta_s - p_s \beta_F \geq 0, \quad \text{if and only if } \beta_s \geq \beta_F; \quad (\text{C.7})$$

$$\frac{\partial \hat{Y}_F(x)}{\partial \alpha_s} = p_s \geq 0; \quad (\text{C.8})$$

and

$$\frac{\partial \hat{Y}_F(x)}{\partial \beta_s} = p_s \mu_s + (x - \bar{\mu}) p_s \sigma_s^2 / \sum_i p_i \sigma_i^2 \geq 0, \quad \text{if } x > \bar{\mu}. \quad (\text{C.9})$$

In (C.7) we see that if bank s has above-average (relative to β_F) sensitivity to feature x , then reducing its average exposure to that feature μ_s reduces loss forecasts for all banks. Equation (C.8) shows a similar overall benefit if bank s improves on the other dimensions captured by α_s . In (C.7) we see that an improvement in risk management at bank s , corresponding to a reduction in β_s , reduces loss forecasts at above-average levels of x . If x is part of the stress scenario, then large values of x are particularly relevant.

The directional effects in (C.7)–(C.9) are fairly simple and reasonable, considering that cross-bank effects are inevitable in an industry model. If the industry improves its performance (perhaps because of improvements at one bank) we generally expect loss forecasts to decrease. (A decrease in a forecast corresponds to a positive derivative because we are considering a decrease μ_s , α_s , or β_s .) Counterparts to (C.7)–(C.9) continue to hold if we replace β_F in (C.6) with any convex combination of the β_s , as in the WATE model. However, the pooled method behaves quite differently.

The pooled forecast $\hat{Y}_P(x)$ can be written in the same form as (C.6) but with β_F replaced by β_{Pool} in (3.11). We now get

$$\frac{\partial \hat{Y}_P(x)}{\partial \mu_s} = p_s (\beta_s - \beta_{Pool}) + (x - \bar{\mu}) \frac{\partial \beta_{Pool}}{\partial \mu_s}.$$

The sign of the last term is not determined by a simple condition, so the overall directional effect is difficult to predict. The sign of

$$\frac{\partial \hat{Y}_P(x)}{\partial \alpha_s} = p_s + p_s \frac{(\mu_s - \bar{\mu})(x - \bar{\mu})}{\sum_s p_s \sigma_s^2 + \text{var}(\mu_S)} \beta_s,$$

depends on the magnitudes of μ_s and x , relative to $\bar{\mu}$. For the sensitivity to β_s , we can write

$$\frac{\partial \hat{Y}_P(x)}{\partial \beta_s} = p_s \mu_s + (x - \bar{\mu}) \frac{\partial \beta_{Pool}}{\partial \beta_s}, \quad \frac{\partial \beta_{Pool}}{\partial \beta_s} = \frac{p_s (\sigma_s^2 + \mu_s (\mu_s - \bar{\mu}))}{\sum_i p_i \sigma_i^2 + \text{var}(\mu_S)}.$$

Among the most troubling aspects of the pooled model is that the last term could be negative: a reduction in β_s could produce an increase in β_{Pool} . In particular, $\sigma_s^2 + \mu_s (\mu_s - \bar{\mu})$ is negative for a bank with below-average exposure to feature x (so $\mu_s < \bar{\mu}$) and low variability σ_s^2 in this exposure. Under the pooled model, it is therefore possible for an improvement in risk management at one bank (a reduction in β_s) to produce an *increase* in loss forecasts at all banks.

The top panel of Table C.1 shows sufficient conditions for positive sensitivities of $\hat{Y}_F(x)$ and $\hat{Y}_P(x)$. The middle and bottom panels show corresponding results for the expected forecasts $E[\hat{Y}_l] = E[\hat{Y}(X_l)]$ and for the bias $E[\hat{Y}(X_l) - Y_l]$. Supporting details for the second and third cases are provided in Section C.3. We have tried to provide simple sufficient conditions, and in most cases the conditions are not necessary. All of the conditions for FEO extend to WATE with β_F replaced by the weighted average coefficient.

Some counterintuitive and undesirable cases can arise at empirically plausible parameter values. For example, in equation (C.12) we derive an expression for $\partial E[\hat{Y}_P(X_l)] / \partial \alpha_s$. Using estimated parameters for the credit card data in Section C.5, we find that this derivative is negative when l is Citigroup and s is JPMorgan Chase. In other words, an improvement at JPMorgan Chase would result in a higher expected loss forecast at Citigroup under the pooled model.

The bias sensitivities in Table C.1 are more complicated than the other cases because the bias involves the difference between the predicted and actual loss rates. A reduction in the predicted loss rate can increase or decrease bias, depending on whether the initial forecast is too low or too high.

$\hat{Y}(x)$	FEO	Pool
$\mu_s \downarrow$	\downarrow iff $\beta_s > \beta_F$	no simple rule
$\alpha_s \downarrow$	\downarrow	\downarrow if $(\mu_s - \bar{\mu})(x - \bar{\mu}) > 0$
$\beta_s \downarrow$	\downarrow if $x > \bar{\mu}$	\downarrow if $[\sigma_s^2 + \mu_s(\mu_s - \bar{\mu})](x - \bar{\mu}) > 0$
E[$\hat{Y}(X_l)$]		
$\mu_s \downarrow$	$l = s: \downarrow$ $l \neq s: \downarrow$ iff $\beta_s > \beta_F$	no simple rule
$\alpha_s \downarrow$	\downarrow	$l = s: \downarrow$ $l \neq s: \downarrow$ if $(\mu_s - \bar{\mu})(\mu_l - \bar{\mu}) > 0$
$\beta_s \downarrow$	\downarrow if $\mu_s + \mu_l > \bar{\mu}$	\downarrow if $[\sigma_s^2 + \mu_s(\mu_s - \bar{\mu})](\mu_l - \bar{\mu}) > 0$ or if μ_s sufficiently large
bias(l)		
$\mu_s \downarrow$	$l = s: \downarrow$ iff $\beta_s < \beta_F$ $l \neq s: \downarrow$ iff $\beta_s > \beta_F$	no simple rule
$\alpha_s \downarrow$	$l = s: \uparrow$ $l \neq s: \downarrow$	$l = s: \text{no simple rule}$ $l \neq s: \downarrow$ if $(\mu_s - \bar{\mu})(\mu_l - \bar{\mu}) > 0$
$\beta_s \downarrow$	$l = s: \uparrow$ if $\mu_s < \bar{\mu}$ $l \neq s: \downarrow$ if $\mu_s + \mu_l > \bar{\mu}$	no simple rule \downarrow if $[\sigma_s^2 + \mu_s(\mu_s - \bar{\mu})](\mu_l - \bar{\mu}) > 0$

Table C.1: Sensitivity of results for bank l in response to a decrease in parameter μ_s , α_s , or β_s for bank s . Sensitivities shown are for predicted loss $\hat{Y}_l(x)$ (top), mean predicted loss $E[\hat{Y}(X_l)]$ (middle), and the bias $E[\hat{Y}(X_l) - Y_l]$.

C.3 Sensitivity Analysis

This section provides supporting details for Section C.2, particularly the conclusions summarized in the middle and bottom panels of Table C.1. We begin with an analysis of forecast bias that is of independent interest.

C.3.1 Forecast Bias

If losses at different banks are described by different models, then forecast bias becomes inevitable when we apply a single model to all banks. But the distribution of bias across banks may differ under different choices of the single model.

Let \hat{Y}_s be any of the forecasts for bank s in Table 3.1, and, as in (3.1), let Y_s denote the actual loss rate for bank s . Both \hat{Y}_s and Y_s are evaluated at X_s . Define the forecast bias for bank s to be

$$\text{bias}(s) = \mathbb{E}[\hat{Y}_s - Y_s]. \quad (\text{C.10})$$

The expectation integrates over the distribution of the error ϵ_s in (3.1) and the features X_s .

Proposition C.3.1. *For each forecast in Table 3.1, the bias is as follows.*

- (i) *Pooled:* $\text{bias}(s) = \mathbb{E}[Y_S] - \mathbb{E}[Y_s] + \beta_{Pool}^\top (\mu_s - \bar{\mu})$;
- (ii) *PTF in (3.18):* $\text{bias}(s) = \mathbb{E}[Y_S] - \mathbb{E}[Y_s]$;
- (iii) *Conditional expectation:* $\text{bias}(s) = \mathbb{E}[\hat{Y}_C(X_s)] - \mathbb{E}[Y_s]$;
- (iv) *FEO:* $\text{bias}(s) = \mathbb{E}[Y_S] - \mathbb{E}[Y_s] + \beta_F^\top (\mu_s - \bar{\mu})$;
- (v) *SEO:* $\text{bias}(s) = \mathbb{E}[Y_S] - \mathbb{E}[Y_s]$.

Proof. For (i), we have, using the definition of α_{Pool} in (3.12),

$$\begin{aligned}
\mathbf{E}[\hat{Y}_s - Y_s] &= \mathbf{E}[\alpha_{Pool} + \beta_{Pool}^\top X_s - Y_s] \\
&= (\mathbf{E}[Y_S] - \beta_{Pool}^\top \bar{\mu}) + \beta_{Pool}^\top \mu_s - \mathbf{E}[Y_s] \\
&= \mathbf{E}[Y_S] - \mathbf{E}[Y_s] + \beta_{Pool}^\top (\mu_s - \bar{\mu}).
\end{aligned}$$

For the PTF forecast, (3.17) and (3.18) yield

$$\mathbf{E}[\hat{Y}_s] = \bar{\alpha}^o = \sum_s p_s \alpha_s^o = \sum_s p_s \mathbf{E}[Y_s] = \mathbf{E}[Y_S],$$

and the bias in (ii) follows. The expression in (iii) holds by definition. The argument for (iv) is the same as the argument for (i). The bias in (v) follows from (iv) because we see from (3.38) that the SEO forecast for bank s subtracts $\beta_F^\top (\mu_s - \bar{\mu})$ from the FEO forecast. \square

In every case of Proposition C.3.1, the average bias $\sum_s p_s \text{bias}(s)$ is zero, but the methods differ in how they distribute bias across banks. We saw previously that the PTF and SEO methods go the farthest in equalizing differences; we now see that the bias for each of these methods is the difference $\mathbf{E}[Y_S] - \mathbf{E}[Y_s]$ between the average loss rate for all banks and the average for an individual bank.

Using the relationship $\beta_{Pool} = \beta_F + \Lambda \delta$ from (3.30), we see that the difference between the expressions in (i) and (iv) is

$$\text{bias}_{Pool}(s) - \text{bias}_{FEO}(s) = \delta^\top \Lambda^\top (\mu_s - \bar{\mu}).$$

In light of the discussion in Section 3.4.2, this difference is the expected disparate impact on bank s of using the pooled model.

C.3.2 Improvement in Intercept α_s

By taking the expectation of (C.6), we get

$$\mathbb{E}[\hat{Y}_F(X_l)] = \sum_s p_s (\alpha_s + \beta_s \mu_s) + \beta_F (\mu_l - \bar{\mu}), \quad (\text{C.11})$$

and the same holds for the expected pooled forecast with β_F replaced by β_{Pool} . It follows that, for any banks s and l ,

$$\frac{\partial \mathbb{E}[\hat{Y}_F(X_l)]}{\partial \alpha_s} = p_s > 0.$$

In other words, all expected forecasts decrease following a reduction in α_s .

In contrast, for the pooled model we get

$$\frac{\partial \mathbb{E}[\hat{Y}_P(X_l)]}{\partial \alpha_s} = p_s - \frac{\partial \text{cov}(\alpha_s, \mu_s) / \partial \alpha_s}{\sum_t p_t \sigma_t^2 + \text{var}(\mu_s)} \beta_s (\bar{\mu} - \mu_l) = p_s + p_s \frac{(\mu_s - \bar{\mu})(\mu_l - \bar{\mu})}{\sum_s p_s \sigma_s^2 + \text{var}(\mu_s)} \beta_s. \quad (\text{C.12})$$

Bank s benefits from its reduction of α_s , in the sense that the derivative with $l = s$ is positive. For $l \neq s$, the sign of (C.12) does not admit a simple description. In particular, it may be negative when μ_s and μ_l are on opposite sides of $\bar{\mu}$, meaning that one bank's loans are riskier than average and the other bank's loans are less risky than average.

For the bias under FEO we have

$$\frac{\partial \text{bias}_F(l)}{\partial \alpha_s} = p_s - \mathbf{1}\{l = s\}$$

It is then immediate that

$$\frac{\partial \text{bias}_F(l)}{\partial \alpha_s} > 0 \text{ if } l \neq s \quad \text{and} \quad \frac{\partial \text{bias}_F(s)}{\partial \alpha_s} < 0.$$

The direction of change makes sense. If the bias for a bank is positive, meaning that the industry model overestimates its losses, then improvements at other banks will reduce loss forecasts and thus reduce the bias. The bank's own improvements will increase the bias by reducing the bank's

own losses by more than they reduce the model's forecasts. The situation is reversed for a bank with a negative bias.

However, for the pooled regression method,

$$\frac{\partial \text{bias}_P(l)}{\partial \alpha_s} = p_s + p_s \frac{(\mu_s - \bar{\mu})(\mu_l - \bar{\mu})}{\sum_s p_s \sigma_s^2 + \text{var}(\mu_S)} \beta_s - \mathbf{1}\{l = s\},$$

and the direction of change is unclear.

C.3.3 Improvement in Loan Quality

Now suppose bank s improves the quality of its loan portfolio, resulting in a smaller μ_s . This has no effect on β_F , which makes sense — changing one bank's loan quality should not change the sensitivity of losses to loan quality. However, it is evident from (3.11) that β_{Pool} does change with μ_s .

Under FEO, the mean the mean predicted loss rate satisfies

$$\frac{\partial \text{E}\hat{Y}_F(X_l)}{\partial \mu_s} = p_s \beta_s + \beta_F (\mathbf{1}\{l = s\} - p_s),$$

which is always positive if $l = s$. This means that an improvement in bank l 's loan quality (a reduction in μ_l) reduces bank l 's mean predicted losses. In the pooled model,

$$\frac{\partial \text{E}\hat{Y}_P(X_l)}{\partial \mu_s} = p_s \beta_s + \beta_{Pool} (\mathbf{1}\{l = s\} - p_s) + (\mu_s - \bar{\mu}) \frac{\partial \beta_{Pool}}{\partial \mu_s};$$

this expression could be negative, even with $l = s$, meaning that a bank could be penalized (through a higher mean predicted loss rate) as a result of improving its loan quality.

The sensitivity of the bias under FEO is given by

$$\frac{\partial \text{bias}_F(l)}{\partial \mu_s} = (\mathbf{1}\{l = s\} - p_s) (\beta_F - \beta_s);$$

in particular, the bias for bank l moves in opposite directions with respect to changes in μ_l and μ_s ,

$s \neq l$. Suppose industry model overestimates bank l 's losses, in the sense that the bias is positive, and suppose the industry model overestimates bank l 's sensitivity to loan quality, in the sense that $\beta_F > \beta_l$. Then bank l will benefit (in the sense of reducing the bias) from improving its loan quality by reducing μ_l .

For the pooled regression,

$$\frac{\partial \text{bias}_P(l)}{\partial \mu_s} = (\beta_{Pool} - \beta_s)(\mathbf{1}\{l = s\} - p_s) + (\mu_s - \bar{\mu}) \frac{\partial \beta_{Pool}}{\partial \mu_s}.$$

The sign of this expression does not admit a simple condition.

C.3.4 Improvement in Loan Management

Now suppose bank s improves its abilities in loan management, resulting in a reduction in β_s .

The mean predicted loss rate under FEO satisfies

$$\frac{\partial \mathbf{E}[\hat{Y}_F(X_l)]}{\partial \beta_s} = p_s(\mu_s + \mu_l - \bar{\mu}),$$

and is positive if $\mu_s + \mu_l > \bar{\mu}$. In the pooled model

$$\frac{\partial \mathbf{E}[\hat{Y}_P(X_l)]}{\partial \beta_s} = p_s \mu_s + \frac{p_s(\sigma_s^2 + \mu_s(\mu_s - \bar{\mu}))}{\sum_i p_i(\sigma_i^2 + \mu_i(\mu_i - \bar{\mu}))}(\mu_l - \bar{\mu}),$$

so $[\sigma_s^2 + \mu_s(\mu_s - \bar{\mu})](\mu_l - \bar{\mu}) > 0$ is a sufficient condition for the sensitivity to be positive. Regardless of the value of μ_l , the sensitivity is positive for all sufficiently large μ_s .

For $l \neq s$, the sensitivity of the bias for bank l with respect to β_s equals the sensitivity of the mean predicted loss because the actual expected loss $\mathbf{E}[Y_l]$ is unaffected by β_s . We therefore focus on the case $l = s$. Under FEO,

$$\frac{\partial \text{bias}_F(s)}{\partial \beta_s} = (p_s - 1)\mu_s + p_s(\mu_s - \bar{\mu}),$$

which is guaranteed to be negative if $\mu_s < \bar{\mu}$. Under the pooled model, the sign of

$$\frac{\partial \text{bias}_P(s)}{\partial \beta_s} = (p_s - 1)\mu_s + \frac{p_s(\sigma_s^2 + \mu_s(\mu_s - \bar{\mu}))}{\sum_i p_i(\sigma_i^2 + \mu_i(\mu_i - \bar{\mu}))}(\mu_s - \bar{\mu})$$

does not admit a simple characterization.

C.4 Convex Combinations of Coefficients

Equation (3.13) aggregates the individual scalar slopes β_s into a single value. We can generalize this perspective and ask what properties we would like in an aggregation function, meaning a function $f : \mathbb{R}^{\bar{S}} \rightarrow \mathbb{R}$,

$$\beta_* = f(\beta_1, \dots, \beta_{\bar{S}}),$$

that combines bank-specific coefficients β_s into an “industry” parameter β_* .

We consider the following properties:

- (i) $f(kb_1, \dots, kb_{\bar{S}}) = kf(b_1, \dots, b_{\bar{S}})$, for all $k, b_1, \dots, b_{\bar{S}} \in \mathbb{R}$;
- (ii) $f(b, \dots, b) = b$, for at least one nonzero $b \in \mathbb{R}$;
- (iii) $b_s > 0$, for all s , implies $f(b_1, \dots, b_{\bar{S}}) \geq 0$;
- (iv) f is differentiable at zero.

Property (i) is needed for the aggregation to perform sensibly under a change of units in the measurement of X_s : if we divide each X_s by k , each β_s increases by a factor of k , and it is natural to require that β_* scale accordingly. Properties (ii) and (iii) are also very modest requirements. Property (iv) is harder to motivate but not unreasonable. These properties constrain the aggregation function as follows:

Proposition C.4.1. *If (i)–(iv) hold, then $f(\beta_1, \dots, \beta_{\bar{S}})$ is a convex combination of its arguments.*

Proof. Fix $\beta \in \mathbb{R}^{\bar{S}}$. Let $g(t) = f(t\beta)$. By condition (iv), $g'(0) = \beta^\top f'(0)$. Condition (i) and (ii) imply $g(t) = tf(\beta)$, so $g'(t) = f(\beta)$ for any t . Thus, $f(\beta) = g'(0) = \beta^\top f'(0) = \sum_{i=1}^{\bar{S}} f'_i(0)\beta_i$.

Condition (ii) now implies $\sum_{i=1}^{\bar{S}} f'_i(0) = 1$, and condition (iii) implies $f'_i(0) \geq 0$, for all i . Thus, $f(\beta) = \sum_{i=1}^{\bar{S}} f'_i(0)\beta_i$ is a convex combination of the components of β . \square

The scalar FEO coefficient in (3.27) is a convex combination of the bank-specific coefficients β_s , but the pooled coefficient (3.13) is generally not. This property of the FEO model extends to the multivariate case under additional conditions. If all the bank-specific covariance matrices Σ_s , $s = 1, \dots, \bar{S}$, coincide, then in (3.25) we get $\beta_F = \mathbb{E}[\beta_S] = \sum_s p_s \beta_s$. If all Σ_s are diagonal (but not necessarily identical), then the representation of the scalar FEO coefficient in (3.27) applies to each coordinate of β_F . If all Σ_s have the same eigenvectors, then we can transform the original features X_s into uncorrelated features using principal components. Using these transformed features, each coordinate of β_F is a convex combination of bank-specific coefficients.

C.5 Empirical Evidence

In this section, we document empirical evidence of heterogeneity in bank-specific models of loss rates, and we examine the implications of this heterogeneity for the choice of an industry-wide model. We find strong evidence of statistically significant differences in model parameters across banks. These differences can lead to material differences between pooled and FEO coefficients in an industry model.

We must emphasize, however, that our investigation is constrained by the very limited information made publicly available by banks about the risk characteristics and losses in their loan portfolios. The Federal Reserve has far more granular information about banks' loans and losses. Our results can therefore provide only a rough indication of the impact of bank heterogeneity in the Fed's stress tests.

C.5.1 Data

We use two types of data and data sources: historical macroeconomic data and loan information for individual banks.

Macroeconomic Data

We use data on seven of the macro variables used in the Federal Reserve’s stress tests: real disposable income growth, real GDP growth, house price index level, inflation rate, unemployment rate, Dow Jones total stock index level, and the Treasury spread. The Federal Reserve provides historical data on its website for all variables used in forming stress scenarios, including these. We use the values reported by the Fed for these variables in the June 2020 stress test; these values run from 1990 through 2019.

We aggregate these variables into a single macro variable by taking the first principal component of their correlation matrix. Table C.2 shows the corresponding loadings. We see that an increase in the principal component corresponds to decreases in income growth and GDP growth and an increase in unemployment, suggesting that this composite variable serves as a reasonable measure of overall economic conditions. Figure C.1 plots the level of this variable over time and shows a sharp climb around 2008 and 2020.¹

Macro Factor	PC1 Loading
Real disposable income growth	-0.229
Real GDP growth	-0.525
Change House Price Index	-0.467
CPI inflation rate	-0.079
Change unemployment	0.529
Change Dow	-0.293
Change Treasury Spread	0.287

Table C.2: Loadings of first principal component on macro variables.

¹The loadings in Table C.2 are calculated using data through 2019, and we use these loadings to extend PC1 through the end of 2021. When we include the COVID period in the calculation of the principal components, PC1 becomes harder to interpret. For example, the coefficients for income growth and unemployment have the same sign.

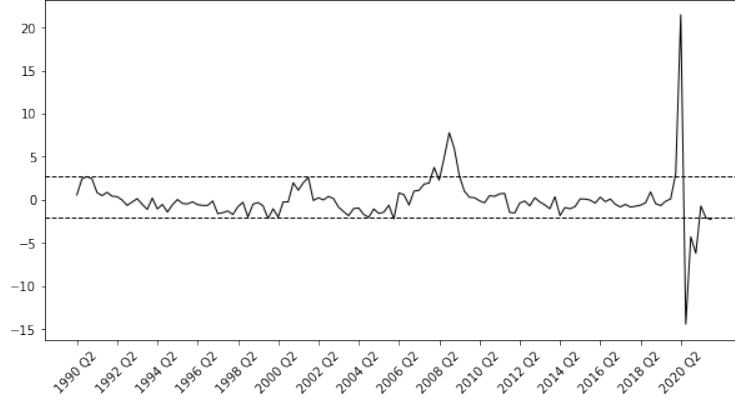


Figure C.1: First principal component (PC1) of macro variables from 1990 Q2 to 2021 Q4. The dashed lines correspond to the 5th and 95th percentiles of PC1.

Loan Information for Individual Banks.

Bank holding companies publicly report financial information quarterly through the Federal Reserve’s form Y-9C. We use these filings to collect information on four loan types that are treated separately in the Fed’s stress tests: credit cards, first lien mortgages, commercial real estate loans, and commercial and industrial loans. For each category, each bank, and each quarter, we collect loan balances, charge-offs, recoveries, and total amounts past due serving as our proxy measures of loan portfolio risk.

We collect this data from 2001 to 2021 for the thirty-five largest banks by total assets (as of December 2021). The banks are listed in Table C.10. The stress test focuses on adverse economic conditions; we weight each observation by the level of stress in each quarter and each bank’s load balance: for each bank s , each quarter t , and loan category p , we weight the observations by

$$w_{s,t}^p = e^{\lambda MacroPC_t} \times Loan_{s,t}^p, \quad (C.13)$$

where $Loan_{s,t}^p$ is the size of the loan portfolio of type p for bank s in quarter t . We choose λ so that for the same loan level, the worst economic quarter (as measured by $MacroPC_t$) is given twice the weight as the best quarter.

We would prefer to conduct our analysis using data from stress periods only, but that would

leave us with too few observations. Weighting by the level of stress in a quarter allows us to approximate the effect of conditioning on stress while making greater use of the available data. This approach relies on the assumption that data from non-stressful periods is relevant to forecasting losses in periods of stress.

We merger-adjust all bank data. For example, Truist Financial, one of the banks in Table C.10, was formed from the 2019 merger of BB&T and SunTrust, so our data for Truist in earlier years combines data from those two banks. We repeat this process as we work backwards in time. We obtain information on mergers and acquisitions from the Federal Financial Institutions Examination Council website. (We have also run our analysis without merger-adjusting the data; doing so does not change our conclusions and generally increases heterogeneity across banks.)

In each loan category, we calculate a loss rate (net charge-off rate) for each bank s and each quarter t as the ratio

$$LossRate_{s,t} = \frac{Charge-offs_{s,t} - Recoveries_{s,t}}{Total\ Loans\ in\ Category_{s,t-1}}. \quad (C.14)$$

This measure is commonly used in stress testing; see, for example, Guerrieri and Welch [3], Hirtle et al. [6], and Kapinos and Mitnik [7]. We similarly normalize the amounts past due to get a $PastDueRate_{s,t}$ for each bank-quarter. We remove values less than -50% or greater than 50% of $LossRate$ and values greater than 20% of $PastDueRate$. We winsorize $PastDueRate$ at the upper and lower 5% levels. To attain a mostly balanced panel for more reliable estimates, in each loan category we include only banks with at least 18 years (72 quarters) of history from 2001 Q1 to 2021 Q4.

Table C.3 shows descriptive statistics for these variables. Loss rates and past due rates are shown by loan category — credit cards (CC), first liens (FL), commercial real estate (CRE), and commercial and industrial (CI). Columns 2–4 of the table summarize time-averaged values across banks. Columns 5–8 summarize observations across all banks and quarters.

	bank averages			all observations			
	min	mean	max	lower 5%	mean	upper 5%	std
Loss Rate: CC	-0.20	2.50	3.36	0.31	3.00	5.88	2.01
Loss Rate: FL	0.01	0.22	0.66	-0.01	0.27	1.03	0.51
Loss Rate: CRE	0.06	0.19	0.46	-0.04	0.16	1.00	0.41
Loss Rate: CI	0.05	0.41	1.00	0.00	0.42	1.56	0.52
Past Due Rate: CC	1.10	2.90	4.23	1.06	3.30	5.74	1.53
Past Due Rate: FL	0.71	4.01	8.02	0.45	6.20	11.80	4.60
Past Due Rate: CRE	1.07	2.15	3.80	0.36	2.12	6.52	1.88
Past Due Rate: CI	0.05	1.65	2.85	0.05	1.74	3.97	1.23

Table C.3: Descriptive statistics in percent. Columns 2–4 are calculated from banks’ time averages, and columns 5–8 are calculated from all observations, with mean and standard deviation are stressed time and loan balance weighted.

Figure C.2 plots the mean past due rate (± 1.96 standard errors) for each bank in each loan category. The banks are identified by their stock tickers. The figure illustrates substantial heterogeneity across banks in their loan portfolios. For example, Bank of America (BAC) has among the highest past due rates for credit card loans, but in the commercial real estate category it has among the lowest. This type of pattern is consistent with the idea that banks have different areas of specialization and may target different markets.

The widths of the bars in Figure C.2 show differences across loan categories and banks in the volatility of their past due rates. We again observe significant heterogeneity among different banks.

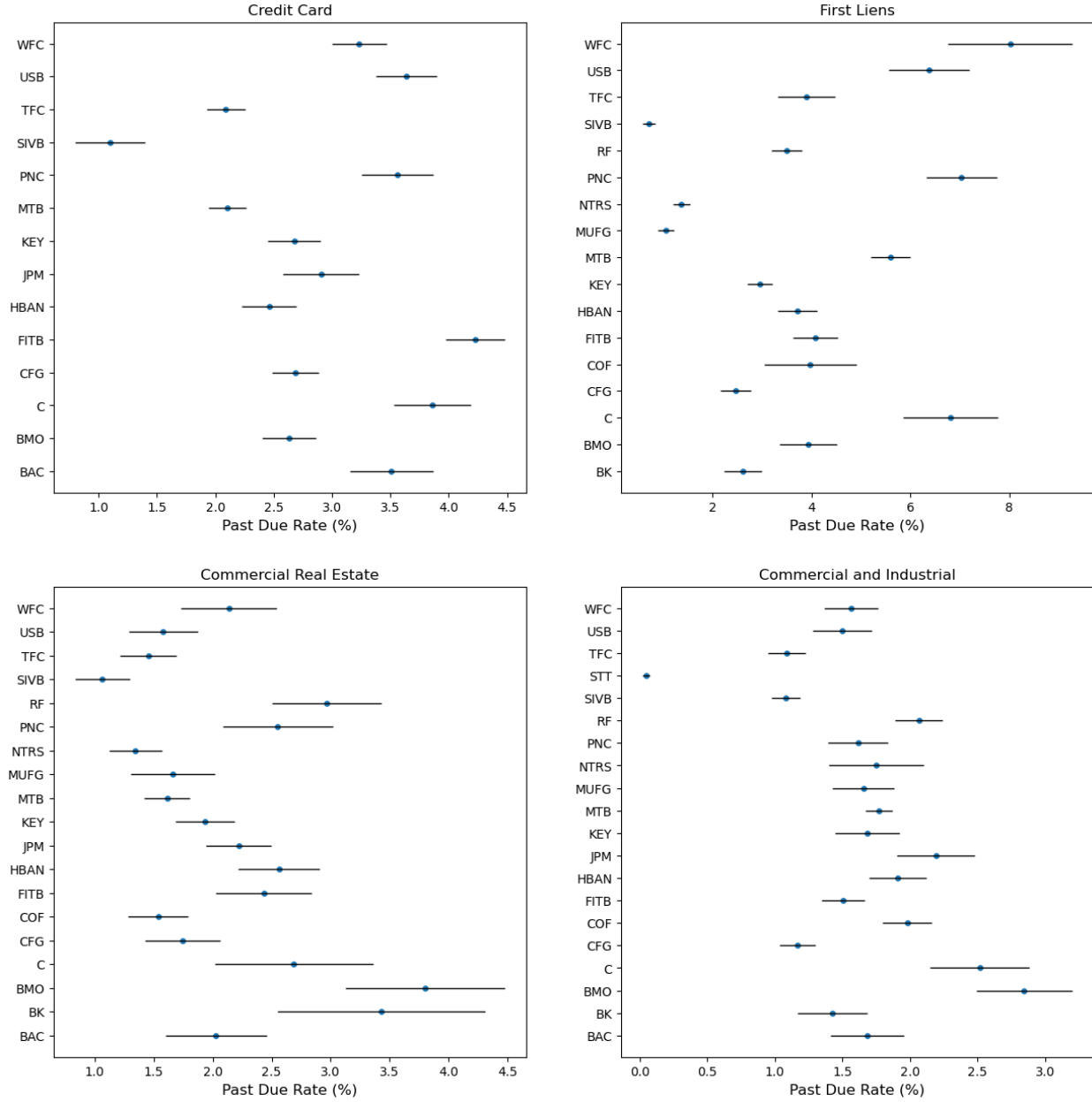


Figure C.2: Past due rates (winsorized) by bank and loan category. The dots show mean values and each horizontal bar corresponds to ± 1.96 standard errors.

C.5.2 Heterogeneity in Slopes and Intercepts

We use the bank data to approximate our theoretical framework through the specification

$$LossRate_{s,t} = \alpha_s + \beta_s PastDueRate_{s,t-l} + \gamma_s MacroPC_{t-l} + \epsilon_{s,t}, \quad (C.15)$$

for bank s in quarter t , where $MacroPC$ is the principal component of the macro variables introduced in Section C.5.1. (In Section C.5.3, we also include allowances in (C.15) as a robustness check.) The lag l is four quarters to mimic the stress testing's forward-looking framework. We estimate separate coefficients for each of the four loan categories, for each bank, and the observations are loan balance and stress weighted using (C.13). Because these are bank-specific regressions, we do not add bank-specific controls.

For each loan category, we want to test for heterogeneity in parameters across banks. When we test for heterogeneity, the null hypothesis states that slopes for all banks are equal,

$$H_0 : \beta_1 = \dots = \beta_{\bar{S}}, \quad (C.16)$$

or that the intercepts are equal,

$$H_0 : \alpha_1 = \dots = \alpha_{\bar{S}}. \quad (C.17)$$

The alternative hypothesis in each case states that the indicated parameters are not identical across banks. We will run these tests with different subsets of the variables in (C.15) included and interpret the coefficients in (C.16) accordingly.

To test these hypotheses for a particular loan category, let \mathbf{X}_s be the n_s by k data matrix for bank s , where n_s is the number of observations for bank s in the loan category, and $k = 1$ or 2 is the number of variables included on the right side of (C.15). Let $\tilde{\mathbf{X}}_s = (\mathbf{1}, \mathbf{X}_s)$ be \mathbf{X}_s concatenated with a column of 1s, and let \mathbf{X}^* be the diagonal block matrix $\mathbf{X}^* = \text{diag}(\tilde{\mathbf{X}}_1, \dots, \tilde{\mathbf{X}}_{\bar{S}})$. Let $\theta^* = (\alpha_1, \beta_1^\top, \dots, \alpha_S, \beta_S^\top)^\top$, $\epsilon^* = (\epsilon_1, \dots, \epsilon_S)$, where ϵ_s is a column vector of length n_s . Our unrestricted model can be written as

$$Y = \mathbf{X}^* \theta^* + \epsilon^*, \quad (C.18)$$

and the restrictions in (C.16) and (C.17) impose linear constraints on the parameter θ^* .

We apply the Wald test to test linear constraints on θ^* in (C.18) under various assumptions on the error covariance matrix. We consider (i) bank-clustered errors, which allows correlation in errors across quarters for each bank, but no correlation across banks; (ii) time-clustered errors,

which allows correlation in errors across banks in each quarter, but no correlation across time.

Table C.4 reports p -values for the tests when different subsets of variables are included on the right side of (C.15), for a forecast horizon of one year. All tests indicate strong evidence of heterogeneity in the intercepts, the coefficients for past due rates, and macro variables.

Next we examine the impact of heterogeneity. Table C.5 compares pooled and FEO coefficients for *PastDueRate* using a one-year lag when *MacroPC* is and is not included in (C.15). We estimate β_{Pool} in a pooled panel regression, and β_F in a panel regression with bank fixed effects included. Both regressions are weighted using (C.13).

In Table C.5, the columns labeled “diff” show the difference in estimates $\beta_F - \beta_{Pool}$, serving as a measure of the impact of addressing heterogeneity in choosing an industry model. The table also shows p -values for tests of $H_0 : \beta_{Pool} = \beta_F$ (or $\gamma_{Pool} = \gamma_F$ in the case of the macro variable). To calculate these p -values, we estimate the pooled and FEO models simultaneously, as follows. Let $\tilde{\mathbf{X}}_{Pool} = (\mathbf{1}, \mathbf{X})$ be \mathbf{X} concatenated with a column of 1s, and let $\tilde{\mathbf{X}}_F = (\mathbf{U}, \mathbf{X})$ be \mathbf{X} concatenated with columns corresponding to centered bank identity variables \mathbf{U} . Let \mathbf{X}_* be the diagonal block matrix $\mathbf{X}_* = \text{diag}(\tilde{\mathbf{X}}_{Pool}, \tilde{\mathbf{X}}_F)$ and $\theta_* = (\alpha_{Pool}, \beta_{Pool}^\top, \delta_1, \delta_2, \dots, \delta_{\bar{s}}, \beta_F^\top)^\top$. Then we have

$$Y = \mathbf{X}_* \theta_* + \epsilon_*,$$

and testing H_0 is equivalent to testing linear constraints on the parameters θ_* , for which we apply the Wald test. The macro variable captures common variability over time, so we cluster errors by bank.

The results in Table C.5 show that the differences between the pooled and FEO estimates are significant in three of the four loan categories. Moreover, the differences can be material. For example, for first lien loans in the top panel of Table C.5, an absolute difference of 0.015 translates to a relative difference of 24% ($= |\beta_{Pool} - \beta_F| / \beta_F$), which can have a large relative impact on predicted losses. From Table C.3, we see that the average bank has a past due rate of 4.01% on FL loans. The difference $0.015 \times 4.01\% = 0.060\%$ is 27% of the average FL loss rate of 0.22% in

Table C.3. The additional capital required to offset the higher predicted loss rate would be 27% of the capital required to offset the average loss rate.

To further analyze the differences in forecasts, we consider the relative prediction differences given by $|(\hat{Y}_{Pool}(x) - \hat{Y}_F(x))/Y(x)|$, where the denominator is the observed loss rate. Table C.6 reports the mean and median of these relative differences in each of the four loan categories, for all banks in all quarters. In each loan category, the mean relative difference is large; moreover, in each case the mean is appreciably larger than the median, reflecting the presence of some very large relative prediction differences, which could be particularly important. To illustrate the differences, Figure C.3 plots histograms of the FEO and pooled fitted values for Citigroup’s first lien loans. The comparison shows, in particular, that the frequency of the largest and smallest predicted loss rates differ between the two methods.

Our main results use data through 2021. As a robustness check, we run our analysis using data through 2019. This truncation serves two purposes. It ensures that our conclusions are not driven by a few extreme values during the COVID period 2020–2021, and it accounts for a change in how banks measure allowances (the Current Expected Credit Losses methodology) beginning at the end of 2019. We also consider including banks’ allowances for losses as another proxy for the portfolio risk. Because allowances are not consistently reported separately by loan category, we use banks’ total allowances across all loan types. That is, for each bank-quarter we calculate $AllowanceRate_{s,t}$ using (C.14), but normalizing by the total loans in all categories. We repeat our tests with pre-COVID data and the addition of allowance rates. The results, reported in Section C.5.3, are similar to those reported in this section.

Our results document evidence of bank heterogeneity and its potential impact on loss forecasts. We have not sought to identify the drivers of heterogeneity; that would require a very different investigation, particularly since some of the most interesting potential drivers are difficult to measure. Guerrieri and Harkrader [2], for example, find that bank-specific factors account for a sizable fraction of the variation in bank performance. But they measure the bank-specific component as the residual in a regression that removes the effect of macroeconomic and banking-wide factors; they

Covariance Estimation	α				β_{PDR}				γ			
	CC	FL	CRE	CI	CC	FL	CRE	CI	CC	FL	CRE	CI
bank clustered	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00				
time clustered	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00				
bank clustered	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
time clustered	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table C.4: P -values for heterogeneity tests. In each loan category, the first two rows are for a model with *PastDueRate* only, and the last two rows are for a model with *PastDueRate* and *MacroPC*. The two rows for each model show results under alternative assumptions on the error covariance matrix.

Loan Type	<i>Past Due Rate</i>				<i>Macro PC</i>			
	β_{Pool}	β_F	diff	p -value	γ_{Pool}	γ_F	diff	p -value
CC	0.782	0.833	-0.050	0.009***				
FL	0.047	0.062	-0.015	0.001***				
CRE	0.131	0.129	0.002	0.464				
CI	0.208	0.219	-0.011	0.040**				
CC	0.774	0.823	-0.049	0.009***	0.040	0.037	0.003	0.000***
FL	0.046	0.061	-0.015	0.001***	0.018	0.019	-0.001	0.176
CRE	0.131	0.129	0.002	0.485	0.011	0.011	0.000	0.813
CI	0.205	0.216	-0.010	0.044**	0.022	0.022	0.000	0.127

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table C.5: Comparison of coefficients for one-year forecasts. We regress *LossRate* on (i) *PastDueRate* and (ii) *PastDueRate* and *MacroPC*. Difference is calculated as $\beta_{Pool} - \beta_F$. p -values test $H_0 : \beta_{Pool} = \beta_F$ for *PastDueRate* or $H_0 : \gamma_{Pool} = \gamma_F$ for *MacroPC*.

do not identify specific bank features that influence performance. Some examples of bank features used as controls in stress testing models can be found in Hirtle et al. [6], Kapinos and Mitnick [7], and Kupiec [8]. These are balance sheet features, and they are usually found to be more relevant to forecasting revenues than losses. We discuss revenue models in Section C.8.

C.5.3 Robustness Checks

We repeat the analysis of Section C.5, limiting the data to 2001–2019. This serves two purposes. It addresses the possibility that our results are driven by a few extreme values during the COVID period 2020–2021. It also accounts for a change in how banks measure allowances (the

	CC	FL	CRE	CI
mean	6.1	510.1	52.0	24.0
median	2.4	89.5	4.2	3.7

Table C.6: Mean and median of relative prediction differences between the pooled and the FEO estimates (in %).

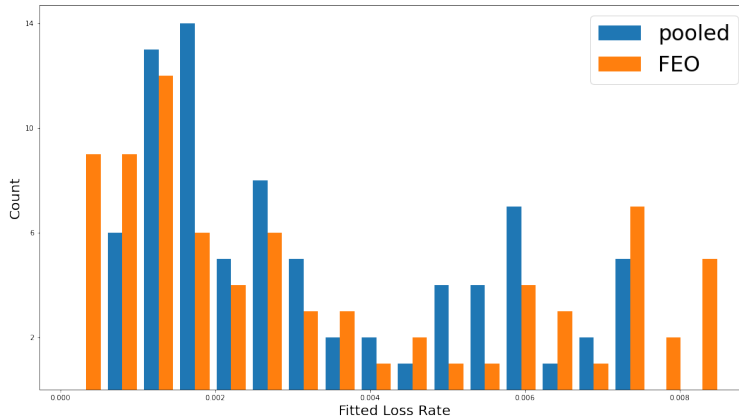


Figure C.3: Pooled and FEO predicted loss rates for Citigroup's first lien loans.

Current Expected Credit Losses methodology) that began to take effect at the end of 2019. We also consider adding allowance rate as another proxy for portfolio risks. Tables C.7 and C.8 report the results for heterogeneity tests and differences of the parameter estimates under this setting. The evidence for heterogeneity and its impact is generally at least as strong using the pre-COVID data as using data through 2021.

Covariance Estimation	α			β_{PDR}			β_{AR}			γ		
	CC	FL	CRE	CI	CC	FL	CRE	CI	CC	FL	CRE	CI
bank clustered	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
time clustered	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
bank clustered	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
time clustered	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
bank clustered	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
time clustered	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
bank clustered	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
time clustered	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table C.7: Heterogeneity tests using pre-COVID data with allowance rate as an additional proxy for banks' portfolio risks.

Loan Type	Past Due Rate			Allowance Rate			Macro PC			
	β_{Pool}	β_F	diff	β_{Pool}	β_F	diff	γ_{Pool}	γ_F	diff	p-value
CC	0.781	0.838	-0.057	0.006***						
FL	0.045	0.061	-0.016	0.002***						
CRE	0.135	0.133	0.002	0.449						
CI	0.207	0.217	-0.011	0.069*						
CC	0.794	0.852	-0.058	0.109	-0.029	-0.032	0.003	0.938		
FL	0.010	0.021	-0.011	0.306	0.335	0.307	0.028	0.424		
CRE	0.095	0.068	0.026	0.052*	0.127	0.191	-0.064	0.090*		
CI	0.178	0.196	-0.018	0.000***	0.064	0.054	0.010	0.514		
CC	0.595	0.633	-0.039	0.001***					0.417	0.397
FL	0.037	0.052	-0.015	0.001***					0.110	0.106
CRE	0.129	0.126	0.002	0.477					0.057	0.057
CI	0.155	0.155	-0.001	0.886					0.190	0.190
CC	0.577	0.618	-0.041	0.046**	0.037	0.032	0.005	0.846	0.420	0.399
FL	0.006	0.018	-0.012	0.146	0.315	0.268	0.047	0.041**	0.095	0.094
CRE	0.097	0.078	0.019	0.182	0.102	0.147	-0.045	0.265	0.052	0.050
CI	0.140	0.153	-0.012	0.035**	0.032	0.007	0.025	0.105	0.189	0.189

*p<0.1; **p<0.05; ***p<0.01

Table C.8: Comparison of pooled and FEO coefficients using pre-COVID data with allowance rate as an additional proxy for banks' portfolio risks.

C.6 Nonlinear Models

Building on Section 3.5, we consider generalized additive models (GAMs) in which the effect of the past due rates and the macro variable are not restricted to be linear. More specifically, we consider the specifications

$$Y_{s,t}^{Pool} = f_0^P + f_1^P(PDR_{s,t-l}) + f_2^P(MacroPC_{t-l}) + f_3^P(PDR_{s,t-l} \times MacroPC_{t-l}) + \epsilon_{s,t}^P \quad (C.19)$$

and

$$Y_{s,t}^F = f_0^F + f_1^F(PDR_{s,t-l}) + f_2^F(MacroPC_{t-l}) + f_3^F(PDR_{s,t-l} \times MacroPC_{t-l}) + f_4^F(s) + \epsilon_{s,t}^F, \quad (C.20)$$

in which $f_i^{P/F}$, $i = 1, 2, 3$, are (possibly nonlinear) centered functions of *PastDueRate*, *MacroPC*, and their interaction *PastDueRate*×*MacroPC*, respectively, and f_4^F measures centered bank fixed effects. The pooled model (C.19) omits f_4^P ; the FEO model (C.20) estimates f_4^F but discards it in forecasting loss rates to satisfy equal treatment. We consider the modeling of loss rates four quarters ahead, so $l = 4$ in both cases.

We use the R package `gam` (Hastie [5]) to fit (C.19) and (C.20), taking the $f_i^{P/F}$, $i = 1, 2, 3$, to be smoothing splines with 4 degrees of freedom. We choose `gam` because it is a direct implementation of the backfitting algorithm in Hastie and Tibshirani [4], which underpins the framework in Section 3.5. In particular, Proposition 3.5.1 applies to FEO forecasts based on (C.20).

Model (C.19) is nested within model (C.20), so we can use an F -test to compare the two. In all four loan categories, the test rejects (with p -values smaller than 0.01) the restriction to equal bank fixed effects ($f_4^P \equiv 0$) imposed in the pooled model (C.19). Figure C.4 plots the centered bank fixed effects f_4^F for all four loan categories, expressed in percent. We observe significant variability within each loan type. For example, for credit card loans, JPM's fixed effect is three percentage points larger than WFC's. We also observe variability across loan types for individual banks. For example, CFG has the highest fixed effect for first lien loans, but one of the lowest

for credit card loans. These observations again reflect the notable heterogeneity among the bank holding companies.

Table C.9 reports the mean and median of the relative prediction differences between the FEO and the pooled predictions, given, as in Section C.5, by $|(\hat{Y}_{Pool}(x) - \hat{Y}_F(x))/Y(x)|$. These summary statistics show that the relative prediction differences can indeed be very large. To further illustrate this point, Figure C.5 contrasts the prediction distributions for Citigroup's first lien loans using the pooled and FEO methods. As in Proposition 3.5.1, the pooled method may yield smaller prediction errors overall, but it does so by implicitly misdirecting legitimate information.

	CC	FL	CRE	CI
mean	25.5	559.0	106.6	88.5
median	10.4	120.8	14.2	9.8

Table C.9: Mean and median of relative prediction differences between the pooled and the FEO estimates (in %) for GAMs.

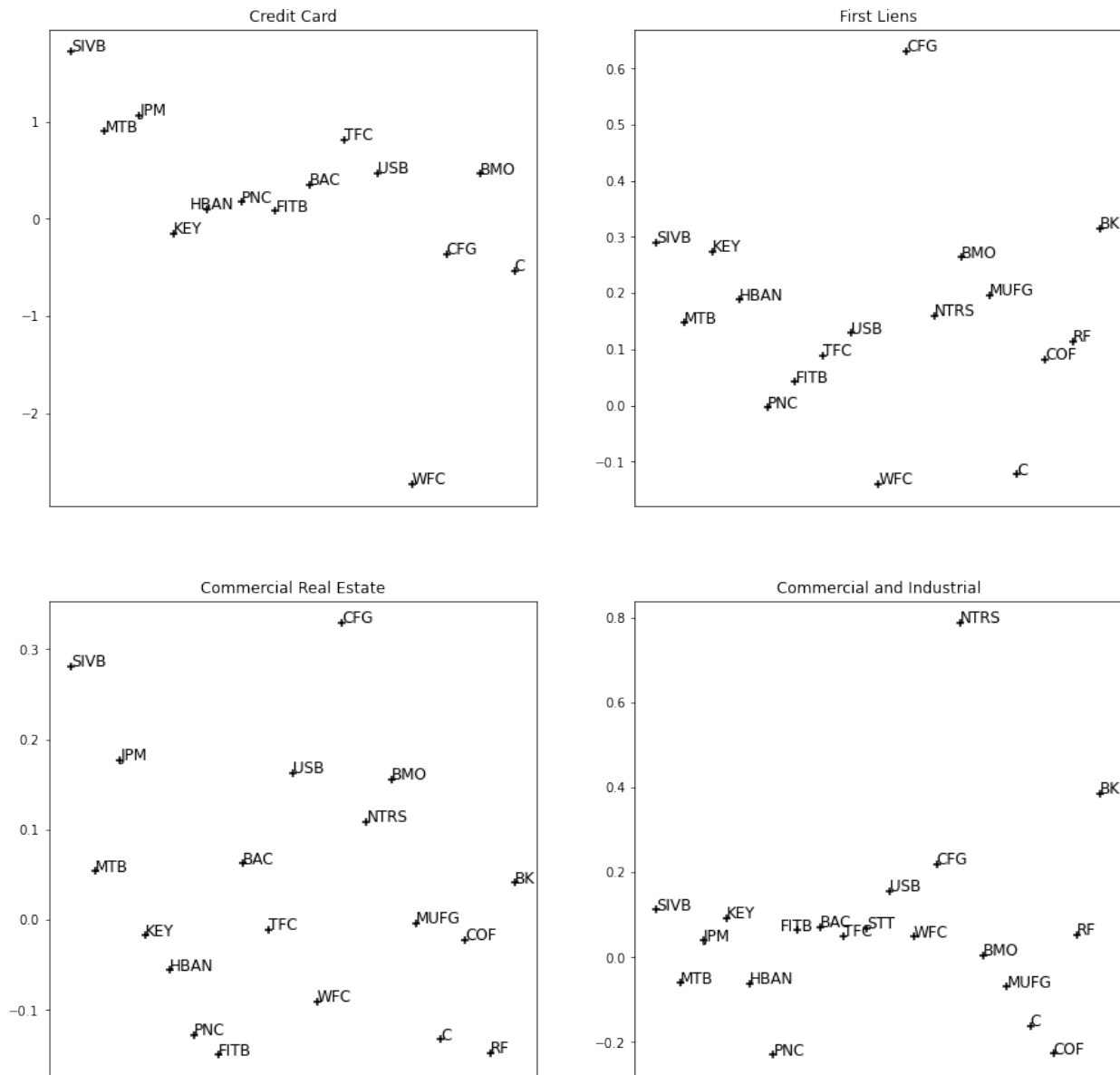


Figure C.4: Banks' generalized fixed effects. Y-axis is in %.

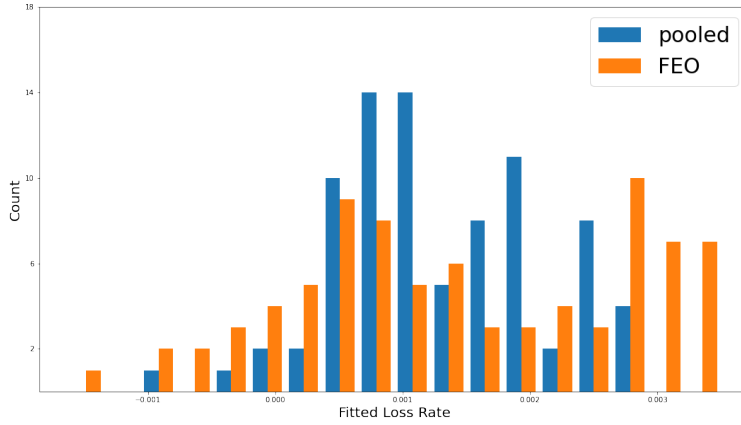


Figure C.5: Pooled and FEO predicted loss rates for Citigroup’s first lien loans.

C.7 Additional Information on Empirical Analysis

Table C.10 lists the bank holding companies included in our empirical analysis and the symbols we use to refer to them. The companies are listed in order of size by total assets.

Ticker	Bank Name
JPM	JPMORGAN CHASE & CO.
BAC	BANK OF AMERICA CORPORATION
C	CITIGROUP INC.
WFC	WELLS FARGO & COMPANY
GS	GOLDMAN SACHS GROUP, INC., THE
MS	MORGAN STANLEY
SCHW	CHARLES SCHWAB CORPORATION, THE
USB	U.S. BANCORP
PNC	PNC FINANCIAL SERVICES GROUP, INC., THE
TFC	TRUIST FINANCIAL CORPORATION
TD	TD GROUP US HOLDINGS LLC
BK	BANK OF NEW YORK MELLON CORPORATION, THE
COF	CAPITAL ONE FINANCIAL CORPORATION
STT	STATE STREET CORPORATION
HSBC	HSBC NORTH AMERICA HOLDINGS INC.
SIVB	SVB FINANCIAL GROUP
FITB	FIFTH THIRD BANCORP
USAA	UNITED SERVICES AUTOMOBILE ASSOCIATION
BMO	BMO FINANCIAL CORP.
CFG	CITIZENS FINANCIAL GROUP, INC.
AXP	AMERICAN EXPRESS COMPANY
KEY	KEYCORP
NTRS	NORTHERN TRUST CORPORATION
ALLY	ALLY FINANCIAL INC.
AMP	AMERIPRISE FINANCIAL, INC.
RY	RBC US GROUP HOLDINGS LLC
HBAN	HUNTINGTON BANCSHARES INCORPORATED
RF	REGIONS FINANCIAL CORPORATION
MUFG	MUFG AMERICAS HOLDINGS CORPORATION
BCS	BARCLAYS US LLC
SAN	SANTANDER HOLDINGS USA, INC.
MTB	M&T BANK CORPORATION
BNPQY	BNP PARIBAS USA, INC.
DB	DB USA CORPORATION
DFS	DISCOVER FINANCIAL SERVICES

Table C.10: Symbols and names of included bank holding companies.

We construct the loss rates, past due rates, and allowance rates using the entries in FR Y-9C forms outlined in Table C.11.

Variables	Loan Types	2007Q1 – Present	2003Q1 – 2006Q4
Loan Amount	CC	BHCKB538	BHCKB538
	FL	BHDM5367	BHDM5367
	CRE	Owned: BHCKF160 Other: BHCKF161	BHDM1480
	CI	BHCK1763	BHCK1763
	Total	BHCK2122	BHCK2122
Charge-Offs	CC	BHCKB514	BHCKB514
	FL	BHCKC234	BHCKC234
	CRE	Owned: BHCKC895 Other: BHCKC897	BHCK3590
	CI	BHCK4645	BHCK4645
Recoveries	CC	BHCKB515	BHCKB515
	FL	BHCKC217	BHCKC217
	CRE	Owned: BHCKC896 Other: BHCKC898	BHCK3591
	CI	BHCK4617	BHCK4617
Past Due: 30-89 days and accruing	CC	BHCKB575	BHCKB575
	FL	BHCKC236	BHCKC236
	CRE	Owned: BHCKF178 Other: BHCKF179	BHCK3502
	CI	BHCK1606	BHCK1606
Past Due: 90 days and accruing	CC	BHCKB576	BHCKB576
	FL	BHCKC237	BHCKC237
	CRE	Owned: BHCKF180 Other: BHCKF181	BHCK3503
	CI	BHCK1607	BHCK1607
Past Due: non-accrual	CC	BHCKB577	BHCKB577
	FL	BHCKC229	BHCKC229
	CRE	Owned: BHCKF182 Other: BHCKF183	BHCK3504
	CI	BHCK1608	BHCK1608

Table C.11: Loan variables and FR Y-9C form correspondence.

C.8 Revenue Models

The Federal Reserve’s stress testing framework includes models of revenues as well as models of losses. We have focused on loan portfolio loss models because they fit most clearly within the Fed’s policy of equal treatment and its preference for industry models. In this section, we show that the heterogeneity documented for loss models in Section C.5 extends to revenue models, referred

to in the Fed’s framework as models of pre-provision net revenue or PPNR.

The Fed uses a suite of PPNR models to forecast difference sources of revenue. These models differ from the loss models in at least two important respects: they are typically autoregressive (AR) models, and, unlike the portfolio loss models, they do not rule out bank fixed effects; see [1]. This feature points to the presence of unmodeled bank heterogeneity in the revenue forecasts. Our goal in this section is to check for heterogeneity in a simple PPNR model and to compare coefficient estimates in the pooled and fixed-effect models.

We consider the modeling of trading revenue, which is one of the PPNR components in the Fed’s framework. We compare AR models with fixed-effects,

$$Y_{s,t}^{FE} = \alpha_s + \rho_{FE}Y_{s,t-1} + \beta_{FE}X_{s,t} + \gamma_{FE}VIX_t + \epsilon_{s,t}^{FE} \quad (C.21)$$

or pooled without fixed effects,

$$Y_{s,t}^P = \alpha_P + \rho_P Y_{s,t-1} + \beta_P X_{s,t} + \gamma_P VIX_t + \epsilon_{s,t}^P. \quad (C.22)$$

In both models, Y is trading revenue normalized by total trading assets; this choice of normalization is consistent with [1]. For the AR term, we use either a one-quarter lag $Y_{s,t-1}$ or a four-quarter average lag, in which case we use $1/4 \sum_{j=1}^4 Y_{s,t-j}$ in place of $Y_{s,t-1}$ in (C.21) and (C.22) to capture average performance over the past year.

For $X_{s,t}$ we use the size of bank s in quarter t , as measured by the log of total assets. For the macro variable, we use the VIX_t , the market volatility index taken from the Federal Reserve’s stress testing historical dataset. Market volatility is expected to have a direct impact on trading revenue, and indeed we observe a more significant effect of VIX_t than $MacroPC_t$ (from Section C.5) in this setting.

Models (C.21) and (C.22) differ in their intercept terms: (C.21) captures banks’ fixed effects, but (C.22) requires the same intercept across all banks. The fixed-effect coefficient estimates ρ_{FE} , β_{FE} and γ_{FE} are identical to those of FEO; the methods differ in their forecasts: FEO uses

the average fixed effect, rather than bank-specific fixed effects in its forecasts.

As in Section C.5, we use Y-9C financial reporting data for the top 35 banks by total asset size (as of year-end 2021), and we include only banks with at least 18 years of data to ensure our panel is mostly balanced. We fit the models using weighted least squares, weighting each observation by quarter stress and bank asset balance. As in (C.13), we choose the weights so that, for the same asset level, the quarters with the highest market volatility get twice the weight as the quarters with the lowest market volatility. As in the AR models in [6], we cluster standard errors by time.

Table C.12 reports the results. The first two columns correspond to the one-quarter AR setting, and the last two columns are the one-year-average AR setting. The numbers in parentheses are standard errors.

As expected, both the lagged response and the *VIX* term are statistically significant. The volatility term is more significant in the one-year-average AR setting, presumably because the market environment changes less over one quarter, and its effect is partly captured by the lagged response.

The pooled and fixed-effect methods result in different estimates of ρ , and the differences in estimates in all settings are more than 20%. We tested the hypotheses that all banks share the same (i) intercept, (ii) AR coefficient term, (iii) *VIX* coefficient, or (iv) coefficients of total asset size, following the approach used in Section C.5.2; all tests strongly reject that banks have identical model coefficients, with *p*-values less than 0.01, extending what we found for loss models. We have also examined the forecasts of trading interest income and trading interest expense (two other components of PPNR), and observed similar patterns across all three components.

Bibliography

- [1] Board of Governors (2021) Dodd-Frank Act Stress Test 2021: Supervisory Stress Test Methodology. Federal Reserve System, Washington, D.C.
- [2] Guerrieri, L., and Harkrader, J. (2021) What drives bank performance?, Working paper 2021-009, Federal Reserve Board, Washington, D.C.
- [3] Guerrieri, L., and Welch, M. (2012) Can macro variables used in stress test forecast the performance of banks? Working paper 2012-49, Federal Reserve Board, Washington, D.C.

<i>Normalized Trading Revenue</i>				
AR	0.466*** (0.134)	0.588*** (0.111)	0.406*** (0.078)	0.596*** (0.094)
VIX	-0.024** (0.011)	-0.022* (0.012)	-0.027*** (0.008)	-0.022*** (0.008)
Log Assets	-0.003 (0.004)	-0.007*** (0.001)	0.000 (0.003)	-0.006*** (0.002)
Bank FE	included	–	included	–

*p<0.1; **p<0.05; ***p<0.01

Table C.12: Coefficient estimates for the AR models. The first two columns use a one-quarter lag, and the last two use a one-year average lag. Columns 1 and 3 correspond to AR models with bank fixed effects, and columns 2 and 4 are pooled AR models.

- [4] Hastie, T., and Tibshirani, R. (1986) Generalized additive models, *Statistical Science* 1(3), 297–318.
- [5] Hastie, T. (2023) gam: Generalized Additive Models, R package version 1.22-1, <https://CRAN.R-project.org/package=gam>.
- [6] Hirtle, B., Kovner, A., Vickery, J., and Bhanot, M. (2016) Assessing financial stability: The Capital and Loss Assessment under Stress Scenarios (CLASS) model. *Journal of Banking and Finance* 69, S35–S55.
- [7] Kapinos, P., and Mitnik, O.A. (2016) A top-down approach to stress-testing banks. *Journal of Financial Services Research* 49, 229–264.
- [8] Kupiec, P. (2020) Policy uncertainty and bank stress testing. *Journal of Financial Stability* 51, 100761.
- [9] Le Gouic, T., Loubes, J.-M., and Rigollet, P. (2020), Projection to fairness in statistical learning, arXiv:2005.11720