ESSAYS ON MACROECONOMICS OF MONETARY AND FISCAL POLICIES

A Dissertation
Presented to the Faculty of the Graduate School
of Cornell University
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by
Yu She
August 2020
My thesis contains three chapters which focus heavily on the macroeconomic policies. The first chapter focuses on the effectiveness of monetary policy on firms with different financial constraints. The second chapter addresses on how would the optimal tax policy change the evolution of inequality. The third chapter emphasizes on how to provide a proxy means testing from a welfare perspective to a transfer program.

In the first chapter, I study the role of financial constraints in the effects of monetary policy on firm investments. I construct a quarterly textual measure of financial constraints from SEC filings using a deep learning model. It improves the prediction accuracy as compared to a Naive Bayes method by capturing the context information, such as grammatical structure and order of words. Firms classified as highly constrained are younger, smaller, have a higher liquidity ratio and higher leverage ratio. However, popular proxies of financial constraints often do not move monotonically with the level of financial constraints. Particularly for the liquidity ratio, it is high for both the least constrained firms, which have ample of cash, and the most constrained firms, which hoard cash due to precautionary saving motives and the high marginal cost of external capital. Using the constructed measure of financial constraints, the investments of financially constrained firms are persistently less responsive to
monetary policy shocks due to high marginal cost of external funds. This implies that monetary policy might be less effective during crisis time due to a larger fraction of constrained firms. My results reconcile previous empirical findings and argue that the seemingly contrary conclusions are, to some extent, consistent with each other.

In the second chapter, it intends to address on the question: how would the optimal taxes change the evolution of wealth inequality? This paper studies this question quantitatively under a standard incomplete market heterogeneous agent model. The benchmark model captures the wealth distribution and its evolution from 1967-2010. Optimal tax policy exercise considers an once-and-for-all tax reform at 1967 accounting for the time varying economic environment and transition dynamics. With a utilitarian social planner, the optimal linear comprehensive income tax leads to a higher level inequality in wealth where top 10% and top 1% gain at least 5% more wealth shares at 2010 compared to benchmark. The optimal tax under a parameterized nonlinear tax function implies a highly progressive tax system which is also highly redistributive compared to the benchmark model. The wealth inequality in this case is increasing from 1960s to mid 1990s and then start to decline to its 1960s level or even lower. At 2010, top 10% remains roughly their wealth holdings at their 1967 level while top 1%, 0.1% and 0.01% wealth holding even decrease on average about 2% compared to their low level at year 1967.

In the last chapter, I propose a new proxy means testing method with minimizing welfare loss as the target instead of traditional targets such as
minimizing consumption loss. In a simple economy with a utilitarian social planner, the welfare approach is equivalent to a weighted logistic regression with inverse consumption as weights. As a result, it focuses mainly on the exclusion error where poor are identified as non-poor and less weights on the inclusion error where non-poor are identified as poor. Using the socio-economic survey data in India in 2011, I compare the targeting performance of the welfare approach to other standard approaches in PMT. It shows that the welfare approach enjoys a lower exclusion error rate by sacrificing the inclusion error rate and does not out-perform the traditional method. It does, on the other hand, provide a welfare foundation for the poverty weighted least square method.
BIOGRAPHICAL SKETCH

Yu She joined the doctoral program in the Department of Economics at Cornell University in the fall of 2014. His broad research interests are in applied macroeconomics, monetary and fiscal policy, and his doctoral research focuses on the heterogeneous impacts of monetary/fiscal policies over the distribution of heterogeneous firms/households. In the fourth year of his PhD period, Mr. She worked on IMF’s engagement on social spending issues using Natural Language Processing as a summer intern with the International Monetary Fund in Washington, D.C. He holds a Bachelor of Science with majors in Economics and Applied Mathematics from University of Michigan, where he also holds a Master of Applied Economics.
This thesis is dedicated to my parents, Zhigang She and Liya Cheng.

For their endless love, support and encouragement.
ACKNOWLEDGEMENTS

My deepest gratitude goes to my committee members and mentors Kristoffer Nimark, Christopher Huckfeldt and Yifei Mao for their invaluable advice and support. I have truly learned a lot from them over the last few years in my development as an economist. I am especially grateful to Kristoffer Nimark, my committee chair, for guiding me through all stages of my dissertation and for his full support and encouragement during my difficult time of finding the right topic. I am indebted to Yifei Mao, without whose insightful advice on empirical works and mental support I could not have finished the dissertation within expected time.

My first encounter of economic research started with Jeffrey Smith in the undergraduate research program back in my undergraduate time at the University of Michigan. His encouragement and mentorship was crucial in my decision of pursuing the PhD degree and becoming an economist. I am thankful to Miles Kimball for introducing and guiding me further to economic research during my time as his research assistant, and to Robert Barsky, who sparked my interest in macroeconomics.

In the course of writing this dissertation, I have greatly benefited from generous advice and discussion from numerous individuals. I sincerely thank Julieta Caunedo, Maxim Troshkin, Mathieu Taschereau-Dumouchel, Penny Sanders and Seth Sanders for several helpful comments. I learned a great deal from the discussion with my peers and friends at Cornell, including Abhishek Ananth, Tilahun Emiru, Sylverie Herbert and Nahim Bin Zahur.
I thank Isha Agarwal, Levon Barseghyan, David Coady, Andrew Fieldhouse, Maura Francese, Malin Hu, Shanjun Li, Cynthia Lin, Michael Lovenheim, Francesca Molinari and seminar participants at Cornell University and International Monetary Fund. I appreciate the hospitality of the International Monetary Fund, where parts of my doctoral studies were enlightened from my time there. The views in this dissertation are those of the authors and do not necessarily reflect the views of the International Monetary Fund.

No words can suffice to thank my parents, Zhigang She and Liya Cheng. Their love and support has been my constant source of strength and happiness.
# TABLE OF CONTENTS

Biographical Sketch ....................................................... iii
Dedication .................................................................iv
Acknowledgements ............................................................. v
Table of Contents ............................................................... vii
List of Tables ................................................................ ix
List of Figures .................................................................. x

1 Financial Constraints, Monetary Policy and Firm Investments: A Text-Based Deep Learning Approach ................................................................. 1
   1.1 Introduction ................................................................. 1
      1.1.1 Related Literature ................................................ 6
   1.2 Data ........................................................................ 9
   1.3 Measure of Financial Constraints ............................... 11
      1.3.1 Construction Strategy ........................................... 11
      1.3.2 MD&A Extraction ............................................... 12
      1.3.3 Training Sample .................................................. 13
      1.3.4 Predicting Model - Transformer Bi-LSTM ............... 16
   1.4 Characteristics of The Financial Constraints ............. 28
      1.4.1 Summary Statistics ............................................. 28
      1.4.2 The Relationship between Proxies and Financial Constraints 31
   1.5 The Effects of Monetary Policy Across Firms ............ 35
      1.5.1 Panel Local Projections with Firm Level Data ....... 37
      1.5.2 Baseline Result of Investment Dynamics to Monetary Policy Shocks .............................................. 39
      1.5.3 Comparison with Different Measures .................. 43
      1.5.4 Robustness of Empirical Results ......................... 46
      1.5.5 Heterogeneous Debt Dynamics to Monetary Policy Shocks ......................................................... 46
   1.6 Conclusion ................................................................. 48

2 Implications of Optimal Tax for Evolution of Wealth Inequality ................................................. 51
   2.1 Introduction ................................................................. 51
   2.2 Literature Review .......................................................... 55
   2.3 Setup ..................................................................... 60
      2.3.1 Consumers ........................................................ 62
      2.3.2 Production and government ................................ 63
      2.3.3 Steady State Equilibrium ..................................... 64
   2.4 Quantitative Model of Wealth Inequality Dynamics .... 65
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.4.1  Calibration</td>
<td>65</td>
</tr>
<tr>
<td>2.4.2  Steady State Equilibrium</td>
<td>70</td>
</tr>
<tr>
<td>2.4.3  Transition dynamics</td>
<td>72</td>
</tr>
<tr>
<td>2.5    Counterfactual Experiments</td>
<td>75</td>
</tr>
<tr>
<td>2.6    Conclusion</td>
<td>84</td>
</tr>
<tr>
<td>3      A Welfare Approach to the Proxy Means Testing</td>
<td>85</td>
</tr>
<tr>
<td>3.1    Introduction</td>
<td>85</td>
</tr>
<tr>
<td>3.2    Targeting Performance and Poverty Measure</td>
<td>88</td>
</tr>
<tr>
<td>3.3    The Welfare Approach</td>
<td>90</td>
</tr>
<tr>
<td>3.3.1  Uniform Transfer Program</td>
<td>90</td>
</tr>
<tr>
<td>3.3.2  Differential Transfer Program</td>
<td>92</td>
</tr>
<tr>
<td>3.4    Data</td>
<td>93</td>
</tr>
<tr>
<td>3.5    PMT Comparison</td>
<td>94</td>
</tr>
<tr>
<td>3.6    Conclusion</td>
<td>98</td>
</tr>
<tr>
<td>A      Chapter 1 of appendix</td>
<td>100</td>
</tr>
<tr>
<td>B      Chapter 2 of appendix</td>
<td>115</td>
</tr>
<tr>
<td>B.1    Calibration Details</td>
<td>115</td>
</tr>
<tr>
<td>B.1.1  Labor Earnings</td>
<td>115</td>
</tr>
<tr>
<td>B.1.2  Tax System</td>
<td>115</td>
</tr>
<tr>
<td>B.2    Computational Algorithms</td>
<td>118</td>
</tr>
<tr>
<td>B.2.1  Algorithm for Solving Equilibrium</td>
<td>118</td>
</tr>
<tr>
<td>B.2.2  Endogenous Grid Method with Policy Function Iteration</td>
<td>119</td>
</tr>
<tr>
<td>B.2.3  Algorithm for Computing Ergodic Distribution</td>
<td>122</td>
</tr>
<tr>
<td>B.2.4  Algorithm for Computing Aggregate Asset Supply and Aggregate Labor Supply</td>
<td>124</td>
</tr>
<tr>
<td>B.2.5  Algorithm for Transition Dynamics</td>
<td>125</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

1.1 Performance Measure for Transformer Bi-LSTM and Naive Bayes Classifier ........................................ 27  
1.2 Summary Statistics .......................................................... 30  
2.1 Matching Wealth Distribution in 1967 as a Steady State ............ 71  
3.1 EER and IER Comparison under Fixed Poverty Rate .................. 96  
3.2 EER and IER Comparison under Fixed Poverty Line ................. 97  
A.1 Financial Constraints and Relation to Proxies ....................... 106  
A.2 Summary Statistics of Federal Funds Future Shocks .................. 110  
A.3 Correlations with Other Measures of Financial Constraints .......... 111
LIST OF FIGURES

1.1 Attention-Based Bidirectional LSTM Model .......................................................... 19
1.2 Word Similarity in Annual/Quarterly Report ......................................................... 20
1.3 Liquidity Ratio Across Level of Financial Constraints ............................................. 33
1.4 Distribution of Leverage Ratio within Constrained Firms ....................................... 34
1.5 Dynamics of Heterogeneous Investment Responses to Monetary Policy Shocks ........... 40
1.6 Naive Bayes: Dynamics of Heterogeneous Investment Responses to Monetary Policy Shocks ......................................................... 45
1.7 Dynamics of Heterogeneous Debt Responses on Monetary Policy Shocks ............... 48

2.1 Labor Earnings Process ......................................................................................... 67
2.2 Imputed Marginal Tax Rate .................................................................................... 69
2.3 Evolution of Top Wealth Shares from 1967-2010 .................................................. 73
2.4 Evolution of Top Income Shares from 1967-2010 ................................................ 75
2.5 Evolution of Volatility of Log Earnings from 1967-2010 ........................................ 76
2.6 Evolution of Top Wealth Share under Optimal Linear Tax ..................................... 79
2.7 Marginal Tax Rate under Optimal Non-linear Tax System ...................................... 82
2.8 Evolution of Top Wealth Share under Optimal Nonlinear Tax ................................. 83

A.1 Heterogeneous Investment Dynamics using Firms with Investment Spells \geq 40 Quarters .......................................................... 107
A.2 Heterogeneous Investment Dynamics with Alternative FFR Shocks ........................ 108
A.3 Heterogeneous Investment Dynamics with Alternative FFR Shocks ........................ 109
A.4 Distribution of Annual/Quarterly Reports .......................................................... 112
A.5 Federal Funds Future Shocks ............................................................................... 113
A.6 Quarterly Monetary Shocks: Weighted vs Simple Sum ......................................... 114
CHAPTER 1
FINANCIAL CONSTRAINTS, MONETARY POLICY AND FIRM
INVESTMENTS: A TEXT-BASED DEEP LEARNING APPROACH

1.1 Introduction

It is of general interest for both policy makers and researchers to understand the role of financial constraints in the effects of monetary policy on firm investments. Not only it is crucial to recognize what types of firms are more influenced by monetary policy, particularly after the financial crisis, but the aggregate efficacy of the investment channel of monetary policy may depend on the fraction of financially constrained firms as well. However, there is no consensus on how and why financial constraints matter for the effects of monetary policy.

The financial accelerator mechanism (Bernanke et al. (1999)) implies that in response to an expansionary monetary policy shock, the collateral value of a firm increases, which directly relaxes the borrowing constraints. As a consequence, financially more constrained firms will be more responsive to monetary policy. On the other hand, financially more constrained firms often face a higher marginal cost of external financing. As a result, they are less responsive to monetary policy due to high marginal cost. Empirical studies typically analyze the role of financial constraints in the effects of monetary policy by using proxies of financial constraints, such as age, size, dividend payment activity, leverage ratio
and liquidity ratio. Their findings differ depending on which proxy they adopt.

My paper addresses the empirical question of which firms’ investments are more responsive to monetary policy shocks given the financial heterogeneity. Instead of using proxies as previous empirical papers, I construct a measure of financial constraints with SEC filings using a deep learning model. The deep learning model raises the prediction accuracy compared to a traditional natural language processing model by capturing the context information, such as grammatical structure and order of words. The firm characteristics of more constrained firms under my measure are younger, smaller, have a higher leverage ratio, a higher liquidity ratio and high Tobin’s q, which suggest they are high-growth firms that desire external funds but face difficulties. However, I find that popular proxies, like firm age and liquidity ratio, often do not move monotonically with the level of financial constraints. This implies that interpretation of the results using proxies as financial constraints should be more careful. Finally, I show that more constrained firms are persistently less responsive to monetary policy due to high marginal cost of external financing. I also contend that the divergent findings in previous papers on the role of financial constraints in the effects of monetary policy can be explained by the non-monotonicity of proxies with the level of financial constraints. It implies that, in a more general settings, firms which rely more on internal funds for investments, such as most constrained firms and most unconstrained firms, are less responsive to monetary policy. In addition, during liquidity shortage crisis, the effectiveness of monetary policy is weaker due to a larger fraction of severely constrained firms, which face
extremely high external cost of capital.

To study the question, my first step is to construct a textual measure of financial constraints by evaluating the information that managers convey in the annual/quarterly reports about their financial conditions. The advantage of exploring text data is that it examines how managers perceive their previous financial conditions without making any assumption of the relationship between the financial constraints and their proxies. To do so, I first construct the training sample by utilizing three constrained words lists suggested by Hoberg and Maksimovic (2014) to determine the most constrained firm-quarter observations and decide on the “unconstrained” firm-quarter observations. By construction of the training sample, financially constrained firms are defined as firms which have robust investment opportunities but face high marginal cost of external financing. I then train a deep learning learning model (Transformer Bi-LSTM) that has been proven to be quite successful in text classifications in recent years on my training sample and score the rest of the data set. As shown later, the deep learning model captures the grammatical structure and order of words, and thus provides a better estimate of financial constraints compared to a Naive Bayes method, which is a traditional language model and also used in Buehlmaier and Whited (2018) to estimate financial constraints.

Under my measure of financial constraints, constrained firms are smaller, younger, have a higher leverage ratios, a higher liquidity ratio, higher Tobin’s q and higher investment rates, which shows the constrained firms are high-growth
firms with a desire for external funds. However, I find that popular proxies often do not change monotonically with the level of financial constraints such as age, size and liquidity ratio. Specifically, the most constrained firms and the least constrained firms tend to have higher a liquidity ratio, though the reasons are different. High liquidity ratio for the least constrained firms, as argued in Jeenas (2018), show that as firms grow out of financial constraints, they reduce the need for external funds and use more liquid assets to finance their investment needs. On the other hand, the most constrained firms also possess a higher liquidity ratio due to precautionary saving motives and high marginal cost of external funds. The fact that many proxies do not change monotonically with financial constraints implies that the interpretation of empirical results using proxies should be done cautiously.

Lastly, with the constructed measure of financial constraints, I employ a panel version of the Jordà (2005)-style local projection method and estimate the investment responses of more constrained firms upon a monetary policy shock identified with the high frequency identification approach. I find that the investments of more constrained firms respond persistently less upon a monetary policy shock. For firms that are one standard deviation more constrained, the investments respond initially less by about 0.2 percentage points and cumulatively less by about 1.5 percentage points at the 10th quarter upon a one standard deviation expansionary monetary policy shocks. In addition, I find that debts of more constrained firms also respond persistently less to monetary policy shocks, thereby providing supporting evidence that constrained firms face high
marginal cost of external funds, and as such, they are less responsive to monetary policy shocks.

The baseline result is consistent with Ottonello and Winberry (2018), though they find such sluggish investment responses of more constrained firms are short-lived for one period using the leverage ratio as a proxy while I establish that it is quite persistent. On the other hand, the baseline result seems to contrast with Jeenas (2018), where he argues that more constrained firms are more responsive to monetary policy shocks using negative liquidity ratio as proxy. I posit that because of the non-monotonicity of the liquidity ratio with the level of financial constraints, his empirical findings that firms with higher a liquidity ratio are less responsive to monetary policy speak to both the most constrained firms and the least constrained firms, given their higher liquidity ratio relative to mid-constrained firms. As a result, his empirical findings are, to some extent, consistent with my findings where constrained firms that have higher a liquidity ratio are less responsive to monetary policy. Using the liquidity ratio as a proxy will miss the fact that highly constrained firms also have a high liquidity ratio. As a result, I argue that the seemingly contrary findings in previous papers are, to some extent, consistent with each other and with my paper.

My findings shed light on the discussion of the effectiveness of monetary policy across firms. While most previous papers focus on the comparison between unconstrained big firms, which have a lot of cash, and constrained firms, which still rely on external funds, my paper focuses on the comparison between most
constrained firms and average constrained firms where both types are somewhat constrained. These results are consistent in a sense that firms that rely more on internal funds, which includes both the unconstrained firms and the most constrained firms, respond less to monetary policy. Thus the responsiveness of aggregate investment to monetary policy will depend on both the fraction of unconstrained firms and the most constrained firms. It also implies that during crisis time when liquidity supply dries up, monetary policy will be less effective due to a much larger fraction of severely financially constrained firms.

1.1.1 Related Literature

This paper contributes to several strands of literature.

First, there is an empirical literature in corporate finance that focuses on providing a measure of financial constraints. Earlier works focus on using accounting data to estimate the financial constraints, including Kaplan and Zingales (1997), Lamont et al. (2001), Whited and Wu (2006), and Hadlock and Pierce (2010). These kinds of measures that employ accounting data have been challenged on their ability to capture financial constraints by Farre-Mensa and Ljungqvist (2016), where they test whether financially constrained firms would be able to raise debt when there is an exogenous state tax reform that increases the incentive of issuing debt. On the other hand, several recent papers use text data to capture financially constrained firms or provide a text-based measure of
financial constraints. Bodnaruk et al. (2015) use the frequency of constraining words in the entire annual report to assess the level of financial constraints. Hoberg and Maksimovic (2014) utilize three constrained word lists to determine the annual reports that are investment-constrained, debt-constrained or equity-constrained and then calculate the document similarities for other reports to provide three measures of financial constraints. Pitschner (2017) use pattern search to identify the most liquidity-constrained firms during normal and crisis times. Buehlmayer and Whited (2018) first build a training sample of constrained and unconstrained firms and then train a Naive Bayes Classifier to predict the level of financial constraints. My work also concentrates on using text data to provide a measure of financial constraints. It contributes to the literature by estimating the textual measure of financial constraints using a deep learning model to increase the prediction accuracy by capturing the context information. And context information is indeed important to predict the level of financial constraints, whereas previous papers using text data fail capture the context information. In addition, compared to Bodnaruk et al. (2015), Hoberg and Maksimovic (2014) and Buehlmayer and Whited (2018), I extend the measure to a quarterly frequency by using both quarterly reports and annual reports.

Second, there is an empirical literature in macroeconomics that focuses on studying the role of financial constraints with respect to the effects of monetary policy on firm investments using proxies of financial constraints including firm size (Gertler and Gilchrist (1994)), age and non-dividend payment (Cloyne et al. (2018)), leverage ratio (Ottonello and Winberry (2018)) and liquidity ratio (Jeenas
While Gertler and Gilchrist (1994), Cloyne et al. (2018) and Jeenas (2018) observe that more constrained firms are more responsive to monetary policy, Ottonello and Winberry (2018) contend that more constrained firms are less responsive to monetary policy. Compared to these papers which use proxies of financial constraints, my work contributes to the literature by leveraging a textual measure of financial constraints instead of proxies. My result is consistent with Ottonello and Winberry (2018) that more constrained firms are less responsive to monetary policy shocks, though the effect is more persistent in my paper than theirs. In addition, I argue that based on the non-monotonicity of some proxies with the level of financial constraints, the previous seemingly contrary results are, to some level, consistent with each other and with my paper. One should be cautious using proxies that may not change monotonically with the level of financial constraints. I also argue that all these results are consistent in a sense that firms which lean on more internal funds for investments are less responsive to monetary policy and thus my paper reconcile previous empirical findings among them and with mine.

Lastly, the policy implication of my paper is related to a literature that suggests that monetary policy is less effective during recessions. Thoma (1994), along with Tenreyro and Thwaites (2016), empirically show that monetary policy is less powerful during recessions than expansions with a time-series model. Vavra (2013) construct a model where monetary policy is less effective during recessions based on the changes in the price change dispersion. Ottonello and Winberry (2018) indicate that changes in the distribution of default risk may be one reason
monetary policy may be less effective during economic downturns. My paper contributes to the literature by suggesting that the changes in the distribution of financial constraints may be one of the explanations of why monetary policy may be less effective during recessions.

**Road Map** The rest of the paper is organized as follows. Section 1.2 describes the text data used to construct the financial constraints index, the firm level data from Compustat and the identification of exogenous monetary policy shocks. Section 1.3 shows the construction of financial constraints index and discuss how the deep learning model can improve the prediction accuracy. Section 1.4 provides the summary statistics of the measure and discusses the non-monotonicity of popular proxies with level of financial constraints. Section 1.5 shows my baseline results and robustness checks. Section 1.6 concludes.

### 1.2 Data

To construct a quarterly financial constraints index and connect it to the firm level data, I use two sources of data: EDGAR from SEC and Compustat Monthly Updates - Fundamentals Quarterly. EDGAR is the primary system for companies and individuals who are required to file information with SEC to submit their documents. For the EDGAR file, I download all the cleaned annual reports (10-K) and the quarterly reports (10-Q) from 1997q1 to 2016q4. Specifically I use the
“stage one parsed file” from where Loughran-McDonald’s website where they remove all the redundant characters such as HTMLChars, ASCIIEncodedChars, XMLChars, etc and include summary information at the beginning of each file. For each firm in a given fiscal year, it files three quarterly reports and one annual report.  

Following Buehlmaier and Whited (2018), I focus on the Management’s Discussion and Analysis section in each quarterly/annual report. Management’s Discussion and Analysis (MD&A) is the part of report where firm will review its past performance, financial conditions and project future performance. SEC requires firms to discuss their liquidity needs and issues in MD&A which is directly related to financial constraints. This is different from Bodnaruk et al. (2015) which focus on the entire 10-K and Hoberg and Maksimovic (2014) which focus on a particular subsection within MD&A. For quarterly report, firms are required to review the past performance of the past quarter while for the annual report, firms are required to review the performance in the past 12 months. To measure the financial constraints of the last fiscal quarter for each firm, I use the content of annual report to construct the index. My assumption of using the annual report as the last fiscal quarter is that firm’s liquidity condition in the last

1The types of annual reports I include are 10-K, 10-KSB, 10-K405 and 10KSB40 following Hoberg and Maksimovic (2014) and the type of quarterly reports I includes are 10-Q and 10QSB. 10-K is the standard annual report required by SEC. Until 2009, 10-KSB is filed by small business that is lack of interest from equity researchers and often is referred as penny stock. 10K405 is the annual report similar to 10-K but the company failed to disclose their insider trading activities within a given period of time. 10KSB40 is the annual report for small business that also have a positive check on item 405. Similarly, 10QSB is the quarterly report for small business and 10-Q is for the rest. See Appendix figure A.4 for a distribution of different file types over time.
fiscal quarter should heavily influence firms’ discussion on financial constraints in
its annual report, though it is required to review all changes of financial conditions
for the past 12 months. Therefore I obtain a list of quarterly frequency MD&A
sections.

For firm level data, I use Compustat Monthly Updates - Fundamentals Quarterly which ranges from 1997q1 to 2016q4. Following the exclusion criteria in Whited and Wu (2006), I excludes the financial industry which has SIC code ranging from 6000-6799 and the utility and transportation industry with SIC code ranging from 4000-4999. To avoid coding errors, I, following Buehlmaier and Whited (2018), deleted firms with total debt smaller than short term debt; firms with an ongoing merger that accounts more than 15% of the book value of its asset; firms with negative or zero total assets, book equity or sales. I then merge the Compustat data with the EDGAR data using the CIK number.

1.3 Measure of Financial Constraints

1.3.1 Construction Strategy

To construct a measure of financial constraints using firm annual/quarterly reports, I pursue the following four steps: (1) extract the MD&A (Management’s Discussion and Analysis) section from each annual/quarterly report, (2) construct a novel training sample that identifies the most constrained firms based on three
constrained-words lists and decide on the unconstrained firms, (3) train my deep learning model (Transformer Bi-LSTM) on the training set and evaluate its performance on the validation set and (4) predict the level of financial constraints for the entire database. I discuss this in more detail below.

1.3.2 MD&A Extraction

First I obtain all the annual/quarterly SEC filings from Loughran-McDonald “Stage One 10-X Parse files” from 1998q1 to 2017q4. Then I extract the entire MD&A section from each annual/quarterly report. MD&A section attempts to give a balanced view of the company’s past and upcoming future performance through the eyes of the company’s management team. It typically contains two parts: Liquidity and Capital Resources and Results of Operations where managers discuss changes of capital structure and changes of real operations respectively. I use the entire MD&A section to construct my index which is similar to Buehlmaier and Whited (2018) while Hoberg and Maksimovic (2014) consider only a subsection of MD&A, Liquidity and Capital Resources. As noted by Buehlmaier and Whited (2018), firms often discuss their issues related to financial constraints outside Liquidity and Capital Resources subsection, thus analyzing the entire MD&A section would capture more information related to firms’

---

Due to low quality of text files, the number of annual/quarterly reports are low prior 1998. Thus I pick the years with more steady number of annual/quarterly reports. Link: https://sraf.nd.edu/textual-analysis/resources/
financial conditions. ³

1.3.3 Training Sample

To predict the level of financial constraints for each firm-quarter observation, I need to construct a training sample to train my model. The key challenge is to identify the most financially constrained firms and I take advantage of three constrained-words lists from Hoberg and Maksimovic (2014). Their constrained-words lists (delay list, debt-focused list and equity-focused list) are used to identify three types of constrained firms: firms which delay in investments, firms which have a need of debt financing and firms which have a need of equity financing. In practice, I find that there are other factors such as weak balance sheet that will also result in a delay of investments besides the inability of raising external funds; firms which mention a need to raise external funds through either debt or equity might not indicate that they are financially constrained, but sometimes, conversely, indicate their capability to raise external funds.

To accurately capture the most constrained firms, I follow a similar strategy in Buehlmaier and Whited (2018) by combining all three lists⁴. For a firm-quarter observation to be classified as financially constrained, it must mention a delay in their investment plan and that is related to debt financing or equity financing.

³See appendix for the distribution of annual/quarterly reports across years. I am able to cover up to 87% of the annual/quarterly reports from 1997 to 2018.

⁴Buehlmaier and Whited (2018) use the lists to construct two measures on debt constrained firms and equity constrained firms respectively.
difficulties. More specifically, a document will have a financial constraints count if it mentions a delay of investments in one sentence and then discuss debt financing or equity financing need in the window sentences (same sentence, one sentence above or below). Thus, the definition of financially constrained firms in my paper are firms which have good investment opportunities but face very high marginal cost of borrowing. Overall, about 6.97% of firm-quarter observations have positive counts\(^5\). To get rid of measurement errors, I select the top 500 firms which have the highest financial constraints count as the most constrained firms. As for the “unconstrained firms”, I choose 1000 firms randomly from the zero-count samples.\(^6\) Thus, my financial constraints measure provides a relative measure whether the financial condition of a firm at a specific quarter is closer to the average firms or the most constrained firms.

To get a sense of how good the training sample is picking up the most constrained firms, here are two examples of the constrained firm-quarter observations in my training sample. The first one is a small electronic manufacturing firm stated in their 2006q3 report:

“We must ...... raise approximately $ 250 million in cash through the issuance of debt ... if we failed ...... , we may be force to delay , alter or abandon our planned expansion .”
- Hoku Scientific inc, 2006q3

\(^5\)See appendix for details on how I count the occurrence of words from the three constrained-words lists.

\(^6\)The final financial constraints measures do not variate much with different sets of randomly selected firms.
Here the managers of the company clearly indicate that the potentially failure of raising debt would result in a delay of investment plan. The second one is a larger oil and gas company states in their 2009q2 report:

“The continued weakness in global financial markets has adversely impacted our ability to fund a growth oriented capital spending program in 2009. ...... we have limited our capital spending in 2009 to more closely mirror internally generated cash flow.” - Penn Virginia Corp, 2009q2

Here the managers point out that the weakness of global financial market will dampens their investment plan and as a result, they will have to cut back on their capital spending program. These two examples demonstrate that the training sample captures well with the most constrained firms.

After constructing my training sample, I label the most constrained firms as 1 and the average firms as 0. Finally, 25% of the documents in the training sample are taken to the validation set to evaluate the out-of-sample prediction of my deep learning model and other language models for comparison. Overall, I have 1125 firm-quarter observations in the training set and 375 firm-quarter observations in the validation set. In the appendix, I also discuss why using financial constraints counts directly to measure the level of financial constraints is not a better method compared to using language model in general.
1.3.4 Predicting Model - Transformer Bi-LSTM

Using the constrained-words lists enable me to identify the small fraction of the most constrained firms, but it gives no information about the level of financial constraints for other observations since 93% of the observations has zero financial constraints count. However, I could utilize the training sample to train a language model that learns the differences in the MD&A section between the constrained firms and the unconstrained firms and predict the level of financial constraints for the observations not in the training sample.

The language model I adopt is a deep learning model called Transformer Bidirectional-Long Short Term Memory model. It belongs to a class of deep learning models that have been proved to be particularly successful on text understanding and text classification in recent years in natural language processing area (for example, Zhou et al. (2016), Vaswani et al. (2017), Yang et al. (2017)).\(^7\) Compared to the traditional language models that take the frequency counts of words as input (the bag-of-words approach), the deep learning model takes into account the connections between words and thus preserves the syntactic and semantic information such as the grammatical structure and orders of words. If context information ever matters for the prediction of financial constraints, theoretically my deep learning model will perform better and I will show later that it indeed matters.

\(^7\)Recent developments in NLP that achieve breakthroughs in question answering such as BERT (Devlin et al. (2018)) and related models are built on the foundation of Transformer model.
For a typical neural network structure, it contains three types of layers. The input layer contains the raw data which is the entire MD&A section in my case. The output layer is the layer that predicts the output which is the level of financial constraints. The hidden layers are the intermediate steps that transform the raw data into something that output layer can use to generate output and there can be multiple hidden layers. The Transformer Bi-LSTM model has two key hidden layers that are particularly useful to capture the grammatical structure and the orders of words that are useful to predict the level of financial constraints.\(^8\)

The first one is Bi-directional LSTM neural layer which reads through each document in both sequential and reverse orders. Take the sequential order for example, each time it reads through a new word, it incorporates the past information from all previous words and the information of current word and decides how much old information to forget for the previous words and how much new information to incorporate with the new word. Together, it preserves context information such as information related to grammatical structure (semantic structure) and the order of words (syntactic structure) that are useful for prediction. Similar case applies when it reads the document in a reverse order. In my case, it will be able to differentiate, for example, whether firms face difficulty of raising external funds or they actually do by the ability to raise enough external funds.

\(^8\)In the NLP literature, such models are often described as having an encoder-decoder structure where encoder is the mechanism of transforming and storing useful information and a decoder converts such information into a corresponding output. The structure gets its name in a seq-to-seq task such as translation. Here I only need to predict one output which is the level of financial constraints. So I abstract away from using the terminology for a clearer illustration.
The second layer is the Transformer layer. Intuitively, it provides several sets of weights on each word which tell the model to focus only on the parts that are related to financial constraints. This mechanism works well with the Bi-LSTM layer which sometimes will encounter difficulties of preserving useful information when the text is long. The average length of a MD&A section is indeed very long with about 6950 words. Thus, the transformer layer is very helpful to keep the model to stay on the most relevant parts for predicts and reduce the noise from redundant information. Together with this two important layers, the deep learning model suits well with my prediction task of financial constraints. Below is a detailed discussion of the model.

**Transformer Bi-LSTM Model**

Transformer Bi-LSTM model contains a total of five layers and the structure is shown in figure 1.1. A layer in neural network is a collection of “nodes” operating together where nodes can be words or vectors.
Figure 1.1: Attention-Based Bidirectional LSTM Model

Notes: This figure graphs all five layers of the deep learning model that I use including input layer, word embedding layer, Bi-LSTM layer, transformer layer and output layer.

**Word Embedding Layer**  To convert the document into useful features that a model could use to make predictions, traditional language models tend to take the frequency counts of words as inputs but it throws away the context information. This step is called feature extraction. For recent language models that use neural networks more often and still want to preserve the context information, one first step is to project the words in the document into a set of vectors that represents word similarities. Then they could extract more information based on the set of vectors. Typically, this first step is called word embedding. Below is the detailed steps of obtaining the word embedding vectors using my text data.
Figure 1.2: Word Similarity in Annual/Quarterly Report

Note: The graph shows word similarity in a two dimensional space for common words in annual/quarterly reports. It is obtained by first conducting a principal component analysis on the word embedding vectors and plotting out the first two most important dimensions.

In the first input layer, for each MD&A section, I treat all the words as a sequence of input \((x_1, x_2, \ldots, x_T)\). For each word \(x_i\), I convert it into its corresponding word embedding vector \(e_i\) and obtain the embedding layer.
(e_1, e_2, ..., e_T). Word embedding projects each word into a vector representation and the dot product of two word embedding vectors gives the similarity between two words. Here we used the pre-trained word embedding vectors from GloVe (Pennington et al. (2014)) that is trained on a large corpus. Specifically, I use the pre-trained word embedding vectors that are trained on Common Crawl which crawls web text and contains 840 billion word tokens with each vector with dimensions of 300. In practice of other NLP applications, word embedding layer is shown be able to capture some levels of semantic and syntactic similarity to other words. Below I show that, word embedding vector provides a good generic measure of word similarity for common words related to financial constraints that are used in annual/quarterly reports.

Figure 1.2 shows the word similarity by conducting a principal component analysis on the word embedding layers and plotting out the first two dimensions. It captures well not only with word similarities such as *limit*, *delay* and *cancel* which are close to each other, but with semantic meanings such as *monetary, fiscal* and *policy* that are also close to each other as well. It is a good first step processing of the word such that its information is ready to be used to extract more features.

**Bi-LSTM Layer** Once the word embedding vectors for each word are obtained, I need a model structure to take this features as input and preserve useful information about the grammatical structure and the order of words that are related to financial constraints. As discussed earlier, the Bi-LSTM layer will serve
such purpose.

The Bi-LSTM layer takes the word embedding vectors as input and read in a sequential order (from $x_1$ to $x_T$) and reverse sequential order (from $x_T$ to $x_1$) respectively and obtain two sets of output vectors that preserve context information, $(\vec{h}_1, \vec{h}_2, ..., \vec{h}_T)$ and $(\overrightarrow{h}_1, \overrightarrow{h}_2, ..., \overrightarrow{h}_T)$. $\vec{h}_i$ is a vector that represents the useful information stored for word $i$ and all previous words and $\overrightarrow{h}_i$ represents the useful information stored for word $i$ and all post words.

LSTM model stands for the long-short term memory model which is one of the recurrent neural networks (RNN) that read through the document in a particular order.\textsuperscript{9} To effectively remember important information in the text data, a standard LSTM model typically is composed of a cell (denoted as $C_t$), an input gate, an output gate and a forget gate. The cell $C_t$ would remember the information preserved up to word $t$. The three gates control the flow of information in and out of the cell. Each time a new word embedding vector $e_t$ is fed into the model, the cell $C_t$ is updated with new information from $e_t$ while selectively preserving old information from cell $C_{t-1}$. It generates an output hidden vector $\vec{h}_t$ which contains information of current word and all previous words and is passed to the next cell which would help to generate $\vec{h}_{t+1}$. For illustration, $\vec{h}_5$ is the output vector that contains information from the first word to the fifth word and $\vec{h}_T$ contains information for the entire documents. Similar illustration for $\overrightarrow{h}_i$ but in reverse order.

\textsuperscript{9}See a great detailed explanation by Colah in https://colah.github.io/posts/2015-08-Understanding-LSTMs/
**Transformer Layer** In Bi-LSTM neural layer, $\vec{h}_t$ and $\vec{h}_t$ store all useful information of the entire documents. If qualities of the stored information in $\vec{h}_t$ and $\vec{h}_t$ are good enough, I could use them to directly predict the level of financial constraints. However, in practice, the output vector that is supposed to store all useful information in the LSTM model store too much information of the newest information but forget the old information fast. As sentence/document gets longer, the model performance drop substantially. In the NLP literature, the transformer layer is designed to resolve such issue and achieve great success.

Transformer layer takes all output vectors from the Bi-LSTM model, $(\vec{h}_1, \vec{h}_2, ..., \vec{h}_T)$ and $(\vec{h}_1, \vec{h}_2, ..., \vec{h}_T)$, and effectively assign several sets of weights on each output vector and use the weighted sum of the output vectors to predict the level of financial constraints. The idea of transformer mechanism is simple: not all parts in the documents are relevant to predict financial constraints and some of them are particularly important. By assigning higher weights on important vectors, it could achieve a much higher accuracy in predicting the level of financial constraints. By having multiple sets of weights, each set of weights would be able to capture a different aspect of the documents. For example, one set of weights focus on investment related issues and another sets of weights focus on the financial difficulties.\(^\text{10}\)

The transformer mechanism has achieved great success in language

\(^{10}\text{For illustration purpose, I abstract away from diving into details of the transformer model such as the multi-head and self-attention mechanism. See Vaswani et al. (2017) for the original paper or https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html for a great intuitive explanation.}\)
understanding in the past few years (Vaswani et al. (2017)), especially in seq2seq task (both input and output are sequences such as language translation). Given the nature that my inputs are sequences of words and the prediction requires text understanding, the overall model seems to be a good fit for my task.

**Naive Bayes Classifier**

For comparison, Naive Bayes Classifier is a traditional, simple and fast language model and used in Buehlmaier and Whited (2018) to predict financial constraints level. It is based on Maximum A Posteriori that models the level of financial constraints as a function of the word counts in each MD&A section with additional assumption that each word’s appearance is independent from each other conditional on the level of financial constraints.\(^{11}\) By design, Naive Bayes Classifier throws away the context information and focus on the information of word frequency counts. However, context information does matter for the prediction and as a result, the deep learning model performs better. In addition, in practice, Naive Bayes method serves as a better classifier and less of importance on providing a score because omitting the context information will often result in an imprecise measure, although the broad classification is correct.

Why Order of Words Matters

In my own practice, I found that Naive Bayes method over-predicts the amount of constrained firms relative to the deep learning model. Here is one example that Naive Bayes method predicts as financially constrained with high score 0.98 whereas the deep learning model predicts as unconstrained with a score 0.01:

“We believe we could fund other acquisitions and repurchase common stock with our internally available cash and investments, cash generated from operations, amounts available under our credit facilities, additional borrowings or from the issuance of additional securities.” - Oracle 2006q1

The managers in Oracle clearly indicate that their capacity of funding other acquisition projects with available external funds is adequate. Naive Bayes method will not be able to distinguish between “could fund” and “could not fund” and thus incorrectly predicts that the firm-quarter observation is financially constrained whereas the deep learning model correctly picks it up. This is a good example demonstrating the importance of context information such as order of words to predict the level of financial constraints. Omitting it could result in the opposite direction of prediction and add additional noise to the measure.

Performance Evaluation on Validation Set Once the model is trained, I evaluate its out-of-sample performance using the validation set. The firm-quarter
observations in the validation set are not used in the training process and thus serve well to evaluate model performance. For comparison, I also test the performance of the Naive Bayes method in the validation set.

Figure 1.1 shows the comparison of precision, recall and f1-score between Transformer Bi-LSTM model and Naive Bayes model in the validation set. Precision, recall and f1-score are the common statistics in machine learning literature that are often used to evaluate the out-of-sample performance of different models. The formulas are:

\[
\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}, \quad \text{Recall} = \frac{\text{True Positive}}{\text{Total Actual Positive}}, \quad f1 - \text{score} = 2 \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.
\]

Precision means out of all predictions, what are the percentages are correct. Recall means out of all the actual positives, what are the percentages are actually predicted. F1-score is just a simple harmonic mean of precision and recall which gives the overall performance of the model with one metric. I am taking macro-average which means that I first calculate the precision, recall and F1-score for both constrained firms and unconstrained firms and take the average across groups. It is important to look at macro-average because I have an unbalanced training set with more unconstrained observations but I cares more about predicting the constrained firms correctly.

The result in figure 1.1 indicates that Transformer Bi-LSTM model does the better than Naive Bayes method in all three metrics. Overall, it has 6% improvement in f1-score compared to Naive Bayes Classifier. The biggest improvement comes from the precision of prediction which is consistent with
previous argument: Naive Bayes method tends to over-predict the amount of constrained firms and results in a lower precision. This is also reflected in the recall rate where Naive Bayes performs well indicating that over-predicting would enable the model to cover more truly constrained firms. Overall, I conclude that the Transformer Bi-LSTM model outperforms the Naive Bayes in the validation set. Given its advantage of preserving the context information, it provides more accurate measure of financial constraints. In the natural language processing literature, a 6% improvement in f1-score is a significant improvement and given the large size of my dataset, it is an important improvement. I will also demonstrate in later section that using Naive Bayes will lead to insignificant results due to the noise it adds to the measure of financial constraints.

Table 1.1: Performance Measure for Transformer Bi-LSTM and Naive Bayes Classifier

<table>
<thead>
<tr>
<th></th>
<th>Transformer Bi-LSTM</th>
<th>Naive Bayes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Macro Precision</td>
<td>0.92</td>
<td>0.81</td>
</tr>
<tr>
<td>Recall</td>
<td>0.86</td>
<td>0.83</td>
</tr>
<tr>
<td>F1-Score</td>
<td>0.88</td>
<td>0.82</td>
</tr>
</tbody>
</table>

Notes: This table shows the model performance on the validation set for both the deep learning model (Transformer Bi-LSTM) and the Naive Bayes method. Precision is the measure of prediction accuracy. Recall measures among the truly constrained firms, what percentage are correctly covered. F1-score is the simple harmonic mean of precision and recall.
1.4 Characteristics of The Financial Constraints

1.4.1 Summary Statistics

Once the measures of financial constraints of all documents are obtained by applying the trained deep learning model to them, I want to conduct a sanity check to see if the characteristics of the constrained firms under my measure are consistent with the characteristics that are typically associated with constrained firms in the literature.

Table 1.2 shows the summary statistics after I merge the financial constraints index with Compustat quarterly data. I divided up firms evenly according to their financial constraints index as highly financially constrained, medium financially constrained and low financially constrained. From the summary statistics, constrained firms under my measure are younger (Cloyne et al. (2018)), smaller (Gertler and Gilchrist (1994)), have higher leverage ratio (Ottonello and Winberry (2018)), lower dividend payments (Cloyne et al. (2018)), higher EBITDA to total asset ratio (which is an alternative way of measuring cash flows), hold more cash (Han and Qiu (2007)) which is consistent with the precautionary saving motives, have higher investment intensity, higher R&D intensity and higher Tobin’s q which reflects good investment opportunities. This results are consistent with the situations where young, small constrained firms which have very good investment opportunities cannot fund themselves through external funds easily.
but instead, need to hoard cash given poor current cash flow conditions. Overall, the characteristics of constrained firms in my financial constraints index are generally consistent with the literature. \(^\text{12}\)

\(^{12}\)The results are generally consistent with the findings in Buehlmaier and Whited (2018), but they only provide the summary statistics for the training sample instead of the whole data set. In addition, they also separate constrained firms into three categories as generally constrained, debt constrained and equity constrained.
<table>
<thead>
<tr>
<th></th>
<th>Low FC (Bottom 33%)</th>
<th>Mid FC</th>
<th>High FC (Top 33%)</th>
<th>Diff</th>
<th>t-stats</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log of Firm Age</td>
<td>2.5734</td>
<td>2.6212</td>
<td>2.2901</td>
<td>-0.2832</td>
<td>-45.07</td>
</tr>
<tr>
<td>Log of Total Asset</td>
<td>5.6409</td>
<td>5.9455</td>
<td>5.1972</td>
<td>-0.4437</td>
<td>-27.23</td>
</tr>
<tr>
<td>Debt/Total Asset</td>
<td>0.2385</td>
<td>0.2358</td>
<td>0.3360</td>
<td>0.0975</td>
<td>17.36</td>
</tr>
<tr>
<td>Dividend/Total Asset</td>
<td>0.0019</td>
<td>0.0021</td>
<td>0.0009</td>
<td>-0.0009</td>
<td>-24.17</td>
</tr>
<tr>
<td>EBITDA/Total Asset</td>
<td>0.0196</td>
<td>0.0234</td>
<td>-0.0019</td>
<td>-0.0215</td>
<td>-35.00</td>
</tr>
<tr>
<td>Cash/Total Asset</td>
<td>0.1532</td>
<td>0.1452</td>
<td>0.1984</td>
<td>0.0451</td>
<td>19.87</td>
</tr>
<tr>
<td>Liquidity Ratio</td>
<td>0.1919</td>
<td>0.1753</td>
<td>0.3031</td>
<td>0.1112</td>
<td>58.73</td>
</tr>
<tr>
<td>Investment Ratio</td>
<td>0.0023</td>
<td>0.0014</td>
<td>0.0034</td>
<td>0.0011</td>
<td>4.34</td>
</tr>
<tr>
<td>Capital Exp/Total Asset</td>
<td>0.0314</td>
<td>0.0318</td>
<td>0.0352</td>
<td>0.0038</td>
<td>10.23</td>
</tr>
<tr>
<td>R&amp;D/Sale</td>
<td>0.7932</td>
<td>0.6529</td>
<td>3.7931</td>
<td>2.9998</td>
<td>20.65</td>
</tr>
<tr>
<td>Tobin’s q</td>
<td>2.2950</td>
<td>2.0863</td>
<td>2.9720</td>
<td>0.6770</td>
<td>23.02</td>
</tr>
<tr>
<td>#</td>
<td>34,830</td>
<td>34,829</td>
<td>34,828</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: This table shows the summary statistics of firm characteristics sorted on my financial constraints measures. The first three columns show unconstrained, mid-constrained and constrained firms. Column 4 shows the t-statistics of the difference between unconstrained firms (average firms) and constrained firms. Liquidity ratio is measured as cash plus short term investments to total asset ratio. Data covers observations from 1999q1 - 2012q4.
1.4.2 The Relationship between Proxies and Financial Constraints

One of the most interesting findings from the summary statistics is that some of the proxies of financial constraints do not change monotonically with the level of financial constraints under my measure such as age, size and liquidity ratio.

Here I want to focus particularly on liquidity ratio. For liquidity ratio, both the least constrained firms and the most constrained firms tend to hold a higher liquidity ratio compared to mid-constrained firms, though the reason could differ. The differences in liquidity ratio across all three groups are pairwise statistically significant. Jeenas (2018) argues that negative liquidity ratio is a proxy for financial constraints because as firms grow out of financial constraints, they use more cash to fund their investment projects and have a higher liquidity ratio as a result. I find that constrained firms also have a high liquidity ratio. Due to the difficulty of raising external funds and high liquidity risks, constrained firms are forced to hoard liquid assets to fund their investment projects and hedge against liquidity risks which results in a higher liquidity ratio.

Figure 1.3 plots the mean of liquidity ratio for each decile of financial constraints and clearly indicates the U-shape in liquidity ratio across the level of financial constraints. The non-monotonicity in the liquidity ratio implies that using a single proxy from accounting data might not correctly pick up the constrained firms because the level of financial constraints does not always move monotonically along with that one dimension. Such findings justify the necessity
of using a composite measure of financial constraints such as my textual measure of financial constraints to avoid such issue. If one still wants to use a single proxy and sometimes it is important to understand the effects of monetary policy along one variable, one should consider estimating a non-linear function of financial constraints with that proxy.

The fact that age and size also don’t move monotonically with financial constraints is consistent with Hadlock and Pierce (2010) in which they estimate the financial constraints as a quadratic function of firm age and size. However, I also show in appendix table A.1 that proxies can only explain a small part of the variations in my measure of financial constraints.
Figure 1.3: Liquidity Ratio Across Level of Financial Constraints

Notes: The figure is a scatter plot of mean liquidity ratio in each decile of the measure of financial constraints connected with line segments. The U-shape indicates that both firms with low financial constraints firms and high financial constraints have higher liquidity ratio compared to firms with mid-financial constraints.

For leverage ratio, it is indeed increasing monotonically as the level of financial constraints increases, given there is no statistically significant differences between low constrained firms and mid-constrained firms. This is consistent with Ottonello and Winberry (2018) where they use leverage ratio as proxy. However, as argued earlier, popular proxies can only explain a small fraction of the variations in my measure of financial constraints which implies that using leverage ratio only will miss important variations in the financial constraints. In addition, as figure 1.4 plots the distribution of leverage ratio among the top 10%
financially constrained firms with the red line indicating Compustat median, it shows that there is a non-trivial amount of firms which are highly constrained, but also with lower than median leverage ratio. Ignoring them implies incorrectly identify them as financially unconstrained. Such findings point to the importance of identifying the constrained firms from multiple aspects instead of using a single proxy.

Figure 1.4: Distribution of Leverage Ratio within Constrained Firms

Notes: The figure plots the distribution of leverage ratio among the top 10% constrained firm in the data. The red line is the Compustat median. Data is winsorized at both 1% and 99% level.

In a similar spirit, Cloyne et al. (2018) identify young and non-dividend payment firms as financially constrained firms where they look at two dimensions
of the firm, though they will still face the issues that proxies do not always move monotonically with the level of financial constraints.

Overall, my composite measure of financial constraints using text data would be able to capture the non-monotonicity of firm characteristics and the multiple dimensions of constrained firms. Most importantly, I will show later that the non-monotonic feature in the liquidity ratio and the monotonic feature in the leverage ratio will reconcile the seemingly contrast findings in Ottonello and Winberry (2018) and Jeenas (2018) on the role of financial constraints in the effects of monetary policy.

1.5 The Effects of Monetary Policy Across Firms

Monetary Shocks

To identify the dynamic causal effects of monetary policy on firm investments, I need exogenous monetary policy shocks that drives investment responses instead of other macro factors that drives both monetary policy and investment responses. For the baseline model, I use monetary policy shocks identified with the high frequency identification approach which is pioneered by Cook and Hahn (1989), Kuttner (2001) and Cochrane and Piazzesi (2002) and is now standard in the literature. It is based on the identification assumption that during the short window of FOMC announcement, there is only the monetary shock that drives
the movement of the federal funds future rates and other macroeconomics shocks are muted. Explicitly I use the federal funds future surprises happened within 30 minute-window of FOMC ottonello2018financial from Nakamura and Steinsson (2018).

The federal funds future shocks are constructed as follows

$$\epsilon^m_t = \frac{D}{D-t}(f f r_{t+\Delta t^+} - f f r_{t-\Delta t^-})$$

where $D$ is the number of day in the month for that FOMC announcement. $t$ is the day when FOMC issues an announcement. $f f r_{t+\Delta t^+}$ is the federal funds future rate 20 minutes after the announcement and $f f r_{t-\Delta t^-}$ is the federal funds future rate 10 minutes before the announcement. $\frac{D}{D-t}$ adjusts for the timing of the announcement within the month. Thus an unanticipated increase of federal funds future rates means a contractionary monetary policy shock. The raw shock series ranges from 1997 to 2012. There are a total of 162 raw shocks.$^{13}$

To merge with my quarterly data, I time aggregate the high frequency shocks following Ottonello and Winberry (2018) by constructing a moving average of the raw shocks weighted by the number of days in the quarter after the FOMC meeting. The quarterly weighted monetary shock is constructed as

$$\epsilon^q_t = \sum_{t \in q} \omega(t) \epsilon^m_t + \sum_{t \in q-1} (1 - \omega(t)) \epsilon^m_t$$

$^{13}$See appendix for a summary statistics of federal funds future shocks.
where \( \omega(t) \equiv \frac{D_T^q - D_{fomc}^q}{D_T^q} \), \( D_T^q \) is the number of days in the quarter of a specific FOMC meeting and \( D_{fomc}^q \) is the day of the FOMC meeting in the quarter. Thus \( \omega(t) \) gives higher weights if the FOMC meeting happened earlier this quarter which gives a longer impact within the quarter. In another word, if I assume that the impact of each monetary shock lasts for a quarter length, the impact of the shock in the current quarter depends on how many days left within the current quarter, the rest of the impact will continue into the next quarter. For robustness, I also take the sum of the raw shocks within a quarter as Wong (2016).

For additional robustness check, I use monetary policy shocks identified with narrative approach in Romer and Romer (2004) where they regress the intended rate changes in FOMC on Greenbook forecasts of economic fundamentals and take the residuals as shocks. The identification assumption is that Federal Reserves responds endogenously to future economic fundamentals and what are left are the exogenous components of monetary policy.

1.5.1 Panel Local Projections with Firm Level Data

To estimate how financial constraints matter for the effects of monetary policy on firm investments, I employ a panel version of Jordà (2005)-style local projection method by regressing relative capital stock changes on the interaction term of financial constraints and monetary policy shocks, along with a set of controls.
The baseline specification is as follows:

\[
\log(k_{i,t+h}) - \log(k_{i,t-1}) = \alpha_i + \beta_{int}^h FC_{i,t-1} \times \epsilon^q_t + \Gamma X_{i,t-1} + \gamma d_{t+h} + e_{j,t+h} \tag{1.1}
\]

where \(k_{i,t+h}\) is the capital stock measured as tangible assets at time \(t\) plus horizon \(h\). \(\log(k_{i,t+h}) - \log(k_{i,t-1})\) is the cumulative investment response of firm \(i\) in time \(t+h\) relative to time \(t-1\). \(FC_{i,t-1} \times \epsilon^q_t\) is the interaction term of financial constraints measure of firm \(i\) at time \(t-1\) and monetary policy shock at time \(t\). \(X_{i,t-1}\) is a set of control variables at time \(t-1\) including firm’s financial constraints level, firm investment, real sales growth, log of firm age and log of total asset. In addition, I also control for the proxies such as the leverage ratio, liquidity ratio and their interactions with monetary policy shock at time \(t\), time fixed effect and firm fixed effect. All firm characteristics are at time \(t-1\) such that everything is fixed prior to the monetary policy shocks. By setup, given monetary policy shocks are time series measure that impact all firms, the average effect of monetary policy is absorbed by time fixed effects. Standard errors are clustered two-way on firms and quarters.

The coefficient of interests is \(\beta_{int}^h\) in front of the interaction term of financial constraints at \(t-1\) and monetary shocks at \(t\). It represents the cumulative investment response at horizon \(h\) for constrained firms in response to a monetary policy shock at horizon \(0\). I will estimate the coefficient of the interaction term for each horizon \(h = 0, 1, \ldots H\). Together, I will have the investment dynamics of constrained firms from horizon \(0\) to \(10\) quarters. To make the result more interpretable, I normalize the shock such that a positive \(\epsilon^q_t\) represents an
expansionary monetary policy shock. In addition, I scaled $\beta_{int}^h$ by one standard deviation monetary policy shocks and one standard deviation of my financial constraints measures. The scaled $\beta_{int}^h$ represents the percentage points response of investment for firms which are one standard deviation more constrained in response to a one standard deviation expansionary monetary policy shock.

1.5.2 Baseline Result of Investment Dynamics to Monetary Policy Shocks

Figure 1.5 shows the relative investment dynamics of more constrained firms in response to a monetary policy shock by plotting the scaled $\beta_{int}^h$ over horizon 0 up to 10th quarter with 95% confidence bands. It shows that investments of financially more constrained firms respond consistently less compared to unconstrained firms upon an expansionary monetary policy shock throughout all horizons.
Figure 1.5: Dynamics of Heterogeneous Investment Responses to Monetary Policy Shocks

Notes: The graph plots the scaled coefficients $\beta_h^{int}$ for each $h = 0, 1...10$ along with 95% confidence interval. The regression specification is $\log(k_{i,t+h}) - \log(k_{i,t-1}) = \alpha_i + \beta_h^{int} FC_{i,t-1} \times \epsilon_i^t + \Gamma X_{i,t-1} + \gamma d_{t+h} + \epsilon_{i,t+h}$. Confidence intervals are constructed based on two-way clustered standard errors at firms and quarters.

Quantitatively, upon a one standard deviation expansionary monetary policy shock at horizon 0, the investments of firms which are one standard deviation more constrained at horizon $-1$ will respond less by about 0.2 percentage point initially and up to 10th quarter, the cumulative investments increase less by about 1.5 percentage point. The result shows that the investments of financially more constrained firms are less responsive to monetary policy shocks and such heterogeneous responses are persistent and statistically significant for all periods.
It shows that the financial constraints of a firm consistently impact firm’s ability to respond to monetary policy. In addition, the results are robust controlling leverage ratio and liquidity ratio while the coefficients of their interaction terms with monetary policy shocks become insignificant.

**Comparison with Similar Studies using Proxies**  
I thus compare my findings to three recent papers, Ottonello and Winberry (2018), Cloyne et al. (2018) and Jeenas (2018), that also focus on the role of financial constraints on the effects of monetary policy. My result is consistent with Ottonello and Winberry (2018) that the investments of financially more constrained firms are less responsive to monetary policy shocks where they use leverage ratio as proxy for financial constraints. The key difference is that they found such heterogeneous responses are short-lived for one period while I found persistent heterogeneous responses up until even 10th quarter. On the other hand, Jeenas (2018) found that firms with low liquidity ratio which is a proxy for financial constraints are more responsive to monetary policy shocks. Cloyne et al. (2018) test various proxies of financial constraints and find that young, non-dividend paying firms are more responsive to monetary shocks and thus argue that constrained firms are more responsive.

Here I argue that my results are not necessarily contrary with Jeenas (2018) due to the non-monotonicity of liquidity ratio with the level of financial constraints. As shown in earlier discussion, high liquidity ratio picks up both the most constrained and the least constrained firms. Essentially, his empirical
results demonstrate that the groups of the most constrained firms and the least constrained firms are less responsive to monetary policy shock compared to mid-constrained firms. Such a result is consistent with my findings that more constrained which have a high liquidity ratio are less responsive to monetary policy shocks. Thus, I intend to reconcile the different findings in previous papers by arguing that they are, to some level, consistent with each other and with my result.

One drawback of my measure is that given the nature of the construction by choosing the average firms as the “unconstrained firms“, it will be hard to identify the most unconstrained firms in the data and thus it groups the most unconstrained firms with the average firms. Otherwise I could assess whether the most unconstrained firms are also less responsive to monetary policy.

A General Discussion on the Effectiveness of Monetary Policy Across Firms
Here I want to provide a more general discussion of the effectiveness of monetary policy. I contend that the findings of many previous papers and my paper are consistent under the framework that firms, which reply more on internal funds for investments, are less responsive to monetary policy. First, the empirical findings are not necessarily contrary to mine as discussed earlier, given the non-monotonicity of proxies. Second, the underling mechanisms they proposed are typically comparing unconstrained firms and the somewhat constrained firms. For example, Cloyne et al. (2018) argue, through the financial accelerator channel,
that the borrowings of the constrained firms are more subject to asset value fluctuations while Jeenas (2018) argues constrained firms are more influenced by monetary policy because funding source of investments for unconstrained firms comes mainly from cash. However, my paper aims at comparing the most constrained firms to the mid-constrained firms and thus speaking to a different type of mechanism.

These mechanisms are consistent if we focus on the funding source of firm investments. For unconstrained firms, it is optimal to use only internal funds for investments because internal funds, such as cash, have a lower cost compared to external funds and these firms have ample of cash. For the most constrained firms, it is optimal for them to rely on internal funds for investments due to high marginal cost of external capital. They would have to sacrifice growth over debt sustainability. For firms which are mid-constrained, it is optimal for them to borrow externally because they want growth and the cost of external funds are bearable. As a result, monetary policy has a larger influence over the mid-constrained firms compared to the most constrained firms and the unconstrained firms.

1.5.3 Comparison with Different Measures

In this section I will discuss the results using measure of financial constraints from Naive Bayes method and also compare my results to Ottonello and Winberry
(2018), Cloyne et al. (2018) and Jeenas (2018) that focus on similar questions.

**Naive Bayes**  As discussed in earlier section that Naive Bayes method provides a noisier measure of financial constraints by over-predicting constrained firms compared to my baseline measure. However, it also captures similar firm characteristics for constrained firms. If it is precise enough to identify the role of financial constraints on the effects of monetary policy, one would prefer a simpler and faster method.
Figure 1.6: Naive Bayes: Dynamics of Heterogeneous Investment Responses to Monetary Policy Shocks

Notes: The graph plots the scaled coefficients $\beta_h^{int}$ for each $h = 0, 1...10$ along with 95% confidence interval with the financial constraints measure from Naive Bayes method. The regression specification is $\log(k_{i,t+h}) - \log(k_{i,t-1}) = \alpha_i + \beta_h^{int}FC_{i,t-1} \times \epsilon^q_t + \Gamma X_{i,t-1} + \gamma d_{t+h} + e_{j,t+h}$. $\beta_h^{int}$ is scaled by one standard deviation of monetary policy shocks and one standard deviation of financial constraints measure. Confidence intervals are constructed based on two-way clustered standard errors at firms and quarters.

Using the measures of financial constraints from the Naive Bayes method, figure 1.6 plots the investment responses of more constrained firms upon a monetary policy shock. The signs of responses are consistent with my baseline result, but the magnitudes are much smaller and are all insignificant. Measures from Naive Bayes method add measurement errors by over-predicting the
constrained firms and thus make the regression results attenuate toward zero and insignificant.

1.5.4 Robustness of Empirical Results

My results are robust to some variations in the empirical approach and the results are shown in the appendix. In the baseline, I already show that the result is robust controlling for leverage ratio, liquidity ratio and their interaction terms with monetary policy shocks. For more robustness checks, I test the results for the following variations: (1) use observations with firm investment spells more than 40 quarters instead of 10.\textsuperscript{14} (2) use the federal funds future shocks with alternative time aggregation of simply summing the raw shocks. (3) use the Romer and Romer (2004) shocks. My baseline result is robust under above different empirical specifications.

1.5.5 Heterogeneous Debt Dynamics to Monetary Policy Shocks

In this section, I study why constrained firms respond less to monetary policy shocks. As pointed out in Ottonello and Winberry (2018), constrained firms respond less due to high marginal cost of external funds. If that is the case,\textsuperscript{14} 

\textsuperscript{14}Both Ottonello and Winberry (2018) and Jeenas (2018) use firms with investment spells more than 40 quarters. The reason why I use 10 instead of 40 quarters is because doing so will result in a smaller dataset, given data lost when merging Compustat data with the financial constraints index.
one should expect to see sluggish responses to monetary policy shocks for more constrained firms on activities related to external funds such as debt issuance, equity issuance and private placements of equity. I focus on debt which is the most popular form of external funds. I employ again the panel version of local projection method and the specification is as follows:

$$\log(b_{i,t+h}) = \alpha_i + \beta_h^{int} FC_{i,t-1} \times \epsilon^q_t + \Gamma X_{i,t-1} + \gamma d_{t+h} + e_{j,t+h}$$

Here $b_{i,t+h}$ is the level of debt for firm $i$ at time $t+h$, $FC_{i,t-1} \times \epsilon^q_t$ is the interaction term of financial constraints of firm $i$ at time $t-1$ and monetary policy shock at time $t$. $X_{i,t-1}$ is the set of control variables at time $t-1$ including firm’s financial constraints level, firm investment, real sales growth, log of firm age and log of total asset. $\alpha_i$ is the firm fixed effects and $d_{t+h}$ is the time fixed effects. Standard errors are clustered two-way on firms and quarters. Again, my coefficients of interests is $\beta_h^{int}$ for $h = 0, 1, \ldots, H$ which represents the debt responses for more constrained firms at horizon $h$ upon a monetary policy shock at horizon 0.

Figure 1.7 shows the debt dynamics of financially more constrained firms upon an expansionary monetary policy shocks. It shows that, quantitatively, the debts for firms which are one standard deviation more constrained respond less, on average, for about 0.7 percentage points in response to a one standard deviation expansionary monetary policy shock. Such sluggish responses are persistent, though it is only statistically significant for some periods. It provides supporting evidence that the sluggish investment responses of constrained firms come from
the high marginal cost of external funds.

Figure 1.7: Dynamics of Heterogeneous Debt Responses on Monetary Policy Shocks

Notes: The graph plots the scaled coefficients $\tilde{\beta}_h$ for each $h = 0, 1...10$ along with 95% confidence interval. The regression specification is $b_{i,t+h} = \alpha_i + \beta_t F C_{i,t-1} \times \epsilon_t + \Gamma X_{i,t-1} + \gamma d_{i,t} + e_{j,t+h}$. $\beta_h$ is scaled by one standard deviation of monetary policy shock and one standard deviation of financial constraints measure. Confidence intervals are constructed based on two-way clustered standard errors at firms and quarters.

1.6 Conclusion

In this paper, I constructed a measure of financial constraints from SEC filings data using a deep learning model. The characteristics of the constrained firms
are consistent with the literature. I argue that using my measure of financial constraints suits better to study the role of financial constraints to monetary policy shocks. The argument is based on the finding that firm characteristics often change non-monotonically with the level of financial constraints. Specifically, I pointed out that liquidity ratio displays a U-shape as the level of financial constraints increases which indicates that both the most constrained firms and the least constrained firms have higher liquidity ratio relative to mid-constrained firms. If the proxy used does not move monotonically with the level of financial constraints, then the empirical results using the proxy should be interpreted with caution.

I found that financially more constrained firms are less responsive to monetary policy shocks due to high marginal cost of external capital. Upon an expansionary monetary policy shock, more constrained firms cannot take advantage of the reduction in the cost of borrowing given the existing high marginal cost of borrowing. A broader inference from this paper for the heterogeneous effects of monetary policy is that firms, which rely more on internal money to fund their investments, are less responsive to monetary policy. This is true for the most constrained firms, which face high marginal cost of external funds, and the least constrained firms, which grow out of financial constraints and have ample of cash. It implies that the effect of monetary policy on aggregate investment would depend on both the fraction of the unconstrained firms and the most constrained firms. During financial crisis when external financing becomes extremely costly and as a result, many firms are financially more constrained, traditional monetary
policy might fail to stimulate investments because the marginal cost of external capital is too high for them to respond.
2.1 Introduction

There has been a rising concern about the inequality in wealth in the United States recently among publics and policy makers. Recent research suggest that wealth inequality in US has experienced a significant upsurge particularly in recent years. Saez and Zucman (2016), using a capitalization method, documented that top 0.1% population has accumulated more than 20% of total wealth at 2010 while more than 40% of wealth is held by top 1% population. Their findings shows a strikingly fast wealth concentration process in the past ten years in the United States.

That gives rise to a natural policy question: how would the optimal taxes change the evolution of wealth inequality over time with a particular focus on the top wealth group? A more progressive tax system certainly reduces the wealth inequality by redistributing wealth. However, the optimal tax policy is often a tradeoff result among redistribution among households, income uncertainty over time and average consumption level. On one hand, more progressive tax system will redistribute the resource more evenly across households and reduce income uncertainty. On the other hand, more progressive tax system will reduce
the incentive to save and invest which will have a negative impact on aggregate output and eventually may harm average consumption level of each household. In reality, policy makers also often need to consider the changing economic environment instead of facing a static economy such as growing labor income risks. Thus, policy makers need to make the tax policy not only based on the status quo, but hinge on a forward-looking consideration on future economic environment as well. However, due to long term policy consideration, it is not desirable to distort economic activity by frequently changing tax system and the slow legislative process in congress will also prevent policy makers to constantly alter the tax policy.

In this paper, I explore the impact of optimal tax policy on wealth evolution in a standard infinite horizon heterogeneous agent incomplete market model while taking into account the above tradeoff and features. I start with building a quantitative heterogeneous incomplete market model that matches the distribution of wealth well at its steady state in 1967. Then I feed in the changing economic environment such as the time varying labor earnings risks and the effective tax system which induce transitional dynamics of the economy from 1967 to 2010. The changes of the top wealth shares generated from the model fits relatively well with data from mid 1960s to 2010. The benchmark model follows closely to the model framework developed in Hubmer et al. (2016). The economy is populated by households who make decisions on consumption and savings, face idiosyncratic labor market shocks, live under a progressive tax system and supply labor. Households are ex ante the same but ex post different due to
labor market shocks and persistent preference shocks. Thus in this economy, people are wealthier due to larger labor income realizations, higher savings. In the benchmark model, my model differs from Hubmer et al. (2016) from two perspectives. First, the tax system is calibrated into a parameterized tax function to match the average tax rate observed in the data. This is different from the Hubmer et al. (2016) where they used a step-wise function. Using the parameterized tax function would allow me to compared the optimal progressivity in the counterfactuals to the progressivities observed in benchmark model. Second, households are allowed to choose labor supply.

For the counterfactual experiment, I consider the case where the planner chooses an once-and-for-all tax reform at 1967 accounting for the full transition dynamics induced by the new tax policy and changing economic environment. The one time unanticipated tax reform at 1967 reflects long term policy consideration. The planner is equipped with a fixed functional form of tax where he/she is free to choose the optimal level or progressivity of the tax system. I consider two counterfactual exercises for a utilitarian social planner: a linear tax reform and a non-linear parameterized tax reform as a replacement for the existing tax system. I found that in the linear tax reform, due to the lack of control over the progressivity of the tax system, the optimal linear tax rate is 33% which is modest compared to the existing tax system. Because planner cannot achieve redistribution and reduce income uncertainty efficiently with linear tax, he/she tends to have a mild level of tax rate to increase the average consumption level to maximize social welfare. Social welfare under the optimal
linear tax is actually 0.4% less than the benchmark model showing that there is a strong restriction on one time change linear tax reform. Wealth inequality enlarges even faster in this counterfactual exercise because the marginal tax rate is significantly lower for the top wealth groups compared to those in the benchmark model. For the non-linear tax reform, Ramsey planner would be equipped with the parameterized tax/transfer function used in Heathcote et al. (2014) which essentially the planner is choosing the optimal progressivity of the tax system. The optimal progressivity is 0.66 which is extremely progressive compared to 0.151, the estimated progressivity level of US current tax system from Heathcote et al. (2014). That shows, in facing an increasing uncertainty of labor earnings and an increasing of top labor earnings share, the motives for redistribution and reduction in income uncertainty dominates the level effect. Even in the benchmark model, the tax system is characterized with the highest progressivity, it is still significantly lower than the optimal level. Wealth inequality under such progressive tax system, however, will reduce eventually at 2010 compared to the benchmark economy. Our result suggested that a utilitarian planner will choose an extremely progressive tax system compared to the current tax system in combat with the enlarging inequality and increasing income uncertainty. It indicates that from a perspective of a utilitarian social planner, the current wealth inequality is concerning and if progressivity is the only policy at hand, it will choose extreme tax measures to bring inequality down despite of dampening the output of the economy.

My paper begins with in Section 2 with a brief review on related literature.
Section 3 describes the model environment in the benchmark economy. Section 4 discusses the calibration strategy, wealth distribution matching at steady state and its evolution in transition dynamics. I conduct two counterfactual experiments and discuss the implication of optimal tax system on evolution of wealth inequality respectively in Section 5 and Section 6 concludes.

2.2 Literature Review

The paper is directly related to a vast literature of optimal fiscal policy in standard incomplete market model under a Ramsey problem. Earlier research focus particularly on maximizing steady state social welfare such as Conesa and Krueger (2006) and Conesa et al. (2009). Conesa and Krueger (2006) try to identify the optimal progressivity of the comprehensive income tax at steady state while Conesa et al. (2009) estimates the optimal capital and labor income tax separately at steady state. Though Conesa et al. (2009) show the transition path from the old steady state to the new one, the optimal policy does not take into account the welfare gain/loss during transition period.

Domeij and Heathcote (2004), Fehr and Kindermann (2015), Bakış et al. (2015) and Pedroni et al. (2016) all consider fiscal policy accounting for the transition dynamics and show the importance of including transitional dynamics in welfare analysis and optimal policy. Domeij and Heathcote (2004), though not studying directly the optimal fiscal policy, consider the welfare implications with zero
capital income tax reform accounting for transition dynamics under a standard incomplete market model. They conclude that such reform would be detrimental when transition dynamics is taken into accounts. Fehr and Kindermann (2015) extends Conesa et al. (2009) by taking into account the transitional dynamics and intragenerational risk sharing. They pointed out that the optimal capital tax would be very different when transition dynamics is considered. They also points out the level of optimal linear tax rate depends crucially on the assumptions of social welfare function. In a similar spirit, Bakış et al. (2015) study the optimal progressivity of the tax system accounting for the transition dynamics under a dynamic overlapping generation model. Instead of considering an once-and-for-all tax reform, Pedroni et al. (2016) consider the optimal fiscal policy path where planner commits to a sequence of fiscal policies they announced at the beginning period. The final steady state is thus determined endogenously. My paper also considers optimal fiscal policy accounting for transition dynamics given its importance and my particular focus on the impact of fiscal policy on the evolution path of wealth distribution along the transition dynamic period. My paper will take a similar experiment that allows only an once-and-for-all tax reform at the beginning period.

Krueger and Ludwig (2013) and Heathcote et al. (2014) characterize the optimal fiscal policy with a particular focus on the tradeoff in the labor markets. Krueger and Ludwig (2013) study how to choose the optimal tax policy jointly with the complementary optimal subsidy policy on college education when taking into account on human capital accumulation. Heathcote et al. (2014) also features
human capital accumulations by adding an endogenous skill investment decision in an incomplete market model. Through a partial insurance structure, they obtain a fully tractable model where households don’t save at equilibrium. That allows them to fully characterize the tradeoff in formulas that determines the optimal progressivity of tax system instead of numerical approximation. As a consequence of the model setup, they abstract away from endogenous saving decisions.

Differing from Krueger and Ludwig (2013) and Heathcote et al. (2014), my paper concentrates mainly on the tradeoff in capital income supply decisions since I am interested in how optimal taxes shape the evolution of wealth distribution. In addition, Saez and Zucman (2016) point out that the top households have a much larger saving rate compared to the bottom 50% and thus contribute significantly to further wealth inequality in the US. In my paper, savings is a key factor related to wealth inequality and richer households are richer partly due to a higher saving rate which captures the data feature. Most importantly, one of the main contributions of my paper is that I consider the optimal tax policy under a time-varying economic environment. By changing economic environment, it means changing labor earnings risk specifically in this paper. Most of earlier papers consider a change of the tax system while assuming other economic environment is static. Thus, I focus on how the optimal tax policy shapes the wealth inequality over its evolution path with a changing economics environment by taking into account the time varying labor income risks.

For the modeling part, it is required to replicate the Pareto tail observed in the
wealth data given my particular focus on them. Thus my modeling part is closely related to the macro-wealth inequality literature. Traditional standard incomplete market heterogeneous agent model (Bewley (1977, 1980, 1983, 1986), Huggett (1993) and Aiyagari (1994)) is well known of having difficulty of matching the right tail of wealth distribution. Benhabib and Bisin (2016) summarize that for an random growth model, in order to match the thick tail in wealth distribution, it is often needed to feature at least one of the following features often observed in the data: skewed labor earnings, increasing saving rates positively associated with wealth level and heterogeneous investment returns. Cagetti and De Nardi (2006) build a quantitative life-cycle model with voluntary bequest of occupational choice that allows for entrepreneurial entry, exit and in investment decision in the presence of financial constraint. Occupational choice along with decreasing return technology differentiates the capital returns where entrepreneurs can use capital to finance profitable production while labors can only save to earn risk free interest rate. Financial constraints on borrowing also creates the increasing saving motives for productive but capital poor entrepreneurs. Benhabib et al. (2011) and Benhabib et al. (2015) emphasize the importance of investment risks in shaping the tail of wealth distribution. Kaplan et al. (2016), though main target is not to match explicitly the wealth distribution, build a two-asset model with transaction cost that matches well for distribution of both liquid and illiquid asset in US. Their model still build in all of the three ingredients: a skewed labor earning process which is left skewed and with high kurtosis, higher saving rate for asset rich due to high transaction cost of illiquid asset and wealth in utility,
endogenous differentiate investment returns based on the two-asset structure. In my benchmark model, I also build in the skewed labor earnings and increasing saving rate to generate realistic wealth distribution under a standard infinite horizon heterogeneous agent incomplete market model.

Most of the macro-inequality literature focus on matching the stationary (long run) wealth distribution but only a few of them focus on quantifying the factors that contributes to the evolution in wealth distribution. Kaymak and Poschke (2016) conduct a quantitative investigation on the contribution to changes in wealth inequality in a life-cycle incomplete market model. They compare the roles of productivity changes which favors skilled workers, flatting tax system and an expansion on social security programs and find that it is the expansion of the social security program that contributes the most in the increase of wealth inequality. They calibrate the model twice to two steady states: one in 1967 and one in 2010 and hence there is no transition dynamics. Hubmer et al. (2016) also concentrate on examining the drivers of evolution of wealth inequality in US using a standard incomplete market model and find that the flatting tax system contributes the most to the sharp rising wealth inequality in the past forty years. The major difference compared to Kaymak and Poschke (2016) is that they only calibrate the model to one steady state in 1967 and then start plugging the time varying labor earnings risk and tax system which induce a transition dynamics between 1967 and 2010. I mainly followed Hubmer et al. (2016) which allows me to match the wealth distribution at the steady-state in 1960s and its evolution from 1960s-2010 in my benchmark model.
A highly complementary literature is the new dynamic public finance literature following Golosov et al. (2003), Farhi and Werning (2013) and Golosov et al. (2016). The literature explores the role of idiosyncratic shocks and the relevant insurance problems using a mechanism design that allows the planner to efficiently extract information on agent’s productivities.

2.3 Setup

Wealth Inequality in Data

Saez and Zucman (2016) document the evolution of top wealth concentration from 1960s to 2010 using a capitalization approach where they back out the wealth of a household through the reported capital income from IRS. Their results indicate that the wealth of the top 1% population in US has sharply increased to 39.5% in 2010 compared to 27.8% in 1967 while the wealth share of the top 0.1% population increases from 9.4% in 1967 to 20.7% in 2010. They point out that two of the main drivers of increasing wealth inequality in the US are the increasing income inequality and an increase in saving rate inequality. Smith et al. (2020) propose a downward revision on these estimates of top wealth shares by accounting the heterogeneity within asset class though there is still a significant increase in the top wealth share. However, they still show that the wealth share of the top 0.1% increases significantly from 7% in 1978 to 14% in 2010. Given that their data on top 1% and top 10% are not available, I will mainly use Saez and Zucman (2016)
as my benchmark data.

**Model Setup**

In this section, I will describe my model economy that captures some data features implied by Saez and Zucman (2016). The model is a standard Bewley-Huggett-Aiyagari heterogeneous incomplete market model in discrete time (SIM model). It is well known that the simple SIM model typically fails to generate realistic wealth distribution that matches US data. Benhabib and Bisin (2016) show that to generate skewed thick tail for wealth distribution, one often needs the following channels: skewed thick tail for labor earnings; differential saving rate (rich households save more) and stochastic investment risks.\(^1\)

To generate realistic income and wealth distribution, I augment the basic SIM model with labor earnings process that consists a persistent component, a transitory component and a Pareto tail and a idiosyncratic stochastic discount factor following closely Hubmer et al. (2016). The persistent and transitory components are standard setup for labor earnings and the Pareto tail is to capture the top earnings share that would otherwise missed by the standard log normal labor earnings process. The stochastic discount factor is a parsimonious to model heterogeneous saving motives in an infinite horizon model. That accounts two channels mentioned in Benhabib and Bisin (2016) and also reflects data features

\(^1\)Note that typical macro-inequality literature features life cycle framework but the main structure is to feature increasing saving rates associated with wealth and the life-cycle labor earnings structure. I use stochastic preferences to model realistic saving rate.
in Saez and Zucman (2016) that heterogeneous savings rate and large volatility in labor earnings are the two main driving factors of wealth inequality. Once solve the steady state at 1967, I then feed in the time-varying labor earnings volatilities and the varying effective tax systems from 1967 to 2010 to obtain the benchmark transitional dynamics of top wealth shares.

2.3.1 Consumers

The economy consists a continuum of measure one of heterogeneous consumers. In each period $t$, each consumer discounts future with an idiosyncratic stochastic factor $\beta_t$ which follows a Markov process. Labor supply is endogenous. Consumer supplies a stochastic amount of efficiency units of labor $l_t(p_t, v_t)$ which contains a persistent component $p_t \sim \Gamma_p(p_t|p_{t-1})$ and a transitory component $v_t \sim \Gamma_v(v_t)$. The gross income which is the sum of labor and capital income are subject to income tax $\tau(.)$ and each consumer receives a lump-sum transfer $T_t$.

Consumer with a CRRA preferences and a disutility of labor supply maximizes present discounted utility subject to budget constraint and borrowing constraint.

$$\max_{(c_t)_{t=0}^\infty} \left\{ \frac{c_{1-\gamma}}{1-\gamma} + E_0 \left[ \sum_{t=1}^\infty \prod_{s=0}^{t-1} \beta_s \frac{c_{1-\gamma}}{1-\gamma} - \chi \frac{l_{1+\varphi}}{1+\varphi} \right] \right\}$$
subject to \( c_t + a_{t+1} = x_t \)
\[
\begin{align*}
    x_t &= a_t + y_t + \tau(y_t) + T_t \\
    y_t &= r_t \eta_t a_t + e_t w_t l_t \\
    a_t &\geq \bar{a}
\end{align*}
\]

where \( \bar{a} \) is the borrowing constraint.

the consumer’s problem could be restated in a simpler recursive form

\[
V(x_t, p_t, \beta_t) = \max_{a_{t+1} \geq \bar{a}} \left\{ u(x_t - a_{t+1}, l_t) + \beta_t E \left[ V_{t+1}(x_{t+1}, p_{t+1}, \beta_{t+1}) | p_t, \beta_t \right] \right\}
\]

subject to \( x_{t+1} = a_{t+1} + y_{t+1} - \tau_{t+1}(y_{t+1}) + T_{t+1} \)
\[
\begin{align*}
    y_{t+1} &= r_{t+1} a_{t+1} + e_{t+1} w_{t+1} l_{t+1}
\end{align*}
\]

where \( x_t \) is the total cash-on-hand at time \( t \).

### 2.3.2 Production and government

Firms are perfectly competitive with CRS technology \( F(K, L) \).

Competitive wage rate per efficiency unit and average market return on capital are
\[ w_t = \frac{\partial F(K_t, L_t)}{\partial L_t} \]

\[ r_t = \frac{\partial F(K_t, L_t)}{\partial K_t} - \delta \]

where \( \delta \) is the depreciation rate of capital.

Government redistributes a fraction \( \lambda \) of total tax revenue by a means of uniform lump-sum payment \( T_t \).

### 2.3.3 Steady State Equilibrium

A steady state equilibrium is characterized by a market clearing level of aggregate capital \( K^* \) and a lump-sum transfer \( T^* \) such that:

1. Factor prices \( w^* = \frac{\partial F(K^*, L^*)}{\partial L} \) and \( r^* = \frac{\partial F(K^*, L^*)}{\partial K} - \delta \) solve firm’s problem.

2. Given \( r^*, w^* \) and \( T^* \), consumers solve their optimal consumption problem \( c^* = g^*(x, p, \beta) \), giving rise to an invariant distribution \( \Gamma(a, p, \beta, v) \).

3. The government redistributes a fraction \( \lambda \) of total tax revenue lump-sum

\[ T^* = \lambda \int \tau(\tau^* \eta a + w^* l(p, v))d\Gamma(a, p, \beta, v) \]

4. Capital market clears

\[ K^* = \int a d\Gamma(a, p, \beta, v) \]
5. Labor market clears

\[ L^* = \int ld\Gamma(a, p, \beta, v) \]

2.4 Quantitative Model of Wealth Inequality Dynamics

2.4.1 Calibration

Earnings Process

Labor earnings process is based on the traditional log-normal framework that earnings process consists of a persistent component which follows an AR(1) process and a transitory component. I use corresponding estimates from Heathcote et al. (2010) from 1967-2010. As well known, the resulting log-normal cross-sectional distribution understates the top labor earning share and the top labor earning share in US displays a Pareto tail. Thus the framework is augmented with a Pareto tail \( \kappa_t \) for the top 10% earners targeting directly to match the top 10%, 1%, 0.1% and 0.01% labor income share from 1967-2000 estimated in Piketty and Saez (2003)) following Hubmer et al. (2016).

The efficient labor unit supplied is

\[ e_t(p_t, l_t) = \psi(p_t) \exp(\nu_t) \]
Where \( \psi(p_t) = \begin{cases} 
\exp(p_t) & \text{if } F_{pt}(p_t) \leq 0.9 \\
F^{-1}_{\text{Pareto}(\kappa_t)}\left(\frac{F_{pt}(p_t) - 0.9}{1 - 0.9}\right) & \text{if } F_{pt}(p_t) > 0.9 
\end{cases} \)

Note that at \( F_{pt}(p_t) = 0.9 \), \( F^{-1}_{\text{Pareto}(\kappa_t)}\left(\frac{0.9 - 0.9}{1 - 0.9}\right) = F^{-1}_{pt}(0.9) < \exp(F^{-1}_{pt}(0.9)) \). Thus the persistent income process is not continuous and non-monotonic at \( F^{-1}_{pt}(0.9) \). The Pareto tail is scaled up by a factor of \( n_t = \frac{\exp(F^{-1}_{pt}(0.9))}{F_{pt}(0.9)} \) to make persistent labor earnings component \( \psi(p_t) \) continuous and monotonic along \( p_t \). The Pareto tail parameter \( \kappa_t \) is calibrated to match the top earner’s labor income share at top 10%, 1%, 0.1% and 0.01% documented in Piketty and Saez(2003). The key reason that allows me to match the top labor earnings share directly with a Pareto tail is the inelastic labor supply since households do not endogenously react to the tail parameter. Another note here is by taking the changing labor earnings risk and Pareto tail as exogenous economic environment, I am also assuming that the optimal tax system cannot change the pre-tax labor earnings risk and the Pareto tail in labor earnings distribution. One of the arguments is some of changes in the labor earning risks are driven by other exogenous factors such as productivity changes which favor skilled labor which is relatively exogenous to tax system.

Figure 2.1 shows both the evolution of cross-sectional standard deviations of persistent component and transitory component and the Pareto tail coefficient. There is a fast rising trend in the innovation of persistent component from 1960s to mid 1990s and then it flattens out. The volatility in transitory component is rising at a much slower pace though it is more volatile. Both show an overall

\footnote{See appendix for detailed matching results.}
Notes: Left figure plots the evolution of the cross-sectional standard deviations of labor earnings components and the right figure plots the calibrated Pareto tail coefficient.

growing earnings risk in the past four decades. For the calibrated Pareto tail coefficient, there is a sharp drop from 1960s to 2000 which shows a fast growing labor earnings share for the top earners. Overall it shows that there is a general rising trend for the volatility of the persistent component of labor earnings and the earnings of the top 10% has been increasing further over time. The enlarging earning inequality will contribute directly to the enlarging wealth inequality.

Tax System

Tax data are taken from Piketty and Saez (2007) where they estimates the federal effective tax rates from 1960-2000. The effective average tax rate comprises four major federal taxes: individual income, corporate income, estate and gift, and
payroll taxes. Effective average tax rates used are the sum of all four major
taxes and are calculated for 11 income groups (percentage: 0-20, 20-40, 40-60,
estimates into my model, tax function is calibrated as a step-wise function. For
each income bracket, the threshold is set to match its income share and marginal
tax are imputed to match its average tax rate in the data.\(^3\).

For setting the income thresholds for each income group, I start assuming
that tax base only applies to non-negative income, so the lowest income group
starts with 0. Then I use a parameterized tax/transfer function to approximate the
cumulative distribution function of total income and then calculate each income
threshold by matching its income share in the data. Once the threshold is set,
one could write out the average tax rate for each income group as a function of
marginal tax rate and income thresholds. The marginal tax rates for each income
group are then solved by solving a system of linear equations.

The tax/transfer system I adopted is from Heathcote et al. (2014) which
specifies the tax function as

\[
y_t - \lambda y_t^{1-\tau}
\]

\[
T(y) = y_t - \lambda y_t^{1-\tau}
\]

Here \(\lambda\) is a proxy measure of average tax and \(\tau\) is a proxy measure of progressivity
of the tax system.

\(^3\)See appendix for construction details.
Figure 2.2 plots the imputed marginal tax rate over time for different income groups. The top rate is high during 1960s and mid 1970s but then declines sharply over time while the mean income marginal tax rate is relatively flat from 1960s-2010s. In Hubmer et al. (2016), it is the sharp decline in the marginal tax rates of high earnings groups that contributes the most to the rising wealth inequality.

**Idiosyncratic Discount Factor**

Standard Aiyagari model has shown great difficulty in matching the wealth distribution, especially for the right tail. Typically the model fails to generate the increasing saving rate along wealth dimension. The macro-inequality literature features various micro-foundations of the differential saving rate such as differential bequest motives, entrepreneurship with borrowing constraint and transaction cost in illiquid asset etc. For simplicity, here we follow the same idea
in Krusell and Smith (1998) by using a persistent idiosyncratic stochastic discount factor. Discount factor $\beta$ follows a Gaussian AR(1) process

$$\beta_t = \rho^\beta \beta_{t-1} + (1 - \rho^\beta) \mu^\beta + \sigma^\beta \epsilon^\beta_t, \epsilon^\beta_t \sim N(0,1)$$

I set $\mu^\beta = 0.92$ to match a ratio of capital to net output of about 4 in steady state. $\rho^\beta = 0.992$ and $\sigma^\beta = 0.0019$ imply the standard deviation of the cross-sectional distribution of discount factors is 0.0148.

**Zero Earning State**

Unemployment is not explicitly modeled here for simplification. A zero earning state is characterized to reflect long term unemployment probability and possibility of temporarily exiting the labor force. The zero earning state happens with probability $\chi_0$ which is independent of realization of $p_t$ and $v_t$.

**2.4.2 Steady State Equilibrium**

Now we calibrate the persistent level of discount factor $\rho^\beta$, standard deviation of discount factor $\sigma^\beta$, standard deviation of investment risk $\sigma^n$, zero earning probability and borrowing constraint $\bar{a}$ to match six main features of wealth distribution in the initial steady state economy in 1967: the share of wealth holdings for top 10%, 1%, 0.1%, bottom 50% and the fraction of negative asset
Table 2.1: Matching Wealth Distribution in 1967 as a Steady State

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\mu^\beta$</th>
<th>$\beta^\beta$</th>
<th>$\sigma^\beta$</th>
<th>$\varphi$</th>
<th>$\chi$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>0.92</td>
<td>0.9926</td>
<td>0.00192</td>
<td>3</td>
<td>17</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Steady-State Wealth Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>Data(1967)</td>
</tr>
<tr>
<td>Model</td>
</tr>
</tbody>
</table>

holdings. Most importantly, given that we have labor supply in the baseline model, we cannot use the earnings volatility directly but instead shrink the earnings volatility by a fixed amount $t = 1.6$ such that the endogenous labor earning volatility matches data. And the shrinking parameter is fixed for the transitional dynamics as well. Table 2.1 summarizes the calibration of the five parameters and the comparison of wealth distribution between the model and actual data. The top wealth share at 1967 is matched closely in the model while the bottom 50% is a bit higher compared to the data. Overall, it shows that my standard heterogeneous agent incomplete market model matches well compared to data.\(^4\)

\(^4\)One could calibrate the borrowing constraint to be negative to match the bottom 50% wealth share better at steady state and we need to revise the parameterized tax function a little bit. It turns out that quantitatively the evolution of top wealth shares under these two cases are not large.
2.4.3 Transition dynamics

In the previous subsection, the benchmark model in steady state matches well with the wealth share in 1967. I now turn to the perfect foresight transition dynamics of the economy under perfect foresight and show that it also matches relatively well on the changes of top wealth share from 1960s to 2010. An economy with myopic agent is also investigated but it doesn’t change much along the transition path for top wealth shares. First, an end-point steady state is needed to compute the transition dynamics. So we are computing a deterministic transition path between two steady states. I assume the volatilities of innovations in both the persistent and transitory components, the Pareto tail of top labor earnings and the tax system to be fixed at 2010 level after 2010 while the parameters in the discount factor, the probability of zero earning state and the borrowing constraint are assumed to be constant through out. An alternative assumption would be the standard deviation of the persistent labor earnings process is constant after 2010. This doesn’t quantitatively change either the number of periods that requires to converge to new steady state or the path of evolution of wealth inequality much. Second, I assume a finite time path for the economy to reach the new steady state. It begins with assuming the economy has reached its new steady state after $T$ periods ($T=300$ in benchmark) and picks a factor price path $\{r_t\}_{t=0+T}$, thus consequently $\{w_t\}_{t=0+T}$, $\{K^b_t\}_{t=0+T}$ is determined. Then start iterating backward from time $T$ using the guessed factor prices back to the beginning time period $t_0$.

---

5A larger $T$ does not change the amount of time converging to new steady state, nor the transition path.
and obtain the policy function for each period. Finally, I iterate the distribution of capital forward using the obtained policy function and factor prices to obtain the implied aggregate \( \{K^f_t\}_{t=0}^{t+T} \). Repeat the algorithms until the transition dynamics of aggregate capital converges, \( \max|K^b_t - K^f_t| < \epsilon \).

Figure 2.3 plots the evolution of top wealth share against its data counterpart in Saez and Zucman (2016). The overall upward trend in top wealth share is picked up by the model: wealth shares in top 10\%, 1\%, 0.1\% and 0.01\% are increasing for the past thirty years showing a sharp increase in wealth concentration which is in line with the data. The overall magnitude changes of wealth shares match well for the top 1\% and top 0.1\% and for the top 10\%, the

73
overall shape looks very similar.

In addition, the short term drops in top wealth shares from 1970s to mid 1980s are not captured by the model which implies it might be caused by factors other than variations in labor earnings and tax system. Hubmer et al. (2016) shows in an extension with partial equilibrium that adding heterogeneous portfolio holdings across wealth groups seems to explain the short term variation in top wealth shares well. Heterogeneous portfolio holdings would also have profound implication to my optimal tax policy as well due to its differential tax incentives across different asset class, but it would require to change the structure completely if we wish to build up a general equilibrium framework with heterogeneous portfolios. The changes of relative magnitude of the top 10% and top 1% are captured well in the model with small percentage errors, but in the further right tail, the changes of relative magnitude in top 0.1% and 0.01% fail far short compared to its data counterparts. It is consistent with the argument in Gabaix et al. (2016) that the transition of wealth inequality under a random growth mechanism would be too slow, especially at the top tail. I would argue that the model picks up the overall trend of increasing concentration of wealth inequality well and the relative change in top 1% and top 10% well but fails to matches the changes in magnitudes for top 0.1% and top 0.01% wealth share. Overall, it is a reasonable fit and it is taken as my benchmark model with transition dynamics that is used to compare to the counterfactual experiment conducted next.

One interesting things to compare is the evolution of labor earnings in the
Figure 2.4: Evolution of Top Income Shares from 1967-2010

model compared to the data given that labor supply is endogenous. 2.4 plots the labor earnings share for top income groups overtime. it shows that overall, it matches the data relatively well for top 10% and top 1%. ?? shows that the endogenous labor earnings volatility matches the data well which is quite surprising.

2.5 Counterfactual Experiments

Given my benchmark model matches relatively well for the changes of the top 1% and top 0.1% from 1960s to 2010 and it also implies that the time varying
tax system contributes significantly to the increasing wealth inequality. Now let’s turn to the optimal tax policy. It is of interest not only to see what the social planner would choose differently compared to the benchmark tax system, but how it would shape the evolution of wealth inequality over time as well. Here I consider two counterfactual experiments with linear income tax and non-linear parameterized income tax respectively for a utilitarian social planner. I am also assuming that the planner can only choose an once-and-for-all tax system due to computing constrications. For the social planner, it mainly face three tradeoffs when it comes to the design of its tax system: the consumption inequality among households, the income uncertainty over time and the average consumption level.
A tighter tax system would increase social welfare by decreasing the inequality among households, reducing the income uncertainty over time but also decrease the social welfare by reducing the average consumption level.

**Linear Comprehensive Income Tax**

I start with a linear tax system for the first counterfactual. The planner chooses and commits to \( \tau_y \), the linear income tax rate to maximize \( SWF = \int \omega_i V_i(\{c_i(t)\}_{t \geq t_0}) di \) starting at year 1967. The linear tax system would still be interesting here because the planner equipped only with a linear tax system would have little impact on redistributing income among households and low impact on income uncertainty but could still influence the level of consumption. So when it comes to progressive tax system, I could compare the difference to figure out how much does the planner care about the inequality among households. Note that transfer system is still here which is funded by 60% of the tax revenue.

Using a global search method, the optimal linear income tax rate is 34% which is relatively higher than 19.2% which is the average tax for mean income from 1967 to 2010. It is quite intuitive here because the average tax rate cannot redistribute and is inefficient at reducing the income uncertainty. To deal with an economy with increasing income uncertainty, the average tax rate has to be raised higher to bring down such uncertainty. However, it is much lower for the top income groups compared to either 71% at their 1967 level or 41% at their 2010 level. As
a result, inequality under linear tax system would be worse than the benchmark. One observation is that the welfare change under such optimal linear tax reform actually decreases by 0.4%. This might due to several reasons. First, the linear tax form restricts a non-progressive tax policy, though the entire government system is slightly progressive after adding the lump-sum transfer system. Second, planner is only allowed to use one tax rate under the reform while tax system varies overtime in the benchmark economy.

Figure 2.6 depicts the evolution of top wealth shares under the optimal linear income tax reform against the benchmark’s results. Without the progressive tax system to restrict wealth accumulation for the top wealth groups, the top wealth share under optimal linear tax increases even faster compared to the benchmark. Humber, Krusell and Smith(2016) show that decreasing tax progressivity not only increases post-tax income but also makes future consumption more attractive by increasing the net-of-tax return. Thus a large drop in top tax rate would significantly boost savings for the top wealth groups and promote wealth inequality further. This is particularly remarkable for the top 1% wealth group where the wealth share of the top 1% in the counterfactual is 10 percentage points higher than the benchmark.
Non-Linear Parameterized Income Tax

Now I turn to the case where planner would be able to use a non-linear parametrized income tax function to maximize social welfare. I consider using the tax/transfer function from Heathcote et al. (2014)(HSV) where they show that their parameterized tax/transfer function fits well with the US data.

In Heathcote et al. (2014), the post-government income takes the form

$$y^d = \lambda(y)^{1-\tau}$$

then the associated tax/transfer function
\[ T(y) = y - \lambda(y)^{1-\tau} \]

where \( \lambda \) denotes the associated level of tax and \( \tau \) denotes the progressivity of tax system. \( \tau = 1 \) means a perfectly progressive tax system and everyone receives post-government income \( \lambda \) while \( \tau = 0 \) indicates a non-progressive system with a flat tax rate at \( (1-\lambda) \). Noted that it is a tax and also a transfer function with a break even point \( y_b = \lambda^{\frac{1}{\tau}} \) where household would pay positive tax if pre-government income is higher than \( y_b \) and receive transfers if pre-government income is lower than \( y_b \). So not only the original tax function is replaced but also the lump-sum transfer system is replaced.

Now consider the utilitarian social planner at 1967 chooses and commits to a progressivity level of the tax/transfer system \( \tau \) to maximize social welfare subject to a constraint that it needs to finance exogenous government spending.

\[
\max_{\tau} \int V_i(k, \beta, p) \Gamma(k, \beta, p) \\
\text{s.t.} \\
\int y_i dF(y_i) = \int (y_i - \lambda(y_i)^{1-\tau}) dF(y_i)
\]

\( g = 0.17 \) is the exogenous proportion of total income spent on public goods and calibration follows Conesa et al. (2009). Here \( \lambda \) is backed out by the government budget balance once \( \tau \) is chosen. I restrict \( g \) to be exogenous because
the public good is not valued by the households thus it is always optimal to choose no public good consumption which implies \( g = 0 \).

The optimal progressivity \( \tau^* \) is 0.66 which is much higher than 0.151 estimated by Heathcote et al. (2014) for US economy. Figure 2.7 plots the marginal tax rate for the optimal nonlinear tax function against the imputed marginal tax rate at 1967. Contrast to the counterfactual experiment with linear tax which features low tax rate and low progressivity for top wealth groups, the optimal tax system under non-linear tax function is much more progressive than the benchmark. Marginal tax rate under optimal tax system is even much higher than the marginal tax rate at 1967 when actual tax system is already very progressive. That shows the redistributive effect and insurance effect likely dominates the level effects where the desire to reduce income uncertainty and existing income inequality is much larger than incentives to increase average consumption. A very progressive system features large redistribution among preexisting inequality and mitigates greatly the uncertainty in incomes while depress output due to its distortion on savings and might potentially reduce average consumption level. Compared to the linear tax policy, the progressive tax policy is an effective tool of for the social planner to dealing with redistributive motives and the insurance motives. To balance out the impact on average consumption, the implied average tax rate is actually close to the average income tax in the benchmark.

Figure 2.8 displays the evolution of top wealth shares under the optimal nonlinear tax system against the benchmark. With such progressive tax system,
wealth inequality is surprisingly at first increasing from 1960s to mid 1990s but eventually decreased significantly at 2010 compared to benchmark economy. The hump-shape of the evolution might mainly due to the sharp rising top labor earnings share(a decrease in Pareto tail) during 1960s to mid 1990s. The rapid increase in top labor earnings would mechanically increase the top labor earnings and their wealth shares. When changes in top labor earnings share stabilize after mid 1990s, with the very progressive tax system, wealth shares start to decline for the top wealth groups. At 2010, top 10% wealth share is around its level at 1967 while for top 1%, 0.1% and 0.01%, their wealth holdings are even lower than their 1967 level. For welfare change, however, it decreases 16% which is even more than previous counterfactual exercise with linear tax which is mainly due to the fact that the planner only commits to an once-and-for-all tax reform starting from 1967 while the benchmark model features time-varying tax system as mentioned

6 labor earnings risk is referred as the standard deviation of the innovation in both persistent and transitory component.
To sum up, even if the planner is constrained with the once-and-for-all tax policy and the earnings risk is only getting larger in the far future but not in 1967, planner would still prefer a much more progressive tax system in the needs of redistribution and uncertainty reduction.
2.6 Conclusion

I characterize the optimal linear and nonlinear taxes in a quantitative standard heterogeneous agent incomplete market model and examine its impact on evolution of wealth distribution. The benchmark economy matches the distribution of wealth well at the steady state in 1967 and reasonable close for its evolution from 1967 to 2010. Optimal linear tax on comprehensive income is too restricted and the wealth inequality over time is enlarged compared to the benchmark model. Optimal non-linear parameterized income tax is very progressive and as a result, the wealth inequality is significantly reduced under such regime. It shows that from a utilitarian social planner’s view, bringing down the increasing inequality is much more of a priority compared to maintaining a high average consumption level giving the increasing earnings risk over time.
CHAPTER 3
A WELFARE APPROACH TO THE PROXY MEANS TESTING

3.1 Introduction

In many developing countries, when it carries out social transfer programs, if it is not a universal basic income scheme, it will perform a means testing to determine the eligible households typically using the consumption or income level as a benchmark. In many cases, the data on income or consumption are not accurately measured due to limited information and it is common to conduct a proxy means testing (PMT) with other information to figure out the targeted households. Traditional methods on the proxy means testing are generally split into two main categories: the categorical targeting and the econometric targeting. In a categorical targeting, it identifies the targeted households by providing a checklist for all households and households are selected once certain items in the list are checked such elderly with age 65 or older, disable people and children under fourteen etc. Econometric targeting refers to any PMT based on a regression model as defined in Brown et al. (2016). Effectively, it predicts the level of household consumption/income by using other covariates. Econometric targeting widely used in plenty of social transfer programs that intend to reduce poverty. Grosh (1994) compared various social programs and concluded that econometric targeting generally produces the best targeting outcomes in terms of reducing inclusion errors where a non-poor household is falsely identified as
poor. However, econometric targeting also suffers a few critics. First, though it outperforms other PMT in many cases, the prediction about which households are poor or not is unsatisfying as pointed out in Kidd and Wylde (2011). In addition, in many applications, the covariates that are used to conduct PMT are often kept secret from the public for the purpose of avoiding manipulations. Even the covariates used are known to the general public, it is still very difficult to explain the intuition of the econometric targeting to the general public such as why some households are targeted while similar households in the same village are not included in the social programs. Roopnaraine and Adato (2004) are concerned that the transparency of such measure is much lower in terms of explaining to the general public compared to the categorical targeting. The study of Cameron and Shah (2014) shows that the difficulty in explanation can lead or contribute to social unrest in some countries.

In the practice of PMT on cash transfer programs that target poverty reduction, typically the econometric targeting is separated into two steps. First it predicts the log consumption per capita through regressions with covariates. Second, it then uses the proxy of consumption to assess the targeting performance based on various different measures such as inclusion error rate, exclusion error rate and normalized targeting differentials etc. However, the evaluation of target performance is detached from the regressions that estimate the consumption. Another common method is to directly predict the probability of being under the poverty line using a probit or logit model in the first step. By doing so, it also effectively deal with the inclusion error and exclusion error. However, Brown
et al. (2016) show that using the binary variable results in even higher targeting errors in all of their cases. Alternatively, the “poverty-weighted least squares” (PLS) often place higher weights on the poorer people in regressions. A typical method is to give zero weights on those above the poverty line and as a result, only running regressions on poor households. PLS seems to perform the best among other methods shown in Brown et al. (2016).

In my paper, I provide a welfare approach to the proxy means testing by deriving the minimization problem from a welfare perspective. With a utilitarian social planner, I start off by evaluating the welfare loss by comparing the welfare gain under a world where poor and non-poor are correctly identified to a welfare gain under an economy with the targeted group identified with proxy of consumption. In a uniform transfer scheme, the welfare approach is equivalent to a weighted logistic regression model with higher weights on poorer households. The intuition is that, under a utilitarian social planner, the marginal benefit of a transfer is higher to the poor than to the non-poor. This effectively shows that the welfare approach would lead to PLS with weights derived from welfare loss. Compared to the standard OLS method using log consumption as dependent variable, the weighted logit model reduces the exclusion error in sacrifice of the exclusion error which is similar to other PLS method.

The contributions of my paper are twofold. First, I show that one can design the PMT from a welfare perspective. Second, I showed that in some settings, the welfare approach would be equivalent to the poverty weighted least square
practices with weights derived from welfare loss. This allows us to better design the PLS method with welfare weights and some transparency in explaining PLS. In addition, I notice that in the previous literature, when estimating the performance of the model, they typically do not separate the data into training and test set. As a result, there is no discussion of out-of-sample prediction accuracy. Given the nature of the problem is all about prediction, I separate the data into training and test set and discuss the results on the test set for different approaches. To be notice, the nature of this particular welfare approach that does not feature a budget constraint leads to an emphasis on the welfare loss on exclusion error (poor identified as non-poor). The overall performance of the welfare approach is similar to the standard PMT with less exclusion error rate and more inclusion error rate.

3.2 Targeting Performance and Poverty Measure

Before I start with the welfare approach, it is important to explain the common measures of targeting. As mentioned earlier, two main measures of the targeting performance will be my focus. The first one is the Inclusion Error Rate (IER). Put it shortly, it is the fraction of the non-poor identified as poor among all the identified poor. It can be written as,

\[
IER = \frac{\sum w_{it} I\{poor|\hat{poor}\}}{\sum w_{it} I\{poor\}}
\]
where $\hat{poor}$ is the predicted status of poor and $poor$ is the true status. $w_{it}$ is the sample weights for household $i$ at time $t$. The IER does not directly address on the impact on poverty reduction as it focuses on the non-poor identified as poor while it is often an indication of the fiscal cost on targeting the wrong group.

The second one is the Exclusion Error Rate (EER). It is the fraction of the poor identified as non-poor among all poor. It can be written as,

$$EER = \frac{\sum w_{it} 1\{non\hat{\text{poor}}|poor\}}{\sum w_{it} 1\{poor\}}$$

where $non\hat{\text{poor}}$ is the predicted status of non-poor. EER does reflect the impact of the poverty reduction. A higher EER reflects a larger proportion of the targeting group not getting the appropriate treatment. In the empirical comparison between the welfare approach and the traditional approach, EER and IER will be the standard of targeting performance used.

The poverty measures will be standard in the literature, I use both a fixed poverty line and a fixed poverty rate. A fixed poverty line is defined by a fixed consumption threshold that defines poor and non-poor. Thus, households with predicted consumption per capita lower than the threshold will be qualified for the social program. On the other hand, a fixed poverty rate states that the government would only consider certain ratio of the populations for the program ranked by predicted consumption. For the fixed poverty line, given that I am using the cross-sectional survey data from India, I pick $\bar{\varepsilon} \equiv F^{-1}(0.3)$ or $\bar{\varepsilon} \equiv F^{-1}(0.7)$ as roughly 70% of Indian population lives under $\$2$ per day and
30% of the Indian population has less than $1.25 per day which are considered extremely poor. Correspondingly, the poverty rates used will be $H = 0.3$ or $H = 0.7$. The choices of poverty line and poverty rate have different results so I will present both results.

3.3 The Welfare Approach

In this section, I demonstrate how the welfare approach directly cope with the inclusion error rate and exclusion error rate and how it is linked to the weighted logistic regression. Here I assume an one-period economy with log utility for all households and a utilitarian social planner who has the social welfare function (SWF):

$$SWF = \int_i \log c_i dF.$$

3.3.1 Uniform Transfer Program

I first consider a classical uniform social transfer program that have been widely adopted in many countries. The uniform social transfer program will transfer a fixed amount $\bar{i}$ to the households with consumption below certain threshold $\bar{z}$ and zero if the consumption of the household is above it. The cash transfer program for each household $i$ can be written as,
Here \( t(c_i) \) is the cash transfer received for \( i \) where \( c_i \) is the true consumption of household \( i \). Using PMT, I obtain an estimation for consumption for each household, \( \hat{c}_i \). As a result, \( t(\hat{c}_i) \) is the transfer received using PMT. With this, I define the social welfare loss as the difference between the post transfer social welfare function with the true log consumption and the post transfer social welfare function with the predicted log consumption, which is written as,

\[
WL = \sum \left[ \frac{\Phi(c_i) \bar{t} y_i}{\text{Marginal Welfare Gain}} - \frac{\Phi(c_i) \bar{t} \hat{y}_i}{\text{Expected Marginal Welfare Gain}} \right]^2
\]

\[
= \sum \bar{t}^2 [\Phi(c_i) (y_i - \hat{y}_i)]^2
\]

\[
= \bar{t}^2 \sum \left[ \frac{1}{c_i} (y_i - \hat{y}_i) \right]^2
\]

Thus immediately, the welfare loss function becomes a weighted least square with weights as the inverse of consumption. Compared to the traditional PMT, the welfare approach would use the binary indicator of poor as dependent variable instead of using the continuous variable, log consumption. By design, it puts a higher weight on the exclusion error. I will compare the out-of-sample performance at the empirical part. The uniform transfer program will be the main part that I focus on in the empirical part.
3.3.2 Differential Transfer Program

Typically, one would discuss the transfer scheme for differential transfers. For demonstration purpose, here I discuss a simple differential transfer scheme just to show derivation of the welfare loss. I assume a simple transfer scheme that intends to provide households with consumption below some threshold \( \bar{z} \) to just reach the threshold post transfer, which can be written as,

\[
t(c_i) = \begin{cases} 
\bar{z} - c_i & \text{if } c_i \leq \bar{z} \\
0 & \text{if } c_i > \bar{z}
\end{cases}
\]

To calculate the welfare loss, I formulate the welfare loss in a slightly different way. For a utilitarian social planner, though the welfare weight for each household is the same but the marginal utility gain of transfer for the poor is higher than the non-poor when utility is strictly concave. In my simple case, the marginal welfare gain under log utility is defined as \( \phi(c_i) = \frac{1}{c_i} \). Notice that the welfare loss function in the uniform transfer scheme in the earlier subsection can be obtained through this formulation. I further assume that the social planner does not want to give extra money to make the poor household’s consumption to surpass the threshold due to some budget constraints which results in a quadratic welfare loss function. This will disincentives the planner to give poor household extra money by under-predict the consumption of the poor. Thus the welfare loss is written as,
\[
WL = \int [\phi(c_i) t(c_i) - \phi(\hat{c}_i) t(\hat{c}_i)]^2 dF
\]

\[
= \left( \int_{c_i \leq \bar{z}} 1\{\hat{c}_i \leq \bar{z}\} \left( \frac{\hat{c}_i - c_i}{c_i} \right)^2 dF \right. \\
\left. + \int_{c_i \leq \bar{z}} 1\{\hat{c}_i > \bar{z}\} \left( \frac{\bar{z} - c_i}{c_i} \right)^2 dF + \int_{c_i > \bar{z}} 1\{\hat{c}_i \leq \bar{z}\} \left( \frac{\bar{z} - \hat{c}_i}{\hat{c}_i} \right)^2 dF + \int_{c_i > \bar{z}} 1\{\hat{c}_i > \bar{z}\} \left( \log \bar{z} - \log c_i \right)^2 dF \right)
\]

The last line of the equation is obtained by using again the first order Taylor expansion when \(\hat{c}_i\) is reasonably close to \(c_i\). The welfare loss obtained is a variation of the quadratic loss of log consumption in econometric targeting that consists three parts: for the poor households identified correctly as poor, it focuses on accurately predicting the level of consumption, for the poor households identified as non-poor, the loss in the exclusion error and the last part is the welfare loss in the inclusion error. Again, econometric targeting using welfare loss is a variation of the weighted logistic regression.

### 3.4 Data

The data for assessing the performance of different PMT methods come from the Socio-Economic Survey in India in 2011. It is a cross-sectional representative data that consists roughly one hundred thousand households in seventy one districts in India. Consumption is measured in local currency Rupee. To rank the households in terms of their consumption level, I use the log consumption
per capita \( \log c_i/HH \), HH is the household size). Two poverty lines are chosen, corresponding to \( H = 0.3 \) and 0.7. The 70\% reflects the general poverty rate in India and 30\% indicates the rate for extreme poor in India. When using the traditional PMT, the log consumption per capita is used as the dependent variable. Under fixed poverty line, the poverty line is backed out by \( z \equiv F^{-1}(0.3) \) or \( z \equiv F^{-1}(0.7) \) using the distribution of surveyed consumption. The poverty line when using the fixed poverty rate is backed out by \( \hat{z} \equiv \hat{F}^{-1}(0.3) \) or \( \hat{z} \equiv \hat{F}^{-1}(0.7) \) where \( \hat{F} \) is the distribution of predicted dependent variable (consumption or probability of being poor).

3.5 PMT Comparison

PMT with welfare approach

The logistic regression for a household \( i \) can be written as,

\[
\log \frac{p_i}{1-p_i} = \alpha + \beta X_i + \epsilon_i
\]

where \( p_i \) is the probability of being poor, \( X_i \) are the covariates used to predict the odds being poor. Variables used include gender, family size, land ownership and land size, martial status and related information to the household head such as age, education level ect. Dummies for categorical variables such as religion, district, household size are also included. For the welfare approach, the logistic
regression is run with weights equal to the inverse of consumption per capita.

For comparison, a standard OLS regression with the log consumption per capita as the dependent variable and the same covariates $X_i$ is conduct. The EER and IER are calculated respectively for both the OLS regression and the weighted logistic regression. In practice, using inverse consumption will put too much weights on the extreme poor. To balance it out, I also provide the results using the log consumption per capita as weights. Table 1 shows EER and IER across all four regression models under a fixed poverty rate. Surprisingly, the baseline welfare approach is the worst among all four by giving an EER and an IER around 31% while the rest of the three seems to produce very close results in EER and IER at around 28% for a fixed poverty rate of 30%. Using the inverse of consumption does seem to be too extreme by putting too much weight on the poor. The adjusted welfare approach which use the inverse of the log consumption seems to be more reasonable but produce very similar results as standard OLS and Logit model.

Table 2 provides the comparisons among these four models under a fixed poverty line. Here the welfare approach seems to provide an extreme focus on the exclusion error and as a result enjoys the lowest exclusion error rate 10.33% for a poverty line that identifies the bottom 30%. However, that comes with a tradeoff by over-predicting the non-poor as poor. As a result, the inclusion error rate is 62.38%. The adjusted welfare approach seems to provide a similar performance compared to the traditional OLS method and the standard Logit model. The main difference is the welfare approach puts higher weights on the poorer households
Table 3.1: EER and IER Comparison under Fixed Poverty Rate

<table>
<thead>
<tr>
<th></th>
<th>EER</th>
<th>IER</th>
<th>EER</th>
<th>IER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Welfare</td>
<td>31.91</td>
<td>31.4</td>
<td>21.12</td>
<td>20.63</td>
</tr>
<tr>
<td>OLS</td>
<td>29.22</td>
<td>28.76</td>
<td>16.54</td>
<td>16.02</td>
</tr>
<tr>
<td>Logit</td>
<td>28.76</td>
<td>28.77</td>
<td>16.23</td>
<td>15.99</td>
</tr>
<tr>
<td>Adjusted Welfare</td>
<td>28.78</td>
<td>28.78</td>
<td>16.55</td>
<td>16.03</td>
</tr>
</tbody>
</table>

Note: this table shows the comparison of target performance in Exclusion Error Rate, Inclusion Error Rate among four PMT methods: weighted logit model with inverse consumption per capita as weights, standard OLS method, standard Logit model, weighted logit model with inverse log consumption per capita. Note that here a fixed poverty line is imposed at $\tilde{z} = F^{-1}(0.3)$ and $\tilde{z} = F^{-1}(0.7)$ respectively.

and as a result, tend to enjoy a lower EER by sacrificing some IER.

A few interesting observations after comparing the results under both scenario. First, the welfare approach, by design, focus too much on the welfare loss incurred from an exclusion of a poor from the transfer program while ignoring the fiscal cost of predicting the non-poor household as poor. In addition, using inverse of consumption per capita in practice does seem to be too extreme because the variation in absolute consumption is way too large and using the log consumption per capita does improve the tradeoff between the EER and IER to be more realistic.

As compared to the results in Brown et al. (2016), it seems that while the
Table 3.2: EER and IER Comparison under Fixed Poverty Line

<table>
<thead>
<tr>
<th></th>
<th>EER</th>
<th>IER</th>
<th>EER</th>
<th>IER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Welfare</td>
<td>10.33</td>
<td>62.38</td>
<td>0</td>
<td>29.56</td>
</tr>
<tr>
<td>OLS</td>
<td>38.27</td>
<td>21.14</td>
<td>15.21</td>
<td>16.75</td>
</tr>
<tr>
<td>Logit</td>
<td>34.27</td>
<td>23.07</td>
<td>11.33</td>
<td>18.78</td>
</tr>
<tr>
<td>Adjusted Welfare</td>
<td>33.28</td>
<td>24.14</td>
<td>9.43</td>
<td>19.78</td>
</tr>
</tbody>
</table>

Note: this table shows the comparison of target performance in Exclusion Error Rate, Inclusion Error Rate among four PMT methods: weighted logit model with inverse consumption per capita as weights, standard OLS method, standard Logit model, weighted logit model with inverse log consumption per capita. Note that here a fixed poverty line is imposed at $\bar{z} = F^{-1}(0.3)$ and $\bar{z} = F^{-1}(0.7)$ respectively.

standard OLS method tends to have very high EER in their practices, the simple welfare approach seems to raise a very high IER by design. It calls for a need to design a more balance welfare approach with budget constraint to strike a useful balance between the welfare loss in EER and the fiscal cost in IER. Also, a standard logit model which directly predicts the probability of being poor in my practice seems to do as good as the standard OLS method which is contrast with their results that using binary dependent variable perform worse. The welfare approach method could be taken as a poverty weighted least square method that is backed with a welfare foundation which would be easier to for interpretation and explanation on the choices of weights. In this sense, it increases
the transparency of PMT at some level.

3.6 Conclusion

In this paper, I explore the connection between the welfare loss of a transfer program and the application of proxy means testing. I show that in a simple example, the welfare loss incurred from PMT can be shown as a weighted logistic regression with inverse consumption as weights. As a result, it puts a higher weights on poorer households. In practice, given that it focuses too much at the lower end, the target performance of the baseline approach is worse than a simple logistic or OLS regression. When using a fixed poverty line, the welfare approach would exclusively intend to reduce the welfare loss incurred from the exclusion error by over-predicting the amount of poor households. As a result, the inclusion error rate is much higher. This is also reflected in design that the fiscal cost is not taken into account. An adjusted version of the baseline using the inverse of the log consumption per capita as the weights does improves the target performance to be similar as a standard Logit model under the fixed poverty rate. Under the fixed poverty line, the adjusted welfare method seems to be balancing the focus between EER and IER and in general, has a lower EER and higher IER by design. Overall, using the welfare approach does connects the welfare loss with the statistics loss function. However, it does not seems to improve the overall performance in a simple case. Future research that intend to derive the statistics
loss function from the welfare loss function should take into account the fiscal cost as a constraint to balance out the EER and IER.
APPENDIX A
CHAPTER 1 OF APPENDIX

Data Series

Data Construction for Firm Level Measure of Financial Constraints

This subsection describes how I process with the SEC filings in details.

1. I extract all the annual/quarterly reports from Loughran-McDonald’s website. Specifically, I use the “stage one parsed file” where they remove all the redundant characters such as HTMLChars, ASCIIEncodedChars, XMLChars, etc and include summary information at the beginning of each file. The annual/quarterly reports include all 10-K, 10-KSB, 10-K405, 10-KSB40, 10-Q and 10-QSB files from 1997q1 - 2016q4.

2. I extract the Management Analysis and Discussion section (MD&A) for each document and keep the ones with more than 5000 words in total and 1500 words in the MD&A section to avoid incomplete documents and extraction errors.

3. For sample selection, I follow closely with both Hoberg and Maksimovic (2014) and Buehlmaier and Whited (2018). I exclude the financial firms (SIC 6000 - 6999) and the regulated firms (SIC 4900 - 4949). In addition, to control
for coding errors, I remove observations with the following conditions: firm-quarter observations report short-term debt larger than total debt ($DLCQ > DLTTQ$); firm-quarter observations have a merger that accounts for more than 15% of the book value of its asset ($AQCY > 0.15 \times ATQ$); firm-quarter observations report zero or negative values for either total assets ($ATQ$), book equity ($PSTKQ + CSTKQ$) or sales ($SALEQ$).

4. A financial constraints count means firms mention a delay of investment and that is related to a debt financing need or equity financing need. Specifically, if firms report a delay of investment from the delay list and the debt financing need from the debt need or the equity financing need from the equity list is mentioned in the window sentences (one sentence above, same sentence or one sentence below), I have one financial constraints count for the quarter-firm observation.

5. For each document, I remove all the stop words (such as “the”, “a”, “an”, “in”), lower cases for all words and stem all words which are all standard data preprocessing steps. Next, I search if the words from the delay list occurs in any particular sentence. If they appear, I then search whether in the same sentence or one sentence above/below, the words from debt-focus list or equity-focus list also appear. I expand the word lists by different combinations of verb and noun phrases from the original list to increase variation. Otherwise, the financial counts will be too little and still sometimes inaccurate.

6. Merge with Compustat quarterly data using Central Index Key (CIK).
Data Construction for Firm Level Variables

Variables

The subsection describes the firm-level variables used in the empirical analysis of my paper which is based on Compustat Monthly Updates - Fundamentals Quarterly. The definition follows standard practices in the literature. (for example, Whited (1992); Gomes (2001); Eisfeldt and Rampini (2006); Clementi and Palazzo (2015), Ottonello and Winberry (2018).

1. Fixed capital stock: defined as \( k_{i,t} \) as the capital stock of firm \( i \) at time \( t \). For each firm, I set the first value of \( k_{i,t} \) to the level of gross property, plant and equipments in the first period of appearance in Compustat (\texttt{ppegtq}, item 118). For later periods, I compute the evolution of fixed capital using the net property, plant and equipments (\texttt{ppentq}, item 42). Specifically,

\[
    k_{i,t+1} = k_{i,t} + PPENTQ_{i,t+1} - PPENTQ_{i,t}
\]

If there is a missing observation of \texttt{ppentq} in between two periods, I interpolate linearly the missing value. If there are more than one consecutive missing observations, I don’t conduct any interpolation.

2. Total debt: defined as \( b_{i,t} \) as the total debt level of firm \( i \) at time \( t \). It is calculated as the sum of total long term debt (\texttt{DLTTQ}, item 51) and debt in current liabilities (\texttt{DLCQ}, item 45)
3. **Leverage ratio**: defined as the ratio of total debt ($DLTTQ + DLCQ$) to total assets ($ATQ$, item 44) which is consistent with Ottonello and Winberry (2018).

4. **Liquidity ratio**: defined as the ratio of cash and short term investment ($CHEQ$, item 36) to total assets ($ATQ$) which is consistent with Jeenas (2018).

5. **Firm age**: defined as the Compustat age which is the difference of current period with the first period when firm first appeared in Compustat.

6. **Real sales growth**: defined as the quarterly growth rate of sales ($SALEQ$, item 2) deflated with CPI ($CONSDEF$ from Federal Reserve Economic Data).

**Data Processing**

1. I winsorize firm characteristics data at level 1% and 99% for each year to mitigate the impact of the outliers.

2. I keep firms with investment spells with more than 10 quarters while Jeenas (2018) and Ottonello and Winberry (2018) keep firms with more than 40 quarters spells. This is because when merging the text-measured financial constraints with Compustat data, it will render a much smaller dataset and keeping only firms with more than 40 firms will result in a small dataset with around 15000 observations over a 15-year period. However, the baseline result is still robust if I only use firms with more than 40-quarter investment spells.
The Relationship of the Textual Measure of Financial Constraints with Proxies

To explore the link between the financial constraints measures and popular proxies used in macroeconomics literature, I regress my financial constraints measure on proxies and firm characteristics that are commonly believed to be related to financial constraints in the corporate finance literature, controlling for firm fixed effects and time fixed effects. Table A.1 shows that financial constraints are highly correlated with firm age, leverage ratio, liquidity ratio and cash flow ratio. The correlations are in the expected directions for most cases, although firm size doesn’t seem to play a role here once other proxies are controlled. For example, more constrained firms also have higher leverage ratio which is used as a proxy for financial constraints in Ottonello and Winberry (2018). However, the negative correlation between the financial constraints and the liquidity ratio (cash plus short term investment/total asset) is in contrast with Jeenas (2018) where he uses the negative liquidity ratio as an indicator of financial constraints. The contrast finding does not necessarily disagree with each other. As discussed in the main text, there is a U-shape in the liquidity ratio along the dimension of the financial constraints measure. The high liquidity ratio with low constrained firms is consistent with Jeenas (2018)’s argument that as firms grow out of financial constraints, they use more cash to fund their investment projects and have a higher liquidity ratio. What is not captured by his paper is that for the constrained firms, due to the difficulty of funding their projects from external financing and
high liquidity risks, firms are forced to hoard large amount of cash or liquid assets to fund their investment projects or hedge against liquidity shocks. Thus both constrained firms and unconstrained firms tend to have a higher liquidity ratio compared to the mid-constrained firms. In addition, the definition of financial constraints is different in his paper where he defines firms which finance their investment projects based mainly on external funds are financially constrained whereas firms fund the projects with mainly cash are financially unconstrained. In my paper, both highly constrained firms and low constrained firms rely on cash to fund investment projects, though the reasons are different. Overall, the proxies can only explain about 10 percent of the variations of the financial constraints measure and it increases to 45% once the firm fixed effects are controlled which captures permanent differences across firms.
Table A.1: Financial Constraints and Relation to Proxies

<table>
<thead>
<tr>
<th>Financial Constraints</th>
<th>Financial Constraints</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log of Firm Age</td>
<td>-0.0207*** (0.00533)</td>
</tr>
<tr>
<td>Log of Total Asset</td>
<td>0.000542 (0.00218)</td>
</tr>
<tr>
<td>Leverage Ratio</td>
<td>0.0200*** (0.00586)</td>
</tr>
<tr>
<td>Liquidity Ratio</td>
<td>0.0175** (0.00800)</td>
</tr>
<tr>
<td>EBITDA/Asset</td>
<td>-0.0337** (0.0165)</td>
</tr>
<tr>
<td>Dividend/Asset</td>
<td>0.104 (0.0764)</td>
</tr>
<tr>
<td>Tobin’s q</td>
<td>-0.000337 (0.000465)</td>
</tr>
<tr>
<td>Real Sales Growth</td>
<td>0.00124** (0.000497)</td>
</tr>
<tr>
<td>Constant</td>
<td>0.0899*** (0.0185)</td>
</tr>
<tr>
<td></td>
<td>0.107*** (0.0200)</td>
</tr>
<tr>
<td>Firm FE</td>
<td>Yes</td>
</tr>
<tr>
<td>Time FE</td>
<td>Yes</td>
</tr>
<tr>
<td>Observations</td>
<td>101,802</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.465</td>
</tr>
</tbody>
</table>

Notes: *** p<0.01, ** p<0.05, * p<0.1. Results from running the regression of financial constraints on a set of proxies of financial constraints and a set of firm characteristics that were used to measure financial constraints in previous literature, controlling for firm fixed effect and time fixed effect. Standard errors are two-way clustered by firms and quarters.
Robustness Check

In this section, the three graphs show the investment responses of more constrained firms upon a monetary policy shocks for three different specifications: (1) use firms with investment spells more than 40 quarters. (2) use alternative Federal Funds future shocks of simply summing the raw shocks within the same quarter. (3) use Romer and Romer (2004) shock.

Figure A.1: Heterogeneous Investment Dynamics using Firms with Investment Spells > =40 Quarters

Notes: The graph plots the scaled coefficients $\beta_{int}^h$ for each $h = 0, 1...10$ along with 95% confidence interval. Data covers firms with investment spells more than 40 quarters. The regression specification is $\log(k_{i,t+h}) - \log(k_{i,t-1}) = \alpha_i + \beta_{int}^h FC_{i,t-1} \times \epsilon_t + \Gamma X_{i,t-1} + \gamma d_{t+h} + \epsilon_{j,t+h}$. Confidence intervals are constructed based on two-way clustered standard errors at firms and quarters.
Figure A.2: Heterogeneous Investment Dynamics with Alternative FFR Shocks

Notes: The graph plots the scaled coefficients $\beta_h^{\text{int}}$ for each $h = 0, 1, ... 10$ along with 95% confidence interval. The monetary policy shock is the alternative aggregation of Federal Funds future shocks of simply summing the raw shocks in the same quarter. The regression specification is $\log(k_{i,t+h}) - \log(k_{i,t-1}) = \alpha_i + \beta_h^{\text{int}} FC_{i,t-1} \times \epsilon_q^t + \Gamma X_{i,t-1} + \gamma d_{t+h} + e_{j,t+h}$. Confidence intervals are constructed based on two-way clustered standard errors at firms and quarters.
Figure A.3: Heterogeneous Investment Dynamics with Alternative FFR Shocks

Notes: The graph plots the scaled coefficients $\beta_{int}^h$ for each $h = 0, 1...10$ along with 95% confidence interval. The monetary policy shock is the Romer and Romer (2004) shock. The regression specification is $\log(k_{i,t+h}) - \log(k_{i,t-1}) = \alpha_i + \beta_{int}^h FC_{i,t-1} \times \epsilon_i^t + \Gamma X_{i,t-1} + \gamma d_{i,t+h} + e_{j,t+h}$. Confidence intervals are constructed based on two-way clustered standard errors at firms and quarters.
### Table A.2: Summary Statistics of Federal Funds Future Shocks

<table>
<thead>
<tr>
<th></th>
<th>raw</th>
<th>weighted</th>
<th>sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>-0.0124</td>
<td>-0.0271</td>
<td>-0.0262</td>
</tr>
<tr>
<td>median</td>
<td>0</td>
<td>-0.0037</td>
<td>0</td>
</tr>
<tr>
<td>std</td>
<td>0.0664</td>
<td>0.0674</td>
<td>0.0944</td>
</tr>
<tr>
<td>min</td>
<td>-0.4125</td>
<td>-0.3274</td>
<td>-0.4094</td>
</tr>
<tr>
<td>max</td>
<td>0.1250</td>
<td>0.1314</td>
<td>0.1746</td>
</tr>
<tr>
<td>#</td>
<td>162</td>
<td>76</td>
<td>77</td>
</tr>
</tbody>
</table>

**Notes.** High frequency shocks are the raw measured interest rate surprises around the FOMC meetings. “Weighted” shocks are time aggregate quarterly shocks with the weighted average. “Sum” shocks are the time aggregate quarterly shocks with simply summing all shocks within the same quarter.
Table A.3: Correlations with Other Measures of Financial Constraints

<table>
<thead>
<tr>
<th></th>
<th>Deep Learning</th>
<th>Naive Bayes</th>
<th>Kaplan and Zingles</th>
<th>Whited and Wu</th>
<th>Hadlock and Pierce</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep Learning</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>0.3005</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kaplan and Zingles</td>
<td>0.1029</td>
<td>0.1413</td>
<td>1.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whited and Wu</td>
<td>-0.0021</td>
<td>0.0032</td>
<td>-0.0054</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>Hadlock and Pierce</td>
<td>0.1148</td>
<td>0.1387</td>
<td>0.3188</td>
<td>0.0074</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Notes: This table shows the correlations among my measure of financial constraints, measures with Naive Bayes method and other population measures estimated with accounting data.
Figure A.4: Distribution of Annual/Quarterly Reports

*Notes.* The graph shows the evolution of the annual report over time. 10KSB is the annual report form used by small business prior 2009. Both 10K405 and 10KSB40 are the forms used prior 2003 to indicate a failure of filing Form 3,4 or 5 which are used to disclose insider trading activity. Overall, the number of public firms in US peaked at early 2000 and continues to decline until 2018.
Notes. The graph plots the raw data on the federal funds future shocks and the monetary news shock. The federal funds future shock is measured as the changes of federal funds future rates around the 30-minute window of the FOMC meetings. Data is taken directly from Nakamura and Steinsson (2018).
Figure A.6: Quarterly Monetary Shocks: Weighted vs Simple Sum

Notes. The graph plots the time aggregated quarterly federal funds future shock and the monetary news shock. The red solid line represents the smoothed/weighted quarterly shock while the blue dot dash line represents the monetary shock aggregated by simple sum within each quarter. The data series covered is from 1996q1 to 2014q1. The raw shocks are taken from Nakamura and Steinsson (2018).
APPENDIX B

CHAPTER 2 OF APPENDIX

B.1 Calibration Details

B.1.1 Labor Earnings

In order to match the top labor earnings share in the data, I followed Hubmer, Krusell and Smith (2016) by adding a Pareto tail for the labor earnings distribution. Pareto coefficient $\kappa$ is calibrated such that the top labor earnings share match closely to the data in Piketty and Saez (2003). Noted that 

$$F_{\text{Pareto}(\kappa_t)}^{-1}\left(\frac{F_{\text{Pareto}(\kappa_t)}(p_t)-0.9}{1-0.9}\right) \neq \exp(p_t) \text{ when } p_t = 0.9$$

in general and often 

$$F_{\text{Pareto}(\kappa_t)}^{-1}\left(\frac{F_{\text{Pareto}(0.9)}(p_t)-0.9}{1-0.9}\right) < \exp(0.9)$$

which results in non-monotonicity in persistent earnings $\psi(p_t)$. To fix that, I scale the Pareto tail by a factor of 

$$s = \frac{\exp(0.9)}{F_{\text{Pareto}(\kappa_t)}^{-1}\left(\frac{F_{\text{Pareto}(0.9)}(p_t)-0.9}{1-0.9}\right)}$$

such that monotonicity is recovered. Given labor supply is endogenous, I divide the labor earnings by a fixed shrinking factor 1.6 such that at 1967 steady state, the endogenous labor earnings volatility matches the data.

B.1.2 Tax System

Tax data are taken from Piketty and Saez (2007) where they estimates the federal effective tax rates from 1960-2000. The effective average tax rate comprises four
major federal taxes: individual income, corporate income, estate and gift, and payroll taxes. Effective average tax rates are calculated for 11 income group (percentage: 0-20, 20-40, 40-60, 60-80, 80-90, 90-95, 95-99, 99-99.5, 99.5-99.9, 99.9-99.99, 99.99-100). Tax function are calibrated as a step-wise function into the model. For each income bracket, the threshold is set to match its income share and marginal tax are imputed to match its average tax rate in the data.

Detailed calibration strategy of tax function:

1. Approximate the cdf of income distribution by a piecewise linear function \( f(y) \) with 12 nodes. It implies that within each income bracket, the pdf is uniformly distributed.

2. Fix the starting lowest income threshold \( y_0 \) to be 0. Use the income share \( y_{ishare} \), the corresponding population share \( y_{ipshare} \) and the aggregate level of income \( \bar{y} \), one could back out the first threshold \( y_1 \) solving

\[
\frac{1}{2}(y_1 + y_0) \times y_{1ipshare} = y_{1ishare} \times \bar{y}
\]

3. Similarly, \( y_2, y_3, ..., y_{11} \) could be solved sequentially using \( \frac{1}{2}(y_{i+1} - y_i) \times y_{i+1ipshare} = y_{i+1ishare} \times \bar{y} \)

4. Once the income thresholds are obtained, calibrate the marginal tax within each income group such that average income tax are matched with the data.

5. Denote the marginal tax rate within each income group as \( \{m_i\}_{i=1}^{11} \), average tax rate for each income group as \( \{\tau_j\}_{j=1}^{11} \). Solving \( m_i \) is solving a system of linear equations with 11 unknowns and 11 equations
6. Obtain the marginal income tax for each income group by simply

\[
\int_{y_0}^{y_1} \frac{(y_i - y_0)m_1}{y_i} \frac{1}{y_1 - y_0} dy_i = \tau_1 \\
\int_{y_1}^{y_2} \frac{(y_i - y_0)m_1 + (y_i - y_1)m_2}{y_i} \frac{1}{y_2 - y_1} dy_i = \tau_2 \\
\int_{y_2}^{y_3} \frac{(y_i - y_0)m_1 + (y_i - y_1)m_2 + (y_i - y_2)m_3}{y_i} \frac{1}{y_3 - y_2} dy_i = \tau_3 \\
\ldots \ldots \\
\int_{y_{10}}^{y_{11}} \frac{(y_i - y_0)m_1 + (y_i - y_1)m_2 + (y_i - y_2)m_3 + \ldots + (y_{10} - y_0)m_{10} + (y_i - y_{10})m_{11}}{y_i} \frac{1}{y_{11} - y_{10}} dy_i = \tau_{11}
\]

rewrite it into matrix multiplication form as \(AM = T\) where \(A\) is the lower triangular matrix, \(M\) is the vector of \(m_i\) and \(T\) is the vector of \(\tau_j\) as following

\[
\begin{bmatrix}
\int_{y_0}^{y_1} \frac{(y_i - y_0)}{y_i} \frac{1}{y_1 - y_0} dy_i & 0 & \ldots & 0 & 0 \\
\int_{y_1}^{y_2} \frac{(y_i - y_0)}{y_i} \frac{1}{y_2 - y_1} dy_i & \int_{y_1}^{y_2} \frac{(y_i - y_1)}{y_i} \frac{1}{y_2 - y_1} dy_i & \ldots & 0 & 0 \\
\int_{y_2}^{y_3} \frac{(y_i - y_0)}{y_i} \frac{1}{y_3 - y_2} dy_i & \int_{y_2}^{y_3} \frac{(y_i - y_1)}{y_i} \frac{1}{y_3 - y_2} dy_i & \int_{y_2}^{y_3} \frac{(y_i - y_2)}{y_i} \frac{1}{y_3 - y_2} dy_i & \ldots & 0 \\
\ldots & \ldots & \ldots & \ldots & \ldots \\
\int_{y_{10}}^{y_{11}} \frac{(y_i - y_0)}{y_i} \frac{1}{y_{11} - y_{10}} dy_i & \int_{y_{10}}^{y_{11}} \frac{(y_i - y_1)}{y_i} \frac{1}{y_{11} - y_{10}} dy_i & \int_{y_{10}}^{y_{11}} \frac{(y_i - y_2)}{y_i} \frac{1}{y_{11} - y_{10}} dy_i & \ldots & \int_{y_{10}}^{y_{11}} \frac{(y_i - y_{10})}{y_i} \frac{1}{y_{11} - y_{10}} dy_i
\end{bmatrix} \begin{bmatrix}
m_1 \\
m_2 \\
m_3 \\
\vdots \\
m_{11}
\end{bmatrix} = \begin{bmatrix}
\tau_1 \\
\tau_2 \\
\tau_3 \\
\vdots \\
\tau_{11}
\end{bmatrix}
\]

6. Obtain the marginal income tax for each income group by simply \(M = A^{-1}T\) for each year from 1967-2001. Note that when putting the stepwise tax function into the model, I assumed that any income lower than \(y_0\) will have a 0 marginal tax and any income higher than \(y_{11}\) will be taxed at the highest
marginal tax rate $m_{11}$. Also the level of $\bar{y}$ only affects the thresholds but have no impact on imputed marginal tax rate.

7. Estimate the tax progressivity using $T(y) = y - \lambda y^{1-\tau}$ such that the average and marginal tax rate matches data.

B.2 Computational Algorithms

B.2.1 Algorithm for Solving Equilibrium

1. Make initial guesses of the interest rate $r_0$.

2. Assume an aggregate labor supply $L$, along with $r_0$, compute the wage rate $w$ and aggregate capital demand $K_d$.

3. Use the given $K_d$ to back out the cut-off thresholds for each income group (11 income groups) and set the marginal tax rate such that the average tax rate in each bracket matches the data. Lump sum transfer $T$ is funded by a fixed share of tax revenue $T = \lambda \tau (wL + r_0K_d)$.

4. Use policy function iteration, endogenous grid method and first order condition on labor supply to compute the policy function $k_t = g(x_t, p_t, \beta_t)$, $l_{t+1} = f(x_{t+1}, p_{t+1}, \beta_{t+1})$ and $l_t = f(x_t, p_t, \beta_t)$.

5. Compute the aggregate labor supply $L^*$, repeat step 2 until the gap between $L$ and $L^*$ is small enough.
6. Compute the steady state distribution of asset for the year 1967 by approximating the conditional CDF of assets $G_{j,m}(k_i)$

7. Use $G_{j,m}(k_i)$ to compute aggregate asset supply $K_s$ and the implied corresponding interest rate $r_1 = F'(K_s, L) - \delta$.

8. If $\text{norm}|r_1 - r_0| < \epsilon$ then stop. Otherwise use bisection method to update the guess $r_0$ and repeat step 2.

B.2.2 Endogenous Grid Method with Policy Function Iteration

Algorithms for EGM method with policy function iteration

1. Create grids for $\{a_{t+1}, p_j, \beta_m\}$ where $a_{t+1}$ is the asset holding at $t + 1$, $p_j$ is the persistent component of labor earnings at $t$ and $\beta_m$ is the discount factor at time $t$. It is called endogenous grid method because we only iterate over next period cash-on-hand $x_{t+1}$ so that the grid of $x_t$ is determined endogenously.

   (a) Here I pick the 100 grids for $a_{t+1}$ as $\{a_i\}_{i=1}^{100}$ from borrowing constraint $a_{\text{min}}$ to a large number $a_{\text{max}}$. $a_{\text{max}}$ is large enough that at equilibrium $Pr(a_i < a_{\text{max}}|p_j, \beta_m) = 1$ is insured. In the model I pick $a_{\text{max}}$ such that it is approximately 1 million times larger than the average asset level.

   (b) The grid for the persistent component of individual productivity $(p_j)_{j=1}^{17}$ is chosen to account for the long right tail in earnings. First, grid points are picked as 0.0001, 0.01, 0.1, 0.25, 0.5, 0.75, 0.9, 0.925, 0.95, 0.975, 0.99,
0.999, ..., and 0.99999999 quantiles of the unconditional \( p \)-distribution.

Second, compute the corresponding efficiency units of labor \((\psi(p_1), ..., \psi(p_{17}))\). Third, compute the Gauss-Hermite integration weights for \( \omega_{j',j}^p \) which is a 17X17 matrix.

(c) Choose the grid points \( \{\beta_m\}_{m=1}^{15} \) as the Gauss-Hermite quadrature points of the unconditional \( \beta \)-distribution. Calculate the Gauss-Hermite integration weights for \( \beta_m' \mid \beta_m \) which is a 15X15 matrix using similar method above.

(d) Computation of \( \omega_{p,j'}^p \mid p_j \) and \( \omega_{m',m}^\beta \): Consider the case of persistent component of earning \( p \). First obtain the Gauss-Hermite quadrature nodes \( \{x_{pn}\}_{n=1}^{15} \) and the associated integration weights \( \{\tilde{\omega}_n\}_{n=1}^{15} \). Conditional on the previous value being \( p_j \) for some fixed \( j \in \{1,...,17\} \), the conditional value of persistent component follows a normal distribution, \( p_{j' \mid j} \sim N(\rho^p p_j, (\sigma^p)^2) \). Given 15 nodes from Gauss-Hermite quadrature, we will evaluate \( \psi(p) \) at \( \{\tilde{p}_n\}_{n=1}^{15} \) for \( j' \), where \( \tilde{p}_n = \rho^p p_j + \sqrt{2}\sigma^p x_{pn} \). In general, \( \tilde{p}_n \) will not coincide with \( p_j \)-grid so we need to interpolate to obtain the integration weights \( \omega_{j',j}^p \) at \( p_j \)-grid. For \( n = 1,...,15 \), locate \( j'(n) \) such that \( p_{j'(n)} \leq \tilde{p}_n \leq p_{j'(n)+1} \) and compute the linear interpolation coefficient in \( \psi(p) \)-space \( \lambda_n \) as

\[
\lambda_n = \frac{\psi(\tilde{p}_n) - \psi(p_{j'(n)})}{\psi(p_{j'(n)+1}) - \psi(p_{j'(n)})}
\]

Then looping over \( n = 1, ..., 15 \), add \( (1 - \lambda_n) \frac{1}{\sqrt{\pi}} \tilde{\omega}_n \) to \( \omega_{j'(n),j}^p \) and \( \lambda_n \frac{1}{\sqrt{\pi}} \tilde{\omega}_n \) to \( \omega_{j'(n)+1,j}^p \). The construction of the integration weights for \( \beta_{m',m} \) is
analogous, except that using $\beta$-space for linear interpolation directly.

2. Make an initial guess of $x_t$ and the policy function, $c_t = g^0(x_t, p_j, \beta_m)$. Here I assume that agent consumes as much as they can $c_t = x_t + |a_{min}|$. Note that I use piecewise linear function for $g^0$.

3. Cash-on-hand $x_{t+1}$ at $t + 1$

$$x_{t+1} = a_{t+1} + y_{t+1} - \tau(y_{t+1}) + T$$

$$= a_{t+1} + \psi(p_j') \exp(v_{t+1}) + r\eta_{t+1}a_{t+1} - \tau(w\psi(p_j) \exp(v_{t+1}) + r\eta_{t+1}a_{t+1}) + T$$

where $v_{t+1}$ is the transitory component of labor earnings at $t + 1$ and $\eta_{t+1}$ is the idiosyncratic i.i.d shock to capital returns, $p_j'$ is the persistent component of labor earnings at $t + 1$.

4. We could calculate directly of the right-hand-side value of Euler equation along with the first order condition on labor supply and back out $c_t$ and $x_t$ without using a root-finding algorithms.

$$u'(c_t) = \beta_mE_t(1 + r\eta_{t+1}(1 - \tau'(y_{t+1})))u'(c_{t+1})$$

$$c_t^{-\sigma} = \beta_mE_t(1 + r\eta_{t+1}(1 - \tau'(y_{t+1})))c_{t+1}^{-\sigma}$$

$$c_t^{-\sigma} = \beta_m \int_p \int_\beta \int_v \int_\eta (1 + r\eta_h(1 - \tau'(y_{t+1})))c_{t+1}^{-\sigma} d\Gamma_\eta(\eta)d\Gamma_v(v)d\Gamma_\beta(\beta'|\beta_m)d\Gamma_p(p'|p_j)$$

$$(x_t - a_{t+1})^{-\sigma} = \beta_m \sum_{j'} \sum_{j} \sum_{q} \sum_{m} \sum_{\beta} \sum_{\eta} \sum_{\psi} \sum_{\omega}(1 + r\eta_h(1 - \tau'(y_{t+1}))) (x_{t+1} - g^0(x_{t+1}, p_j', \beta_m'))$$
\[ \chi l_t^s = c_t^{-\sigma} y_t (1 - T'(y)) \]

In line 4 we use the Gauss-Hermite quadrature to calculate the integration. Call the right-hand-side of the Euler equation \( RHS \), then \( c_t = RHS^{-1/\sigma} \) and \( x_t = c_t + a_{t+1} \).

Note: When interpolating to obtain \( c_{t+1} = g^0(x_{t+1}, p_j', \beta_m') \), we need to consider the case for binding borrowing constraint. For any \( x_{t+1}^* \leq x_{low} \), borrowing constraint is binding and \( c_{t+1}^* = x_{t+1}^* - a_{min} \) where \( x_{low} \) is largest cash-on-hand that will induce borrowing constraint to bind and is obtained from initial guess or last iteration.

5. Update the policy function using \( c_t \) and \( x_t \) obtained, \( c_t = g^1(x_t, p_j, \beta_m) \).

6. If \( |\text{norm}(g^1 - g^0)| < \epsilon \), then stop. Otherwise repeat step 4 with policy function \( g^0 = g^1 \).

### B.2.3 Algorithm for Computing Ergodic Distribution

1. For all \( p_j, \beta_m, v_q \) and \( \eta_h \) and asset grid \( \{a_i\}_{i=1}^{100} \) (same grid as \( a_{t+1} \)), there exists a unique level of asset holdings \( a = s^{-1}(a_i; p_j, \beta_m, v_q, \eta_h) \) such at the saving \( a_i \) is optimal.

\( s^{-1}(a_i; p_j, \beta_m, v_q, \eta_h) \) is defined as the unique \( a \) that solves

\[ x(a_i; p_j, \beta_m) = a + y - \tau(y) + T \]
where \( y = \tau \eta_h a + w \psi(p_j) \exp(v_q). \)

Here we need a root-finding algorithms to solve \( a \) and I used bisection method.

2. We define a much finer grid for asset holdings \((k_i)_{i=1}^{1000}\) and interpolate (linearly) to find the inverse saving function \( s^{-1}(a_i; p_j, \beta_m, v_q, \eta_h). \)

3. Solve for the conditional cdf for the ergodic distribution \( G(k_i; p_j, \beta_m) \equiv \operatorname{Prob}(k < k_i | p = p_j, \beta = \beta_m) \) at grid \((k_i)_{i=1}^{1000}\), \((p_j)_{j=1}^{17}\) and \((\beta_m)_{m=1}^{15}\). Denote \( G_{j,m}(k_i) \) as \( G(k_i; p_j, \beta_m) \) for simplification. The conditional cdf has to satisfy

\[
G_{j,m}(k_i) = \int_p \int_\beta \int_v \int_\eta G(s^{-1}(k_i; p, \beta, v, \eta); p, \beta) \Gamma_\eta(\eta) d\Gamma_v(v) d\Gamma_\beta(\beta | \beta_m) d\Gamma_p(p | p_j)
\]

where \( p_j \) and \( \beta_m \) are realizations in \( t + 1 \) and integration is over the shock values in period \( t \). Given \( p_j \) and \( \beta_m \) follows an AR(1) process, the conditional random variables \( z_t | z_{t+1} \) and \( z_{t+1} | z_t \) follows the same distribution. Starting from some initial distribution \( G_{j,m}^0(k_i) \), we update until convergence according to

\[
G_{j',m'}^1(k_i) = \sum_{j,j'} \omega_{j',j} \sum_{m,m'} \omega_{m,m'}^\beta \sum_q \omega_q^v \sum_h \omega_h^\eta G_{j,m}^0(s^{-1}_{j,m,q,h}(k_i))
\]

where \( G_{j,m}^0(s^{-1}_{j,m,q,h}(k_i)) \) is obtained by linear interpolation. Here given \( z_t | z_{t+1} \) and \( z_{t+1} | z_t \) follows the same distribution, the conditional integration weights for \( p_{j'} \) and \( \beta_{m'} \) is the same as the ones obtained in solving Euler equation.
B.2.4 Algorithm for Computing Aggregate Asset Supply and Aggregate Labor Supply

1. Aggregate wealth supplied is given as

\[ K_s = \int_p \int_\beta (\int_k k dG(k|p, \beta)) d\Gamma_p(p) d\Gamma_\beta(\beta) \]

where \( \Gamma_p(p) \) the unconditional cdf of \( p \) and \( \Gamma_\beta(\beta) \) is the unconditional cdf of \( \beta \).

2. We integrate according to

\[ \hat{K}_s = \sum_{j=1}^{17} \bar{\omega}_j^p \sum_{m=1}^{15} \bar{\omega}_m^\beta \left( k_1 G_{j,m}(k_1) + \sum_{i=2}^{1000} \frac{k_{i-1}+k_i}{2} (G_{j,m}(k_i) - G_{j,m}(k_{i-1})) \right) \]

where \( \bar{\omega}_m^\beta \) is the associated Gauss-Hermite quadrature weights given \( (\beta_m)_{m=1}^{15} \) is chosen as the Gauss-Hermite sample points. To calculate the integration weights for \( p_j \): (1) define a very fine grid of \((p_k)_{k=1}^{100000}\) that covers all \( p_j, j = 1, 2, ..., 17 \); (2) for all \( k = 1, 2, ...100000 \), locate \( j(n) \) and \( \lambda_n \) as above; (3) looping over \( k = 1, 2, ..., N \), add \( 1 - \lambda_n f_p(p_k) \) to \( \bar{\omega}_j^{p(n)} \) and \( \lambda_n f_p(p_k) \) to \( \bar{\omega}_j^{p(n)+1} \) where \( f_p(.) \) is the pdf of \( p \sim N(0, \frac{(\sigma^p)^2}{1-\rho^p^2}) \). (4) finally, normalize such that \( \sum_{j=1}^{17} \bar{\omega}_j^p = 1 \).

Note: Theoretically, applying Gauss-Hermite quadrature should also obtain the same results. But given a concentration on the right tail of earning distribution, using Gauss-Hermite quadrature will often end up a lot of 0s for many \( \bar{\omega}_j^p \) and large calculation error.
3. Similarly, we could solve the aggregate labor supply in a similar manner. The only extra state that needs to be taken into account is $v$.

**B.2.5 Algorithm for Transition Dynamics**

1. Compute the initial steady state $(K^*, T^*)$ and the new steady state $(K^{**}, T^{**})$. We want to solve the transition path from initial steady state to the new steady state.

2. We assume a large number $\Delta t = t_1 - t_0$ such that after $\Delta t$ period $(K_{t_1}, T_{t_1}) \approx (K^{**}, T^{**})$. The task becomes searching for a fixed point in $(K_t, T_t)_{t=t_0+1}$ space.

3. Make an initial guess of $(K^0, T^0)_{t=t_0+1}$ with $K^0_{t_1} = K^{**}$ and $T^0_{t_1} = T^{**}$. The corresponding subsequent prices $\{r_t, w_t\}_{t=t_0+1}$ are calculated as well.

4. Starting from time $t_1$, using the Euler equation and endogenous grid method to iterate backward, the policy functions $c_t = t(x_t, p_j, \beta_m)$ at time $t = t_1 - 1, t_1 - 2, ..., t_0 + 1$ are obtained given prices.

5. Starting from the initial ergodic distribution $G_{jm}(k_i)_{t_0}$ and starting prices, iterate forward using the policy functions solved above and calculate the aggregate asset and corresponding transfers $(\hat{K}_t, \hat{T}_t)_{t=t_0+1}$.

6. Compare $(K^0, T^0)_{t=t_0+1}$ with $(\hat{K}_t, \hat{T}_t)_{t=t_0+1}$, stop if differences are small. Otherwise update a new guess $(K^1, T^1)_{t=t_0+1}$ and repeat step 4.
Aiyagari, S. R.

Bakış, O., B. Kaymak, and M. Poschke

Benhabib, J. and A. Bisin

Benhabib, J., A. Bisin, and M. Luo

Benhabib, J., A. Bisin, and S. Zhu

Bernanke, B. S., M. Gertler, and S. Gilchrist

Bewley, T.
1986. Stationary monetary equilibrium with a continuum of independently
fluctuating consumers. *Contributions to mathematical economics in honor of Gérard Debreu*, 79.

Bodnaruk, A., T. Loughran, and B. McDonald

Brown, C., M. Ravallion, and D. Van de Walle

Buehlmaier, M. M. and T. M. Whited

Cagetti, M. and M. De Nardi

Cameron, L. and M. Shah

Clementi, G. L. and B. Palazzo

Cloyne, J., C. Ferreira, M. Froemel, and P. Surico

Cochrane, J. H. and M. Piazzesi

Conesa, J. C., S. Kitao, and D. Krueger

Conesa, J. C. and D. Krueger

Cook, T. and T. Hahn

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova

Domeij, D. and J. Heathcote
Eisfeldt, A. L. and A. A. Rampini

Farhi, E. and I. Werning

Farre-Mensa, J. and A. Ljungqvist

Fehr, H. and F. Kindermann


Gertler, M. and S. Gilchrist

Golosov, M., N. Kocherlakota, and A. Tsyvinski
Golosov, M., M. Troshkin, and A. Tsyvinski

Gomes, J. F.

Grosh, M. E.

Hadlock, C. J. and J. R. Pierce

Han, S. and J. Qiu

Heathcote, J., K. Storesletten, and G. L. Violante

Heathcote, J., K. Storesletten, and G. L. Violante
Hoberg, G. and V. Maksimovic

Hubmer, J., P. Krusell, and A. A. Smith Jr

Huggett, M.

Jeenas, P.

Jordà, Ò.

Kaplan, G., B. Moll, and G. L. Violante

Kaplan, S. N. and L. Zingales

Kaymak, B. and M. Poschke

Kidd, S. and E. Wylde

Krueger, D. and A. Ludwig

Krusell, P. and A. A. Smith, Jr

Kuttner, K. N.

Lamont, O., C. Polk, and J. Saaá-Requejo

Nakamura, E. and J. Steinsson

Ottonello, P. and T. Winberry

Pedroni, M. Z., S. Dyrda, et al.

Pennington, J., R. Socher, and C. Manning

Piketty, T. and E. Saez

Piketty, T. and E. Saez
Pitschner, S.

Romer, C. D. and D. H. Romer

Roopnaraine, T. and M. Adato

Saez, E. and G. Zucman

Smith, M., O. Zidar, and E. Zwick

Tenreyro, S. and G. Thwaites

Thoma, M. A.

Vavra, J.

Whited, T. M.

Whited, T. M. and G. Wu

Wong, A.

Yang, M., W. Tu, J. Wang, F. Xu, and X. Chen

Zhang, H.