Essays on Econometrics and Economic Theory

by

Kevin Kainan Li

B.S., California Institute of Technology (2015)

Submitted to the Department of Economics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2021

Author . . . .
Department of Economics
January 8, 2021

Certified by . . . .
Alberto Abadie
Professor of Economics
Thesis Supervisor

Certified by . . . .
Drew Fudenberg
Paul A. Samuelson Professor of Economics
Thesis Supervisor

Accepted by . . . .
Amy Finkelstein
John & Jennie S. MacDonald Professor of Economics
Chairman, Department Committee on Graduate Theses

Essays on Econometrics and Economic Theory

by

Kevin Kainan Li

Submitted to the Department of Economics
on January 8, 2021, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

This thesis consists of three chapters. The first chapter extends the recent result of Athey and Wager on the asymptotic normality of random forest estimators to a multivariate setting; in particular, we examine stability properties and bounds for certain classes of tree estimators, and provide guidance and heuristics that make our results useful for practitioners. The second chapter studies a continuous-time principal-agent model for risky projects in which success is binary and not quantifiable. We consider optimal incentive contracts which feature two components: a flow payment that is paid out at each moment and a lump-sum bonus that is paid on project success. We characterize the optimal solution and calibrate our model to data on executive compensation. The final chapter studies speed and competition in trading venues, in particular those for illiquid markets. We show that differences in trading speeds among different venues lead to endogenous market segmentation, thus increasing trading volume and overall liquidity. We also examine equilibria and various notions of welfare in an entry game in which an entrant has the option to create a faster trading venue and compete against an incumbent.

JEL Classifications C14, D82, G14

Thesis Supervisor: Alberto Abadie
Title: Professor of Economics

Thesis Supervisor: Drew Fudenberg
Title: Paul A. Samuelson Professor of Economics

Acknowledgments

It has been a privilege to spend the past five years at MIT. I benefited greatly, both personally and professionally, from my mentors, colleagues, and friends in the Department of Economics, as well as the broader MIT and Cambridge community.

I owe a great deal of thanks to Alberto Abadie: his support and encouragement carried me through the most difficult parts of the PhD program. Alberto was also very generous with his time, and our weekly meetings were tremendously helpful. Drew Fudenberg is another mentor who was also very generous with his time and wisdom: his comments improved every part of the second and third chapters of this thesis. Finally, I would like to thank Victor Chernozhukov, who not only introduced me to econometrics, but was also a constant source of support throughout my PhD career; I am very grateful for his many pointers to avenues of exciting research.

Ali Kakhbod, my coauthor on two of the following chapters, deserves special thanks: our collaboration made this thesis possible, and I could not have done it without him.

I want to thank my cohort as well as other faculty. I learned a lot from the weekly econometrics lunches and seminars, and in particular I would like to thank Max Cytrynbaum, Ben Deaner, Mert Demirer, Yaroslav Mukhin, and Sophie (Liyang) Sun for helpful discussions. I would also like to thank Benjamin Golub, Anna Mikusheva, Whitney Newey, Philippe Rigollet, and Juuso Toikka for memorable conversations and classes.

My friends were an important part of my MIT journey. I would like to thank Sophie Sun for her friendship. I am also very fortunate that my best friends at Caltech, Jake Wellens (MIT) and Jake Marcinek (Harvard University), came to Cambridge with me to pursue their PhDs. It was also a joy to see John Vaughen (Stanford University) and Eric Chen (University of Illinois at Chicago) each time I go home during the holidays: their friendship was a source of happiness during high school and college, and I needed it all the more so while studying at MIT. I also thank their families' hospitality for our numerous holiday gatherings.

I would also like to thank my advisors at Caltech: Professor Matthew Elliott, Professor Federico Echenique, and Professor Kota Saito. Their support and encouragement made pursuing a graduate degree in economics an easy decision. I also want to thank Professor Kim Border—I am saddened to learn of his recent passing. Going back further, I want to thank my math teacher Dr. Ioana Boca at University Laboratory High School, as well as Professor Daniel Shapiro at The Ohio State University for running the Ross Mathematics Program.

My deepest gratitude go to my parents Xinzhong (Jeff) Li and Li (Lisa) Huang and my sister Jade Ruiyang Li for their unwavering support, patience, and love. My parents are first-generation Chinese-Americans, and their perseverance and kindness set a model that I try to live up to every day: this thesis is dedicated to them.

Contents

Chapter 1

Asymptotic Normality for Multivariate Random Forest Estimators

1.1 Introduction

Trees and random forests are non-parametric estimators first introduced by Breiman (2001). Given a feature space $\mathcal{X} \subset \mathbf{R}^p$ and a set of data points $\{(X_i, Y_i)\} \subseteq \mathcal{X} \times \mathbf{R}$, tree estimators recursively partition the feature space into axis-aligned non-overlapping hyperrectangles[1] by repeatedly splitting $\mathcal{X}$ along a given axis. The prediction of the tree estimator at a test point $x \in \mathcal{X}$ is then an aggregate of the targets Y_i's that land in hyperrectangle containing x; when Y_i is continuous, the aggregate is the sample mean and the tree is known as a regression tree. The depth of a tree estimator—defined as the maximal number of splits taken before reaching a terminal hyperrectangle—characterizes its complexity. There are two popular methods for controlling complexity: boosting and bagging. The boosting approach iteratively fits shallow trees to the previous step's residuals, starting with the target function for the initial tree. Complexity is reduced by using a shrinkage factor at each step, as well as by trimming the final tree (i.e., so that predictions are made at non-terminal hyperrectangles). The bagging approach instead grows a collection of deep trees on different subsets of the data, and averages over those trees for the final prediction. The intuition for bagging is that trees grown on different subsets are not perfectly correlated, so that aggregation reduces variance and balances the bias-variance tradeoff. Estimators of the latter type are called

I would like to thank my advisors Alberto Abadie and Victor Chernozhukov for reviewing multiple drafts of this paper. In addition, Sophie (Liyang) Sun, Ben Deaner, participants in the seminar class 14.386 (Spring 2020), and participants of MIT's Econometrics Lunch seminar provided very useful feedback. I would also like to thank Professor Stefan Wager for helping me understand mechanics of covariance estimates in random forest models.

[1] When the feature space $\mathcal{X}$ need not be rectangular, one may always enlarge $\mathcal{X}$ to a rectangular set $\mathcal{X}'$ that is defined to the intersection of all rectangular sets containing $\mathcal{X}$.

random forests, and they are the focus of this paper.

Since their introduction in the early 2000s, random forests have become an increasingly important tool in applied data analysis, owing to a multiple of practical advantages over competing models. First, high-quality random forest libraries are readily available, with popular implementations that scale to hundreds of distributed workers as in Ke et al. (2017); Chen and Guestrin (2016). Moreover, the core algorithm behind tree estimators and random forests are simple enough to allow for rapid prototyping of bespoke implementations, e.g. Athey, Tibshirani and Wager (2017). Another advantage of tree-based methods is that they can ingest real-world data without much issue: continuous, discrete, and ordered categorical features may be freely mixed[2] (c.f., Prokhorenkova et al. (2018)), model estimates are immune to feature outliers (c.f., Biau and Scornet (2015); Maniruzzaman et al. (2018)), and missing data may be easily incorporated (see Tang and Ishwaran (2017) for a survey). The construction of trees also naturally aligns with the spatial locality found in most real world target functions, in which the underlying relationship between Y and X is continuous. Finally, tree models are interpretable, with well-defined notions of feature importance (Gregorutti, Michel and Saint-Pierre (2017); Strobl et al. (2008)) that support their use as model selection tools (Genuer, Poggi and Tuleau-Malot (2010)).

Within economics, random forests may be fruitfully applied to estimate heterogeneous treatment effects. In the potential outcomes framework of Rubin (1974) (see Imbens and Rubin (2015) for an overview), an individual i is associated with two potential outcomes $Y^{(0)}$ and $Y^{(1)}$, with one of the outcomes being realized depending on whether i undergoes treatment. The statistician has access the IID observations $\{(X_i, W_i, Y_i) : 1 \leq i \leq n\}$, where X_i is a vector of observed covariates for individual i, $W_i \in \{0, 1\}$ is (an encoding of) her treatment status, and $Y_i = Y_i^{(W_i)}$ is her realized outcome. One quantity of interest is the treatment effect at x

$$\tau(x) := \mathbf{E}(Y^{(1)} - Y^{(0)} \mid X_i = x).$$

Since only one of $Y_i^{(0)}$ and $Y_i^{(1)}$ is observed, consistent estimation of $\tau(x)$ requires further distributional assumptions. A common assumption is unconfoundedness, i.e., that treatment status W_i is independent of $Y_i^{(1)}$ and $Y_i^{(0)}$ conditional on X_i. Under this assumption,

$$\tau(x) = \mathbf{E}\left[Y_i \left(\frac{W_i}{e(x)} - \frac{1 - W_i}{1 - e(x)} \right) \mid X_i = x \right], \quad \text{where } e(x) = \mathbf{P}(W_i = 1 \mid X_i = x).$$

Here, the key function is $e(x)$, known as the propensity score: it is the probability of treatment for the subpopulation with covariates x (see Hirano, Imbens and Ridder (2003) for the derivation of the preceding formula and its implications). Machine learning methods—including random forests—may be brought to bear on the problem by estimating $e(x)$.

[2] Splits on discrete features partitions that variable into two arbitrary non-empty sets; no changes are needed for ordered categorical features.

Alternatively, unconfoundedness also implies

$$\tau(x) = \mathbf{E}(Y \mid W = 1, X = x) - \mathbf{E}(Y \mid W = 0, X = x),$$

so that $\tau(x)$ may be estimated by fitting two models, one on the subset of the sample in which $W = 1$, and the other on $W = 0$.

In econometric applications, conducting pointwise inference on the target function $f : \mathcal{X} \to \mathbf{R}$ (e.g., to test the null hypothesis $H_0 : f(x) = 0$) requires knowledge about the rate of convergence or asymptotic distribution of the underlying estimator $\hat{f}(x)$, where x specifies a subpopulation. However, *functionals* of the target function are often also of interest: for example, the difference in treatments effects (i.e., $f = \tau$) for two different subpopulations is captured by the quantity

$$f(x) - f(\bar{x}),$$

where x and $\bar{x}$ are covariates describing the two subpopulations. More generally, we might also be interested in a weighted treatment effect, where a subpopulation x is given an importance weight modeled by a density $\mu(x)$. In this case, the corresponding functional of f is

$$\int_{x \in \mathcal{X}} f(x) d\mu, \quad \text{where } \mu \text{ is not necessarily the density of } x,$$

and the integral is taken over the domain $\mathcal{X}$.

Inference on functionals of f requires not only the asymptotic distribution of the point estimate $f(x)$, but also the correlation between estimates at different points $f(x)$ and $f(\bar{x})$. As a concrete example, consider the function $\tau(x)$ and the simple difference $\tau(x) - \tau(\bar{x})$. We have

$$\begin{aligned}\tau(x) - \tau(\bar{x}) &= [\mathbf{E}(Y \mid W = 1, X = x) - \mathbf{E}(Y \mid W = 1, X = \bar{x})] \\ &\qquad - [\mathbf{E}(Y \mid W = 0, X = x) - \mathbf{E}(Y \mid W = 0, X = \bar{x})] \\ &=: A - B.\end{aligned}$$

We may estimate the difference by estimating A and B separately, fitting a random forest model to the two "halves" of the dataset where $W_i = 1$ and $W_i = 0$, as discussed above. The estimators $\hat{A}$ and $\hat{B}$ obtained are thus independent, so that $\mathbf{Var}(\hat{A} - \hat{B}) = \mathbf{Var}\,\hat{A} + \mathbf{Var}\,\hat{B}$. The variances of $\hat{A}$ and $\hat{B}$ then depend on the covariance of their respective random forest estimates at x and $\bar{x}$.

This paper studies the correlation structure of a class of random forest models whose asymptotic distributions were first worked out in Wager and Athey (2018). We find sufficient conditions under which asymptotic covariances of random forest estimates at different points vanish relative to their respective variances; moreover, we provide finite sample heuristics based on our calculations. To the best of our knowledge, this is the first set of results on the correlation structure of random forest estimators.

The present paper builds on and extends the results in Wager and Athey (2018), which in turn builds on related work in Wager and Walther (2015) on general concentration properties of trees and random forest estimators. Another related paper is Athey, Tibshirani and Wager (2019), which extends the random forest model considered here to a broader class of target functions by incorporating knowledge of moment conditions. Similar stability results established in this paper have appeared in Arsov, Pavlovski and Kocarev (2019), which studies notions of algorithmic stability for random forests and logistic regression and derive generalization error guarantees. Also closely related to our paper are Chernozhukov, Chetverikov and Kato (2017) and Chen (2018), concerning finite sample Gaussian approximations of sums and U-statistics in high dimensions, respectively. In this context, our paper provides a stepping stone towards applying the theory of finite sample U-statistics to random forests, where bounds on covariance matrices play a central role.

The paper is structured as follows. In Section 2, we introduce the random forest model and state the assumptions required for our results; Section 3 contains our main theoretical contributions; Section 4 builds on Sections 3 and discusses heuristics useful in finite sample settings; Section 5 concludes. All proofs are found in the appendix.

1.2 Model Setup and Assumptions

1.2.1 Overview of Tree Estimators

The goal of this paper is to study asymptotic Gaussian approximations of random forest estimators. Throughout, we assume that a random sample $\{Z_i = (X_i, Y_i) : 1 \leq i \leq n\} \subseteq \mathcal{X} \times \mathbf{R}$ is given, where each X_i is a vector of *features* or *covariates* belonging to a subset $\mathcal{X} \subseteq \mathbf{R}^p$ of p-dimensional Euclidean space, and $Y_i \in \mathbf{R}$ is the *response* or *target* corresponding to X_i. We will refer to $\mathcal{X}$ as the feature space or the feature domain.

Given the data set $\{Z_i\}_{i=1}^n$, a tree estimator recursively partitions the feature space by making axis aligned splits. Specifically, an axis-aligned split is a pair (j, t) where $j \in \{1, \ldots, p\}$ is the *splitting coordinate* and $t \in \mathbf{R}$ is the *splitting index*; given a subset $\mathcal{R} \subseteq \mathcal{X}$, a split (j, t) divides $\mathcal{R}$ into left and right halves

$$\{x \in \mathcal{R} : x_j < t\} \quad \text{and} \quad \{x \in \mathcal{R} : x_j > t\}, \tag{1.1}$$

where x_j denotes the j-th coordinate of the vector x. Starting with the entire feature space $\mathcal{X}$, the recursive splitting algorithm computes a (axis-aligned) split based on the data $\{Z_i : 1 \leq i \leq n\}$; for example, when the target Y_i is continuous, a popular choice of determining optimal splits is the following rule

$$(j, t) = \underset{\tilde{j}, \tilde{t}}{\arg\min} \sum_{i : X_i \in L} (Y_i - \mu_L)^2 + \sum_{i : X_i \in R} (Y_i - \mu_R)^2, \tag{1.2}$$

where $L = L(\tilde{j}, \tilde{t})$ and $R = R(\tilde{j}, \tilde{t})$ are the two halves of $\mathcal{X}$ obtained by the split $(\tilde{j}, \tilde{t})$, with μ_L and μ_R being the averages of targets Y_i whose corresponding feature X_i land in L and R, respectively.

After the first split, $\mathcal{X}$ is split into two halves L and R. The process is then repeated for L and R separately, in that a split for L is computed by using the subset of the data whose features X_i belong in L, and likewise for R. Each of the halves is then split again, and so on, until a stopping criterion is met. The process completes when the stopping criterion is satisfied for each subset; at this point, the collection of halves forms a partition[3] of $\mathcal{X}$, with each partition—and all the halves that came before it—being the intersection of a hyperrectangle with $\mathcal{X}$. The sequence of splits corresponds to a tree in a natural way; we will call the halfspaces that arise during the splitting process as *nodes*, and elements of the final partition *terminal nodes*.

Given the collection $N_1, \ldots, N_q$ of terminal nodes (which form a partition of $\mathcal{X}$), the prediction of the tree at a generic test point $x \in \mathcal{X}$ is the average of the responses that belong in the same terminal node as x

$$T(x; \xi, Z_1, \ldots, Z_n) = \sum_{j=1}^{q} \mathbf{1}(x \in N_j) \frac{1}{|N_j|} \sum_{i : X_i \in N_j} Y_j, \tag{1.3}$$

where the outer sum runs over observations i for which X_i belongs to the partition N_j, and $|N_j|$ is the number of such observations. The input ξ is an external source of randomization to allow for randomized split selection procedures. Thus, $T(x; \xi, Z_1, \ldots, Z_n)$ refers to the prediction at x for a tree grown using data $\{Z_1, \ldots, Z_n\}$ with randomization parameter ξ. As a function of x, keeping $Z_1, \ldots, Z_n$ and ξ fixed, $T : \mathcal{X} \to \mathbf{R}$ is then a step function, i.e., a linear combination of indicator functions of rectangular sets.

We note here that equations (1.2) and (1.3) are not the only possible choices; in particular, the rule used to choose the optimal split may be path dependent (i.e., dependent on previous splits) as in popular implementations (see Chen and Guestrin (2016), which allows for random forests but uses gradient boosting after the initial split), and the final prediction rule (1.3) may instead do a final linear fit or use a weighted average (c.f., Ke et al. (2017); Chen and Guestrin (2016)). Analysis of tree (and random forest) models may be considerably more complicated in these cases, so the present paper stipulates that the algorithm uses a splitting rule that is similar to (1.2) (c.f. Proposition 1.10) and uses (1.3) as the final prediction. In particular, this implies that the target function estimated by the tree estimator is the regression function $x \mapsto \mathbf{E}(Y \mid X = x)$.

[3] According to (1.1), we exclude edge cases when X_i lead on an "edge" of a hyperrectangle. This is not an issue for continuous variables, while for categorical variables, the definition should be slightly changed so that one of the halves contain $x_j = t$. In this paper, we deal only with continuous features.

1.2.2 From Trees to Random Forests

Given a specific tree estimator T, i.e., given the splitting rule used to build a tree, we define the random forest estimator to be the average of tree estimators across all $\binom{n}{s}$ subsamples, marginalizing over the randomization device ξ. Specifically, the random forest estimate $\mathrm{RF}(x)$ at $x \in \mathcal{X}$, given data $\{Z_1, \ldots, Z_n\}$, is defined to be

$$\mathrm{RF}(x; Z_1, \ldots, Z_n) = \frac{1}{\binom{n}{s}} \sum_{i_1, \ldots, i_s} \mathbf{E}_\xi \, T(x; \xi, Z_{i_1}, \ldots, Z_{i_s}), \tag{1.4}$$

where the summation runs over size-s subsets of $\{1, \ldots, n\}$, and the inner expectation is taken with respect to ξ. Importantly, each tree is grown on a subsample of size $s < n$. We follow Wager and Athey (2018) in assuming that $s \sim n^\beta$ for some β sufficiently close to one; specifically, we assume throughout that the subsample size is chosen as to satisfy the assumptions of Theorem 3 of Wager and Athey (2018), so that—along with other assumptions to be introduced presently—the random forest estimator RF is a consistent estimator of the target function $x \mapsto \mathbf{E}(Y \mid X = x)$.

In keeping with the notation of Wager and Athey (2018), we will write $T(x; Z_1, \ldots, Z_s)$ to mean the expectation of $T(x; \xi, Z_1, \ldots, Z_s)$ over ξ. With this notation, the random forest estimator (at a given point x) defined by (1.4) is a U-statistic with the size s kernel $(Z_1, \ldots, Z_s) \mapsto T(x, Z_1, \ldots, Z_s)$. We will discuss the U-statistic representation of RF in greater detail in Section 3.

1.2.3 Discussion of Model Assumptions

As our results will be an extension of the results in Wager and Athey (2018), we will study the same model of random forests and adopt a similar set of assumptions. The assumptions regarding tree estimators have appeared before in Wager and Walther (2015), while the distributional assumptions on the conditional moments of Y are standard (see e.g., Chapters 7 and 9 in Hastie, Tibshirani and Friedman (2009)).

The first—and most bespoke—assumption is that the tree algorithm is honest. Intuitively, honesty stipulates that knowledge of the tree structure does not affect the conditional distribution of tree estimates when the features are fixed.

Assumption 1.1 (Honesty). *The target Y_i and the tree structure (i.e., the splitting coordinates and splitting indices) are independent conditional on X_i. Specifically, we require*

$$\mathrm{dist}(Y_i \mid X_i, S) = \mathrm{dist}(Y_i \mid X_i), \tag{1.5}$$

for all observations i where Y_i participates in the final prediction, where S is set of splits chosen by the tree algorithm.

There are several ways to satisfy this assumption. The first is to calculate splits based only on the features X_i. This rules out the example splitting rule given in (1.2), so we could instead use its analog in $\mathcal{X}$,

$$(j,t) = \underset{\tilde{j},\tilde{t}}{\arg\min} \sum_{i:X_i \in L} \|X_i - \mu_L\|_2^2 + \sum_{i:X_i \in R} \|X_i - \mu_R\|^2, \tag{1.6}$$

where here μ_L and μ_R denote the average (i.e., center of mass) of the X_i in each halfspace. In this instance, the choice of splits is essentially a clustering algorithm that finds the best division of the sample points into two parts. Another way to satisfy the honesty assumption while still computing splits based on the targets is to use sample splitting. The dataset is partitioned into two parts $\mathcal{I}_1$ and $\mathcal{I}_2$; observations in $\mathcal{I}_1$ and $X_i \in \mathcal{I}_2$ may be freely used during the splitting process, while $Y_i \in \mathcal{I}_2$ are used for the final predictions. In this case, equality in (1.5) is required to hold for $i \in \mathcal{I}_2$. Finally, a third method to satisfy honesty requires the existence of auxiliary data $\{W_i\}$. During the splitting stage ("model fitting"), splits are computed *as if* the response variable is W_i; for example, (1.2) is used with μ_L and μ_R being the averages of $\{W_i\}$. Once the tree is fully grown, predictions ("model inference[4]") are made using Y_i's as usual. The practice of using such *surrogate targets* is especially popular in time-series prediction, where different horizons are used in fitting and inference steps (c.f., Quaedvlieg (2019)).

In the present paper, for simplicity of notation, we shall assume that the first scheme is used to satisfy honesty—namely, splitting decisions are based on the feature vectors X_i only. Our results extend to all three schemes.[5]

Our next assumption will ensure that each one of the p axes is chosen as the splitting coordinate with a probability bounded away from zero.

Assumption 1.2 (Randomized Cyclic Splits). *When computing the optimal split, the algorithm flips a probability δ coin that is entirely independent of everything else. The first time the coin lands heads, the first coordinate is chosen as the splitting coordinate; when the coin lands heads again, the second coordinate is chosen, and so on, such that on the J-th time the coin lands heads, the $(J \text{ mod } p) + 1$-th coordinate is chosen[6]. After the random splitting coordinate is chosen, the splitting index may still be chosen based on the observations.*

This assumption is a modification of the random splitting assumption in Wager and Athey (2018), in which each of the p axes has a probability δ of being chosen at each split. The latter assumption could be directly implemented by flipping a $p\delta$ coin and selecting one of the p coordinates uniformly at random as the splitting coordinate when the coin lands heads. Another method to satisfy the corresponding assumption in Wager and Athey

[4] The terminology 'inference' is used to mean computing the predictions of an existing model, which is unrelated to the typical usage of 'inference' in econometrics. The former terminology is standard in applied settings, used when describing a data pipeline: see the documentation of Google and other contributors (2015); Pedregosa et al. (2011).

[5] As in Wager and Athey (2018), constants appearing in our bounds may change in scheme two.

[6] We adopt the convention that mp mod $p = 0$, hence notation for adding 1 to (J mod p).

(2018), studied in Athey, Tibshirani and Wager (2019) and implemented by popular libraries such as Ke et al. (2017), is randomizing the *number* of available splitting axes in each round. Specifically, a Poisson random variable Q with intensity proportional to $\sqrt{p}$ is first realized (Q is realized independently from round to round). Afterwards, $\min(Q, p)$ many features are uniformly selected as potential candidates[7] for splitting in that round. Clearly, this also yields a positive probability $\delta > 0$ that each coordinate $1 \leq j \leq p$ is chosen, independently of everything else.

In contrast to our assumption, both methods above involve two separate rounds of randomization: first, a random variable encoding the decision to split randomly is made (i.e., the coin or the random variable Q), and second, the splitting axis is then determined. Intuitively, our cyclic splitting assumption above forgoes the second randomization step: in doing so, the variance of the number of times that any coordinate is chosen is reduced. Importantly, the variance will depend only on δ and not on p, which we will exploit in our proofs.

Assumption 1.3 (The Splitting Algorithm is (α, k)-Regular)**.** *There exists some $\alpha \in (0, 1/2)$ such that whenever a split occurs in a node with m sample points, the two resulting hyperrectangles contain at least αm many points each. Moreover, splitting ceases at a node only when it contains fewer than $2k - 1$ points.*

This key assumption is carried over from Wager and Athey (2018) and contains two requirements. The first requires that no split may produce a halfspace containing too few observations, i.e., that both halfspaces are large when measured by the count of observations. As shown in Wager and Walther (2015), this implies that with exponentially small complementary probability, the splitting axis shrinks by a factor between α and $1 - \alpha$, so that both halfspaces are also large in Euclidean volume (with high probability).

The second half of the assumption places an upper bound on the number of observations in terminal nodes. Trees grown under this assumption will necessarily be deeper as the sample size n (and thus, the subsample size $s = n^\beta$) increases. In particular, the predictions at leaf nodes—averages of observations Y_i— will be averages of a bounded number of terms. An important consequence is that the variance of the tree estimator (at any test point x) is bounded below due to the distributional assumption on $\mathbf{Var}(Y \mid X = x)$ (see Assumption 5).

Assumption 1.4 (Predetermined Splits)**.** *The candidate splits considered at each node do not depend on the data $\{Z_i\}$, so that they are fixed ahead of time. Furthermore, the number of candidate splits at each node is finite, and every candidate split shrinks the the length of its splitting axis by at most a factor α.*

The predetermined splitting assumption is specific to our paper. That candidate splits considered at each node are data-independent is "almost" without loss of generality since

[7] Splitting for that node ceases if $Q = 0$.

the feature space $\mathcal{X}$ is fixed. For example, implementations of random forests typically use 32-bit floating point numbers as their splitting index, so that the assumption is automatically satisfied. Furthermore, the assumption allows candidate splits to depend on the location of the hyperrectangle, so that the splitting process is still data-driven insofar as the sequence of splits leading up to a node depends on the observations.

One interpretation of this assumption is that it aids in data compression. For example, suppose that all features are continuous and $\mathcal{X} = [0, 1]^p$, and that all candidate splits have the form $(j, k/2^m)$ for some integers $1 \leq j \leq p$ and $0 \leq k < 2^m$, where $m \geq 1$ is a fixed integer. If a splitting rule such as (1.6) is used, then the optimal split depends only on the values $\{\lfloor 2^m X_{ij} \rfloor : 1 \leq i \leq n, 1 \leq j \leq p\}$ instead of $\{X_{ij}\}$. Each member of the former set is an integer in $\{0, \dots, 2^m - 1\}$ and thus could be represented using m bits. In particular, for a grid resolution of $2^8 = 256$—a fine grid even in moderately large dimensions—each coordinate of the feature vectors X_i may be stored in a single byte. Since modern CPUs and graphics processors store floating point numbers in four or eight bytes, this is a substantial reduction, allowing computation power to scale to larger datasets. The process of encoding features in this way is known as *quantizing*, which is an option supported by popular software packages. In this way, though the predetermined split assumption may seem at first glance restrictive, it aligns our model more closely with practice.

Assumption 1.5 (Distributional Assumptions on the DGP of (X, Y))**.** *The features X_i are supported on the unit cube $\mathcal{X} = [0, 1]^p$ with a density that is bounded away from zero and infinity. Furthermore, the functions $x \mapsto \mathbf{E}(Y \mid X = x)$, $x \mapsto \mathbf{E}(Y^2 \mid X = x)$, and $x \mapsto \mathbf{E}(|Y|^3 \mid X = x)$ are uniformly Lipschitz continuous. Finally, the conditional variance $\mathbf{Var}(Y \mid X = x)$ is bounded away from zero, i.e., $\inf_{x \in \mathcal{X}} \mathbf{Var}(Y \mid X = x) > 0$.*

The continuity and variance bound assumptions are standard. Note that a consequence of continuity and compactness of the hypercube is that the conditional moments up to order three are bounded. Our results will not explicitly depend on knowledge of the density of X: however, the density will affect the implicit constants that we carry throughout our proofs (c.f., Lemma 3.2 and Theorem 3.3 in Wager and Athey (2018)).

1.3 Gaussianity of Multivariate U-Statistics

1.3.1 Test Points and Notational Conventions

We begin investigation of the random forest estimator in this section. As discussed in the model introduction, the random forest estimator $\mathrm{RF}(x)$ at a test point x is a U-statistic whose kernel is the tree estimator $T(x)$ marginalized over external randomizations. This paper studies the *multivariate* distribution of RF, specifically the correlation structure between $\mathrm{RF}(x)$ and $\mathrm{RF}(\bar{x})$ at distinct points x and $\bar{x} \in \mathcal{X}$. Towards that end, we shall fix a collection of q test points $x_1, \dots, x_q \in \mathcal{X}$ throughout the remainder of the paper. As these points will remain fixed, for notational brevity *we will omit their explicit dependence* when writing estimators. Therefore, $\mathrm{RF}(Z_1, \dots, Z_n)$ stands for the q-dimensional estimator that is the random forest evaluated at $x_1, \dots, x_q$, given observations $\{Z_i : 1 \le i \le n\}$. As a consequence of this notation, most of our equations are to be understood in $\mathbf{R}^q$, with equality and arithmetic operations acting coordinate-wise. Finally, when there is no confusion, subscripts (typically k or ℓ) denote a specific coordinate, i.e., the estimate at the k-th or ℓ-th test point; a notable exception is T_1, which refers to a Hajek projection that we now describe.

1.3.2 Hajek Projections

We start by reviewing properties of the Höeffding Decomposition of U-statistics, also known as Hajek projections; see Vaart (1998) for a textbook treatment of the univariate case. Let $f(Z_1, \dots, Z_m) \in \mathbf{R}^q$ be a generic q-dimensional statistic based on m observations. The *Hajek projection* of f is defined to be

$$\overset{\circ}{f}(Z_1, \dots, Z_m) = \sum_{i=1}^{m} \mathbf{E}[f(Z_1, \dots, Z_m) \mid Z_i] - (m-1)\, \mathbf{E}\, f(Z_1, \dots, Z_m).$$

That is, it is the coordinate-wise projection of f to the linear space spanned by functions of the form $\{g(Z_i) : 1 \le i \le m\}$. In particular, when f is symmetric in its arguments and $Z_1, \dots, Z_m$ is an IID sequence, we have

$$\overset{\circ}{f}(Z_1, \dots, Z_m) = \sum_{i=1}^{m} f_1(Z_i) - (m-1)\, \mathbf{E}\, f, \tag{1.7}$$

where $f_1(z)$ is the function such that $f_1(Z_1) = \mathbf{E}(f \mid Z_1)$, i.e., $f_1(z) = \mathbf{E}(f \mid Z_1 = z)$.

In our setting, applying the Hajek projection to the *centered* statistic $\mathrm{RF} - \mu$, where μ is the expectation of RF, yields

$$\overset{\circ}{\mathrm{RF}}(Z_1, \dots, Z_n) - \mu = \sum_{i=1}^{n} \mathbf{E}(\mathrm{RF} - \mu \mid Z_i) = \frac{1}{\binom{n}{s}} \sum_{i=1}^{n} \mathbf{E}\Bigg[\sum_{i_1, \dots, i_s} \mathbf{E}_\xi\, T(\xi, Z_{i_1}, \dots, Z_{i_s}) - \mu \mid Z_i\Bigg],$$

where $i_1, \dots, i_s$ run through the $\binom{n}{s}$ size-s subsets of $\{1, \dots, n\}$. (Recall that RF, μ, and T are all vectors in $\mathbf{R}^q$, with $T(\xi, Z_1, \dots, Z_s)$ denoting the vector of tree estimates produced using data $\{Z_i\}$ with randomization parameter ξ.) Since the samples $Z_1, \dots, Z_n$ are independent, $\mathbf{E}(\mathbf{E}_\xi\, T(\xi, Z_{i_1}, \dots, Z_{i_s}) \mid Z_i) = \mu$ whenever $i \notin \{i_1, \dots, i_s\}$. As $\{i_1, \dots, i_s\}$ runs over the size-s subsets of $\{1, \dots, n\}$, there are exactly $\binom{n-1}{s-1}$ many which contain i. For each of of these subsets,

$$\mathbf{E}(\mathbf{E}_\xi\, T(\xi, Z_{i_1}, \dots, Z_{i_s}) - \mu \mid Z_i) =: T_1(Z_i) - \mu,$$

where $T_1(z) := \mathbf{E}_{\xi, Z_2, \dots, Z_s} T(\xi, z, Z_2, \dots, Z_s)$. Therefore,

$$\mathring{\mathrm{RF}} - \mu = \frac{1}{\binom{n}{s}} \sum_{i=1}^{n} \binom{n-1}{s-1} (T_1(Z_i) - \mu) = \frac{s}{n} \sum_{i=1}^{n} (T_1(Z_i) - \mu). \tag{1.8}$$

The sequence of observations $Z_1, \dots, Z_n$ is assumed to IID, and this property is preserved for the sequence $\{T_1(Z_i) : 1 \leq i \leq n\}$ of projections. It is easily verified that $\mathbf{E}(\mathring{\mathrm{RF}}) = \mu$, and the point of the previous equation is that it expresses the centered statistic $(\mathring{\mathrm{RF}} - \mu)$ as an average of centered IID terms, scaled by s. This will be our main entry point in establishing asymptotic joint normality.

1.3.3 Asymptotic Gaussianity via Hajek Projections

The standard technique in deriving the asymptotic distribution of a U-statistic is to establish a lower bound on the variance of its Hajek projection; this is the approach taken by Wager and Athey (2018) and we follow the approach here. Let V be the variance of $\mathring{\mathrm{RF}}$; using (1.8), we have

$$V = \mathbf{Var}\left[\frac{s}{n} \sum_{i=1}^{n} (T_1(Z_i) - \mu)\right] = \frac{s^2}{n} \mathbf{Var}(T_1(Z_1)) = \frac{s}{n} \mathbf{Var}\left[\sum_{i=1}^{s} T_1(Z_i)\right] = \frac{s}{n} \mathbf{Var}\, \mathring{T} \in \mathbf{R}^{q \times q}, \tag{1.9}$$

where $\mathring{T}$ is the Hajek projection of the statistic T as in (1.7), where $T = \mathbf{E}_\xi T(\xi, Z_1, \dots, Z_s) \in \mathbf{R}^q$.

The assumptions that $\mathbf{Var}(Y \mid X = x)$ is bounded below and that $\mathbf{E}(|Y|^3 \mid X = x)$ is bounded above allow for an easy verification of a multivariate version of the Lyapunov Central Limit Theorem for triangular arrays (see Appendix for details); the proof will demonstrate that under these two assumptions, the multivariate analog of the Lyapunov condition is implied by its univariate version, which was established in Theorem 8 of Wager and Athey (2018). This is the basis of the following lemma.

Lemma 1.6. *Let I be the $q \times q$ identity matrix and let 0 denote the zero vector in $\mathbf{R}^q$. If the assumptions outlined in Section 1.2.3 are all satisfied, then*

$$V^{-1/2}(\mathring{\mathrm{RF}} - \mu) \overset{\text{dist}}{\Longrightarrow} N(0, I).$$

Proof. (All proofs may be found in the Appendix.) □

Remark. We used the boundedness of the third moment to verify that the Lyapunov condition holds with exponent equal to one; in general, this assumption is not necessary, but the verification of the Lyapunov condition (for a smaller exponent) will be considerably more complicated. More recently, triangular array CLTs specific to U-statistics were developed in DiCiccio and Romano (2020), and their conditions are satisfied in our case as well.

The asymptotic normality of the random forest estimator RF can be related to the asymptotic normality of $\overset{\circ}{\mathrm{RF}}$ via

$$V^{-1/2}(\mathrm{RF}-\mu) = V^{-1/2}(\mathrm{RF}-\overset{\circ}{\mathrm{RF}}) + V^{-1/2}(\overset{\circ}{\mathrm{RF}} - \mu).$$

The second summand on the RHS is asymptotically normal by Lemma 1.6; by Slutsky's Theorem, $V^{-1/2}(\mathrm{RF}-\mu)$ is asymptotically normal once we establish the convergence

$$V^{-1/2}(\mathrm{RF}-\overset{\circ}{\mathrm{RF}}) \xrightarrow{\mathbf{P}} 0.$$

The strategy is to show that $e = V^{-1/2}(\mathrm{RF}-\overset{\circ}{\mathrm{RF}})$ converges in squared mean. We may develop its squared norm via

$$\begin{aligned}\mathbf{E}(e^\intercal e) &= \mathbf{E}(\mathrm{RF}-\overset{\circ}{\mathrm{RF}})^\intercal V^{-1}(\mathrm{RF}-\overset{\circ}{\mathrm{RF}}) = \mathbf{E}\,\mathrm{tr}\, V^{-1}(\mathrm{RF}-\overset{\circ}{\mathrm{RF}})(\mathrm{RF}-\overset{\circ}{\mathrm{RF}})^\intercal \\ &= \mathrm{tr}\, V^{-1}\,\mathbf{E}(\mathrm{RF}-\overset{\circ}{\mathrm{RF}})(\mathrm{RF}-\overset{\circ}{\mathrm{RF}})^\intercal = \mathrm{tr}\, V^{-1/2}\,\mathbf{Var}(\mathrm{RF}-\overset{\circ}{\mathrm{RF}})V^{-1/2},\end{aligned} \tag{1.10}$$

where we used the identity $\mathrm{tr}(ABC) = \mathrm{tr}(BCA)$ for conforming matrices A, B, and C.

That the trace on the extreme RHS goes to zero is the natural multivariate generalization of the familiar condition

$$\frac{\mathbf{Var}(f - \mathring{f})}{\mathbf{Var}\,\mathring{f}} \to 0$$

for univariate U-statistics (see Vaart (1998)). In the univariate setting, this condition is checked by considering higher order decompositions of the statistic f; this approach is also valid in the multivariate setting, which we now show. The following proposition defines the proper generalization of higher order decompositions for multivariate statistics.

Proposition 1.7 (Höeffding Decomposition for Multivariate U-statistics)**.** *Fix a positive definite matrix M. Let $f(x_1,\ldots,x_n) \in \mathbf{R}^q$ be a vector-valued function that is symmetric in its arguments and let $X_1,\ldots,X_n$ be a random sample such that $f(X_1,\ldots,X_n)$ has finite variance. Then there exists functions $f_1, f_2, \ldots, f_n$ such that*

$$f(X_1,\ldots,X_n) = \mathbf{E}(f) + \sum_{i=1}^{n} f_1(X_1) + \sum_{i<j} f_2(X_i,X_j) + \cdots + f_n(X_1,\ldots,X_n)$$

where f_k is a function of k arguments, such that

$$\mathbf{E}\, f_k(X_1,\ldots,X_k) = 0 \quad \textit{and} \quad \mathbf{E}[f_k(X_1,\ldots,X_k)^\intercal M f_\ell(X_1,\ldots,X_l)] = 0.$$

By applying Proposition 1.7 to $\mathrm{RF}-\mu$, we may expand $\mathrm{RF}-\overset{\circ}{\mathrm{RF}}$ according to a Höeffding decomposition taken *respect to the matrix* V^{-1},

$$\mathrm{RF} - \mathring{\mathrm{RF}} = \frac{1}{\binom{n}{s}}\left[\sum_{i<j}\binom{n-2}{s-2}(T^{(2)}(Z_i, Z_j) - \mu) + \sum_{i<j<k}\binom{n-3}{s-3}(T^{(3)}(Z_i, Z_j, Z_k) - \mu) + \cdots\right], \tag{1.11}$$

where $T^{(2)}$, $T^{(3)}$, etc. are higher order projections of T (corresponding to f_2, f_3, etc. in the proposition) which satisfy the normal equations

$$\mathbf{E}[(T^{(k)} - \mu)^\intercal V^{-1}(T^{(k')} - \mu)] = 0, \qquad \text{for } k \neq k'.$$

Of course, the higher order terms $T^{(k)}$, being projections of T, also satisfy

$$\mathbf{E}[(T^{(k)} - \mu)^\intercal V^{-1}(T^{(k)} - \mu)] \le \mathbf{E}[(T - \mu)^\intercal V^{-1}(T - \mu)]. \tag{1.12}$$

Next, write

$$V^{-1/2}\,\mathbf{Var}(\mathrm{RF} - \mathring{\mathrm{RF}})V^{-1/2} = \mathbf{Var}(V^{-1/2}(\mathrm{RF} - \mathring{\mathrm{RF}})).$$

The orthogonality conditions in (1.11) show that $V^{-1/2}(\mathrm{RF} - \mathring{\mathrm{RF}})$ is a sum of $(s-1)$ uncorrelated terms, one for each of the higher order projections $T^{(2)}, \ldots, T^{(s)}$. Therefore,

$$\begin{aligned}\mathbf{Var}(V^{-1/2}(\mathrm{RF} - \mathring{\mathrm{RF}})) &= \sum_{i<j}\frac{\binom{n-2}{s-2}^2}{\binom{n}{s}^2}\mathbf{Var}(V^{-1/2}T^{(2)}(Z_i, Z_j)) \\ &\quad + \sum_{i<j<k}\frac{\binom{n-3}{s-3}^2}{\binom{n}{s}^2}\mathbf{Var}(V^{-1/2}T^{(3)}(Z_i, Z_j, Z_k)) \\ &\quad + \ldots\end{aligned}$$

Since the variables $Z_1, \ldots, Z_n$ are IID, the quantities $\mathbf{Var}(V^{-1/2}T^{(2)}(Z_i, Z_j))$ do not depend on i and j; each is equal to $\mathbf{Var}(V^{-1/2}T^{(2)})$; the same is true for other higher-order projections as well. Rewriting (1.12) as the inequality $\operatorname{tr}\mathbf{Var}\,V^{-1/2}T^{(k)} \le \operatorname{tr}\mathbf{Var}\,V^{-1/2}T$ then proves the following

$$\begin{aligned}\mathbf{E}(e^\intercal e) = \operatorname{tr}\mathbf{Var}(V^{-1/2}(\mathrm{RF} - \mathring{\mathrm{RF}})) &\le (\operatorname{tr}V^{-1/2}\,\mathbf{Var}(T)V^{-1/2}) \times \sum_{k=2}^{s}\frac{\binom{n-k}{s-k}^2}{\binom{n}{s}^2} \\ &\le \boxed{\frac{s}{n}\operatorname{tr}\mathbf{Var}\,\mathring{T}^{-1/2}\,\mathbf{Var}\,T},\end{aligned} \tag{1.13}$$

where the final step uses (1.9) to relate V and $\mathbf{Var}\,\mathring{T}$, as well as the fact $\sum_{k=2}^{s}\frac{\binom{n-k}{s-k}^2}{\binom{n}{s}^2} \le \frac{s^2}{n^2}$.

The remainder of this section centers around proving that the boxed quantity in (1.13) converges to zero. For comparison, a central result of Wager and Athey (2018) (using our notation) is a bound on the *diagonal elements* of $\mathbf{Var}\,\mathring{T}$ and $\mathbf{Var}\,T$. Specifically, the authors obtain

$$\frac{(\mathbf{Var}\, T)_{kk}}{(\mathbf{Var}\, \mathring{T})_{kk}} \leq c(\log s)^p, \qquad \text{for each } k = 1, \ldots, q, \tag{1.14}$$

for some constant c. As we will see in the next section, the required bound on the trace will follow from bounds on the *off-diagonal* elements of $\mathbf{Var}\, \mathring{T}$, i.e., bounds on the covariance between random forest estimates at different test points (see discussion following Proposition 1.8).

1.3.4 Covariance Bounds

The aim of this section is to establish asymptotic bounds on the off-diagonal elements of the covariance matrix $\mathbf{Var}\, \mathring{T}$. We shall show that when the tree estimator employs splitting algorithms satisfying suitable *stability conditions*, we have the asymptotic behavior

$$(\mathbf{Var}\, \mathring{T})_{k,l} = o(s^{-\epsilon}) \text{ for all } 1 \leq k \neq l \leq q \text{ and some } \epsilon > 0. \tag{1.15}$$

Before proceeding, we first show that this bound, coupled with control on the diagonal terms, suffices to establish the trace bound in (1.13).

Proposition 1.8. *Suppose the assumptions in Section 1.2.3 hold with $\delta > 1/2$, and further assume that the splitting algorithm is stable (see page 26 for a definition and Proposition 1.10 for a set of sufficient conditions). The entries of* $\mathbf{Var}\, T$ *are bounded and its diagonal entries are bounded away from zero. Furthermore, when* $\mathbf{Var}\, \mathring{T}$ *satisfies the condition in* (1.15),

$$\frac{s}{n} \operatorname{tr}(\mathbf{Var}\, \mathring{T}^{-1}\, \mathbf{Var}\, T) \to 0. \tag{1.16}$$

Remark. The first part of the Proposition, concerning the entries of $\mathbf{Var}\, T$, is a consequence of our (α, k)-regularity assumption and distributional assumptions on $\{Z_i\}$. As discussed in Section 1.2.3, since the number of observations in leaf nodes is bounded above, the (pointwise) variance of the tree estimator at x is bounded below by $\mathbf{Var}(Y \mid X = x)$ up to a constant, and we assumed that the latter function is bounded away from zero. That entries of $\mathbf{Var}\, T$ are bounded is a trivial consequence of the fact that the function $x \mapsto \mathbf{E}(Y^2 \mid X = x)$ is Lipschitz and thus bounded. The techniques we present to bound $\mathbf{Var}\, \mathring{T}$ could also be used to bound $\mathbf{Var}\, T$; it is in fact true that $(\mathbf{Var}\, T)_{k,l} \to 0$ for $k \neq l$, though we will not pursue this further in this paper.

Proposition 1.8 establishes $\mathring{T}$ as the central object of study. Recall that T is the tree estimator while $\mathring{T}$ is its Hajek projection; in other words, $\mathbf{Var}\, \mathring{T}$ is *not* the covariance matrix of tree estimates. However, our result will demonstrate the asymptotic normality of $V^{-1}(\mathrm{RF} - \mu)$, where V, the variance of the Hajek projection $\mathring{\mathrm{RF}}$, is given in terms of $\mathbf{Var}\, \mathring{T}$ (c.f., (1.9)). Therefore, (a rescaled version of) $\mathbf{Var}\, \mathring{T}$ is precisely the object needed to conduct inference on the random forest. In particular, combining (1.14) and (1.15) yields the fact that $\mathbf{Var}\,\mathring{T}$—and hence the asymptotic variance of RF—is diagonally dominant (i.e., tending to a diagonal matrix in the limit).

We may always relabel indices so that the tree is grown on the observations $Z_1, \ldots, Z_s$. To establish the bound (1.16), start with the definition

$$\mathring{T} - \mu = \sum_{i=1}^{s} \mathbf{E}(T \mid Z_i) \quad \text{so that} \quad \mathbf{Var}\,\mathring{T} = s\,\mathbf{Var}(\mathbf{E}(T \mid Z_1)) \text{ due to independence.} \tag{1.17}$$

To develop the term on the RHS, use the orthogonality condition for conditional expectation

$$\mathbf{Var}\,\mathbf{E}(T \mid Z_1) = \mathbf{Var}[\mathbf{E}(T \mid Z_1) - \mathbf{E}(T \mid X_1)] + \mathbf{Var}[\mathbf{E}(T \mid X_1)]. \tag{1.18}$$

Since the tree algorithm is honest, the difference $\mathbf{E}(T \mid Z_1) - \mathbf{E}(T \mid X_1)$ simplifies, so that for each $1 \le k \le q$,

$$\mathbf{E}(T_k \mid Z_1) - \mathbf{E}(T_k \mid X_1) = \mathbf{E}(I_k \mid X_1)(Y_1 - \mathbf{E}(Y_1 \mid I_k = 1, X_1)),$$

where T_k is the tree estimate at x_k, and I_k is the indicator for whether X_1 and x_k belong to the same terminal node. Therefore, the off-diagonal entry at (k, l) of $\mathbf{Var}[\mathbf{E}(T \mid Z_1) - \mathbf{E}(T \mid X_1)]$ is equal to

$$\mathbf{E}[\mathbf{E}(I_k \mid X_1)\,\mathbf{E}(I_l \mid X_1)(Y_1 - \mathbf{E}(Y_1 \mid X_1, I_k = 1)(Y_1 - \mathbf{E}(Y_1 \mid X_1, I_l = 1)]. \tag{1.19}$$

We may expand the terms in the above integrand as follows

$$\begin{aligned}
&\mathbf{E}(I_k \mid X_1)\,\mathbf{E}(I_l \mid X_1)(Y_1 - \mathbf{E}(Y_1 \mid X_1, I_k = 1)(Y_1 - \mathbf{E}(Y_1 \mid X_1, I_l = 1) \\
&\quad = \mathbf{E}(I_k \mid X_1)\,\mathbf{E}(I_l \mid X_1) Y_1^2 - \mathbf{E}(I_k \mid X_1)\,\mathbf{E}(I_l \mid X_1) Y_1\,\mathbf{E}(Y_1 \mid X_1, I_l = 1) + \ldots \\
&\quad = \sum_{t=1}^{4} \mathbf{E}(I_k \mid X_1)\,\mathbf{E}(I_l \mid X_1) \cdot p_t(Y_1, \mathbf{E}(Y_1 \mid X_1, I_k = 1), \mathbf{E}(Y_1 \mid X_1, I_l = 1)),
\end{aligned}$$

for some multinomials $p_1, \ldots, p_4$, each with degree at most two. Since we have assumed that $\mathbf{E}(Y \mid X = x)$ and $\mathbf{E}(Y^2 \mid X = x)$ are continuous and hence bounded, $\mathbf{E}(p_t(\ldots) \mid X_1 = x)$ is also bounded. Therefore, using the Law of Iterated Expectations to evaluate (1.19) shows that it is bounded by a constant times

$$\mathbf{E}[\mathbf{E}(I_k \mid X_1)\,\mathbf{E}(I_j \mid X_1)].$$

Remark. A direct application of the Cauchy-Schwarz inequality, using only that $\mathbf{E}(Y \mid X = x)$ is bounded (i.e., without assuming $\mathbf{E}(Y^2 \mid X = x)$ is bounded), would yield the weaker bound

$$\sqrt{\mathbf{E}[\mathbf{E}(I_k \mid X_1)^2\,\mathbf{E}(I_j \mid X_1)^2]} \le \sqrt{\mathbf{E}[\mathbf{E}(I_k \mid X_1)\,\mathbf{E}(I_j \mid X_1)]},$$

up to a multiplicative constant.

Recall that I_k and I_l are indicator variables for whether X_1 belongs to the same hypercube as x_k and x_l, respectively. Therefore, $\mathbf{E}(I_k \mid X_1)$ is the probability that the first observation is used for the prediction at x_k, and likewise for $\mathbf{E}(I_l \mid X_1)$. Intuitively, this only happens when X_1 is near x_k (respectively, near x_l): since $x_k \neq x_l$, X_1 cannot be near to both, meaning that the product $\mathbf{E}(I_k \mid X_1)\,\mathbf{E}(I_l \mid X_1)$ is small.

Proposition 1.9. *For two points x and $\bar{x} \in \mathcal{X} = [0,1]^p$, define*

$$M(x,\bar{x}) = \mathbf{E}[\mathbf{E}(I \mid X_1)\,\mathbf{E}(\bar{I} \mid X_1)], \tag{1.20}$$

where I and $\bar{I}$ are indicators for X_1 belonging to the same terminal node as x and $\bar{x}$, respectively. If $\delta > 1/2$ and $x \neq \bar{x}$,

$$M(x,\bar{x}) = o(s^{-(1+\epsilon)}) \quad \textit{for some } \epsilon > 0. \tag{1.21}$$

Remark. It is instructive to consider the bound in the preceding display versus $M(x,x)$. It is clear from the definition that $M(x,x) \geq M(x,\bar{x})$ for all $\bar{x}$. In addition, $M(x,x) = \mathbf{E}(\mathbf{E}(I \mid X_1)^2) \leq \mathbf{E}(\mathbf{E}(I \mid X_1)) = \mathbf{E}(I)$. By symmetry, $\mathbf{E}\,I = 1/s$ (up to constant), as the terminal node at x has a bounded number of observations. Therefore, all that the Proposition ensures is that when $x \neq \bar{x}$, the quantity $M(x,\bar{x})$ is smaller than the "trivial" bound $1/s$.

This proposition shows that the contribution of $\mathbf{Var}[\mathbf{E}(T \mid Z_1) - \mathbf{E}(T \mid X_1)]$ to the cross covariances of $\mathbf{Var}\,\mathbf{E}(T \mid Z_1)$ is small, in particular smaller than the required bound $(\log^p s \cdot s)^{-1}$. The requirement that $\delta > 1/2$, while needed for the proof to go through, is almost certainly not needed in practice. The reason is that our proof uses $\delta > 1/2$ to derive a *uniform* bound on the quantity

$$\mathbf{E}(I_k \mid X_1)\,\mathbf{E}(I_l \mid X_1),$$

while the proposition only demands a bound on its expectation. Indeed, in the extreme case $x = 0$ and $\bar{x} = (1,\ldots,1)^\mathsf{T}$, it is easy to see that the expectation meets the required bound even when $\delta \leq 1/2$.

Furthermore, our proof is agnostic to the exact splitting rule used by the base tree learner and uses only "random splits" (c.f., Assumption 2) to derive the required bounds. With a specific splitting rule (e.g., (1.2)) and a specific data distribution, the expectation $M(x,\bar{x})$ will be smaller than that predicted by (1.21). In light of this, an alternative to our cyclic splitting assumption is to assume the *high level* condition that the splitting algorithm and data generating process confer the bound

$$M(x,\bar{x}) = o\left(\frac{1}{\log^p s \cdot s}\right).$$

Bounding $\mathbf{Var}\,\mathbf{E}(T \mid X_1)$

We turn next to bound the off-diagonal terms in $\mathbf{Var}\,\mathbf{E}(T \mid X_1)$. As in the statement of Proposition 1.9, it will be convenient to slightly change notations. We fix $1 \leq k \neq l \leq q$, and use the notation $x \mapsto x_k$, $\bar{x} \mapsto x_l$, and with x_1 denoting the value of X_1 (i.e., x_1 is no longer a test point, that role being played by x and $\bar{x}$). The goal of this section is to establish the bound

$$(\mathbf{Var}\,\mathbf{E}(T \mid X_1))_{kl} = \mathbf{E}[(\mathbf{E}(T_k \mid X_1 = x_1) - \mu_k)(\mathbf{E}(T_l \mid X_1 = x_1) - \mu_l)] = o\left(\frac{\log^p s}{s}\right),$$

where T_k and T_l are the tree estimates at x and $\bar{x}$, and μ_k and μ_l are their unconditional expectations.

The quantity $\mathbf{E}(T_k \mid X_1 = x_1) - \mu_k = \mathbf{E}(T_k \mid X_1 = x_1) - \mathbf{E}(T_k)$ measures the amount of "information" that the location of a single observation X_1 carries for the output of the tree at x. Intuitively, when $X_1 = x_1$ is near x, the effect of X_1 on the leaf node containing x is more pronounced and we expect $\mathbf{E}(T_k \mid X_1 = x_1) - \mu \approx \mathbf{E}(Y \mid X_1 = x_1) - \mu$. Conversely, when $X_1 = x_1$ is far from x, then X_1's effect in determining the location of the leaf node containing x diminishes, and $\mathbf{E}(T \mid X_1 = x_1) - \mu \approx \mathbf{E}(T) - \mu = 0$.

The key in making the above intuition precise is to keep track of when $X_1 = x_1$ leaves the intermediate partition containing x in the splitting process; here, "intermediate partitions" are those nodes created during the splitting process that are not necessarily terminal. Towards this end, fix x and let Π denote the terminal node containing x; Π is a hyperrectangle contained in $\mathcal{X}$ created by axis aligned splits. By Assumption 4, the set of potential splits does not depend on the sample (in particular, it does not depend on X_1). Moreover, splitting ceases after no more than s splits, regardless of the subsample $X_1, \ldots, X_s$, as each split reduces the number of observations in its two child nodes by at least one. Therefore, Π takes on only finitely many possible values, and we may write

$$\mathbf{E}(T) = \sum_{\pi} \mathbf{P}(\Pi = \pi)\mu_\pi \quad \text{and} \quad \mathbf{E}(T \mid X_1 = x_1) = \sum_{\pi} \mathbf{P}(\Pi = \pi \mid X_1 = x)\mu'_\pi \tag{1.22}$$

where $\mu_\pi = \mathbf{E}(T \mid \Pi = \pi)$ and $\mu'_\pi = \mathbf{E}(T \mid \Pi = \pi, X_1 = x)$.

The hyperrectangle Π is determined by the recursive splitting the procedure used to grow the tree, and there is a natural correspondence between (1.22) and a certain "expectation" taken over a *directed acyclic graph* (DAG) defined in the following way. Let $[0, 1]^p$ be the root of the DAG; for every potential split at $[0, 1]^p$, there is a directed edge to a new vertex, where that vertex is whichever one of the left or right hyperrectangles that contains x. If the node represented by a vertex is one of the possible values of Π, then that vertex is a leaf (a "sink") in the DAG and has no outgoing edges; other vertices carry an outgoing edge for each potential split at that node, with each edge going to another vertex which is again a hyperrectangle containing x.

The previous definition determines the DAG recursively: each vertex in the DAG is a node containing x, with terminal vertices corresponding to terminal nodes. To each terminal vertex v, we associate the value $f(v) := \mu_\pi$ as in (1.22). In addition, each edge $e = (v \to w)$ corresponds to a split at a node v producing a halfspace w of v; associate with this edge a "transition probability"

$$p(e) := \mathbf{P}(s \text{ is chosen at } v \mid \text{current node is } v) =: \mathbf{P}(w \mid v).$$

Given the transition probabilities, the value f may be extended to each vertex v recursively (from the bottom up) via the formula

$$f(v) := \sum_{e:v \to w} \mathbf{P}(w \mid v) f(w).$$

We refer to f as the continuation value at v, and by construction we have

$$\mathbf{E}(T) = f(\text{"root"}) = f([0,1]^p).$$

Alternatively, if we had assigned the values $f'(v) = \mu'_v$ to each terminal vertex and used the transition probabilities

$$p'(e) = \mathbf{P}(s \text{ is chosen at } v \mid \text{current node is } v, X_1 = x_1) = \mathbf{P}'(w \mid v),$$

then we recover $\mathbf{E}(T \mid X_1 = x_1) = f'([0,1]^p)$ after extending f' in the same way as f. In other words, bounding $\mathbf{E}(T \mid X_1 = x_1) - \mathbf{E}(T)$ requires bounding the difference between the two types of continuation values.

As we will show presently, the key in bounding the continuation values will be a bound on the differences between the transition probabilities; loosely speaking, we will need to show that $p'(e) \approx p(e)$. Intuitively, that p and p' are close expresses the property that conditioning on a *single* observation will not affect the probability that a particular split is chosen.

This is a natural property in that the optimal split is computed using all the observations in a particular node, so that conditioning on a single observation should have relatively little effect. Of course, whether this property holds will depend on the specifics of the splitting algorithm used to a construct the tree. For this reason, we shall endow this property with a name and specify it as a high level condition; a set of low level sufficient conditions are then given in Proposition 1.10.

Assumption (Splitting Stability). *For any node v, the total variation distance between the distributions $\{p(e)\}_{e:v\to w}$ and $\{p'(e)\}_{e:v\to w}$ is bounded by a scaled volume of v. Specifically, there* *exists some $\epsilon > 0$ such that for all v,*

$$\mathrm{TV}(p, p') \leq \left(\frac{1}{s|v|}\right)^{1+\epsilon} \qquad \textit{(up to a constant).} \tag{1.23}$$

Here, $|v|$ denotes the volume of the hyperrectangle at v, i.e.,

$$|v| = \left|\prod_{j=1}^{p}(a_j, b_j)\right| = \prod_{j=1}^{p}|b_j - a_j|.$$

Remark. Recall that p and p' are discrete probability distributions: thus, if p and p' are written as vectors of probability masses, then the total variation distance is the L_1 norm between the two vectors.

Since the distribution of X has a density that is bounded above and below, a simple Höeffding bound shows that the number of sample points in v is bounded above and below by $s|v|$, with the constants adjusted so that the failure probability is less than[8] $1/s^2$. Since this is smaller than the required bound $(\log^p s \cdot s)^{-1}$, we may interpret $s|v|$ in (1.23) to be the number of samples in $|v|$ without loss of generality. Relatedly, Wager and Walther (2015)'s Lemma 12 (see also the proof of Lemma 2 in Wager and Athey (2018)) extends to this fact to be uniform across nodes.

The stability assumption places a restriction on procedure used to select optimal splits: namely, if the decision is made on the basis of m points, then conditioning on any one of the points changes the optimal split with probability bounded by $m^{-(1+\epsilon)}$. In practice, most splitting procedures satisfy a stronger bound. A set of sufficient conditions is given in the following proposition.

Proposition 1.10. *Assume that the optimal split at a node v is chosen based on quantities of the following form*

$$f_1(\mu_1, \ldots, \mu_Q), \ldots, f_P(\mu_1, \ldots, \mu_Q)$$

for some P and $Q \geq 1$, where $\mu_1, \ldots, \mu_Q$ are the sample averages of the points being split

$$\mu_k = \frac{1}{n_v} \sum_{i: X_i \in v} m_k(X_i)$$

for some functions m_1, ..., m_Q, where the sum runs over points in v, and n_v denote the number of these points.

Specifically, suppose the optimal split is decided by which f_i achieves the largest value, i.e., the value $\arg\max_i f_i(\mu)$. If $f_1, \ldots, f_P$ are Lipschitz, and the functions $m_1, \ldots, m_Q$ are such that $m_k(X)$ is 1 sub-exponential, then the splitting stability assumption is satisfied.

[8] For example, the probability that a binomial random variable $B(n, p)$ deviates from np by more than $\sqrt{n \log n}$ is less than C/s^2 for some constant C.

Remark. Since X are bounded, the requirement that $m_k(X)$ is sub-exponential allows the use of (1.2) to compute the optimal split.

In general, the conditions in Proposition 1.10 are sufficient to guarantee an exponential bound instead of a polynomial one as in (1.23). Thus, Proposition 1.10 should be viewed as simply providing a plausibility argument that stable splitting rules are commonly encountered in practice.

The next proposition shows that bounds on splitting probabilities automatically imply a related bound on the continuation values.

Proposition 1.11. *Suppose the splitting probabilities satisfy a generic bound $\Delta(\cdot)$ in that*

$$\mathrm{TV}(p, p') \le \frac{\Delta(s|v|)}{\log s} \quad \textit{at each node } v.$$

For example, $\Delta(z) = z^{-(1+\epsilon)}$. Then for any node v containing x but not x_1,

$$|f(v) - f'(v)| \le C\Delta(s|v|)$$

for some constant C not depending on v.

The splitting stability assumption stipulates that $\Delta(z) = z^{-(1+\epsilon)}$, where the factor ϵ allows us to ignore the extra logarithm. In that case, we may put the bounds on $\mathrm{TV}(p, p')$ and $|f - f'|$ together and establish required bound on $\mathbf{Var}\,\mathbf{E}(T \mid X_1)$.

Proposition 1.12. *Suppose that the splitting rule is stable as in* (1.23) *and that $\delta > 1 - \alpha$. For $x \neq x_1$,*

$$|\mathbf{E}(T \mid X_1 = x_1) - \mathbf{E}(T)| = o\left(\frac{1}{s^{1+\epsilon}}\right)$$

for some $\epsilon > 0$. In particular, the off-diagonal entries of $\mathbf{Var}\,\mathbf{E}(T \mid X_1)$ *are $o(s^{-(1+\epsilon)})$ as at least one of x and $\bar{x}$ is distinct from x_1.*

Recall that Proposition 1.9 requires $\delta > 1/2$. Since $\alpha < 1/2$ by definition, the requirement that $\delta > 1 - \alpha$ in Proposition 1.12 is more restrictive. Just like Proposition 1.9, we argue that this requirement is plausibly looser in applications. The reason is that it is used to give the following bound on hyperrectangles v created after L splits

$$|v| \ge \alpha^L.$$

The RHS appears since potential splits may reduce the volume of a node by at most α: but only an exponentially small (i.e., 2^{-L}) proportion of nodes is the result of taking the smallest possible split L times! The "average" node at depth L has volume $(1/2)^L$, so that $\delta > 1/2$ may be more appropriate.

1.3.5 Wrapping Up

Combining Propositions 1.9 and 1.12 with equations (1.17) and (1.18) yields the desired bound (1.15) on the off-diagonal terms of $\mathbf{Var}\,\mathring{T}$ discussed at the beginning of this section. Therefore, Proposition 1.8 applies, and the joint normality of the random forest estimator is established.

1.4 Heuristics and Simulations

The previous sections focused on deriving the asymptotic normality result

$$V^{-1/2}(\mathrm{RF}-\mu) \overset{\text{dist}}{\Longrightarrow} N(0,I), \qquad \text{where } V = \mathbf{Var}\,\mathring{\mathrm{RF}} = \frac{s}{n}\,\mathbf{Var}\,\mathring{T}.$$

Recall our standing convention that $\mathrm{RF} \in \mathbf{R}^q$ is the random forest estimate at $x_1, \ldots, x_q$ and μ is its expectation. According to (1.3), the target function of the random forest is actually $m(x) = \mathbf{E}(Y \mid X = x)$. The results in Wager and Athey (2018) show that $(\mathrm{RF}(x) - m(x))/\sqrt{V} \overset{\text{dist}}{\Longrightarrow} N(0,1)$ pointwise for each $x \in \mathcal{X}$; since we have shown that V is diagonally dominant in that its off-diagonal terms vanish relative to the diagonal, the pointwise result carries over to our multivariate setting, and $V^{-1/2}(\mathrm{RF}-m) \overset{\text{dist}}{\Longrightarrow} N(0,I)$, where $m = (m(x_1), \ldots, m(x_q))$.

Moreover, Wager and Athey (2018) proposes a jackknife estimator that can consistently estimate $\sqrt{V}$ in the univariate case. Our diagonal dominance result implies that the random forest estimates at x and $\bar{x} \in [0,1]^p$ are independent in the limit $n \to \infty$,

$$\begin{aligned} \mathbf{Var}(\mathrm{RF}(x) + \mathrm{RF}(\bar{x})) &= \mathbf{Var}(\mathrm{RF}(x)) + \mathbf{Var}(\mathrm{RF}(\bar{x})) + 2\,\mathbf{Cov}(\mathrm{RF}(x) + \mathrm{RF}(\bar{x})) \\ &\approx \mathbf{Var}(\mathrm{RF}(x)) + \mathbf{Var}(\mathrm{RF}(\bar{x})), \end{aligned}$$

so that the jackknife estimator for the scalar case may be fruitfully applied to obtain confidence bands for *functionals* of the random forest estimates (i.e., expressions involving estimates at more than one point). We expand on this point in the remainder of this section.

The accuracy of the approximation above depends on how fast the off-diagonal terms decay. In this section, we provide a "back of the envelope" bound for the covariance term that may be useful for practitioners. We stress that the following calculations are (mostly) heuristics: as we have shown in the previous section, the covariance term depends on quantities such as $M(x, \bar{x})$, which is in turn heavily dependent on the exact mechanics of the underlying splitting algorithm. Since our aim is to produce a "usable" result, we will now dispense with rigorous analysis.

To begin, the proofs of Propositions 1.9 and 1.12 showed that the asymptotic variance V

has off-diagonal terms which are upper bounded by[9]

$$M(x,\bar{x}) + \log^2 s\left(\sum_{\ell=0}^{\infty} p_\ell\right)\left(\sum_{\ell=0}^{\infty} \bar{p}_\ell\right) = M(x,\bar{x}) + \log^2 s\,\mathbf{E}(L)\,\mathbf{E}(\bar{L}),$$

where $p_\ell = \mathbf{P}(L \geq \ell)$ is the probability that x and x_1 are not separated after ℓ splits and likewise for $\bar{p}_\ell = \mathbf{P}(\bar{L} \geq \ell)$. That is, L is the number of splits before which x and x_1 belong to the same partition. If we denote by I (resp. $\bar{I}$) the indicator variable that X_1 is in the terminal node of x (resp. $\bar{x}$), then the events $\{I = 1\}$ and $\{L = \log_2 s\}$ are equal, so that

$$\mathbf{E}(I \mid X_1 = x_1) = \mathbf{P}(L = \log s) \leq \frac{\mathbf{E}\,L}{\log s}.$$

Replacing the inequality with an approximation, we have $\mathbf{E}\,L \approx (\log s)\,\mathbf{E}(I \mid X_1 = x_1)$. Therefore, we have the approximate bound

$$(\log^4 s)\,\mathbf{E}(I \mid X_1 = x_1)\,\mathbf{E}(\bar{I} \mid X_1 = x_1) \approx (\log^4 s) M(x,\bar{x}).$$

Remark. Taken loosely, this heuristic says that the random forest estimator RF, considered as a function on the domain $\mathcal{X}$, is asymptotically Gaussian with covariance process $(\log^4 s) \cdot M(x,\bar{x})$. We stress that this is *not* implied by our theoretical results, as there we kept the number q of test points fixed.

Towards a useful heuristic, we will consider a bound on the correlation instead of the covariance. In our notation, the result of Wager and Athey (2018) lower bounds $M(x,x)$ (and $M(\bar{x},\bar{x})$), while our paper provides an *upper* bound on $M(x,\bar{x})$. Ignoring the logarithmic terms, we have

$$\left|\frac{\mathbf{Cov}(\mathrm{RF}(x), \mathrm{RF}(\bar{x}))}{\sqrt{\mathbf{Var}\,\mathrm{RF}(x) \cdot \mathbf{Var}\,\mathrm{RF}(\bar{x})}}\right| \approx \frac{M(x,\bar{x})}{\sqrt{M(x,x)M(\bar{x},\bar{x})}}.$$

Recall that $M(x,\bar{x}) = \mathbf{E}[\mathbf{E}(I \mid X_1)\,\mathbf{E}(\bar{I} \mid X_1)]$, which decays as $\bar{x}$ moves away from x. Using the previous expression (note that $M(x,x) \approx M(\bar{x},\bar{x})$ due to symmetry between x and $\bar{x}$), we can bound the correlation from purely geometric considerations. Since the integrand

$$\mathbf{E}(I \mid X_1)\,\mathbf{E}(\bar{I} \mid X_1)$$

decays as X_1 moves away from x (and $\bar{x}$), we may imagine that in the integral

$$M(x,x) = \int_{x_1} \mathbf{E}(I \mid X_1 = x_1)^2 dx_1,$$

points x_1 that are near x make the largest contribution, say, those points in a L_∞-box of side lengths d with volume d^p, i.e., $\{y \in [0,1]^p : \|x - y\|_\infty \leq d/2\}$. If we accept this, then

[9] This is a very crude upper bound as we have dropped the quantity $\Delta(\alpha^\ell s)$ from the infinite series.

the contributions for the integral $M(x, \bar{x})$ would come from points that are within $d/2$ of both x *and* $\bar{x}$, and to a first degree approximation, the volume of these points $\{y \in [0,1]^p : \|x - y\|_\infty \leq d/2, \|\bar{x} - y\| \leq d/2\}$ is

$$(d - z_1) \dots (d - z_p) \approx d^p - (z_1 + \dots + z_p) d^{p-1}, \quad \text{where } z_i = |x_j - \bar{x}_j|,$$

where the approximation is accurate if $|z_i| \ll 1$. Dividing through by d^p, the proportion of the volume of the latter set is $1 - \frac{1}{d}\|x - \bar{x}\|_1$, which leads to the heuristic

$$\left| \frac{\mathbf{Cov}(\mathrm{RF}(x), \mathrm{RF}(\bar{x}))}{\sqrt{\mathbf{Var}\,\mathrm{RF}(x) \cdot \mathbf{Var}\,\mathrm{RF}(\bar{x})}} \right| \approx 1 - c\|x - \bar{x}\|_1, \qquad \text{for some constant } c.$$

The RHS has the correct scaling when $x = \bar{x}$, i.e., the correlation equals one when $\|x - \bar{x}\|_1 = 0$. To maintain correct scaling at the other extreme with $\|x - \bar{x}\|_1 = p$, we should take $c = 1/p$, so that

$$\left| \frac{\mathbf{Cov}(\mathrm{RF}(x), \mathrm{RF}(\bar{x}))}{\sqrt{\mathbf{Var}\,\mathrm{RF}(x) \cdot \mathbf{Var}\,\mathrm{RF}(\bar{x})}} \right| \approx 1 - \frac{1}{p} \sum_{i=1}^{p} |x_i - \bar{x}_i|.$$

Of course, this heuristic is trivially incorrect in that it does not depend on s; our theoretical results show that even for non-diametrically opposed points, the correlation drops to zero as $s \to \infty$. Therefore, another recommendation is to use

$$\left| \frac{\mathbf{Cov}(\mathrm{RF}(x), \mathrm{RF}(\bar{x}))}{\sqrt{\mathbf{Var}\,\mathrm{RF}(x) \cdot \mathbf{Var}\,\mathrm{RF}(\bar{x})}} \right| \approx \min\left(1 - \frac{s^\epsilon}{p} \sum_{i=1}^{p} |x_i - \bar{x}_i|,\ 0 \right), \tag{1.24}$$

for some $\epsilon > 0$, where the decay s^ϵ comes from considering the decay of $M(x, \bar{x})$ as $\bar{x}$ moves away from x (c.f. the statement and proof of Proposition 1.9). In any case, our theoretical results also suggest that the sign of the correlation is *positive*: non-zero covariances are driven by the possibility that the two points x and $\bar{x}$ may belong to the same terminal node, in which case perfect correlation obtains as $T(x) = T(\bar{x})$. This intuition is also supported by our simulation results below (c.f., Figure 1).

1.4.1 Confidence Intervals of Sums and Differences

One important implication is that omitting cross covariances will *overestimate* the variance of differences of random forest estimates, while *underestimating* the variance of sums. Coming back to the discussion at the beginning of this section, suppose we are interested in the difference $f(x) - f(\bar{x})$ of our target function f and compute a random forest estimator $\mathrm{RF} = (\mathrm{RF}(x), \mathrm{RF}(\bar{x}))$, along with jackknife estimates of the variances as in Wager and Athey (2018)

$$\hat{V}(x) \approx \mathbf{Var}\,\mathrm{RF}(x) \quad \text{and} \quad \hat{V}(\bar{x}) \approx \mathbf{Var}\,\mathrm{RF}(\bar{x}).$$

Since

$$\mathbf{Var}(\mathrm{RF}(x) - \mathrm{RF}(\bar{x})) = \mathbf{Var}\,\mathrm{RF}(x) + \mathbf{Var}\,\mathrm{RF}(\bar{x}) - 2\rho\sqrt{\mathbf{Var}\,\mathrm{RF}(x)\,\mathbf{Var}\,\mathrm{RF}(\bar{x})}$$

where $\rho > 0$ is the correlation, $\hat{V}(x) + \hat{V}(\bar{x})$ overestimates the variance of the difference. In particular, confidence intervals calculated using the asymptotic Gaussian approximation with standard deviation $\sqrt{\hat{V}(x) + \hat{V}(\bar{x})}$ would be too conservative.

On the other hand, suppose we are interested in computing an average of target function f across different points, i.e., a quantity of the form $\alpha_1 f(x_1) + \cdots + \alpha_q f(x_q)$ where $\alpha_1, \ldots, \alpha_q \geq 0$ and $\sum \alpha_k = 1$. A similar calculation shows that confidence intervals based on the standard deviation

$$\sqrt{\alpha_1^2 \hat{V}(x_1) + \cdots + \alpha_q^2 \hat{V}(x_q)}$$

are too narrow (i.e., their coverage rates are lower than their nominal coverage rates). In this way, heuristics such as (1.24) allow us to construct tighter or more accurate confidence intervals; the second set of the simulation experiments below gauges the effectiveness of (1.24) in doing so.

1.4.2 Simulations

Correlation Structure

In this section, we conduct numerical experiments on the correlation structure of random forests; we set $p = 2$, so that the covariates X are distributed on the unit square. The distribution of X is chosen to be "four-modal"

$$X \sim \begin{cases} \bar{N}(\mu_1, I_2) & \text{with probability } 1/4 \\ \bar{N}(\mu_2, I_2) & \text{with probability } 1/4 \\ \bar{N}(\mu_3, I_2) & \text{with probability } 1/4 \\ \bar{N}(\mu_4, I_2) & \text{with probability } 1/4 \end{cases} \quad \text{where} \quad \begin{cases} \mu_1 = (0.3, 0.3)^\intercal \\ \mu_2 = (0.3, 0.7)^\intercal \\ \mu_3 = (0.7, 0.3)^\intercal \\ \mu_4 = (0.7, 0.7)^\intercal \\ I_2 = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right) \end{cases}$$

and $\bar{N}$ denotes a truncated multivariate Gaussian distribution on the unit square.[10] Thus, X has a bounded density on the unit square, and has four peaks at $\mu_1, \ldots, \mu_4$. The distribution of Y conditional on $X = (x_1, x_2)$ is

$$Y \sim \frac{x_1 + x_2}{2} + \frac{1}{5} N(0, 1).$$

The random splitting probability is to $\delta = 1/2$, and the regularity parameters are $\alpha = 0.01$ and $k = 1$, so that the tree is grown to the fullest extent (i.e., terminal nodes may contain a

[10] That is, $\bar{N}(\mu, \Sigma)$ denotes the conditional distribution of $x \sim N(\mu, \Sigma)$ on the event $x \in [0, 1]^2$.

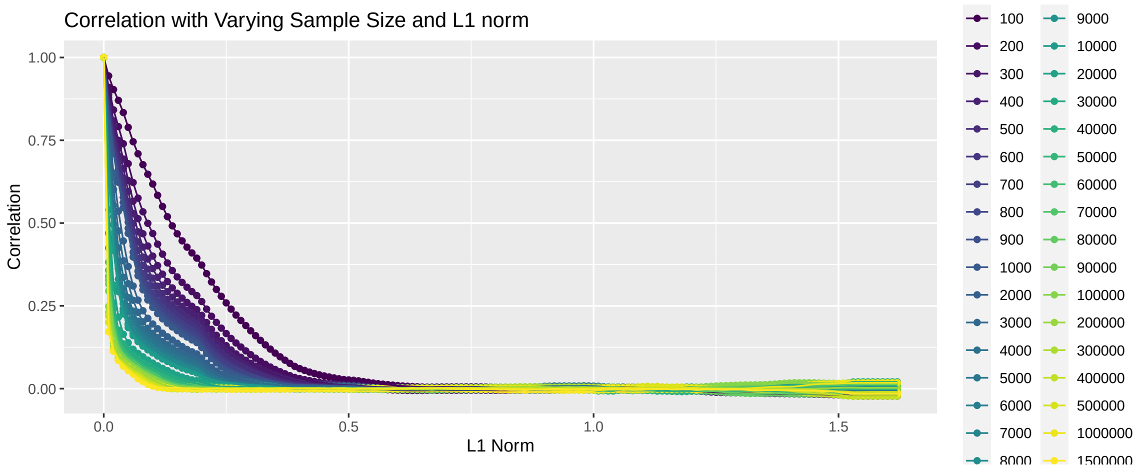

Figure 1-1: Correlation as a function of sample size and L_1 norm.

single observation), with each terminal node lying on the 101×101 grid of the unit square. For each sample size n, five thousand trees are grown, and the estimates are aggregated to compute the correlation.

Figure 1-1 plots the correlation of estimates at x and $\bar{x}$ as a function of the L_1 norm $\|x - \bar{x}\|_1$. The calculation is performed by first fixing x, then calculating the sample correlation (across five thousand trees) as $\bar{x}$ ranges over each cell: the correlation is associated with the L_1 norm $\|x - \bar{x}\|_1$. This process is then repeated by varying the reference point x, and the correlation at $\|x - \bar{x}\|_1$ is the average of the correlations observed. The figure demonstrates that the linear heuristic (1.24) given in the previous section is conservative: it is evident that correlation decreases super-linearly as x and $\bar{x}$ become separated.

Figure 1-2 plots the correlation on a logarithmic scale, which shows that that correlation decay is exponential in a neighborhood of unity. In other words, simulations suggest that the correct heuristic may be of the shape

$$\left| \frac{\mathbf{Cov}(\mathrm{RF}(x), \mathrm{RF}(\bar{x}))}{\sqrt{\mathbf{Var}\,\mathrm{RF}(x) \cdot \mathbf{Var}\,\mathrm{RF}(\bar{x})}} \right| \approx e^{-\lambda \|x - \bar{x}\|_1} \quad \text{for a suitable } \lambda.$$

Confidence Intervals and Coverage Rates

Our second set of experiments examines the coverage rates of confidence intervals computed from the asymptotic Gaussian approximation *with* and *without* taking into account cross covariance terms. The simulation setup is the same as in the previous experiment, except that trees are grown until the number of leaf observations is equal to five instead of one.[11]

[11] This is done for ease of computation: if we were to increase the depth by 1 or 2 in order to have size-1 leaves, the memory requirements would be computationally infeasible. The current set of experiments require a peak memory usage of up to 150 gigabytes when run in parallel.

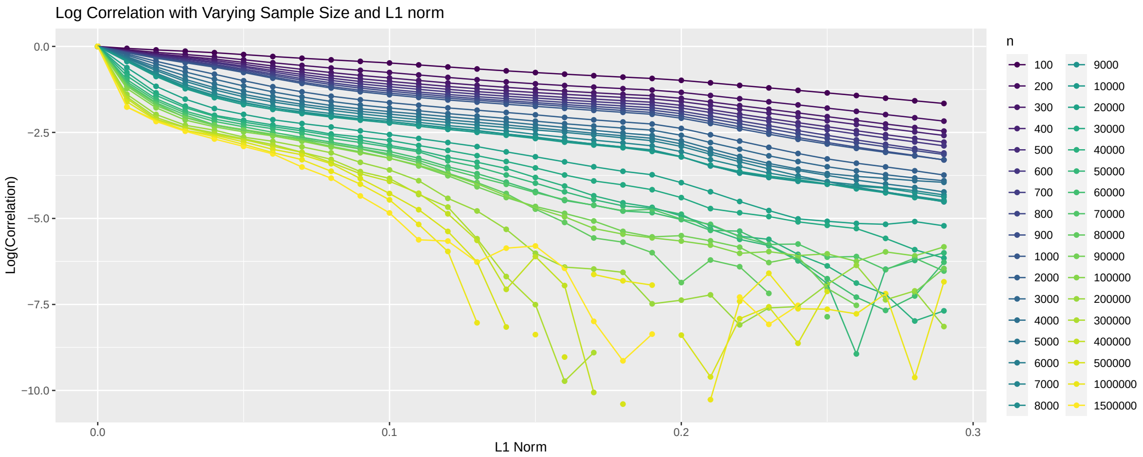

Figure 1-2: Logarithm of correlation as a function of sample size and L_1 norm.

Each random forest is the average of five thousand individual trees, and we present results for sample sizes $n \in \{1000, 2500, 5000, 7500, 10000, 15000\}$ with subsample sizes $n^{0.9}$.

Our goal is to gauge the effect of non-zero covariance terms on the resulting confidence intervals and to gauge the effectiveness of the heuristic (1.24). We fix test points

$$x_1 = (0.45, 0.5), \quad x_2 = (0.50, 0.5), \quad x_3 = (0.80, 0.5)$$

and consider the coverage rates of two kinds of confidence intervals based on the asymptotic Gaussian approximation. The first kind assumes that the cross covariances of the random forest estimate is zero: this is justified theoretically, as the asymptotic covariance matrix is diagonal; the second kind uses the heuristic (1.24) to assign a positive correlation between the estimates, then uses the resulting (adjusted) variance. In both cases, the variances are calculated using[12] the jackknife estimator as in Wager and Athey (2018).

The results are presented Table 1.1, where we report coverage rates for three functionals (the first column). The first and second functionals are differences of the target function; the first functional takes the difference between two nearby points (x_1 and x_2), whereas the second functional uses points farther away (x_1 and x_3). The third functional takes the average between the three points. The third and fourth columns reports the empirical coverage rates for confidence intervals with a *nominal coverage rate* of 95% (i.e., a multiplier of ± 1.96 is used). The first row in each section reports the variance used, where $\hat{V}_1$, $\hat{V}_2$, and $\hat{V}_3$ are jackknife estimators for the variances of $\mathrm{RF}(x_1)$, $\mathrm{RF}(x_2)$, and $\mathrm{RF}(x_3)$, respectively. The numbers ρ_{ij} are given by[13] heuristic in (1.24), specifically

$$\rho_{ij} = \left(1 - n^{0.3} \times \frac{|x_{i1} - x_{j1}| + |x_{i2} - x_{j2}|}{2}\right)^+, \quad \text{where } \eta^+ \text{ denotes } \max(\eta, 0).$$

[12] Specifically, we use a modified version of the GRF package of Athey, Tibshirani and Wager (2017).

[13] The heuristic in (1.24) uses s^ϵ instead of n^ϵ; the two parameterizations are identical because $s = n^\beta$ is itself a fractional power of n.

For example, the covariance in the fourth column of the third functional is

$$\mathbf{Var}\left[\frac{\mathrm{RF}(x_1)+\mathrm{RF}(x_2)+\mathrm{RF}(x_3)}{3}\right] = \sum_{1\le i,j\le 3}\frac{\sigma_{ij}}{9}, \quad \sigma_{ij} = \begin{cases} \hat{V}_i & \text{when } i=j \\ \rho_{ij}\sqrt{\hat{V}_i\hat{V}_j} & \text{when } i\neq j \end{cases} \tag{1.25}$$

Remark. Specifically, we chose the exponent $\epsilon = 0.3$. As in the discussion following (1.24), the exponent corresponds to the decay rate ϵ in Proposition 1.9, and it is easy to see that it should be a small number between 0 and 1. To calibrate the exponent, one method—used in this experiment—is to compute the sample correlation at a single pair of points across the trees in the random forest, and then choosing ϵ to match (1.24) with the sample correlation.

The results make intuitive sense. The coverage rates in the third column are fairly close to the nominal 95% for the second functional: since x_1 and x_3 are far apart, their correlation is close to zero even at modest sample sizes, and ignoring the covariance term is a good approximation. On the other hand, this approximation is not as good for the first and third functionals, both of which involve correlations that are far from zero: for example, the sample correlation of $\mathrm{RF}(x_1)$ and $\mathrm{RF}(x_2)$ is $\hat{\rho}_{12} \approx 70\%$. Since the first functional is a difference, its confidence intervals are too conservative, with coverage rates exceeding 97% at all sample sizes; the third functional, being a sum, has confidence intervals that are too short, with a peak coverage rate of 94%. For all three functionals, the coverage rates approach the nominal coverage rate as the sample size increases, reflecting the fact that covariance tends to a diagonal matrix.

For the first and third functionals, using an adjusted variance (fourth column) that incorporates the heuristic improves the coverage rate. For the first functional, the adjustment narrows the confidence intervals and shrinks the coverage rate towards the nominal 95%, though the intervals remain conservative. The improvement is more evident for the third functional, where the coverage rate is now much closer to the nominal rate at all sample sizes. In addition, the heuristic also maintains good performance for the second functional.

Table 1.1: Coverage rates of confidence intervals with and without cross covariances.

Functional	Sample Size	Coverage Rate (I)	Coverage Rate (II)
$\mathrm{RF}(x_2) - \mathrm{RF}(x_1)$	-	$\hat{V}_1 + \hat{V}_2$	$\hat{V}_1 + \hat{V}_2 - 2\rho_{12}\sqrt{\hat{V}_1\hat{V}_2}$
	1000	99.8%	98.2%
	2500	99.4%	96.9%
	5000	98.9%	98.9%
	7500	98.5%	96.6%
	10000	97.4%	96.0%
	15000	97.3%	96.6%
$\mathrm{RF}(x_3) - \mathrm{RF}(x_1)$	-	$\hat{V}_1 + \hat{V}_3$	$\hat{V}_1 + \hat{V}_3 - 2\rho_{13}\sqrt{\hat{V}_1\hat{V}_3}$
	1000	95.9%	95.9%
	2500	95.8%	95.8%
	5000	95.6%	95.6%
	7500	95.7%	95.7%
	10000	95.7%	95.7%
	15000	95.2%	95.2%
$\dfrac{\mathrm{RF}(x_1) + \mathrm{RF}(x_2) + \mathrm{RF}(x_3)}{3}$	-	$\dfrac{\hat{V}_1 + \hat{V}_2 + \hat{V}_3}{9}$	(see equation (1.25))
	1000	89.4%	93.5%
	2500	91.2%	94.6%
	5000	92.6%	94.9%
	7500	92.9%	94.7%
	10000	94.0%	95.3%
	15000	93.9%	94.5%

In general, this experiment suggests that the covariance adjustment is worth considering, at least for modest sample sizes (say, for $n \leq 5000$). A possible avenue for future work is the development of more accurate heuristics.

1.5 Conclusion

Random forests and tree-based methods form an important part of an applied data analysis toolkit. In this paper, we study the covariance between random forest estimates at multiple points. We develop a novel construction of a directed acyclic graph that keeps track of the splitting probabilities when knowledge of one point is known (Propositions 1.11 and 1.12). As part of the proof, we establish stability properties of a class of splitting rules (see Proposition 1.10). We also identify (Proposition 1.9) $M(x, \bar{x})$, which (roughly) captures the likelihood of two points belonging to the same terminal node, as a key quantity in controlling the off-diagonal terms of the covariance matrix of the multivariate random forest.

In this way, this paper provides a theoretical basis for performing inferences on functionals of the underlying target function (e.g., a heterogeneous treatment effect) when the functional is based on values of the target function at multiple points in the feature space. Specifically, we show that the covariance vanishes in the limit relative to the variance, and provide heuristics on the size of the correlation in finite samples.

We close with discussing a couple avenues for future research. The first is extending our framework to cover categorical or discrete-valued features. Here, new assumptions would be required in order to maintain the guarantee that node sizes are "not too small." Second, our bounds—after potential improvements—on the covariance matrix of the random forest may be used with the recent results of Chernozhukov, Chetverikov and Kato (2017); Chen (2018) in order to provide finite sample Gaussian approximations. This would provide a more sound theoretical underpinning for correlation heuristics, making random forest models more useful and more user-friendly to practitioners.

1.6 Appendix

Proof of Proposition 1.7. For random vectors in $\mathbf{R}^q$, define the inner product

$$\langle X, Y \rangle := \mathbf{E}(X^\top M Y). \tag{1.26}$$

For each subset $A \subseteq \{1, \dots, n\}$, let H_A be the set of square-integrable random vectors of the form $g(X_i : i \in A)$, where g is a function of $|A|$ arguments, satisfying the condition that

$$\mathbf{E}(g(X_i : i \in A) \mid \{X_i : i \in B\}) = 0$$

for all subsets $B \subsetneq A$. It is easy to see that collection H_A are pairwise orthogonal as A ranges over subsets of $\{1, \dots, n\}$. By induction on $r = |A|$, the direct sum $\bigoplus_{B \subseteq A} H_B$ is equal to the set of all statistics which are functions of $\{X_i : i \in A\}$. In particular, $\bigoplus_A H_A$ is the set of all (square-integrable) statistics based on $\{X_1, \dots, X_n\}$. When the variables $\{X_1, \dots, X_n\}$ are IID, then H_A depends only on $|A|$ in that there exist collections of functions $H_0, H_1, \dots, H_n$, where H_k is a collection of k-ary functions, such that

$$H_A = \{g(X_i : i \in A) : g \in H_{|A|}\}.$$

The proof is completed by letting f_k be the projection of f onto H_k according to the inner product given in (1.26). □

The proof of Lemma 1.6 will make use the following multivariate formulation of Lyapunov's CLT for triangular arrays, adapted from Billingsley (2008).

Multivariate Central Limit Theorem for Triangular Arrays. *Let* $\{X_{ni} : 1 \le i \le m_n, n \ge 1\}$ *be a triangular array of random vectors satisfying the following conditions:*

(a) $m_n \to \infty$ *as* $n \to \infty$;

(b) for each $n \ge 1$, *the vectors* $\{X_{ni} : 1 \le i \le m_n\}$ *are jointly independent; and*

(c) there exists some $\eta > 0$ *such that* $\mathbf{E}\,\|X_{ni}\|^{2+\eta} < \infty$ *for all* $n \ge 1$ *and* $1 \le i \le m_n$,

Let μ_{ni} *denote the* (d*-dimensional) mean of* X_{ni}, *and define*

$$s_n^2 = \sum_{i=1}^{m_n} \mathbf{E}(\|X_{ni} - \mu_{ni}\|^2), \quad \nu_n = \sum_{i=1}^{m_n} \mathbf{E}(\|X_{ni} - \mu_{ni}\|^{2+\eta}), \quad \rho_n = \frac{\nu_n}{s_n^{2+\eta}}.$$

If $\rho_n \to 0$ *as* $n \to \infty$ *(the Lyapunov condition) and* $V_n := \mathbf{Var}\sum_{i=1}^{m_n}(X_{ni} - \mu_{ni})$ *is positive definite, then*

$$V_n^{-1/2}\sum_{i=1}^{m_n}(X_{ni} - \mu_{ni}) \xRightarrow{\text{dist}} N(0, I).$$

Proof of Lemma 1.6. We proceed by verifying the hypotheses of the CLT stated above. In our setting, $m_n = n$ and $X_{ni} = \frac{s}{n}T_1(Z_i)$ for $1 \le i \le n$, so that conditions (a) and (b) are automatically satisfied (recall that $\{Z_i : 1 \le i \le n\}$ are IID). Since $\mathbf{E}(|Y|^3 \mid X = x)$ is bounded and T_1 is bounded by $\sup \mathbf{E}(|Y| \mid X = x)$, condition (c) is satisfied as well with

$$\eta = 1.$$

Furthermore, V_n is positive-definite as there are no linear dependence relations among the tree estimates at distinct points.

Therefore, we only need to verify the Lyapunov condition. In other words, we need to show that

$$\frac{\sum_{i=1}^n \mathbf{E}\,\|T_1(Z_i) - \mu\|^{2+\eta}}{\sum_{i=1}^n (\mathbf{E}\,\|T_1(Z_i) - \mu\|^2)^{\frac{2+\eta}{2}}} \to 0, \qquad \text{for } \eta = 1. \tag{1.27}$$

We may develop the integrand in the numerator as follows

$$\|T_1(Z_i) - \mu\|^3 = \left[\sum_{k=1}^q |T_{1k}(Z_i) - \mu_k|^2\right]^{\frac{3}{2}} \le \left[\sum_{k=1}^q |T_{1k}(Z_i) - \mu_k|\right]^3,$$

where the last step uses the subadditivity of the square root. Expanding the sum on the extreme RHS yields

$$\begin{aligned} \|T_1(Z_i) - \mu\|^3 \le &\sum_k |T_{1k} - \mu_k|^3 \\ &+ 2\sum_{k \ne l} |T_{1k} - \mu_k|^2 |T_{1l} - \mu_l| \\ &+ \sum_{k,l,m} |T_{1k} - \mu_k||T_{1l} - \mu||T_{1m} - \mu_m|, \end{aligned} \tag{1.28}$$

where the second summand runs through indices $1 \leq k \neq l \leq q$ and the third summand runs through indices $\{1 \leq k, l, m \leq q\}$ with k, l, and m being pairwise distinct.

We will now bound the second and the third summands in terms of the first summand. For tidiness, fix pairwise distinct indices $1 \leq k, l, m \leq q$ and set

$$A = |T_{1k} - \mu_k|, \quad B = |T_{1l} - \mu_l|, \quad C = |T_{1m} - \mu_m|.$$

Using Hölder's inequality, we have $\mathbf{E}(A^2 B) \leq \mathbf{E}(A^3)^{2/3}\,\mathbf{E}(B^3)^{1/3}$. Since $\mathbf{Var}(Y \mid X = x)$ is bounded below, there is a lower bound on $\mathbf{E}(A^3)$ that is uniform over the location of x_k. To see this, note that Jensen's inequality implies

$$\mathbf{E}(A^2) \leq \mathbf{E}(A^3)^{2/3},$$

where the LHS is bounded below (i.e., it is the variance of the tree estimate at the point x_k) over points in $\mathcal{X}$. Together with the fact that the third conditional moment is bounded above, this implies that there is some constant K—not depending on k, l, and m—for which

$$\mathbf{E}(B^3) \leq K\,\mathbf{E}(A^3).$$

Applying this to the bound from Hölder's inequality above, we have $\mathbf{E}(A^2 B) \leq \mathbf{E}(A^3)$ up to a constant. A similar calculation shows that $\mathbf{E}(ABC) \leq \mathbf{E}(A^3)$ as well.

Substituting the bounds into (1.28) allows us to bound each summand, which yields

$$\mathbf{E}\sum_{i=1}^{n} \|T_1(Z_i) - \mu\|^{2+\eta} \leq \sum_{k=1}^{q}\sum_{i=1}^{n} \mathbf{E}\,|T_{1k}(Z_i) - \mu_k|^{2+\eta}, \quad \text{up to a multiplicative constant.}$$

The univariate version of the Lyapunov condition, namely, that

$$\frac{\sum_{i=1}^{n} \mathbf{E}\,|T_{1k}(Z_i) - \mu_k|^{2+\eta}}{\sum_{i=1}^{n} (\mathbf{E}\,|T_{1k}(Z_i) - \mu_k|^2)^{\frac{2+\eta}{2}}} \to 0$$

was established in Theorem 8 of Wager and Athey (2018). The two previous displays now imply the multivariate analog (1.27), which finishes the proof. □

Proof of Proposition 1.8. We will prove the slightly more general statement that if A_n and B_n are two sequences of square matrices with bounded entries such that

$$B_{ii} > \delta \text{ for some } \delta \text{ for all } n \quad \text{and} \quad A_{ii} \geq \frac{B_{ii}}{\log n}$$

and $A_{ij} = o(1/\log n)$, then $\operatorname{tr}(A^{-1}B) \to 0$. To prove this, start with the determinant formula $\det A = \sum_\pi (-1)^{\operatorname{sgn}\pi} \prod_{i=1}^{q} a_{i\pi_i}$, where the sum runs over permutations π of $\{1, \ldots, n\}$ and $\operatorname{sgn}\pi$ is the sign of the permutation. Since the off-diagonal entries of A_{ij} are assumed to vanish relative to A_{ii}, we have $|\det A| \sim \prod A_{ii}$, where the notation $a \sim b$ stands for $c|b| \leq |a| \leq c'|b|$ for constants c and c' not depending on n. Next, recall Cramer's Rule

$$(A^{-1})_{ii} = \det A_{-i} / \det A,$$

where A_{-i} is the matrix A with its i-th row and i-th column removed. A similar argument shows that $|\det A_{-i}| \sim \prod_{j \neq i} A_{jj}$, whence

$$(A^{-1})_{ii} \sim \frac{1}{A_{ii}}.$$

In particular, the i-th diagonal entry of the matrix $A^{-1}B$ is given by

$$(A^{-1}B)_{ii} = (A^{-1})_{ii} B_{ii} + \sum_{j \neq i} (A^{-1})_{ij} B_{ji} \sim \frac{B_{ii}}{A_{ii}} \leq \log n,$$

where the final relation is due to the fact that $(A^{-1})_{ij}$ is itself a polynomial in the entries of A (viz., the cofactor matrix of A) divided by the determinant. Therefore, the trace of $A^{-1}B$ is on the order of $\log n$, since the dimension $q \times q$ of each matrix is fixed. Using the subsample size $s = n^{\beta}$, so that $s/n = n^{-(1-\beta)}$ completes the proof. □

Proof of Proposition 1.9. Recall that the splitting algorithm has a probability δ chance of splitting on the j-th axis. Since each terminal node contains a bounded number of points, the number of terminal nodes is equal (up to constant) to the subsample size s. Therefore, the number of splits required to reach a terminal node is bounded (by a constant) by $\log_2 s/K = \log s/K$, where $K = 2k - 1$ is the the maximum size of a leaf.

Since $x \neq \bar{x}$, we have

$$0 < \|x - \bar{x}\|_\infty \leq \|x - x_1\|_\infty + \|\bar{x} - x_1\|_\infty$$

for all $x_1 \in \mathcal{X}$. In particular, given any x_1 there exists some $j \in \{1, \ldots, p\}$ and a constant β for which either $|x_j - x_{1j}| > \beta$ or $|\bar{x}_j - x_{1j}| > \beta$. Without loss of generality, we may assume that the former case holds. Certainly, a necessary condition for $X_1 = x_1$ to belong to the same leaf node as x (i.e., a necessary condition for $\{I = 1\}$) is for the length of the first axis of that leaf node to be larger than β.

Let $c_j(x)$ denote the number of splits in coordinate j along the sequence of splits leading to the terminal node containing x. By our randomization assumption, each split has at least an independent chance δ of being chosen, and since we cycle through each coordinate (c.f., Assumption 2),

$$c_j(x) \succeq \frac{1}{p} B\left(\log \frac{s}{K}, \delta\right) \qquad \text{where } \succeq \text{ stands for stochastic dominance.}$$

Per Assumption 4, that each split along the j-th axis decreases its length by a factor of at least $(1-\alpha)$. Since splitting begins in the unit hypercube,

$$(1-\alpha)^{c_1(x)} \geq \beta \implies c_1(x) \leq \frac{\log \beta}{\log(1-\alpha)} =: \rho.$$

Since $\{I = 1\}$ requires that the length of the first axis to exceed β (a constant), this proves

$$\mathbf{E}(I \mid X_1 = x_1) \leq \mathbf{P}\left[B\left(\log \frac{s}{K}, \delta\right) \leq p\rho\right].$$

Since $p\rho$ is a constant, we may conclude

$$\mathbf{P}\left[B\left(\log \frac{s}{K}, \delta\right) \leq p\rho\right] \leq (1-\delta+o(1))^{\log s/K} = \left(\frac{1}{s}\right)^{\log \frac{1}{1-\delta+o(1)}}.$$

Finally, recall base of the logarithm is two since the tree is binary. Therefore, if we choose $\delta > 1/2$, the exponent exceeds 1 and the proof is complete. □

Proof of Proposition 1.10. The easiest case is the splitting decision in the root node $[0,1]^P$, so we start here. We prove the result by introducing a coupling between the splitting decisions with and without conditioning on $X_1 = x_1$

$$\begin{aligned} S &= \arg\max_i f_i\left(\frac{1}{s}\sum_{i=1}^{s} m_1(X_i), \ldots, \frac{1}{s}\sum_{i=1}^{s} m_Q(X_i)\right) =: f_i \\ S' &= \arg\max_i f_i\left(\frac{1}{s}\left[m_1(x_1) + \sum_{i=2}^{s} m_1(X_i)\right], \ldots, \frac{1}{s}\left[m_Q(x_1) + \sum_{i=2}^{s} m_Q(X_i)\right]\right) =: f_i'. \end{aligned} \tag{1.29}$$

Here, S is the split made on the sample $X_1, \ldots, X_s$ and S' is the split made on the sample conditional on $X_1 = x_1$. Note that we may assume without loss of generality that the splits are not randomly chosen, since on that event the splitting probabilities are trivially equal. Clearly, a necessary condition for $S \neq S'$ is the existence of a pair $1 \leq i \neq j \leq P$ for which

$$f_i > f_j \quad \text{but} \quad f_j' > f_i'.$$

Since f is Lipschitz and its arguments are sub-exponential by assumption, the quantities f_i and f_j concentrate around their respective limits $f_i(\mathbf{E}\, m_1, \ldots, \mathbf{E}\, m_Q)$ and $f_j(\mathbf{E}\, m_1, \ldots, \mathbf{E}\, m_Q)$; hence, whenever $f_i(\mathbf{E}\, m_1, \ldots, \mathbf{E}\, m_Q) > f_j(\mathbf{E}\, m_1, \ldots, \mathbf{E}\, m_Q)$ we will have

$$f_i - f_j > \frac{1}{s} \quad \text{with probability at least } 1 - O(e^{-cs}) \text{ for some constant } c.$$

However, the difference of the arguments of f_i in (1.29) differ by at most $1/s$, i.e., the difference coming from $|m_1(x_1) - m_1(X_1)|/s$. By Lipschitz continuity, a change of $1/s$ in the arguments changes the function values by a proportional amount, which renders $f_j' > f_i$ impossible when $f_i - f_j > 1/s$. It follows that $(f_i > f_j, f_j' > f_i')$ occurs with probability at most $O(e^{-cs})$, and we finish by taking a union over the $\binom{P}{2}$ pairs (i,j). This result for $[0,1]^p$ will be referred to as the base case.

Note that the above actually proves something stronger, namely that for every split τ,

$$\mathbf{P}(S \neq s \mid S' = \tau) < e^{-cs} \quad \text{and} \quad \mathbf{P}(S' \neq s \mid S = \tau) < e^{-cs}.$$

It follows that for any τ, the total variation distance between $(X_2, \dots, X_s \mid S = \tau)$ and $(X_2, \dots, X_s \mid S' = \tau)$ is at most e^{-cs}. To see this, note that S and S' are functions of X_i only, so that the densities of the two conditional distributions are

$$p(x) = \mathbf{1}(S(x) = \tau)\frac{p(x)}{\mathbf{P}(S = \tau)} \quad \text{and} \quad p'(x) = \mathbf{1}(S'(x) = \tau)\frac{p(x)}{\mathbf{P}(S' = \tau)},$$

respectively. We may assume without loss of generality that $\mathbf{P}(S' = \tau) \geq \mathbf{P}(S = \tau)$ so that the total variation is

$$\begin{aligned}\int |p(x) - p'(x)| &= \int_{S=\tau} p(x) - p'(x) + \int_{S\neq\tau, S'=\tau} p'(x) \\ &= 1 - \frac{\mathbf{P}(S' = \tau, S = \tau)}{\mathbf{P}(S' = \tau)} + \frac{\mathbf{P}(S' = \tau, S \neq \tau)}{\mathbf{P}(S' = \tau)} = 2\,\mathbf{P}(S \neq \tau \mid S' = \tau) < e^{-cs}.\end{aligned}$$

The upshot is that when considering the splitting probability in the next node, we can ignore the difference in the distribution of $X_2, \dots, X_s$ when conditioning on $S = \tau$ versus conditioning on $S' = \tau$ and pay a cost $O(e^{-cs})$.

Now consider bounding the difference of the splitting probabilities at the next split

$$\mathbf{P}(S_2 = s \mid S_1 = \tau) - \mathbf{P}(S_2 = s \mid S_1 = \tau, X_1 = x_1).$$

Again, the strategy is to find a coupling (S_2, S_2') such that

$$S_2 \sim (S_2 \mid S_1 = \tau) \quad \text{and} \quad S_2' \sim (S_2 \mid S_1 = \tau, X_1 = x_1)$$

with $\mathrm{TV}(S_2, S_2') \leq e^{-s|v|}$, where v is the hyperrectangle corresponding to one of the halfspaces produced by τ. Since the distribution of $X_1, \dots, X_s$ conditional on $S_1 = \tau$ differs from its unconditional distribution by an amount e^{-cs} in the total variation distance, we could use the following coupling

$$S_2 = \arg\max f_i\left(\frac{1}{n_v}\sum_{X_i \in v} m_1(X_i), \dots, \frac{1}{s|v|}\sum_{X_i \in v} m_Q(X_i)\right)$$

$$S_2' = \arg\max f_i\left(\frac{1}{n_v}\sum_{X_i' \in v} m_1(X_i'), \dots, \frac{1}{s|v|}\sum_{X_i \in v} m_Q(X_i')\right)$$

where X_i follows the distribution of $(X_1, \ldots, X_s)$ conditional on $S_1 = \tau$ and X_i' follows the distribution conditional on $S_1 = \tau$ and $X_1 = x_1$. By conditioning on $S_1 = \tau$ instead of $S_1' = \tau$, we increase the total variation by an amount e^{-cs} via the triangle inequality.

Now the rest of the proof is the same as in the base case, noting that with high probability, the number n_v of points in v is equal to $s|v|$ up to an multiplicative constant $(1 - \eta)$ with probability $e^{-s\eta^2}$. The previous bounds are applied recursively at each depth l of the DAG. At depth l, we incur an "approximation cost" from the total variation distance that is bounded by $e^{-|v|s}$. Since $s|v| \geq \alpha^l$, it follows that $l \leq O(\log s|v|)$, whence the cumulative cost at depth l is $O(\log s|v| \cdot e^{-c|v|s})$. Putting everything together, we have proven that

$$\mathrm{TV}(p, p') \leq O(\log(|v|s) \cdot e^{-c|v|s}) \leq o\left(\frac{1}{s|v|}\right)^{1+\epsilon} \quad \text{for some } \epsilon > 0. \qquad \square$$

Proof of Proposition 1.11. The claim is trivially true (by choosing an appropriate constant) if v is a terminal node. Thus, fix a non-terminal node v such that $x_1 \notin v$ and let

$$\mathbf{X} = \mathbf{X}_v = \{X_i : X_i \in v\}$$

denote the set of points landing in v, so that $k := |\mathbf{X}| \in \{1, \ldots, n-1\}$.

Recall that $f = f(v)$ and $f = f'(v)$ are the respective expectations of the tree estimator at x when the sequence of splits is such that v is the current subset of $\mathcal{X}$ containing x, with f' being calculated conditional on $X_1 = x_1$. It follows that f and f' are functions of the distribution of its "input vector" $\mathbf{X}$. In a slight abuse of notation, let Π and Π' denote sequence of splits distributed according to the probabilities p and p'. We will show that, for each $k \in \{1, \ldots, n-1\}$, the total variation distance of

$$(\mathbf{X} \mid |\mathbf{X}| = k, \Pi = v) \quad \text{and} \quad (\mathbf{X} \mid |\mathbf{X}| = k, \Pi' = v, X_1 = x_1) \tag{1.30}$$

is bounded by $(\log s) \cdot \Delta(s|v|)$. This will suffice to bound $|f - f'|$ by the variational definition of total variation distance

$$\mathrm{TV}(p, p') = \sup_{|g| \leq 1} |\mathbf{E}_{A \sim p}\, g(A) - \mathbf{E}_{A \sim p'}\, g(A)|.$$

Since $x_1 \notin v$, X_1 is not an element of $\mathbf{X}$, so that $\mathbf{X}$ and X_1 are independent. Since the split Π' is distributed according to splitting probabilities when $X_1 = x_1$, we have,

$$\mathrm{dist}(\mathbf{X} \mid |\mathbf{X}| = k, \Pi' = v, X_1 = x_1) = \mathrm{dist}(\mathbf{X} \mid |\mathbf{X}| = k, \Pi' = v). \tag{1.31}$$

The depth of v is at most $\log_2 s$, so that $\mathbf{P}(\Pi = \Pi') \geq 1 - (\log_2 s)\Delta(s|v|)$ by applying the splitting stability assumption $\log_2 s$ many times using a union bound. By (1.31) the total variation distance of the distributions in (1.30) differs by $\mathbf{P}(\Pi \neq \Pi')$, and the result follows. □

Poor of Proposition 1.12. The idea is to recursively expand the formulas $\mathbf{E}\,T$ and $\mathbf{E}(T \mid X_1)$ in terms of the directed acyclic graph. We start with

$$
\begin{aligned}
|\mathbf{E}\,T - \mathbf{E}(T \mid X_1 = x_1)| &= \left|\sum_v p(e)f(e) - \sum_v p'(e)f'(e)\right| \\
&\leq \sum_v |p(e) - p'(e)|f'(e) + \left|\sum_v p'(e)(f(e) - f'(e))\right|,
\end{aligned}
$$

where the sum runs over nodes v after the first split, i.e., $[0,1]^p \to v$. The second summand may be split into two terms, one over $x_1 \in v$ and the other $x_1 \notin v$. Due to splitting stability, Proposition 1.11 allows us to bound the second term, so that

$$
\begin{aligned}
|\mathbf{E}\,T - \mathbf{E}(T \mid X_1 = x_1)| &\leq \sum_v |p(e) - p'(e)|f'(e) + \left|\sum_v p'(e)(f(e) - f'(e))\right| \\
&\leq \Delta(\alpha s) + (\log s)\Delta(\alpha s) + \sum_{x_1 \in v} p'(e)|f(e) - f'(e)|,
\end{aligned}
$$

where we used the fact that $|v| \geq \alpha$ for each v in the summand, and Δ is the function $\Delta(z) = z^{-(1+\epsilon)}$. Now, each of the terms $|f(e) - f'(e)|$ may be bounded by $\Delta(\alpha^2 s) + \log(s)\Delta(\alpha^2 s) + \sum_{x_1 \in w}(\cdots)$. Continuing in this way, we have

$$
|\mathbf{E}\,T - \mathbf{E}(T \mid X_1 = x_1)| \leq \log(s)(\Delta(s) + p_1\Delta(\alpha s) + p_2\Delta(\alpha^2 s) + \ldots) = \log s \sum_{\ell=0}^{\infty} p_\ell \Delta(\alpha^\ell s),
$$

where p_ℓ is the probability that x and x_ℓ belong to the same node after ℓ splits. In other words, if we let L be the number of splits after which x and x_1 are separated, then

$$
p_\ell = \mathbf{P}(L \geq \ell).
$$

Since $x \neq x_1$, we may assume without loss of generality that $\|x - x_1\| > \beta$ for some fixed β (c.f. the proof of Proposition 1.9). In particular,

$$
p_\ell \leq (1 - \delta + o(1))^\ell
$$

for sufficiently large ℓ. Moreover, $\Delta(\alpha^\ell s) = \frac{1}{s^{1+\epsilon}}(\frac{1}{\alpha^{1+\epsilon}})^\ell$, whence $\delta > 1 - \alpha^{1+\delta}$ is enough to ensure that infinite series is less than $\frac{1}{s^{1+\epsilon}}$. In particular, as $s \to 0$, we may take $\epsilon \to 0$, so that the restriction is satisfied (after suitable constants) by $\delta > 1 - \alpha$. This completes the proof. □

Chapter 2

Dynamic R&D Contracting

Joint with Ali Kakhbod

2.1 Introduction

Research and development (R&D) play an increasingly large role in the allocation of capital. In contrast to traditional kinds of output studied in principal-agent models, the success of a R&D venture is observed—if at all—only when the venture succeeds. Due to the binary nature of success in R&D (i.e., the "lumpy nature" of the project), the researcher has two sources of compensation: flow payments during the R&D phase, and a lump-sum reward (bonus) upon successful completion of the project.

We embed the R&D process in a continuous-time principal-agent model and study the shape of the flow payment and lump-sum reward in the optimal incentive structure. To make progress on the research project, the agent chooses a level of effort at each moment in time. Agent's effort is costly, modeled by a convex function. The principal rewards the effort with a *two-dimensional incentive-pay contract*: a flow compensation over the course of R&D plus a lump-sum reward (a "bonus") if the R&D is successful. Importantly, we allow the agent's utilities from the immediate flow compensation and the lump-sum reward to modeled by potentially different functions.

We show that when the agent's effort is observable to the principal, the optimal contract—suggested actions, the flow compensations, and the lump-sum bonus— are *constant*, as we might intuitively expect. However, when the principal does not observe the effort and instead observes a signal in the form of a diffusion whose drift is proportional to the effort,

We are grateful to Daron Acemoglu, Ricardo Caballero, Drew Fudenberg, Daniel Garrett, Andrey Malenko, Yuily Sannikov for their help, comments and great suggestions. We also thank seminar participants at Massachusetts Institute of Technology (MIT), Harvard University, and Princeton University for helpful comments. This work was supported by the Hand Foundation Fellowship.

the optimal contract will depend on the agent's continuation value and is therefore non-constant.

We characterize the entire space of incentive compatible contracts; using this result, we explicitly characterize the optimal contract by a solution of an ordinary differential equation (ODE). Our main result shows that the solution to the ODE (i.e., the optimal contract) exists and it is unique.

The duration of the employment is specified by an endogenous threshold $v_{\max} \in \mathbf{R}_+$ such that the principal retires the agent when the agent's continuation exceeds $v_{\max}$. One interesting property of this threshold is that the principal's continuation value may in fact be negative in a neighborhood of $v_{\max}$. The interpretation is that the principal commits to be sufficiently patient such that he continues with the project even while carrying a negative continuation value.

We also show that during employment, the agent always exerts a positive effort whose magnitude evolves over time. There is, however, a unique threshold such that when the agent's continuation value falls below than the threshold, the principal reduces the flow compensation to zero and only provides incentive by a positive lump-sum rewards upon successful completion of the project. Above the threshold, the agent's flow compensation is positive and increasing with the continuation value.

In numerical simulations, we find that the optimal contract features a minuscule level of flow payments, where most of the agent's benefit come from the lump-sum reward when the project is successful. This theoretical feature of our model agrees with empirical evidence that long-term CEO compensation is tied to the success of R&D processes (e.g., Lerner and Wulf (2007)).

Finally, we empirically link our model's theoretical implications with executive compensation data from ExecuComp. We first show that executive compensation is two (or multi) dimensional, consisting of heavy tails in both the salary (flow compensation) and bonuses (lump-sum bonuses), in line with the basic premise of our model. Regression executive bonus pay on company stock growth—a proxy for the successful completion of projects—reveals a statistically significant coefficient, which is consistent with the theoretical optimal contract.

2.1.1 Related Literature

This paper belongs to a fast-growing literature on continuous-time dynamic contract theory. Important works such as Radner (1985), Spear and Srivastava (1987), Fudenberg, Holmstrom and P. (1990), Abreu, Pearce and Stacchetti (1990) and Phelan and Townsend (1991) provide foundations for the analysis of repeated principal agent interactions.[1] Our paper runs parallel

[1] Albuquerque and Hopenhayn (2004), Albuquerque and Hopenhayn (2006) and DeMarzo and Fishman (2007*b*,*a*) apply these theories to dynamic financing. Similar to these papers, we use the continuation utility of the agent as a state variable in characterizing the optimal contract.

to another branch of the dynamic mechanism design literature in which the optimal sequence of decisions depends on the evolution of the agent's hidden type. So-called "local incentive compatibility conditions" for this brand of models were developed by Pavan, Segal and Toikka (2014), while Garrett, Pavan and Toikka (2018) focuses on developing properties of the optimal solution that bypasses direct examination of local IC conditions. In contrast, our model considers a situation in which the both the agent and the principal face uncertainly over the underlying state evolution. Note that there is no hidden type in our model: utilities and costs are common knowledge, as is the status of project completion.

In these settings, the optimal contract can be presented in a recursive form because the agent's effort affects the probability distribution of the signal currently observed by the principal. As a consequence, the agent's continuation value completely summarizes the incentives provided to him by the contract.

Using the recursive structure, Sannikov (2008) in his seminal paper provides a continuous time model of repeated agency, in which it is possible to explicitly characterize the optimal contract using an ordinary differential equation. This approach is applied to several frameworks. For example, security design (DeMarzo and Sannikov (2007), Biais et al. (2007), Piskorski and Tchistyi (2010), Sannikov (2012)), learning (DeMarzo and Sannikov (2017), He et al. (2017)), dynamic compensation (He (2009, 2011, 2012)), risk taking (DeMarzo, Livdan and Tchistyi (2014), Biais et al. (2010)), q-theory and investment (DeMarzo et al. (2012)), dynamic capital budgeting with communication (Malenko (2018)). In this paper we focus on optimal dynamic contracting for R&D projects with *two-dimensional incentive-pay*: a flow of compensations and a lump-sum reward upon successful completion of the project.

A few researchers have investigated the topic of R&D contracting. Manso (2011) studied a two-period model in which a principal provides incentives for an agent not only to work rather than shirk but also to work on exploration of an uncertain technology rather than exploitation of a known technology. Hörner and Samuelson (2013) and Bergemann and Hege (2005) studied contracting problems with dynamic moral hazard and private learning about the quality of the innovation project. Halac, Kartik and Liu (2016) introduced adverse selection about the agent's ability into the problem. In contrast to these important works, this paper is not concerned with experimentation and our focus differs from these in many ways, particularly, general concave payoffs, convex cost of exerting effort, random termination of the job that depends on the agent's effort and the optimal design of bonus for successfully finishing the job.

The rest of the paper is structured as follows. Section 2 sets up the model; section 3 presents our main results, beginning with a characterization of the incentive compatibility constraint in terms of the agent's valuation; section 4 discusses our numerical experiment; section 5 presents our empirical findings; and section 6 concludes.

2.2 The Model

Our model considers a principal who contracts with an outside firm or researcher ("the agent") to finish a research project. The nature of the research project is that such that success is binary: the project is either incomplete or finished. Moreover, contracting ends when (if ever) the project is successful. Thus, our model is most applicable for analyzing "one-off" long-term research or development ventures that have no intermediate quantifiable output and whose success is uncertain. One example would be the successful negotiation of a merger and acquisition—the principal is the board and the agent is the CEO; the deal either goes through or not, and intermediate negotiations bring no benefit to either party.

We work in a continuous time setting, and project success depends on the effort $\{a_t \geq 0 : t \geq 0\}$ exerted by the agent. Each fixed schedule of effort induces an inhomogenous Poisson process with intensity a_t, and the project is successful at the first arrival time. More prosaically, the project succeeds in the time interval $[t, t + dt)$ with independent probability $a_t dt$, and the time interval $dt \to 0$.

While working, the agent incurs a (flow) cost of action of $g(a)dt$ and is compensated by an amount $u(c)dt$. If he is successful, he receives an additional lump sum R, valued according to a utility $\Lambda(R)$. On the other hand, the principal is risk neutral and his only source of benefit is a benefit $\Delta \geq 0$ when (if ever) the project is successful. Everybody discounts at a common rate ρ.

In general, the two-dimensional—flow payments and a lump sum bonus payment—is an importance piece of our model: the optimal contract will identify how best to incentivize research oriented projects. A contract featuring relatively large flow payments corresponds to paying researchers a high salary but low bonus; conversely, a contract with large lump sum reward corresponds to a bonus-heavy compensation structure.

We will assume that the function g is strictly convex and differentiable with boundary conditions

$$g(0) = 0 \quad \text{and} \quad \lim_{a \to \infty} g'(a) = \infty. \tag{2.1}$$

The utility functions u and Λ are allowed to be different; we will assume both functions are strictly increasing and strictly concave, with $u(0) = \Lambda(0) = 0$ and $\lim u'(c) \to \infty$. The main assumption[2] relating u and Λ is that

$$\rho\Lambda(R) \leq u(\rho R). \tag{2.2}$$

The point of this assumption is the following: in order to retire the agent with a constant utility $\Lambda(R)$, the principal could either fire the agent with a lump-sum payment R, or retire the project and pay a flow utility $u^{-1}(\rho\Lambda(R))$ forever after. By assumption, $u^{-1}(\rho\Lambda(R)) \leq \rho R$, so that the principal weakly prefers the latter.

[2] See the Appendix for other standard assumptions on u and Λ.

This assumption implicitly carries with it a couple of modeling assumptions. First, note that to give the agent a utility of $\Lambda(R)$ (when the project is successful, say), the principal could either

- Pay a lump sum R to the agent; or
- pay him a flow utility of $u^{-1}(\rho\Lambda(R))$ forever after.

The former costs the principal R while the latter costs $\frac{u^{-1}(\rho\Lambda(R))}{\rho}$. The assumption above implies that it is (weakly) preferred by the principal to use a constant flow utility instead of a lump-sum bonus. Therefore, an additional assumption of the model is that the contract must terminate (i.e., no further payments) after the project is successful. We contend that imposing this assumption makes our model more closely aligned with real-world R&D contracts.

Another consequence of $\Lambda(R) \leq u(\rho R)/\rho$ is that the agent values her lump-sum bonus R weakly less than the constant consumption stream financed by R at interest rate ρ (i.e., with R, the agent could enjoy a flow consumption of ρR in perpetuity). That is, this is imposes a restriction on the agent's preferences over consumption streams. There are a couple interpretations: the first is that the agent does not have access to a sufficiently long-lived risk-free money market in which she could invest. The second is that the agent is sufficiently risk-averse over fluctuations in long-term interest rates such that she (weakly) prefers a constant rate coupon with infinite maturity over what she could obtain with a lump-sum.

While the principal knows whether and when the project is successful, she cannot directly monitor the agent, i.e., the action a_t is not directly observed. Instead, the principal observes a noisy signal y_t whose law is governed by

$$dy_t = a_t dt + \sigma dB_t \tag{2.3}$$

where σ is a known constant and B_t is the standard Brownian motion.

The principal offers a contract (a, c, R)—the action schedule a, the flow compensation schedule c, and lump-sum bonus schedule R. Each of these three quantities are allowed to be functions of past observed history; the history includes whether the project is successful (a binary event) and on the noisy signal of the action taken

$$dy_t = a'_t dt + \sigma dB_t, \quad \text{where } a'_t \text{ is the effort exerted}$$

and B_t is a standard Brownian motion, σ a known constant.

2.2.1 Stopping Time Process

The stopping time of success, denoted τ, plays a central role in our analysis. Here, we explicitly write down its distribution. For a given schedule of effort $\{a_t\}_{t=0}^{\infty}$, we have

$$\mathbf{P}(\tau = \infty) = \exp\left(-\int_0^{\infty} a_z dz\right). \tag{2.4}$$

Therefore, the CDF of the stopping time conditional on success is

$$\mathbf{P}(\tau < t \mid \tau < \infty) = \frac{1 - \exp(-\int_0^t a_z dz)}{1 - \exp(-\int_0^{\infty} a_z dz)}, \tag{2.5}$$

with corresponding PDF

$$f_a(t) := \frac{d}{dt}\mathbf{P}(\tau < t \mid \tau < \infty) = \frac{a_t \exp(-\int_0^t a_z dz)}{1 - \exp(-\int_0^{\infty} a_z dz)}. \tag{2.6}$$

2.3 Contracting with the Agent

The contract is a tuple (a, c, R) of suggested actions, compensations, and rewards. The feasible sets of the action a, the compensation c, and the lump sum reward R is bounded.

For a fixed contract (a, c, R), the agent's value at time zero is given by

$$v_0 = \rho\,\mathbf{E}\Bigg[\mathbf{P}(\tau < \infty)\int_0^{\infty}\bigg[\int_0^{\tau} e^{-\rho t}(u(c_t) - g(a_t))dt + e^{-\rho\tau}\Lambda(R_\tau)\bigg]f_a(\tau)d\tau \\ + \mathbf{P}(\tau = \infty)\int_0^{\infty} e^{-\rho t}(u(c_t) - g(a_t))dt\Bigg]. \tag{2.7}$$

The integral of $e^{-\rho t}(u(c_t) - g(a_t))$ against the stopping time of success τ (with density $f_a(\tau)$) captures the flow utility while working, while the term $e^{-\rho t}\Lambda(R_\tau)$ is the lump sum reward.[3]

Similarly, the principal's value function at time zero is

$$\Gamma_0^p = \rho\,\mathbf{E}\Bigg[\mathbf{P}(\tau < \infty)\int_0^{\infty}\bigg[\int_0^{\tau} e^{-\rho t}(-c_t)dt + e^{-\rho\tau}(\Delta - R_\tau)\bigg]f_a(\tau)d\tau \\ + \mathbf{P}(\tau = \infty)\int_0^{\infty} e^{-\rho t}(-c_t)dt\Bigg]. \tag{2.8}$$

Here, the principal pays c_t while the agent is working and receives Δ less the lump sum reward when the project succeeds.

We look for the optimal incentive compatible contract with commitment.[4]

[3] We assume the principal and the agent both have the same discount factor ρ. See Farhi and Werning (2006) and DeMarzo and Sannikov (2007) for examples where they have different discount rates.

[4] We note that there is no private saving in our model. For examples with private saving see important works by Werning (2002), Williams (2009) and Di Tella and Sannikov (2016).

2.3.1 Structure of the Optimal Contract

Before proceeding, we give a brief guide for our analytical results and provide some intuition. To start, note that the signal available the principal,

$$dy_t = a_t dt + \sigma dB_t$$

depends only the present time effort a_t and not its past history. Coupled with the fact that the probability of success is independent from one moment to the next, this implies that our model is stationary. Therefore, one way the principal could design the contract is to base the flow compensation c_t on the noisy signal dy_t, and as an additional incentive, provide a lump sum bonus R—constant over time—if the project succeeds. In this case, the contract takes the form

$$c_t = c(dy_t), \qquad R_t = R$$

for some scalar $R \in \mathbf{R}$ and compensation schedule $c : \mathbf{R} \to \mathbf{R}$ (in this section, we will speak loosely and imagine that dt is small but finite, so that dy_t is normally distributed with mean $a_t \cdot dt$.) Since a_t is specified in the contract[5] and thus known (i.e., the revelation principle), the compensation schedule c may be taken to be a function of the noise only, so that without loss of generality,

$$c_t = f\left(\frac{dy_t - a_t dt}{\sigma}\right) = f(dB_t),$$

where $f(z) = c(adt + \sigma z)$.

Fixing a and R, the payment schedule c is constrained by the incentive compatibility condition. By exerting effort $a'_t = a + \sigma\epsilon$, the agent receives a flow compensation equal (in distribution) to $f(dB_t + \epsilon)$, earns the lump-sum bonus R_t (less his continuation value) with additional probability $\sigma\epsilon dt$, while increasing his disutility of effort by approximately $\sigma\epsilon g'(a)dt$. In particular, taking $\epsilon \to 0$, we find that

$$\mathbf{E}\, f'(dB_t) + \sigma(\Lambda'(R) - v_t/\rho)dt = \sigma g'(a)dt,$$

where v_t/ρ is the (scaled) continuation value of the contract. Using $\mathbf{E}\, f'(dB_t) = \sigma\, \mathbf{E}\, c'(dy_t)$, we may cancel σ and recover the following necessary condition that c must satisfy for incentive compatibility

$$\mathbf{E}\, c'(dy_t) = g'(a)dt - (\Lambda'(R) - \frac{v_t}{\rho})dt.$$

The quantity on the LHS involving the derivative of the compensation schedule c is a measure of the sensitivity of the contract to observation noise. A benefit of working in continuous time is that the sensitivity also characterizes the contract insofar as determining

[5] In the present case, clearly a_t should also be constant.

the continuation value of the agent. Our first set of propositions show the previous derivation is necessary and sufficient: any contract (a, c, R) induces a continuation value v_t of the agent that is a diffusion (Lemma 2.1 and Proposition 2.2), and IC holds whenever the diffusion coefficient satisfies an equality involving a and R (Theorem 2.3).

One implication from the above discussion is that when the principal pays the agent according to the stationary contract $c_t = f(dy_t)$ and $R_t = R$, he pays a sensitivity cost to ensure the IC constraint. Indeed, when R decreases, the IC constraint forces an increase in $\mathbf{E}\, c'(dy_t)$; using Stein's Identity,

$$\mathbf{E}\, c'(dy_t) \sim \mathbf{E}[\sigma dB_t \cdot c(dy_t)] \sim \mathbf{Cov}(dB_t, c_t).$$

where $\sim$ means up to constant. This relationship has an intuitive interpretation: the lower the lump-sum bonus, the higher the correlation of compensation with noise. In this way, the principal incurs a "sensitivity cost" if he implements a stationary contract where the agent is incentivized in each period.

Instead, it may be more efficient to aggregate signals and adjust the flow compensation and lump sum bonus based on the entire history y_t instead of its innovation dy_t; for example, the statistic $y_t/t \sim N(a, \sigma^2/t)$ has lower variance for detecting deviations than $dy_t/dt \sim N(a, \sigma^2/dt)$. This suggests that the optimal contract should depend on non-trivially on the past history and is therefore non-stationary. However, in the degenerate cases $\sigma = 0$ and $\sigma = \infty$ where that y_t carries no useful information, the optimal contract is indeed stationary.

In general, the optimal contract is difficult to describe directly as the history $\{y_s : s < t\}$ expands with time. Our first main result (Theorem 2.3) shows that the evolution of the continuation value v_t suffices to characterize the IC condition, while our second main result (Theorem 2.5) establishes that v_t is a valid state variable in that the optimal contract a, c, and R can be written as functions of v_t.

Obviously, the continuation value needs to be non-negative for incentive compatibility: the agent guarantees zero by simply not working. Intuitively, the continuation value cannot be too high either: a low continuation value means the the principal could drive project completion more cheaply as per the IC condition above. The flow compensation and bonus payment c and R act in tandem to constrain the continuation value of the agent: Proposition 2.8 characterizes regimes (stated in terms of the continuation value of the agent) that determine which of c and R is the main driving force of the contract.

Having described the intuitive ideas behind our proof, we now describe our solution to the optimal contracting problem.

2.3.2 Agent's Valuation as a Diffusion

We first rewrite, using integration by parts, the valuations (2.7) and (2.8) as to remove the explicit separation between the events $\{\tau < \infty\}$ and $\{\tau = \infty\}$ and to "smooth out" the timing of the lump sum reward. After, we will realize the agent's valuation as a diffusion.

Lemma 2.1. *For a fixed contract* (a, c, R)*, the principal and the agent's continuations values, conditional on the information available at time* t *are given by, respectively,*

$$\Gamma_t^p(a, c, R) = \rho \mathbf{E}_t^a \bigg[\int_t^\infty e^{-\rho(s-t)} \big\{ e^{-\int_t^s a_z dz}(-c_s)ds + e^{-\int_t^s a_z dz} a_s(\Delta - R_s) \big\} ds \bigg] \tag{2.9}$$

and

$$v_t(a, c, R) = \rho \mathbf{E}_t^a \bigg[\int_t^\infty e^{-\rho(s-t)} \big\{ e^{-\int_t^s a_z dz}(u(c_s) - g(a_s) + a_s \Lambda(R_s)) \big\} ds \bigg]. \tag{2.10}$$

Here, $\mathbf{E}_t^a$ *is the expectation induced by the agent's action* a *conditional on the information available at time* t*.*

(All proofs appear in the appendix.)

The lemma is easiest to interpret if we set $t = 0$. After integrating by parts, the (unconditional) density $\mathbf{P}(\tau < \infty) f_a(\tau)$ and the probability $\mathbf{P}(\tau = \infty)$ combine in such a way as make $e^{-\int_0^s a_z dz}$ the correct time density. More precisely, τ disappears from the expressions above, being carried implicitly by $e^{\int_0^s a_z dz}$. Relatedly, the flow quantity c_s and the lump sum quantities Δ and R_s are now on equal footing, after the latter are multiplied by a_s.

Lemma 2.2. *lemma For any given contract* (a, c, R) *there exists a progressively measurable process* φ_t *with* $\mathbf{E}^a[\int_0^t \varphi_s^2 ds] < \infty$ *such that the agent's continuation value may be written as*

$$\begin{aligned} \rho^{-1} dv_t(a, c, R) = &\bigg[v_t(a, c, R)\Big(1 + \frac{a_t}{\rho}\Big) + g(a_t) - u(c_t) - a_t \Lambda(R_t) \bigg] dt \\ &+ \varphi_t(dy_t - a_t dt). \end{aligned} \tag{2.11}$$

The intuition behind the proof is the following. Let V_t denote the expected value of the agent conditional on information available at time t, taking the contract (a, c, R) as given. Then V_t is a $\mathbf{P}^a$-martingale, and the martingale representation theorem furnishes the process φ_t. Recall that the process $dy_t - a_t dt$ follows a (scaled) Brownian motion, from (2.3) that

$$\frac{1}{\sigma}(dy_t - a_t dt) = dB_t =: dB_t^a. \tag{2.12}$$

In this way, φ_t measures the sensitivity of the agent to the observational noise.

We will see next that the diffusion v_t is actually the correct state variable for the contract, as opposed to the signals y_t. In particular, the signal y_t only matters in so much as the role of φ_t in the diffusion.

2.3.3 Constant Contract as the First Best

Before moving to examine the incentive constraints for the agent and the associated problem for the principal, we first consider the first-best solution. In particular, let us consider maximizing the sum of the principal and the agent's value functions, as defined in Lemma 2.1. Denoting the sum by W_0, and fixing an arbitrary contract (a, c, R), we have

$$W_0 := \Gamma_0^p + v_0 = \rho\, \mathbf{E}_t^a \left[\int_0^\infty e^{-\rho t} e^{-\int_0^t a_z dz} \Big(u(c_t) - c_t + a_t \big(\Delta + \Lambda(R_t) - R_t \big) - g(a_t) \Big) dt \right] \quad (2.13)$$

where we focus on $t = 0$ without loss. Since the effort a_t is non-negative, it is clear from maximizing W_0 pointwise that the optimal contract features a constant compensation and lump-sum reward. Specifically, denoting the optimal first-best contract by (a_t^*, c_t^*, R_t^*), we have

$$c_t^* = c^* \quad \text{and} \quad R_t^* = R^* \qquad \text{for some constants } c^*, R^* \in \mathbf{R} \quad (2.14)$$

where c^* maximizes $u(c) - c$ and R^* maximizes $\Lambda(R) - R$ over feasible compensations c and rewards R, respectively.

The only remaining variable to optimize is the effort a_t. Substituting c^* and R^* into (2.13), the optimal choice of a_t is the solution to the problem

$$\tilde{W}_0 := \max_{\{a_t\}_{t \geq 0} \in A} \int_0^\infty e^{-\rho t} e^{-\int_0^t a_z dz} H(a_t) dt \quad (2.15)$$

where $H(a) \equiv u(c^*) - c^* + a(\Delta + \Lambda(R^*) - R^*) - g(a)$.

Since the choice set A is compact and H is continuous it follows that an optimal solution exists. Moreover, the discounting factor $e^{-\int_0^t a_z dz}$ implies that the optimal a_t is decreasing. Together with the stationary feature[6] of (2.15), we conclude the optimal a is in fact constant, and is the solution to

$$a_t^* = a^* \in \mathbf{R}, \qquad \text{where } a^* \text{ maximizes } \frac{H(a)}{\rho + a}. \quad (2.16)$$

Therefore, we have shown that the optimal first best contract is constant.

[6] It is clear that if $t \mapsto a_t$ is optimal, then for any $t_0 > 0$, the function defined by

$$t \mapsto \begin{cases} a_t & \text{if } t \leq t_0 \\ a_{t-t_0} & \text{if } t > t_0 \end{cases}$$

is also optimal.

2.3.4 Incentive Compatibility Condition

Theorem 2.3. *Let φ_t be the process defined in Lemma 2.1. The prescribed action a_t for the agent is optimal if and only if for all feasible actions a,*

$$a_t \in \arg\max_{a'} a'\left(\varphi_t + \Lambda(R_t) - \frac{v_t(a, c, R)}{\rho}\right) - g(a') =: Z(a') \tag{2.17}$$

for all $0 \leq t < \infty$. Here, the max is taken over all feasible actions a'.

Moreover, when $g(\cdot)$ is differentiable, the prescribed action a_t is optimal if and only if the sensitivity of his continuation value of the observation noise—that is, φ_t—equals the marginal cost of his effort minus the net gain from success, i.e.,

$$\varphi_t = g'(a_t) - \left(\Lambda(R_t) - \frac{v_t}{\rho}\right). \tag{2.18}$$

The only if part of the claim follows from considering the alternative strategy where where the action path $\{a'_z\}$ is used until t, and using $\{a_z\}$ afterwards. The difference $Z(a_t) - Z(a'_t)$ captures the tradeoff between the additional gain of working today and the decrease in the continuation value implied by (2.11). If g is differentiable, then the *"Moreover..."* part of the claim follows immediately by differentiating.

We stipulate that g is indeed differentiable so that (2.18) applies. If we substitute into the diffusion equation (2.11), we see that for incentive compatible contracts, the evolution of v_t depends only on the terms of the contract (and known constants). More precisely, incentive compatibility is satisfied if and only if

$$\rho^{-1} dv_t = \left[v_t\left(1 + \frac{a_t}{\rho}\right) + g(a_t) - u(c_t) - a_t\Lambda(R_t)\right]dt + \sigma\left[g'(a_t) - \frac{v_t}{\rho} - \Lambda(R_t)\right]dB_t^a \tag{2.19}$$

where B_t^a is the Wiener process induced by the action a (refer to the proof for more details).

The previous expression is straightforward to interpret in terms of the original contract (a, c, R). Clearly, every contract (a, c, R) introduces a (discounted) continuation value v_t that that evolves as histories $\{dy_t\}$ are realized over time.

At each instant t, the agent gains $[u(c_t) - g(a_t)]dt$, and succeeds with instantaneous probability $a_t dt$, enjoying payoff $\Lambda(R)$ less the continuation payoff v_t/ρ. Therefore, an amount $u(c_t) - g(a_t) + a_t\Lambda(R) - v_t/\rho$ is subtracted from the continuation value, and exponential discounting explains the rest. The form of the diffusion term is familiar, c.f., discussion at the beginning of this section.

The upshot is that this decouples the principal's optimization problem. In searching for an optimal incentive compatible contract, the correct state variable for a, c, and R is actually v_t instead of y_t.

2.3.5 The Principal's Problem

Having examined the agent's side of the problem, we turn now to the principal. Recall that the principal seeks to maximize

$$F_1 = \max_{a,c,R} \rho\,\mathbf{E}\left[\int_0^\infty e^{-\rho t}\left(e^{-\int_0^t a_s ds}(-c_t)dt + e^{-\int_0^t a_s ds}a_t(\Delta - R)dt\right)\right] \tag{2.20}$$

such that (a, c, R) is incentive compatible.

As before, let v_t denote the expected continuation payoff of the agent under the contract. To solve the principal's problem, we replace the above incentive compatibility constraint with the results of Lemma 2.2 and Theorem 2.3.

Corollary. *Let F_2 be the maximum value of the quantity*

$$\rho\,\mathbf{E}\left[\int_0^\infty e^{-\rho t}\left(e^{-\int_0^t a(v_z)d_z}\big(-c(v_t)\big)dt + e^{-\int_0^t a(v_z)dz}a(v_t)\big(\Delta - R(v_t)\big)dt\right)\right] \tag{2.21}$$

over functions a, c, and R of the agent's valuation $\{v_t\}$, subject to (2.19). *Then $F_1 = F_2$.*

2.3.6 Retirement as an Upper Bound

The first step in the analysis of the principal's problem is to derive an intuitive upper bound from the consideration of the retirement policy.

Suppose the agent's continuation value is v. To retire the agent, the principal could either pay him a reward R so that $\rho\Lambda(R) = v$ or compensate him with c such that $u(c) = v$. Let $\lambda_r(\cdot)$ denote the principal's payoff after retiring the agent. Since the agent exerts no effort after retirement, we have

$$\lambda_r(v) = \max\left\{-u^{-1}(v), -\rho\Lambda^{-1}\left(\frac{v}{\rho}\right)\right\}, \quad v \geq 0. \tag{2.22}$$

Since $\rho\Lambda(R) = v = u(c) \leq u(\rho R)$ by assumption (A1), $c \leq \rho R$. Therefore,

$$\lambda_r(v) = -u^{-1}(v). \tag{2.23}$$

We could now upper-bound the principal's payoff.

Proposition 2.4. *Let $\bar{c}$ denote the maximum feasible compensation, and denote $\bar{v} = u(\bar{c})$. Then under any contract (a, c, R) with $v_0 \geq \bar{v}$, the principal's payoff is at most $\lambda_r(v_0)$. That is,*

$$m(v_0) = \rho\,\mathbf{E}\left[\int_0^\infty e^{-\rho t}\left(e^{-\int_0^t a_z dz}(-c_t)dt + e^{-\int_0^t a_z dz}a_t(\Delta - R_t)dt\right)\right] \leq \lambda_r(v_0). \tag{2.24}$$

Therefore, whenever the agent's continuation value exceeds the maximum possible flow utility, the principal retires the agent to obtain $\lambda_r(\bar{v})$.

Remark. The desired inequality follows from relating the agent's lump sum reward $\Lambda(R_t)$ with $u'(u^{-1}(v_0))$ using (A1), then bounding the relevant quantity by Δ using (A3).

2.3.7 The HJB Equation and associated Boundary Conditions

The main goal of this section is to show that the principal's value function m is defined by the following HJB equation

$$\begin{aligned} m(v) = \max_{a\geq 0, c\geq 0, R\geq 0} \Bigg\{ & -c + a\left(\Delta - R - \frac{m(v)}{\rho}\right) \\ & + m'(v)\left[v - u(c) + g(a) - a\left(\Lambda(R) - \frac{v}{\rho}\right)\right] \\ & + \frac{\rho\sigma^2}{2} m''(v)\left(g'(a) - \Lambda(R) + \frac{v}{\rho}\right)^2 \Bigg\}. \end{aligned} \tag{2.25}$$

Moreover, the maximizers a, c, and R to this equation constitute the optimal contract.

According to Proposition 2.4, we only need to consider v in the interval $v \in [0, \bar{v}]$. Actually, more is true: it turns out that there is some $v_{\max} < \bar{v}$ such that the "correct" boundary equations are

$$m(0) = 0 \quad \text{and} \quad m(v_{\max}) = \lambda_r(v_{\max}), \tag{2.26}$$

together with the smooth pasting condition

$$m'(v_{\max}) = \lambda_r'(v_{\max}). \tag{2.27}$$

Intuitively, $v_{\max}$ is the point above which the agent is retired, under the optimal contract. As suggested by the smooth pasting condition, for $v > v_{\max}$, the value function m decreases faster than the retirement value λ_r; since $m(v_{\max}) = \lambda_r(v_{\max})$ from (2.26), this implies $m(v) < \lambda_r(v)$, making retirement optimal at v. More prosaically, past $v_{\max}$, it becomes too expensive for the principal to continue the contract.

2.3.8 Analysis of the HJB Equation

In this section, we state the main result of the paper; namely, the next three theorems together establish that the HJB equation completely characterizes the solution of the problem.

The first step is to show that the HJB equation (2.25) has a unique solution. It is more convenient to work with another form of the HJB, as stated below.

Theorem 2.5. *Consider a transformed version of the HJB*

$$m''(v) = \min_{a,c,R} \Phi_{a,c,R}\Big(v, m(v), m'(v)\Big) \tag{2.28}$$

where

$$\Phi_{a,c,R}\Big(v,m,m'\Big) = \frac{m + c - a(\Delta - \frac{m}{\rho} - R) - m'(v - u(c) + g(a) - a(\Lambda(R) - \frac{v}{\rho}))}{\frac{1}{2}\rho\sigma^2(g'(a) + v/\rho - \Lambda(R))^2}. \quad (2.29)$$

Here, the minimum is taken over the feasible set.

Then there exists a unique function $m(\cdot)$ *solving* (2.28) *with the property that for some* $v_{\max} \in (0, \bar{v})$*, the following holds*

(i) *For every* $v \in [0, v_{\max}]$*, we have the lower bound* $\lambda_r(v) \leq m(v)$*.*

(ii) *The boundary conditions* $m(0) = 0$*,* $m(v_{\max}) = \lambda_r(v_{\max})$*, and* $m'(v_{\max}) = \lambda_r'(v_{\max})$*.*

The main idea of the theorem and its proof rests on showing that the initial conditions are consistent in that there is some $v_{\max} < \bar{v}$ such that, given the boundary conditions, the solution to m is unique. Along the way we will actually establish that m is strictly concave on its domain $[0, v_{\max}]$.

Having demonstrated a unique solution to the HJB equation, the next theorem shows that the maximizers a, c, and R induce the optimal contract.

Theorem 2.6. *Let* $m(\cdot)$ *be the unique solution to* (2.28) *satisfying the boundary conditions for some* $v_{\max} \in [0, \bar{v}]$*. Let* $a = a(v)$*,* $c = c(v)$*, and* $R = R(v)$ *be the corresponding minimizers. For any* $v_0 \in [0, v_{\max}]$*, the diffusion*

$$\rho^{-1} dv_t = \left[v_t \left(1 + \frac{a(v_t)}{\rho} \right) + g\big(a(v_t)\big) - u\big(c(v_t)\big) - a(v_t)\Lambda\big(R(v_t)\big) \right] dt \\ + \sigma \left[g'\big(a(v_t)\big) + \frac{v_t}{\rho} - \Lambda\big(R(v_t)\big) \right] dB_t^a \quad (2.30)$$

has a unique solution (in the weak probability sense) when $t \in [0, \tau)$*, where* τ *is the (random) retirement time.*

Furthermore, the contract (a, c, R) *defined by*

$$a_t = a(v_t)\mathbf{1}_{t<\tau}, \quad c_t = c(v_t)\mathbf{1}_{t<\tau} - \lambda_r(v_\tau)\mathbf{1}_{t\geq\tau}, \quad R_t = R(v_t)\mathbf{1}_{t<\tau} \quad (2.31)$$

is incentive compatible for the agent. Moreover, it gives payoffs v_0 *to the agent and* $m(v_0)$ *to the principal.*

In terms of the underlying contract, the previous results sets up the lower bound 0 and the upper bound $v_{\max}$ such that the contract is active when the continuation value v lies in $[0, v_{\max}]$. Recall that the agent's continuation value changes over time as a diffusion as the innovations dB_t are realized. The agent becomes increasingly expensive to incentivize as his continuation value increases: working harder and thus finishing the project earlier with higher probability is unattractive because the the agent who would rather collect her promised continuation value. The threshold $v_{\max}$ is the limit beyond which the continuation value is so high that it is not possible to incentivize even a positive amount of effort: thus the agent is retired with a constant payoff via a constant flow compensation.

When retirement obtains, the principal has a negative continuation payoff: he must pay the agent without getting anything in return. That retirement (i.e., contract termination) occurs at a value $v_{\max}$ at which the principal's continuation payoff is strictly negative is an interesting feature of our solution. It implies that there is a neighborhood of $v_{\max}$ for which contracting with the agent continues while the principal's value is negative. The interpretation is that it is optimal for the principal to be patient, and commit to continue the contract, in order to maximally incentivize the agent. (Note that when v is in this neighborhood, it is possible for the continuation value to drop down to a level where the principal again enjoys positive continuation value.)

That the solution to the diffusion exists is a standard result from stochastic differential equations; that v_0 is indeed from the contract (a, c, R) follows from comparing the diffusion to that in Lemma 2.1. Thus, the substantive part of the theorem is the claim verifying that $m(v_0)$ is indeed the valuation of the principal. For this, we verify that

$$F_t = \rho \int_0^t e^{-\rho s} \left(e^{-\int_0^s a_z dz} (-c_s) + e^{-\int_0^s a_z dz} a_s (\Delta - R_s) \right) ds + e^{-\rho t} e^{-\int_0^t a_z dz} m(v_t) \quad (2.32)$$

is a martingale, as its drift is set to zero by the optimality condition of the HJB. In fact, for *any* incentive compatible contract, the process F_t is a super-martingale. That is the basis of the proof of the next theorem, which establishes that the solution to the HJB equation is exactly the principal's valuation.

Theorem 2.7. *Let $m(\cdot)$ be the solution to the HJB equation* (2.28). *Then the payoff to the principal for any incentive compatible contract (a, c, R) starting at v_0 is at most $m\big(v_0(a, c, R)\big)$.*

In particular, this allows for the choice of the optimal starting value v_0 to initialize the agent's continuation. That is, the principal chooses

$$v^* = \arg\max_{v \in [0, v_{\max}]} m(v) \quad (2.33)$$

and sets $v_0 = v^*$ to achieve the maximal payoff with the contract characterized by the HJB equation.

2.3.9 Compensation Cutoff and Positive Action Before Retirement

Given the characterization of the contract above, we can also establish some properties of the contract during employment time. The next proposition shows that in the optimal contract

(a, c, R), the agent always works, i.e.,

$$a(v) > 0 \quad \text{for all } v \in (0, v_{\max}). \tag{2.34}$$

Moreover, when the agent's continuation value is less than the threshold $v^* = \arg\max m(v)$, the agent's compensation is zero and the only incentive is through the lump-sum reward $R(v)$.

Proposition 2.8. *The following properties hold for the optimal contract* (a, c, R).

(i) *The agent always works in that* $a(v) > 0$ *for* $v \in (0, v_{\max})$.

(ii) *The agent's compensation is zero when the continuation value is below the cutoff, i.e.,* $c(v) = 0$ *for* $v \in [0, v^*]$, *where* $v^* = \arg\max m(v)$.

(iii) *The agent's compensation increases monotonically after* v^*, *i.e.,* c *is increasing on* $(v^*, v_{\max})$.

Each of these properties are easy consequences of the structure of the HJB equation. The first property is intuitive given that the agent's flow utility has infinite growth near zero (it is clear that flow compensation drops to zero if the agent exerts no effort). The last two parts of the proposition reflect that compensation is effective (over lump-sum payments) only when the agent has a high enough valuation.

2.4 Illustrations and Simulations

If the utility and cost functions are specified, we may use an ODE solver to compute the principal's value function and the associated optimal contract. For this section, we set

$$u(c) = \sqrt{c}, \quad \Lambda(R) = \sqrt{R}, \ \text{ and } \ g(a) = \frac{1}{2}a^2 + 2a, \tag{2.35}$$

with the parameters $\rho = 0.1$, $\sigma = 1$, and $\Delta = 1.6$. In particular, the first best contract is

$$a^* = c^* = \frac{1}{4}R^* = \frac{1}{4}. \tag{2.36}$$

The principal's value function is shown in Figure 2-1, and the associated optimal contract is shown in Figures 2-2 and 2-3 on page 62.

Recall that the retirement value $v_{\max}$ is endogenously determined by the smooth pasting condition (2.27). The value function is increasing for $v < v^*$ and decreasing for $v > v^*$. This reflects the fact that for small v, growth in the agent's value function is driven by the increase in the probability of R&D success, so that $m(v)$ increases. For larger values of v, the increase is dominated by the increase in compensation (c.f., Figure 2-2 and 2-3), which lowers the principal's value.

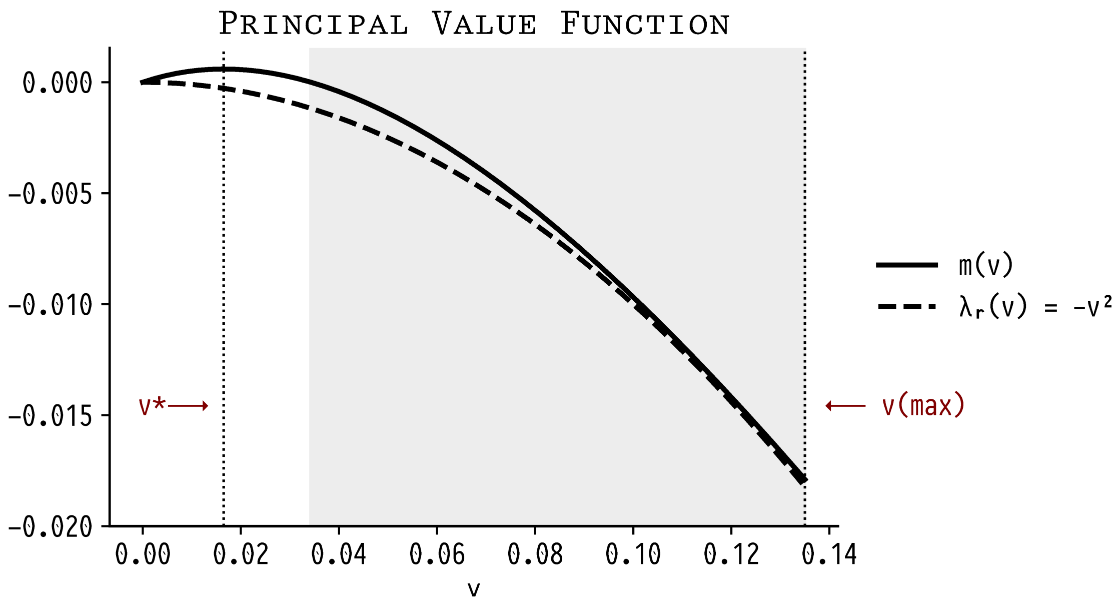

Figure 2-1: The principal's value function $m(v)$ against the retirement value $\lambda_r(v)$. The value $v_{\max}$ is determined when the slopes coincide: $m'(v_{\max}) = \lambda_r'(v)$. For this parameterization, we have $v_{\max} \approx 0.13$, as shown. The dotted vertical line shows the maximizer of m, denoted v^*.

Also interesting is the fact that the magnitude of the flow payment is dwarfed by the lump-sum reward. In words, this means that (for this choice of parameters), the agent is incentivized only by the successful completion of the project. As mentioned in the introduction, this agrees with empirical evidence on long-term CEO compensation.

Finally, we plot the *mean trajectory* of the value functions v_t, a_t, c_t, and R_t. More precisely, we performed following simulation. Let $B = 10^6$ and $\Delta t = 10^{-4}$. For $b = 1, \ldots, B$, we generate a sample path of v_t given by (2.19) where the Brownian motion is approximated with step size Δt.

The diffusion is stopped when v_t falls below zero or v_t exceeds $v_{\max}$ (recall that $v_{\max} \approx 0.13$ for our specific choices of parameters). If the diffusion is stopped because $v_t \leq 0$, then we generate a new sample path v_t, and so on; thus all sample paths of v_t are stopped due to it exceeding $v_{\max}$.

For each of these B sample paths, we record the paths $a(v_t)$, $c(v_t)$, and $R(v_t)$. It is the mean of these quantities that are shown below. To maintain stability, we only consider paths for which the diffusion stops after 700 steps. The result of this simulation is shown in Figure 2-4.

The simulations reveal the fact that conditional on the R&D project being unsuccessful, the lump-sum reward decreases over time. In other words, there is greater incentive to finish the R&D project early (on average). Moreover, the agent puts in less effort if the project fails to succeed before a given time (namely, the time for which a is maximized).

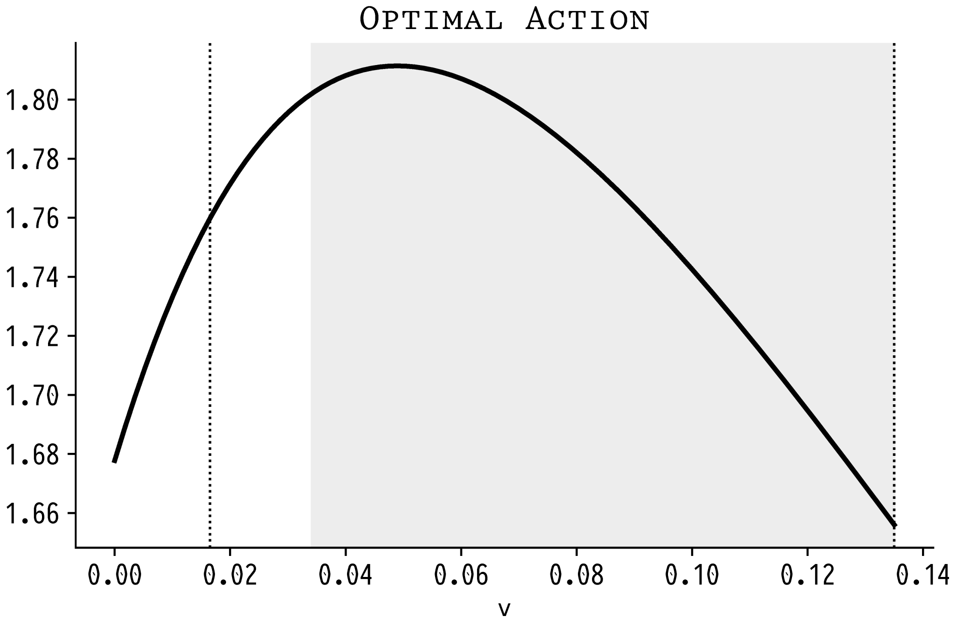

Figure 2-2: The effort $a(v)$ in the optimal contract.

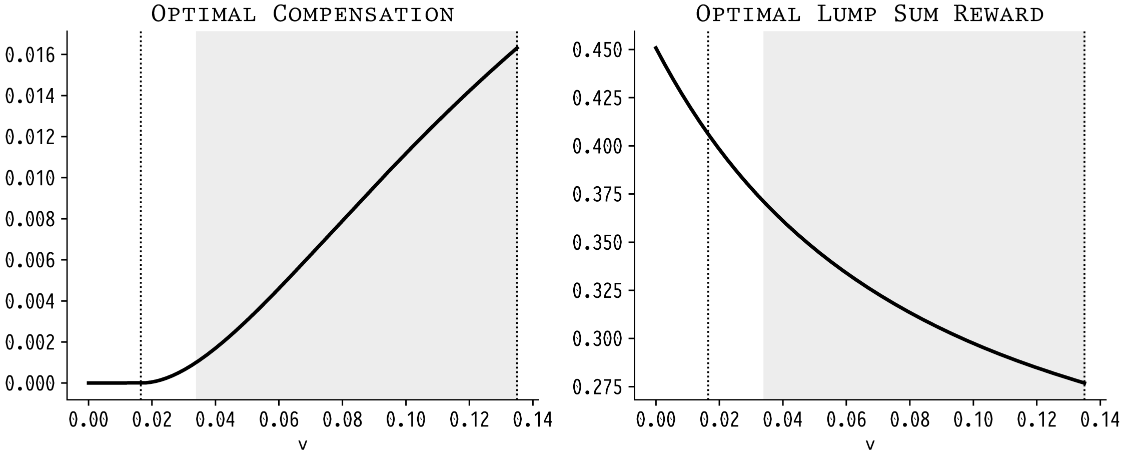

Figure 2-3: The optimal compensation and lump-sum reward.

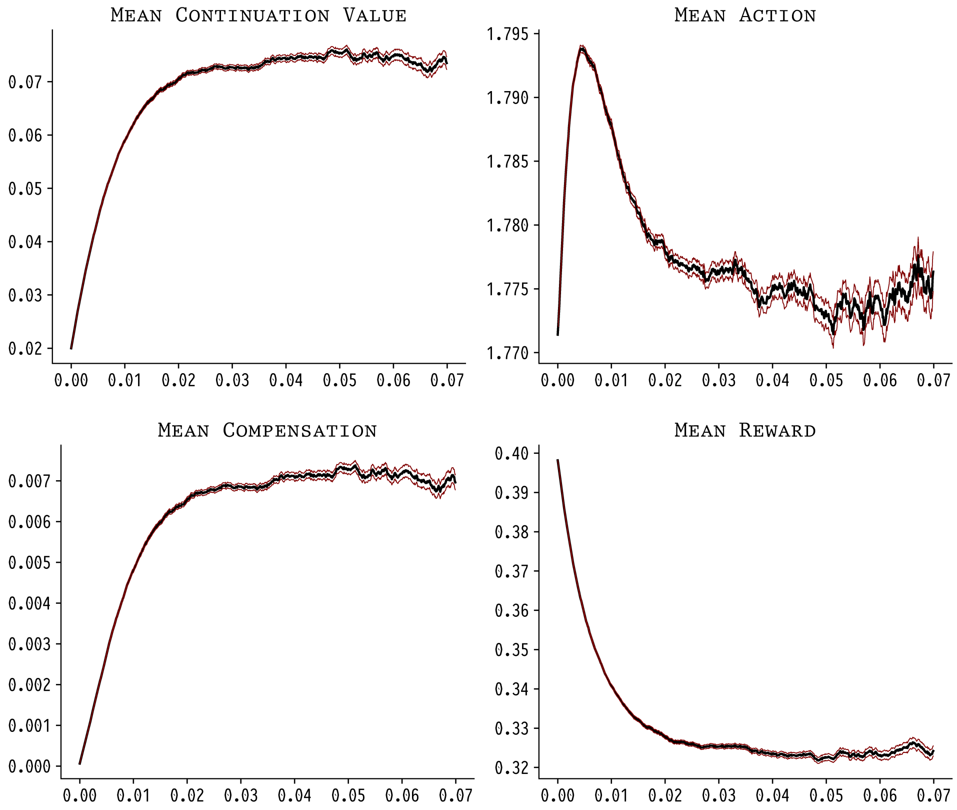

Figure 2-4: Mean trajectories of the agent's value, the optimal action, the optimal compensation, and the optimal reward during the employment time, (the x-axis is the *employment time*. Ninety-five percent confidence bands plotted in maroon color.

2.5 Empirical Findings

We demonstrate the relevance of our two dimensional incentive pay model with a couple empirical exercises. The basic setup is as follows: we map variables of executive pay found in the in the ExecuComp database with our model's compensation plans (the endogenous variable of interest) and map the second finite difference of stock returns to the successful completion of a long term project.

Specifically, the ExecuComp dataset on executive pay decomposes executive compensation into the following categories: salary, bonus, stock awards, option awards, non-equity incentives, pension changes, and all other categories. The salary and non-equity incentives most cleanly map to our model's c and R, respectively; the bonus component is also plausibly mapped to R, though it contributes very little to overall compensation in our dataset that its effects is washed out.

The first point we would like to make is that executive compensation is two-dimensional (or multi-dimensional) in that salary, option and stock rewards, and non-equity incentives each all make up a sizable proportion of the executive's total compensation. For each of the components of executive pay described above, we plot its proportion to overall pay (i.e., a number in $[0, 1]$) in Figure 2-5. We map our model's "flow compensation" to the salary component, and map the "lump-sum compensation" to the sum of the stock awards, option awards, and non-equity incentives. Both groups contribute significantly, with the latter group taking on heavy tails, as shown in Figure 2-6.

To further motivate our model, we gauge the degree to which executive incentive pay is linked to the completion of successful projects. We link the latter with the second finite difference of monthly percentage returns r_m on the stock price,

$$\tilde{a}_m = (r_m - r_{m-1}) - (r_{m-1} - r_{m-2}) \tag{2.37}$$

which we will call *acceleration*. The basic idea behind linking acceleration with the completion of projects is that the success of the latter may result in a new product or profit center which increases the month-over-month growth $r_m - r_{m-1}$. In our dataset, $\tilde{a}_m$ is observed monthly (i.e., m is indexed by months) while executive compensation variables are observed yearly. Therefore, we resample the monthly returns to annual returns by averaging,

$$a_t = \frac{1}{12} \sum_m \tilde{a}_m, \qquad \text{where the sum is taken over the months in year } t. \tag{2.38}$$

To test our model's hypothesis, we regress non-incentive pay, expressed as a proportion of total compensation, on a_t and other controls. In symbols, the estimation equation is

$$y_{it} = \beta a_{it} + \gamma^{\mathsf{T}} x_{it} + \epsilon_{it} \tag{2.39}$$

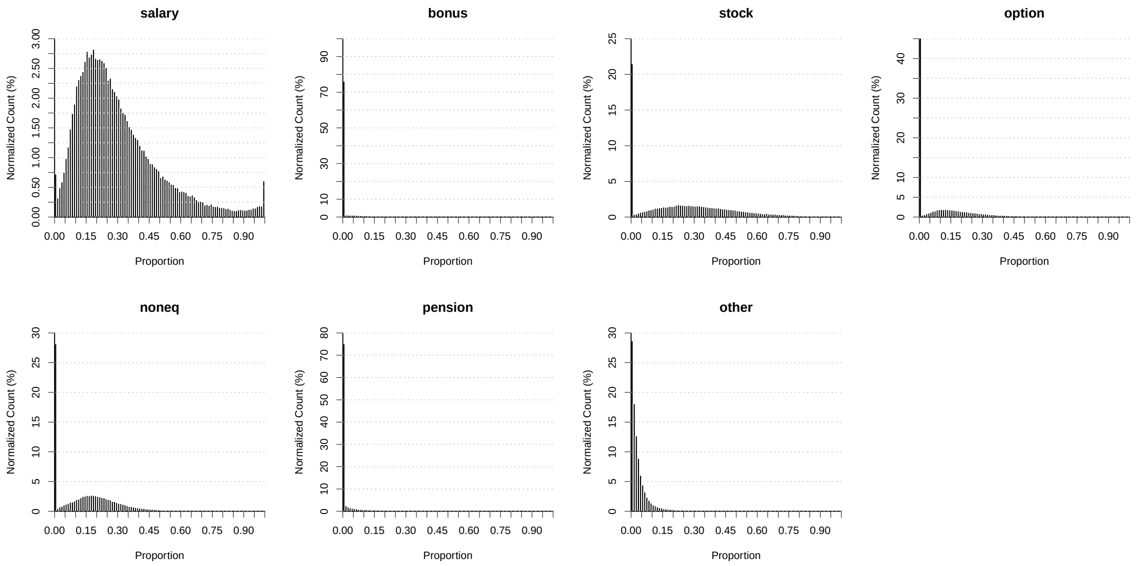

Figure 2-5: The seven components of executive compensation in the ExecuComp dataset. Each histogram plots the proportion of the corresponding component as part of total compensation. Note that sum of stock, option, and non-equity incentives peak around ten percent, far exceeding the salary component, which tops out at around three percent.

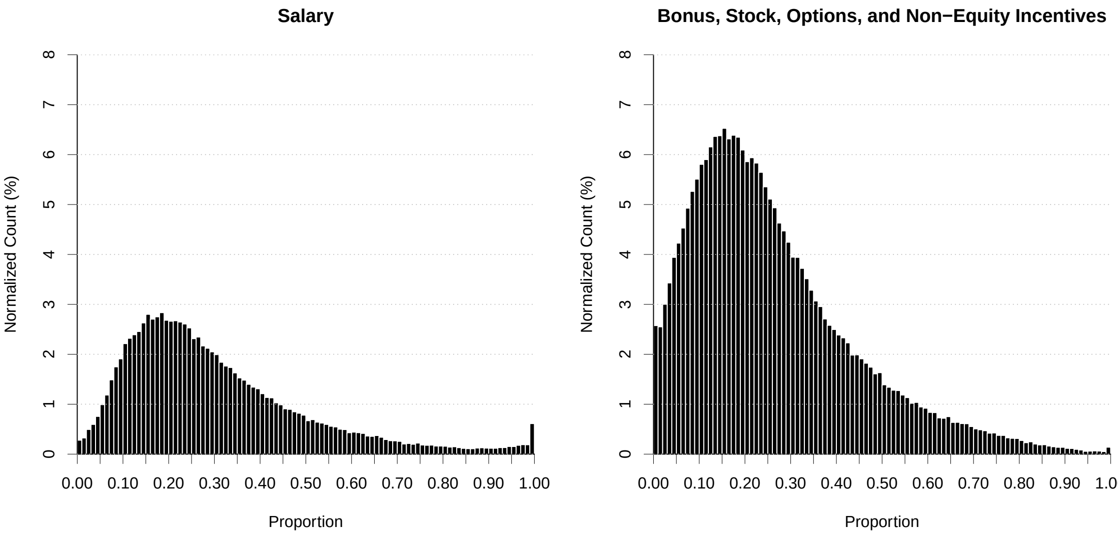

Figure 2-6: Comparison of salary versus bonuses. In general, both distributions exhibit heavy tails, indicating the presence of compensation packages for which the salary component (resp. the bonus component) dominate total compensation. Also note that the distribution of bonuses stochastically dominate that of the salary at every fixed level: for example, there are significantly many more executives whose bonus pay exceeds 70% of total compensation than those whose salary exceeds 70% of total compensation.

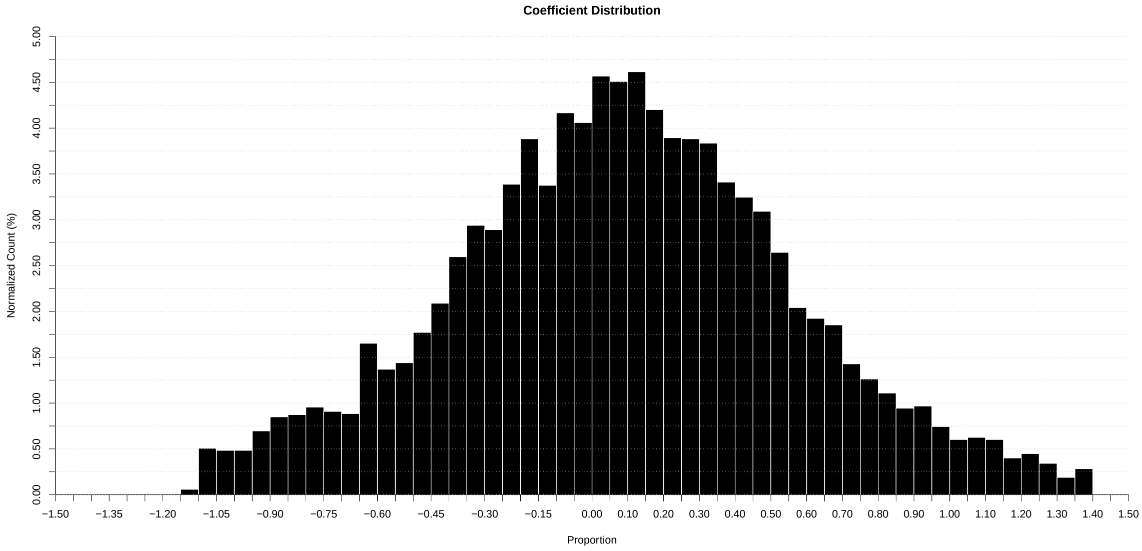

Figure 2-7: Distribution of regression coefficients across each executive; specification (1) is used.

where the indices (i, t) runs over executive-year pairs present in the dataset after implausible outliers are removed. Here, $y_{it} \in [0, 1]$ is proportion of non-equity incentive to total compensation, and x_{it} is a vector of controls depending on the specification. The results of this linear regression is found in Table 2.1; we see that the coefficient of interest β is statistically significant across each specification, consistent with the hypothesis and equilibrium outcome of our model. Recall that acceleration is measured as the change in the increase of returns from one year to the next, a coefficient of 0.16% implies that a acceleration of monthly stock returns by 1% induces a 16% increase in the proportion of non-equity incentive pay in total compensation. (Our specifications focuses on the effect of non-equity incentives since since our measure of project completion (return acceleration) is itself a measure of a stock price.)

Finally, we run the same regression as (2.39) separately for each executive, namely, for each i, we collect the vectors $y_t = \{y_{it}\}$ and $a_t = \{x_{it}\}$ and run the regression using data points in y_t and x_t, producing a coefficient $\beta = \beta_i$ for each executive. The empirical distribution of β_i of specification (1) is shown in Figure 2-7. These regressions roughly correspond to a panel data setup with executive fixed effects; as expected, the distribution of slopes have a positive bias. This is inline with the regressions in Table 2.1.

Table 2.1: Effect of growth acceleration on non-equity incentives.

	Dependent variable:	
	noneq	
	(1)	(2)
acceleration	0.158***	0.165***
	(0.025)	(0.022)
salary	−0.137***	−0.427***
	(0.002)	(0.003)
stock		−0.384***
		(0.002)
option		−0.400***
		(0.003)
Constant	0.206***	0.449***
	(0.001)	(0.001)
Observations	115,958	115,958
R^2	0.031	0.251
Adjusted R^2	0.031	0.251
Residual Std. Error	0.146	0.128

Note. Specification (1) includes salary component as a control, and specification (2) adds stock and option rewards. Here, acceleration is measured as the percentage increase of the change in returns from one year to the next.

2.6 Conclusion

In this paper, we develop a model of contracting a research and development (R&D) project in continuous-time where the agent's efforts are only indirectly observed via a noisy signal. The agent's effort is costly and modeled by a convex function. The principal rewards the effort with a two-dimensional incentive-pay contract: a flow compensation over the course of R&D plus a lump-sum reward if the R&D is successful. When agent's action is fully observable, the optimal (first-best) contract is stationary (i.e., constant). However, when the agency conflict is binding (second-best), the optimal dynamic contract is time varying.

Our first result expresses the agent's continuation value as diffusion where the drift is related to the path of effort levels and the diffusion related to the sensitivity of the agent to the noisy signal. Using this result, we characterize the entire space of incentive compatible contracts in terms of the sensitivity. And, as a consequence, we derive an HJB equation explicitly characterizing the optimal contract. Our main result shows that the solution to the HJB equation (i.e., the optimal contract) exists and it is unique. Time of hiring and firing the agent are both endogenous. Particularly, the nature of the HJB equation furnishes a key quantity $v_{\max}$, the retirement (firing) value of the agent. Before the agent's valuation exceeds $v_{\max}$, the principal retains the agent even if his own value may be negative. The interpretation is that the optimal contract features a patient principal who commits to continue the project when it is close enough to completion. Hiring time is also uniquely pinned down in terms of the agent's valuation, maximizing the principal valuation.

As a consequence of the HJB equation, we also derive some qualitative properties of the optimal contract. Particularly, there is a unique threshold on the agent's valuation so that agent's optimal compensation is zero below the threshold, and increases in the agent's valuation above the threshold. Thus, when agent's value is small enough, the principal provides incentive-pay by only a positive lump-sum bonus upon successful completion of the project. Our key results are valid for a large set of possible utility and cost functions. In particular, the only condition we require outside of the usual concavity and convexity conditions is a lower bound on the slope of the cost function, viz. (A3).

To gauge the practice relevance of our theoretical results, we use executive compensation data from ExecuComp to demonstrate the multi-dimensional nature of compensation. We also show that the bonus component of compensation, in particular the non-equity incentives, is correlated with the acceleration of the growth of the company's stock performance, which we use as a proxy for the completion of long-term projects. This agrees with the qualitative characteristics of the optimal contract discussed above.

2.7 Appendix

2.7.1 Assumptions

We will maintain the following assumptions throughout.

(A1) The utility functions satisfy $\rho\Lambda(R) \leq u(\rho R)$.

(A2) If $\bar{R}$ is the maximum reward possible, and $g'(0) > \Lambda(\bar{R})$.

(A3) The upper bound on c, denoted $\bar{c}$, has the property that $\frac{g(a)}{a} \geq u'(\bar{c})\Delta$. For tidiness, we will write $u'(\bar{c}) = \kappa_0$.

2.7.2 Omitted Proofs

Proof of Lemma 2.1. We start with deriving the principal's value function. We consider Γ_0^P (extension to Γ_t^P is straightforward). It will be convenient to set

$$X_\tau = \int_0^\tau e^{-\rho t}(-c_t)dt + e^{-\rho\tau}(\Delta - R_\tau).$$

With this notation, recall that the principal's value function at time 0 is

$$\Gamma_0^P = \rho\,\mathbf{E}\Big[\Pr(\tau < \infty)\int_0^\infty X_\tau f_a(\tau)d\tau - \Pr(\tau = \infty)\int_0^\infty e^{-\rho t}c_t dt\Big].$$

Thus, plugging $f_a(\tau) = \frac{a_\tau e^{-\int_0^\tau a_z dz}}{1-e^{-\int_0^\infty a_z dz}}$ and rearranging implies

$$\rho^{-1}\Gamma_0^P = \mathbf{E}\Big[\int_0^\infty X_\tau e^{-\int_0^\tau a_z dz}a_\tau d\tau + e^{-\int_0^\infty a_z dz}\int_0^\infty e^{-\rho t}(-c_t)dt\Big].$$

Using Fubini's theorem and change in the order of integrals gives

$$\begin{aligned}\rho^{-1}\Gamma_0^P &= \mathbf{E}\Big[\int_0^\infty\int_t^\infty e^{-\rho t}(-c_t)e^{-\int_0^\tau a_z dz}a_\tau d\tau dt + e^{-\int_0^\infty a_z dz}\int_0^\infty e^{-\rho t}(-c_t)dt\Big]\\ &\quad + \mathbf{E}\Big[\int_0^\infty e^{-\rho\tau}(\Delta - R_\tau)e^{-\int_0^\tau a_z dz}a_\tau d\tau\Big].\end{aligned}$$

Next, we note that $\int_t^\infty e^{-\int_0^\tau a_z dz}a_\tau d\tau = e^{-\int_0^t a_z dz} - e^{-\int_0^\infty a_z dz}$ (which follows by differentiating), as a result we have

$$\begin{aligned}\rho^{-1}\Gamma_0^P &= \mathbf{E}\Big[\int_0^\infty e^{-\rho t}(-c_t)[e^{-\int_0^t a_z dz} - e^{-\int_0^\infty a_z dz}]dt + e^{-\int_0^\infty a_z dz}\int_0^\infty e^{-\rho t}(-c_t)dt\\ &\quad + \int_0^\infty e^{-\rho\tau}(\Delta - R_\tau)e^{-\int_0^\tau a_z dz}a_\tau d\tau\Big]\\ &= \mathbf{E}\Big[\int_0^\infty e^{-\rho t}(e^{-\int_0^t a_z dz}(-c_t)dt + e^{-\int_0^t a_z dz}a_t(\Delta - R_t)dt)\Big].\end{aligned}$$

Similar steps follows to simplify the agent's value function. Her valuation at time zero is

$$
\begin{aligned}
v_0 = \rho \mathbf{E}\Big\{ &\Pr\{\tau < \infty\} \int_0^\infty \left[\int_0^\tau e^{-\rho t}(u(c_t) - g(a_t))dt + e^{-\rho\tau}\Lambda(R_\tau) \right] f_a(\tau)d\tau \\
&+ \Pr\{\tau = \infty\} \int_0^\infty e^{-\rho t}(u(c_t) - g(a_t))dt \Big\}.
\end{aligned}
$$

Thus, plugging in $f_a(\tau) = \frac{a_\tau e^{-\int_0^\tau a_z dz}}{1-e^{-\int_0^\infty a_z dz}}$ and rearranging implies that $\rho^{-1}v_0$ is equal to

$$
\mathbf{E}\left[\int_0^\infty Y_\tau e^{-\int_0^\tau a_z dz} a_\tau d\tau + e^{-\int_0^\infty a_z dz} \int_0^\infty e^{-\rho t}(u(c_t) - g(a_t))dt \right],
$$

where

$$
Y_\tau = \int_0^\tau e^{-\rho t}(u(c_t) - g(a_t))dt + e^{-\rho\tau}\Lambda(R_\tau).
$$

Upon rearranging, this is equal to the sum of the expectations of the quantities

$$
\int_0^\infty \left[\int_0^\tau e^{-\rho t}(u(c_t) - g(a_t))dt \right] e^{-\int_0^\tau a_z dz} a_\tau d\tau, \qquad \int_0^\infty e^{-\rho\tau}\Lambda(R_\tau) e^{-\int_0^\tau a_z dz} a_\tau d\tau,
$$

and

$$
e^{-\int_0^\infty a_z dz} \int_0^\infty e^{-\rho t}(u(c_t) - g(a_t))dt.
$$

Next, change in the order of integrals gives

$$
\begin{aligned}
\rho^{-1}v_0 = \mathbf{E}\Big[&\int_0^\infty e^{-\rho t}(u(c_t) - g(a_t)) \left[\int_t^\infty e^{-\int_0^\tau a_z dz} a_\tau d\tau \right] dt \\
&+ \int_0^\infty e^{-\rho\tau}\Lambda(R_\tau) e^{-\int_0^\tau a_z dz} a_\tau d\tau + e^{-\int_0^\infty a_z dz} \int_0^\infty e^{-\rho t}(u(c_t) - g(a_t))dt \Big].
\end{aligned}
$$

Finally, since $\int_t^\infty e^{-\int_0^\tau a_z dz} a_\tau d\tau = e^{-\int_0^t a_z dz} - e^{-\int_0^\infty a_z dz}$, we have

$$
\rho^{-1}v_0 = \mathbf{E}\left[\int_0^\infty e^{-\rho t}(e^{-\int_0^t a_z dz}(u(c_t) - g(a_t))dt + e^{-\int_0^t a_z dz} a_t \Lambda(R_t)dt) \right],
$$

finishing the proof. □

Proof of Lemma 2.2. Fix the contract (a, c, R). Define V_t to be the expected lifetime utility evaluated conditional on time t information,

$$
\begin{aligned}
V_t = \rho &\int_0^t e^{-\rho s}(e^{-\int_0^s a_z dz}(u(c_s) - g(a_s)) + e^{-\int_0^s a_z dz} a_s \Lambda(R_s))ds \qquad (2.40) \\
&e^{-\rho t} e^{-\int_0^t a_z dz} v_t(a, c, R).
\end{aligned}
$$

Clearly, V_t is a (Doob) martingale under $\mathbf{P}^a$; applying Martingale Representation Theorem,

$$V_t = V_0 + \int_0^t \rho\sigma e^{-\rho s} e^{-\int_0^s a_z dz} \varphi_s dB_s^a, \tag{2.41}$$

where $B_t^a = \frac{1}{\sigma}(y_t - \int_0^t a_z dz)$ is a Wiener process under $\mathbf{P}^a$ and φ_t is $\mathfrak{F}_t$-measurable and $\mathbf{E}^a[\int_0^t \varphi_s^2 ds] < \infty$. Recall that

$$V_t = \rho \int_0^t e^{-\rho s} e^{-\int_0^s a_z dz} \Big(u(c_s) - g(a_s) + a_s \Lambda(R_s)\Big) ds + e^{-\rho t} e^{-\int_0^t a_z dz} v_t(a, c, R),$$

thus differentiating the expression with respect to V_t (using Ito's lemma) gives:

$$\begin{aligned} dV_t &= \rho e^{-\rho t} e^{-\int_0^t a_z dz} \Big(u(c_t) - g(a_t) + a_t \Lambda(R_t)\Big) dt + d\left(e^{-\rho t} e^{-\int_0^t a_z dz} v_t(a, c, R)\right) \\ &= \rho e^{-\rho t} e^{-\int_0^t a_z dz} \Big(u(c_t) - g(a_t) + a_t \Lambda(R_t)\Big) dt + e^{-\rho t} e^{-\int_0^t a_z dz} dv_t(a, c, R) \\ &\quad - (\rho + a_t) e^{-\rho t} e^{-\int_0^t a_z dz} v_t(a, c, R) dt. \end{aligned} \tag{2.42}$$

Also from (2.41) we have

$$dV_t = \sigma\rho e^{-\rho t} e^{-\int_0^t a_z dz} \varphi_t dB_t^a. \tag{2.43}$$

Plugging (2.43) into (2.41) and rearranging gives

$$\rho^{-1} dv_t(a, c, R) = \eta_t dt + \varphi_t (dy_t - a_t dt) = \eta_t dt + \varphi_t \sigma dB_t^a, \tag{2.44}$$

where

$$\eta_t = v_t(a, c, R)\left(1 + \frac{a_t}{\rho}\right) + g(a_t) - u(c_t) - a_t \Lambda(R_t), \tag{2.45}$$

finishing the proof. □

Proof of Theorem 2.3. We prove the **if and only if** parts separately as follows.

If Part. Let us assume that

$$a'\left(\varphi_t + \Lambda(R_t) - \frac{v_t(a, c, R)}{\rho}\right) - g(a') \leq a_t \left(\varphi_t + \Lambda(R_t) - \frac{v_t(a, c, R)}{\rho}\right) - g(a_t),$$

(the IC constraint) holds. Now, consider an alternative strategy $\tilde{a}$. Due to (2.40), we have

$$\tilde{V}_t = \rho \int_0^t e^{-\rho s} \Big(e^{-\int_0^s \tilde{a}_z dz}(u(c_s) - g(\tilde{a}_s)) + e^{-\int_0^s \tilde{a}_z dz} \tilde{a}_s \Lambda(R_s)\Big) ds + e^{-\rho t} e^{-\int_0^t \tilde{a}_z dz} v_t(a, c, R).$$

Then, $\tilde{V}_t$ is a $\mathbf{P}^{\tilde{a}}$-supermartingale for each alternative strategy $\tilde{a}$. Also, $\tilde{V}_t$ is uniformly integrable. Hence

$$\tilde{V}_\infty \equiv \lim_{t\to\infty} \tilde{V}_t = \rho \int_0^\infty e^{-\rho s}\Big(e^{-\int_0^s \tilde{a}_z dz}(u(c_s) - g(\tilde{a}_s)) + e^{-\int_0^s \tilde{a}_z dz}\tilde{a}_s \Lambda(R_s)\Big)ds,$$

that follows by the Doob's convergence theorem Karatzas and Shreve (1991). As a result, the alternative strategy $\tilde{a}$ does not outperform strategy a because:

$$\begin{aligned} v_0(\tilde{a}, c, R) = \mathbf{E}^{\tilde{a}}\left[\tilde{V}_\infty\right] &= \mathbf{E}^{\tilde{a}}\left[\rho \int_0^\infty e^{-\rho s}\Big(e^{-\int_0^s \tilde{a}_z dz}(u(c_s) - g(\tilde{a}_s)) + e^{-\int_0^s \tilde{a}_z dz}\tilde{a}_s \Lambda(R_s)\Big)ds\right] \\ &\le \tilde{V}_0 = v_0(a, c, R). \end{aligned}$$

Only If Part. Let us consider $\tilde{a}$ (an alternative strategy). The expected (average) lifetime utility evaluated conditional time t information if agent uses $\tilde{a}$ for all time t' up to time t (including t) and then follows a after time t, is given by

$$\tilde{V}_t = \rho \int_0^t e^{-\rho s}\Big(e^{-\int_0^s \tilde{a}_z dz}(u(c_s) - g(\tilde{a}_s)) + e^{-\int_0^s \tilde{a}_z dz}\tilde{a}_s \Lambda(R_s)\Big)ds + e^{-\rho t}e^{-\int_0^t \tilde{a}_z dz}v_t(a, c, R).$$

Using Ito's lemma, under probability measure $\mathbf{P}^{\tilde{a}}$, we have

$$\begin{aligned} d\tilde{V}_t &= \rho e^{-\rho t}\Big(e^{-\int_0^t \tilde{a}_z dz}(u(c_t) - g(\tilde{a}_t)) + e^{-\int_0^t \tilde{a}_z dz}\tilde{a}_t \Lambda(R_t)\Big)dt \\ &\quad + d(e^{-\rho t}e^{-\int_0^t \tilde{a}_z dz}v_t(a, c, R)). \end{aligned} \tag{2.46}$$

Applying Ito's lemma implies

$$d(e^{-\rho t}e^{-\int_0^t \tilde{a}_z dz}v_t(a, c, R)) = e^{-\rho t}e^{-\int_0^t \tilde{a}_z dz}dv_t(a, c, R) - (\rho + \tilde{a}_t)e^{-\rho t}e^{-\int_0^t \tilde{a}_z dz}v_t(a, c, R)dt.$$

Hence, above equality and substituting $dv_t(a, c, R)$ (from (2.44)) and rearranging give

$$\begin{aligned} e^{\rho t}e^{\int_0^t \tilde{a}_z dz}d\tilde{V}_t =& \rho\Big(u(c_t) - g(\tilde{a}_t) + \tilde{a}_t\Lambda(R_t)\Big)dt - \tilde{a}_t v_t(a, c, R)dt \\ &+ a_t v_t(a, c, R)dt - \rho\Big(u(c_t) - g(a_t) + a_t\Lambda(R_t)\Big)dt \\ &+ \rho\varphi_t\sigma dB_t^a. \end{aligned} \tag{2.47}$$

Using Girsanov's Change of Measure Theorem (Revuz and Yor (2004)), the Wiener processes under probability measures $\mathbf{P}^a$ and $\mathbf{P}^{\tilde{a}}$ can be linked so that

$$\sigma(dB_t^a - dB_t^{\tilde{a}}) = (\tilde{a}_t - a_t)dt.$$

Hence, substituting $\sigma dB_t^a = (\tilde{a}_t - a_t)dt + \sigma dB_t^{\tilde{a}}$ into (2.47) and rearranging terms,

$$\rho^{-1}e^{\rho t}e^{\int_0^t \tilde{a}_z dz}d\tilde{V}_t = \tilde{\eta}_t dt + \varphi_t \sigma dB_t^{\tilde{a}}, \tag{2.48}$$

where

$$\tilde{\eta}_t = g(a_t) - a_t\left(\varphi_t + \Lambda(R_t) - \frac{v_t(a,c,R)}{\rho}\right) - g(\tilde{a}_t) + \tilde{a}_t\left(\varphi_t + \Lambda(R_t) - \frac{v_t(a,c,R)}{\rho}\right).$$

Now, for the sake of contradiction with the IC constraint, suppose that $\tilde{a}_t$ outperforms a_t on a set of positive measure, so that $\tilde{\eta}_t > 0$. Therefore, under the probability measure $\mathbf{P}^{\tilde{a}}$, the drift of $d\tilde{V}_t$ (see (2.48)) is non-negative almost surely with positive expectation. It follows that there is some t for which $\mathbf{E}^{\tilde{a}}[\tilde{V}_t] > \tilde{V}_0 = v_0(a,c,R)$, which is a contradiction with the optimality of strategy a. To see this, note that the agent receives $\mathbf{E}^{\tilde{a}}[\tilde{V}_t]$ by following $\tilde{a}$ up to time t and then following a afterwards. □

Proof of Proposition 2.4. Suppose $v_0 > \tilde{v} = u(\bar{c})$ and $u(c_1) = v_0$. Recall that $u'(\bar{c}) = \kappa_0$. Therefore, because $u(\cdot)$ is concave and increasing thus $v_0 > \tilde{v}$ implies $c_1 > \bar{c}$ and $u'(c_1) \leq \kappa_0$. Our objective is to show that $m(v_0) \leq -c_1 = -u^{-1}(v_0) = \lambda_r(v_0)$. Using Proposition 2.1 we have $v_0 = \rho\mathbf{E}\left[\int_0^\infty e^{-\rho t}(e^{-\int_0^t a_z dz}\big(u(c_t) - g(a_t)\big)dt + e^{-\int_0^t a_z dz}a_t\Lambda(R_t)dt)\right]$. Recall that (by assumption) $\Lambda(R)\rho \leq u(\rho R)$ therefore

$$T_1 := v_0 \leq \rho\mathbf{E}\left[\int_0^\infty e^{-\rho t}(e^{-\int_0^t a_z dz}\big(u(c_t) - g(a_t)\big)dt + e^{-\int_0^t a_z dz}\frac{a_t}{\rho}u(\rho R_t)dt)\right]$$

Note that $u(\cdot)$ is concave and differentiable. Thus, Taylor expansion of $u(\cdot)$ around c_1 shows that T_1 is at most

$$\rho\mathbf{E}\Bigg[\int_0^\infty e^{-\rho t}(e^{-\int_0^t a_z dz}\big(u(c_1) + u'(c_1)(c_t - c_1) - g(a_t)\big) \\ + e^{-\int_0^t a_z dz}\frac{a_t}{\rho}(u(c_1) + u'(c_1)(\rho R_t - c_1)))dt\Bigg] \tag{2.49}$$

which (by rearranging) is equal to

$$T_2 := \rho\mathbf{E}\left[\int_0^\infty e^{-\rho t}(e^{-\int_0^t a_z dz}(u'(c_1)c_t - g(a_t)) + e^{-\int_0^t a_z dz}a_t u'(c_1)R_t)dt\right] \\ + (u(c_1) - c_1 u'(c_1))\left[\int_0^\infty e^{-\rho t}(e^{-\int_0^t a_z dz}\rho + e^{-\int_0^t a_z dz}a_t)dt\right]$$

Next, note that $\frac{u'(c_1)}{\kappa_0} \leq 1$ therefore $-g(a_t) \leq -g(a_t)\frac{u'(c_1)}{\kappa_0}$. Hence

$$\begin{aligned} T_2 \leq \rho u'(c_1)\,\mathbf{E}\left[\int_0^\infty e^{-\rho t}(e^{-\int_0^t a_z dz}(c_t - \frac{g(a_t)}{\kappa_0}) + e^{-\int_0^t a_z dz} a_t R_t)dt\right] \\ + (u(c_1) - c_1 u'(c_1))\left[\int_0^\infty e^{-\rho t}(e^{-\int_0^t a_z dz}\rho + e^{-\int_0^t a_z dz} a_t)dt\right] =: T_3 \end{aligned}$$

Notice that (by assumption) $-\frac{g(a_t)}{\kappa_0} \leq -\Delta a_t$. Thus replacing $-\frac{g(a_t)}{\kappa_0}$ with $-\Delta a_t$ in T_3 and noting that $\int_0^\infty e^{-\rho t}(e^{-\int_0^t a_z dz}\rho + e^{-\int_0^t a_z dz} a_t)dt \leq 1$ we have

$$T_3 \leq \rho u'(c_1)\,\mathbf{E}\left[\int_0^\infty e^{-\rho t}(e^{-\int_0^t a_z dz} c_t + e^{-\int_0^t a_z dz} a_t(R_t - \Delta))dt\right] + (u(c_1) - c_1 u'(c_1)).$$

Define T_4 to be the RHS; so far we have shown that $v_0 \leq T_1 \leq T_2 \leq T_3 \leq T_4$. To finish the proof, consider $v_0 \leq T_4$. Note that $u(c_1) = v_0$. Hence, rearranging gives

$$\begin{aligned} 0 &\leq -u'(c_1)\left(c_1 + \rho\,\mathbf{E}\left[\int_0^\infty e^{-\rho t}(e^{-\int_0^t a_z dz}(-c_t) + e^{-\int_0^t a_z dz} a_t(\Delta - R_t))dt\right]\right) \\ &= -u'(c_1)(c_1 + m(v_0)) \end{aligned}$$

thus $m(v_0) \leq -c_1 = -u^{-1}(v_0) = \lambda_r(v_0)$, as desired, finishing the proof. □

Proof of Theorem 2.5. The proof follows by proving the following claims. □

To begin, the next claim shows that the solution (2.28) is unique.

Claim 1. The solution to the HJB (2.28) is unique.

Proof of Claim 1. The claim immediately follows because the HJB is uniformly elliptic, (see Section IV.5 in Fleming and Soner (2007)). □

The following claim is about the concavity of the HJB solution, namely that if the solution is concave at one point, then it is concave everywhere.

Claim 2. If there exists $\bar{v}$ such that $m''(\bar{v}) < 0$ then $m''(v) < 0$ for all $v > 0$.

Proof of Claim 2. To prove it we use the uniqueness result established in Claim 1. Let us assume $m''(\bar{v}) < 0$. Rearranging the HJB equation gives

$$\begin{aligned} 0 = {} & c - a(\Delta - R) \\ & - m''(v)\frac{\rho\sigma^2}{2}\left(g'(a) + \frac{v}{\rho} - \Lambda(R)\right)^2 \\ & - m'(v)\left(-u(c) + g(a) - a\Lambda(R) + v\Big(1 + \frac{a}{\rho}\Big)\right) \\ & + m(v)\Big(1 + \frac{a}{\rho}\Big). \end{aligned}$$

Next we show that there is no v such that $m''(v) = 0$. For the sake of contradiction, suppose that there is a v for which $m''(v) = 0$. Then the linear function $m(\tilde{v}) = m(v) + m'(v)(\tilde{v} - v)$ solves the HJB and also $m''(v) = 0$. However, according to Claim 1 the solution to the HJB is unique, meaning that the linear function must be the only solution of the HJB equation. However, this contradicts the concavity at the point $\bar{v}$, contradiction. □

Claim 3. The value function $m(\cdot)$ satisfying the HJB in (2.28) is concave.

Proof of Claim 3. Consider the HJB in (2.28). We have $m(0) = \lambda_r(0)$ and for some $v_{\max} > 0$ with $m(v_{\max}) = \lambda_r(v_{\max})$. Suppose $m'(0) > \lambda_r'(0)$. Since $m(0) = \lambda_r(0)$, $m'(0) > \lambda_r'(0)$, and $m(v_{\max}) = \lambda_r(v_{\max})$, there exists $0 < v < v_{\max})$ for which $m'(v) < \lambda_r'(v)$. Moreover, since $\lambda_r(\cdot)$ is concave $\lambda_r'(v) < \lambda_r'(0)$. All of this shows that,

$$m'(v) < \lambda_r'(v) < \lambda_r'(0) < m'(0),$$

implying that there exists $0 < v'' < v$ for which $m''(v'') < 0$. This in turn implies that there is a point v'' in which $m''(v'') < 0$. Using the result in Claim 2, the function $m(\cdot)$ is concave (strictly) every where in $[0, v_{\max}]$. □

Finally, the last claim establishes that v' is bounded by $\tilde{v}$ (recall that $u(\bar{c}) = \tilde{v}$ and Proposition 2.4).

Claim 4. The endogenous retirement value $v_{\max}$ does not explode with escaping to infinity, precisely, $v_{\max} \leq \tilde{v}$.

Proof of Claim 4. We prove the claim by contradiction. Suppose that $v_{\max} > \tilde{v}$. We also have $m'(0) > \lambda_r'(0)$. By concavity of $\lambda_r(\cdot)$, we then have

$$m'(v_{\max}) < \lambda_r'(v_{\max}) \leq \lambda_r'(\tilde{v}). \tag{2.50}$$

The relation in (2.50) implies that $m'(v'') = \lambda_r'(\tilde{v})$ for some $v'' < v_{\max}$, so that

$$m'(v'') = \lambda_r'(\tilde{v}) = -\frac{1}{u'(u^{-1}(\tilde{v}))} = -\frac{1}{u'(\bar{c})} = -\kappa_0^{-1}.$$

where the last equality follows by Assumption (A3). Moreover, the concavity of $m(\cdot)$ implies

$$m(v'') - v''m'(v'') > m(\tilde{v}) - \tilde{v}m'(v'') > \lambda_r(\tilde{v}) - \tilde{v}m'(v'').$$

Finally, to finish the proof, we show that $m''(v'') > 0$ which in contradiction with Claim 2 (concavity of $m(\cdot)$). To do so, we will show that (c.f., (2.28))

$$m''(v'') = \min_{a\in[0,\bar{a}],c\in[0,\bar{c}],R\in[0,\bar{R}]} \Phi_{a,c,R}(v'', m(v''), m'(v'')) > 0.$$

By fixing the action a (which is arbitrary from $[0,\bar{a}]$), it will suffice to show that

$$\min_{c,R} m(v'')+c-a\left(\Delta-\frac{m(v'')}{\rho}-R\right)-m'(v)\left(v''-u(c)+g(a)-a\left(\Lambda(R)-\frac{v}{\rho}\right)\right)>0. \tag{2.51}$$

Transposing terms of the quantity in the previous display yields

$$\begin{aligned}T_1 &:= m(v'')-m'(v'')\left(v''(1+\frac{a}{\rho})+g(a)\right)\\ &\quad -a\left(\Delta-\frac{m(v'')}{\rho}\right)+a\min_R\{R+m'(v'')\Lambda(R)\}+\min_c\{c+m'(v'')u(c)\}.\end{aligned}$$

Since $\rho\Lambda(R)\le u(\rho R)$ (by Assumption 1), thus replacing $\Lambda(R)$ with $u(\rho R)$ gives

$$\begin{aligned}T_1 &\ge m(v'')-m'(v'')\left(v''(1+\frac{a}{\rho})+g(a)\right)-a\left(\Delta-\frac{m(v'')}{\rho}\right)\\ &\quad +a\min_R\left\{R+m'(v'')\frac{u(\rho R)}{\rho}\right\}+\min_c\left\{c+m'(v'')u(c)\right\}:=T_2\end{aligned}$$

Substituting $m'(v'')=-\frac{1}{u'(\bar{c})}$ and rearranging gives

$$\begin{aligned}T_2 &\ge \left(1+\frac{a}{\rho}\right)(\{m(v'')-m'(v'')v''\}+\{\bar{c}+m'(v'')u(\bar{c})\})-a\Delta-m'(v'')g(a)\\ &> -a\Delta-m'(v'')g(a)=\frac{g(a)}{u'(\bar{c})}-a\Delta=\frac{g(a)}{\kappa_0}-a\Delta\ge 0,\end{aligned}$$

where the last equality follows by Assumption (A3). Putting together, $0<T_2\le T_1$ thus $m''(v'')>0$, contradiction with Claim 2. Hence, there must be that $v_{\max}\le\tilde{v}$. Since $m'(0)>\lambda_r'(0)$ the solution will meet $\lambda_r(\cdot)$ at $v_{\max}$ and for all $0\le v\le v_{\max}$ we then have $\lambda_r(v)\le m(v)$, finishing the proof. □

Proof of Theorem 2.6. Since the drift and the volatility are bounded in (2.19), the regular Lipchitz conditions are satisfied (c.f., Karatzas and Shreve (1991)), and as a result the solution to (2.19) exists and is unique (in the distributional sense). Next, we show v_t is equal to $v_t(a,c,R)$, where $v_t(a,c,R)$ is the agent's continuation value from the contract (a,c,R). Using (2.10) from Proposition 2.1 we have

$$v_{t+z}(a,c,R)-v_{t+z}=e^{\rho z}e^{\int_t^{t+z}a(v_s)ds}\left(v_t(a,c,R)-v_t+\rho\sigma\int_t^{t+z}\frac{\varphi_s-\varphi(v_s)}{e^{\int_t^s\rho dx}e^{\int_t^s a(v_x)dx}}dB_s^a\right).$$

Since $a(v_s)\ge 0$, this shows that

$$\begin{cases}\mathbf{E}_t(v_{t+z}(a,c,R)-v_{t+z})\ge e^{\rho z}(v_t(a,c,R)-v_t) & \text{if } v_{t+z}(a,c,R)-v_{t+z}>0\\ \mathbf{E}_t(v_{t+z}(a,c,R)-v_{t+z})\le e^{\rho z}(v_t(a,c,R)-v_t) & \text{if } v_{t+z}(a,c,R)-v_{t+z}<0\\ \mathbf{E}_t(v_{t+z}(a,c,R)-v_{t+z})=0 & \text{if } v_{t+z}(a,c,R)-v_{t+z}=0.\end{cases}$$

Since v and $v(a, c, R)$ are bounded, we must have $v_t = v_t(a, c, R)$ for all $t \geq 0$. Therefore, the agent receives $v_0 = v_0(a, c, R)$ from the contract. Next, we show that the principal also gets $m(v_0)$, which will follow once we establish that

$$F_t = \int_0^t \rho e^{-\rho s}\Big(e^{-\int_0^s a_z dz}(-c_s) + e^{-\int_0^s a_z dz} a_s(\Delta - R_s)\Big) ds + e^{-\rho t} e^{-\int_0^t a_z dz} m(v_t) \quad (2.52)$$

is a bounded martingale.

Using Ito's lemma we have

$$dF_t = e^{-\rho t}\rho\Big(e^{-\int_0^t a_z dz}(-c_t) + e^{-\int_0^t a_z dz} a_t(\Delta - R_t)\Big) dt + d(e^{-\rho t} e^{-\int_0^t a_z dz} m(v_t)).$$

Applying Ito's lemma one more time to the last term gives

$$d(e^{-\rho t} e^{-\int_0^t a_z dz} m(v_t)) = -(\rho + a_t) e^{-\rho t} e^{-\int_0^t a_z dz} m(v_t) dt + e^{-\rho t} e^{-\int_0^t a_z dz} d(m(v_t))$$

Next, applying Ito's lemma to $d(m(v_t))$ and using (2.11) from Proposition 2.2 to substitute for dv_t yield

$$dF_t = \Upsilon_t dt + \Big(e^{-\rho t} e^{-\int_0^t a_z dz} m'(v_t)\rho\sigma\varphi_t\Big) dB_t^a$$

with drift $\Upsilon_t = \rho e^{-\rho t} e^{-\int_0^t a_z dz} M_t$, where

$$\begin{aligned} M_t = {} & m''(v_t)\frac{\rho\sigma^2}{2}\left(g'(a) - \Lambda(R) + \frac{v}{\rho}\right)^2 \\ & + m'(v_t)\big(g(a_t) - u(c_t) - a_t\Lambda(R_t) + v_t(1 + \frac{a_t}{\rho})\big) \\ & - m(v_t)(1 + \frac{a_t}{\rho}) - c_t + a_t(\Delta - R_t) \end{aligned}$$

Note that the optimality condition of HJB equation (2.25) will imply that $\Upsilon_t = 0$. As a consequence, F_t is a martingale up to the retirement/termination time (i.e., the stopping time τ). Therefore, the principal's value of the (a, c, R) contract is

$$\begin{aligned} \mathbf{E}\, F_\tau &= \mathbf{E}\left[\int_0^\tau \rho e^{-\rho s}\big(e^{-\int_0^s a_z dz}(-c_s) + e^{-\int_0^s a_z dz} a_s(\Delta - R_s)\big) ds + e^{-\rho\tau} e^{-\int_0^\tau a_z dz} \lambda_r(v_\tau)\right] \\ &= F_0 \\ &= m(v_0), \end{aligned}$$

where the last equality follows by the result that F_t is a bounded martingale. To finish the proof, recall that $m(v_\tau) = \lambda_r(v_\tau)$. □

Proof of Theorem 2.7. In the proof we show that any incentive compatible (a, c, R) contract implies, at most, $m(v_0(a,c,R))$ payoff for the principal. To show it, we establish that for any incentive compatible contract (a,c,R), the principal continuation value F_t is a bounded super-martingale.

Let $v_t = v(a,c,R)$ be the agent's payoff (where v_t is the diffusion in (2.11)). The principal's payoff up to time t is given by

$$F_t = \int_0^t \rho e^{-\rho s}(e^{-\int_0^s a_z dz}(-c_s) + e^{-\int_0^s a_z dz} a_s(\Delta - R_s))ds + e^{-\rho t}e^{-\int_0^t a_z dz} m(v_t)$$

Using Ito's lemma, $dF_t = \Upsilon_t dt + (e^{-\rho t}e^{-\int_0^t a_z dz} m'(v_t)\rho\sigma\varphi_t)dB_t^a$ as in the proof of Proposition 2.6; the goal is to show $\Upsilon_t \leq 0$. When $a_t = 0$, the quantity $\rho e^{-\rho t}e^{-\int_0^t a_z dz}(\Upsilon_t\big|_{a_t=0})$ is equal to

$$\begin{aligned} m''(v_t)\frac{\rho\sigma^2}{2}\left(g'(0) - \Lambda(R) + \frac{v}{\rho}\right)^2 &+ m'(v_t)(-u(c_t) + v_t) - m(v_t) - c_t \\ &\leq (m'(v_t)(-u(c_t) + v_t) - m(v_t) - c_t) \\ &= (m'(v_t)(-u(c_t) + v_t) - m(v_t) + \lambda_r(u(c_t))) =: T_1 \end{aligned} \tag{2.53}$$

where the first inequality follows from concavity of $m(\cdot)$, i.e., $m''(v_t) \leq 0$, and the last equality is due to the definition $\lambda_r(x) = -u^{-1}(x)$. Moreover, by Proposition 2.5, part (i), $\lambda_r(v) \leq m(v)$, thus,

$$T_1 \leq \rho e^{-\rho t}e^{-\int_0^t a_z dz}\{m'(v_t)(-u(c_t) + v_t) - m(v_t) + m(u(c_t))\} \leq 0 \tag{2.54}$$

where the last inequality follows because $m(\cdot)$ is concave. Therefore, (2.53) and (2.54) shows that $\Upsilon_t \leq 0$ at $a_t = 0$. For $a_t > 0$, it is also true that $\Upsilon_t \leq 0$. This is due to Proposition 2.2 (see (2.17)) and the HJB equation (2.25) along with the concavity of $m(\cdot)$ (i.e., $m''(v) \leq 0$).

As consequence, F_t is a bounded supermartingale for any incentive compatible contract (a,c,R). This means that the principal's payoff from the (a,c,R) contract

$$\mathbf{E}^a\ F_\tau \leq F_0 = m(v_0),$$

as desired. □

Proof of Proposition 2.8. The proof follows from the HJB equation (2.25). To prove *(i)*, set $a(v) = 0$ in (2.25). Since the principal payoff is concave, i.e., $m''(v) \leq 0$, rearranging gives

$$0 \geq \min_{c\in[0,\bar{c}]} m(v) + c - m'(v)(v - u(c)) \tag{2.55}$$

For brevity, denote $W = u(c)$. Then

$$c = u^{-1}(W) = -\lambda_r(W) \tag{2.56}$$

where the last equality follows by the definition of $\lambda_r(\cdot)$. Also note that during the employment (before retirement), i.e., for any $W \in (0, v_{\max})$, we must have

$$\lambda_r(W) < m(V). \tag{2.57}$$

Now, (2.57) and (2.56) in (2.55) together imply

$$0 \geq \min_{c\in[0,\bar{c}]} m(v) - \lambda_r(W) - m'(v)(v - W) > \min_{c\in[0,\bar{c}]} m(v) - m(W) - m'(v)(v - W) \geq 0,$$

where the last inequality is due to the concavity $m(\cdot)$.

This is a contradiction for optimality in the HJB equation (2.25). Therefore, $a(v) > 0$ for all $v \in (0, v_{\max})$. The rest of the proof follows immediately from the HJB equation (2.25) as well. The first order optimality condition with respect to c gives

$$1 + u'(c)m'(v) = 0.$$

Since $m(\cdot)$ is concave over $(0, v_{\max})$ (i.e., during employment), $c(v) > 0$ when $m'(v) < 0$, which obtains when $v > v^*$ (note that $v^* = \arg\max_{v\in[0,v_{\max}]} m(v)$). Hence,

$$u'(c(v)) = -\frac{1}{m'(v)}.$$

Finally, since $u(\cdot)$ and $m(\cdot)$ are both concave, $c'(v) > 0$ for all $v \in (v^*, v_{\max})$. □

Chapter 3

Speed Competition and Segmentation in Illiquid Markets

Joint with Oğuzhan Çelebi and Ali Kakhbod

3.1 Introduction

In the last decade, there has been significant fragmentation and heterogeneity across trading venues, due to investments in trading infrastructure. These investments reduced latencies in order execution and communication for many different instruments.[1] Investors with different preferences regarding the speed and cost of trading an asset choose venues that serve their needs, while venues compete to attract these investors. The preferences of investors are inherently dynamic and depend on many factors such as their liquidity demand, need for portfolio rebalancing or hedging. How does an investor choose the appropriate venue for her preferences? How does this choice depend on dynamic nature of preferences? How do venues compete in order to attract the investors? What is the effect of differentiation in trading speeds on transaction fees charged by the venues, the trading volume and welfare?

To investigate these issues, we consider a dynamic model where traders with unit demand buy or sell a single security. Traders experience random shocks to their utility of holding the asset and engage in trade in one of the trading venues. They decide which venue to execute their trade depending on their valuation of the asset, transaction costs, and the trading speed of the two venues. In Theorem 3.3, we characterize the equilibrium market structure and show under what conditions there is zero (*no-trade*), one (*no-segmentation*) or two (*market segmentation*) active venues.

We are extremely grateful to Daron Acemoglu, Ricardo Caballero, Drew Fudenberg, Leonid Kogan, Andrey Malenko and Stephen Morris for their great comments and suggestions.

[1] See Pagnotta and Philippon (2018) for a detailed account of these evolutions and recent examples.

Given the structure of the trading equilibrium, we turn our attention to the trading volumes in the venues and show how they depend on the tax and the fees and speeds of the venues (Proposition 3.4). Then, to gain analytical tractability, we restrict attention to the case where the traders' values are distributed uniformly and analyze how venues with different trading speeds compete in fees (Proposition 3.8) and how their equilibrium trading fees depend on their speeds (Proposition 3.9). In particular, we show how differentiation affects competition: when the differentiation between two venues decreases (i.e. slower venue becomes faster or faster venue becomes slower), the trading fees in both venues decrease, whereas when the differentiation between two venues increases (i.e. slower venue becomes slower or faster venue becomes faster), the trading fees in both venues increase.

We then focus on the effect of transaction speed on the trading volume. First, a change in transaction speed of a venue affects the instantaneous trading volume in a venue directly. Second, it affects the fees charged by the venues in equilibrium, thus the market structure itself. We show that both affects are positive for the speed of the slower venue and the total trading volume is increasing in the speed of the slower venue (Proposition 3.10). Next, we show that the effect of an increase in the speed of the fast venue is ambiguous. However, we consider the special case of *full competition*, where the speed difference between the venues is arbitrarily small and show that whenever the rate of preference shock is greater than the trading speed, differentiation increases trading volume (Proposition 3.11). Moreover, if the trading speed is greater than the rate of preference shock, the effect of differentiation on trading volume depends on the discount factor of the traders. We characterize a threshold such that whenever the discount rate of the traders is below that threshold (i.e. the traders are sufficiently patient), differentiation increases trading volume. These result show how the effect of trading speed on trading volume depends on the characteristics of the traders in a market.

Next, we endogenize the entry of a new firm. To this end, we consider an entry model in which a new venue owner makes an entry decision and (conditional on entry) sets a new speed. In the linear cost case, we show that the new firm decides to enter whenever the cost is sufficiently low and the speed of the new venue is decreasing in the operating cost (Proposition 3.12). Then, we characterize a lower bound for the degree of differentiation between the venues (Proposition 3.16) and show that the speed of the new venue is decreasing in the transaction tax.

Fixing the speed of the slow venue, the rest of the paper discusses the entry game from a welfare perspective. We consider different notions of welfare: surplus from trade (Section 3.4.1), trading volume (Section 3.4.2), and trading revenue (Section 3.5). In each of these cases, we consider the optimal regulator's choice for taxing traders, and the resulting optimal choice of speed for the entrant. Specifically, we investigate when the profit maximizing speed is lower than the welfare maximizing speed. We conclude with a short extension of the entry game.

3.1.1 Literature Review

The competition feature of our trading model contributes to the literature on market design. The recent literature has a variety of focuses: optimal design of contests (e.g., Bimpikis, Ehsani and Mostagir (2015)), design of crowdfunding campaigns (e.g., Alaei, Malekian and Mostagir (2016)), inspection and information disclosure (e.g., Papanastasiou, Bimpikis and Savva (2018)), information diffusion in networks (e.g., Acemoglu, Ozdaglar and ParandehGheibi (2010), Ajorlou, Jadbabaie and Kakhbod (2018), Candogan and Drakopoulos (2019)), among others. In contrast to these important works, our paper particularly focuses on how competition design between multiple venue owners (dealers), affect the trading dynamics, trading volume, welfare, liquidity and speed segmentation in illiquid markets.

This work also contributes to the literature on liquidity in market microstructure. There is an earlier literature that analyses the liquidity demand side (e.g., Glosten and Milgrom (1985), Easley and O'Hara (1987), Admati and Pfleiderer (1988)).[2] In contrast to these works we consider how competition between dealers in illiquid markets affect endogenous transaction speeds and liquidity.

The trading framework of this paper is also related to the literature of the market efficiency in strategic informed trading, which dates back to Kyle (1985*a*,*b*)'s seminal articles. Wang (1993, 1994) consider an infinite-horizon model where competitive insiders receive information on a firm's dividends over time in steady-state. They show that risk-neutral competitive insiders will reveal their private information instantly whereas risk-aversion can reduce their trading aggressiveness, leading to a slower information revelation. Back and Pedersen (1998) consider a finite-horizon model with a monopolistic informed insider and show that the insider reveals her information gradually. Chau and Vayanos (2008) consider the market efficiency in an infinite-horizon model with a monopolistic insider trading with competitive dealers and noisy traders as well. They discover that the insider chooses to reveal her information quickly, as the market approaches continuous trading. Similar to these works we also present a fully dynamic trading model, however, in sharp contrast to these important papers, instead of asymmetric information, we focus on impacts of competition between market-makers (dealers, venue owners) on trading volume, tax, transaction costs and welfare.

This paper also contributes to the growing literature on dynamic trading in OTC markets. Duffie (2012) has an excellent review of studies about OTC markets. Recent literature has a variety of focuses. For example, Guerrieri and Shimer (2014) study adverse selection

[2] Most notably, Glosten and Milgrom (1985) analyze transaction prices arising from quotes of competing risk-neutral dealers who are making a market in a single security and facing both privately informed and uninformed traders. They show that, in the zero dealer profit equilibrium, private information induces a positive spread between bid and offer. Easley and O'Hara (1987) theorize that uninformed traders may refrain from trading when they perceive the presence of an informed trader, leading to diminished trading volume. Admati and Pfleiderer (1988) show that if liquidity traders (hedgers) can choose the timing of their transactions strategically, then in equilibrium their trading is relatively more concentrated in periods closer to the realization of their demands.

with search frictions and discrete trading opportunities, Babus and Parlatore (2017) study welfare effects of decentralized trading; Duffie, Gârleanu and Pedersen (2005) and Lagos and Rocheteau (2009) look at random search and matching in large markets with a continuum of traders, Kakhbod and Song (2020) consider how sequential trading with a large informed trader affects price discovery dynamics, Zhu (2014) shows how adding a dark pool improves market price discovery.[3] In contrast to these important papers we consider equilibrium asset prices and trading volumes in multiple OTC venues where speed choices are endogenous (and heterogeneous), and derive necessary and sufficient conditions ensuring segmentation in OTC markets. Finally, the speed competition feature of our model relates to the important work by Pagnotta and Philippon (2018). However, our model distinctly differs from theirs because in our paper, a trader's position between venues is not fixed and dynamically changes over time based on the extent of the trader's preference shocks. This key modeling difference directly affects our analysis for competition between multiple venues (dealers), endogenous market segmentation, transaction speeds and fees, trading volume, optimal regulator's choice for taxing traders and welfare in illiquid asset markets.

3.2 Model

We consider a continuous time model with a unit measure of traders with time-discount factor $\rho > 0$ and a long-lived indivisible asset with supply $Z \in (0, 1)$. Traders have unit demand; a trader who owns the asset is called *holder* and a trader who does not is called *non-holder*. Each trader has a personalized/intrinsic value $\eta \in [\eta_l, \eta_h]$ for holding the asset, which changes over time.

The instantaneous utility of a holder with value η is $u(\eta)$ while a non-holder gets zero instantaneous utility value (regardless of her intrinsic value). The value η changes, independently across the traders, with a Poisson shock with rate γ. Conditional on arrival of a shock, the new valuation is chosen from $[\eta_l, \eta_h]$ according to the CDF $F(\cdot)$ and the PDF $f(\cdot)$.

There are multiple *trading venues*, with different trading speeds and transaction fees (i.e., costs). At any given time, a holder decides to sell or hold the asset and a non-holder decides to buy the asset or do nothing. If a trader decides to sell or buy the asset, she then decides which venue to use.

[3] This paper also contributes to the growing literature on dealer-based OTC markets. Prior literature focuses on the dealers' ability to contract with customers (Grossman and Miller (1988)), discriminate based on order size (Seppi (1990)), welfare effects of decentralized trading (e.g. Malamud and Rostek (2017)), price movements in OTC markets when block orders are large (e.g. Grossman (1992)), searching for good price in OTC markets with multiple dealers (Zhu (2012)), random search and matching in large markets among a continuum of traders (Duffie, Gârleanu and Pedersen (2005); Vayanos and Weill (2008); Duffie, Malamud and Manso (2009); Lagos and Rocheteau (2009)) and adverse selection with search frictions and discrete trading opportunities (Guerrieri, Shimer and Wright (2010), Guerrieri and Shimer (2014)). In sharp contrast to these works our focus here is on the effect of competition between dealers on setting transaction fees and speeds and their consequences on trading volume, welfare and optimal taxes. See also the literature on how illiquidity may be market destabilizing, e.g., Bebchuk and Goldstein (2011); Goldstein (2012); Gorton and Ordonez (2014); Ahnert and Kakhbod (2017).

The venue v, $v \in \{s, f\}$, is characterized by its transaction speed σ_v and transaction fee c_v. We assume venue f is faster, i.e., $\sigma_f \geq \sigma_s$ and $c_f > c_s$. [4] In addition to the transaction costs, traders additionally pay an amount $\theta \geq 0$ per trade. The difference between θ and the transaction fees c_s and c_f arises since venues compete based on speed and costs: there, c_s and c_f are strategically chosen while θ is fixed (c.f. Section 3.3.3 and 3.3.4). We will consider different interpretations of θ in later sections: for example, θ is a tax charged by a regulatory body. In our model, θ is additive instead of multiplicative, a choice we make for a couple reasons. The first is realism: external agencies (i.e., parties who sets fees that are are not c_s and c_f) lack the means and resources to keep track of the prices of any one security during a trading day, especially in illiquid and segmented markets that we consider. As a result, both retail and professional brokers (e.g., Fidelity, Bloomberg IB) charge a per-transaction cost that constant across all securities in an asset class. Moreover, multiplicative taxes arbitrarily disincentivizes trading large stocks such as Amazon over smaller stocks such as Exxon Mobil, two companies with comparable liquidity with drastically different prices per share (Amazon is approximately one hundred times more expensive). The second reason is that our model studies a single asset, so that θ—along with other quantities—could all depend on the security. In this sense, it is simpler and almost without loss to consider additive taxes.

In Section 3.3.1, we analyze the decision of the traders for a given market structure (the trading speed and transaction fees) and characterize the trading equilibrium with two venues; Section 3.3.2 characterizes the trading volume under this trading equilibrium. In Section 3.3.3, we analyze how multiple venues compete in transaction fees in order to maximize their profit for given trading speeds. In Section 3.3.4, we endogenize entry and the trading speed in the new venue. In particular, we assume that there is an existing venue, which we refer as the *old venue*, with fixed trading speed. We then analyze the entry and speed choice of a second firm. In Section 3.4, we use our characterization of the market structure to analyze the profit maximizing speed choice of the new firm and compare it to welfare maximizing and trading volume maximizing alternatives. Finally, in Section 3.5, we consider a counter-party who sets the transaction tax to maximize its revenue analyze how the revenue maximizing tax rate depends on the primitives of the market.

3.3 Analysis

3.3.1 Optimal Trader Decisions

We first characterize the behavior of the traders as a function of speed and transaction fees of the venues. Let $m \in \{0, 1\}$ be a trader's position, where $m = 0$ denotes non-holder and $m = 1$ denotes a holder. At each time t, a holder may Sell (S) or Hold (H) and a non-holder may Buy

[4] Note that $\sigma_f \geq \sigma_s$ is without loss of generality. If $c_f > c_s$ does not hold, then in equilibrium, there is no demand in slower venue. When we endogenize the transaction fees, the slower venue always chooses a lower transaction fee in order to make positive revenue.

(B) or do Nothing (N). The action set of a trader is (H, B^{ν}, S^{ν}, N) where the superscripts denote the venue choices of traders.

We focus on the stationary equilibrium. Let p_{ν} denote the equilibrium asset price in venue $\nu \in \{s, f\}$. Let $V_{m,\eta}$ denote a trader's expected payoff whose current trading position and value are respectively $m \in \{0, 1\}$ and $\eta \in [\eta_l, \eta_h]$. To derive the Hamilton-Jacobi-Bellman (HJB) equations, depending on traders' actions, we define the following value functions

$$\rho V_H(\eta) = \underbrace{u(\eta)}_{\text{flow gain of holding the asset}} + \gamma \underbrace{(\mathbf{E}_{\eta'}[V_{1,\eta'}] - V_H(\eta))}_{\text{net gain of the pref. shock}} \tag{3.1}$$

$$\rho V_S^s(\eta) = u(\eta) + \gamma(\mathbf{E}_{\eta'}[V_{1,\eta'}] - V_S^s(\eta)) + \sigma_s \underbrace{(V_{0,\eta} + p_s - c_s - \theta - V_S^s(\eta))}_{\text{net gain of selling in the slow venue}} \tag{3.2}$$

$$\rho V_S^f(\eta) = u(\eta) + \gamma(\mathbf{E}_{\eta'}[V_{1,\eta'}] - V_S^f(\eta)) + \sigma_f \underbrace{(V_{0,\eta} + p_f - c_f - \theta - V_S^f(\eta))}_{\text{net gain of selling in the fast venue}}. \tag{3.3}$$

For a trader who currently owns the asset, $V_H(\eta)$ denotes his value if he decides to hold it, and $V_S^{\nu}(\eta), \nu \in \{s, f\}$ denotes his value if he decides to sell the asset in venue ν.[5] In a similar fashion, we define:

$$\rho V_N(\eta) = \gamma \underbrace{(\mathbf{E}_{\eta'}[V_{0,\eta'}] - V_N(\eta))}_{\text{net gain of the pref. shock}}, \tag{3.4}$$

$$\rho V_B^s(\eta) = \gamma(\mathbf{E}_{\eta'}[V_{0,\eta'}] - V_B^s(\eta)) + \sigma_s \underbrace{(V_{1,\eta} - p_s - c_s - \theta - V_B^s(\eta))}_{\text{net gain of buying in the slow venue}}, \tag{3.5}$$

$$\rho V_B^f(\eta) = \gamma(\mathbf{E}_{\eta'}[V_{0,\eta'}] - V_B^f(\eta)) + \sigma_f \underbrace{(V_{1,\eta} - p_f - c_f - \theta - V_B^f(\eta))}_{\text{net gain of buying in the fast venue}}. \tag{3.6}$$

For a trader who currently does not own the asset, $V_N(\eta)$ denotes his value if he decides to do nothing, $V_S^{\nu}(\eta), \nu \in \{s, f\}$ denotes his value if he decides to buy the asset in venue ν.[6]

Next, to derive the optimal action of a trader, we note that she has three options: if she is a holder, she chooses between holding the asset, selling it in the slower venue, or selling it in the faster venue. If she is a non-holder, she decides between doing nothing, buying the asset in the slow venue, or buying the asset in the fast venue. Recall that $V_{0,\eta}$ and $V_{1,\eta}$ denote the continuation values (i.e., the expected payoff) for a trader as function of her current

[5] Rearranging (3.1) shows that $V_H(\eta) = \frac{u(\eta)+\gamma\,\mathbf{E}_{\eta'}[V_{1,\eta'}]}{\gamma+\rho}$, $V_S^s(\eta) = \frac{u(\eta)+\sigma_s(V_{0,\eta}+p_s-c_s-\theta)+\gamma\,\mathbf{E}_{\eta'}[V_{1,\eta'}]}{\sigma_s+\gamma+\rho}$, and $V_S^f(\eta) = \frac{u(\eta)+\sigma_f(V_{0,\eta}+p_f-c_f-\theta)+\gamma\,\mathbf{E}_{\eta'}[V_{1,\eta'}]}{\sigma_f+\gamma+\rho}$.

[6] Similarly, (3.1) also shows that $V_N(\eta) = \frac{\gamma\,\mathbf{E}_{\eta'}[V_{0,\eta}]}{\gamma+\rho}$, $V_B^s(\eta) = \frac{\sigma_s(V_{1,\eta}-p_s-c_s-\theta)+\gamma\,\mathbf{E}_{\eta'}[V_{0,\eta'}]}{\sigma_s+\gamma+\rho}$, and $V_B^f(\eta) = \frac{\sigma_f(V_{1,\eta}-p_f-c_f-\theta)+\gamma\,\mathbf{E}_{\eta'}[V_{0,\eta'}]}{\sigma_f+\gamma+\rho}$.

position m and her value η. Given (3.1) and (3.4), we obtain

$$\begin{aligned} V_{0,\eta} &= \max\{V_N(\eta), V_B^s(\eta), V_B^f(\eta)\} \\ V_{1,\eta} &= \max\{V_H(\eta), V_S^s(\eta), V_S^f(\eta)\}. \end{aligned}$$

In order to specify the stationary equilibrium, we need to characterize stationary distribution of values η of holders and non-holders, whose densities we denote as $f_h(\eta)$ and $f_{nh}(\eta)$. We now formally define the stationary equilibrium.

Definition 3.1. *A stationary equilibrium consists of sets N, B_s, B_f, H, S_s, S_f and prices p_s and p_f such that:*

- $f_h(\eta) + f_{nh}(\eta) = f(\eta)$;
- *Traders behave optimally:*

$$\begin{aligned} N &= \{\eta \in [\eta_l, \eta_h] : V_N(\eta) &= \max\{V_B^s(\eta), V_B^f(\eta), V_N(\eta)\}\} \\ B_s &= \{\eta \in [\eta_l, \eta_h] : V_B^s(\eta) &= \max\{V_B^s(\eta), V_B^f(\eta), V_N(\eta)\}\} \\ B_f &= \{\eta \in [\eta_l, \eta_h] : V_B^f(\eta) &= \max\{V_B^s(\eta), V_B^f(\eta), V_N(\eta)\}\} \\ H &= \{\eta \in [\eta_l, \eta_h] : V_H(\eta) &= \max\{V_S^s(\eta), V_S^f(\eta), V_H(\eta)\}\} \\ S_s &= \{\eta \in [\eta_l, \eta_h] : V_S^s(\eta) &= \max\{V_S^s(\eta), V_S^f(\eta), V_H(\eta)\}\} \\ S_f &= \{\eta \in [\eta_l, \eta_h] : V_S^f(\eta) &= \max\{V_S^s(\eta), V_S^f(\eta), V_H(\eta)\}\}; \end{aligned}$$

- *asset market clears*

$$\int_{\eta_l}^{\eta_h} f_h(\eta) d\eta = Z; \tag{3.7}$$

- *fast venue clears;*

$$\int_{B_f} f_{nh}(\eta) d\eta = \int_{S_f} f_h(\eta) d\eta; \tag{3.8}$$

- *slow venue clears*

$$\int_{B_s} f_{nh}(\eta) d\eta = \int_{S_s} f_h(\eta) d\eta; \tag{3.9}$$

For the rest of the paper, we make following normalization:

Assumption 3.2. $\eta_l = 0$ *and* $u(0) = 0$.

We can now characterize the equilibrium. There are 3 main cases: *No trade* where no trader buys or sells the asset, *Market segmentation* where both venues are active, and *No segmentation* where only the slow venue is active. Intuitively, the first case obtains when the transaction fees and tax is prohibitively high so that trading is never profitable, and the third case obtains when the speed advantage of fast venue is small relative to the difference in transaction fees. In Theorem 3.3, we derive conditions which which each of these three cases occur and characterize the resulting equilibrium.

Theorem 3.3. *There are three possible regimes.*

(a) ***No Trade.*** *If $u(\eta_h) \leq 2(c_s + \theta)(\gamma + \rho)$, then there is no equilibrium where positive measure of traders trade.*

(b) ***Market Segmentation.*** *If*

$$u(\eta_h) > \max 2\frac{\sigma_f(\sigma_s + \gamma + \rho)(c_f + \theta) - \sigma_s(\sigma_f + \gamma + \rho)(c_s + \theta)}{\sigma_f - \sigma_s} \tag{3.10}$$

then a positive measure of traders trade in both venues. In particular, the traders' actions (depending on their value η) are uniquely characterized by the following intervals

$$N = [\eta_l, \eta_1],\ B_s = [\eta_1, \eta_2],\ B_f = [\eta_2, \eta_h],$$
$$S_f = [\eta_l, \eta_3],\ S_s = [\eta_3, \eta_4],\ H = [\eta_4, \eta_h],$$

where the equilibrium cutoffs η_1, η_2, η_3 and η_4 satisfy

$$\eta_l < \eta_3 < \eta_4 < \eta_1 < \eta_2 < \eta_h$$

and are uniquely pinned down by the following equations

$$(1 - Z)F(\eta_1) + ZF(\eta_4) = 1 - Z \tag{3.11}$$
$$(1 - Z)F(\eta_2) + ZF(\eta_3) = 1 - Z \tag{3.12}$$

and

$$\frac{u(\eta_1) - u(\eta_4)}{\gamma + \rho} = 2(c_s + \theta) \tag{3.13}$$
$$\frac{u(\eta_2) - u(\eta_3)}{\gamma + \rho} = 2\frac{\sigma_f(c_f + \theta)(\sigma_s + \gamma + \rho) - \sigma_s(c_s + \theta)(\sigma_f + \gamma + \rho)}{(\sigma_f - \sigma_s)(\gamma + \rho)}. \tag{3.14}$$

(c) ***No Segmentation.*** *If $u(\eta_h) > 2(c_s + \theta)(\gamma + \rho)$ and*

$$u(\eta_h) \leq 2\frac{\sigma_f(\sigma_s + \gamma + \rho)(c_f + \theta) - \sigma_s(\sigma_f + \gamma + \rho)(c_s + \theta)}{\sigma_f - \sigma_s},$$

then there is no segmentation and traders only trade in the slow venue. The equilibrium is characterized by cut-offs $\eta_l = \eta_3 < \eta_4 < \eta_1 < \eta_2 = \eta_h$ where

$$N = [\eta_l, \eta_1], \quad B_s = [\eta_1, \eta_h], \quad S_s = [\eta_l, \eta_4], \quad H = [\eta_4, \eta_h],$$

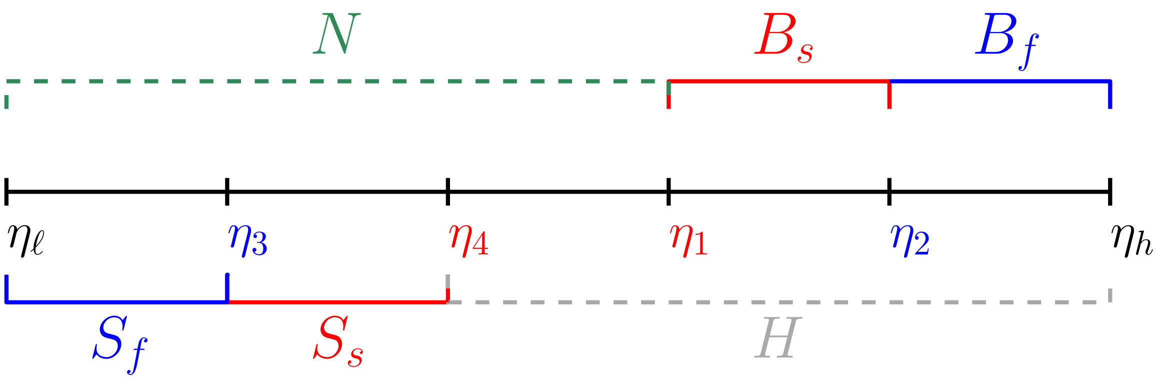

Figure 3-1: Threshold values η_1, ..., η_4 appearing in Theorem 3.3. The symbols S_f and S_s denote types who will sell in the fast and slow venue respectively; B_f and B_s are those will buy in the fast and slow venue, respectively. Types in N will not buy and types in H will hold (i.e., not sell).

and the cut-offs η_1 and η_4 are uniquely pinned down by

$$(1-Z)F(\eta_1) + ZF(\eta_4) = 1-Z \tag{3.15}$$

$$\frac{u(\eta_1) - u(\eta_4)}{\gamma+\rho} = 2(c_s+\theta). \tag{3.16}$$

The theorem characterizes the cutoffs in terms of $u(\eta_h)$; see Figure 3-1 for an illustration when market is segmented.

Theorem 3.3 is the building block for our analysis. Before we proceed with the analysis, we discuss the conditions that give rise to the three main cases. Note that

$$u(\eta_h) = (V_H(\eta_h) - V_H(0)) \cdot (\gamma+\rho) \tag{3.17}$$

This quantity corresponds to the value of a transaction between traders with values η_h and 0. Thus, $u(\eta_h)$ is a measure of the maximum value of a transaction, obtained by a trader with the highest possible valuation η_h buys from a trader with the lowest possible valuation $\eta_l = 0$. Intuitively, if the slow venue is too costly for traders with types η_h and η_l to trade, that is the case for all other traders and there is no trade in any equilibrium. In particular, whenever $2(\gamma+\rho)(c_s+\theta) \geq u(\eta_h)$, the fee of trading is very high compared to the flow payoff of the asset and no trader is willing to trade. It is instructive to express the above condition as:

$$2(c_s+\theta) > V_H(\eta_h) - V_H(0) \tag{3.18}$$

In this form, the equation simply says that total transaction cost (fees and tax) is higher than the value of the most profitable transaction, so there does not exists a price that makes a positive measure of traders in both sides of the market willing to trade.

Another observation is that whenever there is trade, the slow venue is always active. The reason behind this observation is simple: whenever a trader is indifferent between trading fast and holding (or doing nothing), he breaks even when the trade happens. However, as slow venue is cheaper than the fast venue, if that trader trades in the slow venue, he pays a lower transaction fee, thus strictly prefers that outcome to trading in the fast venue or holding/doing nothing. Recall that there is positive trade in the fast venue whenever following condition holds

$$V_H(\eta_h) - V_H(0) > \frac{\sigma_f(\sigma_s + \gamma + \rho)(c_f + \theta) - \sigma_s(\sigma_f + \gamma + \rho)(c_s + \theta)}{(\sigma_f - \sigma_s)(\gamma + \rho)} \tag{3.19}$$

As $c_f > c_s$, the numerator is always positive and bounded away from zero, while denominator goes to zero as the speed difference between the venues vanishes. Thus, the existence of trading in the fast venue depends on the speed advantage of the fast venue and the difference in transaction fees.

Equation 3.13 shows the tradeoff between trading on the slow venue versus no trading. Rewriting that equation, we obtain $V_H(\eta_1) - V_H(\eta_4) = 2(c_s + \theta)$, which relates that the value of a transaction between the cutoff types of the slow venue. Intuitively, the value of a transaction between these two types must be equal to the total transaction cost in equilibrium.

Theorem 3.3 characterizes η_i in terms of σ_f σ_s, c_f, and c_s. We suppress the dependence of $\eta_i(\sigma_f, \sigma_s, c_f, c_s)$ to these parameters to simplify the notation.

Illustration of the Market Structure

In this section we discuss how the equilibrium market structure changes as a function of the model parameters; here, the term 'market structure' refers to the state of market segmentation (i.e., which of the three cases obtains) and the corresponding cutoff thresholds as defined in Theorem 3.3. For simplicity (and to align with the assumptions for later results), we assume that $\eta \sim \textbf{Unif}[\eta_l, \eta_h] = \textbf{Unif}[0, 1]$ is uniformly distributed and the utility function $u(\eta) = \eta$ is linear. These assumptions allow us to derive analytic solutions to the thresholds $\eta_1, \ldots, \eta_4$.

For a generic set of parameters, the thresholds behave as in Figure 3-2, where we plot the four thresholds as a function of a single varying parameter (θ in this case). Recall that $0 \leq \eta_3 < \eta_4 < \eta_1 < \eta_2 \leq 1$: thus, the bottom series is η_3, the middle two series are η_4 and η_1, and the top series is η_2. The thresholds associated with the fast venue—η_3 and η_2—are plotted in blue and the thresholds associated with the slow venue are red.

With $\eta_l = 0$ and $\eta_h = 1$, the types $\eta \in [0, \eta_3] \cup [\eta_2, 1]$ engage in the fast venue. In other words, when $\eta_3 > 0$ and $\eta_2 < 1$, the market is segmented. This is the case in Figure 3-2 when the tax θ is less than 1 (approximately). Once the tax is sufficiently high, the market is no longer segmented: all traders engage only the slow venue. Therefore, a "corner solution" for the thresholds obtains: $\eta_3 = 0$ and $\eta_2 = 1$, and the fast venue disappears. When the tax

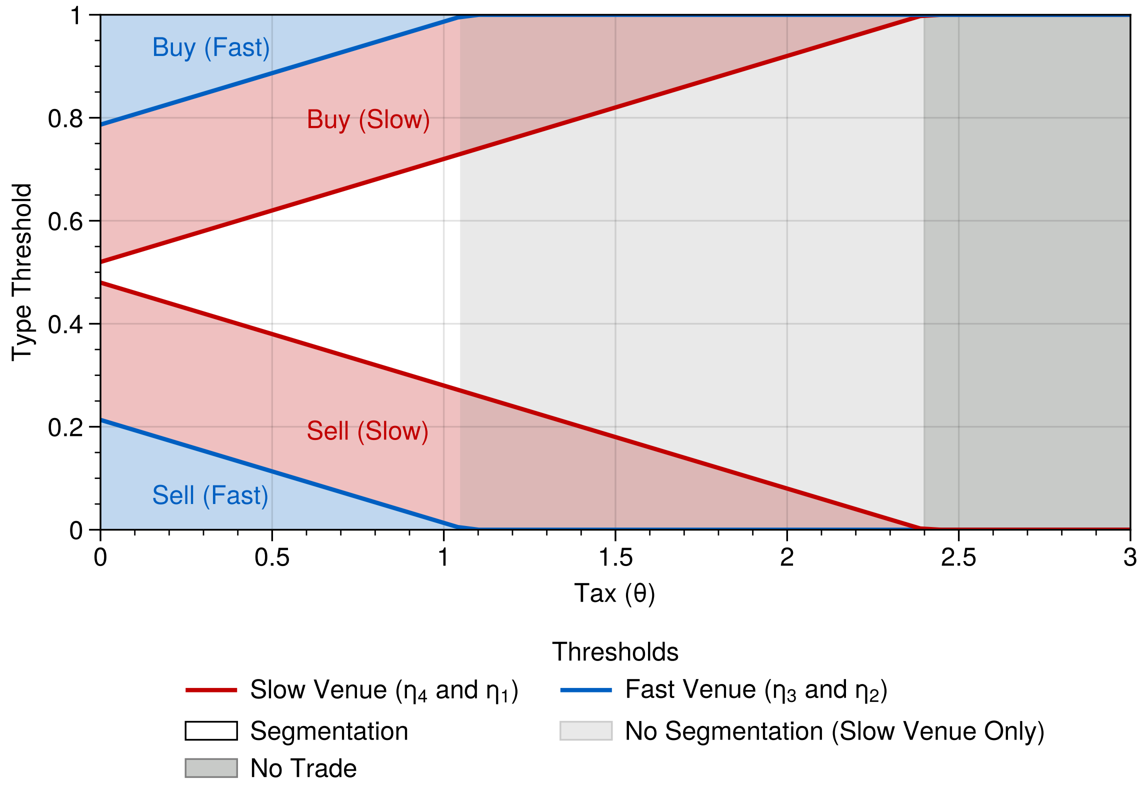

Figure 3-2: Type thresholds $\eta_1, \ldots, \eta_4$ as in Theorem 3.3 as a function of the tax parameter θ. The other parameters are $Z = \frac{1}{2}$; $\rho = \gamma = 0.1$; $c_s = 0.1$ and $\sigma_s = 1$; and $c_f = 0.3$ and $\sigma_f = 10$.

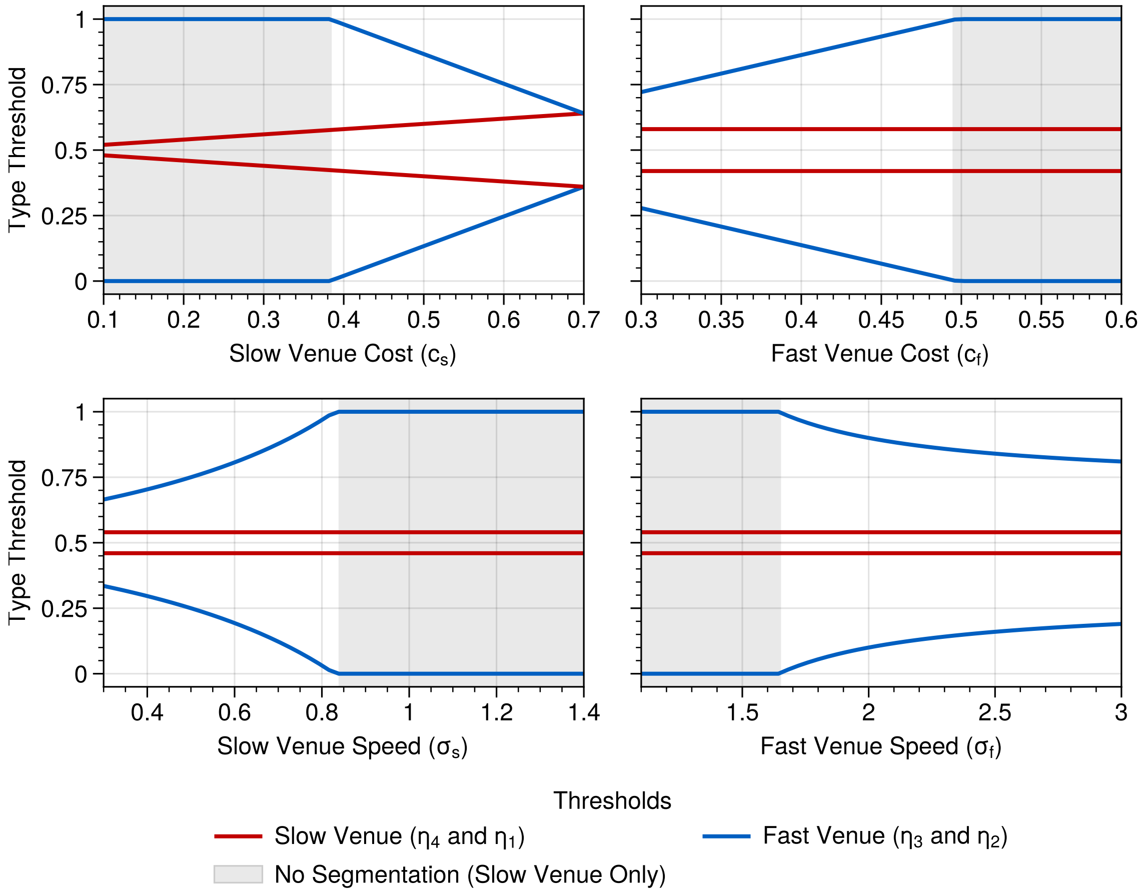

Figure 3-3: Comparative statics for Theorem 3.3. The type thresholds (y-axis) are plotted as a function of changing one single parameter at a time (x-axis); the parameters of interest are the transaction fees and speed offered by the slow and fast venues. An analogous figure for θ is shown in Figure 3-2. As usual, the thresholds are ordered: $\eta_3 < \eta_4 < \eta_1 < \eta_2$; thresholds pertaining to the the fast venue are in blue and thresholds pertaining to the slow venue are in red.

is even higher the no trade condition in Theorem 3-2 obtains and no traders engage in any trading. In the figure, we see that even the slow venue thresholds hit their corner solutions. Thus, the figure showcases each of the three cases in Theorem 3.3.

Comparative Statics of Market Structure

In the previous figure, we illustrated the change in market structure as θ (the external "tax") increases. In general, the relevant exogenous parameters in Theorem 3.3 are (in addition to θ) the transaction fees c_s and c_f; and the speeds σ_s and σ_f. Figure 3-3 illustrates the type thresholds as a function of these parameters.

The behavior of the market structure with respect to each of the parameters is intuitive. Take for example the slow venue transaction fee c_s; as it increases (and holding all other parameters constant), it makes the slow venue less attractive and the fast venue more attractive. For low enough c_s, there is no segmentation as the fast venue is simply too expensive. Note that this is true even when the speed offered by the fast venue is overwhelmingly large; according to Theorem 3.3 segmentation only occurs when

$$u(\eta_h) = u(1) = 1 > 2\frac{\sigma_f(\sigma_s + \gamma + \rho)(c_f + \theta) - \sigma_s(\sigma_f + \gamma + \rho)(c_s + \theta)}{\sigma_f - \sigma_s}. \tag{3.20}$$

Holding all parameters constant (say, with $\sigma_s = 1$) and taking $\sigma_f \to \infty$, the expression on the right hand side approaches $c_f - c_s$. The upshot is market segmentation does not obtain if $c_f - c_s$ is large, no matter how large the speed advantage. This is what we see in the top left panel of Figure 3-3, where $\sigma_f = 10$ is ten times larger than $\sigma_s = 1$. As c_s increases, we see that an increasingly smaller set of types trade in the slow venue, as is expected. Once c_s is large enough, the fast venue is competitive enough that the "extreme types" (i.e., holders whose types η are close to zero or non-holders whose types are close to one) leave the slow venue and begin trading in the fast venue. As c_s increases past this threshold, more and more types engage in trading in the fast venue, and at the limit $c_s = c_f$ the slow venue disappears and all trading occur in the fast venue. In terms of the thresholds, this corner case corresponds to $\eta_3 = \eta_4$ and $\eta_1 = \eta_2$, as depicted in the right edge of the top-left panel in Figure 3-3.

The behavior of the thresholds as the fast venue transaction fee c_f changes is also intuitive; as c_f increases, fewer types engage in the fast venue (i.e., η_3 decreases and η_2 increases). When the fee is sufficiently high, all traders use the slow venue and the market no longer exhibits segmentation. In contrast to changing c_s, changing c_f does not affect the cutoffs for types of traders who engage in the slow venue. Intuitively, the reason is that traders who are on the cutoff are actually indifferent between trading in the slow venue versus not trading at all: the alternative option of trading on the fast venue is not being considered since transaction fee concerns dominate speed concerns. Therefore, these traders remain marginal even as conditions in the fast venue change. On the other hand, there are no marginal traders who are not affected by a change in the slow venue transaction fee: if c_s increases, the scale is tipped in favor of the fast venue for marginal traders between fast and slow venues; similarly, the scale is tipped in favor of not trading for marginal traders between trading slowly and not trading.

Finally, increasing the slow venue speed has a similar effect as increasing the fast venue transaction fee in that the fast venue becomes less attractive. Interestingly, the thresholds for marginal marginal traders using the slow venue does not change: intuitively, they care only about the cost of trading, as explained above. Increasing the speed of fast venue has the opposite effect, though we see that it exhibits "diminishing returns" in that even an infinite speed advantage will not allow the fast venue to capture certain types of traders; see the discussion after equation (3.20) above.

3.3.2 Trading Volume

In this section, we characterize the trading volume for a given market structure. Let f_h and f_{nh} denote the equilibrium densities of the holders of the asset and non-holders of the asset.[7] The measure of traders in each venue is given by the following equations:

$$\begin{aligned} m_f(\sigma_s, \sigma_f, c_s, c_f, \theta) &= \int_{\eta_l}^{\eta_3} f_h(\eta) d\eta + \int_{\eta_2}^{\eta_h} f_{nh}(\eta) d\eta \\ &= F(\eta_3) \frac{\gamma Z}{\gamma + \sigma_f} + (1 - F(\eta_2)) \frac{\gamma(1-Z)}{\gamma + \sigma_f} \end{aligned}$$

and

$$\begin{aligned} m_s(\sigma_s, \sigma_f, c_s, c_f, \theta) &= \int_{\eta_3}^{\eta_4} f_h(\eta) d\eta + \int_{\eta_1}^{\eta_2} f_{nh}(\eta) d\eta \\ &= (F(\eta_4) - F(\eta_3)) \frac{\gamma Z}{\gamma + \sigma_s} + (F(\eta_2) - F(\eta_1)) \frac{\gamma(1-Z)}{\gamma + \sigma_s} \end{aligned}$$

The **trading volume** in slow and fast venues are given by $\text{TV}_s = \sigma_s m_s$ and $\text{TV}_f = \sigma_f m_f$, respectively. We define the total trading volume as $\text{TV} = \text{TV}_s + \text{TV}_f$. Following proposition shows how trading volume in venues depends on prices and speeds :[8]

Proposition 3.4. *The trading volume in the fast venue is increasing in c_s, σ_f and decreasing in c_f, σ_s, θ. The trading volume in the slow venue is increasing in c_f, σ_s, decreasing in c_s, σ_f. The total trading volume is decreasing in θ.*

As expected, the trading volume in a venue is increasing in the trading speed of that venue and transaction fee of the other venue, while it is decreasing in the trading speed of the other venue and transaction fee in that venue. Increasing the transaction fee reduces the trading volume while increases the fee charged per transaction, which is the main trade-off for the firms when they compete in fees. In the next section, we allow firms to compete by setting c_s and c_f to maximize their revenues and analyze the effect of competition has on trading volume.

Illustrations

We illustrate the results of Proposition 3.4 in the following figures; as before, the functional parameters are set to $\eta \sim$ **Unif**$[0, 1]$ and $u(\eta) = \eta$. Figure 3-4 plots the trading volumes in the two venues as the transaction fees in the two venues change and confirms part of the statement in the proposition. The figure also reveals that (at least for this set of functional parameters), the total trading volume decreases in the slow venue fee c_s; in other words, the

[7] The closed form expressions for f_h and f_{nh} and the derivation of the measure of traders are provided in the appendix.

[8] Full set of comparative statics of cut-offs and measure of traders are provided in the appendix.

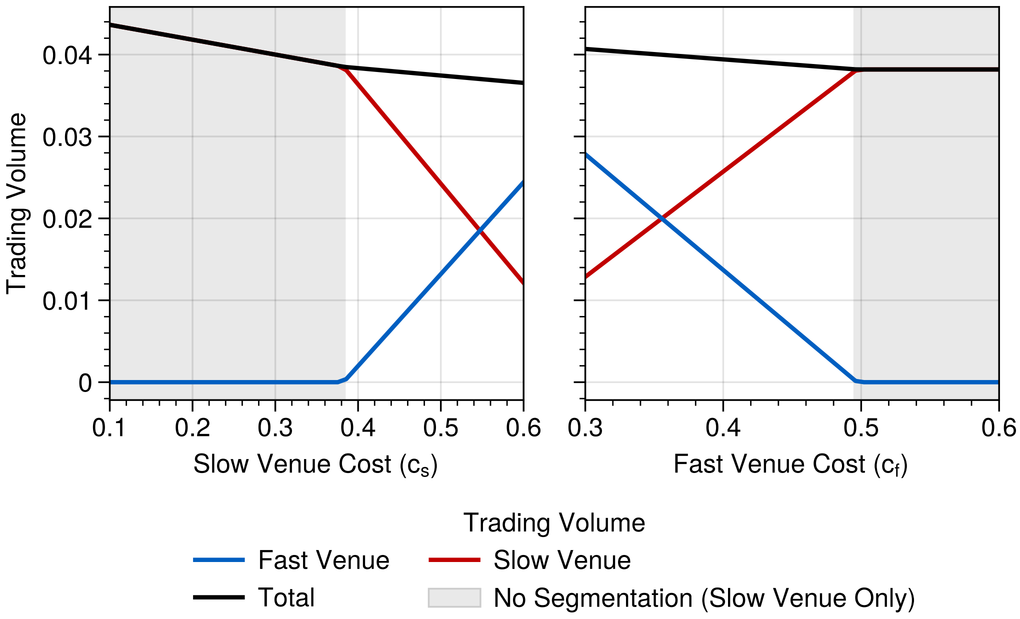

Figure 3-4: Trading volume in the two venues as a function of transaction fees c_s and c_f. The gray region indicates cost regimes in which only the slow venue is active. In particular, whether the fees exceed the threshold values needed for segmentation is the primary driver of the trading volume in either venue.

increase in trading volume in the fast venue does not compensate for the loss in the slow trading venue.

This is in some sense surprising: if it were the case that traders simply migrated from the slow to the fast venue as fees goes up, then total trading volume would increase due to the net increase in speed (i.e., recall that trading volume is speed times the measure of traders). We know, however, that this is not the case—c.f. the top left panel Figure 3-3, where increasing c_s will induce some traders to stop trading. Figure 3-4 shows that the effect carries over even after trading speed is taken into account: total trading volume is decreased.

The behavior when changing the fast venue fee is similar; note however, that after trading in the fast venue ceases, increasing c_f has no more effect on the market structure, so that $\partial \operatorname{TV} / \partial c_f = 0$ for sufficiently large c_f.

Next, consider the effect of increasing the tax θ as show in Figure 3-5; as predicted, it is decreasing. In the case of linear functional parameters $F(\eta)$ and $u(\eta)$, more could be shown. When θ is low enough such that both venues are active (c.f. Theorem 3.3), increasing θ will only decrease trading volume in the fast venue while keeping the slow venue trading volume constant. Note that the set of types engaging in either venue decreases per Figure 3-2; the interpretation is that the inflow of traders from the fast to the slow venue exactly balances out the flow out of the slow venue (flow being measured by speed times measure of traders). Only when the tax is high enough and trading ceases in the fast venue does the trading volume in the slow venue begin to decrease, as there is no inflow to make up for the outflow

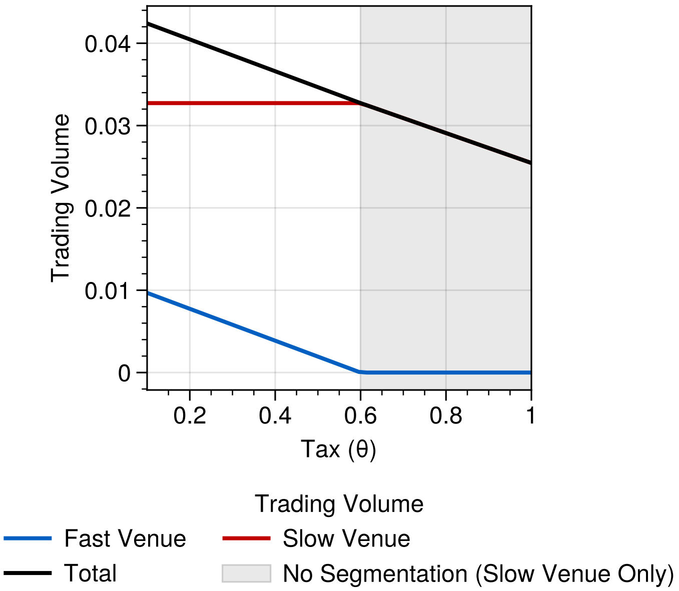

Figure 3-5: Trading volume as a function of the tax θ. The gray region indicates the threshold above which only the slow venue is active. For our linear model, increasing θ has a "layered" effect: as θ increases, traders from the fast venue switch to the slow venue until the fast venue is inactive. From that point, further increases drive traders away from the slow venue to become non-traders.

of traders.

The final set of figures show the effect of increasing venue speed (σ_s and σ_f). It agrees with the proposition that, e.g., $\partial\,\mathrm{TV}_s\,/\partial\sigma_s > 0$, and likewise for the other partial derivatives. Note however, that total trading volume is always increasing in the venue speed.[9] Intuitively, better "technology" (i.e., faster speeds) induces more trading; the results in the next section characterize this behavior in greater detail.

3.3.3 Fee Competition

Having characterized trading volume, we turn next to analyzing the competition between the two venues, starting with transaction fees. The revenues of fast and slow venues for given transaction fees are given by the following expressions.

$$R_f(\sigma_f, c_f, \sigma_s, c_s) = \sigma_f m_f(\sigma_s, \sigma_f, c_s, c_f)c_f \tag{3.21}$$

$$R_s(\sigma_f, c_f, \sigma_s, c_s) = \sigma_s m_s(\sigma_s, \sigma_f, c_s, c_f)c_s \tag{3.22}$$

[9] If σ_f is sufficiently small , then increasing σ_f by a small enough amount will not affect the market structure, in which case $\partial\,\mathrm{TV}_s\,/\partial\sigma_f = \partial\,\mathrm{TV}_f\,/\partial\sigma_f = 0$.

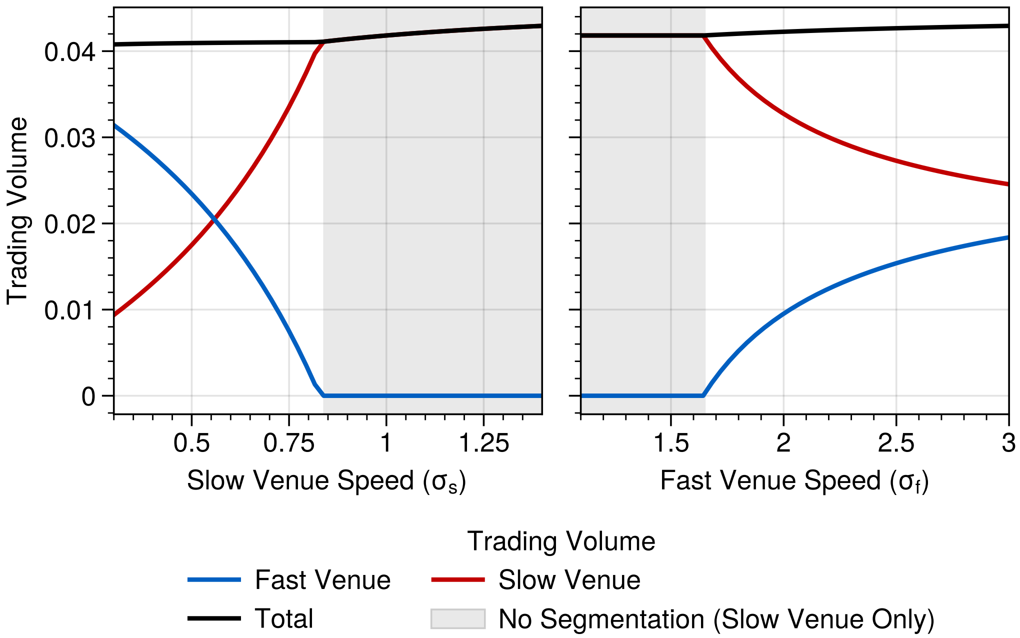

Figure 3-6: Trading volume as a function of venue speeds. The gray region indicates the thresholds beyond which only the slow venue is active.

For the rest of the paper, we keep the following assumption, which guarantees that the tax imposed by government (or external party) is not so high as to prohibit trade *a priori*

Assumption 3.5. $1 > 2\theta(\gamma + \rho)$

An equilibrium is a set of fees c_s, c_f such that $c_s \in \arg\max R_s(\sigma_f, c_f, \sigma_s, c_s)$ $c_f \in \arg\max R_f(\sigma_f, c_f, \sigma_s, c_s)$. For the rest of this section, we keep following assumptions that make the analysis tractable:

Assumption 3.6. *F is uniform over* $[0, 1]$

Assumption 3.7. $u(\eta) = \eta$

Following proposition characterizes the equilibrium prices of two competing venues.

Proposition 3.8. *Under assumptions 3.6 and 3.7, there is a unique equilibrium. The fees in the fast and slow markets are denoted by:*

$$c_f^*(\sigma_f, \sigma_s) = (1 - 2\theta(\gamma + \rho)) \frac{\sigma_f - \sigma_s}{\gamma(4\sigma_f - \sigma_s) + \rho(4\sigma_f - \sigma_s) + 3\sigma_f\sigma_s} \tag{3.23}$$

$$c_s^*(\sigma_f, \sigma_s) = \frac{c_f^*}{2} \tag{3.24}$$

Consequently,

$$\lim_{\sigma_s \to \sigma_f} c_s^*(\sigma_f, \sigma_s) = \lim_{\sigma_s \to \sigma_f} c_f^*(\sigma_f, \sigma_s) = 0.$$

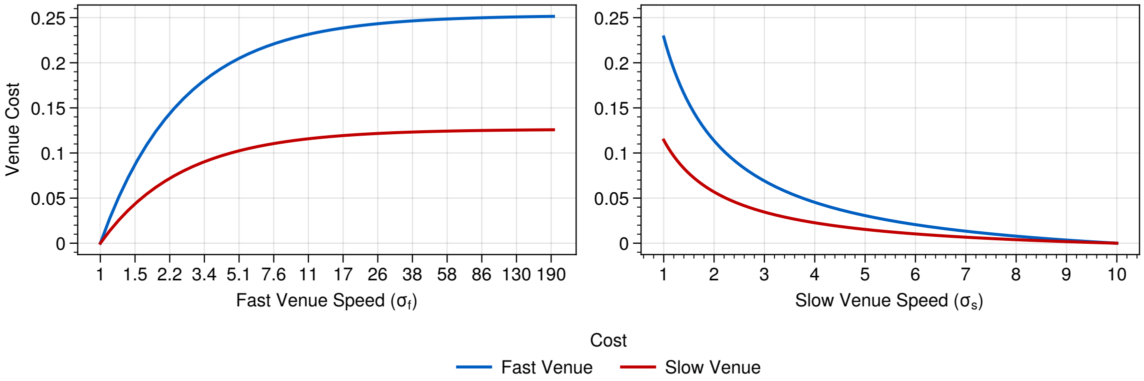

Figure 3-7: Optimal transaction fees (cost) as a function of venue speeds.

Proposition 3.8 characterizes how the venues compete to attract traders. The slower venue always undercuts the fast venue, as otherwise no trader will choose it and the venue would make zero revenue. The slow venue attracts speed-insensitive traders, while the fast venue sets a higher transaction fee and attracts speed sensitive traders. More precisely, for any fixed set of exogenous parameters Z, γ, ρ, θ, the fees c_s^* and c_f^* induced by any pair of speeds $0 < \sigma_s < \sigma_f$ will lead to a segmented market. In other words, the segmentation condition in Theorem 3.3 is satisfied *a posteriori* when the costs are optimally chosen given the other parameters. This captures the "obvious" fact when costs are endogenous, the venues would never set noncompetitive speeds—hence, both venues are active.

Corollary. *Fix parameters $Z \in (0,1)$, $\gamma > 0$, $\rho > 0$, and $\theta \geq 0$. For any pairs of venue speeds satisfying $0 < \sigma_s < \sigma_f$, the segmentation condition*

$$1 = u(\eta_h) = u(1) > \frac{\sigma_f(\sigma_s + \gamma + \rho)(c_f^* + \theta) - \sigma_s(\sigma_f + \gamma + \rho)(c_s^* + \theta)}{\sigma_f - \sigma_s} \tag{3.25}$$

is satisfied, where c_s^* and c_f^* are given by Proposition 3.8.

The following proposition shows how the competition is affected by the speed in each venue.

Proposition 3.9. *Transaction fees c_s and c_f are increasing in σ_f and decreasing in σ_s.*

This result shows the importance of differentiation. When the fast venue becomes faster, the level of differentiation between the venues increases and the effect of competition decreases. This results in higher fees across venues. Conversely, when the slow venue becomes faster, the differentiation between the venues decreases and the effect of competition increases, which result in lower fees. As $\sigma_s \to \sigma_f$, i.e. the venues become similar, the effect of Bertrand competition drives down the transaction fees, hence the revenues goes to zero.

The results above are illustrated in Figure 3-7, where c_s^* and c_f^* are plotted as functions of σ_s and σ_f; the left panel uses $\sigma_s = 1$ and moves σ_f to a very large value to visualize the

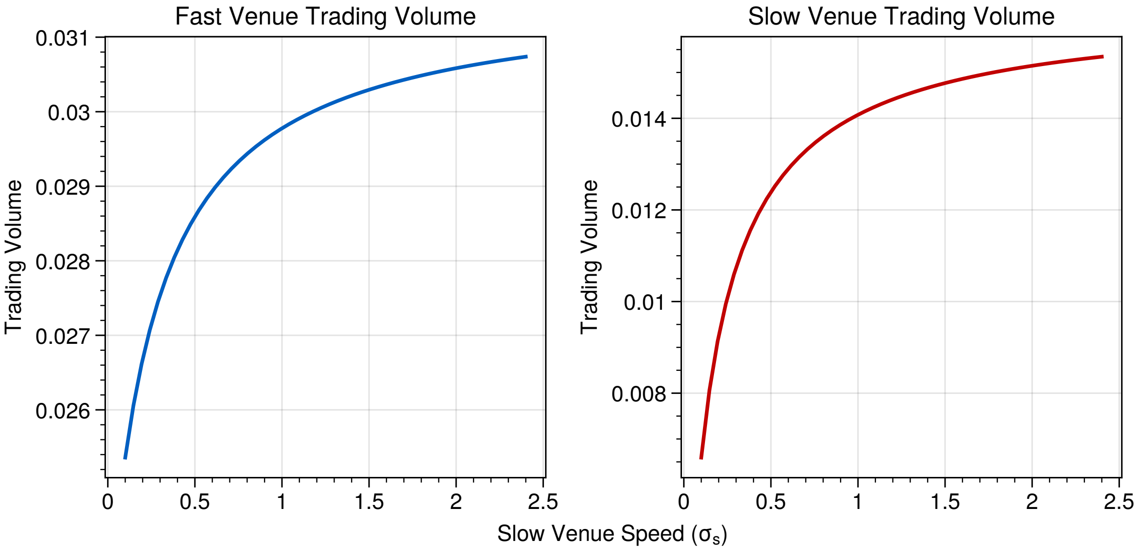

Figure 3-8: Trading volumes in both venues increase as σ_s increases.

behavior when $\sigma_f \to \infty$. As in Proposition 3.8, the transaction fees are driven to zero as σ_s approaches σ_f. Moreover, the function c_f^* (and hence c_s^*) is concave in σ_f: increasing σ_f has diminishing effects on the optimal fee. The intuitive explanation is that that σ_f itself has "diminishing effects" on the market structure as $\sigma_f \to \infty$, and we have

$$\lim_{\sigma_f \to \infty} c_f^*(\sigma_f, \sigma_s = 1) = \frac{1 - 2\theta(\gamma + \rho)}{4(\gamma + \rho) + 3}. \tag{3.26}$$

For the same reason, fees are convex (decreasing) in the slow venue speed σ_s.

Another interesting implication is that traders with different valuations are affected differently from the changes in speeds. For example, an increase in σ_s reduces the transaction fees, so makes everyone better off. On the other hand, an increase in σ_f increases the transaction fees and makes everyone but the traders with most extreme valuations worse off.

Lastly, we analyze the effect of the speed on trading volume. Next proposition shows the effect of the speed of the slow venue

Proposition 3.10. *The trading volume in both venues are increasing in σ_s.*

Increasing σ_s has two effects: first, for given choices of traders, it increases the trading speed in slow venue, which increases the trading volume. Second, it makes slow venue more competitive. Competition lowers the trading fees, which also increases the trading volume. Thus, a speed improvement in slow venue causes an increase in trading volume.

Figure 3-8 plots the trading volumes in either venue as a function of σ_s. Note that (in addition to being increasing) trading volume is a concave function of σ_s, so that increasing σ_s has a diminishing effect.

In fact, the two components of the competition channel driving up the trading volume

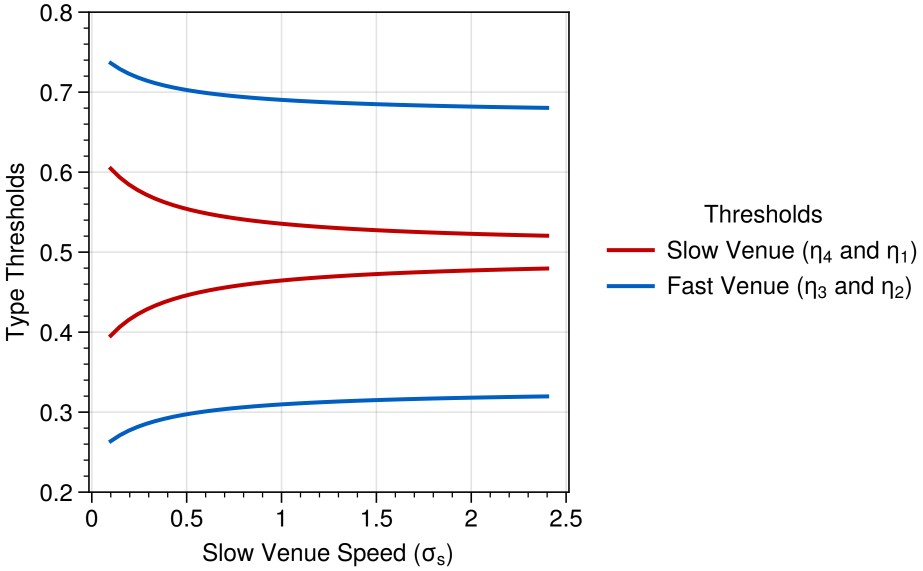

Figure 3-9: Increasing σ_s has diminishing impact on type thresholds.

exhibit diminishing returns. We have already seen in the right panel of Figure 3-7 shows that increasing σ_s^* as diminishing effects on lowering costs. Moreover, Figure 3-9 shows that the impact of competition—namely, inducing former non-traders to trade on the slow venue and former slow venue traders to trade on the fast venue—is also diminishing. The diminishing effect is captured by the thresholds η_1, ..., η_4 being concave; η_3 and η_4 are decreasing, while η_1 and η_2 are increasing. Recall that exogenously, σ_s has no effect on slow venue thresholds η_1 and η_4 (c.f., Figure 3-3); it is through the effect of σ_s on c_s and c_f that induces the change in market structure.

The effect of speed in the fast venue is more complicated. On the one hand it increases the transaction rate in the fast venue, which increases the trading volume. However, it also increases differentiation, thus increases the transaction fees charged in equilibrium. This reduces the incentives of the traders to trade and pay the fee, which reduces the trading volume. In general, the effect of σ_f is ambiguous. We consider an instructive special case, $\sigma_s \to \sigma_f$, which we refer as *full competition benchmark*. Following proposition characterizes the effect of an increase in σ_f when two venues start from the same speed level.

Proposition 3.11. *Let $\sigma_s \to \sigma_f = \sigma$. Then:*

- *If $\gamma > \sigma$, then $\frac{\partial TV}{\partial \sigma_f} > 0$*
- *If $\gamma < \sigma$, then $\frac{\partial TV}{\partial \sigma_f} > 0$ if and only if $\rho < \frac{\gamma^2+\gamma\sigma}{\sigma-\gamma}$*

The proposition shows that, first, if $\sigma < \gamma$, i.e. the trading speed is slow relative to the frequency of preference shock, a speed increase in the fast venue causes an increase in trading volume. This shows that the effect of higher transaction speed dominates the effect of higher fees. Second, if $\gamma < \sigma$, i.e. the trading speed is fast relative to the frequency of preference shock, then the discount factor of the traders become important. In particular, the trading volume is increasing if they are patient relative to the frequency of the preference shock. The reason is that, when the traders are patient and preference shock is not very frequent, the gains from buying the asset quickly is high, so the increased trading speed is more important for the traders compared to the increased transaction costs due to differentiation.

If the traders are impatient, then they are discouraged from trading due to rising transaction fees. In that case, the benefit from buying the asset faster is dominated by the increased prices due to differentiation and trading volume declines. This shows the importance of competition when there are multiple venues: the fast venue becomes faster, but the trading volume decreases due to decreased competition and higher fees.

Following corollary shows how these forces interact if the trading speed is fast:

Corollary. *If $\sigma \to \infty$, then $\frac{\partial TV}{\partial \sigma_f} > 0$ if and only if $\rho < \gamma$*

If the transaction speed is very fast, then the trade-off depends on the comparison between ρ and γ. If $\rho > \gamma$, then there is more benefit from faster transaction as the traders will obtain a higher utility from owning the asset before a preference shock comes. This makes them continue trading under higher fees and increase trading volume. On the other hand, if $\rho < \gamma$, then effect of increased transaction fees dominate faster trading speed and trading volume decreases.

3.3.4 Entry and Speed Choice

Until now, we have abstracted away from potential costs associated with running the markets and the speed choice of the venues. In this section, we analyze the entry decision of a new venue when there is an old venue with fixed trading speed and the entry decision and speed choice of the entrant.

Let σ_o denote the trading speed of the old venue. A new firm may enter and open a new venue. The entrant can choose her speed from the set $[0, \infty]$. There is a cost of operating the new market, which denoted by $\alpha K(\sigma)$, where $K(0) = 0$, $\lim_{\sigma \to \infty} K(\sigma) = \infty$. After entry, both old and new firms compete by announcing their prices, hence the timing of the game is:

1. Entrant decides whether to enter or not.
2. If enters, he decides the speed σ_n at cost $\alpha K(\sigma_n)$. We stipulate that the cost of the entrant must exceed that of the incumbent: that is, σ_n is chosen from $[\sigma_o, \infty]$.
3. Both firms choose c_o and c_n simultaneously.

Our characterization of market cut-offs and trading volume allows us to express the profit of the firm in both cases. The revenue (and profit) of the old venue is $R_o(\sigma_o, \sigma_n, c_o, c_n) =$

$\pi_o(\sigma_o, \sigma_n, c_o, c_n)$ while the profit of the new venue is $\pi_n(\sigma_o, \sigma_n, c_o, c_n) = R_n(\sigma_o, \sigma_n, c_o, c_n) - \alpha K(\sigma_n)$

Let σ^* denote the equilibrium speed choice of the entrant. Following proposition shows how the market structure depends on the technology:

Proposition 3.12. *In the SPE:*

- *There is a threshold α_n^* such that the firm enters whenever $\alpha < \alpha_n^*$.*
- *Conditional on entry, the speed choice of the entrant, $\sigma_n^*(\alpha)$, is strictly decreasing in α.*

The proposition shows that the entrant enters if the cost of opening the new venue is low enough, and the equilibrium speed of the entrant is increases as the cost of speed decreases. This shows that, as expected, entry is easier if the cost of running the market is lower. Moreover, as this costs decreases, the entrant chooses a faster trading speed.

Assumption 3.13. *Going forward, we will assume that $K(\sigma)$ is linear: $\alpha K(\sigma) = \alpha\sigma$.*

Remark. This assumption is not strictly necessary for the results to follow. However, assuming a particular functional form will greatly simplify the proof. In addition, our numerical calculation section will also rely on $K(\sigma) = \sigma$.

One striking property of the entry game is that even though we allow the entrant to choose $\sigma_n \in (\sigma_o, \infty]$ freely, i.e., any improvement over the incumbent is allowed, competition will induce the entrant to choose a substantially higher speed. This is a consequence of the following sequence of propositions. Throughout, we will let $\pi(\sigma) = \pi_n(\sigma_o, \sigma_n, c_o^*, c_n^*)$ denote the profit of the entrant if he enters with $\sigma = \sigma_n$, and c_o and c_n are chosen optimally according to Proposition 3.8. For notational tidiness, we will also let τ denote the fixed parameter σ_o. We begin with straightforward calculation.

Lemma 3.14. *In the notation above, the profit of the entrant after entering with speed σ is $R(\sigma) - \alpha\sigma$, where the revenue is given by the formula*

$$\begin{aligned} R(\sigma) &= \underbrace{4Z(1-Z)\gamma(1-2\theta(\gamma+\rho))^2(\gamma+\rho+\tau)}_{\text{constant not depending on } \sigma} \frac{\sigma^2(\sigma-\tau)}{(\gamma+\sigma)[(4\sigma-\tau)(\gamma+\rho)+3\sigma\tau]^2} \\ &=: C\frac{\sigma^2(\sigma-\tau)}{(\gamma+\sigma)[(4\sigma-\tau)(\gamma+\rho)+3\sigma\tau]^2}, \end{aligned} \tag{3.27}$$

where C stands for the constant in the first line.

Lemma 3.15. *The rational function $R(\sigma)$ is strictly concave on the domain $[\tau, \infty)$. In particular, the profit function $\sigma \mapsto R(\sigma) - \alpha\sigma$ is strictly concave (on the same domain) for all $\alpha \geq 0$.*

Remark. This lemma guarantees the entrant a unique optimal σ conditional on entry.

Corollary. *If $\alpha = 0$, then the optimal choice for the entrant to enter with $\sigma_f = \infty$. Conversely, if it is optimal for the entrant to enter with $\sigma_f = \infty$, then $\alpha = 0$.*

Proposition 3.16. *Fix $\alpha > 0$, as well as other parameters $Z \in (0, 1)$, $\gamma > 0$, $\rho > 0$ and $\theta \geq 0$, and let $\pi(\sigma) := R(\sigma) - \alpha\sigma$. Finally, set $\kappa = \frac{7}{4} = 1 + \frac{3}{4}$. For any $\sigma_f := \sigma \in [\tau, \kappa\tau)$, one of the following conditions hold*

(a) The profit at σ is negative: $\pi(\sigma) < 0$.

(b) The speed σ is not optimal; $\sigma_f = \kappa\tau$ is guaranteed to be strictly better:

$$\pi(\kappa\tau) > \pi(\sigma) \quad \textit{and} \quad \pi(\kappa\tau) > 0. \tag{3.28}$$

Combining the two facts above, we see that for any set of parameters that induces the entrant to enter, the entrant chooses $\sigma_n^* \geq \frac{7}{4}\sigma_o$. That is, the speed of the entrant is at least $(1 + \frac{3}{4})$ times that of the incumbent, conditional on entry, regardless of other parameters. Therefore, the entrant is *substantially* differentiated from her competitor. The intuition is the following: offering a faster speed differentiates the entrant from the incumbent, thus attracting a larger fraction of traders and increasing revenue. Then, if it is profitable to enter at all, then it will be optimal to be maximally differentiable, subject to cost. This leads to the multiplicative gap between the incumbent and the entrant.[10]

Remark. The "growth factor" $\kappa = 1 + \frac{3}{4}$ works but is not optimal; however, it can be shown that the optimal κ is less than 2.

3.3.5 Effect of the Entry Game on Observed Quantities

In this section, we examine the effect of the entry game on the market structure, trading volume, and profits. Specifically, we set parameters $Z = \frac{1}{2}$, $\gamma = \rho = \frac{1}{10}$, normalize $\sigma_s = \sigma_o = 1$ while letting σ_f^* be chosen optimally as α and θ vary; the costs c_s^* and c_f^* will be chosen according Proposition 3.8 when the entrant enters, otherwise, c_s^* be set according to the next lemma.

Lemma 3.17. *If the entrant does not enter, then the optimal cost c_s^* is given by*

$$c_s^* = \frac{1}{4}\left(\frac{1}{\gamma + \rho} - 2\theta\right). \tag{3.29}$$

[10] If we interpret the model (and the proposition) liberally, the behavior as described places a hard cap on the number of entrants, in the following sense. The incumbent has speed $\tau_1 = \tau = 1$, say, and the entrant chooses a speed $\tau_2 \geq \kappa\tau_1 = \kappa$. The second entrant after the first faces a similar scenario: she needs to compete with at least the first entrant and therefore chooses a speed $\tau_3 \geq \kappa\tau_2 \geq \kappa^2\tau_1$, if she enters at all. Continuing in this way, the n-th entrant chooses a speed at least κ^n. If we modify the model as to only allow σ to lie in a bounded region, i.e., σ is limited by available technology such as the speed of fiber optic cable, then the number of entrants can only be grow logarithmically in the upper bound.

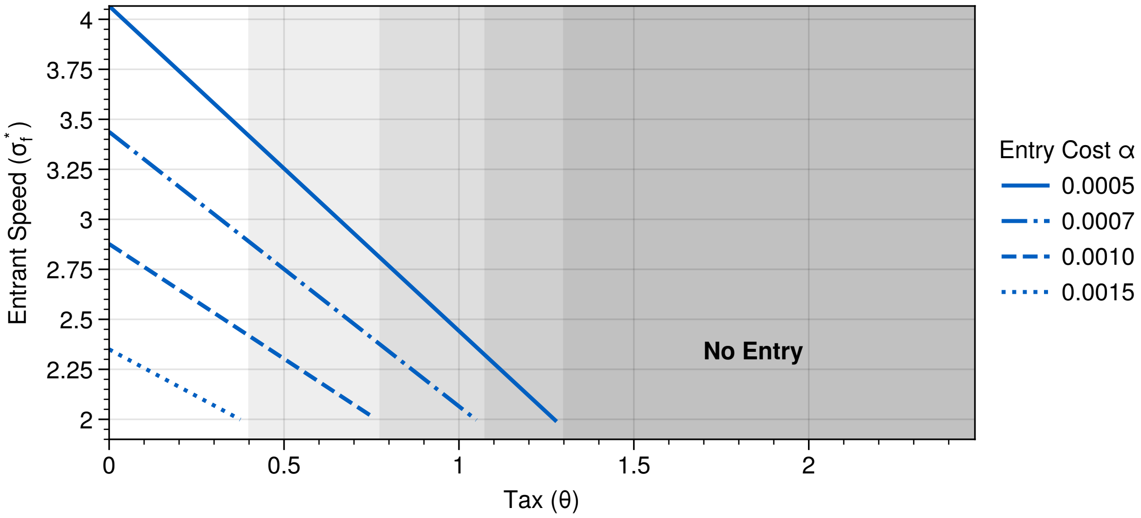

Figure 3-10: The optimal choice σ_f^*; for large enough θ or α, it is optimal for the entrant not to enter, in which case σ_f^* is undefined.

Optimal Entrant Choice

As proven in the corollary in the previous section, we know that $\sigma_f^* = \infty$ at $\alpha = 0$ but is finite for any $\alpha > 0$. It is also true that σ_f^* decreases in θ; this follows from the profit function being supermodular. Likewise, σ_f^* is also decreasing with α.

Proposition 3.18. *Suppose $\theta' > \theta \geq 0$ are such that the fast entrant finds it optimal to enter with speeds σ_f^* and $(\sigma_f^*)'$, respectively. Then $\sigma_f^* > (\sigma_f^*)'$.*

Figure 3-10 shows the optimal σ_f as a function of θ, for different values of α. The shaded regions show the denote values of θ for which the entrant does not enter (the darker regions corresponding to smaller α). As we know from the previous propositions, the optimal σ_f decreases with both θ and α; the effect is linear in θ but non-linear (concave) in α. The scaling factor in Proposition 3.16 is also shown here, since the optimal speed upon entry is at least $\kappa\sigma_n \geq \kappa$, for some $\kappa \approx 2$. The figure also reveals that

$$\frac{\partial}{\partial\alpha}\left(\frac{\partial\sigma^*}{\partial\theta}\right) < 0 \tag{3.30}$$

which means that larger entry costs α makes the effect of trading tax on venue speed more pronounced.

Endogenous Market Structure

We study the effect of market entry on the market structure, i.e., on the thresholds $\eta_1, \ldots, \eta_4$. Recall from Figure 3-2 that the thresholds vary linearly with θ when all parameters are fixed, and both the set of types who trade in the fast venue and the set of traders overall decreases.

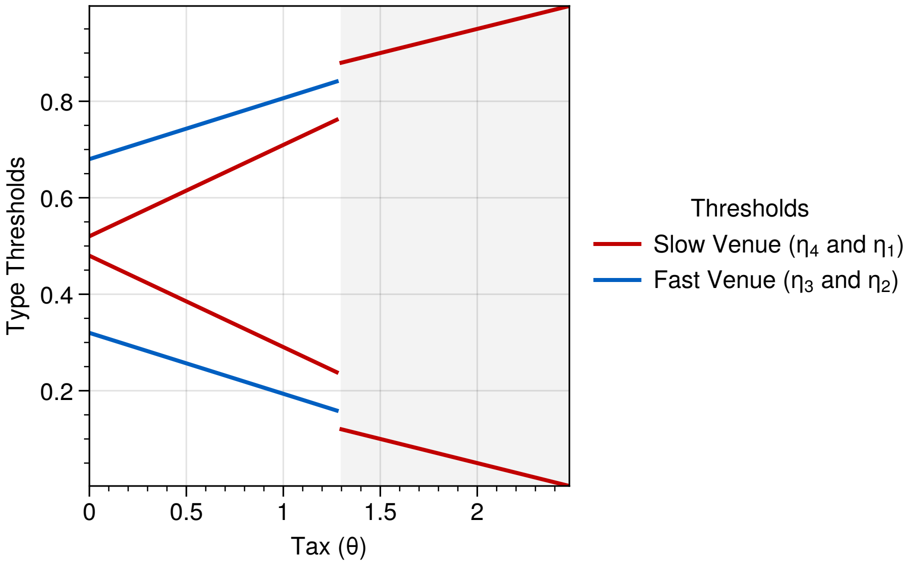

Figure 3-11: Type thresholds as a function of tax in the entry game; the shaded area denotes the region where the entrant does not enter.

As per Figure 3-11, the behavior when when $\sigma_f = \sigma_f^*$, $c_s = c_s^*$, and $c_f = c_f^*$ are chosen endogenously chosen is the same. Surprisingly, however, the effect of θ on the thresholds remains linear, at least in the region where the entrant enters. Since an exogenous shock to σ_f moves η_2 and η_4 nonlinearly, this means the costs being chosen optimally exactly counteracts the nonlinear behavior.

Once θ is sufficiently large, the set of traders who trade exhibit a discontinuous jump: the costs c_s^* increases discontinuously (c.f. the next section) and a fraction of traders—those with moderate types η—could no longer afford to trade on the "slow venue." In fact, c_s^* increases enough that post price increase, the set of traders is even smaller than the set who traded in the fast venue before the increase.

Venue Costs, Trading Volume, and Profits

We plot the effect of θ on (endogenous) venue costs c_s^* and c_f^*; the trading volumes TV_s, TV_f, and TV; and profits $c_f^* \text{TV}_f - \alpha K(\sigma_f^*)$ and $c_s^* \text{TV}_s$ in Figure 3-12. The discontinuous 'jump' of these quantities, occurring when the market structure switches from segmentation (i.e., both venues are active) to non-segmentation (i.e., only the slow venue active), is quite pronounced. The slow venue (i.e., the incumbent) increases its fees more than six-fold when the entrant chooses not to enter. Its trading volume also increases—and so the incumbent enjoys much higher profits in the absence of competition—but the increase in fees bring down total trading volume.

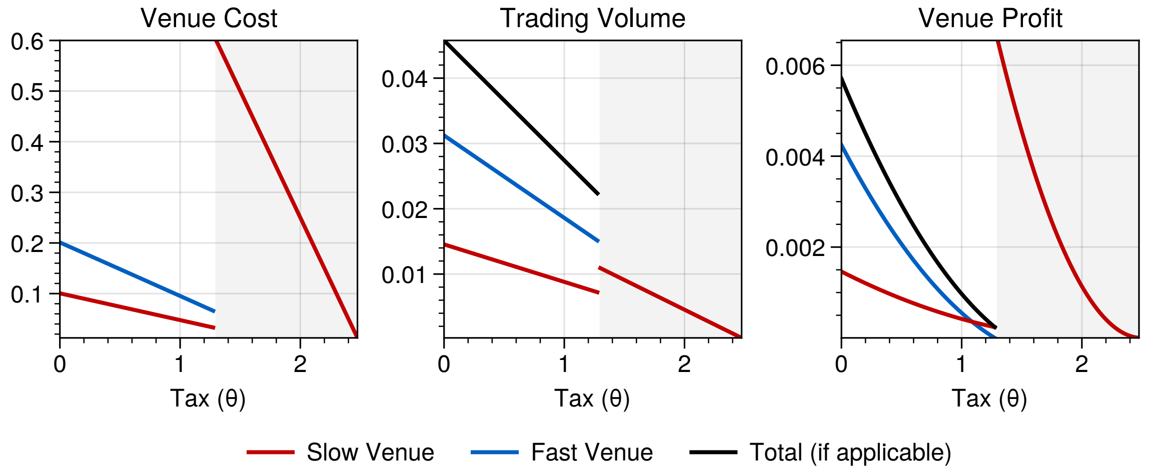

Figure 3-12: Venue costs c_s^* and c_f^*; trading volumes TV_s, TV_f, and TV; and venue profits $c_s^* \,\mathrm{TV}_s$ and $c_f^* \,\mathrm{TV}_f - \alpha K(\sigma_f^*)$. In shaded regions, the entrant does not enter.

3.4 Welfare

In this section, we analyze the model from a welfare perspective. To make the model tractable, we assume $\sigma_f \to \infty$, i.e. the trading happens instantaneously in the fast venue. We can motivate this exercise as the analysis of the equilibrium when the cost of speed in the new venue is very small, as equilibrium speed of fast venue goes to infinity as the cost parameter α goes to 0.

The following equation gives the Welfare (W), as a function of steady state asset ownership and speed of the second market (in case of entry):

$$W = \int_0^1 \eta f_h(\eta) d\eta - K(\sigma) \tag{3.31}$$

We plot the welfare as a function of θ in Figure 3-13, which reveals two important properties: that welfare is decreasing in θ, and there is a discontinuous decrease when θ is high enough to prohibit entry. This is intuitive: higher taxes has a direct negative effect on trading volume via its effect on the thresholds η; the effect is then compounded in the entry game because it dampens competition (i.e., σ_f^* decreases). Therefore, when the regulator picks θ in order to maximize welfare, she chooses $\theta = 0$ in the absence of other considerations. As we saw in previous figures, zero remains the optimal choice if the goal is to maximize trading volume (i.e., maximize liquidity).

Another quantity of interest is the tax revenue $\theta \cdot \mathrm{TV}$, depicted in the right panel. The interpretation of this quantity is that it is either collected by a central regulatory authority, or that the venues' operations are licensed from a central data center, in which case $\theta \cdot \mathrm{TV}$ is the profit of the data center operator.[11] As discussed above, the welfare-maximizing θ is

[11] For example, the New York Stock Exchange (NYSE) operate from multiple data centers to conduct its business.

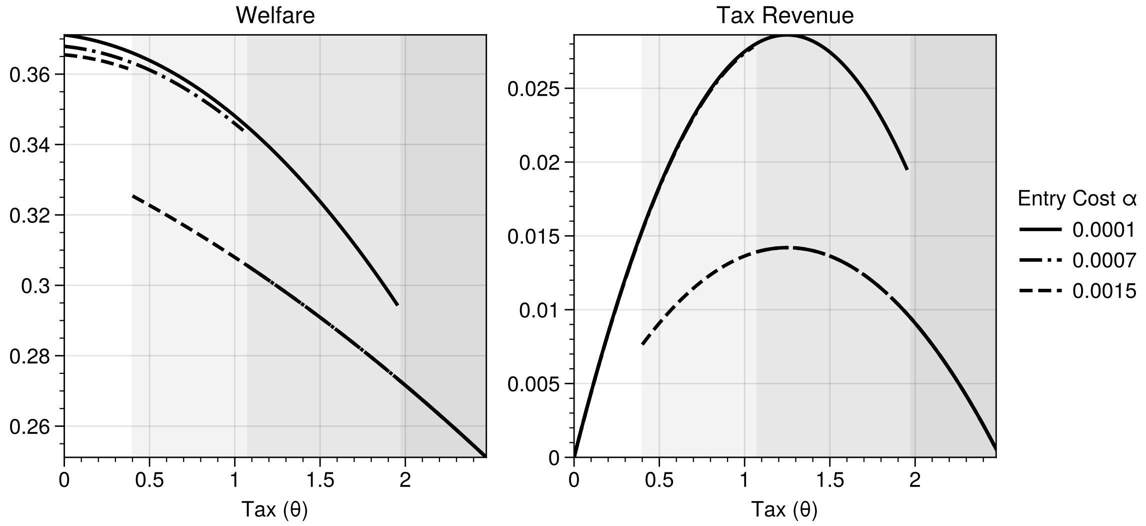

Figure 3-13: Welfare and tax revenue as a function of θ. Shaded regions indicate the entrant does not enter.

zero. However, the optimal θ for tax revenue will always be interior (since tax revenue is *a priori* zero if $\theta = 0$). Even though tax revenue will always have a discontinuous decrease at the pivotal θ, it is possible for the optimal tax revenue θ to induce a monopoly in the general case.

3.4.1 Policy Maximizing Welfare

In this section we consider central authority who wishes to maximize welfare by choosing both θ and σ_f, in contrast to only choosing θ. As discussed in the introduction, we constrain the central authority to a per-trade additive tax in order to align with similar taxes in financial markets. In general, central authorities lack the information and resources to process transactions at a sufficiently granular level in order to implement tax schemes which depend on transaction-specific characteristics. In particular, not being able to keep track of prices or dollar volumes rules out multiplicative taxes.

Throughout, we assume that $\alpha > 0$. We first show that, for fixed θ, the welfare maximizing σ_f is larger than the profit-maximizing σ_f; in particular, competition alone is not sufficient to achieve efficiency.

Instead of taking place on a particular trade floor, electronic trades are received by a data center and then routed to the relevant exchange. For example, the largest data center in the U.S. operate in Secaucus, New Jersey, which serves a majority of stock trades in the U.S.

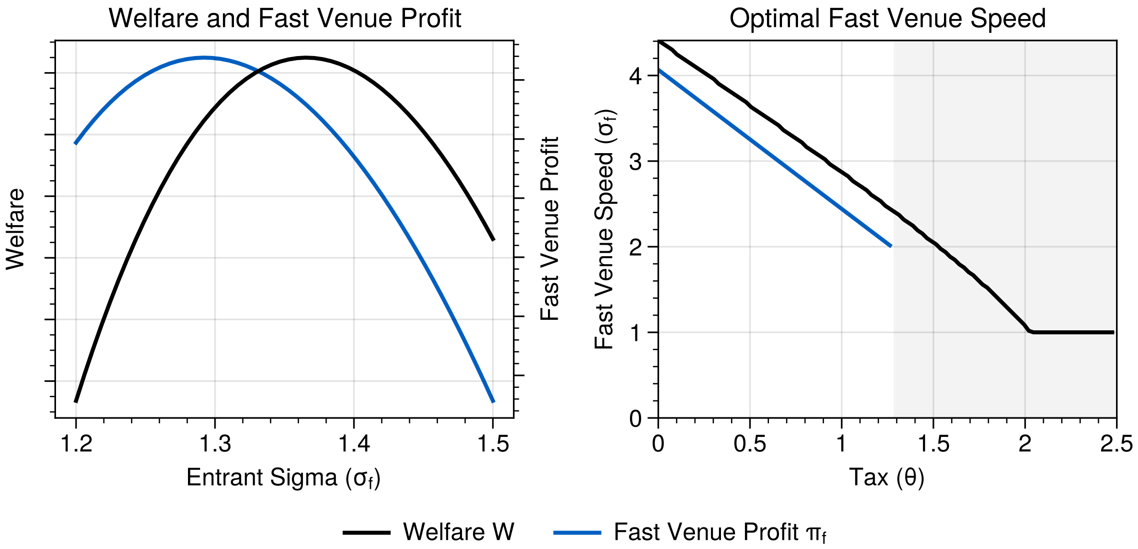

Figure 3-14: Left: welfare and fast venue profit as a function of σ_f, with $\theta = 0$. Two y-axes are used here, and the labels suppressed for tidiness. Right: Welfare and profit optimal σ_f as a function of θ.

First Best Problem. The welfare maximizing σ—denoted σ_f^w—is the solution to

$$\underset{\substack{\theta \geq 0 \\ \sigma \in [\sigma_s, \infty) \cup \{\text{NoEntry}\}}}{\arg\max} \quad W = \underset{\substack{\theta \geq 0 \\ \sigma \in [\sigma_s, \infty) \cup \{\text{NoEntry}\}}}{\arg\max} \int_0^1 f_h(\eta) \eta \, d\eta - \mathbf{1}(\sigma \in \mathbf{R}) \alpha K(\sigma) \tag{3.32}$$

subject to the IR condition of the entrant

$$\sigma = \text{NoEntry} \quad \text{or} \quad \sigma \in \mathbf{R}, \; R(\sigma) - \alpha K(\sigma) \geq 0 \tag{3.33}$$

$R(\sigma) = c_f^* \, \mathrm{TV}_f(\sigma_f)$ being the revenue function of the entrant as discussed in Lemma 3.14. The notation used for σ_f means that it could either be a real number in the interval[12] $[\sigma_s, \infty)$, signifying that the entrant enters; or $\sigma_f = \text{NoEntry}$, signifying that the entrant stays out (and relevant quantities are to be calculated for the monopoly case).

Define $\sigma_f^w(\theta)$ to be the solution of the maximization problem with θ fixed, and as before let $\sigma_f^*(\theta)$ to be the profit-optimal choice of entrant; in general both quantities belong in $[\sigma_s, \infty) \cup \{\text{NoEntry}\}$. The ordering of these two quantities depends on the parameter values, and both cases $\sigma_f^w(\theta) \lesseqgtr \sigma_f^*(\theta)$ are possible. We will refer to these two quantities as the *first-best welfare optimal* speed and the *profit optimal* or *competition* speed, respectively.

The left panel of Figure 3-14 plots the welfare and profit (of the fast venue) as a function of σ_f; for the chosen set of parameters, the welfare-optimal speed is higher than the competition

[12] Strictly speaking, our setup allows $\sigma_f = \infty$, but since $\alpha > 0$ we may assume without loss that σ must be finite.

speed[13]. The figure also shows that the revenue and welfare functions are concave.

The right panel plots functions $\sigma_f^w(\theta)$ and $\sigma_f^*(\theta)$ for $\theta \in [0, \frac{1}{2(\gamma+\rho)}]$ for a different set of parameter values. For these set of parameters, we see that $\sigma_f^w > \sigma_f^*$ whenever the entrant enters (i.e., whenever the latter is a real number).

As discussed previously, $\sigma_f^*(\theta) = \mathsf{NoEntry}$ for sufficiently large θ; which is indicated by the gray region. For these values of θ, there exists no $\sigma_f \in [\sigma_s, \infty)$ that achieves a positive profit: therefore, the IR constraint also forces $\sigma_f^w(\theta) = \mathsf{NoEntry}$. However, ignoring the IR constraint, the welfare optimal σ_f is never NoEntry for these set of parameter values. For example, when θ close to its upper limit, welfare maximization leads to the boundary solution $\sigma_w^f = \sigma_s = 1$. In this case, two venues engage in Bertrand competition and set their fees to zero ($c_f^* = c_s^* = 0$); the market structure has only one venue active with[14] $\eta_4 = \eta_1$, $\eta_3 = 0$, and $\eta_2 = 1$.

In general, the (IR-less) welfare optimal choice involves entry over the entire range of θ whenever the integral $\int_0^1 f_h(\eta)\eta d\eta$ at $\sigma = \sigma_s$ exceeds the cost $\alpha K(\sigma_s)$. Otherwise, if α is sufficiently high, we will have $\sigma_f^w(\theta) = \mathsf{NoEntry}$ for θ in a neighborhood of $\frac{1}{2(\gamma+\rho)}$.

The intuition for which one of the two speeds is faster is the following. For fixed values of other parameters (including θ), increasing σ drives up costs c_f^* and c_s^* (c.f. Figure 3-7). This is a tradeoff for the traders: they pay higher fees but are compensated by faster speeds (i.e., higher liquidity), where the benefit is captured by the increase in $\int f_h(\eta)\eta d\eta$. Profit optimization maximizes the (negative of the) first term, while welfare optimization maximizes the latter; in symbols

$$R(\sigma) = c_f^* \, \mathrm{TV}_f(\sigma) \quad \text{and} \quad W_0(\sigma) = \int_0^1 \eta f_h(\eta) d\eta. \tag{3.34}$$

The formula for $R(\sigma)$ is given in Lemma 3.14, and it is easy to see that W_0 is also a rational function of σ. The optimal points are given by setting the derivatives of these two functions equal to $\alpha K'(\sigma)$, so the determination of $\sigma_f^w \lesseqgtr \sigma_f^*$ depends on

$$R'(\sigma) \lesseqgtr W_0'(\sigma). \tag{3.35}$$

In words, $R'(\sigma) > W_0'(\sigma)$ for all σ means that the entrant fails to fully internalize the externality of his entry on the traders. In this case, the regulator prefers a faster speed, $\sigma_f^w > \sigma_f^*$. On the other hand, if $R'(\sigma) < W_0'(\sigma)$, the entrant sets his speed high enough that larger costs outweigh the liquidity benefits (in aggregate), so the regulator prefers a slower speed.

[13] Though axis labels are suppressed for tidiness, the IR constraint of the entrant is satisfied at the welfare-optimal speed

[14] By solving the equations in Theorem 3.3, we see that $\eta_4 = \eta_1 = Z$.

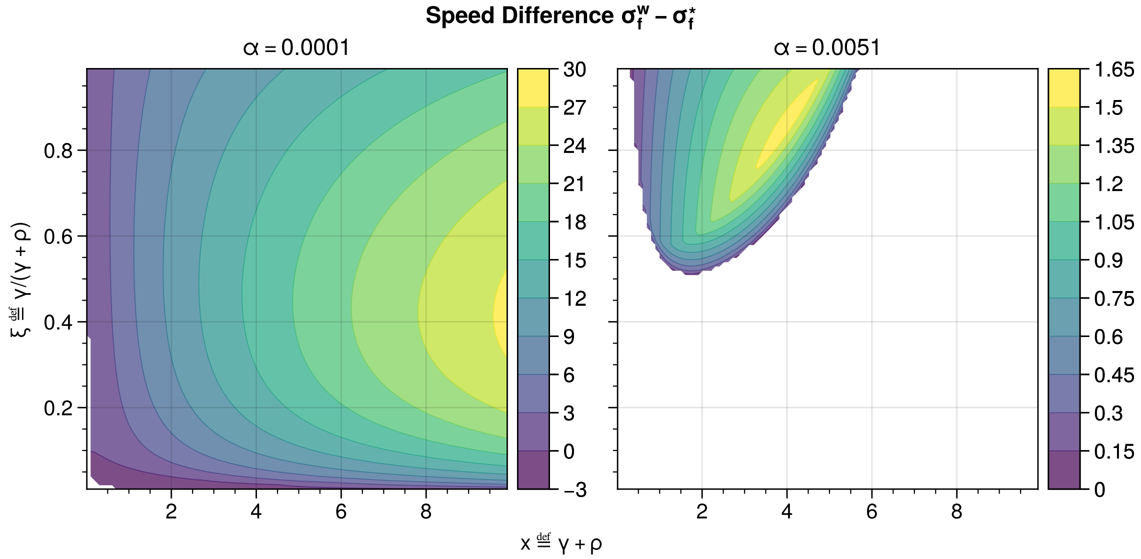

Figure 3-15: Difference between the welfare maximizing venue speed and the profit maximizing venue speed.

3.4.2 Maximizing Trading Volume

Instead of maximizing societal welfare, the central regulatory authority may instead want to maximize total trading volume as defined in section 3.3.2, subject to the IR constraint (3.33).

Figure 3-12 illustrates that trading volume is maximized at $\theta = 0$ when the choice of σ_f is left to the entrant (i.e., when σ_f is the induced competition speed). The figures in the next section illustrate that this remains true when θ and σ_f are jointly chosen subject to the IR condition. Both of these facts are extensions of the result in Proposition 3.4; the proposition states that total trading volume is decreasing in θ when other parameters of fixed. The figures shows that when the other parameters—namely, σ_f, c_s, and c_f—are endogenously chosen, trading volume remains decreasing in θ.

3.4.3 Illustrations

In this section, we compute the solutions to the first best problem and illustrate the possible orderings given in . Throughout, $\sigma_s = 1$ and $Z = \frac{1}{2}$, so that freely varying parameters are γ, ρ, θ, and α. For our figures, it will be convenient to reparameterize our model as follows,

$$x := \gamma + \rho, \quad \xi := \frac{\gamma}{\gamma + \rho}, \tag{3.36}$$

with the new parameters being $x \in (0, \infty)$, $\xi \in (0, 1)$, and $\alpha \in (0, \infty)$.

Figure 3-15 compares the profit-optimal and welfare-optimal venue speeds. The left panel features a relatively low technology cost α; the right panel features a substantially higher α. White regions correspond to combination of parameter values such that $\sigma_f^* = \mathsf{NoEntry}$, in

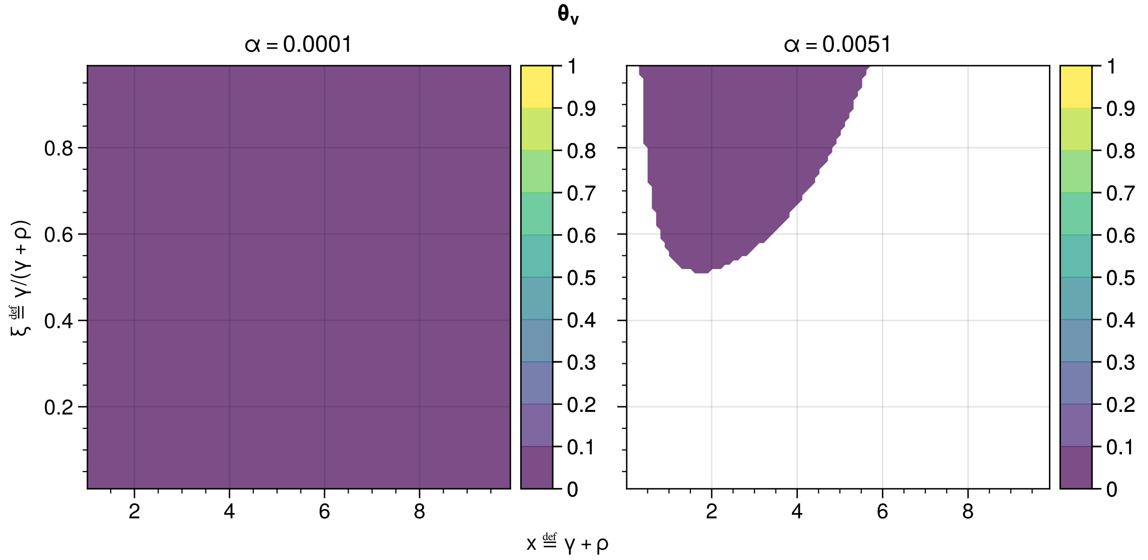

Figure 3-16: The optimal θ—denoted θ_v—that solves the problem of maximizing total trading volume subject to the IR constraint. As shown, θ_v is identically zero whenever the problem is feasible.

which case $\sigma_f^w = \mathsf{NoEntry}$ is forced by the IR condition. In both the low and high cost case, the speed difference $\sigma_f^w - \sigma_f^*$ is usually positive, i.e., the welfare-optimal speed is higher than the profit-optimal speed. However, the qualitative behavior of the difference depends on the size of α.

When α is small, it is easier for the the entrant to achieve a positive profit, all else being equal. Therefore, the left panel depicts a larger region of σ_f^* and σ_f^w in which the entrant chooses to enter. Moreover, the speed difference could be quite large, with a maximum value of approximately 30 depicted in the figure. Recall that $\sigma_f^* \to \infty$ and $\sigma_f^w \to \infty$ as the cost $\alpha \to 0$. The comparison in the figures yields the additional comparison that $\sigma_f^w \to \infty$ at a faster rate than σ_f^*.

In addition, when α is small there exists regions for which $\sigma_f^* > \sigma_f^w$. The reason is that a low α allows for regions with $\gamma + \rho \approx 0$ where the entrant's IR condition could still be satisfied. As α grows, the IR-feasible region shrinks in such a way that $\sigma_f^w > \sigma_f^*$ everywhere. The right panel depicts such α where it is slightly too high to allow for $\sigma_f^* > \sigma_f^w$, i.e., near the boundary of the feasible region, the difference is almost zero.

The next two figures plots the same quantities, but for the problem of maximizing trading volume (subject to the IR condition). Figure 3-16 shows that wherever the IR constraint is feasible, the optimal choice of θ is equal to zero. This agrees with the intuition that taxes—being a source of friction—decreases total trading volume (c.f., Proposition 3.4). Figure 3-17 shows that when $\gamma + \rho$ and γ are sufficiently large or when α is sufficiently large, the trading volume optimal speed σ_f^v is higher than the competitive speed σ_f^*. When the fast venue speed is high, the costs c_f^* will be high as well, driving trading volume from

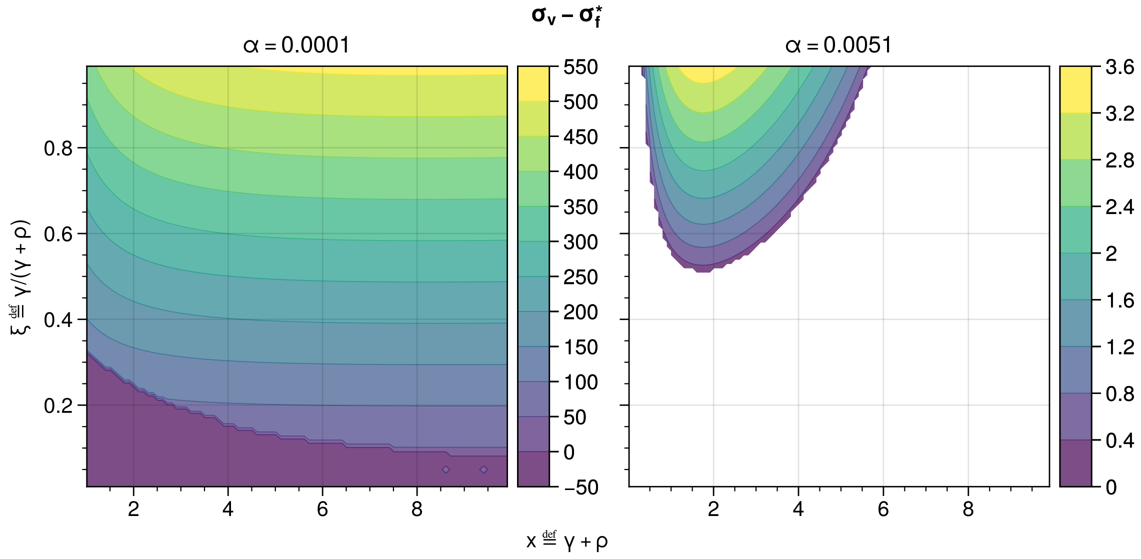

Figure 3-17: Difference between the trading volume maximizing venue speed and the profit maximizing venue speed.

the fast venue to the slow venue. The competitive speed takes this into account, where as maximizing total trading volume is not affected by the increase in cost.

3.5 Trading Revenue Maximization

Another interpretation of the model has θ under the control of another profit-maximizing firm instead of a welfare maximizing entity. In this interpretation, traders are retails traders who trade through brokers, who match them with a counterparty with whom the broker has a professional relationship. The counterparties are usually large market makers for highly liquid instruments, or trading desks at investment banks for less liquid instruments.

Since brokers offers different technologies and have access to different sets of counterparties, trading and execution speed differ among potential brokers. Thus, the brokers serve the role of "venues": they compete on services offered (i.e., fees and trading speed), while retail traders chooses a broker to balance fees and speed. Thus, c_s^* and c_f^* are brokerage fees.

On the other hand, since trading is ultimately routed to a market maker or trading desk, retail traders often need to cross a bid-ask spread for market orders or obtain a less-than-optimal price for limit orders. In other words, every trade incurs an inherent haircut, and profits of these haircuts go to the counterparty. The bid-ask spreads are set by the counterparty, and corresponds to an effective tax, i.e., our model's θ.

Since θ is set by the counterparty, it is chosen to maximize trading revenue $\theta \cdot \text{TV}$ instead of the welfare. In order to engage brokers, the counterparty must not set θ too high; in other words, the problem faced by the counterparty is to maximize $\theta \cdot \text{TV}$ subject to the IR

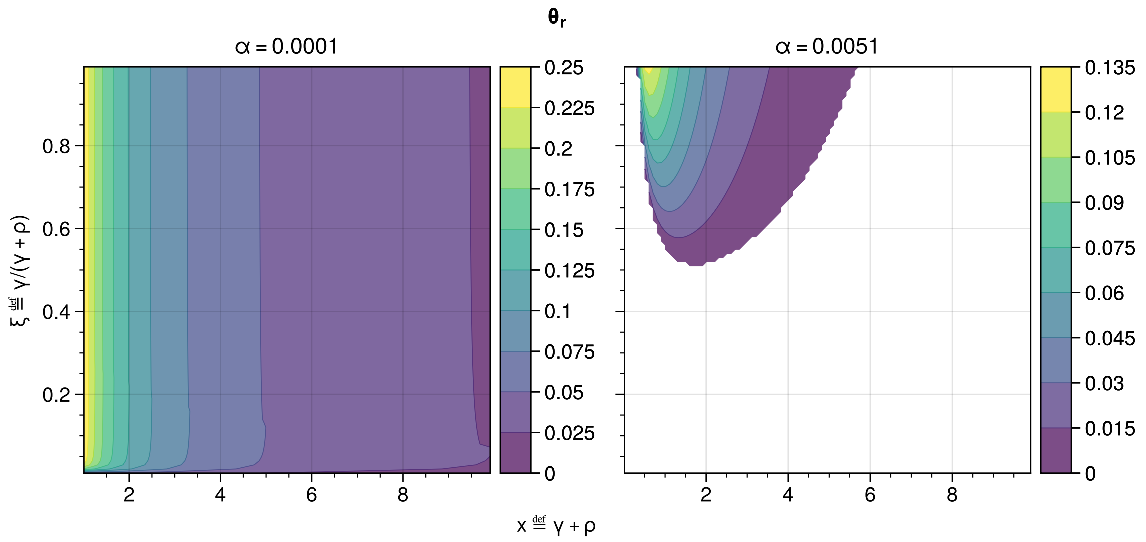

Figure 3-18: Optimal θ set by the counterparty to maximize trading revenue.

constraint (3.33). In contrast to welfare maximization in the previous section, the optimal θ—denoted θ^r—is always nonzero. This is illustrated in Figure 3-18. In general, θ^r increases in $\xi = \gamma/(\gamma + \rho)$, but decreases in $\gamma + \rho$. Moreover, θ^r decreases with the cost α; the intuition is that when α increases, the entrant's speed σ_f decreases, which decreases the utility of trading for traders. Therefore, the trading volume is more sensitive to the haircut fee θ, driving the optimal choice lower.

Let σ_f^r denote the optimal speed under trading revenue maximization. Figure 3-19 depicts the difference $\sigma_f^r - \sigma_f^*$ for the same parameters as Figure 3-15. The behavior of the difference varies drastically between the low and high values of α. For small α, the difference is increasing in $\gamma/(\gamma + \rho)$; for larger α, the opposite is true. Comparing the range of values in the left panels of Figures 3-15 and 3-19, there is no general ordering between the three quantities σ_f^*, σ_f^w, and σ_f^r when α is small; however, when α is sufficiently, we have

$$\sigma_f^r \leq \sigma_f^* \leq \sigma_f^w \tag{3.37}$$

whenever the entrant enters.

3.6 Extension Section

3.6.1 Simultaneous Speed Choice and Segmentation

In this section, we drop the assumption of sequential entry, i.e. there is no old or new venue. We denote the firms by 1 and 2. First, both firms choose to whether to enter or not. After the entry decisions, the firms choose their speeds. If both firms enter, then they compete in prices by choosing their fees c_1 and c_2 simultaneously. Given speeds and fees, the market

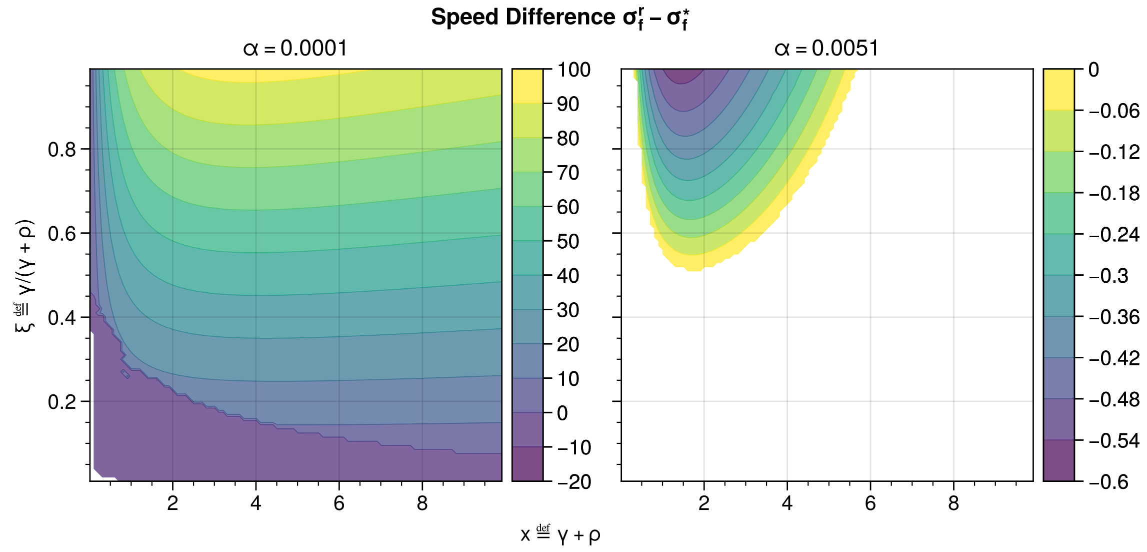

Figure 3-19: Difference between the trading revenue maximizing venue speed and the profit maximizing venue speed.

structure is characterized by Theorem 3.3 and given market structure, fee competition is characterized by section 3.3.3.

In order to make the model tractable, we assume two different speeds, $\sigma_l > \sigma_h$ are available to the entrants. The cost of operating (or setting up) the new venues is denoted by $\alpha K(\sigma)$, which is increasing in σ. We focus on the pure strategy SPE.

Proposition 3.19. *There are two cut-offs $\underline{\alpha}$ and $\overline{\alpha}$ such that in the pure strategy SPE,*

- *If $\alpha < \underline{\alpha}$, both firms enter and choose different speeds.*
- *If $\alpha \in [\underline{\alpha}, \overline{\alpha}]$, then only one firm enters.*
- *If $\alpha > \overline{\alpha}$, both firms stay out.*

Proposition 3.19 echoes our results under sequential entry and shows that segmentation may arise even under simultaneous entry. If the cost of operating the venues α is low enough, then both firms enter and choose different speeds and obtain positive profits due to differentiation. If the cost is high, then neither can operate as even a monopolist cannot make positive profits. In the intermediate range, one of the venues incurs a loss thus there is no differentiation.

3.7 Conclusion

We have presented a general equilibrium model of traders participating in multiple venues. Due to preference shocks, traders continually switch between the two venues, which offer different trading speeds and transaction fees in equilibrium. Our main contribution is to analyze how competition affects endogenous market segmentation, transaction speeds and fees, trading volume, optimal regulator's choice for taxing traders, and welfare in illiquid asset markets. We find that liquidity increases with fragmentation in asset markets, while decreasing with higher regulatory taxes. Importantly, depending on the regulator's objective, the optimal trading tax choice can be zero or strictly positive.

We further analyze how competition between venues (dealers) affect trading fees and speeds. Under competition, the optimal choice of transaction fees are increasing (resp. decreasing) in the speed of the faster (resp. slower) venue. Moreover, with sequential entry, the entrant's optimal speed increases with lowering entry costs and trading taxes.

Finally, we consider different notions of welfare: surplus from trade, trading volume, and trading revenue. Importantly, in each of these cases, we consider the regulator's optimal choice for taxing traders, and the resulting optimal choice of speed for a new entrant. Interestingly, we show that not only does competition lower welfare, but also the speed choice of the entrant can be smaller than the welfare maximizing speed. Furthermore, our results show that lower entry costs reduce these effects and is thus welfare improving.

3.8 Appendix

Proof of Theorem 3.3

The proof will proceed as follows: First, we show in Lemma 3.20 in any equilibrium with positive trade, selling in slow venue leaves the seller with higher revenue than selling the fast venue and buying in the slow venue is cheaper for the buyer compared to buying in the fast venue. Using these facts, Lemmas 3.21 and 3.22 characterize actions of traders after a trade and show that buyers hold the asset after buying and sellers do nothing after selling. Given these actions, Lemma 3.23 characterizes the value functions after any decision (selling, holding, buying or doing nothing) by a trader. Lemmas 3.24, 3.25 and 3.26 characterize the structure of traders' venue choices using a simple cut-off structure. Lemma 3.27 gives the sufficient (which we later show to be necessary) condition for no trade in both venues, proving part (i) of Theorem 3.3. This condition requires the present value of the gains from most profitable trade (between a trader with valuation η_h and a trader with valuation 0) to be larger than total fees and taxes paid by the traders for the trade. Lemma 3.28 gives an analogous condition for existence of trade in the fast venue, which simply requires the value of most profitable (thus most speed sensitive) trade to be larger than some measure of differentiation among the venues. Lemma 3.29 characterizes the equilibrium type distributions of asset holders and non-holders using inflow and outflow equations. Using the distributions characterized in Lemma 3.29 and the venue clearing conditions, in Lemma 3.30 we arrive at two of the four main conditions in part (ii) of Theorem 3.3. Assuming both venues are active (i.e. the condition given in Lemma 3.28 holds), Lemma 3.31 finishes the the characterization of cut-offs structure when the market is segmented. Lemma 3.32 then finishes the proof of part (ii) of Theorem 3.3 by showing the existence and uniqueness of prices. Lastly, we prove part (iii) by using Lemmas 3.29 and 3.30 and showing the existence of prices where there is only demand for the slow venue whenever the condition in Lemma 3.28 does not hold.

Lemma 3.20. *In any equilibrium with positive trading, the following are satisfied:*

1. $p_s - c_s \geq p_f - c_f$*; and*
2. $p_s + c_s \leq p_f + c_f$

Proof. Assume for a contradiction that $p_s - c_s \geq p_f - c_f$ is not satisfied. Then we have $p_s - c_s < p_f - c_f$. For a contradiction, assume that there exists η such that $V_S^s(\eta) \geq V_H(\eta)$ and $V_S^s(\eta) \geq V_S^f(\eta)$. The former implies that $V_{0,\eta} + p_s - c_s - \theta - V_S^s(\eta) > 0$. Note that the values for $V_S^s(\eta)$ and $V_S^f(\eta)$ are given by following equations:

$$\rho V_S^s(\eta) = u(\eta) + \gamma(\mathbf{E}_{\eta'}[V_{1,\eta'}] - V_S^s(\eta)) + \sigma_s(V_{0,\eta} + p_s - c_s - \theta - V_S^s(\eta)) \tag{3.38}$$

$$\rho V_S^f(\eta) = u(\eta) + \gamma(\mathbf{E}_{\eta'}[V_{1,\eta'}] - V_S^f(\eta)) + \sigma_f(V_{0,\eta} + p_f - c_f - \theta - V_S^f(\eta)) \tag{3.39}$$

Looking at each term, as $V_S^s(\eta) \geq V_S^f(\eta)$, we have $\gamma(\mathbf{E}_{\eta'}[V_{1,\eta'}] - V_S^f(\eta)) \geq \gamma(\mathbf{E}_{\eta'}[V_{1,\eta'}] - V_S^s(\eta))$. Moreover, $V_S^s(\eta) \geq V_S^f(\eta)$, $\sigma_f > \sigma_s$, $V_{0,\eta} + p_s - c_s - \theta - V_S^s(\eta) > 0$ together with $p_s - c_s < p_f - c_f$ implies that

$$\sigma_f(V_{0,\eta} + p_f - c_f - \theta - V_S^f(\eta)) > \sigma_s(V_{0,\eta} + p_s - c_s - \theta - V_S^s(\eta)).$$

Thus, we have $V_S^s(\eta) < V_S^f(\eta)$, which is a contradiction. As a result, There is no η such that $V_S^s(\eta) \geq V_H(\eta)$ and $V_S^s(\eta) \geq V_S^f(\eta)$. This means that there is measure 0 of traders who prefer to sell slow. As we assumed there is positive trade, then there must be non-zero measure of traders who prefer to sell fast.

Next, we will show that under $p_s - c_s < p_f - c_f$, there positive demand in selling slow, which is a contradiction as slow venue clearing condition cannot hold in that case. Note that $p_s - c_s < p_f - c_f$ and $c_s < c_f$ implies $p_s + c_s < p_f + c_f$.

In any equilibrium with positive trade, there is a type η that prefers buying fast or slow to doing nothing. If type η prefers buying slow, i.e., $V_B^S(\eta) > \max\{V_S^s(\eta), V_N(\eta)\}$, the continuity of V_N, V_S^s and V_S^f in η implies that there is a positive measure of types around η that prefer to buy slow.[15] However, this will be a contradiction to the slow venue clearing condition

[15] The continuity of V_N and V_H in η follows directly from continuity of $u(\eta)$. The continuity of V_S^s, V_S^f, V_B^s and V_B^f follows from continuity of $u(\eta)$, $V_{1,\eta}$ and $V_{0,\eta}$. To show the continuity of $V_{1,\eta}$ and $V_{0,\eta}$ in any equilibrium, let η and $\eta' = \eta + \epsilon$ with $\epsilon > 0$ denote two different types. First, note that the equilibrium strategy of η is available to η' and $u(\eta') > u(\eta)$, thus η' can guarantee herself a payoff of at least $V_{1,\eta}$ when she owns the asset and $V_{0,\eta}$ when she does not by playing the same strategy. Thus we have $V_{1,\eta'} \geq V_{1,\eta}$ and $V_{0,\eta'} > V_{0,\eta}$. Next, if η plays the equilibrium strategy of η', then $V_{1,\eta'} - V_{1,\eta}$ and $V_{0,\eta'} - V_{0,\eta}$ is bounded above by the expected utility derived from owning the asset until the preference shock strikes. As u is continuous, this bound converges to 0 as ϵ goes to 0, yielding the desired continuity.

and cannot happen under any equilibrium.

Next, assume for a contradiction there are no types that prefer to buy slow. Then each type either prefers doing nothing or buying fast. Let η^* denote the type that is indifferent between buying fast and doing nothing, i.e., $V_N(\eta^*) = V_b^f(\eta^*)$. This means that $V_{1,\eta^*} - p_f - c_f - \theta - V_B^f(\eta^*) = 0$. But as $p_s + c_s < p_f + c_f$ and $V_B^s \leq V_s^f$ we have $V_{1,\eta^*} - p_s - c_s - \theta - V_b^f(\eta^*) > 0$. This implies that $V_B^s(\eta^*) > V_B^f(\eta^*)$. But then, there is a positive measure of traders that prefer to buy slowly. As we have showed there are no traders that prefer to sell slowly, slow venue clearing condition cannot hold and this is a contradiction to the assertion that that we have an equilibrium.

The proof of second part is analogous. Assume for a contradiction that $p_s + c_s \leq p_f + c_f$ is not satisfied. Then we have $p_s + c_s > p_f + c_f$. For a contradiction, assume that there exists η such that $V_B^s(\eta) \geq V_N(\eta)$ and $V_B^s(\eta) \geq V_B^f(\eta)$. The former implies that $V_{1,\eta} - p_s - c_s - \theta - V_B^s(\eta) > 0$. Note that the values for $V_B^s(\eta)$ and $V_B^f(\eta)$ are given by following equations

$$\rho V_B^s(\eta) = \gamma(\mathbf{E}_{\eta'}[V_{0,\eta'}] - V_B^s(\eta)) + \sigma_s(V_{1,\eta} - p_s - c_s - \theta - V_B^s(\eta)) \tag{3.40}$$

$$\rho V_B^f(\eta) = \gamma(\mathbf{E}_{\eta'}[V_{0,\eta'}] - V_B^f(\eta)) + \sigma_f(V_{1,\eta} - p_f - c_f - \theta - V_B^f(\eta)). \tag{3.41}$$

Since $V_B^s(\eta) \geq V_B^f(\eta)$, we have $\gamma(\mathbf{E}_{\eta'}[V_{0,\eta'}] - V_B^f(\eta)) \geq \gamma(\mathbf{E}_{\eta'}[V_{0,\eta'}] - V_B^s(\eta))$. Moreover, $V_B^s(\eta) \geq V_B^f(\eta)$, $\sigma_f > \sigma_s$, $V_{1,\eta} - p_s - c_s - \theta - V_B^s(\eta) > 0$ together with $p_s + c_s > p_f + c_f$ implies that

$$\sigma_f(V_{1,\eta} - p_f - c_f - \theta - V_B^f(\eta)) > \sigma_s(V_{1,\eta} - p_s - c_s - \theta - V_B^s(\eta))$$

Thus, we have $V_B^s(\eta) < V_B^f(\eta)$, which is a contradiction. As a result, There is no η such that $V_B^s(\eta) \geq V_N(\eta)$ and $V_B^s(\eta) \geq V_B^f(\eta)$. This means that there is measure 0 of traders who prefer to buy slow. As we assumed there is positive trade, then there must be non-zero measure of traders who prefer to buy fast.

Next, we will show that under $p_s + c_s > p_f + c_f$, there positive demand in buying slow, which is a contradiction as slow venue clearing condition cannot hold in that case. Note that $p_s + c_s > p_f + c_f$ and $c_s < c_f$ implies $p_s - c_s > p_f - c_f$.

In any equilibrium with positive trade, there is a type η that prefers selling fast or slow to holding. If type η prefers selling slow, i.e., $V_B^S(\eta) > \max\{V_S^s(\eta), V_N(\eta)\}$, the continuity of V_N, V_S^s and V_S^f in η implies that there is a positive measure of types around η that prefer to buy slow. However, this will be a contradiction to the slow venue clearing condition and cannot happen under any equilibrium.

Next, assume for a contradiction there are no types that prefer to sell slow. Then each type either prefers holding or selling fast. Let η^* denote the type that is indifferent between selling fast and holding, i.e. $V_H(\eta^*) = V_S^f(\eta^*)$. This means that $V_{0,\eta^*} + p_f - c_f - \theta - V_B^f(\eta^*) = 0$. But as $p_s - c_s > p_f - c_f$ and $V_S^s \leq V_S^f$ we have $V_{0,\eta^*} + p_s - c_s - \theta - V_b^f(\eta^*) > 0$. This implies that $V_S^s(\eta^*) > V_S^f(\eta^*)$. But then, there is a positive measure of traders that prefer to sell slowly. As we have showed there are no traders that prefer to buy slowly, slow venue clearing condition cannot hold and this is a contradiction to the assertion that that we have an equilibrium. □

Lemma 3.21. *(a)* $S_s \subseteq N \cup B_f$
(b) $S_f \subseteq N \cup B_s$
(c) $B_s \subseteq H \cup S_f$
(d) $B_f \subseteq H \cup S_s$.

Proof. To prove (i) and (iii), assume for a contradiction there exist $\eta \in S_s \cap B_s$. Then we have $V_{1,\eta} = V_S^s(\eta)$ and $V_{0,\eta} = V_B^s(\eta)$. Then by substituting for RHS and transposing terms,

$$(\gamma + \rho)(V_{1,\eta} + V_{0,\eta}) = u(\eta) + \gamma(\mathbf{E}_{\eta'}[V_{1,\eta'}] + \mathbf{E}_{\eta'}[V_{0,\eta'}])) - 2\sigma_s(c_s + \theta). \tag{3.42}$$

From the optimality of selling in the slow venue,

$$(\sigma_s+\gamma+\rho)(\gamma+\rho)[V_S^s(\eta)-V_H] = -\sigma_s(u(\eta)+\gamma\,\mathbf{E}_{\eta'}[V_{1,\eta}])+\sigma_s(\gamma+\rho)(V_{0,\eta}+p_s-c_s-\theta) \geq 0.$$

Rearranging:

$$(\gamma + \rho)V_{0,\eta} \geq u(\eta) + \gamma\,\mathbf{E}_{\eta'}(V_{1,\eta'}) - (\gamma + \rho)(p_s - c_s - \theta) \tag{3.43}$$

From the optimality of buying in the slow venue, using $V_B^s(\eta) - V_N(\eta) > 0$ we obtain:

$$(\gamma + \rho)V_{1,\eta} \geq \gamma\,\mathbf{E}_{\eta'}(V_{0,\eta'}) + (\gamma + \rho)(p_s + c_s + \theta) \tag{3.44}$$

Summing equations 3.43 and 3.44, we obtain

$$(\gamma + \rho)(V_{0,\eta} + V_{1,\eta}) \geq u(\eta) + \gamma\,\mathbf{E}_{\eta'}(V_{1,\eta'}) + \gamma\,\mathbf{E}_{\eta'}(V_{0,\eta'}) + 2(\gamma + \rho)(c_s + \theta),$$

which contradicts equation 3.42. Replacing s with f in above proof proves (ii) and (iv). □

The next Lemma shows that fast sellers does nothing after selling and slow buyers hold after buying.

Lemma 3.22. *The following sets do not intersect*

$$S_f \cap B_s = \emptyset \qquad S_s \cap B_f = \emptyset. \tag{3.45}$$

Proof. To prove (i), assume for a contradiction $\eta \in S_f \cap B_s$. Then as $\eta \in B_s$, we have: the following quantities are all positive: $V_B^s(\eta) - V_N(\eta)$, $(\gamma+\rho)\sigma_s(V_{1,\eta}-p_s-c_s-\theta)-\sigma_s\gamma\,\mathbf{E}_{\eta'}(V_{0,\eta'})$, and $(\gamma + \rho)\sigma_s(V_{1,\eta} - p_s - c_s - \theta) + (\gamma + \rho)\gamma\,\mathbf{E}_{\eta'}(V_{0,\eta'}) - \gamma(\sigma_s + \gamma + \rho)\,\mathbf{E}_{\eta'}(V_{0,\eta'})$.

This implies that

$$\begin{aligned}(\sigma_s + \gamma + \rho)(V_{1,\eta} - p_s - c_s - \theta) &> \sigma_s(V_{1,\eta} - p_s - c_s - \theta) + \gamma\, \mathbf{E}_{\eta'}(V_{0,\eta'}) \\ V_{1,\eta} - p_s - c_s - \theta &> V_B^s(\eta) = V_{0,\eta} \\ V_{1,\eta} &> V_{0,\eta} + p_s + c_s + \theta \qquad (3.46)\end{aligned}$$

This is intuitive as the individual prefers the continuation value while holding the asset and paying p_s to the value of not holding the asset. From $\eta \in S_f$, we have:

$$V_S^f(\eta) - V_H(\eta) > 0$$

Similar calculations as above yield:

$$V_{1,\eta} < V_{0,\eta} + p_f - c_f - \theta \qquad (3.47)$$

Equations 3.46 and 3.47 imply $p_f - c_f > p_s + c_s + 2\theta$. Subtracting $2c_s - 2\theta$ from LHS, we obtain $p_f - c_f > p_s - c_s$ which is a contradiction. Switching s with f in the above proof proves (ii). □

Lemma 3.23. *The value functions are given by the formulas*

$$\begin{aligned}V_S^f(\eta) &= \frac{u(\eta) + \sigma_f(V_N + p_f - c_f - \theta) + \gamma\, \mathbf{E}_{\eta'}(V_{1,\eta'})}{(\sigma_f + \gamma + \rho)} \\ V_S^s(\eta) &= \frac{u(\eta) + \sigma_s(V_N + p_s - c_s - \theta) + \gamma\, \mathbf{E}_{\eta'}(V_{1,\eta'})}{(\sigma_s + \gamma + \rho)} \\ V_B^f(\eta) &= \frac{\sigma_f(V_H(\eta) - p_f - c_f - \theta) + \gamma\, \mathbf{E}_{\eta'}(V_{0,\eta'})}{(\sigma_f + \gamma + \rho)} \\ V_B^s(\eta) &= \frac{\sigma_s(V_H(\eta) - p_s - c_s - \theta) + \gamma\, \mathbf{E}_{\eta'}(V_{0,\eta'})}{(\sigma_s + \gamma + \rho)}.\end{aligned}$$

Furthermore, the following lemma helps us prove the structure of the speed choices.

Lemma 3.24. $\frac{\partial V_{1,\eta}}{\partial \eta} > 0$ *and* $\frac{\partial V_{0,\eta}}{\partial \eta} \geq 0$.

Proof. As $u(\eta)$ is strictly increasing, we have following:

$$\begin{aligned}\frac{\partial V_N(\eta)}{\partial \eta} &= 0 \\ \frac{\partial V_H(\eta)}{\partial \eta} &> 0\end{aligned}$$

Next, assume $\eta \in B_s$. Take any $\eta' > \eta$. We have:

$$V_{0,\eta} \geq V_B^s(\eta') > V_B^s(\eta) = V_{0,\eta},$$

where first inequality follows from the optimality of $V_{0,\eta}$, second follows from the fact $u(\eta)$ is strictly increasing and third equality follows from $\eta \in B_s$.

Similarly, assume $\eta \in B_f$. Take any $\eta' > \eta$. With the exactly same reasoning, we obtain

$$V_{0,\eta} p \geq V_B^f(\eta') > V_B^f(\eta) = V_{0,\eta}.$$

Hence $\frac{\partial V_{0,\eta}}{\partial \eta} \geq 0$, which proves the second claim. Given this, it is immediate to conclude $V_H(\eta), V_S^s(\eta)$ and $V_S^f(\eta)$ are all strictly increasing in η as $u(\eta)$ is strictly increasing. Thus $V_{1,\eta}$, which is their maximum, is strictly increasing:

$$\frac{\partial V_{1,\eta}}{\partial \eta} > 0.$$

This proves the first claim and finishes the proof. □

Lemma 3.25. *There exists cut-offs η_1 and η_2 such that $N = [\eta_l, \eta_1]$, $B_s = [\eta_1, \eta_2]$ and $B_f = [\eta_2, \eta_h]$*

Proof. Let $\eta_1 = \sup\{\eta \in N\}$ and $\eta_2 = \inf\{\eta \in B_f\}$. Notice that given Lemma 3.24, the differences: $V_B^f(\eta) - V_B^s(\eta)$, $V_B^s(\eta) - V_N(\eta)$ and $V_B^f(\eta) - V_N(\eta)$ are all strictly increasing in η.

Then if $\eta \in B_f$, we have $V_B^f(\eta) > V_B^s(\eta)$ and $V_B^f(\eta) > V_N(\eta)$. Then as above differences are increasing in η, $\eta' > \eta$ implies $V_B^f(\eta') > V_B^s(\eta')$ and $V_B^f(\eta') > V_N(\eta')$. Hence $\eta' \in B_f$, which proves that $B_f = [\eta_2, \eta_h]$.

Similarly, if $\eta \in N$, then $V_N(\eta) > V_B^s(\eta)$ and $V_N(\eta) > V_B^f(\eta)$. Then as above differences are increasing in η, $\eta' < \eta$ implies $V_N(\eta') > V_B^s(\eta')$ and $V_N(\eta') > V_B^f(\eta')$. Hence $\eta' \in N$, which proves that $N = [\eta_l, \eta_1]$. The fact that $B_s = [\eta_1, \eta_2]$ follows immediately. □

Lemma 3.26. *There exists cut-offs η_3 and η_4 such that $S_f = [\eta_l, \eta_3]$, $S_s = [\eta_3, \eta_4]$ and $H = [\eta_4, \eta_h]$*

Proof. Let $\eta_3 = \sup\{\eta \in S_f\}$ and $\eta_4 = \inf\{\eta \in H\}$. Notice that the differences $V_H(\eta) - V_S^s(\eta)$, $V_H(\eta) - V_S^f(\eta)$, $V_S^s(\eta) - V_S^f(\eta)$ are all strictly increasing in η.

Then if $\eta \in H$, we have $V_H(\eta) > V_S^s(\eta)$ and $V_H(\eta) > V_S^f(\eta)$. Then as above differences are increasing in η, $\eta' > \eta$ implies $V_H(\eta') > V_S^s(\eta')$ and $V_H(\eta') > V_S^f(\eta)$. Hence $\eta' \in H$, which proves that $H = [\eta_4, \eta_h]$

Similarly, if $\eta \in S_s^f$, then $V_S^f(\eta) > V_H(\eta)$ and $V_S^f(\eta) > V_S^s(\eta)$. Then as above differences are increasing in η, $\eta' < \eta$ implies $V_S^f(\eta') > V_H(\eta')$ and $V_S^f(\eta') > V_S^s(\eta')$. Hence $\eta' \in S_f$, which proves that $S_f = [\eta_l, \eta_3]$. The fact that $S_s = [\eta_3, \eta_4]$ then follows. □

The structure follows from Lemmas 3.25 and 3.26. The fact that $\eta_1 < \eta_4$ follows from $S_f \cap B_s = \emptyset$ and $S_s \cap B_s = \emptyset$. The next lemma proves part (i) of the proposition

Lemma 3.27. *If $u(\eta_h) > 2(c_s + \theta)(\gamma + \rho)$, then there is no trade.*

Proof. Note that following two conditions is necessary for any trader to prefer trade in any equilibrium:

$$V_B^s(\eta_h) > V_N \tag{3.48}$$

$$V_S^s(0) > V_H(0) \tag{3.49}$$

The first of these implies

$$\frac{\sigma_s(V_{1,\eta_h} - p_s - c_s - \theta) + \gamma\, \mathbf{E}_{\eta'}(V_{0,\eta'})}{(\sigma_s + \gamma + \rho)} > \frac{\gamma\, \mathbf{E}_{\eta'}(V_{0,\eta'})}{(\gamma + \rho)},$$

which is equivalent to

$$(\gamma + \rho)\sigma_s(V_{1,\eta_h} - p_s - c_s - \theta) > \sigma_s\gamma\, \mathbf{E}_{\eta'}(V_{0,\eta'}).$$

Therefore,

$$(\gamma + \rho)(V_H(\eta_h) - p_s - c_s - \theta) > (\gamma + \rho)V_N \tag{3.50}$$

From 3.49 and using $u(0) = 0$:

$$\frac{\sigma_s(V_{0,0} + p_s - c_s - \theta) + \gamma\, \mathbf{E}_{\eta'}(V_{1,\eta'})}{(\sigma_s + \gamma + \rho)} > \frac{\gamma\, \mathbf{E}_{\eta'}(V_{1,\eta'})}{(\gamma + \rho)}$$

Therefore,

$$(\gamma + \rho)(V_N + p_s - c_s - \theta) > (\gamma + \rho)V_H(0) \tag{3.51}$$

Summing 3.50 and 3.51:

$$\begin{aligned} V_H(\eta_h) + V_N - 2(c_s + \theta) &> V_N + V_H(0) \\ V_H(\eta_h) - V_H(0) &> 2(c_s + \theta) \\ u(\eta_h) &> 2(c_s + \theta)(\gamma + \rho) \end{aligned}$$

This is what we wanted to show. □

The following lemma characterizes the condition under which fast venue is active in any equilibrium.

Lemma 3.28. *There is positive trade in fast venue only if*

$$u(\eta_h) > 2\frac{\sigma_f(\sigma_s + \gamma + \rho)(c_f + \theta) - \sigma_s(\sigma_f + \gamma + \rho)(c_s + \theta)}{\sigma_f - \sigma_s}.$$

Proof. Given the structure of cut-offs, one necessary condition for positive trade in the fast venue is $V_S^s(0) < V_S^f(0)$. As $u(0) = 0$, this is equivalent to

$$\frac{\sigma_s(V_N + p_s - c_s - \theta) + \gamma\, \mathbf{E}_{\eta'}(V_{1,\eta'})}{(\sigma_s + \gamma + \rho)} < \frac{\sigma_f(V_N + p_f - c_f - \theta) + \gamma\, \mathbf{E}_{\eta'}(V_{1,\eta'})}{(\sigma_f + \gamma + \rho)}$$

After some algebra, we obtain:

$$V_N - V_H(0) > -\frac{\sigma_f(\sigma_s + \gamma + \rho)(p_f - c_f - \theta) - \sigma_s(\sigma_f + \gamma + \rho)(p_s - c_s - \theta)}{(\sigma_f - \sigma_s)(\gamma + \rho)} \tag{3.52}$$

Another necessary condition is $V_B^s(\eta_h) < V_B^f(\eta_h)$. Doing similar algebra as above, we obtain

$$V_H(\eta_h) - V_N > \frac{\sigma_f(\sigma_s + \gamma + \rho)(p_f + c_f + \theta) - \sigma_s(\sigma_f + \gamma + \rho)(p_s + c_s + \theta)}{(\sigma_f - \sigma_s)(\gamma + \rho)} \tag{3.53}$$

Summing equations 3.52 and 3.53 finishes the proof

$$u(\eta_h) > 2\frac{\sigma_f(\sigma_s + \gamma + \rho)(c_f + \theta) - \sigma_s(\sigma_f + \gamma + \rho)(c_s + \theta)}{\sigma_f - \sigma_s}. \qquad \square$$

The next lemma calculates the densities f_h and f_{nh}.

Lemma 3.29.

$$f_{nh}(\eta) = \begin{cases} f(\eta)\frac{\sigma_f + (1-Z)\gamma}{\sigma_f + \gamma} & \textit{if } \eta \in [\eta_l, \eta_3] \\ f(\eta)\frac{\sigma_s + (1-Z)\gamma}{\sigma_s + \gamma} & \textit{if } \eta \in [\eta_3, \eta_4] \\ f(\eta)(1 - Z) & \textit{if } \eta \in [\eta_4, \eta_1] \\ f(\eta)\frac{(1-Z)\gamma}{\sigma_s + \gamma} & \textit{if } \eta \in [\eta_1, \eta_2] \\ f(\eta)\frac{(1-Z)\gamma}{\sigma_f + \gamma} & \textit{if } \eta \in [\eta_2, \eta_h] \end{cases}, \quad f_h(\eta) = \begin{cases} f(\eta)\frac{Z\gamma}{\sigma_f + \gamma} & \textit{if } \eta \in [\eta_l, \eta_3] \\ f(\eta)\frac{Z\gamma}{\sigma_s + \gamma} & \textit{if } \eta \in [\eta_3, \eta_4] \\ f(\eta)Z & \textit{if } \eta \in [\eta_4, \eta_1] \\ f(\eta)\frac{\gamma Z + \sigma_s}{\sigma_s + \gamma} & \textit{if } \eta \in [\eta_1, \eta_2] \\ f(\eta)\frac{\gamma Z + \sigma_f}{\sigma_f + \gamma} & \textit{if } \eta \in [\eta_2, \eta_h] \end{cases}$$

Proof. Let $\eta \in [\eta_l, \eta_3]$. The outflow of asset holders with valuation η is $\gamma f_h(\eta) + \sigma_f f_h(\eta)$, while the inflow is $\gamma Z f(\eta)$, hence we have:

$$\gamma f_h(\eta) + \sigma_f f_h(\eta) = \gamma Z f(\eta) \implies f_h(\eta) = \frac{\gamma Z}{\sigma_f + \gamma} f(\eta)$$

As $f_h(\eta) + f_{nh}(\eta) = f(\eta)$, we have:

$$f_{nh}(\eta) = \frac{\sigma_f + (1 - Z)\gamma}{\sigma_f + \gamma} f(\eta)$$

Let $\eta \in [\eta_3, \eta_4]$. Inflow-outflow balance requires:

$$\gamma f_h(\eta) + \sigma_s f_h(\eta) = \gamma Z f(\eta) \implies f_h(\eta) = \frac{\gamma Z}{\sigma_s + \gamma} f(\eta)$$

As $f_h(\eta) + f_{nh}(\eta) = f(\eta)$, we have:

$$f_{nh}(\eta) = \frac{\sigma_s + (1-Z)\gamma}{\sigma_s + \gamma} f(\eta)$$

Let $\eta \in [\eta_4, \eta_1]$. Inflow outflow balance for holders with valuation η requires:

$$\gamma f_h(\eta) = \gamma Z f(\eta) \implies f_h(\eta) = Z f(\eta)$$

Hence $f_{nh}(\eta) = (1-Z)f(\eta)$.

Let $\eta \in [\eta_1, \eta_2]$. Inflow-outflow balance requires:

$$\gamma f_{nh}(\eta) + \sigma_s f_{nh}(\eta) = (1-Z)\gamma f(\eta) \implies f_{nh}(\eta) = f(\eta)\frac{(1-Z)\gamma}{\gamma + \sigma_s}$$

As $f_h(\eta) + f_{nh}(\eta) = f(\eta)$, we have:

$$f_h(\eta) = f(\eta)\frac{\gamma Z + \sigma_s}{\gamma + \sigma_s}$$

Let $\eta \in [\eta_2, \eta_h]$. Inflow-outflow balance requires:

$$\gamma f_{nh}(\eta) + \sigma_f f_{nh}(\eta) = (1-Z)\gamma f(\eta) \implies f_{nh}(\eta) = f(\eta)\frac{(1-Z)\gamma}{\gamma + \sigma_f}$$

As $f_h(\eta) + f_{nh}(\eta) = f(\eta)$, we have:

$$f_h(\eta) = f(\eta)\frac{\gamma Z + \sigma_f}{\gamma + \sigma_f}. \qquad \square$$

Lemma 3.30. *In any equilibrium, asset market clearing conditions imply:*

$$(1-Z)F(\eta_1) + ZF(\eta_4) = 1 - Z \tag{3.54}$$

$$(1-Z)F(\eta_2) + ZF(\eta_3) = 1 - Z \tag{3.55}$$

Proof. We have two market clearing conditions: one for the slow venue and one for the fast venue. Fast venue clearing condition:

$$\int_{\eta_2}^{\eta_h} f_{nh}(\eta)d\eta = \int_{\eta_l}^{\eta_3} f_h(\eta)d\eta \tag{3.56}$$

$$(1 - F(\eta_2))\frac{\gamma(1-Z)}{\sigma_f + \gamma} = F(\eta_3)\frac{\gamma Z}{\sigma_f + \gamma} \tag{3.57}$$

$$\frac{1-Z}{Z} = \frac{F(\eta_3)}{1 - F(\eta_2)} \tag{3.58}$$

$$ZF(\eta_3) + (1-Z)F(\eta_2) = (1-Z). \tag{3.59}$$

From slow market clearing condition:

$$\int_{\eta_1}^{\eta_2} f_{nh}(\eta)d\eta = \int_{\eta_3}^{\eta_4} f_h(\eta)d\eta \tag{3.60}$$

$$(F(\eta_2) - F(\eta_1))\frac{\gamma(1-Z)}{\sigma_s+\gamma} = (F(\eta_4) - F(\eta_3))\frac{\gamma Z}{\sigma_s+\gamma} \tag{3.61}$$

$$\frac{1-Z}{Z} = \frac{F(\eta_4) - F(\eta_3)}{F(\eta_2) - F(\eta_1)} \tag{3.62}$$

$$(1-Z)F(\eta_1) + ZF(\eta_4) = (1-Z), \tag{3.63}$$

where the last line is obtained by plugging in fast market clearing equality. □

Lemma 3.31. *If $u(\eta_h) > 2\frac{\sigma_f(\sigma_s+\gamma+\rho)(c_f+\theta)-\sigma_s(\sigma_f+\gamma+\rho)(c_s+\theta)}{\sigma_f-\sigma_s}$, then in any equilibrium there is positive trade in both venues. The equilibrium cut-offs are given by:*

$$\frac{u(\eta_1) - u(\eta_4)}{\gamma+\rho} = 2(c_s+\theta) \tag{3.64}$$

$$\frac{u(\eta_2) - u(\eta_3)}{\gamma+\rho} = 2\frac{\sigma_f(c_f+\theta)(\sigma_s+\gamma+\rho) - \sigma_s(c_s+\theta)(\sigma_f+\gamma+\rho)}{(\sigma_f-\sigma_s)(\gamma+\rho)}. \tag{3.65}$$

Proof. Start with the equations

$$\begin{aligned} V_B^s(\eta_1) &= \frac{\sigma_s(V_H(\eta_1) - p_s - c_s - \theta) + \gamma\,\mathbf{E}_{\eta'}(V_{0,\eta'})}{(\sigma_s+\gamma+\rho)} \\ &= \frac{\sigma_s(V_H(\eta_1) - p_s - c_s - \theta) + (\gamma+\rho)V_N}{(\sigma_s+\gamma+\rho)} \\ &= \frac{\sigma_s(V_H(\eta_1) - V_N - p_s - c_s - \theta)}{(\sigma_s+\gamma+\rho)} + V_N \end{aligned}$$

In addition as η_1 is the cut-off type for buying slowly and doing nothing, we have $V_B^s(\eta_1) = V_N$[16]. Combining these:

$$V_H(\eta_1) = V_N + p_s + c_s + \theta \tag{3.66}$$

Similarly, as η_4 is the cut-off type for selling slowly and holding the asset, we have $V_H(\eta_4) = V_S^s(\eta_4)$. Hence

$$\begin{aligned} V_H(\eta_4) &= \frac{u(\eta_4) + \sigma_s(V_N + p_s - c_s - \theta) + \gamma\,\mathbf{E}_{\eta'}(V_{1,\eta'})}{(\sigma_s+\gamma+\rho)} \\ &= \frac{u(\eta_4) + \gamma\,\mathbf{E}_{\eta'}(V_{1,\eta'})}{(\sigma_s+\gamma+\rho)} + \frac{\sigma_s(V_N + p_s - c_s - \theta)}{(\sigma_s+\gamma+\rho)} \\ &= \frac{(\gamma+\rho)V_H(\eta_4)}{(\sigma_s+\gamma+\rho)} + \frac{\sigma_s(V_N + p_s - c_s - \theta)}{(\sigma_s+\gamma+\rho)}. \end{aligned}$$

[16] Notice that $V_N(\eta) = V_N$ for any η

The final equality implies

$$V_H(\eta_4) = V_N + p_s - c_s - \theta. \tag{3.67}$$

Using equations 3.66 and 3.67, we get $V_H(\eta_1) - V_H(\eta_4) = 2(c_s + \theta)$. Therefore

$$V_H(\eta_1) - V_H(\eta_4) = \frac{u(\eta_1) - u(\eta_4)}{\gamma + \rho} = 2(c_s + \theta).$$

Now, the equality $V_S^s(\eta_3) = V_S^f(\eta_3)$ implies

$$\frac{u(\eta_3) + \sigma_s(V_N + p_s - c_s - \theta) + \gamma\,\mathbf{E}_{\eta'}(V_{1,\eta'})}{(\sigma_s + \gamma + \rho)} = \frac{u(\eta_3) + \sigma_f(V_N + p_f - c_f - \theta) + \gamma\,\mathbf{E}_{\eta'}(V_{1,\eta'})}{(\sigma_f + \gamma + \rho)}. \tag{3.68}$$

Upon rearranging terms,

$$\begin{aligned}(\sigma_f - \sigma_s)(u(\eta_3) + \gamma\,\mathbf{E}_{\eta'}(V_{1,\eta'})) = (\sigma_f - \sigma_s)(\gamma + \rho)V_N \\ + \sigma_f(\sigma_s + \gamma + \rho)(p_f - c_f - \theta) \qquad (3.69) \\ - \sigma_s(\sigma_f + \gamma + \rho)(p_s - c_s - \theta). \qquad (3.70)\end{aligned}$$

Dividing by $(\sigma_f - \sigma_s)(\gamma + \rho)$:

$$\begin{aligned}V_N &= \frac{u(\eta_3)}{\gamma + \rho} + \frac{\gamma\,\mathbf{E}_{\eta'}(V_{1,\eta'})}{\gamma + \rho} - \frac{\sigma_f(\sigma_s + \gamma + \rho)(p_f - c_f - \theta) - \sigma_s(\sigma_f + \gamma + \rho)(p_s - c_s - \theta)}{(\sigma_f - \sigma_s)(\gamma + \rho)} \\ &= V_H(\eta_3) - \frac{\sigma_f(\sigma_s + \gamma + \rho)(p_f - c_f - \theta) - \sigma_s(\sigma_f + \gamma + \rho)(p_s - c_s - \theta)}{(\sigma_f - \sigma_s)(\gamma + \rho)} \qquad (3.71)\end{aligned}$$

Similarly, using $V_B^f(\eta_2) = V_B^s(\eta_2)$ yields

$$\frac{\sigma_f(V_H(\eta_2) - p_f - c_f - \theta) + \gamma\,\mathbf{E}_{\eta'}(V_{0,\eta'})}{(\sigma_f + \gamma + \rho)} = \frac{\sigma_s(V_H(\eta_2) - p_s - c_s - \theta) + \gamma\,\mathbf{E}_{\eta'}(V_{0,\eta'})}{(\sigma_s + \gamma + \rho)}$$

Rearranging and dividing by $(\sigma_f - \sigma_s)(\gamma + \rho)$:

$$V_H(\eta_2) - \frac{\sigma_f(\sigma_s + \gamma + \rho)(p_f + c_f + \theta) - \sigma_s(\sigma_f + \gamma + \rho)(p_s + c_s + \theta)}{(\sigma_f - \sigma_s)(\gamma + \rho)} = V_N \tag{3.72}$$

Solving 3.71 and 3.72 gives:

$$\frac{u(\eta_2) - u(\eta_3)}{\gamma + \rho} = 2\frac{\sigma_f(c_f + \theta)(\sigma_s + \gamma + \rho) - \sigma_s(c_s + \theta)(\sigma_f + \gamma + \rho)}{(\sigma_f - \sigma_s)(\gamma + \rho)}$$

Note that if $u(\eta_h) > 2\frac{\sigma_f(\sigma_s+\gamma+\rho)(c_f+\theta) - \sigma_s(\sigma_f+\gamma+\rho)(c_s+\theta)}{\sigma_f - \sigma_s}$, then there exists $\eta_2 < \eta_h$ and $\eta_3 > 0$ such that equation above holds. Moreover, any type $\eta < \eta_3$ and $\eta > \eta_2$ strictly prefers fast venue to slow venue and there is positive trade in the fast venue. □

Next, note that $u(\eta_h) > 2\frac{\sigma_f(\sigma_s+\gamma+\rho)(c_f+\theta)-\sigma_s(\sigma_f+\gamma+\rho)(c_s+\theta)}{\sigma_f-\sigma_s}$ implies $u(\eta_h) > 2(c_s+\theta)(\gamma+\rho)$ whenever $c_f > c_s$, which is assumed. Thus, whenever the former inequality holds, η_1, η_2, η_3 and η_4 that solves equations 3.54, 3.55, 3.64 and 3.65 constitutes equilibrium cut-offs where both venues are active. To finish the proof of part (ii) of theorem 3.3 we characterize the equilibrium prices in case (ii). From equation 3.67, we obtain the characterization of p_s:

$$\begin{aligned} p_s &= V_h(\eta_4) - V_N + c_s + \theta \\ &= \frac{u(\eta_4) + \gamma\,\mathbf{E}_{\eta'}(V_{1,\eta'})}{\gamma+\rho} + c_s + \theta - \frac{\gamma\,\mathbf{E}_{\eta'}(V_{0,\eta'})}{\gamma+\rho} \\ &= \frac{u(\eta_4)}{\gamma+\rho} + c_s + \theta + \frac{\gamma}{\gamma+\rho}(\mathbf{E}_{\eta'}(V_{1,\eta'}) - \mathbf{E}_{\eta'}(V_{0,\eta'})) \end{aligned} \tag{3.73}$$

From equation (3.71):

$$\frac{\sigma_f(\sigma_s+\gamma+\rho)(p_f-c_f-\theta)-\sigma_s(\sigma_f+\gamma+\rho)(p_s-c_s-\theta)}{(\sigma_f-\sigma_s)(\gamma+\rho)} = \frac{u(\eta_3)}{\gamma+\rho} + \frac{\gamma}{\gamma+\rho}(\mathbf{E}_{\eta'}(V_{1,\eta'}) - \mathbf{E}_{\eta'}(V_{0,\eta'})) \tag{3.74}$$

Lemma 3.32. *Equations* (3.73) *and* (3.74) *characterize a unique price vector.*

Proof. We only need to show that (3.73) and (3.74) lead to unique solutions for p_s and p_f. In particular, we will show that the quantity $\mathbf{E}_{\eta'}[V_{1,\eta'}] - \mathbf{E}_{\eta'}[V_{0,\eta'}]$ does not depend on either p_s nor p_f; once this is done, p_s is directly pinned down by (3.73), and after substituting into (3.74), p_f is also uniquely determined. To finish up, we will need to show that the numbers p_s and p_f recovered this way are positive: this is done in the last step of the proof.

To begin, in the following two steps we establish that $\mathbf{E}_{\eta'}[V_{1,\eta'}] - \mathbf{E}_{\eta'}[V_{0,\eta'}]$ only depends on the endogenous thresholds η_1, η_2, η_3 and η_4. First recall that

$$\begin{aligned} V_S^f(\eta) &= \frac{u(\eta) + \sigma_f(V_N + p_f - c_f - \theta) + \gamma\,\mathbf{E}_{\eta'}[V_{1,\eta'}]}{\sigma_f+\rho+\gamma} \\ V_S^s(\eta) &= \frac{u(\eta) + \sigma_s(V_N + p_s - c_s - \theta) + \gamma\,\mathbf{E}_{\eta'}[V_{1,\eta'}]}{\sigma_s+\rho+\gamma} \\ V_B^f(\eta) &= \frac{\sigma_f(V_H(\eta) - p_f - c_f - \theta) + \gamma\,\mathbf{E}_{\eta'}[V_{0,\eta'}]}{\sigma_f+\rho+\gamma} \\ V_B^s(\eta) &= \frac{\sigma_s(V_H(\eta) - p_s - c_s - \theta) + \gamma\,\mathbf{E}_{\eta'}[V_{0,\eta'}]}{\sigma_s+\rho+\gamma} \\ V_H(\eta) &= \frac{u(\eta) + \gamma\,\mathbf{E}_{\eta'}[V_{1,\eta'}]}{\gamma+\rho} \end{aligned}$$

and

$$V_N = \frac{\gamma\, \mathbf{E}_{\eta'}[V_{0,\eta'}]}{\gamma + \rho}. \tag{3.75}$$

Next we derive $\mathbf{E}_{\eta'}[V_{0,\eta'}]$ and $\mathbf{E}_{\eta'}[V_{1,\eta'}]$ in closed form. First, we will compute $\mathbf{E}_{\eta'}[V_{0,\eta'}]$; using the previous displayed equations, we have

$$\frac{dV_{0,\eta}}{d\eta} = \mathbf{1}_{\{\eta\in[\eta_1,\eta_2]\}}\frac{\sigma_s u'(\eta)}{(\sigma_s + \rho + \gamma)(\rho + \gamma)} + \mathbf{1}_{\{\eta\in[\eta_2,\eta_h]\}}\frac{\sigma_f u'(\eta)}{(\sigma_f + \rho + \gamma)(\rho + \gamma)}. \tag{3.76}$$

Integrating (3.76),

$$\begin{aligned} V_{0,\eta} = {} & \mathbf{1}_{\{\eta\in[\eta_l,\eta_1]\}}V_N + \mathbf{1}_{\{\eta\in[\eta_1,\eta_2]\}}\left[V_N + \int_{\eta_1}^{\eta}\frac{\sigma_s u'(\xi)}{(\sigma_s + \rho + \gamma)(\rho + \gamma)}d\xi\right] \\ & + \mathbf{1}_{\{\eta\in[\eta_2,\eta_h]\}}\left[V_N + \int_{\eta_1}^{\eta_2}\frac{\sigma_s u'(\xi)}{(\sigma_s + \rho + \gamma)(\rho + \gamma)}d\xi + \int_{\eta_2}^{\eta}\frac{\sigma_f u'(\xi)}{(\sigma_f + \rho + \gamma)(\rho + \gamma)}d\xi\right]. \end{aligned} \tag{3.77}$$

Finally, taking an expectation from (3.77) gives

$$\begin{aligned} \mathbf{E}_{\eta'}[V_{0,\eta'}] = {} & V_N + \int_{\eta_1}^{\eta_2}(\int_{\eta_1}^{\tilde{\eta}}\frac{\sigma_s u'(\xi)}{(\sigma_s + \rho + \gamma)(\rho + \gamma)}d\xi)dF(\tilde{\eta}) \\ & + (\int_{\eta_1}^{\eta_2}\frac{\sigma_s u'(\xi)}{(\sigma_s + \rho + \gamma)(\rho + \gamma)}d\xi)(1 - F(\eta_2)) \\ & + \int_{\eta_2}^{\eta_h}(\int_{\eta_2}^{\tilde{\eta}}\frac{\sigma_f u'(\xi)}{(\sigma_f + \rho + \gamma)(\rho + \gamma)}d\xi)dF(\tilde{\eta}). \end{aligned} \tag{3.78}$$

Thus, (3.78) shows that $\mathbf{E}_{\eta'}[V_{0,\eta'}]$ only depends on the endogenous thresholds η_1 and η_2.

Next we will compute $\mathbf{E}_{\eta'}[V_{1,\eta'}]$ using (3.75). Start with

$$\frac{dV_{1,\eta}}{d\eta} = \mathbf{1}_{\{\eta\in[\eta_l,\eta_3]\}}\frac{u'(\eta)}{\sigma_f + \rho + \gamma} + \mathbf{1}_{\{\eta\in[\eta_3,\eta_4]\}}\frac{u'(\eta)}{\sigma_s + \rho + \gamma} + \mathbf{1}_{\{\eta\in[\eta_4,\eta_h]\}}\frac{u'(\eta)}{\rho + \gamma}. \tag{3.79}$$

Integrating this formula,

$$\begin{aligned} V_{1,\eta} = {} & \mathbf{1}_{\{\eta\in[\eta_l,\eta_3]\}}\int_{\eta_l}^{\eta}\frac{u'(\xi)}{\sigma_f + \rho + \gamma}d\xi \\ & + \mathbf{1}_{\{\eta\in[\eta_3,\eta_4]\}}(\int_{\eta_l}^{\eta_3}\frac{u'(\xi)}{\sigma_f + \rho + \gamma}d\xi + \int_{\eta_3}^{\eta}\frac{u'(\xi)}{\sigma_s + \rho + \gamma}d\xi) \\ & + \mathbf{1}_{\{\eta\in[\eta_4,\eta_h]\}}(\int_{\eta_l}^{\eta_3}\frac{u'(\xi)}{\sigma_f + \rho + \gamma}d\xi + \int_{\eta_3}^{\eta_4}\frac{u'(\xi)}{\sigma_s + \rho + \gamma}d\xi + \int_{\eta_4}^{\eta}\frac{u'(\xi)}{\gamma + \rho}d\xi). \end{aligned} \tag{3.80}$$

Therefore, taking an expectation from (3.80) gives

$$\begin{aligned}\mathbf{E}_{\eta'}[V_{1,\eta'}] = &\int_{\eta_l}^{\eta_3}\Big(\int_{\eta_l}^{\tilde{\eta}}\frac{u'(\xi)}{\sigma_f+\rho+\gamma}d\xi\Big)dF(\tilde{\eta}) + \Big(\int_{\eta_l}^{\eta_3}\frac{u'(\xi)}{\sigma_f+\rho+\gamma}d\xi\Big)(1-F(\eta_3)) \\ &+\int_{\eta_3}^{\eta_4}(\int_{\eta_3}^{\tilde{\eta}}\frac{u'(\xi)}{\sigma_s+\rho+\gamma}d\xi)dF(\tilde{\eta}) + (1-F(\eta_4))(\int_{\eta_3}^{\eta_4}\frac{u'(\xi)}{\sigma_s+\rho+\gamma}d\xi) \\ &+\int_{\eta_4}^{\eta_h}(\int_{\eta_4}^{\tilde{\eta}}\frac{u'(\xi)}{\gamma+\rho}d\xi)dF(\tilde{\eta})\end{aligned}$$

Thus, (3.81) shows that $\mathbf{E}_{\eta'}[V_{1,\eta'}]$ only depends on the endogenous thresholds η_3 and η_4. Together, (3.78) and (3.81) finish the proof that the difference $\mathbf{E}_{\eta'}[V_{1,\eta'}] - \mathbf{E}_{\eta'}[V_{0,\eta'}]$ does not depend on p_s nor p_f.

To finish the proof of the lemma, note that equations (3.73) and (3.74) yield unique solutions for p_s and p_f. Since $\mathbf{E}_{\eta'}[V_{1,\eta'}] - \mathbf{E}_{\eta'}[V_{0,\eta'}] > 0$, all terms in (3.73) are positive, and hence $p_s > 0$. Since we have assumed that $u(\cdot) \geq 0$, it follows that p_f is also positive, which finishes the lemma. □

To prove part (iii), assume $u(\eta_h) > 2(c_s+\theta)(\gamma+\rho)$ and

$$u(\eta_h) < 2\frac{\sigma_f(\sigma_s+\gamma+\rho)(c_f+\theta) - \sigma_s(\sigma_f+\gamma+\rho)(c_s+\theta)}{\sigma_f-\sigma_s}.$$

From Lemma 3.28, we know that there cannot be any trade in the fast venue, i.e. $\eta_3 = 0$ and $\eta_2 = \eta_h$. Lemmas 3.29 and 3.30 still hold, and given $u(\eta_h) > 2(c_s+\theta)(\gamma+\rho)$, following the same steps in lemma 3.31, we see that the cut-offs η_1 and η_4 are uniquely pinned down by:

$$(1-Z)F(\eta_4) + ZF(\eta_1) = 1 - Z \tag{3.81}$$

$$\frac{u(\eta_1)-u(\eta_4)}{\gamma+\rho} = 2(c_s+\theta) \tag{3.82}$$

As in earlier case, the equilibrium price in the slow venue is given by:

$$p_s = \frac{u(\eta_4)}{\gamma+\rho} + c_s + \theta + \frac{\gamma}{\gamma+\rho}(\mathbf{E}_{\eta'}(V_{1,\eta'}) - \mathbf{E}_{\eta'}(V_{0,\eta'})),$$

To show that this is indeed an equilibrium, we need to find a price p_f such that there is no demand for trade in the fast venue (otherwise, fast venue clearing condition cannot hold.). To see that, there is no demand for selling in the fast venue if $V_S^s(0) - V_S^f(0) \geq 0$. This corresponds to:

$$\begin{aligned}(V_H(0)-V_N)(\sigma_f-\sigma_s)(\gamma+\rho) + \sigma_s(\sigma_f+\gamma+\rho)(p_s-c_s-\theta)& \\ -(p_f-c_f-\theta)\sigma_f(\sigma_s+\gamma+\rho) &\geq 0\end{aligned}$$

Which is equivalent to:

$$\begin{aligned} L = &(V_H(0) - V_N)(\sigma_f - \sigma_s)(\gamma + \rho) + \sigma_s(\sigma_f + \gamma + \rho)(p_s - c_s - \theta) \\ &+ (c_f + \theta)\sigma_f(\sigma_s + \gamma + \rho) \\ &\geq p_f(\sigma_f(\sigma_s + \gamma + \rho)) \end{aligned} \tag{3.83}$$

Similarly, there is no demand for buying in the fast venue if $V^s_B(\eta_h) - V^f_B(\eta_h) \geq 0$. This is equivalent to:

$$\begin{aligned} U = &(V_H(\eta_h) - V_N)(\sigma_f - \sigma_s)(\gamma + \rho) + \sigma_s(\sigma_f + \gamma + \rho)(p_s + c_s + \theta) \\ &- (c_f + \theta)\sigma_f(\sigma_s + \gamma + \rho) \\ &\leq p_f(\sigma_f(\sigma_s + \gamma + \rho)) \end{aligned} \tag{3.84}$$

Note that there exists a p_f such that there is no demand in selling or buying fast if $L > U$. The following condition condition is sufficient to have $L > U$

$$u(\eta_h) < 2\frac{\sigma_f(\sigma_s + \gamma + \rho)(c_f + \theta) - \sigma_s(\sigma_f + \gamma + \rho)(c_s + \theta)}{\sigma_f - \sigma_s}.$$

This condition holds under the assumptions of (iii), thus $L > U$. Hence, under any $p_f \in (U, L)$, there is no demand for fast venue and p_f fast venue clearing condition is satisfied, finishing the characterization of the equilibrium and Theorem 3.3.

Proof of Proposition 3.4

First, we proof a short lemma:

Lemma 3.33. *Let x be a variable. Then if*

$$\frac{\partial(u(\eta_2) - u(\eta_3))}{\partial x} > (<)0$$

Then $\frac{\partial \eta_2}{\partial x} > (<)0$ and $\frac{\partial \eta_3}{\partial x} < (>)0$.

Similarly, if

$$\frac{\partial(u(\eta_1) - u(\eta_4))}{\partial x} > (<)0,$$

then $\frac{\partial \eta_1}{\partial x} > (<)0$ and $\frac{\partial \eta_4}{\partial x} < (>)0$.

Proof. We prove this for the first case, rest is similar. From equation 3.12, we see that if $\frac{\partial(u(\eta_2) - u(\eta_3))}{\partial x} > 0$, this is only possible under $\frac{\partial u(\eta_2)}{\partial x} > 0$ and $\frac{\partial u(\eta_3)}{\partial x} < 0$.[17] Then $\frac{\partial \eta_2}{\partial x} > 0$ and $\frac{\partial \eta_3}{\partial x} < 0$ follows from the fact that u is strictly increasing. □

[17] It is clear that one of these must hold. To see why both are necessary, see that equation 3.12 requires cut-offs to move in opposite direction.

Lemma 3.34. *We have following comparative statics:*

1. $\frac{\partial \eta_1}{\partial \sigma_s} = \frac{\partial \eta_4}{\partial \sigma_s} = 0$, $\frac{\partial \eta_2}{\partial \sigma_s} > 0$, $\frac{\partial \eta_3}{\partial \sigma_s} < 0$
2. $\frac{\partial \eta_4}{\partial \sigma_f} = 0$, $\frac{\partial \eta_1}{\partial \sigma_f} = 0$, $\frac{\partial \eta_2}{\partial \sigma_f} < 0$, $\frac{\partial \eta_3}{\partial \sigma_f} > 0$
3. $\frac{\partial \eta_1}{\partial \gamma} > 0$, $\frac{\partial \eta_4}{\partial \gamma} < 0$, $\frac{\partial \eta_2}{\partial \gamma} > 0$, $\frac{\partial \eta_3}{\partial \gamma} < 0$
4. $\frac{\partial \eta_1}{\partial \rho} > 0$, $\frac{\partial \eta_4}{\partial \rho} < 0$, $\frac{\partial \eta_2}{\partial \rho} > 0$, $\frac{\partial \eta_3}{\partial \rho} < 0$
5. $\frac{\partial \eta_1}{\partial c_s} < 0$, $\frac{\partial \eta_4}{\partial c_s} > 0$, $\frac{\partial \eta_2}{\partial c_s} < 0$, $\frac{\partial \eta_3}{\partial c_s} > 0$
6. $\frac{\partial \eta_1}{\partial c_f} = 0$, $\frac{\partial \eta_4}{\partial c_f} = 0$, $\frac{\partial \eta_2}{\partial c_f} > 0$, $\frac{\partial \eta_3}{\partial c_f} < 0$
7. $\frac{\partial \eta_1}{\partial \theta} < 0$, $\frac{\partial \eta_4}{\partial \theta} > 0$, $\frac{\partial \eta_2}{\partial \theta} > 0$, $\frac{\partial \eta_3}{\partial \theta} < 0$

Proof. **Part 1.** First two equations are trivial as σ_s does not appear in equations that determine η_1 and η_4. For the last two:

$$\frac{\partial(u(\eta_2) - u(\eta_3))}{\partial \sigma_s} = 2\frac{(c_f - c_s)\sigma_f(\sigma_f + \gamma + \rho)}{(\sigma_f - \sigma_s)^2} > 0$$

The result follows from lemma 3.33.

Part 2 First two equations are trivial as σ_s does not appear in equations that determine η_1 and η_4. We have:

$$\frac{\partial(u(\eta_2) - u(\eta_3))}{\partial \sigma_f} = -2\frac{(c_f - c_s)\sigma_f(\sigma_f + \gamma + \rho)}{(\sigma_f - \sigma_s)^2} < 0$$

The result follows from lemma 3.33.

Parts 3 and 4

$$\frac{\partial(u(\eta_2) - u(\eta_3))}{\partial \gamma} = 2\frac{c_f\sigma_f - c_s\sigma_s}{\sigma_f - \sigma_s} > 0$$

$$\frac{\partial(u(\eta_1) - u(\eta_4))}{\partial \gamma} = 2c_s > 0$$

The result follows from lemma 3.33 and the proof for ρ is exactly same.

Part 5

$$\frac{\partial(u(\eta_2) - u(\eta_3))}{\partial c_s} = -2\frac{\sigma_s(\sigma_f + \gamma + \rho)}{(\sigma_f - \sigma_s)(\gamma + \rho)} < 0$$

$$\frac{\partial(u(\eta_1) - u(\eta_4))}{\partial c_s} = 2(\gamma + \rho) > 0$$

The result follows from lemma 3.33.

Part 6

$$\frac{\partial(u(\eta_2) - u(\eta_3))}{\partial c_f} = \frac{\sigma_f(\sigma_s + \gamma + \rho)}{(\sigma_f - \sigma_s)(\gamma + \rho)} > 0$$
$$\frac{\partial(u(\eta_1) - u(\eta_4))}{\partial c_f} = 0$$

The result follows from lemma 3.33.

Part 7

$$\frac{\partial(u(\eta_2) - u(\eta_3))}{\partial \theta} = 2 > 0$$
$$\frac{\partial(u(\eta_1) - u(\eta_4))}{\partial \theta} = 2 > 0$$

Note that $\frac{\partial \eta_2}{\partial c_s} < 0$ and $\frac{\partial \eta_3}{\partial c_s} > 0$ implies m_f is increasing in c_s. Then TV_f is increasing in c_s. The proof for σ_f is exactly same. Likewise, $\frac{\partial \eta_2}{\partial c_f} > 0$ and $\frac{\partial \eta_3}{\partial c_f} < 0$ imply that m_f is decreasing in c_f. Then TV_f is decreasing in c_f. The proof for σ_s and θ is exactly same. Similarly, $\frac{\partial \eta_1}{\partial c_f} < 0$ and $\frac{\partial \eta_4}{\partial c_f} > 0$ imply that m_s is increasing in c_f. Then TV_s is increasing in c_f. The proof for σ_s is exactly same. Finally, $\frac{\partial \eta_1}{\partial c_s} > 0$ and $\frac{\partial \eta_4}{\partial c_s} < 0$ imply m_s is decreasing in c_s. Then TV_s is decreasing in c_s. The proof for σ_f is exactly same.

Part 7 of above the lemma implies that increasing θ results in some types switching from fast venue to slow venue and some types switching from slow venue to no trade. Thus, trading volume is decreasing in θ. □

Proof of Proposition 3.8

Derivatives of the revenue in fast and slow venues are

$$\frac{\partial R_s}{\partial c_s} = \frac{4\gamma\sigma_f\sigma_s(\sigma_s + \gamma + \rho)(1 - Z)Z}{(\gamma + \sigma_f)(\sigma_f - \sigma_s)}(c_f - 2c_s) \tag{3.85}$$
$$\frac{\partial R_f}{\partial c_f} = \frac{\text{numerator}}{\text{denominator}} \tag{3.86}$$

where

$$\text{numerator} = 2\gamma\sigma_f Z(1 - Z)(-\sigma_s + 2\sigma_s(\gamma + \rho)(c_s + \theta) - \sigma_f(4c_f(\sigma_s + \gamma + \rho) + 2(\gamma + \rho)\theta - 1 - 2c_s\sigma_S))$$

and the denominator is $(\gamma + \rho)(\sigma_f - \sigma_s)$.

Next note that $\frac{\partial^2 R_s}{\partial c_s^2} < 0$ and $\frac{\partial^2 R_f}{\partial c_f^2} < 0$, so the fee competition game has unique interior optimum. Moreover, the derivative of R_f evaluated at $c_s = 0$ is equal to

$$\frac{\partial R_f}{\partial c_f}|_{c_s=0} \propto (\sigma_f - \sigma_s)(1 - 2(\gamma + \rho)\theta) - \sigma_f(4c_f(\sigma_s + \gamma + \rho)).$$

The first term is positive, thus $\frac{\partial R_f}{\partial c_f}|_{c_s=0} > 0$ at $c_f = 0$. Thus whenever $c_s = 0$, $c_f > 0$. But $\frac{\partial R_s}{\partial c_s} > 0$ at $c_s = 0$ and $c_f > 0$, so there cannot be any equilibrium where $c_s = 0$. Thus c_s must be interior in any equilibrium. Rearranging 3.85 and suppressing the dependence on primitives,

$$c_s^*(c_f) = \frac{c_f}{2}. \tag{3.87}$$

Thus in any equilibrium, $c_s > 0$ and $c_f > 0$. Solving (3.86) and (3.87) together

$$c_f^*(\sigma_f, \sigma_s) = (1 - 2\theta(\gamma + \rho))\frac{\sigma_f - \sigma_s}{\gamma(4\sigma_f - \sigma_s) + \rho(4\sigma_f - \sigma_s) + 3\sigma_f\sigma_s}$$

As the revenues converge to zero when prices are high, the unique equilibrium is given by c_s^* and c_f^*.

Proof of Proposition 3.9

Taking the derivative of equilibrium fee with respect to σ_f:

$$\frac{\partial c_f^*}{\partial \sigma_f} = -\frac{3\sigma_s(\sigma_s + \gamma + \rho)(-1 + 2(\gamma + \rho)\theta)}{((\gamma + \rho - 3\sigma_f)\sigma_s - 4(\gamma + \rho)\sigma_f)^2} > 0$$

The inequality is due to Assumption 3.5, which guarantees $(-1 + 2(\gamma + \rho)\theta) < 0$. This also shows that $\frac{\partial c_s^*}{\partial \sigma_f} > 0$ as $c_s^* = \frac{c_f^*}{2}$. To prove the second part:

$$\frac{\partial c_f^*}{\partial \sigma_s} = \frac{3\sigma_f(\sigma_f + \gamma + \rho)(-1 + 2(\gamma + \rho)\theta)}{((\gamma + \rho - 3\sigma_f)\sigma_s - 4(\gamma + \rho)\sigma_f)^2} < 0,$$

where the inequality holds for the same reason. Similarly, $\frac{\partial c_s^*}{\partial \sigma_s} < 0$ as $c_s^* = \frac{c_f^*}{2}$.

Proof of Proposition 3.10

$$\frac{\partial TV_s}{\partial \sigma_s} = A\frac{2(1 - Z)Z(1 - 2(\gamma + \rho)\theta)\gamma\sigma_f}{(\sigma_s + \gamma)^2(-3\sigma_f\sigma_s + \gamma(-4\sigma_f + \sigma_s) - 4\sigma_f\rho + \sigma_s\rho)^2}$$

where

$$A = 4\gamma^3\sigma_f + \sigma_s\rho(\sigma_f + \rho) + 8\gamma^2\sigma_f(\sigma_s + \rho) + \gamma(\sigma_s\rho + 4\sigma_f(\sigma_s + \rho)^2).$$

Assumption 3.5 guarantees $(1-2(\gamma+\rho)\theta)>0$, so second factor is positive. Of course, A is also positive, which shows $\frac{\partial TV_s}{\partial \sigma_s}>0$. Likewise, the same assumption also shows

$$\frac{\partial TV_f}{\partial \sigma_s}=\frac{4\gamma\sigma_f^2(\gamma+\rho)(\sigma_f+\gamma+\rho)(1-2(\gamma+\rho)\theta)(1-Z)Z}{(\gamma+\sigma_f)((\gamma+\rho-3\sigma_f)\sigma_s-4(\gamma+\rho)\sigma_f)^2}>0.$$

Proof of Proposition 3.11

$$\frac{\partial TV}{\partial \sigma_f}=-A\frac{2\gamma(\sigma_s+\gamma+\rho)(1-2(\gamma+\rho)\theta)(1-Z)Z}{(\gamma+\sigma_f)^2(\gamma+\sigma_s)(-3\sigma_s\sigma_f+\gamma(\sigma_s-4\sigma_f)-4\sigma_f\rho+\sigma_s\rho)^2}$$

where

$$\begin{aligned}A=\gamma^3(-8\sigma_f^2+4\sigma_f\sigma_s+\sigma_s^2)+3\sigma_f^2\sigma_s^2\rho+\gamma^2(6\sigma_f\sigma_s(-2\sigma_f+\sigma_s)\\+(-8\sigma_f^2+4\sigma_f\sigma_s+\sigma_s^2)\rho)-3\gamma\sigma_f\sigma_s(-2\sigma_s\rho+\sigma_f(\sigma_s+2\rho)).\end{aligned}$$

The term multiplying against A is negative, so the sign of $\partial\,\mathrm{TV}/\partial\sigma_f$ is the sign of $-A$. Let $\sigma_f=\sigma$ and $\sigma_s=\sigma-\epsilon$. After some algebra:

$$\mathrm{sign}(\frac{\partial TV}{\partial \sigma_f})=\mathrm{sign}(\gamma^2+\gamma(\rho+\sigma)-\rho\sigma+\mathcal{O}(\epsilon))$$

Letting $\epsilon\to 0$ and rearranging,

$$\mathrm{sign}(\frac{\partial TV}{\partial \sigma_f})=\mathrm{sign}(\rho(\gamma-\sigma)+\gamma(\gamma+\sigma))).$$

If $\gamma>\sigma$, then clearly $\frac{\partial TV}{\partial \sigma_f}>0$. Conversely, if $\gamma<\sigma$, then

$$\frac{\partial TV}{\partial \sigma_f}>0\iff \rho<\frac{\gamma^2+\gamma\sigma}{\sigma-\gamma}.$$

This finishes the proof.

Proof of Proposition 3.12

Note that conditional on the speed choice of the entrant, the outcome of fee competition stage (equilibrium fees and revenues) is characterized in Section 3.3.3. Let $R(\sigma)$ denote the revenue of the entrant when she chooses σ. Let $\Pi(\sigma,\alpha)=R(\sigma)-\alpha K(\sigma)$ and $\Pi^*(\alpha)=\max_\sigma \Pi(\sigma)$. The firm enters whenever $\Pi^*(\alpha)>0$. For $\alpha=0$, entry is cost-less thus is always optimal and $\Pi(\sigma,\alpha)<0$ for high α. Moreover, $\Pi^*(\alpha)$ is decreasing in α.[18] Thus first part of the result follows.

We first show that the speed is weakly increasing in technological improvements. Let

[18] To see that, notice that if $\alpha_1>\alpha_2$, then $\Pi(\sigma,\alpha_1)<\Pi(\sigma,\alpha_2)$ for all σ.

$R(\sigma_n(\alpha), \alpha)$ denote the equilibrium revenue of the new firm where $\sigma_n(\alpha)$ denotes the optimal speed choice of the firm. $\pi(\sigma_n(\alpha), \alpha)$ denotes the equilibrium profit of the new firm. We will show that, if $\alpha > \alpha'$ then $\sigma(\alpha') > \sigma(\alpha)$. To see that, let $\sigma \in (0, \sigma_n(\alpha))$. Then,

$$\begin{aligned}
\pi(\sigma_n(\alpha), \alpha') - \pi(\sigma, \alpha') &= R(\sigma_n(\alpha), \alpha') - R(\sigma, \alpha') - \alpha(K(\sigma_n(\alpha)) - K(\sigma)) \\
&= R(\sigma_n(\alpha), \alpha) - R(\sigma, \alpha) - \alpha(K(\sigma_n(\alpha)) - K(\sigma)) \\
&> R(\sigma_n(\alpha), \alpha) - R(\sigma, \alpha) - \alpha'(K(\sigma_n(\alpha)) - K(\sigma)) \\
&= \pi(\sigma_n(\alpha), \alpha) - \pi(\sigma, \alpha) > 0.
\end{aligned}$$

This shows that, under α', $\sigma_n(\alpha)$ gives a higher profit than any $\sigma \in (0, \sigma_n(\alpha))$, thus $\sigma_n(\alpha') \geq \sigma_n(\alpha)$

To see why the inequality is strict, notice that at the optimum following FOC must be satisfied:

$$\frac{\partial R(\sigma_n(\alpha), \alpha)}{\partial \sigma} - \alpha \frac{\partial K(\sigma_n(\alpha))}{\partial \sigma} = 0.$$

As $\alpha' < \alpha$, we have

$$\frac{\partial R(\sigma_n(\alpha), \alpha')}{\partial \sigma} - \alpha' \frac{\partial K(\sigma_n(\alpha))}{\partial \sigma} \neq 0.$$

Thus $\sigma_n(\alpha) \neq \sigma_n(\alpha)$. As we have already showed $\sigma_n(\alpha') \geq \sigma_n(\alpha)$, the result follows.

Proof of Lemma 3.14

Solve the linear system of equations for $\eta_1, \ldots, \eta_4$ and substitute into the expression for $R(\sigma) = c_f^*(\sigma, \tau) \cdot \mathrm{TV}_f(\sigma_f)$.

Proof of Lemma 3.15

Lemma 3.14 shows that $R(\sigma)$ is a rational function of σ; thus it is smooth. It also follows that all of its derivatives are rational functions, so that $R''(\sigma) = p(\sigma)/q(\sigma)$ where the polynomials p and q are easily computed. A direct calculation shows $q > 0$ and $p(\sigma) < 0$ for all $\sigma \geq \tau$.[19] This finishes the proof.

Proof of Lemma 3.3.4

The first part is straightforward: when $\alpha = 0$, if the entrant enters at all, she would choose $\sigma_f = \infty$ as there is no downside to doing so and larger σ_f allows for higher costs and therefore higher profits. Thus, the only thing to check is that $\sigma_f = \infty$ leads to a positive profit (i.e., it is better than not entering at all): this is ensured by Theorem 3.3 (see also the corollary following Proposition 3.8).

[19] The calculation for q is immediate; for p, write $\sigma = \lambda\tau$ and note that $p(\lambda\tau) < 0$ for all $\lambda \geq 1$.

To prove the second part of the corollary, note that the formula in 3.14 implies that $R(\sigma)$ tends to finite limit as $\sigma \to \infty$, whence any positive α induces a finite σ_f.

Proof of Proposition 3.16

Suppose (a) does not hold, so that $\pi(\sigma) \geq 0$; the goal is to show that condition (b) holds. Since $\pi(\tau) < 0$ and π is concave (c.f., Lemma 3.15), the optimum σ^* is the unique solution to

$$\pi'(\sigma^*) = 0 \implies R'(\sigma^*) = \alpha.$$

Therefore, concavity of π implies that if $R'(\kappa\tau) \geq \alpha$, then π is strictly increasing on $[\tau, \kappa\tau]$, so that $\pi(\kappa\tau) > \pi(\sigma) \geq 0$ and condition (b) is satisfied.

The remaining case to examine is $R'(\kappa\tau) < \alpha$, which we will rule out by way of contradiction. Towards that goal, suppose there exist parameter values for which $R'(\kappa\tau) < \alpha$ and put $A := \gamma$ and $B = \gamma + \rho$ for tidiness ($0 < A < B$). By Lemma 3.14, we have

$$R'(\kappa\tau) = 112C \cdot \frac{49\tau^2 + 7(7A + 6B)\tau + 48AB}{9(4A + 7\tau)^2(8B + 7\tau)^3}, \tag{3.88}$$

Now, by assumption $\sigma \geq \tau > 0$, whence

$$0 \leq \pi(\sigma) = R(\sigma) - \alpha\sigma < R(\sigma) - R'(\kappa\tau)\sigma.$$

The function $\sigma \mapsto R(\sigma) - R'(\kappa\tau)\sigma$ is strictly concave by Lemma 3.15 and is maximized by setting its derivative to zero, i.e., at $\sigma = \kappa\tau$. Therefore, $0 < R(\sigma) - R'(\kappa\tau)\sigma < R(\kappa\tau) - R'(\kappa\tau)\kappa\tau$; upon transposing terms and using Lemma 3.14,

$$\frac{28C}{3(4A + 7\tau)(8B + 7\tau)^2} = \frac{R(\kappa\tau)}{\kappa\tau} > R'(\kappa\tau).$$

Comparing with the expression in (3.88) and cancelling relevant like terms in the denominator, this is equivalent to

$$3(4A + 7\tau)(8B + 7\tau) > 4(49\tau^2 + 7(7A + 6B)\tau + 48AB).$$

which is impossible as it is equivalent to $49\tau^2 + 112A\tau + 96AB < 0$. This is the desired contradiction.

Proof of Proposition 3.18

The profit $\pi(\sigma) = \mathrm{TV}_f\, c_f^*(\sigma, \sigma_o) - \alpha\sigma$ is supermodular in $(-\theta, \sigma)$.

Proof of Proposition 3.19

First, note that if both firms enter, they choose different speeds in any pure strategy SPE: If firms choose same speed, then both firms obtain 0 revenue. To see why, let c_1 and c_2 be the fees chose by the firms in equilibrium. If $c_1 > c_2$, there is no demand in venue 1, as venue 2 has same speed and lower fee. Thus the revenue of venue 1 is 0. However a deviation to $c_1' = c_2 - \epsilon$ gives a positive market share and revenue to venue 1, thus this cannot be an equilibrium. The case for $c_2 > c_1$ is exactly same. If $c_1 = c_2 > 0$, then at least one of the firms, say firm i, can choose $c_i = c_1 - \epsilon$ to strictly increase its revenue, thus that cannot be an equilibrium. Hence, unique equilibrium in the fee competition stage is $c_1 = c_2 = 0$. The entry is costly ($\alpha > 0$), so this cannot be an equilibrium of the game as firms will decide not to enter.

If they choose different speeds, then the fees and revenues are characterized in Section 3.3.3. Let c_l and c_h denote the equilibrium fees selected by slower and faster venues in equilibrium. Let $R_l = R(\sigma_l, \sigma_h, c_l, c_h)$ denote the revenue of the slower firm and $R_h = R_f(\sigma_l, \sigma_h, c_l, c_h)$ denote the revenue of the faster firm. Note that $R_l > 0$ and $R_h > 0$. As $\alpha K(\sigma) \to 0$ as $\alpha \to 0$, there exists an $\underline{\alpha}$ such that $\min\{R_l - \alpha K(\sigma_l), R_h - \alpha K(\sigma_h)\} = 0$. Thus whenever $\alpha < \underline{\alpha}$, both firms obtain positive revenue after entry in any equilibrium, thus both firms enter in all such equilibria.

Whenever $\alpha > \underline{\alpha}$ at least one of the firms who enter will incur a loss, thus at most one firm can enter. The revenue of the single venue is given by $\Pi^*(\alpha) = \max\{R_l - \alpha K(\sigma_l), R_h - \alpha K(\sigma_h)\} = 0$, which is clearly continuous and decreasing in α. Thus, there exists an $\overline{\alpha}$ such that $\Pi^*(\overline{\alpha}) = 0$. As $\Pi^*(\alpha)$ is decreasing in α, whenever $\alpha > \overline{\alpha}$ entry is not profitable for the firms and in equilibrium no firm enters.

Bibliography

Abreu, D., D. Pearce, and E. Stacchetti. 1990. "Toward a Theory of Discounted Repeated Games with Imperfect Monitoring." *Econometrica,* 58: 1041–1063.

Acemoglu, Daron, Asuman Ozdaglar, and Ali ParandehGheibi. 2010. "Spread of (mis)information in social networks." *Games and Economic Behavior,* 70(2): 194–227.

Admati, A. R., and P. Pfleiderer. 1988. "A Theory of Intraday Patterns: Volume and Price Variability." *Review of Financial Studies,* 1(1): 3–40.

Ahnert, Toni, and Ali Kakhbod. 2017. "Information Choice and Amplification of Financial Crises." *Review of Financial Studies,* 30(6): 2130–2178.

Ajorlou, Amir, Ali Jadbabaie, and Ali Kakhbod. 2018. "Dynamic Pricing in Social Networks: The Word-of-Mouth Effect." *Management Science,* 64(2): 971–979.

Alaei, Saeed, Azarakhsh Malekian, and Mohamed Mostagir. 2016. "A Dynamic Model of Crowdfunding." *Proceedings of the Sixteenth ACM Conference on Economics and Computation, EC' 16, ACM, New York, NY, USA,* 363–363.

Albuquerque, R., and H. A. Hopenhayn. 2004. "Optimal Lending Contracts and Firm Dynamics." *Review of Economic Studies,* 72(2): 285–315.

Albuquerque, R., and H. A. Hopenhayn. 2006. "A Theory of Financing Contracts and Firm Dynamics." *Quarterly Journal of Economics,* 121(2): 229–265.

Arsov, Nino, Martin Pavlovski, and Ljupco Kocarev. 2019. "Stability of decision trees and logistic regression."

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2017. "The GRF Algorithm." *GitHub.*

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. "Generalized random forests." *Ann. Statist.,* 47(2): 1148–1178.

Babus, Ana, and Cecilia Parlatore. 2017. "Strategic Fragmented Markets." *Working paper.*

Back, Kerry, and Hal Pedersen. 1998. "Continuous Auctions and Insider Trading." *Journal of Financial Markets,* 1(3): 385–402.

Bebchuk, Lucian A., and Itay Goldstein. 2011. "Self-fulfilling Credit Market Freezes." *Review of Financial Studies,* 24(11): 3519–3555.

Bergemann, D., and U. Hege. 2005. "The Financing of Innovation: Learning and Stopping." *RAND Journal of Economics,* 36(4): 719–752.

Biais, B., T. Mariotti, G. Plantin, and J. C. Rochet. 2007. "Dynamic Security Design: Convergence to Continuous Time and Asset Pricing Implications." *Review of Economic Studies*, 75(2): 345–390.

Biais, B., T. Mariotti, J. C. Rochet, and S. Villeneuve. 2010. "Large risks, limited liability, and dynamic moral hazard." *Econometrica*, 78(1): 73–118.

Biau, Gérard, and Erwan Scornet. 2015. "A Random Forest Guided Tour." *TEST*, 25.

Billingsley, Patrick. 2008. *Probability and measure.* John Wiley & Sons.

Bimpikis, Kostas, Shayan Ehsani, and Mohamed Mostagir. 2015. "Designing Dynamic Contests." *Proceedings of the Sixteenth ACM Conference on Economics and Computation, EC' 15, ACM, New York, NY, USA*, 281–282.

Breiman, Leo. 2001. "Random Forests." *Machine Learning.*

Candogan, Ozan, and Kimon Drakopoulos. 2019. "Optimal Signaling of Content Accuracy: Engagement vs. Misinformation." *Operations Research*, forthcoming.

Chau, Minh, and Dimitri Vayanos. 2008. "Strong-Form Efficiency with Monopolistic Insiders." *Review of Financial Studies*, 21: 2275–2306.

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." *KDD '16*, 785–794. New York, NY, USA:ACM.

Chen, Xiaohui. 2018. "Gaussian and bootstrap approximations for high-dimensional U-statistics and their applications." *Ann. Statist.*, 46(2): 642–678.

Chernozhukov, Victor, Denis Chetverikov, and Kengo Kato. 2017. "Central limit theorems and bootstrap in high dimensions." *Ann. Probab.*, 45(4): 2309–2352.

DeMarzo, P. M., D. Livdan, and A. Tchistyi. 2014. "Risking other people's money: Gambling, limited liability, and optimal incentives." *Working paper.*

DeMarzo, P. M., and M. J. Fishman. 2007*a*. "Agency and Optimal Investment Dynamics." *Review of Financial Studies*, 20(1): 151–188.

DeMarzo, P. M., and M. J. Fishman. 2007*b*. "Optimal Long-Term Financial Contracting." *Review of Financial Studies*, 20(6): 2079–2128.

DeMarzo, P. M., and Y. Sannikov. 2007. "Optimal Security Design and Dynamic Capital Structure in a Continuous-Time Agency Model." *Journal Finance*, 61(6): 2681–2724.

DeMarzo, P. M., and Y. Sannikov. 2017. "Learning, Termination, and Payout Policy in Dynamic Incentive Contracts." *Review of Economic Studies*, 84(1): 182–236.

DeMarzo, P. M., Z. He, M. Fishman, and N. Wang. 2012. "Dynamic Agency and q Theory of Investment." *Journal of Finance*, 67: 2295–2340.

DiCiccio, Cyrus, and Joseph P Romano. 2020. "CLT for U-statistics with growing dimension." Stanford University.

Di Tella, S., and Y. Sannikov. 2016. "Optimal Asset Management Contracts with Hidden Savings." *Working paper.*

Duffie, Darrell. 2012. *Dark Markets: Asset Pricing and Information Transmission in Over-the-Counter Markets.* Princeton University Press.

Duffie, Darrell, Nicolae Gârleanu, and Lasse Heje Pedersen. 2005. "Over-the-Counter Markets." *Econometrica*, 73(6): 1815–1847.

Duffie, Darrell, Semyon Malamud, and Gustavo Manso. 2009. "Information Percolation with Equilibrium Search Dynamics." *Econometrica*, 77(5): 1513–74.

Easley, D., and M. O'Hara. 1987. "Price, trade size, and information in securities markets." *Journal of Financial Economics*, 19(1): 69–90.

Farhi, E., and I. Werning. 2006. "Inequality, Social Discounting and Progressive Estate Taxation." *MIT Working paper*.

Fleming, W., and H. M. Soner. 2007. *Controlled Markov Processes and Viscosity Solutions.* Springer (2nd edition).

Fudenberg, D., B. Holmstrom, and Milgrom P. 1990. "Short-Term Contracts and Long-Term Agency Relationships." *Journal of Economic Theory*, 51(1): 1–31.

Garrett, Daniel, Alessandro Pavan, and Juuso Toikka. 2018. "Robust predictions in dynamic screening." Tech. rep. Northwestern University.(Cit. on pp. 9, 30, 39).

Genuer, Robin, Jean-Michel Poggi, and Christine Tuleau-Malot. 2010. "Variable selection using random forests." *Pattern recognition letters*, 31(14): 2225–2236.

Glosten, L. R., and P. Milgrom. 1985. "Bid, ask and transaction prices in a specialist market with heterogeneously informed traders." *Journal of Financial Economics*, 14(1): 71–100.

Goldstein, Itay. 2012. "Empirical Literature on Financial Crises: Fundamentals vs. Panic." In *The Evidence and Impact of Financial Globalization.*, ed. Gerard Caprio. Amsterdam:Elsevier.

Google, and other contributors. 2015. "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems." Software available from tensorflow.org.

Gorton, Gary, and Guillermo Ordonez. 2014. "Collateral Crises." *American Economic Review*, 104(2): 343–378.

Gregorutti, Baptiste, Bertrand Michel, and Philippe Saint-Pierre. 2017. "Correlation and variable importance in random forests." *Statistics and Computing*, 27(3): 659–678.

Grossman, Sanford J. 1992. "The Informational Role of Upstairs and Downstairs Trading." *The Journal of Business*, 65(4).

Grossman, Sanford J., and Merton H. Miller. 1988. "Liquidity and Market Structure." *Journal of Finance*, 43(3): 735–777.

Guerrieri, Veronica, and Robert Shimer. 2014. "Dynamic Adverse Selection: A Theory of Illiquidity, Fire Sales, and Flight to Quality." *American Economic Review*, 104(7): 1875–1908.

Guerrieri, Veronica, Robert Shimer, and Randall Wright. 2010. "Adverse Selection in Competitive Search Equilibrium." *Econometrica*, 78(6): 1823–1862.

Halac, M., N. Kartik, and Q. Liu. 2016. "Optimal Contracts for Experimentation." *Review of Economic Studies*, 83: 1040–1091.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media.

He, Z. 2009. "Optimal Executive Compensation when Firm Size Follows Geometric Brownian Motion." *Review of Financial Studies*, 22: 859–892.

He, Z. 2011. "A Model of Dynamic Compensation and Capital Structure." *Journal of Financial Economics*, 100: 351–366.

He, Z. 2012. "Dynamic Compensation Contracts with Private Savings." *Review of Financial Studies*, 25: 1494–1549.

He, Z., B. Wei, J. Yu, and F. Gao. 2017. "Optimal Long-term Contracting with Learning." *Review of Financial Studies*, 30(6): 2006–2065.

Hirano, Keisuke, Guido Imbens, and Geert Ridder. 2003. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica*, 71(4): 1161–1189.

Hörner, J., and L. Samuelson. 2013. "Incentives for experimenting agents." *RAND Journal of Economics*, 44(4): 632–663.

Imbens, Guido, and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press.

Kakhbod, Ali, and Fei Song. 2020. "Dynamic Price Discovery: Transparency vs. Information Design." *Games and Economic Behavior*, 122: 203–232.

Karatzas, I., and S. Shreve. 1991. *Brownian Motion and Stochastic Calculus.* Springer.

Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree." *NIPS'17*, 3149–3157. Red Hook, NY, USA:Curran Associates Inc.

Kyle, Albert S. 1985*a*. "Continuous auctions and insider trading." *Econometrica*, 1315–1335.

Kyle, Albert S. 1985*b*. "Long-lived information and intraday patterns." *Econometrica*, 53(6): 1315–1335.

Lagos, Ricardo, and Guillaume Rocheteau. 2009. "Liquidity in Asset Markets with Search Frictions." *Econometrica*, 77(2): 403–26.

Lerner, J., and J. Wulf. 2007. "Innovation and Incentives: Evidence from Corporate R&D." *Review of Economics and Statistics*, 89(4): 634–644.

Malamud, Semyon, and Marzena Rostek. 2017. "Decentralized Exchange." *American Economic Review*, 107(11): 3320–3362.

Malenko, A. 2018. "Optimal Dynamic Capital Budgeting." *Review of Economic Studies*, Forthcoming.

Maniruzzaman, Md, Md Jahanur Rahman, Md Al-MehediHasan, Harman S Suri, Md Menhazul Abedin, Ayman El-Baz, and Jasjit S Suri. 2018. "Accurate diabetes risk stratification using machine learning: role of missing value and outliers." *Journal of medical systems*, 42(5): 92.

Manso, G. 2011. "Motivating Innovation." *Journal Finance*, 66(5): 1823–1860.

Pagnotta, Emiliano S., and Thomas Philippon. 2018. "Competing on speed." *Econometrica*, 122: 1067–1115.

Papanastasiou, Yiangos, Kostas Bimpikis, and Nicos Savva. 2018. "Crowdsourcing Exploration." *Management Science*, 64(4): 1727–1746.

Pavan, A., I. Segal, and J. Toikka. 2014. "Dynamic Mechanism Design: A Meyersonian Approach." *Econometrica*, 82(2): 601–653.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, 12: 2825–2830.

Phelan, C., and R. Townsend. 1991. "Computing Multi-Period, InformationConstrained Optima." *Review of Economic Studies*, 58: 853–881.

Piskorski, T., and A. Tchistyi. 2010. "Optimal mortgage design." *Review of Financial Studies*, 23(8): 3098–3140.

Prokhorenkova, Liudmila, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. "CatBoost: Unbiased Boosting with Categorical Features." *NIPS'18*, 6639–6649. Red Hook, NY, USA:Curran Associates Inc.

Quaedvlieg, Rogier. 2019. "Multi-horizon forecast comparison." *Journal of Business & Economic Statistics*, 1–14.

Radner, R. 1985. "Repeated Principal-Agent Games with Discounting." *Econometrica*, 53: 1173–1198.

Revuz, D., and M. Yor. 2004. *Continuous Martingales and Brownian Motion.* Springer (3rd edition).

Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology*, 66(5): 688–701.

Sannikov, Y. 2008. "A Continuous-Time Version of the Principal-Agent Problem." *Review of Economic Studies*, 75(3): 957–984.

Sannikov, Y. 2012. "Dynamic Security Design and Corporate Financing." *Handbook of Economics and Finance, Elsevier-Northholland*, 2.

Seppi, Duane J. 1990. "Equilibrium Block Trading and Asymmetric Information." *Journal of Finance*, 45(1): 73–94.

Spear, S. E., and S. Srivastava. 1987. "On Repeated Moral Hazard with Discounting." *Review of Economic Studies*, 54: 599–617.

Strobl, Carolin, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. 2008. "Conditional variable importance for random forests." *BMC bioinformatics*, 9(1): 307.

Tang, Fei, and Hemant Ishwaran. 2017. "Random forest missing data algorithms." *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 10(6): 363–377.

Vaart, A. W. van der. 1998. *Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics*, Cambridge University Press.

Vayanos, Dimitri, and Pierre-Olivier Weill. 2008. "A Search-Based Theory of the On-the-Run Phenomenon." *Journal of Finance*, 63(3): 1361–98.

Wager, Stefan, and Guenther Walther. 2015. "Adaptive Concentration of Regression Trees, with Application to Random Forests." *arXiv: Statistics Theory*.

Wager, Stefan, and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." *Journal of the American Statistical Association*, 113(523): 1228–1242.

Wang, Jiang. 1993. "A Model of Intertemporal Asset Prices under Asymmetric Information." *Review of Economic Studies*, 102: 249–282.

Wang, Jiang. 1994. "A Model of Competitive Stock Trading Volume." *Journal of Political Economy*, 102: 127–168.

Werning, I. 2002. "Repeated Moral-Hazard with Unmonitored Wealth: A Recursive First-Order Approach." *MIT Working paper*.

Williams, N. 2009. "On Dynamic Principal-Agent Problems in Continuous Time." *Working paper*.

Zhu, Haoxiang. 2012. "Finding a Good Price in Opaque Over-the-Counter Markets." *Review of Financial Studies*, 25(4): 1255–1285.

Zhu, Haoxiang. 2014. "Do Dark Pools Harm Price Discovery?" *Review of Financial Studies*, 27(3): 747–789.