JOB MARKET PAPER

Solving combinatorial discrete choice problems in heterogeneous agent models*

Theory and an application to corporate tax harmonization in the European Union

Kathleen Hu† Rowan Shi‡

November 3, 2019

latest version

This paper develops a solution method for computing optimal decisions to combinatorial discrete choice problems (CDCPs) in heterogeneous agent settings. With an arbitrary type distribution over any number of differentiated characteristics, it quickly computes the policy *function* mapping the entire type space to corresponding optimal actions. The binary decisions can display either supermodular or submodular interactions. Problems of this structure arise naturally in economic settings, especially in international trade and industrial organization. The proposed algorithm is particularly well suited for estimating or computing general equilibrium models incorporating heterogeneous agents solving CDCPs, including choices on plant locations, input sourcing partners, or export market entry. As an illustration of the algorithm in practice, the paper then turns to evaluating the effects of a counterfactual policy equalizing corporate tax rates across the European Union using a quantitative general equilibrium model where heterogeneous firms optimally select a set of countries in which to operate affiliates.

1. Introduction

Many problems in economics, especially in international trade and industrial organization, feature optimization over combinatorial discrete choices which tend to be high dimensional

*We are very grateful to our advisor, Stephen Redding, for his tireless guidance and support. We also received helpful comments from Fabian Eckert, Teresa Fort, Gene Grossman, Oleg Itskhoki, Ben Lund, Eduardo Morales, Ezra Oberfield, Richard Rogerson, Esteban Rossi-Hansberg, and seminar participants from the International Economics Section at Princeton University. We acknowledge support from the International Economics Section at Princeton University.

†kh5@princeton.edu, Princeton University International Economics Section

‡hs7@princeton.edu, Princeton University International Economics Section

and difficult to compute. For example, a classic question in international trade is the multinational firm's optimal selection of countries in which to operate plants. This decision effectively encapsulates a collection of discrete binary choices: for each country where it is possible to open a plant, should the firm establish one there or not? If the binary choices are interdependent, that is, opening a plant in one location affects the decision of whether to open a plant in another, then they cannot be considered separately. The multinational is thus left comparing each combination of possible choices on the individual binary decisions. Recasting the problem as selecting the optimal *set* of locations in which a multinational firm operates plants, it is easy to see that the space of possible strategies balloons given the combinatorial nature of the problem. In particular, if there are N available locations, the firm must decide among 2^N possible sets. The plant location problem is not the only one this of type. Selecting partners from whom to source inputs, export markets to enter, and which goods to produce as a multiproduct firm have all been formulated this way. Without a strategy, solving combinatorial discrete choice problems (CDCPs) quickly becomes computationally burdensome or even infeasible due to the choice space's large size and discrete nature.

The relevance of this problem is highlighted by a large body of work in the international trade literature, which emphasizes the importance of heterogeneous firms. In particular, studies in several settings have consistently concluded that multinational firms, importers, exporters, and firms offering many products all tend to be large and to account for disproportionate shares of an economy's total sales, employment, and other aggregate variables. Therefore, understanding how to solve the CDCPs underlying the behavior of these firms is an important research agenda, since their activity can have aggregate implications.

However, a single firm's CDCP solution is often not enough to make aggregate conclusions. To do so, heterogeneous firms solving CDCPs must be embedded in a quantitative general equilibrium. This type of exercise exacerbates the challenge of solving the high dimensional CDCP described previously in three ways. First, if each agent has a different optimal solution to the CDCP, it must now be solved at every point along the heterogeneous distribution. In the plant location context, suppose firms vary by productivity, where firms of different productivity choose different plant locations. The optimal set of plant locations would then need to be solved for each level of productivity to arrive at aggregate variables like labor demand. In the scenario where heterogeneity follows a continuous distribution, there is an infinite number of CDCPs, one for every point within the support. Thus, introducing heterogeneity could greatly multiply the number of CDCPs to be computed. Next, nesting a block of heterogeneous agents, each solving a CDCP, within a general equilibrium magnifies the computational difficulty. The collection of CDCPs must be repeatedly and accurately calculated as the general equilibrium routine searches for a fixed point on aggregates. Lastly, estimating parameters of the general equilibrium model potentially creates another layer of repetition because the fixed point for general equilibrium must be solved for each possible parameter vector. All in all, computing a quantitatively-estimated general equilibrium model incorporating heterogeneous agents, each facing CDCPs, could imply solving them an enormous number of times.

The goal of this paper is to provide a novel way to solve combinatorial discrete choice problems for a distribution of heterogeneous agents, enabling such a block to be feasibly included into a quantitative general equilibrium model. It is applicable to problems where the task is

to find the optimal subset of a defined superset. The superset could be interpreted as the list of all available locations for plants, sourcing partners for inputs, or any similar collection of binary possibilities. In turn, for the plant location problem, the optimal subset would be the set of locations in which a firm should optimally establish plants. A similar interpretation of the optimal subset follows for other binary choice problems, including optimal sourcing partners or export market entry.

The main innovation of this solution method is to find the policy *function* mapping an agent's type to their optimal CDCP solution. Broadly, it does so by computing cutoffs within the range of heterogeneity at which elements are removed or added into the optimal set. With heterogeneous productivity firms making plant location decisions, for example, these would be marginal productivities corresponding to firms indifferent between opening plants in a certain location or not. The intuition of how to find these cutoffs is provided below. If there are multiple dimensions of heterogeneity, the method naturally extends to identifying boundaries within the type space along which agents remain indifferent between including an element or not. Note that, in solving the problem this way, the solutions to a possibly large number of CDCPs implied by allowing for agent heterogeneity are succinctly summarized by a list of cutoffs and the marginal decision at each.

This approach is attractive for a number of reasons. The first is conceptual. Since it calculates the policy function over the entire range of heterogeneity without needing the distribution of agents across the type space, the algorithm can handle distributions of any shape, whether discrete or continuous. It is consequently highly versatile. There are also substantial computational time gains. In a way that will become clear when the solution's intuition is discussed, it is much faster to compute the policy function as a set of indifference points compared to finding optimal sets at multiple discrete agent types. Finally, the algorithm avoids discretizating continuous type spaces, eliminating inaccuracy that would arise from interpolation between gridpoints. Aggregations based on the algorithm's policy functions are instead exact. In summary, the algorithm is well suited for solving CDCPs faced by heterogeneous agents within quantitative general equilibrium settings, since it is flexible on the type distribution, reduces computational time considerably, and maintains full solution accuracy over the type space.

To isolate the cutoffs discussed above, the paper imposes structure on the CDCP by making two assumptions on the marginal value derived from including an item into a set. First, it must be either (weakly) monotonically decreasing or monotonically increasing in the set. In the first case, the gains from including an element into a set will fall as the set grows. For example, if a multinational firm engages in export platform foreign direct investment (FDI), the additional market access provided by any single plant declines as the multinational opens more and more plants. In other words, the elements loosely can be interpreted as substitutes for each other. Conversely, in the monotonically increasing case, an item's additional value to any set rises as the set grows. In this case, each individual item is more beneficial when part of larger groups. The plant location problem with vertical FDI could be an example of this structure, if each facility carries out complementary production tasks. The elements can accordingly be interpreted as being complements for each other. For that reason, the two cases are respectively referred to as the substitutes and complements case in what follows. Either is sufficient to satisfy monotonicity in the set.

The second restriction is (weak) monotonicity along dimensions of heterogeneity, which implies that the marginal value of an additional element's inclusion into any set will increase as the agent's type increases. This assumption specifies how the heterogeneous type interacts with the interdependent binary decision payoffs. Returning to the leading example of firms selecting production locations while varying in productivity, monotonicity in type translates to higher productivity firms benefiting more from opening extra plants than their low productivity counterparts, all else equal. With multiple dimensions of heterogeneity, it is sufficient if an element's marginal value increases in every characteristic. The paper focuses on the montonically increasing case without loss of generality, since the problem can be reformulated in terms of $1/z$ if the marginal value is instead decreasing in a dimension of heterogeneity z. Ultimately, both assumptions appear naturally in CDCPs derived from economic questions since they discipline the marginal value derived from an additional element.

Along with the structure imposed above, the proposed solution leverages a series of necessary conditions on the optimal set. Observe that, if an element is part of the optimal set, it must add weakly positive marginal value in that set. Put another way, removing it from the set cannot increase payoffs. In the plant location example, if a plant is optimally opened in a specific area, it cannot be that closing this plant increases profits. Likewise, if an element does not appear in the optimal set, then it cannot add positive value if included. That is, profits cannot be increased by establishing a plant where optimally no plants are open. Thus, signing an element's marginal value to the optimal set seals its fate. If it is positive, then the element should be included, while if it is negative, the element should be excluded. The algorithm proceeds by signing these marginal values using the two monotonicity assumptions described earlier to rule out many possible combinations from the choice space.

Of course, these sufficient conditions require the optimal set to calculate marginal values, precisely the unknown object being computed. However, monotonicity in the set guarantees that an element's marginal value from inclusion in the optimal set can be bounded from above and below for any type. To demonstrate this statement, first consider the substitutes case. Suppose that a subset of the optimal set is known. Such a set can always be specified since the empty set trivially qualifies. This subset will get more out of any extra element compared to the optimal set, because it contains fewer elements that can serve as substitutes. Then, any element's marginal contribution to the optimal set is bounded from above by its marginal value of inclusion in the subset. The lower bound can be identified in a similar way. Suppose a superset of the optimal set is known, where the full set of items can always serve as a natural starting point. Then, an element's contribution to this superset is smaller than its marginal value from inclusion in the optimal set, since the superset will contain more elements that are substitutable. Thus, the element's marginal value from inclusion in the subset and superset provide upper and lower bounds on its marginal value from inclusion the optimal set, respectively. A reverse argument holds in the complements setting. In that setting, elements add more value in large groups, so the subset and superset provide corresponding lower and upper bounds on any element's marginal value of inclusion into the optimal set. Therefore, in both the substitutes and complements case, monotonicity in the set ensures lower and upper bounds on an element's additional contribution to the optimal set.

Now fix an element and consider for which types its inclusion into the optimal set can be signed. Start with the upper bound. In the substitutes case, it is obtained by finding how

much this element contributes when added to the subset of the optimal set, because there are fewer substitutable items. As discussed above, the lower bound for the complements case is instead derived from the superset, which contains more complementary items. Either way, this maximal marginal value also increases in agent type due to monotonicity in type. Recall the substitutes plant location problem for intuition. Opening plants in a specific country is more valuable when added to emptier sets, and also for more productive firms. If there is a threshold type for which this upper bound crosses zero, then monotonicity in type ensures that it is negative for types below. As a result, these types definitely receive negative marginal reward from including the element in consideration to their optimal sets. Following the argument established above, they should optimally exclude it. In the plant location scenario, this conclusion establishes that firms with productivity lower than some threshold should not open plants in the specific location considered. A similar thread of logic can be made using the lower bound, which is also derived from the element's marginal value from inclusion in the subset or superset of the optimal set. It must likewise increase in agent type, but this time identifying the threshold type where it crosses zero allows implications to be drawn for agents with higher types. In particular, for these types, the lower bound on the element's marginal value from inclusion in the optimal set is positive, implying the element brings positive value to the optimal set. In the horizontal FDI setting, if it is still worth adding a plant to the set with a large number of substitutes for some types, then it must create positive value when added to the smaller optimal set. In this case, agents of these types must include the element in their optimal sets. Finally, the subset and superset sandwiching the optimal set can be updated to incorporate these results, with included elements added into the first and excluded elements discarded from the second. This type of reasoning can then be iteratively applied to all binary decisions for which the optimal choice is not known, and represents the main logical thrust of the algorithm. It crucially harnesses the discipline imposed by monotonicity in type to draw conclusions for large sections of the type space at once, creating the considerable speed gains referenced earlier.

To demonstrate the algorithm's performance in practice, the paper next turns to evaluating the counterfactual effects of corporate tax equalization in the European Union. While the area's economies are integrated by free trade and movement of people, member states currently set corporate tax rates independently. In 2014, a gap exceeding 20 percentage points separated the highest and lowest tax rates in the region. Corporate tax variation among member states has sparked policy debate within the union, with high tax countries concerned that EU multinational firms are incentivized to move production into low tax areas. Relocating production within the EU from high tax to low tax countries could be attractive for multinationals based in the EU because a series of tax treaties imply their profits are largely taxed only in the country of production. If production affiliates were opened in low tax countries, profits would then be taxed at a low rate while goods could still be sold tariff-free anywhere within the EU. High-tax countries could lose jobs and tax revenue as a result of these EU-based multinational firms locating production to low tax regions.

Moreover, cross-member FDI is of particular importance within the EU. In 2014, production by foreign-owned plants owned by companies based in another EU member state accounted for 22.5% of total EU manufacturing turnover. Understanding the behavior of these multinationals is therefore important for understanding aggregate economic outcomes. If EU

firms take tax rates into consideration when establishing plants, then their variation among members countries could affect ultimate production location choices. Resulting relocation into low tax countries could imply fewer jobs, lower wages, and lower tax revenue for the remaining countries. Perhaps for this reason, policy leaders in large high-tax countries, notably Germany and France, have pushed for a EU-wide corporate tax policy harmonization. Unsurprisingly, low-tax countries like Hungary have resisted this proposal.

The effects of a tax rate equalization policy is assessed using a quantitative spatial general equilibrium model. It features rich motives for foreign direct investment (FDI) such as market access, production capability, and tax rates dispersion, as well as competing motives for concentrating production. These are namely fixed costs of plant establishment and losses associated with arms-length production, like communication costs. Multinational activity arises endogenously as firms choose any subset of countries in the EU in which to open plants. The solution method presented in the paper is applied to compute the optimal strategy as a function of firm productivity. The model is then calibrated to aggregate moments from the EU as well as key multinational shares. As a counterfactual exercise, taxes are unified at the rate where EU-wide tax revenues match the total amount from the calibrated model. Once the differences in relative tax rates are eliminated, industry relocates to countries with formerly high rates. Domestic economic activity expands in these countries, as they enjoy more firms, higher real wages, and more workers. All together, these countries host 6.8% more firms, 0.8% more workers, and between 0.2% to 1.9% higher real wages in the counterfactual equilibrium. A mirror effect occurs in the remaining countries, with 9% fewer firms, 0.9% fewer workers, and between 0.3% to 3.7% lower real wages. The model therefore predicts the that policy would trigger industry redistribution towards the large high-tax countries.

This paper is mainly related to three strands of literature. The first encompasses work on CDCPs like those studied here. A foundational paper of this literature is Jia [2008], which proposes a solution method for games with two players over binary choices featuring complementarity. Arkolakis and Eckert [2017] generalize this approach for games with a small number of agents, characterized by either substitutability or complementarity on the set. The method in Jia [2008] is applied in the quantitative exercises of Antràs et al. [2017], which studies combinatorial sourcing decisions of heterogeneous firms. Their computational exercises assume complementarity between choices and proceeds by discretizing a continuous type space to apply the single-type algorithm at each gridpoint. The CDCP represented by the optimal sourcing strategy is further explored in Antrás [2016], which comprehensively studies the sourcing decisions of multinational firms while focusing on the complements case. The key contributions of this paper to the literature is to develop an approach explicitly solving for the policy function as a function of the agent's type. It is flexible enough to allow for a continuum of agent types over an arbitrary number of dimensions heterogeneity in problems featuring either substitutablility or complementarity.

Secondly, this paper builds on the literature on multinational production, summarized by Antrás and Yeaple [2014]. Since this literature is vast, a selection of most closely related papers is highlighted here. First, Helpman et al. [2004] characterizes the selection over firm productivity into exporting or FDI, a key component of many models featuring multinational production, including ours. However, the paper disallows export platforms, limiting the interactions between a firm's decision of how to serve each market. On the other hand,

Ramondo and Rodríguez-Clare [2013] estimate the gains from trade and multinational productions in a Ricardian model based on Eaton and Kortum [2002]. The model in this paper allows for export platform FDI but does not include fixed costs of multinational production, so production occurs wherever variables costs are lowest with no discrete choice over plant locations. Similarly, Arkolakis et al. [2018] includes a model without fixed costs of multinational production to investigate the multinational firm's decision of where to locate production facilities separate from their country of ownership. The paper's main goal is to develop a theory explaining why countries differentially specialize in production activities and headquarter activities. Following the insights from this paper, our paper also allows a country's production capability, which applies to all plants within its borders, to be distinct from the average productivity firms born there. Lastly, with a model most related to ours, Tintelnot [2016] analyzes affiliate location and production choices of German multinationals within a set of twelve countries. The model itself describes heterogeneous firms engaging in horizontal FDI, allowing for export platform FDI while also incorporating a flat fixed cost of multinational production. In our paper, a similar firm decision is expanded to incorporate corporate taxation, then placed within a general equilibrium framework. The choice of production locations is also extended to 27 possibilities to represent the EU, exponentially increasing the number of possible sets in which a firm can place plants as a result. The approach developed in our paper to solve CDCPs over heterogeneous agents greatly reduces this challenge.

Since firms engaging in multinational production tend to be large, this paper is also part of the vast literature on heterogeneous firms in trade. Pioneered by Melitz [2003], this literature emphasizes the importance of firm heterogeneity in the context of international trade. More specifically, Melitz and Redding [2014] reports that firms engaging in a broad class of global activities, such as exporting, importing, or multinational production, all tend to be larger and more productive than the average firm. As a result, globalization can favor these large firms, spurring within-industry reallocation towards them. This within-industry compositional change can in turn contribute to aggregate effects. Our model features heterogeneity in behavior across firms in line with this literature, with more productive firms selecting different sets of production locations from less productive firms. Bernard et al. [2018] develops a unified framework to include an array of decisions faced by these large firms engaged in international activity, including entry into export markets, production locations, sourcing partners, sourced inputs, and goods produced. Each of these decisions lend themselves well to be modeled as a CDCP and solved using the solution method presented in this paper.

Finally, this paper relates to the literature analyzing the effect of tax incentives on the production location choices of firms. For example, Devereux and Griffith [2003] and Buettner and Ruf [2007] establish empirically that tax incentives are linked to FDI and multinational production. Our paper differs because it incorporates a structural model of the multinational firm's FDI decision, which is then used to evaluate the effects of counterfactual policies. On the other hand, Fajgelbaum et al. [2018] also use a structural model to evaluate spatial misallocation across states in the U.S. attributable to tax rate dispersion. While they include many different taxes, including corporate taxes and income taxes, firms in their model are single-location. They therefore do not incorporate the tradeoffs faced by multi-location firms, such as establishing plants both for production capability concerns as well as market access.

The remainder of the paper is organized as follows. Section 2 introduces the plant location

problem, its important structural features, and the intuition of the solution method. Section 3 provides a formalization of the class of problems studied and provides general sufficient conditions to identify when the algorithm can be applied before describing it in detail. The paper then illustrates the performance of the algorithm with an application to multinational plant location decisions in the European Union. A quantitative model, calibration, and results for this exercise are detailed in Section 4. Section 5 concludes.

2. A simple example

Before presenting the formal method in Section 3, this section is dedicated to outlining an illustrative example of the problems addressed in this paper. It considers firms heterogeneous in productivity, each which must establish plants in order to produce and serve destination markets. Section 2.1 formalizes the problem and highlights features of its structure that are used in the solution method. For simplicity, this section only outlines easily interpreted sufficient conditions. Generalized necessary conditions are deferred to Section 3. Section 2.2 then discusses the general intuition used by the solution routine. Finally, a preview of its advantages over existing methods is presented in Section 2.3.

2.1. The problem

This example considers a firm maximizing profits by choosing where to establish affiliates that serve as production facilities and export platforms. The available countries are Germany (G) and Romania (R), with locations interacting in a way that will be defined formally below. This section presents the problem with only two countries for now to remain as simple as possible while building conceptual intuition.

Firms are heterogeneous in productivity, z, and take aggregate conditions as given when deciding where to locate plants. The objective function

$$\pi : \mathscr{P}(\{G,R\}) \times Z \times \boldsymbol{\Upsilon} \to \mathbb{R},$$

is defined over the power set of available locations, the range of productivity $Z \subseteq \mathbb{R}$, and the range of aggregate vectors $\boldsymbol{\Upsilon}$. An aggregate vector $\boldsymbol{v} \in \boldsymbol{\Upsilon}$ includes all relevant aggregate variables, like wages and prices.[1] In this setting, possibilities for the chosen set of a firm are $\{\}$, $\{G\}$, $\{R\}$, or $\{G,R\}$, corresponding to the actions of opening plants nowhere, only in Germany, only in Romania, or in both countries. The task at hand is to find the policy function $\mathscr{J}^*(\cdot;\boldsymbol{v})$, a set-valued function taking firm productivity as its argument and returning one of these possibilities. This policy function is parameterized by the vector of aggregate $\boldsymbol{v} \in \boldsymbol{\Upsilon}$, assumed fixed at an arbitrary value for the rest of the discussion below.

Plants in a country $j \in \{G,R\}$ should be added only if they increase profits. Thus, the key difference considered is

$$D_j(\mathscr{J};z,\boldsymbol{v}) \equiv \begin{cases} \pi(\mathscr{J};z,\boldsymbol{v}) - \pi(\mathscr{J}\setminus\{j\};z,\boldsymbol{v}) & \text{if } j \in \mathscr{J} \\ \pi(\mathscr{J}\cup\{j\};z,\boldsymbol{v}) - \pi(\mathscr{J};z,\boldsymbol{v}) & \text{if } j \notin \mathscr{J} \end{cases},$$

[1] In what follows, vectors are represented with boldface variables.

which is the *marginal value function*. It describes the marginal value from j's inclusion in $\mathscr{J}$.[2] For example, the expression $D_G(\{\};z,\boldsymbol{v}) = \pi(\{G\};z,\boldsymbol{v}) - \pi(\{\};z,\boldsymbol{v})$ captures, for a firm with productivity z, the profit difference between choosing to open plants in only Germany compared to opening no plants anywhere.

Because the marginal value function specifies the set $\mathscr{J}$ to which j is included, it leaves room for interdependence among the choices. For example, in the previous discussion, the value of opening plants in Germany is contextualized with the fact that there are no plants in any other locations. This information is provided by the first argument $\{\}$. In contrast, $D_G(\{R\};z,\boldsymbol{v})$ evaluates how profits would change if plants were opened in Germany, conditional on the firm also establishing plants in Romania. If these two valuations are different, the contribution of Germany changes based on the choice made for the other binary decision, Romania, which reflects interaction between these two choices. Likewise, since z is the second argument, firm productivity could also influence j's marginal value in $\mathscr{J}$. For example, $D_G(\{\};z,\boldsymbol{v})$ varying with z could generate a nontrivial distribution of optimal behavior over the range of firm productivity. The solution relies on the marginal value function exhibiting a specific structure over its two arguments, described below.

2.1.1. Monotonicity in set

The first structural assumption restricts how plant locations interact with each other in the marginal value function $D_j(\cdot;z,\boldsymbol{v})$. To illustrate the intuition, fix a firm productivity z and consider the value of adding plants in Germany. There are two possibilities for Romania: either the firm opens plants there or not. Since the plant locations interact, the marginal value of having plants in Germany depends on the decision for Romania.

Diminishing marginal returns asserts that opening plants in Germany adds more value if there are no plants in Romania. The following inequality holds in this case for all z.

$$D_G(\{\};z,\boldsymbol{v}) = \pi(\{G\};z,\boldsymbol{v}) - \pi(\{\};z,\boldsymbol{v}) \geq \pi(\{G,R\};z,\boldsymbol{v}) - \pi(\{R\};z,\boldsymbol{v}) = D_G(\{R\};z,\boldsymbol{v}) \tag{1}$$

For every productivity level, this inequality ensures that plants in Germany increase profits by more if there are no plants in Romania. A profit function satisfying this condition could emerge from export-platform horizontal FDI models where each plant produces the same product. With no affiliates, the firm cannot serve either the German or Romanian markets. Then, opening German plants would be valuable since they could serve both the domestic consumers as well as those in Romania through trade. If there are instead plants in Romania to begin with, both markets can already be served by them. Adding plants in Germany may lower costs of serving German customers, increasing variable profits, but gains are smaller relative to the first scenario. Intuitively, the condition can be interpreted as corresponding to the case in which the plants are substitutes for each other.

On the other hand, reversing the weak the inequality in equation (1) defines the increasing marginal returns condition. In this case, plants in Germany are more valuable when considered in tandem with plants in Romania. Profit functions emerging from vertical FDI models, where production facilities perform complementary tasks, could satisfy this restriction.

[2] Note that, $\mathscr{J}$ represents a set while $\mathscr{J}^*(\cdot;\boldsymbol{v})$ represents a set-valued function. Functions in the paper will always be denoted this way.

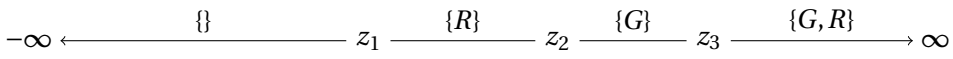


Figure 1: An example of the policy function. At each cutoff z_n, the optimal set of countries $\mathscr{J}^*(\cdot;\boldsymbol{v})$ changes.

Without any other affiliates, a single one in Germany must perform all duties associated with production. The resulting increase in profits gain may be smaller than if there already were a plant in Romania. In that scenario, the Romanian plant is performing all tasks before the introduction of a German plant. After it is opened, both locations could specialize on separate portions of the production process. The additional value of opening the German plant is then larger if a Romanian plant is also established. In this setting, the condition can be interpreted as corresponding to the case in which plants are complementary.

Although the discussion here of the interpretation of monotonicity in set case concentrated on the marginal value of opening plants in Germany, the condition to hold either as the substitutes or complements case, parallel statements also must be true when assessing the marginal value of opening plants in Romania.

For concreteness, the rest of this simple example assumes the substitutes case, in accordance with the application in Section 4. However, problems characterized by either type of monotonicity in set can be solved by the generalized method presented in Section 3.

2.1.2. Monotonicity in type

The second assumption disciplines how firm productivity affects a location's marginal returns. For this assumption, fix a set $\mathscr{J}$ and consider two firms with differing productivities $z_1 < z_2$. Without loss of generality, assuming monotonicity in type involves assuming that marginal returns increase in productivity, or

$$D_G(\mathscr{J}; z_1, \boldsymbol{v}) \leq D_G(\mathscr{J}; z_2, \boldsymbol{v})\,. \tag{2}$$

In words, opening an extra plant in Germany has a larger benefit for a more productive firm than for a less productive firm. The problem can be redefined in $1/z$ if the inequality goes the other way.

2.2. Squeezing the optimal set

Observe that the objective function is continuous in productivity, but could be discontinuous as elements are added to $\mathscr{J}$. If $\mathscr{J}^*(z;\boldsymbol{v})$ is the optimal set of countries in which to open plants for a firm with productivity z, it will also be optimal for firms with productivity in a small enough neighborhood of z. The optimal policy function $\mathscr{J}^*(\cdot;\boldsymbol{v})$ then separates the productivity range into intervals, each associated with a distinct optimal set. Figure 1 provides an example of how $\mathscr{J}^*(\cdot;\boldsymbol{v})$ might look. The strategy in this paper is therefore to solve for these intervals and their associated optimal sets. For the remainder of this section, the $\boldsymbol{v}$ is omitted for notational brevity when unambiguous.

More concretely, consider a firm with productivity z. Initially, the optimal choice on each binary decision is unclear. In other words, for each country, it is unknown whether the firm should open plants there. The goal is to progressively determine z's best choice for each binary decision, narrowing down the possibilities for its optimal set. Formally, define "subset"- and "superset"-valued functions, mapping each z respectively to

$$\begin{aligned} L(z) &= \{j \in J \text{ known to be included in optimal set}\} \\ U(z) &= \{j \in J \textit{ not} \text{ known to be excluded from optimal set}\} \end{aligned}. \tag{3}$$

Crucially, $L(z) \subseteq \mathscr{J}^*(z) \subseteq U(z)$, so the subset is the smallest the optimal set could be, while the superset is the largest. The solution method squeezes $\mathscr{J}^*(\cdot)$ iteratively by including elements into $L(\cdot)$ while excluding elements from $U(\cdot)$. During the iteration, monotonicity in type will be evoked to update the subsets and supersets for many values of productivity at once.

To begin, fix z and consider $D_j(\mathscr{J}^*(z); z)$, a country j's marginal value to the optimal set. Suppose it is negative. Then, it must be the case that $j \notin \mathscr{J}^*(z)$. If it were, using the definition of marginal value implies that $\pi(\mathscr{J}^*(z); z) - \pi(\mathscr{J}^*(z) \setminus \{j\}; z) < 0$, or that the firm is better off removing it. This statement contradicts the fact that $\mathscr{J}^*(z)$ is the optimal set. Then, it must be that j should not be included for firms with productivity z. Conversely, suppose $D_j(\mathscr{J}^*(z); z) > 0$. In this case, it must be that $j \in \mathscr{J}^*(z)$. A similar logic can be applied to arrive at a contradiction if it were not. As a result, the sign of a country j's marginal value to $\mathscr{J}^*(z)$ is directly linked to whether or not it should be optimally included by a firm with productivity z. The squeezing strategy described below attempts to sign $D_j(\mathscr{J}^*(z); z)$ by bounding it from above and below with the marginal value of j's inclusion into $L(z)$ and $U(z)$.

Note that $\{\} \subseteq \mathscr{J}^*(z) \subseteq \{G, R\}$ for all z. Thus, setting $L(z) = \{\}$ and $U(z) = \{G, R\}$ is a valid initialization. Starting with the subset, right away some firms can be identified as better off *excluding* Germany. To do so, consider the productivity value z_G^L such that

$$D_G(\{\}; z_G^L) = \pi(\{G\}; z_G^L) - \pi(\{\}; z_G^L) = 0\,. \tag{4}$$

Monotonicity in type guarantees there is at most one productivity level for which firms are indifferent between opening a plant in Germany or opening no plants at all.

Suppose there is exactly one. Then,

$$z < z_G^L: \qquad D_G(\mathscr{J}^*(z); z) \leq D_G(\{\}; z) \leq D_G(\{\}; z_G^L) = 0\,. \tag{5}$$

Assuming the substitutes case establishes the first inequality. It says that, for any firm productivity z, Germany's inclusion into the optimal set can be (weakly) bounded above by its marginal value when added to $L(z) = \{\}$. The reason is that $L(z)$ is a (weak) subset of the optimal set. If the firm opens plants only in countries specified by $L(z)$, there are fewer plants to serve as substitutes to those opened in Germany than if it had acted according to $\mathscr{J}^*(z)$. Therefore, $D_G(L(z); z)$ represents an upper bound on the value obtained by opening German plants. The productivity level z_G^L in (4) is precisely associated with firms for whom this best-case scenario is zero. The second inequality thereby captures that it is nonpositive for those of lower productivity by monotonicity in type. Overall, the chain of inequalities (5) establishes that any firm ranking below z_G^L in productivity cannot optimally open plants in Germany.

Notably, this cutoff is conservative, as it may be the case that plants in Germany will turn out to be suboptimal for firms with higher productivities. However, it is impossible to conclude that Germany is excluded for firm productivities above z_G^L with the current information.

Now, consider the superset $U(\cdot) = \{G, R\}$. A mirror argument can be used to infer whether some firms are better off including Germany. This time, define z_G^U as

$$D_G(\{G,R\}; z_G^U) = \pi(\{G,R\}; z_G^U) - \pi(\{R\}; z_G^U) = 0 \tag{6}$$

if it exists. A firm with this productivity receives no added value from Germany's inclusion in the full set, and would earn equal profits if it were removed. Using both monotonicity conditions arrives at the conclusion

$$z > z_G^U: \qquad D_G(\mathscr{J}^*(z); z) \leq D_G(\{G,R\}; z) \leq D_G(\{G,R\}; z_G^U) = 0\,. \tag{7}$$

Observe this time that $U(\cdot) = \{G, R\}$ (weakly) contains the optimal set. The relative contribution of plants in Germany must then be (weakly) lower if the firm opens plants in both countries than if it acts optimally. Again, this statement derives from the fact that plants are substitutes for each other. The productivity z_G^U is specifically defined in (6) to identify when Germany's additional value in this "extreme" scenario is zero. Firms with productivity above z_G^U are such that the lower bound on Germany's contribution is still positive as a result. Expression (7) combines these conclusions, ultimately implying that firms exceeding z_G^U in productivity must optimally open plants in Germany.

To summarize the inferences up to this point, Germany should be excluded for firms with productivity below z_G^L and included for firms with productivity above z_G^U. If the subsets and supersets were to be updated at this point, G would be removed from $U(\cdot)$ in the range $(-\infty, z_G^L]$. Likewise, G would be added into $L(\cdot)$ over the range $[z_G^U, \infty)$. Observe that z_G^U necessarily exceeds z_G^L by combining the monotonicity assumptions. Then, for all firms with productivity within (z_G^L, z_G^U), the status of the Germany binary decision is still unclear.

Before they are, however, the same process can be repeated for Romania. Undertaking this analysis results in two additional cutoffs, $z_R^L < z_R^U$, with

$$D_R(\{\}; z_R^L) = 0\,, \qquad\qquad D_R(\{G,R\}; z_R^U) = 0\,.$$

Following the exposition above, all firms with productivities below z_R^L should not open plants in Romania while those above z_R^U should. Using these results, $L(\cdot)$ and $U(\cdot)$ can be updated to incorporate the conclusions concerning both Germany and Romania so far. The real line is thus partitioned into five subintervals defined by $\{z_G^L, z_G^U, z_R^L, z_R^U\}$. Additionally, the pair $L(\cdot)$ and $U(\cdot)$ are constant within each subinterval, but at least one changes from one subinterval to the next. Figure 2 displays a possible outcome of the described procedure.

The logic outlined in this section can now be applied in turn to every generated subinterval. Since the subsets and supersets have been weakly narrowed down in each, the updating process has more information for the second iteration. However, observe that it must be applied to each subinterval separately, because it crucially uses the subset and superset to compute indifference points. Thus, it is only defined over productivity ranges where both $L(\cdot)$ and $U(\cdot)$ are constant. Ultimately, the solution method involves iteratively repeating this logic until the

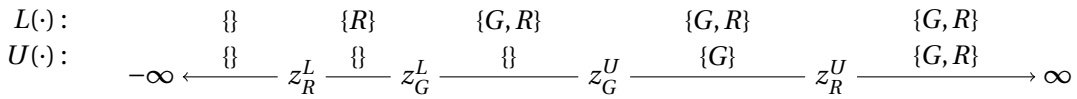


Figure 2: An example of subsets, supersets, and resulting subintervals after one interation.

subset and superset functions no longer change when updated. Describing the full algorithm in depth is covered in Section 3.

Discussion of the simple example concludes for now with the key observation that the optimal sets are not necessarily nested as productivity increases. That is, $z_1 < z_2$ does not imply that $\mathscr{J}^*(z_1) \subseteq \mathscr{J}^*(z_2)$. Consider the case presented in Figure 2 as an example. Within the second subinterval (z_R^L, z_G^L), there could be a value of productivity where Romania is optimally added. Next, examining the fourth subinterval, (z_G^U, z_R^U), reveals that Romania is not necessarily in the optimal set for firms of these productivities. Suppose the extreme case holds where in fact none of these firms should open plants in Romania. Then, Romania is only added back into the optimal set for firms exceeding z_R^U in productivity. There is thus a range of productivity where firms open plants in Romania, then another range later on where they do not, followed by a third range where they will again. Monotonicity in type implies that more productive firms gain more from extra plants, so it is tempting to conclude that they have larger optimal sets. However, substitutability is a force in the opposite direction, since extra elements are less beneficial in larger sets, implying smaller optimal sets. Nested optimal sets consequently only necessarily occur in the complementarities case. In that scenario, both monotonicity in set and type work in the same direction. Higher productivity firms derive more value from additional plants, implying more plants, leading to even higher marginal value from plants from complementarity.

2.3. A preview of method advantages

The proposed solution method contributes to the literature in several ways. While they are discussed in more depth and with more generality in Section 3.3, a preview is provided here in the context of the simple example.

The central contribution is a method to solve explicitly for the policy function, mapping the entire type space to CDCP solutions. Current solution methods for interconnected discrete problems are developed with questions concerning a small number of firms in mind. They therefore solve the CDCP for one firm type at a time using a similar squeezing logic. For example, fixing productivity at a certain z, existing solution methods would check each country's marginal value to the empty set at this productivity. If it were negative for any country, then that country would be discarded from consideration. Similarly, if the marginal value of any country's inclusion into the full set were positive, then plants would definitely be opened there. The process would update subsets and supersets, then iterate. This algorithm is notably proposed by Jia [2008] for the complements case, where she analyzes the behavior of two firms establishing retail locations. Arkolakis and Eckert [2017] expand her approach to include the substitutes possibility but still introduce no dimension of agent heterogeneity.

One obstacle for applying these methods to a heterogeneous agent setting is that they are

fundamentally designed to consider a single type at a time. Direct applicability only occurs when there is a discrete distribution of firms over types. In the illustrative example, this restriction would amount to only allowing a finite number of firm productivities $\{z_1, ..., z_M\}$. Then, the CDCP could be solved for each possible productivity level z_m. A continuous support of firm productivities, common in the international trade literature, must conversely be discretized before using the method in Jia [2008] or Arkolakis and Eckert [2017]. Notably, Antràs et al. [2017] employ this strategy to quantitatively evaluate CDCPs for a continuous distribution of firms in a complements case. However, discretization is unattractive because it introduces computational inaccuracy. If there is little guidance ex ante on where gridpoints should be placed or the correct interpolation method between them, then aggregations over the firm distribution could potentially be very inaccurate. In contrast, the method discussed above remains agnostic on the distribution of firms. By simply developing a mapping from the type space to associated optimal sets, it neither presupposes discrete supports nor does it create discretization error for continuous ones.

Computing the policy function is also faster in general compared to applying a single-type method separately for multiple firm types. Crucially, the updating step in the second strategy only uses the *sign* of an element's marginal value to subsets and supersets. For example, suppose the empty set is chosen as the initial subset for each type when applying the single-type method. The strategy would then proceed by calculating Germany's marginal benefit from inclusion in the empty set for each type, simply to check the sign. Accordingly, monotonicity in type can be leveraged to sign marginal values for entire intervals of the productivity range. Finding a cutoff like z_G^U is therefore equivalent to signing the marginal value of Germany's inclusion into the empty set for *all* productivity values, since monotonicity in type implies that it is negative for all productivities below and positive for all above. The speed gain consequently results for eliminating repeated calculations that would occur otherwise.

Observe that these advantages are closely related when the type space is continuous. If accuracy is very important or little is known ex ante about how the CDCP's solution varies with productivity, then a very dense grid is needed to achieve higher precision. The denser the grid, the slower the discretized alternative, since the CDCP must be solved at more productivity levels. This relationship is quantified for Section 4's general equilibrium model by numerical experiments. The functional approach introduced in this section is therefore valuable in quantitative general equilibrium exercises, because it quickly and exactly characterizes the entire distribution of firm behavior.

3. General solution method

This section formalizes the optimization problems considered in the paper before presenting the algorithm. Section 3.1 begins by defining a combinatorial discrete choice problem and recasts the two assumptions introduced above in a general setting. They turn out to be weaker statements of more standard super- and sub-modularity conditions arising in economic settings, so these sufficient conditions are also discussed. Section 3 details a generalized version of the solution method, defined over a multi-dimensional type space. Then, Section 3.3 discusses its main advantages as well as highlighting nested solution methods.

3.1. Formal overview of CDCPs

The solution method's description begins with a formal definition of a CDCP parameterized by agent heterogeneity. Agents maximize a function

$$\pi : \mathscr{P}(J) \times \mathbb{R}^N_{\mathbf{z}} \times \boldsymbol{\Upsilon} \to \mathbb{R}$$

over its first argument. It is defined over the power set of a known set J, all possible vectors of types $\mathbb{R}^N_{\mathbf{z}} \subseteq \mathbb{R}^N$, and all possible vectors of aggregates $\boldsymbol{\Upsilon}$. While the body of the paper focuses on the problem of finding an optimal subset, which can be cast as a series of 0-1 decisions, Appendix C shows that the problem can be extended to include multiple discrete levels for each decision. Conditional on the objective function, a CDCP can be written as

$$\max_{\mathscr{J} \subseteq J} \pi(\mathscr{J}; \mathbf{z}, \boldsymbol{\upsilon})$$

where $\mathscr{J}$ is a choice of set, $\mathbf{z}$ summarizes individual-level variables which are not choice variables, and $\boldsymbol{\upsilon}$ collects aggregate variables. In the plant location example, π represented a firm's profits, $\mathscr{J}$ the set of countries where the firm opens plants, $\mathbf{z}$ is a scalar denoting the firm's productivity, and $\boldsymbol{\upsilon}$ is the vector containing aggregates such as wages. In most applications, the $\mathbf{z}$ argument will summarize heterogeneous characteristics across agents. Given this interpretation, $\mathbf{z}$ is referred to as the "type vector" for this section. As before, the aim is to find $\mathscr{J}^*(\cdot; \boldsymbol{\upsilon})$, the policy function for each value of type vector conditional on a vector of aggregates. To this end, the marginal value of j's addition into $\mathscr{J}$ is defined as

$$D_j(\mathscr{J}; \mathbf{z}, \boldsymbol{\upsilon}) \equiv \begin{cases} \pi(\mathscr{J}; \mathbf{z}, \boldsymbol{\upsilon}) - \pi(\mathscr{J} \setminus \{j\}; \mathbf{z}, \boldsymbol{\upsilon}) & \text{if } j \in \mathscr{J} \\ \pi(\mathscr{J} \cup \{j\}; \mathbf{z}, \boldsymbol{\upsilon}) - \pi(\mathscr{J}; \mathbf{z}, \boldsymbol{\upsilon}) & \text{if } j \notin \mathscr{J} \end{cases}. \tag{8}$$

Without a strategy, the maximization problem can grow unwieldy quickly. The number of possible choices in contention is $2^{|J|}$ since the maximand is chosen among all of J's possible subsets. Therefore, the choice space grows at an exponential rate in the size of J, presenting a challenge for numerical exercises. Computational time is magnified if CDCPs are defined over heterogeneous types, then included in a general equilibrium model or estimation routine. First, the solution will differ across the type space, so it must be computed at many different points. Second, quantitative exercises usually search for fixed points over aggregates, implying the whole solution distribution must be solved multiple times. However, a growing literature has precisely sought to embed CDCPs into quantitatively estimated general equilibrium frameworks to understand aggregate effects of multinational behavior, differential input sourcing strategies, and other questions in international trade. Third, parameter estimation implies that the general equilibrium fixed point must be solved multiple times for different possible parameter values. The aim of this paper is therefore to address the computational challenge posed by CDCPs in models or estimation routines of this nature.

Next, the two assumptions on the marginal value function introduced in Section 2 are generalized. More easily interpreted sufficient conditions are also discussed for each.

3.1.1. Monotonicity in set

To illustrate the first assumption, consider how much an extra element j would generate if added into two different sets, $\mathscr{J}_1 \subset \mathscr{J}_2$. Then, the monotonicity condition imposes that, for all types $\mathbf{z}$ and vectors of aggregates $\boldsymbol{v}$, either

$$D_j\left(\mathscr{J}_1;\mathbf{z},\boldsymbol{v}\right) \geq D_j\left(\mathscr{J}_2;\mathbf{z},\boldsymbol{v}\right), \text{ or} \tag{MS1}$$

$$D_j\left(\mathscr{J}_1;\mathbf{z},\boldsymbol{v}\right) \leq D_j\left(\mathscr{J}_2;\mathbf{z},\boldsymbol{v}\right). \tag{MS2}$$

The first condition (MS1) asserts that the value of adding an element j to any set falls as the set grows. Conversely, (MS2) implies that the extra element j will become more valuable as the original set grows. As discussed in Section 2, CDCPs where elements are substitutes for each other are broadly linked to property (MS1), while (MS2) generally arises settings where elements complement each other.

This pair of conditions is also outlined in Arkolakis and Eckert [2017], who show that the better-known condition of submodularity on the objective function is sufficient to satisfy (MS1). In this context, submodularity on $\pi(\cdot;\mathbf{z})$ asserts that for any two sets $\mathscr{J}_1$ and $\mathscr{J}_2$,

$$\pi\left(\mathscr{J}_1 \cup \mathscr{J}_2;\mathbf{z},\boldsymbol{v}\right) + \pi\left(\mathscr{J}_1 \cap \mathscr{J}_2;\mathbf{z},\boldsymbol{v}\right) \leq \pi\left(\mathscr{J}_1;\mathbf{z},\boldsymbol{v}\right) + \pi\left(\mathscr{J}_2;\mathbf{z},\boldsymbol{v}\right).$$

In a likewise fashion, supermodularity of the object function $\pi(\cdot;\mathbf{z})$ can be defined by reversing the weak inequality in the above expression. Then, supermodularity of the objective function would be sufficient to guarantee (MS2).

3.1.2. Monotonicity in type

The second condition specifies how the value of an additional element j interacts with types $\mathbf{z}$. To satisfy monotonicity in type, the space

$$\mathbf{Z}_j(\mathscr{J};\boldsymbol{v}) = \left\{\mathbf{z} \in \mathbb{R}^N_{\mathbf{z}} \mid D_j(\mathscr{J};\mathbf{z},\boldsymbol{v}) = \mathbf{0}\right\}$$

must define an $(N-1)$-dimensional subspace partitioning $\mathbb{R}^N_{\mathbf{z}}$ into two cells. In addition, these divisions $\underline{\mathbf{Z}}_j(\mathscr{J};\boldsymbol{v})$ and $\overline{\mathbf{Z}}_j(\mathscr{J};\boldsymbol{v})$ are meaningful in the sense that

$$\begin{cases} D_j(\mathscr{J};\mathbf{z},\boldsymbol{v}) < 0 & \text{if } \mathbf{z} \in \underline{\mathbf{Z}}_j(\mathscr{J};\boldsymbol{v}) \\ D_j(\mathscr{J};\mathbf{z},\boldsymbol{v}) > 0 & \text{if } \mathbf{z} \in \overline{\mathbf{Z}}_j(\mathscr{J};\boldsymbol{v}) \end{cases}. \tag{MT}$$

This statement must hold for any $\mathscr{J} \subseteq J$ and provides a generalization of the condition from the single dimensional case. This condition imposes structure on interactions between the objective function's first and second arguments. Now, there are multiple points $\mathbf{z}$ for which agents are indifferent to j's inclusion into $\mathscr{J}$. However, it is still the case that these indifference points divide the type space $\mathbb{R}^N_{\mathbf{z}}$ into two distinct halves. In turn, the halves directly coincide with the values of $\mathbf{z}$ for which j's inclusion into $\mathscr{J}$ respectively creates positive or negative additional value.

A simple sufficient condition amounts to asserting that $D_j(\mathscr{J};\mathbf{z})$ is monotonically increasing in each dimension of $\mathbf{z}$. Formally, if $\mathbf{z} = [z_1, z_2, ..., z_N]$, it is sufficient that

$$\frac{\partial D_j(\mathscr{J};\mathbf{z},\boldsymbol{v})}{\partial z_i} \geq 0$$

for each i. Observe that if instead marginal value is decreasing in any dimension, the problem can be recast to satisfy the original condition. At its core, this sufficient condition thus only requires that $D_j(\mathscr{J};\cdot,\boldsymbol{v})$ is (weakly) monotonic in every characteristic z_i. For example, a plant location problem where firm heterogeneity arises from varying idiosyncratic plant fixed costs could satisfy this sufficient condition. A concrete example of this problem is discussed in Appendix B, along with weaker sufficient conditions.

If there exist some dimensions of heterogeneity that do not satisfy monotonicity in type, the paper's solution method can still make progress over the dimensions that do. For example, suppose that monotonicity in type holds for the first $M < N$ dimensions of $\mathbb{R}^N_{\mathbf{z}}$, if the remaining ones are fixed at any value. Then, redefine the objective function as

$$\pi(\mathscr{J};\cdot,\cdot,\cdot) : \mathscr{P}(J) \times \hat{\mathbf{Z}} \times \mathbf{X} \times \boldsymbol{\Upsilon} \to \mathbb{R}$$

where $\hat{\mathbf{Z}}$ is the subspace of $\mathbb{R}^N_{\mathbf{z}}$ spanning the first M dimensions, and $\mathbf{X}$ is the subspace spanning the rest. The method can then solve for $\mathscr{J}^*(\cdot;\mathbf{x},\boldsymbol{v})$ over $\hat{\mathbf{Z}}$, at a specific value $\mathbf{x} \in \mathbf{X}$.

3.2. Solution method

This section describes the generalized solution method for a CDCP defined over a type space. To do so, Section 3.2.1 first presents a generalization of the updating procedure described previously. It is then incorporated into an iterative routine for updating the subsets and supersets, which is part of a larger branching procedure. Section 3.2.2 presents this outer branching procedure. Finally, Section 3.2.3 summarizes the full solution method, including the inner iterative routing for updating the subsets and supersets as well as the outer branching procedure.

3.2.1. Iterative step

This section generalizes the procedure used to update subsets and supersets during Section 2's illustrative example. Fix a vector of aggregates $\boldsymbol{v} \in \boldsymbol{\Upsilon}$. For these aggregates, the algorithm takes a connected subspace $\mathbf{Z} \subseteq \mathbb{R}^N_{\mathbf{z}}$ of the type space as well as subset and superset L and U.[3] It returns a partition of $\mathbf{Z}$ into smaller connected subspaces, each with their own (L,U) set pair. The analog in Section 2's single dimensional case would be intervals, each with their own (L,U) pair, as depicted in Figure 2. The iterative step updates $L(\cdot)$ and $U(\cdot)$ by repeating the process on each resulting connected subspace until updating has no effect. Since the rest of the discussion in this section applies for a fixed vector of aggregates $\boldsymbol{v}$, this indexing is dropped for notational brevity when unambiguous.

[3] As before, L and U represent sets, while $L(\cdot)$ and $U(\cdot)$ represent set-valued *functions* defined over portions of the type space.

Specifically, consider an initial subspace $\mathbf{Z}$ associated with a subset and superset (L, U). Assume that these sets sandwich the optimal set over the subspace so that for every $\mathbf{z} \in \mathbf{Z}$, $L \subseteq \mathscr{J}^*(\mathbf{z}) \subseteq U$. This assumption is not binding since it is met trivially by the empty set and full set respectively. Notice that L and U are not parameterized by $\mathbf{z}$, and are therefore constant for all $\mathbf{z} \in \mathbf{Z}$. Then, $U \setminus L$ contains the elements for which the optimal binary decision is still uncertain. The updating step is defined on these decisions as follows.

Algorithm 1 (Updating step)**.** *Given a triple* $(\mathbf{Z}, L, U)$*,*

(i) For each element $j \in U \setminus L$*:*

a. Divide $\mathbb{R}^N_{\mathbf{z}}$ *into two cells* $\underline{\mathbf{Z}}_j(L)$ *and* $\overline{\mathbf{Z}}_j(L)$ *where*

$$\begin{aligned} \mathbf{z} \in \underline{\mathbf{Z}}_j(L) &\quad \Leftrightarrow \quad D_j(L;\mathbf{z}) < 0 \\ \mathbf{z} \in \overline{\mathbf{Z}}_j(L) &\quad \Leftrightarrow \quad D_j(L;\mathbf{z}) > 0 \end{aligned}.$$

Monotonicity in type (MT) guarantees that this partitioning exists.

b. Similarly, divide $\mathbb{R}^N_{\mathbf{z}}$ *into two cells* $\underline{\mathbf{Z}}_j(U)$ *and* $\overline{\mathbf{Z}}_j(U)$ *where*

$$\begin{aligned} \mathbf{z} \in \underline{\mathbf{Z}}_j(U) &\quad \Leftrightarrow \quad D_j(U;\mathbf{z}) < 0 \\ \mathbf{z} \in \overline{\mathbf{Z}}_j(U) &\quad \Leftrightarrow \quad D_j(U;\mathbf{z}) > 0 \end{aligned}.$$

(ii) With diminishing marginal returns (MS1), define $\hat{L}(\cdot)$ *and* $\hat{U}(\cdot)$ *over* $\mathbf{Z}$ *as:*

$$\begin{aligned} \hat{L}(\mathbf{z}) &= L \cup \{j \in U \setminus L \mid \mathbf{z} \in \overline{\mathbf{Z}}_j(U)\} \\ \hat{U}(\mathbf{z}) &= U \setminus \{j \in U \setminus L \mid \mathbf{z} \notin \overline{\mathbf{Z}}_j(L)\} \end{aligned}.$$

Otherwise, with increasing marginal returns (MS2), define $\hat{L}(\cdot)$ *and* $\hat{U}(\cdot)$ *over* $\mathbf{Z}$ *as:*

$$\begin{aligned} \hat{L}(\mathbf{z}) &= L \cup \{j \in U \setminus L \mid \mathbf{z} \notin \underline{\mathbf{Z}}_j(L)\} \\ \hat{U}(\mathbf{z}) &= U \setminus \{j \in U \setminus L \mid \mathbf{z} \in \underline{\mathbf{Z}}_j(U)\} \end{aligned}.$$

(iii) Given the newly defined $\hat{L}(\cdot)$ *and* $\hat{U}(\cdot)$*, partition* $\mathbf{Z}$ *into cells* $\{\mathbf{Z}^e\}_e$ *so that for every* $\mathbf{Z}^e$*,* $\hat{L}(\cdot)$ *and* $\hat{U}(\cdot)$ *are both constant. Let these associated constant subsets and supersets be* L^e *and* U^e *for each* $\mathbf{Z}^e$ *respectively.*

(iv) Return a set of triples $\{(\mathbf{Z}^e, L^e, U^e)\}_e$*.*

Algorithm 1 provides the formal generalization of the logic described in section 2. Part (i) uses monotonicity in type (MT) to divide the characteristic space for each uncertain decision j, according to the provided L and U. Then, part (ii) uses monotonicity in set (MS1)-(MS2) to construct subsets and supersets by either adding items into L or excluding them from U. Finally, the originally given space $\mathbf{Z}$ is partitioned into a collection of subspaces $\{\mathbf{Z}^e\}_e$, each with constant (L^e, U^e), in anticipation of the next iteration.

The proposition below establishes that Algorithm 1 indeed moves the routine closer to the optimal set, without wrongful inclusions or exclusions. Proofs are provided in Appendix A.2.

Proposition 1. *Suppose that* $(\mathbf{Z}, L, U)$ *are such that, for all* $\mathbf{z} \in \mathbf{Z}$*,* $L \subseteq \mathscr{J}^*(\mathbf{z}) \subseteq U$*. Then, for each* $\mathbf{Z}^e$ *returned by Algorithm 1,* $\mathbf{z} \in \mathbf{Z}^e$ *implies* $L \subseteq L^e \subseteq \mathscr{J}^*(\mathbf{z}) \subseteq U^e \subseteq U$*.*

Though short, Proposition 1 states two key results. The first is that the updated subsets and supersets are correct. Alternately put, they truly sandwich the optimal set. The second is that the updating step weakly incorporates more information. The new subsets contain the original subsets while the new supersets are contained in the original supersets. Then, through iteration, Algorithm 1 can close in on the optimal set in a way that takes advantage of the CDCP's structure. The iterative step is now explicitly defined.

Algorithm 2 (Iterative step)**.** *Given an initial triple* $(\mathbf{Z}, L, U)$*:*

(i) Define two collections, C and S. The set C will contain triples to be iterated on ("Continue") while the set S will contain triples where iteration has completed ("Stop").

(ii) Set $C = \{(\mathbf{Z}, L, U)\}$ *and* $S = \{\}$*.*

(iii) Iteratively, for each element $c \in C$*:*

- *a. Remove c from C.*
- *b. Apply Algorithm 1 to c, which will return a collection R of triples ("Returned").*
- *c. If* $R = \{c\}$*, then c cannot be updated further. Thus, it should be added to S. Otherwise, add the elements from R into C for further updating.*

(iv) Stop when C is empty. Return S, which will be a collection of triples $\{(\mathbf{Z}^e, L^e, U^e)\}_e$*. In pariticular,* $\{\mathbf{Z}^e\}_e$ *will partition the characteristic space.*

The outcome of Algorithm 2 will effectively be two functions, $L(\cdot)$ and $U(\cdot)$, summarizing the subsets and supersets. They are defined over the type space so that for any $\mathbf{z} \in \mathbf{Z}$, $L(\mathbf{z}) \subseteq \mathscr{J}^*(\mathbf{z}) \subseteq U(\mathbf{z})$. This result follows directly from Proposition 1. The next proposition confirms that this iteration stops in finite time.

Proposition 2. *With any initialization, Algorithm 2 repeats step (iii) a maximum of* $|J|$ *times.*

There are at most $|J|$ binary decisions to determine, with each instance of step (iii) fixing at least one. If no decisions are determined, then the returned subsets and supersets will not have changed. In this case, the algorithm has converged by step (iii)c., so it stops. Thus, it can continue repeating a maximum of $|J|$ times, if each time only determines one binary decision.

Effectively, the iterative step reduces the dimensionality of a CDCP, returning a smaller problem. To fix ideas, return to the plant location problem with heterogeneity. For each type vector $\mathbf{z}$, $L(\mathbf{z})$ contains countries in which firms of this type should open plants. For example, if Germany is contained in $L(\mathbf{z})$, then the binary decision for Germany is clear. The firm will optimally open plants there. Likewise, $U(\mathbf{z})$ excludes countries where plants should not be opened. If Romania is not an element in this superset, then the binary decision on Romania has also been determined. Overall, the iterative step uses both monotonicity conditions to draw conclusions for some of the binary choices to progress towards the optimal sets.

Nevertheless, the decision of whether or not to open plants is still undetermined for every country contained in $U(\mathbf{z})$ but not in $L(\mathbf{z})$. The iterative step's output remains a CDCP over these. However, by Proposition 1, the resulting CDCP is (weakly) reduced compared to the original in the sense that there are fewer binary decisions to determine. For some type vectors, it may be that $L(\mathbf{z}) = U(\mathbf{z})$. In those cases, the reduced CDCP is trivial, with no undetermined binary decisions left. Then, the optimal set has been found. Conversely, in areas of the type space where $L(\mathbf{z}) \subset U(\mathbf{z})$, a nontrivial CDCP is left.

Observe that Algorithm 2 therefore takes one CDCP and returns a series of simpler ones, some even trivial. Why not simply reapply Algorithm 2 to the nontrivial ones? For these returned CDCPs, the iterative step will not simplify them further. To see why, observe that the simplified CDCPs were themselves returned from Algorithm 2. However, by step (iii)c., these are precisely the instances where the updating step, Algorithm 1, makes no progress. Therefore, for the reduced CDCPs that are nontrivial, another approach must be taken. The next section addresses this issue with a branching step.

3.2.2. Branching step

Before describing the branching step formally, return to Figure 2, which illustrates a possible outcome of the updating step in the simple example. For firm productivities within (z_G^L, z_G^U), neither the subset nor superset were updated. In other words, the binary decisions on both Germany and Romania remain undetermined, posing exactly the challenge described at the end of the previous section.

Falling back on the brute force strategy, while possible, is an unattractive option. Doing so would entail exhaustively guessing all possibilities for both the Germany and Romania decisions, then comparing across combinations. This approach does not take advantage of the CDCP's structure afforded by the monotonicity conditions. Consider instead only guessing for the decision of whether or not to place plants in Germany. The solution method now creates two branches. One corresponds to the scenario where plants are opened in Germany, while the other represents the one where they are not. In either case, the decision on Romanian plants remains undetermined. Rather than guessing for this one as well, as in the brute force method, the updating step can now be reapplied. Each branch will then summarize "locally optimal" behavior, conditional on the guess of whether or not German plants are established. The objective function must then be compared across branches to find the globally optimal solution. Reapplying Algorithm 2 in each branch therefore uses the monotonicity assumptions to draw conclusions for as many choices as possible, minimizing the number determined by guessing.

In practice, the two triples below would be defined, on which Algorithm 2 is reapplied.

$$\left((z_G^L, z_G^U), \{\}, \{R\}\right) \qquad\qquad \left((z_G^L, z_G^U), \{G\}, \{G, R\}\right)$$

While the intervals are the same in both, the subsets and supersets are not. They reflect the guesses for the Germany decision. The first triple corresponds to the branch where no plants are opened in Germany. In that case, the subset contains no countries, while the superset contains only Romania. This definition of the subset and superset rules out opening plants in Germany. On the other hand, the second represents the opposite case. German plants are assumed to be opened, so the subset contains Germany. Since Romanian plants can still be opened, the superset contains both countries. This subset-superset pair likewise disallows the option of not opening plants in Germany. Then, Algorithm 2 can be iterated on both of these branches.

Figure 3 shows a possible outcome from this exercise. The top row represents the policy function if no plant is opened in Germany, $\mathscr{J}^*(\cdot \mid G \to 0)$. Under this assumption, $z_R^{G\to 0}$ emerges as the cutoff value separating firms that should open plants in Romania from those

Figure 3: A possible outcome from applying the recursive step once.

that should not. Likewise, $\mathscr{J}^*(\cdot \mid G \to 1)$ summarizes optimal behavior if instead German plants are opened. This branch identifies a different productivity cutoff $z_R^{G\to 1}$ regarding Romanian plants. The policy function in each branch must now be compared across branches. Concretely, there are three relevant intervals. Each has its own pair of optimal set candidates, one from each branch. As an example, these are $\{G\}$ and $\{R\}$ for firms with productivity within $(z_R^{G\to 0}, z_R^{G\to 1})$. Compared to the brute force strategy, where there are four possible combinations to compare, the branching approach has half the number of combinations.

In general, consider a triple $(\mathbf{Z}, L, U)$ emerging from Algorithm 2 where $L \subset U$.[4] Selecting some undetermined item $j \in U \setminus L$, create two branches. Proceed assuming that j is included in the optimal set along one branch, while the other branch supposes it is not. In each branch, reapply Algorithm 2 accordingly to fix as many of the remaining choices as possible with the economic logic from the previous section. Of course, in either branch, it may be that Algorithm 2 still leaves some undetermined for some intervals. In those intervals, new branches must be created in the same way. Formally, the branching step is defined as follows.

Algorithm 3 (Branching step)**.** *Given a triple $(\mathbf{Z}, L, U)$ that cannot be further updated by Algorithm 2, but with $L \subset U$:*

(i) Define two collections, C and S. The set C will contain triples for recursion ("Continue") while the set S will contain tuples where recursion has completed ("Stop").

(ii) Set $C = \{(\mathbf{Z}, L, U)\}$ and $S = \{\}$.

(iii) For each triple $c \in C$:

a. Remove c from C.

b. Let $c = (\mathbf{Z}^c, L^c, U^c)$. Then, select an item $j \in U^c \setminus L^c$ and create two new triples

$$(\mathbf{Z}^c, L^c, U^c \setminus \{j\}) \qquad\qquad (\mathbf{Z}^c, L^c \cup \{j\}, U^c)\,.$$

c. Apply Algorithm 2 to each, returning corresponding collections $R^{j\to 0}$ and $R^{j\to 1}$.

d. For each $r \equiv (\mathbf{Z}^r, L^r, U^r) \in R^{j\to 0}$, check if $L^r = U^r$. When they are equal, add r into S. If they are not, add r to C. Proceed similarly for all $r \in R^{j\to 1}$.

(iv) Stop when C is empty. S will be a collection of tuples $\{(Z^e, L^e = U^e \equiv \hat{\mathscr{J}}^e)\}_e$. Form an "augmented policy function" returning the optimal sets from the recursion's branches:

$$\hat{\mathscr{J}}^*(\mathbf{z}) = \{\hat{\mathscr{J}}^e \mid \mathbf{z} \in \mathbf{Z}^e\}\,.$$

(v) Partition the original $\mathbf{Z}$ so that $\hat{\mathscr{J}}^*(\cdot)$ is constant within each cell. For each cell, determine $\mathscr{J}^*(\cdot)$ by using the objective function to globally select among the sets in $\hat{\mathscr{J}}^*(\cdot)$.

[4] The e superscript is dropped since these are considered one at a time.

Note that, with any initialization, Algorithm 3 creates at most $2^{|U|-|L|}$ branches. This case corresponds to a worst-case scenario where applying the iterative step in each branch accomplishes nothing. Then, every remaining decision $j \in U \setminus L$ in the reduced CDCP must be determined by guessing, ultimately reducing in this worst-case scenario to brute force.

3.2.3. Full solution method

Section 3.2.1 defined a generalized updating routine Algorithm 1, following the intuition from Section 2. It is the fundamental building block of iterative Algorithm 2 and, in turn, branching Algorithm 3. With these now specified, the solution method can be laid out in its entirety.

Algorithm 4 (Full method)**.** *To find the policy function $\mathscr{J}^*(\cdot)$ for a CDCP defined over $\mathscr{P}(J)$*[5] *and type space $\mathbb{R}^N_{\mathbf{z}}$:*

(i) Perform Algorithm 2 on $(\mathbf{Z}, \{\}, J)$, returning R.

(ii) For each triple $(\mathbf{Z}^r, L^r, U^r) \equiv r \in R$:

(a) Check if $L^r = U^r$. If they are equal, set $\mathscr{J}^*(\mathbf{z}) = L^r$ for all $\mathbf{z} \in \mathbf{Z}^r$.

(b) If $L^r \subset U^r$, use Algorithm 3 on $(\mathbf{Z}^r, L^r, U^r)$ to return $\mathscr{J}^*(\cdot)$ over $\mathbf{Z}^r$.

(iii) Return $\mathscr{J}^*(\cdot) : \mathbf{Z} \to \mathscr{P}(J)$.

At its core, the algorithm harnesses the intuition outlined in Section 2 to update subsets and supersets, thereby ruling out sets that are not optimal at each $\mathbf{z}$. This intuition is formalized and generalized in Algorithm 1. From there, Algorithm 2 simply embeds the updating step in a standard iterative loop which converges to a fixed point on the pair of functions $L(\cdot)$ and $U(\cdot)$. However, it is possible that this fixed point falls short of defining the CDCP's solution. Algorithm 3 includes the iterative loop in a larger branching procedure to handle these cases.

Let the fixed points emerging from Algorithm 2 be $L^\infty(\cdot)$ and $U^\infty(\cdot)$. In the worst-case scenario, Algorithm 2 could make no progress for a region of the type space. That is, there could be some $\tilde{\mathbf{Z}}$ where $L^\infty(\mathbf{z}) = \{\}$ and $U^\infty(\mathbf{z}) = J$ for $\mathbf{z} \in \tilde{\mathbf{Z}}$. Similarly, the worst-case scenario of the recursion defined in Algorithm 3 simply creates a branch for every undetermined decision. Taken together, the overall worst-case for the full solution method amounts to computing the optimal policy function by brute force over $\tilde{\mathbf{Z}}$. In the simulation exercises presented in Section 4.3.2, such worst case scenarios does not occur in conventional settings.

3.3. Method advantages

To the best of our knowledge, this paper is the first to solve CDCPs with heterogeneous agents in either a substitutes or complements case by explicitly solving for the policy function. As discussed in Section 2.3, there are several key advantages to applying this approach when computing or estimating general equilibrium models with agents of different types facing CDCPs. Beyond the two monotonicity assumptions, it imposes no additional restrictions on the problem. In particular, agents can be distributed across the type space in any way. Monotonicity on type is the key condition, allowing the policy function to be solved exactly

[5] If some decisions in J are pretermined for some prior reason, then the problem can be redefined over the smaller set of unknown elements.

and quickly across the entire type space. Finally, the proposed procedure provides a unified method to compute CDCPs in both the substitutes and the complements case.

The solution method's main conceptual contribution is the fact that it returns a policy function defined over a type space of arbitrary dimension. As a result, no assumptions are placed on the distribution of agents across types. It is therefore more flexible compared to the solution method in the closest related paper, Arkolakis and Eckert [2017]. They build on Jia [2008] to develop an algorithm solving CDCPs for a single type, conditional on either (MS1) or (MS2). In the special case where agents are identical so the type space is a single point, the solution method presented in this paper collapses to theirs. However, once there is nontrivial heterogeneity, their method can only be applied type by type. Doing so is possible if there is a finite number of types, and may be computationally feasible for a small number of types, but cannot be generalized over a continuous support. On the other hand, heterogeneous agent models are the central motivation for the solution method developed here. Consequently, the method is broadly applicable to general type distributions, whether discrete or continuous.

The key innovation enabling this flexibility is the monotonicity in type condition. Incorporating an arbitrary number of heterogeneous dimensions requires specifying a condition disciplining the CDCP's interaction with agent type. Otherwise, there is no way to relate an agent's type with their optimal solution to the CDCP. The monotonicity in type condition (MT) is relatively general and is satisfied by existing heterogeneous agent CDCP models in the literature, it also allows a generalized method of solving CDCPs over many dimensions of heterogeneity at once.

The monotonicity on type condition not only enables this conceptual advantage, but also brings dramatic improvements in computational speed. The current numerical alternative is to solve the CDCP at multiple, if not all, points within the type space. For example, Tintelnot [2016] and Antràs et al. [2017] employ this strategy in their quantitative exercises. However, the paper's algorithm is considerably faster if finding the types for which the indifference conditions hold is relatively straightforward. In the language of the simple example from Section 2, these would be the cutoff productivities. Appendix B works through the models considered by the cited studies and shows that the indifference conditions in those studies are also easily identified. Centrally, the sign of an element's additional value from inclusion in subsets or supersets is sufficient for either procedure. In particular, the single-type algorithm might evaluate a single element's additional value to the empty set at a specific type. If the items are substitutes and this marginal value is negative, then it would be discarded. Likewise, in the complements case, a positive marginal value implies the item should be included. In either case, the precise marginal value is not required, only the sign. This calculation would then be carried out for every point in the type space. Thus, the monotonicity on type condition guarantees the sign of these marginal values for the whole type space as soon as the types for which the indifference conditions hold are identified. Considerable speed gains are therefore derived from foregoing performing these calculations for every type.

With continuous type spaces, the solution allows for both speed and accuracy. In contrast, discretization methods entail an accuracy-speed tradeoff, governed by the density of the grid defined over the type space. Section 4.3 provides a numerical illustration of this tradeoff in the application's quantitative model. The key observation is that interpolation approximation is necessary to aggregate a continuous distribution over a collection of discrete

gridpoints. Then, sparser grids are associated with fast compute times but inaccurate aggregations, while the opposite is true for dense grids. However, the proposed solution method leverages monotonicity in type to solve the exact policy function quickly as described above, eliminating the need to choose between accuracy and speed.

4. An application to EU corporate tax equalization

To demonstrate the performance of algorithm in practice, this Section applies it to evaluate the effects of corporate tax equalization in the European Union. In particular, corporate taxes are currently set individually by each member state. The resulting dispersion, combined with the EU's policies of free trade and movement of people, has been cited by high-tax countries as relocating jobs and industry unfairly to their low tax counterparts. Accordingly, they have pushed for a EU-wide corporate tax harmonization. Section 4.1 describes in more detail the current state of corporate taxes in the EU. A quantitative model of plant location decisions within the EU is specified in Section 4.2, while Section 4.3 shows how the multinational firm's choice problem can be solving using the algorithm developed above. A key feature of the model is that firms internalize tax differences when deciding where to locate production. Section 4.4 embeds the firm's problem in general equilibrium. Finally, calibration and policy counterfactuals are conducted in Section 4.5.

4.1. Corporate taxation in the EU

While the EU has an integrated market and currency union, corporate taxation is individually controlled by member states. As a result, significant tax rate dispersion persists across the countries. In 2014, the lowest marginal tax rate was 10% in Bulgaria, compared to the top rate of 34% in Belgium.

Corporate taxation can either be territorial, on profits made within the borders of a country, or residential, on worldwide profits of firms registered in the country. In practice, an extensive set of deferral policies and double taxation treaties between EU states effectively eliminate taxation on profits made in other EU countries. In addition, EU-wide directives waive withholding tax on interest, royalties, and repatriated profits across borders within the EU. The model thus proceeds with the reasonable approximation that profits made by an EU firm are taxed in the country of production only. In most EU countries, corporate tax is assessed at a flat rate, so average tax rate is equal to marginal tax rate. While a handful of countries provide reduced rates for small firms, the top marginal statutory tax rate is used to capture the incentives for multinational enterprises, which tend to be large.

4.2. Modeling the plant location decision

Firms produce a single differentiated product, competing monopolistically in each destination market where goods are substitutable at a constant rate ρ. The economy contains 27 destination markets, one for each EU country.[6] For simplicity, there are no fixed costs of market access, so firms serve all markets. Markets are reached through exporting, supplying

[6] Due to its small size and data limitations, Malta is excluded from the analysis.

from a local plant (horizontal FDI), or shipping from a plant in a third country (export platform FDI). When selecting the set of countries in which to open plants, firms therefore take into account market access, among other factors like production capability. The full set of considerations for each possible plant location will be detailed below. This setting is similar to that of Tintelnot [2016], but the algorithm is harnessed to increase the state space from 12 countries to 27.

Beyond being differentiated by their goods ω, firms also vary by country of birth i and core productivity $z(\omega)$. This core productivity represents the firm's innate production ability, which will affect all plants. For example, a firm may have a particularly efficient management philosophy which is applied at all its plants. Every firm is endowed with a plant in their country of birth, but can also open them in foreign countries at a fixed cost. After firm entry, core productivity is observed, then a set of countries $\mathscr{J} \subseteq J$ in which to establish plants is chosen. Each foreign affiliate costs f_j units of domestic labor to set up, while the domestic plant is endowed upon entry. Thus, $f_{ij} = \mathbb{1}[i \neq j] f_j$ will be used to denote the labor costs of establishment to be paid. The paper follows the convention of indexing country of origin by i, plant location with j, and destination market using k.

Each plant has its own idiosyncratic productivity, which is observed once plants are opened and fixed costs are sunk. These are distinct from the firm's core productivity, summarizing the individual characteristics of plants. For example, each plant may have access to different local technologies, with some being more productive than others. Unit costs are $w_j/\varphi_{ij}(\omega)$ at a plant j with idiosyncratic productivity $\varphi_{ij}(\omega)$, where w_j is the unit cost of labor.

At this stage, firms choose a sourcing strategy for each destination market. Since production is constant returns to scale at each plant and sales in one destination market do not affect demand in another, the firm can determine its sourcing strategy for every destination market separately. In addition to the unit costs of production above, firms also face variable iceberg costs of trade $d_{jk} \geq 1$ and arms-length production $\gamma_{ij} \geq 1$. That is, d_{jk} units of output must be shipped to ensure 1 unit arrives in market k from a plant in j. Likewise, input requirements are magnified by γ_{ij} if a firm born in i produces in j. These costs are meant to capture the various challenges a multinational firm may encounter when producing in a different country, like costs of communication. Taken together, a firm born in i will have marginal cost $w_j \gamma_{ij} d_{jk}/\varphi_{ij}(\omega)$ of serving market k from its plant in j. The sourcing problem for each market k amounts to choosing the plant $j(k)$ generating the highest variable profits.

These variable profits are determined as follows. Given the CES monopolistic competition market structure, prices will be set at a constant markup $\rho/(\rho-1)$ over marginal costs. In addition, variable profits are taxed at a flat rate τ_j in the country of production. Then, for a firm born in i, serving market k from plant j would earn variable profits

$$(1-\tau_j) X_k^T \left(\frac{\rho}{\rho-1} \frac{w_j d_{jk} \gamma_{ij}}{\varphi_j(\omega) P_k^T} \right)^{1-\rho},$$

where X_k^T is total expenditure on traded goods by consumers in k and P_k^T is the CES price index. Market by market, the firm selects $j(k)$ among its plants that maximizes this value. In particular, both marginal cost considerations and tax burden enter into this decision.

Until this point, firms have been restricted to opening only one plant in each country. However, since marginal costs are constant, the firm problem can instead be reinterpreted to allow for multiple plants in any country. In this case, the fixed cost f_{ij} would represent a firm's one-time cost to establish a production presence in a foreign country j. For example, these could be the fees of officially registering the firm with tax authorities. With this interpretation, plant establishment would carry no further fixed costs. Then, without loss of generality, the paper continues assuming there is at most one plant per country.

The discussion above determines the firm's choices once plants are established. Now, taking a step back, consider how firms should choose the set of locations in which to place plants $\mathscr{J}$. Since they do not observe idiosyncratic plant productivities $\varphi_{ij}(\omega)$ until after fixed costs are sunk, they select $\mathscr{J}$ by maximizing the expected value of profits, taken over possible realizations of these plant productivities $\varphi_{ij}(\omega)$. With the local technology interpretation, this assumption means firms do know the exact quality of local technology available to the plant before opening. It then does its best given the known distribution of local technology quality.

To characterize this expectation sharply, idiosyncratic plant productivites are assumed to be drawn from a Fréchet distribution with shape θ and scale $z(\omega)^{1/(\rho-1)}T_j$. Loosely, the Fréchet's shape inversely governs dispersion while scale is related to the mean. While θ is common across countries, the scale contains a country-specific component that is scaled by the firm's core productivity. As an implication, firms with higher core productivity will have more productive plants on average. For example, firms with a better management style may be able utilize any quality local technology more efficiently than firms with a bad management style. Likewise, plants opened in countries with high T_j will be more productive. Then, local technology in countries with high T_j is more efficient, holding firm type constant.

Using the Fréchet properties popularized by Eaton and Kortum [2002], expected profits are

$$\pi_i(\mathscr{J};z,\boldsymbol{v}) = Az\sum_k \frac{X_k^T}{\left(P_k^T\right)^{1-\rho}}\Theta_{ik}(\mathscr{J};\boldsymbol{v})^{\frac{\rho-1}{\theta}} - \sum_j \mathbb{1}[j\in\mathscr{J}]\,w_j f_{ij}\,, \tag{9}$$

where

$$\Theta_{ik}(\mathscr{J};\boldsymbol{v}) = \sum_j \mathbb{1}[j\in\mathscr{J}]\left(1-\tau_j\right)^{\frac{\rho-1}{\theta}}\left(w_j d_{jk}\gamma_{ij}/T_j\right)^{-\theta} \equiv \sum_j \mathbb{1}[j\in\mathscr{J}]\kappa_{ijk} \tag{10}$$

and A is a constant. The vector of aggregates $\boldsymbol{v}$ in this context contains wages, tradables prices, and tradables expenditure in each country $\{X_k^T, P_k^T, w_k\}_k$. The $\Theta_{ik}(\cdot)$ expression here is similar to market access terms found in the gravity literature, with the key difference that it is a *function* of the chosen set $\mathscr{J}$. However, Section 4.4 will show that gravity will not hold for aggregate trade flows.

Since $\Theta_{ik}(\cdot;\boldsymbol{v})$ is a function of $\mathscr{J}$, it describes the (variable) profit potential of a particular set $\mathscr{J}$. Observe that it is obtained by adding together the κ_{ijk} terms associated with each country j in the set $\mathscr{J}$, which incorporate the characteristics of a country salient for profitability. Namely, they are the base cost of production w_j/T_j, trade market access d_{jk}, losses from arms-length production γ_{ij}, and the country's tax rate τ_j. The first three capture marginal cost considerations while the last represents tax burden, so these terms specify exactly how firms trade off a plant's expected marginal cost with its associated tax rate. Note

also $\Theta_{ik}(\mathscr{J};\boldsymbol{v})$ varies by the market k and origin i, due to bilateral costs of trade and arms-length production respectively. Each set is consequently valued differently depending on the firm's country of origin as well as the destination market in question.

There are several properties of the expected profit function of note. First, expected variable profits increase mechanically as a new plant j is added. Recall that for every market k the firm selects the plant delivering the highest variable profits. Then, the addition of another plant to this choice set can only increase the expected value of the maximum. If its ultimate profitability is lower than the maximum from the original set $\mathscr{J}$, it is simply not chosen. Accordingly, the additional plant's value is higher if its marginal costs or tax rate is lower, since it will exert more upward pressure on the maximum.

Next, the profit potential terms are weighted by each market k's relative size then added to arrive at the expected profits in (9). However, they are first raised by the exponent $(\rho-1)/\theta$, the key model component governing the substitutability of the plants. In this context, it is necessary to have $\rho-1<\theta$ for the expectation over $\varphi_{ij}(\omega)$s to exist. Then, inspection of the expected profit function (9) reveals that it is concave in each profit potential term. A direct implication is that plants are substitutes, with the strength of substitutability increasing as the exponent falls.

To clarify this point, consider the marginal value of j's inclusion into $\mathscr{J}$ in this context.

$$D_{ij}(\mathscr{J};z,\boldsymbol{v}) = Az\sum_k \frac{X_k^T}{\left(P_k^T\right)^{1-\rho}}\left[\left(\kappa_{ijk}+\left(\sum_{j\neq j'\in\mathscr{J}}\kappa_{ij'k}\right)\right)^{\frac{\rho-1}{\theta}}-\left(\sum_{j\neq j'\in\mathscr{J}}\kappa_{ij'k}\right)^{\frac{\rho-1}{\theta}}\right]-w_j f_{ij} \tag{11}$$

As the profit potential of the original set $\mathscr{J}$ decreases, concavity implies that a plant j's marginal value rises. Broadly, when the original set is not expected to be profitable for market k, the additional plant will likely be the one offering maximal profits. Then, its presence in the set increases the set's profit potential substantially. On the other hand, if the original set is already quite profitable in expectation, for example through low tax rates, an additional plant will not improve it as much. In this way, the plants are substitutes since only one is ultimately chosen to serve each market. The chosen plant therefore cannibalizes potential profits from the others. In fact, it can be shown that the ex ante probability that $j\in\mathscr{J}$ is chosen to serve market k is $p_{ijk}(\mathscr{J};\boldsymbol{v}) = \mathbb{1}[j\in\mathscr{J}]\kappa_{ijk}/\Theta_{ik}(\mathscr{J};\boldsymbol{v})$. This expression provides another way to see that a specific plant's chance of serving a market is decreasing in the profit potential of the others available to the firm. Formally, the CDCP can be verified to satisfy (MS1).

Observe also that the concavity intensifies when θ increases or ρ decreases. Intuitively, higher θ translates to lower dispersion in plant productivity draws, so that plants are more similar. Likewise, if ρ is low, consumers are less responsive to the price of goods. Then, it is less important for the firm to lower marginal costs, and productivity advantages of one plant over another make little difference. Both comparative statics decrease the dispersion in potential variable profits afforded by different plants, rendering them more substitutable.

Finally, it is immediately clear from (11) that the CDCP satisfies montonicity on type (MT). In fact, the marginal benefit of any j's inclusion into $\mathscr{J}$ linear in core productivity $z(\omega)$. Then, the algorithm developed earlier in the paper can be applied to solve for $\mathscr{J}_i^*(\cdot;\boldsymbol{v})$, the policy function for firms originating from i conditional on aggregates $\boldsymbol{v}$. In particular, numerically

calculating the cutoff productivities in each iteration is straightforward given the linear nature of (11) on firm type.

This section ends with a discussion of the firm's incentives. A plant j's marginal value in $\mathscr{J}$ given by (11) ultimately compares the change in profit potential created by j's inclusion against the flat fixed cost of establishment. Hence, there are both motives for FDI and for concentrating production. These are assessed as a whole to determine optimal behavior $\mathscr{J}_i^*(\cdot;\boldsymbol{v})$. Taxes thus have both a level and dispersion effect on the firm's choices. If there were uniform taxes, they would still appear in (11), but identically for all plants j and sets $\mathscr{J}$. However, tax rate *variation* among countries distorts the firm's choice of which plant will serve market k. This consideration is internalized, so tax rates affect a location j's profit potential in the plant location problem. In particular, countries with comparatively lower taxes have higher profit potential and higher marginal value of inclusion to any set. All else equal, firms will therefore tend towards opening plants there, resulting in a distortion in the CDCP's solution as well.

4.3. The solution method in practice

Algorithm 4 can be applied to find the solution to the problem above for each origin i. Values for parameters are taken from the model's calibration, described in Section 4.5.1. Likewise, general equilibrium aggregates derive from the calibrated model's solution. Then, this section solves for the optimal behavior of German firms $\mathscr{J}_G^*(\cdot;\boldsymbol{v})$ in partial equilibrium.[7] Aggregates are set at their equilibrium values in the calibrated equilibrium. Numerical exercises performed will quantify the accuracy-speed tradeoff in Section 4.3.1 and the role of substitutability in Section 4.3.2.

4.3.1. Accuracy and speed

Sections 2.3 and 3.3 discussed the dual gains from the algorithm to speed and accuracy of solution when dealing with continuous type spaces. In particular, a discretization method imposes a tradeoff between the two that is avoided by the full algorithm. To concretely demonstrate how these two considerations interact, consider the production of an average firm from i. If core productivity z is distributed according to $G_i(\cdot)$, average production is aggregated as

$$A\sum_{j,k}\frac{X_k^T}{\left(P_k^T\right)^{1-\rho}}\int_z z\Theta_{ik}\left(\mathscr{J}_i^*(z)\right)^{\frac{\rho-1}{\theta}}\frac{p_{ijk}\left(\mathscr{J}_i^*(z)\right)}{1-\tau_j}\mathrm{d}G_i(z)\,. \tag{12}$$

How to evaluate this integral will depend on the strategy used to calculate $\mathscr{J}_i^*(\cdot)$. If the policy function is approximated by discretizing the range of productivity and finding the optimal set at each gridpoint, then it likely will not be accurate when interpolated between gridpoints. The approximation becomes more precise the denser the discretization, but then the single-point solution method must be evaluated at more productivity levels. On the other hand, if Algorithm 4 is applied to find $\mathscr{J}_i^*(\cdot)$ accurately on the whole range of z, the integral is greatly simplified.[8] Index the intervals returned by the full solution method by e, so that z_i^e is the

[7] Germany is chosen as a representative origin country, but the arguments made will hold for any origin country.

[8] Since the vector of aggregates $\boldsymbol{v}$ is held at their equilibrium values in the calibrated equilibrium, the $\boldsymbol{v}$ argument is omitted whenever unambiguous for notational brevity.

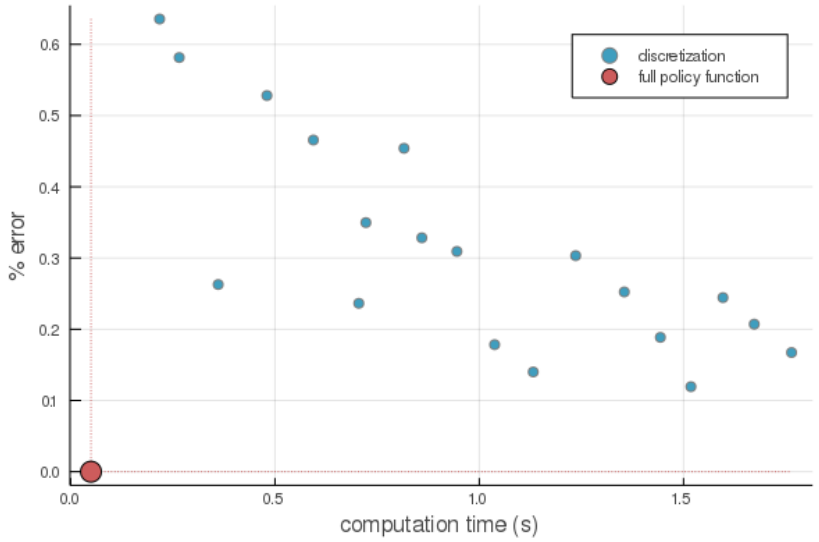


Figure 4: The tradeoff between speed and accuracy in approach based on discretization
Each blue point shows a different density grid. Sparser ones compuate quickly but with more aggregation error. Solving for the policy function with Algorithm 4 shown by the large red point is both quick and exact.

interval's left endpoint and $\mathscr{J}_i^{*e}$ is the associated optimal set. Observe that $\Theta_{ik}(\cdot)$ and $p_{ijk}(\cdot)$ only take the optimal set $\mathscr{J}^{*e}$ as an argument, so they are constant within each interval. Let these values be Θ_{ik}^e and p_{ijk}^e in each e. Then, (12) can be rewritten as

$$A\sum_{j,k}\frac{X_k^T}{\left(P_k^T\right)^{1-\rho}}\sum_e\left(\Theta_{ik}^e\right)^{\frac{\rho-1}{\theta}}\frac{p_{ijk}^e}{1-\tau_j}\int_{z_i^e}^{z_i^{e+1}} z\mathrm{d}G_i(z)\,.$$

When the core productivity distribution $G_i(\cdot)$ can be analytically integrated over z, there is not even a need for numerical integration. The production of an average firm (12) will thus be used as a measurement of accuracy in the numerical exercise below.

To illustrate how the discretization method compares to Algorithm 4 on speed and accuracy, consider the following experiment. Fixing aggregates and parameters, time Algorithm 4 while it recovers the exact policy function $\mathscr{J}_G^*(\cdot)$ and calculates the exact value of the average firms's sales (12). Next, discretize the productivity distribution with a grid of values. Solve the CDCP at each point and numerically approximate the integral in (12) by assigning the relevant mass to each gridpoint. The returned approximation's accuracy can be assessed as percent deviation from the exact value calculated previously. Then, trace out the discretization method's time to accuracy tradeoff by varying the density of gridpoints.

Figure 4 graphs the results from this experiment. To begin, the red point encapsulates the results from using Algorithm 4. It computes the full policy function in under 0.05 seconds, as shown on the x-axis.[9] As discussed above, it also exactly calculates the average firm's pro-

[9]Calculations were done in Julia on a 2018 Intel i7-8550U CPU.

duction, represented by zero error on the y-axis. The points in blue plot the results from ten different discretizations, ranging in density from 50 gridpoints to 500. As expected, there is a negative relationship between accuracy and speed. In particular, sparse grids take a relatively short time to compute, but return a worse approximation for (12). The opposite is true for denser grids. Overall, it takes considerably more computation time to approach the accuracy of the full solution method. The main reason has been alluded to already in Section 2.3. A perfect discretization would place gridpoints precisely according to the cutoffs returned by Algorithm 4. However, it is difficult to predict these values with no prior knowledge, so a dense discretization is necessary to cover as many intervals as possible.

Speed gains and error reduction represented in this exercise are also from computing the average firm's production in partial equilibrium, holding aggregates and parameters fixed. However, they will magnify once the firm block is embedded in a general equilibrium routine. There, optimal firm choices must be computed multiple times while searching for a fixed point on aggregates. Likewise, there is a further layer of amplification if the general equilibrium computation is itself included into a minimization routine for estimating parameters. Finally, the average firm's production was used as the single metric to assess aggregation so far. However, there will be additional aggregate variables requiring similar integration over the entire distribution of firm behavior in the full general equilibrium model. Then, deviations in this integral will appear in many general equilibrium conditions, magnifying the importance of integration accuracy when solving for and estimating numerical general equilibrium. Therefore, the advantages of the algorithm are especially valuable in quantitative general equilibrium models with heterogeneous agents facing CDCPs.

4.3.2. The role of substitutability

As a last note, this section discusses the practical efficiency of the iterative step defined by Algorithm 2. Recall that the iteration does not necessarily fix every binary decision. Undetermined binary decisions dealt with by the branching step Algorithm 3, which guesses a choice for one binary decision at a time. When the iterative algorithm cannot determine many binary choices, the full Algorithm reduces to the brute force method. It is therefore important to characterize the effectiveness of the iterative algorithm at fixing the binary choices.

As the interactions between them become stronger, the iterative step becomes less effective. To see why, consider the updating step. It checks a potential plant j's marginal value from inclusion in subsets and subsets of the optimal set to reach a decision. In the substitutes case (MS1), the subset represents the least amount of possible profit cannibalization among the firm's plants, while the superset features the most. Then, an element is undetermined when it has positive marginal value in the smallest set but negative marginal value in the largest set. In this case, the plant's fixed cost is covered by its contribution to expected variable profits when the firm does not have many other plants to choose from. However, once there are, the extra plant's marginal value falls below the fixed cost. This outcome is more likely if the plants are more substitutable. A similar logic holds in the complements case, with an element possibly adding very little on the margin if included into a small set. However, if complementarities are large, it could be very valuable if added into a large set. Thus, the effectiveness of iterative Algorithm 2 depends on the strength of substitutability

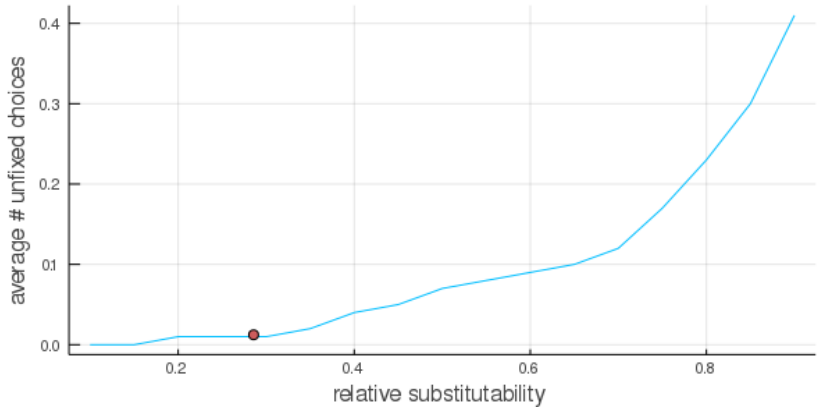


Figure 5: Average number of choices that are undetermined by the iterative step
More binary decisions are undetermined when substitutability among them is high. However, even when substitutability is high, the iterative step solves a majority of firm problems without requiring the branching step.

between plants in this context.

As discussed in the previous section, the strength of substitutability in this model is decreases in the exponent $(\rho-1)/\theta$ from (9). Thus, there is likely a negative relationship between the exponent value and the number of binary decisions determined by the iterative step. To quantify it, the iterative step is run holding aggregates constant for various exponent values. Then, the number of undetermined binary solutions at each productivity level z is calculated to assess the iterative step.

The blue line in Figure 5 graphs the result for exponent values between 0.1 and 0.9. Since the decisions are independent, therefore featuring no substitutability, when the exponent approaches 1, relative substitutabiliy is represented as $1-(\rho-1)/\theta$. As substitutability among plants decreases, the average number of unfixed decisions also decreases. A uniform distribution is chosen for this average since the metric is ultimately meant to speak to computational time, so the density of firms $g_i(\cdot)$ at each productivity does not play a role. The uniform distribution's left bound is taken to be lowest productivity at which a firm opens any plant. Likewise, the right bound is the productivity at which a firm operates a plant in every country. As depicted, the average number of unfixed decisions is small, falling below 0.5. This statement holds even with an exponent value of 0.1, when plants are relatively substitutable. Thus, all binary decisions are determined by the iterative algorithm for most productivity values. The logic of the algorithm eliminates a significant number of possible sets by using both monotonicity conditions.

To contextualize Figure 5, the calibration described in Section 4.5.1 sets $\rho=6$ and $\theta=7$, resulting in an exponent of around 0.71. This point is marked in Figure 5 with a red point, corresponding to an average of 0.012 decisions left unfixed by the iterative procedure. As above, all binary decisions are determined for a large majority of firm productivities, leaving a maximum of four undetermined in rare cases. The CDCP's structure is thus leveraged to determines the optimal set with no need to guess at decisions for large ranges of the produc-

tivity distribution.

4.4. General equilibrium

The plant location problem can now be embedded into a general equilibrium framework. Consumers are assumed to have Cobb-Douglas preferences over the traded basket and nontraded goods. The traded basket has weight $\beta < 1$ and thus receives this share of consumption expenditure. As described above, the differentiated tradable goods of each firm are aggregated into a basket with constant elasticity of substitution ρ.

Consumers are either workers or absentee capitalists. In addition to the consumption goods discussed above, workers value housing. A representative worker living and working in country k has utility

$$u_k \left(\left(c_k^N \right)^{1-\beta} \left(c_k^T \right)^{\beta} \right)^{\eta} h_k^{1-\eta}$$

where c_k^N is quantity of nontraded good, c_k^T is quantity of the traded basket, h_k is quantity of housing, and u_k is a country-specific utility shifter. The final term can be interpreted as measuring amenities in country k. Workers maximize utility subject to their budget constraint $w_k + R_k/L_k$. This expression captures the two sources of worker income: they earn wages and receive a lump sum rebate of total tax revenues raised in their country of residence. Indirect utility of a household in k is thus

$$\Psi \frac{u_k}{L_k} \left(\frac{w_k L_k + R_k}{\left(P_k^N \right)^{1-\beta} \left(P_k^T \right)^{\beta}} \right)^{\eta} H_k^{1-\eta} \equiv \overline{V} \tag{13}$$

where Ψ is a constant, P_k^N is the price of nontradables, and H_k is k's fixed stock of housing. The housing market is assumed to be perfectly competitive in deriving this expression. Workers are freely mobile across countries, so indirect utility must equalize across all k. For example, countries with high amenities, high wages, or low prices will be attractive places to live. Therefore, workers will move there, contributing to congestion in the housing market and decreasing tax revenue per worker, all else equal. Populations in each country k will be endogenously determined by these forces.

On the other hand, absentee capitalists in each country are assumed to be masses so they do not contribute to congestion. Consequently, they only purchase consumption goods and do not consume housing. Capitalists receive a share b_k of a global portfolio as income, which consists of total rents paid to housing as well as firm profits. Put together, their expenditure on consumption goods totals $b_k \sum_j \left(\Pi_j + (1-\eta) \left(w_j L_j + R_j \right) \right)$, where Π_j represents the profits made by firms born in j.

Gathering expenditure by both workers and capitalists together, total sales of traded goods in k can be summarized as

$$X_k^T = \beta\eta \left(w_k L_k + R_k \right) + \beta b_k \sum_j \left(\Pi_j + (1-\eta) \left(w_j L_j + R_j \right) \right). \tag{14}$$

The presence of capitalists who consume goods locally in k but have income from holdings in all countries allows for trade imbalances. Regardless, (14) ensures balance of payments.

A perfectly competitive nontraded sector in each country uses linear technology in labor with country-specific productivity α_k. Consequently, the price of the nontraded good is

$$P_k^N = w_k/\alpha_k\,. \tag{15}$$

Workers are freely mobile across the traded and nontraded sectors, resulting in a single equilibrium wage for each country. This sector is kept parsimonious to focus attention on the traded sector.

Finally, tradables firms must pay a fixed cost of entry f_i^e in labor. After entry, they are endowed with their plant in i and draw a core productivity from a origin-specific distribution, $G_i(\cdot)$. Since the home plant is free, all entering firms participate in the economy. Thus, given a continuous mass of entrants M_i, there will be a similar mass of active firms. The price index of the traded basket can now be written as

$$\left(P_k^T\right)^{1-\rho} = \rho A \sum_{i,j} M_i \int_z z\Theta_{ik}\left(\mathscr{J}_i^*(z)\right)^{\frac{\rho-1}{\theta}} \frac{p_{ijk}\left(\mathscr{J}_i^*(z)\right)}{1-\tau_j}\mathrm{d}G_i(z)\,. \tag{16}$$

This expression integrates over the distribution of core productivities present in each country of origin i, including the profit potential terms $\Theta_{ik}(\mathscr{J}_i^*(\cdot))$ and ex ante probabilities for their plants $p_{ijk}(\mathscr{J}_i^*(\cdot))$. It consequently incorporates the firm's location choice decision, precluding an analytical solution.

Following from this aggregation, sales shares are given by the CES structure of tradables demand. Specifically, the share of tradables sales in k paid to plants in j and owned by firms from i can be written as

$$\lambda_{ijk} = \frac{\rho A M_i}{\left(P_k^T\right)^{1-\rho}} \int_z z\Theta_{ik}\left(\mathscr{J}_i^*(z)\right)^{\frac{\rho-1}{\theta}} \frac{p_{ijk}\left(\mathscr{J}_i^*(z)\right)}{1-\tau_j}\mathrm{d}G_i(z)\,. \tag{17}$$

Because these shares apportion the sales in a market k to a origin-production pair (i,j), observe that $\sum_{i,j}\lambda_{ijk} = 1$ for all k. Then, it is easy to construct trade shares common to the literature as $\sum_i \lambda_{ijk}$. These represent the share of expenditure on tradables in k captured by goods produced in j, no matter the origin of the owner firm. Given these shares, the model departs from the standard gravity form of trade flows.

The λ_{ijk}s are the key intensive margin share of our model, on which many later equilibrium variables depend. For example, the total tax revenues raised in country j are aggregated as

$$R_j = \frac{\tau_j}{\rho}\sum_{i,k}\lambda_{ijk}X_k^T\,. \tag{18}$$

Note that taxes are levied on variable profits attributable to production carried out by plants operating within a country j's borders. Since demand is CES, variable profits are a constant share $1/\rho$ of sales. Total sales by j plants is in turn captured by the summation in (18). It adds over both k and i indices, to account for goods shipped to any destination market and produced by firms of any origin.

While the λ_{ijk} shares in (17) describe intensive margin multinational activity, the second key shares of the model summarize the extensive margin. In particular, aggregate over optimal firm choices to obtain the share of firms born in i with a production facility in j.

$$\mu_{ij} = \int_z \mathbb{1}\left[j \in \mathscr{J}_i^*(z)\right] \mathrm{d}G_i(z)\,. \tag{19}$$

Observe that, since a firm can open multiple plants, these shares do not add to one. In fact, $\sum_j \mu_{ij} \geq 1$ since this sum would double count the firms with more than one affiliate.

There is a free entry into the traded sector at a fixed labor cost f_i^e in the country of origin. Because firms do not observe their core productivity until after entry, they enter until expected profits equal this cost of entry. The free entry condition therefore set total profits to zero and endogenously determines the mass of firms M_i active.

$$\Pi_i = \frac{1}{\rho}\sum_{j,k}\left(1-\tau_j\right)\lambda_{ijk}X_k^T - \sum_j M_i\mu_{ij}w_j f_{ij} - M_i w_i f_i^e = 0 \tag{20}$$

Note that, since there is a convex continuum of firms in each country, the law of large numbers implies that firm-level expectations and probabilities are exactly realized over the measure of firms. For example, among i-firms with the set of affiliates $\mathscr{J}$, the sourcing probability $p_{ijk}(\mathscr{J})$ turns out to be the exact share of firms that serve market k from their j plant. Likewise, the expected firm profits before entry coincide with the average operating firm's profits in each country i.

Finally, labor markets must clear in each country as follows.

$$w_j L_j = \frac{\rho-1}{\rho}\sum_{i,k}\lambda_{ijk}X_k^T + \sum_i M_i\mu_{ij}w_j f_{ij} + M_j w_j f_j^e + \frac{1-\beta}{\beta}X_j^T \tag{21}$$

This condition equates total labor supply in j to labor demand. Labor is used in the tradables sector for variable production, plant setup fixed costs, entry fixed costs, and nontradable production. Similarly to total tax revenue in (18), CES demand fixes the total payments to variable factors of production as a constant share $(\rho-1)/\rho$ of sales. This expression also incorporates the fact that total payments to labor in the nontraded sector is equal to sales because of perfect competition.

General equilibrium can now be defined with the aggregate conditions above.

Equilibrium. *A collection of aggregates* $\boldsymbol{v} = \{\mathbf{L}, \mathbf{X^T}, \mathbf{P^N}, \mathbf{P^T}, \mathbf{R}, \mathbf{M}, \mathbf{w}, \boldsymbol{\lambda}, \boldsymbol{\mu}\}$ *and functions* $\{\mathscr{J}_i^*(\cdot;\boldsymbol{v})\}_i$ *constitute an equilibrium if:*

(i) Given $\{\mathscr{J}_i^*(\cdot;\boldsymbol{v})\}_i$*, aggregates* $\boldsymbol{v}$ *satisfy equations (13) - (21).*

(ii) Given $\boldsymbol{v}$*, functions* $\{\mathscr{J}_i^*(\cdot;\boldsymbol{v})\}_i$ *describe optimal choices for (9).*

The role of the integral appearing in the average firm's production (12) from Section 4.3.1 is now apparent. Ultimately, it integrates over each firm's prices in every market. Combined with the CES nature of demand, this integral also specifies the share of each destination market captured by firms producing in j or originating in i. As such, it appears both in the price index (16) and in the intensive margin market shares (17). In turn, these market shares are

needed to determine production, profits, tax revenue, and labor demand for variable production. Moreover, Section 4.3.1 considered only the average firm's production for simplicity, which sums the integrals over all plant locations $j \in J$. However, some equilibrium conditions like labor market clearing require each j term separately. Accuracy in each term is therefore necessary to ensure these conditions are satisfied in each country. Thus, it is important to calculate the integral over firm behavior accurately when computing general equilibrium. As Section 4.3.1 showed, the algorithm developed in 3 delivers this numerical precision without compromising on solution speed.

4.5. Calibration and counterfactuals

Having specified the structural model, this section outlines quantitative exercises. First, the model is calibrated and estimated in Section 4.5.1, by using the structure of the model in tandem with aggregate EU data. Next, the calibrated model is used to perform a counterfactual exercise evaluating the effects of a tax harmonization policy in the EU. This assessment is described in Section 4.5.2.

4.5.1. Calibration and estimation

The model is now calibrated in anticipation of the policy counterfactual. First, a functional form must be chosen for the distribution of core productivity $G_i(\cdot)$. In the quantitative exercises, $G_i(\cdot)$ is assumed to follow a Pareto distribution with shape ξ and minimum value $z_i^{\min}$. The tail index is constant across origin countries i, but the distribution's minimum value can vary. Then, countries with high $z_i^{\min}$ will have relatively more productive *firms*. Firm productivity directly affects the set of countries selected for production, and will indirectly shift the firm's plant productivity upwards in expectation.

The model contains several parameters and location fundamentals. Values for these are determined with three separate steps. Briefly, several are first calibrated to values from the literature and data, summarized in Table 1. Next, a small set are estimated using the simulated method of moments (SMM), using moments listed in Table 2. Finally, the selection in Table 3 are calculated conditional on the calibrated and estimated values, by inverting a set of equilibrium conditions. Implicitly, these strategies assume the state of the EU represented by the data is in equilibrium. A more detailed description of each of these steps follows.

To begin, some parameters are set to values in the literature or calculated directly in the data. Table 1 summarizes these. In particular, many values from Tintelnot [2016] are used since he estimates these using microdata on European firms. In addition, his model shares many elements with the one in this paper, the most important being that multinationals also solve a plant location CDCP. Other data values are set to EUROSTAT aggregates for 2014.

Next, a collection of model parameters are estimated with SMM. Table 2 summarizes these parameters and corresponding SMM strategies. First, a description of the moments is provided. In particular, the key set targeted is μ_{ij}s, the share of firms from i with an affiliate in j. These moments are available as part of EUROSTAT's Foreign Affiliates Statistics (FATS), which provides the number of establishments operating in each country, by country of ownership.

Parameter	Description	Calibration strategy
$\theta = 7$	Fréchet shape	Tintelnot [2016], Eaton and Kortum [2002]
$\rho = 6$	CES	Tintelnot [2016]
$\xi = 6.436$	Pareto shape	Tintelnot [2016]
$\mathbf{d}$	trade costs	Tintelnot [2016] and CEPII distance measure
$\boldsymbol{\gamma}$	arms-length costs	Tintelnot [2016] and CEPII distance measure
$\eta = 0.75$	C-D share, consumption	EUROSTAT: household consumption
$\beta = 0.7$	C-D share, tradables	EUROSTAT: household consumption
$\mathbf{X^T}$	tradables expenditure	EUROSTAT: manufacturing trade flows
$\mathbf{L}$	workers	EUROSTAT: labor force by country
$\mathbf{M}$	mass of firms	EUROSTAT: number of manufacturing firms
$\boldsymbol{\tau}$	tax rates	statutory corporate tax rates

Table 1: A summary of parameter calibration from literature and data.

Parameter	Description	Estimation strategy
$\mathbf{f}$	plant fixed cost	SMM, EUROSTAT (FATS) $\boldsymbol{\mu}$
$\mathbf{z}^{\min}$	Pareto scale	SMM, EUROSTAT (FATS) $\sum_j \mathbb{1}[j \neq i]\mu_{ij}$
$\mathbf{T}$	country productivity	SMM, EUROSTAT production shares

Table 2: A summary of parameter estimation using a simulated method of moments.

Thus, data equivalents of μ_{ij} are constructed as

$$\mu_{ij}^{\text{data}} \equiv \frac{\#\left(\text{plants in } j \text{ with } i \text{ ownership}\right)}{\#\left(\text{firms owned in } i\right)}$$

for all pairs $i \neq j$. These are matched with the model counterparts μ_{ij}, defined in (19) by integrating over plant location decisions. Because the dataset describes foreign affiliate activity, it does not contain information for domestically owned plants. A combined measure $\sum_j \mathbb{1}[i \neq j]\mu_{ij}^{\text{data}}$, which serves as an overall statistic for foreign activity, will also be matched. Additionally, a second set of moments describing the distribution of manufacturing production within the EU

$$\delta_j^{\text{data}} \equiv \frac{\text{manufacturing production in } j}{\text{manufacturing production in the EU}}$$

are constructed using EUROSTAT data on manufacturing production. This share describes intensive-margin production in each country carried out by both domestic and foreign affiliates. They are likewise matched to the model counterparts, $\delta_j \equiv \sum_{i,k} X_k^T \lambda_{ijk} / \sum_k X_k^T$.

The SMM's general intuition is as follows. Since the μ_{ij}^{data} shares describe extensive-margin moments, they are particularly informative for the fixed costs of plant establishment f_j. In general, when fixed costs to establishing a plant in country j increases, it will be optimal to do so for fewer firms. On the other hand, δ_j^{data} describes intensive margin production in a country so they mainly discipline T_j, which governs the country-level average of plant

Parameter	Description	Estimation strategy
$\mathbf{f}^e$	entry fixed cost	invert (20) with SMM estimates
$\mathbf{u}\boldsymbol{\alpha}^{\eta(\mathbf{1}-\beta)}\mathbf{H}^{-(1-\eta)}$	utility composite	invert (13) with SMM estimates
$\mathbf{b}$	global portfolio share	invert (14) with SMM estimates

Table 3: A summary of parameter estimation from inverting equilibrium conditions.

productivity. With higher T_j, production in a country j increases, all else equal. Finally, the combined measure $\sum_j \mathbb{1}[i \neq j]\mu_{ij}^{\text{data}}$ is primarily targeted by $z_i^{\min}$. Without this measure, the scale of $z_i^{\min}$ cannot be identified separately from the scale of f_j.

In practice, the simulated method of moments searches over composites of equilibrium aggregates and parameters $\{P_j^T, w_j/T_j, w_j f_j\}_j$. Combined with the parameters calibrated in Table 1, these vectors are sufficient to solve the continuum of firm problems (9) for $\{\mathscr{J}_i^*(\cdot)\}_i$. Since every plant is endowed with a plant in their country of origin, the lowest cutoff productivity returned by the algorithm is zero. Then, $\{z_i^{\min}\}_i$ is inferred using the model's structure on $G_i(\cdot)$ by matching $\sum_j \mathbb{1}[i \neq j]\mu_{ij}^{\text{data}}$ exactly. Next, the moments δ_j and μ_{ij} can be calculated in the model and matched to the data. Explicitly, the simulated method of moments minimizes the following mathematical program with equilibrium constraints (MPEC).

$$\begin{aligned}
&\min_{\{P_j^T, w_j/T_j, w_j f_j\}_j} \sum_j \left(\mu_{ij} - \mu_{ij}^{\text{data}}\right)^2 + \sum_j \left(\delta_j - \delta_j^{\text{data}}\right)^2 \\
\text{s.t.} \quad &\left(P_k^T\right)^{1-\rho} = \rho A \sum_{i,j} M_i \int_z z\Theta_{ik} \left(\mathscr{J}_i^*(\cdot)\right)^{\frac{\rho-1}{\theta}} \frac{p_{ijk}\left(\mathscr{J}_i^*(\cdot)\right)}{1-\tau_j} \mathrm{d}G_i(z) \\
&\sum_j \mathbb{1}[j \neq i]\mu_{ij}^{\text{data}} = \sum_j \mathbb{1}[j \neq i]\mu_{ij}
\end{aligned}$$

The equilibrium constraint imposed is the definition of the price index (16). Casting estimation routines of structural models as MPECs was proposed in Su and Judd [2012] as an alternative to the nested fixed point routine. Further, Dubé et al. [2012] test its performance, where they find the MPEC method is faster and more accurate. Note that the SMM routine returns estimates for the composites $w_j f_j$ and w_j/T_j. However, the parameters f_j and T_j are easily separated in what follows.

Finally, a set of location fundamentals are determined by inverting the model after the SMM routine. These are the fixed cost of firm entry f_j^e, a composite $u_j \alpha_j^{\eta(1-\beta)} H_j^{-(1-\eta)}$, and the capitalist claim on the global portfolio b_j. The composite term contains the utility shifter in each country u_j, labor productivity in the nontraded sector α_j, and the country's stock of housing H_j. Recall data for mass of firms, workforce in each country, and tradables expenditure are used as the values of model counterparts. Then, the previous fundamentals can be inferred from inverting the free entry condition, free mobility condition, and balance of payments. The individual elements in the composite term are not separately identified because they enter indirect utility (13) in an identical way. However, the composite term is sufficient to compute equilibrium outcomes of interest in the counterfactual, namely workforce L_j in each country.

As a last step, inverting the labor market clearing condition (21) solves for wages w_j. Then, the fundamentals $\{T_j, f_j\}_j$ can be separated using the composite estimates $\{w_j f_j, w_j / T_j\}$ from the SMM, completing the calibration and estimation procedure.

4.5.2. Counterfactual experiment

With the calibrated quantitative model in hand, a tax policy harmonization can now be evaluated. As discussed in Section 4.2, taxes on firm profits have both a level and dispersion effect. To determine the effect of tax rate dispersion, the counterfactual searches for an equalized tax rate τ generating the same amount of EU-wide tax revenue $\sum_j R_j$ as the calibrated model. The rate is chosen in this way to hold constant the tax rate level effect as much as possible.

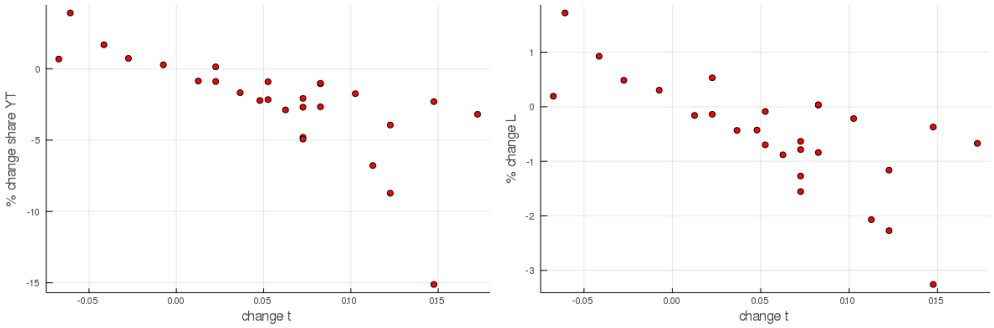
The counterfactual rate turns out to be around 27.26%, implying tax cuts for five countries. In order from largest to smallest cuts, these are Belgium, France, Italy, Germany, and Spain ("cutters"). The remaining countries see tax rate increases, ranging from 1.26 percentage points (Greece) to 17.26 percentage points (Bulgaria). When this counterfactual rate is implemented in the model, aggregate EU production in the tradables sector increases around 0.33%. However, this figure belies considerable redistributive effects across countries. Overall, industry relocates to the cutters, where there are 6.8% more firms, 0.8% more workers, and 0.2% – 1.9% higher real wages. As a mirror effect, there are 9% fewer firms in the other countries, along with 0.9% fewer workers, and 0.3% – 3.7% lower real wages.

It is easy to verify that these changes are due to industry relocating to cutters. Recall that δ_j describes the share of EU-wide tradables production occurring in country j. Figure 6a displays how this share changes country by country, plotted against the change in tax rates. There is a strong negative relationship, with cutters all gaining share at the expense of others. Workers similarly relocate to these formerly high-tax countries, as shown in Figure 6b. At the same time, Figure 6c shows that real wages (deflated by consumables price index) also increase in the cutters. With both quantity and price of labor rising in these, there must have been an increase in labor demand. Figure 6d shows the source. The mass of firms increases in all of the cutters and falls in the other countries, with some cutters seeing over 10% more firms in their borders. Thus, dramatically increased firm entry in cutters coupled with depressed firm entry in other countries accounts for these distributional patterns.

5. Conclusion

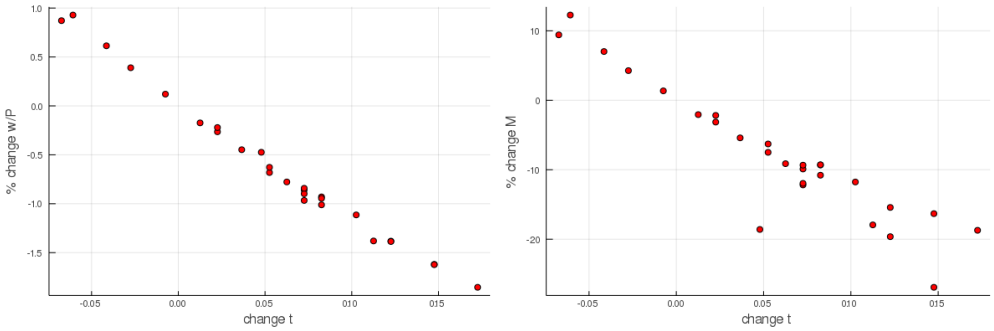
A major challenge in international trade and industrial organization is solving combinatorial discrete choice problems in settings with heterogeneous agents and a high dimensional state space. This paper has presented a unified strategy for solving CDCPs in such heterogeneous agent settings based on explicitly solving for the policy function in terms of an agent's type. This approach flexibly handles any distribution of agents across types, an arbitrary dimensional type space, and either substitutable or complementary discrete choices.

Moreover, it is both faster and more accurate than methods currently employed in the literature. Overall, these advantages render it especially suitable for computing or estimating quantitative general equilibrium models featuring heterogeneous agents facing CDCPs. As this algorithm explicitly solves for the agent's policy function, it is straightforward to embed



(a) change in production share

(b) change in workforce



(c) change in real wage

(d) change in firm mass

Figure 6: With harmonized rate, industry expands in cutters.

within a general equilibrium fixed point problem. The algorithm's computational speed also lends itself to inclusion within parameter estimation routines.

To explicitly solve for the policy function, the paper uses two structural assumptions on the combinatorial discrete choices and how the problem interacts with agent type. These properties have natural economic interpretations, so arise in many economic settings. The first requires the individual discrete choices to interact in a way where they are either substitutes to each other or complements. Respectively, submodularity or supermodularity among the decisions are sufficient. The second asserts that the value of any decision is monotonically related to agent type. Intuitively, this condition assumes a type of supermodularity between the choices and the agent type. Both conditions are satisfied by the discrete choice problems investigated across the literature.

As an illustration of this approach's advantages, the solution method is applied to a heterogeneous firm context where each selects a subset of countries in which to operate plants. Endogenously arising multinational firms therefore trade off a rich set of incentives when making FDI decisions, including a potential plant's expected production capability, market access, tax burden, and fixed cost of establishment. Given the algorithm's speed and precision, both computing numerical general equilibrium and estimating its parameters is feasible with a large set of available plant locations. Crucially, the general equilibrium implications of these heterogeneous firms making high dimensional problems is easily analyzed quantitatively. To illustrate a potential policy question benefiting from this type of analysis, the paper uses the estimated structural general equilibrium model to evaluate the aggregate effects of a EU-wide corporate tax equalization. Distributional effects are large, with some countries seeing significant increases in jobs and real wages, while a mirror effect occurs in others. These effects crucially occur in general equilibrium, governed by aggregate characterizations including worker free mobility and firm free entry.

The method is applicable to many questions in international trade and industrial organization beyond the plant location problem studied in this paper. For example, this method is naturally applicable to quantitatively analyzing the combinatorial decision of entry into markets or product production by multi-good firm. In Appendix C, the method is extended to allow for discrete choices with multiple possible values beyond 0-1 binary decisions. As in the binary case, the key insight is to use monotonicity on the choice set and type to explicitly solve for the agent's policy function. Consequently, it can also be used to analyze more general discrete choice problems of this nature.

References

Pol Antrás. *Global Production: Firms, Contracts, and Trade Structure*. Princeton University Press, 2016.

Pol Antrás and Stephen R. Yeaple. Chapter 2 - multinational firms and the structure of international trade. In Gita Gopinath, Elhanan Helpman, and Kenneth Rogoff, editors, *Handbook of International Economics*, volume 4 of *Handbook of International Economics*, pages 55–130. Elsevier, 2014. doi: 10.1016/B978-0-444-54314-1.00002-1.

Pol Antràs, Teresa C. Fort, and Felix Tintelnot. The margins of global sourcing: Theory and evidence from US firms. *American Economics Review*, 107(9):2514–64, September 2017. doi: 10.1257/aer.20141685.

Costas Arkolakis and Fabian Eckert. Combinatorial discrete choice. *Working paper*, January 2017. doi: 10.2139/ssrn.3455353.

Costas Arkolakis, Natalia Ramondo, Andrés Rodríguez-Clare, and Stephen Yeaple. Innovation and production in the global economy. *American Economic Review*, 108(8):2128–73, August 2018. doi: 10.1257/aer.20141743.

Andrew B. Bernard, J. Bradford Jensen, Stephen J. Redding, and Peter K. Schott. Global firms. *Journal of Economic Literature*, 56(2):565–619, June 2018. doi: 10.1257/jel.20160792.

Thiess Buettner and Martin Ruf. Tax incentives and the location of FDI: Evidence from a panel of german multinationals. *International Tax and Public Finance*, 14(2):151–64, April 2007. doi: 10.1007/s10797-006-8721-5.

Michael P. Devereux and Rachel Griffith. Evaluating tax policy for location decisions. *International Tax and Public Finance*, 10(2):107–26, March 2003. doi: 10.1023/A:1023364421914.

Jean-Pierre Dubé, Jeremy T. Fox, and Che-Lin Su. Improving the numerical performance of static and dynamic aggregate discrete choice random coefficients demand estimation. *Econometrica*, 80(5):2231–67, September 2012. doi: 10.3982/ECTA8585.

Jonathan Eaton and Samuel Kortum. Technology, geography, and trade. *Econometrica*, 70(5): 1741–1779, February 2002. doi: 10.1111/1468-0262.00352.

Pablo D Fajgelbaum, Eduardo Morales, Juan Carlos Suárez Serrato, and Owen Zidar. State taxes and spatial misallocation. *The Review of Economic Studies*, 86(1):333–376, September 2018. doi: 10.1093/restud/rdy050.

Elhanan Helpman, Marc J. Melitz, and Stephen R. Yeaple. Export versus FDI with heterogeneous firms. *American Economic Review*, 94(1):300–316, March 2004. doi: 10.1257/000282804322970814.

Panle Jia. What happens when Wal-Mart comes to town: an empirical analysis of the discount retailing industry. *Econometrica*, 76(6):1263–1316, October 2008. doi: 10.3982/ECTA6649.

Marc J. Melitz. The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica*, 71(6):1695–1725, 2003. doi: 10.1111/1468-0262.00467.

Marc J. Melitz and Stephen J. Redding. Chapter 1 - heterogeneous firms and trade. In Gita Gopinath, Elhanan Helpman, and Kenneth Rogoff, editors, *Handbook of International Economics*, volume 4 of *Handbook of International Economics*, pages 1–54. Elsevier, 2014. doi: 10.1016/B978-0-444-54314-1.00001-X.

Natalia Ramondo and Andrés Rodríguez-Clare. Trade, multinational production, and the gains from openness. *Journal of Political Economy*, 121(2):273–322, 2013. doi: 10.1086/670136.

Che-Lin Su and Kenneth L. Judd. Constrained optimization approaches to estimation of structural models. *Econometrica*, 80(5):2213–30, September 2012. doi: 10.3982/ECTA7925.

Felix Tintelnot. Global production with export platforms. *The Quarterly Journal of Economics*, 132(1):157–209, October 2016. doi: 10.1093/qje/qjw037.

A. Theoretical appendix

During this appendix, the vector of aggregates $\boldsymbol{v} \in \boldsymbol{\Upsilon}$ is assumed to be fixed at an arbitrary value. It is omitted as an argument for notational brevity.

A.1. Monotonicity in type

This section presents a more general sufficiency condition for monotonicity in type than the one in Section 3. First, let each characteristic be indexed by i, so a type vector can be represented as $\mathbf{z} = (z_1, ..., z_i, ..., z_I)$ where there are I total characteristics. Next, define the gradient

$$\frac{\partial D_j(\mathscr{J};\mathbf{z})}{\partial z_i} \equiv \nabla_{\mathbf{z}} D_j(\mathscr{J};\mathbf{z}) .$$

Then, the sufficiency condition can be stated as follows. For all $\mathscr{J}$, j, $\mathbf{z}_1$, and $\mathbf{z}_2$, the inner product $\nabla_{\mathbf{z}} D_j(\mathscr{J};\mathbf{z}_1) \cdot \nabla_{\mathbf{z}} D_j(\mathscr{J};\mathbf{z}_2) \geq 0$.

A.2. Proofs

A.2.1. Proposition 1

Suppose for every $\mathbf{z} \in \mathbf{Z}$, $L \subseteq \mathscr{J}^*(\mathbf{z}) \subseteq U$. We show that if $\mathbf{z} \in \mathbf{Z}^e$, $L^e \subseteq L \subseteq \mathscr{J}^*(\mathbf{z}) \subseteq U^e \subseteq U$ for any $\{\mathbf{Z}^e, L^e, U^e\}_e$ returned by Algorithm 1.

Proof. First, observe that the set of triples $\{\mathbf{Z}^e, L^e, U^e\}_e$ is derived from the functions $\hat{L}(\cdot)$ and $\hat{U}(\cdot)$, which are defined over $\mathbf{Z}$. The triples are a partitioning of $\mathbf{Z}$ such that $\hat{L}(\cdot)$ and $\hat{U}(\cdot)$ are constant over each cell. It is equivalent to show that $L \subseteq \hat{L}(\mathbf{z}) \subseteq \mathscr{J}^*(\mathbf{z}) \subseteq \hat{U}(\mathbf{z}) \subseteq U$ for all $\mathbf{z} \in \mathbf{Z}$.

Step (ii) from Algorithm 1 ensures that $L \subseteq \hat{L}(\mathbf{z})$ and $\hat{U}(\mathbf{z}) \subseteq U$ for every $\mathbf{z} \in \mathbf{Z}$. What remains is to show that $\hat{L}(\mathbf{z}) \subseteq \mathscr{J}^*(\mathbf{z})$ and $\mathscr{J}^*(\mathbf{z}) \subseteq \hat{U}(\cdot)$.

Fix a $\mathbf{z} \in \mathbf{Z}$ and start with $\hat{L}(\mathbf{z})$. Select an arbitrary element $j \in \hat{L}(\mathbf{z})$. If $j \in \mathscr{J}^*(\mathbf{z})$, then $\hat{L}(\mathbf{z}) \subseteq \mathscr{J}^*(\mathbf{z})$ since j was an arbitrary element. Step (ii) constructs $\hat{L}(\mathbf{z})$ as

$$\hat{L}(\mathbf{z}) = \begin{cases} L \cup \{j \in U \setminus L \mid \mathbf{z} \in \overline{\mathbf{Z}}_j(U)\} & \text{if (MS1)} \\ L \cup \{j \in U \setminus L \mid \mathbf{z} \in \underline{\mathbf{Z}}_j(L)\} & \text{if (MS2)} \end{cases} .$$

If $j \in L$, then $j \in L \subseteq J^*(\mathbf{z})$ by assumption on L. The only case remaining is if

$$j \in \begin{cases} \{j \in U \setminus L \mid \mathbf{z} \in \overline{\mathbf{Z}}_j(U)\} & \text{if (MS1)} \\ \{j \in U \setminus L \mid \mathbf{z} \notin \underline{\mathbf{Z}}_j(L)\} & \text{if (MS2)} \end{cases}.$$

By definition on $\overline{\mathbf{Z}}_j(U)$ and $\underline{\mathbf{Z}}_j(L)$, $D_j(U;\mathbf{z}) > 0$ and $D_j(L;\mathbf{z}) \geq 0$ respectively. Then,

$$\begin{aligned} 0 < D_j(U;\mathbf{z}) \leq D_j(\mathscr{J}^*(z);\mathbf{z}) \qquad & \text{when (MS1)} \\ 0 \leq D_j(L;\mathbf{z}) \leq D_j(\mathscr{J}^*(z);\mathbf{z}) \qquad & \text{when (MS2)} \end{aligned}$$

so it must be that $j \in \mathscr{J}^*(\mathbf{z})$.

A similar argument can be made to show $\mathscr{J}^*(\mathbf{z}) \subseteq \hat{U}(\mathbf{z})$ for all $\mathbf{z} \in \mathbf{Z}$. Fix $\mathbf{z} \in \mathbf{Z}$ and select an arbitrary element $j \in \mathscr{J}^*(\mathbf{z})$. Observe that it must be that $D_j(\mathscr{J}^*(\mathbf{z});\mathbf{z}) \geq 0$. Showing $j \in \hat{U}(\mathbf{z})$ is equivalent to showing $\mathscr{J}^*(\mathbf{z}) \subseteq \hat{U}(\mathbf{z})$. By the definition of $\hat{U}(\mathbf{z})$ in step (ii),

$$\hat{U}(\mathbf{z}) = \begin{cases} U \setminus \{j \in U \setminus L \mid \mathbf{z} \notin \overline{\mathbf{Z}}_j(L)\} & \text{if (MS1)} \\ U \setminus \{j \in U \setminus L \mid \mathbf{z} \in \underline{\mathbf{Z}}_j(U)\} & \text{if (MS2)} \end{cases}$$

so it is sufficient to show that

$$j \notin \begin{cases} \{j \in U \setminus L \mid \mathbf{z} \notin \overline{\mathbf{Z}}_j(L)\} & \text{if (MS1)} \\ \{j \in U \setminus L \mid \mathbf{z} \in \underline{\mathbf{Z}}_j(U)\} & \text{if (MS2)} \end{cases}.$$

For a contradiction, suppose that j is in these sets. By the definition of $\overline{\mathbf{Z}}_j(L)$ and $\underline{\mathbf{Z}}_j(U)$, $D_j(L;\mathbf{z}) > 0$ and $D_j(U;\mathbf{z}) < 0$. If $j \in \{U \setminus L \mid \mathbf{z} \notin \overline{\mathbf{Z}}_j(L)\}$ so $D_j(L;\mathbf{z}) \leq 0$. By (MS1), $D_j(\mathscr{J}^*(\mathbf{z});\mathbf{z}) \leq D_j(L;\mathbf{z}) \leq 0$, a contradiction to the initial assertion that $j \in \mathscr{J}^*(\mathbf{z})$. Likewise, if $j \in \{j \in U \setminus L \mid \mathbf{z} \in \underline{\mathbf{Z}}_j(U)\}$, then $D_j(U;\mathbf{z}) < 0$. However, by (MS2), $D_j(\mathscr{J}^*(\mathbf{z});\mathbf{z}) \leq D_j(\mathscr{J}^*(\mathbf{z});\mathbf{z}) < 0$, a contradiction to the assertion that $j \in \mathscr{J}^*(\mathbf{z})$. □

A.2.2. Proposition 2

Consider a triple $(\mathbf{Z}, L, U)$ with $L \subseteq \mathscr{J}^*(\mathbf{z}) \subseteq U$ for all $\mathbf{z} \in \mathbf{Z}$, to which Algorithm 2 is applied.

Let each iteration of step (iii) be indexed by n. Further, let the $C_n = \{(\mathbf{Z}_n^c, L_n^c, U_n^c)\}_c$ be the collection of triples still being iterated on while $R_n = \{(\mathbf{Z}_n^r, L_n^r, U_n^r)\}_r$ is the collection of triples for which iteration has stopped.

Lemma. *For any iteration n, $\mathbb{P}_n \equiv \{\mathbf{Z}_n^c\}_c \cup \{\mathbf{Z}_n^r\}_r$ partitions $\mathbf{Z}$. That is, for each $\mathbf{z} \in \mathbf{Z}$, there exists at most one c with $\mathbf{z} \in \mathbf{Z}_n^c$. If no such c exists, there is a unique r with $\mathbf{z} \in \mathbf{Z}_n^r$.*

Proof. We use induction on n. When $n = 0$, $C_0 = \{(\mathbf{Z}, L, U)\}$ and $R_0 = \{\}$, so $\mathbb{P}_0 = \{\mathbf{Z}\}$. Then, $\mathbb{P}_0$ partitions $\mathbf{Z}$ trivially.

Now suppose $\mathbb{P}_n$ partitions $\mathbf{Z}$. Notice that step (iii) does not affect any triples within R_n. Consider an arbitrary triple $(\mathbf{Z}_n^c, L_n^c, U_n^c) \in C_n$. Step (iii) will apply the updating Algorithm 1 to this triple. If it returns $\{(\mathbf{Z}_n^c, L_n^c, U_n^c)\}$, then substep (iii)c. places this triple into R_{n+1}. Otherwise, it returns $\{\mathbf{Z}_n^{ce}, L_n^{ce}, U_n^{ce}\}_e$, a further partitioning of $\mathbf{Z}_n^c$. Each of these triples is placed into R_{n+1}. Then, each $\mathbf{Z}_n^c$ is (weakly) subpartitioned by step (iii). As a result, $\mathbb{P}_{n+1}$ is still a partitioning of $\mathbf{Z}$. □

By the lemma, for every $\mathbf{z} \in \mathbf{Z}$, there is a unique cell in $\mathbb{P}_n$ to which $\mathbf{z}$ belongs. At each iteration step, the functions $L_n(\cdot)$ and $U_n(\cdot)$ can be defined as follows

$$L_n(\mathbf{z}) = \begin{cases} L_n^c, & \mathbf{z} \in \mathbf{Z}_n^c \\ L_n^r, & \mathbf{z} \in \mathbf{Z}_n^r \end{cases}, \qquad U_n(\mathbf{z}) = \begin{cases} U_n^c, & \mathbf{z} \in \mathbf{Z}_n^c \\ U_n^r, & \mathbf{z} \in \mathbf{Z}_n^r \end{cases}.$$

Another way of interpreting step (ii) is therefore that it updates $L_n(\cdot)$ and $U_n(\cdot)$.

We now prove that the pair $L_n(\cdot)$ and $U_n(\cdot)$ can only be updated a maximum of $|J|$ times.

Proof. Fix $\mathbf{z} \in \mathbf{Z}$ and consider the sequence of $L_n(\mathbf{z})$ and $U_n(\mathbf{z})$ functions for each iteration n. Suppose $L_n(\mathbf{z}) = L_{n+1}(\mathbf{z})$ and likewise $U_n(\mathbf{z}) = U_{n+1}(\mathbf{z})$. Then, the $(n+1)$th iteration of step (iii) has not determined any of the choices that were left undetermined from iteration n. By step (iii)c., the algorithm will stop iteration for $L_n(\mathbf{z})$ and $U_n(\mathbf{z})$.

For continued iteration, it must be that one undetermined decision left from the last repetition is determined. That is, it must be that either $L_n(\mathbf{z}) \subset L_{n+1}(\mathbf{z})$ or $U_{n+1}(\mathbf{z}) \subset U_n(\mathbf{z})$. The algorithm can be initialized with a maximum $|J|$ decisions undetermined. The maximum number of continued iterations on $L_n(\mathbf{z})$ and $U_n(\mathbf{z})$ is therefore $|J|$, if each iteration determines only one decision. □

B. Variable fixed costs

Consider a version of the firm decision presented in Section 4, where the fixed costs of multinational production are idiosyncratic. That is, modify the firm problem so that it maximizes

$$\pi_i(\mathscr{J}; z(\omega), \mathbf{f}(\omega)) = Az(\omega) \sum_k \frac{X_k^T}{\left(P_k^T\right)^{1-\rho}} \Theta_{ik}(\mathscr{J})^{\frac{\rho-1}{\theta}} - \sum_j \mathbb{1}[j \in \mathscr{J}]\, w_j f_j(\omega),$$

where firms are now characterized by core productivity $z(\omega)$ as well as a vector of fixed costs $\{f_j(\omega)\}_j$. The firm problem presented in the quantitative models of Tintelnot [2016] and Antràs et al. [2017] have a similar structure. In this case, it is equivalent to maximize

$$\pi_i(\mathscr{J}; z(\omega), \mathbf{f}(\omega)) = A \sum_k \frac{X_k^T}{\left(P_k^T\right)^{1-\rho}} \Theta_{ik}(\mathscr{J})^{\frac{\rho-1}{\theta}} - \sum_j \mathbb{1}[j \in \mathscr{J}]\, w_j \frac{f_j(\omega)}{z(\omega)},$$

so the firm type can be simplified to a $|J|$-length vector of $f_j(\omega)/z(\omega)$ values.

The algorithm can be applied in this context. First, consider the marginal profit function

$$D_{ij}(\mathscr{J}; \mathbf{f}/z) = A \sum_k \frac{X_k^T}{\left(P_k^T\right)^{1-\rho}} \left[\Theta_{ik}\left(\mathscr{J} \cup \{j\}\right)^{\frac{\rho-1}{\theta}} - \Theta_{ik}\left(\mathscr{J} \setminus \{j\}\right)^{\frac{\rho-1}{\theta}} \right] - w_j \frac{f_j}{z}.$$

Observe that once again this problem satisfies monotonicity in set with the substitutes case (MS1). In order to satisfy monotonicity in type, the firm type can be redefined as the vector of $\nu_j(\omega) = -f_j(\omega)/z(\omega)$ values. These are interpreted as (negative) relative fixed costs. They

are relative since they are adjusted by core productivity. The maximization problem is now defined on

$$\pi_i(\mathscr{J};\boldsymbol{\nu}(\omega)) = A\sum_k \frac{X_k^T}{\left(P_k^T\right)^{1-\rho}}\Theta_{ik}(\mathscr{J})^{\frac{\rho-1}{\theta}} + \sum_j \mathbb{1}[j\in\mathscr{J}]\, w_j\nu_j(\omega)$$

with marginal value function

$$D_{ij}(\mathscr{J};\mathbf{f}/z) = A\sum_k \frac{X_k^T}{\left(P_k^T\right)^{1-\rho}}\left[\Theta_{ik}\left(\mathscr{J}\cup\{j\}\right)^{\frac{\rho-1}{\theta}} - \Theta_{ik}\left(\mathscr{J}\setminus\{j\}\right)^{\frac{\rho-1}{\theta}}\right] + w_j\nu_j\,.$$

The general specification of the algorithm requires the space $\{\nu \mid D_{ij}(\mathscr{J};\nu) = 0\}$ to divide $\mathbb{R}^{|J|}$ into two subspaces, where $D_{ij}(\mathscr{J};\cdot) < 0$ in one and $D_{ij}(\mathscr{J};\cdot) > 0$ in the other. This functional form provides a simple way to find the indifferent types, since $D_{ij}(\cdot;\cdot)$ is linear in ν_j only. For any $\mathscr{J}$ and j, given this linear structure, $D_{ij}(\mathscr{J};\nu_j^*) = 0$ if

$$\nu_j^* = -\frac{A}{w_j}\sum_k \frac{X_k^T}{\left(P_k^T\right)^{1-\rho}}\left[\Theta_{ik}\left(\mathscr{J}\cup\{j\}\right)^{\frac{\rho-1}{\theta}} - \Theta_{ik}\left(\mathscr{J}\setminus\{j\}\right)^{\frac{\rho-1}{\theta}}\right].$$

The space $\{\nu \mid D_{ij}(\mathscr{J};\nu) = 0\}$ comprises simply of all vectors $\boldsymbol{\nu}$ with $\nu_j = \nu_j^*$. All firms with $\nu_j > \nu_j^*$ will derive positive marginal value from j's inclusion in $\mathscr{J}$, while the reverse is true for firms with $\nu_j < \nu_j^*$.

As an example, consider again a two country model with Germany and Romania. In this case, the type space is two-dimensional over the types (ν_G, ν_R). Note that values for both νs will be negative, since they were defined as $\nu_j = -f_j/z$. Working through a similar example as presented in Section 2, consider the conclusions that be drawn after observing that $L \equiv \{\} \subseteq \mathscr{J}_i^*(\boldsymbol{\nu}) \subseteq \{G,R\} \equiv U$ for all $\boldsymbol{\nu}$. In what follows, a single origin i is considered, so the i subscript is dropped for notational brevity.

Consider the marginal value of adding plants in Germany conditional on there being no plants in Romania. The marginal value of doing so is

$$D_G(\{\};\boldsymbol{\nu}) = A\sum_k \frac{X_k^T}{\left(P_k^T\right)^{1-\rho}}\Theta_k\left(\{G\}\right)^{\frac{\rho-1}{\theta}} + w_G\nu_G\,,$$

leading to a cutoff ν_G^L. Firms with ν_G below this cutoff receive negative marginal value from adding Germany to the subset, implying by substitutability among plants that they should not open plants in Germany. Observe in general that $D_G(L;\cdot)$ may include both ν_G and ν_R types. However, with this (relative) fixed cost specification, the only term that enters is ν_G. Since the fixed cost of opening plants in Romania ν_R is irrelevant for the marginal value of opening plants in Germany, it does not appear in the marginal value function for Germany plants. Similarly, the marginal value of Germany's inclusion in the full set is calculated as

$$D_G(\{G,R\};\boldsymbol{\nu}) = A\sum_k \frac{X_k^T}{\left(P_k^T\right)^{1-\rho}}\left[\Theta_k\left(\{G,R\}\right)^{\frac{\rho-1}{\theta}} - \Theta_k\left(\{R\}\right)^{\frac{\rho-1}{\theta}}\right] + w_G\nu_G\,,$$

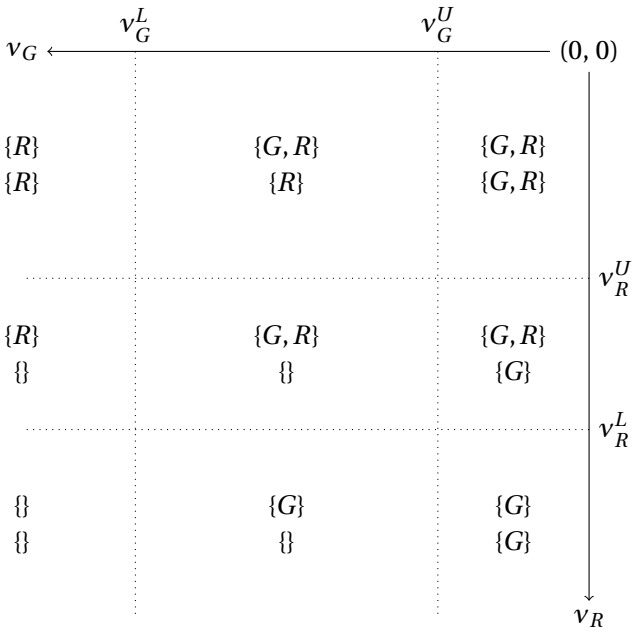


Figure 7: An example of subsets, supersets, and resulting subintervals after one interation. *Supersets are printed above subsets for each region of the type space. Firms above ν_G^U should open German plants while those below ν_G^L should not. Similarly, the logic concludes that firms above ν_R^U should open plants in Romania while those below ν_R^L should not.*

leading to a cutoff ν_G^U above which firms will open German plants. Similarly to before, $\nu_G^L \leq \nu_G^U$ by substitutability.

The key observation is that, while the general formulation of the CDCP given in Section 3 allows for $D_j(\mathscr{J};\cdot)$ to be a function of the full type vector $\boldsymbol{\nu}$, the fixed cost case simplifies to $D_j(\mathscr{J};\cdot)$ being a function only on ν_j. The types indifferent between adding j to $\mathscr{J}$ are therefore represented by a single value for j's relative fixed cost, independently of the fixed cost of other locations $\nu_{j'}$.

Similar cutoff relative fixed costs can be found for Romania, so there will once again be four cutoffs from this procedure $\{\nu_G^L, \nu_G^U, \nu_R^L, \nu_R^U\}$. A possible outcome of the updating step with this fixed cost specification is represented in Figure 7. To summarize the conclusions drawn during this updating procedure, firms with negative relative fixed cost above ν_G^U should open plants in Germany, regardless of their ν_R values. Likewise, firms below ν_G^L should not. Similarly, firms with negative relative fixed cost above ν_R^U have very low productivity-adjusted fixed costs to opening Romania plants. They will optimally open them, regardless of the fixed costs of opening German plants. Those below ν_R^L conversely have very high fixed costs to opening plants in Romania and should not open them.

The regions over which $L(\cdot)$ and $U(\cdot)$ are constant are therefore rectangles, bounded by the appropriate cutoffs on ν_G and ν_R. When there are $|J|$ countries, these regions will be $|J|$-dimensional cubes, each size bounded by an appropriate relative fixed cost cutoff. With the fixed cost CDCP, finding indifferent types is relatively straightforward, as is partitioning the

type space according to $L(\cdot)$ and $U(\cdot)$.

C. Multiple discrete choices

Consider a CDCP where each discrete choice has multiple discrete options. For example, suppose a firm decides whether or not to have zero, one, or two plants in Germany. Likewise, it must decide whether to have zero, one, or two plants in Romania. Then, the full set can be represented as the multi-set $\{G, G, R, R\}$ containing two copies of G and R. The firm then optimizes over all possible subsets of this multi-set. For example, the decision to open two plants in Germany and one in Romania can be represented as choosing the subset $\{G, G, R\}$. The method presented in the paper can then be applied to this problem, conditional on monotonicity in set and type.