Applied Microeconometrics Notes

Vasco Villas-Boas

May 10, 2025

CONTENTS

1 THE RANDOM UTILITY MODEL

Consider the discrete choice problem for an agent. An agent has the choice between $Y+1$ options where I denote $[Y] := \{1, 2, ..., Y\}$ to be the main (inside) Y options. They also have a default (outside) option $y = 0$. I denote $[0:Y] := \{0\} \cup [Y]$.

With each choice is associated a common (across agents) utility value U_y. This is often referred to as the *systematic utility* of the choice y. Further, each agent has an additional individual specific and idiosyncratic component for a choice y, ϵ_{iy}. I will use $\epsilon_i := [\epsilon_{i0}, \epsilon_{i1}, ..., \epsilon_{iY}]$ as a random vector that is drawn from a probability distribution $\mathcal{P}$ with a continuous cumulative distribution function. I normalize $U_0 = 0$, which is without loss of generality since replacing U_y with $U_y - U_0$ is WLOG. Each agent solves:

$$y(\epsilon) := \arg\max_{y \in [0:Y]} U_y + \epsilon_y = \arg\max_{y \in [Y]} \{U_y + \epsilon_y, \epsilon_0\} \tag{1}$$

where $u(\epsilon) := U_{y(\epsilon)} + e_{y(\epsilon)}$ denotes the indirect utility of the agent.

The total welfare in the economy is given by:

$$G(U) := \mathbb{E}_{\mathcal{P}}[\max_{y \in [0:Y]} U_y + \epsilon_y] \tag{2}$$

The interpretation is that this is the expected utility of an agent who's drawn from the population with distribution $\mathcal{P}$.

1.1 Theorem [Daly, Zachary, Williams]

Given U (the vector of systematic utilities) and $\mathcal{P}$, an interesting task is to predict the market shares. I denote $\pi_Y(U) := \Pr(U_y + e_y \geq \max_{z \in [0:Y]} U_z + \epsilon_z)$ as the market share of option y. Note that since the sum of market shares must equal one, we can deduce the market share of the outside option from the market shares of the other options: $\pi_y(0) = 1 - \sum_{y \in [Y]} \pi_{iy}$.

Courtesy Daly, Zachary, and Williams, we know that $\pi_y(U) = \frac{\partial G(U)}{\partial U_y}$.

Proof [Sketch].

Assume for some y that U_y is increased by a small quantity δ. The welfare effect on the group of people who were choosing y is $\pi_y(U) \times \delta$. The welfare effect on people who were not choosing y and are still not choosing y is 0. Finally, the welfare effect on people who were not choosing y but switch to choose y is on the order of $\delta \times \delta$, for δ small enough given our assumption that the CDF of $\mathcal{P}$ is continuous. Rearanging terms, we see the desired claim that $\pi_y(U) = \frac{\partial G(U)}{\partial U_y}$.

1.2 Example [Logit Model]

Let's assume that each ϵ_y is iid drawn from a Gumbel($\mu = 0, \beta = 1$) distribution (ie., each ϵ_y is drawn from a distribution with CDF $F(z) = \exp(-\exp(-z))$). Then, we have the following result.

1.3 Lemma

We have that $\max_{y \in [0:Y]} U_y + \epsilon_y =_D \log(\sum_{y \in [0:Y]} \exp(U_y)) + \epsilon$ where ϵ has the Gumbel distribution.

Proof.

Denote $Z := \max_{y \in [0:Y]} U_y + \epsilon_y$. We wish to show that the CDF of Z is the same as the CDF of $\log(\sum_{y \in [0:Y]} \exp(U_y)) + \epsilon$. Now,

$$\begin{aligned}
F_Z(z) &= \Pr(Z \leq z) \\
&= \Pr(\max_{y \in [0:Y]} U_y + \epsilon_y \leq z) \\
&= \Pr(U_y + \epsilon_y \leq z, \forall y \in [0:Y]) \\
&= \Pi_{y \in [0:Y]} \Pr(\epsilon_y \leq z - U_y) \\
&= \Pi_{y \in [0:Y]} \exp(-\exp(-z + U_y)) \\
&= \exp(-\sum_{y \in [0:Y]} \exp(-z)\exp(U_y)) \\
&= \exp(-\exp(-z + \log(\sum_{y \in [0:Y]} \exp(U_y))))
\end{aligned}$$

which is indeed the CDF of $\log(\sum_{y \in [0:Y]} \exp(U_y)) + \epsilon$ where ϵ has the Gumbel distribution.

1.3.1 Consequence of the Lemma

From the lemma, we have an expression for the welfare function:

$$G(U) := \mathbb{E}_{\mathcal{P}}[\max_{y \in [0:Y]} U_y + \epsilon_y] = \log(\sum_{y \in [0:Y]} \exp(U_y)) + \gamma \tag{3}$$

where γ is Euler's constant. I get this result by using the fact that ϵ which has distribution given by Gumbel$(\mu = 0, \beta = 1)$ has mean γ.

From here, we can deduce the market share map:

$$\pi_y(U) = \frac{\exp(U_y)}{1 + \sum_{y \in [Y]} \exp(U_y)} \tag{4}$$

where I use the fact that $U_0 = 0$ by construction.

1.4 Discrete Choice Inversion

An important task in applied discrete choice theorem is to take market share maps and deduce the systematic utilities given the distribution of idiosyncratic utility modifiers $\mathcal{P}$. More specifically, given π, we aim to find U such that $\pi_y(U) = \pi_y$.

1.4.1 Example [Logit Model]

Using our result from equation 4, we look for U such that $\frac{\exp(U_y)}{1+\sum_{y \in [Y]} \exp(U_y)} = \pi_y$. Rearranging, we have that $U_y - \log(1 + \sum_{z \in [Y]} \exp(U_z)) = \log(\pi_y)$.

Next, taking the expression for $y = 0$, and using the facts that $U_0 = 0$ and $\pi_0 = 1 - \sum_{y \in [Y]} \pi_y$, I get that $-\log(1 + \sum_{y \in [Y]} = \log(\pi_0)$. That implies $U_y = \log(\pi_y) - \log(\pi_0) = \log(\pi_y/\pi_0)$.

1.4.2 General Case

Recall that in the discrete choice inversion problem, we're looking for U such that $\pi_y(U) = \pi_y \implies \pi_y = \frac{\partial G(U)}{\partial U_y} = \pi_y$. As a remark, note that in equation 2, we see that $G(U)$ is the expectation, over the distribution $\mathcal{P}$ of individual shocks ϵ, of the maximum of linear functions and therefore is convex. Thus, we can view U as a solution to the convex optimization problem:

$$\min_{U \in \mathbb{R}^Y} G(U) - \sum_{y \in [Y]} \pi_y U_y = \max_{U \in \mathbb{R}^Y} \sum_{y \in [Y]} \pi_y U_y - G(U) \tag{5}$$

1.4.3 Simulating the Market Shares in Practice

Assume that for I agents, we draw individual shocks ϵ_{iy} for $(i, y) \in [I \times Y]$ from a Gumbel$(\mu = 0, \beta = 1)$ distribution.

We can simulate the market shares and the total welfare as:

$$\tilde{\pi}_y(U) = \frac{1}{|I|} \sum_{i \in [I]} \mathbb{1}_{\{U_y + \epsilon_{iy} \geq U_z + \epsilon_{iz} \forall z \neq y\}}$$
$$\tilde{G}(U) = \frac{1}{|I|} \sum_{i \in [I]} \max_{y \in [0:Y]} U_y + \epsilon_{iy}$$

Thus, the demand inversion problem can be written as:

$$\begin{aligned}\max_{U \in \mathbb{R}^Y} \sum_{y \in [Y]} \tilde{\pi}_y U_y - \tilde{G}(U) &= \max_{U \in \mathbb{R}^Y} \sum_{y \in [Y]} \tilde{\pi}_y U_y - \frac{1}{|I|} \sum_{i \in [I]} \max_{y \in [0:Y]} U_y + \epsilon_{iy} \\ &= \max_{U \in \mathbb{R}^Y, u_i \in \mathbb{R}^I} \sum_{y \in [Y]} \pi U_y - \frac{1}{|I|} \sum_{i \in [I]} u_i \\ &\quad \text{s.t. } u_i \geq U_y + \epsilon_{iy}\end{aligned}$$

This is a linear programming problem that can be solved with efficient solvers such as Gurobi in Python.

1.5 Generalized Entropy of Choice

The market share inversion problem takes a vector of market shares π and looks for systematic utilities U such that $\nabla G(U) = \pi$. We have found a way to solve this problem using a linear programming algorithm. We can define the Legendre-Fenchel Transform of G, G^*, as the value of that linear programming problem at its optimum:

$$G^*(\pi) = \max_{U \in \mathbb{R}^Y} \sum_{y \in [Y]} \pi_y U_y - G(U) \tag{6}$$

Observe that $\frac{\partial G^*(U)}{\partial \pi_y} = U_y$ by the envelope theorem, where U is the optimal search value for this problem. Let's call $G^*(\pi)$ the *generalized entropy of choice* associated with the discrete choice problem.

1.5.1 Example [Logit Model]

In the logit case (with a default), we have by equation (3) that $G(U) = \log(1 + \sum_{y \in [Y]} \exp(U_y)) + \gamma$.

Let's compute $G^*(\pi) = \max_{U \in \mathbb{R}^Y} \sum_{y \in [Y]} \pi_y U_y - \log(1 + \sum_{y \in [Y]} \exp(U_y)) - \gamma$. The FOC with respect to U_y implies: $\pi_y = \frac{\exp(U_y)}{1 + \sum_{z \in [Y]} \exp(U_z)}$. Taking logs on both sides, then multiplying both sides by π_y and finally summing over the options, I get:

$$\sum_{y \in [0:Y]} \pi_y \log(\pi_y) = \sum_{y \in [Y]} \{\pi_y U_y - \pi_y \log[1 + \sum_{z \in [Y]} \exp(U_z)]\}$$

where I cleverly use the fact that $U_0 = 0$. As a result, $G^*(\pi) = \sum_{y \in [0:Y]} \pi_y \log(\pi_y)$ and hence the naming of $G^*(\pi)$ as the (negative) *generalized entropy of choice*.

1.6 Expressing G as a function of G^*

From convex analysis, we know that a convex function is the Legendre-Fenchel transform of its own Legendre-Fenchel transform. That means:

$$\begin{aligned} G(U) &= \max_{\pi \in \mathbb{R}^Y} \sum_{y \in [Y]} \pi_y U_y - G^*(\pi) \\ &= \max_{\pi \in \mathbb{R}^Y} \sum_{y \in [Y]} \pi_y U_y - G^*(\pi) \\ &\quad \text{s.t. } \pi_y > 0, \sum_{y \in [Y]} \pi_y < 1 \end{aligned} \tag{7}$$

In this context, the interpretation is that the social welfare of individuals with random utility consists of maximizing an objective function which is the probability weighted sum of the individual welfares $\sum_{y \in [Y]} \pi_y U_y$ plus a perturbation term equal to $-G^*(\pi)$, which is the generalized entropy of choice. The latter term can be seen as the effect of the random utility terms on the choice problem.

We build on this intuition. Recall the definition of $y(\epsilon)$ from equation 1. Then, for the definition of $G(U)$ in equation 2, I have that:

$$\begin{aligned} G(U) &= \mathbb{E}_{\mathcal{P}}[\max_{y \in [0:Y]} U_y + \epsilon_y] \\ &= \mathbb{E}_{\mathcal{P}}[U_{y(\epsilon)} + \epsilon_{y(\epsilon)}] \\ &= \mathbb{E}_{\mathcal{P}}[U_{y(\epsilon)}] + \mathbb{E}_{\mathcal{P}}[\epsilon_{y(\epsilon)}] \\ &= \sum_{y \in [Y]} \pi_y U_y + \mathbb{E}_{\mathcal{P}}[\epsilon_{y(\epsilon)}] \end{aligned}$$

If we take π, to be that which solves the Legendre-Fenchel Transform problem in equation 7, then we have that $G(U) = \sum_{y \in [Y]} \pi_y U_y - G^*(\pi)$. Hence, we have that $G^*(\pi) = -\mathbb{E}[\epsilon_{y(\epsilon)}]$. As a remark, note that we often choose to assume that $\mathbb{E}[\epsilon_{iy}] = 0 \; \forall y \in [0 : Y]$. Then, why is it that $G^*(\pi) \neq 0$? That is because for the y that is picked, as a function of ϵ, we would tend to expect the mean of the ϵ_y associated with the chosen y to be positive.

1.6.1 Lemma

Let's in fact show that G and G^* are Legendre-Fenchel transforms of each other.

Proof.

By definition, we have $G^*(\pi) = \max_{U \in \mathbb{R}^Y} \sum_{y \in [Y]} \pi_y U_y - G(U)$. As a result, we have that $\forall U, \pi$, $G^*(\pi) \geq \sum_{y \in [Y]} \pi_y U_y - G(U) \implies G(U) \geq \sum_{y \in [Y]} \pi_y U_y - G^*(U) \; \forall U, \pi$. As a result, for any $U \in \mathbb{R}^Y$, we have that $G(U) \geq \max_{\pi \geq 0} \sum_{y \in [Y]} \pi_y U_y - G^*(\pi)$. But then for any $U \in \mathbb{R}^Y$ by taking π to be the market shares which are predicted by U, we see that $G(U) = \max_{\pi \geq 0} \sum_{y \in [Y]} \pi_y U_y - G^*(\pi)$, by the envelope theorem. That concludes the proof.

1.6.2 Binary Example

Suppose that $Y = 1$, meaning that there are 2 choices $y = 0$ and $y = 1$. Assume that $\epsilon_0 = \epsilon_1$. Then, the discrete choice problem yields welfare: $G(U_1) = \max\{0, U_1\}$ and the market shares:

$$\begin{aligned} &\pi_1 = 1 \text{ if } U_1 > 0 \\ &\pi_1 = 0 \text{ if } U_1 < 0 \\ &\pi_1 \in [0, 1] \text{ if } U_1 = 0 \end{aligned}$$

This model is not very rich because for any nondegenerate observed market shares, we will need to pick $U_1 = 0$ to explain the data.

1.7 Example [Logit Model Social Welfare]

In the case of the logit model, we have $G(U) = \max_{\pi \geq 0} \sum_{y \in [Y]} \pi_y U_y - \sum_{y \in [Y]} \pi_y \log(\pi_y)$. By the FOC with respect to π_y, $U_y = \log(\frac{\pi_y}{\pi_0})$, as we've previously seen in the discrete choice inversion problem.

2 Parametric Discrete Choice Problem and Logistic Regression

Consider the discrete choice problem for an agent. An agent has the choice between Y options. For each each agent, each choice has associated with it characteristics $k \in [K]$, that reflect interactions between agent characteristics and raw characteristics of the choice. The agents utility associated with an option y is

$$\sum_{k \in [K]} \phi_{iyk} \lambda_k + \epsilon_{iy}$$

Again, ϵ_{iy} is drawn from a probability distribution $\mathcal{P}$ and reflect individual i's specific shocks associated with the choice y. We can view this utility as a parametrized version of the random utility model covered above where we say that an individuals utility is given by

$$U_{iy} + \epsilon_{iy} \tag{8}$$

where $U_{iy} = \sum_{k \in [K]} \phi_{iyk} \lambda_k = (\phi \lambda)_{iy}$.

To give some examples of plausible ϕ_{iyk}, we can consider an agent i's choice of travel mode $y \in \{\text{car, train, plane}\}$. Some examples of ϕ_{iyk} can be: (a) $\phi_{iy1} :=$ the price of option y times the inverse of the income of y, (b) $\phi_{iy2} :=$ the time taken by option y times a dummy for the consumer being retired.

We say that the probability that an individual $i \in [I]$ chooses an alternative $y \in [Y]$ when the parameter value is given by λ is given by: $\pi^{\lambda}_{iy} = \pi_y((\sum_{k \in [K]} \phi_{iyk} \lambda_k))$.

In a dataset, we know the choice y_i picked by any consumer i. Thus, the log-likelihood of a sample is:

$$l(\lambda) = \sum_{i \in [I]} \log(\pi^{\lambda}_{iy_i})$$

and the associated Maximum Likelihood Estimation Procedure consists of

$$\max_{\lambda \in \mathbb{R}^K} \sum_{i \in [I]} \log(\pi^{\lambda}_{iy_i})$$

Observe that the result from above applies here: $\pi^{\lambda}_{iy_i} = \frac{\partial G((\phi\lambda)_i)}{\partial U_{iy_i}}$.

2.1 Imposing the Gumbel Distribution for ϵ_{iy}

Taking the result from equation (4), with no default option, we have that:

$$\begin{aligned} \pi^{\lambda}_{iy_i} &= \frac{\exp(U_{iy_i})}{\sum_{z \in [Y]} \exp(U_{iz})} \\ &= \frac{\sum_{k \in [K]} \exp(\phi_{iy_i k} \lambda_k)}{\sum_{z \in [Y]} \exp(\sum_{k \in [K]} \phi_{izk} \lambda_k)} \end{aligned} \tag{9}$$

so that the log-likelihood of a sample is:

$$\begin{aligned} l(\lambda) &= \sum_{i\in[I]} \sum_{k\in[K]} \phi_{iy_ik}\lambda_k - \sum_{i\in[I]} \log(\sum_{z\in[Y]} \exp(\sum_{k\in K} \phi_{izk}\lambda_k)) \\ &= \sum_{(i,y,k)\in[I\times Y\times K]} \hat{\pi}_{iy}\phi_{iyk}\lambda_k - \sum_{i\in[I]} \log(\sum_{z\in[Y]} \exp(\sum_{k\in K} \phi_{izk}\lambda_k)) \end{aligned}$$

where $\hat{\pi}_{iy} = \mathbb{1}_{\{y_i=y\}}$.

Again, the Maximum Likelihood Estimation procedure consists of

$$\max_{\lambda\in\mathbb{R}^K} l(\lambda)$$

The first order condition of this problem with respect to λ_k is:

$$\begin{aligned} \sum_{(i,y)\in[I\times Y]} \hat{\pi}_{iy}\phi_{iyk} &= \sum_{i\in[I]} \frac{\sum_{y\in[Y]} \phi_{iyk} \exp(\sum_{k\in[K]} \phi_{iyk}\lambda_k)}{\sum_{y\in[Y]} \exp \sum_{k\in[K]} \phi_{iyk}\lambda_k} \\ &= \sum_{(i,y)\in[I\times Y]} \pi^{\lambda}_{iy}\phi_{iyk} \end{aligned} \tag{10}$$

where I use equation 9 to substitute in for π^{λ}_{iy}. The interpretation of this equation is that I aim to match the predicted expectation of ϕ with the observed average value of ϕ.

In summary, we can write the logistic regresison estimator in matrix form as:

$$\max_{\lambda\in\mathbb{R}^K} \{\hat{\pi}'\Phi\lambda - \sum_{i\in[I]} \log(\sum_{y\in[Y]} \exp((\Phi\lambda)_{iy}))\} \tag{11}$$

2.2 Logistic Regression as a Generalized Linear Model

In the linear regression model, we have a dependent variable $\tilde{\mu}$ and independent variables $\tilde{\rho_1}, ..., \tilde{\rho_p}$ and specify the conditional mean: $\mathbb{E}[\tilde{\mu}|\tilde{\rho_1}, ..., \tilde{\rho_p}] = \sum_{i=1}^{p} \tilde{\rho_i}\theta_i$ where $\{\theta_i\}_{i=1}^{p}$ is a set of parameters. Denoting $\tilde{\rho} = [\rho_1, ..., \rho_p]'$ and $\theta = [\theta_1, ..., \theta_p]'$, we often wish to generalize this into specifying the following conditional mean:

$$f(\mathbb{E}[\mu|\tilde{\rho}]) = \tilde{\rho}'\theta \Leftrightarrow \mathbb{E}[\mu|\tilde{\rho}] = f^{-1}(\tilde{\rho}'\theta)$$

assuming that f is strictly increasing, implying that it is invertible. In this setting f is called a *link function*. Setting $F^*(t) := \int_{-\infty}^{t} f^{-1}(u)du$ as the antiderivative of f^{-1}, we can reinterpret the conditional mean expression as the FOC in

$$\max_{\theta\in\Theta} \mathbb{E}[\tilde{\mu}\tilde{\rho}'\theta - F^*(\tilde{\rho}'\theta)]$$

In order to estimate θ, we write the sample analogue of the previous problem:

$$\max_{\theta\in\Theta} \sum_{\omega\in[\Omega]} [\tilde{\mu}_\omega(\tilde{\rho}'\theta)_\omega - F^*((\tilde{\rho}'\theta)_\omega)]$$

This estimator belongs to the class of *M-Estimators* given by:

$$\hat{\theta}_n := \max_{\theta \in \Theta} \mathbb{E}_{\mathcal{P}_\setminus}[l(z, \theta)]$$

2.2.1 Generalized Linear Model Examples

In least squares, the link function is $f(t) := t$. In Poisson regression, $f(t) = \log(t)$ so that $\mathbb{E}[\tilde{\mu}|\tilde{\rho}] = \exp(\tilde{\rho}'\theta)$. In the case of the Poisson regression, we can reinterpret the conditional mean as the first order conditions to an optimization problem: $\max_{\theta \in \Theta} \mathbb{E}[\tilde{\mu}(\tilde{\rho}'\theta) - \exp(\tilde{\rho}'\theta)]$. Indeed the FOC of this maximization problem is $\mathbb{E}[(\tilde{\mu} - \exp(\tilde{\rho}'\theta))\tilde{\rho}] = 0$, which is matches the conditional mean expression after conditioning on $\tilde{\rho}$ in the expectation.

Now, we can write the Poisson regression as a generalized linear model:

$$\max_{\theta \in \Theta} \sum_{\omega \in [\Omega]} \tilde{\mu}_\omega (\rho' \tilde{\theta})_\omega - \exp((\tilde{\rho}'\theta)_\omega)$$

When one estimates Poisson regression parameters, one typically uses a Maximum Likelihood approach in a Poisson conditional distribution. Recall that a parameter $\tilde{\mu}$ follows a Poisson distribution with parameter λ if $\Pr(\tilde{\mu} = m|\lambda) = \frac{\lambda^m}{m!}\exp(-\lambda)$. When we estimate the parameters of a Poisson conditional distribution, we typically say that $\tilde{\mu}$ conditional on $\tilde{\rho}$ follows a Poisson distribution of parameter $\exp(\tilde{\rho}'\theta)$.

The conditional log-likelihood of the sample, assuming each data-point is iid is:

$$\begin{aligned} l((\tilde{\mu}_\omega, \tilde{\rho}_\omega)_{\omega \in [\Omega]}|\theta) &= \log\{\Pi_{\omega \in [\Omega]} \frac{(\exp((\tilde{\rho}'\theta)_\omega))^{\tilde{\mu}_\omega}}{\tilde{\mu}_\omega!} \exp(-\exp((\tilde{\rho}'\theta)_\omega))\} \\ &= \sum_{\omega \in \Omega} \log(\exp((\tilde{\rho}'\theta)_\omega)^{\tilde{\mu}_\omega}) + \log(\exp(-\exp((\tilde{\rho}'\theta)_\omega))) - \log(\tilde{\mu}_\omega) \\ &= \sum_{\omega \in [\Omega]} \tilde{\mu}_\omega (\tilde{\rho}'\theta)_\omega - \exp((\tilde{\rho}'\theta)_\omega) + t.i.p. \end{aligned}$$

The Maximum Likelihood Estimator for the parameter of the conditional Poisson distribution matches the generalized linear model estimator where one used the $\log(\cdot)$ link function. These two estimators do differ when one does inference. That is because the Maximum Likelihood Estimator imposes distributional and variance assumptions on the model whereas the generalized linear model estimator does not.

2.3 Equivalence between the Poisson Regression and the Multinomial Logistic Regression

We wish to show the equivalence between the Poisson regression:

$$\max_{\theta \in \Theta} \sum_{\omega \in [\Omega]} \tilde{\mu}_\omega (\rho' \tilde{\theta})_\omega - \exp((\tilde{\rho}'\theta)_\omega)$$

and the Multinomial Logistic regression:

$$\max_{\lambda \in \Lambda} \sum_{(i,y,k) \in [I \times Y \times K]} \hat{\mu}_{iy} \phi_{iyk} \lambda_k - \sum_{i \in [I]} \log(\sum_{z \in [Y]} \exp(\sum_{k \in K} \phi_{izk} \lambda_k))$$

where $\hat{\mu}_{iy} = \mathbb{1}_{\{y_i = y\}}$.

We shall see that "Multinomial regression" = "Poisson regression" + "i-fixed effects" by the "Poisson trick."

2.3.1 Derivation

To equate the Poisson regression and the Multinomial Logistic regression, let (a) each $\omega \mapsto iy$, (b) again, $\hat{\mu}_{iy} = \mathbb{1}_{\{y_i=y\}}$, (c) the $\phi_{iyk} + i$-fixed effects go into $\tilde{\rho}$ meaning that $\sum_{p\in[P]}(\tilde{\rho}\theta)_\omega = \sum_{k\in[K]} \phi_{iyk}\lambda_k - u_i$ and $|[P]| = |[K]| + |[I]|$ and $\theta = [\lambda', u']'$.

Let us take a look at the maximization problem of the Poisson regression with the "i-fixed effects":

$$\max_{(\lambda,u)\in(\Lambda,R^I)} \sum_{(i,y)\in[I\times Y]} \hat{\mu}_{iy}(\sum_{k\in[K]} \phi_{iyk}\lambda_k - u_i) - \sum_{(i,y)\in[I\times Y]} \hat{\mu}_{iy} \exp(\sum_{k\in[K]} \phi_{iyk}\lambda_k - u_i) \tag{12}$$

Let's take the first order condition with respect to u_i.

$$\begin{aligned}
0 &= -\sum_{y\in[Y]} \overbrace{\hat{\mu}_{iy}}^{1} + \sum_{y\in[Y]} \exp(\sum_{k\in[K]} \phi_{iyk}\lambda_k - u_i) \\
\implies 1 &= \sum_{y\in[Y]} \exp(\sum_{k\in[K]} \phi_{iyk}\lambda_k - u_i) \\
\implies \exp(u_i) &= \sum_{y\in[Y]} \exp(\sum_{k\in[K]} \phi_{iyk}\lambda_k) \\
\implies u_i &= \log(\sum_{y\in[Y]} \exp(\sum_{k\in[K]} \phi_{iyk}\lambda_k)
\end{aligned}$$

where we can interpret the last term in the expression as the welfare of individual i. Then, substituting this expression for u_i into the maximization problem in equation (12), the maximization problem becomes:

$$\max_{\lambda\in\Lambda} \sum_{(i,y)\in[I\times Y]} \hat{\mu}_{iy}[\sum_{k\in[K]} \phi_{iyk}\lambda_k - \log(\sum_{z\in[Y]} \exp(\sum_{l\in[K]} \phi_{izl}\lambda_l))] - \sum_{(i,y)\in[I\times Y]} \hat{\mu}_{iy} \exp(\sum_{k\in[K]} \phi_{iyk}\lambda_k - \log(\sum_{z\in[Y]} \exp(\sum_{l\in[K]} \phi_{izl}\lambda_l)))$$

This is equivalent to the following problem where the term on the right is simplified by carefully accounting for which terms are independent of the sum:

$$\begin{aligned}
&\max_{\lambda\in\Lambda} \sum_{(i,y)\in[I\times Y]} \hat{\mu}_{iy}[\sum_{k\in[K]} \phi_{iyk}\lambda_k - \log(\sum_{z\in[Y]} \exp(\sum_{l\in[K]} \phi_{izl}\lambda_l))] - I \\
\Leftrightarrow &\max_{\lambda\in\Lambda} \sum_{(i,y,k)\in[I\times Y\times K]} \hat{\mu}_{iy}\phi_{iyk}\lambda_k - \log(\sum_{z\in[Y]} \exp(\sum_{l\in[K]} \phi_{izl}\lambda_l))
\end{aligned}$$

which is exactly the Multinomial Logistic Regression maximization problem.

2.4 The Method of Moments

In the Method of Moments, we again assume the same parametric utility model. We wish to match the predicted moments of ϕ with the observed moments of ϕ $\forall k \in [K]$:

$$\sum_{(i,y)\in[I\times Y]} \hat{\pi}_{iy}\phi_{iyk} = \sum_{(i,y)\in[I\times Y]} \pi^{\lambda}_{iy}\phi_{iyk}$$

Note that this condition is the same as the one I reach in the first order condition for the logistic regression in (10). We wish to generalize this conclusion and rationalize it using this moment matching approach. Recall also our earlier observation that: $\pi_{iy}^{\lambda} = \frac{\partial G((\phi\lambda)_i)}{\partial U_{iy}}$. Then, we recognize the right hand side of the equation as the derivative of $\sum_{i\in[I]} G((\phi\lambda)_i)$ with respect to λ_k and the left hand side as the derivative of $\sum_{(i,y)\in[I\times Y]} \hat{\pi}_{iy}\phi_{iyk}\lambda_k$ with respect to λ_k.

In summary, we see that the Method of Moments is the first order condition of the following maximization problem:

$$\max_{\lambda\in\Lambda} \sum_{(i,y,k)\in[I\times Y\times K]} \hat{\pi}_{iy}\phi_{iyk}\lambda_k - \sum_{i\in[I]} G((\phi\lambda)_i) \tag{13}$$

which very closely resembles equation (6), which defines the generalized entropy of choice.

2.5 The Dual Problem with the Generalized Entropy of Choice

Now we wish to show that the dual to the maximization problem in equation (13) is the following one:

$$\begin{aligned}
&\min_{\pi\geq 0} \sum_{i\in[I]} G^*(\pi_i) \\
&\text{s.t.} \sum_{(i,y)\in[I\times Y]} \pi_{iy}\phi_{iyk} = \sum_{(i,y)\in[I\times Y]} \hat{\pi}_{iy}\phi_{iyk}, \sum_{y\in[Y]} \pi_{iy} = 1 \; \forall i \in [I]
\end{aligned}$$

As intuition for the dual expression, we search for π such that we maximize the (negative) of the generalized entropy of choice, defined in equation (6), while obeying the moments observed in the data.

To show the dual problem, recall that by the Legendre-Fenchel Transform lemma in section 1.6.1, the maximization problem in the method of moments in equation (13) is equivalent to the following one:

$$\max_{\lambda\in\Lambda} \sum_{(i,y,k)\in[I\times Y\times K]} \hat{\pi}_{iy}\phi_{iyk}\lambda_k - \sum_{i\in[I]} \max_{\pi_i\geq 0}\{ \sum_{(y,k)\in[Y\times K]} \pi_{iy}\phi_{iyk}\lambda_k - G^*(\pi_i)\}$$

Next, note that I can introduce the minus sign into the second maximization problem, and thus flip the problem to a minimization problem. From there, I can move the position of the minimization.

$$\max_{\lambda\in\Lambda} \min_{\pi\geq 0} \sum_{(i,y,k)\in[I\times Y\times K]} \hat{\pi}_{iy}\phi_{iyk}\lambda_k - \sum_{(i,y,k)\in[I\times Y\times K]} \pi_{iy}\phi_{iyk}\lambda_k + \sum_{i\in[I]} G^*(\pi_i)$$

Waving my hands a little bit, I switch the order of the maximization problem and the minimization problem to reach the following equivalent problem.

$$\min_{\pi\geq 0} \max_{\lambda\in\Lambda} \sum_{(i,y,k)\in[I\times Y\times K]} \hat{\pi}_{iy}\phi_{iyk}\lambda_k - \sum_{(i,y,k)\in[I\times Y\times K]} \pi_{iy}\phi_{iyk}\lambda_k + \sum_{i\in[I]} G^*(\pi_i)$$

Rearranging, this minimization problem is equivalent to the following one:

$$\min_{\pi \geq 0} \max_{\lambda \in \Lambda} \{ \sum_{k \in [K]} \lambda_k [\sum_{(i,y) \in [I \times Y]} (\hat{\pi}_{iy} - \pi_{iy}) \phi_{iyk}] \} + \sum_{i \in [I]} G^*(\pi_i)$$

Next, note that if for some $k \in [K]$, the following term $\sum_{(i,y) \in [I \times Y]} (\hat{\pi}_{iy} - \pi_{iy}) \phi_{iyk}$ does not equal 0, then the inner maximizer will pick λ_k so as to send its argument to ∞. Thus the minimization problem is equivalent to the following one, as desired:

$$\min_{\pi \geq 0} \sum_{i \in [I]} G^*(\pi_i)$$
$$\text{s.t.} \sum_{(i,y) \in [I \times Y]} \pi_{iy} \phi_{iyk} = \sum_{(i,y) \in [I \times Y]} \hat{\pi}_{iy} \phi_{iyk}, \sum_{y \in [Y]} \pi_{iy} = 1 \; \forall i \in [I]$$

2.6 Macro Logistic Regression

Assume that an individual i has an observable type $x_i \in [X]$. The utility of an individual i for picking choice y, assuming that the type of i is x (ie., $x_i = x$), is given by:

$$\sum_{k \in [K]} \phi_{xyk} \lambda_k + \epsilon_{iy}$$

The assumption here is that the Econometrician cannot distinguish between $i, j \in [X]$. Define these helpful variables: $n_x := \sum_{i \in [I]} \mathbb{1}_{x_i = x}$, $\hat{\mu}_{xy} := \sum_{i \in [I]} \mathbb{1}_{x_i = x} \mathbb{1}_{y_i = y}$, and $m_y := \sum_{i \in [I]} \mathbb{1}_{y_i = y}$.

The macro logistic regression is given by the following maximization problem:

$$\max_{\lambda \in \Lambda} \sum_{(x,y,k) \in [X \times Y \times K]} \hat{\mu}_{xy} \phi_{xyk} \lambda_k - \sum_{x \in [X]} n_x \log(\sum_{y \in [Y]} \exp(\sum_{k \in [K]} \phi_{xyk} \lambda_k))$$

which gives the first order condition for each λ_k:

$$\sum_{(x,y) \in [X \times Y]} n_x \pi^{\lambda}_{xy} \phi_{xyk} = \sum_{(x,y) \in [X \times Y]} \hat{\mu}_{xy} \phi_{xyk}$$
$$\text{with } \pi^{\lambda}_{xy} = \frac{\exp(\sum_{k \in [K]} \phi_{xyk} \lambda_k)}{\sum_{z \in [Y]} \exp(\sum_{k \in K} \phi_{xzk} \lambda_k)}$$

Note that the dual problem of the macro logistic regression is given by:

$$\min_{\pi \geq 0} \sum_{x \in [X]} n_x \sum_{y \in [Y]} \pi_{xy} \log(\pi_{xy})$$
$$\text{s.t.} \sum_{(x,y) \in [X \times Y]} n_x \pi_{xy} \phi_{xyk} = \sum_{(x,y) \in [X \times Y]} \hat{\mu}_{iy} \phi_{xyk}, \sum_{y \in [Y]} \pi_{xy} = 1 \; \forall x \in [X]$$

This dual problem is very similar in spirit to the dual problem of the method of moments covered in section 2.5.

2.7 Multinomial Logistic Regression Alternative Formulation

2.7.1 Prerequisites- Matrix Vectorization and the Kroenecker Product

For a matrix A = $\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$, we vectorize the matrix in *row-major order* as

$$\text{vec}_R(A) := [A_{11}, A_{12}, A_{21}, A_{22}]'$$

For the matrix A defined above and a matrix B, we define the *Kroenecker Product* between A and B as

$$A \otimes B := \begin{pmatrix} A_{11}B & A_{12}B \\ A_{21}B & A_{22}B \end{pmatrix}.$$

Finally, I state a useful fact of the *Kroenecker Product* and *row-major order* vectorization. For conformable matrices A, X, B,

$$\text{vec}_R(AXB) = (A \otimes B')\text{vec}_R(X)$$

2.7.2 On to the Model

The Multinomial Logistic Regression is sometimes presented under a different form. Here, we write that the utility for individual i to choose choice y is $U_{iy} + \epsilon_{iy}$ where

$$U_{iy} = \sum_{p \in [P]} \psi_{iyp}\theta_{py} = (\psi\theta)_{iy}$$

We have that $U = \psi\theta$ where we view U as an $I \times Y$ matrix, ψ as an $I \times P$ matrix, and θ as a $P \times Y$ matrix.

In our original formulation, we wrote $U_{iy} = \sum_{k \in [K]} \phi_{iyk}\lambda_k = (\phi\lambda)_{iy}$. In this setting, we had that $U = \phi\lambda$ where U is viewed as a column vector of size IY, ϕ is viewed as an $IY \times K$ matrix, and λ is viewed as a column vector of size K.

Next, I wish to show that this formulation can be expressed by the original formulation. Write $U = \psi\theta I_Y$. Then, using the useful fact from the prerequisite section, $\text{vec}_R(U) = \text{vec}_R(\psi\theta I_Y) = (\psi \otimes I_Y)\text{vec}_R(\theta)$. Finally, we can define our variables from the original formulation: $\phi := \psi \otimes I_Y$, $\lambda := \text{vec}_R(\theta)$, and a result $K = PY$.

2.8 Existence in Logistic Regression

The existence of a logistic regression estimator is not always guaranteed. In fact, we have the following result:

Theorem 1. *The following conditions are equivalent:*
(i) the logistic regression estimator of equation (11) exists
(ii) there exists $\bar{\pi} \in \mathbb{R}^{I \times Y}$ such that $\bar{\pi}_{iy} > 0 \; \forall (i, y) \in [I \times Y]$ and

$$\sum_{i \in [I]} \bar{\pi}_{iy} = \sum_{i \in [I]} \hat{\pi}_{iy} \;\; \forall y \in [Y] \text{ and } \sum_{(i,y) \in [I \times Y]} \bar{\pi}_{iy}\phi_{iyk} = \sum_{(i,y) \in [I \times Y]} \hat{\pi}_{iy}\phi_{iyk} \; \forall k \in [K]$$

(iii) the value V of the following linear programming problem is finite and strictly positive

$$
\begin{aligned}
V = & \max_{t \geq 0, \pi \in \mathbb{R}_+^{I \times Y}} t \\
& s.t.\ \bar{\pi}_{iy} \geq t \\
& \sum_{i \in [I]} \bar{\pi}_{iy} = \sum_{i \in [I]} \hat{\pi}_{iy} \ \forall y \in [Y] \\
& \sum_{(i,y) \in [I \times Y]} \bar{\pi}_{iy} \phi_{iyk} = \sum_{(i,y) \in [I \times Y]} \hat{\pi}_{iy} \phi_{iyk}
\end{aligned}
$$

As an example, consider the simple case where $I = 2$, $Y = 2$, and the regressors Φ contains only a dummy of the form $\phi_{iy1} = \mathbb{1}_{\{y=2\}}$. Defining $\hat{\pi}_2 := \sum_{i \in [I]} \hat{\pi}_{iy}$, the logistic regression of equation (11) reduces to:

$$\max_{\lambda \in \mathbb{R}} \{\hat{\pi}_2 \lambda - 2 \log(1 + \exp(\lambda))\}$$

Notice that while the objective is concave, it may be unbounded. If $\pi_2 = 0$, then we wish to send $\lambda \to -\infty$, to maximize the objective so that the maximizer isn't achieved. Meanwhile, if $\pi_2 = 2$, then we wish to send $\lambda \to \infty$ to maximize the objective so that again the maximizer isn't achieved.

2.9 Identification in the Logistic Regression

Recall that Poisson regression can be given by the following maximization problem in matrix form:

$$\max_{\theta \in \Theta} \hat{\mu}' R\theta - 1' \exp(R\theta)$$

where $\hat{\mu}$ is a dependent variable that $\Omega \times 1$, θ is a vector of parameters that $P \times 1$, and R is the design matrix that's $\Omega \times P$. In order for θ to be identified, we need R to have full rank (ie., rank$(R) = P$. Otherwise, there would be a nonzero vector δ such that $R\delta = 0$ so that if θ is a solution, then $\theta + \delta$ is also a solution.

Recall that we formulated the logistic regression as a Poisson regression where the design matrix $R := [\phi, -I_I \otimes 1_Y]$. For $\theta := [\lambda, u_i]$ to be identified, we need ϕ to have full rank and Im(ϕ) not to contain any vectors of the form $a \otimes 1_Y$ for some $a \in \mathbb{R}^I \setminus \{0\}$. Intuitively, such a vector would be correlated with the individual fixed effects.

2.10 Goodness of Fit and Interpretation of Parameters

Recall the utility specification of equation (8), that leads to the discrete choice problem

$$\max_{y \in [Y]} \{U_{iy} + \epsilon_{iy}, \epsilon_{i0}\}$$

Note that the the solution(s) y^* to the problem are invariant to adding constants A to each term in the maximization and multiplying each term by a positive scalar value σ. Assume that $U_{iy} = \sum_{k \in [K]} \phi_{iyk} \lambda_k$. We then have that the problem is equivalent to

$$\max_{y \in [Y]} \{\sigma \sum_{k \in [K]} \phi_{iyk} \lambda_k + \sigma \epsilon_{iy}, \sigma \epsilon_{i0}\} \tag{14}$$

Let λ^σ be the estimator of the Logistic regression with $\sigma\epsilon_y$ as the random utility. We have that $\lambda^\sigma = \sigma\lambda^1$, which means that we can only really identify $\frac{\lambda^\sigma}{\sigma}$. Thus, one good measure of the goodness of fit is $||\lambda^1||$, which gives us a sense of the signal-noise ratio. We can also choose to scale the problem in equation (14) by $\lambda_0 > 0$, and call $T = \frac{\sigma}{\lambda_0}$ to get the equivalent problem

$$\max_{y\in[Y]}\{\phi_{iy0} + \sum_{k\geq 1}\phi_{iyk}\frac{\lambda_k}{\lambda_0} + \frac{\sigma}{\lambda_0}\epsilon_{iy}, \frac{\sigma}{\lambda_0}\epsilon_{i0}\}$$

In order to solve this problem, we can solve the standard logistic regression problem

$$\max_{y\in[Y]}\{\phi_{iy0}\lambda_0' + \sum_{k\geq 1}\phi_{iyk}\lambda_k' + \epsilon_{iy}, \epsilon_{i0}\}$$

and then deduce $\lambda_k = \frac{\lambda_k'}{\lambda_0'}, T = \frac{1}{\lambda_0'}$.

2.10.1 Example

Consider a discrete choice model of transportation where individuals pick amongst $y \in [Y]$ transportation alternatives or $y = 0$, which is the choice of staying home. Consider a case where $\phi_{iy0} = -C_y$, which is minus the cost of the trip using alternative y and $\phi_{iy1} = -D_y$, which is the duration of the trip taking alternative y. The discrete choice problem is

$$\max_{y\in[Y]}\{-\lambda_0\phi_{iy0} - \lambda_1\phi_{iy1} + \epsilon_{iy}, \epsilon_{i0}\}$$

Since as shown above, we can only really identify the ratio of coefficients, we can view $\frac{\lambda_1}{\lambda_0}$ as the conversion between time of the trip to dollar cost of the trip. As a result, if we're a policy planner who can spend \$X to reduce the cost of a trip for one alternative by say, subsidizing tickets, or reduce the duration of the trip by improving the quality of the alternative, we can use this coefficient to help determine which is solution is most cost-effective to maximize welfare.

2.10.2 Sample Splitting

Besides looking at the norm of λ, ie., $||\lambda||$, to get a sense of how 'good' our discrete choice model is, we can also consider sample splitting. That is, we can split our sample into a training set and a test set. We can use the training set to estimate $\hat{\lambda}$ and then we can use the test set to evaluate how good our estimated $\hat{\lambda}$ is at predicting the choices made by individuals. Under the assumption that our train and test sets are random samples from the same distribution, this test and various metrics on the test set, will give us a sense of how good our model is at capturing the decision making of individuals in the population.

3 THE GRAVITY EQUATION

The goal of the Gravity equation is to explain the volume of trade between countries. We have data-points $\hat{\mu}_{xy}$ which gives the volume of trade between country x and country y. We wish to explain the volume using characteristics ϕ_{xyk} such as the distance or similarity between country x and country y. The Gravity equation assumes that

$$\mu_{xy}^\lambda = \exp((\phi\lambda)_{xy} - u_x - v_y) \tag{15}$$

where u_x and v_y are exporter and importer fixed effects that adjust for the total volume of exports and imports to a country. As a caveat, we do not allow for self-trade.

Define n_x to be the predicted total exports of country x and m_y to be the predicted total imports of country y so that

$$n_x := \sum_{y \in [Y]} \mu^{\lambda}_{xy} = \sum_{y \neq x} \exp((\phi\lambda)_{xy} - u_x - v_y) \tag{16}$$

$$m_y := \sum_{x \in [Y]} \mu^{\lambda}_{xy} = \sum_{x \neq y} \exp((\phi\lambda)_{xy} - u_x - v_y) \tag{17}$$

We estimate λ, u, v by matching the moments of ϕ for each k, and the total exports and imports. That is, we look for λ, u, v such that:

$$\hat{n}_x = \sum_{y \neq x} \exp((\phi\lambda)_{xy} - u_x - v_y) \tag{18}$$

$$\hat{m}_y = \sum_{x \neq y} \exp((\phi\lambda)_{xy} - u_x - v_y) \tag{19}$$

$$\sum_{(x,y) \in [X \times Y]} \hat{\mu}_{xy} \phi_{xyk} = \sum_{(x,y) \in [X \times Y]} \exp((\phi\lambda)_{xy} - u_x - v_y) \phi_{xyk} \tag{20}$$

We can view these equations as first order conditions of a Logistic Regression problem with y fixed effects. While we don't include the exporter fixed effects u in the regression, we will define them in the first order conditions to match the moments above.

$$\max_{\lambda \in \Lambda, v \in R^Y} \sum_{(x,y) \in [X \times Y]} \hat{\mu}_{xyk} \Big(\sum_{k \in [K]} \phi_{xyk} \lambda_k - v_y \Big) - \sum_{x \in [X]} \hat{n}_x \log \Big(\sum_{y \in [Y]} \exp \Big(\sum_{k \in [K]} \phi_{xyk} \lambda_k - v_y \Big) \Big)$$

The first order conditions of this problem with respect to λ_k and v_y respectively are

$$\sum_{(x,y) \in [X \times Y]} \hat{\mu}_{xy} \phi_{xyk} = \sum_{(x,y) \in [X \times Y]} \mu^{\lambda}_{xy} \phi_{xyk}$$

$$\sum_{x \in [X]} \hat{\mu}_{xy} = \sum_{x \in [X]} \mu^{\lambda}_{xy}$$

$$\text{where } \mu^{\lambda}_{xy} = \hat{n}_x \frac{\exp(\sum_{k \in [K]} \phi_{xyk} \lambda_k - v_y)}{\sum_{z \in [Y]} \exp(\sum_{k \in [K]} \phi_{xzk} - v_z)}$$

Let $u_x := \log(\frac{1}{\hat{n}_x} \sum_{z \in [Y]} \exp(\sum_{k \in [K]} \phi_{xzk} - v_z))$ so that I can rewrite $\mu^{\lambda}_{xy} = \exp((\phi\lambda)_{xy} - u_x - v_y)$ so that it's in the recognizable form. Note that the first FOC directly matches the phi moment in equation (20). To match total exports moment of equation (18), note that $\sum_{y \in [Y]} \mu^{\lambda}_{xy} = \hat{n}_x$. Finally, to match the total imports moment of equation (19), observe that the second FOC directly matches it where by definition $\hat{m}_y = \sum_{x \in [X]} \hat{\mu}_{xy}$.

Finally, I would like to give some intuition for the expression for μ^{λ}_{xy} in equation (15) in light of the first order conditions. Take the perspective of an exporter x. The ϕ_{xyk} are characteristics between x and another country y, such as the distance between them as I describe above. For instance, if y is closer to x, we would expect x to export more to them. Next, consider v_y: if the country y is larger, we would expect x to export more to them. Finally, the purpose of u_x is to simultaneously (a) scale down each of the scores for x's exports to be fractions of its total exports (as does the i-fixed effect in the Poisson regression when equating it to the Logistic Regression in section 2.3) and (b) scale up those fractions to total exports by using $\hat{n}_x$

4 MATCHING WITH TRANSFERABLE UTILITY

Let $x \in [X]$ be the observable type of a worker where there are an exogenously given n_x workers of type x. Also, let $y \in [Y]$ be the observable type of the firm where there are an exogenously given m_y firms of type y. In general, we have that $\sum_{x\in[X]} n_x \neq \sum_{y\in[Y]} m_y$. We consider a model where a worker matches with exactly one firm and a firm hires exactly one worker. Firms and workers have the choice to remain unemployed or not hire, respectively.

4.1 Worker's Problem

Assume that if a worker i of type x matches with a firm j of type y, then i gets utility

$$\alpha_{xy} + w_{xy} + \epsilon_{iy}$$

where α_{xy} is interpreted as the systematic utility of the pairing for the worker, w_{xy} is the agreed upon wage, and ϵ_{iy} is a individual specific random component of utility. We note that the individual specific shock is independent of j so that the worker cannot distinguish between firms of type y. If the worker chooses to remain unemployed, they receive utility ϵ_{i0}.

The worker's indirect utility is given by

$$u_i = \max_{y\in[Y]} \{\alpha_{xy} + w_{xy} + \epsilon_{iy}, \epsilon_{i0}\}$$

From here on out, we also assume that $\epsilon_i = e_i - \gamma$ where $e_i \overset{\text{iid}}{\sim}$ Gumbel$(\mu = 0, \beta = 1)$[1]. Then, where I let u_x be the expected indirect utility for workers of type $x \in [X]$, I have that:

$$\begin{aligned} u_x &= \mathbb{E}_{\mathcal{P}}[\max_{y\in[Y]} \{\alpha_{xy} + w_{xy} + \epsilon_{iy}, \epsilon_{i0}\}] \\ &= \log(1 + \sum_{y\in[Y]} \exp(\alpha_{xy} + w_{xy})) \end{aligned}$$

Similarly, the probability that a worker of type $x \in [X]$ chooses a firm of type y is

$$\begin{aligned} \pi_{xy} &= \frac{\exp(\alpha_{xy} + w_{xy})}{1 + \sum_{y\in[Y]} \exp(\alpha_{xy} + w_{xy})} \\ &= \exp(\alpha_{xy} + w_{xy} - u_x) \end{aligned}$$

Finally, we say that the number of workers of type $x \in [X]$ demanding firms of type y is

$$\begin{aligned} \mu_{xy}^W &= n_x \pi_{xy} \\ &= n_x \exp(\alpha_{xy} + w_{xy} - u_x) \end{aligned}$$

and the number of workers of type $x \in [X]$ that choose to opt out of the labor market is

[1] γ is Euler's constant.

4.2 Firm's Problem

Assume that if a firm j of type y matches with a worker i of type x, then j gets utility

$$\gamma_{xy} - w_{xy} + \eta_{jx}$$

where γ_{xy} is interpreted as the systematic utility of the pairing for the firm, w_{xy}, again, is the agreed upon wage, and η_{jx} is a firm specific random component of utility. We note that the firm specific shock is independent of i so that the worker cannot distinguish between workers of type x. If the firm chooses not to hire, they receive utility η_{j0}.

The firm's indirect utility is given by

$$v_j = \max_{x \in [X]} \{\gamma_{xy} - w_{xy} + \eta_{jx}, \eta_{j0}\}$$

From here on out, we also assume that $\eta_j = e_j - \gamma$ where $e_j \overset{\text{iid}}{\sim} \text{Gumbel}(\mu = 0, \beta = 1)$[2]. Then, where I let v_y be the expected indirect utility for firms of type $y \in [Y]$, I have that:

$$\begin{aligned} v_y &= \mathbb{E}_{\mathcal{P}}[\max_{x \in [X]} \{\gamma_{xy} - w_{xy} + \eta_{jx}, \eta_{j0}\}] \\ &= \log(1 + \sum_{x \in [X]} \exp(\gamma_{xy} - w_{xy})) \end{aligned}$$

Similarly, the probability that a firm of type $y \in [Y]$ chooses a worker of type x is

$$\begin{aligned} \pi_{yx} &= \frac{\exp(\gamma_{xy} - w_{xy})}{1 + \sum_{x \in [X]} \exp(\gamma_{xy} - w_{xy})} \\ &= \exp(\gamma_{xy} - w_{xy} - v_y) \end{aligned}$$

Finally, we say that the number of firms of type $y \in [Y]$ demanding workers of type x is

$$\begin{aligned} \mu^F_{xy} &= m_y \pi_{yx} \\ &= m_y \exp(\gamma_{xy} - w_{xy} - v_y) \end{aligned}$$

and the number of firms of type $y \in [Y]$ that choose to opt out of the labor market is

$$\begin{aligned} \mu_{0y} &= m_y \pi_{y0} \\ &= m_y \exp(-v_y) \end{aligned}$$

[2] Again, here, γ is Euler's constant.

4.3 Competitive Equilibrium

At the competitive equilibrium, by definition, everyone gets their first choice so that

$$\begin{aligned}\mu_{xy} &= \mu_{xy}^W = \mu_{xy}^F \\ &= n_x \exp(\alpha_{xy} + w_{xy} - u_x) = m_y \exp(\gamma_{xy} - w_{xy} - v_y)\end{aligned}$$

By taking the harmonic mean of μ_{xy}^W and μ_{xy}^F, so that the wage cancels, we have that

$$\begin{aligned}\mu_{xy} &= \sqrt{\mu_{xy}^W \mu_{xy}^F} \\ &= \sqrt{n_x m_y \exp(\alpha_{xy} + \gamma_{xy} - u_x - v_y)} \\ &= \sqrt{n_x m_y \exp(\phi_{xy} - u_x - v_y)}\end{aligned}$$

where $\phi_{xy} := \alpha_{xy} + \gamma_{xy}$ is the systematic dollar valuation of a match between types x and y. Let's reformulate the equations of the model, we have:

$$\begin{aligned}n_x &= \mu_{x0} + \sum_{y\in[Y]} \mu_{xy} \\ &= n_x \exp(-u_x) + \sum_{y\in[Y]} \sqrt{n_x m_y \exp(\phi_{xy} - u_x - v_y)} \\ &= \exp(-a_x) + \sum_{y\in[Y]} \sqrt{\exp(\phi_{xy} - a_x - b_y)} \\ &= \exp(-a_x) + \sum_{y\in[Y]} \exp(\frac{\phi_{xy} - a_x - b_y}{2}) \\ m_y &= \mu_{0y} + \sum_{x\in[X]} \mu_{xy} \\ &= m_y \exp(-v_y) + \sum_{x\in[X]} \sqrt{n_x m_y \exp(\phi_{xy} - u_x - v_y)} \\ &= \exp(-b_y) + \sum_{x\in[X]} \sqrt{\exp(\phi_{xy} - a_x - b_y)} \\ &= \exp(-b_y) + \sum_{x\in[X]} \exp(\frac{\phi_{xy} - a_x - b_y}{2})\end{aligned}$$

where I define $a_x := u_x - \log(n_x)$ and $b_y := v_y - \log(m_y)$. For clarity, I can write the two derived equations once more:

$$\begin{aligned}n_x &= \exp(-a_x) + \sum_{y\in[Y]} \exp(\frac{\phi_{xy} - a_x - b_y}{2}) \\ m_y &= \exp(-b_y) + \sum_{x\in[X]} \exp(\frac{\phi_{xy} - a_x - b_y}{2})\end{aligned}$$

These looks a lot like the equations to predict exports (ie., equation (16)) and to predict imports (ie., equation (17)) in the Gravity Equation model. Taking ϕ_{xy} for $x \in [X]$ and $y \in [Y]$ as given, we can show that these are the first order conditions of the following convex optimization problem:

$$\max_{(a,b)\in R^X\times R^Y} \sum_{x\in[X]} n_x(-a_x) + \sum_{y\in[Y]} m_y(-b_y) - \sum_{(x,y)\in[X\times Y]} 2\exp(\frac{\phi_{xy}-a_x-b_y}{2}) - \sum_{x\in[X]}\exp(-a_x) - \sum_{y\in[Y]}\exp(-b_y) \tag{21}$$

4.4 Computing the Equilibirum

To compute the equilibrium, we will take on of three approaches, (1) gradient descent/ ascent, (2) coordinate descent/ ascent, and (3) a generalized linear model regression.

4.4.1 Gradient Descent/ Ascent

To solve the strictly concave problem $\max_{\theta\in\Theta} F(\theta)$ by gradient ascent, we pick an initial θ_0 and then in iteration s update θ_s according to

$$\theta_s := \theta_{s-1} + \eta\nabla F(\theta_{s-1})$$

We terminate the iterative procedure according to a stopping criterion on θ, say $||\theta_s - \theta_{s-1}||_2 < \text{tol}_1$ or $|F(\theta_s) - F(\theta_{s-1})| < \text{tol}_2$ where η is a learning rate and tol_1 or tol_2 is appropriately chosen given the context.

In the context of our problem, $\theta = [a, b]'$ and

$$\begin{aligned} F(\theta) &= F(a,b) \\ &= \sum_{x\in[X]} n_x(-a_x) + \sum_{y\in[Y]} m_y(-b_y) - \sum_{(x,y)\in[X\times Y]} 2\exp(\frac{\phi_{xy}-a_x-b_y}{2}) - \sum_{x\in[X]}\exp(-a_x) - \sum_{y\in[Y]}\exp(-b_y) \end{aligned}$$

Then,

$$\begin{aligned} \frac{\partial F(a,b)}{\partial a_x} &= -n_x + \sum_{y\in[Y]} \exp(\frac{\phi_{xy}-a_x-b_y}{2}) + \exp(-a_x) \\ \frac{\partial F(a,b)}{\partial b_y} &= -m_y + \sum_{x\in[X]} \exp(\frac{\phi_{xy}-a_x-b_y}{2}) + \exp(-b_y) \end{aligned}$$

4.4.2 Coordinate Descent/ Ascent

To solve a strictly concave problem $\max_{(a,b)\in R^X\times R^Y} F(a,b)$ by coordinate ascent, we initialize $(a^{(0)}, b^{(0)})$ and at step s, determine $a^{(s)}$ from $b^{(s-1)}$ by $a^{(s)} := \text{argmin}_{a\in R^X} F(a, b^{(s-1)})$. Then, we determine $b^{(s)}$ from $a^{(s)}$ by $b^{(s)} := \text{argmin}_{b\in R^Y} F(a^{(s)}, b)$. We terminate when $||[a^{(s-1)}, b^{(s-1)}]' - [a^{(s)}, b^{(s)}]'||_2 < \text{tol}_1$ or $|F(a^{(s-1)}, b^{(s-1)}) - F(a^{(s)}, b^{(s)})| < \text{tol}_2$ as we prefer for some appropriately chosen tolerances.

In our setting, when finding $a^{(s)}$ from $b^{(s-1)}$, we have the first order condition (taken from the gradient descent section) for a particular dimension a_x of a:

$$n_x = \sum_{y\in[Y]} \exp(\frac{\phi_{xy}-a_x-b_y^{(s-1)}}{2}) + \exp(-a_x)$$

If we define $K_{xy} := \exp(\frac{\phi_{xy}}{2}), A_x := \exp(-\frac{a_x}{2}), B_y := \exp(-\frac{b_y}{2})$, then we have

$$n_x = \sum_{y\in[Y]} K_{xy}A_xB_y^{(s-1)} + A_x^2 \implies A_x = \sqrt{(\frac{KB_x^{(s-1)}}{2})^2 + n_x} - \frac{KB_x^{(s-1)}}{2}$$

where $KB_x^{(s-1)} = \sum_{y\in[Y]} K_{xy}B_y^{(s-1)}$. Symmetrically,

$$B_y = \sqrt{(\frac{KA_y^{(s)}}{2})^2 + m_y} - \frac{KA_y^{(s)}}{2}$$

where $KA_y^{(s)} = \sum_{x\in[X]} K_{xy}A_x^{(s)}$.

From these expressions for $A_x^{(s)}, B_y^{(s)}$, we can compute a and b and track the progress of our iterative algorithm and terminate appropriately.

4.4.3 Generalized Linear Model for the Method of Moments

We observe $\hat{\mu}_{xy}, \hat{\mu}_{x0}, \hat{\mu}_{0y}$ and parametrize $\phi_{xy}^{\lambda} := \sum_{k\in[K]} \phi_{xyk}\lambda_k$.

Let $\theta := [\lambda, u, v]'$. Adapting to this notation, we have,

$$\begin{aligned}
\mu_{xy}^{\theta} &= \sqrt{n_x m_y}\exp(\frac{\phi_{xy}^{\lambda} - u_x - v_y}{2}) \\
\mu_{x0}^{\theta} &= n_x \exp(-u_x) \\
\mu_{0y}^{\theta} &= m_y \exp(-v_y)
\end{aligned}$$

We determine θ by the equations:

$$\begin{aligned}
n_x &= \mu_{x0}^{\theta} + \sum_{y\in[Y]} \mu_{xy}^{\theta} \\
m_y &= \mu_{0y}^{\theta} + \sum_{x\in[X]} \mu_{xy}^{\theta} \\
&\sum_{(x,y)\in[X\times Y]} \mu_{xy}^{\theta}\phi_{xyk} = \sum_{(x,y)\in[X\times Y]} \hat{\mu}_{xy}\phi_{xyk}
\end{aligned} \tag{22}$$

which are the moment conditions associated with the following minimization problem:

$$\begin{aligned}
\min_{u_x,v_y,\lambda}\{ &\sum_{x\in[X]} n_x u_x + \sum_{y\in[Y]} m_y v_y \\
&+ 2\sum_{(x,y)\in[X\times Y]} \sqrt{n_x m_y \exp(\sum_{k\in[K]} \phi_{xyk}\lambda_k - u_x - v_y)} \\
&+ \sum_{x\in[X]} n_x \exp(-u_x) + \sum_{y\in[Y]} m_y \exp(-v_y) \\
&- \sum_{(x,y,k)\in[X\times Y\times K]} \hat{\mu}_{xy}\phi_{xyk}\lambda_k\}
\end{aligned}$$

This is a weighted Poisson regression that we can clean up a bit by introducing notation: $\tilde{u}_x := u_x - \log(n_x), \tilde{v}_y := v_y - \log(m_y)$. The minimization problem becomes:

$$\begin{aligned}\min_{\tilde{u}_x,\tilde{v}_y,\lambda}\{ & \sum_{x\in[X]} n_x\tilde{u}_x + \sum_{y\in[Y]} m_y\tilde{v}_y \\ & + 2\sum_{(x,y)\in[X\times Y]} \exp(\frac{\sum_{k\in[K]}\phi_{xyk}\lambda_k - \tilde{u}_x - \tilde{v}_y}{2}) \\ & + \sum_{x\in[X]}\exp(-\tilde{u}_x) + \sum_{y\in[Y]}\exp(-\tilde{v}_y) \\ & - \sum_{(x,y,k)\in[X\times Y\times K]} \hat{\mu}_{xy}\phi_{xyk}\lambda_k\}\end{aligned}$$

which can be rewritten once more as

$$\begin{aligned}\min_{\tilde{u}_x,\tilde{v}_y,\lambda}\{ & \sum_{x\in[X]} \hat{\mu}_{x0}\tilde{u}_x + \sum_{y\in[Y]} \hat{\mu}_{0y}\tilde{v}_y \\ & - \sum_{(x,y)\in[X\times Y]} \hat{\mu}_{xy}(\sum_{k\in[K]}\phi_{xyk}\lambda_k - \tilde{u}_x - \tilde{v}_y) \\ & + 2\sum_{(x,y)\in[X\times Y]} \exp(\frac{\sum_{k\in[K]}\phi_{xyk}\lambda_k - \tilde{u}_x - \tilde{v}_y}{2}) \\ & + \sum_{x\in[X]}\exp(-\tilde{u}_x) + \sum_{y\in[Y]}\exp(-\tilde{v}_y)\}\end{aligned}$$

Recall that the (weighted) Poisson regression takes the form:

$$\min_{\theta\in\Theta}\sum_{\omega\in\Omega} w_\omega(\exp(\sum_{p\in[P]} R_{\omega p}\theta_p) - \mu_\omega\sum_{p\in[P]} R_{\omega p}\theta_p) \tag{23}$$

where w_ω is the weight associated with observation ω, $R_{\omega p}$ is the design matrix, θ_p is the vector of parameters, and μ_ω is the dependent variable. In our setting, we can match the problem of the weighted Poisson regression by assigning:

$$\begin{aligned}&\omega \in [xy, x0, 0y] \\ &w_{xy} = 2, w_{x0} = 1, w_{0y} = 1 \\ &R = \begin{pmatrix} (\frac{\phi_{xyk}}{2})_{xy,k} & \frac{-I_X\otimes 1_Y}{2} & \frac{-1_X\otimes I_Y}{2} \\ O_{X\times K} & -I_X & O_{X\times Y} \\ O_{Y\times X} & 0_{Y\times X} & -I_Y \end{pmatrix} \\ &\mu_\omega \in [\hat{\mu}_{xy}, \hat{\mu}_{x0}, \hat{\mu}_{0y}] \\ &\theta = [\lambda, \tilde{u}, \tilde{v}]\end{aligned} \tag{24}$$

4.4.4 Generalized Linear Model for the Maximum Likelihood Estimator

For clarity, I will repeat once more the model's predicted conditions. Letting $\theta = [\lambda, u, v]'$, I have the the equilibrium predicted matches:

$$\begin{aligned}\mu_{xy}^{\theta} &= \sqrt{n_x m_y}\exp(\frac{\phi_{xy}^{\lambda} - u_x - v_y}{2})\\ \mu_{x0}^{\theta} &= n_x \exp(-u_x)\\ \mu_{y0}^{\theta} &= m_y \exp(-v_y)\end{aligned}$$

Letting $a_x := u_x - \log(n_x), b_x := v_y - \log(m_y)$, these equations reformulate as:

$$\begin{aligned}\mu_{xy}^{\theta} &= \exp(\frac{\phi_{xy}^{\lambda} - a_x - b_y}{2})\\ \mu_{x0}^{\theta} &= \exp(-a_x)\\ \mu_{y0}^{\theta} &= \exp(-b_y)\end{aligned}$$

Given the parameter θ, what is the likelihood of observing a specific match $(xy), (x0)$, or $(y0)$ conditional on seeing a match? Denoting $N^{\theta} := \sum_{(x,y)\in[X\times Y]}\mu_{xy}^{\theta} + \sum_{x\in[X]}\mu_{x0}^{\theta} + \sum_{y\in[Y]}\mu_{0y}^{\theta}$, and the predicted probabilities as π^{θ}, we see that:

$$\pi_{xy}^{\theta} = \frac{\mu_{xy}^{\theta}}{N^{\theta}}, \pi_{x0}^{\theta} = \frac{\mu_{x0}^{\theta}}{N^{\theta}}, \pi_{0y}^{\theta} = \frac{\mu_{0y}^{\theta}}{N^{\theta}} \tag{25}$$

where $\hat{\mu}_{xy}$ are the number of matches between x, y, $\hat{\mu}_{x0}$ are the number of unmatched workers, and $\hat{\mu}_{0y}$ are the number of unmatched firms, the log likelihood of the sample of matches is:

$$l(D|\theta) := \sum_{(x,y)\in[X\times Y]}\hat{\mu}_{xy}\log(\pi_{xy}^{\theta}) + \sum_{x\in[X]}\hat{\mu}_{x0}\log(\pi_{x0}^{\theta}) + \sum_{y\in[Y]}\hat{\mu}_{0y}\log(\pi_{0y}^{\theta})$$

The maximum likelihood estimator of θ is given by $\max_{\theta\in\Theta} l(D|\theta)$. Denoting $\hat{N} := \sum_{(x,y)\in[X\times Y]}\hat{\mu}_{xy} + \sum_{x\in[X]}\hat{\mu}_{x0} + \sum_{y\in[Y]}\hat{\mu}_{0y}$, the log likelihood can be equivalently expressed as:

$$l(D|\theta) = \sum_{(x,y)\in[X\times Y]}\hat{\mu}_{xy}\log(\mu_{xy}^{\theta}) + \sum_{x\in[X]}\mu_{x0}^{\theta}\log(\mu_{x0}^{\theta}) + \sum_{y\in[Y]}\mu_{0y}^{\theta}\log(\mu_{0y}^{\theta}) - \hat{N}\log(N^{\theta})$$

The first order conditions of the MLE problem become:

$$\sum_{(x,y)\in[X\times Y]}\hat{\mu}_{xy}\frac{\partial\log(\mu_{xy}^{\theta})}{\partial\theta_p} + \sum_{x\in[X]}\hat{\mu}_{x0}\frac{\partial\log(\mu_{x0}^{\theta})}{\partial\theta_p} + \sum_{y\in[Y]}\hat{\mu}_{0y}\frac{\partial\log(\mu_{0y}^{\theta})}{\partial\theta_p} = \frac{\hat{N}}{N^{\theta}}\frac{\partial N^{\theta}}{\partial\theta_p}$$

Note that by the definition of N^{θ}, and after rearranging, we can transform the first order condition to:

$$\sum_{(x,y)\in[X\times Y]} \frac{\hat{\mu}_{xy}}{\hat{N}} \frac{\partial \log(\mu^{\theta}_{xy})}{\partial \theta_p} + \sum_{x\in[X]} \frac{\hat{\mu}_{x0}}{\hat{N}} \frac{\partial \log(\mu^{\theta}_{x0})}{\partial \theta_p} + \sum_{y\in[Y]} \frac{\hat{\mu}_{0y}}{\hat{N}} \frac{\partial \log(\mu^{\theta}_{0y})}{\partial \theta_p}$$
$$= \sum_{(x,y)\in[X\times Y]} \frac{\mu^{\theta}_{xy}}{N^{\theta}} \frac{\partial \log(\mu^{\theta}_{xy})}{\partial \theta_p} + \sum_{x\in[X]} \frac{\mu^{\theta}_{x0}}{N^{\theta}} \frac{\partial \log(\mu^{\theta}_{x0})}{\partial \theta_p} + \sum_{y\in[Y]} \frac{\mu^{\theta}_{0y}}{N^{\theta}} \frac{\partial \log(\mu^{\theta}_{0y})}{\partial \theta_p}$$

I introduce $\hat{\pi}_{xy} := \frac{\hat{\mu}_{xy}}{\hat{N}}, \hat{\pi}_{x0} := \frac{\hat{\mu}_{x0}}{\hat{N}}$, and $\hat{\pi}_{0y} := \frac{\hat{\mu}_{0y}}{\hat{N}}$. I also introduce $\pi^{\theta}_{xy} := \frac{\mu^{\theta}_{xy}}{N^{\theta}}, \pi^{\theta}_{x0} := \frac{\mu^{\theta}_{x0}}{N^{\theta}}$, and $\pi^{\theta}_{0y} := \frac{\mu^{\theta}_{0y}}{N^{\theta}}$. Then, the first order condition can be equivalently expressed as:

$$\sum_{(x,y)\in[X\times Y]} \hat{\pi}_{xy} \frac{\partial \log(\mu^{\theta}_{xy})}{\partial \theta_p} + \sum_{x\in[X]} \hat{\pi}_{x0} \frac{\partial \log(\mu^{\theta}_{x0})}{\partial \theta_p} + \sum_{y\in[Y]} \hat{\pi}_{0y} \frac{\partial \log(\mu^{\theta}_{0y})}{\partial \theta_p}$$
$$= \sum_{(x,y)\in[X\times Y]} \pi^{\theta}_{xy} \frac{\partial \log(\mu^{\theta}_{xy})}{\partial \theta_p} + \sum_{x\in[X]} \pi^{\theta}_{x0} \frac{\partial \log(\mu^{\theta}_{x0})}{\partial \theta_p} + \sum_{y\in[Y]} \pi^{\theta}_{0y} \frac{\partial \log(\mu^{\theta}_{0y})}{\partial \theta_p}$$

From the definitions of $\mu^{\theta}_{xy}, \mu^{\theta}_{x0}, \mu^{\theta}_{0y}$ at the start of the section, we have that:

$$\log(\mu^{\theta}_{xy}) = \frac{\sum_{k\in[K]} \phi_{xyk}\lambda_k - a_x - b_y}{2}, \log(\mu^{\theta}_{x0}) = -a_x, \log(\mu^{\theta}_{0y}) = -b_x$$

As a result, we have that the first order conditions become

$$[\text{FOC } \theta_p = a_x] : \frac{1}{2} \sum_{y\in[Y]} \hat{\pi}_{xy} + \hat{\pi}_{x0} = \frac{1}{2} \sum_{y\in[Y]} \pi^{\theta}_{xy} + \pi^{\theta}_{x0}$$
$$[\text{FOC } \theta_p = b_y] : \frac{1}{2} \sum_{x\in[X]} \hat{\pi}_{xy} + \hat{\pi}_{0y} = \frac{1}{2} \sum_{x\in[X]} \pi^{\theta}_{xy} + \pi^{\theta}_{0y}$$
$$[\text{FOC } \theta_p = \lambda_k] : \frac{1}{2} \sum_{(x,y)\in[X\times Y]} \hat{\pi}_{xy}\phi_{xyk} = \frac{1}{2} \sum_{(x,y)\in[X\times Y]} \pi^{\theta}_{xy}\phi_{xyk}$$

Recall the general method moments in equation (22),

$$n_x = \hat{\mu}_{x0} + \sum_{y\in[Y]} \hat{\mu}_{xy} = \mu^{\theta}_{x0} + \sum_{y\in[Y]} \mu^{\theta}_{xy}$$
$$m_y = \hat{\mu}_{0y} + \sum_{x\in[X]} \hat{\mu}_{xy} = \mu^{\theta}_{0y} + \sum_{x\in[X]} \mu^{\theta}_{xy}$$
$$\sum_{(x,y)\in[X\times Y]} \hat{\mu}^{\theta}_{xy}\phi_{xyk} = \sum_{(x,y)\in[X\times Y]} \mu^{\theta}_{xy}\phi_{xyk}$$

The top two moments can be adjusted slightly to more precisely mimic the MLE moments more closely:

$$\frac{\hat{\mu}_{x0}}{n_x} + \sum_{y\in[Y]} \frac{\hat{\mu}_{xy}}{n_x} = \frac{\mu^{\theta}_{x0}}{n_x} + \sum_{y\in[Y]} \frac{\mu^{\theta}_{xy}}{n_x}$$
$$\frac{\hat{\mu}_{0y}}{m_y} + \sum_{x\in[X]} \frac{\hat{\mu}_{xy}}{m_y} = \frac{\mu^{\theta}_{0y}}{m_y} + \sum_{x\in[X]} \frac{\mu^{\theta}_{xy}}{m_y}$$

Notice that the moments we match in both problems are very similar except that in the case of the MLE, we weight each match equal whereas in the general method of moments we weight each individual equally. Thus, in the general method of moments the (x, y) matches receive less weight as we incorporate them twice, once in the worker moments and once in the firm moments.

To move towards the generalized linear model for the maximum likelihood estimator, let $A := X \times Y + X + Y$ so that $a \in [A]$ indexes matched workers and firms $(a = xy)$, unmatched workers $(a = x0)$, and unmatched firms $(a = 0y)$. Then, recalling the definition of π^θ from equation (25), we see that the likelihood of the matches are given by:

$$\begin{aligned} l(D|\theta) &= \sum_{a\in[A]} \hat{\mu}_a \log(\frac{\exp((R\theta)_a)}{\sum_{a'\in[A]} \exp((R\theta)_{a'})}) \\ &= \sum_{a\in[A]} \hat{\mu}_a (R\theta)_a - (\sum_{a\in[A]} \hat{\mu}_a) \log(\sum_{a'\in[A]} \exp((R\theta)_{a'})) \end{aligned} \tag{26}$$

where R is defined as in equation (24). this is precisely a logistic regression. The difference from that in equation (11), is that here we have that all alternatives/ matches have market shares across the entire population as opposed to having market shares that are indexed by individuals. For that reason, our normalization term works for all $a \in [A]$ as opposed to in equation (11) where it works for each $i \in [I]$. This can be reformulated as a Poisson regression:

$$\max_{\theta\in\mathbb{R}^K, Z\in\mathbb{R}} \sum_{a\in[A]} \hat{\mu}_a((R\theta)_a - Z) - \sum_{a\in[A]} \exp((R\theta)_a - Z)$$

By the first order condition with respect to Z, we get that

$$\begin{aligned} &\sum_{a\in[A]} \hat{\mu}_a = \sum_{a\in[A]} \exp((R\theta)_a - Z) \\ &\implies Z = \log(\sum_{a\in[A]} \exp((R\theta)_a)) - \log(\sum_{a\in[A]} \hat{\mu}_a) \end{aligned}$$

Plugging the expression for Z into the Poisson regression, I get that after dropping terms that are independent of parameters, the Poisson regression reformulates as

$$\max_{\theta\in\mathbb{R}^K} \sum_{a\in[A]} \hat{\mu}_a((R\theta)_a - \log(\sum_{a'\in[A]} \exp((R\theta)_{a'})))$$

which is precisely the same as maximizing the log likelihood of the data in equation (26).

4.5 Welfare Interpretation of the Matching Optimization Problem

Recall the equilibrium conditions from the worker's problem in section 4.1, the firm's problem in section 4.2, and the competitive equilibrium in 4.3. We had that:

$$\begin{aligned} \mu_{xy} &= \sqrt{n_x m_y} \exp(\frac{\phi_{xy} - u_x - v_y}{2}) \\ \mu_{x0} &= n_x \exp(-u_x) \\ \mu_{0y} &= m_y \exp(-v_y) \end{aligned}$$

where u_x and v_y are determined by the population equations:

$$n_x = n_x \exp(-u_x) + \sum_{y \in [Y]} \sqrt{n_x m_y} \exp(\frac{\phi_{xy} - u_x - v_y}{2})$$
$$m_y = m_y \exp(-m_y) + \sum_{x \in [X]} \sqrt{n_x m_y} \exp(\frac{\phi_{xy} - u_x - v_y}{2})$$

which arise as the first order conditions of the minimization problem

$$\begin{aligned} \min_{u \in \mathbb{R}^X, v \in \mathbb{R}^Y} \{ & \sum_{x \in [X]} n_x u_x + \sum_{y \in [Y]} m_y v_y \\ & + 2 \sum_{(x,y) \in [X \times Y]} \sqrt{n_x m_y} \exp(\frac{\phi_{xy} - u_x - v_y}{2}) \\ & + \sum_{x \in [X]} n_x \exp(-u_x) + \sum_{y \in [Y]} m_y \exp(-v_y) \} \end{aligned} \tag{27}$$

which is identical to the maximization problem in equation (21) recalling that $a_x := u_x - \log(n_x), b_x := v_y - \log(m_y)$. We claim that this quantity in equation (27) is equal to its dual:

$$\begin{aligned} \max_{\mu \in \mathbb{R}^{X \times Y}} \{ & \sum_{(x,y) \in [X \times Y]} \mu_{xy} \phi_{xy} \\ & - \sum_{x \in [X]} n_x (\sum_{y \in [Y]} \frac{\mu_{xy}}{n_x} \log(\frac{\mu_{xy}}{n_x}) + \frac{\mu_{x0}}{n_x} \log(\frac{\mu_{x0}}{n_x})) \\ & - \sum_{y \in [Y]} m_y (\sum_{x \in [X]} \frac{\mu_{xy}}{m_y} \log(\frac{\mu_{xy}}{m_y}) + \frac{\mu_{0y}}{m_y} \log(\frac{\mu_{0y}}{m_y})) \} \end{aligned} \tag{28}$$

4.5.1 Some Notation and Intuition for the Claim

Recall from equations (6) and (7), where we define $G(U) := \mathbb{E}[\max_{y \in [0:Y]} \{U_y + e_y, e_0\}]$ as the welfare function, and $G^*(\pi) := \max_{U \in \mathbb{R}^Y} \sum_{y \in [Y]} \pi_y U_y - G(U)$ as the (negative) generalized entropy of choice, we have $G(U) = \max_{\pi \in \Delta^Y} \sum_{y \in [Y]} \pi_y U_y - G^*(\pi)$. We also had that $\frac{\partial G(U)}{\partial U_y} = \pi_y$ by the DZW theorem in section 1.1 and $\frac{\partial G^*(\pi)}{\pi_y} = U_y$, by the envelope theorem.

Adopting this notation to the matching setting, define:

$$G_x(U) := \mathbb{E}_{\mathcal{P}_x}[\max_{y \in [Y]} \{U_{xy} + e_y, e_0\}]$$
$$G_x^*(\pi) := \max_{(U_{xy})_y \in \mathbb{R}^Y} \sum_{y \in [Y]} \pi_{xy} U_{xy} - G_x((U_{xy})_y)$$
$$H_y(V) := \mathbb{E}_{Q_y}[\max_{x \in [X]} \{V_{xy} + \eta_x, \eta_0\}]$$
$$H_y^*(\pi) := \max_{(U_{xy})_x \in \mathbb{R}^X} \sum_{x \in [X]} \pi_{xy} U_{xy} - G_y((V_{xy})_x)$$

where $G_x(U)$ is the welfare for workers of type x, $G_x^*(\pi)$ is the (negative) generalized entropy of choice for workers of type x, $H_y(V)$ is the welfare for firms of type y, and $H_y^*(\pi)$ is the (negative) generalized entropy of choice for firms of type y.

Then, we can apply the results to have that:

$$\begin{aligned}\frac{\partial G_x(U)}{\partial U_{xy}} &= \pi_{xy} \\ \frac{\partial G_x^*(\pi)}{\partial \pi_{xy}} &= U_{xy} \\ \frac{\partial H_y(V)}{\partial V_{xy}} &= \pi_{yx} \\ \frac{\partial H_y^*(\pi)}{\partial \pi_{yx}} &= V_{xy}\end{aligned}$$

Recall that we defined the systematic component of the welfare for workers of type x matching with firms of type y to be $U_{xy} := \alpha_{xy} + w_{xy}$ and analogously for firms of type y matching with workers of type x: $V_{xy} := \gamma_{xy} - w_{xy}$. Let's introduce the total welfare of workers to be,

$$\begin{aligned}G_T(U) &:= \sum_{x\in[X]} n_x G_x(U) \\ \implies \frac{\partial G_T(U)}{U_{xy}} &= n_x \frac{\partial G_x(U)}{\partial U_{xy}} = n_x \pi_{xy} = \mu_{xy}\end{aligned}$$

And the total (negative) generalized entropy of choice for workers to be

$$\begin{aligned}G_T^*(\mu) :&= \max_{U\in\mathbb{R}^{X\times Y}} \sum_{(x,y)\in[X\times Y]} \mu_{xy}U_{xy} - G_T(U) \\ &= \sum_{x\in[X]} n_x G_x^*((\frac{\mu_{xy}}{n_x})_y)\end{aligned}$$

We then have that $\frac{\partial G_T^*(\mu)}{\partial \mu_{xy}} = \frac{\partial G_x^*((\mu_{xy}/n_x)_y)}{\partial \pi_{xy}} = U_{xy}$.

We can similarly define the aggregate welfare of firms and the total (negative) generalized entropy of choice to be:

$$\begin{aligned}H_T(V) :&= \sum_{y\in[Y]} m_y H_y(V) \\ \implies \frac{\partial H_T(V)}{V_{xy}} &= m_y \frac{\partial H_y(V)}{\partial V_{xy}} = m_y \pi_{yx} = \mu_{xy} \\ H_T^*(\mu) :&= \max_{V\in\mathbb{R}^{X\times Y}} \sum_{(x,y)\in[X\times Y]} \mu_{xy}V_{xy} - H_T(V) \\ &= \sum_{y\in[Y]} m_y H_y^*((\frac{\mu_{xy}}{m_y})_x)\end{aligned}$$

We then also have that $\frac{\partial H_T^*(\mu)}{\partial \mu_{xy}} = \frac{\partial H_y^*((\mu_{xy}/m_y)_x)}{\partial \pi_{xy}} = V_{xy}$.

4.5.2 General Primal Problem for Matching with Transferable Utility

Recall that $\frac{\partial G_T(U)}{U_{xy}} = \mu_{xy}$ and $\frac{\partial H_T(V)}{V_{xy}} = \mu_{xy}$. At equilibrium, these two quantities coincide so that $\frac{\partial G_T(U)}{U_{xy}} = \frac{\partial H_T(V)}{V_{xy}}$. Again, we have by definition that $U_{xy} + V_{xy} = \phi_{xy} \implies \frac{\partial G_T(U)}{\partial U_{xy}} = \frac{\partial H_T(\phi - U)}{\partial V_{xy}}$, which is the first order condition associated with

$$\begin{aligned} &\min_{U \in \mathbb{R}^{X \times Y}} G_T(U) + H_T(\phi - U) \\ \Leftrightarrow &\min_{U \in \mathbb{R}^{X \times Y}, V \in \mathbb{R}^{X \times Y}} G_T(U) + H_T(V) \\ &\text{s.t. } U_{xy} + V_{xy} = \phi_{xy} \forall (x,y) \in [X \times Y] \end{aligned}$$

which is the general primal problem for matching with transferable utility. If we use the results in sections 1.3.1 and 4.5.1, I have that the primal problem in the logit case is:

$$\begin{aligned} &\min_{U \in \mathbb{R}^{X \times Y}, V \in \mathbb{R}^{X \times Y}} \sum_{x \in [X]} n_x \log(1 + \sum_{y \in [Y]} \exp(U_{xy})) + \sum_{y \in [Y]} m_y \log(1 + \sum_{x \in [X]} \exp(V_{xy})) \\ &\text{s.t. } U_{xy} + V_{xy} = \phi_{xy} \forall (x,y) \in [X \times Y] \end{aligned} \tag{29}$$

4.5.3 General Dual Problem for Matching with Transferable Utility

Recall that we defined $\phi_{xy} := U_{xy} + V_{xy}$, which implies that $\frac{\partial G_T^*(\mu)}{\partial \mu_{xy}} + \frac{\partial H_T^*(\mu)}{\partial \mu_{xy}} = \phi_{xy}$ so that we can interpret this as the first order condition associated with the optimization problem:

$$\max_{\mu \in \mathbb{R}^{X \times Y}} \sum_{(x,y) \in [X \times Y]} \mu_{xy} \phi_{xy} - G_T^*(\mu) - H_T^*(\mu) \tag{30}$$

which is the general dual problem for matching with transferable utility. In the case of the logit model, using the result in sections 1.5.1 and 4.5.1, we have

$$\begin{aligned} G_T^*(\mu) &= \sum_{x \in [X]} n_x (\sum_{y \in [Y]} \frac{\mu_{xy}}{n_x} \log(\frac{\mu_{xy}}{n_x}) + \frac{\mu_{x0}}{n_x} \log(\frac{\mu_{x0}}{n_x})) \\ H_T^*(\mu) &= \sum_{y \in [Y]} m_y (\sum_{x \in [X]} \frac{\mu_{xy}}{m_y} \log(\frac{\mu_{xy}}{m_y}) + \frac{\mu_{0y}}{m_y} \log(\frac{\mu_{0y}}{m_y})) \end{aligned}$$

If we plug these two equations into the maximization problem in equation (30), we in fact recover the desired dual maximization problem in equation (28).

4.5.4 Proof of Claim

To complete the proof of the claim, I would like to show that the primal problem in equation (29) is indeed equivalent to the primal problem in equation (27). To that end, start with the problem in equation (29), repeated below:

$$\begin{aligned} &\min_{U \in \mathbb{R}^{X \times Y}, V \in \mathbb{R}^{X \times Y}} \sum_{x \in [X]} n_x \log(1 + \sum_{y \in [Y]} \exp(U_{xy})) + \sum_{y \in [Y]} m_y \log(1 + \sum_{x \in [X]} \exp(V_{xy})) \\ &\text{s.t. } U_{xy} + V_{xy} = \phi_{xy} \forall (x,y) \in [X \times Y] \end{aligned}$$

Introduce extra variables $u_x := \log(1 + \sum_{y\in[Y]} \exp(U_{xy})$ and $v_y := \log(1 + \sum_{x\in[X]} \exp(V_{xy})$ and re-express the problem as

$$\begin{aligned}\min_{U\in\mathbb{R}^{X\times Y}, u\in\mathbb{R}^X, v\in\mathbb{R}^Y} & \sum_{x\in[X]} n_x u_x + \sum_{y\in[Y]} m_y v_y \\ \text{s.t. } u_x &= \log(1 + \sum_{y\in[Y]} \exp(U_{xy}) \forall x \in [X] \\ v_y &= \log(1 + \sum_{x\in[X]} \exp(\phi_{xy} - U_{xy}) \forall y \in [Y]\end{aligned}$$

which can be re-expressed as

$$\begin{aligned}\min_{U\in\mathbb{R}^{X\times Y}, u\in\mathbb{R}^X, v\in\mathbb{R}^Y} & \sum_{x\in[X]} n_x u_x + \sum_{y\in[Y]} m_y v_y \\ \text{s.t. } 1 &= \exp(-u_x) + \sum_{y\in[Y]} \exp(U_{xy} - u_x) \forall x \in [X] \\ 1 &= \exp(-v_y) + \sum_{x\in[X]} \exp(\phi_{xy} - U_{xy} - v_y) \forall y \in [Y]\end{aligned}$$

This problem can be reformulated as a min-max problem, after introducing some variables N_x, M_y. The idea is that if the outer minimizer doesn't set the constraints above equal to each other, then the inner minimizer will send the problem to infinity.

$$\begin{aligned}\min_{U\in\mathbb{R}^{X\times Y}, u\in\mathbb{R}^X, v\in\mathbb{R}^Y} \max_{N\in\mathbb{R}^X, M\in\mathbb{R}^Y} \{ & \sum_{x\in[X]} n_x u_x + \sum_{y\in[Y]} m_y v_y \\ & + \sum_{(x,y)\in[X\times Y]} N_x \exp(U_{xy} - u_x) + \sum_{y\in[Y]} M_y \exp(\phi_{xy} - U_{xy} - v_y) \\ & + \sum_{x\in[X]} N_x \exp(-u_x) + \sum_{y\in[Y]} M_y \exp(-v_y) \\ & - \sum_{x\in[X]} N_x - \sum_{y\in[Y]} M_y \}\end{aligned}$$

waving my hands a little bit and swapping the min and max, the problem becomes:

$$\begin{aligned}\max_{N\in\mathbb{R}^X, M\in\mathbb{R}^Y} \min_{U\in\mathbb{R}^{X\times Y}, u\in\mathbb{R}^X, v\in\mathbb{R}^Y} \{ & \sum_{x\in[X]} n_x u_x + \sum_{y\in[Y]} m_y v_y \\ & + \sum_{(x,y)\in[X\times Y]} N_x \exp(U_{xy} - u_x) + \sum_{(x,y)\in[X\times Y]} M_y \exp(\phi_{xy} - U_{xy} - v_y) \\ & + \sum_{x\in[X]} N_x \exp(-u_x) + \sum_{y\in[Y]} M_y \exp(-v_y) \\ & - \sum_{x\in[X]} N_x - \sum_{y\in[Y]} M_y \}\end{aligned}$$

The first order condition of the inner minimization problem, with respect to U_{xy} is $N_x \exp(U_{xy} - u_x) = M_y \exp(\phi_{xy} - U_{xy} - v_y)$. Then, the square root of the geometric mean of both sides is equal to each side so that:

$$N_x \exp(U_{xy} - u_x) + M_y \exp(\phi_{xy} - U_{xy} - v_y) = 2\sqrt{N_x M_y \exp(\phi_{xy} - u_x - v_y)}$$

The max-min problem above can then be simplified to:

$$\begin{aligned}\max_{N \in \mathbb{R}^X, M \in \mathbb{R}^Y} \min_{u \in \mathbb{R}^X, v \in \mathbb{R}^Y} \{ & \sum_{x \in [X]} n_x u_x + \sum_{y \in [Y]} m_y v_y \\ & + 2 \sum_{(x,y) \in [X \times Y]} \sqrt{N_x M_y \exp(\phi_{xy} - u_x - v_y)} \\ & + \sum_{x \in [X]} N_x \exp(-u_x) + \sum_{y \in [Y]} M_y \exp(-v_y) \\ & - \sum_{x \in [X]} N_x - \sum_{y \in [Y]} M_y \}\end{aligned}$$

The first order conditions of the inner minimization problem with respect to u_x and v_y are

$$\begin{aligned} n_x &= N_x \exp(-u_x) + \sum_{y \in [Y]} \sqrt{N_x M_y \exp(\phi_{xy} - u_x - v_y)} \\ m_y &= M_y \exp(-v_y) + \sum_{x \in [X]} \sqrt{N_x M_y \exp(\phi_{xy} - u_x - v_y)} \end{aligned}$$

Waving our hands once more, let's flip the order of the max and min:

$$\begin{aligned}\min_{u \in \mathbb{R}^X, v \in \mathbb{R}^Y} \max_{N \in \mathbb{R}^X, M \in \mathbb{R}^Y} \{ & \sum_{x \in [X]} n_x u_x + \sum_{y \in [Y]} m_y v_y \\ & + 2 \sum_{(x,y) \in [X \times Y]} \sqrt{N_x M_y \exp(\phi_{xy} - u_x - v_y)} \\ & + \sum_{x \in [X]} N_x \exp(-u_x) + \sum_{y \in [Y]} M_y \exp(-v_y) \\ & - \sum_{x \in [X]} N_x - \sum_{y \in [Y]} M_y \}\end{aligned}$$

The first order conditions of the inner maximization with respect to N_x and M_y are

$$\begin{aligned} 1 &= \exp(-u_x) + \sum_{y \in [Y]} \sqrt{\frac{M_y}{N_x} \exp(\phi_{xy} - u_x - v_y)} \\ 1 &= \exp(-v_y) + \sum_{x \in [X]} \sqrt{\frac{N_x}{M_y} \exp(\phi_{xy} - u_x - v_y)} \end{aligned}$$

Multiplying through the top equation by N_x and the bottom equation by M_y:

$$N_x = N_x \exp(-u_x) + \sum_{y \in [Y]} \sqrt{N_x M_y \exp(\phi_{xy} - u_x - v_y)}$$
$$M_y = M_y \exp(-v_y) + \sum_{x \in [X]} \sqrt{N_x M_y \exp(\phi_{xy} - u_x - v_y)}$$

Comparing these equations to those for the first order conditions with respect to u_x and v_y, one sees that $N_x = n_x$ and $M_y = m_y$. As a result, one can write the min-max problem as

$$\begin{aligned}\min_{u \in \mathbb{R}^X, v \in \mathbb{R}^Y} \{ & \sum_{x \in [X]} n_x u_x + \sum_{y \in [Y]} m_y v_y \\ & + 2 \sum_{(x,y) \in [X \times Y]} \sqrt{n_x m_y \exp(\phi_{xy} - u_x - v_y)} \\ & + \sum_{x \in [X]} n_x \exp(-u_x) + \sum_{y \in [Y]} m_y \exp(-v_y) \\ & - \sum_{x \in [X]} n_x - \sum_{y \in [Y]} m_y \}\end{aligned}$$

Or equivalently, dropping inconsequential terms:

$$\begin{aligned}\min_{u \in \mathbb{R}^X, v \in \mathbb{R}^Y} \{ & \sum_{x \in [X]} n_x u_x + \sum_{y \in [Y]} m_y v_y \\ & + 2 \sum_{(x,y) \in [X \times Y]} \sqrt{n_x m_y \exp(\phi_{xy} - u_x - v_y)} \\ & + \sum_{x \in [X]} n_x \exp(-u_x) + \sum_{y \in [Y]} m_y \exp(-v_y) \}\end{aligned}$$

which exactly matches the problem in equation (27) as desired.

5 Discrete Choice Asymptotics

In this section, I aim to derive asymptotics for various of the estimators we've considered.

5.1 Possible Data Generating Processes

In the setting of discrete processes, it's sometimes subtle what it means to take the size of the sample to infinity. To derive asymptotics, we must posit a data generating process.

(i) Consider the simplest utility model covered in section 1. In this section, we specified that the utility of a choice $y \in [0 : Y]$ for an individual $i \in [I]$ is given by $U_y + \epsilon_{iy}$, where $U_0 = 0$, so that the systematic part of the utility is independent of the individual. In a sample of data, we observe market shares $\hat{\pi} \in \Delta^{Y+1}$. In this case, we can send $I \to \infty$ and derive a distribution for $\hat{\pi}$ and in turn for $\hat{U}$.

(ii) In the macro logistic regression, we specified that the utility for an individual $i \in [I]$ of type $x \in [X]$ of picking some choice $y \in [0 : Y]$ is given by $\sum_{k \in [K]} \phi_{xyk} \lambda_{k \epsilon i y}$, where $\phi_{x0k} = 0 \; \forall (x, k) \in [X \times K]$. Again, the assumption

here is that the econometrician cannot distinguish between $i, j \in [X]$. In a sample of data, we observe market shares $\hat{\pi} \in \Pi_{x\in[X]}\Delta^{Y+1}$. Again, here, we can send $I \to \infty$ and derive an asymptotic distribution for $\hat{\pi}$ and in turn for $\hat{\lambda}$.

(iii) In the matching with transferable utility model, we said that if a worker $i \in [I]$ of type $x \in [X]$ matches with a firm $j \in [J]$ of type $y \in [0 : Y]$, then i gets utility $\alpha_{xy} + w_{xy} + \epsilon_{iy}$ where $\alpha_{x0} = 0, w_{x0} = 0$, and where a worker picking firm of type $y = 0$ means they choose to stay unemployed. Analogously, if a firm $j \in [J]$ of type $y \in [Y]$ matches with a worker $i \in [I]$ of type $x \in [0 : X]$ then j gets utility $\gamma_{xy} - w_{xy} + \eta_j x$ where $\gamma_{0y} = 0, w_{0y} = 0$, and where a firm picking a worker of type $x = 0$ means they choose to not hire anybody. In this setting, in a sample, we observe a vector of "market shares" $\hat{\pi} \in [0 : X] \times [0 : Y] \setminus \{(0, 0)\}$. In this setting, the market shares are interpreted as the fraction 'matches' or 'non-matches' to total 'matches' and 'non-matches'. We send the number of 'matches' or 'non-matches' made to infinity to derive an asymptotic distribution for $\hat{\pi}$. Recall that we ended up parametrizing $\alpha_{xy} + \gamma_{xy} = \phi_{xy} = \sum_{k\in[K]} \phi_{xyk}\lambda_k$ and we had additional firm and worker fixed effects a_x and b_y, respectively. From there, we defined $\theta = [\lambda, a, b]$. After we derive the asymptotic distribution for $\hat{\pi}$, we can also derive it for $\hat{\theta}$.

5.2 MLE Asymptotic Distribution Derivation

For all of these settings, let $a \in [A]$ index $\hat{\pi}$, the design matrix be $R \in \mathbb{R}^{A\times K}$, and the vector of parameters be $\theta \in \mathbb{R}^K$. We estimate $\hat{\theta}$ from the data D of N matches that has 'market shares' $\hat{\pi}$ as

$$\begin{aligned} \max_{\theta\in\Theta} l(D|\theta) &= \max_{\theta\in\Theta} \sum_{a\in[A]} \hat{\pi}_a \log(f_a(\theta)) \\ &= \max_{\theta\in\Theta} \sum_{a\in[A]} \hat{\pi}_a \log\left(\frac{\exp((R\theta)_a)}{\sum_{a'\in[A]}\exp((R\theta)_{a'})}\right) \end{aligned} \tag{31}$$

As a first step in our derivation, we wish to compute $\text{Cov}(\hat{\pi}_a, \hat{\pi}_{a'})$.

$$\begin{aligned} \text{Cov}(\hat{\pi}_a, \hat{\pi}_{a'}) &= \text{Cov}(\frac{1}{N}\sum_{s\in[A]} \mathbb{1}_{\{a_s=a\}}, \frac{1}{N}\sum_{t\in[A]} \mathbb{1}_{\{a_t=a'\}}) \\ &= \frac{1}{N^2}\sum_{(s,t)\in[A\times A]} \text{Cov}(\mathbb{1}_{\{a_s=a\}}, \mathbb{1}_{\{a_t=a'\}}) \\ &= \frac{1}{N^2}\sum_{s\in[A]} \text{Cov}(\mathbb{1}_{\{a_s=a\}}, \mathbb{1}_{\{a_s=a'\}}) \\ &= \frac{1}{N^2}\sum_{s\in[A]} \mathbb{E}[\mathbb{1}_{\{a_s=a\}}\mathbb{1}_{\{a_s=a'\}}] - \mathbb{E}[\mathbb{1}_{\{a_s=a\}}]\mathbb{E}[\mathbb{1}_{\{a_s=a'\}}] \\ &= \frac{1}{N}(\mathbb{1}_{\{a=a'\}}\pi_a - \pi_a\pi_{a'}) \end{aligned}$$

That implies $\text{Cov}(\hat{\pi}) = \frac{1}{N}(diag(\pi) - \pi\pi')$. As a result, by the central limit theorem, $\sqrt{N}(\hat{\pi} - \pi) \xrightarrow{d} \mathcal{N}(0, diag(\pi) - \pi\pi')$. The first order condition of the problem in equation (31) is

$$\sum_{a\in[A]} \hat{\pi}_a \frac{\partial \log(f_a(\hat{\theta}))}{\partial\theta} = 0$$

In the population, assuming correct specification, we also have that

$$\sum_{a\in[A]} \pi_a \frac{\partial \log(f_a(\theta))}{\partial \theta} = 0$$

Writing[3] $\hat{\pi}_a := \pi_a + \delta^\pi \odot \hat{\pi}_a$ and $\hat{\theta} := \theta + \delta^\theta \odot \hat{\theta}$ and taking a first order approximation of $F(\hat{\pi}, \hat{\theta}) := \sum_{a\in[A]} \hat{\pi}_a \frac{\partial \log(f_a(\hat{\theta}))}{\partial \theta}$ around (π, θ), I get that:

$$\begin{aligned}
&\sum_{a\in[A]} \hat{\pi}_a \frac{\partial \log(f_a(\hat{\theta}))}{\partial \theta} \approx \sum_{a\in[A]} \pi_a \frac{\partial \log(f_a(\theta))}{\partial \theta} + \sum_{a\in[A]} \delta^\pi_a \hat{\pi}_a \frac{\partial \log(f_a(\theta))}{\partial \theta} + \sum_{a\in[A]} \pi_a \frac{\partial^2 \log(f_a(\theta))}{\partial \theta \partial \theta'} \delta^\theta \odot \hat{\theta} \\
\implies &\sum_{a\in[A]} \delta^\pi_a \hat{\pi}_a \frac{\partial \log(f_a(\theta))}{\partial \theta} \approx -\sum_{a\in[A]} \pi_a \frac{\partial^2 \log(f_a(\theta))}{\partial \theta \partial \theta'} \delta^\theta \odot \hat{\theta} \\
\implies &\delta^\theta \odot \hat{\theta} \approx \left(-\sum_{a\in[A]} \pi_a \frac{\partial^2 \log(f_a(\theta))}{\partial \theta \partial \theta'} \right)^{-1} \left(\sum_{a\in[A]} \delta^\pi_a \hat{\pi}_a \frac{\partial \log(f_a(\theta))}{\partial \theta} \right)
\end{aligned}$$

Denoting $V := \text{avar}(\sqrt{N}(\sum_{a\in[A]} \delta^\pi_a \hat{\pi}_a \frac{\partial \log(f_a(\theta))}{\partial \theta}))$ and $F := \mathbb{E}[-\frac{\partial^2 \log(f_a(\theta))}{\partial \theta \partial \theta'}] = -\sum_{a\in[A]} \pi_a \frac{\partial^2 \log(f_a(\theta))}{\partial \theta \partial \theta'}$, by the delta method, $\sqrt{N}\delta^\theta \odot \hat{\theta} \xrightarrow{d} \mathcal{N}(0, F^{-1}VF^{-1})$. In this case, by the information matrix equality, $V = F$ so that the asymptotic distribution simplifies to, $\sqrt{N}\delta^\theta \odot \hat{\theta} \xrightarrow{d} \mathcal{N}(0, F^{-1})$.

5.2.1 Information Equality Demonstration

To see the information equality result, first note that

$$\begin{aligned}
\frac{\partial \log(f_a(\theta))}{\partial \theta} &= R_a - \sum_{a'\in[A]} R_{a'} f_{a'}(\theta) \\
\implies \sum_{a\in[A]} \delta^\pi_a \hat{\pi}_a \frac{\partial \log(f_a(\theta))}{\partial \theta} &= R' \delta^\pi \odot \hat{\pi}
\end{aligned}$$

Next, using the result from just above

$$\begin{aligned}
\frac{\partial^2 \log(f_a(\theta))}{\partial \theta \partial \theta'} &= \frac{\partial (R_a - \sum_{a'\in[A]} R_{a'} f_{a'}(\theta))}{\partial \theta'} \\
&= -\sum_{a'\in[A]} R_{a'} \frac{\partial (f_{a'}(\theta))}{\partial \theta'} \\
&= -\sum_{a'\in[A]} R_{a'} \left(\frac{(\sum_{a''\in[A]} \exp((R\theta)_{a''})) \exp((R\theta)_{a'}) R'_{a'} - \exp((R\theta)_{a'})(\sum_{a''\in[A]} \exp((R\theta)_{a''}) R'_{a''})}{(\sum_{a''\in[A]} \exp((R\theta)_{a''}))^2} \right) \\
&= -\sum_{a'\in[A]} R_{a'} \left(f_{a'}(\theta) R'_{a'} - f_{a'}(\theta) \sum_{a''\in[A]} f_{a''}(\theta) R'_{a''} \right) \\
&= -\sum_{a'\in[A]} R_{a'} \left(\pi_{a'} R'_{a'} - \pi_{a'} \sum_{a''\in[A]} \pi_{a''} R'_{a''} \right) \\
&= -R'(diag(\pi) - \pi\pi')R
\end{aligned}$$

[3] I use the symbol $\odot$ to mean the component wise multiplication of the vectors.

where in the second to last step, I use the idea that under correct specification, $\pi_a = f_a(\theta)$.

That implies that

$$-\sum_{a\in[A]} \pi_a \frac{\partial^2 \log(f_a(\theta))}{\partial\theta\partial\theta'} = R'(diag(\pi) - \pi\pi')R$$

So putting this together, what did we find? For one,

$$\begin{aligned} V &= \text{avar}(\sqrt{N}(\sum_{a\in[A]} \delta_a^\pi \hat{\pi}_a \frac{\partial \log(f_a(\theta))}{\partial\theta})) \\ &= \text{avar}(\sqrt{N}R'\delta^\pi \odot \hat{\pi}) \\ &= R'(diag(\pi) - \pi\pi')R \end{aligned}$$

For two, $F = \mathbb{E}[-\frac{\partial^2 \log(f_a(\theta))}{\partial\theta\partial\theta'}] = R'(diag(\pi) - \pi\pi')R$. So that $F = V$ showing the information matrix equality. □

5.3 Weighted Poisson Regression Asymptotics

Recall the weighted Poisson Regression estimator from equation (23). From that equation, there are a few of changes: I relabel Ω to A, I use the fact from section 5.3.1 to transform the sample count μ_ω into the sample frequency $\hat{\pi}_a$, and I multiply the objective by minus one to change the problem from a maximization problem to a minimization problem.

$$\max_{\theta\in\Theta} \sum_{a\in[A]} w_a \left[\hat{\pi}_a (R\theta)_a - \exp((R\theta)_a)\right]$$

where $\hat{\pi}_a$ is the sample frequency of type a and π_a is the population frequency of type a. After setting $\pi_a^\theta := \exp((R\theta)_a)$, we have the first order sufficient and necessary conditions

$$\sum_{a\in[A]} w_a(\hat{\pi}_a - \pi_a^{\hat{\theta}})R_{ap} = 0 \; \forall p \in [P]$$

That is, $\hat{\theta}$ is obtained by

$$R'\Delta_w \pi^{\hat{\theta}} = R'\Delta_w \hat{\pi} \tag{32}$$

where $\Delta_w = diag(w)$. In the population, assuming correct specification, we have that

$$R'\Delta_w \pi^{\theta} = R'\Delta_w \pi \tag{33}$$

Note that linearizing around $\hat{\theta} = \theta$, we have that

$$\begin{aligned}\pi_a^{\hat{\theta}} &= \exp((R\hat{\theta})_a) \\ &\approx \pi_a^{\theta} + \pi_a^{\theta} R(\hat{\theta} - \theta)\end{aligned}$$

Or in matrix form $\pi^{\hat{\theta}} \approx \pi^{\theta} + \Delta_{\pi^\theta} R'(\hat{\theta} - \theta)$ where $\Delta_{\pi^\theta} = diag(\pi^\theta)$. Plugging in this approximation into equation (32) and then subtracting equation (33) from both sides, we have that

$$\begin{aligned}
R'\Delta_w \pi^{\hat{\theta}} &= R'\Delta_w \hat{\pi} \\
\iff R'\Delta_w \pi^{\hat{\theta}} &\approx R'\Delta_w(\pi^\theta + \Delta_{\pi^\theta} R(\hat{\theta} - \theta)) \\
\iff R'\Delta_w(\pi^\theta + \Delta_{\pi^\theta} R(\hat{\theta} - \theta)) &\approx R'\Delta_w \hat{\pi} \\
\iff R'\Delta_w(\pi^\theta + \Delta_{\pi^\theta} R(\hat{\theta} - \theta)) - R'\Delta_w \pi^\theta &\approx R'\Delta_w \hat{\pi} - R'\Delta_w \pi \\
\iff R'\Delta_w \Delta_{\pi^\theta} R(\hat{\theta} - \theta) &\approx R'\Delta_w(\hat{\pi} - \pi) \\
\iff \hat{\theta} - \theta &\approx (R'\Delta_w \Delta_{\pi^\theta} R)^{-1} R'\Delta_w(\hat{\pi} - \pi)
\end{aligned}$$

From there, we deduce that

$$\sqrt{N}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, V)$$

where $V = (R'\Delta_w \Delta_{\pi^\theta} R)^{-1} R'\Delta_w V_\pi \Delta_w R (R'\Delta_w \Delta_{\pi^\theta} R)^{-1}$ and $V_\pi = \Delta_\pi - \pi\pi'$ as computed in section 5.2.

5.3.1 Poisson Regression Homogeneity

The Weighted Poisson regression estimator of equation (23) can be written as

$$\max_{\theta \in \Theta} \hat{\mu}' \Delta_w R\theta - w' \exp(R\theta) \tag{34}$$

where w is a vector of weights, $\Delta_w = diag(w)$, and $R \in \mathbb{R}^{A \times P}$ (assume that R has linearly independent columns). We say that the poisson regression is *homogenous of degree 1* if $1 \in Im(R)$. That means there exists a unique vector $v \in \mathbb{R}^P$ such that $Rv = 1$. In this case, we have that

$$\begin{aligned}\exp(R(\theta + cv)) &= \exp(R\theta + c1_A) \\ &= \exp(c)\exp(R\theta)\end{aligned}$$

Let $I := \text{sum}(\hat{\mu})$ and define $\hat{\pi} := \frac{\hat{\mu}}{I}$. If the Poisson regression is homogenous of degree 1 with vector v, We can equivalently write the problem in equation (34) as

$$\begin{aligned}
&\max_{\theta \in \Theta} \hat{\pi}' \Delta_w R\theta - w' \exp(R\theta - \log(I)) \\
\iff &\max_{\theta \in \Theta} \hat{\pi}' \Delta_w R\theta - w' \exp\left(R\left[\theta - \log(I)v\right]\right)
\end{aligned} \tag{35}$$

which means that if $\hat{\theta}$ solves the problem in equation (34) then $\hat{\theta} - \log(I)v$ solves the problem in equation (35).

This result is useful for deriving asymptotics for the Poisson regression since the quantity $\hat{\mu}$ will blow up as the sample size increases but the quantity $\hat{\pi}$ will not. We note that the R matrix in equation (24) of the transferable utility model does indeed satisfy the homogeneity property since for $v = [0'_{dim(\lambda)}, -1'_X, -1'_Y]'$, $Rv = 1_A$.

6 DYNAMIC DISCRETE CHOICE MODELS

Through this point, we considered static discrete choice models. In other words, a decision maker made a one-time choice based on a payoff associated with each option. In many economic settings, a decision maker needs to make a series of choices where choices made in one period affect the options available in later periods as well as the payoffs associated with these options. The basic idea will be to define a state-space model in which a decision maker find themselves at the beginning of each period and the Markovian transitions between states. The stochastic nature of transitions makes the decision maker need to consider the expected value[4] of the state in the next period conditional on the decision made this period.

6.1 Finite Time Horizon and No Heterogeneity

We assume that each unit of the population can exist in one of X possible states over T periods of time. At initial time $t = 1$, there's a mass N_x of units in state $x \in [X]$ and we let n_{tx} be the mass of units that are in state $x \in [X]$ at time $t \in [T]$ so that the initial condition is $n_{1x} = N_x$ and n_{tx} for $t \geq 2$ are endogenous for each $x \in [X]$. We define $y \in [Y]$ as the set of decisions that can be applied to units. Then, we let μ_{txy} be the mass of units in state $x \in [X]$ at time $t \in [T]$ which make choice $y \in [Y]$ and we will consider this the *vector policy variable*. As there's one decision made per unit, we have the accounting condition $n_{tx} = \sum_{y \in [Y]} \mu_{txy}$ $\forall (t, x) \in [T \times X]$.

We can the stack the elements of μ_{txy} into a single column vector $\mu \in \mathbb{R}_+^{TXY}$, with entries in the order specified in the dimension, so that $n = (I_T \otimes I_X \otimes 1_Y^T)\mu$ reports the total mass of units in each state at each time.

The decisions made can undergo constraints such as the limit on the total number of units making decision y so that $A\mu \leq c$ where $A \in \mathbb{R}^{L \times TXY}$ and $c \in \mathbb{R}^L$.

Let $P_{x'xy}$ for $x, x' \in [X]$ and $y \in [Y]$ be the probability of transitioning from state x to state x' if option y is chosen. Then one has that $n_{tx'} = \sum_{(x,y) \in [X \times Y]} P_{x'xy} \mu_{(t-1)xy}$ $\forall t \in [2 : T]$.

One can then substitute out n to obtain the *evolution equations* of μ

$$\sum_{y \in [Y]} \mu_{1xy} = N_x \; \forall x \in [X]$$
$$\sum_{y \in [Y]} \mu_{txy} - \sum_{(\tilde{x}, y) \in [X \times Y]} P_{x\tilde{x}y} \mu_{(t-1)\tilde{x}y} = 0 \; \forall (t, x) \in [(2 : T) \times X]$$

This vectorizes into $(I_T \otimes I_X \otimes 1_Y^T - J_T \otimes P)\mu = e_T^1 \otimes N$ where J_T is the $T \times T$ lower shift matrix with ones on the subdiagonal and zeroes elsewhere (ie., $(J_T)_{tt'} = \mathbb{1}_{\{t=t'+1\}}$), P is the $X \times (XY)$ matrix whose (x', xy) entry is $P_{x'xy}$, N is the vector of initial conditions for individuals at each state, and e_T^1 is the first vector of the standard basis for $\mathbb{R}^T$.

We define ϕ_{txy} as the immediate payoff associated with making decision y at state x at some time t, measured in units of some numeraire[5].

[4] Presumably there are other models where the agent considers another part of the distribution.

[5] As a special case, if one assumes geometric discounting, one could define $\phi_{txy} = \beta^t \phi_{xy}$ for some time independent payoffs $\phi_{xy} \in \mathbb{R}^{XY}$.

6.1.1 Linear Programming Solution

We can write the primal and dual problem of the inter-temporal decision maker as, respectively

$$\begin{aligned}
&\max_{\mu \geq 0} \mu' \phi \\
&\quad \text{s.t. } (I_T \otimes I_X \otimes 1_Y^T - J_T \otimes P)\mu = e_T^1 \otimes N \\
&\quad\quad A\mu \leq c
\end{aligned}$$

$$\begin{aligned}
&\min_{(u,\tau) \in \mathbb{R}^{TX \times L}} (e_T^1 \otimes N)' u + c'\tau \\
&\quad \text{s.t. } (I_T \otimes I_X \otimes 1_Y - J_T' \otimes P')u \geq \phi
\end{aligned}$$

In coordinates notation, and rearranging one of the constraints a little bit, the dual can be expressed as

$$\begin{aligned}
&\min_{(u,\tau) \in \mathbb{R}^{TX \times L}} \sum_{x \in [X]} N_x u_{1x} + \sum_{l \in [L]} c_l \tau_l \\
&\quad \text{s.t. } u_{tx} \geq \phi_{txy} - \sum_{l \in [L]} A_{l,txy}\tau_l + \sum_{x' \in [X]} P_{x'xy} u_{(t+1)x'} \ \forall (t,x,y) \in [T \times X \times Y] \\
&\quad u_{(T+1)x} = 0 \ \forall x \in [X]
\end{aligned}$$

Since τ is the vector of Lagrangian multipliers associated with the scarcity constraints, $\sum_{l \in [L]} A_{l,txy}\tau_l$ can be interpreted as the shadow cost of option $y \in [Y]$ in state $x \in [X]$ at time $t \in [T]$. If there's a capacity constraint on the mass of individuals that can make choice y at time t, then a mass of agents making a choice y imposes an externality on the other agents. In this formulation, agents will internalize these externalities and act as if there short-term payoff was modified to $\phi^*_{txy} := \phi_{txy} - \sum_{l \in [L]} A_{l,txy}\tau_l^*$ where τ^* is the vector appearing in the optimal solution. It's also easy to see that in any solution to the dual, the constraints will be satisfied at equality.

6.1.2 Bellman Equation Solution

Conditional on knowing τ^*, the optimal value of τ in the dual, or there not existing capacity constraints (ie., $\tau = \mathbf{0}$), it's evident that the vector u can be solved for in closed form by backwards induction

$$\begin{aligned}
&u_{tx} = \max_{y \in [Y]} \{\phi^*_{txy} + \sum_{x' \in [X]} P_{x'xy} u_{(t+1)x'}\} \ \forall (t,x) \in [T \times X] \\
&u_{(T+1)x} = 0 \ \forall x \in [X]
\end{aligned}$$

These are *Bellman Equations* allowing one to interpret u_{tx} in the dual as the value of being in state $x \in [X]$ at time $t \in [T]$. Once the value function has been computed, one can iterate forward to find the policy function μ_{txy}. The algorithm is written in Algorithm (1).

Algorithm 1 Backward-Forward Induction Algorithm

Input: : Transition probabilities $P_{x'xy}$ and one-time payoff function ϕ^*_{txy}
Output: : Value function u_{tx} and policy function μ_{txy}
Backward Phase (value iteration):
$u_{Tx} \leftarrow \max_{y\in[Y]}\{\phi^*_{Txy}\}$
for $t \leftarrow T-1$ **to** 1 **by** -1 **do**
  $u_{tx} \leftarrow \max_{y\in[Y]}\{\phi^*_{(t+1)xy}\}$
end for
Forward Phase (policy iteration):
$n_{1x} = N_x$
$u_{(T+1)x} = 0\ \forall x \in [X]$
for $t \leftarrow 1$ **to** T **do**
  for $x \in [X]$ **do**
    $y^0 = \text{argmax}_{y\in[Y]}\phi^*_{txy} + \sum_{x'\in[X]} P_{x'xy}u_{(t+1)x'}$ ▷ If there are multiple maximizers, pick one arbitrarily.
    $u_{txy} = \mathbb{1}_{\{y=y^0\}}n_{tx}$
  end for
  $n_{(t+1)x} = \sum_{(x',y')\in[X\times Y]} P_{xx'y'}\mu_{tx'y'}$
end for
return μ, u

6.2 Adding Heterogeneity

We now assume that in addition to the single-period payoff ϕ_{txy}, for each unit $i \in [I]$, there's an additional random utility shock at time t associated with every decision $y \in [Y]$ denoted ϵ_{ity}. We will specify that shocks across different agents i, i' are independent but shocks for a given individual i can be related across alternatives at time t: we will allow for dependence between ϵ_{iyt} and $\epsilon_{iy't}$ for $(i, y, y', t) \in [I \times Y \times Y \times T]$. We denote Q_{tx} the distribution of the Y dimensional random vector $\epsilon_t = (\epsilon_{ty})_y$ from which the draws ϵ_{ity} are made.

We make the following critical assumption about Q_{tx}. We assume that the sequence $x_t, (x_t, \epsilon_t), (x_t, y_t), x_{t+1}, ...$ constitutes a Markov chain. Conditional on x_t, a Y dimensional random utility vector ϵ_t is drawn from a distribution Q_{tx_t}, based on (x_t, ϵ_t) the optimal choice y_t is chosen, and finally the next state x_{t+1} is drawn from the transition probabilities $P_{x'xy}$.

In the model with logit heterogeneities, that is when Q_{tx} is the distribution of iid standard Gumbel random variables, the primal problem and dual problem[6] are, respectively

$$\begin{aligned}
\max_{\mu\geq 0} \quad & \sum_{(t,x,y)\in[T\times X\times Y]} \mu_{txy}\phi_{txy} - \sum_{(t,x,y)\in[T\times X\times Y]} \mu_{txy}\log\left(\frac{\mu_{txy}}{\sum_{y'\in[Y]}\mu_{txy'}}\right) \\
\text{s.t.} \quad & \sum_{y\in[Y]} \mu_{1xy} = N_x\ \forall x \in [X] \\
& \sum_{y\in[Y]} \mu_{txy} = \sum_{(\tilde{x},y)\in[X\times Y]} P_{x\tilde{x}y}\mu_{(t-1)\tilde{x}y}\ \forall (t,x) \in [(2:T)\times X] \\
& A\mu \leq c
\end{aligned} \tag{36}$$

$$\begin{aligned}
\min_{(u,\tau)\in\mathbb{R}^{TX\times L}} \quad & \sum_{x\in[X]} N_x u_{1x} + \sum_{l\in[L]} c_l\tau_l \\
\text{s.t.} \quad & \log\left(\sum_{y\in[Y]} \exp\left(\phi_{txy} - \sum_{l\in[L]} A_{l,txy}\tau_l + \sum_{x'\in[X]} P_{x'xy}u_{(t+1)x'}\right)\right) = u_{tx}\ \forall (t,x) \in [T\times X]
\end{aligned}$$

[6] I skip the proof of the dual here as it's quite tedious.

The objective of the primal problem is to maximize social welfare. For logit heterogeneities, we use the result in section 1.7 to specify the objective. We can rewrite the constraint of the dual as $\exp(u_{tx}) = \sum_{y\in[Y]} \exp(\phi_{txy} - \sum_{l\in[L]} A_{l,txy}\tau_l + \sum_{x'\in[X]} P_{x'xy}u_{(t+1)x'}) \implies \sum_{y\in[Y]} \exp(\phi_{txy} - \sum_{l\in[L]} A_{l,txy}\tau_l + \sum_{x'\in[X]} P_{x'xy}u_{(t+1)x'} - u_{tx}) = 1$. Thus, we can define and interpret π_{txy} as the probability that an individual in state x makes choice y at time t

$$\pi_{txy} := \exp(\phi_{txy} - \sum_{l\in[L]} A_{l,txy}\tau_l + \sum_{x'\in[X]} P_{x'xy}u_{(t+1)x'} - u_{tx}) \tag{37}$$

with the accounting equality, given by the dual's constraint, that $\sum_{y\in[Y]} \pi_{txy} = 1 \; \forall (t,x) \in [T \times X]$.

6.3 Parameterizing Systematic Utility with Logit Heterogeneities

We assume a linear parametrization of the per-period utility given by $(\Phi\lambda)_{txy} := \sum_{k\in[K]} \phi_{txyk}\lambda_k$ so that $\Phi \in \mathbb{R}^{TXY\times K}$ and assume no capacity constraints[7]. We assume that we observe a sample $\{(t_j, x_j, y_j) : j \in [D]\}$ of iid draws where for each observation $j \in [D]$, choice y_j was made by the individual at time t_j in state x_j. We denote, for $(t,x,y) \in [T \times X \times Y]$, $\hat{\mu}_{txy} = \sum_{j\in[D]} \mathbb{1}_{\{(t_j,x_j,y_j)=(t,x,y)\}}$. Our goal is to estimate λ from $\hat{\mu}$.

Let u^λ be the optimal vector of utilities given λ with implicitly $u_{(T+1)x} = 0 \; \forall x \in [X]$. Then the conditional likelihood of choosing option $y \in [Y]$ for a unit in state $x \in [X]$ at time $t \in [T]$, as taken from equation (37), is $\pi^\lambda_{txy} = \frac{\mu^\lambda_{txy}}{n^\lambda_{tx}} = \exp((\Phi\lambda)_{txy} + (P'u^\lambda_{t+1})_{xy} - u^\lambda_{tx})$. The conditional log-likelihood is thus $\log(\pi^\lambda_{txy}) = (\Phi\lambda)_{txy} + (P'u^\lambda_{t+1}) - u^\lambda_{tx}$ so that the log-likelihood of the sample is

$$\begin{aligned} l(D|\lambda) &:= \sum_{(t,x,y)\in[T\times X\times Y]} \hat{\mu}_{txy}\left[(\Phi\lambda)_{txy} + (P'u^\lambda_{t+1})_{xy} - u^\lambda_{tx}\right] \\ &= \hat{\mu}'(\Phi\lambda + \Psi u^\lambda) \end{aligned}$$

where $\Psi := J'_T \otimes P' - \Sigma'_Y$ where $\Sigma_Y := I_{TX} \otimes 1'_Y$. We have the accounting constraint that $\sum_{y\in[Y]} \pi^\lambda_{txy} = 1 \; \forall (t,x) \in [T \times X] \implies \sum_{y\in[Y]} \exp((\Phi\lambda)_{txy} + (P'u^\lambda_{t+1})_{xy} - u^\lambda_{tx}) = 1 \; \forall (t,x) \in [T \times X]$. In matrix form

$$\Sigma_Y \exp(\Phi\lambda + \Psi u^\lambda) = 1_{TX}$$

Thus, we can write that the maximum likelihood estimator of (u,λ) solves the problem

$$\begin{aligned} \max_{(u,\lambda)\in\mathbb{R}^{TX\times K}} & \hat{\mu}'(\Phi\lambda + \Psi u) \\ \text{s.t. } & \Sigma_Y \exp(\Phi\lambda + \Psi u) = 1_{TX} \end{aligned}$$

6.3.1 An Aside on Differentials

To solve this model and to compute asymptotics, it will be useful to define a *differential* for tensors. If $f : \mathbb{R}^I \to \mathbb{R}^K$, then one defines $df \in \mathbb{R}^{K\times I}$ as the first differential and $d^2f \in \mathbb{R}^{K\times I^2}$. Evaluated at some point $x \in \mathbb{R}^I$, those differentials are

$$\begin{aligned} df(x) &:= \sum_{(i,k)\in[I\times K]} \frac{\partial f_k(x)}{\partial x_i} e^K_k \otimes \left(e^I_i\right)' \\ d^2f(x) &:= \sum_{(i,j,k)\in[I\times I\times K]} \frac{\partial^2 f_k(x)}{\partial x_i \partial x_j} e^K_k \otimes \left(e^I_i\right)' \otimes \left(e^I_i\right)' \end{aligned}$$

[7] Or alternatively we assume that the surplus over the capacity constraints is parametrized in this way.

Then, $df = Df$, which is the usual Jacobian of f. And if $K = 1$, $(d^2 f)'$ coincides with $\text{vec}_R(D^2 f)$, where $D^2 f$ is the typical Hessian of f.

We also define the operator $R_I(\cdot)$ when $K = 1$ where R_I means reshape into an $I \times I$ matrix, so that

$$R_I(d^2 f) = D^2 f$$

The idea is that R_I is the matrix that turns the $1 \times I^2$ differential into an $I \times I$ Hessian matrix.

For a positive integer A, it's also useful to define the *three-legged monster matrix* $\mathcal{M}_A$ as

$$\mathcal{M}_A := \sum_{a \in [A]} e_a^A \otimes \left(e_a^A\right)' \otimes \left(e_a^A\right)'$$

For matrix $M \in \mathbb{R}^{A \times B}$ and $N \in \mathbb{R}^{A \times C}$, then $\mathcal{M}_A(M \otimes N) \in \mathbb{R}^{A \times BC}$ and the term at (a, bc) is $M_{ab} N_{ac}$.

Indeed, if a function f is given by $f(x) = g(h(x))$ where $h : \mathbb{R}^I \to \mathbb{R}^J$ and $g : \mathbb{R} \to \mathbb{R}$, with the convention that if $y \in \mathbb{R}^Y$, $g(y) = [g(y_1), ..., g(y_Y)]'$. If g and h are twice differentiable, we have that

$$\begin{aligned} \frac{\partial f_j(x)}{\partial x_i} &= \frac{\partial g(h_j(x))}{\partial x_i} \\ &= g'(h_j(x)) \frac{\partial h_j(x)}{x_i} \\ \implies df(x) &= diag(g'(h(x))) dh(x) \end{aligned}$$

$$\begin{aligned} \frac{\partial^2 f_j(x)}{\partial x_i \partial x_k} &= g'(h_j(x)) \frac{\partial^2 g_j(x)}{\partial x_i \partial x_k} + g''(h_j(x)) \frac{\partial g_j(x)}{\partial x_i} \frac{\partial g_j(x)}{\partial x_k} \\ \implies d^2 f(x) &= diag(g'(h(x))) d^2 h(x) + diag(g''(h(x))) \mathcal{M}_J(dh(x) \otimes dh(x)) \end{aligned}$$

6.3.2 Solving Using Gradient Descent

To find the estimator using gradient descent, as hinted above, write u^λ as the vector of utilities such that $\Sigma_Y \exp(\Phi\lambda + \Psi u^\lambda) = 1_{TX}$ given the value of λ. Note that $\Psi \in \mathbb{R}^{TXY \times TX}$ so that in principle we expect u^λ to be uniquely pinned down by λ. It's easy to see that's the case when you look at how u is constructively computed from the systematic utilities in Algorithm 1. The maximization problem rewrites to

$$\max_{\lambda \in \mathbb{R}^K} \hat{\mu}'(\Phi\lambda + \Psi u^\lambda) \tag{38}$$

To compute the gradient of the objective function, we wish to compute $\text{d}_\lambda(u^\lambda)$. Denote $\pi := \exp(\Phi\lambda + \Psi u^\lambda)$ and $\Delta_\pi := diag(\pi)$. Differentiating both side of the constraint with respect to λ, we get that

$$\begin{aligned} &\frac{\partial}{\partial \lambda'}\left(\Sigma_Y \exp(\Phi\lambda + \Psi u^\lambda)\right) = \frac{\partial}{\partial \lambda}\left(1_{TX}\right) \\ \implies &\Sigma_Y \Delta_\pi \Phi + \Sigma_Y \Delta_\pi \Psi \mathrm{d}_\lambda(u^\lambda) = 0 \\ \implies &\mathrm{d}_\lambda(u^\lambda) = -(\Sigma_Y \Delta_\pi \Psi)^{-1} \Sigma_Y \Delta_\pi \Phi \end{aligned} \tag{39}$$

From this point, we can compute the differential of the objective of the problem in equation (38). That is

$$\begin{aligned} \frac{\partial}{\partial \lambda'}\left(\hat{\mu}'(\Phi\lambda + \Psi u^\lambda)\right) &= \hat{\mu}'\Phi + \hat{\mu}'\Psi d_\lambda u^\lambda \\ &= \hat{\mu}'\Phi - \hat{\mu}'\Psi((\Sigma_Y \Delta_\pi \Psi)^{-1}\Sigma_Y \Delta_\pi \Phi) \\ &= \hat{\mu}'(I_{TXY} - (\Sigma_Y \Delta_\pi \Psi)^{-1}\Sigma_Y \Delta_\pi)\Phi \end{aligned} \tag{40}$$

At this point, we have everything necessary to compute run gradient descent to solve the maximization problem in equation (38).

6.3.3 Asymptotics

To find the asymptotic distribution of $\hat{\lambda}$, we need to differentiate equation (39) again with respect to λ.

Using the fact from Section 6.3.1, we get that,

$$\begin{aligned} &\Sigma_Y \Delta_\pi \Psi d^2_{\lambda\lambda} u^\lambda + \Sigma_Y \Delta_\pi \mathcal{M}_{TXY}((\Phi + \Psi d_\lambda u^\lambda) \otimes (\Phi + \Psi d_\lambda u^\lambda)) = 0 \\ \implies &d^2_{\lambda\lambda} u^\lambda = -(\Sigma_Y \Delta_\pi \Psi)^{-1} \Sigma_Y \Delta_\pi \mathcal{M}_{TXY}((\Phi + \Psi d_\lambda u^\lambda) \otimes (\Phi + \Psi d_\lambda u^\lambda)) \end{aligned}$$

From this point, we can compute the second differential of the objective of the problem in equation (38). That is

$$\begin{aligned} \frac{\partial^2}{\partial \lambda' \partial \lambda'}\left(\hat{\mu}'(\Phi\lambda + \Psi u^\lambda)\right) &= \hat{\mu}'\Psi d^2_{\lambda\lambda} u^\lambda \\ &= -\hat{\mu}'\Psi(\Sigma_Y \Delta_\pi \Psi)^{-1}\Sigma_Y \Delta_\pi \mathcal{M}_{TXY}((\Phi + \Psi d_\lambda u^\lambda) \otimes (\Phi + \Psi d_\lambda u^\lambda)) \end{aligned}$$

Properly moving towards the asymptotics, recall that we posited that we observed a sample $\{(t_j, x_j, y_j) : j \in [D]\}$ of iid draws from a population where for each observation $j \in [D]$, choice y_j was made by the individual at time t_j in state x_j. We then define $\hat{\pi} := \frac{\hat{\mu}}{D}$ to be the sample frequency with which each observation is observed. We define

$$F(\hat{\pi}, \hat{\lambda}) = \Phi'\hat{\pi} + (d_\lambda u^{\hat{\lambda}})'\Psi'\hat{\pi}$$

as the gradient of the objective divided by D, almost as in equation (40). The parameter estimate is characterized by $F(\hat{\pi}, \hat{\lambda}) = 0$. We assume that in the population, we have that

$$0 = F(\pi, \lambda) = \Phi'\pi + (d_\lambda u^\lambda)'\Psi'\pi$$

Taking a first order expansion of $F(\cdot,\cdot)$ around (π,λ), and evaluating it at $(\hat{\pi},\hat{\lambda})$, I get that

$$\begin{aligned}
0 &= F(\hat{\pi},\hat{\lambda}) \\
&\approx \cancelto{0}{F(\pi,\lambda)} + F_\pi(\pi,\lambda)(\hat{\pi}-\pi) + F_\lambda(\pi,\lambda)(\hat{\lambda}-\lambda) \\
&= (\Phi' + (d_\lambda u^\lambda)'\Psi')(\hat{\pi}-\pi) + R_K((d_{\lambda\lambda}u^\lambda)'\Psi'\pi)(\hat{\lambda}-\lambda) \\
\implies \hat{\lambda}-\lambda &\approx \left[R_K((d_{\lambda\lambda}u^\lambda)'\Psi'\pi)\right]^{-1}(\Phi' + (d_\lambda u^\lambda)'\Psi')(\hat{\pi}-\pi)
\end{aligned}$$

As a result, assuming correct specification, we have that $\sqrt{D}(\hat{\lambda}-\lambda) \to \mathcal{N}(0, H^{-1}CH^{-1})$ where

$$\begin{aligned}
H &= R_K((d_{\lambda\lambda}u^\lambda)'\Psi'\pi) \\
C &= (\Phi' + (d_\lambda u^\lambda)'\Psi')V_\pi(\Phi' + (d_\lambda u^\lambda)'\Psi')'
\end{aligned}$$

Where $R_K(\cdot)$ is defined in section 6.3.1 and $V_\pi = diag(\pi) - \pi\pi'$ as computed in section 5.2.

6.4 Infinite Time Horizon

We now consider a dynamic discrete choice model but where $T=\infty$. In this case, we assume a constant geometric discounting $\beta \in (0,1)$, so that measured in period-zero equivalent utility, $\phi_{txy} = \beta^{t-1}\tilde{\phi}_{xy}$, where $\tilde{\phi}_{xy}$ is time-invariant. For consistency, we use the same geometric discounting for the random part of the utility so that the total payoff associated with option y chosen at time t for a unit in state x, in terms of time 0 utility is $\phi_{txy}+\epsilon_{txy} = \beta^{t-1}\tilde{\phi}_{xy} + \beta^{t-1}\epsilon_{xy}$. Note that this is different from the assumptions in section 6.2 where we allowed the systematic component of the utility to vary with time and we did not discount the random component of the utility.

Ignoring capacity constraints, and assuming that we have logit heterogeneities, we can write the primal and dual problems as in equation (36) as

$$\begin{aligned}
\max_{\mu\geq 0} & \sum_{(t,x,y)\in[\infty\times X\times Y]} \mu_{txy}\beta^{t-1}\tilde{\phi}_{xy} - \sum_{(t,x,y)\in[\infty\times X\times Y]} \mu_{txy}\beta^{t-1}\log\left(\frac{\mu_{txy}}{\sum_{y'\in[Y]}\mu_{txy'}}\right) \\
\text{s.t. } & \sum_{y\in[Y]} \mu_{1xy} = N_x \;\forall x\in[X] \\
& \sum_{y\in[Y]}\mu_{txy} = \sum_{(\tilde{x},y)\in[X\times Y]} P_{x\tilde{x}y}\mu_{(t-1)\tilde{x}y} \;\forall (t,x)\in[(2:T)\times X]
\end{aligned}$$

$$\begin{aligned}
\min_{u\in\mathbb{R}^{\infty X}} & \sum_{x\in[X]} N_x u_{1x} \\
\text{s.t. } & \log\Big(\sum_{y\in[Y]}\exp(\beta^{t-1}\tilde{\phi}_{xy} + \beta^t \sum_{x'\in[X]} P_{x'xy}u_{(t+1)x'})\Big) = u_{tx} \;\forall (t,x)\in[T\times X]
\end{aligned}$$

We can the define $\tilde{u}_{tx} = \frac{u_{tx}}{\beta^{t-1}}$ to re-express the dual problem as

$$\begin{aligned}
\min_{u\in\mathbb{R}^{\infty X}} & \sum_{x\in[X]} N_x \tilde{u}_{1x} \\
\text{s.t. } & \log\Big(\sum_{y\in[Y]}\exp(\tilde{\phi}_{xy} + \beta \sum_{x'\in[X]} P_{x'xy}\tilde{u}_{(t+1)x'})\Big) = \tilde{u}_{tx} \;\forall (t,x)\in[T\times X]
\end{aligned}$$

For readability, we now drop all of the tildes from the variables and rewrite the dual as

$$\min_{u \in \mathbb{R}^{\infty X}} \sum_{x \in [X]} N_x u_{1x}$$
$$\text{s.t. } \log(\sum_{y \in [Y]} \exp(\phi_{xy} + \beta \sum_{x' \in [X]} P_{x'xy} u_{(t+1)x'})) = u_{tx} \; \forall (t,x) \in [T \times X]$$

6.4.1 Stationary Problem → Stationary Solution

We note that since the systematic utility and discount factor in each period is time-invariant, we claim that there exists a stationary solution u to the dual and that it's unique. More restricted, there exists a unique sequence u so that $u_{tx} = u_x$ $\forall (t,x) \in [\infty \times X]$ and

$$u_x = \log(\sum_{y \in [Y]} \exp(\phi_{xy} + \beta \sum_{x' \in [X]} P_{x'xy} u_{x'})) \tag{41}$$

The idea of the proof is to invoke the *principal of optimality* and argue that we require a stationary solution since the problem faced by an individual in any state $x \in [X]$ is the same at any time $t \in [\infty]$ due to time-invariance[8]. We will argue that the value of u_{tx} remains bounded and that $G : \mathbb{R}^X \to \mathbb{R}^X$, where an arbitrary coordinate $x \in [X]$ of the output is given by

$$G^x(u) = \log(\sum_{y \in [Y]} \exp(\phi_{xy} + \beta \sum_{x' \in [X]} P_{x'xy} u'_x))$$

is a contraction mapping. We note that $G(u) \leq \frac{1}{1-\beta} (||\phi||_\infty + \log(Y))$, where I use the result from equation (3) for zero mean iid Gumbel random variables and $||v||_\infty = \max_x |v_x|$.

Theorem 2 (Blackwell's Theorem)**.** If $F : \mathbb{R}^X \to \mathbb{R}^X$ is unit-additive (ie., $F(u + 1_X) = F(u) + 1_X$), then the following statements are equivalent.

$(i) F$ is order preserving ie., $u \leq u' \implies F(u) \leq F(u')$.
$(ii) F$ is non-explosive for the infinite norm ie., $||F(u) = F(u')||_\infty \leq ||u - u'||_\infty$.

Theorem 3 (Contraction Mapping Theorem, Banach's Fixed Point Theorem)**.** If (S, ρ) is a complete metric space and $T : S \to S$ is a strict contraction mapping with modulus β, then

$(i) T$ has exactly one fixed point $v \in S$.
(ii)For any $v_0 \in S$, $\rho(T^n(v_0), v) \leq \beta^n \rho(v_0, v)$ $\forall n \in \mathbb{N}$

In our setting, we take $G(u) = F(\beta u)$ where $F^x(v) = \log(\sum_{y \in [Y]} \exp(\phi_{xy} + \sum_{x' \in [X]} P_{x'xy} v_{x'}))$. We clearly see that $F(\cdot)$ is order-preserving and can check that it's unit additive, which by Blackwell's Theorem implies that it's not explosive in the infinite norm.

[8]For a sharp treatment of this claim, search for the *Verification Theorem*.

$$
\begin{aligned}
&||F(v) - F(v')||_\infty \leq ||v - v'||_\infty \\
\implies &||F(\beta u) - F(\beta u')||_\infty \leq ||\beta u - \beta u'||_\infty \\
\implies &||G(u) - G(u')||_\infty \leq \beta ||u - u'||_\infty
\end{aligned}
$$

Because $\beta \in (0, 1)$, we see that G is a strict contraction. Applying the Contraction Mapping Theorem on G, which resides in the complete metric space of bounded continuous functions endowed with the sup-norm, we get that G has a unique fixed point u so that $G(u) = u$.

Moving towards a setting relevant for estimation, let's define $\phi_{xy}^\lambda := \Phi\lambda$. We observe that equation (41) implies that $\forall x \in [X]$

$$
1 = \sum_{y \in [Y]} \exp(\phi_{xy}^\lambda + \beta \sum_{x' \in [X]} P_{x'xy} u_{x'} - u_x)
$$

which implies that in vectorized form

$$
\begin{aligned}
1_X &= \Sigma_Y \exp(\Phi\lambda + \beta P' u - \Sigma_Y' u) \\
\implies 1_X &= \Sigma_Y \exp(\Phi\lambda + \Psi' u)
\end{aligned}
$$

where $\Psi := \beta P' - \Sigma_Y'$ and $\Sigma_Y := I_X \otimes 1_Y'$. The likelihood of observing action y in state x is

$$
\begin{aligned}
\pi_{xy} &= \exp(\phi_{xy}^\lambda + \beta \sum_{x' \in [X]} P_{x'xy} u_{x'} - u_x) \\
\implies \pi &= \exp(\Phi\lambda + \Psi u) \in \mathbb{R}^{XY}
\end{aligned}
$$

We can formulate the MLE problem for estimating λ by

$$
\max_{\lambda \in \mathbb{R}^K} \hat{\mu}' (\Phi\lambda + \Psi u^\lambda)
$$

where u^λ is defined by $\Sigma_Y \exp(\Phi\lambda + \Psi u^\lambda) = 1_X$. This problem can be solved using gradient descent in a manner that's very similar to what was done in the finite horizon case (except one needs to compute u^λ using a fixed point algorithm, with some numerical error). It can also be solved using an augmented lagrangian method. The asymptotics here are also totally analogous to the finite horizon case so I will not discuss this further.

6.4.2 Augmented Lagrangian

For functions $f : \mathbb{R}^N \to \mathbb{R}$, $g : \mathbb{R}^N \to \mathbb{R}^M$, consider the minimization problem

$$
\begin{aligned}
&\min_{x \in \mathbb{R}^N} f(x) \\
&\text{s.t. } g(x) = 0_M
\end{aligned}
\tag{42}
$$

There a are a couple more obvious approaches to solve the problem: (1) formulating the problem as a min-max problem and (2) using a penalty approach. Approach (1) is to solve

$$\min_{x\in\mathbb{R}^N}\max_{y\in\mathbb{R}^M} f(x)+y'g(x)$$

The most natural approach is to run the primal dual algorithm. That creates the Lagrangian $\mathcal{L}(x,y)=f(x)+y'g(x)$. The first order conditions are

$$\begin{aligned}0&=\frac{\partial f(x)}{\partial x}+\left(\frac{\partial g(x)}{\partial x'}\right)' y\\ 0&=g(x)\end{aligned}$$

The algorithm then runs gradient descent with respect to x and gradient descent with respect to y. That leads to the update equations, for learning rate α,

$$\begin{aligned}x^{t+1}&=x^t-\alpha(\frac{\partial f(x^t)}{\partial x}+\left(\frac{\partial g(x^t)}{\partial x'}\right)' y^t)\\ y^{t+1}&=y^t+\alpha g(x^{t+1})\end{aligned}$$

To see why this algorithm is very sensitive, consider this algorithm for $f(x)=0$ and $g(x)=x$. The Lagrangian becomes $\mathcal{L}(x,y)=x'y$ and the scheme becomes

$$\begin{aligned}x^{t+1}&=x^t-\alpha y^t\\ y^{t+1}&=y^t+x^{t+1}\\ &=y^t+x^t-\alpha y^t\end{aligned}$$

In matrix form, we can write the update equations as

$$\begin{aligned}\begin{pmatrix}x^{t+1}\\ y^{t+1}\end{pmatrix}&=\begin{pmatrix}1 & -\alpha\\ \alpha & 1-\alpha^2\end{pmatrix}\begin{pmatrix}x^t\\ y^t\end{pmatrix}\\ &=M\begin{pmatrix}x^t\\ y^t\end{pmatrix}\end{aligned}$$

Then, $[x^t,y^t]'=M^t[x^t,y^t]'$. The matrix M needs to have both eigenvalues less than 1 in absolute value for convergence but $det(M)=1$. Since the determinant of a matrix is equal to the product of its eigenvalues, at least one of the eigenvalues of M is at least 1 so that this algorithm won't necessarily converge unless the initial condition is on the eigenspace that corresponds to the potential eigenvalue that's less than 1.

The second approach is to do a penalty approach. That is, for some $\gamma\in\mathbb{R}_{++}$, we can solve the problem

$$\min_{x\in\mathbb{R}^N} f(x) + \frac{\gamma}{2}||g(x)||_2^2$$

The problem converges to the original one as $\gamma \to \infty$, however it's possible the problem becomes numerically unstable as we increase γ and it's not technically equivalent for finite γ. The idea of the augmented lagrangian is to construct the problem, for $\gamma \in \mathbb{R}_{++}$

$$\begin{aligned} &\min_{x\in\mathbb{R}^N} f(x) + \frac{\gamma}{2}||g(x)||_2^2 \\ &\text{s.t. } g(x) = 0 \end{aligned} \tag{43}$$

The solution of the problem is precisely equal to the original problem in equation (42). We can then write the associated min-max problem and create the augmented Lagrangian.

$$\begin{aligned} &\min_{x\in\mathbb{R}^N}\max_{y\in\mathbb{R}^M} f(x) + \frac{\gamma}{2}||g(x)||_2^2 + y'g(x) \\ &\mathcal{L}_\gamma(x,y) = f(x) + \frac{\gamma}{2}||g(x)||_2^2 + y'g(x) \end{aligned}$$

The first order conditions are

$$\begin{aligned} \frac{\partial \mathcal{L}_\gamma(x,y)}{\partial x} &= \frac{\partial f(x)}{\partial x} + \left(\frac{\partial g(x)}{\partial x'}\right)' y + \gamma\left(\frac{\partial g(x)}{\partial x'}\right)' g(x) \\ &= \frac{\partial f(x)}{\partial x} + \left(\frac{\partial g(x)}{\partial x'}\right)' (y + \gamma g(x)) \\ \frac{\partial \mathcal{L}_\gamma(x,y)}{\partial y} &= \gamma g(x) \end{aligned}$$

The augmented Lagrangian algorithm consists of doing the following sequential operations, for some $\gamma_t \to \infty$, till convergence given some initial (x^0, y^0)

$$\begin{aligned} x^{t+1} &= \text{argmin}_{x\in\mathbb{R}^N} f(x) + \frac{\gamma_t}{2}||g(x)||_2^2 + (y^t)'g(x) \\ y^{t+1} &= y^t + \gamma_t g(x^{t+1}) \end{aligned}$$

That is, we're doing gradient ascent on y. As for the determination of x^{t+1}, it's characterized by the first order condition

$$\begin{aligned} 0 &= \frac{\partial \mathcal{L}_{\gamma_t}(x^{t+1}, y^t)}{\partial x} \\ &= \frac{\partial f(x^{t+1})}{\partial x} + \left(\frac{\partial g(x^{t+1})}{\partial x'}\right)' (y^t + \gamma_t g(x^{t+1})) \\ &= \frac{\partial f(x^{t+1})}{\partial x} + \left(\frac{\partial g(x^{t+1})}{\partial x'}\right)' y^{t+1} \\ &= \frac{\partial \mathcal{L}_0(x^{t+1}, y^{t+1})}{\partial x} \end{aligned}$$

Intuitively, at each step t, x^{t+1} is picked so that it anticipates the impact of γ_t on the problem and absorbs that into y.

7 CHARACTERISTICS BASED MODELS

Consider the discrete choice problem in equation (1), replicated here

$$\max_{y\in[Y]} U_y + \epsilon_{iy}$$

We assume that $\epsilon_{iy} = \sum_{k\in[K]} e_{ik}\xi_{ky}$ so that the discrete choice problem becomes

$$\max_{y\in[Y]} U_y + \sum_{k\in[K]} e_{ik}\xi_{ky} \tag{44}$$

Let ξ_y be the vector of characteristics of product y so that $\sum_{k\in[K]} e_{ik}\xi_{ky} = e'\xi_y$. We have that an individual with a vector of random valuations e chooses option $y \in [Y]$ if

$$\begin{aligned} & U_y + \sum_{k\in[K]} e_{ik}\xi_{ky} \geq U_{y'} + \sum_{k\in[K]} e_{ik}\xi_{ky'} \ \forall y' \in [Y] \\ \iff & U_y + e'\xi_y \geq U_{y'} + e'\xi_{y'} \\ \iff & U_y - U_{y'} \geq e'(\xi_y - \xi_{y'}) \end{aligned}$$

In Figure 1, the plot characteristics of 4 alternatives in the characteristics space for the $K = 2$ case. The values ξ_y for $y \in \{1, 2, 3, 4\}$ show the positions of the alternatives. The $\mathcal{E}_i$ for $i \in \{1, 2, 3, 4\}$ show locations of possible random individual valuations that end up with each of the choices respectively.

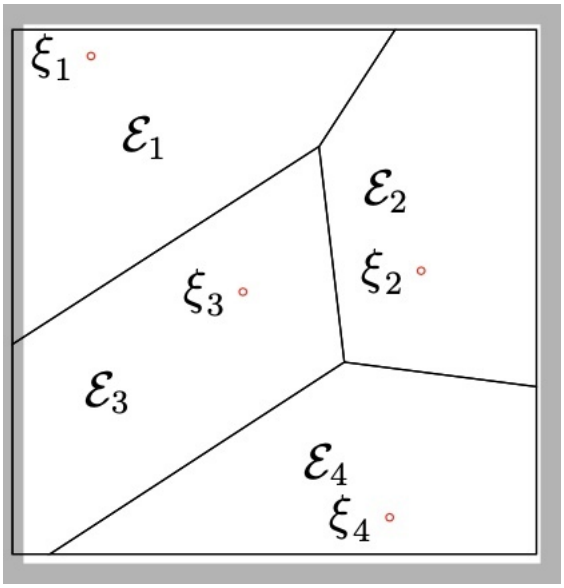

Figure 1. Veronoi Diagram

If we define $D_y = \{e \in \mathbb{R}^K : U_y - U_{y'} \geq e'(\xi_y - \xi_{y'}) \ \forall y' \in [Y]\}$ we get that $\pi_y(U) = \Pr(e \in D_y)$. Recall also the DZW theorem from section 1.1, $\pi_y(U) = \frac{\partial G(U)}{\partial U_y}$ where $G(U) = \mathbb{E}_{\mathcal{P}}[\max_{y\in[Y]}\{U_y + e'\xi_y\}]$.

7.1 Inverse Demand Problem and Optimal Transport Problem

For the inverse demand problem, I observe the market shares $\hat{\pi}_y$ and I wish to compute U so that $\pi_y(U) = \hat{\pi}_y$. That is, I wish to find U such that $\frac{\partial G(U)}{\partial U_y} = \hat{\pi}_y$. We had said in equation (5) that these are the first order conditions associated with

$$\begin{aligned}
&\max_{U \in \mathbb{R}^Y} \sum_{y \in [Y]} \hat{\pi}_y U_y - G(U) \\
\iff &\max_{U \in \mathbb{R}^Y} \sum_{y \in [Y]} \hat{\pi}_y U_y - \mathbb{E}_{\mathcal{P}}[\max_{y' \in [Y]} \{U_{y'} + e' \xi_y\}] \\
\iff &\max_{(U, u_e) \in \mathbb{R}^Y \times \mathbb{R}} \sum_{y \in [Y]} \hat{\pi}_y U_y - \mathbb{E}_{\mathcal{P}}[v_e] \\
&\text{s.t. } v_e = \max_{y' \in [Y]} U_{y'} + e' \xi_y \\
\iff &\max_{(U, u_e) \in \mathbb{R}^Y \times \mathbb{R}} \sum_{y \in [Y]} \hat{\pi}_y U_y - \mathbb{E}_{\mathcal{P}}[v_e] \\
&\text{s.t. } v_e \geq U_y + e' \xi_y \; \forall y \in [Y]
\end{aligned}$$

In the sample, the analogue of the problem, after flipping the problem from a maximization to a minimization is

$$\begin{aligned}
&\min_{(U, v) \in \mathbb{R}^Y \times \mathbb{R}^I} \frac{1}{I} \sum_{i \in [I]} v_i - \sum_{y \in [Y]} \hat{\pi}_y U_y \\
&\text{s.t. } v_i \geq U_y + e_i' \xi_y
\end{aligned}$$

The dual of this problem is

$$\begin{aligned}
&\max_{\pi_{iy} \in \mathbb{R}^{IY}} \sum_{(i,y) \in [I \times Y]} \pi_{iy} e_i' \xi_y \\
&\text{s.t. } \sum_{y \in [Y]} \pi_{iy} = \frac{1}{I} \; \forall i \in [I], \sum_{i \in [I]} \pi_{iy} = \hat{\pi}_y \; \forall y \in [Y] \\
\iff &\max_{\pi_{iy} \in \mathbb{R}^{IY}} \sum_{(i,y) \in [I \times Y]} \pi_{iy} \phi_{iy} \\
&\text{s.t. } \sum_{y \in [Y]} \pi_{iy} = \frac{1}{I} \; \forall i \in [I], \sum_{i \in [I]} \pi_{iy} = \hat{\pi}_y \; \forall y \in [Y]
\end{aligned}$$

where $\phi_{iy} := e_i' \xi_y$. The intuition of this dual problem is that we seek to maximize social welfare subject to each person's welfare counting equally towards the total welfare (first constraint) and each alternative getting chosen by a known fraction of the individuals. This can be viewed as an optimal transport problem from the distribution $e_i \sim \mathcal{P}$ of the individual random valuations to the distribution of the discrete set of alternatives.

7.2 The Random Coefficients Logit Model

Consider the model of equation (44) except where $\epsilon_{iy} = \sum_{k \in [K]} e_{ik} \xi_{ky}$. Let's now introduce some Gumbel noise so that we re-define

$$\epsilon_{iy} = \sum_{k \in [K]} e_{ik} \xi_{ky} + \sigma \eta_{iy}$$

where $e_i \sim \mathcal{P}_e$, a distribution on $\mathbb{R}^K$ and $\eta_i \overset{iid}{\sim} M$ where $M = \text{Gumbel}(\mu = 0, \beta = 1)$, and $e \perp\!\!\!\perp \eta$. We now have the discrete choice problem for agent i is

$$\max_{y \in [Y]} U_y + e_i' \xi_y + \sigma_{\eta i y}$$

Let's compute the usual objects $G(U)$ and $\pi_y(U)$.

$$\begin{aligned} G(U) &= \mathbb{E}_{\mathcal{P}_e, M}[\max_{y \in [Y]} U_y + e_i' \xi_y + \sigma \eta_y] \\ &= \mathbb{E}_{\mathcal{P}_e}[\mathbb{E}_M[\max_{y \in [Y]} U_y + e_i' \xi_y + \sigma \eta_y | e_i]], \text{ by the law of iterated expectations} \\ &= \mathbb{E}_{\mathcal{P}_e}\left[\sigma \log\left(\sum_{y \in [Y]} \exp\left(\frac{U_y + e_i' \xi_y}{\sigma}\right)\right)\right], \text{ since } e \perp\!\!\!\perp \eta \end{aligned}$$

Unless we're lucky (eg., in the case of the nested logit model), we cannot get a closed form for the latter expression so we simulate it. We draw $e_i \sim P_e$ for $i \in [I]$ and we can compute a simulated analogue to G as

$$G_I(U) = \frac{1}{I} \sum_{i \in [I]} \sigma \log\left(\sum_{y \in [Y]} \exp\left(\frac{U_y + e_i' \xi_y}{\sigma}\right)\right)$$

By the theorem of section 1.1, we have that $\pi_y(U) = \frac{\partial G(U)}{\partial U_y}$, which cannot be computed in closed form here since we don't have an expression for G. However, we can compute the simulated analogue

$$\pi_{Iy}(U) = \frac{1}{I} \sum_{i \in [I]} \frac{\exp\left(\frac{U_y + e_i' \xi_y}{\sigma}\right)}{\sum_{y' \in [Y]} \exp\left(\frac{U_{y'} + e_i' \xi_{y'}}{\sigma}\right)}$$

7.2.1 Generalized Entropy of Choice and Demand Inversion

Suppose that we're given $\hat{\pi}$. We'd like to look for U such that $\pi_{Iy}(U) = \hat{\pi}_y \iff \frac{\partial G_I(U)}{\partial U_y} = \hat{\pi}_y$. We can reformulate the search for U as the solution to the problem

$$\max_{U \in \mathbb{R}^Y} \sum_{y \in [Y]} \hat{\pi}_y U_y - G_I(U) \tag{45}$$

The objective is $G^*(\hat{\pi})$, which in section 1.5 we called the *generalized entropy of choice*. This problem can be solved using gradient ascent whereby I take an initial guess $U_y^0 = 0$ and then take the update equations, with learning rate α,

$$\begin{aligned} U_y^{t+1} &= U_y^t + \alpha(\hat{\pi}_y - \frac{\partial G_I(U^t)}{\partial U_y}) \\ &= U_y^t + \alpha(\hat{\pi}_y - \pi_{Iy}(U^t)) \end{aligned} \tag{46}$$

Berry, Levinsohn, and Pakes use an alternative approach based on coordinate ascent. By the 'Poisson trick', the problem in equation (45) is equivalent to the problem

$$\max_{(U,u)\in\mathbb{R}^Y\times\mathbb{R}^I} \sum_{y\in[Y]} \hat{\pi}_y U_y - \frac{1}{I}\sum_{i\in[I]} u_i - \sigma \sum_{(i,y)\in[I\times Y]} \exp\left(\frac{U_y - u_i + e_i'\xi_y}{\sigma}\right) \tag{47}$$

This problem, after flipping the max to the min, is in turn the dual of an optimal transport problem with entropic regularization

$$\begin{aligned}
\max_{\lambda\in\mathbb{R}_+^{IY}} & \sum_{(i,y)\in[I\times Y]} \lambda_{iy}(e_i'\xi_y) - \sigma \sum_{(i,y)\in[I\times Y]} \lambda_{iy}\log(\lambda_{iy}) \\
\text{s.t.} & \sum_{y\in[Y]} \lambda_{iy} = \frac{1}{I} \\
& \sum_{i\in[I]} \lambda_{iy} = \hat{\pi}_y
\end{aligned}$$

The first order conditions of the problem in equation (47) are

$$\begin{aligned}
0 &= \hat{\pi}_y - \sum_{i\in[I]} \exp\left(\frac{U_y - u_i + e_i'\xi_y}{\sigma}\right) \\
\implies \hat{\pi}_y &= \sum_{i\in[I]} \exp\left(\frac{U_y - u_i + e_i'\xi_y}{\sigma}\right) \\
\implies U_y &= -\sigma\log\left(\frac{1}{\hat{\pi}_y}\sum_{i\in[I]} \exp\left(\frac{-u_i + e_i'\xi_y}{\sigma}\right)\right) \\
0 &= \frac{1}{I} - \sum_{y\in[Y]} \exp\left(\frac{U_y - u_i + e_i'\xi_y}{\sigma}\right) \\
\implies \frac{1}{I} &= \sum_{y\in[Y]} \exp\left(\frac{U_y - u_i + e_i'\xi_y}{\sigma}\right) \\
\implies u_i &= \sigma\log\left(I\sum_{y\in[Y]} \exp\left(\frac{U_y + e_i'\xi_y}{\sigma}\right)\right)
\end{aligned}$$

We therefore get the coordinate ascent algorithm where we initialize $U_y^0 = 0, u_i^0 = 0$ and have the update equations,

$$U_y^{t+1} = -\sigma\log\left(\frac{1}{\hat{\pi}_y}\sum_{i\in[I]} \exp\left(\frac{-u_i^t + e_i'\xi_y}{\sigma}\right)\right) \tag{48}$$

$$u_i^{t+1} = \sigma\log\left(I\sum_{y\in[Y]} \exp\left(\frac{U_y^{t+1} + e_i'\xi_y}{\sigma}\right)\right) \tag{49}$$

The coordinate ascent algorithm in the optimal transport context is called *Sinkhorn's algorithm*. There's actually an easier set of update equations for the coordinate ascent algorithm. Defining $A_y = \exp(\frac{U_y}{\sigma}), B_i = \exp(\frac{-u_i}{\sigma})$, and $K_{iy} = \exp(\frac{e_i'\xi_y}{\sigma})$, we have that

$$A_y^{t+1} = \left(\frac{1}{\hat{\pi}_y} \sum_{i \in [I]} K_{iy} B_i^t \right)^{-1}$$
$$B_i^{t+1} = \left(I \sum_{y \in [Y]} K_{iy} A_y^{t+1} \right)^{-1}$$

7.2.2 BLP Contraction Mapping Algorithm

The BLP contraction mapping algorithm consists of expressing U^{t+1} in terms of U^t in the update equation of (48). By equation (49), we have that

$$\exp\left(\frac{u_i^t}{\sigma}\right) = I \sum_{y \in [Y]} \exp\left(\frac{U_y^t + e_i'\xi_y}{\sigma}\right)$$

That implies that

$$\begin{aligned}
U_y^{t+1} &= -\sigma \log\left(\frac{1}{\hat{\pi}_y} \sum_{i \in [I]} \exp\left(\frac{-u_i^t + e_i'\xi_y}{\sigma}\right) \right) \\
&= -\sigma \log\left(\frac{1}{\hat{\pi}_y} \left(I \sum_{y' \in [Y]} \exp\left(\frac{U_{y'}^t + e_i'\xi_{y'}}{\sigma}\right) \right)^{-1} \sum_{i \in [I]} \exp\left(\frac{e_i'\xi_y}{\sigma}\right) \right) \\
&= -\sigma \log\left(\frac{1}{I\hat{\pi}_y} \sum_{i \in [I]} \frac{\exp\left(\frac{e_i'\xi_y}{\sigma}\right)}{\sum_{y' \in [Y]} \exp\left(\frac{U_{y'}^t + e_i'\xi_{y'}}{\sigma}\right)} \right) \\
&= -\sigma \log\left(\exp\left(\frac{-U_y^t}{\sigma}\right) \frac{1}{I\hat{\pi}_y} \sum_{i \in [I]} \frac{\exp\left(\frac{U_y^t + e_i'\xi_y}{\sigma}\right)}{\sum_{y' \in [Y]} \exp\left(\frac{U_{y'}^t + e_i'\xi_{y'}}{\sigma}\right)} \right) \\
&= U_y^t - \sigma \log\left(\frac{\pi_{Iy}(U^t)}{\hat{\pi}_y} \right) \\
&= U_y^t + \sigma(\log(\hat{\pi}) - \log\left(\pi_{Iy}(U^t)\right)
\end{aligned}$$

If we did not have the $\log(\cdot)$ we would get a gradient ascent algorithm with the learning rate of σ. With the $\log(\cdot)$, we get the BLP contraction mapping algorithm. As for intuition for the appearance of the $\log(\cdot)$, recall that in section 1.4.1, we had that $U_y = \log(\pi_y) - \log(\pi_0)$. In other words, we can think that the $\log(\cdot)$ function converts from market shares to the units of utility.

REFERENCES

Alfred Galichon. Applied microeconometrics. PhD Course, NYU Economics, Spring 2025, 2025. Accessed: 2025-05-10. URL: https://alfredgalichon.com/applied-microeconometrics-spring-2025/.