This page is intentionally blank. Proper e-companion title page, with INFORMS branding and exact metadata of the main paper, will be produced by the INFORMS office when the issue is being assembled.

Online Appendix

Estimating Large-Scale Tree Logit Models

Srikanth Jagabathula

Leonard N. Stern School of Business, New York University, New York, NY 10012, sjagabat@stern.nyu.edu

Paat Rusmevichientong

Marshall School of Business, University of Southern California, Los Angeles, CA 90089, rusmevic@marshall.usc.edu

Ashwin Venkataraman

Naveen Jindal School of Management, University of Texas at Dallas, Richardson, TX 75080, axv190029@utdallas.edu

Xinyi Zhao

Amazon Advertising, Palo Alto, CA 94301, xz2197@stern.nyu.edu

Appendix A: Proof of Theorem 2.1 (Random Utility Representation)

Recall that a random variable x follows a Gumbel distribution with a location parameter μ and a scaling parameter β if for all $x \in \mathbb{R}$, $\mathbb{P}\{x \leq x\} = e^{-e^{-(x-\mu)/\beta}}$, and we denote this by $x \sim \mathsf{Gumbel}(\mu, \beta)$. We say that x follows a *standard* Gumbel distribution if $x \sim \mathsf{Gumbel}(0, 1)$. The proof of Theorem 2.1 makes use of the following standard results about the Gumbel distribution (Gumbel 2004).

Lemma A.1 (Gumbel Properties) *Suppose $x \sim \mathsf{Gumbel}(\mu, \beta)$, and let $x_1, \ldots, x_n$ be independent Gumbel random variables with $x_j \sim \mathsf{Gumbel}(\mu_j, \beta)$ for all j. Then,*

1. $\mathbb{E}[x] = \mu + \beta\gamma$ *where* $\gamma = 0.57721...$ *is the Euler-Mascheroni constant and* $\mathsf{Var}[x] = \pi^2\beta^2/6$.
2. *For any* $b > 0$ *and* $a \in \mathbb{R}$, $bx + a \sim \mathsf{Gumbel}(b\mu + a, b\beta)$.
3. *The random variable $x_1 - x_2$ follows a Logistic distribution with a location parameter $\mu_1 - \mu_2$ and scale parameter β; that is, for all $x \in \mathbb{R}$, $\mathbb{P}\{x_1 - x_2 \leq x\} = \frac{1}{1+e^{-[x-(\mu_1-\mu_2)]/\beta}}$.*
4. *The random variable $\max_{i=1,\ldots,n} x_j$ follows Gumbel distribution with location parameter $\beta \ln\left(\sum_{j=1}^n e^{\mu_j/\beta}\right)$ and scale parameter β, and $\mathbb{P}\{x_j > \max_{\ell \neq j} x_\ell\} = \frac{e^{\mu_j/\beta}}{\sum_{\ell=1}^n e^{\mu_\ell/\beta}}$.*
5. *The random variable $\max\{x_1, x_2\}$ is independent of $\mathbb{1}\{x_1 > x_2\}$.*

The next lemma shows that there exists a unique distribution such that when a random variable following this distribution is added to an independent $\mathsf{Gumbel}(0, \beta)$ random variable, with $\beta < 1$, the resulting sum is a standard Gumbel random variable.

Lemma A.2 (Theorem 2.1 in Cardell 1997) *Suppose* $x \sim \mathsf{Gumbel}(0,\beta)$ *with* $0 < \beta < 1$ *and there is another random variable* y *independent of* x*. Then,* $x+y$ *has a standard Gumbel distribution if and only if* y *has a density function* $f_\beta(y) = \frac{1}{\beta}\sum_{k=0}^{\infty} \frac{(-1)^k e^{-ky}}{k!\Gamma(-\beta k)}$ *for all* $y \in \mathbb{R}$*.*

As an immediate corollary, there exists an independent random variable that can be added to another Gumbel random variable to obtain a Gumbel distribution with a higher scaling parameter.

Corollary A.3 (Changing scale through addition) *Suppose* $x \sim \mathsf{Gumbel}(\mu,\nu)$*, with* $\mu \in \mathbb{R}$ *and* $\nu > 0$*, and we are given* $\lambda > 0$ *such that* $\lambda \geq \nu$*. Then, there exists a random variable* y *such that* y *is independent of* x *and* $x + y \sim \mathsf{Gumbel}(\mu,\lambda)$*.*

The requirement that $\nu \leq \lambda$ in the above corollary is necessary. Since x and y are independent, we have that $\mathsf{Var}(x+y) = \mathsf{Var}(x) + \mathsf{Var}(y)$, which implies that $\pi^2\lambda^2/6 = \pi^2\nu^2/6 + \mathsf{Var}(y)$. Because the variance of a random variable is always non-negative, it must always be true that $\lambda \geq \nu$.

For each leaf node $\ell \in \mathcal{N}$ and each node j that is an ancestor of ℓ, recall that $\mathsf{path}(j,\ell]$ denotes the nodes on the unique path from j to ℓ, excluding j. Define

$$z_{j,\ell} = v_j + \sum_{k \in \mathsf{path}(j,\ell]} v_k \; + \; \mu_\ell,$$

where we define $v_{\mathsf{root}} = 0$. The following lemma describes the distribution of $z_{j,\ell}$.

Lemma A.4 *For each leaf node* ℓ *and its ancestor* j *such that* $j \neq \mathsf{root}$*,* $z_{j,\ell} \sim \mathsf{Gumbel}\left(\mu_\ell, \lambda_{\mathsf{pa}(j)}\right)$*.*

Proof: Fix an arbitrary leaf node ℓ. We prove this result by induction on the height[11] of the ancestor j. The base case is when the height of j is zero, which means that $j = \ell$. By definition, $z_{j,\ell} = v_\ell + \mu_\ell \; \sim \; \mathsf{Gumbel}\left(\mu_\ell, \lambda_{\mathsf{pa}(j)}\right)$, which is the desired result.

To establish the induction step, we assume that $z_{j,\ell} \sim \mathsf{Gumbel}(\mu_\ell, \lambda_{\mathsf{pa}(j)})$ for all ancestors j of ℓ whose heights are at most H. Consider an ancestor j with height $H+1$. Let k denote the child of j that is also an ancestor of ℓ; note that the height of k is at most H. By definition, $z_{j,\ell} = v_j + z_{k,\ell}$. Since k has a height of H, it follows from the induction hypothesis that $z_{k,\ell} \sim \mathsf{Gumbel}(\mu_\ell, \lambda_{\mathsf{pa}(k)})$. As j is the parent of k, we have that $\lambda_{\mathsf{pa}(k)} = \lambda_j$ and $z_{k,\ell} \sim \mathsf{Gumbel}(\mu_\ell, \lambda_j)$. The definition of v_j then implies that $z_{j,\ell} = v_j + z_{k,\ell} \sim \mathsf{Gumbel}(\mu_\ell, \lambda_{\mathsf{pa}(j)})$, completing the induction step. This completes the proof. ■

[11] Recall that the height of a node is the number of edges on the longest path between that node and a leaf.

For each $\mathcal{S} \subseteq \mathcal{N}$ and each node $j \in \mathsf{T}[\mathcal{S}]$, let the $Z_j(\mathcal{S})$ denote the maximum of random variables $z_{j,\ell}$ over all the leaf nodes in $\mathsf{T}[\mathcal{S}]$ that are descendants of j; that is,

$$Z_j(\mathcal{S}) \stackrel{\text{def}}{=} \max\{z_{j,\ell} \colon \ell \text{ is a leaf node in } \mathsf{T}_j[\mathcal{S}]\} \ ,$$

where we define the maximum over an empty set to be minus infinity. The next lemma characterizes the distribution of $Z_j(\mathcal{S})$.

Lemma A.5 *For each subset $\mathcal{S} \subseteq \mathcal{N}$ and each node $j \in \mathsf{T}[\mathcal{S}]$ such that $j \neq \mathsf{root}$, $Z_j(\mathcal{S}) \sim \mathsf{Gumbel}\left(W_j(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda}) \ , \ \lambda_{\mathsf{pa}(j)}\right)$.*

Proof: Without loss of generality, we prove the result for the case $\mathcal{S} = \mathcal{N}$, so $\mathsf{T}[\mathcal{S}] = \mathsf{T}$. The proof for a general subset $\mathcal{S}$ follows from an identical argument applied on the sub-tree $\mathsf{T}[\mathcal{S}]$. Since we consider the full assortment, we will drop references to $\mathcal{S}$, and simply write Z_j and T_j. We will prove the result by induction on the height of node j. For the base case, suppose that the height of j is zero. So, j is a leaf node. Then,

$$Z_j = \max\{z_{j,\ell} \colon \ell \text{ is a leaf node in } \mathsf{T}_j\} = z_{j,j} = \mu_j + v_j \ \sim \ \mathsf{Gumbel}\left(W_j(\boldsymbol{\mu},\boldsymbol{\lambda}) \ , \ \lambda_{\mathsf{pa}(j)}\right) \ ,$$

where we use the fact that if j is a leaf node, then $v_j \sim \mathsf{Gumbel}\left(0, \lambda_{\mathsf{pa}(j)}\right)$ and $W_j(\boldsymbol{\mu},\boldsymbol{\lambda}) = \mu_j$. This completes the base case.

To establish the induction step, we assume that the result holds for all nodes with height of at most H. We now consider node j with height of $H+1$. Since $Z_j = \max\{z_{j,\ell} \colon \ell \text{ is a leaf node in } \mathsf{T}_j\}$, it follows that that

$$Z_j \ = \ v_j + \max_{k \in \mathsf{Children}(j)} \{ \ \max\{z_{k,\ell} \colon \ell \text{ is a leaf node in } \mathsf{T}_k\} \ \} \ = \ v_j + \max_{k \in \mathsf{Children}(k)} Z_k.$$

For each $k \in \mathsf{Children}(j)$, the height of k is at most H. So, invoking the induction hypothesis, we obtain that for all $k \in \mathsf{Children}(j)$,

$$Z_k \ = \ \max\{z_{k,\ell} \colon \ell \text{ is a leaf node in } \mathsf{T}_k\} \ \sim \ \mathsf{Gumbel}\left(W_k(\boldsymbol{\mu},\boldsymbol{\lambda}), \lambda_j\right)$$

because $\lambda_{\mathsf{pa}(k)} = \lambda_j$. Further, because the vertex sets of T_k and $\mathsf{T}_{k'}$ are disjoint for any $k \neq k'$ such that $\{k, k'\} \subseteq \mathsf{Children}(j)$, we have that Z_k is independent of $Z_{k'}$. It then follows from the properties of the Gumbel distribution that

$$\max_{k \in \mathsf{Children}(k)} Z_k \sim \mathsf{Gumbel}\left(\lambda_j \log\left(\sum_{k \in \mathsf{Children}(j)} e^{W_k(\boldsymbol{\mu},\boldsymbol{\lambda})/\lambda_j}\right) \ , \ \lambda_j\right) = \mathsf{Gumbel}\left(W_j(\boldsymbol{\mu},\boldsymbol{\lambda}) \ , \ \lambda_j\right),$$

where the equality follows from the definition of $W_j(\boldsymbol{\mu},\boldsymbol{\lambda})$. It thus follows from the defition of v_j that $Z_j \sim \mathsf{Gumbel}(W_j(\boldsymbol{\mu},\boldsymbol{\lambda}), \lambda_{\mathsf{pa}(j)})$. This establishes the induction step, completing the proof. ■

The next lemma allows us to simplify the conditional probability involving $Z_j(\mathcal{S})$.

Lemma A.6 *For each subset $\mathcal{S} \subseteq \mathcal{N}$ and each node $j \in \mathsf{T}[\mathcal{S}]$ such that $j \neq \mathsf{root}$,*

$$\begin{aligned}\mathbb{P}\left\{ Z_j(\mathcal{S}) \;>\; \max_{k \in \mathsf{Sibling}(j)\setminus\{j\}} Z_k(\mathcal{S}) \;\middle|\; Z_i(\mathcal{S}) \;>\; \max_{v \in \mathsf{Sibling}(i)\setminus\{i\}} Z_v(\mathcal{S}) \;\; \forall\; i \in \mathsf{path}\left(\mathsf{root}, \mathsf{pa}(j)\right] \right\} \\ = \;\; \mathbb{P}\left\{ Z_j(\mathcal{S}) \;>\; \max_{k \in \mathsf{Sibling}(j)\setminus\{j\}} Z_k(\mathcal{S}) \right\} \;,\end{aligned}$$

where $\mathsf{Sibling}(j) = \{k \in \mathsf{T}[\mathcal{S}] \;:\; \mathsf{pa}(k) = \mathsf{pa}(j)\}$ denote the siblings of j in the tree $\mathsf{T}[\mathcal{S}]$.

Proof: Without loss of generality, assume that $\mathcal{S} = \mathcal{N}$. The proof for the general $\mathcal{S}$ is essentially the same. Since we consider the full assortment, we will drop references to $\mathcal{S}$. Fix an arbitrary node $j \in \mathsf{T}$ such that $j \neq \mathsf{root}$. Let $x_1 = Z_j$ and $x_2 = \max_{k \in \mathsf{Sibling}(j)\setminus\{j\}} Z_k$. Note that x_1 and x_2 are independent of each other. We will first establish the following claim by induction.

Claim: For every node $i \in \mathsf{path}(\mathsf{root}, \mathsf{pa}(j)]$, there is a deterministic function f_i such that $Z_i = f_i\left(\max\{x_1, x_2\}, \boldsymbol{U}_i\right)$, where the random vector $\boldsymbol{U}_i$ is independent of x_1 and x_2.

We will prove the claim by induction on node i. For the base case, consider $i = \mathsf{pa}(j)$. Then, by definition,

$$Z_{\mathsf{pa}(j)} \;=\; v_{\mathsf{pa}(j)} + \max_{k \in\, \mathsf{Children}(\mathsf{pa}(j))} Z_k \;=\; v_{\mathsf{pa}(j)} + \max\left\{ Z_j \,,\; \max_{k \in \mathsf{Sibling}(j)\setminus\{j\}} Z_k \right\} \;=\; v_{\mathsf{pa}(j)} + \max\{x_1, x_2\} \;,$$

and the result follows because $v_{\mathsf{pa}(j)}$ is independent of x_1 and x_2. This proves the base case.

For the induction step, assume the result holds for some node $i \in \mathsf{path}(\mathsf{root}, \mathsf{pa}(j)]$. We will now prove that it also holds for node $\mathsf{pa}(i)$. By definition,

$$\begin{aligned} Z_{\mathsf{pa}(i)} \;&=\; v_{\mathsf{pa}(i)} + \max_{v \in\, \mathsf{Children}(\mathsf{pa}(i))} Z_v \;=\; v_{\mathsf{pa}(i)} + \max\left\{ Z_i \,,\; \max_{v \in \mathsf{Sibling}(i)\setminus\{i\}} Z_v \right\} \\ &=\; v_{\mathsf{pa}(i)} + \max\left\{ f_i\left(\max\{x_1, x_2\}, \boldsymbol{U}_i\right) \,,\; \max_{v \in \mathsf{Sibling}(i)\setminus\{i\}} Z_v \right\} \end{aligned}$$

and the desired result follows because $v_{\mathsf{pa}(i)}$ is independent of x_1, x_2, and $\boldsymbol{U}_i$. Moreover, $\max_{v \in \mathsf{Sibling}(i)\setminus\{i\}} Z_v$ is also independent of x_1 and x_2 because for each $v \in \mathsf{Sibling}(i) \setminus \{i\}$, the sub-tree T_v rooted at v is completely separate from the sub-tree T_k for all $k \in \mathsf{Sibling}(j)$. This completes the induction, establishing the claim.

It follows from the above claim that

$$\begin{aligned}\mathbb{P}\left\{ Z_j \;>\; \max_{k \in \mathsf{Sibling}(j)\setminus\{j\}} Z_k \;\middle|\; Z_i \;>\; \max_{v \in \mathsf{Sibling}(i)\setminus\{i\}} Z_v \;\; \forall\; i \in \mathsf{path}\left(\mathsf{root}, \mathsf{pa}(j)\right] \right\} \\ = \;\; \mathbb{P}\left\{ x_1 \;>\; x_2 \;\middle|\; f_i\left(\max\{x_1, x_2\}, \boldsymbol{U}_i\right) \;>\; \max_{v \in \mathsf{Sibling}(i)\setminus\{i\}} Z_v \;\; \forall\; i \in \mathsf{path}\left(\mathsf{root}, \mathsf{pa}(j)\right] \right\} \\ = \;\; \mathbb{P}\{x_1 \;>\; x_2\} \;,\end{aligned}$$

where the last equality follows from Lemma A.1, which shows that $\mathbb{1}\{x_1 > x_2\}$ is independent of $\max\{x_1, x_2\}$. Also, note that for all $i \in \mathsf{path}(\mathsf{root}, \mathsf{pa}(j)]$, $\max_{v \in \mathsf{Sibling}(i)\setminus\{i\}} Z_v$ is independent of x_1 and x_2. This completes the induction, proving the desired result.

Finally, here is the proof of Theorem 2.1.

Proof of Theorem 2.1: First note that since $\lambda_j \le \lambda_{\mathsf{pa}(j)}$, it follows from Corollary A.3 that v_j exists for all non-leaf nodes j. Next, fix an arbitrary subset $\mathcal{S}$ and $\ell \in \mathcal{S}$. Recall that $\mathsf{Sibling}(j) = \{k \in \mathsf{T}[\mathcal{S}] \,:\, \mathsf{pa}(k) = \mathsf{pa}(j)\}$ denote the siblings of j in the tree $\mathsf{T}[\mathcal{S}]$. Let $\mathsf{root} \to j_1 \to j_2 \to \cdots \to j_m \to \ell$ denote the unique path from root to ℓ in $\mathsf{T}[\mathcal{S}]$. Note that

$$\{\ell\} \;=\; \mathsf{T}_\ell[\mathcal{S}] \cap \mathcal{S} \;\subseteq\; \mathsf{T}_{j_m}[\mathcal{S}] \cap \mathcal{S} \;\subseteq\; \mathsf{T}_{j_{m-1}}[\mathcal{S}] \cap \mathcal{S} \;\cdots\; \subseteq\; \mathsf{T}_{j_1}[\mathcal{S}] \cap \mathcal{S} \;\subseteq\; \mathsf{T}_{\mathsf{root}}[\mathcal{S}] \cap \mathcal{S} = \mathcal{S}$$

Note that the event $\mathit{utility}_\ell > \max_{k \in \mathcal{S}\setminus\{\ell\}} \mathit{utility}_k$ happens if and only if for every node $j \in \mathsf{path}(\mathsf{root}, \ell]$, we have $\max_{k \in \mathsf{T}_j[\mathcal{S}] \cap \mathcal{S}} \mathit{utility}_k \;>\; \max_{i \in \mathsf{Sibling}(j)\setminus\{j\}} \max_{k \in \mathsf{T}_i[\mathcal{S}] \cap \mathcal{S}} \mathit{utility}_k$. Therefore,

$$\begin{aligned}
&\mathbb{P}\left\{ \mathit{utility}_\ell > \max_{k \in \mathcal{S}\setminus\{\ell\}} \mathit{utility}_k \right\} \\
&= \mathbb{P}\left\{ \max_{k \in \mathsf{T}_j[\mathcal{S}] \cap \mathcal{S}} \mathit{utility}_k \;>\; \max_{i \in \mathsf{Sibling}(j)\setminus\{j\}} \max_{k \in \mathsf{T}_i[\mathcal{S}] \cap \mathcal{S}} \mathit{utility}_k \quad \forall\, j \in \mathsf{path}(\mathsf{root}, \ell] \right\} \\
&= \mathbb{P}\left\{ Z_j(\mathcal{S}) \;>\; \max_{i \in \mathsf{Sibling}(j)\setminus\{j\}} Z_i(\mathcal{S}) \quad \forall\, j \in \mathsf{path}(\mathsf{root}, \ell] \right\} ,
\end{aligned}$$

where the last equality follows because for each $i \in \mathsf{Sibling}(j)$

$$\max_{k \in \mathsf{T}_i[\mathcal{S}] \cap \mathcal{S}} \mathit{utility}_k \;=\; \sum_{v \in \mathsf{path}(\mathsf{root}, \mathsf{pa}(j)]} v_v \;+\; \max_{k \in \mathsf{T}_i[\mathcal{S}] \cap \mathcal{S}} z_{i,k} \;=\; \sum_{v \in \mathsf{path}(\mathsf{root}, \mathsf{pa}(j)]} v_v \;+\; Z_i(\mathcal{S}) \;,$$

and the term $\sum_{v \in \mathsf{path}(\mathsf{root}, \mathsf{pa}(j)]} v_v$ is common for all nodes $i \in \mathsf{Sibling}(j)$.

For any collection of random variables $x_1, \ldots, x_n$, $\mathbb{P}\{x_1, \ldots, x_n\} = \prod_{j=1}^n \mathbb{P}\left\{x_j \;\middle|\; x_{j-1}, \ldots, x_1\right\}$. Thus,

$$\begin{aligned}
&\mathbb{P}\left\{ Z_j(\mathcal{S}) \;>\; \max_{i \in \mathsf{Sibling}(j)\setminus\{j\}} Z_i(\mathcal{S}) \quad \forall\, j \in \mathsf{path}(\mathsf{root}, \ell] \right\} \\
&\quad= \prod_{j \in \mathsf{path}(\mathsf{root}, \ell]} \mathbb{P}\left\{ Z_j(\mathcal{S}) \;>\; \max_{i \in \mathsf{Sibling}(j)\setminus\{j\}} Z_i(\mathcal{S}) \;\middle|\; Z_k > \max_{v \in \mathsf{Sibling}(k)\setminus\{k\}} Z_k \quad \forall\, k \in \mathsf{path}(\mathsf{root}, \mathsf{pa}(j)] \right\} \\
&\quad= \prod_{j \in \mathsf{path}(\mathsf{root}, \ell]} \mathbb{P}\left\{ Z_j(\mathcal{S}) \;>\; \max_{i \in \mathsf{Sibling}(j)\setminus\{j\}} Z_i(\mathcal{S}) \right\}
\end{aligned}$$

where last equality follows from Lemma A.6.

By Lemma A.5, $Z_j(\mathcal{S}) \sim \mathsf{Gumbel}\left(W_j(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda}) \;,\; \lambda_{\mathsf{pa}(j)}\right)$, and for each $i \in \mathsf{Sibling}(j) \setminus \{j\}$, $Z_i(\mathcal{S}) \sim \mathsf{Gumbel}\left(W_i(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda}) \;,\; \lambda_{\mathsf{pa}(i)}\right)$. Since $\lambda_{\mathsf{pa}(i)} = \lambda_{\mathsf{pa}(j)}$ for all $i \in \mathsf{Sibling}(j)$,

$$\mathbb{P}\left\{ Z_j(\mathcal{S}) \;>\; \max_{i \in \mathsf{Sibling}(j)\setminus\{j\}} Z_i(\mathcal{S}) \right\} = \frac{e^{W_j(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda})/\lambda_{\mathsf{pa}(j)}}}{\sum_{i \in \mathsf{Sibling}(j)} e^{W_i(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda})/\lambda_{\mathsf{pa}(j)}}} \;,$$

which implies that

$$\mathbb{P}\left\{\mathit{utility}_\ell > \max_{k\in\mathcal{S}\setminus\{\ell\}} \mathit{utility}_k\right\} \;=\; \prod_{j\in\mathsf{path}(\mathsf{root},\ell]} \frac{e^{W_j(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})/\lambda_{\mathsf{pa}(j)}}}{\sum_{i\in\mathsf{Sibling}(j)} e^{W_i(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})/\lambda_{\mathsf{pa}(j)}}} \;=\; \psi_{\mathsf{root}\to\ell}(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda}) = \mathbb{P}_\ell\left(\mathcal{S}\,;\,\boldsymbol{\mu},\boldsymbol{\lambda}\right)\,,$$

and this completes the proof. ■

Finally, we show that the error terms $(\varepsilon_\ell\colon \ell\in\mathcal{N})$ are identically distributed but are not independent of each other:

Lemma A.7 (Distribution of error terms) *For all $\ell\in\mathcal{N}$, $\varepsilon_\ell\sim\mathsf{Gumbel}(0,\lambda_{\mathsf{root}})$. Moreover, for all $\ell,\ell'\in\mathcal{N}$ such that $\ell\neq\ell'$*

$$\mathsf{Corr}(\varepsilon_\ell,\varepsilon_{\ell'}) = \frac{\mathsf{Cov}(\varepsilon_\ell,\varepsilon_{\ell'})}{\sqrt{\mathsf{Var}(\varepsilon_\ell)\,\mathsf{Var}(\varepsilon_{\ell'})}} \;=\; 1 \;-\; \left(\frac{\lambda_j}{\lambda_{\mathsf{root}}}\right)^2\,,$$

where j is the nearest common ancestor of ℓ and ℓ' in T.

Proof: It follows from the definition of the random variables $(z_{j,\ell}\colon j\in\mathsf{T}\setminus\mathcal{N})$ before Lemma A.4 that $\varepsilon_\ell = z_{\mathsf{root},\ell} - \mu_\ell$ for all $\ell\in\mathcal{N}$. Further since $v_{\mathsf{root}}=0$, it follows from Lemma A.4 that $z_{\mathsf{root},\ell}\sim\mathsf{Gumbel}(\mu_\ell,\lambda_{\mathsf{root}})$. Combining this with property 2 of Lemma A.1, it follows that $\varepsilon_\ell\sim\mathsf{Gumbel}(0,\lambda_{\mathsf{root}})$ for all $\ell\in\mathcal{N}$. This establishes the first part of the lemma.

Next, since $\varepsilon_\ell = \sum_{k\,\in\,\mathsf{path}(\mathsf{root},\ell]} v_k$ for any $\ell\in\mathcal{N}$, and all the v_k's are independent, it follows that $\mathsf{Cov}(\varepsilon_\ell,\varepsilon_{\ell'}) = \sum_{k\in\mathsf{path}(\mathsf{root},j]}\mathsf{Cov}(v_k,v_k) = \sum_{k\in\mathsf{path}(\mathsf{root},j]}\mathsf{Var}(v_k)$, where j is the common ancestor nearest to both ℓ and ℓ'; that is, $\{\ell,\ell'\}\subseteq\mathsf{T}_j$ but $\{\ell,\ell'\}\nsubseteq\mathsf{T}_k$ for all $k\in\mathsf{Children}(j)$. Note that if $j=\mathsf{root}$, then $\mathsf{Cov}(\varepsilon_\ell,\varepsilon_{\ell'})=0$ which is consistent with the expression in the statement of the lemma. So suppose $j\neq\mathsf{root}$. Then, it follows from our definitions that $\mathsf{Var}(v_k)=\pi^2\cdot(\lambda^2_{\mathsf{pa}(k)}-\lambda^2_k)/6$ for each non-leaf node $k\in\mathsf{T}\setminus(\mathsf{root}\cup\mathcal{N})$. Then, we obtain by telescoping that the covariance $\mathsf{Cov}(\varepsilon_\ell,\varepsilon_{\ell'}) = \pi^2\cdot(\lambda^2_{\mathsf{root}}-\lambda^2_j)/6$. Finally, since $\mathsf{Var}(\varepsilon_\ell)=\mathsf{Var}(\varepsilon_{\ell'})=\pi^2\lambda^2_{\mathsf{root}}/6$, the result follows. ■

Appendix B: Properties of the Weight Function

The following lemmas establish important properties of the weight function $W_j(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})$, which we will use repeatedly to establish properties of the negative log-likelihood function. We first state all the lemmas and then provide their proofs. The first lemma establishes positive homogeneity and additivity.

Lemma B.1 (Positive Homogeneity and Additivity) *For each subset $\mathcal{S}\subseteq\mathcal{N}$, node $j\in\mathsf{T}[\mathcal{S}]$, $\alpha>0$, and $\xi\in\mathbb{R}$, $W_j(\mathcal{S};\alpha\boldsymbol{\mu},\alpha\boldsymbol{\lambda}) = \alpha W_j(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})$ and $W_j(\mathcal{S};\boldsymbol{\mu}+\xi\boldsymbol{e},\boldsymbol{\lambda}) = \xi + W_j(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})$, where $\boldsymbol{e}$ is the vector of all ones.*

The next lemma establishes convexity, and under additional assumptions, strict convexity. Recall that for each node $j \in \mathsf{T}[\mathcal{S}]$, $\mathsf{T}_j[\mathcal{S}]$ denotes the sub-tree of $\mathsf{T}[\mathcal{S}]$ rooted at j; that is, $\mathsf{T}_j[\mathcal{S}]$ consists of the node j and all of its descendant in $\mathsf{T}[\mathcal{S}]$.

Lemma B.2 (Convexity and Strict Convexity) *For each subset $\mathcal{S} \subseteq \mathcal{N}$ and $j \in \mathsf{T}[\mathcal{S}]$, the function $(\boldsymbol{\mu}, \boldsymbol{\lambda}) \mapsto W_j(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda})$ satisfies the following properties:*

(a) *It is convex in $(\boldsymbol{\mu}, \boldsymbol{\lambda})$.*

(b) *For each $\boldsymbol{\lambda}$ and $0 < \theta < 1$,*

$$W_j(\mathcal{S}; \theta\boldsymbol{\mu} + (1-\theta)\bar{\boldsymbol{\mu}}, \boldsymbol{\lambda}) = \theta W_j(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda}) + (1-\theta) W_j(\mathcal{S}; \bar{\boldsymbol{\mu}}, \boldsymbol{\lambda})$$

if and only if there exists $\xi \in \mathbb{R}$ such that $\bar{\mu}_\ell = \mu_\ell + \xi$ for all leaf nodes $\ell \in \mathsf{T}_j[\mathcal{S}] \cap \mathcal{S}$.

(c) *For each $a \in \mathbb{R}$ and $\boldsymbol{\lambda}$, the function $\boldsymbol{\mu} \mapsto W_{\mathsf{root}}(\mathcal{N}; \boldsymbol{\mu}, \boldsymbol{\lambda})$ is strictly convex on the set $\{\boldsymbol{\mu} : \mu_1 = a\}$.*

The next lemma establishes the monotonicity of the weight function.

Lemma B.3 (Monotonicity and Strict Monotonicity) *For each $\mathcal{S} \subseteq \mathcal{N}$ and $j \in \mathsf{T}[\mathcal{S}]$, the function $(\boldsymbol{\mu}, \boldsymbol{\lambda}) \mapsto W_j(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda})$ satisfies the following properties.*

(a) *It is increasing in $(\boldsymbol{\mu}, \boldsymbol{\lambda})$; that is, if $(\boldsymbol{\mu}, \boldsymbol{\lambda}) \leq (\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\lambda}})$ where the inequality holds componentwise, then $W_j(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda}) \leq W_j(\mathcal{S}; \bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\lambda}})$.*

(b) *It is strictly increasing in μ_ℓ for each leaf node $\ell \in \mathsf{T}_j[\mathcal{S}] \cap \mathcal{S}$.*

(c) *It is strictly increasing in λ_k for each $k \in \mathsf{T}_j[\mathcal{S}]$ such that k has at least two children in $\mathsf{T}_j[\mathcal{S}]$.*

Finally, the following lemma provides an expression for the derivative of the weight function with respect to the model parameters.

Lemma B.4 (Derivatives of the Weight Functions) *For each $\mathcal{S} \subseteq \mathcal{N}$ and $j \in \mathsf{T}[\mathcal{S}]$,*

$$\begin{aligned}
\frac{\partial W_j}{\partial \mu_\ell}(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda}) &= \psi_{j \to \ell}(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda}) && \forall\ \ell \in \mathcal{N} \\
\frac{\partial W_j}{\partial \lambda_k}(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda}) &= \psi_{j \to k}(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda})\, \Delta_k(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda}) && \forall\ k \in \mathsf{T} \setminus \mathcal{N},\ k \neq \mathsf{root}\ .
\end{aligned}$$

where for each non-leaf node k such that $k \neq \mathsf{root}$,

$$\begin{aligned}
\Delta_k(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda}) &= \log\left(\sum_{i \in \mathsf{Children}(k) \cap \mathsf{T}[\mathcal{S}]} e^{W_i(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda})/\lambda_k}\right) - \frac{\sum_{i \in \mathsf{Children}(k) \cap \mathsf{T}[\mathcal{S}]} e^{W_i(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda})/\lambda_k} \times W_i(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda})}{\lambda_k \sum_{i \in \mathsf{Children}(k) \cap \mathsf{T}[\mathcal{S}]} e^{W_i(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda})/\lambda_k}} \\
&= \frac{W_k(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda}) - \sum_{i \in \mathsf{Children}(k) \cap \mathsf{T}[\mathcal{S}]} \psi_{k \to i}(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda}) \times W_i(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda})}{\lambda_k}\ .
\end{aligned}$$

We now present the proofs of these four lemmas.

B.1 Proof of Lemma B.1

Proof: Consider an arbitrary subset $\mathcal{S}$. We prove both results by induction on the height of node j. For the base case, suppose the height of j is zero. This means that j is a leaf node of $\mathsf{T}[S]$, so $W_j(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda}) = \mu_j$ and the result is trivially true. Suppose that the results holds for all nodes at height at most H. Now consider a vertex $j \in \mathsf{T}[\mathcal{S}]$ of height $H+1$. By the induction hypothesis, the results are true for all children k of node j because the height of k is at most H. Therefore, by definition of W_j and the inductive hypothesis,

$$W_j(\mathcal{S};\alpha\boldsymbol{\mu},\alpha\boldsymbol{\lambda}) \;=\; \alpha\lambda_j \log\left(\sum_{k\in\,\mathsf{Children}(j)\,\cap\,\mathsf{T}[\mathcal{S}]} e^{\alpha W_k(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})/(\alpha\lambda_j)}\right) = \alpha W_j(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda}) \text{ and}$$

$$W_j(\mathcal{S};\boldsymbol{\mu}+\xi\boldsymbol{e},\boldsymbol{\lambda}) \;=\; \lambda_j \log\left(\sum_{k\in\,\mathsf{Children}(j)\,\cap\,\mathsf{T}[\mathcal{S}]} e^{[\xi+W_k(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})]/\lambda_j}\right) = \xi + W_j(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda}).$$

We have thus established the result for nodes at height $H+1$, completing the induction step. The result of the lemma now follows. ■

B.2 Proof of Lemma B.2

Proof: Fix an arbitrary subset $\mathcal{S}$. We prove each of the three parts separately. Throughout this proof, we make use of the following observation: for each $j \in \mathsf{T}[\mathcal{S}]$, the function $W_j(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})$ only depends the parameters associated with nodes in the sub-tree $\mathsf{T}_j[\mathcal{S}]$ rooted at j; that is, $W_j(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda}) \;=\; W_j\left(\mathcal{S};\boldsymbol{\mu}_{\mathsf{T}_j[\mathcal{S}]},\boldsymbol{\lambda}_{\mathsf{T}_j[\mathcal{S}]}\right)$, where $\boldsymbol{\mu}_{\mathsf{T}_j[\mathcal{S}]} = (\mu_\ell : \ell \in \mathsf{T}_j[\mathcal{S}])$ and $\boldsymbol{\lambda}_{\mathsf{T}_j[\mathcal{S}]} = (\lambda_k : k \in \mathsf{T}_j[\mathcal{S}])$.

Proof of part (a): We use induction on the height of node j. For the base case, suppose that the height of j is zero, so j is a leaf node. Then, the result is trivially true by definition. Suppose that $W_k(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})$ is convex in $(\boldsymbol{\mu},\boldsymbol{\lambda})$ for all nodes k with height at most H. Now, consider an arbitrary non-leaf node j at height $H+1$. By the induction hypothesis, $W_k(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})$ is convex in $(\boldsymbol{\mu},\boldsymbol{\lambda})$ for all children k of node j. Then, the function

$$(\boldsymbol{\mu},\boldsymbol{\lambda}) \mapsto \log\left(\sum_{k\in\,\mathsf{Children}(j)\,\cap\,\mathsf{T}[\mathcal{S}]} e^{W_k(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})}\right)$$

must also be convex in $(\boldsymbol{\mu},\boldsymbol{\lambda})$ because it is a composition of the log-sum-exp function, which is increasing and convex, with a collection of convex functions $(\boldsymbol{\mu},\boldsymbol{\lambda}) \mapsto W_k(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})$, for $k \in \mathsf{Children}(j) \cap \mathsf{T}[\mathcal{S}]$. Note that for each $k \in \mathsf{Children}(j) \cap \mathsf{T}[\mathcal{S}]$, the function $W_k(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})$ is *independent* of λ_j because $j \notin \mathsf{T}_k[\mathcal{S}]$, so

$$\log\left(\sum_{k\in\,\mathsf{Children}(j)\,\cap\,\mathsf{T}[\mathcal{S}]} e^{W_k(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})}\right) = \log\left(\sum_{k\in\,\mathsf{Children}(j)\,\cap\,\mathsf{T}[\mathcal{S}]} e^{W_k\left(\mathcal{S};\boldsymbol{\mu}_{\mathsf{T}_k[\mathcal{S}]},\boldsymbol{\lambda}_{\mathsf{T}_k[\mathcal{S}]}\right)}\right) .$$

Therefore, the perspective of the above function is given by

$$(\boldsymbol{\mu},\boldsymbol{\lambda}) \mapsto \lambda_j \log\left(\sum_{k\in \mathsf{Children}(j)\cap \mathsf{T}[\mathcal{S}]} e^{W_k\left(\mathcal{S};\frac{\boldsymbol{\mu}_{\mathsf{T}_k[\mathcal{S}]}}{\lambda_j},\frac{\boldsymbol{\lambda}_{\mathsf{T}_k[\mathcal{S}]}}{\lambda_j}\right)}\right) = \lambda_j \log\left(\sum_{k\in \mathsf{Children}(j)\cap \mathsf{T}[\mathcal{S}]} e^{W_k\left(\mathcal{S};\boldsymbol{\mu}_{\mathsf{T}_k[\mathcal{S}]},\boldsymbol{\lambda}_{\mathsf{T}_k[\mathcal{S}]}\right)/\lambda_j}\right)$$
$$= W_j(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda}) \ ,$$

where the first equality follows from the positive homogeneity property in Lemma B.1. Because the perspective of a convex function is also convex (Boyd and Vandenberghe 2004), it follows that $W_j(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})$ is convex in $(\boldsymbol{\mu},\boldsymbol{\lambda})$, completing the induction step. This proves part (a).

Proof of part (b): We prove part (b) by induction on the height of node j. Consider the base case where j has a height of zero, so $j=\ell$ for some leaf node ℓ in $\mathsf{T}[\mathcal{S}]$. In this case, $\mathsf{T}_\ell[\mathcal{S}]=\{\ell\}$ and $W_\ell(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})=\mu_\ell$. Because $\mathsf{T}_\ell[\mathcal{S}]\cap\mathcal{S}$ contains only the node ℓ in it, the condition $\bar{\mu}_\ell=\mu_\ell+\xi$ can be trivially satisfied by choosing $\xi=\bar{\mu}_\ell-\mu_\ell$. Further, the relationship

$$W_\ell(\mathcal{S};\theta\boldsymbol{\mu}+(1-\theta)\bar{\boldsymbol{\mu}},\boldsymbol{\lambda}) = \theta W_\ell(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda}) + (1-\theta) W_\ell(\mathcal{S};\bar{\boldsymbol{\mu}},\boldsymbol{\lambda})$$

is also trivially satisfied by defnition for all $\theta\in(0,1)$. Therefore, the base case is true.

Suppose that the result holds for all nodes with height at most H. Consider an arbitrary non-leaf node j at height $H+1$. By the induction hypothesis, we have that for each $k\in\mathsf{Children}(j)$,

$$W_k(\mathcal{S};\theta\boldsymbol{\mu}+(1-\theta)\bar{\boldsymbol{\mu}},\boldsymbol{\lambda}) = \theta W_k(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda}) + (1-\theta) W_k(\mathcal{S};\bar{\boldsymbol{\mu}},\boldsymbol{\lambda})$$

if and only if there exists $\xi_k\in\mathbb{R}$ such that $\mu_\ell=\bar{\mu}_\ell+\xi_k$ for all $\ell\in\mathsf{T}_k[\mathcal{S}]\cap\mathcal{S}$. We will now prove the result at node j.

We will first prove the sufficiency. Suppose that there exists $\xi\in\mathbb{R}$ such that $\bar{\mu}_\ell=\bar{\mu}_\ell+\xi$ for all leaf nodes $\ell\in\mathsf{T}_j[\mathcal{S}]\cap\mathcal{S}$, or equivalently $\bar{\boldsymbol{\mu}}_{\mathsf{T}_j[\mathcal{S}]\cap\mathcal{S}}=\boldsymbol{\mu}_{\mathsf{T}_j[\mathcal{S}]\cap\mathcal{S}}+\xi\boldsymbol{e}$, where is $\boldsymbol{e}$ is a vector of ones of an appropriate dimension. Then,

$$\begin{aligned}
W_j(\mathcal{S};\theta\boldsymbol{\mu}+(1-\theta)\bar{\boldsymbol{\mu}},\boldsymbol{\lambda}) &= W_j(\mathcal{S};\boldsymbol{\mu}_{\mathsf{T}_j[\mathcal{S}]}+\xi(1-\theta)\boldsymbol{e},\boldsymbol{\lambda}) && [W_j \text{ only depends on } \boldsymbol{\mu}_{\mathsf{T}_j[\mathcal{S}]}]\\
&= \xi(1-\theta)+W_j(\mathcal{S};\boldsymbol{\mu}_{\mathsf{T}_j[\mathcal{S}]},\boldsymbol{\lambda}) && [\text{by Lemma } B.1]\\
&= \theta W_j(\mathcal{S};\boldsymbol{\mu}_{\mathsf{T}_j[\mathcal{S}]},\boldsymbol{\lambda}) + (1-\theta)\left(\xi+W_j(\mathcal{S};\boldsymbol{\mu}_{\mathsf{T}_j[\mathcal{S}]},\boldsymbol{\lambda})\right)\\
&= \theta W_j(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda}) + (1-\theta) W_j(\mathcal{S};\bar{\boldsymbol{\mu}},\boldsymbol{\lambda}),
\end{aligned}$$

where the last equality follows because $\bar{\boldsymbol{\mu}}_{\mathsf{T}_j[\mathcal{S}]\cap\mathcal{S}}=\boldsymbol{\mu}_{\mathsf{T}_j[\mathcal{S}]\cap\mathcal{S}}+\xi\boldsymbol{e}$. This gives the desired result.

We will now prove the necessity. Suppose that $\boldsymbol{\mu}$, $\bar{\boldsymbol{\mu}}$, and $\theta\in(0,1)$ are such that

$$W_j(\mathcal{S};\theta\boldsymbol{\mu}+(1-\theta)\bar{\boldsymbol{\mu}},\boldsymbol{\lambda}) = \theta W_j(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda}) + (1-\theta) W_j(\mathcal{S};\bar{\boldsymbol{\mu}},\boldsymbol{\lambda}). \tag{EC.1}$$

Our goal is to exhibit a $\xi \in \mathbb{R}$ such that $\bar{\boldsymbol{\mu}}_{\mathsf{T}_j[\mathcal{S}]\cap\mathcal{S}} + \xi = \boldsymbol{\mu}_{\mathsf{T}_j[\mathcal{S}]\cap\mathcal{S}}$. If we have that $\bar{\boldsymbol{\mu}}_{\mathsf{T}_j[\mathcal{S}]\cap\mathcal{S}} = \boldsymbol{\mu}_{\mathsf{T}_j[\mathcal{S}]\cap\mathcal{S}}$, then the result is trivially true. Therefore, we assume that $\bar{\boldsymbol{\mu}}_{\mathsf{T}_j[\mathcal{S}]\cap\mathcal{S}} \neq \boldsymbol{\mu}_{\mathsf{T}_j[\mathcal{S}]\cap\mathcal{S}}$. To simplify notation, let $a = 1/\lambda_j$, $m = |\,\mathsf{Children}(j)|$. Further, let $\mathsf{LSE}\colon \mathbb{R}^m \to \mathbb{R}_{++}$ be defined as $\mathsf{LSE}(\boldsymbol{x}) = \log\left(\sum_{i=1}^m e^{x_i}\right)$, for each $\boldsymbol{x} \in \mathbb{R}^m$. Note that $\mathsf{LSE}(\cdot)$ is the standard log-sum-exp function. Also, define the following three vectors in $\mathbb{R}^m_+$:

$$\begin{aligned}
\boldsymbol{y}^1 &= \left(W_k(\mathcal{S}; \theta\boldsymbol{\mu} + (1-\theta)\bar{\boldsymbol{\mu}}, \boldsymbol{\lambda})\colon k \in \mathsf{Children}(j) \cap \mathsf{T}[\mathcal{S}]\right) \\
\boldsymbol{y}^2 &= \left(W_k(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda})\colon k \in \mathsf{Children}(j) \cap \mathsf{T}[\mathcal{S}]\right) \\
\boldsymbol{y}^3 &= \left(W_k(\mathcal{S}; \bar{\boldsymbol{\mu}}, \boldsymbol{\lambda})\colon k \in \mathsf{Children}(j) \cap \mathsf{T}[\mathcal{S}]\right)
\end{aligned}$$

By convexity of W_k from part (a), we have that $\boldsymbol{y}^1 \leq \theta\boldsymbol{y}^2 + (1-\theta)\boldsymbol{y}^3$, where the inequality holds componentwise. By definition,

$$\begin{aligned}
\frac{1}{\lambda_j} W_j(\mathcal{S}; \theta\boldsymbol{\mu} + (1-\theta)\bar{\boldsymbol{\mu}}, \boldsymbol{\lambda}) &= \mathsf{LSE}\left(a\boldsymbol{y}^1\right) \\
&\leq \mathsf{LSE}\left(\theta a\boldsymbol{y}^2 + (1-\theta)a\boldsymbol{y}^3\right) && [\mathsf{LSE} \text{ is strictly increasing}] \\
&\leq \theta\mathsf{LSE}\left(a\boldsymbol{y}^2\right) + (1-\theta)\mathsf{LSE}\left(a\boldsymbol{y}^3\right) && [\mathsf{LSE} \text{ is convex}] \\
&= \frac{\theta}{\lambda_j} W_j(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda}) + \frac{1-\theta}{\lambda_j} W_j(\mathcal{S}; \bar{\boldsymbol{\mu}}, \boldsymbol{\lambda}) \\
&= \frac{1}{\lambda_j} W_j(\mathcal{S}; \theta\boldsymbol{\mu} + (1-\theta)\bar{\boldsymbol{\mu}}, \boldsymbol{\lambda}) && [\text{by hypothesis (EC.1)}].
\end{aligned}$$

Because the left and right hand side expressions are equal to each other, it must be true that both inequalities hold with equalities. We have thus shown that

$$\mathsf{LSE}\left(a\boldsymbol{y}^1\right) = \mathsf{LSE}\left(\theta a\boldsymbol{y}^2 + (1-\theta)a\boldsymbol{y}^3\right) = \theta\mathsf{LSE}\left(a\boldsymbol{y}^2\right) + (1-\theta)\mathsf{LSE}\left(a\boldsymbol{y}^3\right)$$

We now derive the implications from the above two equalities. Because the LSE function is strictly increasing, the first equality implies that $\boldsymbol{y}^1 = \theta\boldsymbol{y}^2 + (1-\theta)\boldsymbol{y}^3$. In other words, we have that for each $k \in \mathsf{Children}(j)$,

$$y^1_k = \theta y^2_k + (1-\theta)y^3_k \quad \Longleftrightarrow \quad W_k(\mathcal{S}; \theta\boldsymbol{\mu} + (1-\theta)\bar{\boldsymbol{\mu}}, \boldsymbol{\lambda}) = \theta W_k(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda}) + (1-\theta)W_k(\mathcal{S}; \bar{\boldsymbol{\mu}}, \boldsymbol{\lambda})\ .$$

It now follows from the induction hypothesis, applied to W_k, that for each $k \in \mathsf{Children}(j)$, there exists $\xi_k \in \mathbb{R}$ such that $\bar{\mu}_\ell = \mu_\ell + \xi_k$ for all $\ell \in \mathsf{T}_k[\mathcal{S}] \cap \mathcal{S}$. To establish our result, it is now sufficient to show that $\xi_k = \xi_{k'}$ for all $k \neq k'$ such that $\{k, k'\} \subseteq \mathsf{Children}(j)$. For that, we first note that because $\bar{\boldsymbol{\mu}}_{\mathsf{T}_k[\mathcal{S}]\cap\mathcal{S}} = \xi_k + \boldsymbol{\mu}_{\mathsf{T}_k[\mathcal{S}]\cap\mathcal{S}}$, we have that

$$y^3_k = W_k(\mathcal{S}; \bar{\boldsymbol{\mu}}, \boldsymbol{\lambda}) = W_k(\mathcal{S}; \bar{\boldsymbol{\mu}}_{\mathsf{T}_k\cap\mathcal{S}}, \boldsymbol{\lambda}) = \xi_k + W_k(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda}) = \xi_k + y^2_k\ , \tag{EC.2}$$

where third equality follows from Lemma B.1. Now, if $\boldsymbol{y}^2 = \boldsymbol{y}^3$, then it follows that $\xi_k = 0$ for all $k \in \mathsf{Children}(j)$, establishing the result. Therefore, we assume that $\boldsymbol{y}^2 \neq \boldsymbol{y}^3$.

Then, we focus on the second equality above:

$$\mathsf{LSE}\left(\theta a\boldsymbol{y}^2 + (1-\theta)a\boldsymbol{y}^3\right) \;=\; \theta\mathsf{LSE}\left(a\boldsymbol{y}^2\right) + (1-\theta)\mathsf{LSE}\left(a\boldsymbol{y}^3\right). \tag{EC.3}$$

It is a well-known result that the the function $t \mapsto \mathsf{LSE}(\boldsymbol{x} + t\boldsymbol{z})$ is strictly convex if and only if $\boldsymbol{z} \neq \boldsymbol{e}$. In other words, for some $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^m$, and for any $t_1, t_2 \in \mathbb{R}$, we have that

$$\mathsf{LSE}(\boldsymbol{x} + (\theta t_1 + (1-\theta)t_2)\boldsymbol{z}) = \theta\mathsf{LSE}(\boldsymbol{x} + t_1\boldsymbol{z}) + (1-\theta)\mathsf{LSE}(\boldsymbol{x} + t_2\boldsymbol{z}) \quad \text{ if and only if } \quad \boldsymbol{z} = \boldsymbol{e}.$$

Now choosing $\boldsymbol{x}$, $\boldsymbol{z}$, t_1, and t_2 to satisfy

$$\boldsymbol{x} + t_1\boldsymbol{z} \;=\; a\boldsymbol{y}^2 \qquad \text{and} \qquad \boldsymbol{x} + t_2\boldsymbol{z} \;=\; a\boldsymbol{y}^3,$$

we obtain from (EC.3) that $\mathsf{LSE}(\boldsymbol{x} + (\theta t_1 + (1-\theta)t_2)\boldsymbol{z}) = \theta\mathsf{LSE}(\boldsymbol{x} + t_1\boldsymbol{z}) + (1-\theta)\mathsf{LSE}(\boldsymbol{x} + t_2\boldsymbol{z})$. Therefore, we must have that $\boldsymbol{z} = \boldsymbol{e}$. Solving for $\boldsymbol{z}$, we get that $\boldsymbol{e} = \boldsymbol{z} = a(\boldsymbol{y}^2 - \boldsymbol{y}^3)/(t_1 - t_2)$; this equality is well defined because $\boldsymbol{y}^2 \neq \boldsymbol{y}^3$ ensures that $t_1 \neq t_2$. As a result, for an appropriately defined constant δ, we obtain that $y_k^2 - y_k^3 = \delta$ for all children $k \in \mathsf{Children}(j)$. Because we also have that $y_2^k - y_3^k = \xi_k$ from (EC.2), it must be that $\xi_k = \delta$ for all $k \in \mathsf{Children}(j)$. Consequently, we have shown that $\bar{\mu}_\ell = \mu_\ell + \delta$ for all leaf nodes $\ell \in \cup_{k \in \mathsf{Children}(j)} \mathsf{T}_k[\mathcal{S}] \cap \mathcal{S}$. Because $\mathsf{T}_j[\mathcal{S}] \cap \mathcal{S} = \cup_{k \in \mathsf{Children}(j)} \mathsf{T}_k[\mathcal{S}] \cap \mathcal{S}$, we have shown that $\bar{\mu}_\ell = \mu_\ell + \xi$ for all leaf nodes $\ell \in \mathsf{T}_j[\mathcal{S}] \cap \mathcal{S}$ and $\xi = \delta$, which is the desired result. This completes the necessity part and finishes the induction. Therefore, part (b) holds for all nodes $j \in \mathsf{T}[\mathcal{S}]$.

Proof of part (c): Part (c) follows immediately from parts (a) and (b). ■

B.3 Proof of Lemma B.3

Proof: We will prove each of the three parts separately by induction on the height of node j.

Proof of part (a): The base case where j has a height of zero is trivially true by definition. Suppose the result is true for all nodes of height at most H. Then, consider node j with height $H+1$. By definition,

$$W_j(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda}) = \lambda_j \log\left(\sum_{k \in \mathsf{Children}(j) \cap \mathsf{T}[\mathcal{S}]} e^{W_k(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda})/\lambda_j}\right)$$

Note that $W_j(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda})$ only depends on $(\mu_\ell : \ell \in \mathcal{S})$, $\left(\boldsymbol{\lambda}_{\mathsf{T}_k[\mathcal{S}]} : k \in \mathsf{Children}(j)\right)$, and λ_j. By induction, for each $k \in \mathsf{Children}(j)$, $W_j(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda})$ is increasing in $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}_{\mathsf{T}_k[\mathcal{S}]}$. Moreover, taking the derivative of the above expression with respect to λ_j, we get

$$\frac{\partial W_j(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda})}{\partial \lambda_j} = \log\left(\sum_{k \in \mathsf{Children}(j) \cap \mathsf{T}[\mathcal{S}]} e^{W_k(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda})/\lambda_j}\right) - \frac{1}{\lambda_j} \cdot \frac{\sum_{\ell \in \mathsf{Children}(j) \cap \mathcal{S}} W_k(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda}) \times e^{W_k(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda})/\lambda_j}}{\sum_{k \in \mathsf{Children}(j) \cap \mathsf{T}[\mathcal{S}]} e^{W_k(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda})/\lambda_j}} \;\geq\; 0\,,$$

where the inequality follows because $\log\left(\sum_{i=1} e^{x_i}\right) \geq \max_{i=1,\ldots,n} x_i$. This shows that $W_j(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})$ is increasing in $(\boldsymbol{\mu},\boldsymbol{\lambda})$, completing the induction and proving part (a).

Proof of part (b): By using induction on the height on node j, we will establish that $W_j(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})$ is strictly increasing in μ_ℓ for all $\ell \in \mathsf{T}_j[\mathcal{S}] \cap \mathcal{S}$. For the base case, suppose that j has a height of zero, so $j=\ell$ for some leaf node $\ell \in \mathsf{T}[\mathcal{S}] \cap \mathcal{S}$. We have by definition that $W_\ell(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda}) = \mu_\ell$, which is clearly strictly increasing in μ_ℓ. Because $\mathsf{T}_\ell[\mathcal{S}] \cap \mathcal{S} = \{\ell\}$, we have established the base case. Suppose the result is true for all nodes k with height at most H. Consider a non-leaf node j at height $H+1$. By definition, $W_j(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda}) = \lambda_j \log\left(\sum_{k\in \mathsf{Children}(j)\cap \mathsf{T}[\mathcal{S}]} e^{W_k(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})/\lambda_j}\right)$. Now consider a leaf node $\ell \in \mathsf{T}_j[\mathcal{S}] \cap \mathcal{S}$. There exists some child node k of j in $\mathsf{T}_j[\mathcal{S}]$ such that $\ell \in \mathsf{T}_k[\mathcal{S}] \cap \mathcal{S}$. Since the height of $k \in \mathsf{Children}(j)$ is at most H, it follows from the induction hypothesis that $W_k(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})$ is strictly increasing in μ_ℓ. Moreover, because the log-sum-exp function is strictly increasing, it follows from the expression for $W_j(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})$ that it is strictly increasing in μ_ℓ. This completes the induction, establishing the monotonicity in $\boldsymbol{\mu}$ for all nodes $j \in \mathsf{T}[\mathcal{S}]$.

Proof of part (c): Now, consider the last result that $W_j(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})$ is strictly increasing in λ_k if node k has at least two children in $\mathsf{T}_j[\mathcal{S}]$. For the base case, consider node j with height one, so its children are leaf nodes. By definition, $W_j(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda}) = \lambda_j \log\left(\sum_{\ell\in \mathsf{Children}(j)\cap\mathcal{S}} e^{\mu_\ell/\lambda_j}\right)$. Taking the derivative of the above function with respect to λ_j, we get

$$\frac{\partial W_j(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})}{\partial \lambda_j} \;=\; \log\left(\sum_{\ell\in \mathsf{Children}(j)\cap\mathcal{S}} e^{\mu_\ell/\lambda_j}\right) \;-\; \frac{1}{\lambda_j}\cdot\frac{\sum_{\ell\in \mathsf{Children}(j)\cap\mathcal{S}} \mu_\ell \times e^{\mu_\ell/\lambda_j}}{\sum_{\ell\in \mathsf{Children}(j)\cap\mathcal{S}} e^{\mu_\ell/\lambda_j}}\,. \tag{EC.4}$$

Because $|\,\mathsf{Children}(j)\cap\mathcal{S}| \geq 2$, we have that $\log(\sum_{\ell\in \mathsf{Children}(j)\cap\mathcal{S}} e^{\mu_\ell/\lambda_j}) > \max_{\ell\in \mathsf{Children}(j)\cap\mathcal{S}} \mu_\ell/\lambda_j$. Further, since the second term in the expression above is a weighted average of the set of numbers $\{\mu_\ell/\lambda_j : \ell \in \mathsf{Children}(j)\cap\mathcal{S}\}$, it follows that $\partial W_j(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})/\partial\lambda_j > 0$, which establishes the base case.

Suppose that the result is true for all nodes k with height at most H. Consider a non-leaf node j at height $H+1$. Consider an arbitrary non-leaf node $i \in \mathsf{T}_j[\mathcal{S}]$ such that i has at least two children. If $i \neq j$, then $i \in \mathsf{T}_k[\mathcal{S}]$ for some child k of j in $\mathsf{T}[\mathcal{S}]$. Because i has at least two children in $\mathsf{T}[\mathcal{S}]$, it follows from the induction hypothesis that $W_k(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})$ is strictly increasing in λ_i. Moreover, because the log-sum-exp function is strictly increasing, it follows from the expression for $W_j(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})$ that it is also strictly increasing in λ_i.

Now suppose $i=j$. Because $W_k(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})$ are independent of λ_j for all children k of j in sub-tree $\mathsf{T}[\mathcal{S}]$, we can compute the partial derivative of W_j with respect to λ_j, as done for the base case. Using identical arguments, we can conclude that $\partial W_j(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})/\partial\lambda_j > 0$, establishing the induction step. This completes the proof. ■

B.4 Proof of Lemma B.4

Proof: Fix an arbitrary $\mathcal{S} \subseteq \mathcal{N}$. We first consider the derivative with respect to μ_ℓ and prove it using induction on the height of the node. For the base case, we consider a leaf node ℓ of sub-tree $\mathsf{T}[\mathcal{S}]$. We have by definition that $W_\ell(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda}) = \mu_\ell$. Therefore, $\partial W_\ell(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})/\partial\mu_\ell = 1 = \psi_{\ell\to\ell}(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})$. We have thus established the base case.

Now suppose the result is true for all nodes of height at most H. In other words, suppose that $\partial W_v(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})/\partial\mu_\ell = \psi_{v\to\ell}(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})$ for all nodes v of height at most H and all leaf nodes ℓ in $\mathsf{T}[\mathcal{S}]$. Now consider a non-leaf node j at height $H+1$. We have by definition that

$$\frac{\partial W_j(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})}{\partial\mu_\ell} = \frac{\partial}{\partial\mu_\ell}\lambda_j \log\left(\sum_{k\in\mathsf{Children}(j)\cap\mathsf{T}[\mathcal{S}]} e^{W_k(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})/\lambda_j}\right) = \frac{\sum_{k\in\mathsf{Children}(j)\cap\mathsf{T}[\mathcal{S}]} e^{W_k(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})/\lambda_j} \times \frac{\partial W_k(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})}{\partial\mu_\ell}}{\sum_{k\in\mathsf{Children}(j)\cap\mathsf{T}[\mathcal{S}]} e^{W_k(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})/\lambda_j}}$$

Since ℓ is a leaf node in $\mathsf{T}_j[\mathcal{S}]$, there exists exactly one child node $i \in \mathsf{Children}(j)\cap\mathsf{T}[\mathcal{S}]$ such that ℓ is a leaf node in $\mathsf{T}_i[\mathcal{S}]$. For any other child node $k \neq i$, we must have that $\partial W_k(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})/\partial\mu_\ell = 0$. Therefore, we obtain

$$\frac{\partial W_j(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})}{\partial\mu_\ell} = \frac{e^{W_i(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})/\lambda_j}}{\sum_{k\in\mathsf{Children}(j)\cap\mathsf{T}[\mathcal{S}]} e^{W_k(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})/\lambda_j}} \times \frac{\partial W_i(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})}{\partial\mu_\ell} = \psi_{j\to i}(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda}) \times \psi_{i\to\ell}(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda}),$$

where the first term in the second equality above follows from the definition of $\psi_{j\to i}(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})$ and the second term follows from the induction hypothesis. Because $\psi_{j\to i}(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda}) \times \psi_{i\to\ell}(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda}) = \psi_{j\to\ell}(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})$ by definition, we have shown that $\partial W_j(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})/\partial\mu_\ell = \psi_{j\to\ell}(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})$, as desired. We have thus established the induction step. The first result now follows by induction.

We now consider the derivative of $W_j(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})$ with respect to λ_k for some non-leaf nodes j and k in $\mathsf{T}[\mathcal{S}]$. We prove the result by induction on the height of j. For the base case, we start with a non-leaf node j at height one, so all children of j are leaf nodes. As shown in Equation (EC.4) in the proof of Lemma B.3, we have that

$$\frac{\partial W_j(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})}{\partial\lambda_j} = \log\left(\sum_{\ell\in\mathsf{Children}(j)\cap\mathcal{S}} e^{\mu_\ell/\lambda_j}\right) - \frac{1}{\lambda_j}\cdot\frac{\sum_{\ell\in\mathsf{Children}(j)\cap\mathcal{S}} \mu_\ell \times e^{\mu_\ell/\lambda_j}}{\sum_{\ell\in\mathsf{Children}(j)\cap\mathcal{S}} e^{\mu_\ell/\lambda_j}} = \psi_{j\to j}(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda}) \times \Delta_j(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda}),$$

where the second equality follows from the definition of $\Delta_j(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})$ and the fact that $\psi_{j\to j}(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda}) = 1$. We have thus established the base case.

For the induction step, we suppose that $\partial W_j(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})/\partial\lambda_k = \psi_{j\to k}(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda}) \times \Delta_k(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})$ for all nodes j of height at most H and for all non-leaf nodes $k \in \mathsf{T}_j[\mathcal{S}]$. Now consider a node j of height $H+1$. First, suppose that $k \neq j$. We then have by definition that

$$\frac{\partial W_j(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})}{\partial\lambda_k} = \frac{\partial}{\partial\lambda_k}\lambda_j \log\left(\sum_{i\in\mathsf{Children}(j)\cap\mathsf{T}[\mathcal{S}]} e^{W_i(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})/\lambda_j}\right) = \frac{\sum_{i\in\mathsf{Children}(j)\cap\mathsf{T}[\mathcal{S}]} e^{W_i(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})/\lambda_j} \times \frac{\partial W_i(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})}{\partial\lambda_k}}{\sum_{i\in\mathsf{Children}(j)\cap\mathsf{T}[\mathcal{S}]} e^{W_i(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})/\lambda_j}}.$$

Because of the property of a tree, the non-leaf node k belongs to the sub-tree $\mathsf{T}_i[\mathcal{S}]$ of exactly one child node i of node j. For any other child node $v \neq i$, we have that $\partial W_v(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})/\partial\lambda_k = 0$. Therefore, we can now write

$$\frac{\partial W_j(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})}{\partial\lambda_k} = \frac{e^{W_i(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})/\lambda_j}}{\sum_{i\in\,\mathsf{Children}(j)\cap\mathsf{T}[\mathcal{S}]} e^{W_i(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})/\lambda_j}} \times \frac{\partial W_i(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})}{\partial\lambda_k} = \psi_{j\to i}(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda}) \times \psi_{i\to k}(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda}) \times \Delta_k(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda}),$$

where the first term in the second equality above follows from the definition on $\psi_{j\to i}(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})$ and the second and third terms follow from the induction hypothesis. Because $\psi_{j\to k}(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda}) = \psi_{j\to i}(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda}) \times \psi_{i\to k}(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})$ by definition, we have shown that $\partial W_j(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})/\partial\lambda_k = \psi_{j\to k}(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda}) \times \Delta_k(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})$, as desired. This completes the induction step, completing the lemma. The final expression for $\Delta_k(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})$ follows from plugging in the definitions of $W_k(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})$ and $\psi_{k\to i}(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})$ for all $i \in \mathsf{Children}(k)\cap\mathsf{T}[\mathcal{S}]$. ■

It follows from the above lemma that $\frac{\partial W_{\mathsf{root}}}{\partial\mu_\ell}(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda}) = \psi_{\mathsf{root}\to\ell}(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda}) = \mathbb{P}_\ell(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})$, and this is consistent with the result from McFadden (1978), which provides an expression of the choice probability in terms of the derivative of the generating function for Generalized Extreme Value choice models.

Appendix C: Proofs of Theorem 2.2, Theorem 2.4 and Theorem 2.5

C.1 Proof Of Theorem 2.2

The negative log-likelihood value of the data is equal to $\frac{1}{Q}\sum_{q=1}^{Q} -\log\psi^q_{\mathsf{root}\to c^q}(\boldsymbol{\mu},\boldsymbol{\lambda})$. The expression in the statement of the lemma is obtained by invoking the expression in Equation (2) for $-\log\psi^q_{\mathsf{root}\to c^q}$ and then re-arranging the terms. More precisely, we have

$$\begin{aligned}\sum_{q=1}^{Q} -\log\psi^q_{\mathsf{root}\to c^q}(\boldsymbol{\mu},\boldsymbol{\lambda}) &= \sum_{q=1}^{Q}\sum_{j\in\mathsf{path}(\mathsf{root},c^q]} \frac{W^q_{\mathsf{pa}(j)}(\boldsymbol{\mu},\boldsymbol{\lambda}) - W^q_j(\boldsymbol{\mu},\boldsymbol{\lambda})}{\lambda_{\mathsf{pa}(j)}} \\ &= \sum_{q=1}^{Q}\left\{ W^q_{\mathsf{root}}(\boldsymbol{\mu},\boldsymbol{\lambda}) + \sum_{j\in\mathsf{path}(\mathsf{root},c^q)} W^q_j(\boldsymbol{\mu},\boldsymbol{\lambda})\left(\frac{1}{\lambda_j} - \frac{1}{\lambda_{\mathsf{pa}(j)}}\right) - \frac{\mu_{c^q}}{\lambda_{\mathsf{pa}(c^q)}}\right\},\end{aligned}$$

where the last equality follows from a straightforward re-arrangement of the terms of the inner summation. A slightly different re-arrangement yields the following expression:

$$\begin{aligned}\sum_{q=1}^{Q} -\log\psi^q_{\mathsf{root}\to c^q}(\boldsymbol{\mu},\boldsymbol{\lambda}) &= \sum_{q=1}^{Q}\sum_{k\in\mathsf{path}(\mathsf{root},c^q]} \frac{W^q_{\mathsf{pa}(k)}(\boldsymbol{\mu},\boldsymbol{\lambda}) - W^q_k(\boldsymbol{\mu},\boldsymbol{\lambda})}{\lambda_{\mathsf{pa}(k)}} \\ &\overset{(a)}{=} \sum_{q=1}^{Q}\sum_{k\in\mathsf{T}\setminus\{\mathsf{root}\}} \mathbb{1}_{\{c^q\in\mathsf{T}_k\}}\frac{W^q_{\mathsf{pa}(k)}(\boldsymbol{\mu},\boldsymbol{\lambda})}{\lambda_{\mathsf{pa}(k)}} - \sum_{q=1}^{Q}\sum_{k\in\mathsf{T}\setminus\{\mathsf{root}\}} \mathbb{1}_{\{c^q\in\mathsf{T}_k\}}\frac{W^q_k(\boldsymbol{\mu},\boldsymbol{\lambda})}{\lambda_{\mathsf{pa}(k)}}, \\ &\overset{(b)}{=} \sum_{q=1}^{Q}\sum_{j\in\mathsf{T}} \mathbb{1}_{\{c^q\in\mathsf{T}_j\}}\frac{W^q_j(\boldsymbol{\mu},\boldsymbol{\lambda})}{\lambda_j} - \sum_{q=1}^{Q}\sum_{j\in\mathsf{T}} \mathbb{1}_{\{c^q\in\mathsf{T}_j\}}\frac{W^q_j(\boldsymbol{\mu},\boldsymbol{\lambda})}{\lambda_{\mathsf{pa}(j)}},\end{aligned}$$

where the equality (a) follows because $k \in \mathsf{path}(\mathsf{root}, c^q]$ if and only if c^q belongs to the sub-tree rooted at k; that is, $c^q \in \mathsf{T}_k$. The first term in the last equality (b) follows from the change of variable $j = \mathsf{pa}(k)$. With this change of variable, as k is varied over the set of nodes $\mathsf{T} \setminus \{\mathsf{root}\}$, the variable j varies over the set $\mathsf{T} \setminus \mathcal{N}$. Because $\lambda_\ell = +\infty$ for each leaf node $\ell \in \mathcal{N}$, we can extend the summation to cover the entire set of nodes in T, resulting in the first term. The second term is obtained by replacing k with j, and including root in the summation, since $\lambda_{\mathsf{pa}(\mathsf{root})} = +\infty$. ■

C.2 Proof of Theorem 2.4

To facilitate the proof of Theorem 2.4, let us introduce the following function $b : \mathcal{D}_1 \to \mathbb{R}_+$ defined by: for each $\boldsymbol{\zeta} \in \mathcal{D}_1$,

$$b(\boldsymbol{\zeta}) = \max_{q=1,\ldots,Q} \; \max_{\ell \in \mathcal{S}^q \setminus \{c^q\}} \; (\zeta_\ell - \zeta_{c^q}) \;\; ,$$

where we use the convention that the maximum of an empty set is 0. Further, we let $b^* = \min_{\boldsymbol{\zeta} \in \mathcal{D}_1 \cap \Delta} b(\boldsymbol{\zeta})$ where $\Delta = \{\boldsymbol{\mu} \mid \|\boldsymbol{\mu}\|_\infty = 1\}$. In other words, $b(\boldsymbol{\zeta})$ denotes the maximum possible loss in mean utility from the choices made in the data under the mean utility vector $\boldsymbol{\zeta}$. Maximizing the log-likelihood requires us to make this loss negative, *if* possible. If the loss is made negative for some $\boldsymbol{\zeta}$, then scaling all the mean utility values by a contant makes the loss diverge to $-\infty$, resulting in unbounded optimal solutions. As we shown below, the assumption of strong connectedness of the comparison graph prevents this from happening. In particular, strong connectedness ensures that the loss is always positive for any $\boldsymbol{\zeta}$. We formalize this intuition next. The first lemma shows that b^* is positive when the comparison graph Comp is strongly connected. Recall that, given transaction data $\{(S^q, c^q) : q = 1, \ldots, Q\}$, a comparison graph $\mathsf{Comp} = (\mathcal{N}, \mathsf{E})$ is a directed graph whose nodes correspond to the products, and there is a directed edge $(\ell_1, \ell_2) \in \mathsf{E}$ if there exists an offer set $\mathcal{S}^q$ such that $\{\ell_1, \ell_2\} \subseteq \mathcal{S}^q$ and $c^q = \ell_1$.

Lemma C.1 (Positive Gap) *If the comparison graph* Comp *is strongly connected, then* $b^* > 0$.

Proof: We start with the observation that because $\Delta \cap \mathcal{D}_1$ is compact, the minimum $\min_{\boldsymbol{\mu} \in \Delta \cap \mathcal{D}_1} b(\boldsymbol{\mu})$ is attained at some $\boldsymbol{\mu}^* \in \Delta \cap \mathcal{D}_1$, so that $b^* = b(\boldsymbol{\mu}^*)$. Therefore, to show that $b^* > 0$, it is sufficient to show that $b(\boldsymbol{\mu}) > 0$ for every $\boldsymbol{\mu} \in \Delta \cap \mathcal{D}_1$. We prove this result by contradiction. Suppose on the contrary that $b(\boldsymbol{\zeta}) \leq 0$ for some $\boldsymbol{\zeta} \in \Delta \cap \mathcal{D}_1$. Because $b(\boldsymbol{\zeta}) \leq 0$, it follows from definition that $\zeta_\ell \leq \zeta_{c^q}$ for all $\ell \in \mathcal{S}_q$ and for all $q = 1 \ldots, Q$. Consider an arbitrary directed edge (j, k) in Comp. By our construction, there exists a $q \in \{1, \ldots, Q\}$ such that $j = c^q$ and $k \in \mathcal{S}^q \setminus \{c^q\}$. Therefore, we must have that $\zeta_k \leq \zeta_j$. But, Comp is strongly connected, which implies that there is a directed path $k \to s_1 \to s_2 \to \cdots s_m \to j$ in Comp from node k to j. Applying the result that $\zeta_t \leq \zeta_s$ whenever there is a directed edge from s to t, we obtain that $\zeta_j \leq \zeta_{s_m} \leq \cdots \leq \zeta_{s_1} \leq \zeta_k$, which

implies that $\zeta_j \le \zeta_k$. This inequality together with the result that $\zeta_k \le \zeta_j$, implies that $\zeta_k = \zeta_j$. Because j and k are arbitrary nodes and any two nodes are connected by directed paths, we have shown that $\zeta_1 = \zeta_2 = \cdots = \zeta_n$. But $\boldsymbol{\zeta} \in \Delta \cap \mathcal{D}_1$, so $\zeta_1 = 0$, and consequently that $\mathbf{0} \in \Delta$. This is a contradiction! Thus, $b(\boldsymbol{\mu}) > 0$ for all $\boldsymbol{\mu} \in \Delta \cap \mathcal{D}_1$. This completes the proof. ■

The next lemma gives an upper bound on the selection probability in terms of the mean utilities.

Lemma C.2 *For each $\ell_1 \in \mathcal{S}$, $\ell_2 \in \mathcal{S}$, and $(\boldsymbol{\mu}, \boldsymbol{\lambda})$, $-\log \mathbb{P}_{\ell_1}(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda}) \ge \mu_{\ell_2} - \mu_{\ell_1}$.*

Proof: By Equation (2),

$$-\log \mathbb{P}_{\ell_1}(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda}) \;=\; -\log \psi_{\mathsf{root}\to\ell_1}(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda}) \;=\; \sum_{j \in \mathsf{path}(\mathsf{root}, \ell_1]} \frac{W_{\mathsf{pa}(j)}(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda}) - W_j(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda})}{\lambda_{\mathsf{pa}(j)}}$$
$$\ge \sum_{j \in \mathsf{path}(\mathsf{root}, \ell_1]} W_{\mathsf{pa}(j)}(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda}) - W_j(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda}) \;=\; W_{\mathsf{root}}(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda}) - W_{\ell_1}(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda}) \;,$$

where the inequality follows because $W_{\mathsf{pa}(j)}(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda}) \ge W_j(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda})$ and $\lambda_{\mathsf{pa}(j)} \le 1$ for each node j such that $j \ne \mathsf{root}$. The last equality follows from the telescoping sum.

Now, we claim that $W_{\mathsf{root}}(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda}) \ge W_\ell(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda})$ for any leaf node $\ell \in \mathcal{S}$. To see this, consider the path $\mathsf{root} \to j_1 \to \cdots \to j_m \to \ell$ from the root node to the leaf node ℓ in tree $\mathsf{T}[\mathcal{S}]$. Because $W_{\mathsf{pa}(j)}(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda}) \ge W_j(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda})$ for all nodes j, we have the sequence of inequalities $W_{\mathsf{root}}(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda}) \ge W_{j_1}(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda}) \ge \cdots \ge W_{j_m}(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda}) \ge W_\ell(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda})$, which establishes the claim.

By choosing $\ell = \ell_2$, we get

$$-\log \mathbb{P}_{\ell_1}(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda}) \;\ge\; W_{\mathsf{root}}(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda}) - W_{\ell_1}(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda}) \;\ge\; W_{\ell_2}(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda}) - W_{\ell_1}(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\lambda}) \;=\; \mu_{\ell_2} - \mu_{\ell_1} \;,$$

where the last equality follows because both ℓ_1 and ℓ_2 are in $\mathcal{S}$. ■

The next lemma establishes an upper bound on the $\mathsf{NegLog}(\mathbf{0}, \boldsymbol{\lambda})$ for all $\boldsymbol{\lambda}$.

Lemma C.3 *For each $\boldsymbol{\lambda} \in \mathcal{D}_2$, $\mathsf{NegLog}(\mathbf{0}, \boldsymbol{\lambda}) \le |\mathsf{T}| \log |\mathsf{T}|$.*

Proof: Consider an arbitrary $\boldsymbol{\lambda} \in \mathcal{D}_2$ and $\mathcal{S} \subseteq \mathcal{N}$. We will first establish the following claim.

Claim: For each node $j \in \mathsf{T}[\mathcal{S}]$, $W_j(\mathcal{S}; \mathbf{0}, \boldsymbol{\lambda})/\lambda_j \le \log |\mathsf{T}_j[\mathcal{S}]|$.

We will prove the claim by induction on the height of node j. For the base case where node j has a height of zero, then $j = \ell$ for some leaf node $\ell \in \mathcal{S}$. Since $\lambda_\ell = +\infty$ by definition, we have $W_\ell(\mathcal{S}; \mathbf{0}, \boldsymbol{\lambda})/\lambda_\ell = 0 = \log |\mathsf{T}_\ell[\mathcal{S}]|$ because $\mathsf{T}_\ell[\mathcal{S}] = \{\ell\}$. This establishes the base case. Suppose the claim holds for all nodes with height at most H. Consider an arbitrary node j with height $H+1$. By the inductive hypothesis, the claim holds for all children of j; that is, for each $k \in \mathsf{Children}(j) \cap \mathsf{T}[\mathcal{S}]$, $W_k(\mathcal{S}; \mathbf{0}, \boldsymbol{\lambda})/\lambda_k \le \log |\mathsf{T}_k[\mathcal{S}]|$, and thus, $W_k(\mathcal{S}; \mathbf{0}, \boldsymbol{\lambda})/\lambda_j \le \log |\mathsf{T}_k[\mathcal{S}]|$ because $\lambda_k \le \lambda_j$. Therefore,

$$\frac{W_j(\mathcal{S}; \mathbf{0}, \boldsymbol{\lambda})}{\lambda_j} \;=\; \log \left(\sum_{k \in \mathsf{Children}(j) \cap \mathsf{T}[\mathcal{S}]} e^{W_k(\mathcal{S}; \mathbf{0}, \boldsymbol{\lambda})/\lambda_j} \right) \;\le\; \log \left(\sum_{k \in \mathsf{Children}(j) \cap \mathsf{T}[\mathcal{S}]} |\mathsf{T}_k[\mathcal{S}]| \right) \;\le\; \log |\mathsf{T}_j[\mathcal{S}]| \;,$$

which completes the induction argument. So, the claim holds for all nodes $j \in \mathsf{T}[\mathcal{S}]$.

To establish the lemma, note that by Equation (2), for each transaction $(\mathcal{S}^q, c^q)$,

$$-\log \mathbb{P}_{c^q}(\mathcal{S}^q; \mathbf{0}, \boldsymbol{\lambda}) = -\log \psi_{\mathsf{root}\to c^q}(\mathcal{S}^q; \mathbf{0}, \boldsymbol{\lambda}) = \sum_{j \in \mathsf{path}(\mathsf{root}, c^q]} \frac{W_{\mathsf{pa}(j)}(\mathcal{S}^q; \mathbf{0}, \boldsymbol{\lambda}) - W_j(\mathcal{S}^q; \mathbf{0}, \boldsymbol{\lambda})}{\lambda_{\mathsf{pa}(j)}}$$
$$\leq \sum_{j \in \mathsf{path}(\mathsf{root}, c^q]} \frac{W_{\mathsf{pa}(j)}(\mathcal{S}^q; \mathbf{0}, \boldsymbol{\lambda})}{\lambda_{\mathsf{pa}(j)}} \leq \sum_{j \in \mathsf{path}(\mathsf{root}, c^q]} \log \left| \mathsf{T}^q_{\mathsf{pa}(j)} \right| \leq |\mathsf{T}^q| \log |\mathsf{T}^q| \ ,$$

where the first inequality follows because $W_j(\cdot)$ is non-negative, and the second inequality follows from the above claim. Therefore, $\mathsf{NegLog}(\mathbf{0}, \boldsymbol{\lambda}) \leq \frac{1}{Q} \sum_{q=1}^{Q} |\mathsf{T}^q| \log |\mathsf{T}^q| \leq |\mathsf{T}| \log |\mathsf{T}|$, which is the desired result. ■

Here is the proof of Theorem 2.4.

Proof of Theorem 2.4: Fix an arbitrary $\boldsymbol{\lambda} \in \mathcal{D}_2$. Recall that $\mathcal{D}_1 = \{\boldsymbol{\mu} \in \mathbb{R}^n : \mu_1 = 0\}$.

Proof of Necessity: We will first prove the necessity, so suppose that the optimization problem $\min_{\boldsymbol{\mu} \in \mathcal{D}_1} \mathsf{NegLog}(\boldsymbol{\mu}, \boldsymbol{\lambda})$ admits a unique and bounded optimal solution, and let $\boldsymbol{\mu}^* \in \mathcal{D}_1$ denote the unique optimal solution. We will prove by contradiction that the comparison graph Comp must be strongly connected.

Suppose on the contrary that Comp is not strongly connected. This means that there exist vertices ℓ_1 and ℓ_2 such that ℓ_2 is not reachable from ℓ_1. Let A_1 denote the set of vertices that are reachable from ℓ_1, and let $A_2 = \mathcal{N} \setminus A_1$ denote the remaining vertices in Comp. Note that $\ell_1 \in A_1$ and $\ell_2 \in A_2$, so both A_1 and A_2 are nonempty, disjoint, and $A_1 \cup A_2 = \mathcal{N}$. By definition, there is no directed path from a node in A_1 to a node in A_2. Without loss of generality, assume that $1 \in A_1$; the argument for the case where $1 \in A_2$ is analogous.

Consider an arbitrary $\xi > 0$. Define a $\bar{\boldsymbol{\mu}}$ as follows:

$$\bar{\mu}_\ell = \begin{cases} \mu^*_\ell & \text{if} \quad \ell \in A_1, \\ \mu^*_\ell + \xi & \text{if} \quad \ell \in A_2, \end{cases}$$

It is easy to verify that $\bar{\boldsymbol{\mu}} \in \mathcal{D}_1$. Now, consider an arbitrary transaction $(\mathcal{S}^q, c^q)$. There are 3 cases to consider: 1) $\mathcal{S}^q \subseteq A_1$, 2) $\mathcal{S}^q \subseteq A_2$, and 3) $\mathcal{S}^q \cap A_1 \neq \varnothing$ and $\mathcal{S}^q \cap A_2 \neq \varnothing$.

Case 1: $\mathcal{S}^q \subseteq A_1$. Then, we have that $W_j(\mathcal{S}^q; \bar{\boldsymbol{\mu}}, \boldsymbol{\lambda}) = W_j(\mathcal{S}^q; \boldsymbol{\mu}^*, \boldsymbol{\lambda})$ for all $j \in \mathsf{T}[\mathcal{S}^q]$ because the weight function only depends on the utility of the products in the offer set $\mathcal{S}^q$, and by our construction, we have that $\bar{\mu}_\ell = \mu^*_\ell$ for all $\ell \in \mathcal{S}^q \subseteq A_1$. Therefore, $-\log \psi_{\mathsf{root}\to c^q}(\mathcal{S}^q; \bar{\boldsymbol{\mu}}, \boldsymbol{\lambda}) = -\log \psi_{\mathsf{root}\to c^q}(\mathcal{S}^q; \boldsymbol{\mu}^*, \boldsymbol{\lambda})$.

Case 2: $\mathcal{S}^q \subseteq A_2$. By Equation (2), we have that

$$-\log \psi_{\mathsf{root}\to c^q}(\mathcal{S}^q; \bar{\boldsymbol{\mu}}, \boldsymbol{\lambda}) = \sum_{j \in \mathsf{path}(\mathsf{root}, c^q]} \frac{W^q_{\mathsf{pa}(j)}(\bar{\boldsymbol{\mu}}, \boldsymbol{\lambda}) - W^q_j(\bar{\boldsymbol{\mu}}, \boldsymbol{\lambda})}{\lambda_{\mathsf{pa}(j)}}$$

$$
\begin{aligned}
&\overset{(a)}{=} \frac{W_{\mathsf{root}}\left(\mathcal{S}^q;\bar{\boldsymbol{\mu}},\boldsymbol{\lambda}\right)}{\lambda_{\mathsf{root}}} + \sum_{j\in\mathsf{path}(\mathsf{root},c^q]} W_j\left(\mathcal{S}^q;\bar{\boldsymbol{\mu}},\boldsymbol{\lambda}\right)\left(\frac{1}{\lambda_j}-\frac{1}{\lambda_{\mathsf{pa}(j)}}\right)\ ,\\
&\overset{(b)}{=} \frac{W_{\mathsf{root}}\left(\mathcal{S}^q;\boldsymbol{\mu}^*,\boldsymbol{\lambda}\right)}{\lambda_{\mathsf{root}}} + \sum_{j\in\mathsf{path}(\mathsf{root},c^q]} W_j\left(\mathcal{S}^q;\boldsymbol{\mu}^*,\boldsymbol{\lambda}\right)\left(\frac{1}{\lambda_j}-\frac{1}{\lambda_{\mathsf{pa}(j)}}\right)\\
&\qquad + \frac{\xi}{\lambda_{\mathsf{root}}} + \sum_{j\in\mathsf{path}(\mathsf{root},c^q]}\left(\frac{\xi}{\lambda_j}-\frac{\xi}{\lambda_{\mathsf{pa}(j)}}\right)\\
&\overset{(c)}{=} \frac{W_{\mathsf{root}}\left(\mathcal{S}^q;\boldsymbol{\mu}^*,\boldsymbol{\lambda}\right)}{\lambda_{\mathsf{root}}} + \sum_{j\in\mathsf{path}(\mathsf{root},c^q]} W_j\left(\mathcal{S}^q;\boldsymbol{\mu}^*,\boldsymbol{\lambda}\right)\left(\frac{1}{\lambda_j}-\frac{1}{\lambda_{\mathsf{pa}(v)}}\right)\\
&= -\log\psi_{\mathsf{root}\to c^q}\left(\mathcal{S}^q;\boldsymbol{\mu}^*,\boldsymbol{\lambda}\right)\ .
\end{aligned}
$$

where the equality (a) follows from collecting terms and (b) follows because because $\bar{\mu}_\ell = \mu^*_\ell + \xi$ for all $\ell \in A_2$ and from Lemma B.1 on the invariance under translation by a constant, which implies that $W_j\left(\mathcal{S}^q;\bar{\boldsymbol{\mu}},\boldsymbol{\lambda}\right) = \xi + W_j\left(\mathcal{S}^q;\boldsymbol{\mu},\boldsymbol{\lambda}\right)$. The last equality (c) follows by the telescoping sum and the fact that $\lambda_\ell = +\infty$ for all leaf nodes $\ell \in \mathcal{N}$.

Case 3: $\mathcal{S}^q \cap A_1 \neq \varnothing$ and $\mathcal{S}^q \cap A_2 \neq \varnothing$. In this case, $|\mathcal{S}^q| \geq 2$, and thus, by our construction of the comparison graph Comp, there is a directed edge from c^q to all other elements in $\mathcal{S}^q$. Since there is no directed path from A_1 to A_2, it must be the case that $c^q \in A_2$. By re-arranging the terms in Equation (2) and using the fact that $\lambda_\ell = +\infty$ for all leaf nodes $\ell \in \mathcal{N}$, we can write

$$
\begin{aligned}
-\log\psi_{\mathsf{root}\to c^q}\left(\mathcal{S}^q;\bar{\boldsymbol{\mu}},\boldsymbol{\lambda}\right) &= \frac{W_{\mathsf{root}}\left(\mathcal{S}^q;\bar{\boldsymbol{\mu}},\boldsymbol{\lambda}\right)}{\lambda_{\mathsf{root}}} + \sum_{j\in\mathsf{path}(\mathsf{root},c^q)} W_j\left(\mathcal{S}^q;\bar{\boldsymbol{\mu}},\boldsymbol{\lambda}\right)\left(\frac{1}{\lambda_j}-\frac{1}{\lambda_{\mathsf{pa}(j)}}\right) - \frac{W_{c^q}\left(\mathcal{S}^q;\bar{\boldsymbol{\mu}},\boldsymbol{\lambda}\right)}{\lambda_{\mathsf{pa}(c^q)}}\\
&\overset{(a)}{=} \frac{W_{\mathsf{root}}\left(\mathcal{S}^q;\bar{\boldsymbol{\mu}},\boldsymbol{\lambda}\right)}{\lambda_{\mathsf{root}}} + \sum_{j\in\mathsf{path}(\mathsf{root},c^q)} W_j\left(\mathcal{S}^q;\bar{\boldsymbol{\mu}},\boldsymbol{\lambda}\right)\left(\frac{1}{\lambda_j}-\frac{1}{\lambda_{\mathsf{pa}(j)}}\right)\\
&\qquad - \frac{W_{c^q}\left(\mathcal{S}^q;\boldsymbol{\mu}^*,\boldsymbol{\lambda}\right)}{\lambda_{\mathsf{pa}(c^q)}} - \frac{\xi}{\lambda_{\mathsf{pa}(c^q)}}\\
&\overset{(b)}{\leq} \frac{W_{\mathsf{root}}\left(\mathcal{S}^q;\boldsymbol{\mu}^*,\boldsymbol{\lambda}\right)}{\lambda_{\mathsf{root}}} + \sum_{j\in\mathsf{path}(\mathsf{root},c^q)} W_j\left(\mathcal{S}^q;\boldsymbol{\mu}^*,\boldsymbol{\lambda}\right)\left(\frac{1}{\lambda_j}-\frac{1}{\lambda_{\mathsf{pa}(j)}}\right)\\
&\qquad + \frac{\xi}{\lambda_{\mathsf{root}}} + \sum_{j\in\mathsf{path}(\mathsf{root},c^q)}\left(\frac{\xi}{\lambda_j}-\frac{\xi}{\lambda_{\mathsf{pa}(j)}}\right) - \frac{W_{c^q}\left(\mathcal{S}^q;\boldsymbol{\mu}^*,\boldsymbol{\lambda}\right)}{\lambda_{\mathsf{pa}(c^q)}} - \frac{\xi}{\lambda_{\mathsf{pa}(c^q)}}\\
&\overset{(c)}{=} -\log\psi_{\mathsf{root}\to c^q}\left(\mathcal{S}^q;\boldsymbol{\mu}^*,\boldsymbol{\lambda}\right)\ .
\end{aligned}
$$

where (a) follows because $c_q \in A_2$ for all q, so $W_{c^q}\left(\mathcal{S}^q;\bar{\boldsymbol{\mu}},\boldsymbol{\lambda}\right) = \bar{\mu}_{c^q} = \mu^*_{c^q} + \xi$. The inequality (b) above follows from the fact for all $j \in \mathsf{path}(\mathsf{root}, c^q)$,

$$
W_j\left(\mathcal{S}^q;\bar{\boldsymbol{\mu}},\boldsymbol{\lambda}\right)\left(\frac{1}{\lambda_j}-\frac{1}{\lambda_{\mathsf{pa}(j)}}\right) \ \leq\ W_j\left(\mathcal{S}^q;\boldsymbol{\mu}^*,\boldsymbol{\lambda}\right)\left(\frac{1}{\lambda_j}-\frac{1}{\lambda_{\mathsf{pa}(j)}}\right) \ +\ \xi\left(\frac{1}{\lambda_j}-\frac{1}{\lambda_{\mathsf{pa}(j)}}\right)
$$

because $0 < \lambda_j \leq \lambda_{\mathsf{pa}(j)}$ and by Lemma B.3, the weight function is increasing in $\boldsymbol{\mu}$, so for all $j \in \mathsf{T}[\mathcal{S}^q]$, $W_j\left(\mathcal{S}^q;\bar{\boldsymbol{\mu}},\boldsymbol{\lambda}\right) \ \leq\ W_j\left(\mathcal{S}^q;\boldsymbol{\mu}^* + \xi\boldsymbol{e},\boldsymbol{\lambda}\right) \ =\ \xi + W_j\left(\mathcal{S}^q;\boldsymbol{\mu}^*,\boldsymbol{\lambda}\right)$, where the equality follows from the

translation variance property. The final equality (c) above follows from collecting terms in the telescoping sum.

Therefore, we observe that in all three cases,

$$-\log \psi_{\mathsf{root}\to c^q}\left(\mathcal{S}^q;\bar{\boldsymbol{\mu}},\boldsymbol{\lambda}\right) \;\leq\; -\log \psi_{\mathsf{root}\to c^q}\left(\mathcal{S}^q;\boldsymbol{\mu}^*,\boldsymbol{\lambda}\right) \;.$$

Since the transaction $(\mathcal{S}^q, c^q)$ is arbitrary, it follows that $\mathsf{NegLog}(\bar{\boldsymbol{\mu}},\boldsymbol{\lambda}) \leq \mathsf{NegLog}(\boldsymbol{\mu}^*,\boldsymbol{\lambda})$, but this contradicts the fact that $\boldsymbol{\mu}^*$ is the unique optimal solution! Hence, it must be the case that Comp is strongly connected. This completes the proof of the necessity.

Proof of Sufficiency: We will now prove the sufficiency, so assume that the comparison graph Comp is strongly connected. Fix an arbitrary $\boldsymbol{\lambda}$. Consider $\boldsymbol{\mu} \in \mathcal{D}_1$ such that $\|\boldsymbol{\mu}\|_\infty > \frac{Q}{b^*}\mathsf{NegLog}(\mathbf{0},\boldsymbol{\lambda})$. Note that $b^* > 0$ by Lemma C.1. Then, $\boldsymbol{\mu} = \|\boldsymbol{\mu}\|_\infty \boldsymbol{\zeta}$ where $\|\boldsymbol{\zeta}\|_\infty = 1$. By definition, there exists a transaction $(\mathcal{S}^q, c^q)$ such that $b(\boldsymbol{\zeta}) = \zeta_\ell \;-\; \zeta_{c^q}$ for some $\ell \in \mathcal{S}^q \setminus \{c^q\}$. Then, by Lemma C.2

$$\mathsf{NegLog}(\boldsymbol{\mu},\boldsymbol{\lambda}) \;\geq\; -\log \mathbb{P}_{c^q}\left(\mathcal{S}^q;\boldsymbol{\mu},\boldsymbol{\lambda}\right)/Q \;\geq\; \left(\mu_\ell - \mu_{c^q}\right)/Q \;=\; \left(\zeta_\ell - \zeta_{c^q}\right)\|\boldsymbol{\mu}\|_\infty/Q \;=\; b(\boldsymbol{\zeta})\,\|\boldsymbol{\mu}\|_\infty/Q,$$

and it follows that $\mathsf{NegLog}(\boldsymbol{\mu},\boldsymbol{\lambda}) \;\geq\; \|\boldsymbol{\mu}\|_\infty\, b(\boldsymbol{\zeta})/Q \;\geq\; \|\boldsymbol{\mu}\|_\infty\, b^*/Q \;>\; \mathsf{NegLog}(\mathbf{0},\boldsymbol{\lambda})$, which shows that $\min_{\boldsymbol{\mu}\in\mathcal{D}_1}\mathsf{NegLog}(\boldsymbol{\mu},\boldsymbol{\lambda}) = \min\left\{\mathsf{NegLog}(\boldsymbol{\mu},\boldsymbol{\lambda}) \,\middle|\, \boldsymbol{\mu}\in\mathcal{D}_1,\;\; \|\boldsymbol{\mu}\|_\infty \leq \frac{Q}{b^*}\mathsf{NegLog}(\mathbf{0},\boldsymbol{\lambda})\right\}$, so the optimization problem has a bounded solution.

We now show that the optimal solution must be unique by establishing that the mapping $\boldsymbol{\mu} \mapsto \mathsf{NegLog}(\boldsymbol{\mu},\boldsymbol{\lambda})$ is strictly convex over the set $\mathcal{D}_1$. For that, we first note from Theorem 2.2 that

$$\mathsf{NegLog}(\boldsymbol{\mu},\boldsymbol{\lambda}) = \frac{1}{Q}\sum_{q=1}^{Q}\left\{ W^q_{\mathsf{root}}\left(\boldsymbol{\mu},\boldsymbol{\lambda}\right) \;+\; \sum_{j\in\mathsf{path}(\mathsf{root},c^q)} W^q_j\left(\boldsymbol{\mu},\boldsymbol{\lambda}\right)\left(\frac{1}{\lambda_j} - \frac{1}{\lambda_{\mathsf{pa}(j)}}\right) \;-\; \frac{\mu_{c^q}}{\lambda_{\mathsf{pa}(c^q)}}\right\}$$

Because $W^q_j(\boldsymbol{\mu},\boldsymbol{\lambda})$ is convex in $\boldsymbol{\mu}$ for each q and $\boldsymbol{\lambda}$ (by Lemma B.2) and $1/\lambda_j - 1/\lambda_{\mathsf{pa}(j)} \geq 0$ for all $j \in \mathsf{T}\setminus\mathcal{N}$, it follows that the function above is convex in $\boldsymbol{\mu}$.

We are left to show that the convexity is strict, for which it is sufficient to show that the following function is strictly convex: $f(\boldsymbol{\mu}) \stackrel{\text{def}}{=} \sum_{q=1}^{Q} W^q_{\mathsf{root}}(\boldsymbol{\mu},\boldsymbol{\lambda})$. We show strict convexity through a direct application of its definition. The function $f(\cdot)$ is strictly convex if for $\boldsymbol{\mu} \neq \bar{\boldsymbol{\mu}}$ such that $\mu_1 = \bar{\mu}_1 = 0$ and $\theta \in (0,1)$, we have $f(\theta\boldsymbol{\mu} + (1-\theta)\bar{\boldsymbol{\mu}}) < \theta f(\boldsymbol{\mu}) + (1-\theta) f(\bar{\boldsymbol{\mu}})$. To arrive at a contradiction, suppose that for some $\boldsymbol{\mu} \neq \bar{\boldsymbol{\mu}}$ and $\theta \in (0,1)$, we have that

$$\begin{aligned} & f(\theta\boldsymbol{\mu} + (1-\theta)\bar{\boldsymbol{\mu}}) = \theta f(\boldsymbol{\mu}) + (1-\theta) f(\bar{\boldsymbol{\mu}}) \\ \iff\;\; & \sum_{q=1}^{Q} W^q_{\mathsf{root}}(\theta\boldsymbol{\mu} + (1-\theta)\bar{\boldsymbol{\mu}},\boldsymbol{\lambda}) = \sum_{q=1}^{Q} \theta W^q_{\mathsf{root}}(\boldsymbol{\mu},\boldsymbol{\lambda}) + (1-\theta) W^q_{\mathsf{root}}(\bar{\boldsymbol{\mu}},\boldsymbol{\lambda}) \;. \end{aligned} \tag{EC.5}$$

By Lemma B.2, for each $q \in \{1,\ldots,Q\}$, $W^q_{\mathsf{root}}(\theta\boldsymbol{\mu} + (1-\theta)\bar{\boldsymbol{\mu}},\boldsymbol{\lambda}) \leq \theta W^q_{\mathsf{root}}(\boldsymbol{\mu},\boldsymbol{\lambda}) + (1-\theta) W^q_{\mathsf{root}}(\bar{\boldsymbol{\mu}},\boldsymbol{\lambda})$, with equality occurring if and only if $\mu_\ell = \bar{\mu}_\ell + \kappa^q$ for all $\ell \in \mathcal{S}^q$, where $\kappa^q \in \mathbb{R}$ is some constant.

Therefore, for Equation (EC.5) to be satisfied, we must have equality occurring for each $q \in \{1,\ldots,Q\}$. As a result, there exists $\kappa^1,\kappa^2,\ldots,\kappa^Q$ such that for each $q \in \{1,\ldots,Q\}$,

$$\mu_\ell = \bar{\mu}_\ell + \kappa^q \text{ for all } \ell \in \mathcal{S}^q.$$

To arrive at a contradiction, we now establish that $\kappa^q = 0$ for all q, which implies that $\boldsymbol{\mu} = \bar{\boldsymbol{\mu}}$ since $\mathcal{N} = \cup_{q\in[Q]}\mathcal{S}^q$. This contradicts our assumption that $\boldsymbol{\mu} \neq \bar{\boldsymbol{\mu}}$!

To show that $\kappa^q = 0$ for all $q = 1,\ldots,Q$, we first prove the following claim.

Claim: $\kappa^1 = \kappa^2 = \cdots = \kappa^Q$.

To prove the claim, it suffices to show that $\kappa^1 = \kappa^2$. Exactly the same argument applies to show that $\kappa^{q_1} = \kappa^{q_2}$ for all $q_1 \neq q_2$. If $\mathcal{S}^1 \cap \mathcal{S}^2 \neq \varnothing$, the result is trivially true because there exists $\ell \in \mathcal{S}^1 \cap \mathcal{S}^2$, which implies that $\kappa^1 = \mu_\ell - \bar{\mu}_\ell = \kappa^2$, which is the desired result. So, suppose that $\mathcal{S}^1 \cap \mathcal{S}^2 = \varnothing$. Pick $\ell_1 \in \mathcal{S}^1$ and $\ell_2 \in \mathcal{S}^2$. Since Comp is strongly connected, there exists a path $\ell_1 = j_0 \to j_1 \to \cdots \to j_{m-1} \to j_m = \ell_2$ in Comp from ℓ_1 to ℓ_2. By our construction of Comp, for each edge (j_{t-1}, j_t), there exists a set $\mathcal{S}^{h_t} \in \{\mathcal{S}^1,\ldots,\mathcal{S}^Q\}$ such that $\{j_{t-1}, j_t\} \subseteq \mathcal{S}^{h_t}$. Therefore, for each $t = 1,\ldots,m$, $j_t \in \mathcal{S}^{h_t} \cap \mathcal{S}^{h_{t+1}}$, and thus, $\kappa^{h_t} = \mu_{j_t} - \bar{\mu}_{j_t} = \kappa^{h_{t+1}}$. Since this is true for all $t = 1,\ldots,m$, it follows that

$$\kappa^{h_1} = \kappa^{h_2} = \cdots = \kappa^{h_m} \ .$$

Since $\ell_1 \in \mathcal{S}^1 \cap \mathcal{S}^{h_1}$ and $\ell_2 \in \mathcal{S}^2 \cap \mathcal{S}^{h_m}$, we have that $\kappa^1 = \kappa^{h_1} = \kappa^{h_m} = \kappa^2$. This is the desired result, proving the claim.

We will now use the above claim to show that $\kappa^q = 0$ for all q. Consider the subset $\mathcal{S}^q$ that contains product 1. Because $\mu_1 = \bar{\mu}_1 = 0$, it follows that $\kappa^q = 0$. Because $\kappa^1 = \ldots \kappa^Q$, it follows that $\kappa^q = 0$ for all $q = 1,\ldots,Q$. This completes the proof of sufficiency.

■

C.3 Proof of Theorem 2.5

Proof: We will show that for each non-leaf node $j \in \mathsf{T} \setminus \mathcal{N}$ such that $j \neq \mathsf{root}$, the function NegLog function varies with respect to λ_j if and only if there exists a transaction $q \in \{1,\ldots,Q\}$ such that node j has at least two children in sub-tree T^q. Consider an arbitrary non-leaf node $j \neq \mathsf{root}$. The sufficiency follows immediately from Lemma B.3. To establish the necessity, assume that NegLog function varies with respect to λ_j. Suppose, on the contrary, that for every transaction q, node j has at most one child in the tree T^q. Then, by definition, for every q,

$$W_j^q(\boldsymbol{\mu},\boldsymbol{\lambda}) \ = \ \lambda_j \log\left(\sum_{k\in \mathsf{Children}(j)\cap \mathsf{T}^q} e^{W_k(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda})/\lambda_j}\right) = \sum_{k\in \mathsf{Children}(j)\cap \mathsf{T}^q} W_k(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\lambda}) \ ,$$

where the last equality follows because there is at most one term in the summand. Therefore, for every q, $W_j^q(\boldsymbol{\mu},\boldsymbol{\lambda})$ is independent of λ_j, which implies that $-\log \psi^q_{\mathsf{root}\to c^q}(\boldsymbol{\mu},\boldsymbol{\lambda})$ is independent of λ_j, which implies that NegLog function is also independent of λ_j. Contradiction! Therefore, there must be at least one transaction q such that node j has at least two children in the tree T^q. This proves the necessity. ■

Appendix D: Derivatives of the negative log-likelihood function $\mathsf{NegLog}(\boldsymbol{\mu},\boldsymbol{\lambda})$

The following lemmas show that the partial derivatives $(\partial\mathsf{NegLog}(\boldsymbol{\mu},\boldsymbol{\lambda})/\partial\mu_\ell \,:\, \ell\in\mathcal{N})$ and $(\partial\mathsf{NegLog}(\boldsymbol{\mu},\boldsymbol{\lambda})/\partial\lambda_j \,:\, j\in\mathsf{T}\setminus(\mathcal{N}\cup\{\mathsf{root}\}))$ can be computed efficiently via a recursion that starts at the root node and ends at the leaf nodes of the tree T. To facilitate our exposition, for each transaction q, define the set of values $\left(D_j^q(\boldsymbol{\mu},\boldsymbol{\lambda}) \colon j\in\mathsf{T}\right)$ recursively starting from root as follows: Initialize $D^q_{\mathsf{root}}(\boldsymbol{\mu},\boldsymbol{\lambda})=1$ and for all $j\in\mathsf{T}\setminus\{\mathsf{root}\}$,

$$D_j^q(\boldsymbol{\mu},\boldsymbol{\lambda}) \equiv \mathbb{1}_{\{c^q\in\mathsf{T}_j\}}\cdot\left(\frac{1}{\lambda_j}-\frac{1}{\lambda_{\mathsf{pa}(j)}}\right)+\psi^q_{\mathsf{pa}(j)\to j}(\boldsymbol{\mu},\boldsymbol{\lambda})\cdot D^q_{\mathsf{pa}(j)}(\boldsymbol{\mu},\boldsymbol{\lambda})\ .$$

We emphasize that the values $D_j^q(\boldsymbol{\mu},\boldsymbol{\lambda})$ can be computed efficiently through a simple recursion starting from the root node. We will use the values $D_j^q(\boldsymbol{\mu},\boldsymbol{\lambda})$ to compute the derivative of the log-likeihood function; see Lemmas D.2 and D.3. The following lemma gives an equivalent expression for $D_j^q(\boldsymbol{\mu},\boldsymbol{\lambda})$.

Lemma D.1 (Equivalent Expression for $D_j^q(\boldsymbol{\mu},\boldsymbol{\lambda})$) *For each transaction q and for all $k\in\mathsf{T}$, $D_k^q(\boldsymbol{\mu},\boldsymbol{\lambda})=\sum_{j\in\mathsf{T}}\mathbb{1}_{\{c^q\in\mathsf{T}_j\}}\left(\frac{1}{\lambda_j}-\frac{1}{\lambda_{\mathsf{pa}(j)}}\right)\cdot\psi^q_{j\to k}(\boldsymbol{\mu},\boldsymbol{\lambda})$.*

Proof: We prove the result by induction on the height of node k. When $k=\mathsf{root}$, the RHS in the equality above reduces to

$$\begin{aligned}\sum_{j\in\mathsf{T}}\mathbb{1}_{\{c^q\in\mathsf{T}_j\}}\left(\frac{1}{\lambda_j}-\frac{1}{\lambda_{\mathsf{pa}(j)}}\right)\cdot\psi^q_{j\to\mathsf{root}}(\boldsymbol{\mu},\boldsymbol{\lambda}) &= \left(\frac{1}{\lambda_{\mathsf{root}}}-\frac{1}{\lambda_{\mathsf{pa}(\mathsf{root})}}\right)\cdot\psi^q_{\mathsf{root}\to\mathsf{root}}(\boldsymbol{\mu},\boldsymbol{\lambda})\\ &= \psi^q_{\mathsf{root}\to\mathsf{root}}(\boldsymbol{\mu},\boldsymbol{\lambda}) \;=\; 1 \;=\; D^q_{\mathsf{root}}(\boldsymbol{\mu},\boldsymbol{\lambda}),\end{aligned}$$

where the first equality follows since $\psi^q_{j\to\mathsf{root}}(\boldsymbol{\mu},\boldsymbol{\lambda})=0$ for all $j\neq\mathsf{root}$, and the second follows since $\lambda_{\mathsf{pa}(\mathsf{root})}=+\infty$. This establishes the base case. Now, suppose the claim is true for all nodes k of height H. Consider any node k with height $H-1$. Using the definition of $D_k^q(\boldsymbol{\mu},\boldsymbol{\lambda})$, it follows that

$$\begin{aligned}D_k^q(\boldsymbol{\mu},\boldsymbol{\lambda}) &= \mathbb{1}_{\{c^q\in\mathsf{T}_k\}}\cdot\left(\frac{1}{\lambda_k}-\frac{1}{\lambda_{\mathsf{pa}(k)}}\right)+\psi^q_{\mathsf{pa}(k)\to k}(\boldsymbol{\mu},\boldsymbol{\lambda})\cdot D^q_{\mathsf{pa}(k)}(\boldsymbol{\mu},\boldsymbol{\lambda})\\ &= \mathbb{1}_{\{c^q\in\mathsf{T}_k\}}\cdot\left(\frac{1}{\lambda_k}-\frac{1}{\lambda_{\mathsf{pa}(k)}}\right)+\psi^q_{\mathsf{pa}(k)\to k}(\boldsymbol{\mu},\boldsymbol{\lambda})\cdot\left(\sum_{j\in\mathsf{T}}\mathbb{1}_{\{c^q\in\mathsf{T}_j\}}\left(\frac{1}{\lambda_j}-\frac{1}{\lambda_{\mathsf{pa}(j)}}\right)\cdot\psi^q_{j\to\mathsf{pa}(k)}(\boldsymbol{\mu},\boldsymbol{\lambda})\right)\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{1}_{\{c^q \in \mathsf{T}_k\}} \cdot \left(\frac{1}{\lambda_k} - \frac{1}{\lambda_{\mathsf{pa}(k)}}\right) + \psi^q_{\mathsf{pa}(k)\to k}(\boldsymbol{\mu},\boldsymbol{\lambda}) \cdot \left(\sum_{j\in\mathsf{T}\setminus\{k\}} \mathbb{1}_{\{c^q\in\mathsf{T}_j\}} \left(\frac{1}{\lambda_j} - \frac{1}{\lambda_{\mathsf{pa}(j)}}\right) \cdot \psi^q_{j\to\mathsf{pa}(k)}(\boldsymbol{\mu},\boldsymbol{\lambda})\right) \\
&= \mathbb{1}_{\{c^q \in \mathsf{T}_k\}} \cdot \left(\frac{1}{\lambda_k} - \frac{1}{\lambda_{\mathsf{pa}(k)}}\right) + \sum_{j\in\mathsf{T}\setminus\{k\}} \mathbb{1}_{\{c^q\in\mathsf{T}_j\}} \left(\frac{1}{\lambda_j} - \frac{1}{\lambda_{\mathsf{pa}(j)}}\right) \cdot \psi^q_{j\to k}(\boldsymbol{\mu},\boldsymbol{\lambda}) \\
&= \sum_{j\in\mathsf{T}} \mathbb{1}_{\{c^q\in\mathsf{T}_j\}} \left(\frac{1}{\lambda_j} - \frac{1}{\lambda_{\mathsf{pa}(j)}}\right) \cdot \psi^q_{j\to k}(\boldsymbol{\mu},\boldsymbol{\lambda}),
\end{aligned}
$$

where the second equality follows from the induction hypothesis, the third since $\psi^q_{k\to\mathsf{pa}(k)}(\boldsymbol{\mu},\boldsymbol{\lambda}) = 0$, the fourth since $\psi^q_{j\to\mathsf{pa}(k)}(\boldsymbol{\mu},\boldsymbol{\lambda}) \times \psi^q_{\mathsf{pa}(k)\to k}(\boldsymbol{\mu},\boldsymbol{\lambda}) = \psi^q_{j\to k}(\boldsymbol{\mu},\boldsymbol{\lambda})$ for all $j \in \mathsf{path}[\mathsf{root}, k)$, and the final since $\psi^q_{k\to k}(\boldsymbol{\mu},\boldsymbol{\lambda}) = 1$. This completes the induction. ■

Here is the derivative of the negative log-likelihood function with respect to $\boldsymbol{\mu}$.

Lemma D.2 (Derivatives with Respect to $\boldsymbol{\mu}$) *For all $\ell \in \mathcal{N}$,*

$$
\frac{\partial \mathsf{NegLog}(\boldsymbol{\mu},\boldsymbol{\lambda})}{\partial \mu_\ell} \;=\; \frac{1}{Q}\sum_{q=1}^{Q} D^q_\ell(\boldsymbol{\mu},\boldsymbol{\lambda}) \;=\; \frac{1}{Q}\sum_{q=1}^{Q} \sum_{j\in\mathsf{T}\setminus\mathcal{N}} \mathbb{1}_{\{c^q\in\mathsf{T}_j\}} \cdot \left(\frac{1}{\lambda_j} - \frac{1}{\lambda_{\mathsf{pa}(j)}}\right) \cdot \psi^q_{j\to\ell}(\boldsymbol{\mu},\boldsymbol{\lambda}) - \frac{\mathsf{sales}_\ell}{\lambda_{\mathsf{pa}(\ell)}}
$$

Proof. From Theorem 2.2, it follows that

$$
\mathsf{NegLog}(\boldsymbol{\mu},\boldsymbol{\lambda}) = \frac{1}{Q}\sum_{q=1}^{Q} \left\{ \sum_{j\in\mathsf{T}} \mathbb{1}_{\{c^q\in\mathsf{T}_j\}} \left(\frac{1}{\lambda_j} - \frac{1}{\lambda_{\mathsf{pa}(j)}}\right) W^q_j(\boldsymbol{\mu},\boldsymbol{\lambda}) \right\}.
$$

Now, noting that $\partial W^q_j(\boldsymbol{\mu},\boldsymbol{\lambda})/\partial\mu_\ell$ is equal to $\psi^q_{j\to\ell}(\boldsymbol{\mu},\boldsymbol{\lambda})$ (from Lemma B.4 in Appendix B), it follows that for all $\ell \in \mathcal{N}$:

$$
\frac{\partial \mathsf{NegLog}(\boldsymbol{\mu},\boldsymbol{\lambda})}{\partial \mu_\ell} = \frac{1}{Q}\sum_{q=1}^{Q} \left\{ \sum_{j\in\mathsf{T}} \mathbb{1}_{\{c^q\in\mathsf{T}_j\}} \left(\frac{1}{\lambda_j} - \frac{1}{\lambda_{\mathsf{pa}(j)}}\right) \psi^q_{j\to\ell}(\boldsymbol{\mu},\boldsymbol{\lambda}) \right\} = \frac{1}{Q}\sum_{q=1}^{Q} D^q_\ell(\boldsymbol{\mu},\boldsymbol{\lambda}),
$$

where the second equality follows from Lemma D.1. Finally, we note that

$$
\begin{aligned}
\frac{1}{Q}\sum_{q=1}^{Q} D^q_\ell(\boldsymbol{\mu},\boldsymbol{\lambda}) &= \frac{1}{Q}\sum_{q=1}^{Q} \left\{ \sum_{j\in\mathsf{T}} \mathbb{1}_{\{c^q\in\mathsf{T}_j\}} \left(\frac{1}{\lambda_j} - \frac{1}{\lambda_{\mathsf{pa}(j)}}\right) \psi^q_{j\to\ell}(\boldsymbol{\mu},\boldsymbol{\lambda}) \right\} \\
&= \frac{1}{Q}\sum_{q=1}^{Q} \left\{ \sum_{j\in\mathsf{T}\setminus\mathcal{N}} \mathbb{1}_{\{c^q\in\mathsf{T}_j\}} \left(\frac{1}{\lambda_j} - \frac{1}{\lambda_{\mathsf{pa}(j)}}\right) \psi^q_{j\to\ell}(\boldsymbol{\mu},\boldsymbol{\lambda}) + \sum_{j\in\mathcal{N}} \mathbb{1}_{\{c^q\in\mathsf{T}_j\}} \left(\frac{1}{\lambda_j} - \frac{1}{\lambda_{\mathsf{pa}(j)}}\right) \psi^q_{j\to\ell}(\boldsymbol{\mu},\boldsymbol{\lambda}) \right\} \\
&= \frac{1}{Q}\sum_{q=1}^{Q} \left\{ \sum_{j\in\mathsf{T}\setminus\mathcal{N}} \mathbb{1}_{\{c^q\in\mathsf{T}_j\}} \left(\frac{1}{\lambda_j} - \frac{1}{\lambda_{\mathsf{pa}(j)}}\right) \psi^q_{j\to\ell}(\boldsymbol{\mu},\boldsymbol{\lambda}) - \frac{\mathbb{1}_{\{c^q=\ell\}}}{\lambda_{\mathsf{pa}(\ell)}} \right\} \\
&= \frac{1}{Q}\sum_{q=1}^{Q} \left\{ \sum_{j\in\mathsf{T}\setminus\mathcal{N}} \mathbb{1}_{\{c^q\in\mathsf{T}_j\}} \left(\frac{1}{\lambda_j} - \frac{1}{\lambda_{\mathsf{pa}(j)}}\right) \psi^q_{j\to\ell}(\boldsymbol{\mu},\boldsymbol{\lambda}) \right\} - \frac{1}{Q}\sum_{q=1}^{Q} \frac{\mathbb{1}_{\{c^q=\ell\}}}{\lambda_{\mathsf{pa}(\ell)}} \\
&= \frac{1}{Q}\sum_{q=1}^{Q} \sum_{j\in\mathsf{T}\setminus\mathcal{N}} \mathbb{1}_{\{c^q\in\mathsf{T}_j\}} \cdot \left(\frac{1}{\lambda_j} - \frac{1}{\lambda_{\mathsf{pa}(j)}}\right) \cdot \psi^q_{j\to\ell}(\boldsymbol{\mu},\boldsymbol{\lambda}) - \frac{\mathsf{sales}_\ell}{\lambda_{\mathsf{pa}(\ell)}},
\end{aligned}
$$

where the third equality follows since $\lambda_j = +\infty$ for all $j \in \mathcal{N}$ and $\psi^q_{j\to\ell} = 0$ for all $j \in \mathcal{N}; j \neq \ell$, and the last follows from the definition of sales_ℓ. ■

The next lemma gives the derivative with respect to $\boldsymbol{\lambda}$.

Lemma D.3 (Derivatives with Respect to $\boldsymbol{\lambda}$) *For all $k \in \mathsf{T} \setminus (\mathcal{N} \cup \{\mathsf{root}\})$,*

$$\frac{\partial \mathsf{NegLog}(\boldsymbol{\mu}, \boldsymbol{\lambda})}{\partial \lambda_k} = \frac{1}{Q} \sum_{q=1}^{Q} \Delta_k^q(\boldsymbol{\mu}, \boldsymbol{\lambda}) \cdot D_k^q(\boldsymbol{\mu}, \boldsymbol{\lambda}) + E_k^q(\boldsymbol{\mu}, \boldsymbol{\lambda}),$$

where for all q, $\Delta_k^q(\boldsymbol{\mu}, \boldsymbol{\lambda}) := \Delta_k(S^q; \boldsymbol{\mu}, \boldsymbol{\lambda})$ is as defined in Lemma B.4, and

$$E_k^q(\boldsymbol{\mu}, \boldsymbol{\lambda}) = \frac{\sum_{i \in \mathsf{Children}(k)} \mathbb{1}_{\{c^q \in \mathsf{T}_i\}} \log\left(\psi_{k \to i}^q(\boldsymbol{\mu}, \boldsymbol{\lambda})\right)}{\lambda_k}.$$

Proof. From Theorem 2.2, it follows that

$$\mathsf{NegLog}(\boldsymbol{\mu}, \boldsymbol{\lambda}) = \frac{1}{Q} \sum_{q=1}^{Q} \sum_{j \in \mathsf{T}} \mathbb{1}_{\{c^q \in \mathsf{T}_j\}} \frac{W_j^q(\boldsymbol{\mu}, \boldsymbol{\lambda})}{\lambda_j} \quad - \quad \frac{1}{Q} \sum_{q=1}^{Q} \sum_{j \in \mathsf{T}} \mathbb{1}_{\{c^q \in \mathsf{T}_j\}} \frac{W_j^q(\boldsymbol{\mu}, \boldsymbol{\lambda})}{\lambda_{\mathsf{pa}(j)}}$$

From the quotient rule of derivatives, it follows that

$$\frac{\partial}{\partial \lambda_k} \left(\frac{W_j^q(\boldsymbol{\mu}, \boldsymbol{\lambda})}{\lambda_j} \right) = \frac{1}{\lambda_j} \cdot \frac{\partial W_j^q(\boldsymbol{\mu}, \boldsymbol{\lambda})}{\partial \lambda_k} - \mathbb{1}_{\{j=k\}} \frac{W_k^q(\boldsymbol{\mu}, \boldsymbol{\lambda})}{\lambda_k^2}$$

$$\frac{\partial}{\partial \lambda_k} \left(\frac{W_j^q(\boldsymbol{\mu}, \boldsymbol{\lambda})}{\lambda_{\mathsf{pa}(j)}} \right) = \frac{1}{\lambda_{\mathsf{pa}(j)}} \cdot \frac{\partial W_j^q(\boldsymbol{\mu}, \boldsymbol{\lambda})}{\partial \lambda_k} - \mathbb{1}_{\{\mathsf{pa}(j)=k\}} \frac{W_j^q(\boldsymbol{\mu}, \boldsymbol{\lambda})}{\lambda_k^2}$$

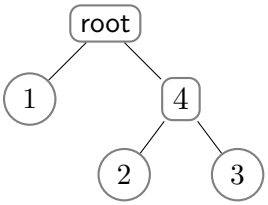
Using the above and the expression for $\frac{\partial W_j^q(\boldsymbol{\mu}, \boldsymbol{\lambda})}{\partial \lambda_k}$ from Lemma B.4, it follows that

$$\begin{aligned}
\frac{\partial \mathsf{NegLog}(\boldsymbol{\mu}, \boldsymbol{\lambda})}{\partial \lambda_k} &= \frac{1}{Q} \sum_{q=1}^{Q} \sum_{j \in \mathsf{T}} \mathbb{1}_{\{c^q \in \mathsf{T}_j\}} \cdot \left(\frac{\psi_{j \to k}^q(\boldsymbol{\mu}, \boldsymbol{\lambda}) \Delta_k^q(\boldsymbol{\mu}, \boldsymbol{\lambda})}{\lambda_j} - \mathbb{1}_{\{k=j\}} \frac{W_k^q(\boldsymbol{\mu}, \boldsymbol{\lambda})}{\lambda_k^2} \right) \\
&\quad - \quad \frac{1}{Q} \sum_{q=1}^{Q} \sum_{j \in \mathsf{T}} \mathbb{1}_{\{c^q \in \mathsf{T}_j\}} \cdot \left(\frac{\psi_{j \to k}^q(\boldsymbol{\mu}, \boldsymbol{\lambda}) \Delta_k^q(\boldsymbol{\mu}, \boldsymbol{\lambda})}{\lambda_{\mathsf{pa}(j)}} - \mathbb{1}_{\{k=\mathsf{pa}(j)\}} \frac{W_j^q(\boldsymbol{\mu}, \boldsymbol{\lambda})}{\lambda_k^2} \right) \\
&= \frac{1}{Q} \sum_{q=1}^{Q} \Delta_k^q(\boldsymbol{\mu}, \boldsymbol{\lambda}) \cdot \left(\sum_{j \in \mathsf{T}} \mathbb{1}_{\{c^q \in \mathsf{T}_j\}} \left(\frac{1}{\lambda_j} - \frac{1}{\lambda_{\mathsf{pa}(j)}} \right) \cdot \psi_{j \to k}^q(\boldsymbol{\mu}, \boldsymbol{\lambda}) \right) \\
&\quad - \quad \frac{1}{Q} \sum_{q=1}^{Q} \left(\mathbb{1}_{\{c^q \in \mathsf{T}_k\}} W_k^q(\boldsymbol{\mu}, \boldsymbol{\lambda}) - \sum_{i \in \mathsf{Children}(k)} \mathbb{1}_{\{c^q \in \mathsf{T}_i\}} W_i^q(\boldsymbol{\mu}, \boldsymbol{\lambda}) \right) \frac{1}{\lambda_k^2} \\
&= \frac{1}{Q} \sum_{q=1}^{Q} \Delta_k^q(\boldsymbol{\mu}, \boldsymbol{\lambda}) \cdot D_k^q(\boldsymbol{\mu}, \boldsymbol{\lambda}) \; - \; \frac{1}{Q} \sum_{q=1}^{Q} \sum_{i \in \mathsf{Children}(k)} \mathbb{1}_{\{c^q \in \mathsf{T}_i\}} \cdot \left(W_k^q(\boldsymbol{\mu}, \boldsymbol{\lambda}) - W_i^q(\boldsymbol{\mu}, \boldsymbol{\lambda}) \right) \cdot \frac{1}{\lambda_k^2} \\
&= \frac{1}{Q} \sum_{q=1}^{Q} \Delta_k^q(\boldsymbol{\mu}, \boldsymbol{\lambda}) \cdot D_k^q(\boldsymbol{\mu}, \boldsymbol{\lambda}) \; - \; \frac{1}{Q} \sum_{q=1}^{Q} \sum_{i \in \mathsf{Children}(k)} \mathbb{1}_{\{c^q \in \mathsf{T}_i\}} \frac{-\log(\psi_{k \to i}^q(\boldsymbol{\mu}, \boldsymbol{\lambda})) \cdot \lambda_k}{\lambda_k^2} \\
&= \frac{1}{Q} \sum_{q=1}^{Q} \Delta_k^q(\boldsymbol{\mu}, \boldsymbol{\lambda}) \cdot D_k^q(\boldsymbol{\mu}, \boldsymbol{\lambda}) \; + \; \frac{1}{Q} \sum_{q=1}^{Q} E_k^q(\boldsymbol{\mu}, \boldsymbol{\lambda})
\end{aligned}$$

where the third equality follows from the expression for $D_k^q(\boldsymbol{\mu}, \boldsymbol{\lambda})$ in Lemma D.1 and the fact that $\mathbb{1}_{\{c^q \in \mathsf{T}_k\}} = \sum_{i \in \mathsf{Children}(k)} \mathbb{1}_{\{c^q \in \mathsf{T}_i\}}$ for all q, the fourth equality follows from the definition of $\psi_{k \to i}^q(\boldsymbol{\mu}, \boldsymbol{\lambda})$ and the last follows from the definition of $E_k^q(\boldsymbol{\mu}, \boldsymbol{\lambda})$ in the statement of the lemma.

Appendix E: Proofs of claims and results from Section 3

E.1 Non-convexity of NegLog in λ



Suppose $\mathcal{N}=\{1,2,3\}$ and the tree structure is shown in the figure to the left, where root has two children: a leaf node 1 and a non-leaf node 4. Node 4 has two children: leaf nodes 2 and 3 . The parameters of the model consist of μ_1, μ_2, μ_3, and λ_4. Suppose we offer the full assortment, and for each $\ell \in \mathcal{N}$, let $s_\ell \in \mathbb{Z}_{++}$ denote the number of customers who select product ℓ. Then,

$$\begin{aligned}
\psi_{\mathsf{root}\to 1}(\boldsymbol{\mu},\lambda_4) &= \psi_{\mathsf{root}\to 1}(\boldsymbol{\mu},\lambda_4) &&= \frac{e^{\mu_1}}{e^{\mu_1}+\left[e^{\mu_2/\lambda_4}+e^{\mu_3/\lambda_4}\right]^{\lambda_4}} \\
\psi_{\mathsf{root}\to 2}(\boldsymbol{\mu},\lambda_4) &= \psi_{\mathsf{root}\to 4}(\boldsymbol{\mu},\lambda_4)\times\psi_{4\to 2}(\boldsymbol{\mu},\lambda_4) &&= \frac{\left[e^{\mu_2/\lambda_4}+e^{\mu_3/\lambda_4}\right]^{\lambda_4}}{e^{\mu_1}+\left[e^{\mu_2/\lambda_4}+e^{\mu_3/\lambda_4}\right]^{\lambda_4}}\times\frac{e^{\mu_2/\lambda_4}}{e^{\mu_2/\lambda_4}+e^{\mu_3/\lambda_4}} \\
\psi_{\mathsf{root}\to 3}(\boldsymbol{\mu},\lambda_4) &= \psi_{\mathsf{root}\to 4}(\boldsymbol{\mu},\lambda_4)\times\psi_{4\to 3}(\boldsymbol{\mu},\lambda_4) &&= \frac{\left[e^{\mu_2/\lambda_4}+e^{\mu_3/\lambda_4}\right]^{\lambda_4}}{e^{\mu_1}+\left[e^{\mu_2/\lambda_4}+e^{\mu_3/\lambda_4}\right]^{\lambda_4}}\times\frac{e^{\mu_3/\lambda_4}}{e^{\mu_2/\lambda_4}+e^{\mu_3/\lambda_4}}\ ,
\end{aligned}$$

which implies that

$$\begin{aligned}
\mathsf{NegLog}(\boldsymbol{\mu},\lambda_4) \;=\; \Big((s_1+s_2+s_3)\log\Big(e^{\mu_1}+\left[e^{\mu_2/\lambda_4}+e^{\mu_3/\lambda_4}\right]^{\lambda_4}\Big) \;+\; (s_2+s_3)(1-\lambda_4)\log\left(e^{\mu_2/\lambda_4}+e^{\mu_3/\lambda_4}\right) \\
-\; s_1\mu_1 \;-\; s_2\frac{\mu_2}{\lambda_4} \;-\; s_3\frac{\mu_3}{\lambda_4}\Big) \Big/ (s_1+s_2+s_3)\ .
\end{aligned}$$

Suppose we have five customers, with $(s_1,s_2,s_3)=(1,1,3)$. For $\boldsymbol{\mu}=(\mu_1,\mu_2,\mu_3)=(0,1,1.03)$, we have $\mathsf{NegLog}(\boldsymbol{\mu},\lambda_4=0.1)=1.0116$ and $\mathsf{NegLog}(\boldsymbol{\mu},\lambda_4=0.3)=1.0381$, and we have that

$$\mathsf{NegLog}(\boldsymbol{\mu},\lambda_4=0.2)=1.0317 \;>\; \frac{\mathsf{NegLog}(\boldsymbol{\mu},\lambda_4=0.1)\;+\;\mathsf{NegLog}(\boldsymbol{\mu},\lambda_4=0.3)}{2} \;=\; 1.0249\ ,$$

which shows that $\lambda_4 \mapsto \mathsf{NegLog}(\boldsymbol{\mu},\lambda_4)$ is not convex in λ_4.

E.2 Proof of Theorem 3.1

The first part of the theorem follows from the Proof of Sufficiency argument in the proof of Theorem 2.4 in Appendix C.2 above.

For the second part, note that from Theorem 2.2, it follows that

$$\begin{aligned}
\mathsf{NegLog}(\boldsymbol{\mu},\boldsymbol{\lambda}) &= \frac{1}{Q}\sum_{q=1}^{Q}\sum_{j\in\mathsf{T}}\mathbb{1}_{\{c^q\in\mathsf{T}_j\}}\frac{W_j^q(\boldsymbol{\mu},\boldsymbol{\lambda})}{\lambda_j} \;-\; \frac{1}{Q}\sum_{q=1}^{Q}\sum_{j\in\mathsf{T}}\mathbb{1}_{\{c^q\in\mathsf{T}_j\}}\frac{W_j^q(\boldsymbol{\mu},\boldsymbol{\lambda})}{\lambda_{\mathsf{pa}(j)}} \\
&= \frac{1}{Q}\sum_{q=1}^{Q}\sum_{j\in\mathsf{T}\setminus\mathcal{N}}\mathbb{1}_{\{c^q\in\mathsf{T}_j\}}\frac{W_j^q(\boldsymbol{\mu},\boldsymbol{\lambda})}{\lambda_j} \;-\; \frac{1}{Q}\sum_{q=1}^{Q}\sum_{j\in\mathsf{T}\setminus\{\mathsf{root}\}}\mathbb{1}_{\{c^q\in\mathsf{T}_j\}}\frac{W_j^q(\boldsymbol{\mu},\boldsymbol{\lambda})}{\lambda_{\mathsf{pa}(j)}},
\end{aligned}$$

where the second equality follows since by definition, $\lambda_\ell=+\infty$ for all $\ell\in\mathcal{N}$ and $\lambda_{\mathsf{pa(root)}}=+\infty$. Then, substituting $\lambda_j=e^{-\delta_{[j]}}$ for all $j\in\mathsf{T}\setminus\mathcal{N}$, it follows that $\mathsf{NegLog}(\boldsymbol{\mu},\boldsymbol{\delta})=F_1(\boldsymbol{\mu},\boldsymbol{\delta})-F_2(\boldsymbol{\mu},\boldsymbol{\delta})$.

Next, we need to show that F_1 and F_2 are strictly convex in $\boldsymbol{\delta}$ for any fixed $\boldsymbol{\mu}$. We first show that the function $F_1(\boldsymbol{\mu},\boldsymbol{\delta})$ is convex in $\boldsymbol{\delta}$ and then establish later that it is strictly convex. By the homogeneity property of the weight function in Lemma B.1 above, we have

$$F_1(\boldsymbol{\mu},\boldsymbol{\delta}) \;=\; \frac{1}{Q}\sum_{q=1}^{Q}\sum_{j\in\mathsf{T}\setminus\mathcal{N}} \mathbb{1}_{\{c^q\in\mathsf{T}_j\}} W_j^q\left(\boldsymbol{\mu},\boldsymbol{\lambda}(\boldsymbol{\delta})\right)\cdot e^{\delta_{[j]}} \;=\; \frac{1}{Q}\sum_{q=1}^{Q}\sum_{j\in\mathsf{T}\setminus\mathcal{N}} \mathbb{1}_{\{c^q\in\mathsf{T}_j\}} W_j^q\left(\boldsymbol{\mu}\cdot e^{\delta_{[j]}},\boldsymbol{\lambda}(\boldsymbol{\delta})\cdot e^{\delta_{[j]}}\right)\;,$$

where recall that the vector $\boldsymbol{\lambda}(\boldsymbol{\delta}) = (\exp\left(-\delta_{[j]}\right) : j\in\mathsf{T}\setminus\mathcal{N})$. Then, for each term in the above summation, the mapping $\boldsymbol{\delta}\mapsto W_j^q\left(\boldsymbol{\mu}\cdot e^{\delta_{[j]}},\boldsymbol{\lambda}(\boldsymbol{\delta})\cdot e^{\delta_{[j]}}\right)$ is convex because it is a composition of an increasing convex function $W_j^q(\cdot,\cdot)$ (by Lemmas B.2 and B.3) with a collection of convex functions, $\mu_\ell\cdot e^{\delta_{[j]}} = \mu_\ell\cdot e^{\sum_{k\in\mathsf{path}[\mathsf{root},j]}\delta_k}$, and $e^{-\delta_{[k]}}\cdot e^{\delta_{[j]}} \;=\; e^{\sum_{h\in\mathsf{path}[\mathsf{root},j]}\delta_h-\sum_{h\in\mathsf{path}[\mathsf{root},k]}\delta_h}$. Note that $\mu_\ell\cdot e^{\delta_{[j]}}$ and $e^{-\delta_{[k]}}\cdot e^{\delta_{[j]}}$ are convex in $\boldsymbol{\delta}$ because the expressions in the exponents are linear functions of $\boldsymbol{\delta}$. Therefore, $F_1(\boldsymbol{\mu},\boldsymbol{\delta})$ is convex in $\boldsymbol{\delta}$.

To establish strict convexity, we show that one of the terms in the summand is strictly convex. Consider the term corresponding to the root node:

$$f(\boldsymbol{\mu},\boldsymbol{\delta}) \;=\; \sum_{q=1}^{Q} W_{\mathsf{root}}^q\left(\boldsymbol{\mu}\cdot e^{\delta_{[\mathsf{root}]}},\boldsymbol{\lambda}(\boldsymbol{\delta})\cdot e^{\delta_{[\mathsf{root}]}}\right) \;=\; \sum_{q=1}^{Q} W_{\mathsf{root}}^q(\boldsymbol{\mu},\boldsymbol{\lambda}(\boldsymbol{\delta})) \quad \text{since} \quad \delta_{[\mathsf{root}]}=\delta_{\mathsf{root}}=0.$$

We now show that $f(\boldsymbol{\mu},\boldsymbol{\delta})$ is strictly convex in $\boldsymbol{\delta}$. For any $\boldsymbol{\delta}\neq\bar{\boldsymbol{\delta}}$ and scalar $x\in(0,1)$, we want to show that

$$f(\boldsymbol{\mu},x\cdot\boldsymbol{\delta}+(1-x)\cdot\bar{\boldsymbol{\delta}}) \;<\; xf(\boldsymbol{\mu},\boldsymbol{\delta})+(1-x)f(\boldsymbol{\mu},\bar{\boldsymbol{\delta}}).$$

Since $W_{\mathsf{root}}^{q'}(\boldsymbol{\mu},\boldsymbol{\lambda}(\boldsymbol{\delta}))$ is convex in $\boldsymbol{\delta}$ for all $q'\in\{1,\ldots,Q\}$, it suffices to exhibit one q such that $W_{\mathsf{root}}^q\left(\boldsymbol{\mu},\boldsymbol{\lambda}(x\boldsymbol{\delta}+(1-x)\bar{\boldsymbol{\delta}})\right) \;<\; xW_{\mathsf{root}}^q\left(\boldsymbol{\mu},\boldsymbol{\lambda}(\boldsymbol{\delta})\right)+(1-x)W_{\mathsf{root}}^q\left(\boldsymbol{\mu},\boldsymbol{\lambda}(\bar{\boldsymbol{\delta}})\right)$.

Now, since $\boldsymbol{\delta}\neq\bar{\boldsymbol{\delta}}$, there must exist a non-leaf node j such that $\delta_{[j]}\neq\bar{\delta}_{[j]}$. From the condition in Theorem 2.5, there exists a $q\in\{1,\ldots,Q\}$ such that j has at least two children in T^q. It then follows from Lemmas B.2 and B.3 that $W_{\mathsf{root}}^q(\boldsymbol{\mu},\boldsymbol{\lambda})$ is convex and increasing in $\boldsymbol{\lambda}$, and strictly increasing in λ_j. Further, because of the strict convexity of the exponential function, we have that $e^{x\cdot(-\delta_{[j]})+(1-x)\cdot(-\bar{\delta}_{[j]})} < xe^{-\delta_{[j]}}+(1-x)e^{-\bar{\delta}_{[j]}}$. Together, these facts imply that

$$\begin{aligned} W_{\mathsf{root}}^q\left(\boldsymbol{\mu},\boldsymbol{\lambda}(x\boldsymbol{\delta}+(1-x)\bar{\boldsymbol{\delta}})\right) \;&<\; W_{\mathsf{root}}^q\left(\boldsymbol{\mu}\;,\; x\boldsymbol{\lambda}(\boldsymbol{\delta})+(1-x)\boldsymbol{\lambda}(\bar{\boldsymbol{\delta}})\right) \\ &\leq\; xW_{\mathsf{root}}^q\left(\boldsymbol{\mu},\boldsymbol{\lambda}(\boldsymbol{\delta})\right)+(1-x)W_{\mathsf{root}}^q\left(\boldsymbol{\mu},\boldsymbol{\lambda}(\bar{\boldsymbol{\delta}})\right)\;. \end{aligned}$$

Consequently, F_1 is strictly convex. The proof of the strict convexity of F_2 follows from an almost identical argument, but with the function f re-defined as $f(\boldsymbol{\mu},\boldsymbol{\delta}) = \sum_{q=1}^{Q}\sum_{j\in\mathsf{Children}(\mathsf{root})\cap\mathsf{T}^q} W_j^q(\boldsymbol{\mu},\boldsymbol{\lambda}(\boldsymbol{\delta}))$. ■

E.3 Step 2 of A-MM algorithm can be solved in closed form

We first show that the partial derivatives $(\partial \mathsf{NegLog}(\boldsymbol{\mu}, \boldsymbol{\delta})/\partial \delta_k \ : \ k \in \mathsf{T} \setminus (\mathcal{N} \cup \{\mathsf{root}\}))$ can also be computed efficiently:

Lemma E.1 *For all $k \in \mathsf{T} \setminus (\mathcal{N} \cup \{\mathsf{root}\})$, it follows that*

$$\frac{\partial \mathsf{NegLog}(\boldsymbol{\mu}, \boldsymbol{\delta})}{\partial \delta_k} = - \sum_{j \in \mathsf{T}_k \setminus \mathcal{N}} \lambda_j \cdot \frac{\partial \mathsf{NegLog}(\boldsymbol{\mu}, \boldsymbol{\lambda})}{\partial \lambda_j},$$

Proof. The result follows by invoking the multi-variable chain rule of derivatives:

$$\begin{aligned}
\frac{\partial \mathsf{NegLog}(\boldsymbol{\mu}, \boldsymbol{\delta})}{\partial \delta_k} &= \sum_{j \in \mathsf{T} \setminus \mathcal{N}} \frac{\partial \lambda_j}{\partial \delta_k} \cdot \frac{\partial \mathsf{NegLog}(\boldsymbol{\mu}, \boldsymbol{\lambda})}{\partial \lambda_j} \\
&= \sum_{j \in \mathsf{T}_k \setminus \mathcal{N}} \frac{\partial \lambda_j}{\partial \delta_k} \cdot \frac{\partial \mathsf{NegLog}(\boldsymbol{\mu}, \boldsymbol{\lambda})}{\partial \lambda_j} \\
&= \sum_{j \in \mathsf{T}_k \setminus \mathcal{N}} -\lambda_j \cdot \frac{\partial \mathsf{NegLog}(\boldsymbol{\mu}, \boldsymbol{\lambda})}{\partial \lambda_j},
\end{aligned}$$

where the second inequality follows since $\lambda_j = e^{-\sum_{i \in \mathsf{path}[\mathsf{root}, j]} \delta_i}$ and the last since $\frac{\partial \lambda_j}{\partial \delta_k} = -\lambda_j$ for all $j \in \mathsf{T}_k \setminus \mathcal{N}$. The claim then follows. ■

Next, define the function $J(\alpha) := H^{(s)}\left(\left\{\boldsymbol{\delta}^{(s)} - \alpha \cdot \nabla_{\boldsymbol{\delta}} \mathsf{NegLog}(\boldsymbol{\mu}^{(s+1)}, \ \boldsymbol{\delta}^{(s)})\right\}^+\right)$ for all $\alpha \in \mathbb{R}_+$. The following lemma establishes the piecewise convexity of $J(\cdot)$:

Lemma E.2 *For each $j \in \mathsf{T} \setminus (\mathcal{N} \cup \{\mathsf{root}\})$, let $d_j = \frac{\partial \mathsf{NegLog}}{\partial \delta_j}\left(\boldsymbol{\mu}^{(s+1)}, \boldsymbol{\delta}^{(s)}\right)$ and define t_j as*

$$t_j = \begin{cases} +\infty & \text{if } d_j \leq 0 \\ \delta_j^{(s)}/d_j & \text{otherwise} \end{cases}.$$

Next, let $0 = t_{(0)} < t_{(1)} < \ldots < t_{(I)} = +\infty$ denote the sorted values in the set $\{t_j \ : \ j \in \mathsf{T} \setminus (\mathcal{N} \cup \{\mathsf{root}\})\} \cup \{0, +\infty\}$.

Then, for each $i \in \{0, 1, \ldots, I-1\}$, the function $J(\alpha)$ is either **constant** *or* **strictly convex** *on the interval $\left[t_{(i)}, t_{(i+1)}\right]$. In particular, $J(\cdot)$ is piecewise convex on $\mathbb{R}_+$ with I pieces, where $I \leq |\mathsf{T} \setminus \mathcal{N}|$.*

Proof. Define the vector $\boldsymbol{M}(\alpha) = \left\{\boldsymbol{\delta}^{(s)} - \alpha \boldsymbol{d}\right\}^+$ for any $\alpha \geq 0$. Then, for any $\alpha \in \left[t_{(i)}, t_{(i+1)}\right]$ and any $j \in \mathsf{T} \setminus (\mathcal{N} \cup \{\mathsf{root}\})$, it can be verified that

$$M_j(\alpha) = \begin{cases} \delta_j^{(s)} - \alpha \cdot d_j & \text{if } t_{(i+1)} \leq t_j \\ 0 & \text{otherwise} \end{cases}. \tag{EC.6}$$

Now, suppose the function $J(\alpha)$ is *not* constant on the interval $\left[t_{(i)}, t_{(i+1)}\right]$. In other words, there exists $\bar{\alpha}, \hat{\alpha} \in \left[t_{(i)}, t_{(i+1)}\right]$ with $\bar{\alpha} \neq \hat{\alpha}$ such that $J(\bar{\alpha}) \neq J(\hat{\alpha})$. Since $J(\alpha) = H^{(s)}\left(\boldsymbol{M}(\alpha)\right)$, we must

have $\boldsymbol{M}(\bar{\alpha}) \neq \boldsymbol{M}(\hat{\alpha})$. From (EC.6), this further implies that there exists $j_i \in \mathsf{T} \setminus (\mathcal{N} \cup \{\mathsf{root}\})$ such that $d_{j_i} \neq 0$ and $M_{j_i}(\alpha) = \delta_{j_i}^{(s)} - \alpha \cdot d_{j_i}$ for all $\alpha \in \left[t_{(i)}, t_{(i+1)}\right]$.

We show that $J(\alpha)$ must be strictly convex on $\left[t_{(i)}, t_{(i+1)}\right]$. For that, consider $\alpha_1, \alpha_2 \in \left[t_{(i)}, t_{(i+1)}\right]$ with $\alpha_1 \neq \alpha_2$ and $w \in (0,1)$. From (EC.6), it can be verified that

$$\boldsymbol{M}(w \cdot \alpha_1 + (1-w) \cdot \alpha_2) = w \cdot \boldsymbol{M}(\alpha_1) + (1-w) \cdot \boldsymbol{M}(\alpha_2) \tag{EC.7}$$

Moreover, since $M_{j_i}(\alpha_1) \neq M_{j_i}(\alpha_2)$ it follows that $\boldsymbol{M}(\alpha_1) \neq \boldsymbol{M}(\alpha_2)$. Then, it follows that

$$\begin{aligned}
J\left(w \cdot \alpha_1 + (1-w) \cdot \alpha_2\right) &= H^{(s)}\left(\boldsymbol{M}(w \cdot \alpha_1 + (1-w) \cdot \alpha_2)\right) \\
&= H^{(s)}\left(w \cdot \boldsymbol{M}(\alpha_1) + (1-w) \cdot \boldsymbol{M}(\alpha_2)\right) \\
&< w \cdot H^{(s)}\left(\boldsymbol{M}(\alpha_1)\right) + (1-w) \cdot H^{(s)}\left(\boldsymbol{M}(\alpha_2)\right) \\
&= w \cdot J(\alpha_1) + (1-w) \cdot J(\alpha_2),
\end{aligned}$$

where the second equality follows from (EC.7) and the inequality follows since $\boldsymbol{M}(\alpha_1) \neq \boldsymbol{M}(\alpha_2)$ and the fact that $H^{(s)}(\cdot)$ is strictly convex as established in Lemma 3.8. This establishes the strict convex of $J(\cdot)$ on $\left[t_{(i)}, t_{(i+1)}\right]$. The result then follows from observing that $\cup_{i=0}^{I-1} \left[t_{(i)}, t_{(i+1)}\right] = \mathbb{R}_+$.

Finally, note that the number of pieces I satisfies

$$\begin{aligned}
I &= \left|\{t_j \ : \ j \in \mathsf{T} \setminus (\mathcal{N} \cup \{\mathsf{root}\})\} \cup \{0, +\infty\}\right| - 1 \\
&\leq \left|\{t_j \ : \ j \in \mathsf{T} \setminus (\mathcal{N} \cup \{\mathsf{root}\})\}\right| + 2 - 1 \\
&= \left|\{t_j \ : \ j \in \mathsf{T} \setminus (\mathcal{N} \cup \{\mathsf{root}\})\}\right| + 1 \\
&\leq \left|\mathsf{T} \setminus \mathcal{N}\right|
\end{aligned}$$

■

The above lemma can be leveraged to efficiently compute the step size $\alpha^{(s)}$ in Step 2 of the A-MM algorithm as follows: we first obtain $\alpha_0, \alpha_1, \ldots, \alpha_{I-1}$ as the following

$$\alpha_i = \underset{\alpha \in \left[t_{(i)}, t_{(i+1)}\right]}{\arg\min} \quad J(\alpha),$$

which can be done efficiently since $J(\alpha)$ is either constant—in which case any α is optimal—or strictly convex—in which case well-known algorithms such as the golden-section search (Kiefer 1953) can be used—on each interval $\left[t_{(i)}, t_{(i+1)}\right]$. Then, we compute $\alpha^{(s)}$ as the following

$$\alpha^{(s)} = \underset{\alpha \in \left\{\alpha_0, \alpha_1, \ldots, \alpha_{I-1}\right\}}{\arg\min} \quad J(\alpha).$$

E.4 Proof of Theorem 3.3

We begin with the following key lemma that will be useful in the proof.

Lemma E.3 *For any $\boldsymbol{\mu} \in \mathsf{Dom}_1$ and $\boldsymbol{\delta} \in \mathsf{Dom}_2$, it follows that*

$$\sum_{\ell \in \mathcal{N}} a_\ell\left(\boldsymbol{\mu},\, \boldsymbol{\delta}\right) \times \exp\left(\delta_{[\mathsf{pa}(\ell)]}\right) = \sum_{\ell \in \mathcal{N}} \mathsf{sales}_\ell \times \exp\left(\delta_{[\mathsf{pa}(\ell)]}\right),$$

where $a_\ell(\boldsymbol{\mu}, \boldsymbol{\delta})$ is as defined in (5).

Proof. We leverage the fact that $\mathsf{NegLog}(\boldsymbol{\mu}, \boldsymbol{\delta})$ is shift-invariant in the variable $\boldsymbol{\mu}$. Specifically, for any $x \in \mathbb{R}$, we have that

$$\mathsf{NegLog}(\boldsymbol{\mu}, \boldsymbol{\delta}) = \mathsf{NegLog}(\boldsymbol{\mu} + x \cdot \boldsymbol{e}, \boldsymbol{\delta}),$$

where $\boldsymbol{e} \in \mathbb{R}^N$ denotes the vector of all ones. In other words, for any $\boldsymbol{\mu}$ and $\boldsymbol{\delta}$, the function $g\colon \mathbb{R} \to \mathbb{R}$ defined as $g(x) = \mathsf{NegLog}(\boldsymbol{\mu} + x \cdot \boldsymbol{e}, \boldsymbol{\delta})$ is a constant as a function of x. Therefore, the derivative $g'(x)$ of $g(x)$ with respect to x is zero everywhere.

We can compute the derivative of $g(\cdot)$ using the chain rule as follows:

$$g'(x) = \frac{\mathrm{d}g}{\mathrm{d}x} = \sum_{\ell \in \mathcal{N}} \frac{\partial \mathsf{NegLog}(\boldsymbol{\mu} + x \cdot \boldsymbol{e}, \boldsymbol{\delta})}{\partial \mu_\ell} \frac{\mathrm{d}}{\mathrm{d}x}\left(\mu_\ell + x\right) = \sum_{\ell \in \mathcal{N}} \frac{\partial \mathsf{NegLog}(\boldsymbol{\mu} + x \cdot \boldsymbol{e}, \boldsymbol{\delta})}{\partial \mu_\ell}.$$

Equating the above derivative to zero at $x = 0$ yields $\sum_{\ell \in \mathcal{N}} \partial \mathsf{NegLog}(\boldsymbol{\mu}, \boldsymbol{\delta}) / \partial \mu_\ell = 0$. We now use the definition of the constant $a_\ell(\boldsymbol{\mu}, \boldsymbol{\delta})$ for each $\ell \in \mathcal{N}$ to obtain

$$0 = \sum_{\ell \in \mathcal{N}} \frac{\partial \mathsf{NegLog}(\boldsymbol{\mu}, \boldsymbol{\delta})}{\partial \mu_\ell} = \sum_{\ell \in \mathcal{N}} \left(a_\ell(\boldsymbol{\mu}, \boldsymbol{\delta}) - \mathsf{sales}_\ell\right) \cdot \exp\left(\delta_{[\mathsf{pa}(\ell)]}\right)$$

The result of the lemma now follows. ∎

We are now ready to prove Theorem 3.3. Below, we denote $\boldsymbol{\lambda}^{(s)} = \boldsymbol{\lambda}(\boldsymbol{\delta}^{(s)})$ to simplify certain expressions. In Theorem 3.7, we show that $\sum_{\ell \in \mathcal{N}} G_\ell\left(\mu_\ell \,|\, \boldsymbol{\mu}^{(s)},\ \boldsymbol{\delta}^{(s)}\right)$ is a majorizing surrogate [see Definition 3.6] for the mapping $\boldsymbol{\mu} \mapsto \mathsf{NegLog}(\boldsymbol{\mu}, \boldsymbol{\delta}^{(s)})$ at $\boldsymbol{\mu}^{(s)}$. From this, it follows that for all $\boldsymbol{\mu} \in \mathsf{Dom}_1$:

$$\begin{aligned}
\mathsf{NegLog}(\boldsymbol{\mu}, \boldsymbol{\delta}^{(s)}) &\le \sum_{\ell \in \mathcal{N}} G_\ell\left(\mu_\ell \,|\, \boldsymbol{\mu}^{(s)},\ \boldsymbol{\delta}^{(s)}\right) \\
&= \sum_{\ell \in \mathcal{N}} \mathsf{C}_\ell(\boldsymbol{\mu}^{(s)},\ \boldsymbol{\delta}^{(s)}) + \sum_{\ell \in \mathcal{N}} a_\ell\left(\boldsymbol{\mu}^{(s)},\, \boldsymbol{\delta}^{(s)}\right) \exp\left(\frac{\mu_\ell - \mu_\ell^{(s)}}{\lambda^{(s)}_{\mathsf{pa}(\ell)}}\right) - \sum_{\ell \in \mathcal{N}} \mathsf{sales}_\ell \cdot \left(\frac{\mu_\ell - \mu_\ell^{(s)}}{\lambda^{(s)}_{\mathsf{pa}(\ell)}}\right)
\end{aligned} \tag{EC.8}$$

and,

$$\mathsf{NegLog}(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) = \sum_{\ell \in \mathcal{N}} \mathsf{C}_\ell(\boldsymbol{\mu}^{(s)},\ \boldsymbol{\delta}^{(s)}) + \sum_{\ell \in \mathcal{N}} a_\ell\left(\boldsymbol{\mu}^{(s)},\, \boldsymbol{\delta}^{(s)}\right) \tag{EC.9}$$

Combining equations (EC.8) and (EC.9), we obtain that for all $\boldsymbol{\mu} \in \mathsf{Dom}_1$:

$$\begin{aligned}
&\mathsf{NegLog}(\boldsymbol{\mu}, \boldsymbol{\delta}^{(s)}) - \mathsf{NegLog}(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) \\
&\leq \sum_{\ell \in \mathcal{N}} a_\ell \left(\boldsymbol{\mu}^{(s)},\, \boldsymbol{\delta}^{(s)}\right) \left(\exp\left(\frac{\mu_\ell - \mu_\ell^{(s)}}{\lambda_{\mathsf{pa}(\ell)}^{(s)}}\right) - 1\right) - \sum_{\ell \in \mathcal{N}} \mathsf{sales}_\ell \cdot \left(\frac{\mu_\ell - \mu_\ell^{(s)}}{\lambda_{\mathsf{pa}(\ell)}^{(s)}}\right)
\end{aligned}$$

Now, noting that $\tilde{\mu}_\ell^{(s+1)} = \mu_\ell^{(s)} + \lambda_{\mathsf{pa}(\ell)}^{(s)} \cdot \log\left(\mathsf{sales}_\ell / a_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)})\right)$ for all $\ell \in \mathcal{N}$, we get (we show in the proof of Theorem 3.7 below that $a_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) > 0$ for all s)

$$\begin{aligned}
&\mathsf{NegLog}(\tilde{\boldsymbol{\mu}}^{(s+1)}, \boldsymbol{\delta}^{(s)}) - \mathsf{NegLog}(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) \\
&\leq \sum_{\ell \in \mathcal{N}} a_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) \left(\frac{\mathsf{sales}_\ell}{a_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)})} - 1\right) - \sum_{\ell \in \mathcal{N}} \mathsf{sales}_\ell \cdot \log\left(\frac{\mathsf{sales}_\ell}{a_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)})}\right) \\
&= \sum_{\ell \in \mathcal{N}} \left(\mathsf{sales}_\ell - a_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) - \mathsf{sales}_\ell \cdot \log\left(\frac{\mathsf{sales}_\ell}{a_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)})}\right)\right). \qquad \text{(EC.10)}
\end{aligned}$$

We first establish that in each iteration, the MM update makes a non-positive improvement (recall that we are minimizing the objective function) by showing that the upper bound in equation (EC.10) above is always non-positive. In fact, we establish a stronger result that each term in the upper bound summation is non-positive; that is, $\mathsf{sales}_\ell - a_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) - \mathsf{sales}_\ell \cdot \log\left(\mathsf{sales}_\ell / a_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)})\right) \leq 0$ for all $\ell \in \mathcal{N}$.

To that end, for each $\ell \in \mathcal{N}$, note that $\tilde{\mu}_\ell^{(s+1)}$ is the minimizer of $G_\ell(\cdot \,|\, \boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)})$ by definition. Therefore, we have that

$$\begin{aligned}
&G_\ell(\tilde{\mu}_\ell^{(s+1)} \,|\, \boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) \leq G_\ell(\mu_\ell^{(s)} \,|\, \boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) \\
\implies &a_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) \exp\left(\frac{\tilde{\mu}_\ell^{(s+1)} - \mu_\ell^{(s)}}{\lambda_{\mathsf{pa}(\ell)}^{(s)}}\right) - \mathsf{sales}_\ell \cdot \frac{\tilde{\mu}_\ell^{(s+1)} - \mu_\ell^{(s)}}{\lambda_{\mathsf{pa}(\ell)}^{(s)}} \leq a_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) \\
\implies &a_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) \frac{\mathsf{sales}_\ell}{a_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)})} - \mathsf{sales}_\ell \cdot \log\left(\frac{\mathsf{sales}_\ell}{a_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)})}\right) \leq a_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) \\
\implies &\mathsf{sales}_\ell - a_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) - \mathsf{sales}_\ell \cdot \log\left(\frac{\mathsf{sales}_\ell}{a_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)})}\right) \leq 0.
\end{aligned}$$

It now follows from the arguments above that

$$\begin{aligned}
&\mathsf{NegLog}(\tilde{\boldsymbol{\mu}}^{(s+1)}, \boldsymbol{\delta}^{(s)}) - \mathsf{NegLog}(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) \\
&\leq \sum_{\ell \in \mathcal{N}} \left(\mathsf{sales}_\ell - a_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) - \mathsf{sales}_\ell \cdot \log\left(\frac{\mathsf{sales}_\ell}{a_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)})}\right)\right) \\
&\leq \sum_{\ell \in \mathcal{N}} \frac{\lambda_{\mathsf{lower}}}{\lambda_{\mathsf{pa}(\ell)}^{(s)}} \left[\mathsf{sales}_\ell - a_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) - \mathsf{sales}_\ell \cdot \log\left(\frac{\mathsf{sales}_\ell}{a_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)})}\right)\right] \\
&= \lambda_{\mathsf{lower}} \cdot \left(\sum_{\ell \in \mathcal{N}} \frac{\mathsf{sales}_\ell - a_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)})}{\lambda_{\mathsf{pa}(\ell)}^{(s)}}\right) - \lambda_{\mathsf{lower}} \cdot \sum_{\ell \in \mathcal{N}} \frac{\mathsf{sales}_\ell}{\lambda_{\mathsf{pa}(\ell)}^{(s)}} \cdot \log\left(\frac{\mathsf{sales}_\ell}{a_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)})}\right)
\end{aligned}$$

$$= \lambda_{\mathsf{lower}} \cdot \left(\sum_{\ell \in \mathcal{N}} \left\{ \mathsf{sales}_\ell - a_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) \right\} \times \exp\left(\delta^{(s)}_{[\mathsf{pa}(\ell)]}\right) \right) - \lambda_{\mathsf{lower}} \sum_{\ell \in \mathcal{N}} \mathsf{sales}_\ell \cdot \exp\left(\delta^{(s)}_{[\mathsf{pa}(\ell)]}\right) \cdot \log\left(\frac{\mathsf{sales}_\ell}{a_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)})} \right)$$

$$= -\lambda_{\mathsf{lower}} \sum_{\ell \in \mathcal{N}} \mathsf{sales}_\ell \cdot \exp\left(\delta^{(s)}_{[\mathsf{pa}(\ell)]}\right) \cdot \log\left(\frac{\mathsf{sales}_\ell}{a_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)})} \right),$$

where the first inequality follows because $\lambda^{(s)}_{\mathsf{pa}(\ell)} = e^{-\sum_{k \in \mathsf{path}[\mathsf{root}, \mathsf{pa}(\ell)]} \delta^{(s)}_k} \geq e^{-(\mathsf{height}(\mathsf{root})-1) \times \delta_{\mathsf{upper}}} := \lambda_{\mathsf{lower}}$ for all $\ell \in \mathcal{N}$ and the last equality follows from the result of Lemma E.3 which implies $\sum_{\ell \in \mathcal{N}} \left\{ \mathsf{sales}_\ell - a_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) \right\} \times \exp\left(\delta^{(s)}_{[\mathsf{pa}(\ell)]}\right) = 0$.

The last expression on the right hand side above has the form of KL-divergence between two probability distributions, except that the corresponding terms in the expressions do not sum to 1. To address that, we normalize $(a_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) \colon \ell \in \mathcal{N})$ and $(\mathsf{sales}_\ell \colon \ell \in \mathcal{N})$ as follows. We let $T(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)})$ denote the sums $\sum_{\ell \in \mathcal{N}} a_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) \times \exp\left(\delta^{(s)}_{[\mathsf{pa}(\ell)]}\right) = \sum_{\ell \in \mathcal{N}} \mathsf{sales}_\ell \times \exp\left(\delta^{(s)}_{[\mathsf{pa}(\ell)]}\right)$ and define

$$\overline{a}_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) = \frac{a_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) \cdot \exp\left(\delta^{(s)}_{[\mathsf{pa}(\ell)]}\right)}{T(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)})}; \quad \overline{\mathsf{sales}}_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) = \frac{\mathsf{sales}_\ell \cdot \exp\left(\delta^{(s)}_{[\mathsf{pa}(\ell)]}\right)}{T(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)})}.$$

It is clear from our definitions that $\overline{\boldsymbol{a}}(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) = \left(\overline{a}_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) \colon \ell \in \mathcal{N} \right)$ and $\overline{\mathbf{sales}} = \left(\overline{\mathsf{sales}}_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) \colon \ell \in \mathcal{N} \right)$ are valid distributions over the set $\mathcal{N}$. It now follows from these definitions that

$$\begin{aligned}
&\mathsf{NegLog}(\tilde{\boldsymbol{\mu}}^{(s+1)}, \boldsymbol{\delta}^{(s)}) - \mathsf{NegLog}(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) \\
&\leq -T(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) \cdot \lambda_{\mathsf{lower}} \sum_{\ell \in \mathcal{N}} \overline{\mathsf{sales}}_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) \log\left(\frac{\overline{\mathsf{sales}}_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)})}{\overline{a}_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)})} \right) \\
&= -T(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) \cdot \lambda_{\mathsf{lower}} \cdot D_{KL}\left(\overline{\mathbf{sales}}(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) \| \overline{\boldsymbol{a}}(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) \right).
\end{aligned}$$

We can simplify the above expression further by invoking Pinsker's inequality (Csiszár and Körner 2011), which states that $D_{KL}\left(\overline{\mathbf{sales}}(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) \| \overline{\boldsymbol{a}}(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) \right) \geq 1/2 \left\| \overline{\mathbf{sales}}(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) - \overline{\boldsymbol{a}}(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) \right\|_1^2$. We now obtain

$$\begin{aligned}
&\mathsf{NegLog}(\tilde{\boldsymbol{\mu}}^{(s+1)}, \boldsymbol{\delta}^{(s)}) - \mathsf{NegLog}(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) \\
&\quad \leq -T(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) \cdot \lambda_{\mathsf{lower}} \cdot \frac{1}{2} \left\| \overline{\mathbf{sales}}(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) - \overline{\boldsymbol{a}}(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) \right\|_1^2 \\
&\quad = -\frac{\lambda_{\mathsf{lower}}}{2 \cdot T(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)})} \left(\sum_{\ell \in \mathcal{N}} \left| \left\{ \mathsf{sales}_\ell - a_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) \right\} \times \exp\left(\delta^{(s)}_{[\mathsf{pa}(\ell)]}\right) \right| \right)^2 \\
&\quad = -\frac{\lambda_{\mathsf{lower}}}{2 \cdot T(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)})} \left(\sum_{\ell \in \mathcal{N}} \left| \frac{\partial \mathsf{NegLog}(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)})}{\partial \mu_\ell} \right| \right)^2 = -\frac{\lambda_{\mathsf{lower}}}{2 \cdot T(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)})} \left\| \nabla_{\boldsymbol{\mu}} \mathsf{NegLog}(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) \right\|_1^2,
\end{aligned}$$

where the second equality follows from the definition of $a_\ell(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)})$.

To complete the proof, we note that

$$T(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) = \sum_{\ell \in \mathcal{N}} \mathsf{sales}_\ell \times \cdot \exp\left(\delta^{(s)}_{[\mathsf{pa}(\ell)]}\right) = \sum_{\ell \in \mathcal{N}} \frac{\mathsf{sales}_\ell}{\lambda^{(s)}_{\mathsf{pa}(\ell)}} \leq \sum_{\ell \in \mathcal{N}} \frac{\mathsf{sales}_\ell}{\lambda_{\mathsf{lower}}} = \frac{\sum_{\ell \in \mathcal{N}} \mathsf{sales}_\ell}{\lambda_{\mathsf{lower}}} = \frac{1}{\lambda_{\mathsf{lower}}}.$$

The inequality follows because $\lambda^{(s)}_{\mathsf{pa}(\ell)} \geq \lambda_{\mathsf{lower}}$ for all $\ell \in \mathcal{N}$. The last equality follows from the definition of sales_ℓ, which denotes the fraction of sales of product ℓ in the dataset, so that $\sum_{\ell \in \mathcal{N}} \mathsf{sales}_\ell = 1$.

Putting everything together, we obtain

$$\mathsf{NegLog}(\tilde{\boldsymbol{\mu}}^{(s+1)}, \boldsymbol{\delta}^{(s)}) - \mathsf{NegLog}(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) \leq \frac{-\lambda^2_{\mathsf{lower}}}{2} \left\| \nabla_{\boldsymbol{\mu}} \mathsf{NegLog}(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) \right\|_1^2.$$

The improvement bound for the MM update follows from observing that $\mathsf{NegLog}(\boldsymbol{\mu}^{(s+1)}, \boldsymbol{\delta}^{(s)}) = \mathsf{NegLog}(\tilde{\boldsymbol{\mu}}^{(s+1)}, \boldsymbol{\delta}^{(s)})$ since NegLog is shift-invariant w.r.t. $\boldsymbol{\mu}$. □

E.5 Proof of Theorem 3.4

We begin with the following lemma:

Lemma E.4 *If $\lambda_j \geq \lambda_{\mathsf{lower}}$ for all $j \in \mathsf{T} \setminus \mathcal{N}$, then the mapping $\boldsymbol{\mu} \mapsto \mathsf{NegLog}(\boldsymbol{\mu}, \boldsymbol{\lambda})$ is L-smooth with $L \leq 1/\lambda^2_{\mathsf{lower}}$.*

Proof. We use the definition of smoothness stated in (Bubeck 2015, Section 3.2). A continuously differentiable function $f\colon \mathbb{R}^n \to \mathbb{R}$ is L-smooth if its gradient ∇f is L-Lipschitz continuous, i.e.

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2 \leq L \|\boldsymbol{x} - \boldsymbol{y}\|_2 \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$$

For a twice continuously differentiable function, the above condition is equivalent to $\nabla^2 f(\boldsymbol{x}) \preceq L \boldsymbol{I}_n$ for all $\boldsymbol{x} \in \mathbb{R}^n$, where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix. In other words, all the eigenvalues of the Hessian are bounded above by L. We will use this definition to derive the smoothness constant L.

In our context, $f(\boldsymbol{\mu}) = \mathsf{NegLog}(\boldsymbol{\mu}, \boldsymbol{\lambda})$. We use the fact that the trace (sum of diagonal elements) of a matrix is equal to the sum of its eigenvalues. Consider the sum of the diagonal entries of the Hessian matrix $\nabla^2 f(\boldsymbol{\mu})$:

$$\begin{aligned}
&\sum_{\ell \in \mathcal{N}} \frac{\partial^2 f(\boldsymbol{\mu})}{\partial \mu_\ell^2} \\
&= \sum_{\ell \in \mathcal{N}} \frac{\partial}{\partial \mu_\ell} \left(\frac{\partial f(\boldsymbol{\mu})}{\partial \mu_\ell} \right) \\
&\overset{\text{(a)}}{=} \sum_{\ell \in \mathcal{N}} \frac{\partial}{\partial \mu_\ell} \left(\frac{1}{Q} \sum_{q=1}^{Q} \sum_{j \in \mathsf{T} \setminus \mathcal{N}} \mathbb{1}_{\{c^q \in \mathsf{T}_j\}} \cdot \left(\frac{1}{\lambda_j} - \frac{1}{\lambda_{\mathsf{pa}(j)}} \right) \cdot \psi^q_{j \to \ell}(\boldsymbol{\mu}, \boldsymbol{\lambda}) - \frac{\mathsf{sales}_\ell}{\lambda_{\mathsf{pa}(\ell)}} \right) \\
&= \sum_{\ell \in \mathcal{N}} \frac{1}{Q} \sum_{q=1}^{Q} \sum_{j \in \mathsf{T} \setminus \mathcal{N}} \mathbb{1}_{\{c^q \in \mathsf{T}_j\}} \cdot \left(\frac{1}{\lambda_j} - \frac{1}{\lambda_{\mathsf{pa}(j)}} \right) \cdot \frac{\partial}{\partial \mu_\ell} \psi^q_{j \to \ell}(\boldsymbol{\mu}, \boldsymbol{\lambda})
\end{aligned}$$

$$
\begin{aligned}
&\overset{\text{(b)}}{=} \sum_{\ell\in\mathcal{N}} \frac{1}{Q}\sum_{q=1}^{Q}\sum_{j\in\mathsf{T}\setminus\mathcal{N}} \mathbb{1}_{\{c^q\in\mathsf{T}_j\}}\cdot\left(\frac{1}{\lambda_j}-\frac{1}{\lambda_{\mathsf{pa}(j)}}\right)\cdot\frac{\partial}{\partial\mu_\ell}\prod_{k\in\mathsf{path}(j,\ell]}\psi^q_{\mathsf{pa}(k)\to k}(\boldsymbol{\mu},\boldsymbol{\lambda})\\
&\overset{\text{(c)}}{=} \sum_{\ell\in\mathcal{N}} \frac{1}{Q}\sum_{q=1}^{Q}\sum_{j\in\mathsf{T}\setminus\mathcal{N}} \mathbb{1}_{\{c^q\in\mathsf{T}_j\}}\cdot\left(\frac{1}{\lambda_j}-\frac{1}{\lambda_{\mathsf{pa}(j)}}\right)\cdot\psi^q_{j\to\ell}(\boldsymbol{\mu},\boldsymbol{\lambda})\cdot\left(\sum_{k\in\mathsf{path}(j,\ell]}\frac{\partial\psi^q_{\mathsf{pa}(k)\to k}(\boldsymbol{\mu},\boldsymbol{\lambda})}{\partial\mu_\ell}\Big/\psi^q_{\mathsf{pa}(k)\to k}(\boldsymbol{\mu},\boldsymbol{\lambda})\right)\\
&\overset{\text{(d)}}{=} \sum_{\ell\in\mathcal{N}} \frac{1}{Q}\sum_{q=1}^{Q}\sum_{j\in\mathsf{T}\setminus\mathcal{N}} \mathbb{1}_{\{c^q\in\mathsf{T}_j\}}\cdot\left(\frac{1}{\lambda_j}-\frac{1}{\lambda_{\mathsf{pa}(j)}}\right)\cdot\psi^q_{j\to\ell}(\boldsymbol{\mu},\boldsymbol{\lambda})\cdot\left(\sum_{k\in\mathsf{path}(j,\ell]}\frac{\psi^q_{k\to\ell}(\boldsymbol{\mu},\boldsymbol{\lambda})-\psi^q_{\mathsf{pa}(k)\to\ell}(\boldsymbol{\mu},\boldsymbol{\lambda})}{\lambda_{\mathsf{pa}(k)}}\right)\\
&\overset{\text{(e)}}{\le} \sum_{\ell\in\mathcal{N}} \frac{1}{Q}\sum_{q=1}^{Q}\sum_{j\in\mathsf{T}\setminus\mathcal{N}} \mathbb{1}_{\{c^q\in\mathsf{T}_j\}}\cdot\left(\frac{1}{\lambda_j}-\frac{1}{\lambda_{\mathsf{pa}(j)}}\right)\cdot\psi^q_{j\to\ell}(\boldsymbol{\mu},\boldsymbol{\lambda})\cdot\left(\sum_{k\in\mathsf{path}(j,\ell]}\frac{\psi^q_{k\to\ell}(\boldsymbol{\mu},\boldsymbol{\lambda})-\psi^q_{\mathsf{pa}(k)\to\ell}(\boldsymbol{\mu},\boldsymbol{\lambda})}{\lambda_{\mathsf{lower}}}\right)\\
&= \sum_{\ell\in\mathcal{N}} \frac{1}{Q}\sum_{q=1}^{Q}\sum_{j\in\mathsf{T}\setminus\mathcal{N}} \mathbb{1}_{\{c^q\in\mathsf{T}_j\}}\cdot\left(\frac{1}{\lambda_j}-\frac{1}{\lambda_{\mathsf{pa}(j)}}\right)\cdot\psi^q_{j\to\ell}(\boldsymbol{\mu},\boldsymbol{\lambda})\cdot\left(\psi^q_{\ell\to\ell}(\boldsymbol{\mu},\boldsymbol{\lambda})-\psi^q_{j\to\ell}(\boldsymbol{\mu},\boldsymbol{\lambda})\right)\Big/\lambda_{\mathsf{lower}}\\
&\overset{\text{(f)}}{\le} \sum_{\ell\in\mathcal{N}} \frac{1}{Q}\sum_{q=1}^{Q}\sum_{j\in\mathsf{T}\setminus\mathcal{N}} \mathbb{1}_{\{c^q\in\mathsf{T}_j\}}\cdot\left(\frac{1}{\lambda_j}-\frac{1}{\lambda_{\mathsf{pa}(j)}}\right)\cdot\psi^q_{j\to\ell}(\boldsymbol{\mu},\boldsymbol{\lambda})\Big/\lambda_{\mathsf{lower}}\\
&= \frac{1}{Q}\sum_{q=1}^{Q}\sum_{j\in\mathsf{T}\setminus\mathcal{N}} \mathbb{1}_{\{c^q\in\mathsf{T}_j\}}\cdot\left(\frac{1}{\lambda_j}-\frac{1}{\lambda_{\mathsf{pa}(j)}}\right)\cdot\frac{\sum_{\ell\in\mathcal{N}}\psi^q_{j\to\ell}(\boldsymbol{\mu},\boldsymbol{\lambda})}{\lambda_{\mathsf{lower}}}\\
&\overset{\text{(g)}}{=} \frac{1}{Q\cdot\lambda_{\mathsf{lower}}}\sum_{q=1}^{Q}\sum_{j\in\mathsf{T}\setminus\mathcal{N}} \mathbb{1}_{\{c^q\in\mathsf{T}_j\}}\cdot\left(\frac{1}{\lambda_j}-\frac{1}{\lambda_{\mathsf{pa}(j)}}\right)\\
&\overset{\text{(h)}}{=} \frac{1}{Q\cdot\lambda_{\mathsf{lower}}}\sum_{q=1}^{Q}\frac{1}{\lambda_{\mathsf{pa}(c^q)}}\\
&\overset{\text{(i)}}{\le} \frac{1}{\lambda^2_{\mathsf{lower}}}
\end{aligned}
$$

where the justifications for the equalities and inequalities in (a) - (i) are given below.

(a) The equality follows from the expression for $\frac{\partial\mathsf{NegLog}(\boldsymbol{\mu},\boldsymbol{\lambda})}{\partial\mu_\ell}$ from Lemma D.2 with $\boldsymbol{\lambda}=\boldsymbol{\lambda}(\boldsymbol{\delta})$.

(b) The equality since $\psi^q_{j\to\ell}(\boldsymbol{\mu},\boldsymbol{\lambda})=\prod_{k\in\mathsf{path}(j,\ell]}\psi^q_{\mathsf{pa}(k)\to k}(\boldsymbol{\mu},\boldsymbol{\lambda})$.

(c) The equality follows from the product rule of derivatives.

(d) The equality follows since $\psi^q_{\mathsf{pa}(k)\to k}(\boldsymbol{\mu},\boldsymbol{\lambda})=e^{-\left(W^q_{\mathsf{pa}(k)}(\boldsymbol{\mu},\boldsymbol{\lambda})-W^q_k(\boldsymbol{\mu},\boldsymbol{\lambda})\right)/\lambda_{\mathsf{pa}(k)}}$ and $\frac{\partial W^q_k(\boldsymbol{\mu},\boldsymbol{\lambda})}{\partial\mu_\ell}=\psi^q_{k\to\ell}(\boldsymbol{\mu},\boldsymbol{\lambda})$ for all $k\in\mathsf{T}\setminus\mathcal{N}$ from Lemma B.4.

(e) The inequality follows since $\lambda_j\ge\lambda_{\mathsf{lower}}$ for all $j\in\mathsf{T}\setminus\mathcal{N}$ by assumption and $\psi^q_{k\to\ell}(\boldsymbol{\mu},\boldsymbol{\lambda})\ge\psi^q_{\mathsf{pa}(k)\to\ell}(\boldsymbol{\mu},\boldsymbol{\lambda})$ for all $k\in\mathsf{path}(j,\ell]$.

(f) The equality follows since $\psi^q_{\ell\to\ell}(\boldsymbol{\mu},\boldsymbol{\lambda})-\psi^q_{j\to\ell}(\boldsymbol{\mu},\boldsymbol{\lambda})\le 1$.

(g) The equality follows since $\sum_{\ell\in\mathcal{N}}\psi^q_{j\to\ell}(\boldsymbol{\mu},\boldsymbol{\lambda})=1$ for any j such that $c^q\in\mathsf{T}_j$.

(h) The equality follows since $\frac{1}{\lambda_{\mathsf{pa}(\mathsf{root})}}=+\infty$.

(i) The inequality follows since $\lambda_{\mathsf{pa}(c^q)}\ge\lambda_{\mathsf{lower}}$ by assumption.

Since $f(\boldsymbol{\mu})$ is convex on $\mathbb{R}^n$ (see proof of Theorem 3.1 above), all the eigenvalues of the Hessian $\nabla^2 f(\boldsymbol{\mu})$ are non-negative, for any $\boldsymbol{\mu} \in \mathbb{R}^n$. Then, it follows that:

$$\begin{aligned}\text{max eigenvalue of } \nabla^2 f(\boldsymbol{\mu}) &\leq \text{sum of eignvalues of } \nabla^2 f(\boldsymbol{\mu}) \\ &= \text{trace of } \nabla^2 f(\boldsymbol{\mu}) \\ &= \sum_{\ell \in \mathcal{N}} \frac{\partial^2 f(\boldsymbol{\mu})}{\partial \mu_\ell^2} \\ &\leq \frac{1}{\lambda_{\mathsf{lower}}^2}\end{aligned}$$

This establishes the smoothness of $f(\boldsymbol{\mu})$. □

We are now ready to prove Theorem 3.4. Since $\delta_j^{(s)} \leq \delta_{\mathsf{upper}}$ for all $j \in \mathsf{T} \setminus (\mathcal{N} \cup \{\mathsf{root}\})$ by assumption, it follows that $\lambda_j^{(s)} \geq \lambda_{\mathsf{lower}}$ for all $j \in \mathsf{T} \setminus \mathcal{N}$, where $\boldsymbol{\lambda}^{(s)} = \boldsymbol{\lambda}(\boldsymbol{\delta}^{(s)})$. Then, it follows from Lemma E.4 that $\mathsf{NegLog}(\boldsymbol{\mu}, \boldsymbol{\delta}^{(s)})$ is L-smooth. We then leverage the standard quadratic upper bound property (Bubeck 2015, Lemma 3.4) for smooth functions to obtain:

$$\begin{aligned}&\mathsf{NegLog}(\boldsymbol{\mu}_{\mathsf{GD}}^{(s+1)}, \boldsymbol{\delta}^{(s)}) \\ &\leq \mathsf{NegLog}(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) + \left\langle \nabla_{\boldsymbol{\mu}} \mathsf{NegLog}(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}), \boldsymbol{\mu}_{\mathsf{GD}}^{(s+1)} - \boldsymbol{\mu}^{(s)} \right\rangle + \frac{L}{2} \left\| \boldsymbol{\mu}_{\mathsf{GD}}^{(s+1)} - \boldsymbol{\mu}^{(s)} \right\|_2^2 \\ &= \mathsf{NegLog}(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) - \frac{1}{L} \left\langle \nabla_{\boldsymbol{\mu}} \mathsf{NegLog}(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}), \nabla_{\boldsymbol{\mu}} \mathsf{NegLog}(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) \right\rangle + \frac{1}{2L} \left\| \nabla_{\boldsymbol{\mu}} \mathsf{NegLog}(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) \right\|_2^2 \\ &\text{(using the definition of } \boldsymbol{\mu}_{\mathsf{GD}}^{(s+1)}) \\ &= \mathsf{NegLog}(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) - \frac{1}{2L} \left\| \nabla_{\boldsymbol{\mu}} \mathsf{NegLog}(\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) \right\|_2^2\end{aligned}$$

■

E.5.1 Tightness of improvement bound. We now present an instance where both the upper bound on the smoothness constant L in Lemma E.4 and the ℓ_1-ℓ_2 norm inequality are tight, so that the improvement bound in Theorem 3.4 is the best possible bound.

For that, we consider the special where each offer set has exactly two products and customers make choices according to the MNL model. Specifically, suppose that we are given the following dataset consisting of $Q = 2 \cdot (n-1)$ transactions:

$$(\mathcal{S}^q, c^q) = \begin{cases} (\{1, q+1\}, q+1) & \text{for } 1 \leq q \leq n-1 \\ (\{1, q-n+2\}, 1) & \text{for } n \leq q \leq Q \end{cases} \tag{EC.11}$$

Tightness of upper bound for L. In the special case when each offer set has exactly two products, the MNL model is equivalent to the Bradley-Terry model. For the Bradley-Terry model, Vojnovic et al. (2020, Lemma 3.1) showed that the negative log-likelihood function is L-smooth on $\mathbb{R}^n$, where $L = \lambda_{\max}(\mathbf{L}_{\mathbf{M}})/4Q$.[12] Here, $\mathbf{M}$ is an $n \times n$ matrix with $M_{i,j}$ equal to the

[12] Vojnovic et al. (2020) consider the un-normalized negative log-likelihood function, and therefore, do not have the normalization by the total number of transactions Q in their result.

number of times offer-set $\{i,j\}$ is observed in the transaction data, and $\lambda_{\max}(\mathbf{L_M})$ is the largest eigenvalue of the Laplacian matrix $\mathbf{L_M} = \mathbf{D_M} - \mathbf{M}$, where $\mathbf{D_M}$ is a diagonal matrix whose diagonal elements are the row sums of $\mathbf{M}$.

For the above dataset, it is easy to check that the product co-occurrence matrix $\mathbf{M} = 2\mathbf{M}'$ where $\mathbf{M}'$ is the adjacency matrix of the undirected version of the comparison graph Comp defined in Definition 2.3, obtained by replacing the pair of directed edges $\{(1,j),(j,1)\}$ by the undirected edge $\{1,j\}$ for all $j \in \mathcal{N} \setminus \{1\}$. We refer to this graph as Comp'. Now, since the degree of the node corresponding to product 1 in Comp' is equal to $n-1$, it follows from Zhang (2011, Theorem 3.19) that $\lambda_{\max}(\mathbf{L_{M'}}) = n$. Further, since $\mathbf{L_M} = 2\mathbf{L_{M'}}$, it follows that $L = \lambda_{\max}(\mathbf{L_M})/4Q = 2\lambda_{\max}(\mathbf{L_{M'}})/4Q = 2n/4Q = \frac{n}{4\cdot(n-1)}$. For large n, we have that $L \approx 1/4$. Finally, note that $\lambda_{\mathsf{lower}} = 1$ for the MNL model. This shows that the upper bound $1/\lambda^2_{\mathsf{lower}}$ is tight up to constant factors.

Tightness of the ℓ_1-ℓ_2 norm inequality. To establish this result, we exhibit an initial solution $\boldsymbol{\mu}^{(0)}$ for which the inequality is tight. First, note that for the dataset described in Equation (EC.11), it is easy to check that $\mathsf{sales}_1 = \frac{n-1}{2\cdot(n-1)} = 1/2$, whereas $\mathsf{sales}_\ell = \frac{1}{2\cdot(n-1)}$ for all $\ell \in \mathcal{N} \setminus \{1\}$. Further, from Lemma D.2 it follows that for all $\ell \in \mathcal{N}$

$$\frac{\partial \mathsf{NegLog}(\boldsymbol{\mu})}{\partial \mu_\ell} \;=\; \frac{1}{Q}\sum_{q=1}^{Q} \psi^q_{\mathsf{root}\to\ell}(\boldsymbol{\mu}) - \mathsf{sales}_\ell, \tag{EC.12}$$

where note that we have dropped the $\boldsymbol{\lambda}$ term because we are working with an MNL model. Now, supposing that n is even, consider an initial solution $\boldsymbol{\mu}^{(0)}$ of the following form:

$$\mu^{(0)}_\ell = \begin{cases} 0 & \text{if } \ell = 1 \\ \log(2n-1) & \text{if } 1 < \ell \le n/2 \\ -\log(2n-1) & \text{o.w.} \end{cases}$$

Next, consider any $1 < \ell \le n/2$. Then, it follows that for the dataset described in in Equation (EC.11):

$$\begin{aligned} \frac{1}{Q}\sum_{q=1}^{Q} \psi^q_{\mathsf{root}\to\ell}(\boldsymbol{\mu}^{(0)}) &= \frac{1}{Q}\cdot\left(\psi^{\ell-1}_{\mathsf{root}\to\ell}(\boldsymbol{\mu}^{(0)}) + \psi^{n+\ell-2}_{\mathsf{root}\to\ell}(\boldsymbol{\mu}^{(0)})\right) \\ &= \frac{2}{Q}\frac{\exp(\mu^{(0)}_\ell)}{1+\exp(\mu^{(0)}_\ell)} \\ &= \frac{2\cdot(2n-1)}{2\cdot(n-1)\cdot 2n} \\ &= \frac{2n-1}{2n\cdot(n-1)} \\ &= \frac{1}{2\cdot(n-1)} + \frac{1}{2n} \\ &= \mathsf{sales}_\ell + \frac{1}{2n} \end{aligned} \tag{EC.13}$$

In a similar fashion, it can be verified that

$$\frac{1}{Q}\sum_{q=1}^{Q}\psi^{q}_{\mathsf{root}\to\ell}(\boldsymbol{\mu}^{(0)}) = \mathsf{sales}_\ell - \frac{1}{2n} \quad \text{for all } n/2 < \ell \le n \tag{EC.14}$$

Then, combining (EC.12), (EC.13), and (EC.14), it follows that

$$\frac{\partial \mathsf{NegLog}(\boldsymbol{\mu}^{(0)})}{\partial \mu_\ell} = \begin{cases} \frac{1}{2n} & \text{if } 1 < \ell \le n/2 \\ -\frac{1}{2n} & \text{if } n/2 < \ell \le n \end{cases} \tag{EC.15}$$

Further, from the proof of Lemma E.3, it follows that

$$\begin{aligned} \sum_{\ell=1}^{n}\frac{\partial \mathsf{NegLog}(\boldsymbol{\mu}^{(0)})}{\partial \mu_\ell} = 0 &\implies \frac{\partial \mathsf{NegLog}(\boldsymbol{\mu}^{(0)})}{\partial \mu_1} + (n/2-1)\cdot 1/2n - (n/2)\cdot(1/2n) = 0 \\ &\implies \frac{\partial \mathsf{NegLog}(\boldsymbol{\mu}^{(0)})}{\partial \mu_1} = \frac{1}{2n} \end{aligned} \tag{EC.16}$$

Finally, given (EC.16) and (EC.15), it can easily be verified that $\left\|\nabla_{\boldsymbol{\mu}}\mathsf{NegLog}(\boldsymbol{\mu}^{(0)})\right\|_2^2 = \frac{1}{n}\left\|\nabla_{\boldsymbol{\mu}}\mathsf{NegLog}(\boldsymbol{\mu}^{(0)})\right\|_1^2$. ■

E.6 Proof of Theorem 3.5

We first establish that the constant D is finite, for which it is sufficient to show that the set $\mathcal{U} = \left\{\boldsymbol{\mu} \in \mathsf{Dom}_1 : \mathsf{NegLog}(\boldsymbol{\mu}) \le \mathsf{NegLog}(\boldsymbol{\mu}^{(0)})\right\}$ is bounded. From the PROOF OF SUFFICIENCY argument in the proof of Theorem 2.4 in Appendix C.2 above, it follows that for all $\boldsymbol{\mu} \in \mathsf{Dom}_1$ (since $\mathsf{Dom}_1 = \mathcal{D}_1$):

$$\mathsf{NegLog}(\boldsymbol{\mu}) \ge \|\boldsymbol{\mu}\|_\infty b^*/Q,$$

where b^* is as defined in Appendix C.2 and $b^* > 0$ by Lemma C.1. Therefore, it follows that for any $\boldsymbol{\mu} \in \mathcal{U}$:

$$\mathsf{NegLog}(\boldsymbol{\mu}) \le \mathsf{NegLog}(\boldsymbol{\mu}^{(0)}) \implies \|\boldsymbol{\mu}\|_\infty \le \frac{\mathsf{NegLog}(\boldsymbol{\mu}^{(0)})\cdot Q}{b^*},$$

and therefore the set $\mathcal{U}$ is bounded.

Next, we derive the convergence rates. We begin with the guarantee for the MM algorithm. The proof follows from existing results, see e.g., Allen-Zhu and Orecchia (2014, Fact B.1 in Appendix B). We reproduce the arguments here for completeness. For each $s' \ge 0$, define $\mathsf{NegLogGap}^{(s')} := \mathsf{NegLog}(\boldsymbol{\mu}^{(s')}) - \mathsf{NegLog}(\boldsymbol{\mu}^*)$. Note that $\mathsf{NegLogGap}^{(s')} \ge 0$ for all $s' \ge 0$. Moreover, since the result is trivially true for any $s \ge 1$ such that $\mathsf{NegLogGap}^{(s)} = 0$, we suppose that $\mathsf{NegLogGap}^{(s')} > 0$ for all $0 \le s' \le s$.

Now, recall that $\mathsf{NegLog}(\boldsymbol{\mu})$ is convex in $\boldsymbol{\mu}$. Then, it follows that for any $s' \ge 0$:

$$\begin{aligned} \mathsf{NegLogGap}^{(s')} &= \mathsf{NegLog}(\boldsymbol{\mu}^{(s')}) - \mathsf{NegLog}(\boldsymbol{\mu}^*) \\ &\le \left\langle \nabla \mathsf{NegLog}(\boldsymbol{\mu}^{(s')}), \boldsymbol{\mu}^{(s')} - \boldsymbol{\mu}^* \right\rangle \\ &\le \left\| \nabla \mathsf{NegLog}(\boldsymbol{\mu}^{(s')}) \right\|_1 \cdot \left\| \boldsymbol{\mu}^{(s')} - \boldsymbol{\mu}^* \right\|_\infty \\ &\le \left\| \nabla \mathsf{NegLog}(\boldsymbol{\mu}^{(s')}) \right\|_1 \cdot D \end{aligned} \tag{EC.17}$$

where the first inequality follows from the subgradient inequality for convex functions (note that we write the gradient as $\nabla\mathsf{NegLog}$ instead of $\nabla_{\boldsymbol{\mu}}\mathsf{NegLog}$), the second follows from the Holder's inequality, and the final follows since the MM algorithm guarantees an improving solution in each iteration and using the definition of D in the statement of the theorem.

Next, by plugging in $\lambda_{\mathsf{lower}} = 1$ in the improvement bound from Theorem 3.3, it follows that for all $0 \le s' < s$:

$$\begin{aligned}
&\mathsf{NegLog}(\boldsymbol{\mu}^{(s'+1)}) - \mathsf{NegLog}(\boldsymbol{\mu}^{(s')}) \le -\frac{1}{2}\left\|\nabla\mathsf{NegLog}(\boldsymbol{\mu}^{(s')})\right\|_1^2 \\
\implies &\mathsf{NegLogGap}^{(s'+1)} - \mathsf{NegLogGap}^{(s')} \le -\frac{1}{2}\left\|\nabla\mathsf{NegLog}(\boldsymbol{\mu}^{(s')})\right\|_1^2 \\
\implies &\mathsf{NegLogGap}^{(s')} - \mathsf{NegLogGap}^{(s'+1)} \ge \frac{1}{2}\left\|\nabla\mathsf{NegLog}(\boldsymbol{\mu}^{(s')})\right\|_1^2 \\
\implies &\mathsf{NegLogGap}^{(s')} - \mathsf{NegLogGap}^{(s'+1)} \ge \frac{\left(\mathsf{NegLogGap}^{(s')}\right)^2}{2D^2} \\
\implies &\frac{\mathsf{NegLogGap}^{(s')} - \mathsf{NegLogGap}^{(s'+1)}}{\mathsf{NegLogGap}^{(s')} \cdot \mathsf{NegLogGap}^{(s'+1)}} \ge \frac{\mathsf{NegLogGap}^{(s')}}{2D^2 \cdot \mathsf{NegLogGap}^{(s'+1)}} \\
\implies &\frac{1}{\mathsf{NegLogGap}^{(s'+1)}} - \frac{1}{\mathsf{NegLogGap}^{(s')}} \ge \frac{\mathsf{NegLogGap}^{(s')}}{2D^2 \cdot \mathsf{NegLogGap}^{(s'+1)}} \\
\implies &\frac{1}{\mathsf{NegLogGap}^{(s'+1)}} - \frac{1}{\mathsf{NegLogGap}^{(s')}} \ge \frac{1}{2D^2},
\end{aligned}$$

where the third implication follows from (EC.17), the fourth follows from diving both sides by $\mathsf{NegLogGap}^{(s')} \cdot \mathsf{NegLogGap}^{(s'+1)}$, and the final follows since $\mathsf{NegLogGap}^{(s')} \ge \mathsf{NegLogGap}^{(s'+1)}$ as the MM algorithm guarantees an improving solution in each iteration.

Summing the last inequality above from $s' = 0$ to $s-1$ (note that we have $s \ge 1$), it follows that

$$\begin{aligned}
\sum_{s'=0}^{s-1}\left(\frac{1}{\mathsf{NegLogGap}^{(s'+1)}} - \frac{1}{\mathsf{NegLogGap}^{(s')}}\right) &\ge \sum_{s'=0}^{s-1}\frac{1}{2D^2} \\
\implies \frac{1}{\mathsf{NegLogGap}^{(s)}} - \frac{1}{\mathsf{NegLogGap}^{(0)}} &\ge \frac{s}{2D^2} \\
\implies \frac{1}{\mathsf{NegLogGap}^{(s)}} &\ge \frac{s}{2D^2} \\
\implies \mathsf{NegLogGap}^{(s)} &\le \frac{2D^2}{s} \\
\implies \mathsf{NegLog}(\boldsymbol{\mu}^{(s)}) - \mathsf{NegLog}(\boldsymbol{\mu}^*) &\le \frac{2D^2}{s}
\end{aligned}$$

The result then follows.

For the GD algorithm, starting from the improvement bound in Theorem 3.4 and using the fact that GD with step size $1/L$ also guarantees an improving solution in each iteration, the above sequence of arguments can be repeated to show that for all $s \ge 1$

$$\mathsf{NegLog}(\boldsymbol{\mu}_{\mathsf{GD}}^{(s)}) - \mathsf{NegLog}(\boldsymbol{\mu}^*) \le \frac{2LnD^2}{s}$$

The result then follows.

E.7 Proof of Theorem 3.7

The proof of the theorem makes use of the following lemma:

Lemma E.5 *Given any $\bar{\boldsymbol{\delta}} \in \mathsf{Dom}_2$, for each $q = 1, \ldots, Q$ and each nonleaf node $j \in \mathsf{T}^q \setminus \mathcal{S}^q$, a majorizing surrogate for the function $\boldsymbol{\mu} \mapsto \dfrac{\sum_{k \in \mathsf{Children}(j) \cap \mathsf{T}^q} e^{W_k^q\left(\boldsymbol{\mu},\, \boldsymbol{\lambda}(\bar{\boldsymbol{\delta}})\right) \times \exp\left(\bar{\delta}_{[j]}\right)}}{\sum_{k \in \mathsf{Children}(j) \cap \mathsf{T}^q} e^{W_k^q\left(\bar{\boldsymbol{\mu}},\, \boldsymbol{\lambda}(\bar{\boldsymbol{\delta}})\right) \times \exp\left(\bar{\delta}_{[j]}\right)}}$ at $\bar{\boldsymbol{\mu}}$ is given by the following separable function:*

$$\boldsymbol{\mu} \;\mapsto\; \mathsf{C}(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\delta}}) \;+\; \exp\left(\bar{\delta}_{[j]}\right) \sum_{\ell \in \mathcal{N}} \exp\left(-\bar{\delta}_{[\mathsf{pa}(\ell)]}\right) \times \psi_{j \to \ell}^q\left(\bar{\boldsymbol{\mu}},\, \boldsymbol{\lambda}(\bar{\boldsymbol{\delta}})\right) \times e^{(\mu_\ell - \bar{\mu}_\ell) \times \exp\left(\bar{\delta}_{[\mathsf{pa}(\ell)]}\right)} \;,$$

where $\mathsf{C}(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\delta}})$ is a constant depending only on $(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\delta}})$ that is irrelevant for our optimization.

Proof. Fix an arbitrary q and let $\bar{\boldsymbol{\lambda}} = \boldsymbol{\lambda}(\bar{\boldsymbol{\delta}})$. We will prove the result by induction on the height of j. For the base case, suppose j has a height of one, that is, it is a parent of some leaf node. In this case, note that for each $\ell \in \mathsf{Children}(j)$, $\exp\left(\bar{\delta}_{[j]}\right) \times \exp\left(-\bar{\delta}_{[\mathsf{pa}(\ell)]}\right) = 1$. Then,

$$\begin{aligned}
&\frac{\sum_{k \in \mathsf{Children}(j) \cap \mathsf{T}^q} e^{W_k^q\left(\boldsymbol{\mu},\, \boldsymbol{\lambda}(\bar{\boldsymbol{\delta}})\right) \times \exp\left(\bar{\delta}_{[j]}\right)}}{\sum_{k \in \mathsf{Children}(j) \cap \mathsf{T}^q} e^{W_k^q\left(\bar{\boldsymbol{\mu}},\, \boldsymbol{\lambda}(\bar{\boldsymbol{\delta}})\right) \times \exp\left(-\bar{\delta}_{[j]}\right)}} \\
&\overset{(a)}{=} \frac{\sum_{\ell \in \mathsf{Children}(j) \cap \mathcal{S}^q} e^{\mu_\ell / \bar{\lambda}_j}}{\sum_{i \in \mathsf{Children}(j) \cap \mathcal{S}^q} e^{\bar{\mu}_i / \bar{\lambda}_j}} \\
&= \sum_{\ell \in \mathsf{Children}(j) \cap \mathcal{S}^q} \frac{e^{\bar{\mu}_\ell / \bar{\lambda}_j}}{\sum_{i \in \mathsf{Children}(j) \cap \mathcal{S}^q} e^{\bar{\mu}_i / \bar{\lambda}_j}} \times e^{(\mu_\ell - \bar{\mu}_\ell)/\bar{\lambda}_{\mathsf{pa}(\ell)}} \\
&= \sum_{\ell \in \mathsf{Children}(j) \cap \mathcal{S}^q} \psi_{j \to \ell}^q\left(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\lambda}}\right) \times e^{(\mu_\ell - \bar{\mu}_\ell)/\bar{\lambda}_{\mathsf{pa}(\ell)}} \\
&\overset{(b)}{=} \sum_{\ell \in \mathcal{N}} \psi_{j \to \ell}^q\left(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\lambda}}\right) \times e^{(\mu_\ell - \bar{\mu}_\ell)/\bar{\lambda}_{\mathsf{pa}(\ell)}} \\
&= \sum_{\ell \in \mathcal{N}} \psi_{j \to \ell}^q\left(\bar{\boldsymbol{\mu}}, \boldsymbol{\lambda}(\bar{\boldsymbol{\delta}})\right) \times e^{(\mu_\ell - \bar{\mu}_\ell) \times \exp\left(\bar{\delta}_{[\mathsf{pa}(\ell)]}\right)} \;,
\end{aligned}$$

where (a) follows from the definition of the weight function at leaf nodes and because $\bar{\boldsymbol{\lambda}} = \boldsymbol{\lambda}(\bar{\boldsymbol{\delta}})$, and (b) follows because if $\ell \notin \mathsf{Children}(j) \cap \mathcal{S}^q$, then $\ell \notin \mathsf{T}^q$, so $\psi_{j \to \ell}^q\left(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\lambda}}\right) = 0$. This completes the base case.

To establish the induction step, suppose that the result holds for all nodes k of height at most H. In other words, for every nonleaf node $k \in \mathsf{T}^q$ of height at most H, there is a constant C_k that depend only on $\bar{\boldsymbol{\mu}}$ and $\bar{\boldsymbol{\delta}}$ such that for all $\boldsymbol{\mu}$,

$$\begin{aligned}
&\frac{\sum_{m \in \mathsf{Children}(k) \cap \mathsf{T}^q} e^{W_m^q\left(\boldsymbol{\mu},\, \boldsymbol{\lambda}(\bar{\boldsymbol{\delta}})\right) \times \exp\left(\bar{\delta}_{[k]}\right)}}{\sum_{m \in \mathsf{Children}(k) \cap \mathsf{T}^q} e^{W_m^q\left(\bar{\boldsymbol{\mu}},\, \boldsymbol{\lambda}(\bar{\boldsymbol{\delta}})\right) \times \exp\left(\bar{\delta}_{[k]}\right)}} \\
&\leq\; \mathsf{C}_k \;+\; \exp\left(\bar{\delta}_{[k]}\right) \sum_{\ell \in \mathcal{N}} \exp\left(-\bar{\delta}_{[\mathsf{pa}(\ell)]}\right) \times \psi_{k \to \ell}^q\left(\bar{\boldsymbol{\mu}}, \boldsymbol{\lambda}(\bar{\boldsymbol{\delta}})\right) \times e^{(\mu_\ell - \bar{\mu}_\ell) \times \exp\left(\bar{\delta}_{[\mathsf{pa}(\ell)]}\right)} \;,
\end{aligned}$$

and the above inequality is tight at $\boldsymbol{\mu} = \bar{\boldsymbol{\mu}}$. Now, consider a nonleaf node j of height $H+1$. Letting C denote a constant that depend only on $(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\delta}})$, we have

$$
\begin{aligned}
&\frac{\sum_{k\in \mathsf{Children}(j)\cap \mathsf{T}^q} e^{W_k^q\left(\boldsymbol{\mu},\, \boldsymbol{\lambda}(\bar{\boldsymbol{\delta}})\right)\times \exp\left(\bar{\delta}_{[j]}\right)}}{\sum_{k\in \mathsf{Children}(j)\cap \mathsf{T}^q} e^{W_k^q\left(\bar{\boldsymbol{\mu}},\, \boldsymbol{\lambda}(\bar{\boldsymbol{\delta}})\right)\times \exp\left(\bar{\delta}_{[j]}\right)}} \\
&\overset{(a)}{=} \frac{\sum_{k\in \mathsf{Children}(j)\cap \mathsf{T}^q} e^{W_k^q\left(\boldsymbol{\mu},\, \bar{\boldsymbol{\lambda}}\right)/\bar{\lambda}_j}}{\sum_{k\in \mathsf{Children}(j)\cap \mathsf{T}^q} e^{W_k^q\left(\bar{\boldsymbol{\mu}},\, \bar{\boldsymbol{\lambda}}\right)/\bar{\lambda}_j}} \\
&\overset{(b)}{=} \sum_{k\in \mathsf{Children}(j)\cap \mathsf{T}^q} \frac{\left(\sum_{m\in \mathsf{Children}(k)\cap \mathsf{T}^q} e^{W_m^q(\boldsymbol{\mu},\bar{\boldsymbol{\lambda}})/\bar{\lambda}_k}\right)^{\bar{\lambda}_k/\bar{\lambda}_j}}{\sum_{s\in \mathsf{Children}(j)\cap \mathsf{T}^q} e^{W_s^q\left(\bar{\boldsymbol{\mu}},\bar{\boldsymbol{\lambda}}\right)/\bar{\lambda}_j}} \\
&\overset{(c)}{\leq} \mathsf{C} + \sum_{k\in \mathsf{Children}(j)\cap \mathsf{T}^q} \frac{\bar{\lambda}_k}{\bar{\lambda}_j} \times \frac{\left(\sum_{m\in \mathsf{Children}(k)\cap \mathsf{T}^q} e^{W_m^q(\bar{\boldsymbol{\mu}},\bar{\boldsymbol{\lambda}})/\bar{\lambda}_k}\right)^{\bar{\lambda}_k/\bar{\lambda}_j}}{\sum_{s\in \mathsf{Children}(j)\cap \mathsf{T}^q} e^{W_s^q\left(\bar{\boldsymbol{\mu}},\bar{\boldsymbol{\lambda}}\right)/\bar{\lambda}_j}} \times \frac{\sum_{m\in \mathsf{Children}(k)\cap \mathsf{T}^q} e^{W_m^q(\boldsymbol{\mu},\bar{\boldsymbol{\lambda}})/\bar{\lambda}_k}}{\sum_{m\in \mathsf{Children}(k)\cap \mathsf{T}^q} e^{W_m^q(\bar{\boldsymbol{\mu}},\bar{\boldsymbol{\lambda}})/\bar{\lambda}_k}} \\
&\overset{(d)}{=} \mathsf{C} + \sum_{k\in \mathsf{Children}(j)\cap \mathsf{T}^q} \frac{\bar{\lambda}_k}{\bar{\lambda}_j} \times \frac{e^{W_k^q(\bar{\boldsymbol{\mu}},\bar{\boldsymbol{\lambda}})/\bar{\lambda}_j}}{\sum_{s\in \mathsf{Children}(j)\cap \mathsf{T}^q} e^{W_s^q\left(\bar{\boldsymbol{\mu}},\bar{\boldsymbol{\lambda}}\right)/\bar{\lambda}_j}} \times \frac{\sum_{m\in \mathsf{Children}(k)\cap \mathsf{T}^q} e^{W_m^q(\boldsymbol{\mu},\bar{\boldsymbol{\lambda}})/\bar{\lambda}_k}}{\sum_{m\in \mathsf{Children}(k)\cap \mathsf{T}^q} e^{W_m^q(\bar{\boldsymbol{\mu}},\bar{\boldsymbol{\lambda}})/\bar{\lambda}_k}} \\
&\overset{(e)}{=} \mathsf{C} + \sum_{k\in \mathsf{Children}(j)\cap \mathsf{T}^q} \frac{\bar{\lambda}_k}{\bar{\lambda}_j} \times \psi_{j\to k}^q(\bar{\boldsymbol{\mu}},\bar{\boldsymbol{\lambda}}) \times \frac{\sum_{m\in \mathsf{Children}(k)\cap \mathsf{T}^q} e^{W_m^q(\boldsymbol{\mu},\bar{\boldsymbol{\lambda}})/\bar{\lambda}_k}}{\sum_{m\in \mathsf{Children}(k)\cap \mathsf{T}^q} e^{W_m^q(\bar{\boldsymbol{\mu}},\bar{\boldsymbol{\lambda}})/\bar{\lambda}_k}} \\
&\overset{(f)}{\leq} \mathsf{C} + \sum_{k\in \mathsf{Children}(j)\cap \mathsf{T}^q} \mathsf{C}_k + \left(\frac{\bar{\lambda}_k}{\bar{\lambda}_j} \times \psi_{j\to k}^q(\bar{\boldsymbol{\mu}},\bar{\boldsymbol{\lambda}}) \times \frac{1}{\bar{\lambda}_k} \sum_{\ell\in\mathcal{N}} \bar{\lambda}_{\mathsf{pa}(\ell)} \times \psi_{k\to\ell}^q\left(\bar{\boldsymbol{\mu}},\bar{\boldsymbol{\lambda}}\right) \times e^{(\mu_\ell-\bar{\mu}_\ell)/\bar{\lambda}_{\mathsf{pa}(\ell)}}\right) \\
&\overset{(g)}{=} \mathsf{C}_j + \sum_{k\in \mathsf{Children}(j)\cap \mathsf{T}^q} \frac{1}{\bar{\lambda}_j} \times \psi_{j\to k}^q(\bar{\boldsymbol{\mu}},\bar{\boldsymbol{\lambda}}) \sum_{\ell\in\mathcal{S}^q\cap \mathsf{T}_k^q} \bar{\lambda}_{\mathsf{pa}(\ell)} \times \psi_{k\to\ell}^q\left(\bar{\boldsymbol{\mu}},\bar{\boldsymbol{\lambda}}\right) \times e^{(\mu_\ell-\bar{\mu}_\ell)/\bar{\lambda}_{\mathsf{pa}(\ell)}} \\
&\overset{(h)}{=} \mathsf{C}_j + \frac{1}{\bar{\lambda}_j} \sum_{\ell\in\mathcal{S}^q\cap \mathsf{T}_j^q} \bar{\lambda}_{\mathsf{pa}(\ell)} \times \psi_{j\to\ell}^q\left(\bar{\boldsymbol{\mu}},\bar{\boldsymbol{\lambda}}\right) \times e^{(\mu_\ell-\bar{\mu}_\ell)/\bar{\lambda}_{\mathsf{pa}(\ell)}} \\
&\overset{(i)}{=} \mathsf{C}_j + \frac{1}{\bar{\lambda}_j} \sum_{\ell\in\mathcal{N}} \bar{\lambda}_{\mathsf{pa}(\ell)} \times \psi_{j\to\ell}^q\left(\bar{\boldsymbol{\mu}},\bar{\boldsymbol{\lambda}}\right) \times e^{(\mu_\ell-\bar{\mu}_\ell)/\bar{\lambda}_{\mathsf{pa}(\ell)}} \\
&\overset{(j)}{=} \mathsf{C}_j + \exp\left(\bar{\delta}_{[j]}\right) \sum_{\ell\in\mathcal{N}} \exp\left(-\bar{\delta}_{[\mathsf{pa}(\ell)]}\right) \times \psi_{j\to\ell}^q\left(\bar{\boldsymbol{\mu}},\boldsymbol{\lambda}(\bar{\boldsymbol{\delta}})\right) \times e^{(\mu_\ell-\bar{\mu}_\ell)\times \exp\left(\bar{\delta}_{[\mathsf{pa}(\ell)]}\right)},
\end{aligned}
$$

where the justifications for the equalities and inequalities in (a) - (j) are given below.

(a) The equality follows because $\bar{\boldsymbol{\lambda}} = \boldsymbol{\lambda}(\bar{\boldsymbol{\delta}})$.

(b) The equality follows from the definition that for each $k \in \mathsf{Children}(j) \cap \mathsf{T}^q$, $W_k^q\left(\boldsymbol{\mu}, \bar{\boldsymbol{\lambda}}\right) = \bar{\lambda}_k \log\left(\sum_{m\in \mathsf{Children}(k)\cap \mathsf{T}^q} e^{W_m^q(\boldsymbol{\mu},\bar{\boldsymbol{\lambda}})/\bar{\lambda}_k}\right)$, so

$$
e^{W_k^q\left(\boldsymbol{\mu},\bar{\boldsymbol{\lambda}}\right)/\bar{\lambda}_j} = \left(\sum_{m\in \mathsf{Children}(k)\cap \mathsf{T}^q} e^{W_m^q(\boldsymbol{\mu},\bar{\boldsymbol{\lambda}})/\bar{\lambda}_k}\right)^{\bar{\lambda}_k/\bar{\lambda}_j}.
$$

(c) For the inequality (c), we apply the following sub-gradient inequality: for all $\alpha \in (0,1]$ and $x, y \in \mathbb{R}_+$, $x^\alpha \leq y^\alpha + \alpha y^{\alpha-1}(x-y) = (1-\alpha)y^\alpha + \alpha y^\alpha \frac{x}{y}$, with equality if and only if $x = y$. The

inequality (c) follows from the application of the sub-gradient inequality where for each $k \in \mathsf{Children}(k) \cap \mathsf{T}^q$, we let

$$x_k = \sum_{m \in \mathsf{Children}(k) \cap \mathsf{T}^q} e^{W_m^q(\boldsymbol{\mu}, \bar{\boldsymbol{\lambda}})/\bar{\lambda}_k}, \qquad y_k = \sum_{m \in \mathsf{Children}(k) \cap \mathsf{T}^q} e^{W_m^q(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\lambda}})/\bar{\lambda}_k}, \qquad \text{and} \qquad \alpha_k = \frac{\bar{\lambda}_k}{\bar{\lambda}_j},$$

and note that $\alpha_k \in (0,1]$ because $\bar{\lambda}_k \le \bar{\lambda}_j$ since $k \in \mathsf{Children}(j)$. Also, the constant C corresponds to $\sum_{k \in \mathsf{Children}(j) \cap \mathsf{T}^q} (1-\alpha_k) y_k^{\alpha_k}$, which only depends on $\bar{\boldsymbol{\mu}}$ and $\bar{\boldsymbol{\lambda}}$.

(d) This follows from the definition of $e^{W_k^q(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\lambda}})/\bar{\lambda}_j}$.

(e) This equality follows from the definition of $\psi_{j \to k}^q(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\lambda}})$.

(f) The inequality (f) follows from an application of the inductive hypothesis to the children of j, and these children have height of at most H, and from collecting terms.

(g) This follows from defining $\mathsf{C}_j = \mathsf{C} + \sum_{k \in \mathsf{Children}(j) \cap \mathsf{T}^q} \mathsf{C}_k$ and use the fact that $\psi_{k \to \ell}^q(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\lambda}}) = 0$ for all $\ell \notin \mathcal{S}^q \cap \mathsf{T}_k^q$.

(h) The equality (h) follows because $\psi_{j \to \ell}^q(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\lambda}}) = \psi_{j \to k}^q(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\lambda}}) \times \psi_{k \to \ell}^q(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\lambda}})$.

(i) The equality (i) follows by noting that for all $\ell \notin \mathcal{S}^q \cap \mathsf{T}_j^q$, $\psi_{j \to \ell}^q(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\lambda}}) = 0$.

(j) Finally, the last equality (j) again follows because $\bar{\boldsymbol{\lambda}} = \boldsymbol{\lambda}(\bar{\boldsymbol{\delta}})$.

This completes the induction step. Therefore, the result holds for all nodes j. ■

We are now ready to prove the theorem. For each $\ell \in \mathcal{N}$, define $G_\ell(\cdot \,|\, \bar{\boldsymbol{\mu}},\ \bar{\boldsymbol{\delta}}) \colon \mathbb{R} \to \mathbb{R}$ as

$$G_\ell(x \,|\, \bar{\boldsymbol{\mu}},\ \bar{\boldsymbol{\delta}}) \ = \ e^{(x - \bar{\mu}_\ell) \times \exp\left(\bar{\delta}_{[\mathsf{pa}(\ell)]}\right)} \cdot a_\ell(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\delta}}) \ - \ (x - \bar{\mu}_\ell) \cdot \exp\left(\bar{\delta}_{[\mathsf{pa}(\ell)]}\right) \cdot \mathsf{sales}_\ell \tag{EC.18}$$

Then, letting $\bar{\boldsymbol{\lambda}} = \boldsymbol{\lambda}(\bar{\boldsymbol{\delta}})$, it follows from Theorem 2.2 that

$$\mathsf{NegLog}(\boldsymbol{\mu}, \bar{\boldsymbol{\delta}}) \ = \ \frac{1}{Q} \sum_{q=1}^{Q} \left\{ W_{\mathsf{root}}^q(\boldsymbol{\mu}, \bar{\boldsymbol{\lambda}}) \ + \sum_{j \in \mathsf{path}(\mathsf{root}, c^q)} W_j^q(\boldsymbol{\mu}, \bar{\boldsymbol{\lambda}}) \cdot \left(1/\bar{\lambda}_j - 1/\bar{\lambda}_{\mathsf{pa}(j)}\right) \ - \ \mu_{c^q} \,/\, \bar{\lambda}_{\mathsf{pa}(c^q)} \right\}.$$

For each nonleaf node $j \in \mathsf{T} \setminus \mathcal{N}$, let $\bar{\zeta}_j = (1/\bar{\lambda}_j) - (1/\bar{\lambda}_{\mathsf{pa}(j)})$ and we set $\bar{\zeta}_{\mathsf{root}} = 1$. Note that $\bar{\zeta}_j \ge 0$ for all j because $0 < \bar{\lambda}_j \le \bar{\lambda}_{\mathsf{pa}(j)}$. Then, we have

$$\begin{aligned}
&\mathsf{NegLog}(\boldsymbol{\mu}, \bar{\boldsymbol{\delta}}) \\
&= \ \frac{1}{Q} \sum_{q=1}^{Q} \left\{ W_{\mathsf{root}}^q(\boldsymbol{\mu}, \bar{\boldsymbol{\lambda}}) \ + \sum_{j \in \mathsf{path}(\mathsf{root}, c^q)} \bar{\zeta}_j \, W_j^q(\boldsymbol{\mu}, \bar{\boldsymbol{\lambda}}) \right\} \ - \ \frac{1}{Q} \sum_{q=1}^{Q} \mu_{c^q} \,/\, \bar{\lambda}_{\mathsf{pa}(c^q)} \\
&\overset{(a)}{=} \ \frac{1}{Q} \sum_{q=1}^{Q} \sum_{j \in \mathsf{T} \setminus \mathcal{N}} \mathbb{1}_{\left\{c^q \in \mathsf{T}_j\right\}} \bar{\zeta}_j \, W_j^q(\boldsymbol{\mu}, \bar{\boldsymbol{\lambda}}) \ - \ \frac{1}{Q} \sum_{\ell \in \mathcal{N}} \left(\mu_\ell \,/\, \bar{\lambda}_{\mathsf{pa}(\ell)}\right) \cdot \left(\sum_{q=1}^{Q} \mathbb{1}_{\{c^q = \ell\}} \right) \\
&\overset{(b)}{=} \ \frac{1}{Q} \sum_{q=1}^{Q} \sum_{j \in \mathsf{T} \setminus \mathcal{N}} \mathbb{1}_{\left\{c^q \in \mathsf{T}_j\right\}} \bar{\zeta}_j \, W_j^q(\boldsymbol{\mu}, \bar{\boldsymbol{\lambda}}) \ - \ \sum_{\ell \in \mathcal{N}} (\mu_\ell - \bar{\mu}_\ell) \cdot \mathsf{sales}_\ell \,/\, \bar{\lambda}_{\mathsf{pa}(\ell)} \ - \ \sum_{\ell \in \mathcal{N}} \left(\bar{\mu}_\ell \,/\, \bar{\lambda}_{\mathsf{pa}(\ell)}\right) \cdot \mathsf{sales}_\ell \,,
\end{aligned}$$

where the equality (a) follows because, for each $j \in \mathsf{T} \setminus \mathcal{N}$, $c^q \in \mathsf{T}_j$ if and only if $j \in \mathsf{path}[\mathsf{root}, c^q)$ and equality (b) follows from the definition of sales_ℓ.

Let $H(\boldsymbol{\mu}) = \frac{1}{Q}\sum_{q=1}^{Q}\sum_{j\in\mathsf{T}\setminus\mathcal{N}} \mathbb{1}_{\{c^q\in\mathsf{T}_j\}}\bar{\zeta}_j\, W_j^q\left(\boldsymbol{\mu},\bar{\boldsymbol{\lambda}}\right)$. We will apply Lemma E.5 to construct a majorizing surrogate function for $H(\boldsymbol{\mu})$. Let C denote a constant that depend only on $(\bar{\boldsymbol{\mu}},\bar{\boldsymbol{\lambda}})$. Then,

$$
\begin{aligned}
& H\left(\boldsymbol{\mu}\right) \\
& \overset{\text{(c)}}{=} \frac{1}{Q}\sum_{q=1}^{Q}\sum_{j\in\mathsf{T}\setminus\mathcal{N}} \mathbb{1}_{\{c^q\in\mathsf{T}_j\}}\,\bar{\zeta}_j\,\bar{\lambda}_j\,\log\left(\sum_{k\in\,\mathsf{Children}(j)\cap\mathsf{T}^q} e^{W_k^q\left(\boldsymbol{\mu},\,\bar{\boldsymbol{\lambda}}\right)/\,\bar{\lambda}_j}\right) \\
& \overset{\text{(d)}}{\leq} \mathsf{C} + \frac{1}{Q}\sum_{q=1}^{Q}\sum_{j\in\mathsf{T}\setminus\mathcal{N}} \mathbb{1}_{\{c^q\in\mathsf{T}_j\}}\,\bar{\zeta}_j\,\bar{\lambda}_j\,\frac{\sum_{k\in\,\mathsf{Children}(j)\cap\mathsf{T}^q}\; e^{W_k^q\left(\boldsymbol{\mu},\,\bar{\boldsymbol{\lambda}}\right)/\,\bar{\lambda}_j}}{\sum_{k\in\,\mathsf{Children}(j)\cap\mathsf{T}^q}\; e^{W_k^q\left(\bar{\boldsymbol{\mu}},\,\bar{\boldsymbol{\lambda}}\right)/\,\bar{\lambda}_j}} \\
& \overset{\text{(e)}}{\leq} \mathsf{C} + \frac{1}{Q}\sum_{q=1}^{Q}\sum_{j\in\mathsf{T}\setminus\mathcal{N}} \mathbb{1}_{\{c^q\in\mathsf{T}_j\}}\,\bar{\zeta}_j\,\bar{\lambda}_j \times \frac{1}{\bar{\lambda}_j}\sum_{\ell\in\mathcal{N}}\bar{\lambda}_{\mathsf{pa}(\ell)} \times \psi_{j\to\ell}^q\left(\bar{\boldsymbol{\mu}},\,\bar{\boldsymbol{\lambda}}\right) \times e^{(\mu_\ell-\bar{\mu}_\ell)\,/\,\bar{\lambda}_{\mathsf{pa}(\ell)}} \\
& \overset{\text{(f)}}{=} \mathsf{C} + \frac{1}{Q}\sum_{\ell\in\mathcal{N}} e^{(\mu_\ell-\bar{\mu}_\ell)\,/\,\bar{\lambda}_{\mathsf{pa}(\ell)}} \times \left[\bar{\lambda}_{\mathsf{pa}(\ell)}\sum_{q=1}^{Q}\sum_{j\in\mathsf{T}\setminus\mathcal{N}} \mathbb{1}_{\{c^q\in\mathsf{T}_j\}}\,\bar{\zeta}_j\,\psi_{j\to\ell}^q\left(\bar{\boldsymbol{\mu}},\,\bar{\boldsymbol{\lambda}}\right)\right] \\
& \overset{\text{(g)}}{=} \mathsf{C} + \frac{1}{Q}\sum_{\ell\in\mathcal{N}} e^{(\mu_\ell-\bar{\mu}_\ell)\,/\,\bar{\lambda}_{\mathsf{pa}(\ell)}} \times \left[\bar{\lambda}_{\mathsf{pa}(\ell)}\sum_{q=1}^{Q}\sum_{j\in\mathsf{T}\setminus\mathcal{N}} \mathbb{1}_{\{c^q\in\mathsf{T}_j\}}\left(1/\bar{\lambda}_j - 1/\bar{\lambda}_{\mathsf{pa}(j)}\right)\,\psi_{j\to\ell}^q\left(\bar{\boldsymbol{\mu}},\,\bar{\boldsymbol{\lambda}}\right)\right]
\end{aligned}
$$

where the justifications for the equalities and inequalities in (c) - (g) are given below.

(c) The equality follows from the definition of $W_j^q\left(\boldsymbol{\mu},\,\bar{\boldsymbol{\lambda}}\right)$.

(d) The inequality follows from applying the subgradient inequality for logarithm, which shows that for all $x\in\mathbb{R}_+$ and $y\in\mathbb{R}_+$, $\log(x)\leq\log(y)+\frac{1}{y}(x-y)=\frac{x}{y}+\log(y)-1$, with equality if and only if $x=y$. Here, for each $q=1,\ldots,Q$ and $j\in\mathsf{T}\setminus\mathcal{N}$,

$$
x_j^q = \sum_{k\in\,\mathsf{Children}(j)} e^{W_k^q\left(\boldsymbol{\mu},\,\bar{\boldsymbol{\lambda}}\right)/\,\bar{\lambda}_j} \qquad \text{and} \qquad y_j^q = \sum_{k\in\,\mathsf{Children}(j)} e^{W_k^q\left(\bar{\boldsymbol{\mu}},\,\bar{\boldsymbol{\lambda}}\right)/\,\bar{\lambda}_j}
$$

and the constant $\mathsf{C}=\sum_{q=1}^{Q}\sum_{j\in\mathsf{T}}\left(\log(y_j^q)-1\right)$, which only depends on $(\bar{\boldsymbol{\mu}},\bar{\boldsymbol{\lambda}})$.

(e) The inequality follows from Lemma E.5, and the constant C is updated accordingly.

(f) The equality follows from algebra and rearranging terms.

(g) The equality follows from the definition of $\bar{\zeta}_j$ and the fact that $\bar{\lambda}_{\mathsf{pa(root)}}=+\infty$.

Putting everything together, we have the following majorizing surrogate for $\boldsymbol{\mu}\mapsto\mathsf{NegLog}(\boldsymbol{\mu},\bar{\boldsymbol{\delta}})$:

$$
\begin{aligned}
\mathsf{C} \;+\; & \frac{1}{Q}\sum_{\ell\in\mathcal{N}} e^{(\mu_\ell-\bar{\mu}_\ell)\,/\,\bar{\lambda}_{\mathsf{pa}(\ell)}} \times \left[\bar{\lambda}_{\mathsf{pa}(\ell)}\sum_{q=1}^{Q}\sum_{j\in\mathsf{T}\setminus\mathcal{N}} \mathbb{1}_{\{c^q\in\mathsf{T}_j\}}\left(1/\bar{\lambda}_j - 1/\bar{\lambda}_{\mathsf{pa}(j)}\right)\,\psi_{j\to\ell}^q\left(\bar{\boldsymbol{\mu}},\,\bar{\boldsymbol{\lambda}}\right)\right] \\
& \qquad - \sum_{\ell\in\mathcal{N}}(\mu_\ell - \bar{\mu}_\ell)\cdot\mathsf{sales}_\ell\,/\,\bar{\lambda}_{\mathsf{pa}(\ell)} \\
= \;\mathsf{C} + & \frac{1}{Q}\sum_{\ell\in\mathcal{N}} e^{\left(\mu_\ell-\bar{\mu}_\ell\right)\times\exp\left(\bar{\delta}_{[\mathsf{pa}(\ell)]}\right)} \times \left[\exp\left(-\bar{\delta}_{[\mathsf{pa}(\ell)]}\right)\sum_{q=1}^{Q}\sum_{j\in\mathsf{T}\setminus\mathcal{N}} \mathbb{1}_{\{c^q\in\mathsf{T}_j\}}\left(\exp\left(\bar{\delta}_{[j]}\right)-\exp\left(\bar{\delta}_{[\mathsf{pa}(j)]}\right)\right)\,\psi_{j\to\ell}^q\left(\bar{\boldsymbol{\mu}},\,\boldsymbol{\lambda}(\bar{\boldsymbol{\delta}})\right)\right] \\
& \qquad - \sum_{\ell\in\mathcal{N}}(\mu_\ell - \bar{\mu}_\ell)\times\exp\left(\bar{\delta}_{[\mathsf{pa}(\ell)]}\right)\times\mathsf{sales}_\ell
\end{aligned}
$$

$$
\begin{aligned}
&= \mathsf{C} + \sum_{\ell \in \mathcal{N}} e^{(\mu_\ell - \bar{\mu}_\ell) \times \exp\left(\bar{\delta}_{[\mathsf{pa}(\ell)]}\right)} \times \exp\left(-\bar{\delta}_{[\mathsf{pa}(\ell)]}\right) \times \left[\left. \frac{\partial \mathsf{NegLog}(\boldsymbol{\mu}, \boldsymbol{\delta})}{\partial \mu_\ell} \right|_{\boldsymbol{\mu} = \bar{\boldsymbol{\mu}},\, \boldsymbol{\delta} = \bar{\boldsymbol{\delta}}} + \mathsf{sales}_\ell \cdot \exp\left(\bar{\delta}_{[\mathsf{pa}(\ell)]}\right) \right] \\
&\qquad\qquad - \sum_{\ell \in \mathcal{N}} (\mu_\ell - \bar{\mu}_\ell) \times \exp\left(\bar{\delta}_{[\mathsf{pa}(\ell)]}\right) \times \mathsf{sales}_\ell \\
&= \mathsf{C} + \sum_{\ell \in \mathcal{N}} e^{(\mu_\ell - \bar{\mu}_\ell) \times \exp\left(\bar{\delta}_{[\mathsf{pa}(\ell)]}\right)} \cdot a_\ell(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\delta}}) - \sum_{\ell \in \mathcal{N}} (\mu_\ell - \bar{\mu}_\ell) \times \exp\left(\bar{\delta}_{[\mathsf{pa}(\ell)]}\right) \times \mathsf{sales}_\ell \\
&= \mathsf{C} + \sum_{\ell \in \mathcal{N}} G_\ell(\mu_\ell \,|\, \bar{\boldsymbol{\mu}},\ \bar{\boldsymbol{\delta}})
\end{aligned}
$$

where the first equality follows since $\bar{\boldsymbol{\lambda}} = \boldsymbol{\lambda}(\bar{\boldsymbol{\delta}})$, the second follows from Lemma D.2, the third from the definition of $a_\ell\left(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\delta}}\right)$, and the final from the definition of $G_\ell(\mu_\ell \,|\, \bar{\boldsymbol{\mu}},\ \bar{\boldsymbol{\delta}})$ in (EC.18).

To show that $G_\ell(\cdot \,|\, \bar{\boldsymbol{\mu}},\ \bar{\boldsymbol{\delta}})$ is strictly convex, it suffices to show that $a_\ell(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\delta}}) > 0$. Using the expression for $\frac{\partial \mathsf{NegLog}(\boldsymbol{\mu}, \boldsymbol{\lambda}(\boldsymbol{\delta}))}{\partial \mu_\ell}$ from Lemma D.2, it follows that

$$
\begin{aligned}
a_\ell\left(\bar{\boldsymbol{\mu}},\, \bar{\boldsymbol{\delta}}\right) &= \frac{\exp\left(-\bar{\delta}_{[\mathsf{pa}(\ell)]}\right)}{Q} \sum_{q=1}^{Q} \sum_{j \in \mathsf{T} \setminus \mathcal{N}} \mathbb{1}_{\{c^q \in \mathsf{T}_j\}} \left(\exp\left(\bar{\delta}_{[j]}\right) - \exp\left(\bar{\delta}_{[\mathsf{pa}(j)]}\right)\right) \psi^q_{j \to \ell}\left(\bar{\boldsymbol{\mu}},\, \boldsymbol{\lambda}(\bar{\boldsymbol{\delta}})\right) \\
&= \frac{\bar{\lambda}_{\mathsf{pa}(\ell)}}{Q} \sum_{q=1}^{Q} \sum_{j \in \mathsf{T} \setminus \mathcal{N}} \mathbb{1}_{\{c^q \in \mathsf{T}_j\}} \left(1/\bar{\lambda}_j - 1/\bar{\lambda}_{\mathsf{pa}(j)}\right) \psi^q_{j \to \ell}\left(\bar{\boldsymbol{\mu}},\, \bar{\boldsymbol{\lambda}}\right) \\
&\geq \frac{\bar{\lambda}_{\mathsf{pa}(\ell)}}{Q} \sum_{q=1}^{Q} \mathbb{1}_{\{c^q \in \mathsf{T}_{\mathsf{root}}\}} \left(1/\bar{\lambda}_{\mathsf{root}} - 1/\bar{\lambda}_{\mathsf{pa}(\mathsf{root})}\right) \psi^q_{\mathsf{root} \to \ell}\left(\bar{\boldsymbol{\mu}},\, \bar{\boldsymbol{\lambda}}\right) \\
&= \frac{\bar{\lambda}_{\mathsf{pa}(\ell)}}{Q} \sum_{q=1}^{Q} \psi^q_{\mathsf{root} \to \ell}\left(\bar{\boldsymbol{\mu}},\, \bar{\boldsymbol{\lambda}}\right) \\
&> 0,
\end{aligned}
$$

where the first inequality follows since each term inside the summation is non-negative, the last equality follows because $\mathbb{1}_{\{c^q \in \mathsf{T}_{\mathsf{root}}\}} = 1$, $\bar{\lambda}_{\mathsf{root}} = 1$ and $\bar{\lambda}_{\mathsf{pa}(\mathsf{root})} = +\infty$, and the final inequality follows because $\bar{\lambda}_{\mathsf{pa}(\ell)} > 0$ and the identifiability condition in Theorem 2.4 ensures that $\ell \in \mathcal{S}^{\hat{q}}$ for some $\hat{q}$ so that $\psi^{\hat{q}}_{\mathsf{root} \to \ell}\left(\bar{\boldsymbol{\mu}},\, \bar{\boldsymbol{\lambda}}\right) > 0$.

Finally, the expression for the minimizer follows from setting $\partial G_\ell(\mu_\ell \,|\, \bar{\boldsymbol{\mu}},\ \bar{\boldsymbol{\delta}})/\partial \mu_\ell = 0$.

E.8 Proof of Lemma 3.8

Recall from Theorem 3.1 that $\mathsf{NegLog}(\bar{\boldsymbol{\mu}}, \boldsymbol{\delta}) = F_1(\bar{\boldsymbol{\mu}}, \boldsymbol{\delta}) - F_2(\bar{\boldsymbol{\mu}}, \boldsymbol{\delta})$ for all $\boldsymbol{\delta} \in \mathsf{Dom}_2$, where both $F_1(\bar{\boldsymbol{\mu}}, \cdot)$ and $F_2(\bar{\boldsymbol{\mu}}, \cdot)$ are strictly convex on Dom_2. By applying the sub-gradient inequality to the strictly convex function $F_2(\bar{\boldsymbol{\mu}}, \boldsymbol{\delta})$ at $\boldsymbol{\delta} = \bar{\boldsymbol{\delta}}$, it follows that

$$
\mathsf{NegLog}(\bar{\boldsymbol{\mu}}, \boldsymbol{\delta}) \leq F_1(\bar{\boldsymbol{\mu}}, \boldsymbol{\delta}) - F_2(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\delta}}) - \sum_{j \in \mathsf{T} \setminus (\mathcal{N} \cup \{\mathsf{root}\})} \frac{\partial F_2}{\partial \delta_j}(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\delta}}) \cdot \left(\delta_j - \bar{\delta}_j\right) \ \ \forall\ \boldsymbol{\delta} \in \mathsf{Dom}_2
$$

$$
\text{with equality if and only if} \quad \boldsymbol{\delta} = \bar{\boldsymbol{\delta}}\ . \qquad \text{(EC.19)}
$$

The result then follows from the definition of a majorizing surrogate. ■

E.9 Improvements via MM and PGD updates for δ

As mentioned in the main text, the improvement guarantee for Step 2 of the A-MM algorithm follows from existing results. In particular, it follows from the standard guarantee for the projected gradient descent (PGD) algorithm for constrained smooth problems, which we reproduce here for completeness. Consider the optimization problem

$$\min_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x}), \tag{EC.20}$$

where f is L-smooth on $\mathcal{X}$ and $\mathcal{X}\subset\mathbb{R}^d$ is a convex set. We define the *gradient mapping* (Nesterov 2013, Section 2.2.3) of f over $\mathcal{X}$, denoted by $\mathcal{G}_{f,\mathcal{X}}\colon \mathcal{X}\to\mathbb{R}^d$ as

$$\mathcal{G}_{f,\mathcal{X}}(\boldsymbol{x}) = L\cdot\left(\boldsymbol{x} - \mathsf{Proj}_{\mathcal{X}}\left(\boldsymbol{x}-\frac{1}{L}\nabla f(\boldsymbol{x})\right)\right), \tag{EC.21}$$

where $\mathsf{Proj}_{\mathcal{X}}(\cdot)$ is the projection operator onto $\mathcal{X}$, i.e. $\mathsf{Proj}_{\mathcal{X}}(\bar{\boldsymbol{x}}) = \arg\min_{\boldsymbol{x}\in\mathcal{X}}\|\boldsymbol{x}-\bar{\boldsymbol{x}}\|_2^2$ for all $\bar{\boldsymbol{x}}\in\mathbb{R}^d$. The gradient mapping plays the role of the gradient in unconstrained optimization problems since it can be shown that $\boldsymbol{x}^*$ is a stationary point for problem (EC.20) if and only if $\|\mathcal{G}_{f,\mathcal{X}}(\boldsymbol{x}^*)\|_2 = 0$.

Now, starting from an initial solution $\boldsymbol{x}^{(0)}\in\mathcal{X}$, consider solving problem (EC.20) using PGD with step size $1/L$, that is, for each $s\geq 0$:

$$\boldsymbol{x}^{(s+1)} = \mathsf{Proj}_{\mathcal{X}}\left(\boldsymbol{x}^{(s)} - \frac{1}{L}\nabla f(\boldsymbol{x}^{(s)})\right)$$

Then, using the quadratic upper bound for smooth functions, it follows that

$$f(\boldsymbol{x}^{(s+1)}) \leq f(\boldsymbol{x}^{(s)}) + \left\langle \nabla f(\boldsymbol{x}^{(s)}), \boldsymbol{x}^{(s+1)}-\boldsymbol{x}^{(s)}\right\rangle + \frac{L}{2}\left\|\boldsymbol{x}^{(s+1)}-\boldsymbol{x}^{(s)}\right\|_2^2 \tag{EC.22}$$

Next, for any $\boldsymbol{x}\in\mathcal{X}$, denote $D(\boldsymbol{x}) = \left\|\boldsymbol{x}-\left(\boldsymbol{x}^{(s)}-\frac{1}{L}\nabla f(\boldsymbol{x}^{(s)})\right)\right\|_2^2$ and note that $\boldsymbol{x}^{(s+1)} = \arg\min_{\boldsymbol{x}\in\mathcal{X}} D(\boldsymbol{x})$. The optimality of $\boldsymbol{x}^{(s+1)}$ implies that there is no descent direction from $\boldsymbol{x}^{(s+1)}$:

$$\left\langle \nabla D(\boldsymbol{x}^{(s+1)}), \boldsymbol{x}-\boldsymbol{x}^{(s+1)}\right\rangle \geq 0 \quad \forall\ \boldsymbol{x}\in\mathcal{X}$$

By plugging in the gradient of $D(\cdot)$ and choosing $\boldsymbol{x}=\boldsymbol{x}^{(s)}$, it follows that

$$\left\langle \nabla f(\boldsymbol{x}^{(s)}), \boldsymbol{x}^{(s+1)}-\boldsymbol{x}^{(s)}\right\rangle \leq -L\left\|\boldsymbol{x}^{(s+1)}-\boldsymbol{x}^{(s)}\right\|_2^2$$

Substituting the above in (EC.22), it follows that

$$\begin{aligned} f(\boldsymbol{x}^{(s+1)}) &\leq f(\boldsymbol{x}^{(s)}) - L\left\|\boldsymbol{x}^{(s+1)}-\boldsymbol{x}^{(s)}\right\|_2^2 + \frac{L}{2}\left\|\boldsymbol{x}^{(s+1)}-\boldsymbol{x}^{(s)}\right\|_2^2 \\ &= f(\boldsymbol{x}^{(s)}) - \frac{L}{2}\left\|\boldsymbol{x}^{(s+1)}-\boldsymbol{x}^{(s)}\right\|_2^2 \\ &= f(\boldsymbol{x}^{(s)}) - \frac{1}{2L}\left\|\mathcal{G}_{f,\mathcal{X}}(\boldsymbol{x}^{(s)})\right\|_2^2, \end{aligned} \tag{EC.23}$$

where the final equality follows from the definition of $\mathcal{G}_{f,\mathcal{X}}(\boldsymbol{x}^{(s)})$.

We adapt the above improvement guarantee to our context. Define the domain $\overline{\mathsf{Dom}_2}$ as:

$$\overline{\mathsf{Dom}_2} = \left\{\boldsymbol{\delta} \in \mathbb{R}^{|\mathsf{T}\setminus(\mathcal{N}\cup\{\mathsf{root}\})|} \;:\; 0 \le \delta_j \le \delta_{\mathsf{upper}} \;\forall\; j \in \mathsf{T}\setminus(\mathcal{N}\cup\{\mathsf{root}\})\right\} \tag{EC.24}$$

Then, we can establish the following:

Lemma E.6 (Improvements via MM and PGD updates for $\boldsymbol{\delta}$) *Suppose $\boldsymbol{\delta}^{(s)} \in \overline{\mathsf{Dom}_2}$ and let $\gamma, \gamma_{\mathsf{GD}} > 0$ denote the smoothness constants of the mappings $\boldsymbol{\delta} \mapsto H\left(\boldsymbol{\delta} \,|\, \boldsymbol{\mu}^{(s+1)},\ \boldsymbol{\delta}^{(s)}\right)$ and $\boldsymbol{\delta} \mapsto \mathsf{NegLog}(\boldsymbol{\mu}^{(s+1)}, \boldsymbol{\delta})$, respectively, over the domain $\overline{\mathsf{Dom}_2}$. If $\boldsymbol{\delta}^{(s+1)} \in \overline{\mathsf{Dom}_2}$, then the MM update guarantees the following improvement in the negative log-likelihood objective:*

$$\mathsf{NegLog}(\boldsymbol{\mu}^{(s+1)}, \boldsymbol{\delta}^{(s+1)}) - \mathsf{NegLog}(\boldsymbol{\mu}^{(s+1)}, \boldsymbol{\delta}^{(s)}) \le -\frac{1}{2\gamma}\left\|\mathcal{G}_{H,\overline{\mathsf{Dom}_2}}(\boldsymbol{\delta}^{(s)})\right\|_2^2$$

Similarly, let $\boldsymbol{\delta}_{\mathsf{GD}}^{(s+1)}$ denote the PGD update with step size $1/\gamma_{\mathsf{GD}}$. If $\boldsymbol{\delta}_{\mathsf{GD}}^{(s+1)} \in \overline{\mathsf{Dom}_2}$, then the PGD update guarantees the following improvement in the negative log-likelihood objective:

$$\mathsf{NegLog}(\boldsymbol{\mu}^{(s+1)}, \boldsymbol{\delta}_{\mathsf{GD}}^{(s+1)}) - \mathsf{NegLog}(\boldsymbol{\mu}^{(s+1)}, \boldsymbol{\delta}^{(s)}) \le -\frac{1}{2\gamma_{\mathsf{GD}}}\left\|\mathcal{G}_{\mathsf{NegLog},\overline{\mathsf{Dom}_2}}(\boldsymbol{\delta}^{(s)})\right\|_2^2$$

Proof. We start with the MM update. First, it can be shown—using an argument analogous to that in the proof of Theorem 3.3 above—that the eigenvalues of the Hessian matrix $\nabla^2 H(\boldsymbol{\delta} \,|\, \boldsymbol{\mu}^{(s+1)},\ \boldsymbol{\delta}^{(s)})$ are bounded above by a constant $\gamma > 0$ that depends on δ_{upper}, for all $\boldsymbol{\delta} \in \overline{\mathsf{Dom}_2}$. This ensures that $H(\boldsymbol{\delta} \,|\, \boldsymbol{\mu}^{(s+1)},\ \boldsymbol{\delta}^{(s)})$ is γ-smooth on $\overline{\mathsf{Dom}_2}$. Now, define $\bar{\boldsymbol{\delta}}^{(s+1)} \in \overline{\mathsf{Dom}_2}$ as follows:

$$\bar{\boldsymbol{\delta}}^{(s+1)} = \mathsf{Proj}_{\overline{\mathsf{Dom}_2}}\left(\boldsymbol{\delta}^{(s)} - \frac{1}{\gamma}\nabla H(\boldsymbol{\delta}^{(s)} \,|\, \boldsymbol{\mu}^{(s+1)},\ \boldsymbol{\delta}^{(s)})\right),$$

Then, it follows from (EC.23) that

$$\begin{aligned} H(\bar{\boldsymbol{\delta}}^{(s+1)} \,|\, \boldsymbol{\mu}^{(s+1)},\ \boldsymbol{\delta}^{(s)}) &\le H(\boldsymbol{\delta}^{(s)} \,|\, \boldsymbol{\mu}^{(s+1)},\ \boldsymbol{\delta}^{(s)}) - \frac{1}{2\gamma}\left\|\mathcal{G}_{H,\overline{\mathsf{Dom}_2}}(\boldsymbol{\delta}^{(s)})\right\|_2^2 \\ &= \mathsf{NegLog}(\boldsymbol{\mu}^{(s+1)}, \boldsymbol{\delta}^{(s)}) - \frac{1}{2\gamma}\left\|\mathcal{G}_{H,\overline{\mathsf{Dom}_2}}(\boldsymbol{\delta}^{(s)})\right\|_2^2, \end{aligned}$$

where the equality follows since $H(\boldsymbol{\delta} \,|\, \boldsymbol{\mu}^{(s+1)},\ \boldsymbol{\delta}^{(s)})$ is a majorizing surrogate. The improvement bound then follows since

$$\mathsf{NegLog}(\boldsymbol{\mu}^{(s+1)}, \boldsymbol{\delta}^{(s+1)}) \le H(\boldsymbol{\delta}^{(s+1)} \,|\, \boldsymbol{\mu}^{(s+1)},\ \boldsymbol{\delta}^{(s)}) \le H(\bar{\boldsymbol{\delta}}^{(s+1)} \,|\, \boldsymbol{\mu}^{(s+1)},\ \boldsymbol{\delta}^{(s)}),$$

where the second inequality follows since $\nabla H(\boldsymbol{\delta}^{(s)} \,|\, \boldsymbol{\mu}^{(s+1)},\ \boldsymbol{\delta}^{(s)}) = \nabla_{\boldsymbol{\delta}}\mathsf{NegLog}(\boldsymbol{\mu}^{(s+1)}, \boldsymbol{\delta}^{(s)})$ and the fact that the MM update $\boldsymbol{\delta}^{(s+1)} \in \overline{\mathsf{Dom}_2}$ corresponds to the optimal step size.

The improvement bound for the GD update $\boldsymbol{\delta}_{\mathsf{GD}}^{(s+1)}$ follows in an identical fashion by leveraging the smoothness of $\mathsf{NegLog}(\boldsymbol{\mu}^{(s+1)}, \boldsymbol{\delta})$. In particular, we can show that $\mathsf{NegLog}(\boldsymbol{\mu}^{(s+1)}, \boldsymbol{\delta})$ is γ_{GD}-smooth on $\overline{\mathsf{Dom}_2}$, for some constant $\gamma_{\mathsf{GD}} > 0$ that depends on δ_{upper}. The result then follows from (EC.23). ■

E.10 Sublinear convergence rate of the A-MM algorithm

As stated in the main text, the improvements bounds from Theorem 3.3 and Lemma E.6 can be combined to establish a sublinear rate of convergence of the A-MM algorithm, as shown in the following theorem, where we leverage the notations introduced above:

Theorem E.7 (Sublinear convergence of A-MM to a stationary point) *Suppose that* $\boldsymbol{\delta}^{(s)} \in \overline{\mathsf{Dom}_2}$ *for all* $s \geq 0$. *Then, the sequence of iterates* $\left((\boldsymbol{\mu}^{(s)}, \boldsymbol{\delta}^{(s)}) : s \geq 0\right)$ *converges to a stationary point of the* MLE problem. *Moreover, we can establish the following guarantee*

$$\min_{0 \leq s' \leq s} \left[\left\| \nabla_{\boldsymbol{\mu}} \mathsf{NegLog}\left(\boldsymbol{\mu}^{(s')}, \boldsymbol{\delta}^{(s')}\right) \right\|_1^2 + \left\| \mathcal{G}_{H, \overline{\mathsf{Dom}_2}}(\boldsymbol{\delta}^{(s')}) \right\|_2^2 \right] \leq \frac{2 \cdot \left(\mathsf{NegLog}(\boldsymbol{\mu}^{(0)}, \boldsymbol{\delta}^{(0)}) - \mathsf{NegLog}^* \right)}{R \cdot (s+1)},$$

where NegLog^* *is the optimal objective for the* MLE problem, *and* $R = \min\left(\lambda_{\mathsf{lower}}^2, 1/\gamma\right)$.

Proof. The first part of the theorem follows from the fact that the A-MM algorithm guarantees an improving solution as long as the current solution is not a stationary point. For the second part, we leverage the improvements from Theorem 3.3 and Lemma E.6 to obtain, for each $s' \geq 0$:

$$\begin{aligned}
&\mathsf{NegLog}\left(\boldsymbol{\mu}^{(s'+1)}, \boldsymbol{\delta}^{(s'+1)}\right) - \mathsf{NegLog}\left(\boldsymbol{\mu}^{(s')}, \boldsymbol{\delta}^{(s')}\right) \\
&= \left[\mathsf{NegLog}\left(\boldsymbol{\mu}^{(s'+1)}, \boldsymbol{\delta}^{(s'+1)}\right) - \mathsf{NegLog}\left(\boldsymbol{\mu}^{(s'+1)}, \boldsymbol{\delta}^{(s')}\right)\right] + \left[\mathsf{NegLog}\left(\boldsymbol{\mu}^{(s'+1)}, \boldsymbol{\delta}^{(s)}\right) - \mathsf{NegLog}\left(\boldsymbol{\mu}^{(s')}, \boldsymbol{\delta}^{(s')}\right)\right] \\
&\leq -\frac{1}{2\gamma} \left\| \mathcal{G}_{H, \overline{\mathsf{Dom}_2}}(\boldsymbol{\delta}^{(s')}) \right\|_2^2 - \frac{\lambda_{\mathsf{lower}}^2}{2} \left\| \nabla_{\boldsymbol{\mu}} \mathsf{NegLog}(\boldsymbol{\mu}^{(s')}, \boldsymbol{\delta}^{(s')}) \right\|_1^2 \\
&\leq -\frac{R}{2} \left[\left\| \mathcal{G}_{H, \overline{\mathsf{Dom}_2}}(\boldsymbol{\delta}^{(s')}) \right\|_2^2 + \left\| \nabla_{\boldsymbol{\mu}} \mathsf{NegLog}(\boldsymbol{\mu}^{(s')}, \boldsymbol{\delta}^{(s')}) \right\|_1^2 \right],
\end{aligned}$$

where the second inequality follows from the definition of R. Summing the above from 0 to any iteration $s \geq 0$, we get

$$\sum_{s'=0}^{s} \left[\mathsf{NegLog}\left(\boldsymbol{\mu}^{(s'+1)}, \boldsymbol{\delta}^{(s'+1)}\right) - \mathsf{NegLog}\left(\boldsymbol{\mu}^{(s')}, \boldsymbol{\delta}^{(s')}\right)\right] \leq -\frac{R}{2} \sum_{s'=0}^{s} \left[\left\| \mathcal{G}_{H, \overline{\mathsf{Dom}_2}}(\boldsymbol{\delta}^{(s')}) \right\|_2^2 + \left\| \nabla_{\boldsymbol{\mu}} \mathsf{NegLog}(\boldsymbol{\mu}^{(s')}, \boldsymbol{\delta}^{(s')}) \right\|_1^2 \right]$$

The left hand side of the inequality above is a telescoping sum, and it is equal to $\mathsf{NegLog}\left(\boldsymbol{\mu}^{(s+1)}, \boldsymbol{\delta}^{(s+1)}\right) - \mathsf{NegLog}\left(\boldsymbol{\mu}^{(0)}, \boldsymbol{\delta}^{(0)}\right)$, which is bounded below by $\mathsf{NegLog}^* - \mathsf{NegLog}\left(\boldsymbol{\mu}^{(0)}, \boldsymbol{\delta}^{(0)}\right)$. We thus have

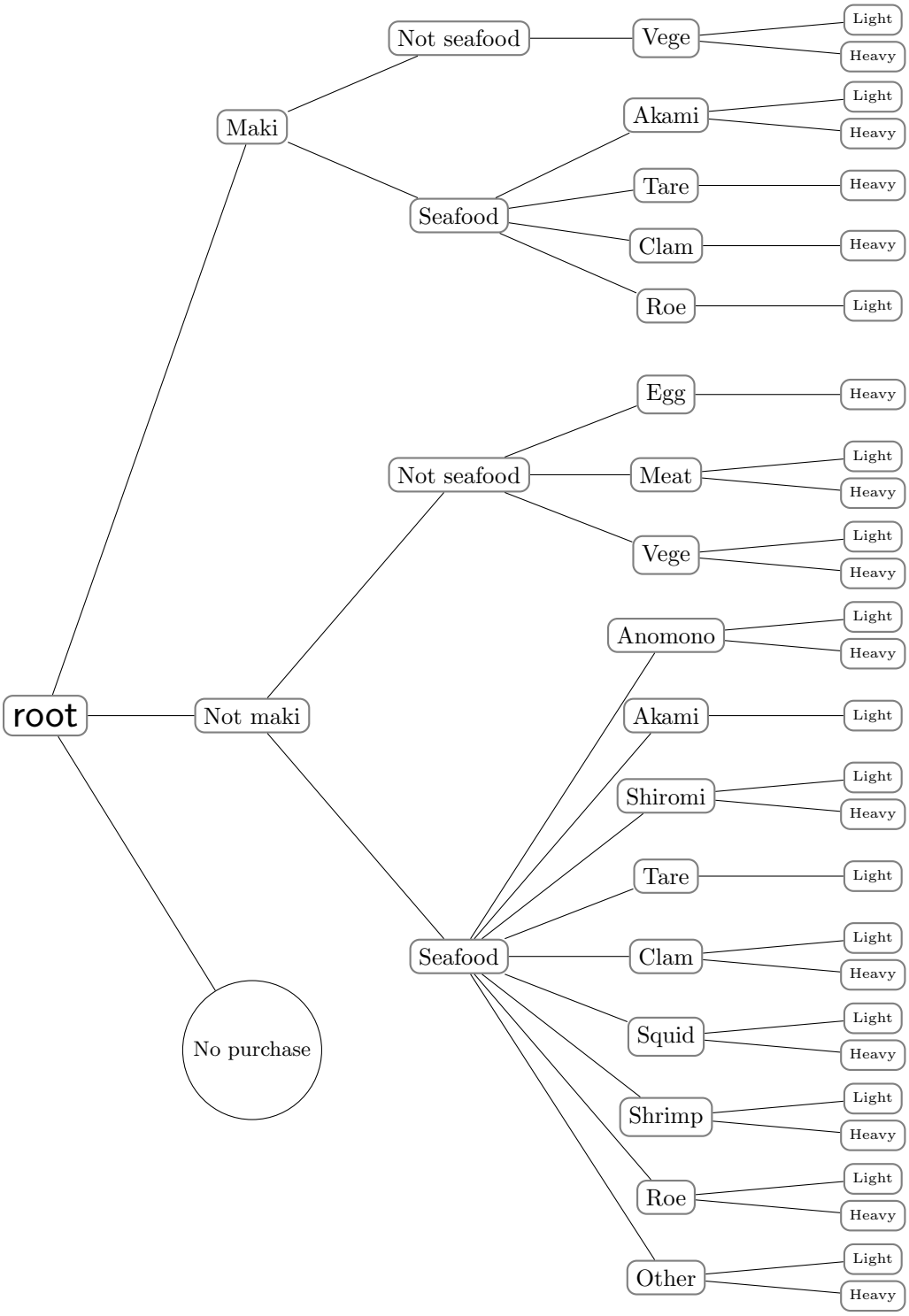
$$\mathsf{NegLog}^* - \mathsf{NegLog}\left(\boldsymbol{\mu}^{(0)}, \boldsymbol{\delta}^{(0)}\right) \leq -\frac{R}{2} \sum_{s'=0}^{s} \left[\left\| \mathcal{G}_{H, \overline{\mathsf{Dom}_2}}(\boldsymbol{\delta}^{(s')}) \right\|_2^2 + \left\| \nabla_{\boldsymbol{\mu}} \mathsf{NegLog}(\boldsymbol{\mu}^{(s')}, \boldsymbol{\delta}^{(s')}) \right\|_1^2 \right] \quad \text{(EC.25)}$$

Finally, using (EC.25) it follows that

$$\begin{aligned}
&\min_{0 \leq s' \leq s} \left[\left\| \nabla_{\boldsymbol{\mu}} \mathsf{NegLog}\left(\boldsymbol{\mu}^{(s')}, \boldsymbol{\delta}^{(s')}\right) \right\|_1^2 + \left\| \mathcal{G}_{H, \overline{\mathsf{Dom}_2}}(\boldsymbol{\delta}^{(s')}) \right\|_2^2 \right] \\
&\leq \frac{1}{s+1} \sum_{s'=0}^{s} \left[\left\| \nabla_{\boldsymbol{\mu}} \mathsf{NegLog}\left(\boldsymbol{\mu}^{(s')}, \boldsymbol{\delta}^{(s')}\right) \right\|_1^2 + \left\| \mathcal{G}_{H, \overline{\mathsf{Dom}_2}}(\boldsymbol{\delta}^{(s')}) \right\|_2^2 \right] \\
&\leq \frac{2}{R} \cdot \frac{1}{s+1} \left[\mathsf{NegLog}\left(\boldsymbol{\mu}^{(0)}, \boldsymbol{\delta}^{(0)}\right) - \mathsf{NegLog}^* \right].
\end{aligned}$$

■

Appendix F: Tree structure used in SUSHI Dataset study



The figure above depicts the tree structure employed in our real-world study on the SUSHI Preference Dataset; since there are 100 products (= the number of sushi varieties), we only show the non-leaf nodes. At the root node, the customer either decides to purchase a sushi variety or leave without a purchase. If the customer decides to make a purchase, she first chooses the *style*

of the sushi type: maki or not maki. Then, for each style, she decides to purchase a sushi variety based on whether or not it contains *seafood*. If the sushi contains seafood, she further chooses from nine minor groups: aomono (blue-skinned fish), akami (red meat fish), shiromi (white-meat fish), tare (something like baste; for eel or sea eel), clam or shell, squid or octopus, shrimp or crab, roe, and other seafood. Otherwise, she chooses from three minor groups: egg, meat other than fish, and vegetables. Lastly, she determines whether to purchase a sushi type with heavy/oily or light taste.

Appendix G: Additional Performance Measure

Here, we compare the PGD, Knitro and A-MM benchmarks on the root mean square error (RMSE) metric. For the simulation study, let $(\mathcal{S}_i^m : i = 1, 2, \ldots, 700)$ denote the sampled offer-sets for instance m. Then, we compute the average RMSE value across all the 100 instances for each algo $\in \{\text{A-MM}, \texttt{Knitro}, \text{PGD}\}$ as follows:

$$\text{RMSE}^{\text{algo}} = \frac{1}{100}\sum_{m=1}^{100}\sqrt{\frac{1}{60}\sum_{i=1}^{60}\frac{1}{|\mathcal{S}_i^m|}\sum_{\ell\in\mathcal{S}_i^m}\left(\mathbb{P}_\ell\left(\mathcal{S}_i^m;\boldsymbol{\mu}^{\text{true},m},\boldsymbol{\lambda}^{\text{true},m}\right)-\mathbb{P}_\ell\left(\mathcal{S}_i^m;\boldsymbol{\mu}^{\text{algo},m},\boldsymbol{\lambda}^{\text{algo},m}\right)\right)^2},$$

where for each instance m, $\mathbb{P}_\ell\left(\mathcal{S}_i^m;\boldsymbol{\mu}^{\text{true},m},\boldsymbol{\lambda}^{\text{true},m}\right)$ is the ground-truth purchase probability, and $\mathbb{P}_\ell\left(\mathcal{S}_i^m;\boldsymbol{\mu}^{\text{algo},m},\boldsymbol{\lambda}^{\text{algo},m}\right)$ is the estimated purchase probability by method algo, of product ℓ in offer-set $\mathcal{S}_i^m$. We report the RMSE values as well as the percentage improvements $100\times\left(\text{RMSE}^{\text{algo}}-\text{RMSE}^{\text{A-MM}}\right)/\text{RMSE}^{\text{algo}}$ for algo $\in \{\text{PGD}, \texttt{Knitro}\}$ in Figure EC.1. Similar to the NegLogGap values in the main text, the A-MM method achieves significantly lower RMSE than the benchmarks for all ground-truth problem sizes except the two smallest ones.

For the Sushi dataset study, let $(\mathcal{S}_i^m : i = 1, 2, \ldots, 700)$ denote the training offer-sets for instance m, and $(\mathcal{S}_i^m : i = 701, 702, \ldots, 1000)$ denote the test offer-sets. Then, we compute the average RMSE value on the training and test offer-sets for each method algo $\in \{\text{A-MM}, \text{PGD}, \texttt{Knitro}\}$ as follows:

$$\text{RMSE}^{\text{algo}}_{\text{Train}} = \frac{1}{400}\sum_{m=1}^{400}\sqrt{\frac{1}{700}\sum_{i=1}^{700}\frac{1}{|\mathcal{S}_i^m|}\sum_{\ell\in\mathcal{S}_i^m}\left(\text{sales}_\ell(\mathcal{S}_i^m)-\mathbb{P}_\ell\left(\mathcal{S}_i^m;\boldsymbol{\mu}^{\text{algo},m},\boldsymbol{\lambda}^{\text{algo},m}\right)\right)^2}$$

$$\text{RMSE}^{\text{algo}}_{\text{Test}} = \frac{1}{400}\sum_{m=1}^{400}\sqrt{\frac{1}{300}\sum_{i=701}^{1000}\frac{1}{|\mathcal{S}_i^m|}\sum_{\ell\in\mathcal{S}_i^m}\left(\text{sales}_\ell(\mathcal{S}_i^m)-\mathbb{P}_\ell\left(\mathcal{S}_i^m;\boldsymbol{\mu}^{\text{algo},m},\boldsymbol{\lambda}^{\text{algo},m}\right)\right)^2}$$

where $\text{sales}_\ell(\mathcal{S}_i^m)$ is the fraction of observed sales for product ℓ in offer set $\mathcal{S}_i^m$, and $\mathbb{P}_\ell\left(\mathcal{S}_i^m;\boldsymbol{\mu}^{\text{algo},m},\boldsymbol{\lambda}^{\text{algo},m}\right)$ is the probability that a customer purchases product ℓ from the offer set $\mathcal{S}_i^m$ under the parameters estimated by algo in instance m. We report the RMSE numbers as well as the percentage improvements $100\times\left(\text{RMSE}^{\text{algo}}_{\text{Train}}-\text{RMSE}^{\text{A-MM}}_{\text{Train}}\right)/\text{RMSE}^{\text{algo}}_{\text{Train}}$ and $100\times\left(\text{RMSE}^{\text{algo}}_{\text{Test}}-\text{RMSE}^{\text{A-MM}}_{\text{Test}}\right)/\text{RMSE}^{\text{algo}}_{\text{Test}}$ for algo $\in \{\text{PGD}, \texttt{Knitro}\}$ in Figure EC.2. Similar to the NegLogImp values in the main text, the A-MM method achieves lower RMSE values than the benchmarks on both training and test data, and is robust to initialization.

Degree	Height	# Prods.	# Nodes	λ_{lower}	RMSE ($\times 10^{-5}$)			RMSE Impr.		% better	
					A-MM	PGD	Knitro	over PGD	over Knitro	over PGD	over Knitro
5	4	625	781	0.50	5.02	6.52	7.21	23.0%	30.4%	72	91
				0.10	8.60	7.73	9.32	-11.2%	7.8%	40	74
				0.01	9.75	9.49	9.89	-2.8%	1.4%	50	68
	5	3,125	3,906	0.50	1.18	2.40	2.12	50.6%	44.1%	100	99
				0.10	2.15	2.87	3.45	25.1%	37.5%	95	100
				0.01	2.68	3.53	3.95	24.1%	32.2%	100	100
6	4	1,296	1,555	0.50	2.38	5.54	5.89	56.9%	59.5%	100	100
				0.10	4.32	6.45	6.95	33.0%	37.8%	90	100
				0.01	5.04	8.15	7.45	38.2%	32.4%	100	100
	5	7,776	9,331	0.50	0.46	1.57	3.19	70.8%	85.6%	100	100
				0.10	0.87	1.84	6.01	52.5%	85.5%	100	100
				0.01	1.09	2.32	6.57	53.2%	83.5%	100	100
7	4	2,401	2,801	0.50	1.26	4.42	2.99	71.6%	58.0%	100	100
				0.10	2.36	5.62	6.22	58.1%	62.1%	90	100
				0.01	2.80	6.39	6.56	56.2%	57.4%	100	100
	5	16,807	19,608	0.50	0.20	0.99	—	79.8%	—	100	—
				0.10	0.40	1.11	—	63.8%	—	100	—
				0.01	0.50	1.38	—	63.7%	—	100	—
8	4	4,096	4,681	0.50	0.85	3.59	—	76.3%	—	100	—
				0.10	1.37	4.28	—	68.0%	—	100	—
				0.01	1.63	5.07	—	67.9%	—	100	—
	5	32,768	37,449	0.50	0.10	0.62	—	83.9%	—	100	—
				0.10	0.17	0.78	—	78.2%	—	100	—
				0.01	0.23	0.91	—	74.7%	—	100	—

Figure EC.1 Comparison of the performances of PGD, `Knitro`, and our proposed A-MM method in fitting tree logit models to choice data. The columns "Degree", "Height", and λ_{lower} report the degree of each non-leaf node, the height of the tree, and the lower bound on the nest dissimilarity parameters, respectively. The columns "# Prods." and "# Nodes" report the number of products and the number of nodes in the tree, respectively. The columns under A-MM, PGD and `Knitro` report the average RMSE for each method. The columns under "RMSE Impr." reports the percentage improvement in the average RMSE value that our A-MM method achieves over the benchmarks. Finally, the columns under "% better" reports the percentage of instances in which the A-MM method obtains a lower RMSE value than the corresponding benchmark. The `Knitro` benchmark is unable to complete even a single iteration for large problem sizes, and we use "—" to denote such instances. All numbers under the "RMSE Impr." columns are significantly different from zero at the 1% significance level under a paired samples t-test.

Initialization	Train RMSE ($\times 10^{-4}$)			Test RMSE ($\times 10^{-4}$)			Impr. over PGD		Impr. over Knitro	
	A-MM	Knitro	PGD	A-MM	Knitro	PGD	Train	Test	Train	Test
0/1 start	13.64	13.72	14.35	13.69	13.78	14.40	5.23%	5.20%	0.59%	0.61%
warm start	13.64	13.67	13.79	13.69	13.73	13.85	1.16%	1.15%	0.27%	0.27%

Figure EC.2 Comparison of the performances of PGD, Knitro and our proposed A-MM method in fitting tree logit models to the Sushi Preference Dataset. The first (resp. second) row reports the performance when the methods are initialized using *0/1 start* (resp. *warm start*). The second, third and fourth columns report the RMSE value of A-MM, Knitro and PGD on the training data, while the fifth, sixth and seventh columns report the corresponding RMSE values on the test data. The eight and tenth columns report the percentage improvement in the average RMSE of the A-MM method over the PGD and Knitro benchmarks, respectively, on the training data. The corresponding improvements on the test data are reported in columns nine and eleven. All numbers under the "Impr." columns are significantly different from zero at the 1% significance level under a paired sample t-test.

Appendix H: Performance of Alternating GD method

As mentioned in the main text, we evaluate the performance of our method, A-MM, against the alternating gradient descent (A-GD) method to tease apart the effect of the MM approach from effect the variable transformation procedure on the overall performance. Starting from an initial solution $\left(\boldsymbol{\mu}^{(0)}_{\text{A-GD}}, \boldsymbol{\delta}^{(0)}_{\text{A-GD}}\right)$, the A-GD method performs the following two updates in each iteration $s \geq 0$:

$$\text{(i)} \quad \tilde{\boldsymbol{\mu}}^{(s+1)}_{\text{A-GD}} = \boldsymbol{\mu}^{(s)}_{\text{A-GD}} - \beta^{(s)} \cdot \nabla_{\boldsymbol{\mu}} \mathsf{NegLog}(\boldsymbol{\mu}^{(s)}_{\text{A-GD}}, \boldsymbol{\delta}^{(s)}_{\text{A-GD}}) \quad \text{for some step size } \beta^{(s)} > 0;$$
$$\left(\boldsymbol{\mu}^{(s+1)}_{\text{A-GD}}\right)_\ell = \left(\tilde{\mu}^{(s+1)}_{\text{A-GD}}\right)_\ell - \left(\tilde{\mu}^{(s+1)}_{\text{A-GD}}\right)_1 \quad \forall\ \ell \in \mathcal{N}$$
$$\text{(ii)} \quad \boldsymbol{\delta}^{(s+1)}_{\text{A-GD}} = \left\{\boldsymbol{\delta}^{(s)}_{\text{A-GD}} - \gamma^{(s)} \cdot \nabla_{\boldsymbol{\delta}} \mathsf{NegLog}(\boldsymbol{\mu}^{(s+1)}_{\text{A-GD}}, \boldsymbol{\delta}^{(s)}_{\text{A-GD}})\right\}^+ \quad \text{for some step size } \gamma^{(s)} > 0$$

We use backtracking linesearch to choose the step sizes in each iteration and run 300 iterations of the A-GD method starting from the initial solution $\boldsymbol{\mu}^{(0)}_{\text{A-GD}} = \mathbf{0}$, $\boldsymbol{\delta}^{(0)}_{\text{A-GD}} = \mathbf{0}$, which resulted in comparable runtime to the A-MM method (avg. 6-hr runtime per instance for A-GD compared to 5.9-hr for A-MM for instances with degree 8 and height 5).

The NegLogGap values for both methods are presented in a table shown in Figure EC.3. While the magnitude of improvement is lower compared to the Knitro and the PGD benchmarks in Section 4, our proposed A-MM method still outperforms A-GD across all problem sizes, obtaining significantly lower negative log-likelihood values for the (harder) problem instances with smaller values of the nest dissimilarity parameters.

To further understand the source of these improvements, in Figure EC.4, we plot the NegLogGap values (on a log-scale) for both methods as a function of the iteration number on a sample problem

Degree	Height	# Prods.	# Nodes	λ_{lower}	NegLogGap		NegLog	% better
					A-MM	A-GD	Impr.	
5	4	625	781	0.50	2.6	2.7	0.1	64
				0.10	7.8	15.5	7.7	100
				0.01	9.9	27.2	17.3	100
	5	3,125	3,906	0.50	3.6	4.5	0.9	100
				0.10	13.2	24.3	11.1	100
				0.01	21.6	49.0	27.4	100
6	4	1,296	1,555	0.50	2.5	2.7	0.2	70
				0.10	8.5	16.1	7.6	100
				0.01	11.1	30.8	19.7	100
	5	7,776	9,331	0.50	3.3	3.8	0.5	100
				0.10	12.8	19.6	6.8	100
				0.01	20.9	43.0	22.1	100
7	4	2,401	2,801	0.50	2.4	2.9	0.5	99
				0.10	8.6	14.7	6.1	100
				0.01	11.4	29.6	18.2	100
	5	16,807	19,608	0.50	2.9	3.3	0.4	100
				0.10	12.1	16.7	4.6	100
				0.01	19.0	37.8	18.8	100
8	4	4,096	4,681	0.50	2.2	2.6	0.4	100
				0.10	8.0	13.0	5.0	100
				0.01	10.5	28.3	17.8	100
	5	32,768	37,449	0.50	2.5	2.8	0.3	100
				0.10	10.7	14.6	3.9	100
				0.01	17.2	34.2	17.0	100

Figure EC.3 Comparison of the performances of the A-GD benchmark and our proposed A-MM method in fitting tree logit models to choice data. The columns "Degree", "Height", and λ_{lower} report the degree of each non-leaf node, the height of the tree, and the lower bound on the nest dissimilarity parameters, respectively. The columns "# Prods." and "# Nodes" report the number of products and the number of nodes in the tree, respectively. The columns under A-MM and A-GD report the average NegLogGap for each method. Recall that smaller values for the gaps are preferred. The "NegLog Impr." column reports the average improvement in the negative log-likelihood value that our A-MM method achieves over the A-GD benchmark, and the "% better" column reports the percentage of instances in which the A-MM method obtains a lower NegLog value than A-GD. All the "NegLog Impr." numbers are significantly different from zero at the 1% significance level under a paired samples t-test.

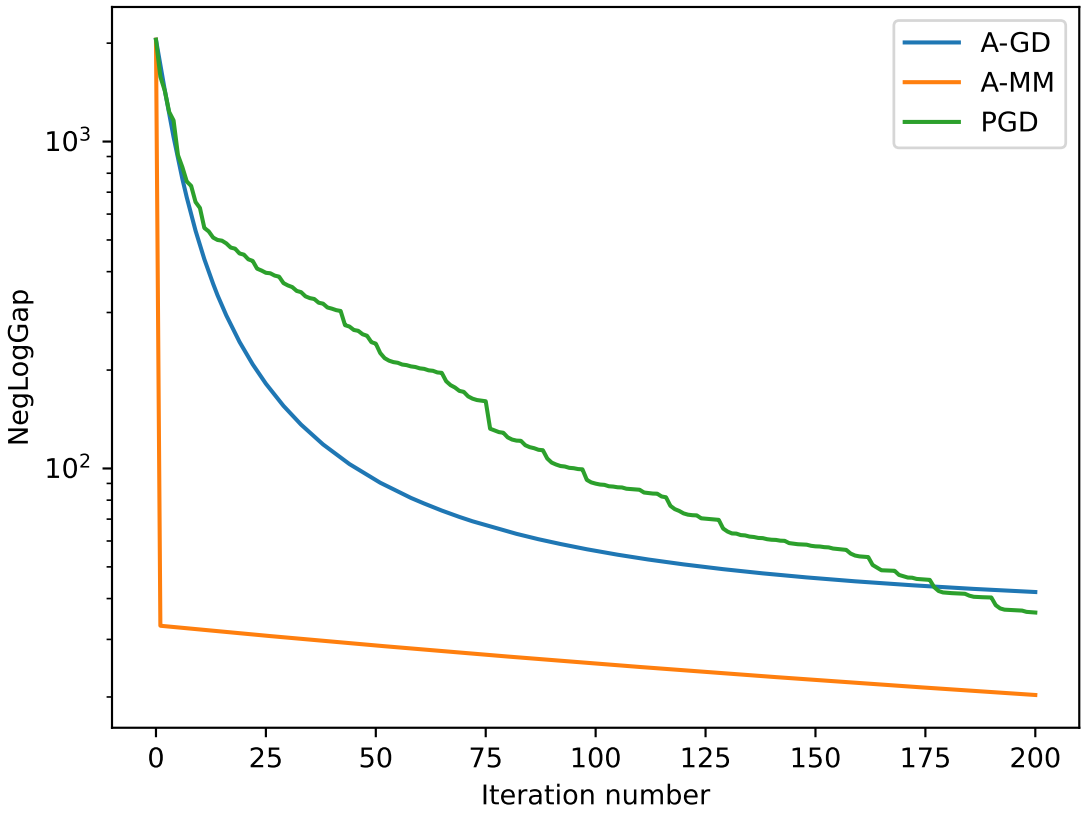


Figure EC.4 NegLogGap as a function of the iteration number for our proposed A-MM algorithm as well as the A-GD and PGD benchmarks on a sample problem instance under ground-truth parameters degree = 8, height = 5, and $\lambda_{\text{lower}} = 0.01$. Note the y-axis is on a log-scale.

instance under ground-truth parameters $r = 8$ (degree), $H = 5$ (height), and $\lambda_{\text{lower}} = 0.01$. The NegLogGap values achieved by the PGD method are also shown for comparison. It can be seen that the A-MM method consistently achieves a lower NegLogGap compared to both benchmarks. In particular, our proposed method makes rapid progress during the first few iterations, which is in line with the theoretical results. In contrast, the two benchmarks make more gradual progress leading to a larger NegLogGap for the same number of iterations. Moreover, an interesting thing to note is that the A-GD method converges faster than PGD, although PGD eventually achieves a lower NegLogGap value for this instance. This could be because PGD utilizes the "full" gradient of the negative log-likelihood, whereas A-GD alternates between gradient descent steps for the mean utility and nest dissimilarity parameters. Future work can investigate this difference in more detail.

Appendix I: Code and Data

The data and source code used in all the numerical studies in the paper can be downloaded at `https://github.com/ashwin90/TreeLogitEstimation`.

References

Allen-Zhu, Zeyuan, Lorenzo Orecchia. 2014. Linear coupling: An ultimate unification of gradient and mirror descent. *arXiv preprint arXiv:1407.1537* .

Boyd, S., L. Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press, Cambridge, UK.

Bubeck, S. 2015. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning* **8**(3-4) 231–357.

Cardell, N. S. 1997. Variance components structures for the extreme-value and logistic distributions with applications to models of heterogeneity. *Economic Theory* **13**(2) 185–213.

Csiszár, Imre, János Körner. 2011. *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press.

Gumbel, E. J. 2004. *Statistics of Extremes*. Dover Publications, Meneola, NY.

Kiefer, J. 1953. Sequential minimax search for a maximum. *Proceedings of the American Mathematical Society* **4**(3) 502–506.

McFadden, D. 1978. Modeling the choice of residential location. *Transportation Research Record* (672) 72–77.

Nesterov, Y. 2013. *Introductory Lectures on Convex Optimization: A Basic Course*, vol. 87. Springer Science & Business Media.

Vojnovic, Milan, Se-Young Yun, Kaifang Zhou. 2020. Convergence rates of gradient descent and mm algorithms for bradley-terry models. *International Conference on Artificial Intelligence and Statistics*. PMLR, 1254–1264.

Zhang, Xiao-Dong. 2011. The laplacian eigenvalues of graphs: a survey. *arXiv preprint arXiv:1111.2897* .