Online Appendix to "Flexible Data Aggregation for Prediction and Decision Making with Contextual Information: Applications in Retailing"

Zhenkang Peng, Chengzhang Li, Ying Rong, Zichao Luo, Guangrui Ma, Mingyong Zhao.

Appendix

A. Proofs of formal results

Proof of Proposition 1. When $s = K-1$, then regardless of α, we have, $\text{MSE}(\alpha) = \text{MSE}(0) = \text{MSE}(1)$ holds.

For $0 \le s < K-1$, the out-of-sample cost $\text{MSE}(\alpha)$ can be decomposed as:

$$\begin{aligned}
\text{MSE}(\alpha) &= \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{\mathcal{S}_k}\left[\mathbb{E}_{\epsilon_k}\left[\left(\alpha\hat{\mu}_k + (1-\alpha)\hat{T}_0(k) - \mu_k - \epsilon_k\right)^2\right]\right] \\
&= \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{\mathcal{S}_k}\left[\left(\alpha\hat{\mu}_k + (1-\alpha)\hat{T}_0(k) - \mu_k\right)^2 + \sigma_k^2\right] \\
&= \frac{1}{K}\mathbb{E}_{\mathcal{S}_k}\left[\sum_{k\in\mathbb{S}_s}(\hat{\mu}_k-\mu_k)^2 + \sum_{k\in\mathbb{S}_{K-s}}\left(\alpha\hat{\mu}_k + (1-\alpha)\hat{\mu}_p - \mu_k\right)^2\right] + \frac{1}{K}\sum_{k=1}^{K}\sigma_k^2 \\
&= \frac{1}{K}\mathbb{E}_{\mathbb{S}_s}\left[\mathbb{E}_{\mathcal{S}_k}\left[\sum_{k\in\mathbb{S}_s}(\hat{\mu}_k-\mu_k)^2 + \sum_{k\in\mathbb{S}_{K-s}}\left(\alpha\hat{\mu}_k + (1-\alpha)\hat{\mu}_p - \mu_k\right)^2\Big|\mathbb{S}_s\right]\right] + \frac{1}{K}\sum_{k=1}^{K}\sigma_k^2 \\
&= \frac{1}{K}\mathbb{E}_{\mathbb{S}_s}\left[\sum_{k\in\mathbb{S}_s}\frac{\sigma_k^2}{N} + \alpha^2\sum_{k\in\mathbb{S}_{K-s}}\frac{\sigma_k^2}{N} + (1-\alpha)^2\mathbb{E}_{\mathcal{S}_k}\left[\sum_{k\in\mathbb{S}_{K-s}}(\hat{\mu}_p-\mu_k)^2\right]\right] + \frac{1}{K}\sum_{k=1}^{K}\sigma_k^2 \\
&= \frac{1}{K}\mathbb{E}_{\mathbb{S}_s}\left[\sum_{k\in\mathbb{S}_s}\frac{\sigma_k^2}{N} + \alpha^2\sum_{k\in\mathbb{S}_{K-s}}\frac{\sigma_k^2}{N} + (1-\alpha)^2\sum_{k\in\mathbb{S}_{K-s}}\left(\frac{\sigma_k^2}{N(K-s)} + (\mu_k-\mu_p)^2\right)\right] + \frac{1}{K}\sum_{k=1}^{K}\sigma_k^2.
\end{aligned}$$

One can easily verify that $\text{MSE}(\alpha)$ is a convex function with respect to α. By the first-order condition, we obtain the optimal α^*, which is given by

$$\alpha^* = \frac{\mathbb{E}_{\mathbb{S}_s}\left[\sum_{k\in\mathbb{S}_{K-s}}\left(\frac{\sigma_k^2}{N(K-s)} + (\mu_k-\mu_p)^2\right)\right]}{\mathbb{E}_{\mathbb{S}_s}\left[\sum_{k\in\mathbb{S}_{K-s}}\frac{\sigma_k^2}{N} + \sum_{k\in\mathbb{S}_{K-s}}\left(\frac{\sigma_k^2}{N(K-s)} + (\mu_k-\mu_p)^2\right)\right]}. \tag{EC.1}$$

By the optimality of $\alpha^* \in (0,1)$, we have

$$\text{MSE}(0) - \text{MSE}(\alpha^*) \ge 0, \quad \text{MSE}(1) - \text{MSE}(\alpha^*) \ge 0.$$

Equivalently, we have,

$$\text{MSE}(\alpha^*) \le \min\{\text{MSE}(0), \text{MSE}(1)\}. \tag{EC.2}$$

The completes the proof. □

Proof of Lemma 1. We can compute the shared OLS estimator $\hat{\boldsymbol{\beta}}_0$ as follows:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_0 &= \left(\hat{X}^\top\hat{X}\right)^{-1}\hat{X}^\top\hat{D} \\
&= \Big(\sum_{k\in[K]}\hat{X}_k^\top\hat{X}_k\Big)^{-1}\cdot\Big(\sum_{k\in[K]}\hat{X}_k^\top\hat{D}_k\Big) \\
&= \Big(\sum_{k\in[K]}\hat{X}_k^\top\hat{X}_k\Big)^{-1}\cdot\sum_{k\in[K]}\left(\hat{X}_k^\top\hat{X}_k\boldsymbol{\beta}_k + \hat{X}_k^\top\hat{X}_k\hat{\boldsymbol{\epsilon}}_k\right) \\
&= \Big(\frac{1}{K}\sum_{k\in[K]}\hat{X}_k^\top\hat{X}_k\Big)^{-1}\left(\frac{1}{K}\sum_{k\in[K]}\hat{X}_k^\top\hat{X}_k\boldsymbol{\beta}_k\right) + \Big(\frac{1}{K}\sum_{k\in[K]}\hat{X}_k^\top\hat{X}_k\Big)^{-1}\left(\frac{1}{K}\sum_{k\in[K]}\hat{X}_k^\top\hat{X}_k\hat{\boldsymbol{\epsilon}}_k\right).
\end{aligned}$$

For the term, $\left(\frac{1}{K}\sum_{k\in[K]}\hat{X}_k^\top\hat{X}_k\right)^{-1}\left(\frac{1}{K}\sum_{k\in[K]}\hat{X}_k^\top\hat{X}_k\boldsymbol{\beta}_k\right)$, we denote $\hat{X}_k^\top\hat{X}_k = [a_{k,1}^\top,\cdots,a_{k,F}^\top]$, then $\hat{X}_k^\top\hat{X}_k\boldsymbol{\beta}_k = [a_{k,1}^\top\boldsymbol{\beta}_k,\cdots,a_{k,F}^\top\boldsymbol{\beta}_k]$, and

$$\frac{1}{K}\sum_{k\in[K]}\hat{X}_k^\top\hat{X}_k\boldsymbol{\beta}_k = \big[\frac{1}{K}\sum_{k\in[K]}a_{k,1}^\top\boldsymbol{\beta}_k,\cdots,\frac{1}{K}\sum_{k\in[K]}a_{k,F}^\top\boldsymbol{\beta}_k\big].$$

For any component $i\in[F]$, we have that $a_{k,i}^\top\boldsymbol{\beta}_k = \sum_{j\in[F]}a_{k,i,j}\beta_{k,j}$. Based on Assumption 2 and $|\beta_{k,j}|\le\beta_{\max}$, we can deduce that $a_{k,i,j}\beta_{k,j}$ follows sub-Gaussian distribution with parameter at most $\sqrt{N_k x_{\max}^2\beta_{\max}^2}$. Thus, we have,

$$\mathbb{P}\Big(\Big|\frac{1}{K}\sum_{k\in[K]}a_{k,i,j}(\beta_{k,j}-\beta_{0,j})\Big|>t\Big)\le 2\exp\Big(-\frac{2Kt^2}{N_{\max}x_{\max}^2\beta_{\max}^2}\Big),$$

where $N_{\max}=\max_{k\in[K]}N_k$. Then, we can conclude that, as $K\to\infty$, $\frac{1}{K}\sum_{k\in[K]}a_{k,i,j}(\beta_{k,j}-\beta_{0,j})\to_p 0$. By Slutsky's Theorem, we can obtain that $\frac{1}{K}\sum_{k\in[K]}a_{k,i}^\top(\boldsymbol{\beta}_k-\boldsymbol{\beta}_0)\to_p 0$, that is, $\frac{1}{K}\sum_{k\in[K]}a_{k,i}^\top\boldsymbol{\beta}_k\to_p\frac{1}{K}\sum_{k\in[K]}a_{k,i}^\top\boldsymbol{\beta}_0$.

Thus, we can know that,

$$\frac{1}{K}\sum_{k\in[K]}\hat{X}_k^\top\hat{X}_k\boldsymbol{\beta}_k = \Big[\frac{1}{K}\sum_{k\in[K]}a_{k,1}^\top\boldsymbol{\beta}_k,\cdots,\frac{1}{K}\sum_{k\in[K]}a_{k,F}^\top\boldsymbol{\beta}_k\Big]\to_p\Big[\frac{1}{K}\sum_{k\in[K]}a_{k,1}^\top\boldsymbol{\beta}_0,\cdots,\frac{1}{K}\sum_{k\in[K]}a_{k,F}^\top\boldsymbol{\beta}_0\Big]=\frac{1}{K}\sum_{k\in[K]}\hat{X}_k^\top\hat{X}_k\boldsymbol{\beta}_0.$$

Thus, we have,

$$\Big(\frac{1}{K}\sum_{k\in[K]}\hat{X}_k^\top\hat{X}_k\Big)^{-1}\Big(\frac{1}{K}\sum_{k\in[K]}\hat{X}_k^\top\hat{X}_k\boldsymbol{\beta}_k\Big)\to_p\Big(\frac{1}{K}\sum_{k\in[K]}\hat{X}_k^\top\hat{X}_k\Big)^{-1}\Big(\frac{1}{K}\sum_{k\in[K]}\hat{X}_k^\top\hat{X}_k\boldsymbol{\beta}_0\Big)=\boldsymbol{\beta}_0.$$

For the term, $\big(\frac{1}{K}\sum_{k\in[K]}\hat{X}_k^\top\hat{X}_k\big)^{-1}\big(\frac{1}{K}\sum_{k\in[K]}\hat{X}_k^\top\hat{X}_k\hat{\boldsymbol{\epsilon}}_k\big)$, due to each element in $\hat{\boldsymbol{\epsilon}}_k$ follows a normal distribution with zero mean and variance of σ_k^2 , which is also sub-Gaussian distribution. Thus, we similarly show that, as $K\to\infty$,

$$\Big(\frac{1}{K}\sum_{k\in[K]}\hat{X}_k^\top\hat{X}_k\Big)^{-1}\Big(\frac{1}{K}\sum_{k\in[K]}\hat{X}_k^\top\hat{X}_k\hat{\boldsymbol{\epsilon}}_k\Big)\to_p 0.$$

Based on the above analysis, by Slutsky's Theorem, we can conclude that:

$$\hat{\boldsymbol{\beta}}_0\to_p\boldsymbol{\beta}_0. \tag{EC.3}$$

This completes the proof. □

Proof of Proposition 2. We take the derivative of $\mathcal{L}^{\text{FlexDA}}(\alpha;\boldsymbol{\beta}_0)$ with respect to α,

$$\frac{\partial\mathcal{L}^{\text{FlexDA}}(\alpha;\boldsymbol{\beta}_0)}{\partial\alpha}=2\alpha\frac{1}{K}\sum_{k\in[K]}\boldsymbol{x}_k^\top\Big[\mathbb{E}_{S_k}\Big[(\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k)(\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k)^\top\Big]+(\boldsymbol{\beta}_0-\boldsymbol{\beta}_k)(\boldsymbol{\beta}_0-\boldsymbol{\beta}_k)^\top\Big]\boldsymbol{x}_k-2\frac{1}{K}\sum_{k\in[K]}\boldsymbol{x}_k^\top(\boldsymbol{\beta}_0-\boldsymbol{\beta}_k)(\boldsymbol{\beta}_0-\boldsymbol{\beta}_k)^\top\boldsymbol{x}_k.$$

With the first-order condition, the optimal parameter can be computed as:

$$\alpha^*=\frac{\sum_{k\in[K]}\boldsymbol{x}_k^\top(\boldsymbol{\beta}_0-\boldsymbol{\beta}_k)(\boldsymbol{\beta}_0-\boldsymbol{\beta}_k)^\top\boldsymbol{x}_k}{\sum_{k\in[K]}\boldsymbol{x}_k^\top\Big[\mathbb{E}_{S_k}\Big[(\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k)(\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k)^\top\Big]+(\boldsymbol{\beta}_0-\boldsymbol{\beta}_k)(\boldsymbol{\beta}_0-\boldsymbol{\beta}_k)^\top\Big]\boldsymbol{x}_k}<1. \tag{EC.4}$$

We plug the optimal parameter into the cost function associated with the FlexDA estimator, and we have,

$$\mathcal{L}^{\text{FlexDA}}(\alpha^*;\boldsymbol{\beta}_0)=\alpha^*\cdot\frac{1}{K}\sum_{k\in[K]}\boldsymbol{x}_k^\top\mathbb{E}_{S_k}\Big[(\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k)(\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k)^\top\Big]\boldsymbol{x}_k+\frac{1}{K}\sum_{k\in[K]}\sigma_k^2.$$

Thus, we obtain,

$$\mathcal{L}(\hat{\boldsymbol{\beta}}^{\text{ols}})-\mathcal{L}^{\text{FlexDA}}(\alpha^*;\boldsymbol{\beta}_0)=\frac{1-\alpha^*}{K}\sum_{k\in[K]}\boldsymbol{x}_k^\top\sigma_k^2(\hat{X}_k^\top\hat{X}_k)^{-1}\boldsymbol{x}_k,$$
$$\mathcal{L}(\boldsymbol{\beta}_0)-\mathcal{L}^{\text{FlexDA}}(\alpha^*;\boldsymbol{\beta}_0)=\frac{\alpha^*}{K}\sum_{k\in[K]}\boldsymbol{x}_k^\top(\boldsymbol{\beta}_0-\boldsymbol{\beta}_k)(\boldsymbol{\beta}_0-\boldsymbol{\beta}_k)^\top\boldsymbol{x}_k.$$

This completes the proof. □

Proof of Proposition 3. To prove $\hat{\alpha} \to_p \alpha^*$, we first show that:

$$\begin{aligned}
&\left|\frac{1}{K}\sum_{k\in[K]}\left[(\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k-\hat{\boldsymbol{\beta}}_0))^2-\mathbb{E}\left[(\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_0))^2\right]\right]\right| \\
=&\left|\frac{1}{K}\sum_{k\in[K]}\left[(\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k-\hat{\boldsymbol{\beta}}_0))^2-(\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_0))^2+(\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_0))^2-\mathbb{E}\left[(\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_0))^2\right]\right]\right| \\
\le&\underbrace{\left|\frac{1}{K}\sum_{k\in[K]}\left[(\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k-\hat{\boldsymbol{\beta}}_0))^2-(\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_0))^2\right]\right|}_{\text{A}}+\underbrace{\left|\frac{1}{K}\sum_{k\in[K]}\left[(\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_0))^2-\mathbb{E}\left[(\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_0))^2\right]\right]\right|}_{\text{B}}.
\end{aligned}$$

For part A,

$$\begin{aligned}
\left|\frac{1}{K}\sum_{k\in[K]}\left[(\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k-\hat{\boldsymbol{\beta}}_0))^2-(\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_0))^2\right]\right| =&\left|\frac{1}{K}\sum_{k\in[K]}\left[(\boldsymbol{\beta}_0-\hat{\boldsymbol{\beta}}_0)^\top\boldsymbol{x}_k\boldsymbol{x}_k^\top(2\hat{\boldsymbol{\beta}}_k-\hat{\boldsymbol{\beta}}_0-\boldsymbol{\beta}_0)\right]\right| \\
=&\left|(\boldsymbol{\beta}_0-\hat{\boldsymbol{\beta}}_0)^\top\frac{1}{K}\sum_{k\in[K]}\left[2\boldsymbol{x}_k\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k)+\boldsymbol{x}_k\boldsymbol{x}_k^\top(2\boldsymbol{\beta}_k-\hat{\boldsymbol{\beta}}_0-\boldsymbol{\beta}_0)\right]\right| \\
\le&\|\boldsymbol{\beta}_0-\hat{\boldsymbol{\beta}}_0\|\cdot\left\|\frac{1}{K}\sum_{k\in[K]}\left[2\boldsymbol{x}_k\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k)+\boldsymbol{x}_k\boldsymbol{x}_k^\top(2\boldsymbol{\beta}_k-\hat{\boldsymbol{\beta}}_0-\boldsymbol{\beta}_0)\right]\right\| \\
\le&\|\boldsymbol{\beta}_0-\hat{\boldsymbol{\beta}}_0\|\cdot\left(\left\|\frac{1}{K}\sum_{k\in[K]}2\boldsymbol{x}_k\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k)\right\|+\left\|\frac{1}{K}\sum_{k\in[K]}\boldsymbol{x}_k\boldsymbol{x}_k^\top(2\boldsymbol{\beta}_k-\hat{\boldsymbol{\beta}}_0-\boldsymbol{\beta}_0)\right\|\right).
\end{aligned}$$

The first inequality holds by the Cauchy-Schwarz inequality, and the second by the triangle inequality of norms.

Firstly, we know that when $K\to\infty$, $\hat{\boldsymbol{\beta}}_0 \to_p \boldsymbol{\beta}_0$, which implies that $\|\boldsymbol{\beta}_0-\hat{\boldsymbol{\beta}}_0\| \to_p 0$. Thus, we next need to prove that $\|\frac{1}{K}\sum_{k\in[K]}2\boldsymbol{x}_k\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k)\|$ and $\|\frac{1}{K}\sum_{k\in[K]}\boldsymbol{x}_k\boldsymbol{x}_k^\top(2\boldsymbol{\beta}_k-\hat{\boldsymbol{\beta}}_0-\boldsymbol{\beta}_0)\|$ are bounded.

For the term, $\|\frac{1}{K}\sum_{k\in[K]}2\boldsymbol{x}_k\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k)\|$, we know that $\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k=(\hat{X}_k^\top\hat{X}_k)^{-1}\hat{\boldsymbol{\epsilon}}_k$. Thus, under Assumption 2, each element of $2\boldsymbol{x}_k\boldsymbol{x}_k^\top(\hat{X}_k^\top\hat{X}_k)^{-1}\hat{\boldsymbol{\epsilon}}_k$ follows a Gaussian distribution with a mean of 0 but with bounded variance. Thus, for each element of $\frac{1}{K}\sum_{k\in[K]}2\boldsymbol{x}_k\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k)$ will converge to zero as $K\to\infty$, which makes $\|\frac{1}{K}\sum_{k\in[K]}2\boldsymbol{x}_k\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k)\|\to_p 0$.

For the term, $\|\frac{1}{K}\sum_{k\in[K]}\boldsymbol{x}_k\boldsymbol{x}_k^\top(2\boldsymbol{\beta}_k-\hat{\boldsymbol{\beta}}_0-\boldsymbol{\beta}_0)\|$, we have, as $K\to\infty$, $\hat{\boldsymbol{\beta}}_0\to_p\boldsymbol{\beta}_0$ is bounded according to Assumption 2(c). Thus, combining with Assumption 2(b), we can conclude that $\|\frac{1}{K}\sum_{k\in[K]}\boldsymbol{x}_k\boldsymbol{x}_k^\top(2\boldsymbol{\beta}_k-\hat{\boldsymbol{\beta}}_0-\boldsymbol{\beta}_0)\|$ is bounded when $K\to\infty$. Thus, when $K\to\infty$, the value of part A will converge to zero in probability.

For part B,

$$\begin{aligned}
\left|\frac{1}{K}\sum_{k\in[K]}\left[(\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_0))^2-\mathbb{E}[(\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_0))^2]\right]\right| =&\left|\frac{1}{K}\sum_{k\in[K]}\left[(\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k))^2-\mathbb{E}[(\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k))^2]+2(\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k))(\boldsymbol{x}_k^\top(\boldsymbol{\beta}_k-\boldsymbol{\beta}_0))\right]\right| \\
\le&\left|\frac{1}{K}\sum_{k\in[K]}\left[(\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k))^2-\mathbb{E}[(\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k))^2]\right]\right|+\left|\frac{1}{K}\sum_{k\in[K]}2(\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k)^\top\boldsymbol{x}_k\boldsymbol{x}_k^\top(\boldsymbol{\beta}_k-\boldsymbol{\beta}_0)\right|.
\end{aligned}$$

We know that $\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k)$ follows Gaussian distribution with zero mean and bounded variance, thus $(\boldsymbol{x}_k^\top((\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k))^2$ follows sub-exponential distribution with some bounded parameter. Then, by the property of sub-exponential distribution, we have, as $K\to\infty$:

$$\frac{1}{K}\sum_{k\in[K]}\left[(\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k))^2-\mathbb{E}[(\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k))^2]\right]\to_p 0.$$

Similarly, we can know that $2(\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k)^\top\boldsymbol{x}_k\boldsymbol{x}_k^\top(\boldsymbol{\beta}_k-\boldsymbol{\beta}_0)$ follows Gaussian distribution with zero mean and bounded variance. Then, we have, as $K\to\infty$,

$$\frac{1}{K}\sum_{k\in[K]}2(\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k)^\top\boldsymbol{x}_k\boldsymbol{x}_k^\top(\boldsymbol{\beta}_k-\boldsymbol{\beta}_0)\to_p 0.$$

Combining parts A and B, we have, as $K\to\infty$,

$$\frac{1}{K}\sum_{k\in[K]}(\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k-\hat{\boldsymbol{\beta}}_0))^2\to_p\frac{1}{K}\sum_{k\in[K]}\mathbb{E}\left[(\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_0))^2\right]. \tag{EC.5}$$

In addition, s_k^2 is proportional to the chi-squared distribution (see Page 14 in Amemiya 1985),

$$s_k^2 \sim \frac{\sigma_k^2}{N_k - F} \cdot \mathcal{X}^2_{N_k - F}. \tag{EC.6}$$

Thus, $\mathbb{E}[s_k^2] = \sigma_k^2$ and s_k^2 follows sub-exponential distribution (see Example 2.4 in Wainwright 2019). Therefore, we have,

$$\mathbb{E}\Big[\frac{1}{K}\sum_{k\in[K]} s_k^2 \boldsymbol{x}_k^\top (\hat{X}_k^\top \hat{X}_k)^{-1}\boldsymbol{x}_k\Big] = \frac{1}{K}\sum_{k\in[K]} \sigma_k^2 \boldsymbol{x}_k^\top (\hat{X}_k^\top \hat{X}_k)^{-1}\boldsymbol{x}_k.$$

Since $s_k^2 \boldsymbol{x}_k^\top (\hat{X}_k^\top \hat{X}_k)^{-1}\boldsymbol{x}_k$ follows a sub-exponential distribution, as $K \to \infty$, we have,

$$\frac{1}{K}\sum_{k\in[K]} s_k^2 \boldsymbol{x}_k^\top (\hat{X}_k^\top \hat{X}_k)^{-1}\boldsymbol{x}_k \to_p \frac{1}{K}\sum_{k\in[K]} \sigma_k^2 \boldsymbol{x}_k^\top (\hat{X}_k^\top \hat{X}_k)^{-1}\boldsymbol{x}_k. \tag{EC.7}$$

Finally, based on the following decomposition,

$$\frac{1}{K}\sum_{k\in[K]} \mathbb{E}\big[(\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_0))^2\big] = \frac{1}{K}\sum_{k\in[K]} \mathbb{E}\big[(\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k] - \boldsymbol{\beta}_k))^2 + \frac{1}{K}\sum_{k\in[K]} \boldsymbol{x}_k^\top(\boldsymbol{\beta}_k - \boldsymbol{\beta}_0)^2,$$

by Slutsky's Theorem and Equations (EC.5) and (EC.7), we have,

$$\frac{1}{K}\sum_{k\in[K]} (\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k - \hat{\boldsymbol{\beta}}_0))^2 - \frac{1}{K}\sum_{k\in[K]} s_k^2 \boldsymbol{x}_k^\top (\hat{X}_k^\top \hat{X}_k)^{-1}\boldsymbol{x}_k \to_p \frac{1}{K}\sum_{k\in[K]} (\boldsymbol{x}_k^\top(\boldsymbol{\beta}_k - \boldsymbol{\beta}_0))^2,$$
$$\frac{1}{K}\sum_{k\in[K]} (\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k - \hat{\boldsymbol{\beta}}_0))^2 \to_p \frac{1}{K}\sum_{k\in[K]} \mathbb{E}\big[(\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k))^2\big] + \frac{1}{K}\sum_{k\in[K]} (\boldsymbol{x}_k^\top(\boldsymbol{\beta}_k - \boldsymbol{\beta}_0))^2.$$

Thus, by Slutsky's Theorem and definitions of $\hat{\alpha}$ and α^*, we have $\hat{\alpha} \to \alpha^*$. This completes the proof. □

Proof of Theorem 1. Given any historical dataset $\{\mathcal{S}_1, \mathcal{S}_2, \cdots, \mathcal{S}_k\}$, we define a difference term, denoted by Ψ, which is given by,

$$\Psi = \frac{1}{K}\sum_{k\in[K]} \Big(\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k^{\text{FlexDA}}(\hat{\alpha}, \hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_k) - \boldsymbol{\beta}_{\boldsymbol{k}})(\hat{\boldsymbol{\beta}}_k^{\text{FlexDA}}(\hat{\alpha}, \hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_k) - \boldsymbol{\beta}_{\boldsymbol{k}})^\top \boldsymbol{x}_k - \boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k^{\text{FlexDA}}(\alpha^*, \boldsymbol{\beta}_0, \hat{\boldsymbol{\beta}}_k) - \boldsymbol{\beta}_{\boldsymbol{k}})(\hat{\boldsymbol{\beta}}_k^{\text{FlexDA}}(\alpha^*, \boldsymbol{\beta}_0, \hat{\boldsymbol{\beta}}_k) - \boldsymbol{\beta}_k)^\top \boldsymbol{x}_k\Big).$$

By the definition of the FlexDA estimator, we can decompose Ψ as follows,

$$\begin{aligned}
|\Psi| &= \left|\frac{1}{K}\sum_{k\in[K]} \left(\Big(\hat{\alpha}\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k) + (1-\hat{\alpha})\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_k)\Big)^2 - \Big(\alpha^*\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k) + (1-\alpha^*)\boldsymbol{x}_k^\top(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_k)\Big)^2\right)\right| \\
&= \Bigg|\frac{1}{K}\sum_{k\in[K]} \Big((\hat{\alpha}^2 - (\alpha^*)^2)(\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k))^2 + 2\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k)\boldsymbol{x}_k^\top\Big(\hat{\alpha}(1-\hat{\alpha})(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_k) - \alpha^*(1-\alpha^*)(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_k)\Big) \\
&\qquad + (1-\hat{\alpha})^2(\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_k))^2 - (1-\alpha^*)^2(\boldsymbol{x}_k^\top(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_k))^2\Big)\Bigg| \\
&\le (\hat{\alpha}^2 - (\alpha^*)^2)\left|\frac{1}{K}\sum_{k\in[K]} (\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k))^2\right| + \left|\frac{1}{K}\sum_{k\in[K]} \Big((1-\hat{\alpha})^2(\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_k))^2 - (1-\alpha^*)^2(\boldsymbol{x}_k^\top(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_k))^2\Big)\right| \\
&\qquad + \left|\frac{1}{K}\sum_{k\in[K]} 2\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k)\boldsymbol{x}_k^\top\Big(\hat{\alpha}(1-\hat{\alpha})(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_k) - \alpha^*(1-\alpha^*)(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_k)\Big)\right| \\
&\le (\hat{\alpha}^2 - (\alpha^*)^2)\left|\frac{1}{K}\sum_{k\in[K]} (\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k))^2\right| + \left|\frac{1}{K}\sum_{k\in[K]} ((1-\hat{\alpha})^2(\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_k))^2 - (1-\alpha^*)^2(\boldsymbol{x}_k^\top(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_k))^2)\right| \\
&\qquad + \sqrt{\frac{1}{K}\sum_{k\in[K]} (2\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k))^2} \cdot \sqrt{\frac{1}{K}\sum_{k\in[K]} \Big(\boldsymbol{x}_k^\top(\hat{\alpha}(1-\hat{\alpha})(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_k) - \alpha^*(1-\alpha^*)(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_k))\Big)^2}.
\end{aligned}$$

The first inequality is due to the triangle inequality of norms, and the second inequality is due to the Cauchy-Schwarz inequality.

Similar to the proof of Proposition 3, we can show that $(\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k))^2$ follows a sub-exponential distribution with some bounded parameter, which makes $\left|\frac{1}{K}\sum_{k\in[K]} \Big((1-\hat{\alpha})^2(\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_k))^2 - (1-\alpha^*)^2(\boldsymbol{x}_k^\top(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_k))^2\Big)\right|$ converge to some bounded value as $K \to \infty$. We also have, $\hat{\alpha} \to_p \alpha^*$, $\hat{\boldsymbol{\beta}}_0 \to_p \boldsymbol{\beta}_0$. Thus, by Slutsky's Theorem and the continuous mapping theorem, we have, as $K \to \infty$,

$$\hat{\alpha}(1-\hat{\alpha})(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_k) - \alpha^*(1-\alpha^*)(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_k) \to_p 0,$$
$$(1-\hat{\alpha})^2(\boldsymbol{x}_k^\top(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_k))^2 - (1-\alpha^*)^2(\boldsymbol{x}_k^\top(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_k))^2 \to_p 0.$$

Thus, combining with all results, we have shown that, as $K \to \infty$, $\Psi \to_p 0$, which concludes the proof. □

The proof of Theorem 2 requires the following definitions and theorem EC.1 and lemma EC.1.

First, we introduce the definitions of the packing number and covering number (see Definitions 5.4 and 5.1 of Wainwright (2019)).

DEFINITION EC.1 (PACKING NUMBER). A δ-packing of a set $\mathbb{T}$ with respect to a metric ρ is a set $\theta^1, \cdots, \theta^M \subset \mathbb{T}$ such that $\rho(\theta^i, \theta^j) > \delta$ for all distinct $i, j \in \{1, 2, \cdots, M\}$. The δ-packing number $D(\delta, \mathbb{T}, \rho)$ is the cardinality of the largest δ-packing.

DEFINITION EC.2 (COVERING NUMBER). A δ-covering of a set $\mathbb{T}$ with respect to a metric ρ is a set $\theta^1, \cdots, \theta^M \subset \mathbb{T}$ such that $\rho(\theta^i, \theta^j) \le \delta$ for all distinct $i, j \in \{1, 2, \cdots, M\}$. The δ-covering number $N(\delta, \mathbb{T}, \rho)$ is the cardinality of the smallest δ-cover.

All subsequent analyses are based on the Euclidean metric, hence we omit the notation ρ for brevity.

THEOREM EC.1 (**A Maximal Inequality; Pollard (1990)**). *Let $W(t) = (W_1(t), \dots, W_K(t)) \in \mathbb{R}^K$ be a stochastic process indexed by $t \in \mathcal{T}$, and let $\overline{W}_K(t) = \frac{1}{K}\sum_{k=1}^K W_k(t)$. Let $\boldsymbol{V} \in \mathbb{R}^K$ be a random variable such that $|W_k(t)| \le V_k$ for all $t \in \mathcal{T}, k = 1, \dots, K$. Finally, define the random variable*

$$J \equiv J(\{W(t) : t \in \mathcal{T}\}, \boldsymbol{V}) \equiv 9\|\boldsymbol{V}\|_2 \int_0^1 \sqrt{\log D(\|\boldsymbol{V}\|_2 \mu, \{W(t) : t \in \mathcal{T}\})} du. \tag{EC.8}$$

Then, for any $p \ge 1$ and any $0 < \delta < 1$, with probability at least $1 - \delta$,

$$\sup_{t \in \mathcal{T}} |\overline{W}_K(t) - \mathbb{E}[\overline{W}_K(t)]| \le 5^{1/p}\sqrt{p}\|J\|_p K^{-1}\delta^{-1/p}.$$

LEMMA EC.1 (**Chen (1975)**). *For any $f : \mathbb{Z}_+ \to \mathbb{R}$, if W follows a Poisson distribution with parameter λ, then we have:*

$$\lambda\mathbb{E}[f(W+1)] = \mathbb{E}[Wf(W)].$$

We first define the FlexDA estimator and the corresponding out-of-sample cost under general cost functions and nonlinear models. Under Assumption 3 (ii), for any product k, there exist at most η possible realized values $\{(\boldsymbol{a}_k^i, b_k^i) : i \in [\eta]\}$, where $\boldsymbol{a}_k^i$ represents a possible realized value of feature vector $\boldsymbol{x}_k$ and b_k^i represents a possible realized value of random noise ϵ_k. We denote the corresponding probability distribution as $\boldsymbol{p}_k = [p_{k,1}, \dots, p_{k,\eta}]$. Notably, for a product k with fewer than η possible values, we can freely assign values of $\boldsymbol{a}_k^i$ and b_k^i and regulate their effects by setting the corresponding probability, $p_{k,i}$, to zero.

Thus, under the cost function $c_k(\cdot)$, the set of all possible realized costs for any decision function $g \in \mathcal{H}$ can be denoted as $\boldsymbol{c}_k(g) = \{c_{k,i} = c_k(g(\boldsymbol{a}_k^i), h_k(\boldsymbol{a}_k^i) + b_k^i) : i \in [\eta]\}$. In a data-driven setting, we denote $\hat{\boldsymbol{m}}_k = (\hat{m}_{k,1}, \cdots, \hat{m}_{k,\eta})$ as the count number for $\{(\boldsymbol{a}_k^i, b_k^i) : i \in [\eta]\}$ and $\hat{N}_k = \sum_{i=1}^{\eta} \hat{m}_{k,i}$. With Assumption 3(i), we have,

$$\hat{m}_{k,i} \sim \text{Poisson}(m_{k,i}), \quad \text{where } m_{k,i} = N \cdot p_{k,i}. \tag{EC.9}$$

Thus, given the above notations, the decoupled estimator, which is defined in Equation (19), can be written as follows,

$$\hat{h}_k(\cdot) \in \operatorname*{argmin}_{g \in \mathcal{H}} \hat{\boldsymbol{m}}_k^\top \boldsymbol{c}_k(g). \tag{EC.10}$$

The shared estimator which is defined in Equation (20) can be rewritten as follows,

$$\hat{h}_0(\cdot) \in \operatorname*{argmin}_{g \in \mathcal{H}} \sum_{k=1}^K \hat{\boldsymbol{m}}_k^\top \boldsymbol{c}_k(g). \tag{EC.11}$$

The FlexDA estimator improves the trade-off between the bias and variance with a parameter α as follows,

$$\hat{h}_k^{\text{FlexDA}}(\boldsymbol{x}; \hat{h}_k, \hat{h}_0, \alpha) = \alpha \cdot \hat{h}_k(\boldsymbol{x}) + (1-\alpha) \cdot \hat{h}_0(\boldsymbol{x}), \tag{EC.12}$$

where $\hat{\boldsymbol{m}} = \{\hat{\boldsymbol{m}}_k, k \in [K]\}$ denote the set of count numbers. Given the decoupled and shared estimators, $\hat{h}_k$ and $\hat{h}_0$, the out-of-sample cost function of the FlexDA estimator is computed as,

$$Z(\alpha) = \frac{1}{K}\sum_{k \in [K]} Z_k(\alpha, \hat{h}_k, \hat{h}_0),$$
$$Z_k(\alpha, \hat{h}_k, \hat{h}_0) = \boldsymbol{p}_k^\top \boldsymbol{c}_k(\hat{h}_k^{\text{FlexDA}}(\cdot; \hat{h}_k, \hat{h}_0, , \alpha)).$$

Suppose the probabilities $\boldsymbol{p}_k, k \in [K]$ are known, the optimal weight α can be derived as follows,

$$\alpha^{\mathrm{OR}} \in \operatorname*{argmin}_{\alpha \in [0,1]} Z(\alpha).$$

However, the above probability distribution is unknown. As described in Section 4, we adopt the leave-one-out (LOO) validation to tune the weight parameter α. In particular, for each fold of LOO validation, we reserve one data point for validation, resulting in a training set removing that particular data point. We denote $\boldsymbol{e}_i$ as an η-dimensional vector with 1 in the i-th position and zeros elsewhere. Thus, $\hat{\boldsymbol{m}}_k - \boldsymbol{e}_i$ represents the training set with the i-th data point removed. With a slight abuse of notation, in the following discussion, we replace $\hat{h}_k$ by $\hat{\boldsymbol{m}}_k$ or $(\hat{\boldsymbol{m}}_k - \boldsymbol{e}_i)$ in $\hat{h}_k^{\mathrm{FlexDA}}$ to explicitly indicate the dependence on the observed data points. The resulting data-driven cost function for selecting α can then be written as follows,

$$Z^{\mathrm{LOO}}(\alpha) = \frac{1}{K} \sum_{k \in [K]} Z_k^{\mathrm{LOO}}(\alpha),$$
$$Z_k^{\mathrm{LOO}}(\alpha) = \frac{1}{N} \sum_{i=1}^{\eta} \hat{m}_{k,i} c_{k,i}(\hat{h}_k^{\mathrm{FlexDA}}(\hat{\boldsymbol{m}}_k - \boldsymbol{e}_i, \hat{h}_0, \alpha)).$$

The data-driven weight $\hat{\alpha}$ is obtained by optimizing the LOO objective,

$$\hat{\alpha} \in \operatorname*{argmin}_{\alpha \in [0,1]} Z^{\mathrm{LOO}}(\alpha). \tag{EC.13}$$

In the following, we will derive an upper bound on the optimality gap associated with the data-driven weight alpha $\hat{\alpha}$. The key step is to characterize the behaviors of the stochastic process associated with the cost functions $Z_k(\alpha)$ and $Z_k^{\mathrm{LOO}}(\alpha)$. We will apply Theorem EC.1 to characterize the maximal deviations of these processes with the corresponding packing numbers.

For ease of exposition, for any α, we define two cost vectors as follows,

$$\boldsymbol{Z}(\alpha) = (Z_1(\alpha), \cdots, Z_K(\alpha)),$$
$$\boldsymbol{Z}^{\mathrm{LOO}}(\alpha) = (Z_1^{\mathrm{LOO}}(\alpha), \cdots, Z_K^{\mathrm{LOO}}(\alpha)).$$

Based on Assumption 3 (iv), the vector $\boldsymbol{F} = [C, C, \cdots, C] \in \mathbb{R}^K$ is an envelope for $\boldsymbol{Z}(\alpha)$, and $\boldsymbol{F}^{\mathrm{LOO}} = \frac{1}{N}[\hat{N}_1 C, \hat{N}_2 C, \cdots, \hat{N}_K C] \in \mathbb{R}^K$ is an envelope for $\boldsymbol{Z}^{\mathrm{LOO}}$. We then derive the following lemma that characterizes the packing numbers of the cost vectors.

LEMMA EC.2. *Under Assumption 3, we have, for any* $0 < \xi \le 1$,

$$D(\xi \|\boldsymbol{F}\|_2, \{\boldsymbol{Z}(\alpha) : 0 \le \alpha \le 1\}) \le 1 + \frac{4L\Pi_1}{C\xi}, \tag{EC.14}$$
$$D(\xi \|\boldsymbol{F}^{\mathrm{LOO}}\|_2, \{\boldsymbol{Z}^{\mathrm{LOO}}(\alpha) : 0 \le \alpha \le 1\}) \le 1 + \frac{4L\Pi_1}{C\xi}. \tag{EC.15}$$

Proof of Lemma EC.2. According to Lemma 5.5 of Wainwright (2019), we know that $D(\delta, \mathbb{T}) \le N(\delta/2, \mathbb{T})$. Therefore, in the subsequent analysis, we complete the proof by obtaining an upper bound on the covering number. We first prove Equation (EC.14). Let $\{\alpha_1, \cdots, \alpha_M\}$ be a $\frac{C\xi}{4L\Pi_1}-$ covering of $[0,1]$. Note that, $M \le 1 + \frac{4L\Pi_1}{C\xi}$. For any $\alpha \in [0,1]$, let α_j be the nearest element of α. Then, we have,

$$\begin{aligned}
|Z_k(\alpha) - Z_k(\alpha_j)| &\le \sum_{i=1}^{\eta} p_{k,i} \| c_{k,i}(\hat{h}_k^{\mathrm{FlexDA}}(\boldsymbol{a}_k^i; \hat{\boldsymbol{m}}_k, \hat{h}_0, \alpha)) - c_{k,i}(\hat{h}_k^{\mathrm{FlexDA}}(\boldsymbol{a}_k^i; \hat{\boldsymbol{m}}_k, \hat{h}_0, \alpha_j)) \|_2 \\
&= \sum_{i=1}^{\eta} L p_{k,i} \| \alpha \hat{h}_k(\boldsymbol{a}_k^i; \hat{\boldsymbol{m}}_k) + (1-\alpha)\hat{h}_0(\boldsymbol{a}_k^i) - \alpha_j \hat{h}_k(\boldsymbol{a}_k^i; \hat{\boldsymbol{m}}_k) - (1-\alpha_j)\hat{h}_0(\boldsymbol{a}_k^i) \|_2 \\
&\le \sum_{i=1}^{\eta} L p_{k,i} |\alpha - \alpha_j| \cdot \| \hat{h}_k(\boldsymbol{a}_k^i; \hat{\boldsymbol{m}}_k) - \hat{h}_0(\boldsymbol{a}_k^i) \|_2 \\
&\le \sum_{i=1}^{\eta} p_{k,i} |\alpha - \alpha_j| (2L\Pi_1)
\end{aligned}$$

$$\leq \frac{C\xi}{4L\Pi_1}(2L\Pi_1) = \frac{C\xi}{2}.$$

The third inequality holds by Assumption 3(iii), and the fourth inequality holds due to the definition of covering numbers. Then, we have,

$$\|\boldsymbol{Z}(\alpha) - \boldsymbol{Z}(\alpha_j)\|_2 = \|[Z_1(\alpha) - Z_1(\alpha_j), Z_2(\alpha) - Z_1(\alpha_2), \cdots, Z_K(\alpha) - Z_K(\alpha_j)]\|_2 \leq \frac{\xi}{2}\|\boldsymbol{F}\|_2.$$

Thus, $\{\boldsymbol{Z}(\alpha) : \alpha \in \{\alpha_1, \cdots, \alpha_M\}\}$ can cover the set of all possible realizations of the true cost function with $\delta = \xi\|\boldsymbol{F}\|_2$. This concludes the proof for Equation (EC.14).

Next, we prove Equation (EC.15) via a similar analysis. Let $\{\alpha_1, \cdots, \alpha_M\}$ be a $\frac{C\xi}{4L\Pi_1}$-covering of $[0,1]$. Note that, $M \leq 1 + \frac{4L\Pi_1}{C\xi}$. For any $\alpha \in [0,1]$, let α_j be the nearest element of α. Then, we have,

$$\begin{aligned}
|Z_k^{\text{LOO}}(\alpha) - Z_k^{\text{LOO}}(\alpha_j)| &\leq \frac{1}{N}\sum_{i=1}^{\eta} \hat{m}_{k,i}\|c_{k,i}(\hat{h}_k^{\text{FlexDA}}(\boldsymbol{a}_k^i; \hat{\boldsymbol{m}}_k - \boldsymbol{e}_i, \hat{h}_0, \alpha)) - c_{k,i}(\hat{h}_k^{\text{FlexDA}}(\boldsymbol{a}_k^i; \hat{\boldsymbol{m}}_k - \boldsymbol{e}_i, \hat{h}_0, \alpha_j))\|_2 \\
&\leq \frac{1}{N}\sum_{i=1}^{\eta} \hat{m}_{k,i} L\|\alpha\hat{h}_k(\boldsymbol{a}_k^i; \hat{\boldsymbol{m}}_k - \boldsymbol{e}_i) + (1-\alpha)\hat{h}_0(\boldsymbol{a}_k^i) - \alpha_j\hat{h}_k(\boldsymbol{a}_k^i; \hat{\boldsymbol{m}}_k - \boldsymbol{e}_i) - (1-\alpha_j)\hat{h}_0(\boldsymbol{a}_k^i)\|_2 \\
&\leq \frac{1}{N}\sum_{i=1}^{\eta} \hat{m}_{k,i}|\alpha - \alpha_j| \cdot L\|(\hat{h}_k(\boldsymbol{a}_k^i; \hat{\boldsymbol{m}}_k - \boldsymbol{e}_i) - \hat{h}_0(\boldsymbol{a}_k^i))\|_2 \\
&\leq |\alpha - \alpha_j|\frac{\hat{N}_k 2L\Pi_1}{N} \leq \frac{\xi C\hat{N}_k}{2N}.
\end{aligned}$$

Then, we have,

$$\|\boldsymbol{Z}^{\text{LOO}}(\alpha) - \boldsymbol{Z}^{\text{LOO}}(\alpha_j)\|_2 \leq \frac{\xi}{2}\|\boldsymbol{F}^{\text{LOO}}\|_2$$

Thus, $\{\boldsymbol{Z}^{\text{LOO}}(\alpha) : \alpha \in \{\alpha_1, \cdots, \alpha_M\}\}$ can cover the set of all possible realizations of the true cost function with $\delta = \xi\|\boldsymbol{F}^{\text{LOO}}\|_2$. This concludes the proof of Equation (EC.15). □

Using Theorem EC.1 and the packing numbers we characterize in Lemma EC.2, we can bound the maximal deviations of the average out-of-sample or LOO costs, which are formally presented in the following lemma.

LEMMA EC.3 (**Bounding the Maximal Deviations**). *Under Assumption 3, for any $0 < \delta \leq 1/2$, the following two statements each hold with probability at least $1 - \delta$:*

$$\sup_{\alpha}\left|\frac{1}{K}\sum_{k\in[K]}\Big(Z_k(\alpha) - \mathbb{E}[Z_k(\alpha)]\Big)\right| \leq 9\sqrt{2}e^2 C\left(\frac{\sqrt{\pi}}{2} + \sqrt{\log\left(1 + \frac{4L\Pi_1}{C}\right)}\right)\frac{\log^{1/2}(1/\delta)}{K},$$

$$\sup_{\alpha}\left|\frac{1}{K}\sum_{k\in[K]}\Big(Z_k^{\text{LOO}}(\alpha) - \mathbb{E}[Z_k^{\text{LOO}}(\alpha)]\Big)\right| \leq 108\sqrt{2}e^{2.5} C\left(\frac{\sqrt{\pi}}{2} + \sqrt{\log\left(1 + \frac{4L\Pi_1}{C}\right)}\right)\frac{\log^{3/2}(1/\delta)\log(K)}{K}.$$

Proof of Lemma EC.3. To prove the first inequality, we will apply Theorem EC.1 to the process $\{\boldsymbol{Z}(\alpha, \hat{f}_0) : 0 \leq \alpha \leq 1\}$. To that end, we first bound the variable J in Equation (EC.8). In particular, we have,

$$J \leq 9C\int_0^1 \sqrt{\log\left(1 + \frac{4L\Pi_1}{C\xi}\right)}d(\xi) \leq 9C\int_0^1 \sqrt{\log\left(\frac{1}{\xi} + \frac{4L\Pi_1}{C\xi}\right)}d(\xi) = 9C\int_0^1 \sqrt{\log\left(\frac{1}{\xi}\left(1 + \frac{4L\Pi_1}{C}\right)\right)}d(\xi). \tag{EC.16}$$

The first inequality follows from Lemma EC.2, and the second holds because $\frac{1}{\xi} \geq 1$ for $\xi \in (0,1)$.

Furthermore, by setting $u = \sqrt{2\log\left(\frac{1}{\xi}\left(1 + \frac{4L\Pi_1}{C}\right)\right)}$, we have,

$$\begin{aligned}
\int_0^1 \sqrt{\log\left(\frac{1}{\xi}\left(1 + \frac{4L\Pi_1}{C}\right)\right)}d(\xi) &= \frac{1 + \frac{4L\Pi_1}{C}}{\sqrt{2}}\int_{\sqrt{2\log\left(1+\frac{4L\Pi_1}{C}\right)}}^{\infty} u^2 e^{-u^2/2}d(u) \\
&= \frac{1 + \frac{4L\Pi_1}{C}}{\sqrt{2}}\int_{\sqrt{2\log\left(1+\frac{4L\Pi_1}{C}\right)}}^{\infty}(-u)d(e^{-u^2/2})
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\sqrt{2}}{2}(1+\frac{4L\Pi_1}{C})ue^{-u^2/2}\Big|_{\infty}^{\sqrt{2\log\left(1+\frac{4L\Pi_1}{C}\right)}} + \frac{1+\frac{4L\Pi_1}{C}}{\sqrt{2}}\int_{\sqrt{2\log\left(1+\frac{4L\Pi_1}{C}\right)}}^{\infty} e^{-u^2/2}d(u)\\
&= \sqrt{\log\left(1+\frac{4L\Pi_1}{C}\right)} + \frac{1+\frac{4L\Pi_1}{C}}{\sqrt{2}}\int_{\sqrt{2\log\left(1+\frac{4L\Pi_1}{C}\right)}}^{\infty} e^{-u^2/2}d(u)\\
&\le \sqrt{\log\left(1+\frac{4L\Pi_1}{C}\right)} + \frac{1}{\sqrt{2}}\int_0^{\infty} e^{-u^2/2}d(u)\\
&= \frac{\sqrt{\pi}}{2} + \sqrt{\log\left(1+\frac{4L\Pi_1}{C}\right)}.
\end{aligned}
$$

The third equality follows from integration by parts, and the first inequality holds because the function $x\int_{\sqrt{2\log(x)}}^{\infty} e^{-u^2/2}\,du$ is non-increasing in x, and $1+\frac{4L\Pi_1}{C} > 1$. Thus, we have,

$$
J \le 9C\left(\frac{\sqrt{\pi}}{2} + \sqrt{\log\left(1+\frac{4L\Pi_1}{C}\right)}\right). \tag{EC.17}
$$

Thus, according to Theorem EC.1, we have,

$$
\sup_{\alpha}\left|\frac{1}{K}\sum_{k\in[K]}\left(Z_k(\alpha,\hat{f}_0)-\mathbb{E}[Z_k(\alpha,\hat{f}_0)]\right)\right| \le 5^{1/p}\sqrt{p}\|J\|_p K^{-1}\delta^{-1/p} \le 9C\left(\frac{\sqrt{\pi}}{2} + \sqrt{\log\left(1+\frac{4L\Pi_1}{C}\right)}\right)5^{1/p}\sqrt{p}K^{-1}\delta^{-1/p}.
$$

For any $0<\delta<1/2$, by taking $p = 2\log(1/\delta)\ge 1$, we have $5^{1/p}\delta^{-1/p} = (\frac{5}{\delta})^{1/(2\log(1/\delta))}$. Thus,

$$
\left(\frac{5}{\delta}\right)^{1/\left(2\log(1/\delta)\right)} = \exp\left(\frac{\log(5/\delta)}{2\log(1/\delta)}\right) = \exp\left(\frac{\log(5)+\log(1/\delta)}{2\log(1/\delta)}\right) = \exp\left(\frac{\log(5)}{2\log(1/\delta)}+\frac{1}{2}\right) \le \exp\left(\frac{\log(5)}{2\log(2)}+\frac{1}{2}\right) \le e^2.
$$

Thus, we have,

$$
\sup_{\alpha}\left|\frac{1}{K}\sum_{k\in[K]}\left(Z_k(\alpha)-\mathbb{E}[Z_k(\alpha)]\right)\right| \le 9\sqrt{2}e^2 C\left(\frac{\sqrt{\pi}}{2} + \sqrt{\log\left(1+\frac{4L\Pi_1}{C}\right)}\right)\frac{\log^{1/2}(1/\delta)}{K}.
$$

We can prove the second inequality of Lemma EC.3 via a similar analysis. We denote $\hat{N}_{\max} = \max_k \hat{N}_k$. First, we bound the variable J in Equation (EC.8) as follows,

$$
J \le 9C\frac{\hat{N}_{\max}}{N}\int_0^1\sqrt{\log\left(1+\frac{4L\Pi_1}{C\xi}\right)}d(\xi) \le 9C\frac{\hat{N}_{\max}}{N}\int_0^1\sqrt{\log\left(\frac{1}{\xi}\left(1+\frac{4L\Pi_1}{C}\right)\right)}d(\xi) \le 9\frac{\hat{N}_{\max}}{N}C\left(\frac{\sqrt{\pi}}{2} + \sqrt{\log\left(1+\frac{4L\Pi_1}{C}\right)}\right). \tag{EC.18}
$$

Thus, we have,

$$
\|J\|_p \le \|\hat{N}_{\max}\|_p\frac{9C}{N}\left(\frac{\sqrt{\pi}}{2} + \sqrt{\log\left(1+\frac{4L\Pi_1}{C}\right)}\right). \tag{EC.19}
$$

By Lemma B.5(v) of Gupta and Kallus (2022), we have,

$$
\|\hat{N}_{\max}\|_p \le 6^{1/p}\left(\frac{6p}{e}\right)N\log(K). \tag{EC.20}
$$

Combining the above relationships, we have,

$$
\|J\|_p \le 6^{1/p}\left(\frac{6p}{e}\right)\log(K)9C\left(\frac{\sqrt{\pi}}{2} + \sqrt{\log\left(1+\frac{4L\Pi_1}{C}\right)}\right). \tag{EC.21}
$$

For any $0<\delta<1/2$, by taking $p = 2\log(1/\delta)\ge 1$, we have,

$$
6^{1/p} = \exp\left(\frac{\log(6)}{2\log(1/\delta)}\right) \le \exp\left(\frac{\log(6)}{2\log(2)}\right) \le e^{1.5}. \tag{EC.22}
$$

We then deduce,

$$
\|J\|_p \le 108e^{0.5}\log(1/\delta)\log(K)C\left(\frac{\sqrt{\pi}}{2} + \sqrt{\log\left(1+\frac{4L\Pi_1}{C}\right)}\right). \tag{EC.23}
$$

Thus, by Theorem EC.1, we have,

$$
\sup_{\alpha}\left|\frac{1}{K}\sum_{k\in[K]}\left(Z_k^{\mathrm{LOO}}(\alpha)-\mathbb{E}[Z_k^{\mathrm{LOO}}(\alpha)]\right)\right| \le 108\sqrt{2}e^{2.5}C\left(\frac{\sqrt{\pi}}{2} + \sqrt{\log\left(1+\frac{4L\Pi_1}{C}\right)}\right)\frac{\log^{3/2}(1/\delta)\log(K)}{K}.
$$

This concludes the proof. □

Proof of Theorem 2. We can decompose the optimality gap as follows,

$$\begin{aligned}
\mathrm{SubOpt}_K(\hat{\alpha}) =& \frac{1}{K}\sum_{k\in[K]} Z_k(\hat{\alpha}) - \frac{1}{K}\sum_{k\in[K]} Z_k(\alpha^{\mathrm{OR}}) \\
\le& \frac{1}{K}\sum_{k\in[K]} Z_k(\hat{\alpha}) - \frac{1}{K}\sum_{k\in[K]} Z_k(\alpha^{\mathrm{OR}}) + \frac{1}{K}\sum_{k\in[K]} Z_k^{\mathrm{LOO}}(\alpha^{\mathrm{OR}}) - \frac{1}{K}\sum_{k\in[K]} Z_k^{\mathrm{LOO}}(\hat{\alpha}) \\
\le& 2\sup_{\alpha}\left|\frac{1}{K}\sum_{k\in[K]} Z_k(\alpha) - \frac{1}{K}\sum_{k\in[K]} Z_k^{\mathrm{LOO}}(\alpha)\right| \\
\le& 2\sup_{\alpha}\left|\frac{1}{K}\sum_{k\in[K]} Z_k(\alpha) - \frac{1}{K}\sum_{k\in[K]} \mathbb{E}[Z_k(\alpha)]\right| + 2\sup_{\alpha}\left|\frac{1}{K}\sum_{k\in[K]} Z_k^{\mathrm{LOO}}(\alpha) - \frac{1}{K}\sum_{k\in[K]} \mathbb{E}[Z_k^{\mathrm{LOO}}(\alpha)]\right| \\
&+ 2\sup_{\alpha}\left|\frac{1}{K}\sum_{k\in[K]} \mathbb{E}[Z_k(\alpha)] - \frac{1}{K}\sum_{k\in[K]} \mathbb{E}[Z_k^{\mathrm{LOO}}(\alpha)]\right| \\
=& 2\sup_{\alpha}\left|\frac{1}{K}\sum_{k\in[K]} Z_k(\alpha) - \frac{1}{K}\sum_{k\in[K]} \mathbb{E}[Z_k(\alpha)]\right| + 2\sup_{\alpha}\left|\frac{1}{K}\sum_{k\in[K]} Z_k^{\mathrm{LOO}}(\alpha) - \frac{1}{K}\sum_{k\in[K]} \mathbb{E}[Z_k^{\mathrm{LOO}}(\alpha)]\right|.
\end{aligned}$$

The first inequality holds because $\frac{1}{K}\sum_{k\in[K]} Z_k^{\mathrm{LOO}}(\alpha^{\mathrm{OR}}) - \frac{1}{K}\sum_{k\in[K]} Z_k^{\mathrm{LOO}}(\hat{\alpha}) > 0$, which follows from the definition of $\hat{\alpha}$. The last equality holds because $\mathbb{E}[Z_k(\alpha)] = \mathbb{E}[Z_k^{\mathrm{LOO}}(\alpha)]$, which is based on Lemma EC.1.

Then, according to Lemma EC.3, for any $0 < \delta \le 1/2$, the following statement holds with probability at least $1 - 2\delta$:

$$\begin{aligned}
&2\sup_{\alpha}\left|\frac{1}{K}\sum_{k\in[K]} Z_k(\alpha) - \frac{1}{K}\sum_{k\in[K]} \mathbb{E}[Z_k(\alpha)]\right| + 2\sup_{\alpha}\left|\frac{1}{K}\sum_{k\in[K]} Z_k^{\mathrm{LOO}}(\alpha) - \frac{1}{K}\sum_{k\in[K]} \mathbb{E}[Z_k^{\mathrm{LOO}}(\alpha)]\right| \\
&\le 432\sqrt{2}e^{2.5}C\left(\frac{\sqrt{\pi}}{2} + \sqrt{\log\left(1 + \frac{4L\Pi_1}{C}\right)}\right)\frac{\log^{3/2}(1/\delta)\log(K)}{K}.
\end{aligned}$$

Thus, we can know that for any $0 < \delta \le 1$, the following statement holds with probability at least $1 - \delta$,

$$\mathrm{SubOpt}_K(\hat{\alpha}) \le 432\sqrt{2}e^{2.5}C\left(\frac{\sqrt{\pi}}{2} + \sqrt{\log\left(1 + \frac{4L\Pi_1}{C}\right)}\right)\frac{\log^{3/2}(2/\delta)\log(K)}{K}. \tag{EC.24}$$

This completes the proof. □

B. Empirical validation of FlexDA for Prediction and Decision-making

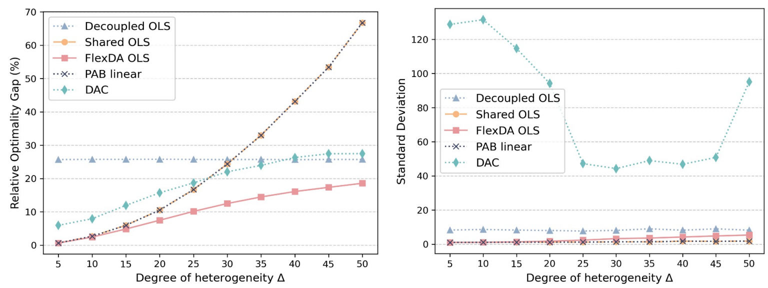
In this section, we present the empirical validation of FlexDA for prediction and decision-making with synthetic and real data.

B.1. Empirical Validation of FlexDA for Prediction with Synthetic Data

In this section, we conducted a comprehensive numerical experiment to validate the effectiveness of the proposed method by varying the parameters associated with the data-generating process. In addition to the conventional approaches, such as Decoupled OLS and Shared OLS, we also implement the Data Aggregation with Clustering (DAC) proposed by Cohen et al. (2022b) that employs hypothesis testing to determine the clustering structure of model parameters and estimate the parameters with linear models based on the cluster structure. The pooling and boosting (PAB) approach proposed by Lei et al. (2025) first constructs proxy data at the category level. The PAB approach then fits regression models to the original data with the estimates obtained from the proxy data as the regularizer. To compare the performance of the proposed method with the benchmark ones with a scale-independent measure, we compute the relative optimality gap for each approach $\mathcal{M}$ with respect to the optimal out-of-sample cost, which is computed as, $(\mathcal{L}_{\mathcal{M}} - \mathcal{L}^*)/\mathcal{L}^* \times 100\%$, where $\mathcal{L}_{\mathcal{M}}$ represents the expected average out-of-sample cost for method $\mathcal{M}$, and $\mathcal{L}^*$ denotes the optimal averaged out-of-sample cost achieved when the true parameters $\beta_{k,0}$ and $\beta_{k,1}$ are known. The parameter setting introduced in Section 5 is set as the basic case. We vary the degree of heterogeneity across products to investigate its impact on the performance of the implemented approaches. Specifically, the true parameters $\beta_{k,0}$ and $\beta_{k,1}$ are drawn from a uniform distribution $[50 - \Delta, 50 + \Delta]$, where Δ, indicating the degree of heterogeneity across products, ranges from 5 to 50. The results are reported in Figure EC.1, where each result is the average of 100 instances.

The FlexDA OLS approach outperforms the benchmark ones with different levels of heterogeneity, except for the case with $\Delta = 5$, where the relative optimality gaps of FlexDA OLS and Shared OLS are 0.75% and 0.70%, respectively. With a relatively small Δ, the

Figure EC.1 Relative optimality gaps for prediction.



(a) Averaged results (b) Standard deviation

Note. $D_k = \beta_{k,0} + \beta_{k,1} x_k + \epsilon_k$, where $\epsilon_k \sim \mathcal{N}(0, 50^2)$. The true parameters $\beta_{k,0}$ and $\beta_{k,1}$ are drawn from a uniform distribution $[50 - \Delta, 50 + \Delta]$, where Δ ranges from 5 to 50. The total number of products $K = 1000$ and the sample size $N_k = 10$ for each product. Panel (a) presents the averaged results 100 instances and Panel (b) presents the standard deviation of the relative optimality gaps 100 instances.

products can be regarded as almost homogeneous, which suggests that the prediction generated by the shared OLS should be close to the demand mean of each product. In this case, the performance of FlexDA OLS is comparable to that of shared OLS. A further investigation shows that FlexDA OLS outperforms Shared OLS if the number of problems K increases from 1,000 to 10,000 even when $\Delta = 5$. In this setting, the relative optimality gaps for FlexDA OLS and Shared OLS are 0.657% and 0.664%, respectively. As the degree of heterogeneity increases, while the decoupled OLS approach remains insensitive to such changes, the performance of approaches involving certain forms of data aggregation, such as shared OLS, FlexDA, DAC, and PAB, deteriorates. This is consistent with our intuition that the effectiveness of data aggregation is damped by a pronounced product heterogeneity. It is worth noting that when the degree of heterogeneity is relatively high, the simple decoupled OLS can even outperform the modern approaches such as DAC and PAB, as the benefit of data aggregation becomes marginal. However, our proposed FlexDA approach, which carefully combines individual datasets and the aggregated dataset, can flexibly capture different degrees of heterogeneity across products, resulting in superior performance compared to decoupled OLS in this case.

Impacts of sample size and problem size. We vary the sample size N from 10 to 25 and the number of products K from 100 to 1000 to validate the performance of the proposed FlexDA approach. As demonstrated in Figure EC.2 (a), the performances of approaches that assume a shared model across products, such as Shared OLS and PAB, are consistent with different sample sizes as the total sample size of the fully aggregated dataset is at least 10,000. While the relative optimality gaps of data aggregation approaches decrease as the sample size increases, the advantage of the FlexDA approach is more pronounced when the sample size is relatively small, which suggests its applicability to address Meituan's data scarcity issue. Furthermore, we find that the proposed approach outperforms the benchmarks with different numbers of products. Surprisingly, even when the number of products is relatively small (e.g., 100), the FlexDA approach remains to result in the smallest relative optimality gap. This implies that the proposed FlexDA approach can still be effective in situations when the firm needs to manage a moderate amount of products.

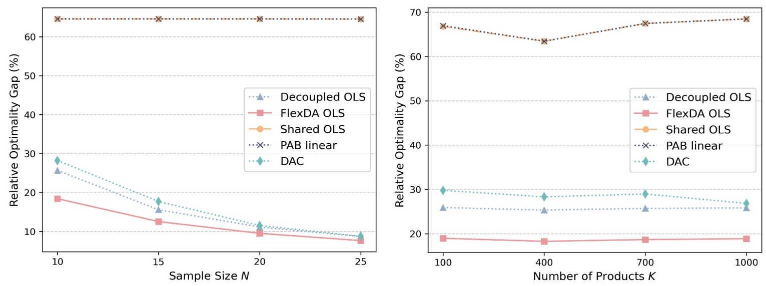
B.2. Empirical Validation of FlexDA for Prediction with Real Data

In this section, we validate the performance of the FlexDA method using sales data from the Rossmann store, a European pharmacy chain operating in seven countries, and compare its prediction power with various state-of-the-art models. The detailed data description and preprocessing are articulated in Appendix C. All relevant data and code can be found at GitHub.[1].

Regarding the implementation of the FlexDA method, we employ the two-level FlexDA estimators that build on two extreme-case partitions, which are discussed in Section 5. In particular, we propose three variants of the two-level FlexDA methods as summarized in Table EC.1, which differ in the model classes used to construct the submodels in two levels. We denote the FlexDA $(M_1 + M_2)$ as the generic two-level FlexDA model in which the first level adopts model class M_1, and the second level adopts model class M_2.

[1] https://github.com/Zhenkang-Peng/FlexDA.

Figure EC.2 The relative optimality gap under different N and K.



(a) Sample size N (b) Number of products K

Note. $D_k = \beta_{k,0} + \beta_{k,1} x_k + \epsilon_k$, where $\epsilon_k \sim \mathcal{N}(0, 50^2)$. The true parameters $\beta_{k,0}$ and $\beta_{k,1}$ are drawn from a uniform distribution $[0, 100]$. Each reported result is the average of 100 instances.

Specifically, the first variant, referred to as "FlexDA (linear + linear)", adopts linear models for both the decoupled estimators and the shared estimators. Thus, we can compute the closed-form weight according to Proposition 3. The second and third FlexDA models, referred to as "FlexDA (linear + random forest)" and "FlexDA (SAA + linear)", will obtain the weight by implementing Algorithm 1.

In addition to the benchmark approaches introduced in Section B.1, we adopt the Shrunken SAA, proposed by Gupta and Kallus (2022), that shrinks the SAA solutions by an anchor distribution. Since the proposed Shrunken SAA approach did not consider feature information, we excluded feature information from the implementation of Shrunken SAA. The implemented benchmarks are summarized in the following table EC.1.

Table EC.1 Summary of the implemented methods.

Model	Description
FlexDA (linear + linear):	A two-level FlexDA with linear models at both levels.
FlexDA (linear + random forest):	A two-level FlexDA with a linear model (random forecast) at the first (second) level.
FlexDA (SAA + linear):	A two-level FlexDA with SAA at the first level and a linear model at the second.
Decoupled OLS:	Train a linear model for each product using its own dataset.
Shared OLS:	Train a shared linear model for all products using the fully aggregated data.
Shared Random Forest:	Train a shared random forest for all products using the fully aggregated data.
Shared Random Forest (product index):	Shared Random Forest with the product index as an additional feature.
Shrunken SAA:	The shrunken SAA approach without accounting for feature information.
DAC:	The data aggregation with clustering (DAC) approach assuming linear models.
PAB linear:	The pooling and boosting (PAB) approach assuming linear models.
PAB tree:	Similar to "PAB linear", but employs XGBoost to estimate the shared model.

According to the above model description, we would like to highlight that the three FlexDA models, the decoupled OLS and the shrunken SAA, take into consideration the heterogeneity among products, while the shared OLS, the shared random forest, the PAB linear, and the PAB tree focus on the homogeneity among products. The DAC approach may either emphasize heterogeneity or homogeneity across products, depending on the estimated cluster structure, which prescribes the degree of data aggregation.

We adopt *rolling validation* to evaluate and compare the performance of implemented models. In particular, when predicting the demand at period t, the sales data recorded from period $t - N$ to $t - 1$ are used as the training data, and the realized demand at period t is reserved as the testing data. For example, if we set $N = 15$ in this experiment, then we can construct 614 pairs of training and test datasets, as a total of 629 days of sales for each store are available.[2]

We investigate the performance of the proposed approach with various sample sizes N, which range from 10 to 30. As shown in Table EC.2, the FlexDA-based approaches consistently outperform the existing ones. Interestingly, when the sample size ranges from 15 to 30, the simple decoupled OLS outperforms all other approaches except for the FlexDA-based approaches. This might be due to the relatively high heterogeneity across products in this real dataset, as discussed in Section B.1. In a similar vein, we observe that approaches that train a shared model with fully aggregated data for all products, such as PAB linear or tree, shared random forest, and shared OLS,

[2] Note that, although we allocate N data samples for each store as training data, the actual number of data samples for implementation can be different due to the issue of missing data.

Table EC.2 Out-of-sample averaged MSE (standard deviation) (10^6) for all methods under different sample sizes N.

Models	$N = 10$	$N = 15$	$N = 20$	$N = 25$	$N = 30$
FlexDA (linear + linear)	6.43 (32.39)	**2.47** (3.95)	**1.97** (2.39)	**1.79** (2.14)	**1.50** (1.75)
FlexDA (linear + Random Forest)	5.94 (33.37)	2.53 (4.81)	1.99 (2.84)	1.84 (2.51)	1.53 (1.80)
FlexDA (SAA + linear)	**2.16** (2.93)	2.63 (3.35)	2.47 (2.91)	2.49 (3.04)	2.48 (2.96)
Decoupled OLS	11.31 (84.46)	2.86 (6.83)	2.15 (4.00)	1.90 (2.67)	1.55 (1.88)
Shrunken SAA	2.76 (3.51)	3.00 (3.92)	2.89 (3.38)	2.78 (3.45)	2.83 (3.35)
DAC	7.29 (19.64)	4.08 (3.31)	3.44 (2.54)	3.24 (2.61)	2.62 (2.28)
PAB linear	7.79 (4.04)	7.41 (3.69)	7.15 (3.17)	7.08 (3.32)	6.92 (3.15)
Shared Random Forest	6.95 (4.57)	7.77 (6.22)	7.63 (5.22)	7.63 (5.48)	7.54 (4.49)
Shared Random Forest (product index)	6.67 (4.56)	5.70 (4.09)	5.27 (3.82)	4.98 (3.82)	4.82 (3.73)
Shared OLS	8.91 (14.34)	7.42 (3.77)	7.15 (3.19)	7.08 (3.33)	6.92 (3.16)
PAB tree	6.99 (4.85)	8.59 (14.89)	8.13 (8.68)	7.98 (9.51)	8.69 (18.23)

perform significantly worse than methods that attempt to capture heterogeneity among products, such as FlexDA, Decoupled OLS, and Shrunken SAA.

In the case where $N = 10$, the FlexDA approach with SAA as the decoupled model and a shared linear model performs the best, while the shrunken SAA, which ignores contextual information, ranks second. With limited observations ($N = 10$), from the perspective of the bias-variance tradeoff, the variance reduction outweighs the bias introduced by ignoring features. In other words, a simpler model may avoid the risk of overfitting with small samples. Without leveraging variance reduction through data aggregation, the decoupled OLS, which suffers from highly variable estimates, results in the highest out-of-sample cost in this case. In general, FlexDA-based approaches are the best, demonstrating their robustness under different degrees of data availability.

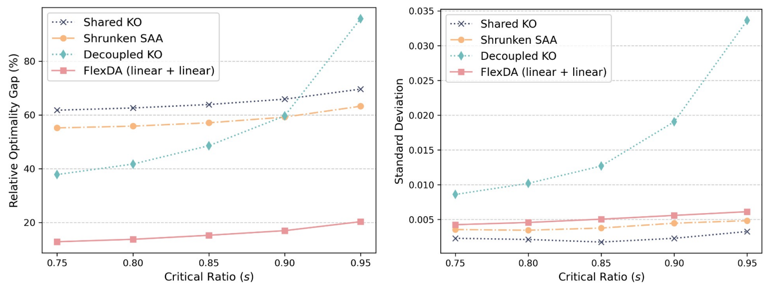
B.3. Empirical Validation of FlexDA for Decision Making with Synthetic Data

In this section, we will demonstrate the effectiveness of our FlexDA method with general cost functions via synthetic data and further conduct the empirical validation with real data in Section B.4. Specifically, we consider the newsvendor loss function, i.e., $c_k(q_k, d_k) = c_h(q_k - d_k)^+ + c_b(d_k - q_k)^+$, where c_h denotes the per-unit holding cost and c_b denotes the per-unit lost-sale penalty. Thus, the critical ratio for deriving the optimal newsvendor solution is computed as, $s = \frac{c_b}{c_b + c_h}$. In the following experiments, we vary the critical ratio s from 0.75 to 0.95 with $c_b + c_h = 1$ fixed. In terms of the data generating process, we adopt the same parameter setting introduced in Section 5 to generate feature vectors and random demands.

For the FlexDA approach, we map features to decisions directly with linear functions by setting the cost function as the newsvendor cost in Algorithm 1 and tuning the weight parameter through leave-one-out cross-validation as described in Algorithm 1. For brevity, we denote this FlexDA approach as “FlexDA (linear + linear)”. We also implement the Shrunken SAA approach, proposed by Gupta and Kallus (2022), and the Kernel-weighted Optimization (KO) approach, proposed by Ban and Rudin (2019), to derive the order quantities for a newsvendor system in a data-driven setting. For the KO approach, we consider two variants: one that solves each product individually, denoted as “Decoupled KO” hereafter, and another that pools all data together, denoted as “Shared KO” hereafter. However, the DAC and PAB approaches, which are designed to focus on demand prediction, are not implemented in this section, as it remains unclear how to extend them to derive data-driven decisions with a general operational cost function.

We report the average relative optimality gaps of the implemented approaches in Figure EC.3, where each result is the average of 100 instances. We find that the proposed FlexDA approach outperforms the benchmark approaches regardless of the critical ratio, with significant reductions in the relative optimality gap. The Decoupled KO approach performs better than the Shared KO approach when the critical ratio is moderate, as it more effectively captures product-level differences with pronounced product heterogeneity. However, its performance deteriorates relative to the Shared KO approach when the critical ratio increases to 0.95. This decline is due to the fact that accurately estimating the high quantile associated with a large critical ratio requires significantly more data (Besbes and Mouchtaki 2023). Consequently, the Shared KO approach suffers from higher variance in such settings, which ultimately results in poor performance. At a critical ratio of 0.95, the Shrunken SAA approach outperforms the Decoupled KO approach. This improvement arises from the shrinkage mechanism in Shrunken SAA, which helps mitigate the high variance caused by small data in high-quantile estimation, effectively addressing the challenge that hinders the Decoupled KO approach.

Figure EC.3 The averaged relative optimality gap and standard deviation for all methods in the newsvendor cost.



(a) Averaged results (b) Standard deviation

B.4. Empirical Validation of FlexDA for Decision Making with Real Data

In this section, we implement FlexDA for decision making in a Newsvedor system and validate its performance with the real data introduced in Section B.2. Similar to Section B.2, since the true data-generating process of the real data is unknown, we also consider three variants of the FlexDA approach, as listed in Table EC.3. For FlexDA approaches based on linear and SAA, we map the feature vector to the decisions in an end-to-end manner with newsvendor cost as the training objective. However, when implementing the FlexDA with random forecasts, it is unclear how to split the tree structure with newsvendor objectives. Thus, we take a two-step procedure that predicts the demand means and prescribes the order quantities using the residuals for the random-forest-based FlexDA approach.

We adopt the rolling validation to calculate the out-of-sample cost with a sample size $N = 15$. We vary the critical ratio s from 0.75 to 0.95, and report the results in Table EC.3. We find that the three FlexDA-based approaches perform the best regardless of critical ratios. We also observe that FlexDA (linear + linear) performs the best, which is consistent with the prediction results. This indicates that the relationships between the outputs and features are more likely to be linear.

Table EC.3 Out-of-sample average newsverdor costs (standard deviation) under different critical ratios.

Models	$s = 0.75$	$s = 0.80$	$s = 0.85$	$s = 0.90$	$s = 0.95$
FlexDA (linear + linear)	481.65 (253.47)	428.60 (229.06)	392.63 (216.93)	311.20 (167.65)	190.76 (90.99)
FlexDA (linear + Random Forest)	499.23 (286.97)	444.14 (264.15)	412.96 (268.40)	330.02 (222.54)	199.37 (124.38)
FlexDA (SAA + linear)	517.51 (327.09)	474.63 (286.49)	406.08 (280.88)	316.81 (203.53)	191.26 (95.36)
Shrunken SAA	556.41 (437.11)	504.78 (414.74)	436.34 (403.50)	346.84 (330.09)	205.95 (161.39)
Decoupled KO	555.65 (643.57)	537.52 (691.03)	507.07 (726.38)	474.89 (781.56)	405.45 (815.25)
Shared KO	944.48 (373.63)	867.41 (347.06)	759.10 (304.39)	610.41 (244.88)	399.48 (156.81)

C. Description and Preprocessing the Public Rossmann Dataset

The dataset is obtained from a Kaggle competition (https://www.kaggle.com/competitions/rossmann-store-sales), and it has been commonly adopted for numerical studies in existing literature (e.g., Gupta and Kallus 2022). Different from this numerical experiment, Gupta and Kallus (2022) did not leverage the feature information. Specifically, the dataset initially consists of 1,782,871 entries, covering 1,115 stores. Each entry includes 7 relevant features, with approximately 17% of the data indicating that the store was closed, resulting in a demand of 0. When the stores were open, the average demand was 6,955, with the minimum demand being 46 and the maximum demand reaching 41,551.

Table EC.4 Description of features in the Rossmann sales dataset.

Feature	Description
Store	A unique index for each store
SchoolHoliday	Indicator of whether public schools were closed on the day sales were recorded
Promo	Indicates whether a store was on promotion on the day sales were recorded
Day Of Week	Indicator of the day of the Week
Date	The date of recorded sales

Following the data pre-processing steps outlined in Gupta and Kallus (2022), we excluded all weekends and data from the month of December from the dataset. This was motivated by the fact that fewer than 2.5% of stores are open on weekends, and many stores engage in periodic promotional activities throughout December leading up to Christmas. Consequently, we collected historical sales data spanning 629 days across 1105 stores, and treated each store as a product in this numerical experiment. Subsequently, we retained the data in its original format, different from Gupta and Kallus (2022), which processed the sales data with additional steps. Different from Gupta and Kallus (2022), we did not remove temporal correlations within the sales data or correlations between stores, nor did we discretize the data using support information inferred from the whole dataset.

In terms of feature selection, we first exclude the "Customers" feature since customer counts before store opening are not observable. Second, "StateHoliday", a dummy variable indicating whether the day is a state holiday or not, is 0 for 96.94% of data samples, demonstrating minimal variability. Since we are using rolling validation for evaluation (see Appendix B.2 and B.4), and the training data for each store in a single test consists of only 15 data points, the feature "StateHoliday" remains 0 in the training set for most trials and fails to contribute to the prediction. Therefore, we decided to remove this feature. The description of the remaining features is presented in Table EC.4.

D. Choice of Partition Levels under the FlexDA Framework

In this section, we demonstrate that the FlexDA approach can automatically select appropriate partition levels via a numerical experiment. We begin by introducing the data generation process adopted in our experiment. Specifically, we consider a set of K products. Each product k is characterized by a continuous feature $\boldsymbol{x}_k$ and a set of eight categorical indicators $\{z_k^j : j = 1, \ldots, 8\}$, where each $z_k^j \in \{0, 1\}$ denotes whether the product belongs to category j with equal probability. We assume that products belonging to the same category exhibit similar responses to the feature x_k. The data-generating process (DGP) for the demand of product k is specified as follows:

$$\begin{aligned} d_{k,i} &= \alpha_k + \beta_k x_{k,i} + z_k^1(\alpha_k^1 + \beta_k^1 x_{k,i}) + z_k^2(\alpha_k^2 + \beta_k^2 x_{k,i}) + z_k^3(\alpha_k^3 + \beta_k^3 x_{k,i}) + z_k^4(\alpha_k^4 + \beta_k^4 x_{k,i}) + \epsilon_{k,i} \\ &= \alpha_k + z_k^1\alpha_k^1 + z_k^2\alpha_k^2 + z_k^3\alpha_k^3 + z_k^4\alpha_k^4 + (\beta_k + z_k^1\beta_k^1 + z_k^2\beta_k^2 + z_k^3\beta_k^3 + z_k^4\beta_k^4)x_{k,i} + \epsilon_{k,i},\ i = 1, 2, \ldots, N, \end{aligned}$$

where $\alpha_k, \beta_k \sim \mathcal{N}(\mu_0, \sigma^2)$, $\alpha_k^j, \beta_k^j \sim \mathcal{N}(\mu_j, \sigma^2)$, $j = 1, \ldots, 4$. In addition, we assume that $\mu_1^2, \mu_2^2, \mu_3^2, \mu_4^2 > 2\sigma^2$.

For ease of demonstration, we assume that each level can be partitioned using only a single indicator. For example, if z_k^1 represents the product's category and z_k^2 represents the location of the RDC to which the product belongs, then we can use either category or location for data aggregation, but not a combination of the two indicators. Therefore, with eight indicators, and including both no aggregation and full aggregation, we obtain a total of ten levels at most. Based on the above data-generating process, aggregating datasets based on the first four categorical variables $\{z_k^j : j = 1, \cdots, 4\}$ may help to estimate the coefficients (α_k^j, β_k^j), $j = 1, \cdots, 4$. However, aggregating datasets based on the last four categorical variables $\{z_k^j : j = 5, \cdots, 8\}$ should yield no additional values. We unfold the intuition in detail as follows.

For any product k, the expectation and variance for the intercept term can be computed as:

$$\begin{aligned} &\mathbb{E}\left[\alpha_k + z_k^1\alpha_k^1 + z_k^2\alpha_k^2 + z_k^3\alpha_k^3 + z_k^4\alpha_k^4\right] = \mu_0 + \frac{\mu_1 + \mu_2 + \mu_3 + \mu_4}{2}, \\ &\mathbb{V}\left[\alpha_k + z_k^1\alpha_k^1 + z_k^2\alpha_k^2 + z_k^3\alpha_k^3 + z_k^4\alpha_k^4\right] = \sigma^2 + \left(0.5\sigma^2 + 0.25\mu_1^2\right) + \left(0.5\sigma^2 + 0.25\mu_2^2\right) + \left(0.5\sigma^2 + 0.25\mu_3^2\right) + \left(0.5\sigma^2 + 0.25\mu_4^2\right). \end{aligned}$$

Intuitively, the suitable partition level should group products with higher similarity together. For example, if we use the value of z_k^1 to partition the products, then for the cluster of products where $z_k^1 = 1$, the demand model of $d_{k,i}$ can be presented as:

$$d_{k,i} = \alpha_k + \alpha_k^1 + z_k^2\alpha_k^2 + z_k^3\alpha_k^3 + z_k^4\alpha_k^4 + \left(\beta_k + \beta_k^1 + z_k^2\beta_k^2 + z_k^3\beta_k^3 + z_k^4\beta_k^4\right)x_{k,i} + \epsilon_{k,i},\ i = 1, \cdots, N.$$

The variance for this cluster of products is computed as:

$$\mathbb{V}[\alpha_k + \alpha_k^1 + z_k^2\alpha_k^2 + z_k^3\alpha_k^3 + z_k^4\alpha_k^4] = \sigma^2 + \sigma^2 + \left(0.5\sigma^2 + 0.25\mu_2^2\right) + \left(0.5\sigma^2 + 0.25\mu_2^3\right) + \left(0.5\sigma^2 + 0.25\mu_2^4\right).$$

Similarly, for the cluster of products where $z_k^1 = 0$, the demand model of $d_{k,i}$ can be presented as:

$$d_{k,i} = \alpha_k + z_k^2\alpha_k^2 + z_k^3\alpha_k^3 + z_k^4\alpha_k^4 + \left(\beta_k + z_k^2\beta_k^2 + z_k^3\beta_k^3 + z_k^4\beta_k^4\right)x_{k,i} + \epsilon_{k,i},\ i = 1, \cdots, N.$$

The variance for this cluster is computed as:

$$\mathbb{V}[\alpha_k + z_k^2\alpha_k^2 + z_k^3\alpha_k^3 + z_k^4\alpha_k^4] = \sigma^2 + \left(0.5\sigma^2 + 0.25\mu_2^2\right) + \left(0.5\sigma^2 + 0.25\mu_2^3\right) + \left(0.5\sigma^2 + 0.25\mu_2^4\right).$$

As one can easily observe, the variance of the parameters in both clusters decreases compared to that of no aggregation. Thus, the aggregation based on the indicator z_k^1 can lead to greater similarity among products within each cluster. This variance reduction effect also holds for the indicators z_k^2, z_k^3, z_k^4. However, for the indicators $z_k^5, z_k^6, z_k^7, z_k^8$, a similar analysis can be conducted to show that they don't bring any information and fail to reduce the variance by aggregation.

To demonstrate the effectiveness of FlexDA in selecting appropriate data aggregation levels, we implement FlexDA under three different choices of partitions, resulting in three corresponding FlexDA models:

1. *DGP-inspired*: This specification includes all informative categorical variables, leading to six partition levels: no aggregation, full aggregation, and four intermediate levels obtained by partitioning based on the first four categorical variables.
2. *Misspecified*: This specification omits one informative categorical variable, resulting in five partition levels: no aggregation, full aggregation, and three intermediate levels obtained by partitioning based on the first three categorical variables.
3. *All-inclusive*: This specification incorporates all available domain knowledge, yielding ten partition levels: no aggregation, full aggregation, and eight intermediate levels obtained by partitioning based on all eight categorical variables.

In this numerical experiment, we set $N = 10$, $\mu_0 = 0$, $\mu_1 = \mu_2 = \mu_3 = \mu_4 = 3$, $\sigma = 2$, and $\epsilon_{k,i} \sim \mathcal{N}(0, 10^2)$. We vary the number of products across two levels, specifically $K = 100$ and $K = 1000$. To accelerate computation, we did not directly apply Algorithm 1 in the main text to estimate the submodel weights. Instead, similar to the implementation at Meituan, we used eight data points per product to train the submodels and reserved the remaining two data points to train the meta-model weights. Based on 100 experimental instances, the average estimated weights of each FlexDA model across different partition levels are reported in Table EC.5. Meanwhile, we also present the correlation of submodel weights at corresponding data partition levels across 100 instances between the DGP-inspired and All-inclusive approaches. The results are shown in Table EC.6.

Table EC.5 Estimated weight parameters of meta-model for different choices of partition levels.

				Informative Indicators				Noninformative Indicators			
K	Methods	No Agg.	Full Agg.	1	2	3	4	5	6	7	8
100	DGP-inspired	0.47	$1.51e^{-14}$	0.14	0.14	0.12	0.13	N/A	N/A	N/A	N/A
	Misspecified	0.48	$4.03e^{-3}$	0.19	0.17	0.16	N/A	N/A	N/A	N/A	N/A
	All-inclusive	0.47	$3.84e^{-14}$	0.13	0.13	0.11	0.13	$7.34e^{-3}$	$5.42e^{-3}$	$9.64e^{-3}$	0.01
1000	DGP-inspired	0.49	$1.25e^{-11}$	0.13	0.12	0.13	0.12	N/A	N/A	N/A	N/A
	Misspecified	0.50	$1.19e^{-11}$	0.17	0.16	0.17	N/A	N/A	N/A	N/A	N/A
	All-inclusive	0.49	$1.81e^{-11}$	0.13	0.12	0.13	0.12	$1.79e^{-11}$	$1.79e^{-11}$	$1.81e^{-11}$	$1.77e^{-11}$

According to Table EC.5, we first observe that, compared with the DGP-inspired partitions, omitting one valuable indicator leads to substantial changes in the corresponding coefficients across different aggregation levels. In contrast, for the all-inclusive partitions, which are more relevant to practical implementation, especially when $K = 1000$, the meta-model can automatically learn and assign near-zero weights to submodels associated with the last four categorical variables. Meanwhile, the coefficients for the informative indicators remain consistent with those obtained under the DGP-inspired partitions. Furthermore, combining with the results in Table EC.6, we find that for the coefficients of the informative indicators, the covariance coefficients between the two methods are close to 1. This clearly indicates that the similarity between the two methods is not only at the average level but also consistent across individual instances. This result highlights the flexibility and adaptivity of FlexDA in effectively identifying and leveraging the most valuable partitions.

The effectiveness of FlexDA in identifying informative partitions also depends on having a sufficiently large sample size for training the meta-model. In our setting, although only two data points per product are used to train the weight parameter α, this still provides $2K$ observations for estimating α, that is, 200 when $K = 100$ and 2,000 when $K = 1000$. Since at most ten parameters need to be estimated, the sample size is much larger than the dimensionality of the parameter space. As a result, FlexDA can automatically discern the most

Table EC.6 Correlation of submodel weights at corresponding data partition levels between the DGP-inspired and All-inclusive models.

	No Agg.	1	2	3	4
$K = 100$	0.9997	0.9432	0.9840	0.9484	0.9971
$K = 1000$	1.0000	1.0000	1.0000	1.0000	1.0000

informative partition levels even when all potential partitions are considered. Moreover, when $K = 1000$, the increased data availability further enhances the model's performance, explaining the superior results observed in this case. Therefore, in large-scale applications, as long as the amount of domain knowledge is not excessively large, it is advisable to incorporate all available contextual information and rely on the data to assign appropriate weights adaptively.

E. Table of Notations

Table EC.7 Table of Notations.

Notation	Definition
K	number of products
F	number of features
N_k	number of data samples for each product k
p_k, q_k	price and ordering quantity for product k
$c_k(\cdot)$	cost function for prodcut k
$\hat{\boldsymbol{x}}_k, \boldsymbol{x}_k$	historical feature vector and future feature vector for product k
$x_{\max}$	maximum feature value
$\hat{\boldsymbol{X}}_k$	historical feature matrix for product k
$\hat{d}_k, d_k$	historical demand and future demand for product k
$\hat{D}_k$	historical demand vector for product k
$\mathcal{S}_k$	historical dataset for product k
$\mathcal{S}$	total dataset for all products
$\boldsymbol{\beta}_k, \hat{\boldsymbol{\beta}}_k$	true parameters and OLS estimated parameters for product k
$\hat{\boldsymbol{\beta}}^{\text{ols}} = \{\hat{\boldsymbol{\beta}}_k : k \in [K]\}$	the set for OLS estimated parameters for all products
$\boldsymbol{\beta}_0, \hat{\boldsymbol{\beta}}_0$	mean of distribution of true parameters and estimated shared parameters
$\beta_{\max}$	maximum parameter value
$\alpha, \alpha^*, \hat{\alpha}$	any given weight parameter, optimal weight parameter and estimated weight parameter
ϵ_k, σ_k^2	random noise and the variance of the random noise for product k
$\hat{\boldsymbol{\beta}}_k^{\text{FlexDA}}(\alpha, \hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_k)$	the FlexDA estimated parameter given the weight parameter α and shared model $\boldsymbol{\beta}_0$
$\mathcal{L}(\hat{\boldsymbol{\beta}}^{\text{ols}})$	the total averaged expected out-of-sample cost with OLS estimated parameters
$\mathcal{L}^{\text{FlexDA}}(\alpha, \boldsymbol{\beta}_0)$	the total averaged expected out-of-sample cost by using FlexDA method with weight parameter α and shared model $\boldsymbol{\beta}_0$

Appendix References

Amemiya T (1985) *Advanced econometrics*. Harvard university press, Cambridge, MA, USA.

Ban GY, Rudin C (2019) The big data newsvendor: Practical insights from machine learning. *Operations Research* 67(1):90–108.

Besbes O, Mouchtaki O (2023) How big should your data really be? data-driven newsvendor: learning one sample at a time. *Management Science* 69(10):5848–5865.

Chen LH (1975) Poisson approximation for dependent trials. *The Annals of Probability* 3(3):534–545.

Cohen MC, Zhang R, Jiao K (2022) Data aggregation and demand prediction. *Operations Research* 70(5):2597–2618.

Gupta V, Kallus N (2022) Data pooling in stochastic optimization. *Management Science* 68(3):1595–1615.

Lei D, Qi Y, Liu S, Geng D, Zhang J, Hu H, Shen ZJM (2025) Pooling and boosting for demand prediction in retail: A transfer learning approach. *Manufacturing & Service Operations Management* 27(6):1779–1794.

Pollard D (1990) *Empirical Processes: Theory and Applications*. Institute of Mathematical Statistics, Hayward, CA.

Wainwright MJ (2019) *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, Cambridge, UK, 21–57.