Integrating Inference and Experimental Design for Contextual Behavioral Model Learning (Supplemental Material)

Gongtao Zhou, Haoran Yu

A Derivation of EIG

We derive Expected Information Gain (EIG) and its approximation. Recall that Information Gain (IG) is defined as follows:

$$\begin{aligned}\mathrm{IG}(\boldsymbol{d},\boldsymbol{y}) &\triangleq H[p(\boldsymbol{\theta})] - H[p(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{y},\boldsymbol{d})] \\ &= \int p(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{y},\boldsymbol{d})\log p(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{y},\boldsymbol{d})d\boldsymbol{\theta} - \int p(\boldsymbol{\theta})\log p(\boldsymbol{\theta})d\boldsymbol{\theta}.\end{aligned}$$

Information Gain captures the change in information entropy under the conditions where the platform selects $\boldsymbol{d}$ and the investor chooses $\boldsymbol{y}$. The investor's choice $\boldsymbol{y}$ is determined by the context $\boldsymbol{X}$ and the design $\boldsymbol{d}$. Given $\boldsymbol{X}$ and $\boldsymbol{d}$, its distribution is given by $p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{d}) = \mathbb{E}_{p(\boldsymbol{\theta})}[p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{d},\boldsymbol{\theta})]$. We can further take the expectation of $\mathrm{IG}(\boldsymbol{d},\boldsymbol{y})$ with respect to the uncertainty of $\boldsymbol{y}$ to obtain EIG:

$$\begin{aligned}\mathrm{EIG}(\boldsymbol{d}) &\triangleq \mathbb{E}_{p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{d})}[\mathrm{IG}(\boldsymbol{d},\boldsymbol{y})] \\ &= \sum_{\boldsymbol{y}}\int p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{d})p(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{y},\boldsymbol{d})\log p(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{y},\boldsymbol{d})d\boldsymbol{\theta} \\ &\quad - \sum_{\boldsymbol{y}} p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{d})\int p(\boldsymbol{\theta})\log p(\boldsymbol{\theta})d\boldsymbol{\theta} \\ &= \sum_{\boldsymbol{y}}\int p(\boldsymbol{\theta},\boldsymbol{y}|\boldsymbol{X},\boldsymbol{d})\log p(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{y},\boldsymbol{d})d\boldsymbol{\theta} - \int p(\boldsymbol{\theta})\log p(\boldsymbol{\theta})d\boldsymbol{\theta} \\ &= \sum_{\boldsymbol{y}}\int p(\boldsymbol{\theta},\boldsymbol{y}|\boldsymbol{X},\boldsymbol{d})\log p(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{y},\boldsymbol{d})d\boldsymbol{\theta} - \sum_{\boldsymbol{y}}\int p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta},\boldsymbol{d})p(\boldsymbol{\theta})\log p(\boldsymbol{\theta})d\boldsymbol{\theta} \\ &\overset{(a)}{=} \sum_{\boldsymbol{y}}\int p(\boldsymbol{\theta},\boldsymbol{y}|\boldsymbol{X},\boldsymbol{d})\log\frac{p(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{y},\boldsymbol{d})}{p(\boldsymbol{\theta})}d\boldsymbol{\theta} \\ &\overset{(b)}{=} \sum_{\boldsymbol{y}}\int p(\boldsymbol{\theta},\boldsymbol{y}|\boldsymbol{X},\boldsymbol{d})\log\frac{p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta},\boldsymbol{d})}{p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{d})}d\boldsymbol{\theta} \\ &= \sum_{\boldsymbol{y}}\int p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta},\boldsymbol{d})\log\frac{p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta},\boldsymbol{d})}{p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{d})}d\boldsymbol{\theta}.\end{aligned}$$

We use the fact that $\boldsymbol{\theta}$ is independent of $\boldsymbol{X}$ and $\boldsymbol{d}$ (i.e., $p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{d})$) to derive equation (a), and use Bayes' theorem to derive equation (b).

Since there is no closed-form solution for $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{d})$ and $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{d})$, obtaining the exact computation of the EIG is intractable. Therefore, we approximate the value using the nested Monte Carlo method:

$$\mathrm{EIG}(\boldsymbol{d}) \approx \widehat{\mathrm{EIG}}(\boldsymbol{d}) = \frac{1}{I}\sum_{i=1}^{I} \log \frac{p(\boldsymbol{y}^i|\boldsymbol{X}, \boldsymbol{d}, \boldsymbol{\theta}^{i,0})}{\frac{1}{J}\sum_{j=1}^{J} p(\boldsymbol{y}^i|\boldsymbol{X}, \boldsymbol{d}, \boldsymbol{\theta}^{i,j})},$$

where $\boldsymbol{\theta}^{i,0} \sim p(\boldsymbol{\theta})$, $\boldsymbol{y}^i \sim p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta} = \boldsymbol{\theta}^{i,0}, \boldsymbol{d})$, $\boldsymbol{\theta}^{i,j} \sim p(\boldsymbol{\theta})$. The outer sampling estimates $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{d})$ and the inner sampling estimates $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{d})$. Parameters I and J denote the number of samples of the outer and inner loops.

B Updating $\boldsymbol{\phi}^\tau$ in Iterative Optimization

We introduce the optimization of $\boldsymbol{\phi}^\tau$ in the iterative optimization of our **I-ID-LP** method.

$$\begin{aligned}
\boldsymbol{\phi}^\tau &= \arg\max_{\boldsymbol{\phi}} \int q(\boldsymbol{\theta}|\boldsymbol{\phi}) \log \frac{p(\boldsymbol{\theta}|\mathcal{H}_{t-1}) g(\boldsymbol{\theta}, \boldsymbol{d}_t^{\tau-1})}{q(\boldsymbol{\theta}|\boldsymbol{\phi})} d\boldsymbol{\theta} \\
&= \arg\min_{\boldsymbol{\phi}} \int q(\boldsymbol{\theta}|\boldsymbol{\phi}) \log \frac{q(\boldsymbol{\theta}|\boldsymbol{\phi})}{p(\boldsymbol{\theta}|\mathcal{H}_{t-1}) g(\boldsymbol{\theta}, \boldsymbol{d}_t^{\tau-1})} d\boldsymbol{\theta} \\
&\overset{(a)}{=} \arg\min_{\boldsymbol{\phi}} \int q(\boldsymbol{\theta}|\boldsymbol{\phi}) \log \frac{q(\boldsymbol{\theta}|\boldsymbol{\phi}) p(\mathcal{H}_{t-1})}{p(\mathcal{H}_{t-1}|\boldsymbol{\theta}) p(\boldsymbol{\theta})} d\boldsymbol{\theta} - \int q(\boldsymbol{\theta}|\boldsymbol{\phi}) \log g(\boldsymbol{\theta}, \boldsymbol{d}_t^{\tau-1}) d\boldsymbol{\theta} \\
&\overset{(b)}{=} \arg\min_{\boldsymbol{\phi}} \int q(\boldsymbol{\theta}|\boldsymbol{\phi}) \log \frac{q(\boldsymbol{\theta}|\boldsymbol{\phi})}{p(\boldsymbol{\theta})} d\boldsymbol{\theta} - \int q(\boldsymbol{\theta}|\boldsymbol{\phi}) \log p(\mathcal{H}_{t-1}|\boldsymbol{\theta}) d\boldsymbol{\theta} \\
&\qquad - \int q(\boldsymbol{\theta}|\boldsymbol{\phi}) \log g(\boldsymbol{\theta}, \boldsymbol{d}_t^{\tau-1}) d\boldsymbol{\theta} \\
&= \arg\min_{\boldsymbol{\phi}} \mathrm{KL}[q(\boldsymbol{\theta}|\boldsymbol{\phi}) || p(\boldsymbol{\theta})] - \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\phi})}[\log p(\mathcal{H}_{t-1}|\boldsymbol{\theta})] - \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\phi})}[\log g(\boldsymbol{\theta}, \boldsymbol{d}_t^{\tau-1})].
\end{aligned}$$

We use the fact that $p(\boldsymbol{\theta}|\mathcal{H}_{t-1}) = \frac{p(\mathcal{H}_{t-1}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{H}_{t-1})}$ to derive equation (a), and the fact that $p(\mathcal{H}_{t-1})$ is independent of $\boldsymbol{\theta}$ to derive equation (b).

C Data Settings

We introduce the remaining four data settings.

Setting B: *We use a three-layer neural network to generate data. Specifically, we compute*

$$\big(\bar{\lambda}(\boldsymbol{x}); \bar{r}(\boldsymbol{x})\big) = \mathrm{Sigmoid}(\boldsymbol{w}_3^T \mathrm{Sigmoid}(\boldsymbol{w}_2^T \mathrm{Sigmoid}(\boldsymbol{w}_1^T \boldsymbol{x} + \boldsymbol{b}_1) + \boldsymbol{b}_2) + \boldsymbol{b}_3),$$

where $\boldsymbol{w}_1 \in \mathbb{R}^{16\times 6}$, $\boldsymbol{w}_2 \in \mathbb{R}^{6\times 4}$, $\boldsymbol{w}_3 \in \mathbb{R}^{4\times 2}$ and $\boldsymbol{b}_1 \in \mathbb{R}^6$, $\boldsymbol{b}_2 \in \mathbb{R}^4$, $\boldsymbol{b}_3 \in \mathbb{R}^2$ are randomly chosen. Then, we normalize $\bar{r}(\boldsymbol{x}), \bar{\lambda}(\boldsymbol{x})$ across all data to get $r(\boldsymbol{x}), \lambda(\boldsymbol{x})$.

Setting C: *We use a two-layer network to generate data. We compute*

$$\big(\bar{\lambda}(\boldsymbol{x}); \bar{r}(\boldsymbol{x})\big) = \text{Sigmoid}(\boldsymbol{w}_2^T(\text{Sigmoid}(\boldsymbol{w}_1^T\boldsymbol{x}) + \boldsymbol{b}_1) + \boldsymbol{b}_2 + \boldsymbol{\epsilon}),$$

where $\boldsymbol{w}_1 \in \mathbb{R}^{16\times 16}$, $\boldsymbol{w}_2 \in \mathbb{R}^{16\times 2}$ and $\boldsymbol{b}_1 \in \mathbb{R}^{16}$, $\boldsymbol{b}_2 \in \mathbb{R}^2$ are randomly chosen. $\boldsymbol{\epsilon} \in \mathbb{R}^2$ represents the noise vector. We normalize $\bar{r}(\boldsymbol{x}), \bar{\lambda}(\boldsymbol{x})$ across all data to get $r(\boldsymbol{x}), \lambda(\boldsymbol{x})$.

Setting D: *Setting D employs the same network structure as Setting C to generate data. The difference is that setting D normalizes the values of* $\bar{\lambda}(\boldsymbol{x})$ *to a smaller range to get* $\lambda(\boldsymbol{x})$*. Table 1 in the paper shows that our methods can more accurately predict investor behavior, compared with Setting C.*

Setting E: *We generate data according to*

$$\big(\bar{\lambda}(\boldsymbol{x}); \bar{r}(\boldsymbol{x})\big) = \text{Sigmoid}(\boldsymbol{w}_2^T(\text{Sigmoid}(\boldsymbol{w}_1^T\boldsymbol{x}) + \boldsymbol{b}_1) + \boldsymbol{b}_2),$$

where $\boldsymbol{w}_1 \in \mathbb{R}^{16\times 16}$, $\boldsymbol{w}_2 \in \mathbb{R}^{16\times 2}$ and $\boldsymbol{b}_1 \in \mathbb{R}^{16}$, $\boldsymbol{b}_2 \in \mathbb{R}^2$ are randomly chosen. Then, we normalize $\bar{r}(\boldsymbol{x}), \bar{\lambda}(\boldsymbol{x})$ across all data to get $r(\boldsymbol{x}), \lambda(\boldsymbol{x})$. Compared with Setting D, Setting E does not include any noise in the data generation.

D Comparison of ID Methods

We show the comparison of ID methods under data settings C∼E as follows.

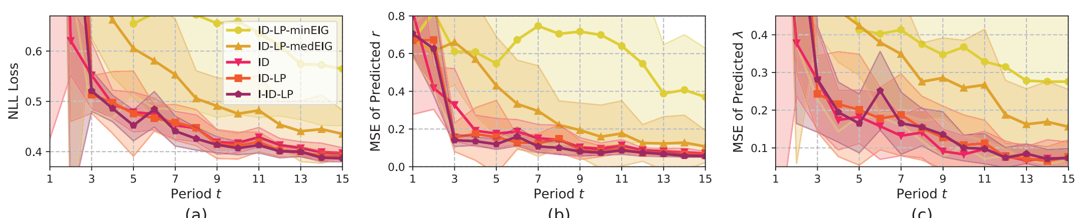

Figure 1: Comparison of ID Methods Under Setting C.

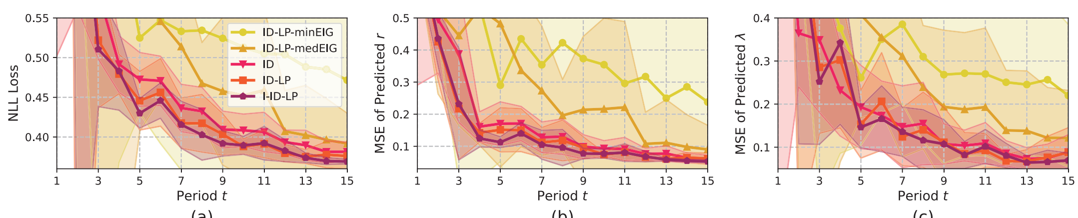

Figure 2: Comparison of ID Methods Under Setting D.

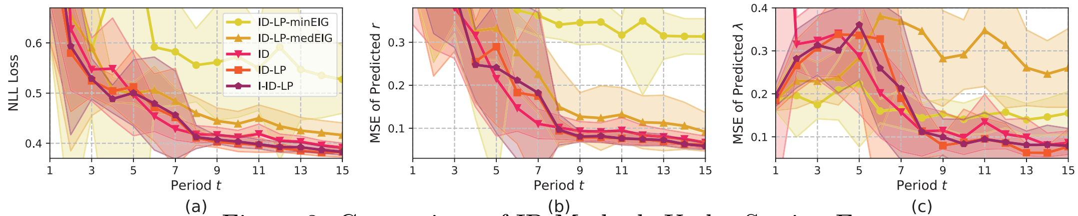

Figure 3: Comparison of ID Methods Under Setting E.

E Mean Squared Errors of Predicted r and λ

We evaluate the performance of predicting r and λ based on the mean squared errors (MSEs). We show the MSEs achieved by different methods under settings A$\sim$E as follows.

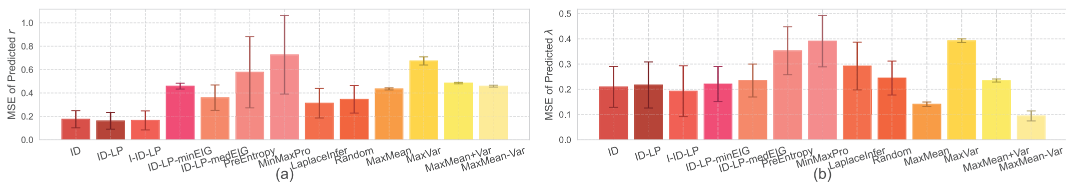

Figure 4: Mean Squared Errors of Predicted r and λ Under Setting A.

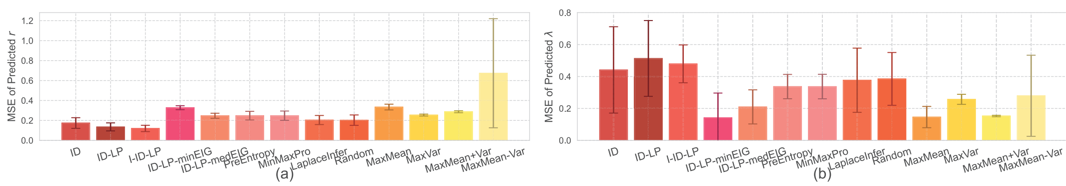

Figure 5: Mean Squared Errors of Predicted r and λ Under Setting B.

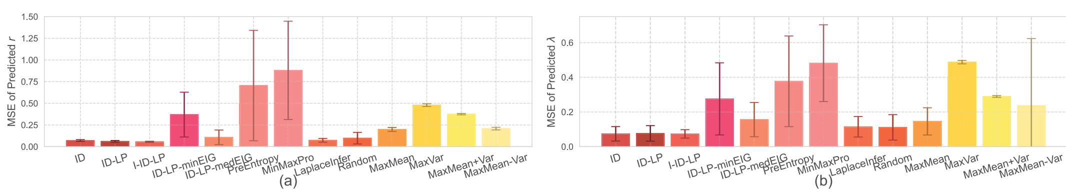

Figure 6: Mean Squared Errors of Predicted r and λ Under Setting C.

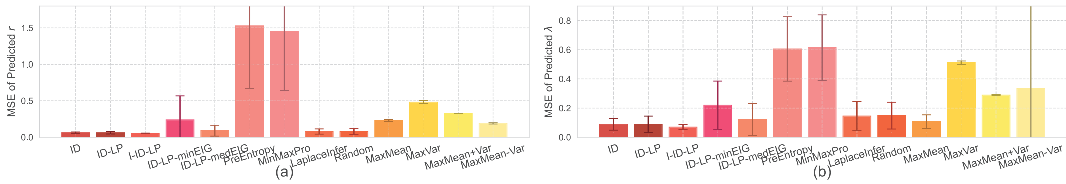

Figure 7: Mean Squared Errors of Predicted r and λ Under Setting D.

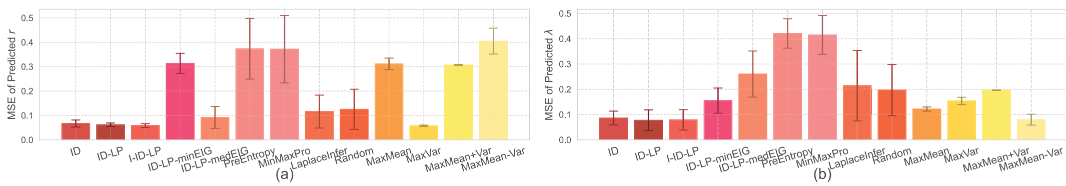

Figure 8: Mean Squared Errors of Predicted r and λ Under Setting E.