ORF 418 Final Project: Online Learning in the South Korean Idol Industry

Peter Chen
Princeton University
xi@princeton.edu

Tor Nitayanont
Princeton University
torpongn@princeton.edu

1 Narrative

The Korean popular music industry, K-pop for short, is characterized by major entertainment agencies that manage all aspects of their signed artists' professional careers. Known as "idols," these musical artists receive extensive training for several years in singing, rapping, dancing, acting and/or languages before debuting, usually in a group. In our problem, we are the company, and our goal is to maximize profits out of an idol. In 2009, South Korea's Fair Trade Commission imposed a limit of seven years to entertainment contracts. As a result, a major learning problem that faces each entertainment agency is the optimal use of their artists' schedules to maximize profits over the course of the contract period. We simplify the problem by generalizing four categories of activities for each time period of 3 months: producing an album, touring, acting, or recording variety shows.

2 Mathematical Model

Belief Model

We will be using a hybrid model consisting of a lookup table for the four categories of activities, with each category then having its own separate parametric function. The lookup table distribution will be normally distributed around the maximum of the activities within each category. Our decision variable C^n will be the category for our lookup table and x^n the specific activity within a chosen category for our parametric, as follows:

Album	**Touring**	**Acting**	**Variety**
Rap/R&B	Korea	Drama	Competition
Ballad	Japan	Comedy	Travel
Dance	International	History	Reality

At the end of each time period, we observe the profit W^{n+1}. For each observation of profit for an activity within a given category:

1. The parametric belief model for the given category will be updated.
2. The lookup table belief model will be updated.
3. According to this updated model, a category will be chosen using lookup table policies.
4. Within the category, a specific activity will be chosen using parametric model policies.

This process is repeated for each observation until the end of our time horizon. At the end of each time period of 3 months, we observe our profits, then we choose the next category/activity that would maximize our cumulative reward at the end of the seven year contract period. As a result, this is an **online learning** problem.

2.1 State Variables

State variables S^n describe our belief and contain all information needed to model the system forward in time after n experiments. We are using a hybrid belief model combining lookup tables (for selecting categories) and the parametric model (for selecting the specific activity within a category).

$$S^n = (S^{LT,n}, S^{Par,n})$$

where $S^{LT,n}$ denotes the state variable or belief of the lookup table and $S^{Par,n}$ denotes the state variable of the parametric model.

In the lookup table, $S^{LT,n}$ includes uncorrelated beliefs of different categories, which are distributed normally centered on the maximum of its activities, given by the parametric model.

$$S^{LT,n} = (\bar{\mu}_i^n, \bar{\sigma}_i^n)_{i\in\{1,2,3,4\}}$$

where $(\bar{\mu}_i^n, \bar{\sigma}_i^n)$ are our estimated expected revenue and its standard deviation, given that we choose an activity from category C_i, after n observations.

The state variable within each category involves correlated parametric belief models. This parametric model is the model of profit as a function of related features. The model of each activity in category C_i, for $i \in \{1, 2, 3, 4\}$, involves K features. We let $\phi_{i,k}$ be the k-th feature of an activity in the i-th category where $i \in \{1, 2, 3, 4\}$ and $k \in \{1, 2, ..., K\}$, and let $\theta_{i,k}$ be its coeffcient in the parametric model and $\theta_{i,0}$ is the intercept. Therefore, for each activity x in the category C_i, we can write its parametric model as

$$\mu_x = \theta_{i,0} + \sum_{k=1}^{K} \theta_{i,k}\phi_{i,k}(x)$$

$$S^{Par,n} = (\bar{\theta}_{i,0}^n, \bar{\theta}_{i,1}^n, ..., \bar{\theta}_{i,K}^n, \Sigma_i^n)_{i\in\{1,2,3,4\}}$$

($\phi_{i,k}(x)$ might be used interchangeably with $x_{i,k}$ later in this report.) where $\bar{\theta}_{i,k}^n$ is our belief about the coefficient of the k-th attribute of activities in category C_i and Σ^n is our belief covariance matrix of the coefficients in category C_i after n observations. Below is the description of attributes in each category.

For albums, its profit μ_1 will be determined by the reputation of its producer $\theta_{1,1}$, lyricist $\theta_{1,2}$, and choreographer $\theta_{1,3}$. For tours, its profit μ_2 will be determined by the artist popularity $\theta_{2,1}$, quality of promoter $\theta_{2,2}$, and quality of venue $\theta_{2,3}$. For acting, its profits μ_3 will be determined by the reputation of the director $\theta_{3,1}$, cast $\theta_{3,2}$, and writer $\theta_{3,3}$. For variety, its profits μ_4 will simliarly be determined by the reputation of the director $\theta_{4,1}$ cast $\theta_{4,2}$, and writer $\theta_{4,3}$. Each reputation/popularity measure $x_{i,k}$ will range from 1-10. Our parametric model will be as follows:

$$\mu_1 = \theta_{1,0} + \theta_{1,1}x_{1,1} + \theta_{1,2}x_{1,2} + \theta_{1,3}x_{1,3}$$
$$\mu_2 = \theta_{2,0} + \theta_{2,1}x_{2,1} + \theta_{2,2}x_{2,2} + \theta_{2,3}x_{2,3}$$
$$\mu_3 = \theta_{3,0} + \theta_{3,1}x_{3,1} + \theta_{3,2}x_{3,2} + \theta_{3,3}x_{3,3}$$
$$\mu_4 = \theta_{4,0} + \theta_{4,1}x_{4,1} + \theta_{4,2}x_{4,2} + \theta_{4,3}x_{4,3}$$

2.1.1 Prior Construction

At the beginning, we have the prior belief $(\bar{\mu}_i^0, \bar{\sigma}_i^0)$ in our lookup table and the prior belief $(\bar{\theta}_{i,0}^0, \bar{\theta}_{i,1}^0, ..., \bar{\theta}_{i,K}^0, \Sigma^0)$ for each category C_i. Note that $\phi_{i,k}(x)$ or $x_{i,k}$ is a fixed value for each i, k and x. To construct the prior beliefs, we begin with the parametric model within each category.

In our experiment, each activity has 3 features ($K = 3$). Each features takes value as an integer from 1 to 10. We assume that they follow discrete uniform distribution. Therefore, there are $10^3 = 1000$ possible combinations of features for each activity in each category, with equal probability. These 1000 combinations include $(2, 8, 3)$, $(1, 5, 7)$ and $(6, 3, 1)$, for instance.

For each activity, we randomly sample 250 combinations/permutations out of 1000. We let this 250-by-4 matrix be X_L, where the first column is all 1 and the last 3 columns are the attributes combinations that we have randomly chosen.

Since we have the true θ's, which remain hidden during the experiment, but we, as researchers, secretly know them, we can generate profits corresponding to each combination of features by using the true parametric model equations, and add some uncertainty $\epsilon \sim \mathcal{N}(0, \sigma_w^2)$ to them. σ_w captures the noise in our ability to observe the true value. We denote these revenues as Y_L, which is a vector of 250 revenues. Now that we have obtained the corresponding revenues from all 250 permutations, we regress them back on their 250 combinations of features to obtain our prior belief $\bar{\theta}^0$.

$$\bar{\theta}^0 = [X_L^T X_L]^{-1} X_L^T Y_L$$

We can compute the standard deviation of the revenues Y_L to be σ_ϵ. From this, we can construct the covariance matrix, following the equation (7.9) of the textbook.

$$\Sigma_i^0 = [X_L^T X_L]^{-1} \sigma_\epsilon^2$$

Therefore we have obtained $(\bar{\theta}_{i,0}^0, \bar{\theta}_{i,1}^0, \bar{\theta}_{i,2}^0, \bar{\theta}_{i,3}^0, \Sigma_i^0)$, which are our belief model or state variable of our parametric model at the beginning, or after 0 experiment (before the first experiment).

Now that we have constructed the prior belief of the parametric model, we can compute the highest expected revenue of each activity in each category. For category C_i, we take the highest expected profit among the 3 choices to be $\bar{\mu}_i^0$. $\bar{\sigma}_i^0$ can be easily computed from Y_L that we obtained earlier for each category C_i, whic is basically σ_ϵ. Therefore, we have $(\bar{\mu}_i^0, \bar{\sigma}_i^0)$, which is the state variable or belief model of our lookup table at the beginning.

2.2 Decision Variables

At the n-th experiment, we have to make 2 decisions: which category to choose and which activity within the chosen category to do.

Assume that we choose the category C^n with policy π_1 and choose the activity x^n from the set of choices $\{x_{C^n,1}, x_{C^n,2}, x_{C^n,3}\}$ within the category C^n with policy π_2. Then, we could write our decisions in terms of the updated state variables after the n-th experiment as

$$C^n = X^{\pi_1}(S^{LT,n});\ C^n \in \{1, 2, 3, 4\}$$

and

$$x^n = X^{\pi_2}(S_{C^n}^{Par,n});\ x^n \in \{x_{C^n,1}, x_{C^n,2}, x_{C^n,3}\}$$

2.3 New Information

After we choose the action x^n from the category C^n, we observe the profit W^{n+1} that the action x^n yields.

$$W^{n+1} = \theta^T X^n + \epsilon^n$$

where θ is the coefficient vector of our choice x^n. ϵ^n is the uncertainty, which follows $\mathcal{N}(0, \sigma_w^2)$

2.4 Transition function

Our transition functions are Bayesian updating equations of the belief models. This could be written as the following:

$$\begin{aligned} S^{n+1} &= (S^{LT,n+1}, S^{Par,n+1}) \\ &= S^M(S^n, x^n, C^n, W^{n+1}) \\ &= S^M(S^{LT,n}, S^{Par,n}, x^n, C^n, W^{n+1}) \end{aligned}$$

We will use Bayesian's updating equations for correlated beliefs to update our belief of the lookup table after each experiment.

$$\bar{\mu}^{n+1}(x) = \bar{\mu}^n + \frac{W^{n+1} - \bar{\mu}_x^n}{\lambda^W + \Sigma_{xx}^n}\Sigma^n e_x$$

$$\Sigma^{n+1}(x) = \Sigma^n - \frac{\Sigma^n e_x (e_x)^T \Sigma^n}{\lambda^W + \Sigma_{xx}^n}$$

where $x \in \{C_1, C_2, C_3, C_4\}$

However, since the beliefs in our lookup table are uncorrelated, Σ^n is basically a diagonal matrix. Σ_{xx}^n when $x = C_i$ refers to the i-th entry on the diagonal of Σ^n.

Regarding the parametric model,if the activity tested in the recent experiment is in the category C_i , then we update $\bar{\theta}_i^n$ and Σ_i^n as follows:

$$\bar{\theta}^{n+1} = \bar{\theta}^n + \frac{1}{\gamma^{n+1}\sigma_\epsilon^2}\Sigma_i^n x^{n+1}\epsilon^{n+1}$$

$$\epsilon^{n+1} = W^{n+1} - \bar{\theta}^n x^n$$

$$\Sigma_i^{n+1} = \Sigma_i^n - \frac{1}{\gamma^{n+1}}(\Sigma_i^n x^{n+1}(x^{n+1})^T \Sigma_i^n)$$

$$\gamma^{n+1} = \sigma_\epsilon^2 + (x^{n+1})^T \Sigma_i^n x^{n+1}$$

(These updating equations are from $(7.3) - (7.6)$ of the textbook.)

2.5 Objective Function

2.5.1 Performance Metric

Our performance metric of each experiment is the reward or profit of our choice.

$$\mu_{X^{\pi_1,\pi_2}(S^n)} = C(S^{LT,n}, S^{Par,n}, X^{\pi_1}(S^{LT,n}), X^{\pi_2}(S^{Par,n}))$$

2.6 Cumulative Reward

Our performance metric over the time horizon is the cumulative reward $\mu_{X^\pi(S^n)}$ observed at the end of each 3 month time period. We are interested in cumulative profits at the end of our seven year time horizon.

$$\max_{\pi_1,\pi_2} \mathbb{E}_\mu \mathbb{E}_{W^1,\ldots,W^N|\mu} \sum_{n=0}^{N-1} \mu_{X^{\pi_1,\pi_2}(S^n)}$$

3 Policies

3.1 Policy search

We will be using **pure exploitation**:

$$X^{IE}(S^n|\theta^{IE}) = \underset{x}{\text{argmax}}(\bar{\mu}_x^n)$$

We will also be using **interval estimation**:

$$X^{IE}(S^n|\theta^{IE}) = \underset{x}{\text{argmax}}(\bar{\mu}_x^n + \theta^{IE}\bar{\sigma}_x^n)$$

where θ^{IE} is a tunable parameter. It has been reported in literature that values around 2 or 3 work the best, but we can try anything from 0.1 up to 100.

3.2 Lookahead policy

We will be using the **knowledge gradient policy**:

$$X_x^{KG,n} = \underset{x}{\operatorname{argmax}}(\mathbb{E}\{\max_{x' \in \mathcal{X}} \bar{\mu}_{x'}^{n+1}(x)|S^n\} - V^n(S^n))$$

When we use this policy to choose an activity, we rely on independent beliefs. We will use correlated beliefs only when updating.

3.2.1 Computation of Knowledge Gradient

The computation of knowledge gradient could be slightly complicated, especially when we implement the policy in two different models: the lookup table, which contains uncorrelated belief, and the parametric model, which has correlated belief. We implement the knowledge gradient policy on both models, as explained in the section 3.3.

When implementing the knowledge gradient policy on the lookup table with uncorrelated belief, we compute the knowledge gradient of each choice

$$\nu_x^{KG,n} = \tilde{\sigma}_x^n f(\xi_x^n)$$

where $f(\xi) = \xi\Phi(\xi) + \phi(\xi)$. The functions $\Phi(\xi)$ and $\phi(\xi)$ are the cumulative standard normal distribution and the standard normal density, respectively. We compute ξ_x^n by:

$$\xi_x^n = -\left|\left|\frac{\bar{\mu_x^n} - max_{x' \neq x}\bar{\mu}_x^n}{\tilde{\sigma}_x^n}\right|\right|$$

and

$$\tilde{\sigma}_x^n = \frac{\bar{\sigma}_x^{2,n}}{1 + \frac{\sigma_W^2}{\bar{\sigma}_x^{2,n}}}$$

When we implement the knowledge gradient policy on the parametric model with correlated belief, we have already chosen the category from the lookup table, from which we choose an activity. In that chosen category C_i, we have our belief of coefficient $\bar{\theta}^n = \{\bar{\theta}_{i,0}^n, \bar{\theta}_{i,1}^n, \bar{\theta}_{i,2}^n, \bar{\theta}_{i,3}^n\}$ According to the knowledge gradient policy,

$$X^{KG}(S^n) = \underset{x}{\operatorname{argmax}}\, h(a, b(x))$$

To make the equations easier to understand, we rewrite this as

$$X^{KG}(S^n) = \underset{k}{\operatorname{argmax}}\, h(a, b(k))$$

where k is the index of x of the previous equation, in our chosen category.

$$h(a, b(k)) = \sum_{i=1}^{M-1} (b_{i+1}(k) - b_i(k)) f(-|c_i(k)|)$$

and

$$c_i(k) = \frac{a_i - a_{i+1}}{b_{i+1}(k) - b_i(k)}$$

In our parametric model, a_i refers to $\bar{\theta}^n x_i$ where x_i is a choice (or more precisely, its attributes) in the chosen category, and $b_i(k)$ is the i-th entry of $\tilde{\sigma}^n(k) = X_k \Sigma^n X^T$ where X is the matrix of attributed of choices in te category, and X_k is the k-th row of X (an extension of the equation (7.11) from the textbook).

3.3 Implementation

Because we are using a hybrid belief model consisting of an outer lookup table and inner parametric model, we test different combinations of policies to determine the best one. The policiy combinations that we are testing are:

- Pure exploitation lookup table, pure exploitation parametric
- Interval estimation lookup table, interval estimation parametric
- Knowledge gradient lookup table, knowledge gradient parametric
- Pure exploitation lookup table, interval estimation parametric
- Pure exploitation lookup table, knowledge gradient parametric
- Interval estimation lookup table, pure exploitation parametric
- Interval estimation lookup table, knowledge gradient parametric
- Knowledge gradient lookup table, pure exploitation parametric
- Knowledge gradient lookup table, interval estimation parametric

We will be tuning the parameter θ^{IE} for each individual case (if applicable).

3.4 Tuning parameters

We need to tune the parameter θ^{IE} of the interval estimation policy. Best value may be anywhere from 0.1 to 100, with 2 or 3 working best for many applications and high values arising when the priors are really poor. However, we try θ^{IE} from as low as 10^{-4} up to 20. The values that we experiment with are $10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 5, 10, 20$. We choose the one that gives the highest cumulative reward.

4 Numeric Testing

For each of the 500 priors, we run 100 experiments. We report the mean and variance values of our cumulative rewards below:

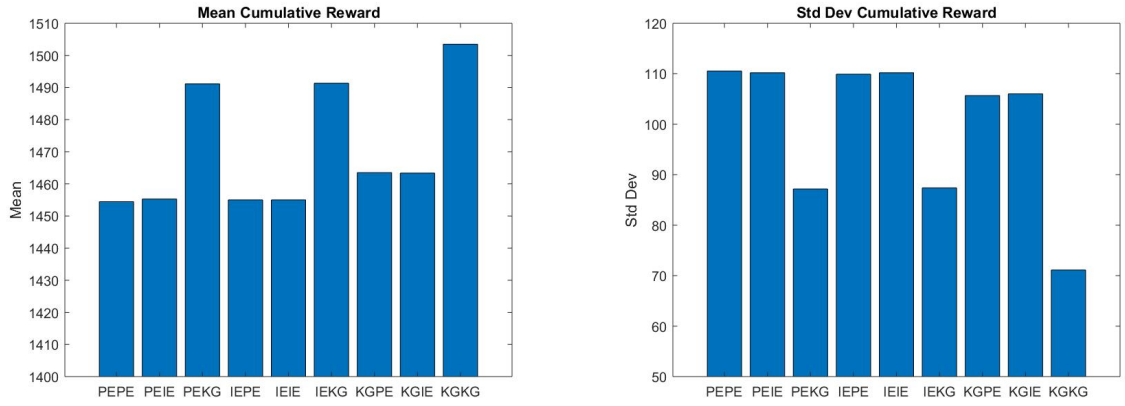
Lookup Table ↓ Parametric →	**Pure Exploitation**	**Interval Estimation**	**Knowledge Gradient**
Pure Exploitation	1454.4585 (110.5554)	1455.2850 (110.2070)	1491.1718 (87.1835)
Interval Estimation	1455.0185 (109.9202)	1455.0231 (110.2073)	1491.3486 (87.3944)
Knowledge Gradient	1463.4993 (105.6865)	1463.3635 (106.0316)	1503.5122 (71.1252)

Table 1: Average (standard deviation) of cumulative reward over 50000 samples (500 priors, each with 100 experiments)

Histograms showing distributions (Figure 3) is included at the end of report.

The rewards from the same policy implemented on the lookup table go to the same row whereas the rewards from the same policy implemented on the parametric model go to the same column. Results of the interval estimation policy are from our tuned θ^{IE} i.e. they are the highest numbers among all numbers from different θ^{IE}'s when tested with the same combination of policies.

We might notice that the highest rewards that interval estimation yields are just as high as those of pure exploitation. Compared to interval estimation and pure exploitation, knowledge gradient performs better. We can see that the result is highest when we use knowledge gradient policy on both the lookup table and the parametric model.



(a) (b)

Figure 1: Graphical comparison of cumulative reward mean (a) and std dev (b) across policies

The knowledge gradient policy does not only yield a high average cumulative reward but also results in a lower standard deviation, which implies a lower risk.

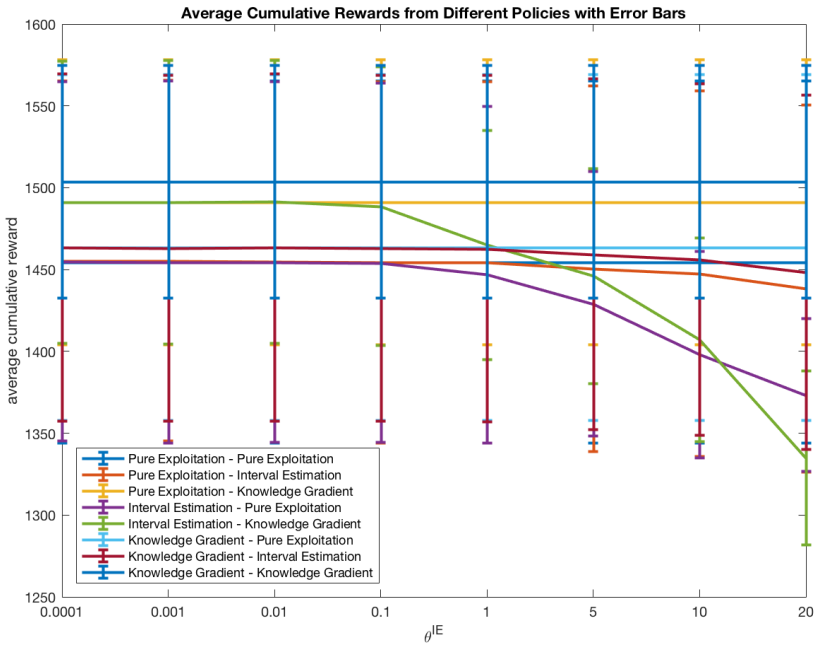


Figure 2: Comparison of cumulative rewards across different θ^{IE}

From Figure 2, the horizontal lines are the mean cumulative rewards of the policies that do not vary over different θ^{IE} such as pure exploitation-pure exploitation, pure exploitation-knowledge gradient, knowledge gradient-pure exploitation, knowledge gradient-knowledge gradient. The vertical bars are the error bars. Half of the widths of the bars are equal to the standard deviations of their results.

θ^{IE}	PE IE	IE PE	IE KG	KG IE
0.0001	1455.1944 (110.0942)	1455.0185 (109.9202)	1491.0180 (86.1583)	1463.3635 (106.0316)
0.001	1455.2850 (110.2070)	1454.4908 (110.6617)	1491.2028 (86.5896)	1463.0855 (105.7557)
0.01	1454.7057 (110.0389)	1454.5097 (110.0756)	1491.3486 (86.5907)	1463.3529 (105.9970)
0.1	1454.3846 (110.6015)	1454.1379 (109.8508)	1488.6300 (85.2062)	1462.9885 (105.6170)
1	1454.4356 (110.3184)	1446.8372 (102.6875)	1465.1617 (70.0066)	1462.7064 (105.8672)
5	1450.4836 (111.4905)	1428.9502 (80.7871)	1446.0340 (65.6678)	1459.3111 (107.2636)
10	1447.5186 (111.5519)	1398.0276 (62.9572)	1407.0924 (62.4000)	1456.0212 (107.3982)
20	1438.3869 (112.2980)	1373.2607 (46.7215)	1334.8264 (53.1559)	1448.3906 (108.1118)

Table 2: Mean (with standard deviation in the parentheses) cumulative returns for each tuned θ^{IE}

Lookup Table $\downarrow$ Parametric $\rightarrow$	**Pure Exploitation**	**Interval Estimation**	**Knowledge Gradient**
Pure Exploitation	NA	.001	NA
Interval Estimation	.0001	.0001, .0001	.01
Knowledge Gradient	NA	.0001	NA

Table 3: Tuned θ^{IE}

From the plot in Figure 2, it is clear that the interval estimation policy does not outperform pure exploitation policy. We can see that it only converges to the performance level of pure exploitation when θ^{IE} decreases. Nevertheless, the horizontal line at the top shows the highest cumulative reward, which is from the knowledge gradient-knowledge gradient policy.

From Table 2 and Table 3, we might see that even though the values of the best θ^{IE} might vary slightly between $0.0001, 0.001, 0.01$, when considering one particular policy combination, the rewards that these three values of θ^{IE} are about the same since, at that point, the performance of interval estimation already converges to that of pure exploitation.

This is also true when we look at the performances of the policy where we apply interval estimation on both lookup table and parametric model(or the interva; estimation-interval estimation policy). The rewards are shown in the table below.

Parametric $\rightarrow$ Lookup Table $\downarrow$	**0.0001**	**0.001**	**0.01**	**0.1**	**1**	**5**
0.0001	1455.0231 (110.2073)	1454.3691 (110.3197)	1454.9905 (111.2217)	1454.5528 (110.2825)	1454.2103 (110.4848)	1450.9644 (111.3094)
0.001	1455.2873 (110.0830)	1453.8843 (110.0900)	1455.2645 (110.0333)	1454.9141 (110.2297)	1454.3196 (110.4960)	1450.4274 (111.7191)
0.01	1454.8328 (110.4060)	1454.7104 (109.9761)	1454.7204 (109.5976)	1454.1003 (110.4977)	1454.1651 (110.5466)	1450.8295 (110.9958)
0.1	1454.3452 (110.8702)	1454.2572 (109.8935)	1454.2677 (109.8768)	1453.9479 (109.8129)	1453.2707 (109.7138)	1450.2111 (110.5081)
1	1446.8095 (102.8702)	1446.8645 (102.3117)	1446.6496 (102.0840)	1447.0391 (102.1367)	1445.9597 (101.9520)	1443.7618 (103.5687)
5	1428.7350 (80.8279)	1429.0107 (79.1278)	1428.6745 (78.9020)	1428.5173 (79.2659)	1428.4427 (79.9380)	1424.4726 (81.9285)

Table 4: Average (with standard deviation in the parentheses) cumulative returns for each tuned θ^{IE} using interval estimation for both lookup table and parametric

From the plot of distribution of each policy combination (in Figure 3 at the end of the report), we might notice that most of the distributions actually consist of two sub-distributions, one with a centered at a lower cumulative reward, one at a higher cumulative reward. This is a bad signal since it means that there is a considerable chance that, at any experiment, our result might fall into

the lower sub-distribution and we receive a low reward. However, the results are different for the policy combinations where the knowledge gradient policy is implemented on the parametric model. In those 3 cases, there seems to be just one left-skewed distribution with a high mean. This means that there is less risk when using the knowledge gradient policy on the parametric model. (These is no lower distribution that our result might fall into.) The knowledge gradient policy does not only yield a high average cumulative reward but also results in a lower standard deviation, which implies a lower risk.

5 Extensions

As an extension, we consider a true model that does not have the same structure as our belief model. We do this by retaining the same belief model while adding an additional parameter (a latent variable we fail to account for in our belief) to the true model as follows:

$$\mu_1 = \theta_{1,0} + \theta_{1,1}x_{1,1} + \theta_{1,2}x_{1,2} + \theta_{1,3}x_{1,3} + \theta_{1,4}x_{1,4}$$
$$\mu_2 = \theta_{2,0} + \theta_{2,1}x_{2,1} + \theta_{2,2}x_{2,2} + \theta_{2,3}x_{2,3} + \theta_{2,4}x_{2,4}$$
$$\mu_3 = \theta_{3,0} + \theta_{3,1}x_{3,1} + \theta_{3,2}x_{3,2} + \theta_{3,3}x_{3,3} + \theta_{3,4}x_{3,4}$$
$$\mu_4 = \theta_{4,0} + \theta_{4,1}x_{4,1} + \theta_{4,2}x_{4,2} + \theta_{4,3}x_{4,3} + \theta_{4,4}x_{4,4}$$

We obtain the following results:

Lookup Table ↓ Parametric →	**Pure Exploitation**	**Knowledge Gradient**
Pure Exploitation	1575.5802 (28.1014)	1549.9680 (42.5592)
Knowledge Gradient	1587.4215 (20.9610)	1539.6378 (62.1170)

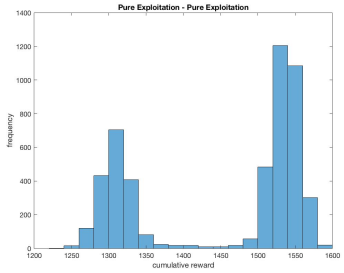
Table 5: Average (with standard deviation in the parentheses) cumulative reward

Histograms showing distributions (figure 4) is included at the end of report.

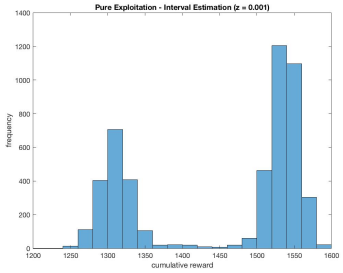
It appears that pure exploitation in the parametric (where we have an additional variable) works better than knowledge gradient with higher cumulative returns and lower standard deviations. We also observe one normal distribution (as opposed to two normal sub-distributions that we observe from our original model) in these cases. For further study, we could investigate why pure exploitation works better when the truth involves hidden features or when we are uncertain whether our belief model captures enough information about our choices.

6 Conclusion

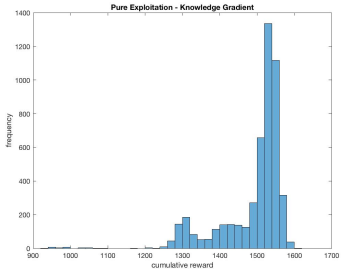
From our results, we conclude that policies using knowledge gradient perform best. The knowledge gradient policy does not only yield a high average cumulative reward but also a lower standard deviation. We also conclude that the interval estimation policy does not outperform pure exploitation policy. This is evident as our tuned θ^{IE} tend to converge to zero, in which case the policy becomes pure exploitation. One assumption is that interval estimation encourages exploration of the options that we are uncertain about. In the online learning setting, it might hurt us when we observe a poor choice since we are collecting cumulative reward. Looking at the distributions of cumulative reward, we notice that most distributions consist of two Gaussian distributions, one centered at a lower cumulative reward, one at a higher cumulative reward (and any given experiment may fall into either distribution). We notice, however, that implementing knowledge gradient policy on the parametric model produces one left-skewed distributions with a high mean and no lower distribution, indicating smaller risk of low reward. Knowledge gradient policies in both components of our hybrid model produce the best results and is what we would recommend. This is the intuitive result we expected because the knowledge gradient policy is both myopically optimal and asymptotically optimal. Since we were not sure in the beginning whether our budget N is large enough for our model to collect information and learn fast enough, using knowledge gradient would guarantee that in either case, we will obtain a sufficiently good result. However, when our belief model fails to capture the truth, we find that pure exploitation may perform better (only in the parametric case).



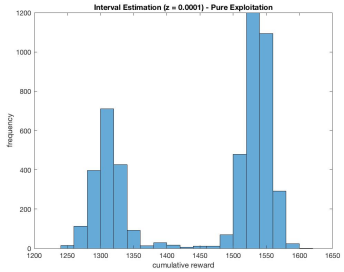
(a) Lookup Table: Pure Exploitation, Parametric: Pure Exploitation



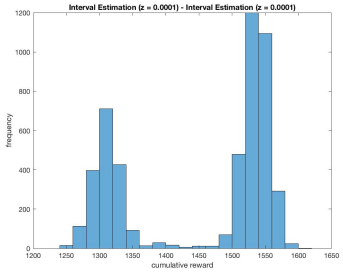
(b) Lookup Table: Pure Exploitation, Parametric: Interval Estimation



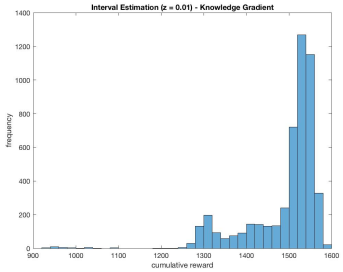
(c) Lookup Table: Pure Exploitation, Parametric: Knowledge Gradient



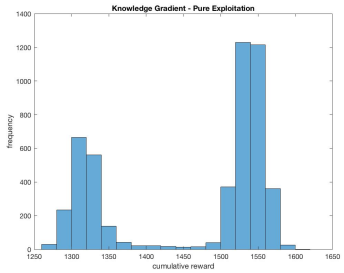
(d) Lookup Table: Interval Estimation, Parametric: Pure Exploitation



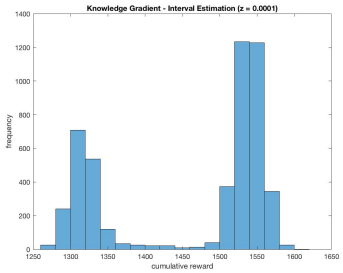
(e) Lookup Table: Interval Estimation, Parametric: Interval Estimation



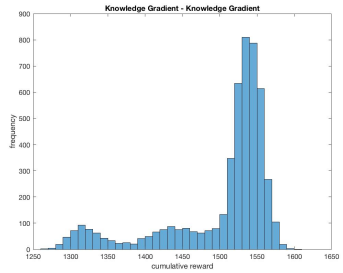
(f) Lookup Table: Interval Estimation, Parametric: Knowledge Gradient



(g) Lookup Table: Knowledge Gradient, Parametric: Pure Exploitation

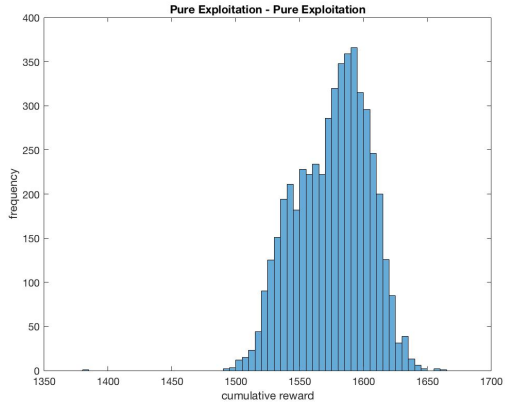


(h) Lookup Table: Knowledge Gradient, Parametric: Interval Estimation

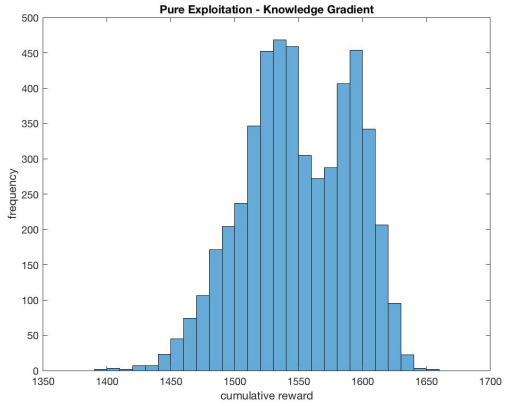


(i) Lookup Table: Knowledge Gradient, Parametric: Knowledge Gradient

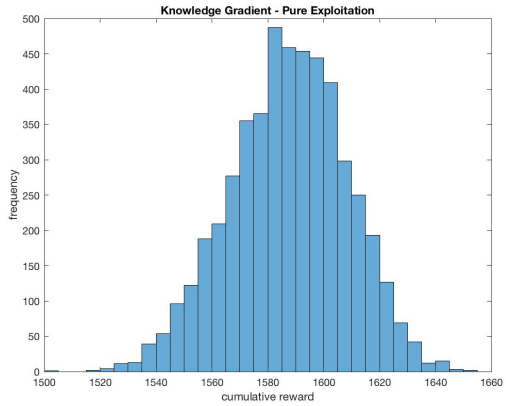
Figure 3: Distribution of cumulative reward over 50000 samples (500 priors each with 100 experiments)



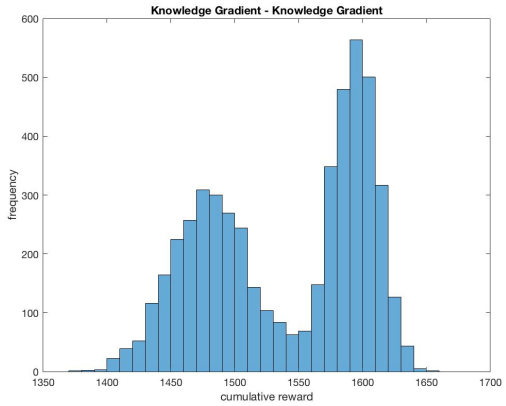
(a) Lookup Table: Pure Exploitation, Parametric: Pure Exploitation



(b) Lookup Table: Pure Exploitation, Parametric: Knowledge Gradient



(c) Lookup Table: Knowledge Gradient, Parametric: Pure Exploitation



(d) Lookup Table: Knowledge Gradient, Parametric: Knowledge Gradient

Figure 4: Extensions: Distribution of cumulative reward when we add an additional latent variable that we fail to account for in our belief model