Written Preliminary Examination Simulation Study

Peter P. Norwood
Department of Statistics, North Carolina State University

1 Introduction

This simulation study analyzes different sequential experimental designs, specifically adaptive designs that balance optimization with information gain. The study evaluates how well different methods simultaneously learn model parameters and maximize reward. The framework is a K-arm contextual bandit where reward distributions depend on both the context and interventions. The primary endpoints of interest are cumulative in-experiment reward and parameter convergence. Additionally, based on the data from each experiment and a batch of new subjects, we examine the proportion of out-of-experiment subjects assigned the optimal intervention and the out-of-experiment regret.

In Section 2, we detail the data generation and experimentation methods. In Section 3, we present the metrics of interest and the corresponding results. We include tables detailing the results across different methods and scenarios. Brief descriptions highlighting the main results are given. We conclude with discussion of the results and topics for future research in Section 4.

2 Methodology

2.1 Data Generation

For this study, the reward is the outcome, which is a function of the context, $\mathbf{X}_t \in \mathbb{R}^p$, the discrete intervention, $A_t \in \{1, \ldots, K\}$, and mean parameter $\theta^* = (\theta_1^*, \ldots, \theta_K^*)^T$ where $\theta_k^* = (\theta_{k,0}^*, \ldots, \theta_{k,p}^*)^T$. The model takes the form $Y_t = \phi(\mathbf{X}_t, A_t)^T\theta^* + \varepsilon_t$ and can be represented as,

$$Y_t = \sum_{k=1}^{K} \mathbb{I}(A_t = k)\left(\theta_{k,0}^* + \theta_{k,1}^* X_{t,1} + \cdots + \theta_{k,p}^* X_{t,p}\right) + \varepsilon_t,$$

where $\varepsilon_t \overset{iid}{\sim} N\left(0, 0.25\sqrt{p}\right)$ and $\mathbb{I}(\cdot)$ is the indicator function.

Each Monte Carlo replicate $j = 1, \ldots, M$ has its own θ^*, which is drawn randomly from $MVN\left(\mathbf{0}_{K(p+1)}, \mathbf{I}_{K(p+1)}\right)$, where $\mathbf{I}_d$ is a $d \times d$ identity matrix. Each $t = 1, \ldots, T$ experimental subject has context $\mathbf{X}_t \sim MVN(\mathbf{0}_p, 0.25 \times \mathbf{I}_p)$, truncated at -1 and 1. In each Monte Carlo replicate, each method assigns interventions to the same group of subjects (same context). To ensure $\mathbf{M}_t = \sum_{i=1}^{t} \phi(\mathbf{X}_i, A_i)\phi(\mathbf{X}_i, A_i)^T$ is positive definite, a burn-in period of $t_{min} = 3K(p+1)$ of simple randomization is used before the adaptive methods begin. After the burn-in period, we estimate θ^* using ordinary least squares (OLS) for all methods. A simulation study in Bastani et al. (2017) motivates this framework.

2.2 Methods

Based on the literature review, we consider the following methods. We first consider an ϵ-Greedy (EG) method with exploration probability $\epsilon = 0.05$. We also consider a simple greedy bandit (GB).

Next is the greedy-first (GF) algorithm from Bastani et al. (2017). For $t \ni t_{min} < t < t_0$,

GF is equivalent to GB. For $t > t_0$, if $\lambda_{min}\mathbf{M}_t > (8t_0)^{-1}t * \lambda_{min}\mathbf{M}_{t_0}$, GF continues with greedy optimization (λ_{min} denotes the minimum eigenvalue). If $\lambda_{min}\mathbf{M}_t$ is less than or equal to the threshold, GF switches to ϵ-Greedy with $\epsilon = 0.05$ for the remainder of the experiment. For all GF situations, we take $t_0 = t_{min} + 100$.

We now consider the methodology from Pronzato (2000) (PR). We choose

$$A_t = \underset{a\in(1,\dots,K)}{\arg\max}\, \phi(\mathbf{X}_t, A_t = a)^T\widehat{\theta}_{t-1} + \frac{\alpha_t}{t}\phi(\mathbf{X}_t, A_t = a)^T\mathcal{I}_{t-1}^{-1}\phi(\mathbf{X}_t, A_t = a),$$

where $\mathcal{I}_t = \mathbf{M}_t/t$. We set $\alpha_t = 0.01\sqrt{t}\log(t)$; this is chosen to satisfy assumptions in Theorem 1 in the literature review and based on preliminary simulations to ensure a balance of reward and information gain.

Next, we consider two variants of Information-Directed Sampling (IDS) (Russo and Van Roy, 2018). The first is a contextual bandit generalization of Algorithm 6 in the reference. At time step t, we have OLS estimators $\widehat{\theta}_{t-1}$ and $\widehat{\Sigma}_{t-1}$ and the goal is to generate $\Psi_t(a) = \Delta_t(a)^2/g_t(a)$, the information ratio if we selection intervention $A_t = a$. For both variants of IDS, we take $A_t = \arg\min_{a\in(1,\dots,K)} \Psi_t(a)$.

We refer to the first variant as IDS-B since it approximates Bayesian information gain. Algorithm 1 details the computation of of $\Psi_t(a)$ for IDS-B. In Algorithm 1, if $|\widehat{\Theta}_a| = 0$, we simply sum over the a such that $|\widehat{\Theta}_a| > 0$ in step 5. Additionally, if $|\widehat{\Theta}_a| = B$ for some a, then we choose $A_t = a$. Here we use $MVN(\widehat{\theta}_{t-1}, \widehat{\Sigma}_{t-1})$ as a surrogate posterior distribution for θ; one can easily specify priors and sample from the posterior distribution of θ instead. For the simulations, we take $B = 100$.

The next variant, IDS-D, uses the determinant of $\mathbf{M}_t$ to measure information gain. Algorithm 2 details the computation of $\Psi_t(a)$ for IDS-D. In these simulations, we take $c = 1$ in step 2 of algorithm 2.

Algorithm 1: Approximate Information Ratio via IDS-B

Sample $\left(\tilde{\theta}^1, \ldots, \tilde{\theta}^B\right)$ from $MVN(\widehat{\theta}_{t-1}, \widehat{\Sigma}_{t-1})$

$\widehat{\mu} \leftarrow B^{-1} \sum_{b=1}^{B} \tilde{\theta}^b$

$\widehat{\Theta}_a \leftarrow \left\{ b : \phi(\mathbf{X}_t, A_t = a)^T \tilde{\theta}^b = \max_{a' \in (1,\ldots,K)} \phi(\mathbf{X}_t, A_t = a')^T \tilde{\theta}^b \right\}, \forall a \in (1, \ldots, K)$

$\widehat{p}^*(a) \leftarrow |\widehat{\Theta}_a|/B, \forall a \in (1, \ldots, K)$

$\widehat{\mu}_a \leftarrow \sum_{\tilde{\theta}^b \in \widehat{\theta}_a} \tilde{\theta}^b / |\widehat{\Theta}_a|, \forall a \in (1, \ldots, K)$

$\widehat{\mathbf{L}} \leftarrow \sum_{a=1}^{K} \widehat{p}^*(a)(\widehat{\mu}_a - \widehat{\mu})(\widehat{\mu}_a - \widehat{\mu})^T$

$\rho^* \leftarrow \sum_{a=1}^{K} \widehat{p}^*(a) \phi(\mathbf{X}_t, A_t = a)^T \widehat{\mu}_a$

$\rho^* \leftarrow \sum_{a=1}^{K} \widehat{p}^*(a) \phi(\mathbf{X}_t, A_t = a)^T \widehat{\mu}_a$

$\Delta_t(a) \leftarrow \phi(\mathbf{X}_t, A_t = a)^T \widehat{\mathbf{L}} \phi(\mathbf{X}_t, A_t = a), \forall a \in (1, \ldots, K)$

$g_t(a) \leftarrow \rho^* - \phi(\mathbf{X}_t, A_t = a)^T \widehat{\mu}, \forall a \in (1, \ldots, K)$

$\Psi_t(a) \leftarrow \Delta_t(a)^2 / g_t(a), \forall a \in (1, \ldots K)$

Algorithm 2: Approximate Information Ratio via IDS-D

$\widehat{\mu}_a \leftarrow \phi(\mathbf{X}_t, A_t = a)^T \widehat{\theta}_{t-1}, \forall a \in (1, \ldots, K)$

$\rho^* \leftarrow \max_a \widehat{\mu}_a + c, c > 0$

$\Delta_t(a) \leftarrow \rho^* - \widehat{\mu}_a, \forall a \in (1, \ldots, K)$

$\mathbf{M}_t^a \leftarrow \mathbf{M}_{t-1} + \phi(\mathbf{X}_t, A_t = a) \phi(\mathbf{X}_t, A_t = a)^T, \forall a \in (1, \ldots, K)$

$g_t(a) \leftarrow \det \mathbf{M}_t^a / \det \mathbf{M}_{t-1}, \forall a \in (1, \ldots, K)$

$\Psi_t(a) \leftarrow \Delta_t(a) / g_t(a), \forall a \in (1, \ldots K)$

2.3 Simulation Parameters

We vary both the number of actions (K) and the dimension of the context (p), leading to four different scenarios. The first has $K = 2, p = 2$, the second has $K = 2, p = 8$, the third has $K = 8, p = 2$, and the fourth has $K = 8, p = 8$. In all scenarios, $T = 1000$ subjects complete the trial.

A primary endpoint is the mean cumulative regret at $T = 1000$. Based on preliminary situations, we choose a Monte Carlo sample size of $M = 5000$; this sample size ensures a

maximum Monte Carlo standard error smaller than 0.50 for mean cumulative regret.

3 Results

3.1 Cumulative Regret

The first endpoint we examine is the mean cumulative regret at $T = 1000$ (equation 2.4 in the review). The Monte Carlo mean and the standard deviation is the sample mean and sample standard deviation of all Regret(T) for a certain method and scenario. The Monte Carlo standard error is the Monte Carlo standard deviation divided by $\sqrt{5000}$. We report these means and standard errors (in parentheses) across the different methods and scenarios in Table 1.

When $K = 2$, variants of IDS have the lowest cumulative regret; IDS-D minimizes for $K = 2, p = 2$ and IDS-B for $K = 2, p = 8$. In both $K = 8$ cases, GF has the lowest regret with GB close behind. When $K = 8, p = 2$, variants of IDS are close behind GB and GF. When $K = 8, p = 8$, GB and GF have much lower regret than other methods. In all scenarios, EG performs dramatically worse than all other methods.

Table 1: Mean Cumulative Regret at $t = 1000$

	$K = 2, p = 2$	$K = 2, p = 8$	$K = 8, p = 2$	$K = 8, p = 8$
EG	37.655 (0.262)	64.576 (0.239)	63.885 (0.312)	142.167 (0.357)
GB	5.392 (0.188)	21.710 (0.152)	21.199 (0.390)	91.984 (0.361)
GF	7.171 (0.147)	21.956 (0.152)	**21.003** (0.371)	**91.772** (0.359)
IDS-B	5.394 (0.067)	**21.119** (0.142)	23.682 (0.159)	105.889 (0.322)
IDS-D	**4.585** (0.061)	21.997 (0.147)	23.025 (0.161)	97.492 (0.334)
PR	5.261 (0.068)	21.806 (0.143)	29.635 (0.151)	104.024 (0.323)

3.2 Convergence of Parameter Estimates

To measure the parameter convergence, we calculate the euclidean norm between $\widehat{\theta}_T$ and θ^*:

$$\text{Norm}(T) = \left\{ \sum_{k=1}^{K} \sum_{j=0}^{p} \left(\widehat{\theta}_{T,k,j} - \theta^*_{k,j} \right)^2 \right\}^{1/2}.$$

In Table 2, we report the Monte Carlo mean norms and standard errors. To gauge how much information was lost from the adaptive methods, we calculate the relative efficiency compared to simple randomization (SR). The relative efficiency is the Monte Carlo mean norm from SR divided by the Monte Carlo mean norm from the method of interest; higher values indicate greater efficiency. Table 3 reports the different relative efficiencies across different scenarios and methods.

In all scenarios, GB and GF are less efficient than the other methods. When $K = 2$, the forced exploration of EG leads to the most efficient estimation. When $K = 8$, PR is the most efficient method. For all scenarios, variants of IDS perform fairly well, modestly less efficient than the leaders, but more efficient than GB and GF. In general, more information is lost compared to SR when $K = 2$ than when $K = 8$.

Table 2: Mean Parameter Convergence: $||\widehat{\theta}_T - \theta^*||_2$

	$K = 2, p = 2$	$K = 2, p = 8$	$K = 8, p = 2$	$K = 8, p = 8$
EG	**0.251** (0.002)	**0.629** (0.003)	1.594 (0.006)	3.171 (0.006)
GB	0.517 (0.005)	0.723 (0.004)	1.968 (0.008)	3.285 (0.007)
GF	0.393 (0.005)	0.723 (0.003)	1.969 (0.008)	3.289 (0.007)
IDS-B	0.323 (0.003)	0.686 (0.003)	1.615 (0.006)	3.073 (0.006)
IDS-D	0.341 (0.003)	0.707 (0.003)	1.590 (0.004)	3.156 (0.006)
PR	0.326 (0.003)	0.701 (0.003)	**1.427** (0.004)	**3.065** (0.006)

While exceptions exist, there is generally a negative relationship between cumulative

Table 3: Relative Efficiency Compared to Simple Randomization

	$K = 2, p = 2$	$K = 2, p = 8$	$K = 8, p = 2$	$K = 8, p = 8$
EG	**0.528**	**0.800**	0.351	0.701
GB	0.317	0.735	0.284	0.676
GF	0.337	0.735	0.284	0.675
IDS-B	0.409	0.776	0.347	0.723
IDS-D	0.388	0.752	0.352	0.704
PR	0.407	0.758	**0.392**	**0.725**

regret and efficiency. It is important to gauge the cost (regret) per amount of information gained (relative efficiency) in the experiment, as with the information ratio from Russo and Van Roy (2018). To do this, we calculate cost as the mean regret at $T = 1000$ divided by the relative efficiency for each method under each scenario. Table 4 presents these ratios for different methods and scenarios.

IDS-D has the lowest cost relative to its efficiency when $K = 2, p = 2$ and $K = 8, p = 2$. PR has the lowest ratio with $K = 2, p = 8$. When $K = 8, p = 8$, GF has the lowest ratio. While EG is more efficient in the $K = 2$ cases, it pays a considerable price in terms of cumulative regret – EG has the highest ratio in all scenarios.

Table 4: Ratio of Cumulative Regret to Relative Efficiency

	$K = 2, p = 2$	$K = 2, p = 8$	$K = 8, p = 2$	$K = 8, p = 8$
EG	71.325	80.741	181.804	202.949
GB	16.998	29.529	74.513	136.034
GF	21.291	29.864	73.796	**135.882**
IDS-B	13.182	29.797	68.311	146.491
IDS-D	**12.084**	29.226	**65.377**	138.519
PR	12.941	**28.755**	75.517	143.569

3.3 Out-of-Experiment Performance

The next section analyzes out-of-experiment performance. Using all the information from the experiment, we hope methods can assign a high proportion of subjects the optimal intervention. If they fail to assign the optimal intervention, we hope they assign an intervention that has a similarly high reward (low regret). To analyze this out-of-experiment performance, we generate $i = 1, \ldots, 500$ new subjects with the same context distribution as the experimental subjects. Using all the information from the experiment, we assign the estimated optimal intervention, $A_i = \arg\max_{a \in (1,\ldots,K)} \phi(\mathbf{X}_i, A_i = a)^T \widehat{\theta}_T$ to each member of the new group. For each Monte Carlo replicate from each scenario and method, the proportion of optimal assignment is:

$$\text{Prop} = \sum_{i=1}^{500} \mathbb{I} \left\{ \arg\max_{a \in (1,\ldots,K)} \phi(\mathbf{X}_i, A_i = a)^T \widehat{\theta}_T = \arg\max_{a \in (1,\ldots,K)} \phi(\mathbf{X}_i, A_i = a)^T \theta^* \right\}.$$

Table 5 describes the Monte Carlo mean and standard errors for the proportions across different scenarios and methods. We include SR for comparison. We also calculate the regret for the 500 out-of-experiment subjects:

$$\text{Regret} = \sum_{i=1}^{500} \max_{a \in (1,\ldots,K)} \phi(\mathbf{X}_i, A_i = a)^T \theta^* - \phi(\mathbf{X}_i, A_i)^T \widehat{\theta}_T.$$

Table 6 includes the Monte Carlo means and standard errors for the out-of-experiment regret.

When $K = 2$, SR has the strongest out-of-experiment performance in both metrics. When $K = 2, p = 2$, the improvements are most dramatic for SR. Interestingly, when $K = 8$, the adaptive methods generally outperform SR. When $K = 8, p = 2$, IDS and PR perform better than others. When $K = 8, p = 8$, the adaptive methods all perform

similarly.

Table 5: Mean Proportion of Out-of-Experiment Subjects Assigned Optimal Intervention

	$K = 2, p = 2$	$K = 2, p = 8$	$K = 8, p = 2$	$K = 8, p = 8$
EG	0.976 (< 0.001)	0.944 (< 0.001)	0.907 (0.001)	0.773 (0.001)
GB	0.971 (0.001)	0.943 (< 0.001)	0.902 (0.001)	0.774 (0.001)
GF	0.971 (0.001)	0.943 (< 0.001)	0.903 (0.001)	0.774 (0.001)
IDS-B	0.976 (< 0.001)	0.945 (< 0.001)	**0.919** (0.001)	**0.777** (0.001)
IDS-D	0.975 (< 0.001)	0.943 (< 0.001)	0.916 (0.001)	0.776 (0.001)
PR	0.975 (< 0.001)	0.943 (< 0.001)	0.916 (0.001)	0.776 (0.001)
SR	**0.981** (< 0.001)	**0.950** (< 0.001)	0.903 (0.001)	0.753 (0.001)

Table 6: Mean Regret of Out-of-Experiment Subjects

	$K = 2, p = 2$	$K = 2, p = 8$	$K = 8, p = 2$	$K = 8, p = 8$
EG	1.132 (0.030)	7.423 (0.070)	7.471 (0.164)	56.556 (0.277)
GB	1.829 (0.100)	7.835 (0.075)	8.642 (0.209)	56.837 (0.281)
GF	1.850 (0.080)	7.871 (0.075)	8.399 (0.200)	56.559 (0.280)
IDS-B	1.080 (0.028)	7.335 (0.069)	**5.120** (0.088)	**53.919** (0.264)
IDS-D	1.147 (0.029)	7.706 (0.074)	5.409 (0.088)	55.048 (0.271)
PR	1.147 (0.031)	7.553 (0.070)	5.356 (0.085)	54.693 (0.273)
SR	**0.664** (0.018)	**5.964** (0.059)	7.039 (0.100)	65.255 (0.317)

Our last metric of interest is the ratio of in-experiment cumulative regret to the out-of-experiment proportion of subjects assigned the optimal intervention. Similar to the previous ratio, this gives us one way to understand the cost of exploration in the experiment. Table 7 details these ratios.

In all scenarios except $K = 2, p = 2$, GB and GF have similarly low ratios compared to others. When $K = 2, p = 2$, IDS-D has the lowest ratio. Once again, EG severely under-performs the other methods as its in-experiment regret is much higher.

Table 7: Ratio of Cumulative Regret to Proportion Assigned Optimal Intervention

	$K = 2, p = 2$	$K = 2, p = 8$	$K = 8, p = 2$	$K = 8, p = 8$
EG	38.593	68.380	70.428	183.807
GB	5.555	**23.022**	23.507	118.846
GF	7.389	23.286	**23.256**	**117.554**
IDS-B	5.530	24.474	25.774	136.205
IDS-D	**4.803**	23.319	25.131	125.712
PR	5.397	23.103	32.336	134.035

4 Conclusion

The purpose of this simulation study is to examine how different sequential experimental designs balance in-experiment optimization and information gain. Attention is placed on methods that consider if, when, and how to explore. We are interested in how they compare to less sophisticated methods that never explore or only randomly explore with no direction. Performance is measured by how well they minimize in-experiment regret, learn the parameters, and assign interventions to new, out-of-experiment subjects.

When optimizing within the experiment, we find that the methods that consider if, when, and how to explore fare better than others. With two arms, we find variants of IDS minimize cumulative regret compared to others. When we have eight arms, GF, which can switch away from GB based on the amount of information collected, best minimizes cumulative regret.

With parametric models, we can measure information gain by how well we learn the parameters throughout the experiment. With two-armed bandits, the forced, random exploration of EG leads to more efficient parameter estimation. When we have eight arms, however, the targeted exploration of IDS and PR is more efficient. This highlights the importance of how we explore, not just how often.

Another way to measure information gain is out-of-experiment performance. Using all

the information from an experiment, we wish to assign new subjects effective interventions. Generally, variants of IDS perform well out of the experiment. With two arms, SR has lower regret and assigns higher proportions of subjects to the optimal intervention. Outside of the simplest $K = 2, p = 2$ case, the random exploration of EG does not translate into better out of trial performance compared to targeted exploration. While GB and GF suffer when $p = 2$, they both perform quite well when $p = 8$. This is due to the additional randomness from the context, which leads to natural exploration when choosing greedily.

Interestingly, with eight arms, the adaptive methods generally outperform SR out of the experiment. This may seem counter-intuitive given that SR is more efficient in these cases. However, SR is more efficient for *all* of the parameters. The adaptive methods spend more time sampling from the potentially optimal arms, so they can better identify which arms are optimal. Furthermore, targeted sampling of high information gain actions can improve out-of-experiment performance over greedy optimization and random exploration.

By looking at the ratio of cumulative regret to relative efficiency and proportion of post-experiment optimal assignment, we gauge the balance of in-experiment optimization and information gain. Across different scenarios, IDS-D and GF perform particularly well. Overall, EG suffers considerably. This furthers the point that major gains can be made by carefully considering if, when, and how to explore.

Of the methods that combine optimization and exploration (EG, IDS, GF, PR), EG is clearly the worst. Between IDS, GF, and PR, we find PR suffering some with both ratios. In general, variants of IDS appear to strike the best balance when $K = 2$ and GF when $K = 8$.

This simulation study verifies the need to look further into sequential experimentation methods that focus on exploration. Future work includes examining methods in different modeling scenarios, developing methods that ensure valid statistical inference, and looking

at other ways to evaluate how a method balances optimization and information gain (e.g., power analysis).

References

Bastani, H., Bayati, M., and Khosravi, K. (2017). Mostly exploration-free algorithms for contextual bandits.

Pronzato, L. (2000). Adaptive optimization and d-optimum experimental design. *Ann. Statist.*, 28(6):1743–1761.

Russo, D. and Van Roy, B. (2018). Learning to optimize via information-directed sampling. *Operations Research*, 66(1):230–252.

5 Appendix

All simulation code can be found at `https://github.com/peterpnorwood/WrittenPrelim`. A README detailing the repository is included.