A PROOFS

Proof of Theorem 1.

We first provide two lemmas for verifying condition equation 10 in Lemma 1.

Lemma 2 (Li et al. (2017) Proposition 1)**.** *Define $\Sigma_t = \sum_{s=1}^{t} X_s X_s^T$, where X_s is drawn i.i.d. from some distribution ν supported on d-dimensional unit ball $\mathcal{B}^d := \{x \in \mathbb{R}^d : \|x\| \leq 1\}$. Furthermore, let $\Sigma := \mathbb{E}[X_s X_s^T]$ be the second moment matrix, and B and $\delta > 0$ be two positive constants. Then, there exist positive, universal constants C_1 and C_2 such that $\lambda_{\min}(\Sigma_t) \geq B$ with probability at least $1 - \delta$, as long as*

$$t \geq \left(\frac{C_1\sqrt{d} + C_2\sqrt{\log(1/\delta)}}{\lambda_{\min}(\Sigma)} \right)^2 + \frac{2B}{\lambda_{\min}(\Sigma)}.$$

Lemma 3 (Cesa-Bianchi & Fischer (1998) Lemma 3)**.** *For index set $\mathcal{J} \subseteq \{1, 2, \cdots\}$, define $\Sigma(\mathcal{J}) := \sum_{s \in \mathcal{J}} X_s X_s^T$. If $\mathcal{J}_2 \subseteq \mathcal{J}_1$, then $\lambda_{\min}(\Sigma(\mathcal{J}_2)) \leq \lambda_{\min}(\Sigma(\mathcal{J}_1))$.*

For $W_s \in \mathcal{B}^{2d_1}$ and $V_s \in \mathcal{B}^{3d_2}$, denote $G_t = \sum_{s=1}^{t} W_s W_s^T$, $H_t = \sum_{s=1}^{t} V_s V_s^T$, $\Sigma_1 = \mathbb{E}[W_s W_s^T]$ and $\Sigma_2 = \mathbb{E}[V_s V_s^T]$. Recall $\tau = m \times K$ denote the number of initialization rounds. When $\tau = \max\left\{ \left(\frac{C_1\sqrt{2d_1} + C_2\sqrt{\log(2/\delta)}}{\lambda_{\min}(\Sigma_1)} \right)^2 + \frac{2B_1}{\lambda_{\min}(\Sigma_1)}, \left(\frac{C_1\sqrt{3d_2} + C_2\sqrt{\log(2/\delta)}}{\lambda_{\min}(\Sigma_2)} \right)^2 + \frac{2B_2}{\lambda_{\min}(\Sigma_2)} \right\}$, then by Lemma 2, w.p. at least $1 - \delta$, $\lambda_{\min}(G_\tau) \geq B_1$ and $\lambda_{\min}(H_\tau) \geq B_2$ hold. Here we may take $B_1 = \max\left\{1, \frac{512 M_g^2 \sigma_y^2}{\kappa_y^4}\left(4d_1^2 + \log\frac{6}{\delta}\right)\right\}$ and $B_2 = \max\left\{1, \frac{512 M_h^2 \sigma_z^2}{\kappa_z^4}\left(9d_2^2 + \log\frac{6}{\delta}\right)\right\}$.

Note that for $t \geq \tau$, when $\lambda_{\min}(G_\tau) \geq B_1$ and $\lambda_{\min}(H_\tau) \geq B_2$, by Lemma 3, $\lambda_{\min}(G_t) \geq B_1$ and $\lambda_{\min}(H_t) \geq B_2$ are always true.

Let $\omega_s = I(A_t \neq \hat{k}_t)$, $s > \tau$. Then $\omega_s = 0$ when the algorithm chooses the best arm estimated by $\hat{\eta}_{t-1}$ and $\omega_s = 1$ if the algorithm explores other arms. Further, let $\mathcal{T}(t) = \{\tau < s \leq t : \omega_s = 1\}$, then the cumulative regret up to time t can be decomposed as

$$\begin{aligned}
R_t &= \sum_{s=1}^{t} \{q_s^* - q_{A_s,s} + \lambda(p_{A_s,s} - \theta)_+ - \lambda(p_s^* - \theta)_+\} \\
&= \sum_{s=1}^{t} \left\{h(V_{s^*}^T\gamma) - h(V_s^T\gamma) + \lambda(g(W_s^T\beta) - \theta)_+ - \lambda(g(W_{s^*}^T\beta) - \theta)_+\right\} \\
&= \sum_{s=1}^{\tau} \left\{h(V_{s^*}^T\gamma) - h(V_s^T\gamma) + \lambda(g(W_s^T\beta) - \theta)_+ - \lambda(g(W_{s^*}^T\beta) - \theta)_+\right\} \qquad (12) \\
&+ \sum_{s=\tau+1, s\in\mathcal{T}(t)}^{t} \left\{h(V_{s^*}^T\gamma) - h(V_s^T\gamma) + \lambda(g(W_s^T\beta) - \theta)_+ - \lambda(g(W_{s^*}^T\beta) - \theta)_+\right\} \qquad (13) \\
&+ \sum_{s=\tau+1, s\notin\mathcal{T}(t)}^{t} \left\{h(V_{s^*}^T\gamma) - h(V_s^T\gamma) + \lambda(g(W_s^T\beta) - \theta)_+ - \lambda(g(W_{s^*}^T\beta) - \theta)_+\right\} \qquad (14)
\end{aligned}$$

here we use $W_{s^*} := (\Phi(X_s)^T, u_{k^*(s)}\Phi(X_s)^T)^T$ and $V_{s^*} := (\Psi(X_s)^T, u_{k^*(s)}\Psi(X_s)^T, u_{k^*(s)}^2 \Psi(X_s)^T)^T$ as short-hand notations.

Now we bound the three parts equation 12, equation 13 and equation 14 separately.

The first part

$$equation\ 12 \leq \sum_{s=1}^{\tau}(1+\lambda) \leq (1+\lambda)\tau, \tag{15}$$

since $h(x) \in [0,1]$ and $(g(x)-\theta)_+ \in [0,1]$.

The second part $equation\ 13 \leq (1+\lambda)\sum\limits_{s=\tau+1}^{t}\omega_s$. Since $\omega_s \overset{ind}{\sim} Bernoulli(\frac{K-1}{K}\epsilon_s)$, by Hoeffding's inequality, for fixed $\delta > 0$,

$$P\left(\sum_{s=\tau+1}^{t}\omega_s \leq \sum_{s=\tau+1}^{t}\mathbb{E}\omega_s + \sqrt{\frac{t-\tau}{2}\log\frac{1}{\delta}}\right) \geq 1-\delta,$$

and $\sum\limits_{s=\tau+1}^{t}\mathbb{E}\omega_s = \sum\limits_{s=\tau+1}^{t}\frac{K-1}{K}\min\left\{1, C\frac{\log s}{s}\right\} \leq C\left\{\frac{(\log t)^2}{2} - \frac{(\log\tau)^2}{2}\right\}$.

Therefore, w.p. at least $1-\delta$,

$$equation\ 13 \leq (1+\lambda)\left(C\frac{(\log t)^2}{2} - C\frac{(\log\tau)^2}{2} + \sqrt{\frac{t-\tau}{2}\log\frac{1}{\delta}}\right) \tag{16}$$

The next lemma provides a bound on the third part of regret.

Lemma 4 (Li et al. (2017) Lemma 2)**.** *Let $\{X_t\}_{t=1}^{\infty}$ be a sequence in $\mathbb{R}^d$ satisfying $\|X_t\| \leq 1$. Define $\Sigma_t = \sum\limits_{s=1}^{t} X_s X_s^T$. Suppose there is an integer n such that $\lambda_{\min}(\Sigma_n) \geq 1$, then for all $t > n$,*

$$\sum_{s=n+1}^{t}\|X_s\|_{\Sigma_{s-1}^{-1}} \leq \sqrt{2(t-n)d\log\frac{t}{d}}.$$

Therefore, we can bound term equation 14 by Lemma 1 and Lemma 4.

When $A_s = \hat{k}_s$, since the algorithm is greedy, we know $h(V_s^T\hat{\gamma}_{s-1}) - \lambda(g(W_s^T\hat{\beta}_{s-1})-\theta)_+ \geq h(V_{s^*}^T\hat{\gamma}_{s-1}) - \lambda(g(W_{s^*}^T\hat{\beta}_{s-1})-\theta)_+$, therefore,

$$\begin{aligned}
equation\ 14 \leq & \sum_{s=\tau+1, s\notin\mathcal{T}(t)}^{t}\Big\{h(V_{s^*}^T\gamma) - h(V_s^T\gamma) + \lambda(g(W_s^T\beta)-\theta)_+ - \lambda(g(W_{s^*}^T\beta)-\theta)_+ \\
& + h(V_s^T\hat{\gamma}_{s-1}) - \lambda(g(W_s^T\hat{\beta}_{s-1})-\theta)_+ - h(V_{s^*}^T\hat{\gamma}_{s-1}) + \lambda(g(W_{s^*}^T\hat{\beta}_{s-1})-\theta)_+\Big\} \\
= & \sum_{s=\tau+1, s\notin\mathcal{T}(t)}^{t}\Big\{h(V_{s^*}^T\gamma) - h(V_{s^*}^T\hat{\gamma}_{s-1}) \\
& \qquad + h(V_s^T\hat{\gamma}_{s-1}) - h(V_s^T\gamma) \\
& \qquad + \lambda(g(W_s^T\beta)-\theta)_+ - \lambda(g(W_s^T\hat{\beta}_{s-1})-\theta)_+ \\
& \qquad + \lambda(g(W_{s^*}^T\hat{\beta}_{s-1})-\theta)_+ - \lambda(g(W_{s^*}^T\beta)-\theta)_+\Big\} \\
\leq & \sum_{s=\tau+1, s\notin\mathcal{T}(t)}^{t}\Big\{L_h|V_{s^*}^T(\hat{\gamma}_{s-1}-\gamma)| + L_h|V_s^T(\hat{\gamma}_{s-1}-\gamma)| + \lambda L_g|W_s^T(\hat{\beta}_{s-1}-\beta)| + \lambda L_g|W_{s^*}^T(\hat{\beta}_{s-1}-\beta)|\Big\},
\end{aligned} \tag{17}$$

where the last inequality follows from $0 \leq h'(x) \leq L_h$, $0 \leq g'(x) \leq L_g$ and observing than $(g(x)-\theta)_+$ is also an L_g-Lipschitz continuous function.

When $\lambda_{\min}(G_\tau) \geq B_1$ and $\lambda_{\min}(H_\tau) \geq B_2$, by Lemma 1, for any $w \in \mathbb{R}^{2d_1}$, $v \in \mathbb{R}^{3d_2}$ and $s > \tau$,

$$P\left(|w^T(\hat{\beta}_s-\beta)|\le\frac{\sigma_y}{\kappa_y}\sqrt{\log(6/\delta)}\|w\|_{G_s^{-1}}\Big|\lambda_{\min}(G_s)\ge B_1\right)\ge 1-\delta/2,$$

and

$$P\left(|v^T(\hat{\gamma}_s-\gamma)|\le\frac{\sigma_z}{\kappa_z}\sqrt{\log(6/\delta)}\|v\|_{H_s^{-1}}\Big|\lambda_{\min}(H_s)\ge B_2\right)\ge 1-\delta/2.$$

Thus,

$$P\left(|w^T(\hat{\beta}_s-\beta)|\le\frac{\sigma_y}{\kappa_y}\sqrt{\log(6/\delta)}\|w\|_{G_s^{-1}},|v^T(\hat{\gamma}_s-\gamma)|\le\frac{\sigma_z}{\kappa_z}\sqrt{\log(6/\delta)}\|v\|_{H_s^{-1}}\right)\ge 1-2\delta.$$

With probability at least $1-2\delta$, we can bound equation 17 by

$$\begin{aligned}
equation\ 17 \le & \sum_{s=\tau+1,s\notin\mathcal{T}(t)}^{t}\Bigg\{L_h\frac{\sigma_z}{\kappa_z}\sqrt{\log(6/\delta)}\|V_{s^*}\|_{H_{s-1}^{-1}}+L_h\frac{\sigma_z}{\kappa_z}\sqrt{\log(6/\delta)}\|V_s\|_{H_{s-1}^{-1}}\\
&+\lambda L_g\frac{\sigma_y}{\kappa_y}\sqrt{\log(6/\delta)}\|W_{s^*}\|_{G_{s-1}^{-1}}+\lambda L_g\frac{\sigma_y}{\kappa_y}\sqrt{\log(6/\delta)}\|W_s\|_{G_{s-1}^{-1}}\Bigg\}\\
&\le 2L_h\frac{\sigma_z}{\kappa_z}\sqrt{\log(6/\delta)}\sum_{s=\tau+1}^{t}\|V_s\|_{H_{s-1}^{-1}}+2\lambda L_g\frac{\sigma_y}{\kappa_y}\sqrt{\log(6/\delta)}\sum_{s=\tau+1}^{t}\|W_s\|_{G_{s-1}^{-1}}\\
&\le 2L_h\frac{\sigma_z}{\kappa_z}\sqrt{\log(6/\delta)}\sqrt{6d_2(t-\tau)\log\frac{t}{3d_2}}+2\lambda L_g\frac{\sigma_y}{\kappa_y}\sqrt{\log(6/\delta)}\sqrt{4d_1(t-\tau)\log\frac{t}{2d_1}}.
\end{aligned}\tag{18}$$

The last inequality follows from Lemma 4.

Combining equation 15, equation 16 and equation 18 together, we can show the regret bound in Theorem 1 holds w.p at least $1-3\delta$.

□

B ADDITIONAL ALGORITHM AND RESULTS

One criticism for ϵ-greedy is that the exploration is carried out uniformly among all the arms and cannot adapt to the difference in sample size or uncertainty for different arms like UCB or Thompson sampling. To improve the statistical accuracy of $\hat{\eta}_t$ in the trial, at the exploration step, the arms should be allocated so that the new data obtained from the chosen arm can minimize the variance of $\hat{\eta}_t$. A brief introduction to optimal design of experiment can be found in Fedorov (2010).

At time t, say the algorithm decides to explore based on history data $X_{1:t-1}, A_{1:t-1}, Y_{1:t-1}, Z_{1:t-1}$ and new context X_t. A design $D_t=\begin{Bmatrix}1, & 2, & \cdots, & K\\ \pi_1, & \pi_2, & \cdots, & \pi_K\end{Bmatrix}$ is to choose $A_t=k$ with probability π_k. Here the dependence of π_k on X_t and history data up to time $t-1$ is suppressed for simplicity of notation. For MLE $\hat{\eta}_{t-1}=\underset{\eta=(\beta,\gamma)}{\arg\max}\sum_{s=1}^{t-1}\{\log f(Y_s|X_s,A_s;\beta)+\log f(Z_s|X_s,A_s;\gamma)\}$, its observed information matrix is given by

$$\hat{I}(\hat{\eta}_{t-1})=\begin{pmatrix}\nabla_\beta\ell_{t-1}(\nabla_\beta\ell_{t-1})^T & \nabla_\beta\ell_{t-1}(\nabla_\gamma\ell_{t-1})^T\\ \nabla_\gamma\ell_{t-1}(\nabla_\beta\ell_{t-1})^T & \nabla_\gamma\ell_{t-1}(\nabla_\gamma\ell_{t-1})^T\end{pmatrix},$$

where

$$\nabla_\beta\ell_{t-1}=\sum_{s=1}^{t-1}\nabla\log f(Y_s|X_s,A_s;\beta)|_{\hat{\beta}_{t-1}}$$

$$= \sum_{s=1}^{t-1} \phi(Y_s - p(X_s, A_s; \beta)) \nabla \zeta(X_s, A_s; \beta)|_{\hat{\beta}_{t-1}},$$

and

$$\begin{aligned} \nabla_\gamma \ell_{t-1} &= \sum_{s=1}^{t-1} \nabla \log f(Z_s | X_s, A_s; \gamma)|_{\hat{\gamma}_{t-1}} \\ &= \sum_{s=1}^{t-1} \phi(Z_s - q(X_s, A_s; \gamma)) \nabla \xi(X_s, A_s; \gamma)|_{\hat{\gamma}_{t-1}}, \end{aligned}$$

We are interested in a design that chooses an arm with probability 1.

Write $D_{t,k} = \begin{Bmatrix} 1, & 2, & \cdots, & k, & \cdots, & K \\ 0, & 0, & \cdots, & 1, & \cdots, & 0 \end{Bmatrix}$, then the expected information for the new observation under $D_{t,k}$ is

$$\begin{aligned} I(D_{t,k}, \hat{\eta}_{t-1}) &= \mathbb{E}\{\nabla \log f(Y_t, Z_t | X_t, k; \eta)|_{\hat{\eta}_{t-1}} (\nabla \log f(Y_t, Z_t | X_t, k; \eta)|_{\hat{\eta}_{t-1}})^T\} \\ &= \begin{pmatrix} I_Y & 0 \\ 0 & I_Z \end{pmatrix}, \end{aligned}$$

where $I_Y = \phi^2 \mathbb{E}(Y_t - p(X_t, k; \beta))^2 \{\nabla \zeta(X_t, k; \beta)|_{\hat{\beta}_{t-1}} (\nabla \zeta(X_t, k; \beta)|_{\hat{\beta}_{t-1}})^T\}$ and $I_Z = \phi^2 \mathbb{E}(Z_t - q(X_t, k; \gamma))^2 \{\nabla \xi(X_t, k; \gamma)|_{\hat{\gamma}_{t-1}} (\nabla \xi(X_t, k; \gamma)|_{\hat{\gamma}_{t-1}})^T\}$.

Note in generalized linear models, $\mathbb{E}(Y_t - p(X_t, k; \beta))^2 = \frac{1}{\phi} m_1''(\zeta(X_t, k; \beta))$ and $\mathbb{E}(Z_t - q(X_t, k; \gamma))^2 = \frac{1}{\phi} m_2''(\xi(X_t, k; \gamma))$, so the expectations can be calculated explicitly.

The D-optimum design is to minimize the determinant of the inverse information matrix, i.e., we the arm chosen according to optimal design is

$$\tilde{k}_t = \arg\min_k \det \left(\{\hat{I}(\hat{\eta}_{t-1}) + I(D_{t,k}, \hat{\eta}_{t-1})\}^{-1} \right).$$

The information matrix depends on the true value of η which is unknown. In the MAB algorithm, we can plug in the estimated parameter $\hat{\eta}_{t-1}$ at the tth round, and this only gives us a local optimum design.

In the additional experiment results, we also provide the effect of λ-value on regret and safety feature. $\lambda = 1, 2, 5$ and 10 are considered. The data generation process is the same as described in Section 5. It is not a surprise that no matter how large a λ-value we use, the method "Ignore harm" will not learn the harm effect and induce a large regret as well as violate the safety constraint. Hence, to keep a proper scale of the plots, the results for method "Ignore harm" are omitted. Figure 3 show the results for different λ-value, including the "Optimal design" method.

Note since we do not use a hard constraint, even the oracle method (true models are known) will sometimes pick an arm with $p > \theta$. Nevertheless, the chosen p is usually not much larger than θ. The average value for $(p - \theta)_+$ is from the oracle method is given in Table 2

However, the simulation for exploration with optimal design("Optimal design") did not show an advantage over the original ϵ-greedy algorithm ("Varying coefficient model"). The optimal design tends to collect more data on the extreme arms $k = 1$ or $k = K$ to improve the statistical accuracy for parameter estimates. Thus, the regret is higher than the original ϵ-greedy algorithm. Compared to the original ϵ-greedy algorithm, optimal design is able to identify the best arm correctly slightly more often across all the rounds. Optimal design also identifies the best arm correctly the most often at the end of a trial. The chance of each method choosing the best arm is given in Table 3 and 4.

With more careful integration of the optimal design of experiment into ϵ-greedy framework, we believe the effort to improve statistical accuracy will pay off, and this approach is worth investigating in the future.

Algorithm 2: ϵ-greedy with optimal exploration

input: Time horizon T, exploration rate $\epsilon_t \in (0,1)$, penalty $\lambda \in (0,\infty)$, harm threshold $\theta \in (0,1)$, initialization rounds m

for $t = 1, \cdots, m \times K$ **do**
  Sample from each arm m times, record context X_t and response Y_t, Z_t
end
for $t = m \times K + 1, \cdots, T$ **do**
  Estimate $\hat{\eta}_{t-1}$ based on $X_{1:t-1}, A_{1:t-1}, Y_{1:t-1}, Z_{1:t-1}$
  Identify best arm $\hat{k}_t = \arg\max_k q(X_t, k; \hat{\gamma}_{t-1}) - \lambda(p(X_t, k; \hat{\beta}_{t-1}) - \theta)_+$
  if *A random* $w \sim Unif(0,1) < \epsilon_t$ **then**
    Find the best exploration choice according to optimal design
    $\tilde{k}_t = \arg\min_k \det\left(\{\hat{I}(\hat{\eta}_{t-1}) + I(D_{t,k}, \hat{\eta}_{t-1})\}^{-1}\right)$
    Sample from $A_t = \tilde{k}_t$
  else
    Sample from $A_t = \hat{k}_t$
  end
  Record context X_t, choice A_t and responses Y_t, Z_t
end

output: Parameter estimates $\hat{\eta}_T$

λ	1	2	5	10
Avg. $(p-\theta)_+$	0.014	0.007	0.055	0.053

Table 2: The harm chosen by oracle method does not exceed threshold $\theta = 0.33$ a lot.

C ADDITIONAL COMPUTATION DETAILS

Base `R` function `glm.fit` is used to estimate the parameters. It uses iteratively reweighted least squares to find the MLE for given input data and distribution.

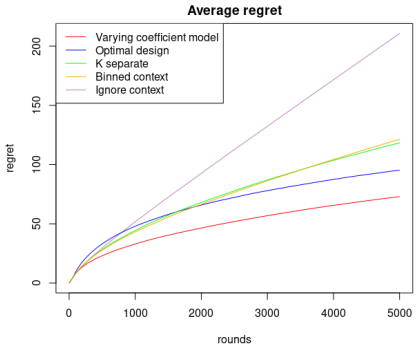
When the number of initialization rounds τ is not large enough, there is a small chance that the parameter $\hat{\eta}_\tau$ cannot be estimated stably. To prevent the algorithm from getting stuck on poor parameter estimates, we use the `cv.glmnet` function in `glmnet` package to fit a GLM with elasticnet regularization when the estimated parameter has an entry with magnitude larger than 100. This regularization only works as a measure of caution, and dose not affect parameter estimates when t is sufficiently large. Choosing a large τ can prevent unstable estimates too. If we use a rule of thumb from linear regression, we need approximately $\tau = 10 \times \#$ of parameters. In our experiment, initialization for the varying coefficient model or K separate models takes $\tau = 63$ rounds ($m = 9$). For binned context model, we sample each arm 3 times for each category so that the total number of initialization rounds is also 63. The exploration rate $\epsilon_t = 11.43\frac{\log t}{t}$. The whole experiment of 100 trails can be finished in several hours on a computer using Intel Xeon Gold 6244 CPU with 32 cores.

λ	1	2	5	10
Varying coefficient	0.56	0.69	0.69	0.72
Optimal design	0.61	0.72	0.72	0.73
K separate	0.37	0.47	0.47	0.48
Binned context	0.38	0.50	0.50	0.50
Ignore Context	0.16	0.14	0.14	0.14

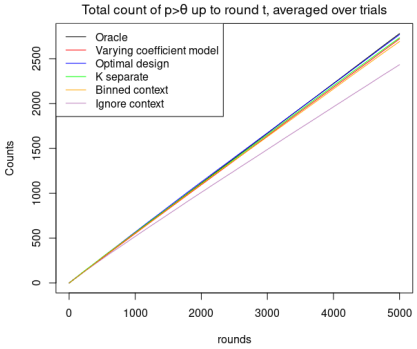
Table 3: The overall chance of each method choosing the best arm among all the 100 trials and 5,000 rounds.

λ	1	2	5	10
Varying coefficient	0.72	0.81	0.81	0.80
Optimal design	0.74	0.86	0.86	0.86
K separate	0.43	0.53	0.53	0.55
Binned context	0.43	0.55	0.55	0.51
Ignore Context	0.12	0.17	0.17	0.17

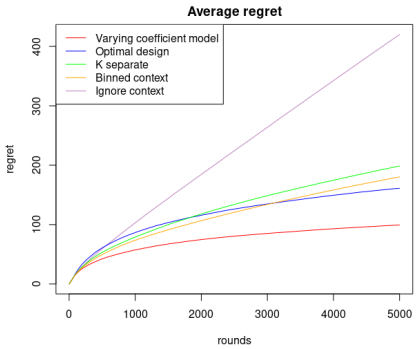
Table 4: The chance of each method choosing the best arm among all the 100 trials at the last (5,000th) round.



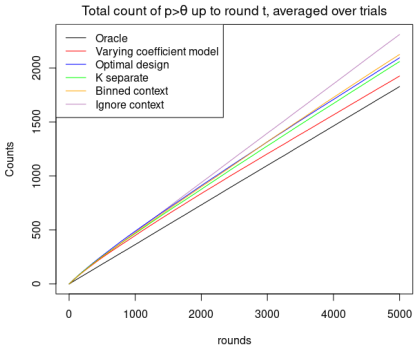
(a) $\lambda = 1$



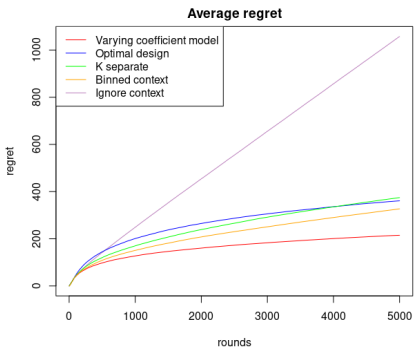
(b) $\lambda = 1$



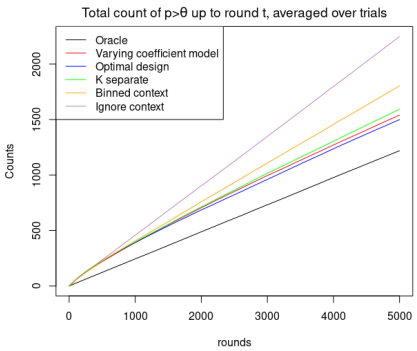
(c) $\lambda = 2$



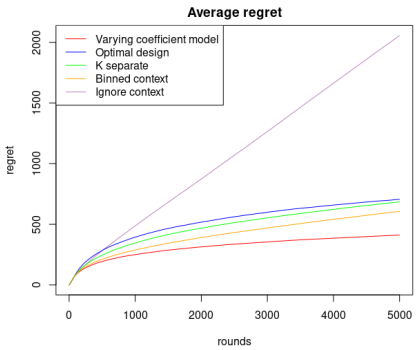
(d) $\lambda = 2$



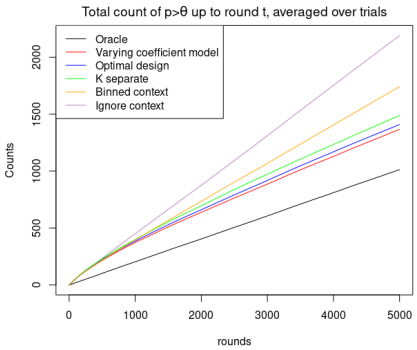
(e) $\lambda = 5$



(f) $\lambda = 5$



(g) $\lambda = 10$



(h) $\lambda = 10$

Figure 3: (a,c,e, g)The cumulative regret averaged across 100 trials. (b,d,f, h) To reflect the safety feature of each method, we calculate how many times the method chooses an arm with harm probability $p > \theta$ up to round t. Then the count is averaged over 100 trials. As penalty λ gets larger, this count decreases.