Online Companion: Appendices

Appendix A summarizes notation. Appendix B provides proofs for mathematical claims in the main paper. Appendix C describes the implementation issues related to our new indices. Appendix D comments on further practical issues for multi-arm highly adaptive trials and related research questions.

Appendix A: Summary of Principal Notation from the Main Paper.

Table EC.1 summarizes the principal notation in the manuscript.

Table EC.1 Principal Notation.

Symbol	*Definition*
M	Number of arms
$\mathcal{M}$	The set of arms, $\{1,2,\ldots,M\}$
c_i	Marginal cost of sampling from arm i during the trial
P	Adopting population size (number of patients affected by the implementation decision)
I_i	One-time cost of implementing arm i at the end of the trial
$t=0,1,\ldots$	Time index; number of patients from whom a sample has been observed
T	The time at which the trial is stopped to select an arm
$T_i, T_{i,l}$	Stopping times used by the adaptive lookahead policies to compute indices
Y_i^t	Random reward obtained from arm i observed at time $t=1,2,\ldots$
θ_i	Unknown mean for the reward of arm i
$\boldsymbol{\theta}$	Column vector $(M\times 1)$ of unknown means
λ_i	Known sampling variance for the reward of arm i
$\boldsymbol{\Lambda}$	Diagonal matrix $(M\times M)$ of sampling variances
$\boldsymbol{\mu}^0$	Mean vector $(M\times 1)$ of the prior distribution for $\boldsymbol{\theta}$
$\boldsymbol{\Sigma}^0$	Variance-covariance matrix $(M\times M)$ of the prior distribution for $\boldsymbol{\theta}$
$\boldsymbol{\mu}^t$	Posterior mean vector of $\boldsymbol{\theta}$ given $t=0,1,\ldots$ samples have been observed
$\boldsymbol{\Sigma}^t$	Posterior variance-covariance matrix of $\boldsymbol{\theta}$ given $t=0,1,\ldots$ samples have been observed
μ_i^t	Posterior mean for arm i at time t
$\Sigma_{i,j}^t$	Posterior covariance between arms i and j at time t
u^t	Action chosen at time after $t=0,1,\ldots$ samples have been observed
$\mathcal{U}$	The set of available actions
$\mathcal{D}$	Arm selected for implementation at the stopping time T
π	A policy that gives a sequence $\{u^0,u^1,\ldots\}$, a stopping time T, and an arm $\mathcal{D}$
Π	The set of all nonanticipating policies
$\Delta\in(0,1]$	Discount factor

Appendix B: Mathematical Results

B.1. Proof of Prop. 1

We begin by noting that we can replace $\mathbb{E}\left[Y_{\mathcal{D}}^{T+1}\middle|\boldsymbol{\mu}^T,\boldsymbol{\Sigma}^T\right]$ in (2) with $\theta_{\mathcal{D}}$: $\mathbb{E}\left[Y_{\mathcal{D}}^{T+1}\middle|\boldsymbol{\mu}^T,\boldsymbol{\Sigma}^T\right] = \mathbb{E}\left[\mathbb{E}\left[Y_{\mathcal{D}}^{T+1}\middle|\theta_{\mathcal{D}}\right]\middle|\boldsymbol{\mu}^T,\boldsymbol{\Sigma}^T\right]=\mathbb{E}\left[\theta_{\mathcal{D}}\middle|\boldsymbol{\mu}^T,\boldsymbol{\Sigma}^T\right]$, and $\mathbb{E}_\pi\left[\mathbb{E}\left[\theta_{\mathcal{D}}\middle|\boldsymbol{\mu}^T,\boldsymbol{\Sigma}^T\right]\middle|\boldsymbol{\mu}^0,\boldsymbol{\Sigma}^0\right]=\mathbb{E}_\pi\left[\theta_{\mathcal{D}}\middle|\boldsymbol{\mu}^0,\boldsymbol{\Sigma}^0\right]$.

We can then characterize the value function of the optimal policy, V^*, using Bellman's equation. To that end, we first define the expected net reward of selecting the best arm for implementation, given perfect information about the means, to be $\mathbb{E}\left[\max_{j\in\mathcal{M}}\{P\theta_j-I_j\}\middle|\boldsymbol{\mu}^0,\boldsymbol{\Sigma}^0\right]$. This term does not depend on the policy adopted. To link this expected value with the value function of a given policy, $V^\pi(\boldsymbol{\mu}^0,\boldsymbol{\Sigma}^0)$, we

define the opportunity cost of selecting a potentially suboptimal arm $\mathcal{D}$ to be $L_{\mathcal{D}} = \max_{j\in\mathcal{M}}\{P\theta_j - I_j\} - \Delta^T(P\theta_{\mathcal{D}} - I_{\mathcal{D}})$. Then, the value function in (2) can be rewritten as

$$V^{\pi}(\boldsymbol{\mu}^0, \boldsymbol{\Sigma}^0) = \mathbb{E}\left[\max_{j\in\mathcal{M}}\{P\theta_j - I_j\} \,\middle|\, \boldsymbol{\mu}^0, \boldsymbol{\Sigma}^0\right] - \mathbb{E}_{\pi}\left[\sum_{t=0}^{T-1}\Delta^t c_{u^t} + L_{\mathcal{D}} \,\middle|\, \boldsymbol{\mu}^0, \boldsymbol{\Sigma}^0\right]. \tag{EC.1}$$

Again note that $\mathbb{E}\left[\max_{j\in\mathcal{M}}\{P\theta_j - I_j\} \,\middle|\, \boldsymbol{\mu}^0, \boldsymbol{\Sigma}^0\right]$ does not depend on the policy π, and the expectation is finite independent of $\Delta \in (0,1]$. Therefore, $V^{\pi}(\boldsymbol{\mu}^0, \boldsymbol{\Sigma}^0)$ can be maximized by minimizing the second expectation, $\mathbb{E}_{\pi}\left[\sum_{t=0}^{T-1}\Delta^t c_{u^t} + L_{\mathcal{D}} \,\middle|\, \boldsymbol{\mu}^0, \boldsymbol{\Sigma}^0\right]$. Let $\mathcal{K}_{\pi} = \sum_{t=0}^{T-1}\Delta^t c_{u^t}$ be the discounted sampling cost, and $\mathcal{L}_{\pi} = L_{\mathcal{D}}$ be the expected opportunity cost. We can rewrite the second expectation: $\mathbb{E}_{\pi}\left[\sum_{t=0}^{T-1}\Delta^t c_{u^t} + L_{\mathcal{D}} \,\middle|\, \boldsymbol{\mu}^0, \boldsymbol{\Sigma}^0\right] = \mathbb{E}_{\pi}\left[\mathcal{K}_{\pi} + \mathcal{L}_{\pi} \,\middle|\, \boldsymbol{\mu}^0, \boldsymbol{\Sigma}^0\right]$. Since our problem is infinite horizon, and $\mathcal{K}_{\pi}$ and $\mathcal{L}_{\pi}$ are non-negative by definition, the (P) property of Bertsekas and Shreve (1978), chapter 9, is satisfied for the minimization of $\mathbb{E}_{\pi}\left[\mathcal{K}_{\pi} + \mathcal{L}_{\pi} \,\middle|\, \boldsymbol{\mu}^0, \boldsymbol{\Sigma}^0\right]$. Minimizing $\mathbb{E}_{\pi}\left[\mathcal{K}_{\pi} + \mathcal{L}_{\pi} \,\middle|\, \boldsymbol{\mu}^0, \boldsymbol{\Sigma}^0\right]$ is equivalent to maximizing $V^{\pi}(\boldsymbol{\mu}^0, \boldsymbol{\Sigma}^0)$ (as shown for the case of two arms in the proofs of Prop. 2 and 3 of Chick et al. 2017).

Prop. 9.1 of Bertsekas and Shreve (1978) shows that, given (P), an additional dependence of the state evolution on the past cannot bring additional expected reward and justifies our restriction of the policy set to Markov policies. Prop. 9.8 of Bertsekas and Shreve (1978) shows that the optimal policy in (3) satisfies Bellman's recursion,

$$V^*(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t) = \max\left\{\max_{j\in\mathcal{M}} -c_j + \Delta\mathbb{E}\left[V^*(\boldsymbol{\mu}^{t+1}, \boldsymbol{\Sigma}^{t+1}) \,\middle|\, \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t\right], \max_{j\in\mathcal{M}}\mathbb{E}\left[P\theta_j - I_j \,\middle|\, \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t\right]\right\}, \tag{EC.2}$$

and given the optimality of Markov policies $\max_{j\in\mathcal{M}}\mathbb{E}\left[P\theta_j - I_j \,\middle|\, \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t\right] = \max_{j\in\mathcal{M}}\left\{P\mu_j^t - I_j\right\}$. □

B.2. Proof of Prop. 2, Discussion on the Impact of Patient Pool Size and Fixed Costs

Note that this proposition allows for any value of $\Delta \in (0,1]$. First, we rewrite (12) using (9)

$$V_i^*(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t, c_i, P, \boldsymbol{I}, \Delta) = \sup_{\pi_i}\mathbb{E}_{\pi_i}\left[\sum_{r=0}^{T_i-1} -\Delta^r c_i + \Delta^{T_i}\max_{j\in\mathcal{M}}\left\{P\left(\mu_j^t + \frac{\Sigma_{i,j}^t}{\Sigma_{i,i}^t}Z_i^{T_i}\right) - I_j\right\} \,\middle|\, \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t\right]. \tag{EC.3}$$

Second, we rearrange terms to obtain

$$V_i^*(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t, c_i, P, \boldsymbol{I}, \Delta) = P\times\left\{\sup_{\pi_i}\mathbb{E}_{\pi_i}\left[\sum_{r=0}^{T_i-1} -\Delta^r\frac{c_i}{P} + \Delta^{T_i}\max_{j\in\mathcal{M}}\left\{\mu_j^t + \frac{\Sigma_{i,j}^t}{\Sigma_{i,i}^t}Z_i^{T_i} - \frac{I_j}{P}\right\} \,\middle|\, \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t\right]\right\} \tag{EC.4}$$

$$= P\times\left\{\sup_{\pi_i}\mathbb{E}_{\pi_i}\left[\sum_{r=0}^{T_i-1} -\Delta^r\frac{c_i}{P} + \Delta^{T_i}\max_{j\in\mathcal{M}}\left\{\left(\mu_j^t - \frac{I_j}{P}\right) + \frac{\Sigma_{i,j}^t}{\Sigma_{i,i}^t}Z_i^{T_i}\right\} \,\middle|\, \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t\right]\right\} \tag{EC.5}$$

$$= P\times V_i^*\left(\boldsymbol{\mu}^t - \frac{\boldsymbol{I}}{P}, \boldsymbol{\Sigma}^t, \frac{c_i}{P}, 1, [0,0,\ldots,0], \Delta\right). \quad \square$$

We note that the cost per sample is effectively smaller when $P > 1$, since the sampling cost is now divided across the adopting population. In Chick et al. (2019), we observed that correlation affects the optimal stopping boundary through division of the sampling cost by $b_{(2)} - b_{(1)}$. Let $\tilde{c}_i = c_i/(b_{(2)} - b_{(1)})$ be the effective sampling cost. We showed in Chick et al. (2019) that when $\tilde{c}_i < c_i$, it is optimal to stop later than in the independent case. We see that $P > 1$ similarly implies a decrease in effective sampling cost, $c_i/P < c_i$. Therefore, as P increases, the continuation region enlarges, and it is optimal to stop later.

We also observe that the fixed implementation cost of an arm impacts the problem through the expected reward of that arm. When $I_j > 0$ for arm j, the mean reward of j decreases by I_j/P. In other words, the individual level benefit of arm j decreases, and the amount of decrease equals to the fixed cost of implementation per person.

B.3. Proof of Prop. 3

Williams (1991) defines two conditions for Z_i^τ to be an uniformly integrable (UI) martingale: (i) Z_i^τ is a martingale, and (ii) $\{Z_i^\tau : \tau \in \mathbb{Z}^+\}$ is a UI family. Condition (i) holds because (10) implies $\mathbb{E}[Z_i^{\tau+1} \,|\, \mu_i^\tau] = 0$. To show that condition (ii) holds, it is sufficient to prove that Z_i^τ is bounded in $\mathcal{L}^p$, for some constant $K < \infty$; that is $\mathbb{E}[|Z_i^\tau|^p] < K$, for some $p > 0$ and for all t (Williams 1991). We pick $p = 2$, and we will show that $\mathbb{E}[|Z_i^\tau|^2] < \Sigma_{i,i}^t$ for all τ. For any normal random variable $Z \sim \mathcal{N}(\mu, \sigma^2)$, $\mathbb{E}[|Z|^2] = \mathbb{E}[Z^2] = \sigma^2 + \mu^2$. Using the distribution of Z_i^τ given in (10), $\mathbb{E}[|Z_i^\tau|^2] = \mathbb{E}[(Z_i^\tau)^2] = \sigma^2_{Z_i^\tau} = \frac{\lambda_i \tau}{n_i^t(n_i^t+\tau)} = \frac{\lambda_i}{n_i^t(n_i^t/\tau+1)}$. As $\tau \to \infty$, $\sigma^2_{Z_i^\tau}$ increases and converges to $\Sigma_{i,i}^t$. Then, for any τ, $\mathbb{E}[|Z_i^\tau|^2] = \sigma^2_{Z_i^\tau} < \Sigma_{i,i}^t < \infty$.

For the second part of the lemma, let T be a stopping time for Z_i^τ. Using the optional stopping theorem for UI martingales (Williams 1991), we can show that $\mathbb{E}_T\left[Z_i^T \,|\, \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t\right] = 0$. □

B.4. Proof of Theorem 1

Our approach to Theorem 1 takes advantage of the machinery used to prove Theorem 1 of Xie et al. (2016) (called XFC below), which demonstrates the consistency of cKG-style allocation policies. Our approach bounds $\nu_i^*(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$ below and above with the EVI's of related cKG policies, follows the proof arguments in XFC to demonstrate that limiting properties of the bounding policies follow those in XFC, and finally demonstrates that these desired properties carry over from the bounding policies to cPDE itself.

Lower and upper bounds for cPDE allocation index. We can bound $\nu_i^*(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$ below and above with the EVI of two variants of the cKG_τ policy defined in (23). In these modified policies one pays for exactly one sample but can use the information from $\tau \geq 1$ samples, where τ is fixed *a priori*, and we call the policies $\mathrm{cKG}_{1:\tau}$. As with $\nu_i^*(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$ we normalize sampling costs by dividing bi c_i:

$$\nu_i^{\mathrm{cKG}_{1:\tau}}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t) = -1 + \frac{1}{c_i}\,\mathbb{E}\left[\max_j \left\{\mu_j^t + \frac{\Sigma_{i,j}^t}{\Sigma_{i,i}^t} Z_i^\tau\right\} \;\middle|\; \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t\right] - \frac{1}{c_i}\max_j\{\mu_j^t\} \tag{EC.6}$$

$$= -1 + \frac{1}{c_i}\left[\sum_{l=1}^{M'-1} \left(b_{(l+1)} - b_{(l)}\right) \sigma_{Z_i^\tau}\, \psi\left(\frac{|d_{(l)}|}{\sigma_{Z_i^\tau}}\right)\right]. \tag{EC.7}$$

When $\tau \equiv 1$, $\nu_i^{\mathrm{cKG}_{1:\tau}}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$ is equivalent to that of the cKG_1 policy in (23) for $\tau = 1$, and we will refer to it as cKG_1. For $\tau > 1$ we will use the name $\mathrm{cKG}_{1:\tau}$.

We also note that our proofs make use of the definition of $\nu_i^*(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$ in (20), as well as of an alternative form that is derived from (14) by setting $\Delta = 0$, taking the supremum over $T_i \geq 1$, and dividing by c_i:

$$\nu_i^*(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t) = \sup_{T_i \geq 1} \mathbb{E}\left[-T_i + \frac{1}{c_i}\max_j \left\{\mu_j^t + \frac{\Sigma_{i,j}^t}{\Sigma_{i,i}^t} Z_i^{T_i}\right\} \;\middle|\; \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t\right] - \frac{1}{c_i}\max_j\{\mu_j^t\}. \tag{EC.8}$$

With these definitions, we can delineate and demonstrate the validity of our bounds.

LEMMA EC.1. *Let* $\bar{\tau} = \left\lceil \sqrt{\frac{2\min_j\{\lambda_j\}}{\pi}} \frac{\max_j\{\Sigma_{j,j}^0\}}{\min_j\{c_j\}} \right\rceil$. *Then* $\nu_i^{\mathrm{cKG}_1}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t) \leq \nu_i^*(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t) \leq \nu_i^{\mathrm{cKG}_{1:\bar{\tau}}}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$.

Proof. Beginning with (EC.8), we use (EC.6) to derive a simple lower bound as follows.

$$\begin{aligned}
\nu_i^*(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t) &= \sup_{T_i \geq 1} \mathbb{E}\left[-T_i + \frac{1}{c_i}\max_j \left\{\mu_j^t + \frac{\Sigma_{i,j}^t}{\Sigma_{i,i}^t} Z_i^{T_i}\right\} \;\middle|\; \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t\right] - \frac{1}{c_i}\max_j\{\mu_j^t\} \\
&\geq \mathbb{E}\left[-1 + \frac{1}{c_i}\max_j \left\{\mu_j^t + \frac{\Sigma_{i,j}^t}{\Sigma_{i,i}^t} Z_i^1\right\} \;\middle|\; \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t\right] - \frac{1}{c_i}\max_j\{\mu_j^t\} \\
&= \nu_i^{\mathrm{cKG}_1}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t),
\end{aligned}$$

since $T_i \equiv 1$ is a special case of the random stopping time T_i over which we take the original supremum.

We next derive a deterministic, finite stopping time, $\bar{\tau}$, for our upper bound. We begin by considering the case in which $b_i = b_j$ for all $j \in \{1, \ldots, M\}$. Since $b_j = \Sigma^t_{i,j}/\Sigma^t_{i,i}$ for all j, including i, this implies that the arm $j^* = \arg\max\{\mu^t_j\}$ maximizes $\max_j \left\{\mu^t_j + \frac{\Sigma^t_{i,j}}{\Sigma^t_{i,i}} z\right\}$ for all z, and $M' = 1$. In this case $\nu_i^{\mathrm{cKG}_1}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t) = \nu_i^*(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t) = \nu_i^{\mathrm{cKG}_{1:\tau}}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t) = -1$ for all $\tau \geq 1$, so $\nu_i^{\mathrm{cKG}_{1:\bar{\tau}}}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$ represents an upper bound on $\nu_i^*(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$ for any $\bar{\tau} \geq 1$.

For the case in which there exist terms $b_{(l+1)} - b_{(l)} > 0$, so $M' > 1$, we have $\mathbb{E}\left[\max_j \left\{\mu^t_j + \frac{\Sigma^t_{i,j}}{\Sigma^t_{i,i}} z\right\}\right] > \max_j \left\{\mu^t_j\right\}$, and we consider the Bellman equation associated with the index. Here,

$$\nu_i^*(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t) \;=\; -1 \;+ \max \begin{cases} \frac{1}{c_i}\left(\mathbb{E}\left[\max_j \left\{\mu^t_j + \frac{\Sigma^t_{i,j}}{\Sigma^t_{i,i}} Z^1_i\right\}\right] - \max_j \left\{\mu^t_j\right\}\right) \\ \mathbb{E}\left[\nu_i^*(\boldsymbol{\mu}^{t+1}, \boldsymbol{\Sigma}^{t+1})\right], \end{cases} \tag{EC.9}$$

where $\nu_i^*(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t) > -1$.

To derive $\bar{\tau}$, we develop an upper bound for the expected value of stopping, the upper maximand in (EC.9), that has three properties. First, it depends only on initial problem data and the period, t, and not the data or associated posterior statistics associated with any given sample path. Second, the upper bound is strictly decreasing in t. Finally, its limit equals zero as $t \to \infty$.

These three properties ensure that, no matter what the starting period and state $(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$, there exists some finite number of additional samples, $\bar{\tau}$, after which the upper maximand in (EC.9) falls between zero and one for all $\tau \geq \bar{\tau}$. This in turn implies that the normalized cost of sampling of -1 per period will thereafter dominate any possible improvement in the expected value of sampling, which itself must be bounded above by one. Thus, there cannot be a sample path on which it is of value to sample beyond $t + \bar{\tau}$.

We begin by constructing an upper bound on the expression within the expectation in upper maximand in (EC.9) as the expression varies with z:

$$\max_j \left\{\mu^t_j + \frac{\Sigma^t_{i,j}}{\Sigma^t_{i,i}} z\right\} \leq \max_j \left\{\mu^t_j\right\} + \left(\frac{\max_j |\Sigma^t_{i,j}|}{\Sigma^t_{i,i}}\right) |z|.$$

The right-hand side of the inequality constructs upper bounds on both the intercept and the (absolute value of the) slope of the left-hand side's convex function, as z varies above and below zero.

Next we use the upper bound on the expression to develop an analogous inequality for the expected value of stopping and proceed to derive a bound that is depends only on initial problem data and on $\Sigma^t_{i,i}$.

$$\begin{aligned}
\mathbb{E}\left[\max_j \left\{\mu^t_j + \frac{\Sigma^t_{i,j}}{\Sigma^t_{i,i}} Z^1_i\right\}\right] \;-\; \max_j \left\{\mu^t_j\right\} &\leq \mathbb{E}\left[\max_j \left\{\mu^t_j\right\} + \left(\frac{\max_j |\Sigma^t_{i,j}|}{\Sigma^t_{i,i}}\right) \left|Z^1_i\right|\right] \;-\; \max_j \left\{\mu^t_j\right\} \\
&= \mathbb{E}\left[\left|Z^1_i\right|\right] \left(\frac{\max_j |\Sigma^t_{i,j}|}{\Sigma^t_{i,i}}\right) \\
&= \mathbb{E}\left[|Z|\right] \left(\frac{\max_j |\Sigma^t_{i,j}|}{\Sigma^t_{i,i}}\right) \left(\frac{\Sigma^t_{i,i}}{\lambda_i/\Sigma^i_{i,i} + 1}\right) \\
&= \sqrt{\frac{2}{\pi}} \left(\frac{\max_j |\Sigma^t_{i,j}|}{\Sigma^t_{i,i}}\right) \left(\frac{\Sigma^t_{i,i}}{\lambda_i/\Sigma^i_{i,i} + 1}\right) \\
&\leq \sqrt{\frac{2}{\pi}} \max_j \left\{\sqrt{\Sigma^t_{j,j}}\right\} \left(\frac{\sqrt{\Sigma^t_{i,i}}}{\lambda_i/\Sigma^i_{i,i} + 1}\right)
\end{aligned}$$

$$\leq \sqrt{\frac{2}{\pi}} \max_j \left\{ \sqrt{\Sigma^0_{j,j}} \right\} \left(\frac{\sqrt{\Sigma^t_{i,i}}}{\lambda_i/\Sigma^i_{i,i}+1} \right)$$
$$\leq \sqrt{\frac{2}{\pi}} \left(\frac{\max_j \left\{ \sqrt{\Sigma^0_{j,j}} \right\}}{\min_j\{\lambda_j\}} \right) \left(\Sigma^t_{i,i}\right)^{3/2}. \tag{EC.10}$$

The first equality nets out the two $\max_j \{\mu^t_j\}$ terms. Given the standard normal random variable, Z, the second equality follows from the definition of Z^1_i in (10), and the third calculates $\mathbb{E}[|Z|] = \sqrt{2/\pi}$. By definition, the absolute value of the correlation coefficient is $\frac{|\Sigma^t_{i,j}|}{\sqrt{\Sigma^t_{i,i}\,\Sigma^t_{j,j}}} \leq 1$, so that $|\Sigma^t_{i,j}| \leq \sqrt{\Sigma^t_{i,i}\,\Sigma^t_{j,j}}$, from which the second inequality follows. Lemma 4 in XFC shows that, for any arm j, $\Sigma^{t+1}_{j,j} \leq \Sigma^t_{j,j}$ for all t which implies the third inequality. To obtain the last inequality, we drop the one in the denominator of the last term of the right-hand side, substitute $\min_j\{\lambda_j\}$ for λ_i, and rearrange terms.

Thus, with the exception of $\Sigma^t_{i,i}$, our upper bound (EC.10) depends only on initial problem data and is independent of i. Furthermore, given we sample exclusively from arm i, (8) implies that $\Sigma^{t+\tau}_{i,i} = \Sigma^t_{i,i}\left(\frac{\min_j\{\lambda_j\}}{\tau\Sigma^t_{i,i}+\min_j\{\lambda_j\}}\right)$, which is strictly decreasing in τ and converges to zero as $\tau \to \infty$. Together (EC.9) and these facts imply that it cannot be optimal to continue sampling after $t+\tau$.

Using the tighter bound $\min_j\{c_j\}$, we can then find an upper bound on the maximum number of periods to sample, $\bar{\tau}$, that is independent of the sampling cost. We then look for the smallest τ such that

$$\sqrt{\frac{2}{\pi}} \left(\frac{\max_j \left\{ \sqrt{\Sigma^0_{j,j}} \right\}}{\lambda_i} \right) \left(\Sigma^{t+\tau}_{i,i}\right)^{3/2} \leq \min_j\{c_j\}.$$

Then substituting $\Sigma^{t+\tau}_{i,i} = \Sigma^t_{i,i}\left(\frac{\min_j\{\lambda_j\}}{\tau\Sigma^t_{i,i}+\min_j\{\lambda_j\}}\right)$ and rearranging terms. we equivalently look for the smallest τ so that

$$\frac{\tau\,\Sigma^t_{i,i} + \min_j\{\lambda_j\}}{\Sigma^t_{i,i}\,\min_j\{\lambda_j\}} \geq \sqrt{\frac{2}{\pi}} \left(\frac{\max_j \left\{ \sqrt{\Sigma^0_{j,j}} \right\}}{\lambda_i} \right).$$

If we drop the $\min_j\{\lambda_j\}$ from the numerator of the left-hand side, we tighten the constraint and generate a sufficient condition for a minimum number of samples that is independent of $\Sigma^t_{i,i}$ as well. Carrying the $\min_j\{\lambda_j\}$ to the right-hand side and taking the ceiling we find our desired upper uniform bound on the number of periods,

$$\bar{\tau} \equiv \left\lceil \sqrt{\frac{2\,\min_j\{\lambda_j\}}{\pi}} \frac{\max_j \left\{ \Sigma^0_{j,j} \right\}}{\min_j\{c_j\}} \right\rceil, \tag{EC.11}$$

beyond which we do not need to sample to calculate $\nu^*_i(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$.

With the upper bound $\bar{\tau}$ in hand, we proceed to construct our upper bound on $\nu^*_i(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$. Beginning with (EC.8) we have

$$\begin{aligned}
\nu^*_i(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t) &= \sup_{T_i \geq 1} \mathbb{E}\left[-T_i + \frac{1}{c_i} \max_j \left\{ \mu^t_j + \frac{\Sigma^t_{i,j}}{\Sigma^t_{i,i}} Z^{T_i}_i \right\} \;\middle|\; \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t \right] - \frac{1}{c_i} \max_j\{\mu^t_j\} \\
&= \sup_{1 \leq T_i \leq \bar{\tau}} \mathbb{E}\left[-T_i + \frac{1}{c_i} \max_j \left\{ \mu^t_j + \frac{\Sigma^t_{i,j}}{\Sigma^t_{i,i}} Z^{T_i}_i \right\} \;\middle|\; \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t \right] - \frac{1}{c_i} \max_j\{\mu^t_j\} \\
&\leq -1 \;+ \sup_{1 \leq T_i \leq \bar{\tau}} \frac{1}{c_i} \mathbb{E}\left[\max_j \left\{ \mu^t_j + \frac{\Sigma^t_{i,j}}{\Sigma^t_{i,i}} Z^{T_i}_i \right\} \;\middle|\; \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t \right] - \frac{1}{c_i} \max_j\{\mu^t_j\} \\
&= -1 \;+ \frac{1}{c_i} \mathbb{E}\left[\max_j \left\{ \mu^t_j + \frac{\Sigma^t_{i,j}}{\Sigma^t_{i,i}} Z^{\bar{\tau}}_i \right\} \;\middle|\; \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t \right] - \frac{1}{c_i} \max_j\{\mu^t_j\} \\
&= \nu^{\mathrm{cKG}_{1:\bar{\tau}}}_i(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t).
\end{aligned}$$

The second equality holds because, beyond $\bar{\tau}$, the added sampling costs exceed the expected gains from additional information, so $T_i > \bar{\tau}$ will be realized with probability zero. The inequality holds because the sampling cost is increasing in T_i, and the third equality holds because the expected value of sampling, the term in the square brackets, is increasing in T_i. □

Convergence results for the indices of the bounding policies. The setting and policies we consider for our index bounds represent special cases of those considered in XFC, which analyzes cKG-style policies for which the number of samples used to construct an index – what they call β_n – is positive and finite. The policies used in our bounds sample either once or $1 \leq \bar{\tau} < \infty$ times by construction. While XFC considers policies for which arms need only have a positive probability of being included in the consideration set for sampling in any period, we assume that the probability equals 1 in every period, and while XFC allows for sampling from one arm or a pair of arms in any period – and constructs allocation indices for both cases – we allow only for sampling from a single arm.

One small remaining difference between the two models regards the treatment of sampling costs. Specifically, our $\nu_i^{\text{cKG}_{1:\bar{\tau}}}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$ subtracts the normalized cost of one sample, -1, from the analogous cKG indices used in XFC. This difference uniformly shifts the value of all indices we compute by -1, and rather than comparing the indices to a lower bound of 0, as done by XFC, we compare our index bounds to a benchmark of -1. With this modification, the proofs of convergence for the indices analyzed in XFC apply directly to the lower and upper bounds we have constructed above.

The first result, proven in the previous paper, provides a convergence result for the upper bound.

LEMMA EC.2. *If arm i is sampled infinitely often, then $\lim_{T\to\infty} \Sigma^T_{i,i} = 0$. In turn, if $\lim_{T\to\infty} \Sigma^T_{i,i} = 0$, then $\lim_{T\to\infty} \nu_i^{cKG_{1:\bar{\tau}}}(\boldsymbol{\mu}^T, \boldsymbol{\Sigma}^T) = -1$ as well.*

Because the previous paper considers policies that can use common random numbers to simultaneously sample more than one arm, the second result requires some additional care.

LEMMA EC.3. *If $\liminf_{T\to\infty} \nu_i^{cKG_1}(\boldsymbol{\mu}^T, \boldsymbol{\Sigma}^T) = -1$ for all i, then $\lim_{T\to\infty} \Sigma^T_{i,i} = 0$ for all i as well.*

Proof. The lemma's proof follows that in XFC, with three modifications. First, when sampling from arm i, equations (16) and (17) in the current paper defines $a_{(l)} = \mu^t_{(l)}$ and $b_{(l)} = \Sigma^t_{i,(l)}/\Sigma^t_{i,i}$, and $d_{(l)} = |(a_{(l)} - a_{(l+1)})/(b_{(l+1)} - b_{(l)})|$ so that the summand in (23) is

$$\left(b_{(l+1)} - b_{(l)}\right) \sigma_{Z_i^\tau} \psi\left(\left|\frac{a_{(l)} - a_{(l+1)}}{b_{(l+1)} - b_{(l)}}\right| / \sigma_{Z_i^\tau}\right).$$

In contrast, equation (11) in XFC defines analogous b_j's that include the standard deviation of the predictive posterior,

$$b_j = \frac{\Sigma^t_{i,j}}{\Sigma^t_{i,i}} \sigma_{Z_i^\tau} = \frac{\Sigma^t_{i,j}}{\Sigma^t_{i,i}} \frac{\Sigma^t_{i,i}}{\sqrt{\lambda_i/\tau + \Sigma^t_{i,i}}} = \frac{\Sigma^t_{i,j}}{\sqrt{\lambda_i/\tau + \Sigma^t_{i,i}}}.$$

Second, the policies in XFC also allow for simultaneous sampling of pairs of arms using simulation with common random numbers, and its equation (12) defines analogous terms for paired samples

$$\begin{bmatrix} b_i \\ b_j \end{bmatrix} = \begin{bmatrix} \Sigma^t_{i,i} - \Sigma^t_{i,j} \\ \Sigma^t_{i,j} - \Sigma^t_{j,j} \end{bmatrix} \times \frac{1}{\sqrt{P/\tau + Q^t}},$$

where $Q^t = \Sigma^t_{i,i} + \Sigma^t_{j,j} - 2\Sigma^t_{i,j}$ and where, in our setting, with sampling errors that are independent across arms, $P = \lambda_i + \lambda_j$. In both cases the analogue of our summand in (23) then becomes

$$\left(b_{(l+1)} - b_{(l)}\right)\, \psi\left(\left|\frac{a_{(l)} - a_{(l+1)}}{b_{(l+1)} - b_{(l)}}\right|\right).$$

Third, we modify the proof of Lemma 8 in XFC to use only samples of individual arms, rather than pairwise samples. In particular, at the top of the left column of page 556, the paper's proof uses statistics regarding the pairwise sample to prove that, if $\lim_{T\to\infty} \Sigma^T_{i,i} > 0$ and $\lim_{T\to\infty} \Sigma^T_{j,j} = 0$, then there exists a t^* so that $|b_i - b_j| > 0$ for all $t \geq t^*$. This fact, in turn, implies that the relevant EVI for arm i is strictly positive. In our case, samples are taken one arm at a time, so the analogous proof expressions are both simpler and slightly different.

Using the notation for b_j's provided in XFC, we have the following

$$b_i - b_j \;=\; \frac{\Sigma^t_{i,i} - \Sigma^t_{i,j}}{\sqrt{\lambda_i/\tau + \Sigma^t_{i,i}}}. \tag{EC.12}$$

By definition, the absolute value of correlation coefficient is $\left|\frac{\Sigma^t_{i,j}}{\sqrt{\Sigma^t_{i,i}\,\Sigma^t_{j,j}}}\right| \leq 1$, so for any t we have $|\Sigma^t_{i,j}| \leq \sqrt{\Sigma^t_{i,i}\,\Sigma^t_{j,j}}$, and the fact that $\lim_{T\to\infty} \Sigma^T_{i,i} > 0$ while $\lim_{T\to\infty} \Sigma^T_{j,j} = 0$ implies that both the numerator and the denominator of equation (EC.12) converge to strictly positive quantities. With this fact, rest of the lemma's proof then follows. □

Proof of the Main Result. Lemma EC.2, together with the fact that $\nu_i^{\mathrm{cKG}_{1:\bar{\tau}}}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t) \geq \nu_i^*(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$, implies the following.

LEMMA EC.4. *If arm i is sampled infinitely often, then $\lim_{T\to\infty} \Sigma^T_{i,i} = 0$. In turn, if $\lim_{T\to\infty} \Sigma^T_{i,i} = 0$, then $\lim_{T\to\infty} \nu_i^*(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t) = -1$ as well.*

Lemma EC.3, together with the fact that $\nu_i^*(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t) \geq \nu_i^{\mathrm{cKG}_1}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$, implies the following.

LEMMA EC.5. *If $\liminf_{T\to\infty} \nu_i^*(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t) = -1$ for all i, then $\lim_{T\to\infty} \Sigma^T_{i,i} = 0$ for all i as well.*

Together Lemma EC.4 and Lemma EC.5 allow us to directly apply the arguments of Theorem 1 in XFC to demonstrate the asymptotic consistency of the cPDE allocation policy. The normality of the positive definite prior $\boldsymbol{\Sigma}^0$ implies that $\arg\max\{\theta_i\}$ is uniquely defined (no ties) with probability 1. □

B.5. Proof of Prop. 4

We explicitly prove the result for $\mathrm{EVI}_i^*(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$, as defined in (19), and note that, if instead we begin with $\nu_i^*(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$, as defined in (20), we can also use the argument below to construct $\underline{\nu}_i(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$.

We first fix l. Each of the $M' - 1$ terms of the summand inside of the expectation of (19) is non-negative by construction. Thus, if we set $T_i = T_{i,l}$, the entire term in the expectation in (21) is less than or equal to the entire term in the expectation in (19). Because both of these equations take suprema over the same set of policies $\pi_i \in \Pi_i$, just using a different naming for stopping times T_i and $T_{i,l}$, and because both sample only from arm i, the inequality holds for the given l. The claimed bound then holds because the inequality is true for all $l = 1, 2, \ldots, M' - 1$. □

B.6. Proof of Prop. 5

We explicitly prove the result for $\mathrm{EVI}_i^*(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$, as defined in (19), and note that, if instead we begin with $\nu_i^*(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$, as defined in (20), we can also use the argument described below to construct $\overline{\nu}_{i,\boldsymbol{\alpha}}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$.

We proceed directly from the definition of $\mathrm{EVI}_i^*(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$.

$$\begin{aligned}
\mathrm{EVI}_i^*(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t) &= \sup_{T_i \geq 0} \mathbb{E}_{T_i}\left[-c_i T_i \;+\; \sum_{l=1}^{M'-1} \left(b_{(l+1)} - b_{(l)}\right)\left(-|d_{(l)}| + Z_i^{T_i}\right)^+ \;\middle|\; \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t\right] \\
&= \sup_{T_i \geq 0} \mathbb{E}_{T_i}\left[\sum_{l=1}^{M'-1} -c_i \alpha_l T_i \;+\; \left(b_{(l+1)} - b_{(l)}\right)\left(-|d_{(l)}| + Z_i^{T_i}\right)^+ \;\middle|\; \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t\right] \\
&= \sup_{T_i \geq 0} \sum_{l=1}^{M'-1} \mathbb{E}_{T_i}\left[-c_i \alpha_l T_i \;+\; \left(b_{(l+1)} - b_{(l)}\right)\left(-|d_{(l)}| + Z_i^{T_i}\right)^+ \;\middle|\; \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t\right] \\
&\leq \sum_{l=1}^{M'-1} \sup_{T_{i,l} \geq 0} \mathbb{E}_{T_{i,l}}\left[-c_i \alpha_l T_i \;+\; \left(b_{(l+1)} - b_{(l)}\right)\left(-|d_{(l)}| + Z_i^{T_i}\right)^+ \;\middle|\; \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t\right] \\
&= \overline{\mathrm{EVI}}_{i,\boldsymbol{\alpha}}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)
\end{aligned}$$

The first line holds by definition. The second line follows because $\sum_{l=1}^{M'-1} \alpha_l = 1$. The third line follows from the linearity of expectation. The inequality of the fourth line holds because the supremum of a sum is less than or equal to the sum of the individual suprema. Furthermore, for each summand, samples are only taken from arm i, the Z_i are independent, and the value of l is fixed. So we can label the stopping time T_i by $T_{i,l}$ in this setting. The sixth line is by definition. □

B.7. Proof of Prop. 6

We explicitly prove the result for $\overline{\mathrm{EVI}}_{i,\boldsymbol{\alpha}}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$ and note that, if instead we begin with $\overline{\nu}_{i,\boldsymbol{\alpha}}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$, we can also use the argument below to prove the convexity of $\overline{\nu}_{i,\boldsymbol{\alpha}}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$ with respect to $\boldsymbol{\alpha}$.

Let the function $f(\boldsymbol{\alpha}) \equiv \overline{\mathrm{EVI}}_{i,\boldsymbol{\alpha}}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$ for any $(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$, then

$$f(\boldsymbol{\alpha}) \;=\; \sum_{l=1}^{M'-1} \sup_{T_{i,l}} \mathbb{E}_{T_{i,l}}\left[-c_i \alpha_l T_{i,l} \;+\; \left(b_{(l+1)} - b_{(l)}\right)\left(-|d_{(l)}| + Z_i^{T_{i,l}}\right)^+ \;\middle|\; \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t\right].$$

Suppose we have two cost allocations, $\boldsymbol{\alpha^1}$ and $\boldsymbol{\alpha^2}$, along with a weighting $0 < \gamma < 1$. Then

$$\begin{aligned}
&f(\gamma\boldsymbol{\alpha^1} + (1-\gamma)\boldsymbol{\alpha^2}) \\
&\quad = \sum_{l=1}^{M'-1} \sup_{T_{i,l}} \mathbb{E}_{T_{i,l}}\left[-(\gamma\alpha_l^1 + (1-\gamma)\alpha_l^2) c_i T_{i,l} \;+\; \left(b_{(l+1)} - b_{(l)}\right)\left(-|d_{(l)}| + Z_i^{T_{i,l}}\right)^+ \;\middle|\; \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t\right].
\end{aligned}$$

Let $\{T_l^\gamma \mid l = 1, 2, \ldots, M'-1\}$ be the set of stopping times that maximize $f(\gamma\boldsymbol{\alpha^1} + (1-\gamma)\boldsymbol{\alpha^2})$, so that

$$\begin{aligned}
f(\gamma\boldsymbol{\alpha^1} + (1-\gamma)\boldsymbol{\alpha^2}) &= \sum_{l=1}^{M'-1} \mathbb{E}\left[-(\gamma\alpha_l^1 + (1-\gamma)\alpha_l^2) c_i T_l^\gamma + \left(b_{(l+1)} - b_{(l)}\right)\left(-|d_{(l)}| + Z_i^{T_l^\gamma}\right)^+ \;\middle|\; \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t\right] \\
&= \sum_{l=1}^{M'-1} \mathbb{E}\left[-(\gamma\alpha_l^1 + (1-\gamma)\alpha_l^2) c_i T_l^\gamma \;+\; \left(\gamma(b_{(l+1)} - b_{(l)}) + (1-\gamma)(b_{(l+1)} - b_{(l)})\right)\left(-|d_{(l)}| + Z_i^{T_l^\gamma}\right)^+ \;\middle|\; \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t\right] \\
&= \sum_{l=1}^{M'-1} \mathbb{E}\left[-\gamma\alpha_l^1 c_i T_l^\gamma \;+\; \gamma\left(b_{(l+1)} - b_{(l)}\right)\left(-|d_{(l)}| + Z_i^{T_l^\gamma}\right)^+ \;\middle|\; \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t\right]
\end{aligned}$$

$$
\begin{aligned}
&+ \sum_{l=1}^{M'-1} \mathbb{E}\left[-(1-\gamma)\alpha_l^2 c_i T_l^{\gamma} + (1-\gamma)(b_{(l+1)} - b_{(l)})\left(-|d_{(l)}| + Z_i^{T_l^{\gamma}}\right)^+ \;\middle|\; \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t\right] \\
&= \gamma \sum_{l=1}^{M'-1} \mathbb{E}\left[-\alpha_l^1 c_i T_l^{\gamma} \;+\; (b_{(l+1)} - b_{(l)})\left(-|d_{(l)}| + Z_i^{T_l^{\gamma}}\right)^+ \;\middle|\; \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t\right] \\
&\quad + (1-\gamma) \sum_{l=1}^{M'-1} \mathbb{E}\left[-\alpha_l^2 c_i T_l^{\gamma} + (b_{(l+1)} - b_{(l)})\left(-|d_{(l)}| + Z_i^{T_l^{\gamma}}\right)^+ \;\middle|\; \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t\right] \\
&\le \gamma \sum_{l=1}^{M'-1} \sup_{T_{i,l}} \mathbb{E}_{T_{i,l}}\left[-\alpha_l^1 c_i T_{i,l} \;+\; (b_{(l+1)} - b_{(l)})\left(-|d_{(l)}| + Z_i^{T_{i,l}}\right)^+ \;\middle|\; \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t\right] \\
&\quad + (1-\gamma) \sum_{l=1}^{M'-1} \sup_{T_{i,l}} \mathbb{E}_{T_{i,l}}\left[-\alpha_l^2 c_i T_{i,l} + (b_{(l+1)} - b_{(l)})\left(-|d_{(l)}| + Z_i^{T_{i,l}}\right)^+ \middle|\; \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t\right] \\
&= \gamma f(\boldsymbol{\alpha^1}) + (1-\gamma) f(\boldsymbol{\alpha^2}) \quad \square.
\end{aligned}
$$

We note that the proof does not depend on the discretization of the stopping times.

B.8. Proof of Corollary 1

The bounding strategy of Theorem 1 is robust and easily can be adapted to prove analogous consistency results for the cPDELower and cPDEUpper allocation policies as well. In particular, we can use transformations of the cKG-style policies used to prove Theorem 1 to generate a suitable lower bound for the allocation index of cPDELower and a suitable upper bound for the allocation index of cPDEUpper. With these bounds we can again apply our proof approach to show that cPDELower's and cPDEUpper's indices inherit the properties of the bounds.

Bounds on the allocation indices of cPDELower and cPDEUpper. To prove that cPDELower and cPDEUpper are asymptotically consistent allocation policies, we develop an upper and lower bound on their allocation indices that scale both one-period sampling costs and the information gains from sampling, based on the number of arms, M. We keep track of these changes by extending the naming conventions for cPDE and the cKG policies, used in the proof of Theorem 1, to include scaling factors.

$$
\begin{aligned}
\nu_i^{\mathrm{cKG}_{1\cdot x:\tau}}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t) &= -x \;+\; \frac{1}{c_i}\,\mathbb{E}\left[\max_j \left\{\mu_j^t + \frac{\Sigma_{i,j}^t}{\Sigma_{i,i}^t} Z_i^{\tau}\right\} \;\middle|\; \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t\right] - \frac{1}{c_i}\max_j\{\mu_j^t\} && \text{(EC.13)} \\
&= -x \;+\; \frac{1}{c_i}\left[\sum_{l=1}^{M'-1} \left(b_{(l+1)} - b_{(l)}\right) \sigma_{Z_i^{\tau}}\, \psi\left(\frac{|d_{(l)}|}{\sigma_{Z_i^{\tau}}}\right)\right], && \text{(EC.14)}
\end{aligned}
$$

where the normalized one-period sampling cost, -1, is scaled by x. Again, when $\tau \equiv 1$, $\nu_i^{\mathrm{cKG}_{1\cdot x:\tau}}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$ is equivalent to that of the cKG_{τ} policy in (23) with $\tau = 1$ and per-period sampling cost up by a factor of x, and we will refer to it as $\mathrm{cKG}_{1\cdot x}$. For $\tau > 1$ we will use the name $\mathrm{cKG}_{1\cdot x:\tau}$.

In deriving our bounds we will also use a version of (EC.8) with scaled sampling costs

$$
\nu_i^{*\cdot x}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t) \;\equiv\; \sup_{\tau \ge 1} \mathbb{E}\left[-x\cdot\tau + \frac{1}{c_i}\max_j\left\{\mu_j^t + \frac{\Sigma_{i,j}^t}{\Sigma_{i,i}^t} Z_i^{\tau}\right\} \;\middle|\; \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t\right] - \frac{1}{c_i}\max_j\{\mu_j^t\}. \qquad \text{(EC.15)}
$$

With these definitions, we proceed to construct bounds for $\underline{\nu}_i(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$ and $\overline{\nu}_{i,\boldsymbol{\alpha}}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$.

LEMMA EC.6. *Let* $\bar{\tau} = \left\lceil \sqrt{\frac{2\min_j\{\lambda_j\}}{\pi}}\, \frac{\max_j\{\Sigma_{j,j}^0\}}{\min_j\{c_j/M\}} \right\rceil$. *Then*

$$
\frac{1}{M}\cdot \nu_i^{cKG_{1\cdot M}}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t) \le \underline{\nu}_i(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t) \le \overline{\nu}_{i,\boldsymbol{\alpha}}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t) \le M \cdot \nu_i^{cKG_{1\cdot \frac{1}{M}:\bar{\tau}}}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t).
$$

Proof. We begin with $\underline{\nu}_i(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$. The allocation-index analogue of (21) is

$$\begin{aligned}
\underline{\nu}_i(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t) &= \max_{l=1,\ldots,M'-1} \left\{ \sup_{T_{i,l}\geq 1} \mathbb{E}_{T_{i,l}} \left[-T_{i,l} \; + \; \frac{1}{c_i} \left(b_{(l+1)} - b_{(l)}\right) \left(-|d_{(l)}| + Z_i^{T_{i,l}}\right)^+ \;\middle|\; \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t \right] \right\} \\
&\geq \left(\frac{1}{M'-1}\right) \sum_{l=1}^{M'-1} \left\{ \sup_{T_{i,l}\geq 1} \mathbb{E}_{T_{i,l}} \left[-T_{i,l} \; + \; \frac{1}{c_i} \left(b_{(l+1)} - b_{(l)}\right) \left(-|d_{(l)}| + Z_i^{T_{i,l}}\right)^+ \;\middle|\; \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t \right] \right\} \\
&\geq \left(\frac{1}{M'-1}\right) \sup_{T_i\geq 1} \sum_{l=1}^{M'-1} \mathbb{E}_{T_i} \left[-T_i \; + \; \frac{1}{c_i} \left(b_{(l+1)} - b_{(l)}\right) \left(-|d_{(l)}| + Z_i^{T_i}\right)^+ \;\middle|\; \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t \right] \\
&= \left(\frac{1}{M'-1}\right) \sup_{T_i\geq 1} \mathbb{E}_{T_i} \left[-(M'-1)\, T_i \; + \; \frac{1}{c_i} \sum_{l=1}^{M'-1} \left(b_{(l+1)} - b_{(l)}\right) \left(-|d_{(l)}| + Z_i^{T_i}\right)^+ \;\middle|\; \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t \right] \\
&\geq \left(\frac{1}{M}\right) \sup_{T_i\geq 1} \mathbb{E}_{T_i} \left[-M\, T_i \; + \; \frac{1}{c_i} \sum_{l=1}^{M'-1} \left(b_{(l+1)} - b_{(l)}\right) \left(-|d_{(l)}| + Z_i^{T_i}\right)^+ \;\middle|\; \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t \right] \\
&\equiv \frac{1}{M}\, \nu_i^{*\cdot M}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t),
\end{aligned}$$

where $\nu_i^{*\cdot M}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$ is the allocation index of cPDE when normalized sampling costs are multiplied by M.

The first inequality follows from the fact that the average of all $M'-1$ expectations cannot be greater than max of those expectations. The second inequality reflects the fact that optimizing the sum of the expectations with a single stopping time cannot outperform the sum of optimizing each expectation with a separate stopping time. The second equality is due to the additive nature of expectations. The third inequality holds because $\frac{1}{M'-1}\,(M'-1)\,T_i = \frac{1}{M}\,M\,T_i$, while $\frac{1}{M'-1}\,\frac{1}{c_i}\sum_{l=1}^{M'-1}(b_{(l+1)}-b_{(l)})\left(-|d_{(l)}|+Z_i^{T_i}\right)^+ \geq \frac{1}{M}\,\frac{1}{c_i}\sum_{l=1}^{M'-1}(b_{(l+1)}-b_{(l)})\left(-|d_{(l)}|+Z_i^{T_i}\right)^+$.

As in Lemma EC.1 we then have

$$\begin{aligned}
\frac{1}{M}\,\nu_i^{*\cdot M}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t) &= \frac{1}{M} \left(\sup_{\tau\geq 1} \mathbb{E} \left[-M\, T_i \; + \; \frac{1}{c_i} \max_j \left\{ \mu_j^t + \frac{\Sigma_{i,j}^t}{\Sigma_{i,i}^t} Z_i^\tau \right\} \;\middle|\; \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t \right] - \frac{1}{c_i} \max_j \{\mu_j^t\} \right) \\
&\geq \frac{1}{M} \left(\mathbb{E} \left[-M \; + \; \frac{1}{c_i} \max_j \left\{ \mu_j^t + \frac{\Sigma_{i,j}^t}{\Sigma_{i,i}^t} Z_i^1 \right\} \;\middle|\; \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t \right] - \frac{1}{c_i} \max_j \{\mu_j^t\} \right) \\
&\equiv \frac{1}{M} \cdot \nu_i^{\mathrm{cKG}_{1\cdot M}}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t),
\end{aligned}$$

$1/M$ times the allocation index for cKG_1, given normalized sampling costs that are scaled up by a factor of M, our lower bound on $\underline{\nu}_i(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$.

We continue with the upper bound for the allocation index and begin by assuming that $M' > 1$. Here, the allocation-index analogue of (22) is

$$\begin{aligned}
\overline{\nu}_{i,\boldsymbol{\alpha}}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t) &= \sum_{l=1}^{M'-1} \left\{ \sup_{T_{i,l}\geq 1} \mathbb{E}_{T_{i,l}} \left[-\alpha_i\, T_{i,l} \; + \; \frac{1}{c_i} \left(b_{(l+1)} - b_{(l)}\right) \left(-|d_{(l)}| + Z_i^{T_{i,l}}\right)^+ \;\middle|\; \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t \right] \right\} \\
&\leq \sum_{l=1}^{M'-1} \left\{ \sup_{T_{i,l}\geq 1} \mathbb{E}_{T_{i,l}} \left[-\alpha_i\, T_{i,l} \; + \; \frac{1}{c_i} \sum_{k=1}^{M'-1} \left(b_{(k+1)} - b_{(k)}\right) \left(-|d_{(k)}| + Z_i^{T_{i,l}}\right)^+ \;\middle|\; \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t \right] \right\} \\
&= (M'-1) \sup_{T_i\geq 1} \mathbb{E}_{T_i} \left[-\frac{1}{M'-1}\, T_i \; + \; \frac{1}{c_i} \sum_{k=1}^{M'-1} \left(b_{(k+1)} - b_{(k)}\right) \left(-|d_{(k)}| + Z_i^{T_i}\right)^+ \;\middle|\; \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t \right] \\
&\leq M \cdot \left(\sup_{T_i\geq 1} \mathbb{E} \left[-\frac{T_i}{M} \; + \; \frac{1}{c_i} \max_j \left\{ \mu_j^t + \frac{\Sigma_{i,j}^t}{\Sigma_{i,i}^t} Z_i^{T_i} \right\} \;\middle|\; \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t \right] - \frac{1}{c_i} \max_j \{\mu_j^t\} \right) \\
&\equiv M \cdot \nu_i^{*\cdot \frac{1}{M}}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t),
\end{aligned} \tag{EC.16}$$

M times the allocation index of cPDE, given normalized sampling costs that are scaled by a factor $\frac{1}{M}$.

The first inequality holds because, as we introduce the inner summation, we are adding positive summands. The second equality follows from the fact that, given the inclusion of the inner summation, each $\sup_{T_{i,l}\geq 1}$ within the outer summation optimizes over the same inner summation, and we can use single stopping time, T_i, for all $(M'-1)$ identical optimization problems within the outer summation. The outer summation then sums the per-period sampling costs across their α_i's and scales up the inner summation by $(M'-1)$. The second inequality follows from the fact that $(M'-1)\frac{T_i}{M'-1} = M\frac{T_i}{M}$, while $(M'-1)\left(\max_j\left\{\mu_j^t + \frac{\Sigma_{i,j}^t}{\Sigma_{i,i}^t}Z_i^\tau\right\} - \max_j\{\mu_j^t\}\right) \leq M\left(\max_j\left\{\mu_j^t + \frac{\Sigma_{i,j}^t}{\Sigma_{i,i}^t}Z_i^\tau\right\} - \max_j\{\mu_j^t\}\right)$.

To determine an upper bound on the maximum number of periods to sample, we scale unit sampling costs by $x = \frac{1}{M}$ in (EC.15). Then using the same argument that generated (EC.11), we have

$$\bar{\tau} \equiv \left\lceil \sqrt{\frac{2\min_j\{\lambda_j\}}{\pi}}\frac{\max_j\{\Sigma_{j,j}^0\}}{\min_j\{c_j/M\}} \right\rceil \tag{EC.17}$$

for $\overline{\nu}_{i,\boldsymbol{\alpha}}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$.

In turn, we continue with (EC.16) and use our definition of $\bar{\tau}$ to proceed as in Lemma EC.1.

$$\begin{aligned}
M\cdot\nu_i^{*,\frac{1}{M}}(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t) &= M\cdot\left(\sup_{T_i\geq 1}\mathbb{E}\left[-\frac{T_i}{M} + \frac{1}{c_i}\max_j\left\{\mu_j^t + \frac{\Sigma_{i,j}^t}{\Sigma_{i,i}^t}Z_i^{T_i}\right\} \;\middle|\; \boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t\right] - \frac{1}{c_i}\max_j\{\mu_j^t\}\right)\\
&= M\cdot\left(\sup_{1\leq T_i\leq\bar{\tau}}\mathbb{E}\left[-\frac{T_i}{M} + \frac{1}{c_i}\max_j\left\{\mu_j^t + \frac{\Sigma_{i,j}^t}{\Sigma_{i,i}^t}Z_i^{T_i}\right\} \;\middle|\; \boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t\right] - \frac{1}{c_i}\max_j\{\mu_j^t\}\right)\\
&\leq M\cdot\left(\sup_{1\leq T_i\leq\bar{\tau}}\mathbb{E}\left[-\frac{1}{M} + \frac{1}{c_i}\max_j\left\{\mu_j^t + \frac{\Sigma_{i,j}^t}{\Sigma_{i,i}^t}Z_i^{T_i}\right\} \;\middle|\; \boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t\right] - \frac{1}{c_i}\max_j\{\mu_j^t\}\right)\\
&\leq M\cdot\left(\mathbb{E}\left[-\frac{1}{M} + \frac{1}{c_i}\max_j\left\{\mu_j^t + \frac{\Sigma_{i,j}^t}{\Sigma_{i,i}^t}Z_i^{\bar{\tau}}\right\} \;\middle|\; \boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t\right] - \frac{1}{c_i}\max_j\{\mu_j^t\}\right)\\
&\equiv M\cdot\nu_i^{\mathrm{cKG}_{1\cdot\frac{1}{M}:\bar{\tau}}}(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t).
\end{aligned}$$

The second equality holds because, beyond $\bar{\tau}$, the added sampling costs exceed the expected gains from additional information, so $T_i > \bar{\tau}$ will be realized with probability zero. The first inequality reflects the fact that sampling costs are increasing in T_i. The second inequality follows from the fact that the sampling cost is fixed, for a single period, while the expected information benefit of sampling is increasing in T_i.

Finally, as in Lemma EC.1, we consider the case in which $M'=1$. Here, $b_i = b_j$ for all $j\in\{1,\ldots,M\}$. Since $b_j = \Sigma_{i,j}^t/\Sigma_{i,i}^t$ for all j, including i, the arm $j^* = \arg\max\{\mu_j^t\}$ maximizes $\max_j\left\{\mu_j^t + \frac{\Sigma_{i,j}^t}{\Sigma_{i,i}^t}z\right\}$ for all z, $M'=1$, and

$$\frac{1}{M}\cdot\nu_i^{\mathrm{cKG}_{1\cdot M}}(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t) = \underline{\nu}_i(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t) = \overline{\nu}_{i,\boldsymbol{\alpha}}(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t) = M\cdot\nu_i^{\mathrm{cKG}_{1\cdot\frac{1}{M}:\bar{\tau}}}(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t) = -1$$

for all $\bar{\tau}\geq 1$. Thus, the bounds are trivially satisfied in this case as well. □

Limiting Behavior and Proof of Corollary 1. Having constructed appropriate lower and upper bounds, the limiting results we developed for $\nu_i^*(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t)$ carry over to $\underline{\nu}_i(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t)$ and $\overline{\nu}_{i,\boldsymbol{\alpha}}(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t)$ as well.

LEMMA EC.7.

i) If arm i is sampled infinitely often, then $\lim_{T\to\infty}\Sigma_{i,i}^T = 0$. In turn, if $\lim_{T\to\infty}\Sigma_{i,i}^T = 0$, then $\lim_{T\to\infty} M\cdot\nu_i^{cKG_{1\cdot\frac{1}{M}:\bar{\tau}}}(\boldsymbol{\mu}^T,\boldsymbol{\Sigma}^T) = -1$ as well.

ii) If $\liminf_{T\to\infty} \frac{1}{M}\cdot \nu_i^{cKG_{1\cdot M}}(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t) = -1$ *for all* i*, then* $\lim_{T\to\infty}\Sigma^T_{i,i}=0$ *for all* i *as well.*

Proof. For part (i) we note that the arguments of Lemma EC.4 directly apply to $\nu_i^{\mathrm{cKG}_{1\cdot\frac{1}{M}:\bar{\tau}}}(\boldsymbol{\mu}^T,\boldsymbol{\Sigma}^T)$ to demonstrate that $\lim_{T\to\infty}\nu_i^{\mathrm{cKG}_{1\cdot M:\bar{\tau}}}(\boldsymbol{\mu}^T,\boldsymbol{\Sigma}^T) = -\frac{1}{M}$ as well. If instead we multiply $\nu_i^{\mathrm{cKG}_{1\cdot\frac{1}{M}:\bar{\tau}}}(\boldsymbol{\mu}^T,\boldsymbol{\Sigma}^T)$ by M and let $T\to\infty$ we obtain the desired limit.

Analogously, for part (ii) we note that the arguments of Lemma EC.5 directly apply to $\nu_i^{\mathrm{cKG}_{1\cdot M}}(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t)$ to demonstrate that, if $\liminf_{T\to\infty}\nu_i^{\mathrm{cKG}_{1\cdot M}}(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t) = -M$ for all i, then $\lim_{T\to\infty}\Sigma^T_{i,i}=0$ for all i as well. If instead we divide $\nu_i^{\mathrm{cKG}_{1\cdot M}}(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t)$ by M and let $t\to\infty$ we obtain the desired limit. □

As before, the bounds of Lemma EC.6 imply that the limiting behavior of Lemma EC.7 also carries over to cPDEUpper and cPDELower.

Lemma EC.8.

i) If arm i *is sampled infinitely often, so* $\lim_{T\to\infty}\Sigma^T_{i,i}=0$*, then* $\lim_{T\to\infty}\overline{\nu}_{i,\boldsymbol{\alpha}}(\boldsymbol{\mu}^T,\boldsymbol{\Sigma}^T) = -1$ *as well.*

ii) If $\liminf_{T\to\infty}\overline{\nu}_{i,\boldsymbol{\alpha}}(\boldsymbol{\mu}^T,\boldsymbol{\Sigma}^T) = -1$ *for all* i*, then* $\lim_{T\to\infty}\Sigma^T_{i,i}=0$ *for all* i *as well.*

iii) If arm i *is sampled infinitely often, so* $\lim_{T\to\infty}\Sigma^T_{i,i}=0$*, then* $\lim_{T\to\infty}\underline{\nu}_i(\boldsymbol{\mu}^T,\boldsymbol{\Sigma}^T) = -1$ *as well.*

iv) If $\liminf_{T\to\infty}\underline{\nu}_i(\boldsymbol{\mu}^T,\boldsymbol{\Sigma}^T) = -1$ *for all* i*, then* $\lim_{T\to\infty}\Sigma^T_{i,i}=0$ *for all* i *as well.*

Proof. Parts (i) and (iii) follow part (i) of Lemma EC.7, along with the fact that $\underline{\nu}_i(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t) \le \overline{\nu}_{i,\boldsymbol{\alpha}}(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t) \le M\cdot\nu_i^{\mathrm{cKG}_{1\cdot\frac{1}{M}:\bar{\tau}}}(\boldsymbol{\mu}^T,\boldsymbol{\Sigma}^T)$. Parts (ii) and (iv) follow part (ii) of Lemma EC.7, along with the fact that $\frac{1}{M}\cdot\nu_i^{\mathrm{cKG}_{1\cdot M}}(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t) \le \underline{\nu}_i(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t) \le \overline{\nu}_{i,\boldsymbol{\alpha}}(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t)$. □

Together parts (i) and (ii) of Lemma EC.8 can be used directly in the proof of Theorem 1 in XFC to prove the claimed consistency results for cPDEUpper. Similarly, parts (iii) and (iv) of Lemma EC.8 can be used in the proof of Theorem 1 in XFC to prove the consistency of cPDELower. □

B.9. Proof of Prop. 7

The proof follows directly by observing that the set of KG policies, indexed by $\tau = 1,2,\ldots$, is a subset of nonanticipative policies. □

B.10. Proof of Corollary 2

Recall from Assumption 1 that each arm $i\in\{1,\ldots,M\}$ need not be considered for sampling in every period, only that, almost surely, it be considered infinitely often as the stopping horizon increases without bound: $\mathbb{P}\{\lim_{T\to\infty}\sum_{t=1}^T \mathbb{1}\{i\in\mathcal{M}_t\}=\infty\}=1$. The allocation policies cPDE, cPDEUpper, and cPDELower include all arms in every $\mathcal{M}_t$, so they trivially satisfy the condition.

Randomized versions of these policies can similarly satisfy the inclusion condition. One simple approach for deciding which arms should be a part of $\mathcal{M}_t$ for any of the allocation policies is to calculate each arm's allocation index *before* deciding which to include in $\mathcal{M}_t$.

Consider the first randomization scheme of Section 7.2. At each time t we use random allocation to assign exactly one arm to $\mathcal{M}_t$. With probability p we make a random allocation at time t so that, given a random allocation, each of the M arm has probability $1/M$ of being selected randomly for inclusion in $\mathcal{M}_t$. In addition, with probability $(1-p)$ we select an arm with a maximal allocation index at time t, and, given an

index-maximizing allocation, each of the $m^t_{\max} \in \{1,\ldots,M\}$ maximizing arms has a probability $1/m^t_{\max}$ of being selected.

All together, in each period, t, the randomization scheme includes each arm i as the one and only element of $\mathcal{M}_t$ with probability

$$p^t_i = \begin{cases} \frac{p}{M} & \text{if } i \text{ does not have a maximal allocation index in period } t \\ \frac{p}{M} + \frac{1-p}{m^t_{\max}} & \text{if } i \text{ has a maximal allocation index in period } t, \end{cases}$$

and then samples that arm with probability one. Thus, every arm has a positive probability of at least p/M of being included in each consideration set $\mathcal{M}_t$. In fact, each arm has a strictly positive probability of being sampled in each period t.

By the second Borel-Cantelli Lemma, $\lim_{T\to\infty}\sum_{t=1}^{T}\frac{p}{M}=\infty$ implies $\mathbb{P}\left\{\lim_{T\to\infty}\sum_{t=1}^{T}\mathbb{1}\{i\in\mathcal{M}_t\}=\infty\right\}=1$, satisfying part (iii) of Assumption 1. Therefore, the randomized cPDE, cPDEUpper, and cPDELower allocation policies all continue to satisfy the conditions that assure consistency.

B.11. Consistency of TTVS Versions of cPDE, cPDELower, and cPDEUpper

We consider the application of top-two value-sampling (TTVS) randomization (Russo 2020) to cPDE, cPDE-Lower, and cPDEUpper allocation policies and demonstrate that it maintains property (iii) of Assumption 1, that $\mathbb{P}\left\{\lim_{T\to\infty}\sum_{t=1}^{T}\mathbf{1}\{i\in\mathcal{M}_t\}=\infty\right\}=1$. This implies that the TTVS versions of these policies are asymptotically consistent. Here, we explicitly demonstrate the consistency of the cPDE allocation index, $\nu^*_i(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t)$, and we note that the same argument holds for the other indices as well.

We operationalize TTVS randomization as follows. We first calculate the cPDE allocation indices of all M arms and order them from largest to smallest, breaking ties lexicographically. We call the ordered indices $\nu^*_{(1)}(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t)\geq\nu^*_{(2)}(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t)\geq\cdots\geq\nu^*_{(M)}(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t)$. We then remove the top-ranked arm, with index $\nu^*_{(1)}(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t)$, from the consideration set $\mathcal{M}_t$ with probability p, and we leave it in with probability $1-p$. Thus, every arm has a probability of at least $1-p$ of inclusion in every period.

In turn, for $p<1$ we have $1-p>0$, and $\lim_{T\to\infty}\sum_{t=1}^{T}(1-p)=\infty$, and the second Borel-Cantelli Lemma implies that $\mathbb{P}\left\{\lim_{T\to\infty}\sum_{t=1}^{T}\mathbb{1}\{i\in\mathcal{M}_t\}=\infty\right\}=1$, satisfying part (iii) of Assumption 1. Therefore, the TTVS randomized cPDE allocation policy continues to satisfy the conditions that assure consistency, and the same argument applies to cPDEUpper and cPDELower as well. □

Appendix C: Implementation Details: New Allocation Indices and Stopping Indices

This section provides additional details for the implementation for our new allocation indices and stopping indices. Appendix C.1 discusses the selection of optimal weights to find the smallest value of cPDEUpper, the useful upper bound on cPDE. Appendix C.2 compares the values of the EVI's used in the EVI-based allocation indices and stopping indices (cPDE-type and cKG-type policies), to give a sense of the tightness of the various bounds which define those policies. Appendix C.3 discusses computation times to solve the partial differential equation (PDE) to compute the EVI and allocation index for cPDE, and offers an algorithm to speed the computation of the cPDE stopping time. Appendix C.4 gives the algorithm we used to determine the Gauassian process regression estimate using data from a pilot study, to support the selection of a prior distribution for the dose-finding trial in Section 6.4.

C.1. Weights for the cPDEUpper Policy

Prop. 5 shows that $\overline{\text{EVI}}_{i,\boldsymbol{\alpha}}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$ is an upper bound on $\text{EVI}_i^*(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$ for any set of weights, $\boldsymbol{\alpha} \geq 0$, that sum to 1. Prop. 6 shows there is a least upper bound $\boldsymbol{\alpha}^*$ obtainable by the convexity of $\overline{\text{EVI}}_{i,\boldsymbol{\alpha}}$. The results of this section suggest that optimizing $\boldsymbol{\alpha}^*$ does not necessarily offer large gains relative to using equal weights with $\alpha_l = \alpha_k$ for $l, k \in \{1, \ldots, M'\}$, as this optimization adds computation time to the algorithm without significant benefit in a numerical application. It would only be recommended to consider optimizing $\boldsymbol{\alpha}$ for settings where observations are particularly expensive.

We test different weights to calculate the EVI of cPDEUpper and present results for one of the experiments detailed in Section 6.2. We set $M = 80$, $P = 10^4$, $I_i = 0$, $\lambda_i = 0.01$, $\mu_i^0 = 0$, and $c_i = 1$ for all i, with prior covariance determine by $\sigma^2 = 0.5$, $\zeta = 100/(80-1)^2$.

Table EC.2 The expected sample size, E[T], expected opportunity cost, E[OC], and expected total cost, E[TC], for the cPDEUpper allocation policy with equal weights and weights $\boldsymbol{\alpha}^*$ for the problem in Section 6.2 with $\zeta = 100/(80-1)^2$ and $P = 10^4$ for all i.

Allocation	Stopping	E[T] ± S.E.	E[OC] ± S.E.	E[TC] ± S.E.
cPDEUpper ($\boldsymbol{\alpha}^*$)	cKG$_*$	12 ± 0.20	96 ± 7.44	108 ± 7.46
cPDEUpper (Equal)	cKG$_*$	14 ± 0.28	91 ± 6.97	105 ± 6.97
cPDEUpper ($\boldsymbol{\alpha}^*$)	cPDELower	17 ± 0.36	68 ± 5.60	86 ± 5.63
cPDEUpper (Equal)	cPDELower	26 ± 0.68	61 ± 5.64	87 ± 5.64

Table EC.2 presents the results of experiments that aim to measure the value of optimization over $\boldsymbol{\alpha}$. The cPDEUpper (Equal) allocation index uses equal weights to allocate sampling costs. To obtain the optimal weight vector $\boldsymbol{\alpha}^*$ for cPDEUpper ($\boldsymbol{\alpha}^*$), we use Matlab function *fmincon*. We experiment with different numbers of optimization iterations and found that 8 iterations generated results that are close to the optimal weights for this particular problem. At each time $t = 0, 1, \ldots$, we calculate weights using 8 iterations of fmincon function and use resulting weight vector as $\boldsymbol{\alpha}^*$ for the cPDEUpper ($\boldsymbol{\alpha}^*$) allocation index. The optimization adds considerably to run time, however. Table EC.2 shows that the performance does not significantly improve beyond that with equal weights. Therefore, we use cPDEUpper (Equal) in our experiments.

C.2. Comparison of the EVSI for Several Stopping Times

Each heuristic introduced in Section 4.1 (cPDE, cPDELower, cPDEUpper, cKG$_*$, and cKG$_1$) uses a different set of lookaheads to approximate the EVSI of further sampling. In this section, we compare the values of the EVSI-based stopping indices of these policies to understand how they differ from each other and, in particular, how tightly cPDELower and cPDEUpper may bound cPDE.

We present three sets of experiments. Each uses a different numerical example and provides slightly different insights into differences among the indices and how their relative values may affect trial performance.

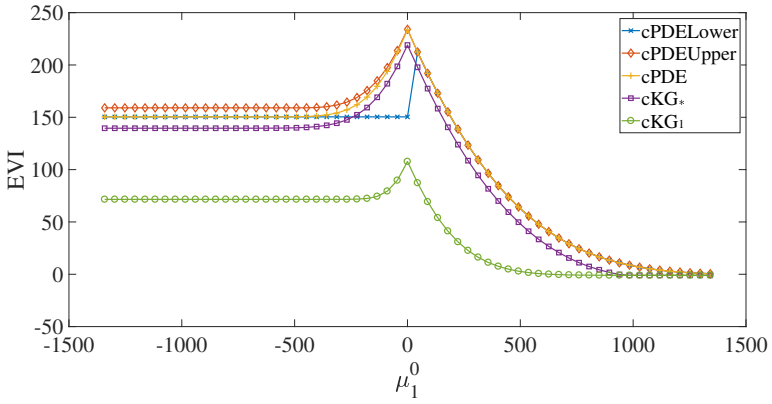
Our first set of experiments uses a stylized 3-arm problem to compare how indices vary with changes in the value of a prior mean. We set $M = 3$, $P = 1$, $I_i = 0$, $c_i = 1$, $\lambda_i = 10^6$ and $n_i^0 = 5$ for all $i = 1, 2, 3$. Thus $\Sigma_{i,i}^0 = \lambda_i / n_i^0 = 2 \times 10^5$. The correlation matrix is

$$\tilde{\boldsymbol{\rho}}^0 = \begin{bmatrix} 1 & 0.5 & -0.5 \\ 0.5 & 1 & 0.25 \\ -0.5 & 0.25 & 1 \end{bmatrix}, \tag{EC.18}$$

and we fix the prior means for arms 2 and 3 to be 0.5 and 0, respectively, so that differentiating the three means becomes difficult for μ_0^1 near zero. We then record how arm 1's indices change with its prior mean.

Figure EC.1 plots the results for the first set of experiments. Its horizontal axis marks the prior mean of arm 1 for each experiment, as it ranges from -3 to 3 times the standard deviation $\sqrt{\Sigma_{i,i}^0}$, and the vertical axis plots the heuristics' indices for arm 1 at $t=0$ for that prior mean.

Figure EC.1 EVI for different stopping indices for arm $i=1$ as a function of arm $i=1$'s prior mean.



For the problem used to generate Figure EC.1, the number of undominated arms is $M'=2$ for values of μ_0^1 lower than 44.72 and is $M'=3$ for the values greater than 44.72. When $M'=2$, the indices of cPDELower and cPDEUpper are very close to that of cPDE. When $M'>2$, the index of cPDEUpper is somewhat higher than that of cPDE. The index of cPDELower is close to that of cPDE, and the difference between the indices of cPDELower and cPDE is largest when μ_0^1 is around zero and differentiating among 3 arms is difficult. Since cPDELower takes only the maximum value among undominated arms, it is possible that cPDELower's EVI estimate falls farther away from cPDE's estimate as M' increases.

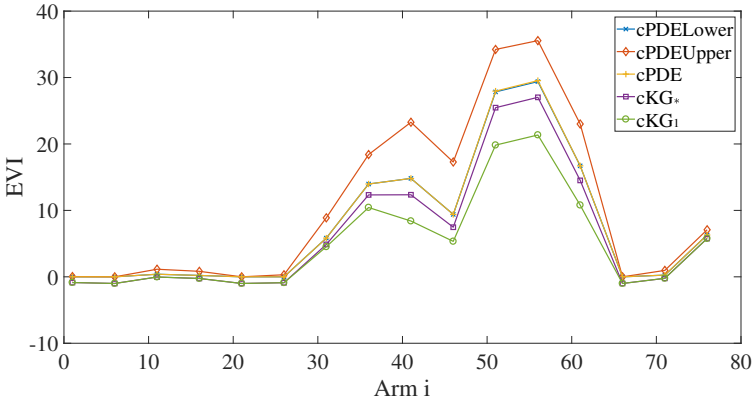
Figure EC.1 also shows that indices from cKG_1 and cKG_* are lower than those from cPDE, cPDELower and cPDEUpper for all prior mean values we tested. We observe that, for high μ_0^1, indices from cKG_1 and cKG_* are below zero, while indices from cPDELower, cPDEUpper and cPDE are positive. Since index-based policies stop the trial when indices from all arms fall below zero, we expect cKG-based stopping indices to stop much earlier than cPDE-based stopping indices (and observe this in experiments).

In summary, cKG_1 severely and cKG_* slightly underestimates EVI for all mean prior values, cPDEUpper overestimates EVI slightly only when $M'>2$, and cPDELower offers a good estimate for EVI except when $M'>2$ and the prior means of arms are close in value.

The second set of experiments resembles those found in Frazier et al. (2009). Here, we explore how indices vary across different arms at a given period during the realization of a sample path. We set $M=80$, $P=10^4$, $I_i=0$, $c_i=1$ and $\lambda_i=0.01$ for all i. The covariance across arms is determined by $\zeta=16/(M-1)^2$ and $\sigma^2=0.5$. We start with a prior mean of 0 for all arms and then randomly sample from 10 arms to obtain a new prior mean. Figure EC.2 shows the index values of heuristics for arms $i=1,6,\ldots,76$.

In contrast with Figure EC.1, the indices for cPDELower and cPDE are similar for all arms in Figure EC.2. As expected, the indices for cKG_1 and cKG_* underestimate the index of cPDE, which in turn is less than

Figure EC.2 EVI for different stopping indices for several arms during a realization of a sample path.



that of cPDEUpper. Importantly, the ratio of these indices is not constant across arms, so there does not seem to be an obvious heuristic 'fudge factor' which one might use to convert from one index to another.

When we sort arms in decreasing order of their EVI estimates, the rankings obtained for cPDELower and cPDE results are identical here, while those for cPDEUpper, cKG_*, and cKG_1 differ from that for cPDE. The state $(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$ for which this figure was drawn, therefore, illustrates a situation where the allocation policies associated with cPDELower and cPDE would choose the same arm, but the other associated allocation policies may differ in their choice.

Our third set of experiments provides insight into the behavior of indices as the number of arms in a problem increase. We run a set of experiments in which we vary the number of arms from $M = 5$ to $M = 100$. We set $P = 10^4$, $I_i = 0$, $c_i = 1$, and $\lambda_i = 0.01$ for all i. The covariance across arms is determined by $\zeta = 100/(M-1)^2$ and $\sigma^2 = 0.5$. We start with a prior mean of 0 for all arms and then randomly sample from 5 arms to obtain a set of posterior means.

Figure EC.3 The EVI_i^* less the EVI for different stopping indices for arm $i = 5$ as a function of number of arms.

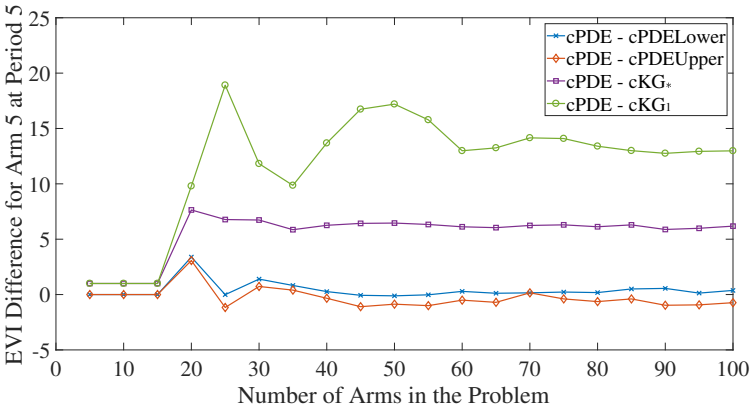


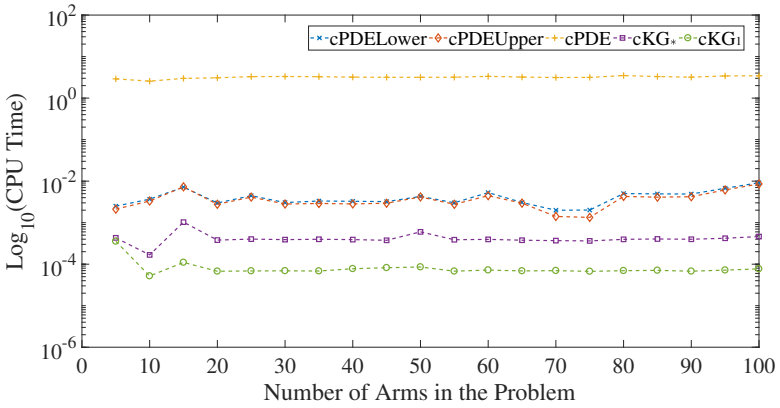
Figure EC.3 shows the difference between EVI estimates for cPDE with the EVI of several different stopping indices. The vertical axis represents the difference in EVI between cPDE and the approximate EVIs from the cPDELower, cPDEUpper, cKG_*, and cKG_1 stopping times. The horizontal axis shows the number of arms in the problem. The plot does not indicate a particular relation between the number of

arms and the accuracy with the four approximate EVIs and the EVI of cPDE. The bounds cPDELower and cPDEUpper on cPDE both are reasonably close, as compared to the approximation based on fixed-duration lookaheads, cKG_*. As expected, cKG_1 has the poorest approximation of the EVI of cPDE, and is therefore not recommended as a stopping time at all. In an analogous figure with $P = 10^6$, the EVI approximation for cKG_* approached the quality of that of cPDELower, which were both modestly less good than cPDEUpper's EVI as an approximation to the EVI of cPDE.

C.3. Computational Improvement for the cPDE Stopping Time.

A naive implementation of the cPDE stopping time would recompute $\text{EVI}_i^*(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$, the solution of a PDE, for each arm at each time step. Such repeated computation would be tedious. To be more precise, Figure EC.4 depicts the log of the CPU times for computing different indices as a function of the number of arms, M. The average CPU time required to compute the EVI of the cPDE stopping index, $\text{EVI}_i^*(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$, is orders of magnitude larger (3-5 seconds to compute the index for each arm) than for the EVI of other policies. The average CPU per index does not strongly depend on the number of arms because they require computation proportional to the number of arms which might become best, M', which tends to be smaller than M.

Figure EC.4 Average CPU times, per arm, to compute the indices of different stopping times, as a function of number of arms.



Fortunately, the lower and upper bounds of $\text{EVI}_i^*(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$ introduced in Section 4.1 can be used to decrease the computation time of the cPDE stopping time. The indices for cPDELower and cPDEUpper can be found by grid interpolation of a standardized PDE which need be pre-computed only once prior to sampling. Thus, we can reduce the number of PDEs solutions with this technique if the bounds on cPDE justify the question of whether to continue or not. The numerical results in Section 6.3 provide empirical evidence that CPU time can be improved dramatically.

Algorithm 1 uses the following three facts: (1) if an upper bound on the value of continuing with arm i suggests that there is no value in sampling from arm i, then there is no value in sampling from arm i: $\overline{\text{EVI}}_{i,\boldsymbol{\alpha}}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t) \leq 0$ implies $\text{EVI}_i^*(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t) \leq 0$; (2) if a lower bound on the value of continuing with arm i justifies continuing, then arm i justifies continuing: $\underline{\text{EVI}}_i(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t) > 0$ implies $\text{EVI}_i^*(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t) > 0$; and (3) if at least one arm $i \in \mathcal{M}$ has $\text{EVI}_i^*(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t) > 0$ then it is optimal to continue.

Algorithm 1: cPDE stopping time: version to avert computing unneeded PDE solutions.

Result: `Return Boolean variable` ***stop*** `with true to stop sampling, false to continue.`

stop $\leftarrow 0$; `% Initialize Boolean variable` ***stop*** `= 0 to continue;`

for *i in* $1,\ldots,M$ **do**
  Calculate the cPDELower stopping index, $\underline{\text{EVI}}_i(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t)$;
  if $\underline{\text{EVI}}_i(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t)>0$ **then**
    return *stop*; `% If any` $\underline{\text{EVI}}_i>0$ `then return` ***stop*** `= 0 to continue`
  end
end

for *i in* $1,\ldots,M$ **do**
  Calculate the cPDEUpper stopping index, $\overline{\text{EVI}}_{i,\boldsymbol{\alpha}}(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t)$;
  if $\overline{\text{EVI}}_{i,\boldsymbol{\alpha}}(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t)>0$ **then**
    Calculate the cPDE stopping index, $\text{EVI}_i^*(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t)$;
    if $\text{EVI}_i^*(\boldsymbol{\mu}^t,\boldsymbol{\Sigma}^t)>0$ **then**
      return *stop*; `% If any` $\text{EVI}_i^*>0$ `then return` ***stop*** `= 0 to continue.`
    end
  end
end

return *stop* $\leftarrow 1$; `% Nothing justified continuation, so return` ***stop*** `= 1 to stop`

We use this algorithm to drastically reduce the time to compute the cPDE stopping time in applications. It initially assumes that continuing should happen until proven otherwise. It first checks if any index for cPDELower would suggest continuing. If no such index exists, it then checks if any index for cPDEUpper suggests stopping. Only if such an upper bound suggests that continuing may be justified is the PDE computation for cPDE's index, EVI_i^*, needed.

C.4. Computations for the Gaussian Process Prior.

Section 5.2 presents a method to use data from a pilot study to develop an empirical Bayes prior distribution for the unknown mean rewards of arms, specified by $\boldsymbol{\mu}^{0,\text{GPR}},\boldsymbol{\Sigma}^{0,\text{GPR}}$, as well as an estimate for sampling variance, specified by Λ^{GPR}. Section 6.4 illustrates the use of such a prior distribution. Algorithm 2 gives pseudocode that describes how this was done for our numerical experiments.

In summary, the trial manager first specifies a subset of arms to use in the pilot study, an initial number of patient observations to make for each arm. The trial manager also specifies a functional form for the estimated responses of each arm as a function of pilot data. We chose a Gaussian process regression model (GPR) with squared exponential kernel from (25) and assumed a constant sampling variance across arms (so that Λ^{GPR} is a diagonal matrix with common diagonal $\lambda_i^{\text{GPR}}=\lambda^{\text{GPR}}$).

Then, the trial manager uses standard GPR tools (in our case, Matlab's `fitrgp` with MLE parameter fitting) to compute the GPR estimate for the unknown means and sampling variance. This results in GPR estimates $\boldsymbol{\mu}^{0,\text{GPR}},\boldsymbol{\Sigma}^{0,\text{GPR}}$ and λ^{GPR}.

Algorithm 2: Determine a prior distribution $\mathcal{N}(\boldsymbol{\mu}^{0,\mathrm{GPR}}, \boldsymbol{\Sigma}^{0,\mathrm{GPR}})$ for the unknown mean rewards, $\boldsymbol{\theta}$, and an estimate for sampling variances Λ^{GPR} based on a pilot study with Gaussian process regression (GPR).

Result: `Return prior mean and variance,` $(\boldsymbol{\mu}^{0,\mathrm{GPR}}, \boldsymbol{\Sigma}^{0,\mathrm{GPR}})$, `for unknown mean rewards,` $\boldsymbol{\theta}$, `and sampling variances,` Λ^{GPR}.

Inputs: A subset of arms, $\mathcal{M}_0$, to be tested in pilot study; an initial number of observations, n_0, per arm in the pilot study; and model for GPR and sampling with parameter $\Gamma_{0,\mathrm{GPR}}$. (We assumed the kernel in (25) and common sampling variance, $\lambda_i^{\mathrm{GPR}} = \lambda^{\mathrm{GPR}}$, so that $\Gamma_{0,\mathrm{GPR}} = (\sigma^2, \zeta, \lambda^{\mathrm{GPR}})$) ;

Initial sampling: Observe outcomes from n_0 patients for each arm $j \in \mathcal{M}_0$ in the pilot study and compute their sample means, $\bar{y}_{pilot,j}$; Initialize number of samples per arm in the pilot, $n \leftarrow n_0$;

Fit: Use Matlab's `fitrgp` function to estimate model parameters $\Gamma_{0,\mathrm{GPR}}$ using the MLE option; Compute $(\boldsymbol{\mu}^{0,\mathrm{GPR}}, \boldsymbol{\Sigma}^{0,\mathrm{GPR}})$ and Λ^{GPR} from $\Gamma_{0,\mathrm{GPR}}$;

Diagnostic check: *If* the effective sample size, $\lambda^{\mathrm{GPR}}/\Sigma_{j,j}^{0,\mathrm{GPR}}$, for all arms j is less than the number of observations in the pilot so far, *Then* go to Step *Return* ;

Check pilot size: **if** $n < 2n_0$ **then**
  increment n by 1, observe one more outcome from each arm in the trial;
  update the sample means, $\bar{y}_{pilot,j}$, for each $j \in \mathcal{M}_0$;
  go to Step *Fit*
else
  go to Step *Alternative Fit*
end

Alternative Fit: Use Matlab's `fitrgp` function to estimate model parameters $\Gamma_{0,\mathrm{GPR}}$ using the Bayesian optimization option; Compute $(\boldsymbol{\mu}^{0,\mathrm{GPR}}, \boldsymbol{\Sigma}^{0,\mathrm{GPR}})$ and Λ^{GPR} from $\Gamma_{0,\mathrm{GPR}}$;

Return GPR estimate for prior $(\boldsymbol{\mu}^{0,\mathrm{GPR}}, \boldsymbol{\Sigma}^{0,\mathrm{GPR}})$ and sampling variances Λ^{GPR} ;

In some simulated pilot studies, the GPR estimate exhibited poor fit based on these initial samples. For example, the estimated effective sample size for some of the arms (estimated values of $\lambda^{\mathrm{GPR}}/\Sigma_{j,j}^{0,\mathrm{GPR}}$) exceeded the number of observations in the pilot, a result associated with extremely high estimated correlations across arms. In such cases, where the fit of the GPR estimate was poor, we obtained one more observation from each arm in the pilot study, with the goal of obtaining a better fit. Additional observations were obtained until the fit was good, or until an upper limit on the number of observations was reached. In our case, we set that limit to be twice the initial sample size in the pilot. If after reaching that limit, diagnostics tests were still not satisfied, we used an alternative method for fitting the GPR model. In particular, we used the Bayesian optimization option of Matlab's `fitrgp` function. In all 1000 sample paths, we found that the addition of a few sample points, along with the use of two alternative Gaussian process regression fitting procedures, was able to result in a GPR prior that satisfied diagnostic tests for goodness of fit.

We found those diagnostic tests to be practically important. Without accounting for both of them, correlation across arms might be significantly overestimated (for 2-4% of simulated pilot studies in our experiments),

resulting in unjustified overconfidence about the values of the mean rewards as well as strong under-sampling. The diagnostic checks that result in additional pilot study samples (for a subset of sample paths) and refitting of the model avoids those problems which might otherwise arise due to numerical issues unrelated to the value of the arms of the trial.

Appendix D: Further Comments on Practical and Research Questions

Regulators expect frequentist operating characteristics of a clinical trial to be reported. In Appendix D.1 we illustrate Monte Carlo estimates of statistical power for some of our proposed policies, consistent with guidance by the FDA (2019) for complex innovative trials. Appendix D.2 discusses a non-exhaustive set of interesting issues related to our approach to clinical trial designs.

D.1. Frequentist Performance Measures

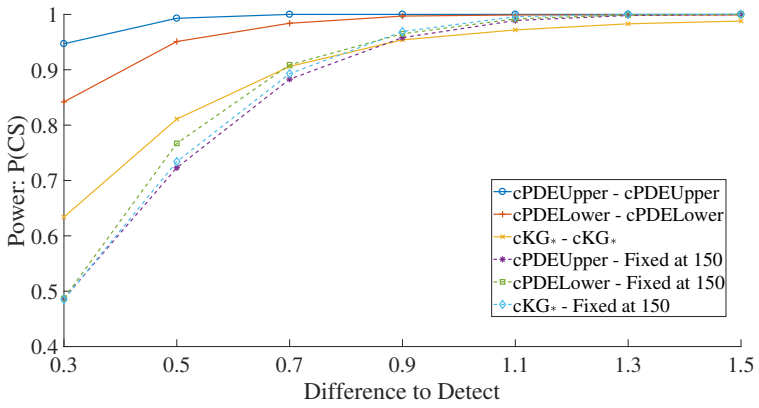
We use a synthetic Phase II/III dose-finding example motivated by Section 6.4 but with some differences to allow us to control the gap between the performance of the best arm and that of other arms. To control the minimal difference to detect, we used a triangular shaped curve for the patient responses in which we denote the difference between the response for the best arm and for the second best arms to be δ. We set $\theta_{11} = 0$ and $\theta_j = -|j-11|\delta 2000$ for $j = \{1,\ldots,21\}/\{11\}$. For a dose finding trial described by Bornkamp et al. (2007), the difference to detect for the best arm is 1.3, which guided the range of δ values we tested.

Figure EC.5 shows power curves for six policies, including the Fixed stopping time with sample size 150 (not atypical sample size for such a trial, Bornkamp et al. 2007, Huang et al. 2015). The horizontal axis marks the difference to detect, δ. The vertical axis reports power, the fraction of 1000 simulation trials in which the policy correctly identifies the best arm.

In Figure EC.5, the cPDELower-cPDELower and cPDEUpper-cPDEUpper policies exceed the commonly-used 80% power threshold for all δs. The cKG$_*$-cKG$_*$ policy's power is not as high for $\delta = 0.3$ and $\delta = 0.5$, and it attains a power higher than 80% for larger δs. When our heuristics used as allocation policies and paired with the Fixed stopping time with a sample size of 150, they satisfy the 80% power requirement for a difference to detect of 0.7 or more.

Thus, our new allocation and stopping indices have appropriate power in this example. More generally, power is larger for larger P, and power is smaller for larger c_i.

Figure EC.5 Power curve for different policies



D.2. Population Size, Recruitment Rates, Precision Medicine and Other Points

There are many other interesting and valuable issues for pushing the value-based MAMS approach, or for trial designs in general. We comment on a few of them here, and note their potential for future research.

Adopting population size. The adopting population size, P, is assumed to be a fixed constant, and to not explicitly depend on the stopping time of the trial, T, nor on the posterior mean responses $\boldsymbol{\mu}^T$ above and beyond the dependence of the selected arm for implementation, arm $\mathcal{D}$, on $\boldsymbol{\mu}^T$. This makes sense in the context of many nonpharmaceutical trials and ensuing technology assessment decisions (NICE 2014), or for pharmaceutical trials with market exclusivity agreements whose duration is of a fixed length (FDA 2015). For a fixed horizon for exploitation of pharmaceuticals, for example associated with patent protection, it may be useful to model an adopting population size $P(T)$ that is decreasing in T. Also, it may be useful to allow for the size of the adopting population to depend more strongly on the mean reward of the arm selected for adoption, so that $P(\mu_{\mathcal{D}}^T, T)$ depends on T and the mean reward of the selected arm. Such influence might come from a greater fraction of adoptions for 'better' arms, for example. There are some interesting cases where this phenomenon has been modeled, such as Willan and Eckermann (2010). Strong empirical evidence in general for the best form for $P(\mu_{\mathcal{D}}^T, T)$ is to be determined, although Gaessler and Wagner (2019) present interesting and relevant data for patent and data protection in the time to exploit a pharmaceutical technology, when taken from a firm-perspective point of view. At present, we note that such generality can be obtained within our social welfare maximizing framework by putting $P(\mu_{\mathcal{D}}^T, T)$ in for P in the main reward function of (2), with similar changes elsewhere. Analysis of this more general formulation and further empirical study of the effects of such influence of T and $\mu_{\mathcal{D}}^T$ on P are interesting topics for further work.

QALY and cost information collection. Our base model also presumes that health outcomes that are convertible to money (such as QALYs) and treatment costs can be collected for each patient during the trial and monitored sequentially. This may represent an additional burden for many trials, even if outcomes are already monitored for safety. That said, QALY information is already collected in many trials (Gold et al. 1996, Angus et al. 2001, Ferguson et al. 2013, Flight et al. 2019, Karakike et al. 2019) and such information may be needed anyway in a health-economic assessment for a technology adoption decision that follows the trial (NICE 2014). Although QALY estimates are sometimes assessed with delays on the order of a year or two (Forster et al. 2019), it may be sufficient to have estimators which have the same bias across arms for the purpose of allocating arms. The inference of such potential future QALYs and costs using surrogate measures during a longitudinal study represents an area of further interest.

Online and offline learning. Ahuja and Birge (2016) look at an adaptive design with a fraction of patients allocated to each arm and aim to improve outcomes of those in the trial while finding the most effective alternative with high probability, a so-called online learning approach. Their model assumes two arms and Bernoulli outcomes. Bandit problem-based approaches also model the rewards of patients in a trial (Williamson and Villar 2020) but might not model patients affected after the adoption decision is made. This would seem implementable with related work for online learning with an EVSI framework (Ryzhov et al. 2010, Chick et al. 2017).

Prior distribution selection. Alternative methods for specifying the prior distribution may exist. For example, Qu et al. (2015) propose methods to infer an unknown covariance structure and provide convergence results. They do not guarantee asymptotic consistency of their estimators.

Escalation studies for dose-range assessment. For dose finding, Phase I/IIa trials may use dose escalation studies to find a maximum tolerable dose, so as to avoid the risk of excess toxicity from high dosing in later stages of a trial. Toxicity may lead to side effects, which would reduce health benefits, thus lowering the overall effectiveness (explaining our dose-response curves in Section 6.4 which have initially increasing health benefits in dose followed by a decreasing health benefit at higher doses, due to side effects and/or the costs of higher dosing). For the pilot study in Section 6.4, we presume that all doses tested have the same sample size. In practice, we may wish to order the assignment of doses to patients during the pilot study using standard techniques for dose escalation, with lower doses tested first, such as the well-known 3+3 design, or some alternative designs which have been found to be more effective (e.g. Huang et al. 2015, Wheeler et al. 2019). Thus, our pilot study can be run in a way consistent with these practical considerations.

References

Ahuja V, Birge JR (2016) Response-adaptive designs for clinical trials: Simultaneous learning from multiple patients. *European Journal of Operational Research* 248(2):619–633.

Angus D, Musthafa A, et al. (2001) Quality-adjusted survival in the first year after the acute respiratory distress syndrome. *Am J Respir Crit Care Med.* 163(6):1389–94.

Bertsekas D, Shreve S (1978) *Stochastic Optimal Control: The Discrete Time Case* (Academic Press).

Bornkamp B, et al. (2007) Innovative approaches for designing and analyzing adaptive dose-ranging trials. *Journal of Biopharmaceutical Statistics* 17(6):965–995.

Chick SE, Forster M, Pertile P (2017) A Bayesian decision-theoretic model of sequential experimentation with delayed response. *Journal of the Royal Statistical Society: Series B* 79(5):1439–1462.

Chick SE, Gans N, Yapar O (2019) Sequential, value-based designs for certain clinical trials with multiple arms having correlated rewards. *2019 Winter Simulation Conference (WSC)*, 1032–1043 (IEEE).

FDA (2015) Patents and exclusivity. US Food and Drug Administration, https://www.fda.gov/downloads/drugs/developmentapprovalprocess/smallbusinessassistance/ucm447307.pdf.

FDA (2019) Interacting with the FDA on complex innovative trial designs for drugs and biological products. US Food and Drug Administration, https://www.fda.gov/media/130897/download.

Ferguson ND, et al. (2013) Integrating mortality and morbidity outcomes using quality-adjusted life years in critical care trials. *Am J Respir Crit Care Med.* 187(3):256–261.

Flight L, Arshad F, Barnsley R, Patel K, Julious S, Brennan A, Todd S (2019) A review of clinical trials with an adaptive design and health economic analysis. *Value in Health* 22:391–398.

Forster M, et al. (2019) Cost-effective clinical trial design: Application of a Bayesian sequential stopping rule to the profher pragmatic trial. Discussion Papers 19/01, Department of Economics, University of York.

Frazier PI, Powell W, Dayanik S (2009) The knowledge-gradient policy for correlated normal beliefs. *INFORMS Journal on Computing* 21(4):599–613.

Gaessler F, Wagner S (2019) Patents, data exclusivity, and the development of new drugs. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3401226.

Gold MR, Siegel JE, et al. (1996) *Cost-effectiveness in Health and Medicine* (OUP Oxford).

Huang JH, et al. (2015) Sample sizes in dosage investigational clinical trials: a systematic evaluation. *Drug design, development and therapy* 9:305.

Karakike E, et al. (2019) Exploring alternative trial designs for pragmatic clinical studies: A Bayesian decision-theoretic model applied on a real ongoing one-stage trial, Abstract from 39th International Symposium on Intensive Care and Emergency Medicine. *Critical Care* 23(72).

NICE (2014) Interim methods guide for developing service guidance, modeling and health economics considerations. UK National Inst. for Health and Care Excellence.

Qu H, Ryzhov IO, Fu MC (2015) Sequential selection with unknown correlation structures. *Operations Research* 63(4):931–948.

Russo D (2020) Simple Bayesian Algorithms for Best Arm Identification. *Operations Research* to appear.

Ryzhov IO, Frazier PI, Powell WB (2010) On the robustness of a one-period look-ahead policy in multi-armed bandit problems. *Procedia Computer Science* 1(1):1635–1644.

Wheeler G, Mander A, Bedding A, et al. (2019) How to design a dose-finding study using the continual reassessment method. *BMC Med Res Methodol* 19(18), https://doi.org/10.1186/s12874-018-0638-z.

Willan AR, Eckermann S (2010) Optimal clinical trial design using value of information methods with imperfect implementation. *Health economics* 19(5):549–561.

Williams D (1991) *Probability with Martingales* (Cambridge University Press).

Williamson SF, Villar SS (2020) A response-adaptive randomization procedure for multi-armed clinical trials with normally distributed outcomes. *Biometrics* 76(1):197–209.

Xie J, Frazier PI, Chick SE (2016) Bayesian optimization via simulation with pairwise sampling and correlated prior beliefs. *Operations Research* 64(2):542–559.