informs
https://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH

Vol. 00, No. 0, Xxxx 0000, pp. 000–000

Submitted to *Operations Research*

ISSN 0030-364X, EISSN 1526-5463

Module Design for Pilots Training with Full Flight Simulator

Abstract. This study addresses a practical problem in the design of training modules for pilots with Full Flight Simulators (FFS), using operations research techniques. FFSs are highly precise tools that accurately replicate real-world flight scenarios, making them indispensable for pilot training and evaluation. Given their high cost and limited availability, airlines must strategically design modules for pilots' periodic recurrent assessments. To enhance aviation safety, we formulate a combinatorial optimization objective function to select a set of modules that minimizes a pilot's lowest score, enabling early detection of potential weaknesses. Our research underscores the importance of accounting for correlations in pilots' performance across modules. To this end, we develop a suite of optimization methods and provide corresponding managerial insights. Extensive experimental evaluations demonstrate that our approach significantly outperforms traditional methods reliant on intuition and experience.

Funding: This research was supported by [grant number, funding agency].

Key words: module design, Full Flight Simulators, simulation optimization

1. Background

The Full Flight Simulator (FFS) is a highly sophisticated flight training device designed to replicate the cockpit environment and operational dynamics of a specific aircraft model with exceptional fidelity (ICA 2015). It is capable of simulating the entire spectrum of flight operations, from takeoff and cruise to landing, encompassing both routine procedures and complex scenarios such as engine failures, hydraulic system malfunctions, severe weather conditions, and even rare events like bird strikes or runway contamination (Adv 2024). The simulation integrates visual, auditory, and tactile feedback, alongside precise instrument readouts and control responses, to provide an immersive experience that closely mirrors actual flight conditions. Certified to the highest standards, such as Level D by aviation authorities like the Federal Aviation Administration (FAA) or the European Union Aviation Safety Agency (EASA), the FFS achieves unparalleled precision, enabling zero flight time training where pilots can qualify without operating an actual aircraft (Federal Aviation Administration 1995, European Union Aviation Safety Agency 2020). However, this extraordinary

accuracy comes at a significant cost, with a single FFS unit priced in the tens of millions of dollars, rendering its manufacture and maintenance a substantial financial commitment (CAE 2023). Most airlines typically own only a limited number of FFS units outright or instead lease them from manufacturers such as CAE or L3Harris and dedicated flight training centers (IBA 2023).

The primary application of the FFS lies in pilot training and qualification maintenance, with a particular emphasis on mandatory recurrent training to renew licenses. Pursuant to regulations set by the International Civil Aviation Organization (ICAO) and national aviation authorities, pilots are required to undergo recurrent training every six to twelve months, typically conducted in an FFS, to validate their proficiency (Fli 2024).

During qualification assessments, pilots are evaluated through a series of subjects designed to test their ability to manage diverse operational challenges. These subjects are carefully crafted based on real-world flight data and historical incident analyses(National Transportation Safety Board 2012). Common subjects include "Engine Failure on Takeoff," where pilots must stabilize the aircraft and execute a safe return to the airport following an engine malfunction immediately after liftoff; "Low Visibility Approach," which tests precision navigation and landing in simulated adverse weather; and "Emergency Descent and Decompression," requiring rapid altitude reduction and oxygen management in response to cabin pressure loss (Pil 2023). These assessments, which may span several hours, evaluate technical proficiency, decision-making speed, and crew coordination. Successful completion of these subjects results in license renewal, while deficiencies necessitate further training, ensuring consistent competency across the pilot workforce (European Union Aviation Safety Agency 2020).

Given the high cost of FFS and the critical need for regular pilot assessments, airlines must carefully balance budgetary constraints with the demand for efficient and effective evaluations. However, the current practice of selecting assessment subjects is often random or reliant on airlines' experiential judgment, which can fail to identify specific deficiencies or redundantly evaluate areas of established proficiency. This work aims to develop dedicated optimization approaches to enhance the selection of assessment subjects, maximizing the effectiveness of the evaluation process.

Suppose there is a pool of M different training modules from which m are selected in one training session due to cost restriction. Ideally, these m subjects should collectively cover a diverse range of skills to efficiently identify a pilot's weaknesses.

Unlike conventional approaches, our strategy incorporates the correlations and variability in pilot performance across these modules, driven by individual differences in skills such as hand-eye coordination, spatial awareness, and stress resilience. For instance, pilots with strong coordination skills

often excel in subjects like "Precision Approach" and "Instrument Flying," as both require similar competencies, including accurate interpretation of flight instruments, stable aircraft control, and anticipation of spatial positioning (Damos 2003). As a result, high performance in one module frequently predicts success in the other due to shared skill requirements. Conversely, pilots with superior stress management capabilities may outperform in complex emergency scenarios such as "Multi-Engine Failure", but struggle in subjects like "Crosswind Landing" if their coordination is less developed (Szczepanska et al. 2025). These correlations arise from overlapping skill sets and selecting multiple highly correlated subjects may lead to redundant testing.Accounting for these correlations is therefore critical in optimizing the selection of assessment modules to enhance training efficiency and safety (Duruaku et al. 2024).

We illustrate this with a simple example, where we need to select two subjects out of three for a pilot's assessment. From an aviation safety perspective, we aim to select the two subjects that best identify the subject with the lowest score, and thus the assessment metric is the minimum score of the two selected subjects. From an optimization perspective, the following objective function is considered:

$$\min_{x_1,x_2} \mathbb{E}[\min\{Y(x_1),...,Y(x_2)\}],$$

where x_1, x_2 are two selected subjects from the subjects pool $\{A, B, C\}$ and $Y(x_i)$ is the score on x_i. The scores of each pilot on these three subjects follow a discrete distribution with three scenarios (shown in Table 1), each with probability 1/3. From the table, we can observe that the correlation

Table 1 Discrete distribution of the scores

Scenario	A score	B score	C score
1	9	9.5	7.5
2	6	5	8
3	10	9.5	10
Expectation	8.33	8	8.5

between scores for subjects A and B is significantly higher than that between A and C or B and C, due to the differing skill requirements for the subjects, as previously discussed. If correlations are ignored, the first two subjects might be chosen (due to their lower mean scores). However, calculations reveal that the latter two subjects should be selected. This is precisely because their lower correlation allows them to more independently yield lower scores, increasing the likelihood of

identifying the pilot's weakest performance. Therefore, we need to design an optimization method that fully accounts for the correlations between scores of different subjects.

Addressing this problem presents multiple challenges. First, computing the expected minimum value of several correlated distributions is inherently complex, particularly when selecting an optimal subset of m modules from a larger set to minimize this value. Second, in practical applications, the expected values of these distributions are likely unknown. While we can infer correlations between scores based on the skill requirements of different subjects, estimating the mean scores of pilots across these subjects is considerably more difficult. Hence, an approach that simultaneously learns from data of previous test results and optimizes the selection process is necessary.

To address the first challenge, we assume that the scores follow a multivariate normal distribution (a reasonable model for characterizing examination scores (Ross 2014) and has been validated through real-world FFS data). By analyzing the impact of parameters—such as means, variances, and correlations—on the target function, we derive insights to guide module selection. Furthermore, we identify the target function as a submodular problem, enabling the use of a greedy algorithm for approximate computation. To improve computational efficiency, we propose a heuristic algorithm informed by these analytical results. To tackle the second challenge, we develop a combinatorial multi-armed bandit (MAB) algorithm based on a Bayesian framework. This algorithm assigns a prior distribution to the score means, which is continuously updated through iterative testing. In each iteration, an upper confidence bound approach is used to select the module combination to be tested, and we prove the algorithm's regret bound to ensure its theoretical robustness.

This paper addresses a problem of significant practical relevance, representing an innovative effort to apply optimization and operations research algorithms to real-world challenges. Previously, the design of pilot training modules has relied heavily on intuition and experience. Beyond its practical significance, the optimization problem explored here also carries substantial theoretical value, presenting multiple challenges as outlined earlier. Moreover, while related to existing research domains, it exhibits notable distinctions. Our work primarily aligns with two categories of optimization studies. The first is simulation-based optimization (Fu 2015), which employs simulation models to support decision-making in real-world systems (FFS is a highly precise simulation tool). Certain simulation optimization studies account for correlations in experimental results when selecting multiple design alternatives (Xie et al. 2016). These correlations primarily arise from the use of common random numbers in computer simulations. However, no prior work in this field has targeted the same objective function as ours, as most focus on evaluating a single design's

performance without an internal minimization operation. The second category is combinatorial multi-armed bandit research (Chen et al. 2013a), where some studies select multiple alternatives and use the best as the evaluation metric, akin to our objective function (1). Yet, to the best of our knowledge, none have addressed correlations among alternatives. These correlations significantly complicate the problem while offering unique managerial insights.

The remainder of the paper is organized as follows. Section 2 reviews the related works. Section 3 formulates the objective function. Section 4 offers some insights and heuristic algorithms for scenarios where expectations are known. Section 5 presents an algorithm that learns expectations during the optimization process, supported by theoretical guarantees. Section 6 reports numerical experiments. Finally, Section 7 concludes the paper.

2. Related works

We review relevant works in three aspects: existing approaches for pilot training modules design, simulation-based optimization, and combinatorial MAB.

Combinatorial MAB Our module selection problem naturally fits into the combinatorial multi-armed bandit (CMAB) framework with semi-bandit feedback. In this framework, an agent sequentially selects a subset of base arms (referred to as a super arm) and observes a reward determined by the outcomes of all selected arms (Chen et al. 2013b), generalizing the classical multi-armed bandit problem to combinatorial action spaces and complex reward structures.

The CMAB literature predominantly relies on two common modeling assumptions. A substantial body of work assumes that base arms are *mutually independent* (Zhang and Combes 2025, Combes et al. 2015, Cuvelier et al. 2021). Alternatively, another line of research—including some works that incorporate dependencies—often restricts the reward function to be *linear* with respect to the base arm outcomes (Degenne and Perchet 2016, Kveton et al. 2014a,b), or models the expected reward as a function of only the *marginal means*, while our problem setting requires accounting for the influence of the full joint distribution on the reward function (Demirel and Tekin 2021, Kveton et al. 2015).

Our work is most closely related to (Chen et al. 2016), who also consider the expected minimum (or maximum) of the selected arms as the reward function. A key distinction, however, is that their model does *not* incorporate correlations between base arms, whereas our work explicitly addresses the impact of the joint distribution on the reward. (Wang et al. 2023) also examine a minimum-based reward, but under a more restrictive *full-bandit feedback* setting where only the maximum

value and its originating arm index are observed, rather than the outcomes of all base arms in the super arm. This limited feedback significantly increases the learning difficulty. Moreover, their approach still assumes independent base arms, and their algorithm is confined to problems where the base arm outcomes lie in a *finite, discrete set*, while our setting considers continuous distributions where the joint structure plays a crucial role. In a somewhat orthogonal direction, (Degenne and Perchet 2016) consider correlated base arms with a known covariance matrix under a sub-Gaussian assumption, yet their reward function depends only on the *means* of the arms, while our problem requires considering the effect of the joint distribution on the reward function.

3. problem formulation

Our goal is to efficiently identify pilots' deficiencies through rational module design, with these deficiencies measured by their scores during the module evaluation process. Suppose there are M different training modules from which m are selected in one training session due to the time and cost restriction. Then, we select m modules to maximally expose a pilot's deficiencies, such that the minimum score across these m modules is as low as possible. The optimization problem can be formulated as follows:

$$\min_{x_1,...,x_m} \mathbb{E}_\xi[\min\{Y(x_1,\xi),...,Y(x_m,\xi)\}], \tag{1}$$

where $x_1,...,x_m$ represent the m modules in the training session, taken from the module set $\mathcal{X} = \{x_1,...,x_M\}$, and $Y(x_1,\xi)$ is the score the pilot obtain in module x_i. It is evaluated through simulation at FFS. In this model, we assume the score vector in the population $\{Y(x_1),...,Y(x_M)\}$ follows a multivariate normal distribution. For each specific person, his score is a sample from this distribution. In equation (1), the random vector ξ is used to illustrate the randomness in this multivariate normal distribution. The inner minimization is to find the minimum score one pilot obtain across the given m training modules. We take the minimum value as we need to find the weakness of the pilots considering flight safety. The outer minimization is the find the best set of training sessions in detecting the pilot inadequacy.

4. Insights and Algorithm with Known Score Means

We first consider an ideal scenario for solving Problem (1), where all parameters of the score distributions for all modules are fully known. That is, we aim to select a limited number of dimensions from a completely known multivariate normal distribution such that the expected value of the minimum score across these dimensions is minimized. It should be noted that, in general, this objective

function cannot be simplified into a more computationally convenient form. Existing studies provide closed-form solutions for the expected minimum of two normal distributions (Nadarajah and Kotz 2008). However, for more than two normal distributions, prior research remains limited. Clark (1961) relied on numerical approximation, that recursively applies a three-variate formula, to compute the expected value of their minimum and has analyzed the accuracy of these approximations through numerical validation. Such computational methods are time-consuming, particularly in our optimization scenario, where we need to repeatedly select m modules and compute the expected value of their minimum scores. Therefore, we aim to derive managerial insights from simple, analytically tractable examples and leverage these insights to develop efficient heuristic methods. This section is divided into two parts: in Section 4.1, we use two simplified examples to obtain managerial insights; in Section 4.2, we demonstrate that the objective function is submodular, enabling the use of a greedy algorithm to achieve an efficient and theoretically guaranteed approximation; in Section 4.3, we utilize these insights to develop an efficient heuristic method based on greedy approach.

4.1. Insights from two simple examples

This subsection examines two simplified examples to investigate how parameters in a multivariate normal distribution—specifically the mean, correlation, and variance—affect the expected minimum value. The examples include a bivariate normal distribution (Section 4.1.1) and a multi-variate scenario with same means, variances, and correlations (Section 4.1.2). Despite their simplicity, the conclusions and insights from these examples align with intuition and are corroborated by subsequent numerical experiments.

4.1.1. Bivariate normal example For any two Gaussian random number $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$ with correlation ρ, the expectation $\mathrm{E}[\min\{X_1, X_2\}]$ takes the following form (Clark 1961):

$$\mathrm{E}[\min\{X_1, X_2\}] = \mu_1 \Phi(\frac{\mu_2 - \mu_1}{\theta}) + \mu_2 \Phi(\frac{\mu_1 - \mu_2}{\theta}) - \theta \phi(\frac{\mu_2 - \mu_1}{\theta}),$$

where ϕ, Φ are cdf and pdf for standard normal distribution, respectively, and $\theta = \sqrt{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} = \sqrt{\mathrm{var}(X_1 - X_2)}$.

Denote $f = \mathrm{E}[\min\{X_1, X_2\}]$. We have:

$$\frac{\partial f}{\partial \theta} = -\phi(\frac{\mu_2 - \mu_1}{\theta}) < 0, \tag{2}$$

$$\frac{\partial f}{\partial \rho} = \frac{\partial f}{\partial \theta}\frac{\partial \theta}{\partial \rho} = \frac{\sigma_1\sigma_2}{\theta}\phi(\frac{\mu_2-\mu_1}{\theta}) > 0, \tag{3}$$

$$\frac{\partial f}{\partial \sigma_1} = \frac{\partial f}{\partial \theta}\frac{\partial \theta}{\partial \sigma_1} = -\frac{\sigma_1-\rho\sigma_2}{\theta}\phi(\frac{\mu_2-\mu_1}{\theta}), \tag{4}$$

$$\frac{\partial f}{\partial \mu_1} = \Phi(\frac{\mu_2-\mu_1}{\theta}) > 0. \tag{5}$$

We summarize the conclusions is the following lemma:

LEMMA 1. *In the bivariate normal example, the expectation of the minimum value is smaller when the two variables have smaller expectations and correlations.*

The relation between f and the variance for each individual variable is not monotone. We can, however, derive some special cases:

1. When $\rho < \frac{\sigma_1}{\sigma_2}$ (specifically when $\rho < 0$), a larger value of σ_1 is preferred.

2. When $\rho > \frac{\sigma_1}{\sigma_2}$, smaller value of σ_1 is preferred. This result may appear counter-intuitive. To gain insight, consider a special case where $\rho = 1$ and $\sigma_1 < \sigma_2$. In this scenario, X_1 and X_2 are positively linear dependent, with $X_1 = \frac{\sigma_1}{\sigma_2}X_2$. Consequently, X_1 and X_2 always share the same sign. When $X_1, X_2 > 0$, the minimum is $\min\{X_1, X_2\} = X_1 > 0$, which increases with σ_1. When $X_1, X_2 < 0$, the minimum is $\min\{X_1, X_2\} = X_2 < 0$, which is independent of σ_1. Thus, a larger σ_1 increases the objective function value in this case. Although a larger σ_1 reduces X_1 when negative, these negative values do not contribute to the objective function as long as $\sigma_1 < \sigma_2$. This effect diminishes slightly as ρ deviates from 1, but negative values of X_1 remain less likely to influence the objective function until $\rho < \frac{\sigma_1}{\sigma_2}$.

4.1.2. Multi-variate normal example We consider the following scenario: m normal random numbers $X_1, ..., X_m$ with equal mean μ and equal variance σ^2. The correlation between each pair of variables is $\rho \geq 0$. We would like to explore the relationship between the objective function $f = \mathrm{E}[\min\{X_1, ..., X_m\}]$ and ρ, μ, and σ.

In this case, each random variable can be represented as $X_i = \mu + \sigma(\sqrt{\rho}W + \sqrt{1-\rho}N_i)$, where W and $N_1, ..., N_m$ are i.i.d. standard normal random variables. Then,

$$f = \mathrm{E}[\min\{X_1, ..., X_m\}] = \mathrm{E}[\min\{\mu + \sigma(\sqrt{\rho}W + \sqrt{1-\rho}N_1), ..., \mu + \sigma(\sqrt{\rho}W + \sqrt{1-\rho}N_m)\}]$$

$$= \mu + \sigma\mathrm{E}[\min\{\sqrt{\rho}W + \sqrt{1-\rho}N_1, ..., \sqrt{\rho}W + \sqrt{1-\rho}N_m\}]$$

$$= \mu + \sigma\mathrm{E}[\sqrt{\rho}W + \sqrt{1-\rho}\min\{N_1, ..., N_m\}] = \mu + \sigma\sqrt{1-\rho}\mathrm{E}[\min\{N_1, ..., N_m\}].$$

Noted that $\mathrm{E}[\min\{N_1, ..., N_m\}]$ is a constant for a given m. It is negative and can be approximated as

$$-\sqrt{2\ln m} + \frac{\ln\ln m + \ln 4\pi}{2\sqrt{2\ln m}}.$$

Hence, $f \approx \mu - \sigma\sqrt{1-\rho}\sqrt{2\ln m}$.

We summarize the insights from this example in the following lemma.

LEMMA 2. *In the simple multi-variate normal example, the expectation of the minimum value is smaller with smaller μ, larger σ or smaller ρ.*

For a general multi-variate normal distribution, analyzing the impact of the covariance matrix on the objective function f using simple algebra is challenging. However, we can derive the following lemma regarding the influence of the means.

LEMMA 3. *In a general multi-variate normal distribution, if the covariance matrix is fixed, the value of f reduces if the mean vector gets smaller pointwisely.*

We summarize the managerial insights derived from these examples for decision-making as follows. To achieve a smaller value for the objective function, we should generally select:

1. m Gaussian variables with smaller means;
2. variables with lower correlations;
3. variables with larger variances.

According to Lemma 3, when other factors remain constant, a smaller mean vector directly reduces the expected minimum value. Furthermore, regarding the other two points, the lower the correlations and the larger the variances of the m variables, the more independently each variable can take potentially smaller values, thereby reducing the minimum value.

4.2. Demonstration of Submodular Objective Function and Greedy Algorithm

The analytical insights from Section 4.1 provide intuitive guidelines for module selection. In this subsection, we unveil a deeper, more powerful mathematical property of our objective function—submodularity. This property allows us to tackle the complex combinatorial optimization problem using a greedy algorithm with a provable performance guarantee.

We first recall the definition of submodularity. For a finite set $\mathcal{V}$ (the set of all training modules $\mathcal{X}$ in our context), a set function $F : 2^{\mathcal{V}} \to \mathbb{R}$ is **submodular** if it satisfies the diminishing returns property: for any subsets $S \subseteq T \subseteq \mathcal{V}$ and any element $x \in \mathcal{V} \setminus T$,

$$F(S \cup \{x\}) - F(S) \geq F(T \cup \{x\}) - F(T).$$

This means the marginal gain of adding a new element x to a set S is at least as large as the marginal gain of adding the same element to a superset T of S.

In our problem, we define the set function for a module subset $S \subseteq \mathcal{X}$ as $R(S) = \mathbb{E}[\min\{Y(x) : x \in S\}]$. Our goal is to minimize $R(S)$. To frame this for a greedy algorithm that maximizes a utility, we define a utility function $U(S) = y_{\max} - R(S)$. Here $y_{\max}$ is the maximum score a pilot could achieve on any module. Then we give the following theorem:

THEOREM 1. *Given that the module score vector* $\mathbf{Y} = (Y(x_1, \xi), ..., Y(x_m, \xi))$ *follows a multivariate normal distribution* $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. *For* $S \subseteq \mathcal{X}$, *the set function* $U(S) = y_{max} - R(S) = y_{max} - \mathbb{E}[\min\{Y(x) : x \in S\}]$ *is:*

1. ***Monotone Increasing****: For any* $S \subseteq T \subseteq \mathcal{X}$, $U(S) \leq U(T)$.
2. ***Submodular****: For any* $S \subseteq T \subseteq \mathcal{X}$ *and any module* $x \in \mathcal{X} \setminus T$, $U(S \cup \{x\}) - U(S) \geq U(T \cup \{x\}) - U(T)$.

Theorem 1 establishes that $U(S)$ is a monotone increasing, submodular function. Therefore, the problem of selecting a set S of size m to maximize $U(S)$ is an instance of maximizing a monotone submodular function under a cardinality constraint.

For this class of problems, the classical greedy algorithm provides a strong approximation guarantee. The algorithm starts with an empty set $S_0 = \emptyset$. At each step $k = 1, ..., m$, it adds the module x that provides the highest marginal gain:

$$x_k = \underset{x \in \mathcal{X} \setminus S_{k-1}}{\arg\max} \left[U(S_{k-1} \cup \{x\}) - U(S_{k-1}) \right], \tag{6}$$

and then updates $S_k = S_{k-1} \cup \{x_k\}$.

THEOREM 2 **(Nemhauser et al. (1978))**. *Let* S^* *be the optimal set of size m that maximizes* $U(S)$, *and let* S_m^{greedy} *be the set of size m obtained by the greedy algorithm. If U is a monotone, non-negative, submodular function, then the greedy algorithm achieves the following performance guarantee:*

$$U(S_m^{greedy}) \geq \left(1 - \frac{1}{e}\right) U(S^*) \approx 0.632 \cdot U(S^*).$$

This guarantee holds regardless of the order of the modules added after the first optimal choice, making the greedy algorithm highly effective for large-scale problems where exhaustive search is infeasible.

This theoretical guarantee is significant. It ensures that even without computing the computationally expensive exact optimal solution, the greedy heuristic yields a solution that is provably within at least approximately 63% of the optimal value in the worst case. This result justifies the development of greedy-inspired heuristic methods in the next section. Although calculating the exact marginal gain $U(S \cup \{x\}) - U(S)$ for each candidate module x may require numerical integration of the expected minimum (as discussed in Section 4.1), the submodularity property allows us to use the insights derived in Section 4.1—favoring modules with lower means, higher variances, and lower correlations—as effective proxies for the true marginal gain when designing efficient heuristics.

4.3. Heuristic algorithms

The greedy algorithm established in the preceding section provides a theoretically sound framework for module selection. A direct implementation of this algorithm, however, requires the computation of the marginal gain $x_k = \arg\max_{x \in \mathcal{X} \setminus S_{k-1}} [U(S_{k-1} \cup \{x\}) - U(S_{k-1})]$ at each iteration. Calculating this expectation of the minimum of correlated normal distributions poses a significant computational challenge, as it lacks a closed-form solution for the general case. While numerical techniques, such as Sample Average Approximation (SAA), can be employed to estimate this value, they often require a substantial number of Monte Carlo simulations to achieve acceptable accuracy. Consequently, for large-scale instances of the problem where the module pool size M is substantial, the computational burden of repeatedly evaluating the marginal gain for each candidate module becomes prohibitive. This computational intractability motivates the development of efficient heuristic methods.

4.3.1. Improvement-based approach As established in Section 4.1.1, the expected value of the minimum of two module scores, $\mathbb{E}[\min\{Y(x_1), Y(x_2)\}]$, admits an exact closed-form solution under the bivariate normal distribution assumption. Leveraging this analytical tractability, we define a pairwise measure of deficiency enhancement for any two modules x_1 and x_2. Specifically, the **improvement** (Ipr) of incorporating module x_2 given a baseline module x_1 is defined as:

$$\mathrm{Ipr}(x_2|x_1) = \mathbb{E}[\min\{Y(x_1), Y(x_2)\}] - \mathbb{E}[Y(x_1)].$$

This quantity captures the expected reduction in the minimum score by adding x_2 to an assessment already containing x_1. It follows directly from the definition that $\mathrm{Ipr}(x_2|x_1) = 0$ when x_1 and x_2 are identical.

Next, consider an existing set of selected modules $S = \{x_2, x_3, x_4\}$. We can extend the pairwise improvement concept by defining the marginal improvement of adding a new module x_1 to the set S as:

$$\mathrm{Ipr}(x_1|S) = \mathbb{E}\left[\min\left\{Y(x) : x \in S \cup \{x_1\}\right\}\right] - \mathbb{E}\left[\min\left\{Y(x) : x \in S\right\}\right].$$

This metric quantifies the expected reduction in the overall minimum score resulting from the inclusion of x_1 into the current assessment portfolio S.

Now let's consider an extreme case, assuming x_1 and x_2 are exactly the same (i.e. $\mathrm{Ipr}(x_1|x_2) = 0$), regardless of the value of $\mathrm{Ipr}(x_1|x_3)$ and $\mathrm{Ipr}(x_1|x_4)$, the value of $\mathrm{Ipr}(x_1|S)$ is always 0. This inspires us to speculate that the value of the improvement of module x_1 over set S is dominated by pairwise minima: $\min_{x \in S} \mathrm{Ipr}(x_1|x)$. So we can define the contribution of module x_1 to set S:

$$\mathrm{Ctr}(x_1|S) = \min_{x \in S} \mathrm{Ipr}(x_1|x)$$

Here, the contribution measure $\mathrm{Ctr}(x_1|S)$ represents quantifies the reduction potential in the minimum score when adding module x to set S.

Based on this we can get our improvement-based approach as follows:

Algorithm 1 Improvement-based approach

1: **Input:** module set $\mathcal{X}$, number of selected modules m.
2: Initialize $S = \{x_1, x_2\}$, where $x_1 = \min_{x \in \mathcal{X}} \mathbb{E}[Y(x)]$, $x_2 = \max_{x \in \mathcal{X}} \mathrm{Ipr}(x_2|x_1)$.
3: **for** $i = 3$ **to** m **do**
4: Choose $x_i = \arg\max_{x \in \mathcal{X}/S} \mathrm{Ctr}(x|S)$
5: Update $S = S \cup x_i$
6: **end for**
7: **return** S

Algorithm 1 likewise employs a greedy strategy for problem resolution, yet it achieves significantly higher computational efficiency compared to the SAA-based greedy algorithm, rendering it suitable for large-scale applications. However, this gain in efficiency comes at the cost of solution quality for more complex problem instances, where the SAA-based algorithm frequently yields superior optimization results. To harness the complementary strengths of both methodologies—the computational efficiency of the heuristic and the solution quality of the SAA-based approach—we propose a novel hybrid greedy algorithm. This mixed strategy, which synergistically combines both approaches, is detailed in the following section.

4.3.2. Mixed approach The workflow for Mixed approach is described below(**Algorithm 2**). In Step 4, the algorithm identifies the q modules that offer the highest contribution to the current selected modules S. This step functions as a preliminary screening phase, designed to efficiently prune the candidate pool and mitigate the computational burden associated with evaluating all M modules. The parameter q is user-defined; a larger value for q yields a more comprehensive search, leading to a solution that more closely approximates the one obtained by the full SAA-based greedy algorithm. Our numerical experiments presented in Section 6 shows that a value of $q = 10$ can be sufficient for achieving near-optimal performance in most practical instances, effectively balancing computational efficiency with solution quality.

Algorithm 2 Mixed approach

1: **Input:** module set $\mathcal{X}$, number of selected modules m, size of candidate set $q = 10$.
2: Initialize $S = \{x_1, x_2\}$, where $x_1 = \min_{x\in\mathcal{X}} \mathbb{E}[Y(x)]$, $x_2 = \max_{x\in\mathcal{X}} \mathrm{Ipr}(x_2|x_1)$.
3: **for** $i = 3$ **to** m **do**
4: Choose candidate set $C \in \mathcal{X}$, $s.t\ C = \arg\max_{|C|=q} \sum_{x\in C} \mathrm{Ctr}(x|S)$
5: Choose $x_i = \arg\max_{x\in C} [U(S \cup \{x\}) - U(S)]$ by SAA
6: Update $S = S \cup x_i$
7: **end for**
8: **return** S

5. A Combinatorial MAB Algorithm for Unknown Score Means

5.1. Model Formulation

We assume that the score has Gaussian noise: $Y(x) = Z(x) + \mathcal{N}(0, \sigma(x))$ and the joint distribution for $Y(x)_{x\in\{\mathcal{X}\}}$ is a multivariate normal distribution with mean vector $\theta = [Z(x_1), ..., Z(x_M)]^T$ and covariance matrix Λ. We use a Bayesian approach to model θ and assume a normal prior:

$$\theta \sim \mathcal{N}(\mu_0, \Sigma_0).$$

As with Xie et al. (2016), we assume Λ, μ_0, Σ_0 are known.

The ith entry of a vector v is denoted as $v(i)$ and the (i, j)th entry of a matrix M is denoted as $M(i, j)$. For an ordered collection of m alternatives $\mathbf{x} = (x_1, ..., x_m)$ with element $x_i \in \{1, ..., M\}$ for each i, we use $v(\mathbf{x})$ to denote the subvector of v with the ith entry equal to $v(x_i)$. We further denote by $M(\mathbf{x}, \mathbf{x}')$ the m-by-m submatrix of M with the (i, j)th entry equal to $M(x_i, x'_j)$.

We consider a situation where in each iteration t, one pilot will attend m different modules $\mathbf{x}_t = (x_{t,1}, ..., x_{t,m})^T$ and we obtain his score vector $\mathbf{y}_t = (Y(x_{t,1}), ..., Y(x_{t,m}))^T$. The conditional distribution of $\mathbf{y}_t$ is:

$$\mathbf{y}_t|\theta, \mathbf{x}_t \sim \mathcal{N}(\theta(\mathbf{x}_t), \Lambda(\mathbf{x}_t, \mathbf{x}_t)).$$

Let $\mathbb{X}_t = (\mathbf{x}_1^T, ..., \mathbf{x}_t^T)$ denote the concatenation of the design points of the previous t iteration and similarly $\mathbb{Y}_t = (\mathbf{y}_1^T, ..., \mathbf{y}_t^T)^T$. Then, the posterior distribution for θ is:

$$\theta_n|\mathbb{X}_t, \mathbb{Y}_t \sim \mathcal{N}(\mu_t, \Sigma_t),$$

where for any vector $\mathbf{x} = (x_1, x_2, ..., x_m)$,

$$\mu_t(\mathbf{x}) = \mu_0(\mathbf{x}) + \Sigma_0(\mathbf{x}, \mathbb{X}_\mathbf{t})(\mathbf{\Sigma_0}(\mathbb{X}_\mathbf{n}, \mathbb{X}_\mathbf{t}) + \mathbf{\Gamma_t})^{-\mathbf{1}}(\mathbb{Y}_\mathbf{t} - \mu_\mathbf{0}(\mathbb{X}_\mathbf{t})),$$

$$\Sigma_n(\mathbf{x}, \mathbf{x}) = \Sigma_0(\mathbf{x}, \mathbf{x}) - \Sigma_0(\mathbf{x}, \mathbb{X}_\mathbf{t})(\mathbf{\Sigma_0}(\mathbb{X}_\mathbf{t}, \mathbb{X}_\mathbf{t}) + \mathbf{\Gamma_t})^{-\mathbf{1}}\mathbf{\Sigma_0}(\mathbb{X}_\mathbf{t}, \mathbf{x}),$$

where Γ_t can be calculated from the observed samples since we know the covariance matrix Λ.

5.2. The GP-UCB Algorithm

To address the challenge of unknown score means, we propose a Bayesian optimization algorithm based on the Gaussian Process Upper Confidence Bound (GP-UCB) principle. This algorithm iteratively maintains a posterior distribution over the mean scores $\boldsymbol{\theta}$ and leverages the heuristic algorithms from Section 4.3 as a selection oracle.

The algorithm's core mechanism involves using an optimistic estimate of the mean scores to guide exploration. Specifically, at each iteration t, we construct the lower confidence bound (LCB) vector $\boldsymbol{\mu}_{t-1} - \boldsymbol{\beta}_t \circ \boldsymbol{\sigma}_{t-1}$, where $\boldsymbol{\sigma}_{t-1}$ contains the posterior standard deviations. This LCB vector represents an optimistic estimate of the true means from the perspective of minimizing the expected minimum score. We then pass this optimistic estimate to a heuristic oracle from Section 4.3, which returns a module set S_t for evaluation. The observed scores from testing S_t are used to update our posterior belief about $\boldsymbol{\theta}$. The complete procedure is detailed in Algorithm 4. Note that in Step 14, the exact computation of $R(S_t)$ remains analytically intractable so we use SAA to estimate it.

Algorithm 3 GP-UCB Algorithm with Efficient Initialization

Require: Module set $\mathcal{X}$; Covariance matrix $\mathbf{\Lambda}$; Prior parameters $\boldsymbol{\mu}_0$, $\mathbf{\Sigma}_0$; Oracle O; Total iterations N

1: Initialize $\boldsymbol{\mu}_0$, $\mathbf{\Sigma}_0$ and count vector $\mathbf{T}_0 \leftarrow \mathbf{0}$
2: $N_0 \leftarrow \lceil M/m \rceil$ ▷ Calculate minimum number of super arms needed
3: **for** $i = 1$ to N_0 **do** ▷ Efficient initial phase: cover all modules with minimal super arms
4: $S_i \leftarrow$ Select a super arm that maximizes the number of unobserved modules
5: Play super arm S_i, observe $\mathbf{y}_i$
6: Update $(\boldsymbol{\mu}, \mathbf{\Sigma}, \mathbf{T}_i)$, Calculate $R(S_i)$ ▷ Same as steps 12, 13, and 14
7: **end for**
8: **for** $t = N_0 + 1$ to N **do** ▷ Main optimization phase
9: $\boldsymbol{\sigma}_{t-1} \leftarrow \sqrt{\mathrm{diag}(\mathbf{\Sigma}_{t-1})}$ ▷ Compute posterior standard deviations
10: $\boldsymbol{\beta}_t[x] \leftarrow \sqrt{6 \ln t / \mathbf{T}_{t-1}[x]}$ for each $x \in \mathcal{X}$ ▷ $\mathbf{T}_{t-1}[x]$: observation count of module x
11: $\tilde{\boldsymbol{\mu}}_t \leftarrow \boldsymbol{\mu}_{t-1} - \boldsymbol{\beta}_t \circ \boldsymbol{\sigma}_{t-1}$ ▷ Construct LCB-based mean estimate
12: $S_t \leftarrow O(\tilde{\boldsymbol{\mu}}, \mathbf{\Lambda})$ ▷ Query selection oracle
13: Observe score vector $\mathbf{y}_t$ for modules S_t on a new pilot
14: Update $(\boldsymbol{\mu}_t, \mathbf{\Sigma}_t)$ with $(\mathbf{y}_t, S_t)$; Update $\mathbf{T}_t[x] \leftarrow \mathbf{T}_t[x] + 1$ for each $x \in S_t$
15: Calculate $R(S_t) = \mathbb{E}[\min\{Y(x) : x \in S_t\}]$, track minimal $R(S_t)$ and corresponding S_t
16: **end for**
17: **return** S^* with minimal $R(S^*)$

5.3. Regret bounds

We establish both distribution-dependent and distribution-independent regret bounds for Algorithm 4, proving $O(\log T)$ and $O(\sqrt{T \log T})$ rates respectively.

For any approximation parameter $0 < \alpha < 1$, we define a super arm S as *bad* if $R(S) > \alpha R(S^*)$, where S^* denotes the true optimal super arm. The suboptimality gap for each super arm S is defined as:

$$\Delta_S = \max\{R(S) - \alpha \cdot R(S^*), 0\}.$$

Let $\mathcal{F}_B = \{S \in \mathcal{F} \mid \Delta_S > 0\}$ denote the set of all bad super arms, and let $E_B \subseteq \mathcal{X}$ be the set of modules contained in at least one bad super arm. For each module $i \in E_B$, we define its minimal suboptimality gap as:

$$\Delta_{i,\min} = \min\{\Delta_S \mid S \in \mathcal{F}_B, i \in S\}.$$

Let σ^2 represent the maximum variance among all modules (the largest diagonal element of covariance matrix $\mathbf{\Lambda}$). Recall that $m = |S|$ is the cardinality of the selected super arm, and $y_{\max}$ is the maximum possible score for any module.

THEOREM 3. *The α-approximation regret of Algorithm 4 over T rounds satisfies the following upper bounds:*

1. ***Distribution-dependent bound:***

$$\sigma^2 m \sum_{i \in E_B} \frac{2136}{\Delta_{i,\min}} \ln T + \left(\frac{\pi^2}{6} + 1\right) \alpha y_{max} M.$$

2. ***Distribution-independent bound:***

$$93\sigma\sqrt{MmT \ln T} + \left(\frac{\pi^2}{6} + 1\right) \alpha y_{max} M.$$

Proof sketch: To establish Theorem 3, we begin by defining the high-probability concentration event $\varepsilon_t = \left\{\text{there exists } i \in [m] \text{ such that } |\hat{\mu}^i_{T_{i,t-1}} - \mu^i| < \sigma(i)\sqrt{\frac{6\ln(t)}{T_{i,t-1}}}\right\}$. We then decompose the α-regret $\text{Reg}_\alpha(T)$ into three components: $\text{Reg}_\alpha(T) = \mathbb{E}[\sum_{t=1}^{m} \Delta_{S_t}] + \mathbb{E}[\sum_{t=m+1}^{T} \mathbb{I}\{\neg\varepsilon_t\}\Delta_{S_t}] + \mathbb{E}[\sum_{t=m+1}^{T} \mathbb{I}\{\varepsilon_t\}\Delta_{S_t}]$. The first term corresponds to the initialization phase and is bounded trivially. The second term is controlled using concentration inequalities. For the third term, we identify certain events that must occur under ε_t and demonstrate that these events occur infrequently, thereby bounding their cumulative contribution. The overall proof idea aligns with the methodologies employed in references (Kveton et al. 2014b), (Degenne and Perchet 2016), and (Chen et al. 2016).

6. Numerical Experiments

We conduct extensive numerical experiments to evaluate the performance of the proposed heuristic algorithms under the known-means scenario and GP-UCB algorithm under the unknown-means scenario. Our experiments are designed to systematically assess the impact of problem scale and correlation structure on algorithmic performance.

6.1. Experimental Setup

6.1.1. Problem Scales We examine four distinct problem scales to represent varying levels of computational complexity:

- **Medium Scale I:** Selecting 8 from 50 modules ($m = 8, M = 50$)
- **Medium Scale II:** Selecting 16 from 50 modules ($m = 16, M = 50$)
- **Large Scale I:** Selecting 10 from 200 modules ($m = 10, M = 200$)
- **Large Scale II:** Selecting 20 from 200 modules ($m = 20, M = 200$)

6.1.2. Correlation Structures For each problem scale, we investigate two fundamental correlation structures:

1. Grouped Structure: Modules are partitioned into groups with distinctive correlation patterns.

- For $M = 50$: 5 groups of 10 modules each
- For $M = 200$: 8 groups of 25 modules each
- Within-group correlation: $\rho_{\text{in}} \approx 0.8$ (highly correlated)
- Between-group correlation: $\rho_{\text{out}} \approx 0$ (weakly correlated)

This structure is further divided into two scenarios:

- **Worst Case:** Modules with the smallest mean scores are concentrated in one group. Additionally, we introduce a positive correlation between module means and noise levels: modules with smaller means exhibit lower assessment variance, while those with larger means have higher variance.
- **Best Case:** Mean scores are distributed evenly across all groups, and noise levels are independent of mean scores.

2. Random Structure: All pairwise correlations are randomly generated with non-negative values, representing an unstructured scenario without inherent module clustering. Noise levels are generated independently of mean scores.

6.1.3. Data Generation Procedure The ground-truth mean score vector $\boldsymbol{\mu}_{\text{true}} \in \mathbb{R}^M$ is generated by sampling from a multivariate normal distribution $\mathcal{N}(\mathbf{70}, \boldsymbol{\Sigma}_0)$, where $\mathbf{70}$ is a vector of baseline scores. The covariance matrix $\boldsymbol{\Lambda}$ is then constructed based on the specified correlation structure:

- For grouped structures, $\boldsymbol{\Lambda}$ is constructed such that modules within the same group exhibit high correlations ($\rho \approx 0.8$), while modules from different groups show negligible correlations ($\rho \approx 0$)
- For the **Worst Case** scenario, the diagonal elements of $\boldsymbol{\Lambda}$ (variances) are set to be positively correlated with the mean scores: $\boldsymbol{\Lambda}_{ii} \propto \boldsymbol{\mu}_{\text{true},i}$, which creates a more challenging environment
- For random structures, $\boldsymbol{\Lambda}$ is randomly generated with non-negative entries

To comprehensively evaluate the performance of our proposed algorithms (Improvement, Mixed, and SAA), we compare them against two intuitive and straightforward benchmark policies:

1. **Smallest Mean (SM)**: This naive baseline selects the m modules with the smallest mean scores, entirely ignoring the covariance structure (variances and correlations) among them.

2. **Clustering-based Selection (CLS)**: This policy first clusters the M modules into m groups based on their correlation matrix using spectral clustering, aiming to group highly correlated modules together. It then selects the module with the smallest mean score from each cluster. This approach seeks diversity by explicitly minimizing correlation within the selected set.

Table 2 Summary of Experimental Configurations

Scale	Structure	Configuration
50/8	Grouped	5 groups; $\rho_{\text{in}} \approx 0.8$; $\rho_{\text{out}} \approx 0$; Worst/Best Case
50/16	Grouped	5 groups; $\rho_{\text{in}} \approx 0.8$; $\rho_{\text{out}} \approx 0$; Worst/Best Case
200/10	Grouped	8 groups; $\rho_{\text{in}} \approx 0.8$; $\rho_{\text{out}} \approx 0$; Worst/Best Case
200/20	Grouped	8 groups; $\rho_{\text{in}} \approx 0.8$; $\rho_{\text{out}} \approx 0$; Worst/Best Case
50/8	Random	Random non-negative correlations
50/16	Random	Random non-negative correlations
200/10	Random	Random non-negative correlations
200/20	Random	Random non-negative correlations

6.2. Results and Discussion

6.2.1. known-means scenario The experimental results across all 12 scenarios (3 structures $\times$ 4 scales) are succinctly presented in Table 3. The best objective value(expected minimum score $\mathbb{E}[\min\{Y(x_1), ..., Y(x_m)\}]$) achieved in each scenario is highlighted in **bold**. We also show the running time of the algorithms.

The comprehensive evaluation reveals that the SAA-based greedy algorithm achieves the best results in most cases, while the Mixed algorithm produces nearly identical results with minimal optimality gaps, demonstrating their superior performance. In contrast, both SM and CLS exhibit significant performance degradation compared to SAA in the challenging Worst Case and Random Case configurations, although the Improvement algorithm shows relatively better performance than these benchmarks while still falling short of Mixed and SAA. For the Best Case conditions, all algorithms except CLS yield comparable results. Overall, the proposed algorithms (Improvement, Mixed, and SAA) consistently outperform the benchmark approaches, confirming the value of systematic optimization; however, it is important to note that SAA's computational time increases substantially with problem scale, particularly as the total number of modules M grows, while the performance differences between all algorithms diminish as the selection size m increases. These results collectively indicate that while SAA provides optimal performance, Mixed offers an attractive balance between solution quality and computational efficiency for practical applications.

References

(2015) *Manual of Criteria for the Qualification of Flight Simulation Training Devices - Volume I - Aeroplanes*. International Civil Aviation Organization, 4th edition, URL `https://store.icao.int/en/manual-of-criteria-for-the-qualification-of-flight-simulation-training-devices-volume-i-aeroplanes`

Table 3 Experimental Results: Comparison of Objective Value and Computation Time

Scale	Case	SM		CLS		Improvement		Mixed		SAA	
		Obj.	Time(s)	Obj.	Time(s)	Obj.	Time(s)	Obj.	Time(s)	Obj.	Time(s)
50/8	Best	56.77	0.0003	57.07	0.0008	56.76	0.10	**56.52**	3.01	56.53	13.47
	Worst	60.72	0.0002	58.59	0.0011	57.81	0.11	57.75	2.97	**57.61**	13.15
	Random	54.28	0.0001	54.41	0.0007	53.00	0.10	52.68	2.90	**52.67**	12.62
50/16	Best	55.59	0.0002	55.62	0.0005	55.63	0.37	55.41	7.17	**55.36**	27.23
	Worst	57.90	0.0001	56.72	0.0027	56.64	0.38	56.59	7.03	**56.10**	27.44
	Random	52.53	0.0002	53.17	0.0005	52.48	0.36	**52.21**	7.16	52.22	27.22
200/10	Best	48.77	0.0001	48.87	0.0021	48.77	0.69	**48.76**	4.48	48.80	72.13
	Worst	52.28	0.0001	53.18	0.0017	51.45	0.68	51.22	4.50	**51.12**	73.72
	Random	47.59	0.0002	47.09	0.0017	46.19	0.71	45.88	4.57	**45.84**	72.85
200/20	Best	**47.87**	0.0001	48.33	0.0019	47.97	2.86	47.91	11.83	47.94	167.31
	Worst	51.90	0.0001	51.76	0.0020	50.38	2.81	50.12	12.07	**49.97**	172.30
	Random	44.31	0.0001	45.18	0.0029	44.23	2.83	44.08	11.69	**43.98**	169.09

(2023) Cae 7000xr series level d full-flight simulator. URL https://www.cae.com/civil-aviation/aviation-simulation-equipment/training-equipment/full-flight-simulators/cae7000xr/.

(2023) Engine failure on takeoff. URL https://pilotworkshop.com/tips/engine-fail-takeoff/.

(2023) Meeting the growing demand for flight simulators. URL https://www.iba.aero/resources/articles/meeting-the-growing-demand-for-flight-simulators/.

(2024) Advancements in flight simulation visual systems: An in-depth analysis of variable collimation display development and testing. *AIAA SCITECH 2024 Forum* URL http://dx.doi.org/10.2514/6.2024-4205.

(2024) Keeping fresh: what is involved in pilot recurrent training URL https://www.flightradar24.com/blog/aviation-explainer-series/pilot-recurrent-training/.

Chen W, Hu W, Li F, Li J, Liu Y, Lu P (2016) Combinatorial multi-armed bandit with general reward functions. *ArXiv* abs/1610.06603, URL https://api.semanticscholar.org/CorpusID:865427.

Chen W, Wang Y, Yuan Y (2013a) Combinatorial multi-armed bandit: General framework and applications. *Proceedings of the 30th International Conference on Machine Learning* 28(1):151–159, URL http://proceedings.mlr.press/v28/chen13.html.

Chen W, Wang Y, Yuan Y (2013b) Combinatorial multi-armed bandit: General framework and applications. Dasgupta S, McAllester D, eds., *Proceedings of the 30th International Conference on Machine Learning*, vol-

ume 28 of *Proceedings of Machine Learning Research*, 151–159 (Atlanta, Georgia, USA: PMLR), URL https://proceedings.mlr.press/v28/chen13a.html.

Clark CE (1961) The greatest of a finite set of random variables. *Operations Research* 9(2):145–162.

Combes R, Talebi MS, Proutière A, Lelarge M (2015) Combinatorial bandits revisited. *Neural Information Processing Systems*, URL https://api.semanticscholar.org/CorpusID:9697688.

Cuvelier T, Combes R, Gourdin E (2021) Statistically efficient, polynomial-time algorithms for combinatorial semi-bandits. *Abstract Proceedings of the 2021 ACM SIGMETRICS / International Conference on Measurement and Modeling of Computer Systems*, 5–6, SIGMETRICS '21 (New York, NY, USA: Association for Computing Machinery), ISBN 9781450380720, URL http://dx.doi.org/10.1145/3410220.3453926.

Damos DL (2003) Pilot abilities and performance URL https://www.researchgate.net/publication/252442125_PILOT_ABILITIES_AND_PERFORMANCE.

Degenne R, Perchet V (2016) Combinatorial semi-bandit with known covariance. *Neural Information Processing Systems*, URL https://api.semanticscholar.org/CorpusID:2716853.

Demirel I, Tekin C (2021) Combinatorial gaussian process bandits with probabilistically triggered arms. *International Conference on Artificial Intelligence and Statistics*, URL https://api.semanticscholar.org/CorpusID:233236143.

Duruaku F, Nguyen B, Sonnenfeld NA, Jentsch F (2024) Unveiling virtual reality overheads and their potential impact on flightcrew training. *The International Journal of Aerospace Psychology* URL https://www.tandfonline.com/doi/full/10.1080/24721840.2024.2396147.

European Union Aviation Safety Agency (2020) Certification specifications for aeroplane flight simulation training devices. Technical report, EASA, Cologne, URL https://www.easa.europa.eu/en/downloads/110530/en.

Federal Aviation Administration (1995) Airplane simulator qualification. Technical Report AC 120-40B, URL https://www.faa.gov/documentLibrary/media/Advisory_Circular/120-40B1.pdf.

Fu MC (2015) *Handbook of Simulation Optimization* (New York, NY: Springer), URL https://link.springer.com/book/10.1007/978-1-4939-1384-8.

Kveton B, Wen Z, Ashkan A, Eydgahi H, Eriksson B (2014a) Matroid bandits: Fast combinatorial optimization with learning. *Conference on Uncertainty in Artificial Intelligence*, URL https://api.semanticscholar.org/CorpusID:7411576.

Kveton B, Wen Z, Ashkan A, Szepesvari C (2014b) Tight regret bounds for stochastic combinatorial semi-bandits. *ArXiv* abs/1410.0949, URL https://api.semanticscholar.org/CorpusID:6152788.

Kveton B, Wen Z, Ashkan A, Szepesvari C (2015) Combinatorial cascading bandits. *ArXiv* abs/1507.04208, URL https://api.semanticscholar.org/CorpusID:1083153.

Nadarajah S, Kotz S (2008) Exact distribution of the max/min of two gaussian random variables. *IEEE Transactions on very large scale integration (VLSI) systems* 16(2):210–212.

National Transportation Safety Board (2012) Review of u.s. civil aviation accidents, calendar year 2010. Technical report, NTSB, Washington, D.C., URL `https://www.ntsb.gov/safety/safety-studies/Documents/ARA1201.pdf`.

Nemhauser GL, Wolsey LA, Fisher ML (1978) An analysis of approximations for maximizing submodular set functions–i. *Math. Program.* 14(1):265–294, ISSN 0025-5610, URL `http://dx.doi.org/10.1007/BF01588971`.

Ross SM (2014) *Introduction to Probability Models* (San Diego, CA: Academic Press), 10th edition.

Szczepanska A, et al. (2025) The role of temperament and personality traits in assessing reaction efficiency in pilot candidates. *Personality and Individual Differences* URL `https://www.sciencedirect.com/science/article/abs/pii/S0191886925000431`.

Wang Y, Chen W, Vojnovi'c M (2023) Combinatorial bandits for maximum value reward function under max value-index feedback. *ArXiv* abs/2305.16074, URL `https://api.semanticscholar.org/CorpusID:258887807`.

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

Zhang R, Combes R (2025) Thompson sampling for combinatorial bandits: polynomial regret and mismatched sampling paradox. *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24 (Red Hook, NY, USA: Curran Associates Inc.), ISBN 9798331314385.

Online Appendix

EC.1. Proof of Theorem 1

Proof. Monotone Increasing is quite obvious. Here we show that $U(\cdot)$ is submodular. i.e., for any $S \subseteq T \subseteq \mathcal{X}$ and $x \in \mathcal{X} \setminus T$:

$$U(S \cup \{x\}) - U(S) \geq U(T \cup \{x\}) - U(T).$$

which is equivalent to:

$$R(S) - R(S \cup \{x\}) \geq R(T) - R(T \cup \{x\}).$$

Let $Y_S = \min_{i \in S} Y(i, \xi)$, $Y_T = \min_{i \in T} Y(i, \xi)$, $Y_x = Y(x, \xi)$. Then $Y_T \leq Y_S$. We have:

$$R(S) - R(S \cup \{x\}) = \mathbb{E}\left[Y_S - \min\{Y_S, Y_x\}\right],$$

$$R(T) - R(T \cup \{x\}) = \mathbb{E}\left[Y_T - \min\{Y_T, Y_x\}\right].$$

Consider the function $g(y) = y - \min\{y, x\}$. For fixed x:

- If $y \leq x$, then $g(y) = 0$;
- If $y > x$, then $g(y) = y - x$.

Thus, $g(y)$ is non-decreasing in y. Since $Y_A \geq Y_B$, we have:

$$Y_S - \min\{Y_S, Y_x\} \geq Y_T - \min\{Y_T, Y_x\}.$$

Taking expectation:

$$\mathbb{E}\left[Y_S - \min\{Y_S, Y_x\}\right] \geq \mathbb{E}\left[Y_T - \min\{Y_T, Y_x\}\right],$$

which implies:

$$U(S \cup \{x\}) - U(S) \geq U(T \cup \{x\}) - U(T).$$

Therefore, $U(\cdot)$ is submodular. ∎

EC.1.1. Proof of Theorem 3

LEMMA EC.1. *For any $\epsilon > 0$ and any $n = T_{i,t-1} \in \mathbb{Z}_+$, we have :*

$$\Pr[|\mu_{t-1}(i) - \mu(i)| \geq \epsilon_i] \leq \exp(-\frac{n\epsilon_i^2}{2\sigma^2(i)})$$

Proof. Fix $t \geq 1$ and $i \in [m]$. Conditioned on $(\mathbf{y}_1, ..., \mathbf{y}_{t-1}), \{S_1, ..., S_{t-1}\}$ are deterministic, and $\mu(i) \sim N(\mu_{t-1}(i), \sigma^2_{t-1}(i))$. Now, if $r \sim N(0,1)$,then:

$$\begin{aligned}\Pr\{r > c\} &= e^{-c^2/2}(2\pi)^{-1/2} \int e^{-(r-c)^2/2 - c(r-c)} dr \\ &\leq e^{-c^2/2} \Pr\{r > 0\} = (1/2)e^{-c^2/2}\end{aligned}$$

for $c > 0$, since $e^{-c(r-c)} \leq 1$ for $r \geq c$. Therefore, $\Pr\{|\mu_{t-1}(i) - \mu(i)| > \lambda_i \sigma_{t-1}(i)\} \leq e^{-\lambda_i/2}$

Assume that arm i has been observed n times, yielding n observations. Further assume that these n observations are our only observational data. Using GP modeling, we obtain:

$$\sigma^2_n(i) = k(i,i) - \mathbf{k}_n^\top(i)(\mathbf{K}_n + \sigma^2(i)\mathbf{I}_n)^{-1}\mathbf{k}_n(i)$$

where:

- $\mathbf{k}_n(i) = [k(i,i), \ldots, k(i,i)]^\top \in \mathbb{R}^n$
- $\mathbf{K}_n = \sigma_0^2 \mathbf{1}_n \mathbf{1}_n^\top$ $(\sigma_0^2 = k(i,i))$

For $\mathbf{A} = \sigma^2(i)\mathbf{I}_n$, $\mathbf{u} = \sigma_0^2 \mathbf{1}_n$, $\mathbf{v} = \mathbf{1}_n$, by **Sherman-Morrison** formula:

$$\begin{aligned}(\mathbf{K}_n + \sigma^2(i)\mathbf{I}_n)^{-1} &= \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^\top\mathbf{A}^{-1}}{1 + \mathbf{v}^\top\mathbf{A}^{-1}\mathbf{u}} \\ &= \frac{1}{\sigma^2(i)}\mathbf{I}_n - \frac{\sigma_0^2/\sigma(i)^4 \cdot \mathbf{1}_n\mathbf{1}_n^\top}{1 + n\sigma_0^2/\sigma^2(i)}\end{aligned}$$

So we have:

$$\mathbf{k}_n^\top(i)(\mathbf{K}_n + \sigma^2(i)\mathbf{I}_n)^{-1}\mathbf{k}_n(i) = \frac{\sigma_0^4 n}{\sigma^2(i) + n\sigma_0^2}$$

Substitute the results into the posterior variance formula:

$$\begin{aligned}\sigma^2_n(i) &= \sigma_0^2 - \frac{\sigma_0^4 n}{\sigma^2(i) + n\sigma_0^2} = \frac{\sigma_0^2\sigma^2(i)}{\sigma^2(i) + n\sigma_0^2} \\ &= \frac{n\sigma_0^2}{\sigma^2(i) + n\sigma_0^2} \cdot \frac{\sigma^2(i)}{n} \leq \frac{\sigma^2(i)}{n}\end{aligned}$$

Considering that adding other observation points will not increase the uncertainty of arm i, so we have $\sigma_{t-1}(i) \leq \sigma_n(i) \leq \frac{\sigma(i)}{\sqrt{n}}$. Therefore:

$$\Pr\left\{|\mu_{t-1}(i) - \mu(i)| > \lambda_i \frac{\sigma(i)}{\sqrt{n}}\right\} \leq \Pr\{|\mu_{t-1}(i) - \mu(i)| > \lambda_i \sigma_{t-1}(i)\} \leq e^{-\lambda_i/2}$$

Denote $\epsilon_i = \lambda_i \frac{\sigma(i)}{\sqrt{n}}$, then we have $\Pr[|\mu_{t-1}(i) - \mu(i)| \geq \epsilon_i] \leq \exp(-\frac{n\epsilon_i^2}{2\sigma^2(i)})$ ■

LEMMA EC.2. *If for any $i \in [m]$ we have $\mu_1(i) \geq \mu_2(i)$ (Abbreviated as $\boldsymbol{\mu}_1 \geq \boldsymbol{\mu}_2$), then for any super arm $S \in \mathcal{F}$,we have:*

$$r_{P_1}(S) \geq r_{P_2}(S)$$

LEMMA EC.3. *If for any $i \in [m]$ we have $\mu_1(i) - \mu_2(i) \leq \Lambda_i (\Lambda_i > 0)$ ($\boldsymbol{\mu}_1 \geq \boldsymbol{\mu}_2$). For any super arm $S \in \mathcal{F}$, let $\Lambda_S = \max_{i \in S} \Lambda_i$, then we have:*

$$r_{P_1}(S) - r_{P_2}(S) \leq \Lambda_S$$

Proof. For any super arm $S = \{i_1, ..., i_n\}$, denote the corresponding mean vectors in distributions P_1 and P_2 as $\boldsymbol{\mu}_{1,S} = [\mu_1(i_1), ..., \mu_1(i_n)]^T$, $\boldsymbol{\mu}_{2,S} = [\mu_1(i_2), ..., \mu_2(i_n)]^T$. Note that P_1 and P_2 are both multivariate normal distributions, and have same covariance matrices Σ_S.

Now, let $\mathbf{z} = [z_1, ..., z_n]^T \sim N(0, \Sigma_S)$. We have:

$$r_{P_1}(S) = \int \cdots \int_{\mathbf{z}} \min\{\mu_1(i_1) + z_1, \mu_1(i_2) + z_2, \ldots, \mu_1(i_n) + z_n\}\, dz_1 \ldots dz_k$$

$$r_{P_2}(S) = \int \cdots \int_{\mathbf{z}} \min\{\mu_2(i_1) + z_1, \mu_2(i_2) + z_2, \ldots, \mu_2(i_n) + z_n\}\, dz_1 \ldots dz_k$$

Denote $f(\boldsymbol{\mu}_{1,S} + \mathbf{z}) = \min\{\mu_1(i_1) + z_1, \mu_1(i_2) + z_2, \ldots, \mu_1(i_n) + z_n\}$ and $f(\boldsymbol{\mu}_{2,S} + \mathbf{z}) = \min\{\mu_2(i_1) + z_1, \mu_2(i_2) + z_2, \ldots, \mu_2(i_n) + z_n\}$. We want to show that $f(\mu_{1,S} + \mathbf{z}) - f(\mu_{2,S} + \mathbf{z}) \leq \Lambda_S$.

If there exists $i, j \in S = \{i_1, ..., i_n\}$ s.t. $\mu_1(i) + z_i = f(\mu_{1,S} + \mathbf{z})$, $\mu_2(j) + z_j = f(\mu_{2,S} + \mathbf{z})$, $f(\mu_{1,S} + \mathbf{z}) - f(\mu_{2,S} + \mathbf{z}) > \Lambda_S$. Since $\mu_1(i) + z_i = f(\mu_{1,S} + \mathbf{z}) = \min\{\mu_1(i_1) + z_1, \mu_1(i_2) + z_2, \ldots, \mu_1(i_n) + z_n\}$, we have: $\mu_1(j) + z_j > \mu_1(i) + z_i$.

Thus $\mu_1(j) - \mu_2(j) = \mu_1(j) + z_j - (\mu_2(j) + z_j) > \mu_1(i) + z_i - (\mu_2(j) + z_j) > \Lambda_S$, which is contradict with $\boldsymbol{\mu}_1 \geq \boldsymbol{\mu}_2$.

Therefore

$$\begin{aligned} r_{P_1}(S) - r_{P_2}(S) &= \int \cdots \int_{\mathbf{z}} f(\boldsymbol{\mu}_{1,S} + \mathbf{z}) - f(\boldsymbol{\mu}_{2,S} + \mathbf{z})\, dz_1 \ldots dz_k \\ &\leq \int \cdots \int_{\mathbf{z}} \Lambda_S\, dz_1 \ldots dz_k = \Lambda_S \end{aligned}$$

■

LEMMA EC.4. *Define an event in each round $t (m + 1) \leq t \leq T$:*

$$\mathcal{H}_t = \left\{ 0 < \Delta_{S_t} \leq 2 \max_{i \in S_t} \sigma(i) \sqrt{\frac{6 \ln(t)}{T_{i,t-1}}} \right\}$$

Then the α-approximation regret in T rounds is at most

$$\mathbb{E}[\sum_{t=m+1}^{T}\mathbb{I}\{\mathcal{H}_t\}\Delta_{S_t}]+(\frac{\pi^2}{6}+1)\alpha Mm.$$

Proof. For each super arm S, we define:

$$\Delta_S=\max\{r_P(S)-\alpha\cdot r_P(S^*),0\}$$

From $m+1\le t\le T$, Define an event:

$$\varepsilon_t=\left\{\text{there exists } i\in[m] \text{ such that } |\hat{\mu}^i_{T_{i,t-1}}-\mu^i|\ge\sigma(i)\sqrt{\frac{6\ln(t)}{T_{i,t-1}}}\right\}$$

We bound the α-approximation regret as:

$$\begin{aligned}\mathrm{Reg}_{P,\alpha}(T)&=\sum_{t=1}^{T}\mathbb{E}\left[\alpha r_P(S^*)-r_P(S_t)\right]\\&\le\sum_{t=1}^{T}\mathbb{E}[\Delta_{S_t}]\\&=\mathbb{E}[\sum_{t=1}^{m}\Delta_{S_t}]+\mathbb{E}[\sum_{t=m+1}^{T}\mathbb{I}\{\varepsilon_t\}\Delta_{S_t}]\\&+\mathbb{E}[\sum_{t=m+1}^{T}\mathbb{I}\{\neg\varepsilon_t\}\Delta_{S_t}]\end{aligned}\tag{EC.1}$$

(a) the first term

Let $M=100$, The first term can be trivially bounded as:

$$\mathbb{E}[\sum_{t=1}^{m}\Delta_{S_t}]\le\sum_{t=1}^{m}\alpha\cdot r_P(S^*)\le m\cdot\alpha M\tag{EC.2}$$

(b) the second term

By Lemma 1 we know that for any $i\in[m], t\ge m+1$, denote $c_i=\sigma(i)\sqrt{\frac{6\ln(t)}{T_{i,t-1}}}$, we have:

$$\Pr[|\hat{\mu}_{t-1}(i)-\mu(i)|\ge c_i]\le\exp(-\frac{nc_i^2}{2\sigma^2(i)})=t^{-3}$$

Therefore

$$\begin{aligned}\mathbb{E}[\sum_{t=m+1}^{T}\mathbb{I}\{\varepsilon_t\}]&\le\sum_{t=m+1}^{T}\sum_{i=1}^{m}\sum_{l=1}^{t-1}t^{-3}\\&\le m\sum_{t=m+1}^{T}t^{-2}\\&\le\frac{\pi^2}{6}m\end{aligned}\tag{EC.3}$$

and then the second term in (6) can be bounded as

$$\mathbb{E}[\sum_{t=m+1}^{T} \mathbb{I}\{\varepsilon_t\}\Delta_{S_t}] \leq \frac{\pi^2}{6} m \cdot (\alpha \cdot r_P(S^*)) \leq \frac{\pi^2}{6} \alpha M m \tag{EC.4}$$

(c) the third term

We fix $t > m$ and first asuume $\neg\varepsilon_t$ happens. For each $i \in [m]$, since $\neg\varepsilon_t$ happens, we have:

$$|\hat{\mu}_{T_{i,t-1}}(i) - \mu(i)| < c_i \quad \forall i \in [m] \tag{EC.5}$$

Recall that in round t, the input to the oracle is $\underline{P} = \mathcal{N}(\underline{\boldsymbol{\mu}}, \Sigma)$, where the mean vector of $\underline{P}$ is:

$$\underline{\mu}(i) = \hat{\mu}(i) - c_i \quad \forall i \in [m] \tag{EC.6}$$

From (10) and (11) we konw that $\mu(i) > \underline{\mu}(i) > \mu(i) - 2c_i$ for all $i \in [m]$. Thus, from Lemma 2 we have:

$$r_{\underline{P}}(S) \leq r_P(S) \quad \forall S \in \mathcal{F}. \tag{EC.7}$$

and from Lemma 3 we have:

$$r_{\underline{P}}(S) \geq r_P(S) - 2\Lambda_S \quad \forall S \in \mathcal{F}. \tag{EC.8}$$

where $\Lambda = \max_i c_i$

Also, from the fact that the algorithm chooses S_t in the t-th round, we have:

$$r_{\underline{P}}(S_t) \leq \alpha \cdot \min_{S \in \mathcal{F}} r_{\underline{P}}(S) \leq \alpha \cdot r_{\underline{P}}(S^*). \tag{EC.9}$$

From (12),(13) and (14) we have:

$$r_P(S) - 2\Lambda_S \leq r_{\underline{P}}(S_t) \leq \alpha \cdot r_{\underline{P}}(S^*) \leq \alpha \cdot r_P(S^*) \tag{EC.10}$$

which implies:

$$\Delta_{S_t} \leq 2\Lambda_S$$

Therefore, when $\neg\varepsilon_t$ happens, we always have $\Delta_{S_t} \leq 2 \max_{i \in S_t} c_i$.

This implies:

$$\{\neg\varepsilon_t, \Delta_{S_t} > 0\} \Longrightarrow \{0 < \Delta_{S_t} \leq 2 \max_{i \in S_t} c_i\} = \mathcal{H}_t.$$

Hence, the third term in (6) can be bounded as:

$$\begin{aligned}\mathbb{E}[\sum_{t=m+1}^{T}\mathbb{I}\{\neg\varepsilon_t\}\Delta_{S_t}] &= \mathbb{E}[\sum_{t=m+1}^{T}\mathbb{I}\{\neg\varepsilon_t.\Delta_{S_t}>0\}\Delta_{S_t}] \\ &\leq \mathbb{E}[\sum_{t=m+1}^{T}\mathbb{I}\{\mathcal{H}_t\}\Delta_{S_t}]\end{aligned} \tag{EC.11}$$

Finally, by combining (6),(7),(9),(16) we have:

$$\mathrm{Reg}_{P,\alpha}(T) \leq \mathbb{E}[\sum_{t=m+1}^{T}\mathbb{I}\{\mathcal{H}_t\}\Delta_{S_t}] + (\frac{\pi^2}{6}+1)\alpha Mm$$

■

Now it remains to bound $\mathbb{E}[\sum_{t=m+1}^{T}\mathbb{I}\{\mathcal{H}_t\}\Delta_{S_t}]$.

Define two decreasing sequences of positive contants:

$$1=\beta_0>\beta_1>\beta_2>\ldots$$

$$\alpha_1>\alpha_2>\ldots$$

such that $\lim_{k\to\infty}\alpha_k = \lim_{k\to\infty}\beta_k = 0$. We choose $\{\alpha_k\}$ and $\{\beta_k\}$ as in Theorem 4 of (Kveton et al. 2014b), which satisfy:

$$\sqrt{6}\sum_{k=1}^{\infty}\frac{\beta_{k-1}-\beta_k}{\sqrt{\alpha_k}} \leq 1 \tag{EC.12}$$

and

$$\sum_{k=1}^{\infty}\frac{\alpha_k}{\beta_k} < 267. \tag{EC.13}$$

For $t\in\{m+1,...,T\}$ and $k\in\mathbb{Z}_+$, let

$$m_{k,t} = \begin{cases}\alpha_k\left(\frac{2\sigma K}{\Delta_{S_t}}\right)^2\cdot\ln(T) & \Delta_{S_t}>0,\\ +\infty & \Delta_{S_t}=0,\end{cases}$$

and

$$A_{k,t} = \{i\in S_t | T_{i,t-1}\leq m_{k,t}\}.$$

Then we define an event:

$$\mathcal{G}_{k,t} = \{|A_{k,t}|\geq\beta_k K\},$$

which means "in the $t-th$ round, at least $\beta_k K$ arms in S_t had been observed at most $m_{k,t}$ times."

LEMMA EC.5. *In the $t-th$ round, at least $\beta_k K$ arms in S_t had been observed at most $m_{k,t}$ times.*

Proof. Assume that $\mathcal{H}_t$ happens and that none of $\mathcal{G}_{1,t}, \mathcal{G}_{2,t}, \ldots$ happens. Then $|A_{k,t}| < \beta_k K$ for all $k \in \mathbb{Z}_+$.

Let $A_{0,t} = S_t$ and $\bar{A}_{k,t} = S_t \setminus A_{k,t}$ for $k \in \mathbb{Z}_+ \cup \{0\}$. It is easy to see $\bar{A}_{k-1,t} \subseteq \bar{A}_{k,t}$ for all $k \in \mathbb{Z}_+$. Note that $\lim_{k\to\infty} m_{k,t} = 0$. Thus there exists $N \in \mathbb{Z}_+$ such that $\bar{A}_{k,t} = S_t$ for all $k \geq N$, and then we have $S_t = \bigcup_{k=1}^{\infty} \left(\bar{A}_{k,t} \setminus \bar{A}_{k-1,t}\right)$. Finally, note that for all $i \in \bar{A}_{k,t}$, we have $T_{i,t-1} > m_{k,t}$. Therefore

$$\begin{aligned}
\sum_{i\in S_t} \frac{1}{\sqrt{T_{i,t-1}}} &= \sum_{k=1}^{\infty} \sum_{i\in \bar{A}_{k,t}\setminus\bar{A}_{k-1,t}} \frac{1}{\sqrt{T_{i,t-1}}} \leq \sum_{k=1}^{\infty} \sum_{i\in \bar{A}_{k,t}\setminus\bar{A}_{k-1,t}} \frac{1}{\sqrt{m_{k,t}}} \\
&= \sum_{k=1}^{\infty} \frac{\left|\bar{A}_{k,t} \setminus \bar{A}_{k-1,t}\right|}{\sqrt{m_{k,t}}} = \sum_{k=1}^{\infty} \frac{\left|A_{k-1,t} \setminus A_{k,t}\right|}{\sqrt{m_{k,t}}} = \sum_{k=1}^{\infty} \frac{\left|A_{k-1,t}\right| - \left|A_{k,t}\right|}{\sqrt{m_{k,t}}} \\
&= \frac{|S_t|}{\sqrt{m_{1,t}}} + \sum_{k=1}^{\infty} |A_{k,t}| \left(\frac{1}{\sqrt{m_{k+1,t}}} - \frac{1}{\sqrt{m_{k,t}}}\right) \\
&< \frac{K}{\sqrt{m_{1,t}}} + \sum_{k=1}^{\infty} \beta_k K \left(\frac{1}{\sqrt{m_{k+1,t}}} - \frac{1}{\sqrt{m_{k,t}}}\right) \\
&= \sum_{k=1}^{\infty} \frac{(\beta_{k-1} - \beta_k)K}{\sqrt{m_{k,t}}}.
\end{aligned}$$

Note that we asuume $\mathcal{H}_t$ happens. Denote $\sigma = \max_{i\in S_t} \sigma(i)$, then we have:

$$\begin{aligned}
\Delta_{S_t} \leq 2\Lambda_S &= 2 \max_{i\in S_t} \sigma(i) \sqrt{\frac{6\ln(t)}{T_{i,t-1}}} \\
&\leq 2\sigma \cdot \sqrt{6\ln(T)} \cdot \max_{i\in S_t} \frac{1}{\sqrt{T_{i,t-1}}} \\
&\leq 2\sigma \cdot \sqrt{6\ln(T)} \cdot \sum_{i\in S_t} \frac{1}{\sqrt{T_{i,t-1}}} \\
&< 2\sigma \cdot \sqrt{6\ln(T)} \cdot \sum_{k=1}^{\infty} \frac{(\beta_{k-1} - \beta_k)K}{\sqrt{m_{k,t}}} \\
&= \sqrt{6} \sum_{k=1}^{\infty} \frac{\beta_{k-1} - \beta_k}{\sqrt{\alpha_k}} \cdot \Delta_{S_t} \leq \Delta_{S_t},
\end{aligned}$$

We reach a contradiction here. The proof of lemma 5 is completed. ∎

By Lemma 5 we have

$$\sum_{t=m+1}^{T} \mathbb{I}\{\mathcal{H}_t\}\Delta_{S_t} \leq \sum_{k=1}^{\infty} \sum_{t=m+1}^{T} \mathbb{I}\{\mathcal{G}_{k,t}, \Delta_{S_t} > 0\}\Delta_{S_t}.$$

For $i \in [m], k \in \mathbb{Z}_+, t \in \{m+1, ..., T\}$, define an event $Gi, k, t = Gk, tiSt, Ti, t1mk, t.\mathcal{G}_{i,k,t} = \mathcal{G}_{k,t} \wedge \{i \in S_t, T_{i,t-1} \leq m_{k,t}\}$.

Then by the definitions of $\mathcal{G}_{k,t}$ and $\mathcal{G}_{i,k,t}$ we have

$$\mathbb{I}\{\mathcal{G}_{k,t}, \Delta_{S_t} > 0\} \leq \frac{1}{\beta_k K} \sum_{i \in E_{\mathrm{B}}} \mathbb{I}\{\mathcal{G}_{i,k,t}, \Delta_{S_t} > 0\}.$$

Therefore

$$\sum_{t=m+1}^{T} \mathbb{I}\{\mathcal{H}_t\} \Delta_{S_t} \leq \sum_{i \in E_{\mathrm{B}}} \sum_{k=1}^{\infty} \sum_{t=m+1}^{T} \mathbb{I}\{\mathcal{G}_{i,k,t}, \Delta_{S_t} > 0\} \frac{\Delta_{S_t}}{\beta_k K}.$$

For each arm $i \in E_B$, suppose i is contained in N_i bad super arms $S^{\mathrm{B}}_{i,1}, S^{\mathrm{B}}_{i,2}, \ldots, S^{\mathrm{B}}_{i,N_i}$. Let $\Delta_{i,l} = \Delta_{S^{\mathrm{B}}_{i,l}} (l \in [N_i])$. Without loss of generality, we assume $\Delta_{i,1} \geq \Delta_{i,2} \geq \ldots \geq \Delta_{i,N_i}$. Note that $\Delta_{i,N_i} = \Delta_{i,\min}$. For convenience, we also define $\Delta_{i,0} = +\infty,$ i.e., $\alpha_k \left(\frac{2\sigma K}{\Delta_{i,0}}\right)^2 = 0$. Then we have

$$\begin{aligned}
&\sum_{t=m+1}^{T} \mathbb{I}\{\mathcal{H}_t\} \Delta_{S_t} \\
&\leq \sum_{i \in E_{\mathrm{B}}} \sum_{k=1}^{\infty} \sum_{t=m+1}^{T} \sum_{l=1}^{N_i} \mathbb{I}\{\mathcal{G}_{i,k,t}, S_t = S^{\mathrm{B}}_{i,l}\} \frac{\Delta_{S_t}}{\beta_k K} \\
&\leq \sum_{i \in E_{\mathrm{B}}} \sum_{k=1}^{\infty} \sum_{t=m+1}^{T} \sum_{l=1}^{N_i} \mathbb{I}\{T_{i,t-1} \leq m_{k,t}, S_t = S^{\mathrm{B}}_{i,l}\} \frac{\Delta_{i,l}}{\beta_k K} \\
&= \sum_{i \in E_{\mathrm{B}}} \sum_{k=1}^{\infty} \sum_{t=m+1}^{T} \sum_{l=1}^{N_i} \mathbb{I}\left\{T_{i,t-1} \leq \alpha_k \left(\frac{2\sigma K}{\Delta_{i,l}}\right)^2 \ln T, S_t = S^{\mathrm{B}}_{i,l}\right\} \frac{\Delta_{i,l}}{\beta_k K} \\
&= \sum_{i \in E_{\mathrm{B}}} \sum_{k=1}^{\infty} \sum_{t=m+1}^{T} \sum_{l=1}^{N_i} \sum_{j=1}^{l} \mathbb{I}\left\{\alpha_k \left(\frac{2\sigma K}{\Delta_{i,j-1}}\right)^2 \ln T < T_{i,t-1} \leq \alpha_k \left(\frac{2\sigma K}{\Delta_{i,j}}\right)^2 \ln T, S_t = S^{\mathrm{B}}_{i,l}\right\} \frac{\Delta_{i,l}}{\beta_k K} \\
&\leq \sum_{i \in E_{\mathrm{B}}} \sum_{k=1}^{\infty} \sum_{t=m+1}^{T} \sum_{l=1}^{N_i} \sum_{j=1}^{l} \mathbb{I}\left\{\alpha_k \left(\frac{2\sigma K}{\Delta_{i,j-1}}\right)^2 \ln T < T_{i,t-1} \leq \alpha_k \left(\frac{2\sigma K}{\Delta_{i,j}}\right)^2 \ln T, S_t = S^{\mathrm{B}}_{i,l}\right\} \frac{\Delta_{i,j}}{\beta_k K} \\
&\leq \sum_{i \in E_{\mathrm{B}}} \sum_{k=1}^{\infty} \sum_{t=m+1}^{T} \sum_{l=1}^{N_i} \sum_{j=1}^{N_i} \mathbb{I}\left\{\alpha_k \left(\frac{2\sigma K}{\Delta_{i,j-1}}\right)^2 \ln T < T_{i,t-1} \leq \alpha_k \left(\frac{2\sigma K}{\Delta_{i,j}}\right)^2 \ln T, S_t = S^{\mathrm{B}}_{i,l}\right\} \frac{\Delta_{i,j}}{\beta_k K} \\
&\leq \sum_{i \in E_{\mathrm{B}}} \sum_{k=1}^{\infty} \sum_{t=m+1}^{T} \sum_{j=1}^{N_i} \mathbb{I}\left\{\alpha_k \left(\frac{2\sigma K}{\Delta_{i,j-1}}\right)^2 \ln T < T_{i,t-1} \leq \alpha_k \left(\frac{2\sigma K}{\Delta_{i,j}}\right)^2 \ln T\right\} \frac{\Delta_{i,j}}{\beta_k K}
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{i\in E_{\mathrm{B}}}\sum_{k=1}^{\infty}\sum_{j=1}^{N_i}\left(\alpha_k\left(\frac{2\sigma K}{\Delta_{i,j}}\right)^2\ln T-\alpha_k\left(\frac{2\sigma K}{\Delta_{i,j-1}}\right)^2\ln T\right)\frac{\Delta_{i,j}}{\beta_k K}\\
&=4\sigma^2 K\left(\sum_{k=1}^{\infty}\frac{\alpha_k}{\beta_k}\right)\ln T\cdot\sum_{i\in E_{\mathrm{B}}}\sum_{j=1}^{N_i}\left(\frac{1}{\Delta_{i,j}^2}-\frac{1}{\Delta_{i,j-1}^2}\right)\Delta_{i,j}\\
&\leq 1068\sigma^2 K\ln T\cdot\sum_{i\in E_{\mathrm{B}}}\sum_{j=1}^{N_i}\left(\frac{1}{\Delta_{i,j}^2}-\frac{1}{\Delta_{i,j-1}^2}\right)\Delta_{i,j},
\end{aligned}
$$

where the last inequality is due to (18).

Finally, for each $i\in E_B$ we have

$$
\begin{aligned}
\sum_{j=1}^{N_i}\left(\frac{1}{\Delta_{i,j}^2}-\frac{1}{\Delta_{i,j-1}^2}\right)\Delta_{i,j}&=\frac{1}{\Delta_{i,N_i}}+\sum_{j=1}^{N_i-1}\frac{1}{\Delta_{i,j}^2}(\Delta_{i,j}-\Delta_{i,j+1})\\
&\leq\frac{1}{\Delta_{i,N_i}}+\int_{\Delta_{i,N_i}}^{\Delta_{i,1}}\frac{1}{x^2}\mathrm{d}x\\
&=\frac{2}{\Delta_{i,N_i}}-\frac{1}{\Delta_{i,1}}\\
&<\frac{2}{\Delta_{i,\min}}.
\end{aligned}
$$

It follows that

$$
\begin{aligned}
\sum_{t=m+1}^{T}\mathbb{I}\{\mathcal{H}_t\}\Delta_{S_t}&\leq 1068\sigma^2 K\ln T\cdot\sum_{i\in E_{\mathrm{B}}}\frac{2}{\Delta_{i,\min}}\\
&=\sigma^2 K\sum_{i\in E_{\mathrm{B}}}\frac{2136}{\Delta_{i,\min}}\ln T.
\end{aligned}
\tag{EC.14}
$$

Combining (19) with Lemma 4, the distribution-dependent regret bound in Theorem 1 is proved.

To prove the distribution-independent bound, we decompose $\sum_{t=m+1}^{T}\mathbb{I}\{\mathcal{H}_t\}\Delta_{S_t}$ into two parts:

$$
\begin{aligned}
\sum_{t=m+1}^{T}\mathbb{I}\{\mathcal{H}_t\}\Delta_{S_t}&=\sum_{t=m+1}^{T}\mathbb{I}\{\mathcal{H}_t,\Delta_{S_t}\leq\epsilon\}\Delta_{S_t}\\
&\quad+\sum_{t=m+1}^{T}\mathbb{I}\{\mathcal{H}_t,\Delta_{S_t}>\epsilon\}\Delta_{S_t}\\
&\leq\epsilon T+\sum_{t=m+1}^{T}1\{\mathcal{H}_t,\Delta_{S_t}>\epsilon\}\Delta_{S_t},
\end{aligned}
\tag{EC.15}
$$

where $\epsilon>0$ is a constant to be determined. The second term can be bounded in the same way as in the proof of the distribution-dependent regret bound, except that we only consider the case

$\Delta_{S_t} > \epsilon$. Thus we can replace (19) by

$$\begin{aligned}
&\sum_{t=m+1}^{T} \mathbb{I}\{\mathcal{H}_t, \Delta_{S_t} > \epsilon\}\Delta_{S_t} \\
&\leq \sigma^2 K \sum_{i \in E_{\mathrm{B}}, \Delta_{i,\min} > \epsilon} \frac{2136}{\Delta_{i,\min}} \ln T \\
&\leq \sigma^2 K m \frac{2136}{\epsilon} \ln T.
\end{aligned} \tag{EC.16}$$

It follows that

$$\sum_{t=m+1}^{T} \mathbb{I}\{\mathcal{H}_t\}\Delta_{S_t} \leq \epsilon T + \sigma^2 K m \frac{2136}{\epsilon} \ln T.$$

Finally, letting $\epsilon = \sqrt{\frac{2136\sigma^2 K m \ln T}{T}}$, we get

$$\sum_{t=m+1}^{T} \mathbb{I}\{\mathcal{H}_t\}\Delta_{S_t} \leq 2\sqrt{2136\sigma^2 K m T \ln T} < 93\sigma\sqrt{mKT \ln T}.$$

Combining this with Lemma 4, we conclude the proof of the distribution-independent regret bound in Theorem 1. ■