Multiplier Bootstrap-based Exploration

Runzhe Wan * [1] **Haoyu Wei** * [2 3] **Branislav Kveton** [1] **Rui Song** [1]

Abstract

Despite the great interest in the bandit problem, designing efficient algorithms for complex models remains challenging, as there is typically no analytical way to quantify uncertainty. We propose Multiplier Bootstrap-based Exploration (`MBE`), a novel exploration strategy that is applicable to any reward model amenable to weighted loss minimization. We prove both instance-dependent and instance-independent rate-optimal regret bounds for `MBE` in sub-Gaussian multi-armed bandits. With extensive simulation and real-data experiments, we show the generality and adaptivity of `MBE`.

1. Introduction

The bandit problem has found wide applications in various areas such as clinical trials (Durand et al., 2018), finance (Shen et al., 2015), recommendation systems (Zhou et al., 2017), among others. Accurate uncertainty quantification is the key to address the exploration-exploitation trade-off. Most existing bandit algorithms critically rely on certain analytical property of the imposed model (e.g., linear bandits) to quantify the uncertainty and derive the exploration strategy. Thompson Sampling (TS, Thompson, 1933) and Upper Confidence Bound (UCB, Auer et al., 2002) are two prominent examples, which are typically based on explicit-form posterior distributions or confidence sets, respectively.

However, in many real problems, the reward model is fairly complex: e.g., a general graphical model (Chapelle & Zhang, 2009) or a pipeline with multiple prediction modules and manual rules. In these cases, it is typically impossible to quantify the uncertainty in an analytical way, and frameworks such as TS or UCB are either methodologically not applicable or computationally infeasible.

*Equal contribution [1]Amazon [2]Department of Economics, University of California San Diego [3]Department of Statistics, North Carolina State University. Correspondence to: Runzhe Wan <runzhe.wan@gmail.com>, Rui Song <songray@gmail.com>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

Motivated by the real needs, we are concerned with the following question:

Can we design a practical bandit algorithm framework that is general, adaptive, and computationally tractable, with certain theoretical guarantee?

A straightforward idea is to apply the bootstrap method (Efron, 1992), a widely applicable data-driven approach for measuring uncertainty. However, as discussed in Section 2, most existing bootstrap-based bandit algorithms are either heuristic without a theoretical guarantee, computationally intensive, or only applicable in limited scenarios.

Contribution. Our contributions are three-fold. First, to address the aforementioned limitations, we propose a general-purpose bandit algorithm framework, *Multiplier Bootstrap-based Exploration* (`MBE`). `MBE` is based on *multiplier bootstrap* (Van Der Vaart & Wellner, 1996), an easy-to-adapt bootstrap framework that only requires randomly weighted data points. We further show that a naive application of multiplier bootstrap may result in linear regret, and we introduce a suitable way to add additional perturbations for sufficient exploration. The main advantage of `MBE` is that it is *general*: it is applicable to any reward model amenable to weighted loss minimization, without need of analytical-form uncertainty quantification or case-by-case algorithm design. As a data-driven exploration strategy, `MBE` is also *adaptive* to different environments.

Second, theoretically, we prove near-optimal regret bounds for `MBE` under sub-Gaussian multi-armed bandits (MAB), in both the instance-dependent and the instance-independent sense. Compared with all existing results for bootstrap-based bandit algorithms, our result is strictly more general (see Table 1), since existing results only apply to some special cases of sub-Gaussian distributions. To overcome the technical challenges, we proved a novel concentration inequality for some function of sub-exponential variables, and also developed the first *finite-sample* concentration and anti-concentration analysis for multiplier bootstrap, to the best of our knowledge. Given the broad applications of multiplier bootstrap in statistics and machine learning, our theoretical analysis is of independent interest.

This work does not relate to the positions of Runzhe Wan, Branislav Kveton and Rui Song at Amazon.

Table 1: Comparisons between several bootstrap- and perturbation-based bandit algorithms. All papers derive near-optimal regret bounds in MAB, with different reward distribution requirements. To compare the computational cost, we focus on MAB to illustrate, and consider Algorithm 2 for MBE. See Section 2 for more details of discussions in this table.

	Exploration Source	Methodology Generality	Theory Requirement	Computation Cost
MBE (this paper)	intrinsic & extrinsic	general	sub-Gaussian	$\mathcal{O}(KT)$
GIRO (Kveton et al., 2019b)	intrinsic & extrinsic	general	Bernoulli	$\mathcal{O}(T^2)$
ReBoot (Wang et al., 2020; Wu et al., 2022)	intrinsic & extrinsic	fixed & finite set of arms	Gaussian	$\mathcal{O}(KT)$
PHE (Kveton et al., 2019a; 2020a;b)	only extrinsic	general	bounded	$\mathcal{O}(KT)$

Third, with extensive simulation and real-data experiments, we demonstrate that MBE yields comparable performance with existing algorithms in different MAB settings and three real-world problems (online learning to rank, online combinatorial optimization, and dynamic slate optimization). This supports that MBE is easily generalizable, as it requires minimal modifications and derivations to match the performance of those near-optimal algorithms specifically designed for each problem. Moreover, we also show that MBE adapts to different environments and is relatively robust, due to its data-driven nature.

2. Related Work

The most popular bandit algorithms, arguably, include ϵ-greedy (Watkins, 1989), TS, and UCB. ϵ-greedy is simple and thus widely used. However, its exploration strategy is not aware of the uncertainty in data and thus is known to be statistically sub-optimal. TS and UCB rely on posteriors and confidence sets, respectively. Yet, their closed forms only exist in limited cases, such as MAB or linear bandits. For a few other models (such as generalized linear model or neural nets), we know how to construct the *approximate* posteriors or confidence sets (Filippi et al., 2010; Li et al., 2017; Phan et al., 2019; Kveton et al., 2020a;b) with theoretical guarantees, though the corresponding algorithms are usually costly or conservative. In more general cases, it is often not clear how to adapt UCB and TS in a valid and efficient way. Although approximate TS can be done via approximate posterior inference methods (e.g., particle filtering, Markov chain Monte Carlo, or variational inference) (Gopalan et al., 2014; Kawale et al., 2015; Wan et al., 2021; Urteaga & Wiggins, 2018; Yu et al., 2020), they do not come with guarantees. Moreover, the dependency on the probabilistic model assumptions (e.g., the reward distribution family or the noise level) also pose challenges to being robust.

To enable wider applications of bandit algorithms, several bootstrap-based (and related perturbation-based) methods have been proposed in the literature. Most algorithms are TS-type, by replacing the posterior with a bootstrap distribution. We next review the related papers, and summarize those with near-optimal asymptotic regret bounds in Table 1. We divide the sources of exploration into (i) leveraging the *intrinsic* randomness in the observed data (e.g., by randomizing the subset of history used for training) and (ii) manually adding *extrinsic* perturbations that are independent of the observed data (e.g., adding additive Gaussian noise to observed rewards).

Arguably, the non-parametric bootstrap is the most well-known bootstrap method, which works by re-sampling data with replacement. Vaswani et al. (2018) propose a version of non-parametric bootstrap with forced exploration to achieve a $\mathcal{O}(T^{2/3})$ regret bound in Bernoulli MAB. GIRO proposed in Kveton et al. (2019b) successfully achieves a rate-optimal regret bound in Bernoulli MAB, by adding Bernoulli perturbations to non-parametric bootstrap. However, due to the re-sampling nature of non-parametric bootstrap, it is challenging to implement it efficiently beyond Bernoulli MAB (see Section 4.3). Specifically, the computational cost of re-sampling scales quadratically in T. Riou & Honda (2020) apply Bayesian bootstrap (Rubin, 1981), which is a smooth version of non-parametric Bootstrap. An asymptotically optimal regret bound is proved for MAB with bounded rewards. However, only MAB is studied and similar computational challenge exists in general cases.

Another line of research is the residual bootstrap-based approach (ReBoot) (Hao et al., 2019; Wang et al., 2020; Tang et al., 2021; Wu et al., 2022). For each arm, ReBoot randomly perturbs the residuals of the corresponding observed rewards with respect to the estimated model to quantify the uncertainty for its mean reward. Although these methods also use random weights, they are applied to residuals, and thus are fundamentally different from our work. The limitation is that, by design, this approach is only applicable to problems with a *fixed* and *finite* set of arms, since the residuals are attached closely to each arm (see Appendix

A.4 for more details).

The perturbed history exploration (PHE) algorithm (Kveton et al., 2019a; 2020a;b) is also related. PHE works by adding additive noise to the observed rewards. Osband et al. (2019) apply similar ideas to reinforcement learning. However, PHE has two main limitations. First, for models where adding additive noise is not feasible (e.g., decision trees), PHE is not applicable. Second, as demonstrated in both Wang et al. (2020) and our experiments, the fact that PHE relies on only the extrinsically injected noise for exploration makes it less robust. For a complex structured problem, it may not be clear how to add the noise in a sound way (Wang et al., 2020). In contrast, it is typically more natural (and hence easier to be accepted) to leverage the intrinsic randomness in the observed data.

Finally, we note that multiplier bootstrap has been considered in the bandit literature, mostly as a computationally efficient approximation to non-parametric bootstrap studied in those papers. Eckles & Kaptein (2014) study the direct adaption of multiplier bootstrap (see Section 4.1) in simulation, and its empirical performance in contextual bandits is studied later (Tang et al., 2015; Elmachtoub et al., 2017; Riquelme et al., 2018; Bietti et al., 2021). However, no theoretical guarantee is provided in these works. In fact, as demonstrated in Section 4.1, such a naive adaptation may have a linear regret. Osband & Van Roy (2015) show that, in Bernoulli MAB, a variant of multiplier bootstrap is mathematically equivalent to TS. No further theoretical or numerical results are provided except for this special case. Our work is the first systematic study of multiplier bootstrap in bandits. Our unique contributions include: we identify the potential failure of naively applying multiplier bootstrap, highlight the importance of additional perturbations, design a general algorithm framework to make this heuristic idea concrete, provide the first theoretical guarantee in general MAB settings, and conduct extensive numerical experiments to study its generality and adaptivity.

3. Preliminary

Setup. We consider a general stochastic bandit problem. For any positive integer M, let $[M] = \{1, \dots, M\}$. At each round $t \in [T]$, the agent observes a context vector $\boldsymbol{x}_t$ (it is empty in non-contextual problems) and an action set $\mathcal{A}_t$, then chooses an action $A_t \in \mathcal{A}_t$, and finally receives the corresponding reward $R_t = f(\boldsymbol{x}_t, A_t) + \epsilon_t$, Here, f is an unknown function and ϵ_t is the noise term. Without loss of generality, we assume $f(\boldsymbol{x}_t, A_t) \in [0, 1]$. We note that the realized reward R_t does not need to be bounded. The goal is to minimize the cumulative regret

$$\text{Reg}_T = \sum_{t=1}^{T} \mathbb{E}\big[\max_{a \in \mathcal{A}_t} f(\boldsymbol{x}_t, a) - f(\boldsymbol{x}_t, A_t)\big].$$

At the end of round t, with an existing dataset $\mathcal{D}_t = \{(\boldsymbol{x}_l, A_l, R_l)\}_{l \in [t]}$, to decide the action A_{t+1}, most algorithms typically first estimate f in some function class $\mathcal{F}$ by solving a weighted loss minimization problem (also called weighted empirical risk minimization or cost-sensitive training)

$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{t} \sum_{l=1}^{t} \omega_l \mathcal{L}\big(f(\boldsymbol{x}_l, A_l), R_l\big) + J(f). \quad (1)$$

Here, $\mathcal{L}$ is a loss function (e.g., ℓ_2 loss or negative log-likelihood), ω_l is the weight of the lth data point, and J is an optional penalty function. We consider the weighted problem as it is general and related to our proposal below. One can just set $\omega_l \equiv 1$ to get the unweighted problem. As the simplest example, consider the K-armed bandit problem where $\boldsymbol{x}_l$ is empty and $\mathcal{A}_l = [K]$. Let $\mathcal{L}$ be the ℓ_2 loss, $J \equiv 0$, and $f(\boldsymbol{x}_l, A_l) \equiv r_{A_l}$ where r_k is the mean reward of the k-th arm. Then, (1) reduces to $\arg\min_{\{r_1,\dots,r_K\}} \sum_{l=1}^{t} \omega_l (R_l - r_{A_l})^2$, which gives the estimator $\widehat{r}_k = (\sum_{l:A_l=k} \omega_l)^{-1} \sum_{l:A_l=k} \omega_l R_l$, i.e., the arm-wise weighted average. Similarly, in linear bandits, (1) reduces to the weighted least-square problem (see Appendix A.2 for details).

Challenges. The estimation of f, together with the related uncertainty quantification, forms the foundation of most bandit algorithms. In the literature, $\mathcal{F}$ is typically a class of models that permit closed-form uncertainty quantification (e.g., linear models, Gaussian processes, etc.). However, in many real applications, the reward model can yield a fairly complicated structure, e.g., a hierarchical pipeline with both classification and regression modules. Manually specified rules are also common part of the model. It is challenging to quantify the uncertainty of these complicated models in analytical forms. Even when feasible, the dependency on the probabilistic model assumptions also pose challenges to being robust.

Therefore, in this paper, we focus on the bootstrap-based approach due to its generality and data-driven nature. Bootstrapping, as a general approach to quantify the model uncertainty, has many variants. The most popular one, arguably, is non-parametric bootstrap (used in GIRO), which constructs bootstrap samples by re-sampling the dataset with replacement. However, due to the re-sampling nature, it is computationally intense (see Section 4.3 for more discussions). In contrast, multiplier bootstrap (Van Der Vaart & Wellner, 1996), as an efficient and easy-to-implement alternative, is popular in statistics and machine learning.

Multiplier bootstrap. The main idea of multiplier bootstrap is to learn the model using randomly weighted data points. Specifically, given a multiplier weight distribution $\rho(\omega)$, for every bootstrap sample, we first randomly sample $\{\omega_l^{MB}\}_{l=1}^{t} \sim \rho(\omega)$ at round t, and then solve (1) with $\omega_l = \omega_l^{MB}$ to obtain $\widehat{f}^{MB}$. Repeat the procedure and the distribution of $\widehat{f}^{MB}$ forms the *bootstrap distribution* that quantifies our uncertainty over f. The popular choices of $\rho(\omega)$ include $\mathcal{N}(1, \sigma_\omega^2)$, Exp$(1)$, Poisson$(1)$, and the double-or-nothing distribution $2 \times$ Bernoulli(0.5).

4. Multiplier Bootstrap-based Exploration

4.1. Failure of the Naive Adaption of Multiplier Bootstrap

To design an exploration strategy based on multiplier bootstrap, a natural idea is to replace the posterior distribution in TS with the bootstrap distribution. Specifically, at every time point, we sample a function $\widehat{f}$ following the multiplier bootstrap procedure as described in Section 3, and then take the greedy action $\arg\max_{a \in \mathcal{A}_t} \widehat{f}(\boldsymbol{x}_t, a)$. However, perhaps surprisingly, such an adaptation may not be valid. The main reason is that the intrinsic randomness in a finite dataset is, in some cases, not enough to guarantee sufficient exploration. For example, the support of the bootstrap distribution cannot go outside the convex hull of the observed rewards. We illustrate this further with the following toy example.

Example 1. *Consider a two-armed Bernoulli bandit. Let the mean rewards of the two arms be p_1 and p_2, respectively. Without loss of generality, assume $1 > p_1 > p_2 > 0$. Let $\mathbb{P}(\omega = 0) = 0$. Then, with non-zero probability, an agent following the naive adaption of multiplier bootstrap (breaking ties randomly and initializing in an optimistic way; see Algorithm 5 in Appendix A.3 for details) pulls arm 1 only once. Therefore, the agent suffers a linear regret.*

Proof. *We first define two events*

$$\mathcal{E}_1 = \{A_t = 1, R_1 = 0\}, \mathcal{E}_2 = \{A_2 = 2, R_2 = 1\}.$$

By design, at time $t = 1$, the agent randomly choose an arm and hence will pull arm 1 with probability 0.5. Then the observed reward R_1 is 0 with probability $1 - p_1$. Therefore, $\mathbb{P}(\mathcal{E}_1) = 0.5(1 - p_1)$. Conditioned on $\mathcal{E}_1$, at $t = 2$, the agent will pull arm 2 (since multiplying $R_1 = 0$ with any weight always gives 0), then it will observe reward $R_2 = 1$ with probability p_2. Conditioned on $\mathcal{E}_1 \cap \mathcal{E}_2$, by induction, the agent will pull arm 2 for any $t > 2$. This is because the only reward record for arm 1 is $R_1 = 0$ and hence its weighted average is always 0, which is smaller than the weighted average for arm 2, which is at least positive. In conclusion, with probability at least $0.5 \times (1 - p_1) \times p_2 > 0$, the algorithm takes the optimal arm 1 only once.

4.2. Main Algorithm

The failure of the naive application of multiplier bootstrap implies that some additional randomness is needed to ensure sufficient exploration. In this paper, we consider achieving that by adding *pseudo-rewards*, an approach that proves its effectiveness in a few other setups (Kveton et al., 2019b; Wang et al., 2020). The intuition is as follows. The under-exploration issue happens when, by randomness, the observed rewards are in the low-value region (compared with the expected reward). Therefore, if we can blend in some data points with rewards that have a relatively wide coverage, then the agent would have a higher chance to explore.

These discussions motivate the design of our main algorithm, Multiplier Bootstrap-based Exploration (`MBE`), as in Algorithm 1. Specifically, at every round, in addition to the observed reward, we additionally add two pseudo-rewards with value 0 and 1. The pseudo-rewards are associated with the pulled arm and the context (if exists). Then, we solve a weighted loss minimization problem to update the model estimation (line 8). The weights are first sampled from a multiplier distribution (line 7), and then those of pseudo-rewards are additionally multiplied by a tuning parameter λ. In MAB, the estimates are arm-wise weighted average of all (observed or pseudo-) rewards

$$\overline{Y}_k = \frac{\sum_{\ell: A_\ell = k} (\omega_\ell R_{k,\ell} + \lambda \omega'_\ell \times 1 + \lambda \omega''_\ell \times 0)}{\sum_{\ell: A_\ell = k} (\omega_\ell + \lambda \omega'_\ell + \lambda \omega''_\ell)} \quad (2)$$

for $k \in [K]$, where multiplier weights $\{\omega_\ell, \omega'_\ell, \omega''_\ell\}_{l=1}^t \sim \rho(\omega)$. See Appendix A.1 for details.

We make three remarks on the algorithm design. First, we choose to add pseudo-rewards at the boundaries of the mean reward range (i.e., $[0, 1]$), since such a design naturally induces a high variance (and hence more exploration). Adding pseudo-rewards in other manners is also possible. Second, the tuning parameter λ controls the amount of extrinsic perturbation and determines the degree of exploration (together with the dispersion of $\rho(\omega)$). In Section 5, we give a theoretically valid range for λ. Finally and critically, besides guaranteeing sufficient exploration, we need to make sure the optimal arm can still be identified (asymptotically) after adding the pseudo-rewards. Intuitively, this is guaranteed, since we shift and scale the (asymptotic) mean reward from $f(\boldsymbol{x}, a)$ to $\big(f(\boldsymbol{x}, a) + \lambda\big)/(1 + 2\lambda) = f(\boldsymbol{x}, a)/(1 + 2\lambda) + \lambda/(1 + 2\lambda)$, which preserves the order between arms. A detailed analysis for MAB can be found in Appendix A.1.

We conclude this section by re-visiting Example 1 to provide some insights into how the pseudo-rewards help.

Example 1 (Continued)**.** *Even under the event $\mathcal{E}_1 \cap \mathcal{E}_2$, Algorithm 1 explores. To see this, consider an example with multiplier distribution is $2 \times$ Bernoulli(0.5). Then*

$$\begin{aligned}
\mathbb{P}(A_3 = 1) &\geq \mathbb{P}(\overline{Y}_1 > \overline{Y}_0) \\
&= \mathbb{P}\left(\frac{\lambda \omega'_1}{\omega_1 + \lambda \omega'_1 + \lambda \omega''_1} > \frac{\omega_2 + \lambda \omega'_2}{\omega_2 + \lambda \omega'_2 + \lambda \omega''_2}\right) \\
&\geq \mathbb{P}(\omega'_1 = 2, \omega_1 = \omega''_1 = \omega_2 = \omega'_2 = \omega''_2 = 0) \\
&= (1/2)^6.
\end{aligned}$$

Therefore, the agent can still choose the optimal arm.

Algorithm 1: General Template for MBE

Data: Function class $\mathcal{F}$, loss function $\mathcal{L}$, (optional) penalty function J, multiplier weight distribution $\rho(\omega)$, tuning parameter λ

2 Initialize $\widehat{f}$

3 **for** $t = 1, \dots, T$ **do**

4 Observe context $\boldsymbol{x}_t$ and action set $\mathcal{A}_t$

5 Take action $A_t = \arg\max_{a \in \mathcal{A}_t} \widehat{f}(\boldsymbol{x}_t, A)$ (break ties randomly)

6 Observe reward R_t

7 Sample the multiplier weights $\{\omega_l, \omega_l', \omega_l''\}_{l=1}^t \sim \rho(\omega)$

8 Solve the weighted loss minimization problem

$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \sum_{l=1}^{t} \Big[\omega_l \mathcal{L}\big(f(\boldsymbol{x}_l, A_l), R_l\big) + \lambda \omega_l' \mathcal{L}\big(f(\boldsymbol{x}_l, A_l), 0\big) + \lambda \omega_l'' \mathcal{L}\big(f(\boldsymbol{x}_l, A_l), 1\big)\Big] + J(f).$$

9

10 **end**

4.3. Computationally-Efficient Implementation

Efficient computation is critical for real applications of bandit algorithms. One potential limitation of Algorithm 1 is the computational burden: at every decision point, we need to re-sample the weights for all historical observations (line 8). This leads to a total computational cost of order $\mathcal{O}(T^2)$, similar to GIRO.

Fortunately, one prominent advantage of multiplier bootstrap over other bootstrap methods (such as non-parametric bootstrap or residual bootstrap) is that the (approximate) bootstrap distribution can be efficiently updated in an online manner, so that the per-round computation cost does not grow over time. Suppose we have a dataset $\mathcal{D}_t$ at time t, and denote $\mathcal{B}(\mathcal{D}_t)$ as the corresponding bootstrap distribution for f. With multiplier bootstrap, it is feasible to update $\mathcal{B}(\mathcal{D}_{t+1})$ approximately based on $\mathcal{B}(\mathcal{D}_t)$. We detail the procedure below and elaborate more in Algorithm 2.

Specifically, we maintain B different models $\{\widehat{f}_{b,t}\}_{b=1}^B$ and the corresponding observed history

$$\mathcal{H}_b = \{(\boldsymbol{x}_l, A_l, R_l, \omega_{l,b})\}_{l=1}^t$$

and pseudo-history

$$\mathcal{H}_b' = \{(\boldsymbol{x}_l, A_l, 0, \omega_{l,b}')\}_{l=1}^t \cup \{(\boldsymbol{x}_l, A_l, 1, \omega_{l,b}'')\}_{l=1}^t$$

for every $b \in [B]$. $\{\widehat{f}_{b,t}\}_{b=1}^B$ can be regarded as sampled from $\mathcal{B}(\mathcal{D}_t)$ and hence the empirical distribution over them is an approximation to the bootstrap distribution. At every time point t, for each replicate b, we only need to sample one weight for the new data point and then update $\widehat{f}_{b,t}$ as $\widehat{f}_{b,t+1}$. Then, $\{\widehat{f}_{b,t+1}\}_{b=1}^B$ are still B valid samples from $\mathcal{B}(\mathcal{D}_{t+1})$ and hence still a valid approximation. We note that, since we only have one new data point, the updating of f can typically be done efficiently (e.g., with closed-form updating or via online gradient descent). The per-round computational cost is hence independent of t.

Such an approximation is a common practice in the online bootstrap literature and can be regarded as an ensemble sampling-type algorithm (Lu & Van Roy, 2017; Qin et al., 2022). The hyper-parameter B is typically not treated as a tuning parameter but depends on the available computational resource (Hao et al., 2019). In our numerical experiments, this practical variant shows desired performance with $B = 50$. Moreover, the algorithm is embarrassingly parallel and also easy to implement: given an existing implementation for estimating f (i.e., solving (1)), the major requirement is to replicate it for B times and use random weights for each. This feature is attactive in real applications.

Algorithm 2: Practical Implementation of MBE

Data: Number of bootstrap replicates B, function class $\mathcal{F}$, loss function $\mathcal{L}$, (optional) penalty function J, weight distribution $\rho(\omega)$, tuning parameter λ

Let $\mathcal{H}_b = \{\}$ be the history and $\mathcal{H}_b' = \{\}$ be the pseudo-history, for any $b \in [B]$

Initialize $\widehat{f}_{b,0}$ for any $b \in [B]$

for $t = 1, \dots, T$ **do**

Observe context $\boldsymbol{x}_t$ and action set $\mathcal{A}_t$

Sample an index b_t uniformly from $\{1, \dots, B\}$

Offer $A_t = \arg\max_{A \in \mathcal{A}_t} \widehat{f}_{b_t, t-1}(\boldsymbol{x}_t, A)$ (break ties randomly)

Observe reward R_t

for *b = 1, …,* B **do**

Sample the weights $\omega_{t,b}, \omega_{t,b}', \omega_{t,b}'' \sim \rho(\omega)$.

Update $\mathcal{H}_b = \mathcal{H}_b \cup \big\{(\boldsymbol{x}_t, A_t, R_t, \omega_{t,b})\big\}$ and $\mathcal{H}_b' = \mathcal{H}_b' \cup \big\{(\boldsymbol{x}_t, A_t, 0, \omega_{t,b}'), (\boldsymbol{x}_t, A_t, 1, \omega_{t,b}'')\big\}$

Solve the weighted loss minimization problem

$$\widehat{f}_{b,t} = \arg\min_{f \in \mathcal{F}} \sum_{l=1}^{t} \Big[\omega_{l,b} \mathcal{L}\big(f(\boldsymbol{x}_l, A_l), R_l\big) + \lambda \omega_{l,b}' \mathcal{L}\big(f(\boldsymbol{x}_l, A_l), 0\big) + \lambda \omega_{l,b}'' \mathcal{L}\big(f(\boldsymbol{x}_l, A_l), 1\big)\Big] + J(f).$$

end

end

5. Regret Analysis

In this section, we provide the regret bound for Algorithm 1 under MAB with sub-Gaussian rewards. We regard this as the first step towards the theoretical understanding of MBE, and leave the analysis of more general settings for future work. We call a random variable X as σ-sub-Gaussian if $\mathbb{E}\exp\{t(X-\mathbb{E}X)\}\le\exp\{t^2\sigma^2/2\}$ for any $t\in\mathbb{R}$. The instantiation of Algorithm 1 under MAB is presented as Algorithm 3 in the appendix.

Theorem 5.1. *Consider a K-armed bandit, where the reward distribution of arm k is 1-sub-Gaussian with mean μ_k. Suppose arm 1 is the unique best arm that has the highest mean reward and $\Delta_k=\mu_1-\mu_k$. Take the multiplier weight distribution as $\mathcal{N}(1,\sigma_\omega^2)$. Let the tuning parameters satisfy*

$$\lambda\ge\big(1+\sigma_\omega^2/4+4/\sigma_\omega\big)+\sqrt{4\left(1+4/\sigma_\omega\right)/\sigma_\omega}.$$

Then, the problem-dependent regret is upper bounded by

$$\mathrm{Reg}_T\le\sum_{k=2}^{K}\Big\{7\Delta_k+\frac{55\big[C_1^*(\lambda,\sigma_\omega)+C_2^*(\lambda,\sigma_\omega)\big]}{\Delta_k}\log T\Big\},$$

and the problem-independent regret is bounded by

$$\mathrm{Reg}_T\le 7K\mu_1+C_1^*(\lambda,\sigma_\omega)K\log T+2\sqrt{C_2^*(\lambda,\sigma_\omega)KT\log T},$$

where

$$C_1^*(\lambda,\sigma_\omega)=8\sqrt{2}C_3^*(\lambda,\sigma_\omega)+38\sigma_\omega^2,$$
$$C_2^*(\lambda,\sigma_\omega)=6\lambda^2+\left[45(3+\sigma_\omega^2)\lambda^4C_3^*(\lambda,\sigma_\omega)+38\sigma_\omega^2\right],$$
$$C_3^*(\lambda,\sigma_\omega)=\frac{\log\left[(1+15\sigma_\omega^{-2}+3\sigma_\omega+10\sigma_\omega^2)\lambda^2\right]}{3\log 2}+1.$$

The two regret bounds are known as near-optimal (up to a logarithm term) in both the problem-dependent and problem-independent sense (Lattimore & Szepesvári, 2020). Notably, recall that the Gaussian distribution and all bounded distributions belong to the sub-Gaussian class. Therefore, as reviewed in Table 1, our theory is strictly more general than all existing results for bootstrap-based MAB algorithms.

Technical challenges. It is particularly challenging to analyze MBE for two reasons. First, the probabilistic analysis of multiplier bootstrap itself is technically challenging, since the same random weights appear in both the denominator and the numerator (recall that MBE uses the weighted averages (2) to select actions in MAB). It is notoriously complicated to analyze the ratio of random variables, especially when they are correlated. Besides, existing bootstrap-based papers rely on the properties of specific *parametric* reward classes (e.g., Bernoulli in Kveton et al. (2019b) and Gaussian in Wang et al. (2020)), while we lose these nice structures when considering sub-Gaussian rewards.

To overcome these challenges, we denote the first s rewards from pulling arm k as $\mathcal{H}_{k,s}$ with the i-th observation denoted as $R_{k,i}$, and start with carefully defining two good events $G_{k,s}$ and $A_{k,s}$. Here, $G_{k,s}$ denotes the event that the weighted average $\overline{Y}_{k,s}=\sum_{i=1}^s\left[\omega_iR_{k,i}+\omega_i'(1\times\lambda)+\omega_i''(0\times\lambda)\right]/\sum_{i=1}^s(\omega_i+\lambda\omega_i'+\lambda\omega_i'')$ is close to the unweighted average (with pseudo-rewards) $\overline{R}_{k,s}^*=\sum_{i=1}^s(R_{k,i}+1\times\lambda+0\times\lambda)/\sum_{i=1}^s(1+\lambda+\lambda)$, and $A_{k,s}$ represents the event that $\overline{R}_{k,s}^*$ is close to its population mean $(\mu_k+\lambda)/(1+2\lambda)$. It is worthy to note that $\{\omega_i,\omega_i',\omega_i''\}_{i=1}^s$ are resampled from $\rho(\omega)$ at each round. To bound the probability of and the regret conditioned on the bad event, we face two major technical challenges. First, when transforming the ratio into an analyzable form, a summation of correlated sub-Gaussian and sub-exponential variables appears and is hard to analyze. We carefully design and analyze a novel event to remove the correlation and the sub-Gaussian terms (see proof of Lemma D.3). Second, the proof needs a new concentration inequality for functions of sub-exponential variables that does not exist in the literature. We obtain such a new concentration inequality (Lemma E.4) via careful analysis of sub-exponential distributions.

To the best of our knowledge, our proof provides the first *finite-sample* concentration and anti-concentration analysis for multiplier bootstrap, which has broad applications in statistics and machine learning.

Extension. Theorem 5.1 is proved with Gaussian weights to simplify analysis. Indeed, to analyze multiplier Bootstrap, we need to analyze $\sum_{\ell:A_\ell=k}\omega_\ell R_{k,\ell}$ conditioned on the reward history, which is the sum of scaled i.i.d. variables, and Gaussian distribution has nice analytical properties for us to derive the bound. We hypothesize our result can be extended to other weight distributions that satisfy similar anti-concentration and concentration properties, such as $\underline{C}\exp\{-t^2/(2\underline{\sigma}^2)\}\le\mathbb{P}(|\omega_i-1|>t)\le\overline{C}\exp\{-t^2/(2\overline{\sigma}^2)\}$ with positive $\underline{C},\overline{C},\underline{\sigma},\overline{\sigma}$. However, we expect some analytical challenges.

Tuning parameters. In Theorem 5.1, MBE has two tuning parameters, λ and σ_ω. Intuitively, λ controls the amount of external perturbation and σ_ω controls the magnitude of exploration from bootstrap. In general, higher values of these two parameters facilitate exploration but also lead to a slower convergence. The condition on λ in Theorem 5.1 requires that (i) λ is not too small and (ii) the joint effect of λ and σ_ω is not too small. Both are intuitive. In practice, this could be loose: e.g., it requires $\lambda\ge5.25+2\sqrt{5}$ when $\sigma_\omega=1$. As we observe in Section 6, MBE with a smaller λ (e.g., 0.5) still empirically performs well.

6. Experiments

In this section, we empirically evaluate MBE with both simulation (Section 6.1) and real datasets (Section 6.2).

6.1. MAB Simulation

We first experiment with simulated MAB instances. The goal is to (i) further validate our theoretical findings, (ii) check whether MBE can yield comparable performance with standard methods, and (iii) study the robustness and adaptivity of MBE. We also experimented with linear bandits and the main findings are similar. To save space, we defer these results to Appendix B.1.

We compare MBE with TS (Thompson, 1933), PHE (Kveton et al., 2019a), ReBoot (Wang et al., 2020), and GIRO (Kveton et al., 2019b). The last three algorithms are the existing bootstrap- or perturbation-type algorithms reviewed in Section 2. Specifically, PHE explores by perturbing observed rewards with additive noise, without leveraging the intrinsic uncertainly in the data, ReBoot explores by perturbing the residuals of the rewards observed for each arm, and GIRO re-samples observed data points with replacement. In all experiments below, the weights of MBE are sampled from $\mathcal{N}(1, \sigma_\omega^2)$ [1]. We fix $\lambda = 0.5$ and run MBE with three different values of σ_ω^2: $0.5, 1$ and 1.5. We also compare with the naive adaption of multiplier bootstrap (i.e., no pseudo-rewards; denoted as Naive MB). We run Algorithm 2 with $B = 50$ replicates.

We first study 10-armed bandits, where the mean reward of each arm is independently sampled from Beta$(1, 8)$. We consider three reward distributions, including Bernoulli, Gaussian, and exponential. For Gaussian MAB, the reward noise is sampled from $\mathcal{N}(0, 1)$. The other two distributions are determined by their means. For TS, we always use the correct reward distribution class and its conjugate prior. The prior mean and variance are calibrated using the true model. Therefore, TS is a strong baseline. For GIRO and ReBoot, we use the default implementations as they work well. For PHE, the original paper adds Bernoulli perturbation since it only studies bounded reward distributions. We extend PHE by sampling additive noise from the same distribution family as the true rewards, as done in Wu et al. (2022). GIRO, ReBoot, and PHE all have one tuning parameter to control the degree of exploration. We tune it over $\{2^{k-4}\}_{k=0}^{6}$ and report the best performance for each method. Without tuning, these algorithms generally do not perform well as originally proposed, due to differences in the settings. We tuned Naive MB as well.

Results. Results over 100 runs are reported in Figure 1. Our findings can be summarized as follows. First, without knowledge of the problem settings (e.g., the reward distribution family and its parameters, and the prior distribution) and without heavy tuning, MBE performs favorably and close to TS. Second, pseudo-rewards are indeed important in exploration, otherwise the algorithm suffers a linear regret. Third, MBE has a stable performance with different σ_ω while the other methods are tuned for their best performance. This is thanks to the data-driven nature of MBE. Finally, the other three general-purpose exploration strategies perform reasonably after tuning, as expected. However, GIRO is computationally intense. For example, in Gaussian bandits, the time cost of GIRO is 2 minutes while all the other algorithms can complete within 10 seconds. The computational burden is due to the limitation of non-parametric bootstrap (see Section 4.3). ReBoot also performs reasonably, yet by design it is not easy to extend to more complex problems (e.g., problems in Section 6.2).

[1] We also experimented with other weight distributions with similar main conclusions. Using Gaussian weights allows us to study impact of different multiplier magnitudes more clearly.

Adaptivity. PHE relies on sampling additive noise from an appropriate distribution, and TS can be viewed similarly. In the results above, we provide auxiliary information about the environment to them and need to modify their implementation in different setups. In contrast, MBE automatically adapts to these problems. As argued in Section 2, one main advantage of MBE over them is its adaptiveness. To see this, we consider the following procedure: we run the Gaussian versions of TS and PHE in Bernoulli MAB, and run their Bernoulli versions in Gaussian MAB. We also run MBE with $\sigma_\omega^2 = 0.5$. MBE does not require any modifications across the two problems. The results presented in Figure 2 clearly demonstrate that MBE adapts to reward distributions.

Similarly, in Figure 3, we also studied the adaptivity of these methods against the reward distribution scale (the standard deviation of the Gaussian noise, σ) and the task distribution (we sample the mean rewards from Beta$(\alpha, 8)$ and vary the parameter α). For all settings, we use the algorithms tuned for Figure 1. MBE shows impressive adaptivity, while PHE and TS may not perform well when the environment is not close to the one they are tuned for. Recall that, in real applications, heavy tuning is not possible without the ground truth. This demonstrates the adaptivity of MBE, as a data-driven exploration strategy.

Additional results. In Appendix B.2, we also try different values of λ and B for MBE. We also repeat the main experiment with $K = 25$. Our main observations still hold, and MBE is relatively robust to its tuning parameters.

6.2. Real-Data Applications

The main benefit of MBE is that it easily generalizes to complex models. In this section, we use real datasets to demonstrate this property. Specifically, we test if MBE can achieve comparable performance with strong problem-specific base-

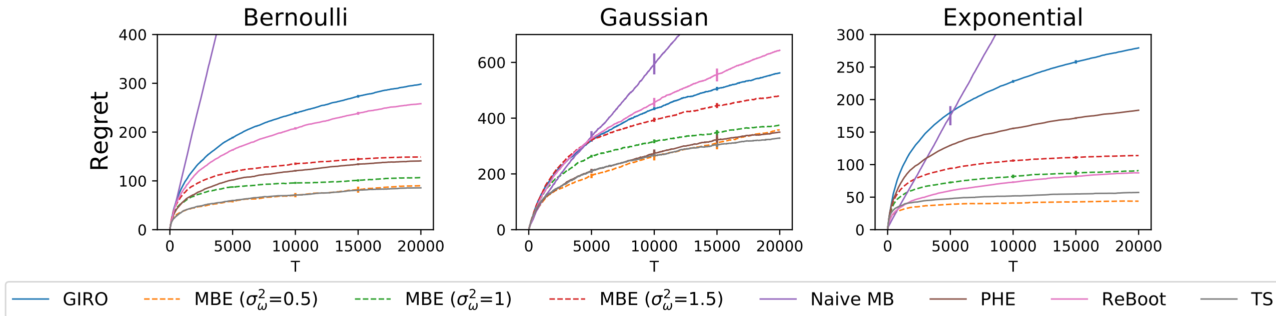


Figure 1: Simulation results under MAB. The error bars indicate the standard errors, which may not be visible when the width is small.

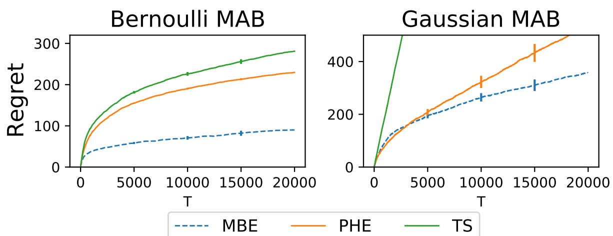
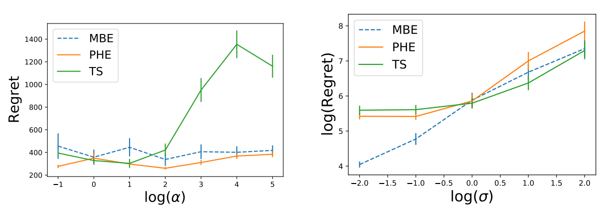


Figure 2: Robustness results, to the reward distribution class.



(a) Trend with α **(b)** Trend with σ

Figure 3: Results with different reward variances and task distributions. For the x-axis in both figures and the y-axis in the second one, we plot at the logarithmic scale for better visualization.

lines proposed in the literature, without problem-specific algorithm design and heavy tuning.

Domain-specific models. We study the three problems considered in Wan et al. (2022), including cascading bandits for online learning to rank (Kveton et al., 2015), combinatorial semi-bandits for online combinatorial optimization (Chen et al., 2013), and multinomial logit (MNL) bandits for dynamic slate optimization (Agrawal et al., 2017; 2019). All these are practical and important problems in real life. Yet, these domain models all have unique structures and require a case-by-case algorithm design. For example, the rewards in MNL bandits follow multinomial distributions that have complex dependency with the pulled arms. To derive the posterior or confidence bound, one has to use a delicately designed epoch-type procedure (Agrawal et al., 2019).

Datasets. We use the three datasets studied in Wan et al. (2022). Specifically, we use the Yelp rating dataset (Zong et al., 2016) to recommend and rank K restaurants, use the Adult dataset (Dua & Graff, 2017) to send advertisements to $K/2$ men and $K/2$ women (a combinatorial semi-bandit problem with continuous rewards), and use the MovieLens dataset (Harper & Konstan, 2015) to display K movies. In our experiments, we fix $K = 4$ and randomly sample 30 items from the dataset to choose from. We provide a summary of these datasets and problems in Appendix B.3, and refer interested readers to Wan et al. (2022) and references therein for more details.

Baselines. We compared `MBE` with state-of-the-art baselines in the literature, including `TS-Cascade` (Zhong et al., 2021) and `CascadeKL-UCB` (Kveton et al., 2015) for cascading bandits, `CUCB` (Chen et al., 2016) and `CTS` (Wang & Chen, 2018) for semi-bandits, and `MNL-TS` (Agrawal et al., 2017) and `MNL-UCB` (Agrawal et al., 2019) for MNL bandits. To save space, we denote the TS-type algorithms by `TS` and the UCB-type ones by `UCB`. We also study `PHE` and ϵ-greedy (`EG`) as two other general-purpose exploration strategies.

Tuning. For the baseline methods, as in Section 6.1, we either use the default hyperparameters in Wan et al. (2022) or tune them extensively via grid search and present their best performance. For `EG`, we choose the exploration rate as $\epsilon_t = \min(1, a/2\sqrt{t})$ with tuning parameter a, following Kveton et al. (2020a). For `MBE`, with every bootstrap sample, we estimate the reward model via maximum weighted likelihood estimation, which yields closed-form solution that allows online updating in all three problems. The other implementation details are similar to Section 6.1.

Results. We present the results in Figure 4. The overall findings are consistent with Section 6.1. First, without any additional derivations or algorithm design, `MBE` matches the performance of problem-specific algorithms. Second, pseudo-rewards are important to guarantee sufficient exploration, and naively applying multiplier bootstrap may fail. Third, `MBE` has relatively stable performance with varying σ_ω, since its exploration is mostly data-driven. In contrast, the hyper-parameters of `PHE` and `EG` have to be carefully tuned, since they rely on externally added perturbation or

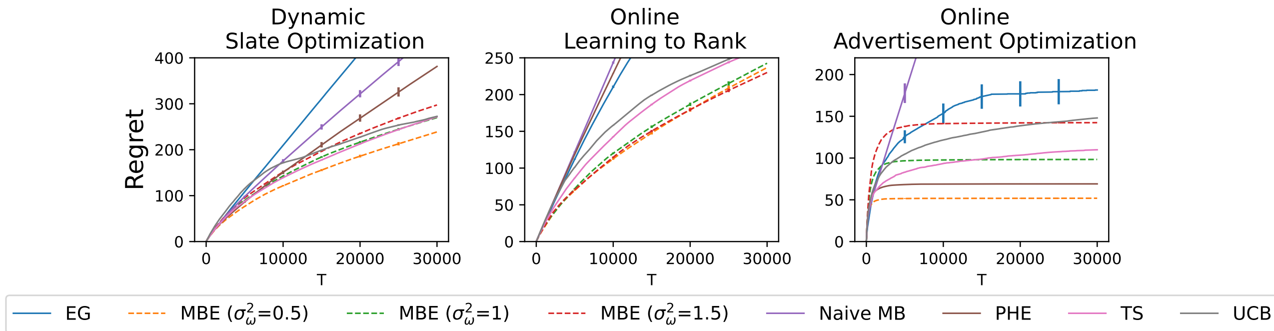


Figure 4: Real-data results for three structured bandit problems that need domain-specific models.

forced exploration. For example, the best parameters for EG in the three problems are $a = 5$, 0.1 and 0.5. Finally, PHE does not perform well in MNL and cascading bandits, where the outcomes are binary. We investigated this trend and found that the response rates (i.e., the probabilities for the binary outcome to be 1) in the two datasets are low. In this case, PHE introduces too much noise to explore, which slows down the estimation convergence.

7. Conclusion

In this paper, we propose a new bandit exploration strategy, Multiplier Bootstrap-based Exploration (MBE). The main advantage of MBE is its generality: for any reward model that can be estimated via weighted loss minimization, the idea of MBE is applicable, and requires minimal efforts on derivation or implementation of the exploration mechanism. As a data-driven method, MBE also shows nice adaptivity. We prove near-optimal regret bounds for MBE in the sub-Gaussian MAB setup, which is more general than in other bootstrap-based bandit papers. Numerical experiments demonstrate that MBE is general, efficient, and adaptive.

There are a few meaningful future extensions. First, the regret analysis for MBE (and more generally, other bootstrap-based bandit methods) in more complicated setups would be valuable. The main challenge comes from analyzing the finite-sample property of multiplier bootstrap.

Second, adding pseudo-rewards at every round is needed for the analysis. We hypothesize that there exists a more adaptive and efficient way of introducing extrinsic perturbation, such that we have sufficient exploration while avoiding over-exploration.

Third, the practical implementation of MBE relies on an ensemble of models to approximate the bootstrap distribution and the online regression oracle to update the model estimation. Both parts lead to approximation and also correlation over time. Our numerical experiments show that such an approach works well empirically, but it would be still meaningful to have more theoretical understanding.

Lastly, in this paper, we present MBE assuming the knowledge of a generative model of the rewards (i.e., assuming the existence of a regression oracle). The idea can be naturally generalized to the policy-based setting, where we assume the existence of a classification oracle that can compute the optimal policy within a pre-specified policy class (see, e.g., Agarwal et al. (2014)). We leave this to future study.

References

Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L., and Schapire, R. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pp. 1638–1646. PMLR, 2014.

Agrawal, S., Avadhanula, V., Goyal, V., and Zeevi, A. Thompson sampling for the mnl-bandit. *arXiv preprint arXiv:1706.00977*, 2017.

Agrawal, S., Avadhanula, V., Goyal, V., and Zeevi, A. Mnl-bandit: A dynamic learning approach to assortment selection. *Operations Research*, 67(5):1453–1485, 2019.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

Bietti, A., Agarwal, A., and Langford, J. A contextual bandit bake-off. *J. Mach. Learn. Res.*, 22:133–1, 2021.

Chapelle, O. and Zhang, Y. A dynamic bayesian network click model for web search ranking. In *Proceedings of the 18th international conference on World wide web*, pp. 1–10, 2009.

Chen, W., Wang, Y., and Yuan, Y. Combinatorial multi-armed bandit: General framework and applications. In *International Conference on Machine Learning*, pp. 151–159. PMLR, 2013.

Chen, W., Wang, Y., Yuan, Y., and Wang, Q. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *The Journal of Machine Learning Research*, 17(1):1746–1778, 2016.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Durand, A., Achilleos, C., Iacovides, D., Strati, K., Mitsis, G. D., and Pineau, J. Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis. In *Machine learning for healthcare conference*, pp. 67–82. PMLR, 2018.

Eckles, D. and Kaptein, M. Thompson sampling with the online bootstrap. *arXiv preprint arXiv:1410.4009*, 2014.

Efron, B. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pp. 569–593. Springer, 1992.

Elmachtoub, A. N., McNellis, R., Oh, S., and Petrik, M. A practical method for solving contextual bandit problems using decision trees. *arXiv preprint arXiv:1706.04687*, 2017.

Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. Parametric bandits: The generalized linear case. *Advances in Neural Information Processing Systems*, 23, 2010.

Gopalan, A., Mannor, S., and Mansour, Y. Thompson sampling for complex online problems. In *International conference on machine learning*, pp. 100–108. PMLR, 2014.

Hao, B., Abbasi Yadkori, Y., Wen, Z., and Cheng, G. Bootstrapping upper confidence bound. *Advances in Neural Information Processing Systems*, 32, 2019.

Harper, F. M. and Konstan, J. A. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

Kawale, J., Bui, H. H., Kveton, B., Tran-Thanh, L., and Chawla, S. Efficient thompson sampling for online matrix-factorization recommendation. *Advances in neural information processing systems*, 28, 2015.

Kveton, B., Szepesvari, C., Wen, Z., and Ashkan, A. Cascading bandits: Learning to rank in the cascade model. In *International Conference on Machine Learning*, pp. 767–776. PMLR, 2015.

Kveton, B., Szepesvári, C., Ghavamzadeh, M., and Boutilier, C. Perturbed-history exploration in stochastic multi-armed bandits. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 2786–2793, 2019a.

Kveton, B., Szepesvari, C., Vaswani, S., Wen, Z., Lattimore, T., and Ghavamzadeh, M. Garbage in, reward out: Bootstrapping exploration in multi-armed bandits. In *International Conference on Machine Learning*, pp. 3601–3610. PMLR, 2019b.

Kveton, B., Szepesvári, C., Ghavamzadeh, M., and Boutilier, C. Perturbed-history exploration in stochastic linear bandits. In *Uncertainty in Artificial Intelligence*, pp. 530–540. PMLR, 2020a.

Kveton, B., Zaheer, M., Szepesvari, C., Li, L., Ghavamzadeh, M., and Boutilier, C. Randomized exploration in generalized linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 2066–2076. PMLR, 2020b.

Lai, C. D. and Balakrishnan, N. *Continuous bivariate distributions*. Springer, 2009.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Li, L., Lu, Y., and Zhou, D. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pp. 2071–2080. PMLR, 2017.

Lu, X. and Van Roy, B. Ensemble sampling. *Advances in neural information processing systems*, 30, 2017.

Osband, I. and Van Roy, B. Bootstrapped thompson sampling and deep exploration. *arXiv preprint arXiv:1507.00300*, 2015.

Osband, I., Van Roy, B., Russo, D. J., Wen, Z., et al. Deep exploration via randomized value functions. *J. Mach. Learn. Res.*, 20(124):1–62, 2019.

Phan, M., Abbasi Yadkori, Y., and Domke, J. Thompson sampling and approximate inference. *Advances in Neural Information Processing Systems*, 32, 2019.

Qin, C., Wen, Z., Lu, X., and Roy, B. V. An analysis of ensemble sampling. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=c6ibx0yl-aG`.

Riou, C. and Honda, J. Bandit algorithms based on thompson sampling for bounded reward distributions. In *Algorithmic Learning Theory*, pp. 777–826. PMLR, 2020.

Riquelme, C., Tucker, G., and Snoek, J. Deep bayesian bandits showdown: An empirical comparison of bayesian deep networks for thompson sampling. *arXiv preprint arXiv:1802.09127*, 2018.

Rubin, D. B. The bayesian bootstrap. *The annals of statistics*, pp. 130–134, 1981.

Shen, W., Wang, J., Jiang, Y.-G., and Zha, H. Portfolio choices with orthogonal bandit learning. In *Twenty-fourth international joint conference on artificial intelligence*, 2015.

Tang, L., Jiang, Y., Li, L., Zeng, C., and Li, T. Personalized recommendation via parameter-free contextual bandits. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pp. 323–332, 2015.

Tang, Q., Xie, H., Xia, Y., Lee, J., and Zhu, Q. Robust contextual bandits via bootstrapping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 12182–12189, 2021.

Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Urteaga, I. and Wiggins, C. Variational inference for the multi-armed contextual bandit. In *International Conference on Artificial Intelligence and Statistics*, pp. 698–706. PMLR, 2018.

Van Der Vaart, A. W. and Wellner, J. A. Weak convergence. In *Weak convergence and empirical processes*, pp. 16–28. Springer, 1996.

Vaswani, S., Kveton, B., Wen, Z., Rao, A., Schmidt, M., and Abbasi-Yadkori, Y. New insights into bootstrapping for bandits. *arXiv preprint arXiv:1805.09793*, 2018.

Wan, R., Ge, L., and Song, R. Metadata-based multi-task bandits with bayesian hierarchical models. *Advances in Neural Information Processing Systems*, 34:29655–29668, 2021.

Wan, R., Ge, L., and Song, R. Towards scalable and robust structured bandits: A meta-learning framework. *arXiv preprint arXiv:2202.13227*, 2022.

Wang, C.-H., Yu, Y., Hao, B., and Cheng, G. Residual bootstrap exploration for bandit algorithms. *arXiv preprint arXiv:2002.08436*, 2020.

Wang, S. and Chen, W. Thompson sampling for combinatorial semi-bandits. In *International Conference on Machine Learning*, pp. 5114–5122. PMLR, 2018.

Watkins, C. J. C. H. Learning from delayed rewards. 1989.

Wen, Z., Kveton, B., and Ashkan, A. Efficient learning in large-scale combinatorial semi-bandits. In *International Conference on Machine Learning*, pp. 1113–1122. PMLR, 2015.

Wu, S., Wang, C.-H., Li, Y., and Cheng, G. Residual bootstrap exploration for stochastic linear bandit. *arXiv preprint arXiv:2202.11474*, 2022.

Yu, T., Kveton, B., Wen, Z., Zhang, R., and Mengshoel, O. J. Graphical models meet bandits: A variational thompson sampling approach. In *International Conference on Machine Learning*, pp. 10902–10912. PMLR, 2020.

Zhang, H. and Chen, S. Concentration inequalities for statistical inference. *Communications in Mathematical Research*, 37(1):1–85, 2021.

Zhang, H. and Wei, H. Sharper sub-weibull concentrations. *Mathematics*, 10(13):2252, 2022.

Zhong, Z., Chueng, W. C., and Tan, V. Y. Thompson sampling algorithms for cascading bandits. *Journal of Machine Learning Research*, 22(218):1–66, 2021.

Zhou, Q., Zhang, X., Xu, J., and Liang, B. Large-scale bandit approaches for recommender systems. In *International Conference on Neural Information Processing*, pp. 811–821. Springer, 2017.

Zong, S., Ni, H., Sung, K., Ke, N. R., Wen, Z., and Kveton, B. Cascading bandits for large-scale recommendation problems. *arXiv preprint arXiv:1603.05359*, 2016.

A. Additional Method Details

A.1. MBE for MAB

In this section, we present the concrete form of MBE when being applied to MAB. Recall that $\boldsymbol{x}_t$ is null, $A_t \in [K]$, and r_k is the mean reward of the k-th arm. We define $f(\boldsymbol{x}_t, A_t; \boldsymbol{r}) = r_{A_t}$, where the parameter vector $\boldsymbol{r} = (r_1, \ldots, r_K)^\top$. We define the loss function as

$$\frac{1}{t'} \sum_{t=1}^{t'} \omega_t (r_{A_t} - R_t)^2.$$

The solution is then $(\widehat{r}_1, \ldots, \widehat{r}_K)^\top$ with $\widehat{r}_k = (\sum_{t:A_t=k} \omega_t)^{-1} \sum_{t:A_t=k} \omega_t R_t$, i.e., the arm-wise weighted average. After adding the pseudo rewards, we can give algorithm for MAB in Algorithm 3.

Next, we provide intuitive explanation on why Algorithm 3 works. Indeed, denote $s := |\mathcal{H}_{k,T}|$, where $\mathcal{H}_{k,T}$ is the set of observed rewards for the k-th arm up to round T. Let $R_{k,l}$ be the l-th element in $\mathcal{H}_{k,T}$. Then

$$\begin{aligned}
\overline{Y}_{k,s} &= \frac{\sum_{i=1}^s \omega_i R_{k,i} + \lambda \sum_{i=1}^s \omega_i'}{\sum_{i=1}^s \omega_i + \lambda \sum_{i=1}^s \omega_i' + \lambda \sum_{i=1}^s \omega_i''} \\
&= \frac{s^{-1} \sum_{i=1}^s \omega_i (R_{k,i} - \mu_k) + s^{-1} \sum_{i=1}^s (\omega_i - 1) + \lambda s^{-1} \sum_{i=1}^s (\omega_i' - 1) + \mu_k + \lambda}{s^{-1} \sum_{i=1}^s (\omega_i - 1) + \lambda s^{-1} \sum_{i=1}^s (\omega_i' - 1) + \lambda s^{-1} \sum_{i=1}^s (\omega_i'' - 1) + 1 + 2\lambda} \xrightarrow{\mathbb{P}} \frac{\mu_k + \lambda}{1 + 2\lambda}
\end{aligned}$$

by using the law of large numbers. Then, by Slutsky's theorem,

$$\sqrt{s}\left[\overline{Y}_{k,s} - \frac{\mu_k + \lambda}{1 + 2\lambda}\right] = \frac{1}{1+2\lambda}\left[\frac{1}{\sqrt{s}} \sum_{i=1}^s \omega_i (R_{k,i} - \mu_k) + \frac{1}{\sqrt{s}} \sum_{i=1}^s (\omega_i - 1) + \frac{\lambda}{\sqrt{s}} \sum_{i=1}^s (\omega_i' - 1)\right] + o_p(1)$$

will weakly converge to a mean-zero Gaussian distribution $\mathcal{N}\left(0, \frac{\sigma_k^2 + 2}{(1+2\lambda)^2} \sigma_\omega^2\right)$. Therefore, our algorithm preserves the order of the arms for any $\lambda > 0$.

Algorithm 3: MBE for MAB with sub-Gaussian rewards with mean bounded in $[0, 1]$

Data: number of arms K, multiplier weight distribution $\rho(\omega)$, tuning parameter λ

Set $\mathcal{H}_k = \{\}$ be the history of the arm k and $\overline{Y}_k = +\infty, \forall k \in [K]$

for $t = 1, \ldots, T$ **do**

  Pull $A_t = \arg\max_{k \in [K]} \overline{Y}_k$ (break tie randomly),

  Observe reward R_t

  Set $\mathcal{H}_k = \mathcal{H}_k \cup \{R_t\}$

  for $k = 1, \ldots, K$ **do**

    if $|\mathcal{H}_k| > 0$ **then**

      Sample the multiplier weights $\{\omega_l, \omega_l', \omega_l''\}_{l=1}^{|\mathcal{H}_k|} \sim \rho(\omega)$.

      Update the mean reward

$$\overline{Y}_k = \left(\sum_{\ell=1}^{|\mathcal{H}_k|} (\omega_\ell \cdot R_{k,\ell} + \omega_\ell' \cdot 1 \times \lambda + \omega_\ell'' \cdot 0 \times \lambda)\right) / \left(\sum_{\ell=1}^{|\mathcal{H}_k|} (\omega_\ell + \lambda \omega_\ell' + \lambda \omega_\ell'')\right),$$

      where $R_{k,l}$ is the l-th element in $\mathcal{H}_k$.

    end

  end

end

A.2. MBE for stochastic linear bandits

In this section, we derive the form of MBE when applied to stochastic linear bandits. We focus on the setup where $\boldsymbol{x}_t$ is empty and $A_t \in \mathbb{R}^p$ is a linear feature vector, and other setups of linear bandits can be formulated similarly. In this case,

Algorithm 4: MBE for linear bandits.

Data: number of arms K, multiplier weight distribution $\rho(\omega)$, tuning parameter λ

Set $\mathcal{H}_k = \{\}$ be the history of the arm k, set $A_0 = \mathbf{0}$, $\widehat{\boldsymbol{\theta}}_0 = \mathbf{0}$ with $b_0 = \mathbf{0}$, and $V_0 = (1+\xi)I_p$.

if $t = 1, \ldots, p$ **then**

  Offer $A_t = t$.

end

for $t = p+1, \ldots, T$ **do**

  Offer $A_t = \arg\max_{\mathbf{a} \in \mathcal{A}_t} \mathbf{a}^\top \boldsymbol{\theta}_t$ (break tie randomly)

  Observe reward R_t

  Set $\mathcal{H}_k = \mathcal{H}_k \cup \{R_t\}$

  for $k = 1, \ldots, K$ **do**

    if $|\mathcal{H}_k| > 0$ **then**

      Sample the multiplier weights $\{\omega_l, \omega_l', \omega_l''\}_{l=1}^{|\mathcal{H}_k|} \sim \rho(\omega)$.

      Update the following quantities:

      - $V_{t+1} = V_t + \omega_t A_t A_t^\top + \lambda \omega_t' 0 I_d + \lambda \omega_t'' I_d$;
      - $b_{t+1} = b_t + A_t\big(\omega_t R_t + \lambda \omega_t' 0 + \lambda \omega_t'' 1\big)$;
      - Refresh the parameter as $\widehat{\boldsymbol{\theta}}_{t+1} = V_{t+1}^{-1} b_{t+1}$.

    end

  end

end

$f(\boldsymbol{x}_t, A_t; \boldsymbol{\theta}) = A_t^\top \boldsymbol{\theta}$ where the parameter vector is $\boldsymbol{\theta} \in \mathbb{R}^p$. Then, the weighted loss function is

$$\sum_{t=1}^{T} \omega_t \big(A_t^\top \boldsymbol{\theta} - R_t\big)^2 + \frac{\xi}{2} \|\boldsymbol{\theta}\|_2^2,$$

where $\xi \geq 0$ is a penalty tuning parameter. The solution is the standard weighted ridge regression estimator and can be updated in the following way:

0. Initialization: $A_0 = \mathbf{0}$, $\widehat{\boldsymbol{\theta}}_0 = \mathbf{0}$ with $b_0 = \mathbf{0}$, and $V_0 = (\xi+1) I_{\dim(A_t)}$.
1. $\widehat{\boldsymbol{\theta}}_t = V_t^{-1} A_t$;
2. $V_{t+1} = V_t + \omega_t A_t A_t^\top$, $b_{t+1} = b_t + \omega_t R_t A_t$, and hence update

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_{t+1} = V_{t+1}^{-1} b_{t+1} &= \big(V_t + \omega_t A_t A_t^\top\big)^{-1} \big(b_t + \omega_t R_t A_t\big) \\ &= V_t^{-1} - V_t^{-1} A_t (\omega_t^{-1} + A_t^\top V_t^{-1} A_t)^{-1} A_t^\top V_t^{-1} \end{aligned}$$

3. Take the action $A_{t+1} = \arg\max_{\mathbf{a} \in \mathcal{A}_t} \mathbf{a}^\top \boldsymbol{\theta}_{t+1}$.

The MBE algorithm for linear bandits is presented in Algorithm 4.

A.3. Naive Adaptation of the Multiplier Bootstrap

We present the naive multiplier bootstrap-based exploration algorithm in Algorithm 5. Specifically, there is no pseudo-rewards added.

A.4. ReBoot

For completeness, we introduce the details of ReBoot in this section and discuss its generalizability. More details can be found in the original papers (Wang et al., 2020; Wu et al., 2022).

Algorithm 5: A Naive Design of `MBE`

Data: Function class $\mathcal{F}$, loss function $\mathcal{L}$, (optional) penalty function J, multiplier weight distribution $\rho(\omega)$, tuning parameter λ

Set $\mathcal{H} = \{\}$ be the history be the pseudo-history

Initialize $\widehat{f}$ in an optimistic way

for $t = 1, \ldots, T$ **do**

  Observe context $\boldsymbol{x}_t$ and action set $\mathcal{A}_t$

  Offer $A_t = \arg\max_{a \in \mathcal{A}_t} \widehat{f}(\boldsymbol{x}_t, a)$ (break tie randomly)

  Observe reward R_t

  Update $\mathcal{H} = \mathcal{H} \cup \{(\boldsymbol{x}_t, A_t, R_t)\}$

  Sample the multiplier weights $\{\omega_l\}_{l=1}^t \sim \rho(\omega)$

  Solve the weighted loss minimization problem to update $\widehat{f}$ as

$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{t} \sum_{l=1}^{t} \omega_l \mathcal{L}\big(f(\boldsymbol{x}_l, A_l), R_l\big) + J(f).$$

end

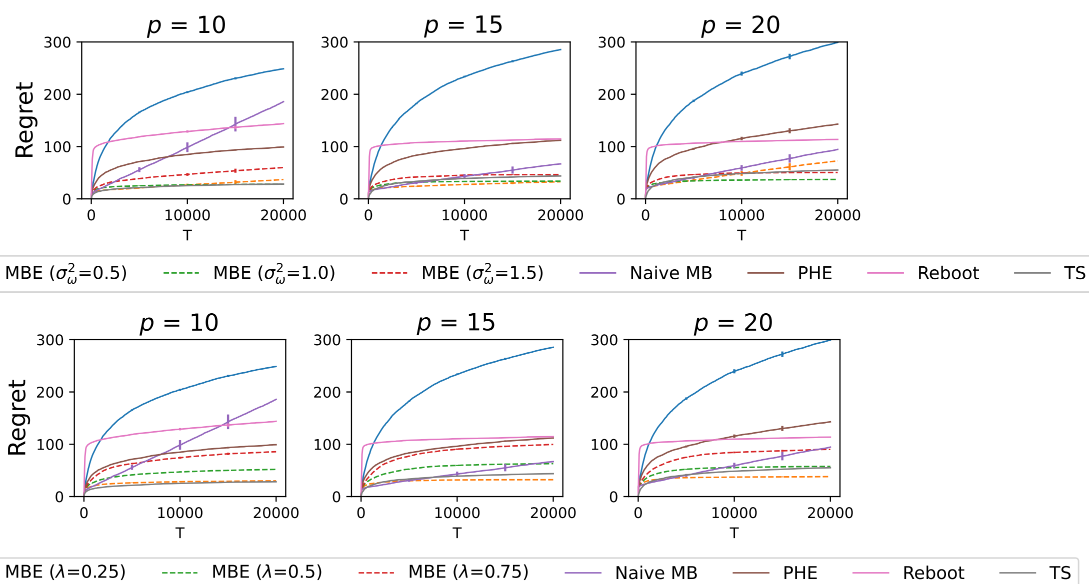


Figure 5: Performance of `MBE` in three linear bandit problems.

Consider a stochastic bandit problem with a fixed and finite set of arm $\mathcal{A}$. Every arm $a \in \mathcal{A}$ may have a fixed feature vector (which with slight overload of notation, we also denote as a). The mean reward of arm a is $f(a)$. At each round t', `ReBoot` first fit the model f as $\widehat{f}$ using all data. Then for each arm a, `ReBoot` first computes the corresponding residuals using rewards related to that arm as $\{\epsilon_t = R_t - \widehat{f}(a)\}_{t:A_t=a, t\le t'}$, then perturbs these residuals with random weights as $\{\omega_t\epsilon_t = R_t - \widehat{f}(a)\}_{t:A_t=a,t\le t'}$ (`ReBoot` also adds pseudo-residuals, which we omit for ease of notations), and finally use $\widehat{f}(a) + |\{t : A_t = a, t \le t'\}|^{-1} \sum_{t:A_t=a,t\le t'} \omega_t\epsilon_t$ as the perturbed estimation of the mean reward of arm a. By design, it can be seen that `ReBoot` critically relies on the reward history of each *fixed* arm. Therefore, to the best of our understanding, it is not easy to extend `ReBoot` to problems with either changing (e.g., contextual problems) or infinite arms.

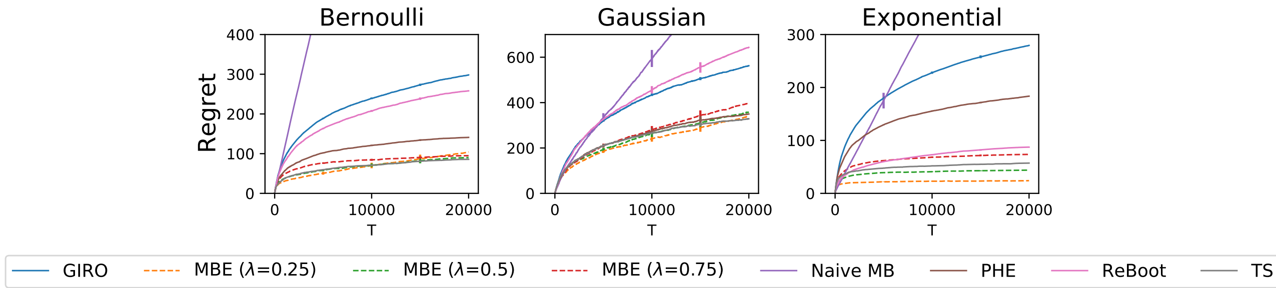


Figure 6: Performance of MBE with different values of λ in MAB.

B. More Experiment Results and Details

B.1. Results for linear bandits

We also consider the linear bandit problem. The linear bandit version of MBE is presented in Appendix A.2. We experiment with several dimensions $p = 10, 15, 20$. The number of arms is $K = 100$. The feature vector $x_k \in \mathbb{R}^p$ of arm k is generated as follows. For the last 10 arms, the features are drawn uniformly at random from $(0, 1)$. For the first 90 arms, we consider a practical setup where they are low-rank: we first generate a loading metric $A = (a_{ij}) \in \mathbb{R}^{p\times 5}$ from Uniform$(0, 1)$, then sample $b \in \mathbb{R}^5$ from Uniform$(0, 1)$, and finally constructs $x_k = Ab$. The parameter vector $\theta \in \mathbb{R}^p$ is uniformly sampled from $[0, 1]^p$. We normalize the feature vectors such that the mean reward $\mu_k = x_k^\top \theta$ falls within the interval $[0, 1]$. The rewards of arm k are drawn i.i.d. from Bernoulli$(x_k^\top \theta)$.

We still compare MBE with the method for linear bandit version of GIRO, PHE, and ReBoot with tuning to their best performance over the hyper-parameter set $\{2^{k-4}\}_{k=0}^6$ and report the best performance of each method. For TS, we use Gaussian for both its reward and prior distribution, and calibrate their parameters using the true model. The total rounds are $T = 20000$ and our results are averaged over 50 randomly chosen problems. Most other details are similar to our MAB experiments.

We present the results in Figure 5, where we vary either σ_ω or λ in the two subplots. We can see that naive MB leads to a linear regret. Hence, the pseudo-reward also matters in this problem. MBE achieves comparable performance with strong baselines such as TS. Another finding is that MBE is robust to its tuning parameters. Finally, Reboot needs to pull K times to initialize (the linear regret part in the first K rounds) due to the nature of its design. In contrast, most other linear bandit algorithms typically only need p rounds of forced exploration. This shows the limitation of the ReBoot framework (See Appendix A.4).

B.2. Additional results

In this section, we study the performance of MBE with respect to a few other hyper-parameters.

We first study the robustness to another tuning parameter λ. Recall that λ controls the amount of external perturbation. Specifically, we repeat the experiment in 6.1 with σ_ω^2 fixed as 0.5 and with different values of λ $(0.25, 0.5, 0.75)$. From Figure 6, it can be seen that a small amount of pseudo-rewards ($\lambda = 0.25$) seems sufficient in these settings, and the results are fairly stable. We believe this is because the exploration of MBE is main driven by the internal randomness in the data.

In Figure 7, we repeat our main experiments with K changed to 25. We can see that our main conclusions still hold.

Finally, in Figure 8, we implement MBE with different number of replicates B. As expected, more replicates does help exploration due to a better approximation to the whole bootstrapping distribution. Yet, we find that $B = 50$ suffices to generate comparable performance with TS and the performance of MBE becomes relatively stable for larger values of B.

B.3. Details of the real-data experiments

Our real-data experiments closely follow Wan et al. (2022). For completeness, we provide information of the three problems here, and refer interested readers to Wan et al. (2022) and references therein.

In an online learning to rank problem, we aim to select and rank K items from a pool of L ones. We iteratively interacts with users to learn about their preferences. The cascading model is popular in learning to rank (Kveton et al., 2015), which

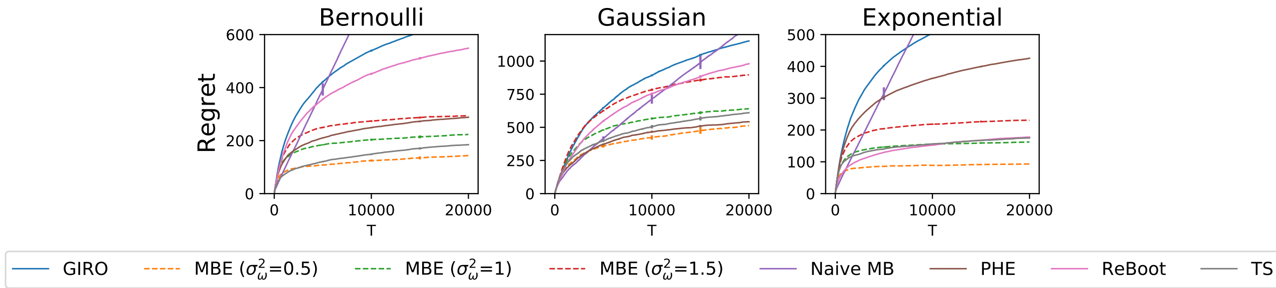


Figure 7: Performance of MBE with $K = 25$.

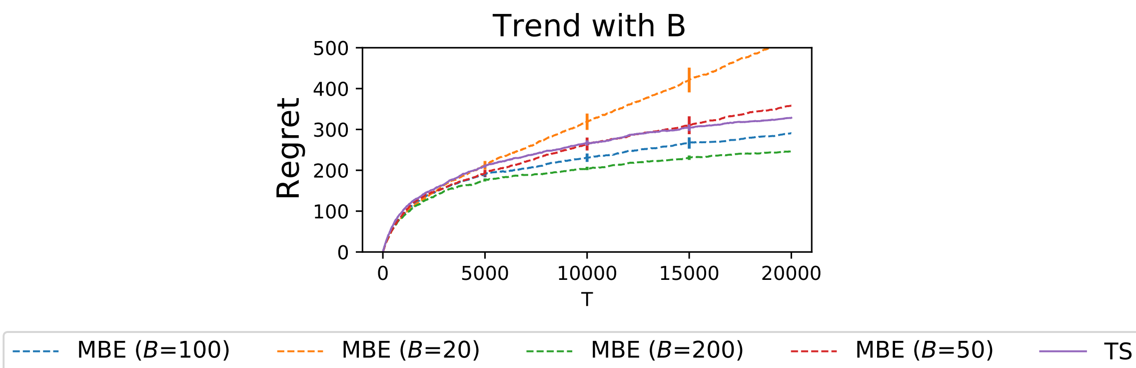


Figure 8: Performance of MBE with different number of replicates B.

models the user behaviour as glancing from top to bottom (like a cascade) and choose to click an item following a Bernoulli distribution when she looks at that item. Therefore, we will binary outcomes for all items that the user has examined, and there are complex dependency between them.

In a slate optimization (or called assortment optimization) problem, we aim to offer K items from a pool of L ones, especially when there exist substitution effects. The Multinomial Logit (MNL) model characterizes the choice behaviour as a multinomial distribution based on the attractiveness of each item. Since the offer subset changes over rounds, the joint likelihood is actually complex. To get the posterior or confidence bounds, one has to resort to an epoch-type offering schedule (Agrawal et al., 2017).

Online combinatorial optimization also has numerous applications (Wen et al., 2015), including maximum weighted matching, ads allocation, webpage optimization, etc. It is common that every chosen item will generate a separate observation, known as the semi-bandit problem. We consider a special problem in our experiments, where we need to choose K persons from a pool under constraints.

The three datasets we used (and related problem setups) are studied in corresponding TS papers in the literature. To general random rewards, we need to either generate from a real-data-calibrated model or by directly sampling from the dataset. We follow Wan et al. (2022) and references therein. For cascading or MNL bandits, we split the dataset into a training and a testing set, use the training to estimate the reward model, and compare on the testing set. For semi-bandits, we sample rewards from the dataset.

C. Main Proof

This section gives the proof of the main regret bound (Theorem 5.1). Section D gives the major lemmas required to bound the regret components used in this section. Section E lists all supporting technical lemmas, including the lower bound of the Gaussian tail and some novel results on the concentration property of sub-Gaussian and sub-exponential distributions.

Before beginning our proof, we first provide the definition of sub-Gaussian and sub-exponential variables: A mean-zero random variable X is called sub-Gaussian with variance proxy σ^2 if $\mathbb{E}\exp\{tX\} \leq \exp\{t^2\sigma^2/2\}$ for any $t \in \mathbb{R}$, and we denote it as $X \sim \mathrm{subG}(\sigma^2)$. A mean-zero variable X is called sub-exponential with parameters λ and α if

$$\mathbb{E}\exp\{tX\} \leq \exp\left\{\frac{t^2\lambda^2}{2}\right\}, \qquad |t| \leq \frac{1}{\alpha}.$$

We denote as $X \sim \mathrm{subE}(\lambda, \alpha)$ if sub-exponential X has parameters λ and α. For simplicity, we denote $\mathrm{subE}(\lambda) := \mathrm{subE}(\lambda, \lambda)$. Sub-Gaussian and sub-exponential variables play important roles in bandit problems and exhibit various concentration properties. For more details, please refer to Zhang & Chen (2021) and Zhang & Wei (2022).

Notations: *Let $\mathbb{P}_\xi(A) = \int_A \mathrm{d}F_\xi(x)$ denote the probability of event A, where $F_\xi(x)$ is the distribution function of the random variable ξ. Similarly, let $\mathbb{E}_\xi f(\xi) = \int f(x)\,\mathrm{d}F_\xi(x)$ represent the expectation. We write two functions $a(s,T) \lesssim b(s,T)$ if $a(s,T) \leq c \times b(s,T)$ for some constant c independent of s and T. We write $a(s,T) \asymp b(s,T)$ if both $a(s,T) \lesssim b(s,T)$ and $a(s,T) \gtrsim b(s,T)$. Furthermore, we define $a \vee b = \max\{a,b\}$ and $a \wedge b = \min\{a,b\}$ for any real numbers a and b. Similarly, we define $a \vee b \vee c = \max\{a,b,c\}$ and $a \wedge b \wedge c = \min\{a,b,c\}$ for any $a,b,c \in \mathbb{R}$.*

We will present a comprehensive version of the `MBE` theory under MAB, as stated in Theorem C.1, along with its proof. In this version, we allow for arbitrary variance proxies σ_k^2 instead of constraining them to be equal to one in Theorem 5.1.

Theorem C.1. *Consider a K-armed bandit, where the reward distribution of arm k is $\mathrm{subG}(\sigma_k^2)$ with mean μ_k. Suppose $\mu_1 = \max_{k\in[K]} \mu_k$ and $\Delta_k = \mu_1 - \mu_k$. Take the multiplier weight distribution as $\mathcal{N}(1, \sigma_\omega^2)$ in Algorithm 3. Let the tuning parameters satisfy $\lambda \geq \left(1 + \frac{\sigma_\omega^2}{4} + \frac{4\sigma_1}{\sigma_\omega}\right) + \sqrt{\frac{4\sigma_1}{\sigma_\omega}\left(\frac{4\sigma_1}{\sigma_\omega}+1\right)}$, Then the problem-dependent regret is upper bounded by*

$$\mathrm{Reg}_T \leq \sum_{k=2}^{K} \Delta_k \left[7 + \left\{C_1(\sigma_1, \sigma_k, \lambda, \sigma_\omega) + \frac{C_2(\sigma_1, \sigma_k, \lambda, \sigma_\omega)}{\Delta_k^2}\right\}\log T\right],$$

where

$$C_1(\sigma_1, \sigma_k, \lambda, \sigma_\omega) = 55\left[8\sqrt{2}\max_{k\in[K]} \sigma_k^2 \left(\frac{\log D_2(\sigma_1, \sigma_k, \lambda, \sigma_\omega)}{3\log 2} + 1\right) + 38\sigma_\omega^2\right],$$

and

$$C_2(\sigma_1, \sigma_k, \lambda, \sigma_\omega) = 310\lambda^2\sigma_k^2 + 55\left[D_1(\sigma_1, \sigma_k, \lambda, \sigma_\omega)\left(\frac{\log D_2(\sigma_1, \sigma_k, \lambda, \sigma_\omega)}{3\log 2} + 1\right) + 38\sigma_\omega^2\right],$$

with

$$D_1(\sigma_1, \sigma_k, \lambda, \sigma_\omega) = \left[\left(1 + 8\sqrt{2}\max_{k\in[K]} \sigma_k^2\right)\left(16 + \frac{\sigma_\omega^2}{\sigma_1^2}\right) + 16\sigma_1^4 + 3\frac{\sigma_\omega^2}{\sigma_1^2} + 3\sigma_\omega^2 + 1\right]\sigma_\omega^2\lambda^4,$$

and

$$D_2(\sigma_1, \sigma_k, \lambda, \sigma_\omega) = 3\left[1 + \frac{3\sqrt{\pi}\sigma_\omega^2}{2\sigma_1}\left(\frac{\sigma_1}{\sigma_\omega} + 3\lambda\right) + 16\sqrt{\pi}\max_{k\in[K]} \sigma_k^2\left(\frac{\sigma_\omega^2}{16\sigma_1^4} + \frac{1}{\sigma_\omega^2}\right) + \frac{\lambda^2\sigma_\omega^2}{4\sigma_1^4}\right].$$

Furthermore, the problem-independent regret is upper bounded by

$$\mathrm{Reg}_T \leq 7K\mu_1 + \max_{k\in[K]\setminus\{1\}} C_1(\sigma_1, \sigma_k, \lambda)K\log T + 2\sqrt{\max_{k\in[K]\setminus\{1\}} C_2(\sigma_1, \sigma_k, \lambda)KT\log T}.$$

Proof. We denote the first s rewards from pulling arm k as $\mathcal{H}_{k,s}$, with the i-th observations denoted as $R_{k,i}$. Let $Q_{k,s}(\tau) = \mathbb{P}\big(\overline{Y}_{k,s} > \tau \mid \mathcal{H}_{k,s}\big)$ be the tail probability that $\overline{Y}_{k,s}$, conditioned on history $\mathcal{H}_{k,s}$ is at least τ. Further, let $N_{k,s}(\tau) = 1/Q_{k,s}(\tau) - 1$ represent the expected number of rounds in which the arm k is underestimated, given s sample rewards. Here

$$\overline{Y}_{k,s} = \frac{\sum_{i=1}^{s}\left[\omega_i R_{k,i} + \omega_i' \cdot (1\times\lambda) + \omega_i'' \cdot (0\times\lambda)\right]}{\sum_{i=1}^{s}(\omega_i + \lambda\omega_i' + \lambda\omega_i'')}$$

is the objective function defined in Algorithm 3.

Step 0: Decomposition of the regret bound. Our proof relies on the following decomposition of the cumulative regret.

Lemma C.2 (Theorem 1, Kveton et al. (2019b)). *Suppose in MAB we select arms according to the rule* $A_t = \arg\max_{k\in[K]} \overline{Y}_{k,t}$ *with* $\overline{Y}_{k,t}$ *defined in Algorithm 3. Then for any* $\{\tau_k\}_{k=2}^K \subseteq \mathbb{R}$*, the expected* T*-round regret can be bounded above as*

$$\mathrm{Reg}_T \leq \sum_{k=2}^{K} \Delta_k \big(a_k + b_k\big),$$

where $a_k = \sum_{s=0}^{T-1} a_{k,s}$ *and* $b_k = \sum_{s=0}^{T-1} b_{k,s} + 1$*, and* $a_{k,s} = \mathbb{E}\big[N_{1,s}(\tau_k) \wedge T\big]$ *and* $b_{k,s} = \mathbb{P}\big(Q_{k,s}(\tau_k) > T^{-1}\big)$*.*

Recall the summation index s is the number of times we pull the k-th arm. In the proof, we will fix $\tau_k \in \Big(\frac{\mu_k+\lambda}{1+2\lambda}, \frac{\mu_1+\lambda}{1+2\lambda}\Big)$. The definitions of a_k and b_k have important meanings: a_k represents the expected number of rounds that optimal arm 1 has been being underestimated, whereas b_k is the probability that the suboptimal arm k is being overestimated. Here we only need to consider the lower bound of the tail of the distribution of the rewards from the optimal arm. The intuition behind this is twofold: (i) we only need the rewards from the optimal arm taking a relatively large value with a probability that is not too small; (ii) we do not care about the negligible probability of receiving a large reward from suboptimal arms.

Therefore, our target is then to bound a_k and b_k for any $k \geq 2$. These are completed in Step 1 and Step 2 below, respectively.

Step 1: Bounding a_k.

We first provide a roadmap for proving a_k is bounded by a term of $O(\log T)$ order: For a given constant level τ_k, the probability of the optimal arm 1 being underestimated given s rewards is $1 - Q_{1,s}(\tau_k)$. If we pick the level to satisfy $\tau_k < \frac{\mu_1+\lambda}{1+2\lambda}$, the theory of large deviation gives

$$\lim_{s\to\infty} Q_{1,s}(\tau_k) = 1.$$

Hence the expected number of rounds to observe a not-underestimated case $N_{1,s}(\tau_k) = \frac{1}{Q_{1,s}(\tau_k)} - 1$ has the property $\lim_{s\to\infty} N_s(\tau_k) = 0$ as the number of pulls s grows to infinity. Thus, given the round T, there exists a constant $s_{a,k}(T)$ such that $N_s(\tau_k) \leq T^{-1}$ for all s over $s_{a,k}(T)$. Consequently, the quantity a_k in regret bound will be bounded by

$$a_k \leq \sum_{s=0}^{s_{a,k}(T)} \mathbb{E}\left[N_{1,s}(\tau_k) \wedge T\right] + 1.$$

The fact that the constant $s_{a,k}(T)$ is at the order of $\log T$ will be shown in Lemma D.2, Lemma D.3, and Lemma D.4. For small number of pulls $s \leq s_{a,k}(T)$, we show in Lemma D.1 that $\mathbb{E}\left[N_{1,s}(\tau_k) \wedge T\right] \leq 4 + 16e^{9/8}$ for any $s \in \mathbb{N}$. Thus, it is enough to conclude that a_k can be bounded by a term of $O(\log T)$ order.

To formally bound a_k in the non-asymptotic sense following the intuition above, we need to decompose a_k. For the decomposition, a common approach is to use indicators on *good* events. Denote the shifted (sample-) mean reward as

$$\overline{R}^*_{k,s} = \frac{\sum_{i=1}^s R_{k,i} + \lambda s}{s(1+2\lambda)} = \frac{\lambda}{1+2\lambda} + \frac{1}{1+2\lambda}\overline{R}_{k,s}.$$

where $\overline{R}_{k,s} = \frac{1}{s}\sum_{i=1}^s R_{k,i}$ is the mean reward of the k-th arm. Then we can define the following *good* events for l-th arm as

$$A_{l,s} = \left\{-C_1\Delta_k < \overline{R}^*_{l,s} - \frac{\mu_l+\lambda}{1+2\lambda} < C_1\Delta_k\right\},$$

and

$$G_{l,s} = \left\{-C_2\Delta_k \leq \overline{Y}_{l,s} - \overline{R}^*_{l,s} \leq C_2\Delta_k\right\}.$$

The definitions of $A_{l,s}$ and $G_{l,s}$ are intuitive: $A_{l,s}$ represents the sample mean does not deviate excessively from the true mean, and events on $G_{l,s}$ means $\overline{Y}_{l,s}$ is not too far away from the scaled-shifted sample mean $\overline{R}^*_{l,s}$. Here C_1, C_2 are constants belonging to the interval $(0,1)$.

Therefore, by using these *good* events, we decompose $a_{k,s}$ into the following three parts:

$$a_{k,s,1} = \mathbb{E}\Big[\big(N_{1,s}(\tau_k) \wedge T\big)\mathbb{I}(A^c_{1,s})\Big], \tag{3}$$

$$a_{k,s,2} = \mathbb{E}\Big[\big(N_{1,s}(\tau_k) \wedge T\big)\mathbb{I}(A_{1,s})\mathbb{I}(G_{1,s}^c)\Big], \tag{4}$$

and

$$a_{k,s,3} = \mathbb{E}\Big[\big(N_{1,s}(\tau_k) \wedge T\big)\mathbb{I}(A_{1,s})\mathbb{I}(G_{1,s})\Big]. \tag{5}$$

Let $C_1 = \frac{1}{6\lambda}$ and $C_2 = \frac{1}{12\lambda}$ with fixed $\lambda > 1$, then $C_1, C_2 \in (0,1)$. Consider the case when $T \geq 2$. Define

$$s_{a,k,j}(T) := \max\{s : a_{k,s,j} \geq T^{-1}\}, \qquad k = 2, \ldots, K, \qquad j = 1, 2, 3.$$

Lemma D.2 in Appendix D demonstrates that by taking

$$s \geq s_{a,k,1}(T) := \frac{144\lambda^2\sigma_1^2}{(1+2\lambda)^2\Delta_k^2}\log T,$$

we will have $a_{k,s,1} \leq T^{-1}$. Lemma D.3 and Lemma D.4 in Appendix D say that: if we choose

$$s \geq s_{a,k,2}(T) := \left\{\left[\left(\Omega_1 + \mathfrak{a}_1 + \mathfrak{b}_1 + \Omega_1\mathfrak{a}_1\right)\left(\frac{1}{3}\log^{-1}2 \times \log\left\{3\left[1 + \frac{\sqrt{\pi}\mathfrak{b}_1}{2} + \frac{\sqrt{2\pi}\Omega_1\mathfrak{a}_1}{2\mathfrak{b}_1^2} + \frac{2\mathfrak{a}_1}{\mathfrak{b}_1}\right]\right\} + 1\right)\right] \vee \frac{18\sigma_\omega^2(2\mu_1 - 1)^2}{(\lambda^2 + \lambda + 1/4)^2} \vee \frac{2\sigma_\omega^2(1+2\lambda^2)}{(1+2\lambda)^2}\right\} \times 3\log T$$

and

$$s \geq s_{a,k,3}(T) := \left\{\left[\left(\Omega_1 + \mathfrak{a}_2 + \mathfrak{b}_2 + \Omega_1\mathfrak{a}_2\right)\left(\frac{1}{3}\log^{-1}2 \times \log\left\{3\left[1 + \frac{\sqrt{\pi}\mathfrak{b}_1}{2} + \frac{\sqrt{2\pi}\Omega_1\mathfrak{a}_2}{2\mathfrak{b}_2^2} + \frac{2\mathfrak{a}_2}{\mathfrak{b}_2}\right]\right\} + 1\right)\right] \vee \frac{18\sigma_\omega^2(2\mu_k - 1)^2}{(\lambda^2 + \lambda + 1/4)^2} \vee \frac{2\sigma_\omega^2(1+2\lambda^2)}{(1+2\lambda)^2}\right\} \times 3\log T,$$

then we have $a_{k,s,2} \leq T^{-1}$ and $a_{k,s,3} \leq T^{-1}$, respectively, where

$$\mathfrak{a}_1 = \frac{192\sigma_\omega^2\lambda^4}{3(1+2\lambda)^2\Delta_k^2}, \qquad \mathfrak{b}_1 = \frac{2\sigma_\omega^2\left[96\lambda^4\sigma_1^2 + (1+2\lambda)^2C_2^2\Delta_k^2\right]}{3(1+2\lambda)^2\Delta_k^2},$$

$$\mathfrak{a}_2 = \frac{36\sigma_\omega^2\lambda^4}{3(\lambda-1)^2\Delta_k^2}, \qquad \mathfrak{b}_2 = \frac{\sigma_\omega^2\left[72\lambda^4\sigma_1^2 + 25(1+2\lambda)^2\Delta_k^2\right]}{6(\lambda-1)^2\Delta_k^2}, \qquad \text{and} \qquad \Omega_k = 8\sqrt{2}\sigma_k^2, \qquad k \in [K].$$

Let $\Omega_{\max} = \max_{k\in[K]}\Omega_k$. Then, for any

$$s \geq s_{a,k}(T) = \frac{144\lambda^2\sigma_1^2}{(1+2\lambda)^2\Delta_k^2}\log T + \left\{\left[\left(\Omega_{\max} + (\mathfrak{a}_1 + \mathfrak{a}_2) + (\mathfrak{b}_1 + \mathfrak{b}_2) + \Omega_{\max}(\mathfrak{a}_1 + \mathfrak{a}_2)\right)\left(\frac{1}{3}\log^{-1}2 \times \log\left\{3\left[1 + \frac{\sqrt{\pi}}{2}\left(\sqrt{\mathfrak{b}_1} + \sqrt{\mathfrak{b}_2}\right) + \frac{\sqrt{2\pi}\Omega_{\max}}{2}\left(\frac{\mathfrak{a}_1}{\mathfrak{b}_1^2} + \frac{\mathfrak{a}_2}{\mathfrak{b}_2^2}\right) + 2\left(\frac{\mathfrak{a}_1}{\mathfrak{b}_1} + \frac{\mathfrak{a}_2}{\mathfrak{b}_2}\right)\right]\right\} + 1\right)\right] \vee \frac{18\sigma_\omega^2(2\mu_1 - 1)^2 \vee (2\mu_k - 1)^2}{(\lambda^2 + \lambda + 1/4)^2} \vee \frac{2\sigma_\omega^2(1+2\lambda^2)}{(1+2\lambda)^2}\right\} \times 3\log T,$$

it holds that $s \geq \max_{j=1,2,3} s_{a,k,j}(T)$ because

$$s_{a,k}(T) = s_{a,k,1}(T) + \max_{j=2,3} s_{a,k,j}(T) \geq \max_{j=1,2,3} s_{a,k,j}(T).$$

Hence, for any $s \geq s_{a,k}(T)$, we have

$$a_{k,s} = a_{k,s,1} + a_{k,s,2} + a_{k,s,3} \leq 3T^{-1}.$$

Finally, Lemma D.1 in Appendix D guarantees that if we choose $\lambda \geq \left(1+\frac{\sigma_\omega^2}{4}+\frac{4\sigma_1}{\sigma_\omega}\right)+\sqrt{\frac{4\sigma_1}{\sigma_\omega}\left(\frac{4\sigma_1}{\sigma_\omega}+1\right)}$, the component $a_{k,s} \leq 4+16e^{9/8}$ for any $s \geq 0$. Thus, by setting $\lambda \geq \left(1+\frac{\sigma_\omega^2}{4}+\frac{4\sigma_1}{\sigma_\omega}\right)+\sqrt{\frac{4\sigma_1}{\sigma_\omega}\left(\frac{4\sigma_1}{\sigma_\omega}+1\right)}$, we have

$$\begin{aligned}
a_k &= \sum_{s=0}^{T-1} a_{k,s} \\
&\leq \sum_{s<s_{a,k}(T)} \max_{s\in\{0,1,\ldots,T-1\}} a_{k,s} + \sum_{s_{a,k}(T)\leq s<T} 3T^{-1} \\
&= \max_{s\in\{0,1,\ldots,T-1\}} a_{k,s} \times s_{a,k}(T) + 3T^{-1}\times\big(T-s_{a,k}(T)\big) \\
&\leq 4(1+4e^{9/8})\Bigg\{\frac{144\lambda^2\sigma_1^2}{(1+2\lambda)^2\Delta_k^2}\log T + \Bigg[\Bigg(\big[\Omega_{\max}+(\mathfrak{a}_1+\mathfrak{a}_2)+(\mathfrak{b}_1+\mathfrak{b}_2)+\Omega_{\max}(\mathfrak{a}_1+\mathfrak{a}_2)\big] \\
&\qquad \left(\frac{\log\left\{3\left[1+\frac{\sqrt{\pi}}{2}\left(\sqrt{\mathfrak{b}_1}+\sqrt{\mathfrak{b}_2}\right)+\frac{\sqrt{2\pi}\Omega_{\max}}{2}\left(\frac{\mathfrak{a}_1}{\mathfrak{b}_1^2}+\frac{\mathfrak{a}_2}{\mathfrak{b}_2^2}\right)+2\left(\frac{\mathfrak{a}_1}{\mathfrak{b}_1}+\frac{\mathfrak{a}_2}{\mathfrak{b}_2}\right)\right]\right\}}{3\log 2}+1\right)\Bigg) \\
&\qquad \vee \frac{18\sigma_\omega^2(2\mu_1-1)^2\vee(2\mu_k-1)^2}{(\lambda^2+\lambda+1/4)^2} \vee \frac{2\sigma_\omega^2(1+2\lambda^2)}{(1+2\lambda)^2}\Bigg]\Bigg\}\log T + 3
\end{aligned}$$

for any $k \in \{2,\ldots,K\}$ and $T \geq 2$.

Step 2: Bounding b_k.

Again, we will provide a roadmap for proving that b_k is bounded by a term of order $O(\log T)$. Similar to Step 1, we set a fixed level τ_k such that $\tau_k > \frac{\mu_k+\lambda}{1+2\lambda}$. Then, according to the theory of large deviations, we have

$$\lim_{s\to\infty} Q_{k,s}(\tau_k) = 0.$$

Thus, given the time horizon T, there exists a constant $s_{b,k}(T)$ such that $Q_{k,s}(\tau_k) \leq T^{-1}$ for all s beyond $s_{b,k}(T)$. As a result, the event $\{Q_{k,s}(\tau_k) > T^{-1}\}$ is empty if the number of pulls s exceeds $s_{b,k}(T)$. Consequently,

$$b_k \leq \sum_{s=0}^{s_{b,k}(T)} \mathbb{P}\left(Q_{k,s}(\tau_k) > T^{-1}\right).$$

We will demonstrate that the constant $s_{b,k}(T)$ is of order $O(\log T)$ in Lemma D.5, Lemma D.6, and Lemma D.7. For a small number of pulls, $s \leq s_{b,k}(T)$, we apply a trivial bound $\mathbb{P}\left(Q_{k,s}(\tau_k) > T^{-1}\right) \leq 1$ that holds for any s. Therefore, it is sufficient to conclude that b_k can be bounded by a term of order $O(\log T)$.

It should be noted that $b_{k,s}$ is naturally bounded by the constant 1. Similar to Step 1, we decompose $b_{k,s} = \mathbb{P}\big(Q_{k,s}(\tau_k) > T^{-1}\big)$ into $b_k = b_{k,s,1} + b_{k,s,2} + b_{k,s,3}$, with

$$b_{k,s,1} = \mathbb{E}\big[\mathbb{I}(Q_{k,s}(\tau_k) > T^{-1})\mathbb{I}(A_{k,s}^c)\big], \tag{6}$$

$$b_{k,s,2} = \mathbb{E}\big[\mathbb{I}(Q_{k,s}(\tau_k) > T^{-1})\mathbb{I}(A_{k,s})\mathbb{I}(G_{k,s}^c)\big], \tag{7}$$

and

$$b_{k,s,3} = \mathbb{E}\big[\mathbb{I}(Q_{k,s}(\tau_k) > T^{-1})\mathbb{I}(A_{k,s})\mathbb{I}(G_{k,s})\big]. \tag{8}$$

Again, we define $s_{b,k,j} := \max\{s : b_{k,s,j} \geq T^{-1}\}$ for $j = 1,2,3$. Lemma D.5 in Appendix D guarantees that when $s \geq s_{b,k,1}(T)$, with

$$s_{b,k,1}(T) = \frac{72\lambda^2\sigma_k^2}{(1+2\lambda)^2\Delta_k^2}\log T,$$

we have $b_{k,s,1} \le T^{-1}$. Considering $T \ge 2$ and letting $C_1 = \frac{1}{6\lambda}$ and $C_2 = \frac{1}{12\lambda}$ with fixed $\lambda > 1$ as in Step 1, Lemma D.6 and Lemma D.7 proves that: if we take

$$s \ge s_{b,k,2}(T) = \left\{ \left[\big(\Omega_k + \mathfrak{a}_1 + \mathfrak{b}_1 + \Omega_k \mathfrak{a}_1\big) \left(\frac{1}{3} \log^{-1} 2 \times \log \left\{ 3 \left[1 + \frac{\sqrt{\pi} \mathfrak{b}_1}{2} + \frac{\sqrt{2\pi} \Omega_k \mathfrak{a}_1}{2\mathfrak{b}_1^2} + \frac{2\mathfrak{a}_1}{\mathfrak{b}_1} \right] \right\} + 1 \right) \right] \vee \frac{18\sigma_\omega^2 (2\mu_k - 1)^2}{(\lambda^2 + \lambda + 1/4)^2} \vee \frac{2\sigma_\omega^2 (1 + 2\lambda^2)}{(1 + 2\lambda)^2} \right\} \times 3 \log T,$$

and

$$s \ge s_{b,k,3}(T) = \left\{ \left[\big(\Omega_k + \mathfrak{a}_2 + \mathfrak{b}_2 + \Omega_k \mathfrak{a}_2\big) \left(\frac{1}{3} \log^{-1} 2 \times \log \left\{ 3 \left[1 + \frac{\sqrt{\pi} \mathfrak{b}_2}{2} + \frac{\sqrt{2\pi} \Omega_k \mathfrak{a}_2}{2\mathfrak{b}_2^2} + \frac{2\mathfrak{a}_2}{\mathfrak{b}_2} \right] \right\} + 1 \right) \right] \vee \frac{18\sigma_\omega^2 (2\mu_k - 1)^2}{(\lambda^2 + \lambda + 1/4)^2} \vee \frac{2\sigma_\omega^2 (1 + 2\lambda^2)}{(1 + 2\lambda)^2} \right\} \times 3 \log T,$$

then we have $b_{k,s,2} \le T^{-1}$ and $b_{k,s,3} \le T^{-1}$, respectively, where $\mathfrak{a}_1, \mathfrak{a}_2, \mathfrak{b}_1, \mathfrak{b}_2$, and Ω_k are already defined in Step 1. Let

$$\begin{aligned} s_{b,k}(T) := \frac{72\lambda^2 \sigma_k^2}{(1+2\lambda)^2 \Delta_k^2} \log T + \Bigg\{ \Bigg[&\big(\Omega_{\max} + (\mathfrak{a}_1 + \mathfrak{a}_2) + (\mathfrak{b}_1 + \mathfrak{b}_2) + \Omega_{\max}(\mathfrak{a}_1 + \mathfrak{a}_2)\big) \\ &\left(\frac{1}{3} \log^{-1} 2 \times \log \left\{ 3 \left[1 + \frac{\sqrt{\pi}}{2} \left(\sqrt{\mathfrak{b}_1} + \sqrt{\mathfrak{b}_2} \right) + \frac{\sqrt{2\pi}\Omega_{\max}}{2} \left(\frac{\mathfrak{a}_1}{\mathfrak{b}_1^2} + \frac{\mathfrak{a}_2}{\mathfrak{b}_2^2} \right) + 2 \left(\frac{\mathfrak{a}_1}{\mathfrak{b}_1} + \frac{\mathfrak{a}_2}{\mathfrak{b}_2} \right) \right] \right\} + 1 \right) \Bigg] \\ &\vee \frac{18\sigma_\omega^2 (2\mu_1 - 1)^2 \vee (2\mu_k - 1)^2}{(\lambda^2 + \lambda + 1/4)^2} \vee \frac{2\sigma_\omega^2 (1 + 2\lambda^2)}{(1 + 2\lambda)^2} \Bigg\} \times 3 \log T. \end{aligned}$$

So for any $s \ge s_{b,k}(T)$, we have $s \ge \max_{j=1,2,3} s_{b,k,j}(T)$ since $s_{b,k}(T) = s_{b,k,1}(T) + \max_{j=2,3} s_{b,k,j}(T)$. Therefore, $b_{k,s,1} + b_{k,s,2} + b_{k,s,3} \le 3T^{-1}$ for any $s \ge s_{b,k}(T)$. Note that $b_{k,s} = \mathbb{P}\big(Q_{k,s}(\tau_k) > T^{-1}\big) \le 1$ for any $s \ge 0$, then

$$\begin{aligned} b_k &= 1 + \sum_{s=0}^{T-1} b_{k,s,1} \\ &\le 1 + 1 \times s_{b,k}(T) + 3T^{-1} \times \big(T - s_{b,k}(T)\big) \\ &\le \Bigg\{ \frac{72\lambda^2 \sigma_k^2}{(1+2\lambda)^2 \Delta_k^2} + \Bigg[\Bigg(\big[\Omega_{\max} + (\mathfrak{a}_1 + \mathfrak{a}_2) + (\mathfrak{b}_1 + \mathfrak{b}_2) + \Omega_{\max}(\mathfrak{a}_1 + \mathfrak{a}_2)\big] \\ &\qquad \left(\frac{1}{3} \log^{-1} 2 \times \log \left\{ 3 \left[1 + \frac{\sqrt{\pi}}{2} \left(\sqrt{\mathfrak{b}_1} + \sqrt{\mathfrak{b}_2} \right) + \frac{\sqrt{2\pi}\Omega_{\max}}{2} \left(\frac{\mathfrak{a}_1}{\mathfrak{b}_1^2} + \frac{\mathfrak{a}_2}{\mathfrak{b}_2^2} \right) + 2 \left(\frac{\mathfrak{a}_1}{\mathfrak{b}_1} + \frac{\mathfrak{a}_2}{\mathfrak{b}_2} \right) \right] \right\} + 1 \right) \Bigg) \\ &\qquad \vee \frac{18\sigma_\omega^2 (2\mu_1 - 1)^2 \vee (2\mu_k - 1)^2}{(\lambda^2 + \lambda + 1/4)^2} \vee \frac{2\sigma_\omega^2 (1 + 2\lambda^2)}{(1 + 2\lambda)^2} \Bigg] \Bigg\} \log T + 4 \end{aligned}$$

for any $k \in \{2, \ldots, K\}$ and $T \ge 2$.

Step 3: Aggregating results.

Let us define

$$d_1 = \Delta_k^2 \big\{ (\mathfrak{a}_1 + \mathfrak{a}_2) + (\mathfrak{b}_1 + \mathfrak{b}_2) + \Omega_{\max}(\mathfrak{a}_1 + \mathfrak{a}_2) \big\}$$

and

$$d_2 = 3 \left[1 + \frac{\sqrt{\pi}}{2} \left(\sqrt{\mathfrak{b}_1} + \sqrt{\mathfrak{b}_2} \right) + \frac{\sqrt{2\pi}\Omega_{\max}}{2} \left(\frac{\mathfrak{a}_1}{\mathfrak{b}_1^2} + \frac{\mathfrak{a}_2}{\mathfrak{b}_2^2} \right) + 2 \left(\frac{\mathfrak{a}_1}{\mathfrak{b}_1} + \frac{\mathfrak{a}_2}{\mathfrak{b}_2} \right) \right],$$

so the bounds in Step 1 and Step 2 yield

$$
\begin{aligned}
a_k \leq 4(1+4e^{9/8})\Bigg\{ & \frac{144\lambda\sigma_1^2}{(1+2\lambda)^2\Delta_k^2} \\
& + \left[(\Omega_{\max}+d_1)\left(\frac{1}{3}\frac{\log d_2}{\log 2}+1\right) \vee \frac{18\sigma_\omega^2(2\mu_1-1)^2 \vee (2\mu_k-1)^2}{(\lambda^2+\lambda+1/4)^2} \vee \frac{2\sigma_\omega^2(1+2\lambda^2)}{(1+2\lambda)^2}\right]\Bigg\} \log T + 3.
\end{aligned}
$$

and

$$
\begin{aligned}
b_k \leq \Bigg\{ & \frac{72\lambda^2\sigma_k^2}{(1+2\lambda)^2\Delta_k^2} \\
& + \left[(\Omega_{\max}+d_1)\left(\frac{1}{3}\frac{\log d_2}{\log 2}+1\right) \vee \frac{18\sigma_\omega^2(2\mu_1-1)^2 \vee (2\mu_k-1)^2}{(\lambda^2+\lambda+1/4)^2} \vee \frac{2\sigma_\omega^2(1+2\lambda^2)}{(1+2\lambda)^2}\right]\Bigg\} \log T + 4
\end{aligned}
$$

for any $k \in \{2, \ldots, K\}$ and $T \geq 2$. Therefore, by applying Lemma C.2, we obtain the following inequality

$$
\begin{aligned}
\mathrm{Reg}_T &\leq \sum_{k=2}^{K} \Delta_k (a_k + b_k) \\
&\leq \sum_{k=2}^{K} \Delta_k \left[7 + \left\{c_1(\mu_1, \sigma_1, \mu_k, \sigma_k, \lambda, \sigma_\omega) + c_2(\mu_1, \sigma_1, \mu_k, \sigma_k, \lambda, \sigma_\omega)\Delta_k^{-2}\right\} \log T\right]
\end{aligned} \tag{9}
$$

for any $T \geq 2$, where

$$
\begin{aligned}
& c_1(\mu_1, \sigma_1, \mu_k, \sigma_k, \lambda) \\
&\quad = (5+16e^{9/8})\left[\left(\Omega_{\max}\left(\frac{\log d_2}{3\log 2}+1\right)\right) \vee \frac{18\sigma_\omega^2\{(2\mu_1-1)^2 \vee (2\mu_k-1)^2\}}{(\lambda^2+\lambda+1/4)^2} \vee \frac{2\sigma_\omega^2(1+2\lambda^2)}{(1+2\lambda)^2}\right],
\end{aligned} \tag{10}
$$

and

$$
\begin{aligned}
& c_2(\mu_1, \sigma_1, \mu_k, \sigma_k, \lambda) \\
&= \frac{72\lambda^2\sigma_k^2}{(1+2\lambda)^2}(9+32e^{9/8}) \\
&\quad + (5+16e^{9/8})\left[\left(d_1\left(\frac{\log d_2}{3\log 2}+1\right)\right) \vee \frac{18\sigma_\omega^2\{(2\mu_1-1)^2 \vee (2\mu_k-1)^2\}}{(\lambda^2+\lambda+1/4)^2} \vee \frac{2\sigma_\omega^2(1+2\lambda^2)}{(1+2\lambda)^2}\right],
\end{aligned} \tag{11}
$$

for $k = 2, \ldots, K$. When the total round $T = 1$, the bound (9) still holds because $\mathrm{Reg}_T \leq \max_{k=2,\ldots K} \Delta_k \leq 7\sum_{k=2}^{K} \Delta_k$. Finally, by utilizing the bound in (9) and the bounds for $c_1(\mu_1, \sigma_1, \mu_k, \sigma_k, \lambda, \sigma_\omega)$ and $c_2(\mu_1, \sigma_1, \mu_k, \sigma_k, \lambda, \sigma_\omega)$ in Lemma D.8, the following inequality holds for the problem-dependent case:

$$
\begin{aligned}
\mathrm{Reg}_T &\leq \sum_{k=2}^{K} \Delta_k \left[7 + \left\{c_1(\mu_1, \sigma_1, \mu_k, \sigma_k, \lambda, \sigma_\omega) + c_2(\mu_1, \sigma_1, \mu_k, \sigma_k, \lambda, \sigma_\omega)\Delta_k^{-2}\right\} \log T\right] \\
&\leq \sum_{k=2}^{K} \Delta_k \left[7 + \left\{C_1(\sigma_1, \sigma_k, \lambda, \sigma_\omega) + C_2(\sigma_1, \sigma_k, \lambda, \sigma_\omega)\Delta_k^{-2}\right\} \log T\right].
\end{aligned}
$$

For proving the problem-independent case in Theorem C.1, let us denote

$$
J_{k,T} := \sum_{t=1}^{T} \mathbb{I}(I_t = k) = \sum_{k=2}^{K} (a_k + b_k),
$$

and then we have

$$
\begin{aligned}
\mathrm{Reg}_T &= \sum_{\Delta_k:\Delta_k<\Delta} \Delta_k \mathbb{E} J_{k,T} + \sum_{\Delta_k:\Delta_k\geq\Delta} \Delta_k \mathbb{E} J_{k,T} \\
&\overset{\text{by the bounds of } a_k, b_k}{\leq} T\Delta + \sum_{\Delta_k:\Delta_k\geq\Delta} \Delta_k \Big[7 + \big\{ C_1(\sigma_1,\sigma_k,\lambda,\sigma_\omega) + C_2(\sigma_1,\sigma_k,\lambda,\sigma_\omega)\Delta_k^{-2} \big\} \log T \Big] \\
&= T\Delta + 7 \sum_{\Delta_k:\Delta_k\geq\Delta} \Delta_k + \sum_{\Delta_k:\Delta_k\geq\Delta} C_1(\sigma_1,\sigma_k,\lambda,\sigma_\omega)\Delta_k \log T + \sum_{\Delta_k:\Delta_k\geq\Delta} \frac{C_2(\sigma_1,\sigma_k,\lambda,\sigma_\omega)}{\Delta_k} \log T \\
&\overset{\text{by } \Delta_k\leq\mu_1}{\leq} T\Delta + 7K\mu_1 + \max_{k\in[K]\setminus\{1\}} C_1(\sigma_1,\sigma_k,\lambda,\sigma_\omega) K \log T + \max_{k\in[K]\setminus\{1\}} C_2(\sigma_1,\sigma_k,\lambda,\sigma_\omega) \frac{K\log T}{\Delta},
\end{aligned}
\tag{12}
$$

for any previously specified $\Delta \in (0,1)$. Taking $\Delta = \sqrt{\max_{k\in[K]\setminus\{1\}} C_2(\sigma_1,\sigma_k,\lambda,\sigma_\omega) K \log T / T}$, we obtain

$$
\mathrm{Reg}_T \leq 7K\mu_1 + \max_{k\in[K]\setminus\{1\}} C_1(\sigma_1,\sigma_k,\lambda,\sigma_\omega) K \log T + 2\sqrt{\max_{k\in[K]\setminus\{1\}} C_2(\sigma_1,\sigma_k,\lambda,\sigma_\omega) K T \log T}.
$$

Thus, we complete the proof of Theorem C.1.

Lastly, to prove Theorem 5.1, we set $\sigma_k \equiv 1$ for $k \in [K]$ and utilize the fact

$$
C_1(1,1,\lambda,\sigma_\omega) + C_2(1,1,\lambda,\sigma_\omega)\Delta_k^{-2} \leq \frac{C_1(1,1,\lambda,\sigma_\omega) + C_2(1,1,\lambda,\sigma_\omega)}{\Delta_k^2}.
$$

Then applying the bounds for $C_1(1,1,\lambda,\sigma_\omega)$ and $C_2(1,1,\lambda,\sigma_\omega)$ from Lemma D.9, we obtain the first result in Theorem 5.1. By employing the same technique as in (12), we obtain the problem-independent regret we sought in Theorem 5.1. □

D. Lemmas on Bounding Regret Components

D.1. Lemmas on bounding a_k.

Lemma D.1 (Bounding $a_{k,s}$ for any $s > 0$)**.** *Set*

$$
\lambda \geq \left(1 + \frac{\sigma_\omega^2}{4} + \frac{4\sigma_1}{\sigma_\omega}\right) + \sqrt{\frac{4\sigma_1}{\sigma_\omega}\left(\frac{4\sigma_1}{\sigma_\omega} + 1\right)},
$$

then

$$
a_{k,s} \leq 4 + 16 e^{9/8}
$$

for any $k \in \{2,\ldots,K\}$ and $s \geq 0$.

Proof. Note that if we take

$$
\tau_k \leq \frac{\mu_1 + \lambda}{1 + 2\lambda}, \tag{13}
$$

then we can ensure that the bound

$$
\begin{aligned}
a_{k,s} &= \mathbb{E}\left[\left(\frac{1}{Q_{1,s}(\tau_k)} - 1 \right) \wedge T \right] \\
&\leq \mathbb{E}\left[\frac{1}{Q_{1,s}(\tau_k)} \right] \\
&\overset{\text{by } Q_{1,s}(\cdot) \text{ is decreasing}}{\leq} \mathbb{E} Q_{1,s}^{-1}\left(\frac{\mu_1 + \lambda}{1 + 2\lambda} \right),
\end{aligned}
\tag{14}
$$

holds. Hence, we need to find a lower bound of the tail probability $Q_{1,s}\left(\frac{\mu_1+\lambda}{1+2\lambda}\right) = \mathbb{P}\left(\overline{Y}_{1,s} > \frac{\mu_1+\lambda}{1+2\lambda} \mid \mathcal{H}_{1,s}\right)$. Throughout the proof, we will use the choice of (13) for τ_k, and we can take advantage of the bound (14).

For further analyze (14), we will express it as the probability with respect to the weighted random summation of the Gaussian random variables $\{\omega_i, \omega_i', \omega_i''\}$. Let $x_i := (2\lambda+1)R_{1,i} - (\mu_1+\lambda)$, $y_i := \lambda(\mu_1 - 1 - \lambda)$, $z_i := \lambda(\lambda+\mu_1)$ for $i \in [s]$, then

$$S_x = \sum_{i=1}^{s} x_i = s\big((2\lambda+1)\overline{R}_{1,s} - (\mu_1+\lambda)\big), \quad S_y = \sum_{i=1}^{s} y_i = s\lambda(\mu_1 - 1 - \lambda), \quad S_z = \sum_{i=1}^{s} z_i = s\lambda(\lambda+\mu_1),$$

with $S_x - S_y - S_z = s(2\lambda+1)(\overline{R}_{1,s} - \mu_1)$ and

$$T_x = \sum_{i=1}^{s} x_i^2 = \sum_{i=1}^{s} \left[(2\lambda+1)R_{1,i} - (\mu_1+\lambda)\right]^2, \quad T_y = \sum_{i=1}^{s} y_i^2 = s\lambda^2(\mu_1 - 1 - \lambda)^2, \quad T_z = \sum_{i=1}^{s} z_i^2 = s\lambda^2(\lambda+\mu_1)^2,$$

with $T_x + T_y + T_z = \sum_{i=1}^{s} \left[(2\lambda+1)\overline{R}_{1,s} - (\mu_1+\lambda)\right]^2 + s\big[\lambda^2(\mu_1 - 1 - \lambda)^2 + \lambda^2(\lambda+\mu_1)^2\big]$. Denote

$$Z_1 := \frac{\sum_{i=1}^{s} x_i\omega_i - \sum_{i=1}^{s} y_i\omega_i' - \sum_{i=1}^{s} z_i\omega_i'' - (S_x - S_y - S_z)}{\sigma_\omega\sqrt{T_x + T_y + T_z}}$$

and

$$Z_2 := \frac{\sum_{i=1}^{s} \omega_i + \lambda\sum_{i=1}^{s} \omega_i' + \lambda\sum_{i=1}^{s} \omega_i'' - (1+2\lambda)s}{\sigma_\omega\sqrt{(1+2\lambda^2)s}}.$$

Then

$$\begin{aligned}
&Q_{1,s}\left(\frac{\mu_1+\lambda}{1+2\lambda}\right) \\
=&\mathbb{P}\left(\overline{Y}_{1,s} > \frac{\mu_1+\lambda}{1+2\lambda} \mid \mathcal{H}_{1,s}\right) \\
\geq&\mathbb{P}_{\omega,\omega',\omega''}\left(\left\{\sum_{i=1}^{s} x_i\omega_i - \sum_{i=1}^{s} y_i\omega_i' - \sum_{i=1}^{s} z_i\omega_i'' > 0\right\} \cap \left\{\sum_{i=1}^{s}\omega_i + \lambda\sum_{i=1}^{s}\omega_i' + \lambda\sum_{i=1}^{s}\omega_i'' > 0\right\}\right) \\
=&\mathbb{P}_{Z_1,Z_2}\left(\left\{Z_1 > \frac{-(S_x - S_y - S_z)}{\sigma_\omega\sqrt{T_x+T_y+T_z}}\right\} \cap \left\{Z_2 > \frac{-(1+2\lambda)\sqrt{s}}{\sigma_\omega\sqrt{1+2\lambda^2}}\right\}\right) \\
=&\mathbb{P}_{Z_1,Z_2}\left(\left\{Z_1 > \frac{-(S_x - S_y - S_z)}{\sigma_\omega\sqrt{T_x+T_y+T_z}}\right\} \cap \left\{Z_2 > \frac{-(1+2\lambda)\sqrt{s}}{\sigma_\omega\sqrt{1+2\lambda^2}}\right\}\right)\mathbb{I}\left(\overline{R}_{1,s} - \mu_1 \geq 0\right) \\
&+\mathbb{P}_{Z_1,Z_2}\left(\left\{Z_1 > \frac{-(S_x - S_y - S_z)}{\sigma_\omega\sqrt{T_x+T_y+T_z}}\right\} \cap \left\{Z_2 > \frac{-(1+2\lambda)\sqrt{s}}{\sigma_\omega\sqrt{1+2\lambda^2}}\right\}\right)\mathbb{I}\left(\overline{R}_{1,s} - \mu_1 < 0\right).
\end{aligned} \tag{15}$$

Note that $(Z_1, Z_2)^\top \in \mathbb{R}^2$ is the mean-zero bivariate Gaussian distribution with covariance $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, where

$$\begin{aligned}
\rho &= \mathbb{E}_{\omega,\omega',\omega''}[Z_1 Z_2] - 0 \\
&= \frac{\mathbb{E}_{\omega,\omega',\omega''}\Big\{\left[\sum_{i=1}^{s} x_i\omega_i - \sum_{i=1}^{s} y_i\omega_i' - \sum_{i=1}^{s} z_i\omega_i'' - (S_x - S_y - S_z)\right]\left[\sum_{i=1}^{s}\omega_i + \lambda\sum_{i=1}^{s}\omega_i' + \lambda\sum_{i=1}^{s}\omega_i'' - (1+2\lambda)s\right]\Big\}}{\sigma_\omega^2\sqrt{s(1+2\lambda^2)(T_x+T_y+T_z)}} \\
&= \frac{\sigma_\omega^2(S_x - \lambda S_y - \lambda S_z) - (S_x - S_y - S_z)(1+2\lambda)s}{\sigma_\omega^2\sqrt{s(1+2\lambda^2)(T_x+T_y+T_z)}}.
\end{aligned}$$

The sign of ρ depends on the sign of $\check{\rho}$, where $\check{\rho}$ is defined as

$$\begin{aligned}
\check{\rho} &:= \frac{1}{s}\left[\sigma_\omega^2(S_x - \lambda S_y - \lambda S_z) - (S_x - S_y - S_z)(1+2\lambda)s\right] \\
&= \sigma_\omega^2\Big[(2\lambda+1)\overline{R}_{1,s} - \mu_1(2\lambda^2+1) + \lambda^2 - \lambda\Big] - 2(1+2\lambda)^2 s(\overline{R}_{1,s} - \mu_1) \\
&= (1+2\lambda)\big[-(\overline{R}_{1,s} - \mu_1)\big]\left[2(1+2\lambda)s - \sigma_\omega^2\right] + \sigma_\omega^2\lambda(\lambda-1)(1-2\mu_1).
\end{aligned}$$

Thus, for the event $\{\rho < 0\}$, if we take $\lambda > \frac{\sigma_\omega^2}{4} - \frac{1}{2}$, we will have

$$\begin{aligned}\{\rho < 0\} &= \{\breve{\rho} < 0\} \\ &= \left\{ -(\overline{R}_{1,s} - \mu_1) < \frac{\sigma_\omega^2 \lambda(\lambda-1)(2\mu_1 - 1)}{(1+2\lambda)\big[2(1+2\lambda)s - \sigma_\omega^2\big]} : s \geq 1 \right\} \\ &= \left\{ \overline{R}_{1,s} - \mu_1 > -\frac{\sigma_\omega^2 \lambda(\lambda-1)(2\mu_1 - 1)}{(1+2\lambda)\big[2(1+2\lambda)s - \sigma_\omega^2\big]} : s \geq 1 \right\}.\end{aligned}$$

This furthermore implies

$$\begin{aligned}&\{\rho < 0\} \cap \{\overline{R}_{1,s} - \mu_1 < 0\} \\ =&\left\{ (\overline{R}_{1,s} - \mu_1) > -\frac{\sigma_\omega^2 \lambda(\lambda-1)(2\mu_1 - 1)}{(1+2\lambda)\big[2(1+2\lambda)s - \sigma_\omega^2\big]} : s \geq 1 \right\} \cap \{\overline{R}_{1,s} - \mu_1 < 0\} \\ =&\begin{cases} \varnothing, & \text{if } 2\mu_1 - 1 \leq 0 \\ \left\{ -\frac{\sigma_\omega^2 \lambda(\lambda-1)(2\mu_1-1)}{(1+2\lambda)\left[2(1+2\lambda)s - \sigma_\omega^2\right]} < \overline{R}_{1,s} - \mu_1 < 0 : s \geq 1 \right\}, & \text{if } 2\mu_1 - 1 > 0. \end{cases} \\ \subseteq&\begin{cases} \varnothing, & \text{if } 2\mu_1 - 1 \leq 0 \\ \mathcal{A}_s =: \left\{ -\frac{\sigma_\omega^2 \lambda(\lambda-1)(2\mu_1-1)}{2(1+2\lambda)^2 s} < \overline{R}_{1,s} - \mu_1 < 0 : s \geq 1 \right\}, & \text{if } 2\mu_1 - 1 > 0. \end{cases}\end{aligned}$$

Now, we can decompose the second probability in (15) as

$$\begin{aligned}&\mathbb{P}_{Z_1,Z_2}\left(\left\{ Z_1 > \frac{-(S_x - S_y - S_z)}{\sigma_\omega\sqrt{T_x + T_y + T_z}} \right\} \cap \left\{ Z_2 > \frac{-(1+2\lambda)\sqrt{s}}{\sigma_\omega\sqrt{1+2\lambda^2}} \right\}\right) \mathbb{I}\left(\overline{R}_{1,s} - \mu_1 < 0\right) \\ =&\mathbb{P}_{Z_1,Z_2}\left(\left\{ Z_1 > \frac{-(S_x - S_y - S_z)}{\sigma_\omega\sqrt{T_x + T_y + T_z}} \right\} \cap \left\{ Z_2 > \frac{-(1+2\lambda)\sqrt{s}}{\sigma_\omega\sqrt{1+2\lambda^2}} \right\}\right) \mathbb{I}\left(\overline{R}_{1,s} - \mu_1 < 0\right) \mathbb{I}\left(\rho \geq 0\right) \\ &+ \mathbb{P}_{Z_1,Z_2}\left(\left\{ Z_1 > \frac{-(S_x - S_y - S_z)}{\sigma_\omega\sqrt{T_x + T_y + T_z}} \right\} \cap \left\{ Z_2 > \frac{-(1+2\lambda)\sqrt{s}}{\sigma_\omega\sqrt{1+2\lambda^2}} \right\}\right) \mathbb{I}\left(\overline{R}_{1,s} - \mu_1 < 0\right) \mathbb{I}\left(\rho < 0\right) \\ =&\mathbb{P}_{Z_1,Z_2}\left(\left\{ Z_1 > \frac{-(S_x - S_y - S_z)}{\sigma_\omega\sqrt{T_x + T_y + T_z}} \right\} \cap \left\{ Z_2 > \frac{-(1+2\lambda)\sqrt{s}}{\sigma_\omega\sqrt{1+2\lambda^2}} \right\}\right) \mathbb{I}\left(\overline{R}_{1,s} - \mu_1 < 0\right) \mathbb{I}\left(\rho \geq 0\right) \\ &+ \mathbb{P}_{Z_1,Z_2}\left(\left\{ Z_1 > \frac{-(S_x - S_y - S_z)}{\sigma_\omega\sqrt{T_x + T_y + T_z}} \right\} \cap \left\{ -Z_2 < \frac{(1+2\lambda)\sqrt{s}}{\sigma_\omega\sqrt{1+2\lambda^2}} \right\}\right) \mathbb{I}\left(\overline{R}_{1,s} - \mu_1 < 0\right) \mathbb{I}\left(-\rho > 0\right) \\ =&\mathbb{P}_{Z_1,Z_2}\left(\left\{ Z_1 > \frac{-(S_x - S_y - S_z)}{\sigma_\omega\sqrt{T_x + T_y + T_z}} \right\} \cap \left\{ Z_2 > \frac{-(1+2\lambda)\sqrt{s}}{\sigma_\omega\sqrt{1+2\lambda^2}} \right\}\right) \mathbb{I}\left(\overline{R}_{1,s} - \mu_1 < 0\right) \mathbb{I}\left(\rho \geq 0\right) \\ &+ \mathbb{P}_{Z_1,Z_3}\left(\left\{ Z_1 > \frac{-(S_x - S_y - S_z)}{\sigma_\omega\sqrt{T_x + T_y + T_z}} \right\} \cap \left\{ Z_3 < \frac{(1+2\lambda)\sqrt{s}}{\sigma_\omega\sqrt{1+2\lambda^2}} \right\}\right) \mathbb{I}\left(\overline{R}_{1,s} - \mu_1 < 0\right) \mathbb{I}\left(\varrho > 0\right),\end{aligned}$$

where $Z_3 = -Z_2$ and $\varrho = -\rho$ is the correlation coefficient between Z_1 and Z_3. We will utilize the lower bound of the tail probability for bivariate Gaussian distribution with a positive correlation coefficient.

First,

$$\begin{aligned}
&\mathbb{P}_{Z_1,Z_2}\left(\left\{Z_1>\frac{-(S_x-S_y-S_z)}{\sigma_\omega\sqrt{T_x+T_y+T_z}}\right\}\cap\left\{Z_2>\frac{-(1+2\lambda)\sqrt{s}}{\sigma_\omega\sqrt{1+2\lambda^2}}\right\}\right)\mathbb{I}\left(\overline{R}_{1,s}-\mu_1\le 0\right)\mathbb{I}\left(\rho\ge 0\right)\\
&\ge\mathbb{P}_{Z_1,Z_2}\Bigg(Z_1>\max\left\{\frac{-(S_x-S_y-S_z)}{\sigma_\omega\sqrt{T_x+T_y+T_z}},\frac{-(1+2\lambda)\sqrt{s}}{\sigma_\omega\sqrt{1+2\lambda^2}}\right\},\\
&\qquad\qquad Z_2>\max\left\{\frac{-(S_x-S_y-S_z)}{\sigma_\omega\sqrt{T_x+T_y+T_z}},\frac{-(1+2\lambda)\sqrt{s}}{\sigma_\omega\sqrt{1+2\lambda^2}}\right\}\Bigg)\mathbb{I}\left(\overline{R}_{1,s}-\mu_1\le 0\right)\mathbb{I}\left(\rho\ge 0\right)\\
&=\mathbb{P}_{Z_1,Z_2}\left(Z_1>\frac{-(S_x-S_y-S_z)}{\sigma_\omega\sqrt{T_x+T_y+T_z}},\,Z_2>\frac{-(S_x-S_y-S_z)}{\sigma_\omega\sqrt{T_x+T_y+T_z}}\right)\mathbb{I}\left(\overline{R}_{1,s}-\mu_1\le 0\right)\mathbb{I}\left(\rho\ge 0\right)\\
&\ge\mathbb{P}_{Z_1}\left(Z_1>\frac{-(S_x-S_y-S_z)}{\sigma_\omega\sqrt{T_x+T_y+T_z}}\right)\times\mathbb{P}_{Z_2}\left(Z_2>\sqrt{\frac{1-\rho}{1+\rho}}\cdot\frac{-(S_x-S_y-S_z)}{\sigma_\omega\sqrt{T_x+T_y+T_z}}\right)\mathbb{I}\left(\overline{R}_{1,s}-\mu_1\le 0\right)\mathbb{I}\left(\rho\ge 0\right)\\
&\overset{\text{by }\sqrt{\frac{1-\rho}{1+\rho}}\le 1}{\ge}\left[\mathbb{P}_Z\left(Z>\frac{-(S_x-S_y-S_z)}{\sigma_\omega\sqrt{T_x+T_y+T_z}}\right)\right]^2\mathbb{I}\left(\overline{R}_{1,s}-\mu_1\le 0\right)\\
&=\left[\mathbb{P}_Z\left(Z>\frac{-(S_x-S_y-S_z)}{\sigma_\omega\sqrt{T_x+T_y+T_z}}\right)\mathbb{I}\left(\overline{R}_{1,s}-\mu_1\le 0\right)\right]^2,
\end{aligned}$$

where Z is the standard Gaussian distribution. Throughout the appendix, we will always use Z to represent the standard Gaussian distribution. The first equality is due to $-(S_x-S_y-S_z)=s(1+2\lambda)\big[-(\overline{R}_{1,s}-\mu_1)\big]\ge 0\ge-(1+2\lambda)\sqrt{s}$ when $\overline{R}_{1,s}-\mu_1\le 0$, and the second inequality is by (11.44) of (Lai & Balakrishnan, 2009) for bivariate Gaussian distributions. By applying the first inequality in Lemma E.1 and using the fact that $T_x+T_y+T_z\ge T_z\ge s\lambda^4$, we can get that

$$\mathbb{P}_Z\left(Z>\frac{-(S_x-S_y-S_z)}{\sigma_\omega\sqrt{T_x+T_y+T_z}}\right)\mathbb{I}\left(\overline{R}_{1,s}-\mu_1\le 0\right)\ge\frac{1}{4}\exp\left\{-\frac{s(2\lambda+1)^2(\overline{R}_{1,s}-\mu_1)^2}{\sigma_\omega^2\lambda^4}\right\}\mathbb{I}\left(\overline{R}_{1,s}-\mu_1\le 0\right),$$

then

$$\begin{aligned}
&\mathbb{P}_{Z_1,Z_2}\left(\left\{Z_1>\frac{-(S_x-S_y-S_z)}{\sigma_\omega\sqrt{T_x+T_y+T_z}}\right\}\cap\left\{Z_2>\frac{-(1+2\lambda)\sqrt{s}}{\sigma_\omega\sqrt{1+2\lambda^2}}\right\}\right)\mathbb{I}\left(\overline{R}_{1,s}-\mu_1\le 0\right)\mathbb{I}\left(\rho\ge 0\right)\\
&\ge\frac{1}{16}\exp\left\{-\frac{2s(2\lambda+1)^2(\overline{R}_{1,s}-\mu_1)^2}{\sigma_\omega^2\lambda^4}\right\}\mathbb{I}\left(\overline{R}_{1,s}-\mu_1\le 0\right)\mathbb{I}\left(\rho\ge 0\right).
\end{aligned}\tag{16}$$

Second, we also have

$$\begin{aligned}
&\mathbb{P}_{Z_1,Z_3}\left(\left\{Z_1>\frac{-(S_x-S_y-S_z)}{\sigma_\omega\sqrt{T_x+T_y+T_z}}\right\}\cap\left\{Z_3<\frac{(1+2\lambda)\sqrt{s}}{\sigma_\omega\sqrt{1+2\lambda^2}}\right\}\right)\mathbb{I}\left(\overline{R}_{1,s}-\mu_1<0\right)\mathbb{I}\left(\varrho>0\right)\\
&\ge\mathbb{P}_{Z_1,Z_3}\left(\left\{Z_1>\frac{-(S_x-S_y-S_z)}{\sigma_\omega\sqrt{T_x+T_y+T_z}}\right\}\cap\left\{Z_3<\frac{(1+2\lambda)\sqrt{s}}{\sigma_\omega\sqrt{1+2\lambda^2}}\right\}\right)\mathbb{I}\left(\overline{R}_{1,s}-\mu_1<0\right)\mathbb{I}\left(\varrho>0\right)\\
&=\mathbb{P}_{Z_1,Z_3}\big(Z_1>h(s),Z_3<k(s)\big)\mathbb{I}\left(\overline{R}_{1,s}-\mu_1<0\right)\mathbb{I}\left(\varrho>0\right)\\
&=\mathbb{P}_{Z_1,Z_3}\big(Z_1>h(s),Z_3<k(s)\big)\mathbb{I}\left(\overline{R}_{1,s}-\mu_1<0\right)\mathbb{I}\left(\varrho>0\right)\mathbb{I}(2\mu_1-1>0)
\end{aligned}$$

where

$$h(s):=\frac{-(S_x-S_y-S_z)}{\sigma_\omega\sqrt{T_x+T_y+T_z}},\qquad k(s):=\frac{(1+2\lambda)\sqrt{s}}{\sigma_\omega\sqrt{1+2\lambda^2}}$$

are positive when $\overline{R}_{1,s}-\mu_1<0$. Note that

$$\mathbb{P}_{Z_1,Z_3}\big(Z_1>h(s),Z_3<k(s)\big)=\mathbb{P}_{Z_1}\big(Z_1>h(s)\big)-\mathbb{P}_{Z_1,Z_3}\big(Z_1>h(s),Z_3>k(s)\big)$$

and

$$\begin{aligned}&\mathbb{P}_{Z_1,Z_3}\big(Z_1 > h(s), Z_3 > k(s)\big)\mathbb{I}\big(\varrho > 0\big)\\ \leq&\Bigg\{\Phi\big(-k(s)\big) - \left[\Phi\left(\frac{\varrho k(s) - h(s)}{\sqrt{1-\varrho^2}}\right) + \varrho \exp\left[\frac{k^2(s) - h^2(s)}{2}\right]\Phi\left(\frac{\varrho h(s) - k(s)}{\sqrt{1-\varrho^2}}\right)\right]\Bigg\}\mathbb{I}\big(\varrho > 0\big)\end{aligned}$$

by (11.42) in Lai & Balakrishnan (2009). Next, by combining the above two inequalities, we have

$$\begin{aligned}&\mathbb{P}_{Z_1,Z_3}\left(\left\{Z_1 > \frac{-(S_x - S_y - S_z)}{\sigma_\omega\sqrt{T_x+T_y+T_z}}\right\} \cap \left\{Z_3 < \frac{(1+2\lambda)\sqrt{s}}{\sigma_\omega\sqrt{1+2\lambda^2}}\right\}\right)\mathbb{I}\left(\overline{R}_{1,s} - \mu_1 < 0\right)\mathbb{I}\left(\varrho > 0\right)\\ \geq&\Bigg\{\Big[\Phi\big(-h(s)\big) - \Phi\big(-k(s)\big)\Big]\\ &\quad + \left[\Phi\left(\frac{\varrho k(s) - h(s)}{\sqrt{1-\varrho^2}}\right) + \varrho\exp\left[\frac{k^2(s) - h^2(s)}{2}\right]\Phi\left(\frac{\varrho h(s) - k(s)}{\sqrt{1-\varrho^2}}\right)\right]\Bigg\}\mathbb{I}\left(\overline{R}_{1,s} - \mu_1 < 0\right)\mathbb{I}\big(\varrho > 0\big)\\ =&\Bigg\{\Big[\Phi\big(-h(s)\big) - \Phi\big(-k(s)\big)\Big]\\ &\quad + \left[\Phi\left(\frac{\varrho k(s) - h(s)}{\sqrt{1-\varrho^2}}\right) + \varrho\exp\left[\frac{k^2(s) - h^2(s)}{2}\right]\Phi\left(\frac{\varrho h(s) - k(s)}{\sqrt{1-\varrho^2}}\right)\right]\Bigg\}\mathbb{I}\Big(\big\{\overline{R}_{1,s} - \mu_1 < 0\big\} \cap \{\varrho > 0\}\Big).\end{aligned}$$

Note that $\big\{\overline{R}_{1,s} - \mu_1 < 0\big\} \cap \{\varrho > 0\} \subseteq \mathcal{A}_s$ always hold, we have

$$0 < -s(1+2\lambda)(\overline{R}_{1,s} - \mu_1) < \frac{\sigma_\omega^2\lambda(\lambda-1)(2\mu_1 - 1)}{2(1+2\lambda)}. \tag{17}$$

Then using the inequality $T_x + T_y + T_z \geq s\lambda^4$, we have

$$\begin{aligned}h(s) &= \frac{-(S_x - S_y - S_z)}{\sigma_\omega\sqrt{T_x+T_y+T_z}}\\ &\overset{\text{by (17)}}{<} \frac{\sigma_\omega\lambda(\lambda-1)(2\mu_1-1)}{\sqrt{s}\lambda^2(1+2\lambda)}\\ &\overset{\text{by } \mu_1 \leq 1}{\leq} \frac{\sigma_\omega(\lambda-1)}{\lambda(1+2\lambda)}\\ &\leq \frac{\sigma_\omega}{1+2\lambda}\\ &\overset{\text{by (19)}}{\leq} \frac{(1+2\lambda)}{2\sigma_\omega\lambda}\\ &\overset{\text{by } \lambda > 1}{\leq} \frac{(1+2\lambda)}{\sigma_\omega\sqrt{1+2\lambda^2}} \leq \frac{(1+2\lambda)\sqrt{s}}{\sigma_\omega\sqrt{1+2\lambda^2}} = k(s),\end{aligned} \tag{18}$$

where fourth inequality follows from the fact that

$$2\sigma_\omega^2\lambda < (1+2\lambda)^2 \tag{19}$$

provided that

$$\lambda\begin{cases}\in \mathbb{R} & \text{if } \sigma_\omega < 2,\\ > \frac{2-\sigma_\omega+\sigma_\omega\sqrt{\sigma_\omega^2-4}}{4}, & \text{if } \sigma_\omega \geq 2\end{cases}.$$

Note that when $\sigma_\omega > 2$, the expression $\frac{2-\sigma_\omega+\sigma_\omega\sqrt{\sigma_\omega^2-4}}{4}$ is negative. Thus, the inequality (18) will hold for any $\lambda > 1$. This implies that

$$\Big[\Phi\big(-h(s)\big) - \Phi\big(-k(s)\big)\Big]\mathbb{I}(\mathcal{A}_s) > 0.$$

whenever $\lambda > \frac{\sigma_\omega^2}{4} - \frac{1}{2}$. On the other hand, note that $-\check{\rho} \geq \sigma_\omega^2 \lambda(\lambda-1)(2\mu_1 - 1)$, then on the event $\mathcal{A}_s$, we know that

$$\varrho = \frac{-s\check{\rho}}{\sigma_\omega^2 \sqrt{s(1+2\lambda^2)(T_x+T_y+T_z)}} \geq \frac{s\sigma_\omega^2 \lambda(\lambda-1)(2\mu_1-1)}{\sigma_\omega^2 \sqrt{s(1+2\lambda^2)(T_x+T_y+T_z)}}$$

and thus

$$\begin{aligned}
\frac{\varrho k(s)}{h(s)} &= \frac{(1+2\lambda)\sqrt{s}}{\sigma_\omega \sqrt{1+2\lambda^2}} \times \frac{\varrho}{h(s)} \\
&\overset{\text{by the bound of } \varrho}{\geq} \frac{(1+2\lambda)\sqrt{s}}{\sigma_\omega \sqrt{1+2\lambda^2}} \times \frac{s\sigma_\omega^2 \lambda(\lambda-1)(2\mu_1-1)}{s(2\lambda+1)\left[-(\overline{R}_{1,s}-\mu_1)\right]} \\
&\overset{\text{by the bound in (17)}}{\geq} \frac{(1+2\lambda)\sqrt{s}}{\sigma_\omega \sqrt{1+2\lambda^2}} \times \left[s\sigma_\omega^2 \lambda(\lambda-1)(2\mu_1-1)\right] \times \frac{2(1+2\lambda)}{\sigma_\omega \lambda(\lambda-1)(2\mu_1-1)} \\
&= \frac{2(1+2\lambda)^2}{\sqrt{1+2\lambda^2}} \sqrt{s}s > 1,
\end{aligned}$$

i.e., $\varrho k(s) > h(s)$. Combining the above analysis, we have

$$\begin{aligned}
&\mathbb{P}_{Z_1,Z_3}\left(\left\{Z_1 > \frac{-(S_x-S_y-S_z)}{\sigma_\omega\sqrt{T_x+T_y+T_z}}\right\} \cap \left\{Z_3 < \frac{(1+2\lambda)\sqrt{s}}{\sigma_\omega\sqrt{1+2\lambda^2}}\right\}\right) \mathbb{I}\left(\overline{R}_{1,s}-\mu_1<0\right)\mathbb{I}\left(\varrho>0\right) \\
\geq& \Big[\Phi\big(-h(s)\big) - \Phi\big(-k(s)\big)\Big] \mathbb{I}\left(\overline{R}_{1,s}-\mu_1<0\right)\mathbb{I}\left(\varrho>0\right) \\
&\qquad + \Phi\left(\frac{\varrho k(s)-h(s)}{\sqrt{1-\varrho^2}}\right)\mathbb{I}\left(\overline{R}_{1,s}-\mu_1<0\right)\mathbb{I}\left(\varrho>0\right) \\
>& 0\mathbb{I}\left(\overline{R}_{1,s}-\mu_1<0\right)\mathbb{I}\left(\varrho>0\right) + \Phi(0)\mathbb{I}\left(\overline{R}_{1,s}-\mu_1<0\right)\mathbb{I}\left(\varrho>0\right) \\
=& \frac{1}{2}\mathbb{I}\left(\overline{R}_{1,s}-\mu_1<0\right)\mathbb{I}\left(\varrho>0\right).
\end{aligned} \tag{20}$$

Besides, note that when $\overline{R}_{1,s} - \mu_1 > 0$, we have $-(S_x - S_y - S_z) < 0$, and then

$$\begin{aligned}
&\mathbb{P}_{Z_1,Z_2}\left(\left\{Z_1 > \frac{-(S_x-S_y-S_z)}{\sigma_\omega\sqrt{T_x+T_y+T_z}}\right\} \cap \left\{Z_2 > \frac{-(1+2\lambda)\sqrt{s}}{\sigma_\omega\sqrt{1+2\lambda^2}}\right\}\right) \mathbb{I}\left(\overline{R}_{1,s}-\mu_1>0\right) \\
\geq& \frac{1}{2} \times \mathbb{I}\left(\overline{R}_{1,s}-\mu_1>0\right).
\end{aligned} \tag{21}$$

Plugging (16), (20), and (21) into (15), we obtain that

$$\begin{aligned}
&Q_{1,s}\left(\frac{\mu_1+\lambda}{1+2\lambda}\right) \\
\geq& \frac{1}{16}\exp\left\{-\frac{2s(2\lambda+1)^2(\overline{R}_{1,s}-\mu_1)^2}{\sigma_\omega^2\lambda^4}\right\}\mathbb{I}\left(\overline{R}_{1,s}-\mu_1\leq 0\right)\mathbb{I}(\rho\geq 0) \\
&\qquad + \frac{1}{2}\times\mathbb{I}\left(\overline{R}_{1,s}-\mu_1\leq 0\right)\mathbb{I}(\rho<0) + \frac{1}{2}\times\mathbb{I}\left(\overline{R}_{1,s}-\mu_1>0\right).
\end{aligned}$$

As a result, with the fact (14), the upper bound for $a_{k,s}$ now is

$$\begin{aligned}
a_{k,s} &= \mathbb{E}Q_{1,s}^{-1}\left(\frac{\mu_1+\lambda}{1+2\lambda}\right) \\
&\leq 16\mathbb{E}\exp\left\{\frac{2s(2\lambda+1)^2(\overline{R}_{1,s}-\mu_1)^2}{\sigma_\omega^2\lambda^4}\right\}\mathbb{I}\left(\overline{R}_{1,s}-\mu_1\leq 0\right)\mathbb{I}(\rho\geq 0) \\
&\qquad + 2\mathbb{E}\mathbb{I}\left(\overline{R}_{1,s}-\mu_1\leq 0\right)\mathbb{I}(\rho<0) + 2\mathbb{E}\mathbb{I}\left(\overline{R}_{1,s}-\mu_1>0\right) \\
&\leq 16\mathbb{E}\exp\left\{\frac{2s(2\lambda+1)^2(\overline{R}_{1,s}-\mu_1)^2}{\sigma_\omega^2\lambda^4}\right\} + 4.
\end{aligned}$$

where the first inequality is due to that the several indicators are based on mutually exclusive events.

From the above expression, to prove that $a_{k,s}$ is bounded by a constant independent of s, it remains to show that the expectation

$$\mathbb{E} \exp\left\{\frac{2s(2\lambda+1)^2(\overline{R}_{1,s}-\mu_1)^2}{\sigma_\omega^2\lambda^4}\right\} \tag{22}$$

can be bounded below some constants free of s.

Applying Lemma E.2, we know that if we take

$$\frac{2s(2\lambda+1)^2}{\sigma_\omega^2\lambda^4} \leq \frac{s}{8\sigma_1^2},$$

then (22) can be upper bounded by $e^{9/8}$. A sufficient condition for the above inequality is

$$\frac{(2\lambda+1)^2}{\sigma_\omega^2\lambda^4} \leq \frac{1}{16\sigma_1^2}, \qquad \text{i.e.} \qquad \lambda^2 - \frac{8\sigma_1}{\sigma_\omega}\lambda - \frac{4\sigma_1}{\sigma_\omega} \geq 0,$$

in other words, this can be expressed as:

$$\lambda \geq \frac{4\sigma_1}{\sigma_\omega} + \sqrt{\frac{4\sigma_1}{\sigma_\omega}\left(\frac{4\sigma_1}{\sigma_\omega}+1\right)}. \tag{23}$$

Therefore, if we take the tuning parameters λ and σ_ω as specified in (23), then $a_{k,s}$ will be bounded by a constant such that

$$a_{k,s} \leq 4 + 16e^{9/8}. \tag{24}$$

It is important to note that this lemma establishes the tuning conditions for the tuning parameters in Theorem C.1. □

Lemma D.2 (Bounding $a_{k,s,1}$ at (3))**.** *Take*

$$s_{a,k,1}(T) := \frac{4\sigma_1^2}{C_1^2(1+2\lambda)^2\Delta_k^2} \times \log T.$$

Then for any $k \in \{2,\ldots,K\}$ and $s \geq s_{a,k,1}(T)$,

$$a_{k,s,1} \leq T^{-1},$$

when $T \geq 2$.

Proof. We will bound $a_{k,s,1}$ by bounding the probability of the event $A_{k,s}^c$. Write

$$a_{k,s,1} = \mathbb{E}\Big[\big(N_{1,s}(\tau_k) \wedge T\big)\mathbb{I}(A_{1,s}^c)\Big] \leq \mathbb{E}\Big[T\mathbb{I}(A_{1,s}^c)\Big] = T\mathbb{P}(A_{1,s}^c). \tag{25}$$

Since the summation of independent sub-Gaussian variables is still sub-Gaussian, (40) in Lemma E.2 gives $\overline{R}_{1,s} - \mu_1 \sim \operatorname{subG}(\sigma_1^2/s)$. By applying the concentration inequality of the sub-Gaussian variable, we have

$$\begin{aligned}\mathbb{P}(A_{1,s}^c) &= \mathbb{P}\left(\left|\overline{R}_{1,s}-\mu_1\right| \geq (1+2\lambda)C_1\Delta_k\right) \\ &\leq 2\exp\left\{-\frac{(1+2\lambda)^2C_1^2\Delta_k^2 s}{2\sigma_1^2}\right\} \\ &\overset{\text{take } s \geq s_{a,1}(T) \geq \frac{2\sigma_1^2}{C_1^2(1+2\lambda)^2\Delta_k^2} \times \log(\sqrt{2}T)}{\leq} \frac{1}{T^2}.\end{aligned}$$

Hence, we obtain that

$$a_{k,s,1} \leq T\mathbb{P}(A_{1,s}^c) \leq T^{-1}, \qquad \text{for any} \qquad s \geq s_{a,1}(T).$$

□

Lemma D.3 (Bounding $a_{k,s,2}$ at (4))**.** *Take*

$$s_{a,k,2}(T) := \left\{ \left[\left(\Omega_1 + \mathfrak{a}_1 + \mathfrak{b}_1 + \Omega_1 \mathfrak{a}_1\right) \left(\frac{1}{3} \log^{-1} 2 \times \log \left\{ 3 \left[1 + \frac{\sqrt{\pi \mathfrak{b}_1}}{2} + \frac{\sqrt{2\pi}\Omega_1 \sigma_1^2 \mathfrak{a}_1}{2\mathfrak{b}_1^2} + \frac{2\mathfrak{a}_1}{\mathfrak{b}_1} \right] \right\} + 1 \right) \right] \vee \frac{18\sigma_\omega^2 (2\mu_1 - 1)^2}{(\lambda^2 + \lambda + 1/4)^2} \vee \frac{2\sigma_\omega^2 (1 + 2\lambda^2)}{(1 + 2\lambda)^2} \right\} \times 3 \log T,$$

where

$$\mathfrak{a}_1 = \frac{4\sigma_\omega^2 \lambda^2}{3C_2^2 (1+2\lambda)^2 \Delta_k^2}, \qquad \mathfrak{b}_1 = \frac{2\sigma_\omega^2 \left[2\lambda^2 \sigma_1^2 + (1+2\lambda)^2 C_2^2 \Delta_k^2\right]}{3C_2^2 (1+2\lambda)^2 \Delta_k^2}, \qquad \textit{and} \qquad \Omega_1 = 8\sqrt{2}\sigma_1^2.$$

Then for any $s \geq s_{a,k,2}(T)$,

$$a_{k,s,2} \leq T^{-1},$$

for any $k \in \{2, \dots, K\}$ and $T \geq 2$.

Proof. We will bound the probability of $G_{1,s}^c$ conditioning on $\mathcal{H}_{1,s}$ to prove this lemma. Note that

$$\begin{aligned} &\mathbb{E}_{\omega,\omega',\omega''} \left[\sum_{i=1}^s \left[(1+2\lambda)R_{1,i} - (\overline{R}_{1,s} + \lambda)\right]\omega_i + \sum_{i=1}^s \lambda(1 + \lambda - \overline{R}_{1,s})\omega_i' - \sum_{i=1}^s \lambda(\overline{R}_{1,s} + \lambda)\omega_i'' \right] \\ =& s\big((1+2\lambda)\overline{R}_{1,s} - (\overline{R}_{1,s} + \lambda) + \lambda(1 + \lambda - \overline{R}_{1,s}) - \lambda(\overline{R}_{1,s} + \lambda)\big) = 0, \end{aligned}$$

we can bound the tail probability $\mathbb{P}(G_{1,s}^c \mid \mathcal{H}_{1,s})$ by

$$\begin{aligned} &\mathbb{P}(G_{1,s}^c \mid \mathcal{H}_{1,s}) \\ =&\mathbb{P}\Big(|\overline{Y}_{1,s} - \overline{R}_{1,s}^*| > C_2 \Delta_k \mid \mathcal{H}_{1,s}\Big) \\ =&2\mathbb{P}_{\omega,\omega',\omega''} \left\{ \left| \frac{\sum_{i=1}^s \left[(1+2\lambda)R_{1,i} - (\overline{R}_{1,s} + \lambda)\right]\omega_i + \sum_{i=1}^s \lambda(1 + \lambda - \overline{R}_{1,s})\omega_i' - \sum_{i=1}^s \lambda(\overline{R}_{1,s} + \lambda)\omega_i''}{(1+2\lambda)(\sum_{i=1}^s \omega_i + \lambda \sum_{i=1}^s \omega_i' + \lambda \sum_{i=1}^s \omega_i'')} \right| > C_2 \Delta_k \right\} \\ \overset{\text{by Lemma E.3}}{\leq}& 2\mathbb{P}_{\omega,\omega',\omega''} \left\{ \sum_{i=1}^s \left[(1+2\lambda)R_{1,i} - (\overline{R}_{1,s} + \lambda)\right]\omega_i + \sum_{i=1}^s \lambda(1 + \lambda - \overline{R}_{1,s})\omega_i' \right. \\ & \left. - \sum_{i=1}^s \lambda(\overline{R}_{1,s} + \lambda)\omega_i'' - C_2(1+2\lambda)\Delta_k \left[\sum_{i=1}^s \omega_i + \lambda \sum_{i=1}^s \omega_i' + \lambda \sum_{i=1}^s \omega_i'' \right] > 0 \right\} \\ & + \mathbb{P}_{\omega,\omega',\omega''} \left\{ \sum_{i=1}^s \omega_i + \lambda \sum_{i=1}^s \omega_i' + \lambda \sum_{i=1}^s \omega_i'' < 0 \right\} = 2I + II, \end{aligned} \tag{26}$$

where

$$\begin{aligned} I = \mathbb{P}_{\omega,\omega',\omega''} &\left\{ \sum_{i=1}^s \left[(1+2\lambda)R_{1,i} - (\overline{R}_{1,s} + \lambda)\right]\omega_i + \sum_{i=1}^s \lambda(1 + \lambda - \overline{R}_{1,s})\omega_i' \right. \\ & \left. - \sum_{i=1}^s \lambda(\overline{R}_{1,s} + \lambda)\omega_i'' - C_2(1+2\lambda)\Delta_k \left[\sum_{i=1}^s \omega_i + \lambda \sum_{i=1}^s \omega_i' + \lambda \sum_{i=1}^s \omega_i'' \right] > 0 \right\}, \end{aligned}$$

and

$$II = \mathbb{P}_{\omega,\omega',\omega''} \left\{ \sum_{i=1}^s \omega_i + \lambda \sum_{i=1}^s \omega_i' + \lambda \sum_{i=1}^s \omega_i'' < 0 \right\}.$$

To bound $\mathbb{P}(G_{1,s}^c \mid \mathcal{H}_{1,s})$, it is sufficient to bound I and II separately. We first examine II. Write

$$\begin{aligned}
II &= \mathbb{P}_{\omega,\omega',\omega''}\left(\sum_{i=1}^{s}\omega_i + \lambda\sum_{i=1}^{s}\omega_i' + \lambda\sum_{i=1}^{s}\omega_i'' < 0\right) \\
&= \mathbb{P}_{\omega,\omega',\omega''}\left(\frac{\sum_{i=1}^{s}(\omega_i - 1) + \lambda\sum_{i=1}^{s}(\omega_i' - 1) + \lambda\sum_{i=1}^{s}(\omega_i'' - 1)}{\sqrt{s\sigma_\omega^2(1+2\lambda^2)}} < -\frac{(1+2\lambda)s}{\sqrt{s\sigma_\omega^2(1+2\lambda^2)}}\right) \\
&= \mathbb{P}_{\omega,\omega',\omega''}\left(Z < -\frac{(1+2\lambda)}{\sigma_\omega\sqrt{1+2\lambda^2}} \times \sqrt{s}\right) \\
&\overset{\text{by Lemma E.1}}{\leq} \frac{1}{2}\exp\left\{-\frac{s(1+2\lambda)^2}{2\sigma_\omega^2(1+2\lambda^2)}\right\}.
\end{aligned}$$

For studying I, we define the following functions of $\{R_{1,i}\}_{i=1}^s$:

$$f\big(\{R_{1,i}\}_{i=1}^s\big) = f_1\big(\{R_{1,i}\}_{i=1}^s\big) + f_2\big(\{R_{1,i}\}_{i=1}^s\big) - f_3\big(\{R_{1,i}\}_{i=1}^s\big),$$

where

$$f_1\big(\{R_{1,i}\}_{i=1}^s\big) = \sum_{i=1}^{s}\big[(R_{1,i} - \overline{R}_{1,s}) + \lambda(2\overline{R}_{1,i} - 1) - (1+2\lambda)C_2\Delta_k\big]\omega_i,$$

$$f_2\big(\{R_{1,i}\}_{i=1}^s\big) = \sum_{i=1}^{s}\big[\lambda(\lambda + 1 - \overline{R}_{1,s}) - (1+2\lambda)C_2\Delta_k\big]\omega_i',$$

and

$$f_3\big(\{R_{1,i}\}_{i=1}^s\big) = \sum_{i=1}^{s}\big[\lambda(\overline{R}_{1,s} + \lambda) + (1+2\lambda)C_2\Delta_k\big]\omega_i''.$$

Then we can write $I = \mathbb{P}_{\omega,\omega',\omega'}(f_1 + f_2 - f_3 > 0)$. Given that f_1, f_2, and f_3 are mutually independent conditioning on $\mathcal{H}_{1,s}$, the expectation is

$$\mathbb{E}\big[f_1 + f_2 - f_3 \mid \mathcal{H}_{1,s}\big] = -3C_2(1+2\lambda)\Delta_k s,$$

and the variance is

$$\begin{aligned}
&\operatorname{var}\big(f_1 + f_2 - f_3 \mid \mathcal{H}_{1,s}\big) \\
=&\sigma_\omega^2\Bigg[\sum_{i=1}^{s}\big[(R_{1,i} - \overline{R}_{1,s}) + \lambda(2\overline{R}_{1,i} - 1) - (1+2\lambda)C_2\Delta_k\big]^2 \\
&\qquad + \sum_{i=1}^{s}\big[\lambda(\lambda + 1 - \overline{R}_{1,s}) - (1+2\lambda)C_2\Delta_k\big]^2 + \sum_{i=1}^{s}\big[\lambda(\overline{R}_{1,s} + \lambda) + (1+2\lambda)C_2\Delta_k\big]^2\Bigg] \\
=&\sigma_\omega^2\Big[V_1 + V_2 + V_3\Big],
\end{aligned} \tag{27}$$

where

$$V_1 = \sum_{i=1}^{s}\Big\{\big[(R_{1,i} - \overline{R}_{1,s}) + \lambda(2\overline{R}_{1,s} - 1)\big]^2 + \lambda^2(\lambda + 1 - \overline{R}_{1,s})^2 + \lambda^2(\overline{R}_{1,s} + \lambda)^2\Big\},$$

$$V_2 = \sum_{i=1}^{s}\Big\{-2(1+2\lambda)C_2\Delta_k\big[(R_{1,i} - \overline{R}_{1,s}) + \lambda(2\overline{R}_{1,s} - 1)\big] - 2(1+2\lambda)C_2\Delta_k\lambda(\lambda + 1 - \overline{R}_{1,s}) + 2(1+2\lambda)C_2\Delta_k(\overline{R}_{1,s} + \lambda)\Big\},$$

and

$$V_3 = \sum_{i=1}^{s}\big[(1+2\lambda)^2C_2^2\Delta_k^2 + (1+2\lambda)C_2^2\Delta_k^2 + (1+2\lambda)C_2^2\Delta_k^2\big].$$

For bounding the conditional variance above, we will calculate its components as follows.

$$\begin{aligned}
V_1 &= \sum_{i=1}^{s}\left[(R_{1,i}-\overline{R}_{1,s})+\lambda(2\overline{R}_{1,i}-1)\right]^2 + s\lambda^2(\lambda+1-\overline{R}_{1,s})^2 + s\lambda^2(\overline{R}_{1,s}+\lambda)^2 \\
&= \sum_{i=1}^{s}(R_{1,i}-\overline{R}_{1,s})^2 + 2s\lambda^2\big(3\overline{R}_{1,s}^2 - 3\overline{R}_{1,s}+\lambda^2+\lambda+1\big) \\
&= \sum_{i=1}^{s}(R_{1,i}-\mu_1)^2 + (6\lambda^2-1)s(\overline{R}_{1,s}-\mu_1)^2 + 6s\lambda^2(2\mu_1-3)(\overline{R}_{1,s}-\mu_1) + 2s\lambda^2\big[\lambda^2+\lambda+1+3\mu_1(\mu_1-1)\big] \\
&\overset{\text{Cauchy inequality}}{\leq} 6\lambda^2\sum_{i=1}^{s}(R_{1,i}-\mu_1)^2 + 6s\lambda^2(2\mu_1-1)(\overline{R}_{1,s}-\mu_1) + 2s\lambda^2\big(\lambda^2+\lambda+1-\mu_1(1-\mu_1)\big) \\
&= 6\lambda^2\sum_{i=1}^{s}(R_{1,i}-\mu_1)^2 + 6s\lambda^2(2\mu_1-1)(\overline{R}_{1,s}-\mu_1) + 2s\lambda^2\left(\lambda^2+\lambda+1-\mu_1(1-\mu_1)\right),
\end{aligned} \tag{28}$$

$$\begin{aligned}
V_2 &= -2(1+2\lambda)C_2\Delta_k\left[\sum_{i=1}^{s}\big((R_{1,i}-\overline{R}_{1,s})+\lambda(2\overline{R}_{1,i}-1)\big) + s\lambda(\lambda+1-\overline{R}_{1,s}) - s\lambda(\overline{R}_{1,s}+\lambda)\right] \\
&= -2s(1+2\lambda)C_2\Delta_k\left[\big(\overline{R}_{1,s}-\overline{R}_{1,s})+\lambda(2\overline{R}_{1,i}-1)+\lambda(\lambda+1-\overline{R}_{1,s})\big) - \lambda(\overline{R}_{1,s}+\lambda)\right] = 0,
\end{aligned}$$

and

$$V_3 = 3s(1+2\lambda)^2C_2^2\Delta_k^2.$$

Therefore, the conditional variance in (27) is bounded by

$$\begin{aligned}
\sigma_\omega^{-2}\operatorname{var}\big(f_1+f_2-f_3 \mid \mathcal{H}_{1,s}\big) &= V_1+V_2+V_3 \\
&\leq \underbrace{6\lambda^2\sum_{i=1}^{s}(R_{1,i}-\mu_1)^2 + 6s\lambda^2(2\mu_1-1)\left(\overline{R}_{1,s}-\mu_1\right)}_{\text{the random part}} \\
&\qquad + \underbrace{2s\lambda^2\left(\lambda^2+\lambda+1-3\mu_1(1-\mu_1)\right) + 3s(1+2\lambda)^2C_2^2\Delta_k^2}_{\text{the determined part}} \\
&= \mathfrak{R}_1+\mathfrak{R}_2+\mathfrak{D},
\end{aligned} \tag{29}$$

where

$$\mathfrak{R}_1 = 6\lambda^2\sum_{i=1}^{s}(R_{1,i}-\mu_1)^2, \qquad \mathfrak{R}_2 = 6s\lambda^2(2\mu_1-1)\left(\overline{R}_{1,s}-\mu_1\right) + 2s\lambda^2\left(\lambda^2+\lambda+1-3\mu_1(1-\mu_1)\right),$$

and

$$\mathfrak{D} = 3s(1+2\lambda)^2C_2^2\Delta_k^2.$$

It is clear that both $\mathfrak{R}_1$ and $\mathfrak{R}_2$ are strictly positive. Additionally, it can be shown with high probability that $\mathfrak{R}_2$ is non-positive. Indeed, note that $0 \leq 3\mu_1(1-\mu_1) \leq \frac{3}{4}$, we have

$$\begin{aligned}
\mathbb{P}(\mathfrak{R}_2>0) &= \mathbb{P}\left(3(2\mu_1-1)\left(\overline{R}_{1,s}-\mu_1\right) > \lambda^2+\lambda+1-3\mu_1(1-\mu_1)\right) \\
&\leq \mathbb{P}\left(\left|3(2\mu_1-1)\left(\overline{R}_{1,s}-\mu_1\right)\right| > \left|\lambda^2+\lambda+1-3\mu_1(1-\mu_1)\right|\right)\mathbb{I}\left(\mu_1\neq\frac{1}{2}\right) + \mathbb{P}(\varnothing)\times\mathbb{I}\left(\mu_1=\frac{1}{2}\right) \\
&\overset{0\leq 3\mu_1(1-\mu_1)\leq\frac{3}{4}}{\leq} \mathbb{P}\left(\left|\overline{R}_{1,s}-\mu_1\right| > \frac{\lambda^2+\lambda+1/4}{|3(2\mu_1-1)|}\right)\mathbb{I}\left(\mu_1\neq\frac{1}{2}\right) + 0\times\mathbb{I}\left(\mu_1=\frac{1}{2}\right) \\
&\overset{\text{by sub-Gaussian inequality}}{\leq} \exp\left\{-\frac{(\lambda^2+\lambda+1/4)^2 s}{18\sigma_\omega^2(2\mu_1-1)^2}\right\}\times\mathbb{I}\left(\mu_1\neq\frac{1}{2}\right) + 0\times\mathbb{I}\left(\mu_1=\frac{1}{2}\right).
\end{aligned} \tag{30}$$

Then by applying Lemma E.1 again,

$$
\begin{aligned}
\mathbb{P}(G_{1,s}^c \mid \mathcal{H}_{1,s}) &= 2I + II \\
&= 2\mathbb{P}\left(\frac{f_1 + f_2 - f_3 - \mathbb{E}\big(f_1 + f_2 - f_3 \mid \mathcal{H}_{1,s}\big)}{\sqrt{\operatorname{var}\big(f_1 + f_2 - f_3 \mid \mathcal{H}_{1,s}\big)}} > -\frac{\mathbb{E}\big(f_1 + f_2 - f_3 \mid \mathcal{H}_{1,s}\big)}{\sqrt{\operatorname{var}\big(f_1 + f_2 - f_3 \mid \mathcal{H}_{1,s}\big)}}\right) \\
&\quad + \frac{1}{2}\exp\left\{-\frac{s(1+2\lambda)^2}{2\sigma_\omega^2(1+2\lambda^2)}\right\} \\
&\leq 2\exp\left\{-\frac{\mathbb{E}^2\big(f_1 + f_2 - f_3 \mid \mathcal{H}_{1,s}\big)}{2\operatorname{var}\big(f_1 + f_2 - f_3 \mid \mathcal{H}_{1,s}\big)}\right\} + \frac{1}{2}\exp\left\{-\frac{s(1+2\lambda)^2}{2\sigma_\omega^2(1+2\lambda^2)}\right\}.
\end{aligned}
$$

Therefore, combining with (29), $a_{k,s,2}$ will be bounded by

$$
\begin{aligned}
a_{k,s,2} &\leq \mathbb{E}\Big[T\mathbb{I}(G_{1,s}^c)\Big] = T\mathbb{E}\Big[\mathbb{P}(G_{1,s}^c \mid \mathcal{H}_{1,s})\Big] \\
&\overset{\text{by (26)}}{\leq} \leq T\big[2I + II\big] \\
&\leq 2T\mathbb{E}\exp\left\{-\frac{\mathbb{E}^2\big(f_1 + f_2 - f_3 \mid \mathcal{H}_{1,s}\big)}{2\operatorname{var}\big(f_1 + f_2 - f_3 \mid \mathcal{H}_{1,s}\big)}\right\} + \frac{T}{2}\exp\left\{-\frac{s(1+2\lambda)^2}{2\sigma_\omega^2(1+2\lambda^2)}\right\} \\
&= T\mathbb{E}\exp\left\{-\frac{\mathbb{E}^2\big(f_1 + f_2 - f_3 \mid \mathcal{H}_{1,s}\big)}{2\operatorname{var}\big(f_1 + f_2 - f_3 \mid \mathcal{H}_{1,s}\big)}\right\}\big(\mathbb{I}(\mathfrak{R}_2 \leq 0) + \mathbb{I}(\mathfrak{R}_2 > 0)\big) + \frac{T}{2}\exp\left\{-\frac{s(1+2\lambda)^2}{2\sigma_\omega^2(1+2\lambda^2)}\right\} \\
&\overset{\text{by the decomposition (29)}}{\leq} T\mathbb{E}\exp\left\{-\frac{\mathbb{E}^2\big(f_1 + f_2 - f_3 \mid \mathcal{H}_{1,s}\big)}{2\sigma_\omega^2\big(\mathfrak{R}_1 + \mathfrak{R}_2 + \mathfrak{D}\big)}\right\}\mathbb{I}(\mathfrak{R}_2 \leq 0) + T\mathbb{P}(\mathfrak{R}_2 > 0) + \frac{T}{2}\exp\left\{-\frac{s(1+2\lambda)^2}{2\sigma_\omega^2(1+2\lambda^2)}\right\} \\
&\leq T\mathbb{E}\exp\left\{-\frac{\mathbb{E}^2\big(f_1 + f_2 - f_3 \mid \mathcal{H}_{1,s}\big)}{2\sigma_\omega^2\big(\mathfrak{R}_1 + \mathfrak{D}\big)}\right\} + T\mathbb{P}(\mathfrak{R}_2 > 0) + \frac{T}{2}\exp\left\{-\frac{s(1+2\lambda)^2}{2\sigma_\omega^2(1+2\lambda^2)}\right\}.
\end{aligned}
$$

Recall $R_{1,i} - \mu_1 \sim \mathrm{subG}(\sigma_1^2)$, and then $\xi_i := (R_{1,i} - \mu_1)^2 - \operatorname{var}(R_{1,i}) \sim \mathrm{subE}(8\sqrt{2}\sigma_1^2)$. Therefore, for furthermore bounding $\mathfrak{R}_1 + \mathfrak{D}$, we have

$$
\begin{aligned}
\mathfrak{R}_1 + \mathfrak{D} &= 6\lambda^2\left[\sum_{i=1}^{s}(R_{1,i} - \mu_1)^2 - s\operatorname{var}(R_{1,i})\right] + \left[s\operatorname{var}(R_{1,i}) + 3s(1+2\lambda)^2 C_2^2\Delta_k^2\right] \\
&= 6\lambda^2\sum_{i=1}^{s}\xi_i + s\left[6\lambda^2\operatorname{var}(R_{1,i}) + 3(1+2\lambda)^2 C_2^2\Delta_k^2\right] \\
&\overset{\text{by } \operatorname{var}(R_{1,i}) \leq \sigma_1^2}{\leq} 6\lambda^2\sum_{i=1}^{s}\xi_i + 3s\left[2\lambda^2\sigma_1^2 + (1+2\lambda)^2 C_2^2\Delta_k^2\right] \\
&= 3s\left[2\lambda^2\overline{\xi} + 2\lambda^2\sigma_1^2 + (1+2\lambda)^2 C_2^2\Delta_k^2\right],
\end{aligned}
\tag{31}
$$

where $\overline{\xi} = \frac{1}{s}\sum_{i=1}^{s}\xi_i$. Then

$$\begin{aligned}
a_{k,s,2} &\leq T\mathbb{E}\exp\left\{-\frac{\mathbb{E}^2\big(f_1+f_2-f_3 \mid \mathcal{H}_{1,s}\big)}{2\sigma_\omega^2\big(\mathfrak{R}_1+\mathfrak{D}\big)}\right\} + T\mathbb{P}(\mathfrak{R}_2>0) + \frac{T}{2}\exp\left\{-\frac{s(1+2\lambda)^2}{2\sigma_\omega^2(1+2\lambda^2)}\right\} \\
&\overset{\text{by (31) and (30)}}{\leq} T\mathbb{E}\exp\left\{-\frac{9C_2^2(1+2\lambda)^2\Delta_k^2 s^2}{6\sigma_\omega^2 s\left[2\lambda^2\overline{\xi}+2\lambda^2\sigma_1^2+(1+2\lambda)^2C_2^2\Delta_k^2\right]}\right\} \\
&\quad + T\left[\exp\left\{-\frac{(\lambda^2+\lambda+1/4)^2 s}{18\sigma_\omega^2(2\mu_1-1)^2}\right\}\times\mathbb{I}\left(\mu_1\neq\frac{1}{2}\right)+0\times\mathbb{I}\left(\mu_1=\frac{1}{2}\right)\right] \\
&\quad + \frac{T}{2}\exp\left\{-\frac{s(1+2\lambda)^2}{2\sigma_\omega^2(1+2\lambda^2)}\right\} \\
&= T\mathbb{E}\exp\left\{-\frac{3C_2^2(1+2\lambda)^2\Delta_k^2 s}{2\sigma_\omega^2\left[2\lambda^2\overline{\xi}+2\lambda^2\sigma_1^2+(1+2\lambda)^2C_2^2\Delta_k^2\right]}\right\} \\
&\quad + T\left[\exp\left\{-\frac{(\lambda^2+\lambda+1/4)^2 s}{18\sigma_\omega^2(2\mu_1-1)^2}\right\}\times\mathbb{I}\left(\mu_1\neq\frac{1}{2}\right)+0\times\mathbb{I}\left(\mu_1=\frac{1}{2}\right)\right] \\
&\quad + \frac{T}{2}\exp\left\{-\frac{s(1+2\lambda)^2}{2\sigma_\omega^2(1+2\lambda^2)}\right\}.
\end{aligned} \tag{32}$$

The next step is to apply Lemma E.4. Let

$$\mathfrak{a}_1 = \frac{4\sigma_\omega^2\lambda^2}{3C_2^2(1+2\lambda)^2\Delta_k^2}, \qquad \mathfrak{b}_1 = \frac{2\sigma_\omega^2\left[2\lambda^2\sigma_1^2+(1+2\lambda)^2C_2^2\Delta_k^2\right]}{3C_2^2(1+2\lambda)^2\Delta_k^2}, \qquad \lambda_i \equiv 8\sqrt{2}\sigma_1^2, \qquad i\in[s],$$

and $\Omega_1 = (s^{-1}\sum_{i=1}^{s}\lambda_i^2)^{1/2} = 8\sqrt{2}\sigma_1^2$, then Lemma E.4 gives

$$\begin{aligned}
&\mathbb{E}\exp\left\{-\frac{3C_2^2(1+2\lambda)^2\Delta_k^2 s}{2\sigma_\omega^2\left[2\lambda^2\overline{\xi}+2\lambda^2\sigma_1^2+(1+2\lambda)^2C_2^2\Delta_k^2\right]}\right\} \\
&= \mathbb{E}\exp\left\{-\frac{s}{\frac{4\sigma_\omega^2\lambda^2}{3C_2^2(1+2\lambda)^2\Delta_k^2}\overline{\xi}+\frac{2\sigma_\omega^2\left[2\lambda^2\sigma_1^2+(1+2\lambda)^2C_2^2\Delta_k^2\right]}{3C_2^2(1+2\lambda)^2\Delta_k^2}}\right\} = \mathbb{E}\exp\left\{-\frac{s}{\mathfrak{a}_1\overline{\xi}+\mathfrak{b}_1}\right\} \\
&\leq \left[1+\frac{\sqrt{\pi\mathfrak{b}_1}}{2}+\frac{\sqrt{2\pi}\Omega_1\mathfrak{a}_1}{2\mathfrak{b}_1^2}+\frac{2\mathfrak{a}_1}{\mathfrak{b}_1}\right]\exp\left\{-\frac{s}{\sqrt{2\Omega_1\mathfrak{a}_1}\vee(\mathfrak{b}_1+\Omega_1\mathfrak{a}_1)}\right\} \\
&\overset{\text{by } \sqrt{2\Omega_1\mathfrak{a}_1}\vee(\mathfrak{b}_1+\Omega_1\mathfrak{a}_1)\leq\mathfrak{b}_1+\Omega_1\mathfrak{a}_1+\Omega_1+\mathfrak{a}_1}{\leq} \left[1+\frac{\sqrt{\pi\mathfrak{b}_1}}{2}+\frac{\sqrt{2\pi}\Omega_1\mathfrak{a}_1}{2\mathfrak{b}_1^2}+\frac{2\mathfrak{a}_1}{\mathfrak{b}_1}\right] s\exp\left\{-\frac{s}{\Omega_1+\mathfrak{a}_1+\mathfrak{b}_1+\Omega_1\mathfrak{a}_1}\right\}.
\end{aligned} \tag{33}$$

Hence, by taking

$$\begin{aligned}
s &\geq \left[\Omega_1+\mathfrak{a}_1+\mathfrak{b}_1+\Omega_1\mathfrak{a}_1\right]\left(\log^{-1}2\times\log\left\{3\left[1+\frac{\sqrt{\pi\mathfrak{b}_1}}{2}+\frac{\sqrt{2\pi}\Omega_1\mathfrak{a}_1}{2\mathfrak{b}_1^2}+\frac{2\mathfrak{a}_1}{\mathfrak{b}_1}\right]\right\}+3\right)\log T \\
&\geq \left[\Omega_1+\mathfrak{a}_1+\mathfrak{b}_1+\Omega_1\mathfrak{a}_1\right]\log\left\{3\left[1+\frac{\sqrt{\pi\mathfrak{b}_1}}{2}+\frac{\sqrt{2\pi}\Omega_1\mathfrak{a}_1}{2\mathfrak{b}_1^2}+\frac{2\mathfrak{a}_1}{\mathfrak{b}_1}\right]T^3\right\},
\end{aligned}$$

we have

$$\mathbb{E}\exp\left\{-\frac{3C_2^2(1+2\lambda)^2\Delta_k^2 s}{2\sigma_\omega^2\left[2\lambda^2\overline{\xi}+2\lambda^2\sigma_1^2+(1+2\lambda)^2C_2^2\Delta_k^2\right]}\right\} \leq \frac{s}{3T^3} \leq \frac{1}{3T^2}.$$

Similarly, the inequality

$$T\left[\exp\left\{-\frac{(\lambda^2+\lambda+1/4)^2 s}{18\sigma_\omega^2(2\mu_1-1)^2}\right\}\times\mathbb{I}\left(\mu_1\neq\frac{1}{2}\right)+0\times\mathbb{I}\left(\mu_1=\frac{1}{2}\right)\right] \leq \frac{1}{3T},$$

and

$$\frac{T}{2}\exp\left\{-\frac{s(1+2\lambda)^2}{2\sigma_\omega^2(1+2\lambda^2)}\right\}\leq\frac{1}{3T}$$

have the solution

$$s\geq\frac{18\sigma_\omega^2(2\mu_1-1)^2}{(\lambda^2+\lambda+1/4)^2}\left[\log 3+2\log T\right].$$

and

$$s\geq\frac{2\sigma_\omega^2(1+2\lambda^2)}{(1+2\lambda)^2}\left[\log(3/2)+2\log T\right],$$

respectively. Therefore, we have $a_{k,s,2}\leq T^{-1}$ for any $s\geq s_{a,k,2}(T)$ by taking

$$s_{a,k,2}(T)=\Bigg[\left[\Omega_1+\mathfrak{a}_1+\mathfrak{b}_1+\Omega_1\mathfrak{a}_1\right]\left(\frac{1}{3}\log^{-1}2\times\log\left\{3\left[1+\frac{\sqrt{\pi\mathfrak{b}_1}}{2}+\frac{\sqrt{2\pi}\Omega_1\mathfrak{a}_1}{2\mathfrak{b}_1^2}+\frac{2\mathfrak{a}_1}{\mathfrak{b}_1}\right]\right\}+1\right)$$
$$\vee\frac{18\sigma_\omega^2(2\mu_1-1)^2}{(\lambda^2+\lambda+1/4)^2}\vee\frac{2\sigma_\omega^2(1+2\lambda^2)}{(1+2\lambda)^2}\Bigg]\times 3\log T$$

for any $T\geq 2$. □

Lemma D.4 (Bounding $a_{k,s,3}$ at (5))**.** *Take*

$$s_{a,k,3}(T)=\Bigg\{\Bigg[\left(\Omega_1+\mathfrak{a}_2+\mathfrak{b}_2+\Omega_1\mathfrak{a}_2\right)\left(\frac{1}{3}\log^{-1}2\times\log\left\{3\left[1+\frac{\sqrt{\pi\mathfrak{b}_2}}{2}+\frac{\sqrt{2\pi}\Omega_1\mathfrak{a}_2}{2\mathfrak{b}_2^2}+\frac{2\mathfrak{a}_2}{\mathfrak{b}_2}\right]\right\}+1\right)\Bigg]$$
$$\vee\frac{18\sigma_\omega^2(2\mu_1-1)^2}{(\lambda^2+\lambda+1/4)^2}\vee\frac{2\sigma_\omega^2(1+2\lambda^2)}{(1+2\lambda)^2}\Bigg\}\times 3\log T$$

where

$$\mathfrak{a}_2=\frac{\sigma_\omega^2\lambda^2}{3(\lambda-1)^2C_1^2\Delta_k^2},\qquad\mathfrak{b}_2=\frac{\sigma_\omega^2\left[2\lambda^2\sigma_1^2+25(1+2\lambda)^2C_1^2\Delta_k^2\right]}{6(\lambda-1)^2C_1^2\Delta_k^2},\qquad\text{and}\qquad\Omega_1=8\sqrt{2}\sigma_1^2.$$

Then for any $s\geq s_{a,k,3}(T)$ and $\lambda>1$,

$$a_{k,s,3}\leq T^{-1}.$$

for any $k\in\{2,\ldots,K\}$ and $T\geq 2$.

Proof. Unlike the proofs for bounding $a_{k,s,1}$ and $a_{k,s,2}$, which involve controlling the probability of *bad* events, the definition of $a_{k,s,3}$ is based on *good* events instead. Therefore, we require a different technique to handle $a_{k,s,3}$. Observe that

$$\begin{aligned}\{N_{1,s}(\tau_k)<T^{-1}\}&\supseteq\{N_{1,s}(\tau_k)<(T^2-1)^{-1}\}\\&\overset{\text{by }(T^2-1)^{-1}\leq T^{-1}}{=}\left\{Q_{1,s}\left(\tau_k\right)^{-1}<1+\left(T^2-1\right)^{-1}\right\}\\&=\left\{Q_{1,s}(\tau_k)^{-1}<(1-T^{-2})^{-1}\right\}=\left\{\left[1-Q_{1,s}(\tau_k)\right]<T^{-2}\right\},\end{aligned}\tag{34}$$

and $a_{k,s,3}\leq\mathbb{E}\big[N_{1,s}(\tau_k)\mathbb{I}(A_{1,s}\cap G_{1,s})\big]$. Thus, as suggested by Wang et al. (2020), we can bound $a_{k,s,3}\leq T^{-1}$ for $s\geq s_{a,k,3}$ by finding $s\geq s_{a,k,3}$ such that $\left\{[1-Q_{1,s}(\tau_k)]<T^{-2}\right\}$ holds on the event $\{A_{1,s}\cap G_{1,s}\}$.

We can express the term

$$\left[1-Q_{1,s}(\tau_k)\right]\mathbb{I}(A_{1,s})\mathbb{I}(G_{1,s})=\mathbb{P}\big(\overline{Y}_{1,s}-\overline{R}_{1,s}^*\leq\Gamma_1\,|\,\mathcal{H}_{1,s}\big)\mathbb{I}(A_{1,s}\cap G_{1,s}),$$

where we define $\Gamma_1:=\Gamma_1(\tau_1)=\tau_1-\overline{R}_{1,s}^*$ as the difference between $\overline{Y}_{1,s}$ and $\overline{R}_{1,s}^*$. On the event $\{A_{1,s}\cap G_{1,s}\}$, Γ_1 can be bounded within an interval. To see this,

$$\begin{aligned}A_{1,s}&=\left\{\overline{R}_{1,s}^*-\frac{\mu_1+\lambda}{1+2\lambda}>-C_1\Delta_k\right\}\cap\left\{\overline{R}_{k,s}^*-\frac{\mu_k+\lambda}{1+2\lambda}<C_1\Delta_k\right\}\\&=\left\{\Gamma_1<C_1\Delta_k-\frac{\mu_1+\lambda}{1+2\lambda}+\tau_1\right\}\cap\left\{\Gamma_1>-C_2\Delta_k-\frac{\mu_1+\lambda}{1+2\lambda}+\tau_1\right\}\\&=\left\{\tau_1-C_1\Delta_k-\frac{\mu_1+\lambda}{1+2\lambda}<\Gamma_1<\tau_1+C_1\Delta_k-\frac{\mu_1+\lambda}{1+2\lambda}\right\},\end{aligned}$$

and

$$\begin{aligned}G_{1,s} &= \big\{-C_2\Delta_k \le \overline{Y}_{1,s} - \overline{R}^*_{1,s} \le C_2\Delta_k\big\} \\ &= \big\{\tau_1 - C_2\Delta_k \le \overline{Y}_{1,s} + \Gamma_1 \le \tau_1 + C_2\Delta_k\big\} \\ &= \big\{\tau_1 - C_2\Delta_k - \overline{Y}_{1,s} \le \Gamma_1 \le \tau_1 + C_2\Delta_k - \overline{Y}_{1,s}\big\} \\ &\overset{\text{by } -C_2\Delta_k + \overline{R}^*_{1,s} \le \overline{Y}_{1,s} \le C_2\Delta_k + \overline{R}^*_{1,s}}{\subseteq} \big\{\tau_1 - 2C_2\Delta_k - \overline{R}^*_{1,s} \le \Gamma_1 \le \tau_1 + 2C_2\Delta_k - \overline{R}^*_{1,s}\big\}.\end{aligned}$$

Let us combine the previous two results, given by

$$C_1 = 2C_2, \tag{35}$$

which yields

$$\begin{aligned}&A_{1,s} \cap G_{1,s} \\ &\subseteq \left\{\tau_1 - C_1\Delta_k - \overline{R}^*_{1,s} \wedge \frac{\mu_1+\lambda}{1+2\lambda} \le \Gamma_1 \le \tau_1 + C_1\Delta_k - \overline{R}^*_{1,s} \vee \frac{\mu_1+\lambda}{1+2\lambda}\right\} \\ &\overset{\text{by } \frac{\mu_1+\lambda}{1+2\lambda} - C_1\Delta_k < \overline{R}^*_{1,s} < \frac{\mu_1+\lambda}{1+2\lambda} + C_1\Delta_k}{\subseteq} \left\{\tau_1 - 2C_1\Delta_k - \frac{\mu_1+\lambda}{1+2\lambda} \le \Gamma_1 \le \tau_1 + 2C_1\Delta_k - \frac{\mu_1+\lambda}{1+2\lambda}\right\}.\end{aligned} \tag{36}$$

Now let

$$\tau_1 := \frac{\mu_1+\lambda}{1+2\lambda} - \frac{6\lambda C_1\Delta_k}{1+2\lambda}.$$

This choice of τ_1 satisfies (13) as $\tau_1 \le \frac{\mu_1+\lambda}{1+2\lambda}$. We can furthermore reduce on $A_{1,s} \cap G_{1,s}$ that

$$\begin{aligned}\Gamma_1 &\le \tau_1 + 2C_1\Delta_k - \frac{\mu_1+\lambda}{1+2\lambda} \\ &= 2C_1\Delta_k - \frac{6\lambda C_1\Delta_k}{1+2\lambda} = -2C_1\Delta_k\frac{\lambda-1}{2\lambda+1} \\ &\overset{\text{by } \lambda\ge 1}{\le} -\frac{2C_1(\lambda-1)\Delta_k}{2\lambda+1} < 0,\end{aligned}$$

and

$$\begin{aligned}0 > \Gamma_1 &\ge \tau_1 - 2C_1\Delta_k - \frac{\mu_1+\lambda}{1+2\lambda} \\ &= -2C_1\Delta_k - \frac{6\lambda C_1\Delta_k}{1+2\lambda} = -2C_1\Delta_k\left[1 + \frac{3\lambda}{1+2\lambda}\right] \\ &\overset{\text{by } \frac{3\lambda}{1+2\lambda}\le\frac{3}{2}}{\ge} -2C_1\Delta_k \times \frac{5}{2} = -5C_1\Delta_k.\end{aligned}$$

Thus, we obtain $-\Gamma_1 \in \left(\frac{2C_1(\lambda-1)\Delta_k}{2\lambda+1}, 5C_1\Delta_k\right) \subseteq (0,+\infty)$. Return to the quantity we are interested in, $\big[1 - Q_{1,s}(\tau_k)\big]\mathbb{I}(A_{1,s})\mathbb{I}(G_{1,s})$. We can express it as follows:

$$\begin{aligned}&\big[1 - Q_{1,s}(\tau_k)\big]\mathbb{I}(A_{1,s})\mathbb{I}(G_{1,s}) \\ =&\mathbb{P}\big(\overline{Y}_{1,s} - \overline{R}^*_{1,s} \le \Gamma_1 \mid \mathcal{H}_{1,s}\big)\mathbb{I}(A_{1,s} \cap G_{1,s}) \\ =&\mathbb{E}\Bigg[\mathbb{P}_{\omega,\omega',\omega''}\Bigg\{\sum_{i=1}^s (R_{1,i} - \overline{R}_{1,s})\omega_i + \lambda\sum_{i=1}^s (2R_{1,i}\omega_i - \overline{R}_{1,s}\omega_i' - \overline{R}_{1,s}\omega_i'') \\ &\qquad + \lambda\sum_{i=1}^s(\omega_i' - \omega_i) + \lambda^2\sum_{i=1}^s(\omega_i' - \omega_i'') \\ &\qquad - \Gamma_1(1+2\lambda)\left[\sum_{i=1}^s \omega_i + \lambda\sum_{i=1}^s\omega_i' + \lambda\sum_{i=1}^s\omega_i''\right] < 0\Bigg\}\mathbb{I}(A_{1,s} \cap G_{1,s})\Bigg].\end{aligned}$$

Similar to the technique we used to bound the conditional probability $\mathbb{P}(G_{1,s}^c \mid \mathcal{H}_{1,s})$, we just need to replace $C_2\Delta_k$ by $-\Gamma_1$ in (32) and obtain that

$$\begin{aligned}
&\left[1 - Q_{1,s}(\tau_k)\right]\mathbb{I}(A_{1,s})\mathbb{I}(G_{1,s}) \\
&\overset{\text{apply steps in (26)}}{\leq} T\mathbb{E}\left[\exp\left\{-\frac{9(-\Gamma_1)^2(1+2\lambda)^2 s^2}{6\sigma_\omega^2 s\left[2\lambda^2\overline{\xi} + 2\lambda^2\sigma_1^2 + (-\Gamma_1)^2(1+2\lambda)^2\right]}\right\}\mathbb{I}(A_{1,s} \cap G_{1,s})\right] \\
&\qquad + T\left[\exp\left\{-\frac{(\lambda^2+\lambda+1/4)^2 s}{18\sigma_\omega^2(2\mu_1-1)^2}\right\}\times\mathbb{I}\left(\mu_1 \neq \frac{1}{2}\right) + 0\times\mathbb{I}\left(\mu_1 = \frac{1}{2}\right)\right] \\
&\qquad + \frac{T}{2}\exp\left\{-\frac{s(1+2\lambda)^2}{2\sigma_\omega^2(1+2\lambda^2)}\right\} \\
&\overset{\text{by the bound of } -\Gamma_1}{\leq} T\mathbb{E}\exp\left\{-\frac{36(\lambda-1)^2C_1^2\Delta_k^2 s^2}{6\sigma_\omega^2 s\left[2\lambda^2\overline{\xi} + 2\lambda^2\sigma_1^2 + 25(1+2\lambda)^2C_1^2\Delta_k^2\right]}\right\} \\
&\qquad + T\left[\exp\left\{-\frac{(\lambda^2+\lambda+1/4)^2 s}{18\sigma_\omega^2(2\mu_1-1)^2}\right\}\times\mathbb{I}\left(\mu_1 \neq \frac{1}{2}\right) + 0\times\mathbb{I}\left(\mu_1 = \frac{1}{2}\right)\right] \\
&\qquad + \frac{T}{2}\exp\left\{-\frac{s(1+2\lambda)^2}{2\sigma_\omega^2(1+2\lambda^2)}\right\}.
\end{aligned} \tag{37}$$

Similarly, we define the following expressions

$$\mathfrak{a}_2 = \frac{\sigma_\omega^2\lambda^2}{3(\lambda-1)^2C_1^2\Delta_k^2}, \qquad \text{and} \qquad \mathfrak{b}_2 = \frac{\sigma_\omega^2\left[2\lambda^2\sigma_1^2 + 25(1+2\lambda)^2C_1^2\Delta_k^2\right]}{6(\lambda-1)^2C_1^2\Delta_k^2}.$$

Consider the value of s such that

$$s \geq \left[\Omega_1 + \mathfrak{a}_2 + \mathfrak{b}_2 + \Omega_1\mathfrak{a}_2\right]\left(\log^{-1}2 \times \log\left\{3\left[1 + \frac{\sqrt{\pi\mathfrak{b}_2}}{2} + \frac{\sqrt{2\pi}\Omega_1\mathfrak{a}_2}{2\mathfrak{b}_2^2} + \frac{2\mathfrak{a}_2}{\mathfrak{b}_2}\right]\right\} + 3\right)\log T.$$

By applying Lemma E.4 and following the steps in (33) again, we obtain

$$\begin{aligned}
&T\mathbb{E}\exp\left\{-\frac{36(\lambda-1)^2C_1^2\Delta_k^2 s^2}{6\sigma_\omega^2 s\left[2\lambda^2\overline{\xi} + 2\lambda^2\sigma_1^2 + 25(1+2\lambda)^2C_1^2\Delta_k^2\right]}\right\} \\
&= T\mathbb{E}\exp\left\{-\frac{6(\lambda-1)^2C_1^2\Delta_k^2 s}{\sigma_\omega^2\left[2\lambda^2\overline{\xi} + 2\lambda^2\sigma_1^2 + 25(1+2\lambda)^2C_1^2\Delta_k^2\right]}\right\} \\
&= T\mathbb{E}\exp\left\{-\frac{s}{\frac{\sigma_\omega^2\lambda^2}{3(\lambda-1)^2C_1^2\Delta_k^2}\overline{\xi} + \frac{\sigma_\omega^2\left[2\lambda^2\sigma_1^2+25(1+2\lambda)^2C_1^2\Delta_k^2\right]}{6(\lambda-1)^2C_1^2\Delta_k^2}}\right\} \\
&= T\mathbb{E}\exp\left\{-\frac{s}{\mathfrak{a}_2\overline{\xi} + \mathfrak{b}_2}\right\} \\
&\overset{\text{by applying steps in (33)}}{\leq} \frac{1}{3T}.
\end{aligned}$$

Therefore, to determine the value of s such that $[1 - Q_{1,s}(\tau_k)] < T^{-2}$ on $\{A_{1,s} \cap G_{1,s}\}$, we can define

$$\begin{aligned}
s \geq s_{a,k,3}(T) = \Bigg[&\left[\Omega_1 + \mathfrak{a}_2 + \mathfrak{b}_2 + \Omega_1\mathfrak{a}_2\right]\left(\frac{1}{3}\log^{-1}2 \times \log\left\{3\left[1 + \frac{\sqrt{\pi\mathfrak{b}_2}}{2} + \frac{\sqrt{2\pi}\Omega_1\mathfrak{a}_2}{2\mathfrak{b}_2^2} + \frac{2\mathfrak{a}_2}{\mathfrak{b}_2}\right]\right\} + 1\right) \\
&\vee \frac{18\sigma_\omega^2(2\mu_1-1)^2}{(\lambda^2+\lambda+1/4)^2} \vee \frac{2\sigma_\omega^2(1+2\lambda^2)}{(1+2\lambda)^2}\Bigg] \times 3\log T.
\end{aligned}$$

By choosing this value for $s_{a,k,3}$ then we get that $a_{s,k,3} \leq T^{-1}$ whenever $s \geq s_{a,k,3}$. □

D.2. Lemmas on bounding b_k.

Lemma D.5 (Bounding $b_{k,s,1}$ at (6))**.** *Consider*

$$s_{b,k,1}(T) = \frac{2\sigma_k^2}{C_1^2(1+2\lambda)^2\Delta_k^2} \times \log T.$$

For any $s \geq s_{b,k,1}(T)$, we have

$$b_{k,s,1} \leq T^{-1}, \qquad \forall\, k \in \{2, \ldots, K\}$$

provided that $T \geq 2$.

Proof. Similar to Lemma D.5 and noting

$$s_{b,k,1}(T) \geq \frac{\sigma_k^2}{C_1^2(1+2\lambda)^2\Delta_k^2} \times \log(2T),$$

we apply the Hoeffding inequality, which gives

$$\begin{aligned} b_{k,s,1} &= \mathbb{E}\big[\mathbb{I}(Q_{k,s}(\tau_k) > T^{-1})\mathbb{I}(A_{k,s}^c)\big] \\ &\leq \mathbb{P}(A_{k,s}^c) \\ &= \mathbb{P}\left(|\overline{R}_{k,s} - \mu_k| \geq (1+2\lambda)C_1\Delta_k\right) \\ &\overset{\text{by Hoeffding inequality}}{\leq} T^{-1}. \end{aligned}$$

□

Lemma D.6 (Bounding $b_{k,s,2}$ at (7))**.** *Consider*

$$s_{b,k,2}(T) = \left\{\left[\left(\Omega_k + \mathfrak{a}_1 + \mathfrak{b}_1 + \Omega_k\mathfrak{a}_1\right)\left(\frac{1}{3}\log^{-1}2 \times \log\left\{3\left[1 + \frac{\sqrt{\pi\mathfrak{b}_1}}{2} + \frac{\sqrt{2\pi}\Omega_k\mathfrak{a}_1}{2\mathfrak{b}_1^2} + \frac{2\mathfrak{a}_1}{\mathfrak{b}_1}\right]\right\} + 1\right)\right] \vee \frac{18\sigma_\omega^2(2\mu_k-1)^2}{(\lambda^2+\lambda+1/4)^2} \vee \frac{2\sigma_\omega^2(1+2\lambda^2)}{(1+2\lambda)^2}\right\} \times 3\log T,$$

where

$$\mathfrak{a}_1 = \frac{4\sigma_\omega^2\lambda^2}{3C_2^2(1+2\lambda)^2\Delta_k^2}, \qquad \mathfrak{b}_1 = \frac{2\sigma_\omega^2\left[2\lambda^2\sigma_1^2 + (1+2\lambda)^2C_2^2\Delta_k^2\right]}{3C_2^2(1+2\lambda)^2\Delta_k^2}, \qquad \textit{and} \qquad \Omega_k = 8\sqrt{2}\sigma_k^2.$$

For any $s \geq s_{b,k,2}(T)$, we have

$$b_{k,s,2} \leq T^{-1}, \qquad \forall\, k \in \{2, \ldots, K\}$$

provided that $T \geq 2$.

Proof. Similar to bounding $a_{k,s,2}$ in Lemma D.3, we can show that if we take

$$s_{b,k,2}(T) = \left\{\left[\left(\Omega_k + \mathfrak{a}_1 + \mathfrak{b}_1 + \Omega_k\mathfrak{a}_1\right)\left(\frac{1}{3}\log^{-1}2 \times \log\left\{3\left[1 + \frac{\sqrt{\pi\mathfrak{b}_1}}{2} + \frac{\sqrt{2\pi}\Omega_k\mathfrak{a}_1}{2\mathfrak{b}_1^2} + \frac{2\mathfrak{a}_1}{\mathfrak{b}_1}\right]\right\} + 1\right)\right] \vee \frac{18\sigma_\omega^2(2\mu_k-1)^2}{(\lambda^2+\lambda+1/4)^2} \vee \frac{2\sigma_\omega^2(1+2\lambda^2)}{(1+2\lambda)^2}\right\} \times 3\log T,$$

with $\Omega_k = 8\sqrt{2}\sigma_k^2$, we have

$$\begin{aligned} b_{k,s,2} &\leq \mathbb{E}\big[\mathbb{I}(G_{k,s}^c)\big] \\ &\leq \mathbb{E}\exp\left\{-\frac{3C_2^2(1+2\lambda)^2\Delta_k^2 s}{2\sigma_\omega^2\left[2\lambda^2\overline{\zeta^{(k)}} + 2\lambda^2\sigma_k^2 + (1+2\lambda)^2C_2^2\Delta_k^2\right]}\right\} \\ &\quad + \left[\exp\left\{-\frac{(\lambda^2+\lambda+1/4)^2 s}{18\sigma_\omega^2(2\mu_k-1)^2}\right\} \times \mathbb{I}\left(\mu_k \neq \frac{1}{2}\right) + 0 \times \mathbb{I}\left(\mu_k = \frac{1}{2}\right)\right] + \frac{T}{2}\exp\left\{-\frac{s(1+2\lambda)^2}{2\sigma_\omega^2(1+2\lambda^2)}\right\} \leq T^{-2}, \end{aligned}$$

where $\overline{\zeta^{(k)}} = \frac{1}{s}\sum_{i=1}^{s}\zeta_i^{(k)}$, with $\{\zeta_i^{(k)}\}_{i=1}^{s}$ being i.i.d independent sub-Exponential variables such that $\zeta_i^{(k)} \sim \text{subE}(8\sqrt{2}\sigma_k^2)$. Then $b_{k,s,2} \leq T^{-2} \leq T^{-1}$ for any $s \geq s_{b,k,2}$ and $k \in \{2, \ldots, K\}$. $\square$

Lemma D.7 (Bounding $b_{k,s,3}$ at (8))**.** *Consider*

$$s_{b,k,3}(T) = \left\{ \left[\left(\Omega_k + \mathfrak{a}_2 + \mathfrak{b}_2 + \Omega_k \mathfrak{a}_2\right) \left(\frac{1}{3}\log^{-1} 2 \times \log \left\{ 3 \left[1 + \frac{\sqrt{\pi \mathfrak{b}_2}}{2} + \frac{\sqrt{2\pi}\Omega_k \mathfrak{a}_2}{2\mathfrak{b}_2^2} + \frac{2\mathfrak{a}_2}{\mathfrak{b}_2} \right] \right\} + 1 \right) \right] \vee \frac{18\sigma_\omega^2(2\mu_k - 1)^2}{(\lambda^2 + \lambda + 1/4)^2} \vee \frac{2\sigma_\omega^2(1 + 2\lambda^2)}{(1 + 2\lambda)^2} \right\} \times 3\log T,$$

where

$$\mathfrak{a}_2 = \frac{\sigma_\omega^2 \lambda^2}{3(\lambda - 1)^2 C_1^2 \Delta_k^2}, \qquad \mathfrak{b}_2 = \frac{\sigma_\omega^2 \left[2\lambda^2\sigma_1^2 + 25(1 + 2\lambda)^2 C_1^2 \Delta_k^2\right]}{6(\lambda - 1)^2 C_1^2 \Delta_k^2}, \qquad \textit{and} \qquad \Omega_k = 8\sqrt{2}\sigma_k^2.$$

For any $s \geq s_{b,k,3}(T)$ and $\lambda > 1$,

$$b_{k,s,3} \leq T^{-1}, \qquad \forall\, k \in \{2, \ldots, K\}$$

provided that $T \geq 2$.

Proof. The basic idea is the same as bounding $a_{k,s,3}$. The only difference is that we replace $\left[1 - Q_{k,s}(\tau_k)\right]$ with $Q_{k,s}(\tau_k)$. Again, we will first bound $\Gamma_k = \overline{Y}_{k,s} - \overline{R}^*_{k,s}$ on the event $A_{k,s} \cap G_{k,s}$. Exactly as before, we let

$$\tau_k = \frac{\mu_k + \lambda}{1 + 2\lambda} + \frac{6\lambda C_1 \Delta_k}{1 + 2\lambda}$$

which satisfies $\tau_k \geq \frac{\mu_k + \lambda}{1 + 2\lambda}$. To ensure $\tau_k \leq \frac{\mu_1 + \lambda}{1 + 2\lambda}$, one just needs to take $C_1 = \frac{1}{6\lambda}$, and then $C_2 = \frac{1}{2}C_1 = \frac{1}{12\lambda}$ by (35). Next, we will obtain the range for Γ_k on the event $A_{k,s} \cap G_{k,s}$ as

$$\begin{aligned} \Gamma_k &\geq \tau_k - 2C_1\Delta_k - \frac{\mu_k + \lambda}{1 + 2\lambda} \\ &= 2C_1\Delta_k \frac{\lambda - 1}{2\lambda + 1} \\ &\overset{\text{by } \lambda > 1}{>} 0, \end{aligned}$$

and

$$\begin{aligned} \Gamma_k &\leq \tau_k + 2C_1\Delta_k - \frac{\mu_k + \lambda}{1 + 2\lambda} \\ &= 2C_1\Delta_k\left(1 + \frac{3\lambda}{1 + 2\lambda}\right) \\ &\overset{\text{by } \frac{3\lambda}{1+2\lambda} \geq \frac{3}{2}}{\leq} 2C_1\Delta_k\left(1 + \frac{3}{2}\right) = 5C_1\Delta_k, \end{aligned}$$

i.e., $\Gamma_k \in \left(\frac{2C_1(\lambda-1)\Delta_k}{2\lambda+1}, 5C_1\Delta_k\right) \subseteq (0,\infty)$. Therefore,

$$
\begin{aligned}
& Q_{k,s}(\tau_k)\mathbb{I}(A_{k,s})\mathbb{I}(G_{k,s}) \\
&= \mathbb{P}\left(\overline{Y}_{k,s} - \overline{R}^*_{k,s} \ge \Gamma_k \mid \mathcal{H}_{k,s}\right)\mathbb{I}(A_{k,s} \cap G_{k,s}) \\
&\overset{\text{apply steps in (26)}}{\le} T\mathbb{E}\left[\exp\left\{-\frac{9\Gamma_k^2(1+2\lambda)^2 s^2}{6\sigma_\omega^2 s\left[2\lambda^2\overline{\zeta^{(k)}} + 2\lambda^2\sigma_1^2 + \Gamma_k^2(1+2\lambda)^2\right]}\right\}\mathbb{I}(A_{k,s} \cap G_{k,s})\right] \\
&\quad + T\left[\exp\left\{-\frac{(\lambda^2+\lambda+1/4)^2 s}{18\sigma_\omega^2(2\mu_k-1)^2}\right\} \times \mathbb{I}\left(\mu_k \ne \frac{1}{2}\right) + 0 \times \mathbb{I}\left(\mu_k = \frac{1}{2}\right)\right] \\
&\quad + \frac{T}{2}\exp\left\{-\frac{s(1+2\lambda)^2}{2\sigma_\omega^2(1+2\lambda^2)}\right\} \\
&\overset{\text{by the bound of } \Gamma_k}{\le} T\mathbb{E}\exp\left\{-\frac{36(\lambda-1)^2 C_1^2\Delta_k^2 s^2}{6\sigma_\omega^2 s\left[2\lambda^2\overline{\zeta^{(k)}} + 2\lambda^2\sigma_1^2 + 25(1+2\lambda)^2 C_1^2\Delta_k^2\right]}\right\} \\
&\quad + T\left[\exp\left\{-\frac{(\lambda^2+\lambda+1/4)^2 s}{18\sigma_\omega^2(2\mu_k-1)^2}\right\} \times \mathbb{I}\left(\mu_k \ne \frac{1}{2}\right) + 0 \times \mathbb{I}\left(\mu_k = \frac{1}{2}\right)\right] \\
&\quad + \frac{T}{2}\exp\left\{-\frac{s(1+2\lambda)^2}{2\sigma_\omega^2(1+2\lambda^2)}\right\},
\end{aligned}
$$

where $\overline{\zeta^{(k)}} = \frac{1}{s}\sum_{i=1}^s \zeta_i^{(k)}$ with $\{\zeta_i^{(k)}\}_{i=1}^s$ are i.i.d. independent sub-Exponential variables such that $\zeta_i^{(k)} \sim \mathrm{subE}(8\sqrt{2}\sigma_k^2)$. Let

$$
s \ge \left[\Omega_k + \mathfrak{a}_2 + \mathfrak{b}_2 + \Omega_k\mathfrak{a}_2\right]\left(\frac{1}{3}\log^{-1}2 \times \log\left\{3\left[1 + \frac{\sqrt{\pi}\mathfrak{b}_2}{2} + \frac{\sqrt{2\pi}\Omega_k\mathfrak{a}_2}{2\mathfrak{b}_2^2} + \frac{2\mathfrak{a}_2}{\mathfrak{b}_2}\right]\right\} + 1\right).
$$

By applying Lemma E.4 again, we have

$$
\begin{aligned}
& T\mathbb{E}\exp\left\{-\frac{36(\lambda-1)^2 C_1^2\Delta_k^2 s^2}{6\sigma_\omega^2 s\left[2\lambda^2\overline{\zeta^{(k)}} + 2\lambda^2\sigma_1^2 + 25(1+2\lambda)^2 C_1^2\Delta_k^2\right]}\right\} \\
=& T\mathbb{E}\exp\left\{-\frac{s}{\mathfrak{a}_2\overline{\zeta^{(k)}} + \mathfrak{b}_2}\right\} \le \frac{1}{3T}.
\end{aligned}
$$

Therefore, we have $b_{k,s,3} \le T^{-1}$ for any $k \in \{2,\ldots,K\}$ if we choose

$$
\begin{aligned}
s_{b,k,3}(T) = \Bigg\{\Bigg[&\left(\Omega_k + \mathfrak{a}_2 + \mathfrak{b}_2 + \Omega_k\mathfrak{a}_2\right)\left(\frac{1}{3}\log^{-1}2 \times \log\left\{3\left[1 + \frac{\sqrt{\pi}\mathfrak{b}_2}{2} + \frac{\sqrt{2\pi}\Omega_k\mathfrak{a}_2}{2\mathfrak{b}_2^2} + \frac{2\mathfrak{a}_2}{\mathfrak{b}_2}\right]\right\} + 1\right)\Bigg] \\
&\vee \frac{18\sigma_\omega^2(2\mu_k-1)^2}{(\lambda^2+\lambda+1/4)^2} \vee \frac{2\sigma_\omega^2(1+2\lambda^2)}{(1+2\lambda)^2}\Bigg\} \times 3\log T,
\end{aligned}
$$

□

D.3. Lemmas on simplifying bounds.

Lemma D.8. *Assuming the conditions are identical to those in Theorem C.1, the constants $c_1(\mu_1,\sigma_1,\mu_k,\sigma_k,\lambda)$ and $c_2(\mu_1,\sigma_1,\mu_k,\sigma_k,\lambda)$, as defined in* (10) *and* (11)*, can be upper-bounded by $C_1(\sigma_1,\sigma_k,\lambda,\sigma_\omega)$ and $C_2(\sigma_1,\sigma_k,\lambda,\sigma_\omega)$ specified in Theorem C.1, respectively.*

Proof. First, we establish bounds for the components d_1 and d_2 in $c_1(\mu_1,\sigma_1,\mu_k,\sigma_k,\lambda)$ and $c_2(\mu_1,\sigma_1,\mu_k,\sigma_k,\lambda)$. Observe that

$$\begin{aligned}d_1 =&\Delta_k^2\{(\mathfrak{a}_1+\mathfrak{a}_2)+(\mathfrak{b}_1+\mathfrak{b}_2)+\Omega_{\max}(\mathfrak{a}_1+\mathfrak{a}_2)\}\\ \overset{\text{by } \Delta_k\le 1}{\le}&\frac{1+\Omega_{\max}}{3}\sigma_\omega^2\left[\frac{192\lambda^4}{(1+2\lambda)^2}+\frac{36\lambda^4}{(\lambda-1)^2}\right]\\ &+\sigma_\omega^2\left[\frac{2\left[96\lambda^4\sigma_1^2+(1+2\lambda)^2/(36\lambda^2)\right]}{3(1+2\lambda)^2}+\frac{72\lambda^4\sigma_1^2+25(1+2\lambda)^2}{6(\lambda-1)^2}\right].\end{aligned}$$

Given $\lambda\ge\left(1+\frac{\sigma_\omega^2}{4}+\frac{4\sigma_1}{\sigma_\omega}\right)+\sqrt{\frac{4\sigma_1}{\sigma_\omega}\left(\frac{4\sigma_1}{\sigma_\omega}+1\right)}>\frac{4\sigma_1}{\sigma_\omega}+1$, we can infer that $(\lambda-1)^2\ge 16\sigma_1^2/\sigma_\omega^2$, and consequently,

$$\frac{192\lambda^4}{(1+2\lambda)^2}+\frac{36\lambda^4}{(\lambda-1)^2}\le\frac{192\lambda^4}{4\lambda^2}+\frac{36\lambda^4}{16\sigma_1^2/\sigma_\omega^2}=48\lambda^2+3\frac{\sigma_\omega^2}{\sigma_1^2}\lambda^4.$$

Similarly,

$$\begin{aligned}&\frac{2\left[96\lambda^4\sigma_1^2+(1+2\lambda)^2/(36\lambda^2)\right]}{3(1+2\lambda)^2}+\frac{72\lambda^4\sigma_1^2+25(1+2\lambda)^2}{6(\lambda-1)^2}\\ \le&\frac{2\times 96\lambda^4\sigma_1^2}{3\times 4\lambda^2}+\frac{2}{3\times 36\lambda^2}+\frac{72\lambda^4\sigma_1^2+25(1+2\lambda)^2}{6\times 16\sigma_1^2/\sigma_\omega^2}\\ \le&16\lambda^2\sigma_1^2+\frac{1}{48}+\lambda^4\sigma_\omega^2+\frac{25\times 9\lambda^2}{6\times 16\sigma_1^2/\sigma_\omega^2}\le 16\lambda^2\sigma_1^2+\frac{1}{48}+3\lambda^4\sigma_\omega^2+3\lambda^2\sigma_\omega^2/\sigma_1^2.\end{aligned}$$

Thus,

$$\begin{aligned}d_1 =&\Delta_k^2\{(\mathfrak{a}_1+\mathfrak{a}_2)+(\mathfrak{b}_1+\mathfrak{b}_2)+\Omega_{\max}(\mathfrak{a}_1+\mathfrak{a}_2)\}\\ \le&\frac{(1+\Omega_{\max})\sigma_\omega^2}{3}\left[48\lambda^2+3\frac{\sigma_\omega^2}{\sigma_1^2}\lambda^4\right]+\sigma_\omega^2\left[16\lambda^2\sigma_1^2+\frac{1}{48}+3\lambda^4\sigma_\omega^2+3\frac{\sigma_\omega^2}{\sigma_1^2}\lambda^2\right]\\ =&\sigma_\omega^2\left\{\left[16(1+\Omega_{\max})+16\sigma_1^4+3\frac{\sigma_\omega^2}{\sigma_1^2}\right]\lambda^2+\left[(1+\Omega_{\max})\frac{\sigma_\omega^2}{\sigma_1^2}+3\sigma_\omega^2\right]\lambda^4+\frac{1}{48}\right\}\\ \le&\sigma_\omega^2\left[16(1+\Omega_{\max})+16\sigma_1^4+3\frac{\sigma_\omega^2}{\sigma_1^2}+(1+\Omega_{\max})\frac{\sigma_\omega^2}{\sigma_1^2}+3\sigma_\omega^2+1\right]\lambda^4\\ \le&\left[(1+\Omega_{\max})\left(16+\frac{\sigma_\omega^2}{\sigma_1^2}\right)+16\sigma_1^4+3\frac{\sigma_\omega^2}{\sigma_1^2}+3\sigma_\omega^2+1\right]\sigma_\omega^2\lambda^4.\end{aligned}\tag{38}$$

Define

$$D_1(\sigma_1,\sigma_k,\lambda,\sigma_\omega):=\left[(1+\Omega_{\max})\left(16+\frac{\sigma_\omega^2}{\sigma_1^2}\right)+16\sigma_1^4+3\frac{\sigma_\omega^2}{\sigma_1^2}+3\sigma_\omega^2+1\right]\sigma_\omega^2\lambda^4,$$

then $d_1\le D_1(\sigma_1,\sigma_k,\lambda,\sigma_\omega)$. On the other hand, regarding d_2, we have the following:

$$\begin{aligned}d_2 =&3\left[1+\frac{\sqrt{\pi}}{2}\left(\sqrt{\mathfrak{b}_1}+\sqrt{\mathfrak{b}_2}\right)+\frac{\sqrt{2\pi}\Omega_{\max}}{2}\left(\frac{\mathfrak{a}_1}{\mathfrak{b}_1^2}+\frac{\mathfrak{a}_2}{\mathfrak{b}_2^2}\right)+2\left(\frac{\mathfrak{a}_1}{\mathfrak{b}_1}+\frac{\mathfrak{a}_2}{\mathfrak{b}_2}\right)\right]\\ \le&3\left[1+\frac{\sqrt{\pi}}{2}\left(\sqrt{\frac{2\sigma_\omega^2(1+2\lambda)^2\Delta_k^2/(144\lambda^2)}{3(1+2\lambda)^2\Delta_k^2}}+\sqrt{\frac{25\sigma_\omega^2(1+2\lambda)^2\Delta_k^2}{6(\lambda-1)^2\Delta_k^2}}\right)\right.\\ &+\frac{\sqrt{2\pi}\Omega_{\max}}{2}\left(\frac{(1+2\lambda)^2}{64\lambda^4\sigma_1^4}+\frac{(\lambda-1)^2}{12\lambda^4\sigma_1^2\sigma_\omega^2}\right)\\ &\left.+2\left(\frac{96\lambda^4}{96\lambda^4\sigma_1^2+(1+2\lambda)^2/(36\lambda^2)}+\frac{72\lambda^4}{72\lambda^4\sigma_1^2+25(1+2\lambda)^2}\right)\right].\end{aligned}$$

To simply the bound above, we can use the fact $\lambda>1$ to derive the following inequalities:

$$\begin{aligned}\sqrt{\frac{2\sigma_\omega^2(1+2\lambda)^2\Delta_k^2/(144\lambda^2)}{3(1+2\lambda)^2\Delta_k^2}}+\sqrt{\frac{25\sigma_\omega^2(1+2\lambda)^2\Delta_k^2}{6(\lambda-1)^2\Delta_k^2}}&\le\frac{\sigma_\omega}{12\lambda}+\frac{5\sigma_\omega(1+2\lambda)}{2(\lambda-1)}\\ &\le\sigma_\omega+3\frac{\sigma_\omega^2}{\sigma_1}(1+2\lambda)\le\frac{3\sigma_\omega^2}{\sigma_1}\left[\frac{\sigma_1}{\sigma_\omega}+3\lambda\right],\end{aligned}$$

$$
\begin{aligned}
\frac{(1+2\lambda)^2}{64\lambda^4\sigma_1^4} + \frac{(\lambda-1)^2}{12\lambda^4\sigma_1^2\sigma_\omega^2} &\le \frac{9\lambda^2}{64\lambda^4\sigma_1^4} + \frac{16\sigma_1^2/\sigma_\omega^2}{12\lambda^4\sigma_1^2\sigma_\omega^2} \\
&\le \frac{9\lambda^2}{64\lambda^4\sigma_1^4} + \frac{16\sigma_1^2/\sigma_\omega^2}{12\lambda^2\sigma_1^2\sigma_\omega^2} \\
&\le 2\left[\frac{1}{\sigma_1^2\lambda^2} + \frac{1}{\sigma_\omega^2\lambda^2}\right] \le 2\left[\frac{\sigma_\omega^2}{16\sigma_1^4} + \frac{1}{\sigma_\omega^2}\right],
\end{aligned}
$$

and

$$
\begin{aligned}
&\frac{96\lambda^4}{96\lambda^4\sigma_1^2 + (1+2\lambda)^2/(36\lambda^2)} + \frac{72\lambda^4}{72\lambda^4\sigma_1^2 + 25(1+2\lambda)^2} \\
\le&\frac{96\lambda^4}{96\lambda^4\sigma_1^2} + \frac{72\lambda^4}{72\lambda^4\sigma_1^2} \\
=&\frac{2\lambda^2}{\lambda^2\sigma_1^2} \\
\le&\frac{2\lambda^2}{16\sigma_1^4/\sigma_\omega^2} = \frac{\lambda^2\sigma_\omega^2}{8\sigma_1^4}.
\end{aligned}
$$

Thus, d_2 is upper-bound by

$$
\begin{aligned}
d_2 &= 3\left[1 + \frac{\sqrt{\pi}}{2}\left(\sqrt{\mathfrak{b}_1} + \sqrt{\mathfrak{b}_2}\right) + \frac{\sqrt{2\pi}\Omega_{\max}}{2}\left(\frac{\mathfrak{a}_1}{\mathfrak{b}_1^2} + \frac{\mathfrak{a}_2}{\mathfrak{b}_2^2}\right) + 2\left(\frac{\mathfrak{a}_1}{\mathfrak{b}_1} + \frac{\mathfrak{a}_2}{\mathfrak{b}_2}\right)\right] \\
&\le 3\left[1 + \frac{3\sqrt{\pi}\sigma_\omega^2}{2\sigma_1}\left(\frac{\sigma_1}{\sigma_\omega} + 3\lambda\right) + \sqrt{2\pi}\Omega_{\max}\left(\frac{\sigma_\omega^2}{16\sigma_1^4} + \frac{1}{\sigma_\omega^2}\right) + \frac{\lambda^2\sigma_\omega^2}{4\sigma_1^4}\right].
\end{aligned}
\tag{39}
$$

Once again, we define

$$
D_2(\sigma_1, \sigma_k, \lambda, \sigma_\omega) := 3\left[1 + \frac{3\sqrt{\pi}\sigma_\omega^2}{2\sigma_1}\left(\frac{\sigma_1}{\sigma_\omega} + 3\lambda\right) + \sqrt{2\pi}\Omega_{\max}\left(\frac{\sigma_\omega^2}{16\sigma_1^4} + \frac{1}{\sigma_\omega^2}\right) + \frac{\lambda^2\sigma_\omega^2}{4\sigma_1^4}\right],
$$

then $d_2 \le D_2(\sigma_1, \sigma_k, \lambda, \sigma_\omega)$.

Next, we can provide simple bounds for $c_1(\mu_1, \sigma_1, \mu_k, \sigma_k, \lambda, \sigma_\omega)$ and $c_2(\mu_1, \sigma_1, \mu_k, \sigma_k, \lambda, \sigma_\omega)$. Using the fact that $a \vee b \le a + b$ and $a \vee b \vee c = [(a \vee b) \vee c]$, we obtain that

$$
\begin{aligned}
&\left(\Omega_{\max}\left(\frac{\log d_2}{3\log 2} + 1\right)\right) \vee \frac{18\sigma_\omega^2\{(2\mu_1-1)^2 \vee (2\mu_k-1)^2\}}{(\lambda^2+\lambda+1/4)^2} \vee \frac{2\sigma_\omega^2(1+2\lambda^2)}{(1+2\lambda)^2} \\
\le&\left(\Omega_{\max}\left(\frac{\log d_2}{3\log 2} + 1\right)\right) \vee \frac{72\sigma_\omega^2\{(2\mu_1-1)^2 + (2\mu_k-1)^2\}}{(1+2\lambda)^2} \vee \frac{2\sigma_\omega^2(1+2\lambda^2)}{(1+2\lambda)^2} \\
\le&\left(\Omega_{\max}\left(\frac{\log d_2}{3\log 2} + 1\right)\right) \vee \frac{72\sigma_\omega^2 \times 2 + 2\sigma_\omega^2(1+2\lambda^2)}{(1+2\lambda)^2} \\
\le&\left(\Omega_{\max}\left(\frac{\log d_2}{3\log 2} + 1\right)\right) \vee \frac{144\sigma_\omega^2 + 2\sigma_\omega^2 + 4\lambda^2\sigma_\omega^2}{4\lambda^2} \\
\le&\left[\Omega_{\max}\left(\frac{\log d_2}{3\log 2} + 1\right)\right] + 38\sigma_\omega^2 \le \Omega_{\max}\left(\frac{\log D_2(\sigma_1, \sigma_k, \lambda, \sigma_\omega)}{3\log 2} + 1\right) + 38\sigma_\omega^2.
\end{aligned}
$$

and

$$
\begin{aligned}
&\left(d_1\left(\frac{\log d_2}{3\log 2} + 1\right)\right) \vee \frac{18\sigma_\omega^2\{(2\mu_1-1)^2 \vee (2\mu_k-1)^2\}}{(\lambda^2+\lambda+1/4)^2} \vee \frac{2\sigma_\omega^2(1+2\lambda^2)}{(1+2\lambda)^2} \\
\le&D_1(\sigma_1, \sigma_k, \lambda, \sigma_\omega)\left(\frac{\log D_2(\sigma_1, \sigma_k, \lambda, \sigma_\omega)}{3\log 2} + 1\right) + 38\sigma_\omega^2,
\end{aligned}
$$

where the last steps in two inequalities above are by (38) and (39). Finally, note that

$$
5 + 16e^{9/8} \le 55, \qquad \frac{72\lambda^2\sigma_k^2}{(1+2\lambda)^2}(9 + 32e^{9/8}) \le \frac{72\lambda^2\sigma_k^2}{25}(9 + 32e^{9/8}) \le 310\lambda^2\sigma_k^2,
$$

which implies $c_1(\mu_1, \sigma_1, \mu_k, \sigma_k, \lambda, \sigma_\omega) \leq C_1(\sigma_1, \sigma_k, \lambda, \sigma_\omega)$ and $c_2(\mu_1, \sigma_1, \mu_k, \sigma_k, \lambda, \sigma_\omega) \leq C_2(\sigma_1, \sigma_k, \lambda, \sigma_\omega)$. □

Lemma D.9. *Assuming the conditions the same as in Lemma D.8, we have*

$$C_1(1, 1, \lambda, \sigma_\omega) \leq 55\Bigg[8\sqrt{2}\left(\frac{\log\left[(1 + 15\sigma_\omega^{-2} + 3\sigma_\omega + 10\sigma_\omega^2)\lambda^2\right]}{3\log 2} + 1\right) + 38\sigma_\omega^2\Bigg]$$

and

$$C_2(1, 1, \lambda, \sigma_\omega) \leq 330\lambda^2 + 55\left[45(3 + \sigma_\omega^2)\left(\frac{\log\left[(1 + 15\sigma_\omega^{-2} + 3\sigma_\omega + 10\sigma_\omega^2)\lambda^2\right]}{3\log 2} + 1\right)\lambda^4 + 38\sigma_\omega^2\right].$$

Proof. We have

$$\begin{aligned} D_1(1, 1, \lambda, \sigma_\omega) &= \Big[\big(1 + 8\sqrt{2}\big)\big(16 + \sigma_\omega^2\big) + 16 + 3\sigma_\omega^2 + 3\sigma_\omega^2 + 1\Big]\sigma_\omega^2\lambda^4 \\ &\leq \Big[13\big(16 + 3\sigma_\omega^2\big) + 6\sigma_\omega^2 + 17\Big]\sigma_\omega^2\lambda^4 = 45(3 + \sigma_\omega^2)\lambda^4 \end{aligned}$$

and

$$\begin{aligned} D_2(1, 1, \lambda, \sigma_\omega) &= 3\left[1 + \frac{3\sqrt{\pi}\sigma_\omega^2}{2}\left(\frac{1}{\sigma_\omega} + 3\lambda\right) + 8\sqrt{\pi}\left(\frac{\sigma_\omega^2}{16} + \frac{1}{\sigma_\omega^2}\right) + \frac{\lambda^2\sigma_\omega^2}{4}\right] \\ &\leq 3\left[1 + \frac{3\sqrt{\pi}\sigma_\omega^2}{2}\left(\frac{1}{\sigma_\omega} + 3\right) + 8\sqrt{\pi}\left(\frac{\sigma_\omega^2}{16} + \frac{1}{\sigma_\omega^2}\right) + \frac{\sigma_\omega^2}{4}\right]\lambda^2 \leq \big(1 + 15\sigma_\omega^{-2} + 3\sigma_\omega + 10\sigma_\omega^2\big)\lambda^2. \end{aligned}$$

Therefore, we have

$$\begin{aligned} C_1(1, 1, \lambda, \sigma_\omega) &= 55\left[8\sqrt{2}\left(\frac{\log D_2(1, 1, \lambda, \sigma_\omega)}{3\log 2} + 1\right) + 55\sigma_\omega^2\right] \\ &\leq 55\Bigg[8\sqrt{2}\left(\frac{\log\left[(1 + 15\sigma_\omega^{-2} + 3\sigma_\omega + 10\sigma_\omega^2)\lambda^2\right]}{3\log 2} + 1\right) + 38\sigma_\omega^2\Bigg], \end{aligned}$$

and

$$\begin{aligned} C_2(1, 1, \lambda, \sigma_\omega) &= 310\lambda^2 + 55\left[D_1(1, 1, \lambda, \sigma_\omega)\left(\frac{\log D_2(1, , 1, \lambda, \sigma_\omega)}{3\log 2} + 1\right) + 38\sigma_\omega^2\right] \\ &< 330\lambda^2 + 55\left[45(3 + \sigma_\omega^2)\left(\frac{\log\left[(1 + 15\sigma_\omega^{-2} + 3\sigma_\omega + 10\sigma_\omega^2)\lambda^2\right]}{3\log 2} + 1\right)\lambda^4 + 38\sigma_\omega^2\right]. \end{aligned}$$

□

E. Technical Lemmas

Lemma E.1. *Let Z be a standard Gaussian variable, then the tail probability $\mathbb{P}(Z > x)$ satisfies*

$$\frac{1}{4}\exp(-x^2) < \mathbb{P}(Z > x) \leq \frac{1}{2}\exp(-x^2/2)$$

for any $x \geq 0$.

Proof. Let $Q(x) := \mathbb{P}(Z > x)$ for $x > 0$ represent the tail probability for a standard Gaussian variable Z. Let Z_1 and Z_2 are two independent standard Gaussian random variables. Then

$$\begin{aligned} \mathbb{P}(Z_1 \leq x,\, Z_2 \leq x) &= \int_{-x}^{x}\int_{-x}^{x}\frac{1}{2\pi}\exp\left[\left(-z_1^2 - z_2^2\right)/2\right]\,\mathrm{d}z_1\mathrm{d}z_2 \\ &\leq \int_{0}^{\sqrt{2}x}\int_{0}^{2\pi}\frac{1}{2\pi}\exp\left(-r^2/2\right)r\,\mathrm{d}\theta\,\mathrm{d}r \\ &= 1 - \exp(-x^2), \end{aligned}$$

which implies that $[1-2Q(x)]^2 \le 1-\exp(-x^2)$ for $x \ge 0$, or, equivalently, $\exp(-x^2) \le 4Q(x) - 4Q^2(x)$. However, since $4Q^2(x) > 0$ for all x, we obtain that

$$Q(x) > \frac{1}{4}\exp(-x^2)$$

which provides the lower bound in the lemma. For the upper bound, we observe that

$$\begin{aligned}
\exp\left(x^2/2\right) Q(x) &= \exp\left(x^2/2\right) \int_x^\infty (2\pi)^{-1/2} \exp\left(-t^2/2\right) \mathrm{d}t \\
&= \int_x^\infty (2\pi)^{-1/2} \exp\left(-\left(t^2 - x^2\right)/2\right) \mathrm{d}t \\
&< \int_x^\infty (2\pi)^{-1/2} \exp\left(-(t-x)^2/2\right) \mathrm{d}t = \frac{1}{2},
\end{aligned}$$

which establishes the upper bound. □

Lemma E.2. *Let $\{X_i\}_{i=1}^n$ be a sequence of i.i.d. mean-zero sub-Gaussian random variables with variance proxy σ^2, and let $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Then*

$$\mathbb{E}\exp\left(\lambda \overline{X}^2\right) \le e^{9/8}$$

holds for any $|\lambda| \le \frac{n}{8\sigma^2}$.

Proof. Note that if $X - \mu \sim \mathrm{subG}(\sigma^2)$, then

$$\overline{X} - \mu \sim \mathrm{subG}(\sigma^2/n) \tag{40}$$

follows from the fact that

$$\begin{aligned}
\mathbb{E}\exp\left\{s(\overline{X}-\mu)\right\} &= \prod_{i=1}^n \exp\left\{\frac{s}{n}(X_i - \mu)\right\} \\
(\text{by } X_i - \mu \text{ is sub-Gaussian}) &\le \prod_{i=1}^n \exp\left\{\frac{s^2}{n^2}\frac{\sigma^2}{2}\right\} = \exp\left\{\frac{s^2(\sigma^2/n)}{2}\right\}.
\end{aligned}$$

Then, by Proposition 4.3 in Zhang & Chen (2021), $(\overline{X}-\mu)^2 - n^{-1}\operatorname{var}(X) \sim \mathrm{subE}(8\sqrt{2}\sigma^2/n, 8\sigma^2/n)$. Therefore,

$$\begin{aligned}
\mathbb{E}\exp\left\{\lambda(\overline{X}-\mu)^2\right\} &= \mathbb{E}\exp\left\{\lambda\big((\overline{X}-\mu)^2 - n^{-1}\operatorname{var}(X)\big)\right\} \mathbb{E}\exp\{\lambda n^{-1}\operatorname{var}(X)\} \\
&\le \exp\left\{\frac{(8\sqrt{2}\sigma^2/n)^2\lambda^2}{2}\right\} \cdot \exp\left\{\lambda n^{-1}\sigma^2\right\} \\
&= \exp\left\{\frac{64\lambda^2\sigma^4}{n^2} + \frac{\lambda\sigma^2}{n}\right\} \le e^{9/8}
\end{aligned}$$

for any $\lambda \le \frac{n}{8\sigma^2}$. □

Lemma E.3. *Suppose X and Y are Gaussian variables with $\mathbb{E}X = 0$ and $\mathbb{E}Y > 0$. Then*

$$\mathbb{P}\left(\frac{|X|}{|Y|} > c\right) \le 2\mathbb{P}(X > cY) + \mathbb{P}(Y < 0)$$

for any $c > 0$.

Proof. We have

$$\begin{aligned}
\mathbb{P}\left(\frac{|X|}{|Y|} > c\right) &= \mathbb{P}\left(|X| > c|Y|\right) \\
&= \mathbb{P}\left(\{X > cY\} \cap \{X > 0\} \cap \{Y > 0\}\right) + \mathbb{P}\left(\{-X > cY\} \cap \{X \le 0\} \cap \{Y > 0\}\right) \\
&\quad + \mathbb{P}\left(\{X > -cY\} \cap \{X > 0\} \cap \{Y \le 0\}\right) + \mathbb{P}\left(\{-X > -cY\} \cap \{X \le 0\} \cap \{Y \le 0\}\right) \\
&\le \mathbb{P}\left(\{X > cY\} \cap \{X > 0\} \cap \{Y > 0\}\right) + \mathbb{P}\left(\{-X > cY\} \cap \{-X \ge 0\} \cap \{Y > 0\}\right) + \mathbb{P}(Y \le 0) \\
&\overset{\text{by using } X \text{ is symmetric about } 0.}{\le} 2\mathbb{P}(X > cY) + \mathbb{P}(Y \le 0)
\end{aligned}$$

for any $c > 0$. $\square$

Lemma E.4 (sub-Exponential concentration)**.** *Consider mean-zero independent random variables $X_i \sim \mathrm{subE}(\lambda_i)$, and positive constants a and b such that $a\overline{X} + b > 0$, where the sample mean $\overline{X} := \frac{1}{s}\sum_{i=1}^{s} X_i$. Then,*

$$\mathbb{E}\exp\left\{-\frac{s}{a\overline{X}+b}\right\} \le \left[1 + \frac{\sqrt{\pi b}}{2} + \frac{\sqrt{2\pi}\lambda a}{2b^2} + \frac{2a}{b}\right] s \exp\left\{-\frac{s}{\sqrt{2\lambda a} \vee (b+\lambda a)}\right\},$$

for $s \in \mathbb{N}$ such that $2\lambda as/b \ge 1$, where $\lambda := \left(\frac{1}{s}\sum_{i=1}^{s}\lambda_i^2\right)^{1/2}$. Specifically, we have $\log\mathbb{E}\exp\left\{-\frac{s}{a\overline{X}+b}\right\} \lesssim -s$.

Proof. Denote $Y := a\overline{X} + b$ as a strictly positive random variable. For any non-negative strictly increasing function $f(\cdot)$, we have

$$\mathbb{E}f(Y) = \int_0^\infty \mathbb{P}\left(f(Y) > r\right)\,\mathrm{d}r = \int_0^\infty \mathbb{P}\left(Y > f^{-1}(r)\right)\,\mathrm{d}r.$$

Then for any fixed $s \in \mathbb{N}$,

$$\begin{aligned}
\mathbb{E}\exp\left\{-\frac{s}{a\overline{X}+b}\right\} &= \mathbb{E}\exp\left\{-\frac{s}{Y}\right\} \\
&= \int_0^\infty \mathbb{P}\left(\exp\left\{-\frac{s}{Y}\right\} > r\right)\mathrm{d}r \\
&= \int_0^1 \mathbb{P}\left(Y > \frac{s}{\log r^{-1}}\right)\mathrm{d}r \\
&= \int_0^1 \mathbb{P}\left(\overline{X} > \frac{1}{a}\left[\frac{s}{\log r^{-1}} - b\right]\right)\mathrm{d}r \\
&= \int_0^{\exp\{-\frac{s}{b}\}} \mathbb{P}\left(\overline{X} > \underbrace{\frac{1}{a}\left[\frac{s}{\log r^{-1}} - b\right]}_{\le 0}\right)\mathrm{d}r + \int_{\exp\{-\frac{s}{b}\}}^1 \mathbb{P}\left(\overline{X} > \underbrace{\frac{1}{a}\left[\frac{s}{\log r^{-1}} - b\right]}_{>0}\right)\mathrm{d}r \\
&\le \exp\left\{-\frac{s}{b}\right\} \times 1 + \int_{\exp\{-\frac{s}{b}\}}^1 \mathbb{P}\left(\overline{X} > \underbrace{\frac{1}{a}\left[\frac{s}{\log r^{-1}} - b\right]}_{>0}\right)\mathrm{d}r \\
&\overset{\text{by letting } u = \frac{s}{\log r^{-1}}}{=} \exp\left\{-\frac{s}{b}\right\} + \int_b^\infty \mathbb{P}\left(\overline{X} > \underbrace{\frac{u-b}{a}}_{>0}\right)\exp\left\{-\frac{s}{u}\right\}\frac{s}{u^2}\,\mathrm{d}r.
\end{aligned}$$

Recall that $\overline{X}$ is the average of independent mean-zero sub-exponential variables, and $\lambda = \left(\frac{1}{s}\sum_{i=1}^{s}\lambda_i^2\right)^{1/2}$. Then, we have

$$\mathbb{P}\left(\overline{X} > t\right) \le \exp\left\{-\frac{1}{2}\left(\frac{st^2}{\lambda^2} \wedge \frac{st}{\lambda}\right)\right\}$$

for any $t > 0$ by Corollary 4.2.(c) in Zhang & Chen (2021). Therefore, we can further bound the expectation as

$$\begin{aligned}
\mathbb{E}\exp\left\{-\frac{s}{a\overline{X}+b}\right\} &\overset{\text{by sub-Exponential inequality}}{\le} \exp\left\{-\frac{s}{b}\right\} + \int_b^\infty \frac{s}{u^2}\exp\left\{-\frac{s}{u}\right\}\exp\left\{-\frac{1}{2}\frac{s(u-b)^2}{\lambda^2a^2} \wedge \frac{s(u-b)}{\lambda a}\right\}\mathrm{d}u \\
&\le \exp\left\{-\frac{s}{b}\right\} + \frac{s}{b^2}\int_b^\infty \exp\left\{-\left[\frac{s}{u} + \frac{1}{2}\frac{s(u-b)^2}{\lambda^2a^2} \wedge \frac{s(u-b)}{\lambda a}\right]\right\}\mathrm{d}u \\
&= \exp\left\{-\frac{s}{b}\right\} + \left[\int_b^{b+\lambda a} + \int_{b+\lambda a}^\infty\right]\frac{s}{u^2}\exp\left\{-\left[\frac{s}{u} + \frac{1}{2}\frac{s(u-b)^2}{\lambda^2a^2} \wedge \frac{s(u-b)}{\lambda a}\right]\right\}\mathrm{d}u \\
&=: \exp\left\{-\frac{s}{b}\right\} + \left[I + II\right].
\end{aligned}$$

The next step is bounding both I and II. For I, we have

$$
\begin{aligned}
I &= \int_b^{b+\lambda a} \frac{s}{u^2} \exp\left\{-\left[\frac{s}{u} + \frac{s(u-b)^2}{2\lambda^2 a^2}\right]\right\} \mathrm{d}u \\
&\leq \int_b^{b+\lambda a} \frac{s}{b^2} \exp\left\{-\left[\frac{s}{b+\lambda a} + \frac{s(u-b)^2}{2\lambda^2 a^2}\right]\right\} \mathrm{d}u \\
&= \frac{s}{b^2} \exp\left\{-\frac{s}{b+\lambda a}\right\} \int_b^{b+\lambda a} \exp\left\{-\frac{s(u-b)^2}{2\lambda^2 a^2}\right\} \mathrm{d}u \\
&\leq \frac{s}{b^2} \exp\left\{-\frac{s}{b+\lambda a}\right\} \int_b^{\infty} \exp\left\{-\frac{(u-b)^2}{2\lambda^2 a^2/s}\right\} \mathrm{d}u = \frac{\sqrt{2\pi}\lambda a}{2b^2} \exp\left\{-\frac{s}{b+\lambda a}\right\}
\end{aligned}
$$

For II, we decompose it as

$$
\begin{aligned}
II &= \int_{b+\lambda a}^{\infty} \frac{s}{u^2} \exp\left\{-\left[\frac{s}{u} + \frac{s(u-b)}{2\lambda a}\right]\right\} \mathrm{d}u \\
&= \left[\int_{b+\lambda a}^{\frac{2\lambda a s}{b}} + \int_{\frac{2\lambda a c}{b}}^{\infty}\right] \frac{s}{u^2} \exp\left\{-\left[\frac{s}{u} + \frac{s(u-b)}{2\lambda a}\right]\right\} \mathrm{d}u \\
&= \int_{b+\lambda a}^{\frac{2\lambda a s}{b}} \frac{s}{u^2} \exp\left\{-\left[\frac{s}{u} + \frac{s(u-b)}{2\lambda a}\right]\right\} \mathrm{d}u + \exp\left\{-\frac{s}{b}\right\} \int_{\frac{2\lambda a s}{b}}^{\infty} \frac{s}{u^2} \exp\left\{-\frac{s(u-b)(u-\frac{2\lambda a s}{b})}{2\lambda a u}\right\} \mathrm{d}u \\
&=: II_1 + II_2.
\end{aligned}
$$

For II_1, we have

$$
\begin{aligned}
II_1 &= \int_{b+\lambda a}^{\frac{2\lambda a s}{b}} \frac{s}{u^2} \exp\left\{-\left[\frac{s}{u} + \frac{s(u-b)}{2\lambda a}\right]\right\} \mathrm{d}u \\
&= \int_{b+\lambda a}^{\frac{2\lambda a s}{b}} \frac{s}{u^2} \exp\left\{-s \times \left[\frac{1}{u} + \frac{(u-b)}{2\lambda a}\right]\right\} \mathrm{d}u \\
&\overset{g(u)=\frac{1}{u}+\frac{(u-b)}{2\lambda a} \text{ is increasing on } u \geq \sqrt{2\lambda a}}{\leq} \int_{b+\lambda a}^{\frac{2\lambda a s}{b}} \frac{s}{u^2} \exp\left\{-s \times \left[\frac{1}{\sqrt{2\lambda a} \vee (b+\lambda a)} + \frac{(\sqrt{2\lambda a} \vee (b+\lambda a) - b)}{2\lambda a}\right]\right\} \mathrm{d}u \\
&\leq \frac{s}{(b+\lambda a)} \left[\frac{2\lambda a s}{b} - \big(b+\lambda a\big)\right] \exp\left\{-\frac{s}{\sqrt{2\lambda a} \vee (b+\lambda a)}\right\}.
\end{aligned}
$$

For II_2, if $\frac{2\lambda a s}{b} \geq 1$, i.e. $\lambda s \geq \frac{b}{2a}$, we obtain that

$$
\begin{aligned}
II_2 &= \exp\left\{-\frac{s}{b}\right\} \int_{\frac{2\lambda a s}{b}}^{\infty} \frac{s}{u^2} \exp\left\{-\frac{s(u-b)(u-\frac{2\lambda a s}{b})}{2\lambda a u}\right\} \mathrm{d}u \\
&\leq \exp\left\{-\frac{s}{b}\right\} \int_{\frac{2\lambda a s}{b}}^{\infty} \frac{s}{u^2} \exp\left\{-\frac{s(u-\frac{2\lambda a s}{b})^2}{2\lambda a u}\right\} \mathrm{d}u \\
&\overset{\text{by } \frac{2\lambda a s}{b} \geq 1}{\leq} s \exp\left\{-\frac{s}{b}\right\} \int_{\frac{2\lambda a s}{b}}^{\infty} \frac{1}{u^{3/2}} \exp\left\{-\frac{(u-\frac{2\lambda a s}{b})^2}{2\lambda a u}\right\} \mathrm{d}u \\
&\overset{\text{by } \frac{2\lambda a s}{b} \geq 1}{\leq} s \exp\left\{-\frac{s}{b}\right\} \sqrt{\frac{2\lambda a s}{b}} \int_{\frac{2\lambda a s}{b}}^{\infty} \frac{1}{u^{3/2}} \exp\left\{-\frac{(u-\frac{2\lambda a s}{b})^2}{2\lambda a u}\right\} \mathrm{d}u \\
&= s \exp\left\{-\frac{s}{b}\right\} \sqrt{\frac{2\lambda a s}{b}} \frac{b}{2s} \sqrt{\frac{\pi}{2\lambda a}} \left[1 - e^{4s/b} \frac{2}{\sqrt{\pi}} \int_{2\sqrt{s/b}}^{\infty} e^{-t^2} \, \mathrm{d}t\right] \\
&\overset{\text{by Lemma E.5}}{\leq} s \exp\left\{-\frac{s}{b}\right\} \sqrt{\frac{2\lambda a s}{b}} \frac{b}{2s} \sqrt{\frac{\pi}{2\lambda a}} (1-0) = \frac{\sqrt{\pi b s}}{2} \exp\left\{-\frac{s}{b}\right\}.
\end{aligned}
$$

By Combining the results obtained for I, II_1, and II_2, we can derive the following bound:

$$\begin{aligned}
\mathbb{E}\exp\left\{-\frac{s}{aX+b}\right\} &\le \exp\left\{-\frac{s}{b}\right\} + I + II_1 + II_2 \\
&\le \exp\left\{-\frac{s}{b}\right\} + \frac{\sqrt{2\pi}\lambda as}{2b^2}\exp\left\{-\frac{s}{b+\lambda a}\right\} \\
&\quad + \frac{s}{(b+\lambda a)}\left[\frac{2\lambda as}{b} - \big(b+\lambda a\big)\right]\exp\left\{-\frac{s}{\sqrt{2\lambda a}\vee(b+\lambda a)}\right\} + \frac{\sqrt{\pi bs}}{2}\exp\left\{-\frac{s}{b}\right\} \\
&\le \left[1+\frac{\sqrt{\pi bs}}{2}\right]\exp\left\{-\frac{s}{b}\right\} + \left[\frac{\sqrt{2\pi}\lambda as}{2b^2} + \frac{2\lambda as^2}{b(b+\lambda s)}\right]\exp\left\{-\frac{s}{\sqrt{2\lambda a}\vee(b+\lambda a)}\right\} \\
&\le \left[1+\frac{\sqrt{\pi bs}}{2}\right]\exp\left\{-\frac{s}{b}\right\} + \left[\frac{\sqrt{2\pi}\lambda a}{2b^2} + \frac{2a}{b}\right]s\exp\left\{-\frac{s}{\sqrt{2\lambda a}\vee(b+\lambda a)}\right\} \\
&\overset{\text{by } b\le\sqrt{2\lambda a}\vee(b+\lambda a)}{\le} \left[1+\frac{\sqrt{\pi b}}{2} + \frac{\sqrt{2\pi}\lambda a}{2b^2} + \frac{2a}{b}\right]s\exp\left\{-\frac{s}{\sqrt{2\lambda a}\vee(b+\lambda a)}\right\},
\end{aligned}$$

which gives the result we need. □

Lemma E.5. *For any $x \ge 0$,*

$$\frac{2e^{x^2}}{\sqrt{\pi}}\int_x^\infty e^{-t^2}\,\mathrm{d}t \le 1.$$

Proof.

$$\begin{aligned}
\frac{2e^{x^2}}{\sqrt{\pi}}\int_x^\infty e^{-t^2}\,\mathrm{d}t &= \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-(t^2-x^2)}\,\mathrm{d}t \\
&\le \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-(t-x)^2}\,\mathrm{d}t \\
&= \frac{2}{\sqrt{\pi}}\int_0^\infty \frac{e^{-t^2}}{2\times\frac{1}{2}}\,\mathrm{d}t = \frac{2}{\sqrt{\pi}}\frac{\sqrt{2\pi\times\frac{1}{2}}}{2} = 1.
\end{aligned}$$

□