

Notation	Meaning
K	# of arms
a_i	arm i
$\hat{\mu}_i^{\text{off}}, \hat{\mu}_i$	the estimated mean reward of a_i , offline and online respectively
$\mu_i^{\text{off}}, \mu_i$	the ground truth mean reward of a_i , for offline and online respectively
$p_{i,j}, p_{i,j}^{\text{off}}$	ground truth Ber(p) s.t. a_i wins against a_j for offline and online data respectively
$\hat{p}_{i,j}$	the estimated value of $p_{i,j}$ only using online data
$\hat{p}_{i,j}^{\text{off}}$	the estimated value of $p_{i,j}$ using offline data
$\hat{p}_{i,j}^{\text{hyb}}$	the estimated value of $p_{i,j}$ using offline and online data(hybrid)
Δ_i	$p_{1,i} - \frac{1}{2}$, a_1 refers to the Condorcet winner
$T_{i,j}$	# of times that played (a_i, a_j) online, especially, we use $T_{i,j}(t)$ to refer the # of times that play (a_i, a_j) at the start of t
$T_i(t)$	# of times that a_i is selected for online comparison
N_i	# of times that played a_i in the offline dataset
$X_{i,k}$	the k^{th} reward of a_i
$Y_{i,j,k}$	the k -th feedback of (a_i, a_j)
$V_{i,j}$	a input valid bound for pairwise data, where $V_{i,j} \geq p_{i,j}^{\text{off}} - p_{i,j} $
\mathcal{E}_t	good event, $\mathcal{E}_t = \bigcap_{i,j \in [K]} \left(\mathcal{E}_t(p_{i,j}) \cap \mathcal{E}_t^{\text{hyb}}(p_{i,j}) \right)$
$\omega_{i,j}$	$V_{i,j} + p_{i,j}^{\text{off}} - p_{i,j}$
UCB, UCB ^{hyb}	the UCB using pure online data and hybrid data
$\sigma(\cdot)$	sigmoid function

780 **B More Detail Comparisons with Existing Works**

781 **B.1 Comparison with Cheung and Lyu (2024)**

782 Cheung and Lyu (2024) proposed Policy MIN-UCB algorithm, which is the first to deal with bias dataset under
 783 MAB setting with input valid bias bound $V(a)$, our work extends it to heterogeneous data setting and refine its
 784 regret upper bound. According to Lemma 4.8 in Cheung and Lyu (2024),

$$N_t(a) \geq 32 \cdot \frac{\log(4Kt^4)}{\Delta(a)^2} - T_S(a) \cdot \max \left\{ 1 - \frac{\omega(a)}{\Delta(a)}, 0 \right\}^2,$$

785 where $N_t(a)$ is the number of times arm a is selected online, and $T_S(a)$ denotes its offline sample size. Following
 786 the derivation in Appendix B.2 of Cheung and Lyu (2024), and adapting it to our setting, we simplify the above
 787 inequality into the following form, as used in our Lemma 5 and 6:

$$T' \geq 8 \frac{\delta'}{\Delta'} - N' \cdot \max \left\{ 1 - \frac{\omega'}{\Delta'}, 0 \right\}^2.$$

788 One difference in our setting lies in the preference-based feedback model, which assumes a $1/2 \cdot \sqrt{1/T_{i,j}}$
 789 sub-Gaussian distribution. This changes the confidence radius from $\sqrt{2 \log(1/\delta_t)/T_i}$ to $\sqrt{\log(1/\delta_t)/(2T_{i,j})}$,
 790 leading to a different constant factor in the inequality.

791 However, the theoretical analysis in Cheung and Lyu (2024) is limited in that it only examines the Saving when
 792 there is no online data at all (Case 1a in Appendix B.2, where it directly scales online data $N_t(a) = 0$). In our
 793 analysis, we detailly analyzed when the saving exists (Lemma 5), and when the Saving is big enough that no
 794 online data are needed (Lemma 6). Furthermore, The Case 1a in Cheung and Lyu (2024) is actually a subcase of
 795 our more general analysis, as the following chain of inequalities shows:

$$\begin{aligned} N' \left(1 - \frac{\omega'}{\Delta'} \right)^2 &\geq \frac{8\delta'}{\Delta'^2} \\ &\geq \frac{4\delta'}{\Delta'^2} \cdot \frac{2\Delta' - \omega'}{\Delta'} \\ &= \frac{4\delta'(2\Delta' - \omega')}{\Delta'^3}, \end{aligned}$$

where the first inequality corresponds to the condition in Case 1a in Cheung and Lyu (2024), the second holds due to the assumption $\Delta' \geq \omega'$, and the resulting expression aligns with the condition stated in our Lemma 5.

B.2 Comparison with Wang et al. (2025a)

Wang et al. (2025a) proposed an online hybrid MAB setting where the agent could observe both rewards and pairwise preferences per round. To be specific, Wang et al. (2025a) proposed two algorithms to address the hybrid feedback multi-armed bandit (MAB) problem: *Elimination Fusion* (EIMFUSION) and *Decomposition Fusion* (DECOFUSION). EIMFUSION eliminates an arm a_j if either $\text{UCB}(a_j) \leq \text{LCB}(a_i)$ or $\text{UCB}(a_j, a_i) < \frac{1}{2}$ holds, this method closely related to our HybElimUCB-RA. DECOFUSION employs a more complex and cooperative policy between dueling and reward-based feedback, randomly using one of the feedback to explore the arm space and the other to exploit. Under the regret definition $R_T = \alpha R_T^R + (1 - \alpha) R_T^D$, where D refers "dueling" and R refers "reward", and α is an input parameter. The regret upper bound of EIMFUSION and DECOFUSION algorithms are:

$$\mathbb{E}[R_T] \leq O \left(\sum_{k \neq 1} \frac{(\alpha \Delta_k^{(R)} + (1 - \alpha) \Delta_k^{(D)}) \log T}{\max\{(\Delta_k^{(R)})^2, (\Delta_k^{(D)})^2 / K\}} \right)$$

and

$$\mathbb{E}[R_T] \leq O \left(\sum_{k \neq 1} \frac{\log T}{\max\{\Delta_k^{(R)} / \alpha, \Delta_k^{(D)} / (1 - \alpha)\}} \right)$$

respectively.

In our setting, the fusion arises from the integration of offline and online data. Within each source (offline or online), the data is homogeneous; therefore, in the online learning phase, only one type of feedback is available. To effectively leverage both sources, we adopt the Bradley–Terry model and construct a unified estimator for $p_{i,j}$, capturing the pairwise preference probability. In contrast, the method in Wang et al. (2025a) treats the estimation of each arm or arm pair independently, the joint utility of heterogeneous data comes by constructing candidate set together.

B.3 Comparison with Zoghi et al. (2014)

Zoghi et al. (2014) proposed the RUCB algorithm to address preference-based feedback without relying on the strong stochastic transitivity and stochastic triangle inequality assumptions required by earlier works such as Yue et al. (2012). However, RUCB does not fully exploit the information in its candidate set, as it selects the second arm c (corresponding to $A_2(t)$ in HybUCB-AR) uniformly at random. As a result, it can only guarantee a regret bound of

$$O \left(\sum_{i \leq j} \frac{(\Delta_i + \Delta_j) \log T}{\min\{\Delta_i^2, \Delta_j^2\}} \right).$$

In this work, we improve upon this by maximizing informative pair in the candidate set:

$$(A_1(t), A_2(t)) = \arg \max_{a_i, a_j \in \mathcal{C}_t \times \mathcal{C}_t} \text{UCB}(a_i, a_j).$$

This change allows us to better utilize candidate set information, leading to a tighter regret bound of

$$O \left(\sum_{i \leq j} \frac{(\Delta_i + \Delta_j) \log T}{\max\{\Delta_i^2, \Delta_j^2\}} \right),$$

which, to the best of our knowledge, is the first result of this kind in the dueling bandit setting without strong structural assumptions.

B.4 Comparison with Qu et al. (2024)

Qu et al. (2024) proposed a hybrid transfer reinforcement learning (HTRL) algorithm, HySRL, which selectively uses historical data exhibiting shifted dynamics to reduce the sample complexity of online reinforcement learning. Similar to the findings in Cheung and Lyu (2024), it proved in general HTRL, when no additional knowledge or restriction is applied to historical dataset, the sample complexity could not be improved.

To address the distributional shift between offline and online samples, HySRL introduces the concept of β -separable shift, which classifies offline data as either distributionally identical to or different from the online environment. For offline samples deemed identical, the algorithm directly incorporates them into online

estimation. In contrast, samples identified as distributionally different are entirely excluded during the online learning phase. Under this framework, the sample complexity of HySRL is:

$$\tilde{O}\left(\min\left(\frac{H^3 S A}{\epsilon^2}, \frac{H^3 |\mathcal{B}|}{\epsilon^2} + \frac{H^2 S^2 A}{(\sigma\beta)^2}\right)\right),$$

where α, β, ϵ are their input parameters, H is the length of episode, S, A refer to state and action space respectively.

In Qu et al. (2024), the distribution shift region is detected online, and under the theoretical sample complexity bounds, the shifted region can be correctly identified with high probability. Despite its flexibility, this approach relies on strong assumptions regarding the β -separability definition and the choice of β . Moreover, it does not account for the sample size of the historical dataset. In contrast, our method assumes a valid bias bound as input, which eliminates the need for warm-start estimation of the shift region, but requires stronger prior knowledge about the arms or the transition probability functions.

C Proof for Theorem 1

C.1 Sub-Gaussian Properties

Lemma 1 (sub-Gaussian Properties of Estimators). *Let $\hat{p}_{i,j}$ denote an estimator of the preference probability between arms i and j . The sub-Gaussian parameter of $\hat{p}_{i,j}$ depends on the type of data used for estimation:*

1. For relative preference data, where

$$\hat{p}_{i,j} = \frac{1}{T_{i,j}} \sum_{k=1}^{T_{i,j}} Y_{i,j,k},$$

the estimator is $\frac{1}{2} \sqrt{\frac{1}{T_{i,j}}}$ sub-Gaussian.

2. For stochastic utility data, where

$$\hat{p}_{i,j} = \sigma \left(\frac{1}{N_i} \sum_{k=1}^{N_i} X_{i,k} - \frac{1}{N_j} \sum_{k=1}^{N_j} X_{j,k} \right),$$

the estimator is $\frac{1}{2} \sqrt{\frac{N_i + N_j}{N_i N_j}}$ sub-Gaussian.

3. For a hybrid estimator combining both types of data:

$$\hat{p}_{i,j}^{hyb} = \frac{1}{T_{i,j} + \frac{N_i N_j}{N_i + N_j}} \left(\sum_{k=1}^{T_{i,j}} Y_{i,j,k} \right) + \frac{\frac{N_i N_j}{N_i + N_j}}{T_{i,j} + \frac{N_i N_j}{N_i + N_j}} \sigma \left(\frac{1}{N_i} \sum_{k=1}^{N_i} X_{i,k} - \frac{1}{N_j} \sum_{k=1}^{N_j} X_{j,k} \right),$$

the sub-Gaussian parameter is $\frac{1}{2} \sqrt{\frac{1}{T_{i,j} + \frac{N_i N_j}{N_i + N_j}}}$.

Proof. We prove each case separately.

1. **Pure Relative Preference Estimator.** Let $Y \sim \text{Bernoulli}(p)$ with values in $[0, 1]$. By Hoeffding's Lemma, for any $\lambda \in \mathbb{R}$,

$$\mathbb{E}[\exp(\lambda(Y - \mathbb{E}[Y]))] \leq \exp\left(\frac{\lambda^2(1-0)^2}{8}\right) = \exp\left(\frac{\lambda^2}{8}\right).$$

This corresponds to a $\frac{1}{2}$ sub-Gaussian variable, since

$$\frac{\lambda^2}{8} = \frac{\lambda^2(1/2)^2}{2}.$$

As $\hat{p}_{i,j}$ is the average of $T_{i,j}(t)$ independent Bernoulli samples, Lemma 5.4 in Lattimore and Szepesvári (2020) imply that $\hat{p}_{i,j}$ is $\frac{1}{2} \sqrt{\frac{1}{T_{i,j}(t)}}$ sub-Gaussian.

860 **2. Pure Stochastic Utility Estimator.** Assume that each $X_{i,k}$ and $X_{j,k}$ is 1 sub-Gaussian. Since
 861 sub-Gaussianity is preserved under averaging, we have:

$$\frac{1}{N_i} \sum_{k=1}^{N_i} X_{i,k} \sim \frac{1}{\sqrt{N_i}} \text{ sub-Gaussian}, \quad \frac{1}{N_j} \sum_{k=1}^{N_j} X_{j,k} \sim \frac{1}{\sqrt{N_j}} \text{ sub-Gaussian}.$$

862 Therefore, the difference of these averages is sub-Gaussian with parameter $\sqrt{1/N_i + 1/N_j}$. Since
 863 sigmoid function is 1/4-Lipschitz continuous, applying Lemma 10 yields the final sub-Gaussian
 864 parameter:

$$2 \cdot \frac{1}{4} \cdot \sqrt{\frac{1}{N_i} + \frac{1}{N_j}} = \frac{1}{2} \sqrt{\frac{1}{N_i} + \frac{1}{N_j}}.$$

865 **3. Hybrid Estimator.** Let

$$\alpha := T_{i,j}, \quad \beta := \frac{N_i N_j}{N_i + N_j}, \quad Z := \sigma \left(\frac{1}{N_i} \sum_{k=1}^{N_i} X_{i,k} - \frac{1}{N_j} \sum_{k=1}^{N_j} X_{j,k} \right).$$

866 The estimator is a convex combination:

$$\hat{p}_{i,j}^{\text{hyb}} = \frac{1}{\alpha + \beta} \left(\sum_{k=1}^{\alpha} Y_{i,j,k} + \beta Z \right).$$

867 Since $\sum_{k=1}^{\alpha} Y_{i,j,k}$ is $\frac{1}{2} \sqrt{\alpha}$ sub-Gaussian and Z is $\frac{1}{2} \sqrt{\frac{1}{N_i} + \frac{1}{N_j}}$ sub-Gaussian, by Lemma 5.4 Latti-
 868 more and Szepesvári (2020), the entire sum has a sub-Gaussian parameter:

$$\sigma = \frac{1}{\alpha + \beta} \sqrt{\alpha^2 \cdot \frac{1}{4\alpha} + \beta^2 \cdot \frac{1}{4} \left(\frac{1}{N_i} + \frac{1}{N_j} \right)} = \frac{1}{2\sqrt{\alpha + \beta}}.$$

869 Substituting back yields the desired result.

870 □

871 C.2 Justification of Valid Bias Bound

872 **Lemma 2.** Given $V_i \geq |\mu_i^{\text{off}} - \mu_i|, \forall i \in [K]$, under the Bradley-Terry model, we could derive the valid bias
 873 upper bound for pairwise term as $V_{i,j} \geq |p_{i,j}^{\text{off}} - p_{i,j}|, \forall i, j \in [K]$.

874 *Proof.* Since $V_i \geq \mu_i - \mu_i^{\text{off}}$ and $V_j \geq \mu_j^{\text{off}} - \mu_j$, we have $V_i + V_j \geq (\mu_i - \mu_j) - (\mu_i^{\text{off}} - \mu_j^{\text{off}})$, which further
 875 derives

$$\sigma(V_i + V_j) \geq \sigma((\mu_i - \mu_j) - (\mu_i^{\text{off}} - \mu_j^{\text{off}})) \geq \sigma(\mu_i - \mu_j) - \sigma(\mu_i^{\text{off}} - \mu_j^{\text{off}}) = p_{i,j} - p_{i,j}^{\text{off}}.$$

876 The first inequality is by the monotonically increasing property of sigmoid function, and the second inequality is
 877 by the property of sigmoid function where $\sigma(x + y) \leq \sigma(x) + \sigma(y), \forall x, y \in \mathbb{R}$.

878 Similarly, we could derive $\sigma(V_i + V_j) \geq p_{i,j}^{\text{off}} - p_{i,j}$, hence $\sigma(V_i + V_j) \geq |p_{i,j}^{\text{off}} - p_{i,j}|$, completes the proof. □

879 C.3 Good Event and High Probability Bound

880 Define good event $\mathcal{E}_t = \bigcap_{i,j \in [K]} \left(\mathcal{E}_t(p_{i,j}) \cap \mathcal{E}_t^{\text{hyb}}(p_{i,j}) \right)$ where

$$\mathcal{E}_t(p_{i,j}) = \left\{ p_{i,j} \leq \text{UCB}(a_i, a_j) \leq p_{i,j} + 2\sqrt{\frac{\log(1/\delta_t)}{2T_{i,j}}} \right\} \quad (11)$$

$$\mathcal{E}_t^{\text{hyb}}(p_{i,j}) = \left\{ p_{i,j} \leq \text{UCB}^{\text{hyb}}(a_i, a_j) \leq p_{i,j} + 2\sqrt{\frac{\log(1/\delta_t)}{2(T_{i,j}(t) + \frac{N_i N_j}{N_i + N_j})}} + \frac{\frac{N_i N_j}{N_i + N_j}}{T_{i,j}(t) + \frac{N_i N_j}{N_i + N_j}} \omega_{i,j} \right\} \quad (12)$$

881 with $\omega_{i,j} := V_{i,j} + p_{i,j}^{\text{off}} - p_{i,j}$, we have the following lemma:

882 **Lemma 3.** The good event satisfies the following lower bound: $\Pr(\mathcal{E}_t) \geq 1 - 2K(K+1)\delta_t$.

883 *Proof.* This proof largely follows the argument of Appendix B.1 in Cheung and Lyu (2024). We begin by
 884 applying the Chernoff bound for sub-Gaussian random variables. Consider the estimator $\hat{p}_{i,j}$ of the true pairwise
 885 comparison probability $p_{i,j}$. Since the difference $X = \hat{p}_{i,j} - p_{i,j}$ is sub-Gaussian with parameter $\frac{1}{2}\sqrt{1/T_{i,j}(t)}$
 886 by Lemma 1, we can apply the Chernoff inequality (see Lemma 8) to obtain:

$$\Pr\left(|\hat{p}_{i,j} - p_{i,j}| \geq \sqrt{\frac{\log(1/\delta_t)}{2T_{i,j}(t)}}\right) \leq 2\delta_t.$$

887 This establishes the condition for the 'bad event' $\mathcal{E}_t^c(p_{i,j})$ by rearranging the inequality above.

888 Next, we analyze the 'bad event' $\mathcal{E}_t^{\text{hyb},c}(p_{i,j})$, which occurs when either of the following two conditions is
 889 violated. The first case corresponds to a violation of the first inequality in Equation (12). For this case, we aim
 890 to derive an upper bound on the probability of violation:

$$\Pr\left(p_{i,j} \geq \text{UCB}^{\text{hyb}}(a_i, a_j)\right),$$

891 where

$$\text{UCB}^{\text{hyb}}(a_i, a_j) = \frac{T_{i,j}(t)\hat{p}_{i,j} + \frac{N_i N_j}{N_i + N_j}\sigma(\hat{X}_i - \hat{X}_j)}{T_{i,j}(t) + \frac{N_i N_j}{N_i + N_j}} + \sqrt{\frac{\log(1/\delta_t)}{2\left(T_{i,j}(t) + \frac{N_i N_j}{N_i + N_j}\right)}} + \frac{\frac{N_i N_j}{N_i + N_j}}{T_{i,j}(t) + \frac{N_i N_j}{N_i + N_j}} V_{i,j}.$$

892 This probability can be rewritten as:

$$\begin{aligned} & \Pr\left(\frac{T_{i,j}(t)p_{i,j} + \frac{N_i N_j}{N_i + N_j}p_{i,j}^{\text{off}}}{T_{i,j}(t) + \frac{N_i N_j}{N_i + N_j}} + \frac{\frac{N_i N_j}{N_i + N_j}(p_{i,j} - p_{i,j}^{\text{off}})}{T_{i,j}(t) + \frac{N_i N_j}{N_i + N_j}} \geq \right. \\ & \quad \left. \frac{T_{i,j}(t)\hat{p}_{i,j} + \frac{N_i N_j}{N_i + N_j}\sigma(\hat{X}_i - \hat{X}_j)}{T_{i,j}(t) + \frac{N_i N_j}{N_i + N_j}} + \sqrt{\frac{\log(1/\delta_t)}{2\left(T_{i,j}(t) + \frac{N_i N_j}{N_i + N_j}\right)}} + \frac{\frac{N_i N_j}{N_i + N_j}}{T_{i,j}(t) + \frac{N_i N_j}{N_i + N_j}} V_{i,j} \right) \\ & \leq \Pr\left(\frac{T_{i,j}(t)p_{i,j} + \frac{N_i N_j}{N_i + N_j}p_{i,j}^{\text{off}}}{T_{i,j}(t) + \frac{N_i N_j}{N_i + N_j}} \geq \frac{T_{i,j}(t)\hat{p}_{i,j} + \frac{N_i N_j}{N_i + N_j}\sigma(\hat{X}_i - \hat{X}_j)}{T_{i,j}(t) + \frac{N_i N_j}{N_i + N_j}} + \sqrt{\frac{\log(1/\delta_t)}{2\left(T_{i,j}(t) + \frac{N_i N_j}{N_i + N_j}\right)}} \right) \\ & \leq \delta_t, \end{aligned}$$

893 where the penultimate inequality uses the fact that $V_{i,j} \geq |p_{i,j}^{\text{off}} - p_{i,j}|$, and the last inequality comes by applying
 894 the Chernoff inequality to the hybrid estimator, which is $\frac{1}{2}\sqrt{\frac{1}{T_{i,j}(t) + \frac{N_i N_j}{N_i + N_j}}}$ sub-Gaussian (by Lemma 1).

895 For the second case, consider:

$$\Pr\left(\text{UCB}^{\text{hyb}}(a_i, a_j) \geq p_{i,j} + 2\sqrt{\frac{\log(1/\delta_t)}{2\left(T_{i,j}(t) + \frac{N_i N_j}{N_i + N_j}\right)}} + \frac{\frac{N_i N_j}{N_i + N_j}}{T_{i,j}(t) + \frac{N_i N_j}{N_i + N_j}}(V_{i,j} + p_{i,j}^{\text{off}} - p_{i,j})\right).$$

896 This can be written as:

$$\begin{aligned} & \Pr\left(\frac{T_{i,j}(t)\hat{p}_{i,j} + \frac{N_i N_j}{N_i + N_j}\sigma(\hat{X}_i - \hat{X}_j)}{T_{i,j}(t) + \frac{N_i N_j}{N_i + N_j}} \geq p_{i,j} + \sqrt{\frac{\log(1/\delta_t)}{2\left(T_{i,j}(t) + \frac{N_i N_j}{N_i + N_j}\right)}} + \frac{\frac{N_i N_j}{N_i + N_j}}{T_{i,j}(t) + \frac{N_i N_j}{N_i + N_j}}(p_{i,j}^{\text{off}} - p_{i,j})\right) \\ & = \Pr\left(\frac{T_{i,j}(t)\hat{p}_{i,j} + \frac{N_i N_j}{N_i + N_j}\sigma(\hat{X}_i - \hat{X}_j)}{T_{i,j}(t) + \frac{N_i N_j}{N_i + N_j}} \geq \frac{T_{i,j}(t)p_{i,j} + \frac{N_i N_j}{N_i + N_j}p_{i,j}^{\text{off}}}{T_{i,j}(t) + \frac{N_i N_j}{N_i + N_j}} + \sqrt{\frac{\log(1/\delta_t)}{2\left(T_{i,j}(t) + \frac{N_i N_j}{N_i + N_j}\right)}} \right) \\ & \leq \delta_t. \end{aligned}$$

897 The final inequality is by applying the Chernoff inequality.

898 Since there are $\frac{K(K+1)}{2}$ distinct pairs of arms, and for each pair the probability of \mathcal{E}_t is at most $4\delta_t$, applying
 899 this to all arms gives:

$$\begin{aligned}
1 - \Pr(\mathcal{E}_t) &= \Pr(\mathcal{E}_t^c) \\
&\leq \sum_{i=1}^K \sum_{i \leq j} \Pr(\mathcal{E}_t^c(p_{i,j}) + \mathcal{E}_t^{\text{hyb},c}(p_{i,j})) \quad (\text{union bound}) \\
&\leq 2K(K+1)\delta_t \quad \left(\text{since } \Pr(\mathcal{E}_t^c(p_{i,j}) + \mathcal{E}_t^{\text{hyb},c}(p_{i,j})) \leq 4\delta_t \right)
\end{aligned}$$

and hence,

$$\Pr(\mathcal{E}_t) \geq 1 - 2K(K+1)\delta_t,$$

which concludes the proof. \square

C.4 Property of Algorithm 1

Lemma 4. (HybUCB-AR's property) Conditioned on \mathcal{E}_t , the arm pair (a_i, a_j) will **not** be selected in round t if it satisfies the following Inequality:

$$\max\{\Delta_i, \Delta_j\} > \min \left\{ 4\sqrt{\frac{\log(1/\delta_t)}{2T_{i,j}(t)}}, 4\sqrt{\frac{\log(1/\delta_t)}{2\left(T_{i,j}(t) + \frac{N_i N_j}{N_i + N_j}\right)}} + 2\frac{\frac{N_i N_j}{N_i + N_j}}{T_{i,j}(t) + \frac{N_i N_j}{N_i + N_j}} \omega_{i,j} \right\}.$$

Proof. Conditioned on \mathcal{E}_t , Suppose arm pair (a_i, a_j) is selected, there are two possible cases for this selection:

1. $(A_1(t), A_2(t)) = (a_i, a_j)$,
2. $(A_1(t), A_2(t)) = (a_j, a_i)$.

We analyze the first case, while the second case follows symmetrically. By HybUCB-AR's selection criterion (Line-7), when (a_i, a_j) is chosen, it must hold that:

$$1 - p_{i,j} + 2\sqrt{\frac{\log(1/\delta_t)}{2T_{i,j}(t)}} = p_{j,i} + 2\sqrt{\frac{\log(1/\delta_t)}{2T_{i,j}(t)}} \geq \text{UCB}(a_j, a_i) \geq \frac{1}{2}, \quad (13)$$

and

$$1 - p_{j,i} + 2\sqrt{\frac{\log(1/\delta_t)}{2T_{i,j}(t)}} = p_{i,j} + 2\sqrt{\frac{\log(1/\delta_t)}{2T_{i,j}(t)}} \geq \text{UCB}(a_i, a_j) \geq \frac{1}{2}. \quad (14)$$

Equation (13) and (14) implies that

$$p_{i,j} \leq \frac{1}{2} + 2\sqrt{\frac{\log(1/\delta_t)}{2T_{i,j}(t)}} \quad \text{and} \quad p_{j,i} \leq \frac{1}{2} + 2\sqrt{\frac{\log(1/\delta_t)}{2T_{i,j}(t)}}. \quad (15)$$

Given $\Delta_i > 4\sqrt{\frac{\log(1/\delta_t)}{2T_{i,j}(t)}}$, we have the following:

$$\text{UCB}(a_i, a_j) \leq p_{i,j} + 2\sqrt{\frac{\log(1/\delta_t)}{2T_{i,j}(t)}} \leq \frac{1}{2} + 4\sqrt{\frac{\log(1/\delta_t)}{2T_{i,j}(t)}} < p_{1,i} \leq \min \left\{ \text{UCB}(a_1, a_i), \text{UCB}^{\text{byb}}(a_1, a_i) \right\}, \quad (16)$$

The first and fourth inequalities in Equation (16) follow from condition \mathcal{E}_t , while the second inequality is derived from the first inequality in Equation (15). The third inequality holds by the given condition $\Delta_i > 4\sqrt{\frac{\log(1/\delta_t)}{2T_{i,j}(t)}}$. Recalling that HybUCB-AR's selection criterion (Line 10) satisfies $(a_i, a_j) = \arg \max_{a_i, a_j \in \mathcal{C}_t} \min \{ \text{UCB}(a_i, a_j), \text{UCB}^{\text{hyb}}(a_i, a_j) \}$, we observe that (16) leads to a contradiction.

Following the same reasoning, the condition $\Delta_j > 4\sqrt{\frac{\log(1/\delta_t)}{2T_{i,j}(t)}}$ would also lead to an analogous contradiction. This implies that the arm pair (a_i, a_j) cannot be selected when the following holds:

$$\max\{\Delta_i, \Delta_j\} > 4\sqrt{\frac{\log(1/\delta_t)}{2T_{i,j}(t)}}. \quad (17)$$

Since the candidate set is constructed by $\mathcal{C}_t = \mathcal{C}_t^{\text{on}} \cap \mathcal{C}_t^{\text{hyb}}$, applying the same argument to $\mathcal{C}_t^{\text{hyb}}$ yields that (a_i, a_j) will also not be selected if:

$$\max\{\Delta_i, \Delta_j\} > 4\sqrt{\frac{\log(1/\delta_t)}{2\left(T_{i,j} + \frac{N_i N_j}{N_i + N_j}\right)}} + 2\frac{\frac{N_i N_j}{N_i + N_j}}{T_{i,j} + \frac{N_i N_j}{N_i + N_j}} \omega_{i,j}. \quad (18)$$

Combining Equation (17) and (18) leads to the desired result. By symmetry, the analysis for the second case $(A_1(t), A_2(t)) = (a_j, a_i)$ mirrors that of (a_i, a_j) and leads to the same conclusion, as desired. \square

C.5 Proof of Instance Dependent Regret Upper Bound

Before we provide the proof, we present a general lemma concerning the solution to the key inequalities derived in Lemma 4.

Lemma 5. Consider the notation N' , T' , ω' , δ' , and Δ' , these variables are defined to simplify the structure of the equations while preserving their form, enhancing clarity and generality. Let T' be the smallest integer satisfying

$$\Delta' > \min \left\{ 2\sqrt{\frac{\delta'}{T'}}, 2\sqrt{\frac{\delta'}{T' + N'}} + \frac{N'}{T' + N'}\omega' \right\}, \quad (19)$$

then

$$T' > \begin{cases} \max \left\{ \frac{2\delta' + N'\Delta'(\omega' - \Delta') + \sqrt{4\delta'^2 + 4\delta'N'\omega'\Delta'}}{\Delta'^2}, 0 \right\} & \text{if Condition Saving,} \\ \frac{4\delta'}{\Delta'^2} & \text{otherwise,} \end{cases}$$

where:

$$\text{Condition Saving} = (2\omega' \leq \Delta') \text{ or } \left(N' \geq \frac{4\delta'(2\omega' - \Delta')}{\Delta'(\omega' - \Delta')^2} \right).$$

Proof. As established, Equation (19) represent bounds derived from the HybUCB-AR's framework, corresponding to pure online and pure hybrid data scenarios, respectively. Our proposed algorithm requires that T' satisfies at least one of these conditions. The use of offline data reduces regret when the smallest T' satisfying

$$\Delta' > 2\sqrt{\frac{\delta'}{T'}} \quad (20)$$

is less than that required by

$$\Delta' > 2\sqrt{\frac{\delta'}{T' + N'}} + \frac{N'}{T' + N'}\omega'. \quad (21)$$

For Equation (20), direct manipulation yields:

$$T' > \frac{4\delta'}{\Delta'^2}. \quad (22)$$

For Equation (21), define $\sqrt{T' + N'} = x$, $A = 2\sqrt{\delta'}$, and $B = N'\omega'$. The inequality becomes:

$$\frac{A}{x} + \frac{B}{x^2} < \Delta'.$$

Solving this for $x \geq 0$, we obtain:

$$\begin{aligned} x &> \frac{A + \sqrt{A^2 + 4B\Delta'}}{2\Delta'} \\ x^2 &> \frac{A^2 + 2B\Delta' + \sqrt{A^4 + 4A^2B\Delta'}}{2\Delta'^2}. \end{aligned}$$

This implies:

$$T' > \frac{2\delta' + N'\omega'\Delta' + \sqrt{4\delta'^2 + 4\delta'N'\omega'\Delta'}}{\Delta'^2} - N' = \frac{2\delta' + N'\Delta'(\omega' - \Delta') + \sqrt{4\delta'^2 + 4\delta'N'\omega'\Delta'}}{\Delta'^2}.$$

Since $T' \geq 0$, we have:

$$T' > \max \left\{ \frac{2\delta' + N'\Delta'(\omega' - \Delta') + \sqrt{4\delta'^2 + 4\delta'N'\omega'\Delta'}}{\Delta'^2}, 0 \right\}. \quad (23)$$

Combining these, the general bound is:

$$T' > \min \left\{ \frac{4\delta'}{\Delta'^2}, \max \left\{ \frac{2\delta' + N'\Delta'(\omega' - \Delta') + \sqrt{4\delta'^2 + 4\delta'N'\omega'\Delta'}}{\Delta'^2}, 0 \right\} \right\}. \quad (24)$$

941 Next, we identify conditions under which Equation (21) yields a smaller T' than Equation (20). Define the
 942 Saving as:

$$\text{Saving} = \frac{4\delta' - \left(2\delta' + N'\Delta'(\omega' - \Delta') + \sqrt{4\delta'^2 + 4\delta'N'\omega'\Delta'}\right)}{\Delta'^2}.$$

943 From Equation (24), Saving are positive only if $\Delta' > \omega'$, since $\omega' - \Delta' < 0$ is the sole negative contribution,
 944 potentially reducing T' . This condition also ensures the left-hand side of the subsequent second inequality
 945 (Equation (25)) remains non-negative.

946 We require Saving ≥ 0 :

$$\begin{aligned} \frac{4\delta' - 2\delta' - N'\Delta'(\omega' - \Delta') - \sqrt{4\delta'^2 + 4\delta'N'\omega'\Delta'}}{\Delta'^2} &\geq 0 \\ 2\delta' - N'\Delta'(\omega' - \Delta') &\geq \sqrt{4\delta'^2 + 4\delta'N'\omega'\Delta'} \\ N'^2\Delta'^2(\omega' - \Delta')^2 - 4\delta'N'\Delta'(\omega' - \Delta') &\geq 4\delta'N'\omega'\Delta' \\ N'^2\Delta'^2(\omega' - \Delta')^2 &\geq 4\delta'N'\Delta'(2\omega' - \Delta'). \end{aligned} \quad (25)$$

947 Hence the final inequality, guaranteeing that saving ≥ 0 , holds if either

$$\begin{aligned} \text{(i)} \quad 2\omega' &\leq \Delta'; \\ \text{or (ii)} \quad 2\Delta' &> 2\omega' > \Delta' \quad \text{and} \quad N' \geq \frac{4\delta'(2\omega' - \Delta')}{\Delta'(\omega' - \Delta')^2}. \end{aligned}$$

948 These conditions together validate the case distinction stated in the lemma. \square

949 *Proof.* (Complete Proof of Instance Dependent Regret Upper Bound in Theorem 1)

950 By Lemma 3 and Lemma 4, take $\Delta' = \frac{1}{2} \max\{\Delta_i, \Delta_j\}$, $\delta' = \frac{1}{2} \log(1/\delta_t)$, $\omega' = w_{i,j}$, $N' = \frac{N_i N_j}{N_i + N_j}$, $T' =$
 951 $T_{i,j}(t)$, under the good event \mathcal{E}_t , the number of times the arm pair (a_i, a_j) is played satisfies:

952 If $2 \max\{\Delta_i, \Delta_j\} > 2\omega_{i,j} > \max\{\Delta_i, \Delta_j\}$ and $\frac{N_i N_j}{N_i + N_j} \geq \frac{2 \log(1/\delta_t)(2\omega_{i,j} - \max\{\Delta_i, \Delta_j\})}{\max\{\Delta_i, \Delta_j\}(\omega_{i,j} - \max\{\Delta_i, \Delta_j\})^2}$, then:

$$T_{i,j}(t) \leq \max \left\{ \frac{4 \log(1/\delta_t) + \frac{N_i N_j}{N_i + N_j} \max\{\Delta_i, \Delta_j\} (2\omega_{i,j} - \max\{\Delta_i, \Delta_j\}) + \sqrt{D}}{\max\{\Delta_i^2, \Delta_j^2\}}, 0 \right\}, \quad (26)$$

953 where $D = \log^2(1/\delta_t) + \log(1/\delta_t) \cdot \frac{N_i N_j}{N_i + N_j} \omega_{i,j} \max\{\Delta_i, \Delta_j\}$,

954 otherwise:

$$T_{i,j}(t) \leq \frac{8 \log(1/\delta_t)}{\max\{\Delta_i^2, \Delta_j^2\}}. \quad (27)$$

955 Define $g(a_i, a_j)$ and $f(a_i, a_j)$ as follows. Here, $g(a_i, a_j)$ denotes the upper bound of $T_{i,j}(t)$ in the presence
 956 of Saving, while $f(a_i, a_j)$ characterizes the condition under which Saving occur. Specifically, Saving exist if
 957 $f(a_i, a_j) < 0$.

$$\begin{aligned} g(a_i, a_j) &= \max \left\{ \frac{4 \log(1/\delta_t)}{\max\{\Delta_i^2, \Delta_j^2\}} + \frac{\frac{N_i N_j}{N_i + N_j} \max\{\Delta_i, \Delta_j\} \cdot (2\omega_{i,j} - \max\{\Delta_i, \Delta_j\}) + \sqrt{D}}{\max\{\Delta_i^2, \Delta_j^2\}}, 0 \right\}, \\ f(a_i, a_j) &= \begin{cases} -1 & \text{if } 4\omega_{i,j} \leq \max\{\Delta_i, \Delta_j\}, \\ \frac{4 \log(1/\delta_t)(4\omega_{i,j} - \max\{\Delta_i, \Delta_j\})}{\max\{\Delta_i, \Delta_j\}(2\omega_{i,j} - \max\{\Delta_i, \Delta_j\})^2} - N_{i,j} & \text{if } 2 \max\{\Delta_i, \Delta_j\} > 4\omega_{i,j} > \max\{\Delta_i, \Delta_j\}, \\ 1 & \text{otherwise.} \end{cases} \end{aligned} \quad (28)$$

958 The expected regret is:

$$\begin{aligned}
\text{Reg}(T) &= \sum_{t=1}^T \mathbb{E} \left[\left(\frac{\Delta_{A_1(t)} + \Delta_{A_2(t)}}{2} \right) \cdot [\mathbf{1}(\mathcal{E}_t) + \mathbf{1}(\mathcal{E}_t^c)] \right] \\
&\leq \sum_{t=1}^T \mathbb{E} \left[\left(\frac{\Delta_{A_1(t)} + \Delta_{A_2(t)}}{2} \right) \cdot \mathbf{1}(\mathcal{E}_t) \right] + \Delta_{\max} \sum_{t=1}^T \mathbb{E} [\mathbf{1}(\mathcal{E}_t^c)] \\
&= \sum_{i \leq j} \frac{\Delta_i + \Delta_j}{2} \mathbb{E}[T_{i,j}(T) \cdot \mathbf{1}(\mathcal{E}_t)] + \Delta_{\max} \sum_{t=1}^T \mathbb{E} [\mathbf{1}(\mathcal{E}_t^c)],
\end{aligned}$$

Let $\delta_t = 1/(2K(1+K)t^2)$, under the bad event, the regret is:

$$\Delta_{\max} \sum_{t=1}^T \mathbb{E} [\mathbf{1}(\mathcal{E}_t^c)] = \Delta_{\max} \sum_{t=1}^T \mathbb{E} [\mathbf{1}(\mathcal{E}_t)] = \sum_{t=1}^T 2K(1+K)\delta_t \Delta_{\max} = \sum_{t=1}^T \Delta_{\max} \frac{1}{t^2} \leq \frac{\pi^2}{6} \Delta_{\max}, \quad (29)$$

under the good event, the regret is:

$$\begin{aligned}
\sum_{i \leq j} \frac{\Delta_i + \Delta_j}{2} \mathbb{E}[T_{i,j}(T) \cdot \mathbf{1}(\mathcal{E}_t)] &= \sum_{\substack{i \leq j, \\ f(a_i, a_j) < 0}} g(a_i, a_j) \frac{\Delta_i + \Delta_j}{2} + \sum_{\substack{i \leq j, \\ f(a_i, a_j) \geq 0}} \frac{8(\Delta_i + \Delta_j) \log(\sqrt{2K(1+K)}T)}{\max\{\Delta_i^2, \Delta_j^2\}}.
\end{aligned} \quad (30)$$

Combining Equation (29) and (30), we obtain the desired result. The simplified regret upper bound is obtained by applying the following inequality to $g(a_i, a_j)$ in Equation (30):

$$\sqrt{x^2 + 2ax} < a + x, \quad (31)$$

where $x = \log(1/\delta_t)$, $a = \frac{N_i N_j}{2(N_i + N_j)} \max\{\Delta_i, \Delta_j\} \omega_{i,j}$, we derive:

$$\text{Reg}(T) \leq \frac{\pi^2}{6} \Delta_{\max} + \sum_{i \leq j} \max \left\{ \frac{\Delta_i + \Delta_j}{2} \left[\frac{16 \log(\sqrt{2K(K+1)}T)}{\max\{\Delta_i^2, \Delta_j^2\}} - \frac{N_i N_j}{N_i + N_j} \frac{\max\{\max\{\Delta_i, \Delta_j\} - 4\omega_{i,j}, 0\}}{\max\{\Delta_i, \Delta_j\}} \right], 0 \right\}. \quad (32)$$

□

C.6 Saving: Further Analysis

Remark 7. To facilitate concise notation, we denote by $T_{ij}(\text{UCB})$ and $T_{ij}(\text{UCB}^{\text{hyb}})$ the upper bounds on T_{ij} under the UCB and UCB^{hyb} selection rules, respectively. Specifically,

$$\begin{aligned}
T_{ij}(\text{UCB}) &= \frac{4\delta'}{\Delta'^2}, \\
T_{ij}(\text{UCB}^{\text{hyb}}) &= \max \left\{ \frac{2\delta' + N' \Delta' (\omega' - \Delta') + \sqrt{4\delta'^2 + 4\delta' N' \omega' \Delta'}}{\Delta'^2}, 0 \right\}.
\end{aligned}$$

This is the result we derived from Equation (22) and (23) in Lemma 5. In the context of Theorem 1, $T_{ij}(\text{UCB})$ and $T_{ij}(\text{UCB}^{\text{hyb}})$ becomes Equation (27) and (26) respectively.

After establishing Theorem 1, we provided a brief analysis of the "Saving" term in its regret bound in Remark 3. Note that Theorem 1 is a simplified version derived by applying Equation (31) to Equations (29) and (30). Below, we present a more precise analysis of the conditions under which "Saving" exists. By Equation (28), we could divide the condition into three different cases:

1. **Significant Saving:** When $\max\{\Delta_i, \Delta_j\} \geq 4\omega_{i,j}$, offline data guarantees regret reduction.
2. **Data-Dependent Saving:** If $4\omega_{i,j} > \max\{\Delta_i, \Delta_j\} > 2\omega_{i,j}$, Saving exist only when the offline data size satisfies:
$$\frac{N_i N_j}{N_i + N_j} \geq \frac{4 \log(1/\delta_t) (4\omega_{i,j} - \max\{\Delta_i, \Delta_j\})}{\max\{\Delta_i, \Delta_j\} (2\omega_{i,j} - \max\{\Delta_i, \Delta_j\})^2}.$$
3. **No Saving:** For $\max\{\Delta_i, \Delta_j\} \leq 2\omega_{i,j}$, offline data provides no benefit.

This result aligns with our Saving analysis in Section 3, which shows that Saving depends on the similarity between offline and online data distributions, with the magnitude of Saving influenced by the offline data size. But the detail analysis further shows there exist an intermediate phase: where the existence of Saving is data dependent. Here, we provide deeper insight into the three regimes. When the distributions are sufficiently close ($\max\{\Delta_i, \Delta_j\} \geq 4\omega_{i,j}$), the $T_{ij}(\text{UCB})$ is consistently at most that of $T_{ij}(\text{UCB}^{\text{hyb}})$, ensuring Saving regardless of the offline data size. As $\omega_{i,j}$ increases, indicating greater distributional divergence, the $T_{ij}(\text{UCB})$ grows, potentially exceeding that of $T_{ij}(\text{UCB}^{\text{hyb}})$. In this intermediate regime ($4\omega_{i,j} > \max\{\Delta_i, \Delta_j\} > 2\omega_{i,j}$), Saving is possible only if the offline data size $N_{i,j}$ is large enough to reduce $T_{ij}(\text{UCB}^{\text{hyb}})$, compensating for the increased $\omega_{i,j}$. However, when the distributions are too dissimilar ($\max\{\Delta_i, \Delta_j\} \leq 2\omega_{i,j}$), $T_{ij}(\text{UCB}^{\text{hyb}})$ is dominated by $\omega_{i,j}$, and no amount of offline data can yield Saving. These cases illustrate a spectrum of outcomes driven by distributional similarity and the size of offline data.

In Lemma 5, we established conditions for the existence of Saving, but their magnitude remains unanalyzed. For instance, when $\Delta' > 2\omega'$, Saving are guaranteed regardless of the offline dataset size N' . However, if N' is small, the confidence radius may remain large, resulting in negligible Saving. To further quantify the saving term, we analyze a specific setting where inferior arms can be identified without online data, thereby maximizing Saving.

Given the following Lemma 6, this implies no online exploration is needed for HybUCB-AR when the following condition holds:

$$\frac{N_i N_j}{N_i + N_j} \geq \frac{16 \log(1/\delta_t)(\max\{\Delta_i, \Delta_j\} - \omega_{i,j})}{\max\{\Delta_i, \Delta_j\}(2\omega_{i,j} - \max\{\Delta_i, \Delta_j\})^2} \quad \text{and} \quad 2\omega_{i,j} < \max\{\Delta_i, \Delta_j\}.$$

Lemma 6. Suppose $\omega' < \Delta'$ and

$$N' \geq \frac{4\delta'(2\Delta' - \omega')}{\Delta'(\Delta' - \omega')^2}.$$

Then, under the good event, no online exploration is needed for this suboptimal arm.

Proof. Continuing from Equation (23), we aim to identify the condition under which the following expression becomes non-positive:

$$\frac{2\delta' + N'\Delta'(\omega' - \Delta') + \sqrt{4\delta'^2 + 4\delta'N'\omega'\Delta'}}{\Delta'^2} \leq 0.$$

It is clear that this inequality can only hold when $\omega' < \Delta'$; otherwise, all terms in the numerator would be positive and the entire expression strictly positive.

Under the substitution from Equation (23), we consider the equivalent condition:

$$N'\Delta'(\Delta' - \omega') - 2\delta' \geq \sqrt{4\delta'^2 + 4\delta'N'\omega'\Delta'}.$$

Squaring both sides yields:

$$N'^2\Delta'^2(\Delta' - \omega')^2 - 4\delta'N'\Delta'(\Delta' - \omega') \geq 4\delta'N'\omega'\Delta'.$$

Rearranging terms, we obtain:

$$N' \geq \frac{4\delta'(2\Delta' - \omega')}{\Delta'(\Delta' - \omega')^2},$$

as desired. \square

Remark 8. Specifically, when $\omega' = 0$, the condition simplifies to $N' \geq 8\delta'/\Delta'^2$, which degenerate to the algorithm property of vanilla UCB. In contrast, under the Saving condition in Lemma 5 ($2\Delta' > 2\omega' > \Delta'$), Saving begin when the offline data size $N_i N_j / (N_i + N_j)$ meets a certain threshold, but fully replacing online exploration requires a larger $N_i N_j / (N_i + N_j)$. The difference in $N_i N_j / (N_i + N_j)$ between these two thresholds is given by:

$$\frac{4\delta'(2\Delta' - \omega')}{\Delta'(\Delta' - \omega')^2} - \frac{4\delta'(2\omega' - \Delta')}{\Delta'(\omega' - \Delta')^2} = \frac{4\delta'(3\Delta' - 3\omega')}{\Delta'(\Delta' - \omega')^2}.$$

C.7 Proof of Instance Independent Regret Upper Bounds

As shown in Theorem 1, we provide two instance-independent regret upper bounds, and the final bound is obtained by taking the minimum of the two. The details of these bounds are presented in the following two Sections (C.7.1 and C.7.2).

1015 C.7.1 Analysis 1

1016 *Proof.* We derive the instance independent upper bound through modification of the instance dependent regret
 1017 upper bound. Let Δ be a variable between $(0, 1)$. For $(a_i, a_j) \in \mathcal{A} \times \mathcal{A}$ such that $\max\{\Delta_i, \Delta_j\} \geq \Delta$, take
 1018 $\delta_t = \frac{1}{2K(K+1)t^2}$, the core term in the instance regret upper bound (Equation (32), e.g., the leading Δ -dependent
 1019 term) is bounded by:

$$\frac{4(\Delta_i + \Delta_j) \log(1/\delta_t)}{\max\{\Delta_i^2, \Delta_j^2\}} \leq \frac{8 \log(1/\delta_t)}{\max\{\Delta_i, \Delta_j\}} < \frac{8 \log(1/\delta_T)}{\Delta} = \frac{16 \log(\sqrt{2K(K+1)}T)}{\Delta},$$

1020 where the first inequality comes by taking $\Delta_i + \Delta_j \leq 2 \max\{\Delta_i, \Delta_j\}$.

1021 For any arm pair $(a_i, a_j) \in \mathcal{A} \times \mathcal{A}$ satisfying $\max\{\Delta_i, \Delta_j\} < \Delta$, the total regret incurred by these sub-
 1022 optimal pairs is upper bounded by $T\Delta$. Let $\text{Saving}(a_i, a_j) = \frac{\Delta_i + \Delta_j}{2} \cdot \frac{N_i N_j}{N_i + N_j} \cdot \frac{\max\{\max\{\Delta_i, \Delta_j\} - 4\omega_{i,j}, 0\}}{\max\{\Delta_i, \Delta_j\}}$,
 1023 and $\text{Saving}'(a_i, a_j) = \min \left\{ \frac{\Delta_i + \Delta_j}{2} \cdot \frac{N_i N_j}{N_i + N_j} \cdot \frac{\max\{\max\{\Delta_i, \Delta_j\} - 4\omega_{i,j}, 0\}}{\max\{\Delta_i, \Delta_j\}}, \frac{8(\Delta_i + \Delta_j) \log(\sqrt{2K(K+1)}T)}{\max\{\Delta_i^2, \Delta_j^2\}} \right\}$ for
 1024 all $i, j \in [K]$. Combining these results yields:

$$\begin{aligned} \text{Reg}(T) &\leq \frac{\pi^2}{6} \Delta_{\max} + \sum_{i \leq j} \max \left\{ \frac{16 \log(\sqrt{2K(K+1)}T)}{\max\{\Delta_i^2, \Delta_j^2\}} - \text{Saving}(a_i, a_j), 0 \right\} \\ &\leq \frac{\pi^2}{6} \Delta_{\max} + T\Delta + \frac{8K(K+1) \cdot \log(\sqrt{2K(K+1)}T)}{\Delta} - \sum_{i \leq j} \text{Saving}'(a_i, a_j) \\ &\leq \frac{\pi^2}{6} \Delta_{\max} + 4\sqrt{2K(K+1) \cdot \log(\sqrt{2K(K+1)}T)} - \sum_{i \leq j} \text{Saving}'(a_i, a_j), \end{aligned}$$

1025 where the last inequality follows by taking $\Delta = \sqrt{\frac{8K(K+1) \cdot \log(\sqrt{2K(K+1)}T)}{T}}$, as desired. \square

1026 C.7.2 Analysis 2

1027 This theoretical analysis extends the framework of Appendix B.3 in Cheung and Lyu (2024), generalizing its
 1028 regret analysis from the classical multi-armed bandit (MAB) setting to a preference-based model.

1029 **Lemma 7.** Let $\{T_{i,j}^*\}_{i \leq j \in [K]}$ is the optimal solution to the following maximization problem:

$$\begin{aligned} \max_{(T_{i,j})_{i \leq j}} & \sum_{i \leq j} \sum_{t=1}^{T_{i,j}} \sqrt{\frac{1}{t + \frac{N_i N_j}{N_i + N_j}}}, \\ \text{s.t.} & \sum_{i \leq j} T_{i,j} = T, \\ & T_{i,j} \in \mathbb{N}_{\geq 0}, \forall i \leq j \in [K]. \end{aligned}$$

1030 then it must hold that $T_{i,j}^* \leq \max\{\lceil \tau_* \rceil - \frac{N_i N_j}{N_i + N_j}, 0\}$ for all $a_i, a_j \in \mathcal{A}$.

1031 *Proof.* We proof this lemma by contradiction. Suppose there's an arm pair (a_i, a_j) such that $T_{i,j}^* \geq \max\{\lceil \tau_* \rceil -$
 1032 $\frac{N_i N_j}{N_i + N_j}, 0\} + 1$, then it must hold that there exist at least one arm pair $(a_{i'}, a_{j'})$ such that $T_{i',j'}^* \leq \max\{\lceil \tau_* \rceil -$
 1033 $\frac{N_{i'} N_{j'}}{N_{i'} + N_{j'}}, 0\} - 1$, otherwise we have: $T_{i'',j''}^* \geq \max\{\lceil \tau_* \rceil - \frac{N_{i''} N_{j''}}{N_{i''} + N_{j''}}, 0\}$ for all $i'', j'' \in \mathcal{A}$. Particularly,
 1034 $T_{i,j}^* \geq \max\{\lceil \tau_* \rceil - \frac{N_i N_j}{N_i + N_j}, 0\} + 1$, this implies:

$$\sum_{i \leq j} T_{i,j}^* > \sum_{i \leq j} \max\{\lceil \tau_* \rceil - \frac{N_i N_j}{N_i + N_j}, 0\} \geq \sum_{i \leq j} \max\{\tau_* - \frac{N_i N_j}{N_i + N_j}, 0\} = \sum_{i \leq j} t_{i,j}^* = T, \quad (33)$$

1035 where the penultimate equality comes by the property of the optimal solution of the Linear Programming
 1036 problem in Theorem 1: Since by the Linear Programming problem, $\forall k, l \in [K]$, $k \leq l$, let $\epsilon_{k,l} = t_{k,l}^* -$
 1037 $\max\{\tau_* - \frac{N_k N_l}{N_k + N_l}, 0\} \geq 0$. If $\exists k', l' \in [K]$ such that $\epsilon_{k',l'} > 0$, then the solution $(\tilde{\tau}, \{\tilde{t}_{k,l}\}_{k,l \in [K]})$ defined
 1038 as $\tilde{\tau} = \tau_* + \frac{2\epsilon_{k',l'}}{K(K+1)}$, $\tilde{t}_{k,l} = t_{k,l}^* + \frac{2\epsilon_{k',l'}}{K(K+1)}$ for all $(k, l) \in [K] \times [K] \setminus \{(k', l'), (l', k')\}$, and $\tilde{t}_{k',l'} =$
 1039 $t_{k',l'}^* - \frac{K(K+1)-2}{K(K+1)} \epsilon_{k',l'}$ could also be a feasible solution to the linear programming problem. This implies
 1040 $\tilde{\tau} > \tau_*$, hence the optimal case could only have $\epsilon = 0$, implying the establishment of the equality.

1041 Hence, Equation (33) violated the feasible region of the Linear Programming problem in Lemma 7. Thereby, we
 1042 have two distinct arm pairs (a_i, a_j) and $(a_{i'}, a_{j'})$ such that:

$$\begin{aligned} T_{i,j}^* + \frac{N_i N_j}{N_i + N_j} &\geq \max\{\lceil \tau_* \rceil, \frac{N_i N_j}{N_i + N_j}\} + 1, \text{ in particular } T_{i,j}^* \geq 1, \\ T_{i',j'}^* + \frac{N_{i'} N_{j'}}{N_{i'} + N_{j'}} &\leq \max\{\lceil \tau_* \rceil, \frac{N_{i'} N_{j'}}{N_{i'} + N_{j'}}\} - 1 = \lceil \tau_* \rceil - 1. \end{aligned}$$

1043 To establish the contradiction argument, let $k \leq l$ with $k, l \in [K] \times [K]$, consider another feasible solution:

$$\tilde{T}_{k,l} = \begin{cases} \tilde{T}_{k,l}^* - 1 & \text{if } (k, l) = (i, j), \\ \tilde{T}_{k,l}^* + 1 & \text{if } (k, l) = (i', j'), \\ \tilde{T}_{k,l}^* & \text{otherwise.} \end{cases}$$

1044 but then we have:

$$\begin{aligned} &\sum_{k \leq l} \sum_{t_{k,l}=1}^{\tilde{T}_{k,l}} \sqrt{\frac{1}{t_{k,l} + \frac{N_k N_l}{N_k + N_l}}} - \sum_{k \leq l} \sum_{t_{k,l}=1}^{T_{k,l}^*} \sqrt{\frac{1}{t_{k,l} + \frac{N_k N_l}{N_k + N_l}}} \\ &= \sqrt{\frac{1}{T_{i',j'}^* + \frac{N_{i'} N_{j'}}{N_{i'} + N_{j'}} + 1}} - \sqrt{\frac{1}{T_{i,j}^* + \frac{N_i N_j}{N_i + N_j}}} \\ &\geq \sqrt{\frac{1}{\lceil \tau_* \rceil}} - \sqrt{\frac{1}{\max\{\lceil \tau_* \rceil, \frac{N_i N_j}{N_i + N_j}\} + 1}} > 0, \end{aligned}$$

1045 this implies $T_{i,j}^*$ is not the optimal solution, thereby $T_{i,j}^* \leq \max\{\lceil \tau_* \rceil - \frac{N_i N_j}{N_i + N_j}, 0\}$, as desired. \square

1046 *Proof.* (Complete Proof of Instance Independent Upper Bound) By Lemma 4, arm pair (a_i, a_j) has the chance
 1047 to be selected only if:

$$\max\{\Delta_i, \Delta_j\} \leq \min \left\{ 4\sqrt{\frac{\log(1/\delta_t)}{2T_{i,j}}}, 4\sqrt{\frac{1}{2\left(T_{i,j} + \frac{N_i N_j}{N_i + N_j}\right) \log((1/\delta_t))}} + 2\frac{\frac{N_i N_j}{N_i + N_j}}{T_{i,j} + \frac{N_i N_j}{N_i + N_j}} \omega_{i,j} \right\}.$$

1048 Denote $\text{rad}_{i,j} = \sqrt{\frac{\log(1/\delta_t)}{2T_{i,j}}}$ and $\text{rad}_{i,j}^{\text{hyb}} = \sqrt{\frac{\log(1/\delta_t)}{2\left(T_{i,j} + \frac{N_i N_j}{N_i + N_j}\right)}} + \frac{\frac{N_i N_j}{N_i + N_j}}{T_{i,j} + \frac{N_i N_j}{N_i + N_j}} V_{i,j}$, hence we have:

$$\begin{aligned} \text{Reg}(T) &= \sum_{t=1}^T \frac{\Delta_{A_1(t)} + \Delta_{A_2(t)}}{2} \\ &\leq \sum_{t=1}^T \max\{\Delta_{A_1(t)}, \Delta_{A_2(t)}\} \\ &\stackrel{(*)}{\leq} \sum_{t=1}^T \min\{4\text{rad}_{A_1(t), A_2(t)}, 4\text{rad}_{A_1(t), A_2(t)}^{\text{hyb}}\} \\ &\leq \min \left\{ 4 \sum_{i \leq j} \sum_{t=1}^{T_{i,j}} \text{rad}_{i,j}, 4 \sum_{i \leq j} \sum_{t=1}^{T_{i,j}} \text{rad}_{i,j}^{\text{hyb}} \right\}. \end{aligned}$$

1049 The Inequality $(*)$ established due to Lemma 4 and the fact that $\omega_{i,j} = V_{i,j} + p_{i,j}^{\text{off}} - p_{i,j} \leq 2V_{i,j}$. For the first
 1050 term $4 \sum_{i \leq j} \sum_{t=1}^{T_{i,j}} \text{rad}_{i,j}$, we have:

$$\begin{aligned} 4 \sum_{i \leq j} \sum_{t=1}^{T_{i,j}} \sqrt{\frac{\log(1/\delta_t)}{2T_{i,j}}} &\leq 8 \sum_{i \leq j} \sqrt{T_{i,j} \log(1/\delta_T)} = 8\sqrt{\log(1/\delta_T)} \sum_{i \leq j} 1 \cdot \sqrt{T_{i,j}} \\ &\leq 8\sqrt{\log(1/\delta_T)} \sqrt{\sum_{i \leq j} 1^2} \sqrt{\sum_{i \leq j} T_{i,j}} = 8\sqrt{\frac{K(K+1)}{2}} T \log(1/\delta_T). \quad (34) \end{aligned}$$

1051 The last inequality follows from applying the Cauchy–Schwarz inequality. Since this bound corresponds to the
 1052 vanilla UCB setting without offline data, it is already covered by the result provided in Section C.7.1, and is
 1053 therefore omitted from the final regret bound.

1054 For the Second term $4 \sum_{i \leq j} \sum_{t=1}^{T_{i,j}} \text{rad}_{i,j}^{\text{hyb}}$, we have:

$$\begin{aligned} & \sum_{i \leq j} \sum_{t=1}^{T_{i,j}} 4 \sqrt{\frac{\log(1/\delta_T)}{2 \left(t + \frac{N_i N_j}{N_i + N_j}\right)}} + 4 \frac{\frac{N_i N_j}{N_i + N_j}}{t + \frac{N_i N_j}{N_i + N_j}} V_{i,j} \\ & \leq 4 \max_{i \leq j} V_{i,j} \cdot T + \sqrt{\log \frac{1}{\delta_T}} \sum_{i \leq j} \sum_{t=1}^{T_{i,j}} 4 \sqrt{\frac{1}{2 \left(t + \frac{N_i N_j}{N_i + N_j}\right)}}. \end{aligned} \quad (35)$$

1055 Let $T_{i,j}^*$ be the optimal solution of the LP problem in Lemma 7. By applying Lemma 7, we obtain:

$$\sum_{i \leq j} \sum_{t=1}^{T_{i,j}} \sqrt{\frac{1}{t + \frac{N_i N_j}{N_i + N_j}}} \leq \sum_{i \leq j} \sum_{t=1}^{T_{i,j}^*} \sqrt{\frac{1}{t + \frac{N_i N_j}{N_i + N_j}}} \leq \sum_{i \leq j} \sum_{t=1}^{\max\{\lceil \tau_* \rceil - \frac{N_i N_j}{N_i + N_j}, 0\}} \sqrt{\frac{1}{t + \frac{N_i N_j}{N_i + N_j}}},$$

1056 thus:

$$\begin{aligned} & \sum_{i \leq j} \sum_{t=1}^{\max\{\lceil \tau_* \rceil - \frac{N_i N_j}{N_i + N_j}, 0\}} \sqrt{\frac{1}{t + \frac{N_i N_j}{N_i + N_j}}} \leq \sum_{i \leq j} \frac{\max\{\lceil \tau_* \rceil - \frac{N_i N_j}{N_i + N_j}, 0\}}{\lceil \tau_* \rceil} \sum_{t=1}^{\lceil \tau_* \rceil} \sqrt{\frac{1}{t}} \\ & \leq \sum_{i \leq j} \max\{\lceil \tau_* \rceil - \frac{N_i N_j}{N_i + N_j}, 0\} \cdot \frac{2}{\sqrt{\tau_*}} \\ & \leq \sum_{i \leq j} \frac{4t_{ij}^*}{\sqrt{\tau_*}} = \frac{4T}{\sqrt{\tau_*}}. \end{aligned}$$

1057 where the last inequality follows by applying the feasibility of the LP problem in Theorem 2.

1058 Replacing term $\sum_{i \leq j} \sum_{t=1}^{T_{i,j}} 4 \sqrt{\frac{1}{2 \left(t + \frac{N_i N_j}{N_i + N_j}\right)}}$ with $\frac{16T}{\sqrt{\tau_*}}$ in Equation (35) leads to the desired result. \square

1059 D Proof of Theorem 2

1060 The proof of the regret lower bound builds upon the approach in Cheung and Lyu (2024), which we have
1061 extended to accommodate the hybrid setting.

1062 **Theorem 4.** Let \mathbb{P}, \mathbb{Q} be probability distributions on (Ω, \mathcal{F}) . For an event $E \in \sigma(\mathcal{F})$, it holds that

$$\Pr_{\mathbb{P}}(E) + \Pr_{\mathbb{Q}}(E^c) \geq \frac{1}{2} \exp(-KL(\mathbb{P}, \mathbb{Q})).$$

1063 **Theorem 5.** Consider two instances \mathcal{I}_P and \mathcal{I}_Q that share the same arm set \mathcal{A} , online phase horizon T ,
1064 and offline sample size $\{N_i\}_{i \in [K]}$, but have different reward distributions $P = (P_i, P_i^{\text{off}})_{i \in [K]}$ and $Q =$
1065 $(Q_a, Q_a^{\text{off}})_{a \in [K]}$. For any non-anticipatory policy π , it holds that

$$KL(P, Q) = \sum_{i \leq j} \mathbb{E}_{P, \pi}[T_{i,j}(T)] \cdot KL(P_{i,j}, Q_{i,j}) + \sum_{i \in [K]} N_i \cdot KL(P_i^{\text{off}}, Q_i^{\text{off}}).$$

1066 Theorem 4 is a direct restatement of Lemma 15.1 from Lattimore and Szepesvári (2020), while Theorem 5 is
1067 adapted from Theorem C.2 in Cheung and Lyu (2024). The only difference in our setting is that the relative
1068 feedback data follows a discrete distribution; nonetheless, by following the original proof strategy, we are able to
1069 derive Theorem 5 accordingly.

1070 **Claim 1.** For $P_i = \mathcal{N}(\mu_i, \sigma^2)$, where $i \in \{1, 2\}$, we have

$$KL(P_1, P_2) = \frac{(\mu_1 - \mu_2)^2}{2\sigma^2}.$$

1071 **Claim 2.** For $P_i = \text{Bernoulli}(p_i)$, where $i \in \{1, 2\}$, we have

$$KL(P_1, P_2) = p_1 \log \frac{p_1}{p_2} + (1 - p_1) \log \left(\frac{1 - p_1}{1 - p_2} \right) \leq \frac{(p_1 - p_2)^2}{p_2(1 - p_2)},$$

1072 where the last inequality of Claim 2 comes by applying $\log x \leq x - 1$.

1073 D.1 Instance dependent lower bound

1074 *Proof.* We begin by denoting μ_i^{off} and μ_i as the means of the Gaussian distributions of P_i^{off} and P_i , respectively,
 1075 for each arm $a \in \mathcal{A}$. Assume μ_1 corresponds to the optimal arm, while the remaining arms are suboptimal, with
 1076 $\mu_i = \mu_j$ for all $i, j \neq 1$. The gap between the optimal arm a_1 and the suboptimal arms (a_2, \dots, a_K) is denoted
 1077 by Δ , where $\Delta = \mu_1 - \mu_i \in (0, \frac{1}{2})$.

1078 Next, we introduce an alternative distribution Q . Let k be a fixed arm in $\{2, \dots, K\}$ and $T_i(t)$ represents the
 1079 number of times arm i is selected in pairs and pulled up to time t . For all $i \in [K] \setminus \{k\}$, define $Q_i^{\text{off}} = P_i^{\text{off}}$ and
 1080 $Q_i = P_i$. For arm k , let $Q_k = \mathcal{N}(\mu_k + 2\Delta, 1)$. Define Q_k^{off} as follows:

$$Q_k^{\text{off}} = \begin{cases} \mathcal{N}(\mu_k^{\text{off}}, 1) & \text{if } \mu_k^{\text{off}} \geq \mu_k + 2\Delta - V_k, \\ \mathcal{N}(\mu_k + 2\Delta - V_k, 1) & \text{if } \mu_k^{\text{off}} < \mu_k + 2\Delta - V_k, \end{cases} \quad (36)$$

1081 where $V_k \geq |\mu_k^{\text{off}} - \mu_k|$. The construction of Q_k^{off} in Equation (36) is governed by V_k . The high-level idea
 1082 behind it is to maximize the regret lower bound by minimizing the KL divergence between Q_k^{off} and P_k^{off} .
 1083 However, directly setting $Q_k^{\text{off}} = P_k^{\text{off}}$ may violate the constraint $Q \in \mathcal{I}_V$. When the constraint is violated
 1084 (e.g. $\mu_k^{\text{off}} < \mu_k + 2\Delta - V_k$), we instead define $Q_k^{\text{off}} = \mathcal{N}(\mu_k + 2\Delta - V_k, 1)$, which yields the smallest KL
 1085 divergence possible while preserving $Q \in \mathcal{I}_V$.

1086 Under the dueling bandit setting with the Bradley–Terry model, the pairwise preference gaps under distribution
 1087 P is given by:

$$\Delta_i = p_{1,i} - \frac{1}{2} = \sigma(\Delta) - \frac{1}{2}, \forall i \in [K].$$

1088 Under the perturbed distribution Q , the pairwise gaps are defined as

$$\Delta_{k,i} = p_{k,i} - \frac{1}{2} = \sigma(2\Delta) - \frac{1}{2}, \forall i \in [K] \setminus \{1, k\}, \quad \text{and} \quad \Delta_{k,1} = \sigma(\Delta) - \frac{1}{2}.$$

1089 Using the definition of regret under distribution P , we derive:

$$\begin{aligned} \text{Reg}_P(T) &= T \left(\sigma(\Delta) - \frac{1}{2} \right) - \frac{(\sigma(\Delta) - \frac{1}{2})}{2} \mathbb{E}_P [T_1(T) \mathbf{1}_{T_1(T) \leq T} + T_1(T) \mathbf{1}_{T_1(T) > T}] \\ &\geq T \left(\sigma(\Delta) - \frac{1}{2} \right) - \frac{(\sigma(\Delta) - \frac{1}{2})}{2} \mathbb{E}_P [T_1(T) \mathbf{1}_{T_1(T) \leq T}] - \frac{(\sigma(\Delta) - \frac{1}{2})}{2} \mathbb{E}_P [T_1(T) \mathbf{1}_{T_1(T) > T}] \\ &= T \left(\sigma(\Delta) - \frac{1}{2} \right) - \frac{T(\sigma(\Delta) - \frac{1}{2})}{2} \Pr_P(T_1(T) \leq T) - T \left(\sigma(\Delta) - \frac{1}{2} \right) \Pr_P(T_1(T) > T) \\ &= \frac{T(\sigma(\Delta) - \frac{1}{2})}{2} \Pr_P(T_1(T) \leq T). \end{aligned}$$

1090 For distribution Q , the regret is lower bounded as:

$$\text{Reg}_Q(T) = \sum_{i=1}^K \frac{\Delta_i}{2} \mathbb{E}_Q(T_i(T)) \geq \frac{\sigma(\Delta) - \frac{1}{2}}{2} \mathbb{E}_Q(T_1(T)) \geq \frac{T(\sigma(\Delta) - \frac{1}{2})}{2} \Pr_Q(T_1(T) > T).$$

1091 Combining these, we obtain:

$$\begin{aligned} 2CT^p &\geq \text{Reg}_P(T) + \text{Reg}_Q(T) \\ &\geq \frac{T(\sigma(\Delta) - \frac{1}{2})}{2} \left(\Pr_P(T_1(T) \leq T) + \Pr_Q(T_1(T) > T) \right) \\ &\geq \frac{T(\sigma(\Delta) - \frac{1}{2})}{4} \exp(-KL(P, Q)) \\ &\geq \frac{T\Delta}{32} \exp(-KL(P, Q)). \end{aligned} \quad (37)$$

1092 where the first inequality follows from the definition of Cp-consistent, the penult inequality applies Theorem 4,
 1093 and the final inequality comes by our restriction on Δ : Since $\Delta \in (0, \frac{1}{2})$, we have $\sigma(\Delta) \geq \frac{1}{8}\Delta + \frac{1}{2}$.

1094 By Theorem 5, the KL divergence is:

$$\text{KL}(P, Q) = \sum_{i \leq j} \mathbb{E}_P[T_{i,j}(T)] \cdot \text{KL}(P_{i,j}, Q_{i,j}) + \sum_{i \in [K]} N_i \cdot \text{KL}(P_i^{\text{off}}, Q_i^{\text{off}}).$$

1095 For the first term, using $d(p, q) = p \log \left(\frac{p}{q} \right) + (1 - p) \log \left(\frac{1-p}{1-q} \right) \leq \frac{(p-q)^2}{q(1-q)}$, we have:

$$\begin{aligned}
\sum_{i,j} \mathbb{E}_P[T_{i,j}(T)] \text{KL}(P_{i,j}, Q_{i,j}) &= \sum_{i=2}^K \mathbb{E}_P[T_{k,i}(T)] d\left(\frac{1}{2}, \sigma(2\Delta)\right) + \mathbb{E}_P[T_{k,1}(T)] d(1 - \sigma(\Delta), \sigma(\Delta)) \\
&\stackrel{(a)}{\leq} \sum_{i=2}^K \mathbb{E}_P[T_{k,i}(T)] d\left(\frac{1}{2}, \frac{1}{2}\Delta + \frac{1}{2}\right) + \mathbb{E}_P[T_{k,1}(T)] d\left(\frac{1}{2} - \frac{1}{4}\Delta, \frac{1}{2} + \frac{1}{4}\Delta\right) \\
&\leq \sum_{i=2}^K \mathbb{E}_P[T_{k,i}(T)] \frac{\Delta^2}{1 - \Delta^2} + \mathbb{E}_P[T_{k,1}(T)] \frac{\Delta^2}{1 - \frac{1}{4}\Delta^2} \\
&\leq \mathbb{E}_P[T_k(T)] \frac{\Delta^2}{1 - \Delta^2} \\
&\leq \mathbb{E}_P[T_k(T)] 2\Delta^2,
\end{aligned} \tag{38}$$

1096 where inequality (a) follows by $\sigma(\Delta) \leq \frac{1}{4}\Delta + \frac{1}{2}$, this implies:

$$d\left(\frac{1}{2}, \sigma(\Delta)\right) \leq d\left(\frac{1}{2}, \frac{1}{4}\Delta + \frac{1}{2}\right) \quad \text{and} \quad d(1 - \sigma(\Delta), \sigma(\Delta)) \leq d\left(\frac{1}{2} - \frac{1}{4}\Delta, \frac{1}{2} + \frac{1}{4}\Delta\right).$$

1097 For the second term, by our setup, we have:

$$\sum_{i \in [K]} N_i \cdot \text{KL}(P_i^{\text{off}}, Q_i^{\text{off}}) = N_k \frac{\max\{2\Delta - \omega_k, 0\}^2}{2}, \tag{39}$$

1098 where $\omega_k = V_k + \mu_k^{\text{off}} - \mu_k$. Combining Inequality (37), (38) and Equation (39), we get:

$$2CT^p \geq \frac{T\Delta}{32} \exp \left\{ -\mathbb{E}_P[T_k(T)] 2\Delta^2 - \frac{\max\{2\Delta - \omega_k, 0\}^2}{2} N_k \right\}.$$

1099 To isolate $\mathbb{E}_P[T_k(T)]$, we take the natural logarithm of both sides and rearrange the inequality, yielding:

$$\mathbb{E}[T_k(T)] \geq \frac{1}{2\Delta^2} \left((1-p) \log T + \log \frac{\Delta}{64C} - \frac{\max\{2\Delta - \omega_k, 0\}^2}{2} N_k \right).$$

1100 Summing over all arms $i \in [K]$, the total regret lower bound becomes:

$$\sum_{i \in [K]} \frac{\Delta}{2} \mathbb{E}[T_i(T)] \geq \frac{K}{4\Delta} \left((1-p) \log T + \log \frac{\Delta}{64C} - \sum_{i \in [K]} \frac{\max\{2\Delta - \omega_k, 0\}^2}{2} N_i \right),$$

1101 as desired. □

1102 D.2 Instance independent regret lower bound

1103 *Proof.* The divide the proof into three distinct cases:

1104 **Case 1:** $2\sqrt{KT} > T \cdot (V_{\max} + \sqrt{2/\tau_*})$, and $V_{\max} \leq 1/\sqrt{\tau_*}$, we derive a regret lower bound of:

$$\Omega \left(\min \left\{ \sqrt{KT}, \left(\sqrt{\frac{1}{\tau_*}} + V_{\max} \right) \cdot T \right\} \right) = \Omega \left(\frac{T}{\sqrt{\tau_*}} \right).$$

1105 At this point, we consider a setting where the bias between offline and online feedback is negligible. We define
1106 the Gaussian reward distributions as follows, assuming no discrepancy between offline and online feedback:

$$P_i^{\text{off}} = P_i = \begin{cases} N(\Delta, 1) & \text{if } i = 1, \\ N(0, 1) & \text{if } i \neq 1. \end{cases} \quad Q_i^{\text{off}} = Q_i = \begin{cases} N(\Delta, 1) & \text{if } i = 1, \\ N(2\Delta, 1) & \text{if } i = k, \\ N(0, 1) & \text{if } i \in [K] \setminus \{k, 1\}. \end{cases}$$

1107 where $\Delta = 1/\sqrt{\tau_*}$ and $k = \arg \min_{i \in [K]} N_i + \mathbb{E}_P(T_i)$.

1108 Obviously $P, Q \in \mathcal{I}_V$, since $|\mu_i^{\text{off}} - \mu_i| = 0 \leq V_{\max}$. Without loss of generality, we assume that arm 1 has the
1109 largest offline sample size, that is, $1 = \arg \max_{i \in [K]} N_i$.

1110 **Subcase 1.1:** $k = 1$.

1111 In this case, we have:

$$\mathbb{E}_P[T_1] + N_1 \leq \mathbb{E}_P[T_i] + N_i \quad \text{for all } i \in [K] \setminus \{1\}.$$

1112 Since we assume $1 = \arg \max_{i \in [K]} N_i$, it follows that:

$$\mathbb{E}_P[T_1] \leq \mathbb{E}_P[T_i] + N_i - N_1 \leq \mathbb{E}_P[T_i], \quad \text{for all } i \in [K] \setminus \{1\}.$$

1113 Therefore, arm 1 has both the largest offline sample size and the smallest expected number of online pulls among
 1114 all arms. By the pigeonhole principle, this implies that the optimal arm $i \neq 1$ is pulled at most $2T/K$ times.
 1115 Thereby, the expected cumulative regret can then be bounded as:

$$\text{Reg}(T) \geq \left(2T - \frac{2T}{K}\right) \cdot \frac{\Delta}{2} = \frac{(K-1)\Delta T}{K} = \frac{(K-1)T}{K\sqrt{\tau_*'}}.$$

1116 **Subcase 1.2:** $k \neq 1$:

$$\begin{aligned} \text{Reg}_P(T) + \text{Reg}_Q(T) &\geq \frac{1}{2} \left[\sigma(\Delta) - \frac{1}{2} \right] \left(\Pr(\mathbb{E}[T_1] \leq T) + \Pr(\mathbb{E}[T_1] > T) \right) \\ &\geq \frac{T \left[\sigma(\Delta) - \frac{1}{2} \right]}{4} \cdot \exp \left(- \sum_{i=1}^K \mathbb{E}[T_i] \cdot \text{KL}(P_{k,i}, Q_{k,i}) - N_k \cdot \text{KL}(P_i^{\text{off}}, Q_i^{\text{off}}) \right) \\ &\stackrel{(1)}{\geq} \frac{T\Delta}{32} \exp \left(-2\mathbb{E}[T_k]\Delta^2 - 2N_k\Delta^2 \right) \\ &\stackrel{(2)}{\geq} \frac{T\Delta}{32} \exp \left(-2\tau_*'\Delta^2 \right) \\ &\stackrel{(3)}{=} \frac{1}{32e^2} \cdot \frac{T}{\sqrt{\tau_*'}}, \end{aligned}$$

1117 where inequality (1) follows by applying the same technique in D.1, inequality (2) hold since $k =$
 1118 $\arg \min_{i \in [K]} N_i + \mathbb{E}_P[T_i] \leq \tau_*'$, and the final equality (3) comes by applying $\Delta = 1/\sqrt{\tau_*'}$.

1119 **Case 2:** $2\sqrt{KT} > T \cdot (V_{\max} + \sqrt{2/\tau_*'})$, and $V_{\max} \geq 1/\sqrt{\tau_*'}$, we derive a regret lower bound of:

$$\Omega \left(\min \left\{ \sqrt{KT}, \left(\sqrt{\frac{1}{\tau_*'}} + V_{\max} \right) \cdot T \right\} \right) = \Omega(T \cdot V_{\max}).$$

1120 We construct another instance such that:

$$P_i^{\text{off}} = \mathcal{N}(0, 1), \forall i \in [K], \quad P_i = \begin{cases} \mathcal{N}(\Delta, 1) & \text{if } i = 1, \\ \mathcal{N}(0, 1) & \text{if } i \neq 1. \end{cases}$$

1121

$$Q_i^{\text{off}} = \mathcal{N}(0, 1), \forall i \in [K], \quad Q_i = \begin{cases} \mathcal{N}(\Delta, 1) & \text{if } i = 1, \\ \mathcal{N}(2\Delta, 1) & \text{if } i = k, \\ \mathcal{N}(0, 1) & \text{if } i \in [K] \setminus \{k, 1\}. \end{cases}$$

1122 where $k = \arg \min_{i \in [K]} \mathbb{E}[T_i]$, and take $\Delta = 1/2 \cdot V_{\max}$.

1123 Since $|\mu_i^{\text{off}} - \mu_i| \leq 2\Delta = V_{\max} = V_i$ for all $i \in [K]$, this implies $P, Q \in \mathcal{I}_V$. Furthermore:

$$\begin{aligned} \text{Reg}_P(T) + \text{Reg}_Q(T) &\geq \frac{1}{2} \left[\sigma(\Delta) - \frac{1}{2} \right] \left(\Pr(\mathbb{E}[T_1] \leq T) + \Pr(\mathbb{E}[T_1] > T) \right) \\ &\geq \frac{T \left[\sigma(\Delta) - \frac{1}{2} \right]}{4} \cdot \exp \left(- \sum_{i=1}^K \mathbb{E}[T_i] \cdot \text{KL}(P_{k,i}, Q_{k,i}) - N_k \cdot \text{KL}(P_i^{\text{off}}, Q_i^{\text{off}}) \right) \\ &\geq \frac{T\Delta}{32} \exp \left(-2\mathbb{E}[T_k]\Delta^2 \right) \\ &\geq \frac{TV_{\max}}{64} \exp \left(-\frac{TV_{\max}^2}{K-1} \right) \\ &\geq \frac{1}{64e^8} TV_{\max}, \end{aligned}$$

1124 where the last inequality comes by the condition $2\sqrt{KT} > TV_{\max} + \sqrt{2/\tau_*'} > TV_{\max}$, this implies
 1125 $TV_{\max}^2/(K-1) \leq 4K/(K-1) \leq 8$.

1126 **Case 3:** When $2\sqrt{KT} \leq T \cdot (V_{\max} + \sqrt{2/\tau_*'})$, we derive a lower bound of

$$\Omega \left(\min \left\{ \sqrt{KT}, \left(\sqrt{\frac{1}{\tau_*'}} + V_{\max} \right) \cdot T \right\} \right) = \Omega(\sqrt{KT}).$$

1127 The analysis for this case largely follows that of **Case 2**. We use the same construction of reward distributions P ,
 1128 Q , but now take $\Delta = \sqrt{(K-1)/(4T)}$.

1129 From the properties of the associated Linear Program problem in Theorem 2, we know that $\tau'_* \geq 2T/K$, this
 1130 implies: $\sqrt{2T}/\sqrt{\tau'_*} \leq \sqrt{KT}$, them by the condition under this case, we get:

$$\sqrt{KT} \leq TV_{\max} + \sqrt{2T}/\sqrt{\tau'_*} - \sqrt{KT} \leq TV_{\max},$$

1131 which further leads to:

$$|\mu_i^{\text{off}} - \mu_i| \leq 2\Delta = \sqrt{\frac{K-1}{T}} \leq \frac{1}{T} \cdot \sqrt{KT} \leq V_{\max}, \quad \forall i \in [K].$$

1132 Hence, the constructed distributions P, Q satisfy the bias constraint and belong to the class \mathcal{I}_V .

1133 Following the same procedure in Case 2, we have:

$$\begin{aligned} \text{Reg}_P(T) + \text{Reg}_Q(T) &\geq \frac{1}{2} \left[\sigma(\Delta) - \frac{1}{2} \right] \left(\Pr_P(\mathbb{E}[T_1] \leq T) + \Pr_Q(\mathbb{E}[T_1] > T) \right) \\ &\geq \frac{T\Delta}{32} \exp(-2\mathbb{E}[T_k]\Delta^2) \\ &\geq \frac{1}{64e} \sqrt{(K-1)T}. \end{aligned}$$

1134 The Theorem is proved. \square

1135 E Proof of Theorem 3 and Supplementary Regret Bounds

1136 E.1 Instance Dependent Regret Upper Bound

1137 The analysis of this algorithm is similar to Algorithm 1, the only key change here is the position of the relative
 1138 data and the stochastic data, therefore, the UCB^{hyb} , UCB and good event \mathcal{E} changed to the structure below. Note
 1139 that elimination based algorithm should consider all cases from $t = 1, \dots, T$, first we define good events for
 1140 both hybrid and online UCBs, let

$$\begin{aligned} \mathcal{E}_t^{\text{hyb}}(p_{i,j}) &= \left\{ p_{i,j} \leq \text{UCB}^{\text{hyb}}(a_i, a_j) \leq p_{i,j} + 2\sqrt{\frac{\log(1/\delta_t)}{2(N_{i,j} + \frac{T_i(t)T_j(t)}{T_i(t)+T_j(t)})}} + \frac{N_{i,j}}{N_{i,j} + \frac{T_i(t)T_j(t)}{T_i(t)+T_j(t)}}\omega_{i,j} \right\}, \\ \mathcal{E}_t(p_{i,j}) &= \left\{ \mu_i \leq \text{UCB}(a_i, a_j) \leq p_{i,j} + 2\sqrt{\frac{\log(1/\delta_t)}{2} \cdot \frac{T_i + T_j}{T_i T_j}} \right\}. \end{aligned}$$

1141 with

$$\mathcal{E} = \bigcap_{t \in [T]} \bigcap_{i,j \in [K]} \left(\mathcal{E}_t(p_{i,j}) \cap \mathcal{E}_t^{\text{hyb}}(p_{i,j}) \right).$$

1142 Following the proof of Lemma 3 will derive $\Pr(\mathcal{E}) \geq 1 - 2K(K+1) \sum_{t=1}^T \delta_t$.

1143 Then we move to the analysis of the property of this algorithm, which is to detect the point of which an arm is
 1144 eliminated at time t . we claim the result as follows:

1145 **Claim 3.** For Algorithm 2 under good event \mathcal{E} , arm a_i will be eliminated if it satisfies the following inequality:

$$\Delta_i > \min \left\{ 2\sqrt{\frac{\log(1/\delta_t)}{2N_{1,i} + T_i(t)}} + \frac{2N_{1,i}}{2N_{1,i} + T_i(t)}\omega_{1,i}, 2\sqrt{\frac{\log(1/\delta_t)}{T_i(t)}} \right\}.$$

1146 *Proof of the claim.* According to the elimination criterion of HybUCB-RA (Line 6), arm a_i will be eliminated
 1147 if either of the following inequalities holds:

- 1148 1. $\text{UCB}(a_i, a_j) < \frac{1}{2}, \forall a_j \in \mathcal{A}$,
- 1149 2. $\text{UCB}^{\text{hyb}}(a_i, a_j) < \frac{1}{2}, \forall a_j \in \mathcal{A}$.

1150 This also includes:

$$\text{UCB}(a_i, a_1) < \frac{1}{2} \quad \text{and} \quad \text{UCB}^{\text{hyb}}(a_i, a_1) < \frac{1}{2}.$$

1151 When $\Delta_i > 2\sqrt{\frac{\log(1/\delta_t)}{T_i(t)}}$, and the event \mathcal{E} holds, we derive

$$\text{UCB}(a_i, a_1) \leq p_{i,1} + 2\sqrt{\frac{\log(1/\delta_t)}{2} \cdot \frac{T_i + T_j}{T_i T_j}} \approx p_{i,1} + 2\sqrt{\frac{\log(1/\delta_t)}{T_i(t)}} < \frac{1}{2},$$

1152 where the first inequality follows from the definition of event \mathcal{E} , and the final inequality results from the
 1153 condition $\Delta_i > 2\sqrt{\frac{\log(1/\delta_t)}{T_i(t)}}$. The approximation holds because when arm a_i remains active in round t , we have
 1154 $|T_i(t) - T_1(t)| \leq 1$, making $\frac{T_i(t)T_1(t)}{T_i(t)+T_1(t)} = \frac{T_i(t)}{2} + O(1)$. The case for $\text{UCB}^{\text{hyb}}(a_1, a_i)$ follows symmetrically,
 1155 we derive when $\Delta_i > 2\sqrt{\frac{\log(1/\delta_t)}{2N_{1,i}+T_i(t)}} + \frac{2N_{1,i}}{2N_{1,i}+T_i(t)}\omega_{i,1}$, the following condition holds:

$$\text{UCB}^{\text{hyb}}(a_i, a_1) \lesssim 2\sqrt{\frac{\log(1/\delta_t)}{2N_{1,i}+T_i(t)}} + \frac{2N_{1,i}}{2N_{1,i}+T_i(t)}\omega_{i,1} < \frac{1}{2}.$$

1156 Combining them together completes the proof. \square

1157 Now we move on to the complete proof of this algorithm.

1158 *Proof.* (Complete Proof of Algorithm 2) Define

$$g(a_i) = \max \left\{ \frac{1}{\Delta_i^2} \left[2\log(1/\delta_t) + 2N_{i,1}\Delta_i(\omega_{i,1} - \Delta_i) + \sqrt{4\log^2(1/\delta_t) + 8\log(1/\delta_t)N_{i,1}\omega_{i,1}\Delta_i} \right], 0 \right\},$$

$$f(a_i) = \begin{cases} -1 & \text{if } 2\omega_{i,1} \leq \Delta_i, \\ \frac{2\log(1/\delta_t)(2\omega_{i,1}-\Delta_i)}{\Delta_i(\omega_{i,1}-\Delta_i)} - N_{i,1} & \text{if } 2\Delta_i > 2\omega_{i,1} > \Delta_i, \\ 1 & \text{otherwise.} \end{cases}$$

1159 By Lemma 5 and Claim 3, taking $\Delta' = \Delta_i$, $\omega' = \omega_{i,1}$, $N' = 2N_{i,1}$, $T' = T_i$, $\delta' = 1/\delta_t$ we could derive:

$$T_i(t) \leq \begin{cases} \frac{4\log(1/\delta_t)}{\Delta_i^2} & \text{if } f(a_i) \geq 0, \\ g(a_i) & \text{otherwise.} \end{cases}$$

1160 Take $\delta_t = \frac{1}{2K(K+1)T^2}$, the expected regret for bad event becomes:

$$T\Delta_{\max} \times \Pr(\mathcal{E}^c) \leq T\Delta_{\max} \cdot 2K(K+1)T \frac{1}{2K(K+1)T^2} = \Delta_{\max}.$$

1161 Combining the good event together, the regret is:

$$\begin{aligned} \text{Reg}_T(T) &= \sum_{t=1}^T \mathbb{E} [\Delta_i \cdot [\mathbf{1}(\mathcal{E}_t) + \mathbf{1}(\mathcal{E}_t^c)]] \\ &\leq \sum_{t=1}^T \mathbb{E} [\Delta_i \cdot \mathbf{1}(\mathcal{E}_t)] + \Delta_{\max} \sum_{t=1}^T \mathbb{E} [\mathbf{1}(\mathcal{E}_t^c)] \\ &\leq \Delta_{\max} + \sum_{f(a_i) \leq 0} g(a_i)\Delta_i + \sum_{f(a_i) > 0} \frac{8\log(\sqrt{2K(K+1)}T)}{\Delta_i}. \end{aligned}$$

1162 Applying $\sqrt{x^2 + 2ax} < a + x$ to the final regret leads to the simplified version, which is:

$$\text{Reg}(T) = \Delta_{\max} + \sum_{i \in [K]} \max \left\{ \frac{8\log(\sqrt{2K(K+1)}T)}{\Delta_i} - 2N_{i,1} \max\{\Delta_i - 2\omega_{i,1}, 0\}, 0 \right\}.$$

1163 \square

1164 E.2 Instance Independent Regret Upper Bound

1165 **Theorem 6.** Choosing $\delta_t = \frac{1}{2K(K+1)T^2}$, the gap independent bound of HybElimUCB-RA satisfies:

$$O \left(\min \left\{ \sqrt{KT \log(T)} - \sum_{i \in [K]} \text{Saving}(a_i), \left(\sqrt{\frac{\log(T)}{\tau_*}} + V_{\max} \right) \cdot T \right\} \right).$$

1166 where $\text{Saving}(a_i) = \min \left\{ \frac{8 \log(\sqrt{2K(K+1)}T)}{\Delta_i}, 2N_{i,1} \max\{\Delta_i - 2\omega_{i,1}, 0\} \right\}$, and τ_* is the optimal solution to
 1167 the below linear programming problem:

$$\begin{aligned} & \max_{\tau, t_i} \quad \tau, \\ & \text{s.t.} \quad \tau \leq t_i/2 + N_{i,1} \quad \forall i \in [K], \\ & \quad \sum_{i \in [K]} t_i = T, \\ & \quad \tau \geq 0, t_i \geq 0 \quad \forall i, j \in [K]. \end{aligned}$$

1168 *Proof.* The proof of this instance regret lower bound is symmetric to the proof of the instance dependent bound
 1169 of HybUCB-AR, which is in C.7.1 and C.7.2. By the definition of regret bound and Claim 3, the expected
 1170 regret is upper bounded by:

$$\begin{aligned} \text{Reg}(T) &= \sum_{t=1}^T \Delta_{(A_t)} \\ &\leq \sum_{t=1}^T \min \left\{ 2\sqrt{\frac{1}{2N_{1,A(t)} + T_{A_t}(t)} \log\left(\frac{1}{\delta_t}\right)} + \frac{2N_{1,A(t)}}{2N_{1,A(t)} + T_{A_t}(t)} \omega_{1,A(t)}, 2\sqrt{\frac{\log(1/\delta_t)}{T_{A_t}(t)}} \right\} \\ &\leq \sum_{t=1}^T \min \left\{ 2\sqrt{\frac{1}{2N_{1,A(t)} + T_{A_t}(t)} \log\left(\frac{1}{\delta_t}\right)} + \frac{4N_{1,A(t)}}{2N_{1,A(t)} + T_{A_t}(t)} V_{1,A(t)}, 2\sqrt{\frac{\log(1/\delta_t)}{T_{A_t}(t)}} \right\}. \end{aligned} \tag{40}$$

1171 The second term of Equation (40) follows the conventional regret analysis, as provided in Appendix C.7.2,
 1172 Equation (34). The upper bound for the first term can be derived similarly by following the same technique
 1173 in Section C.7.2, where we replace the term $N_i N_j / (N_i + N_j)$ with $N_{1,i}$ and $t_{i,j}$ with $t_i/2$. For the sake of
 1174 conciseness, we omit the detailed of the proof here. \square

1175 E.3 Regret Lower Bound

1176 **Theorem 7.** Let $\Delta \in (0, \frac{1}{2})$ be the gap between the optimal arm and all suboptimal arms. The regret lower
 1177 bounds of HybElimUCB-RA satisfy:

1178 (a) *Instance-dependent bound:*

$$\Omega \left(\frac{8K}{\Delta} \left((1-p) \log T + \log \frac{\Delta}{64C} - \sum_{i \in [K]} \frac{\max\{2\Delta - \omega_i, 0\}^2}{3} N_i \right) \right),$$

1179 where C, p are constants from Definition 1, and $\omega_i = V_i + \mu_i^{\text{off}} - \mu_i$ with $V_i \geq |\mu_i^{\text{off}} - \mu_i|$.

1180 (b) *Instance-independent bound:*

$$\Omega \left(\min \left\{ \sqrt{KT}, \left(\sqrt{\frac{1}{\tau_*'}} + V_{\max} \right) \cdot T \right\} \right).$$

1181 where τ_*' is the optimal solution to the following Linear Programming problem:

$$\begin{aligned} & \max_{\tau', t_i, j} \quad \tau', \\ & \text{subject to} \quad \tau' \leq t_i + N_i, \quad \forall i \in [K], \\ & \quad \sum_{i \in [K]} t_i = T, \quad \tau' \geq 0, t_i \geq 0. \end{aligned}$$

1182 *Proof.* This proof follows a similar structure to the lower bound argument presented in Appendix D. Specifically,
 1183 let k be a fixed arm in $\{2, \dots, K\}$, we construct two instances P and Q such that

$$P_i^{\text{off}} = N(\mu^{\text{off}}, 1) \quad \forall i \in [K], \quad P_i = \begin{cases} N(\mu + \Delta, 1) & \text{if } i = 1, \\ N(\mu, 1) & \text{if } i \neq 1. \end{cases}$$

1184

$$Q_i^{\text{off}} = \begin{cases} \mathcal{N}(\mu + 2\Delta - V_k, 1) & \text{if } i = k \text{ and } \mu^{\text{off}} < \mu + 2\Delta - V_k, \\ \mathcal{N}(\mu^{\text{off}}, 1) & \text{otherwise,} \end{cases} \quad Q_i = \begin{cases} \mathcal{N}(\mu + \Delta, 1) & \text{if } i = 1, \\ \mathcal{N}(\mu + 2\Delta, 1) & \text{if } i = k, \\ \mathcal{N}(\mu, 1) & \text{if } i \in [K] \setminus \{1, k\}. \end{cases}$$

1185 where μ^{off} and μ could be any values such that $|\mu^{\text{off}} - \mu| \leq V_i$ for all $i \in [K] \setminus \{1\}$ and $|\mu^{\text{off}} - (\mu + \Delta)| \leq V_1$.
 1186 In addition, $P \in \mathcal{I}_V$ implies $Q \in \mathcal{I}_V$. Following the conventional analysis, we have:

$$\begin{aligned} \text{Reg}_P(T) &= T \left(\sigma(\Delta) - \frac{1}{2} \right) - \left(\sigma(\Delta) - \frac{1}{2} \right) \mathbb{E}_P [T_1(T) \mathbf{1}_{T_1(T) \leq T/2} + T_1(T) \mathbf{1}_{T_1(T) > T/2}] \\ &\geq T \left(\sigma(\Delta) - \frac{1}{2} \right) - \left(\sigma(\Delta) - \frac{1}{2} \right) \mathbb{E}_P [T_1(T) \mathbf{1}_{T_1(T) \leq T/2}] - \left(\sigma(\Delta) - \frac{1}{2} \right) \mathbb{E}_P [T_1(T) \mathbf{1}_{T_1(T) > T/2}] \\ &= T \left(\sigma(\Delta) - \frac{1}{2} \right) - \frac{T}{2} \left(\sigma(\Delta) - \frac{1}{2} \right) \Pr(T_1(T) \leq T/2) - T \left(\sigma(\Delta) - \frac{1}{2} \right) \Pr(T_1 > T/2) \\ &= \frac{T \left(\sigma(\Delta) - \frac{1}{2} \right)}{2} \Pr(T_1(T) \leq T/2). \end{aligned}$$

1187 and

$$\text{Reg}_Q(T) = \sum_{i=1}^K \Delta_i \mathbb{E}_Q [T_i(T)] \geq (\sigma(\Delta) - \frac{1}{2}) \mathbb{E}_Q [T_1(T)] \geq \frac{T \left(\sigma(\Delta) - \frac{1}{2} \right)}{2} \Pr(T_1(T) > T/2).$$

1188 Following the same procedure we obtain:

$$2CT^p \geq \text{Reg}_P(T) + \text{Reg}_Q(T) \geq \frac{T\Delta}{32} \exp(-KL(P, Q)), \quad (41)$$

1189 where

$$KL(P, Q) = \sum_{i \in [K]} \mathbb{E}_P [T_i(T)] \cdot KL(P_i, Q_i) + \sum_{i \leq j} N_{i,j} \cdot KL(P_{i,j}^{\text{off}}, Q_{i,j}^{\text{off}}).$$

1190 For the first term, we have:

$$\sum_{i \in [K]} \mathbb{E}_P [T_i(T)] \cdot KL(P_i, Q_i) = \mathbb{E}_P [T_k(T)] \frac{(\sigma(2\Delta) - 1/2)^2}{2} \leq \frac{\mathbb{E}_P [T_k(T)] \Delta^2}{8}. \quad (42)$$

1191 For the second term, let N_k denotes the number of time we selected a_k in each pair, we have:

$$\begin{aligned} \sum_{i \leq j} N_{i,j} \cdot KL(P_{i,j}^{\text{off}}, Q_{i,j}^{\text{off}}) &= \sum_{i=2}^K N_{k,i} \cdot d\left(\frac{1}{2}, \max\left\{\frac{1}{2}, \sigma(2\Delta - \omega_k)\right\}\right) \\ &\leq N_k \cdot d\left(\frac{1}{2}, \max\left\{\frac{1}{2}, \frac{2\Delta - \omega_k}{4} + \frac{1}{2}\right\}\right) \\ &\leq N_k \frac{\max\{2\Delta - \omega_k, 0\}^2}{4 - (\max\{2\Delta - \omega_k, 0\})^2} \\ &\leq \frac{1}{3} N_k \max\{2\Delta - \omega_k, 0\}. \end{aligned} \quad (43)$$

1192 Putting Equation (42) and Inequality (43) and (41) together, we have:

$$2C^p \geq \text{Reg}_P(T) + \text{Reg}_Q(T) \geq \frac{T\Delta}{32} \exp\left(-\frac{\mathbb{E}_P [T_k(t)] \Delta^2}{8} - \frac{\max\{2\Delta - \omega_k, 0\}^2}{3} N_k\right).$$

1193 reorganize the structure of the above equation following Appendix D.1 leads to the gap independent bound,
 1194 which is:

$$\frac{8K}{\Delta} \left((1-p) \log T + \log \frac{\Delta}{64C} - \sum_{i \in [K]} \frac{\max\{2\Delta - \omega_i, 0\}^2}{3} N_i \right).$$

1195 as desired. \square

1196 *Gap-Independent Regret Bound.* The proof of the gap-independent lower bound largely follows the argument
 1197 in Appendix D.2, with the only difference being the switch of the offline and online data. Therefore, we
 1198 provide only a sketch of the modified proof. In **Case 1**, since the offline and online share the same distributions,
 1199 the switch does not affect the analysis. In **Case 2** and **Case 3**, where the discrepancy between P and Q
 1200 appears only in the online distributions, the KL divergence term is updated from the original expression to
 1201 $\sum_{i=1}^K T_i \cdot KL(P_i, Q_i) = 2T_k \cdot \Delta$, but the conclusion remains unchanged. The only modification lies in the
 1202 boundary condition for each case, which now compares $2\sqrt{KT}$ with $T \cdot (V_{\max} + 1/\sqrt{T_*})$. This adjustment
 1203 arises because, in the online-stochastic setting, we have $\sum_{i=1}^K T_i = T$ rather than $2T$. \square

F Stochastic Bandits with Offline Relative Feedback: An Alternative Approach

As previously discussed, preference data typically provide less information compared to stochastic feedback. Estimating the underlying utility values μ_i for all $i \in [K]$ solely from preference data is extremely challenging, unless additional structural assumptions are imposed. For the sake of theoretical completeness, we still adopt a preference-based model in Algorithm 2 and build upon the elimination framework. However, this class of algorithms cannot match the performance of vanilla UCB when offline data is absent (i.e., when the offline dataset is empty) or provides no additional information.

Algorithm 3 Offline Relative Online Stochastic: An alternative Approach

Require: an arm set \mathcal{A} , offline dataset $\mathcal{D} = \{(A_i, A_j, Y_{i,j,k}), i, j \in [K], k \in [N_{i,j}]\}$ hyperparameter δ_t and estimated bias $V_{i,j}$ for all (a_i, a_j) pairs.

Initialization: Let $\mathcal{C} = \mathcal{A}$, $T_i = 0, \forall i \in [K]$, and for all $i, j \in [K]$, let $\text{UCB}(a_i, a_j) = +\infty$, $\text{UCB}^{\text{hyb}}(a_i, a_j) = \hat{p}_{i,j}^{\text{off}} + \sqrt{\frac{\log 1/\delta_t}{N_{i,j}}}$.

- 1: **for** $t = 1, \dots, T$ **do**
 - 2: Select Action $A(t) = \arg \max_{a_i \in \mathcal{C}} \text{UCB}(a_i)$.
 - 3: Update selection times $T_{A(t)} = T_{A(t)} + 1$.
 - 4: Record observation $X_{i,k}$ for $i = A(t)$, $k = T_{A(t)}$.
 - 5: For all pairs $i, j \in [K]$, update UCB, UCB^{hyb} according to equation (9), (10).
 - 6: $\mathcal{C} = \mathcal{C} \setminus \{a_i \in \mathcal{C}, \exists a_j \in \mathcal{C} \setminus a_i \text{ s.t. } \text{UCB}^{\text{hyb}}(a_i, a_j) < \frac{1}{2}\}$.
 - 7: **end for**
-

In practice, more flexible algorithmic designs can be considered. In Algorithm 3, we present an alternative approach that aligns with vanilla UCB by utilizing UCB and UCB^{hyb} separately. Specifically, we construct the candidate set $\mathcal{C}_t^{\text{hyb}}$ using hybrid data, and then apply the standard UCB selection rule to choose $A(t)$ within this set.

While such variants may perform well empirically, they lack the strong theoretical guarantees provided by Algorithm 2. In particular, these methods no longer ensure balanced exploration across arms during the online phase. As a result, the sample counts $T_i(t)$ for each arm $a_i \in \mathcal{C}_t$ can become highly imbalanced.

Moreover, since vanilla UCB does not guarantee a theoretical lower bound on $\min\{T_i(t)\}$ for all $i \in [K]$, the confidence radius in UCB^{hyb} —given by $\sqrt{1/(N_{i,j} + \frac{T_i(t)T_j(t)}{T_i(t)+T_j(t)})}$ —can degrade to $\sqrt{1/N_{i,j}}$ in the worst case.

Thereby, the theoretical analysis falls back to a naive two-stage procedure: (i) using offline data to eliminate suboptimal arms, followed by (ii) applying vanilla UCB to the surviving set. Under this simplified strategy, the regret bound becomes:

$$\text{Reg}(T) = O\left(\sum_{a_i \in \mathcal{C}} \frac{\log T}{\Delta_i}\right),$$

where $\mathcal{C} \subseteq \mathcal{A}$ denotes the set of arms retained after offline filtering. Compared to Algorithm 1, such two-phase strategies lack a continuous and elegant theoretical guarantee throughout the learning process. Bridging this gap—by designing practical algorithms that maintain strong regret guarantees even under hybrid and heterogeneous feedback—remains an important direction for future work.

G Supplemental Experiments

We conduct a comprehensive evaluation of HybUCB-AR and HybElimUCB-RA in hybrid bandit settings with heterogeneous offline-online feedback. The experiments aim to demonstrate the algorithms' robustness and superiority over established baselines across diverse scenarios. Our evaluation encompasses two primary environments: (1) synthetic datasets, where we assess scalability with varying numbers of arms K and sensitivity to key parameters, (2) real-world datasets, where we validate practical efficacy using authentic data. For all experiments, we set a set of random seeds to ensure the reproducibility of the experiments. Our experiments were implemented in Python.

G.1 Synthetic Data Experiments

To evaluate the performance of HybUCB-AR and HybElimUCB-RA in controlled settings, we conduct experiments on synthetic datasets. In the first set of experiments, we assess the algorithms' scalability by varying the

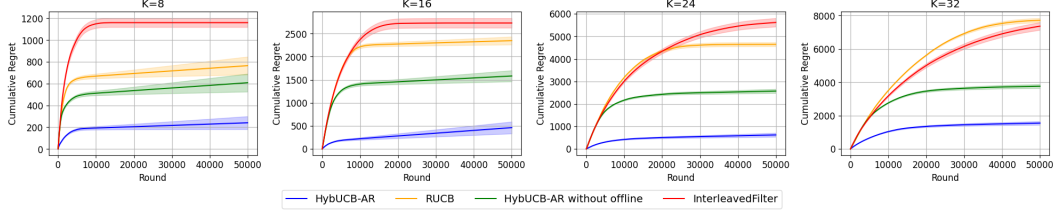


Figure 2: Average cumulative regret of HybUCB-AR for $K = 8, 16, 24, 32$ over 100 runs, with shaded area indicating standard deviation.

number of arms K , comparing their performance against established baselines to demonstrate their effectiveness across different scales. In the second set of experiments, we analyze the sensitivity of HybUCB-AR and HybElimUCB-RA to key parameters (N_i , Δ , and V_i), focusing on how different parameter values affect their convergence behavior. The synthetic environment allows precise control over reward distributions and feedback mechanisms, enabling robust analysis of algorithmic behavior.

G.1.1 Data Generation

We generate a synthetic K -armed bandit environment to simulate the heterogeneous feedback setting. For each arm $a_i \in \mathcal{A}$, we generate the offline and online reward means, μ_i^{off} and μ_i , as follows. The offline mean of the best arm, μ_1^{off} , is drawn uniformly from $[0.5 + \Delta, 1]$, where $\Delta \in (0, 0.5)$ denotes the gap between μ_1^{off} and μ_2^{off} ; we then set $\mu_2^{\text{off}} = \mu_1^{\text{off}} - \Delta$. For the remaining $K - 2$ arms, μ_i^{off} is drawn uniformly from $[0, \mu_2^{\text{off}}]$. The online reward mean is defined as

$$\mu_i = \mu_i^{\text{off}} + d_i \cdot \text{bias}, \quad \text{where } d_i \in \{-1, 1\} \text{ is chosen uniformly at random, } \forall i \in [K].$$

All absolute feedback $\{i, X_i\}$ for each $i \in [K]$ is generated from a Gaussian distribution $\mathcal{N}(\mu_i, 1)$, and relative feedback $\{i, j, Y_{i,j}\}$ for all $i, j \in [K]$ is generated from a Bernoulli distribution $\text{Bernoulli}(p_{i,j})$ according to the Bradley-Terry model.

G.1.2 Experiments for HybUCB-AR

HybUCB-AR leverages offline absolute rewards to enhance online dueling bandit learning. We compare it against three baselines: HybUCB-AR without offline data ($N_i = 0$), Relative Upper Confidence Bound (RUCB) (Zoghi et al., 2014), and Interleaved Filter 2 (IF2) (Yue et al., 2012).

Performance Comparison and Scalability ($K = 8, 16, 24, 32$) Setup. We evaluate HybUCB-AR's scalability across $K = 8, 16, 24, 32$ arms. We set the sub-optimal gap $\Delta = 0.1$, the offline-online bias to 0.1, and the number of offline samples $N_i = 500$. Each algorithm runs for $T = 50,000$ rounds with 100 independent trials. We set $\delta_t = 0.02$, for RUCB, we set the confidence parameter $\alpha = 0.51$, following Zoghi et al. (2014).

Results. Figure 2 shows the average cumulative regret over 100 trials, with shaded area indicating standard deviation. Without offline data, HybUCB-AR, by maximizing the informative pair in the confidence set, achieves a regret reduction of 15–40% compared to RUCB across all K . The performance advantage grows with larger K , highlighting its scalability. And across all K , HybUCB-AR surpass baseline when heterogeneous offline data is included.

Parameter Sensitivity Analysis Setup. We analyze HybUCB-AR's sensitivity to three parameters: For all $i \in [K]$, we set the number of offline samples, $N_i \in \{100, 300, 500\}$ the sub-optimal gap $\Delta \in \{0.05, 0.1, 0.2\}$, and the offline-online bias, $V_i \in \{0.01, 0.05, 0.1\}$. We fix $K = 20$, run for 30,000 rounds with 100 trials, and use default values ($N_i = 500$, $\Delta = 0.1$, $V_i = 0$) for all parameters unless otherwise specified.

Results. Figure 3 shows the cumulative regret. As N_i increases, the algorithm converges faster, leading to lower cumulative regret. Smaller Δ increases regret due to harder arm differentiation. Larger V_i reduces the ability to utilize offline data, resulting in slower convergence. One key insight is that the accumulated regret does not decrease linearly with the size of the offline data, while the increase in regret caused by bias shift grows approximately linearly with the magnitude of the bias. This observation is consistent with our theoretical regret analysis in Theorem 1.

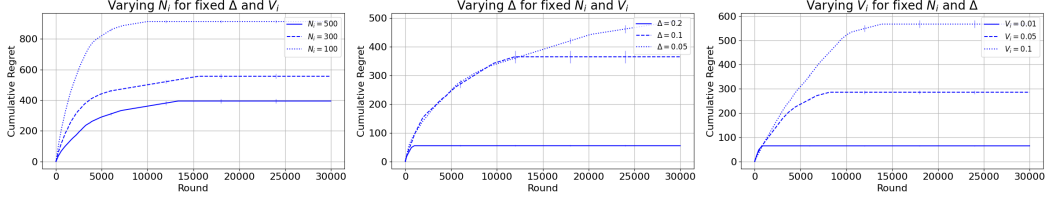


Figure 3: Parameter sensitivity of HybUCB-AR, showing effects of N_i , Δ , and V_i on cumulative regret, with vertical lines indicating standard deviation every 6,000 rounds.

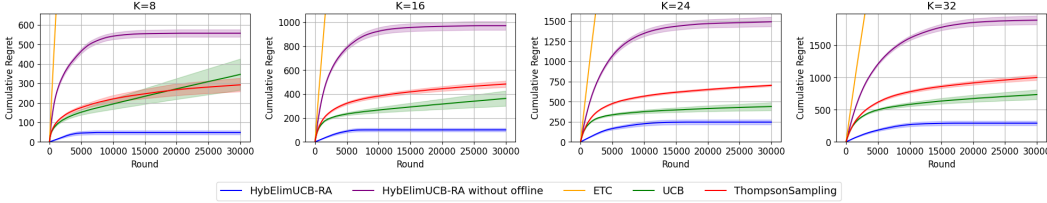


Figure 4: Average cumulative regret of HybElimUCB-RA for $K = 8, 16, 24, 32$ over 100 trials, with shaded area indicating standard deviation.

1277 G.1.3 Experiments for HybElimUCB-RA

1278 HybElimUCB-RA uses offline relative preferences to improve online stochastic learning. We compare it against
 1279 four baselines: HybElimUCB-RA without offline data (with $N_i = 0$), Explore-Then-Commit (ETC) (Robbins,
 1280 1952), Upper Confidence Bound (UCB) (Lai and Robbins, 1985), and Thompson Sampling (TS) (Agrawal and
 1281 Goyal, 2013).

1282 **Performance Comparison and Scalability ($K = 8, 16, 24, 32$) Setup.** We evaluate HybElimUCB-
 1283 RA’s scalability across $K = 8, 16, 24, 32$ arms. The sub-optimal gap Δ is set to 0.1, the offline-online bias to
 1284 0.01, and the number of offline samples to $N_i = 500$. Each algorithm runs for $T = 30,000$ rounds with 100
 1285 independent trials. HybElimUCB-RA and UCB use $\delta_t = 0.05$, ETC uses an exploration phase of 500, and TS
 1286 assumes a Gaussian prior with variance 1.

1287 **Results.** Figure 4 shows the average cumulative regret over 100 trials, with shaded area indicating standard
 1288 deviation. When leveraging offline relative preferences, HybElimUCB-RA outperforms all baselines. The
 1289 performance advantage grows with larger K , highlighting its scalability. Without offline data, HybElimUCB-RA
 1290 matches the performance of elimination-based UCB.

1291 **Parameter Sensitivity Analysis Setup.** We analyze HybElimUCB-RA’s sensitivity to three parameters:
 1292 For all $i \in [K]$, we set the number of offline samples $N_i \in \{100, 200, 500\}$, the sub-optimal gap $\Delta \in$
 1293 $\{0.05, 0.1, 0.2\}$, and the offline-online bias, $V_i \in \{0.01, 0.05, 0.1\}$. We fix $K = 10$, run for 25,000 rounds
 1294 with 100 trials, and use default values ($N_i = 100$, $\Delta = 0.1$, $V_i = 0.01$) for all parameters unless otherwise
 1295 specified.

1296 **Results.** Figure 5 shows the cumulative regret, with vertical lines indicating standard deviation every 6,000
 1297 rounds. As N_i increases, the algorithm converges faster, leading to lower cumulative regret. Smaller Δ increases
 1298 regret due to harder arm differentiation. Larger V_i reduces the ability to utilize offline data, resulting in slower
 1299 convergence. This result is similar to the result we obtained in Figure 3, demonstrating the consistent of our
 1300 algorithm.

1301 G.2 Real Data Experiments

1302 We evaluate HybUCB-AR and HybElimUCB-RA on MovieLens-20M and Yelp datasets to validate their
 1303 performance in real-world hybrid bandit settings. Experiments compare both algorithms against established
 1304 baselines, leveraging offline data to enhance online learning. All experiments are implemented in Python.

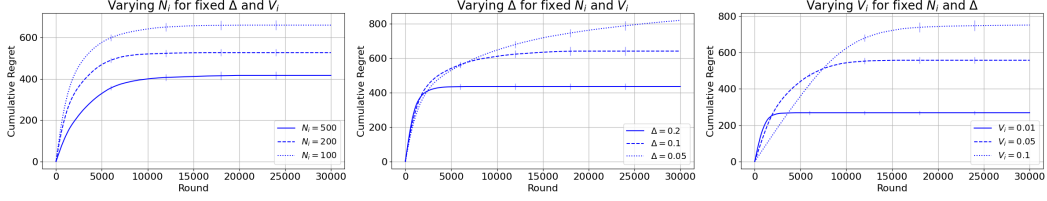


Figure 5: Parameter sensitivity of HybElimUCB-RA, showing effects of N_i , Δ , and V_i on cumulative regret, with vertical lines indicating standard deviation every 6,000 rounds.

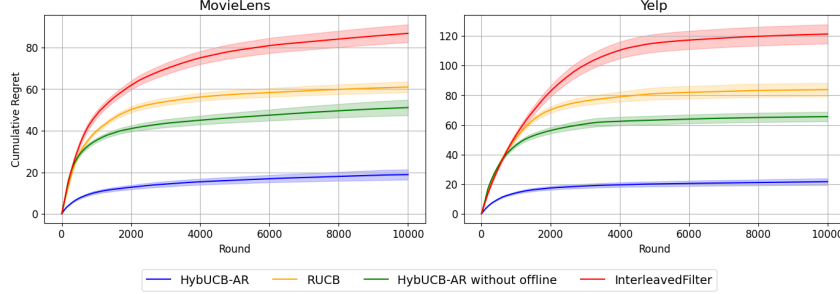


Figure 6: Performance of HybUCB-AR on MovieLens 20M ($K = 10$), with shaded areas showing standard deviation across 50 trials.

G.2.1 Data Preparation

The **MovieLens-20M** dataset contains 20,000,000 ratings (1–5) for movies. We normalize ratings to $[0, 1]$ by dividing by 5, select movies with at least 100 ratings, take the top 100 by rating count, and randomly sample $K = 10$ movies as arms. For HybUCB-AR, offline data consists of 100,000 normalized ratings. For HybElimUCB-RA, offline data includes 1,000 preference duels per arm pair, generated by sampling ten ratings per arm and comparing their means.

The **Yelp** dataset, sourced from the Yelp Academic Dataset, includes business reviews with star ratings (1–5). We normalize ratings to $[0, 1]$, select businesses with at least 100 ratings, take the top 100 by rating count, and randomly sample $K = 10$ businesses. Offline data follows the same structure as MovieLens.

Online feedback is generated using the environment module designed by us. For HybUCB-AR, duels sample 3 ratings per arm, with the higher average rating winning (ties resolved randomly). For HybElimUCB-RA, rewards are the average of 30 sampled ratings per arm.

G.2.2 Experiments for HybUCB-AR

HybUCB-AR leverages offline absolute rewards to improve online dueling bandit learning, using a V-matrix defined as $V = (V_{i,j})_{i,j \in [K]}$. We compare it against HybUCB-AR without offline data ($N_i = 0$), RUCB (Zoghi et al., 2014), and Interleaved Filter 2 (IF2) (Yue et al., 2012).

Performance Comparison Setup. We set $K = 10$, run each algorithm for 10,000 rounds over 50 trials. The environment module samples 3 ratings per arm for duels. We set $\delta_t = 0.02$. Offline data comprises 100,000 normalized ratings.

Results. Figures 6 show the average cumulative regret, with shaded areas indicating standard deviation. Without offline data, HybUCB-AR outperforms RUCB by 15–25% on both datasets. HybUCB-AR shows the same good performance as in the synthetic experiments.

G.2.3 Experiments for HybElimUCB-RA

HybElimUCB-RA uses offline relative preferences to enhance online absolute reward learning, employing the same V-matrix as HybUCB-AR. We compare it against HybElimUCB-RA without offline data ($N_i = 0$), Explore-Then-Commit (ETC) (Robbins, 1952), Upper Confidence Bound (UCB) (Lai and Robbins, 1985), and Thompson Sampling (TS) (Agrawal and Goyal, 2013).

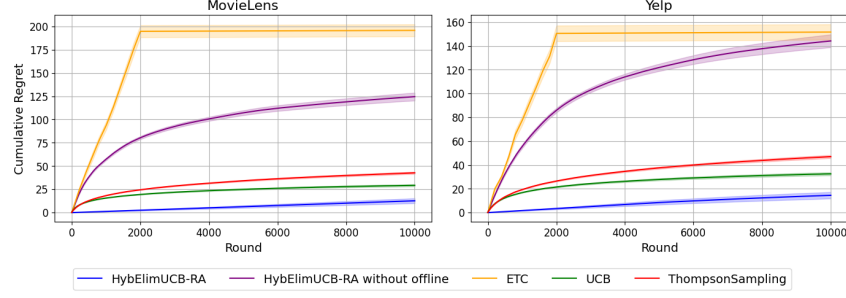


Figure 7: Performance of HybElimUCB-RA on MovieLens-20M and Yelp with ($K = 10$). Shaded areas showing standard deviation across 50 runs.

Performance Comparison Setup. We set $K = 10$, run each algorithm for 10,000 total arm pulls over 50 trials. The environment module samples 30 ratings per arm to compute rewards. HybElimUCB-RA and UCB use $\delta_t = 0.05$, ETC uses an exploration phase of 200 pulls per arm, and TS assumes a Gaussian prior with mean 0.5 and variance 1. Offline data consists of 1,000 relative preference duels per arm pair.

Results. Figure 7 shows the average cumulative regret. In two real data environments, HybElimUCB-RA combined with offline data can achieve good results and is better than the baseline.

H Useful Lemmas

Lemma 8 (Chernoff Inequality, Theorem 5.3 in Lattimore and Szepesvári (2020)). *If X is σ sub-Gaussian, then for any $\epsilon > 0$,*

$$\Pr(X \geq \epsilon) \leq \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right).$$

Lemma 9 (Hoeffding's Lemma). *Let X be any real-valued random variable such that $a \leq X \leq b$ almost surely, i.e., with probability one. Then, for all $\lambda \in \mathbb{R}$,*

$$\mathbb{E}\left[e^{\lambda X}\right] \leq \exp\left(\lambda \mathbb{E}[X] + \frac{\lambda^2(b-a)^2}{8}\right),$$

or equivalently,

$$\mathbb{E}\left[e^{\lambda(X - \mathbb{E}[X])}\right] \leq \exp\left(\frac{\lambda^2(b-a)^2}{8}\right).$$

Lemma 10 (sub-Gaussian Property of Lipschitz Transformations). *Let X be a random variable that is σ sub-Gaussian, meaning $\mathbb{E}[e^{\lambda X}] \leq e^{\lambda^2 \sigma^2 / 2}$ for all $\lambda \in \mathbb{R}$. If $f : \mathbb{R} \rightarrow \mathbb{R}$ is L -Lipschitz continuous, then $f(X)$ is $2L\sigma$ sub-Gaussian.*

Proof. A related result appears in Theorem 5.5 of Boucheron et al. (2013), which shows that $f(X)$ is $L\sigma$ sub-Gaussian when X is strictly Gaussian. For a general σ sub-Gaussian X , we establish that $f(X)$ is $2L\sigma$ sub-Gaussian.

Let Y be an independent copy of X . We aim to bound the moment-generating function of $(f(X) - \mathbb{E}[f(X)])^2$. Consider a constant $c > 0$:

$$\begin{aligned}
\mathbb{E} \left[\exp \left(\frac{(f(X) - \mathbb{E}[f(X)])^2}{c^2} \right) \right] &= \mathbb{E} \left[\exp \left(\frac{(f(X) - \mathbb{E}[f(Y)])^2}{c^2} \right) \right] \\
&= \int_{\mathbb{R}} P(X \in dx) \exp \left(\frac{(f(x) - \mathbb{E}[f(Y)])^2}{c^2} \right) \\
&\leq \int_{\mathbb{R}} P(X \in dx) \mathbb{E} \left[\exp \left(\frac{(f(x) - f(Y))^2}{c^2} \right) \right] \\
&= \mathbb{E} \left[\exp \left(\frac{(f(X) - f(Y))^2}{c^2} \right) \right] \\
&\leq \mathbb{E} \left[\exp \left(\frac{L^2(X - Y)^2}{c^2} \right) \right] \\
&\leq \mathbb{E} \left[\exp \left(\frac{2L^2X^2 + 2L^2Y^2}{c^2} \right) \right] \\
&= \left(\mathbb{E} \left[\exp \left(\frac{2L^2X^2}{c^2} \right) \right] \right)^2 \\
&\leq \mathbb{E} \left[\exp \left(\frac{4L^2X^2}{c^2} \right) \right] \leq 2.
\end{aligned}$$

1352 Here, the second equality uses the law of total expectation; the first inequality applies Jensen's inequality;
1353 the second inequality uses the L -Lipschitz property, $|f(x) - f(y)| \leq L|x - y|$; the third inequality uses
1354 $(X - Y)^2 \leq 2X^2 + 2Y^2$; and the final steps exploit the independence of X and Y . The last inequality holds
1355 because X is σ sub-Gaussian, so $\mathbb{E}[e^{X^2/\sigma^2}] \leq 2$. Setting $4L^2/c^2 = 1/\sigma^2$, we get $c = 2L\sigma$. Thus, $f(X)$ is
1356 $2L\sigma$ sub-Gaussian, as desired. \square

1357 **Lemma 11 (MVUE).** Let \hat{X}_1, \hat{X}_2 be 2 independent estimator of μ , with variance σ_1^2 and σ_2^2 , then the minimum
1358 variance unbiased estimator of X is:

$$\hat{X} = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \hat{X}_1 + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} \hat{X}_2,$$

1359 with variance $\frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2}$.