Safely Optimal: Pure Exploration in Bandits with Unknown Linear Constraints

Udvas Das [1] **Achraf Azize** [2] **Debabrota Basu** [1]

Abstract

In real-world sequential decision making tasks, such as adaptive clinical trials, user studies for recommender systems, we aim to identify a reward-maximising policy over possible decisions while abiding by safety and resource constraints, which might not be exactly known *a priori*. We study these problems in terms of identifying a safe and optimal policy with a fixed confidence level in the multi-armed bandit setting, where pulling each arm yields correlated, noisy reward and cost signals. First, we show the statistical instability of exactly identifying the safely optimal policy. Then, we derive two information-theoretic lower bounds for identifying a safe and ϵ-optimal policy for unstructured and linear bandits with multiple linear constraints. These results significantly extend the existing lower bound on identifying the unique best arm under a single constraint. Then, we design an algorithmic framework, PRUNE, that plugs-in the sequentially estimated reward and cost means in the lower bounds, optimises them efficiently to find a sampling strategy, and deploys two novel stopping rules to output accurate policy recommendations. Finally, we prove that PRUNE is asymptotically optimal for both the unstructured and linear bandits with unknown linear constraints.

1. Introduction

Sequential decision making under uncertainty is a central problem across science, economics, and engineering. Stochastic multi-armed bandits (Lattimore & Szepesvári, 2020) offer a clean formalisation of this problem, where at every round an agent pulls one of K decisions (called arms), receives a noisy reward, and must balance gathering information (*exploration*) against acting on it (*exploitation*). For the fixed-confidence *pure exploration* (or *best-arm identification*, in brief BAI) tasks, the agent collects evidences by strategically pulling the arms and stops as soon as it can certify validity of its recommended policy being a reward-maximising policy (or arm) with a pre-decided confidence level $1-\delta$ (Even-Dar et al., 2006; Bubeck et al., 2009; Kaufmann et al., 2016).

Motivation. In many consequential applications the reward-maximising arm is not a valid recommendation on its own. The agent must ultimately commit to a *policy*: a distribution over arms that simultaneously optimises reward *and* satisfies operational constraints. For example, in an *adaptive clinical trial* (Amani et al., 2019; Thomas et al., 2015), we test different dosage levels, each corresponding to an individual arm, with incoming patients to recommend a certified and efficacy maximising treatment policy that maps patient profiles to doses while keeping the *average toxicity* below a statutory threshold. The toxicity coefficients are unknown a priori and also be estimated from noisy side-effect measurements that are *correlated* with the efficacy signal, as applying a dosage simultaneously reveals information about both benefit and harm. Crucially, the answer the clinician must certify is not a single arm but a mixture policy $\boldsymbol{p} \in \triangle_K$ satisfying $B\boldsymbol{p} \leq 0$, where B is the unknown mean cost (toxicity) matrix. Analogous problems appears in *energy dispatch* (Pacchiano et al., 2020; Bui et al., 2024; Kumari & Tanwar, 2021), where a grid operator must commit to an allocation policy that has to respect supply, capacity, and fairness limits, and in *portfolio construction* (Li et al., 2024; Wang & Liu, 2025) with linear budget and risk constraints. In all of these problems, the agent plays arms to finally identify a policy that is simultaneously *safe* (feasible under the true constraints) and *optimal* (close to the best feasible reward).

Related Work. Though a rich line of work studies how to minimise cumulative regret while respecting safety or resource budgets during learning (Pacchiano et al., 2020, 2024; Gangrade et al., 2024a), pure exploration with fixed confidence demands a different algorithm design.

Pure exploration under constraints. (Carlsson et al., 2024) study this problem of identifying the safely optimal policy for *known* constraints B. They derive a lower bound and a Track-and-Stop style algorithm extending the tech-

[1]Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 - CRIStAL, Lille, France [2]FairPlay Joint Team, CREST, ENSAE, IP Paris. Correspondence to: Udvas Das <udvas.das@inria.fr>.

ICML 2026 Workshop on Hypothesis Testing, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

niques of Garivier & Kaufmann (2016). Our work affirmatively resolves their open problem, i.e. *extending the lower bound and algorithm design to the practically important case where B is estimated from data*. Previously, Das & Basu (2026) proposed a novel Lagrangian relaxations of the lower bound. But this inflates the achieved sample complexity by the condition number of B– a price that one can avoid by deriving and solving the exact lower bound.

BAI with an unknown constraint. Lardy et al. (2025) study fixed confidence BAI with a *single* unknown constraint, and design an asymptotically optimal algorithm enabled by a lower bound. Their problem setup considers the optimal policy to be only an arm and they conjecture that handling $d > 1$ unknown constraints would incur exponential complexity in d. We resolve both the limitations.

Feasibility testing. Gangrade et al. (2024b) study feasibility testing of a linear program $B\mathbf{z} \leq 0$. They derive a lower bound sensitive to B's slack, and design efficient algorithms. But *identifying the safely optimal policy is a strictly harder problem*. It requires certifying both feasibility *and* optimality, and inherits a similar instability when the optimal policy is close to constrained boundary.

In this work, we ask: *what is the fundamental hardness of identifying the safely optimal arm under unknown linear constraints, and can we design a statistically and computationally efficient algorithm to achieve this limit?*

Contributions. We affirmatively solve this question with the following four axes.

1. Statistical instability of exact identification (Lemma 1). We show that when the optimal policy $\boldsymbol{p}^\star(\boldsymbol{\nu})$ has at least one tight constraint, $\mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta] \to \infty$ for any δ-correct algorithm, making exact identification statistically intractable. This echoes the feasibility-testing hardness of (Gangrade et al., 2024b) and the slack assumption ubiquitous in constrained regret minimisation (Pacchiano et al., 2020, 2024), and motivates a natural relaxation to ε-optimal identification.

2. Novel information-theoretic lower bounds. We derive tight lower bounds on the expected stopping time for identifying an ϵ-optimal safe policy, for both unstructured multi-armed bandits with independent arms (Theorem 1) and linear bandits (Theorem 2), each under d unknown linear constraints. Both bounds expose a fundamental two-way trade-off: the agent must simultaneously defend against an adversary that renders the recommended policy infeasible (*feasibility pressure*) and one that reveals a superior alternative (*optimality pressure*), with the coupling between reward and cost observations entering through their joint covariance. We provide closed-form expressions for jointly multivariate Gaussian distributions (Corollary 1) and for the single-constraint linear bandit (Corollary 2), from which we recover the lower bounds of Carlsson et al. (2024) for known constraints, Lardy et al. (2025) for BAI under a single unknown constraint, and the classic BAI lower bound for linear bandits (Soare, 2015).

3. Algorithm Design: PRUNE (Algorithm 1). We propose PRUNE, which combines three components. First, a *Frank-Wolfe allocation player* (Wang et al., 2021a) that plugs empirical reward-cost statistics into the lower bound and optimises over the sampling simplex. We carefully construct a sub-differential set (Equation (8)) to handle the double non-smoothness arising from the minimum operator between reward and constraint terms, and show that gradient and curvature are bounded, establishing convergence. Second, two novel *Chernoff stopping rules* (Lemma 2) for multivariate Gaussian and linear bandit settings, grounded in the mixture-martingale technique of Kaufmann & Koolen (2021), with thresholds that are valid for all t and δ. Third, a *recommendation rule* that solves the estimated (SOP) or (SOP-lin) at stopping time. We additionally show that the GLRs in our stopping rules, when normalised by the runtime t, yield valid e-values under constraints, connecting PRUNE to the sequential testing literature (Clerico, 2024; Agrawal & Ramdas, 2025).

4. Asymptotic optimality and computational efficiency. We prove that PRUNE is asymptotically optimal in both settings (Theorem 3), and derive a non-asymptotic sample complexity bound (Appendix K). Per-round, PRUNE incurs $\mathcal{O}(K \max\{K, d\})$ and $\mathcal{O}(d_{\text{lin}}^2 \max\{K, d_{\text{lin}}, d\})$ operations for unstructured and linear bandits, respectively i.e., *polynomial* in all problem dimensions. This directly refutes the conjecture of (Lardy et al., 2025) that handling d unknown constraints requires exponential cost in d, while simultaneously closing the open problem of (Carlsson et al., 2024) and improving over the condition-number-inflated complexity of Das & Basu (2026).

2. Problem Formulation

We formalise pure exploration under unknown linear constraints for unstructured bandits with independent arms and linear bandits. In both the settings, the reward and cost signals are correlated.

2.1. Unstructured Bandits with Independent Arms and Linear Constraints

An unstructured bandit instance $\boldsymbol{\nu}$ consists of K independent arms (Lattimore & Szepesvári, 2020). Each arm $a \in [K]$ represents a joint distribution over reward and cost $(P_a, C_a) \in M \subseteq \mathcal{P}(\mathbb{R}^{d+1})$. Here, $P_a \in \mathcal{P}(\mathbb{R})$ and and $C_a \in \mathcal{P}(\mathbb{R}^d)$ denote the reward and d-dimensional cost distributions, respectively. Further, $\boldsymbol{\mu} \in \mathbb{R}^K$ is the mean rewards of the arms and $B \in \mathbb{R}^{d\times K}$ denotes the mean costs, i.e. $\boldsymbol{\mu}_a \triangleq \mathbb{E}_{R\sim P_a}[R] \in \mathbb{R}$ and $B_a \triangleq \mathbb{E}_{C\sim C_a}[C] \in \mathbb{R}^d$. The goal is to identify the unique safely optimal policy (SOP)

for the given instance:

$$\boldsymbol{p}^\star(\boldsymbol{\nu}) \triangleq \underset{\boldsymbol{p}\in\triangle_K : B\boldsymbol{p}\leq 0}{\arg\max}\ \boldsymbol{\mu}^\top \boldsymbol{p}\,. \tag{SOP}$$

When the (SOP) reduces to a pure strategy, i.e. a specific arm, this setting reduces to the constrained best-arm identification problem (Lardy et al., 2025) with d-constraints. At each time $t \in \mathbb{N}$, an agent pulls an arm $A_t \in [K]$ according to a history-dependent sampling (or allocation) policy $\boldsymbol{\omega} \in [K]$. Pulling A_t yields a scalar reward $R_t \sim P_{A_t}$ and a cost vector $C_t \sim C_{A_t}$, which are jointly conditionally independent of the observed history $H_{t-1} \triangleq \{(A_t, R_t, C_t)\}_{t=1}^{\tau}$. Here, H_{t-1} denotes the history of arm pulls, rewards, and cost vectors generated by the sampling policies till time t, and $\mathbb{P}_{\boldsymbol{\nu},\boldsymbol{\pi}}$ is the marginal distribution of history. Following the sampling policy to collect evidences for identifying (SOP), the agent decides to stop at step τ to recommend a policy $\hat{\boldsymbol{p}} \in \triangle_K$.

Assumption 1 (Bounded domains)**.** *Reward and constraint means lie in bounded domains, i.e.* $\boldsymbol{\mu} \in \mathcal{D}_{\boldsymbol{\mu}}$ *and* $B_j \in \mathcal{D}_B, \forall j = [d]$. *Additionally, w.l.o.g* $\|B_j\|_2 \leq 1 \forall j \in [d]$.

2.2. Linear Bandits with Linear Constraints

Linear bandits (Soare, 2015; Fiez et al., 2019) generalise the unstructured bandit setting by assuming that each arm $a \in [K]$ corresponds to a feature vector $\mathbf{x}_a \in \mathbb{R}^{d_{\text{lin}}}$ and the rewards of the arms depend on a linear regression parameter $\boldsymbol{\theta} \in \mathbb{R}^{d_{\text{lin}}}$. In our constrained problem, a bandit *environment* is a pair $\boldsymbol{\nu} \triangleq (\boldsymbol{\theta}, B)$, where $\boldsymbol{\theta} \in \mathcal{D}_{\boldsymbol{\theta}} \subseteq \mathbb{R}^{d_{\text{lin}}}$ governs rewards and $B \in \mathcal{D}_B \subseteq \mathbb{R}^{d\times d_{\text{lin}}}$ is the unknown constraint matrix. At each time step, pulling an arm A_t at round t yields:

$$r_t = \langle \mathbf{x}_{A_t}, \boldsymbol{\theta}\rangle + \eta_t, \quad \mathbf{c}_t = B\mathbf{x}_{A_t} + \xi_t\,. \tag{1}$$

We assume that the noises are zero centered and jointly distributed: $\gamma_t \triangleq (\eta_t, \xi_t) \sim \mathcal{N}(\mathbf{0}^{d+1}, \Sigma)$, where the covariance matrix $\Sigma \triangleq \begin{bmatrix} \sigma_r^2 & \Sigma_{rc}^\top \\ \Sigma_{rc} & \Sigma_c \end{bmatrix}$ is positive definite. The goal is to sequentially and adaptively sample from these arms for identifying the SOP:

$$\boldsymbol{p}^*(\boldsymbol{\theta}, B) \triangleq \underset{\boldsymbol{p}\in\triangle_K :\ B\mathbb{X}\boldsymbol{p}\leq 0}{\arg\max}\ \boldsymbol{p}^\top \mathbb{X}^\top \boldsymbol{\theta}\,. \tag{SOP-lin}$$

Here, $\mathbb{X} \triangleq [\mathbf{x}_1, \ldots, \mathbf{x}_K] \in \mathbb{R}^{d_{\text{lin}}\times K}$ denotes the feature matrix of all the arms. Note that this is a special setting of transductive linear bandits where the answer set is the (SOP-lin) (Fiez et al., 2019).

Assumption 2 (Bounded domains)**.** $\mathcal{D}_{\boldsymbol{\theta}}$ *and* $\mathcal{D}_B$ *are bounded:* $\boldsymbol{\theta} \in \mathcal{D}_{\boldsymbol{\theta}}$ *and* $B_j \in \mathcal{B}_j$ *for every row* $j \in [d]$, *where* $\|B_j\|_2 \leq 1$. *constraints at* $\boldsymbol{p}^\star(\boldsymbol{\theta})$.

Goal: Efficiency and Correctness. The goal is to design a pure exploration algorithm that is *efficient*– samples different arms as little as possible (in expectation), and *correct*– stops at time τ_δ to recommend a policy $\hat{\boldsymbol{p}}$ that matches the SOP with probability at least $(1-\delta)$.

Definition 1 (δ-correctness)**.** *A pure exploration algorithm* $\boldsymbol{\pi}$ *is* δ-correct *over a class* $\mathcal{E}$ *of bandit environments if, for all* $\boldsymbol{\nu} \in \mathcal{E}$, $\mathbb{P}_{\boldsymbol{\nu},\boldsymbol{\pi}}[\hat{\boldsymbol{p}} \neq \boldsymbol{p}^\star(\boldsymbol{\theta})] \leq \delta$. *Additionally, it is* δ-feasible *over* $\mathcal{E}$ *if, for all* $\boldsymbol{\nu} \in \mathcal{E}$, $\mathbb{P}_{\boldsymbol{\nu},\boldsymbol{\pi}}[B\hat{\boldsymbol{p}} \not\leq 0] \leq \delta$. *Here,* $X \not\leq 0 \Leftrightarrow \exists j : X_j > 0$.

3. Fundamental Hardness of Identifying Safely Optimal Policies

The fundamental hardness of a fixed-confidence pure exploration is quantified by a lower bound on its expected stopping time, $\mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta]$, to yield a δ-correct recommendation of the true answer, i.e., the (SOP) (or (SOP-lin)). Specifically, for the aforementioned problem of pure exploration under linear constraints, we can apply the generic lower bound of (Garivier & Kaufmann, 2016) to observe that the expected stopping time to exactly identify the (SOP) is $\mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta] \geq T^\star(\boldsymbol{\nu})\,\mathrm{KL}(\delta \parallel 1-\delta)$, where the inverse *characteristic time* is $(T^\star(\boldsymbol{\nu}))^{-1} \triangleq$

$$\sup_{\boldsymbol{\omega}\in\triangle_K}\ \inf_{\boldsymbol{\lambda}\in\Lambda_M(\boldsymbol{\nu})} \sum_{a=1}^{K} \boldsymbol{\omega}_a \mathrm{KL}\left(P_a, C_a \parallel P'_a, C'_a\right)\,, \tag{2}$$

and $\Lambda_M(\boldsymbol{\nu}) \triangleq \{\boldsymbol{\lambda} \triangleq (P'_a, C'_a)_{a\in[K]} \mid \boldsymbol{p}^\star(\boldsymbol{\lambda}) \neq \boldsymbol{p}^\star(\boldsymbol{\nu})\}$ denotes the set of *alternative instances* that consists of the same family of reward-cost distributions as $\boldsymbol{\nu}$ but have a different (SOP). This lower bound further suggests that an agent should sample the arms according to the oracle allocation policy $\boldsymbol{\omega}^*(\boldsymbol{\nu})$ supremising Equation (2) to simulate a small enough stopping time.

3.1. Statistical Instability of Exact Identification

While aiming to solve Equation (2), we observe that one can always find an alternative instance consisting of constraint means B', which are infinitesimally close to the true constraints B while exhibit a different SOP. Such instances can be constructed by changing one of the active constraints in B by an infinitesimal perturbation ϵ. This leads to instability in any exact identification algorithm in this setting. The detailed proof is in Appendix B.

Lemma 1 (Statistical instability of exact identification)**.** *Let there exists at least one tight constraint at the optimal solution* $\boldsymbol{p}^\star(\boldsymbol{\nu})$, *i.e.,* $\exists j \in [d]$ *such that* $\sum_{a=1}^{K} B_{j,a}\boldsymbol{p}_a^\star(\boldsymbol{\nu}) = 0$. *Then given a bandit instance* $\boldsymbol{\nu} \triangleq (\boldsymbol{\mu}_a, B_a)_{a=1}^K$, *and* $\epsilon > 0$, *the expected stopping time* $\mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta] \to \infty$.

This result echoes the sample complexity lower bound for identifying feasibility of a linear program: whether there exists $\mathbf{z} \in \mathbb{R}^K$ such that $B\mathbf{z} \leq 0$ or not (Gangrade et al., 2024b). Gangrade et al. show that complexity of this test depends on Γ^{-2}, where $\Gamma \triangleq \min_{\mathbf{z}} \max_j B_j^\top \mathbf{z}$ denotes the

minimum slack of the constraint matrix. Thus, even to check feasibility of a policy under linear constraints, we need the slack with respect to $\boldsymbol{p}^*(\boldsymbol{\nu})$ to be away from zero. Identifying (SOP) being a harder problem of identifying both feasible and optimal policy inherits similar instability. In a similar spirit, it is common in the regret minimisation literature for bandits with constraints to assume existence of at least a safe arm with non-zero slack (Pacchiano et al., 2020, 2024) for obtaining tractable lower bounds and algorithms. (Lardy et al., 2025) do not explicitly need this assumption as they aim to identify the safely optimal arm that already has a non-zero slack.

3.2. Lower Bound on Sample Complexity of ϵ-optimal Identification

Alarmed by the instability result, we relax the problem and aim to identify an ϵ-optimal but safe policy (Das & Basu, 2026).

Definition 2 (SϵOP Set)**.** *The set of ϵ-optimal safe policies is defined as* $\Pi(\boldsymbol{\nu}, \epsilon) \triangleq \{\boldsymbol{p} \in \triangle_K \mid \boldsymbol{\mu}^\top \boldsymbol{p} + \epsilon \geq \boldsymbol{\mu}^\top \boldsymbol{p}^\star(\boldsymbol{\nu}) \wedge B\boldsymbol{p} \leq 0\}$ *for some* $\epsilon > 0$.

Thus, the modified aim of a pure exploration algorithm is to sample the arms enough to collect reward-cost feedbacks to identify one of the policies in $\Pi(\boldsymbol{\nu}, \epsilon)$ with confidence level $(1-\delta)$. Now, we show that with this relaxed goal the lower bound of Equation (2) can be reduced to a stable optimisation problems involving tractable but nested $\max$ and $\min$ problems.

I. Unstructured Bandits. For bandits with independent arms, we obtain the following lower bound.

Theorem 1 (Lower bound on expected stopping time for unstructured bandits)**.** *For a bandit instance $\boldsymbol{\nu}$ and $\delta \in (0,1)$, any δ-correct pure explorer for identifying an ϵ-optimal safe policy satisfies* $\mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta] \geq T^\star_P(\boldsymbol{\nu}, \epsilon) \log(1/(2.4\delta))$ *, such that* $(T^\star(\boldsymbol{\nu}, \epsilon))^{-1} =$

$$\max_{\boldsymbol{\omega} \in \triangle_K} \max_{\boldsymbol{p} \in \mathcal{E}(\Pi(\boldsymbol{\nu}, \epsilon))} \min \left\{ D^r(\boldsymbol{\nu}, \epsilon, \boldsymbol{\omega}, \boldsymbol{p}), D^c(B, \boldsymbol{\omega}, \boldsymbol{p}) \right\}, \quad (3)$$

where $D^r(\boldsymbol{\nu}, \epsilon, \boldsymbol{\omega}, \boldsymbol{p}) \triangleq$

$$\min_{\boldsymbol{q} \in \triangle_K} \min_{\substack{\boldsymbol{\mu}' : \boldsymbol{\mu}'^\top (\boldsymbol{q} - \boldsymbol{p}) = \epsilon \\ B' : B'\boldsymbol{q} = 0}} \sum_{a=1}^{K} \boldsymbol{\omega}_a \mathrm{KL}\left(\boldsymbol{\mu}_a, B_a \parallel \boldsymbol{\mu}'_a, B'_a\right),$$

$D^c(B, \boldsymbol{\omega}, \boldsymbol{p}) \triangleq \min_{j \in [d]} \min_{B'_j : B'_j \boldsymbol{p} = 0} \sum_{a=1}^{K} \boldsymbol{\omega}_a \mathrm{KL}\left(B_a \parallel B'_a\right)$, *and* $\mathcal{E}(\Pi(\boldsymbol{\nu}, \epsilon))$ *are extreme points.*

Proof Sketch. The lower bound includes three main steps. The detailed proof is in Appendix C.

1. *Decomposition of alt-set.* First, in Lemma 4 we reduce the alt-set as union of $d+1$ convex sets.

2. *Inner optimisation.* By the virtue of closed and compactness of $\Lambda_M(\boldsymbol{\nu})$, in Corollary 3 the inner $\inf - \inf$ optimisation is reduced to a $\min - \min$ problem. Additionally, in Proposition 1-2, we show the answer to these optimisation i.e., $(\boldsymbol{q}^*, \boldsymbol{\mu}'^*, B'^*)$ is unique and absolutely continuous.

3. *Outer optimisation.* The objective with respect to $\boldsymbol{p}$ is quasi-convex. In Proposition 4 we show that the answer set to the outer $\sup_{\boldsymbol{p}}$ is finite, and spanned by all the extremal points in $\Pi(\boldsymbol{\nu}, \epsilon)$. Thus, for a given allocation $\boldsymbol{\omega}$, $\sup_{\boldsymbol{p}}$ reduces to $\max_{\boldsymbol{p}}$. Finally, we optimise the allocation in $\triangle_K$, which is a compact and convex subset in $\mathbb{R}^K$. Additionally, the objective is strictly concave in $\boldsymbol{\omega}$. Thus, the outer most optimisation $\sup_{\boldsymbol{\omega}}$ also reduces to a $\max$ problem.

Consequences. This lower bound implies that identifying a policy in SϵOP encounters two choices: changing the constraint to turn a policy infeasible, and changing both the correlated reward and constraint means to exclude $\boldsymbol{p}$ from the SϵOP. While the former reduces to testing the feasibility for the loosest policy in SϵOP (given by$D^c(B, \boldsymbol{\omega}, \boldsymbol{p})$), the later encodes the impact of correlated reward-cost observations in this pure exploration problem (given by $D^r(\boldsymbol{\nu}, \epsilon, \boldsymbol{\omega}, \boldsymbol{p})$). If the rewards and costs were uncorrelated, we could decompose $D^r(\boldsymbol{\nu}, \epsilon, \boldsymbol{\omega}, \boldsymbol{p})$ into two decoupled terms dependent on reward and cost, respectively. Thus, we could optimise them separately and $D^c(B, \boldsymbol{\omega}, \boldsymbol{p})$ would emerge as a subpart of it consuming feasibility testing as a sub-problem.

Case study: Jointly multivariate Gaussian rewards and costs. To gain further insights on the constraints' impacts on the lower bound, we consider the special case where all arms are $d+1$-variate Gaussian distributions $\mathcal{N}\left([\boldsymbol{\mu}_a \; B_a]^\top, \Sigma\right)$ with known covariance matrix $\begin{bmatrix} \sigma_r^2 & \Sigma_{rc}^\top \\ \Sigma_{rc} & \Sigma_c^{d \times d} \end{bmatrix}$, where $\Sigma_{rc} = \mathbb{R}^d$ is a non-zero vector encoding the reward-cost correlations. Since the covariance Σ is common across instances, the KL divergence between the true and the alternative of arm a reduces to the Mahalanobis distance between their reward-cost means: $\mathrm{KL}\left(\boldsymbol{\mu}_a, B_a \parallel \boldsymbol{\mu}'_a, B'_a\right) = \frac{1}{2} \begin{pmatrix} \boldsymbol{\mu}_a - \boldsymbol{\mu}'_a \\ B_a - B'_a \end{pmatrix}^\top \Sigma^{-1} \begin{pmatrix} \boldsymbol{\mu}_a - \boldsymbol{\mu}'_a \\ B_a - B'_a \end{pmatrix}$. Enabled by this structure, we now derive an analytical bound.

Corollary 1 (Analytical LB for multivariate Gaussians)**.** *The inverse characteristic time for jointly $d+1$-variate Gaussian rewards and costs is* $(T^\star(\boldsymbol{\nu}, \epsilon))^{-1} =$

$$\max_{\boldsymbol{\omega} \in \triangle_K} \max_{\boldsymbol{p} \in \mathcal{E}(\Pi(\boldsymbol{\nu}, \epsilon))} \min \Bigg\{ \min_{\boldsymbol{q} \in \triangle_K} \frac{\beta_{cc}\left(\boldsymbol{\mu}^\top(\boldsymbol{q} - \boldsymbol{p}) - \epsilon\right)^2 - \beta_{cc}^{-1} \|B\boldsymbol{q}\|^2_{\Sigma_c^{-1}\Sigma_{rc}\Sigma_{rc}^\top\Sigma_c^{-1}}}{2(\sigma_r^2 \beta_{rr} \beta_{cc} - (\Sigma_{rc}^\top \Sigma_c^{-1} \Sigma_{rc})\beta_{rc}^2)} + \frac{1}{2\beta_{cc}} \|B\boldsymbol{q}\|^2_{\Sigma_c^{-1}}, \min_{j \in [d]} \frac{(B\boldsymbol{p})_j^2}{2(\Sigma_c)_{j,j} \|\boldsymbol{p}\|^2_{W^{-1}}} \Bigg\}, \quad (4)$$

where $\beta_{rr} \triangleq \sum_{a=1}^{K} \frac{(\boldsymbol{q}_a - \boldsymbol{p}_a)^2}{\boldsymbol{\omega}_a}, \beta_{rc} \triangleq \sum_{a=1}^{K} \frac{\boldsymbol{q}_a(\boldsymbol{q}_a - \boldsymbol{p}_a)}{\boldsymbol{\omega}_a}, \beta_{cc} \triangleq \sum_{a=1}^{K} \frac{\boldsymbol{q}_a^2}{\boldsymbol{\omega}_a}$.

Discussions: *1. Reduction for known constraints.* For known constraints, we can ignore the cost observations and their coupling with rewards. Thus, putting $\beta_{rc} = 0$, we get the inverse characteristics time to be $\max_{\boldsymbol{\omega}} \max_{\boldsymbol{p}} \min_{\boldsymbol{q}} \frac{\left(\boldsymbol{\mu}^\top(\boldsymbol{q}-\boldsymbol{p})-\epsilon\right)^2}{2\sigma_r^2 \beta_{rr}}$. For $\epsilon = 0$, this exactly retrieves the lower bound for identifying the unique (SOP) for bandits with known linear constraints (Carlsson et al., 2024).

2. Reduction for BAI with single unknown constraint. In case of BAI, we only consider the pure strategies, i.e. playing individual arms, as $\boldsymbol{p}$ and $\boldsymbol{q}$. Thus, the allocation policy has mass over only two arms, while we can set $\epsilon = 0$. Under this reduction, we retrieve the lower bound of (Lardy et al., 2025) for Gaussian reward-costs with known covariance. Note that, we do not consider Lardy et al. (2025)'s additional case when no arm is feasible as it is the same in both settings.

3. Testing feasibility vs. safe optimality. While the second term in Equation (4) represents the cost of feasibility testing of $\boldsymbol{p}$, the first term is more nuanced. We observe that it has two parts: $\frac{1}{2\beta_{cc}}\|B\boldsymbol{q}\|^2_{\Sigma_c^{-1}}$ checks feasibility of the alternative policy $\boldsymbol{q}$, and the other term looks into the effective reward gap-square between $\boldsymbol{p}$ and $\boldsymbol{q}$ recalibrated with the conditional variance of reward, given cost. Thus, the trade-off between weighted slack of $\boldsymbol{q}$ plus the effective gap between $\boldsymbol{p}$ and $\boldsymbol{q}$, and the weighted slack of $\boldsymbol{p}$ dictates the hardness of pure exploration task. v

II. Linear Bandits. We further derive the lower bound for linear bandits with linear constraints. First, let us define the conditional variance of the reward given the cost: $S \triangleq \sigma_r^2 - \Sigma_{rc}^\top \Sigma_c^{-1} \Sigma_{rc} > 0$.

Theorem 2 (Lower Bound for Linear Bandits under Multiple Constraints)**.** *Under Assumption 2, the expected stopping time* $\mathbb{E}[\tau_\delta] \geq T^*_{\text{lin}}(\boldsymbol{\nu}, \epsilon)\text{KL}\left(\delta \parallel 1-\delta\right)$, *such that* $(T^\star_{\text{lin}}(\boldsymbol{\nu}, \epsilon))^{-1} =$

$$\max_{\boldsymbol{\omega}\in\triangle_K} \max_{\boldsymbol{p}\in\mathcal{E}(\Pi(\boldsymbol{\nu},\epsilon))} \min_{(\boldsymbol{\theta}',B')\in\Lambda(\boldsymbol{\nu},\boldsymbol{p},\epsilon)} \left(\frac{\|\varepsilon_{\boldsymbol{\theta}} - \Sigma_{rc}^\top \Sigma_c^{-1/2} \varepsilon_B\|^2_{V(\boldsymbol{\omega})}}{S} + \sum_{i,j\in[d]} (\varepsilon_B^i)^\top V(\boldsymbol{\omega}) \varepsilon_B^j \right). \tag{5}$$

We observe that similar to the unstructured bandits, the alternative set $\Lambda(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon) \triangleq \{\tilde{\boldsymbol{\nu}} = (\boldsymbol{\theta}, B') : \boldsymbol{p} \notin \Pi(\tilde{\boldsymbol{\nu}}, \epsilon)\}$ consists of two types of alternatives: $\Lambda_M^c(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon) \triangleq \bigcup_{j\in[d]} \left\{\boldsymbol{\lambda} \in (\boldsymbol{\mu}'_a, B'_a)_{a=1}^K : B'_j \mathbb{X}\boldsymbol{p} > 0\right\}$, and $\Lambda_M^r(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon) \triangleq \bigcup_{\boldsymbol{q}\in\triangle_K} \left\{\boldsymbol{\lambda} \in (\boldsymbol{\mu}'_a, B'_a)_{a=1}^K : \boldsymbol{\theta}'^\top \mathbb{X}(\boldsymbol{q}-\boldsymbol{p}) > \epsilon \wedge B'\mathbb{X}\boldsymbol{q} \leq 0 \wedge \boldsymbol{q} \in \arg\max_{\boldsymbol{v}} \boldsymbol{\theta}'^\top \mathbb{X}\boldsymbol{v}\right\}$. $\Lambda_M^c(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon)$ consists of instances, where $\boldsymbol{p}$ is not feasible. $\Lambda_M^r(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon)$ consists of instances that has a safely optimal policy $\boldsymbol{q}$ that does not belong to the ϵ-close policy set built around $\boldsymbol{p}$. The proof is in Appendix D.1, and hereafter, follows similar steps as that of unstructured bandits.

Explicit lower bound for single constraints: Though the lower bound in Equation (5) is solvable by deploying convex and quasi-convex optimisation algorithms, to gain further insight into the involved geometry, we instantiate it for the single correlated constraint case. In this case, an arm $\mathbf{x}$ yields at t

$$r_t = \langle \mathbf{x}, \boldsymbol{\theta}_r \rangle + \eta_t\,,$$
$$\mathbf{c}_t = \langle \mathbf{x}, \boldsymbol{\theta}_c \rangle + \xi_t,\ (\eta_t, \xi_t) \sim \mathcal{N}(\mathbf{0}^2, \Sigma),$$

where σ_r^2 and σ_c^2 are the noise variances, and their Pearson correlation coefficient is $\rho \in [-1, 1]$.

Corollary 2 (Lower Bound for Linear Bandits under Single Constraint)**.** *Let* $V_{\mathbb{X},\boldsymbol{\omega}} \triangleq \mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}$. *The inverse characteristics time of a linear bandit with a single linear constraint is* $(T^\star(\boldsymbol{\nu}, \epsilon))^{-1} =$

$$\max_{\boldsymbol{\omega}\in\triangle_K} \max_{\boldsymbol{p}\in\mathcal{E}(\Pi(\boldsymbol{\nu},\epsilon))} \min \Bigg\{ \min_{\boldsymbol{q}\in\triangle_K} \frac{c^2(a-\rho b)}{(ac-\rho^2 b^2)^2} \left(\frac{\boldsymbol{\theta}_r^\top \mathbb{X}(\boldsymbol{q}-\boldsymbol{p}) - \epsilon}{\sigma_r}\right)^2 + \frac{a}{(ac-\rho^2 b^2)} \left(\frac{\boldsymbol{\theta}_c^\top \mathbb{X}\boldsymbol{q}}{\sigma_c}\right)^2 - \frac{2(\rho b)}{(ac-\rho^2 b^2)} \left(\frac{\boldsymbol{\theta}_c^\top \mathbb{X}\boldsymbol{q}}{\sigma_c}\right) \left(\frac{\boldsymbol{\theta}_r^\top \mathbb{X}(\boldsymbol{q}-\boldsymbol{p}) - \epsilon}{\sigma_r}\right), \frac{(\boldsymbol{\theta}_c^\top \mathbb{X}\boldsymbol{p})^2}{\sigma_c^2(1-\rho^2)d} \Bigg\}, \tag{6}$$

where $a \triangleq \|\boldsymbol{q}-\boldsymbol{p}\|^2_{V_{\mathbb{X},\boldsymbol{\omega}}}$, $b \triangleq (\boldsymbol{q}-\boldsymbol{p})^\top V_{\mathbb{X},\boldsymbol{\omega}}\boldsymbol{q}$, $c \triangleq \|\boldsymbol{q}\|^2_{V_{\mathbb{X},\boldsymbol{\omega}}}$, $d \triangleq \|\boldsymbol{p}\|^2_{V_{\mathbb{X},\boldsymbol{\omega}}}$.

Consequences: *1. Feasibility vs. safely optimal policy identification, and impact of correlation.* We observe that, like multivariate Gaussians, the lower bound for lineart bandits has two parts with similar appearance of conditional variances of costs w.r.t. rewards and rewards w.r.t. costs. In the second term, we test feasibility of the loosest $\boldsymbol{p}$ in the ϵ-optimal policy set, while the price of correlation appears through $(1-\rho^2)^{-1}$. On the other hand, the price of identifying safely optimal arm depends on the coupling of the renormalised reward gap $\frac{1}{\sigma_r}(\boldsymbol{\theta}_r^\top \mathbb{X}(\boldsymbol{q}-\boldsymbol{p}) - \epsilon)$ and constraint violation gap $\frac{1}{\sigma_c}(\boldsymbol{\theta}_c^\top \mathbb{X}\boldsymbol{q})$. In absence of correlation, the lower bound further reduces to a minimum between the weighted reward-constrained gap between $\boldsymbol{p}$ and $\boldsymbol{q}$, and feasibility gap of $\boldsymbol{q}$. One can directly use this reduced bound if the constraints are known.

2. Retrieving BAI lower bound. If we further assume both $\boldsymbol{p}$ and $\boldsymbol{q}$ to be pure strategies and make the single con-

straint to be the simplex constraint, we retrieve the classic lower bound for BAI in linear bandits with linear constraints (Soare, 2015).

3. Reduction to testing linear programs (LPs). (Gangrade et al., 2024b) provide the lower bound for testing feasibility of an LP $B\mathbf{x} \geq 0$ over a d-dimensional real space as $d\Gamma^{-2}$, where Γ is the minimum slack. The second term i.e., $\max_{\boldsymbol{\omega}\in\triangle_K} \max_{\boldsymbol{p}\in\mathcal{E}(\Pi(\boldsymbol{\nu},\epsilon))} \frac{(\boldsymbol{\theta}_c^\top \mathbb{X}\boldsymbol{p})^2}{\sigma_c^2(1-\rho^2)d}$ encodes the hardness of the same but in the ϵ-optimal policy set. While Gangrade et al. (2024b)'s result motivates a confidence interval based algorithm for further sampling, here, we can use the lower bound to select the sampling policy at every step, and design a Track-and-Stop (Garivier & Kocák, 2021) type algorithm for feasible policy testing with multiple arms. This technique can yield asymptotically optimal algorithm for feasibility testing with empirical constraint estimates, while (Gangrade et al., 2024b) designs a special tail bound controlled estimator to achieve the $d\Gamma^{-2}$ sample complexity.

4. Algorithm Design : PRUNE

Now, we leverage the lower bounds for unstructured and linear bandits to design an algorithmic framework, PRUNE, that achieves asymptotically optimal sample complexities in both the settings. Specifically, PRUNE being a pure exploration algorithm consists of three main components:

1. **Sampling/allocation policy:** While sampling, at every step t, PRUNE plugs in the empirical statistics of rewards and costs in the lower bounds, and solves the lower bounds to obtain a sampling/allocation policy $\boldsymbol{\omega}_t$. Instead of using an linear programming solver, PRUNE uses an extension of the Frank-Wolfe algorithm (Jaggi, 2013; Wang et al., 2021a) to optimise the lower bound computationally efficiently. Then, it applies C-tracking (Kaufmann et al., 2016) with the computed allocation policy to select an arm A_t to be pulled.

In case of linear bandits, we have to additionally estimate the regression parameters $\boldsymbol{\theta}$ and B. We deploy a ridge regressor (Tikhonov & Arsenin, 1977) that uses the history of reward-cost observations across arms to estimate the parameters as

$$\begin{pmatrix}\hat{\boldsymbol{\theta}}_t & \hat{B}_t\end{pmatrix} = \left(\gamma I_{d_{\text{lin}}} + \sum_{s=1}^{t-1} \mathbf{x}_s \mathbf{x}_s^\top\right)^{-1} \left(\sum_{s=1}^{t-1} \mathbf{x}_s \begin{pmatrix} r_s & \mathbf{c}_s \end{pmatrix}^\top\right). \tag{7}$$

Here, $\gamma > 0$ is the regularisation parameter.

2. **Stopping rule:** After sampling and collecting reward-cost observations at each step, PRUNE computes a Generalised Likelihood Ratio (GLR) score and checks whether it crosses a statistically sound threshold. If it does, PRUNE infers that it has collected enough evidence and stops. The threshold depends on confidence level δ. We propose two new stopping rules (Lemma 2) for multivariate Gaussian reward-costs and linear bandits with multiple linear constraints.

3. **Recommendation rule:** Finally, PRUNE uses the samples collected till the stopping time τ_δ and use them to compute the empirical reward-cost means or the regression parameters for linear bandits. Then, we plug-in the estimated parameters in (SOP) and (SOP-lin), and use linear programming solvers to yield a valid recommendation $\hat{\boldsymbol{p}}$.

Now, we elaborate the specific technical hurdles and solutions for designing the allocation policy and the stopping time.

4.1. Computing Allocation Policies with Frank-Wolfe

Though Frank-Wolfe (FW) algorithm was originally proposed for optimising smooth functions, (Wang et al., 2021a) extended it to computing the allocation policies $\{\boldsymbol{\omega}_s\}_{s=1}^{\tau_\delta}$ though the lower bound is not smooth on $\boldsymbol{\omega}$. The non-smoothness arises due to the fact the gradients of the lower bound with respect to $\boldsymbol{\omega}$ blows up at the boundary of the simplex. In this problem, we observe two additional degeneracies:

- both $\nabla_{\boldsymbol{\omega}} D^r(\boldsymbol{\nu}, \epsilon, \boldsymbol{\omega}, \boldsymbol{p})$ and $\nabla_{\boldsymbol{\omega}} D^c(B, \boldsymbol{\omega}, \boldsymbol{p})$ blow up at the boundary of the feasible polytope,
- $F(\boldsymbol{\omega}, \boldsymbol{p}) \triangleq \min\{D^r(\boldsymbol{\nu}, \epsilon, \boldsymbol{\omega}, \boldsymbol{p}), D^c(B, \boldsymbol{\omega}, \boldsymbol{p})\}$ is not smooth whenever $D^r(\boldsymbol{\nu}, \epsilon, \boldsymbol{\omega}, \boldsymbol{p}) = D^c(B, \boldsymbol{\omega}, \boldsymbol{p})$. This creates an additional problem than the existing FW-type algorithms for pure exploration, which are smooth at least if we could plug-in all the inner optima (Wang et al., 2021a; Das et al., 2025). But if any two of them become equal, we still encounter non-smoothness in this unknown constraint setting.

Sub-differential set. Thus, following (Wang et al., 2021a), we carefully design sub-differential sets. First, from Theorem 6 and Lemma 5, we observe that $F(\boldsymbol{\omega}, \boldsymbol{p}) \triangleq \min\{D^r(\boldsymbol{\nu}, \epsilon, \boldsymbol{\omega}, \boldsymbol{p}), D^c(B, \boldsymbol{\omega}, \boldsymbol{p})\} = \min_{j\in\mathcal{J}_{\boldsymbol{p}}} f_j(\boldsymbol{\omega}, \boldsymbol{p})$, where $\mathcal{J}_{\boldsymbol{p}} \triangleq \{0, 1, \ldots, d\}$, $f_0(\boldsymbol{\omega}, \boldsymbol{p}) \triangleq D^r(\boldsymbol{\nu}, \epsilon, \boldsymbol{\omega}, \boldsymbol{p})$, and $f_j(\boldsymbol{\omega}, \boldsymbol{p}) \triangleq \min_{B'_j : B_j \boldsymbol{p} = 0} D^c(B, \boldsymbol{\omega}, \boldsymbol{p})$ for $j \in [d]$. Thus, we define the r-subdifferential subspace for $F(\boldsymbol{\omega}) \triangleq \max_{\boldsymbol{p}} F(\boldsymbol{\omega}, \boldsymbol{p})$ as

$$H_{F(\cdot)}(\boldsymbol{\omega}, r) \triangleq \operatorname{conv}\Big\{\nabla_{\boldsymbol{\omega}} f_j(\boldsymbol{\omega}, \boldsymbol{p}) \mid \boldsymbol{p} \in \mathcal{P}(\boldsymbol{\omega}), j \in \mathcal{J}_{\boldsymbol{p}}, f_j(\boldsymbol{\omega}, \boldsymbol{p}) > F(\boldsymbol{\omega}, \boldsymbol{p}) - r\Big\}. \tag{8}$$

This subdifferential set separately takes care of each of the components of the lower bounds due to rewards and the constraints.

Algorithm 1 PRUNE - **P**ure explo**R**ation **U**nder u**N**known lin**E**ar constraints

Input: Confidence level $\delta \in (0,1)$ and sequence $\{r_t\}_{t\geq 1} = t^{-0.9}/K$
Initialise: For $t \in [K]$, sample each arm once s.t. $\boldsymbol{\omega}_K = (1/K, \cdots, 1/K)$ to warm-up reward mean estimate $\hat{\boldsymbol{\mu}}_K$ and constraint estimate $\hat{B}_K$
while STOPPING RULE (Lemma 2) is **FALSE do**
  if $\sqrt{t/K} \in \mathbb{N}$ or $(\hat{\boldsymbol{\mu}}_t, \hat{B}_t) \notin (\mathcal{D}_{\boldsymbol{\mu}}, \mathcal{D}_B)$ (or $(\hat{\boldsymbol{\theta}}_t, \hat{B}_t) \notin (\mathcal{D}_{\boldsymbol{\theta}}, \mathcal{D}_B)$ for linear bandits) **then**
    Forced Exploration: $\boldsymbol{\omega}_t \leftarrow (1/K, \ldots, 1/K)$
  else
    $\boldsymbol{\lambda}$-player: Calculate confusing instance using closed forms in Appendix D.
    Sub-Differential Set: $H_{F(\cdot)}(\boldsymbol{\omega}_{t-1}, r_t)$ using Equation (8)
    FW-Update: $\boldsymbol{\omega}_{t+1} \leftarrow \mathrm{FW}(\boldsymbol{\omega}_t, r_t)$ (Equation (9))
  end if
  C-TRACKING: (i)*Unstructured bandits.* Play $A_t \in \arg\min_{a\in[K]} N_{a,t} - \sum_{s=1}^{t+1} \boldsymbol{\omega}_s$, and (ii) *Linear bandits.* Play $\mathbf{x}_t \in \arg\min_{\mathbf{x}\in\mathcal{X}} N_{\mathbf{x},t} - \sum_{s=1}^{t+1} \boldsymbol{\omega}_s$ (ties broken arbitrarily)
  Feedback: $(R_t, C_t) \in \mathbb{R}^{d+1}$
  Update parameters: (i) *Unstructured.* Update $\hat{\boldsymbol{\mu}}_{t+1} \leftarrow \hat{\boldsymbol{\mu}}_t$ and $\hat{A}_{t+1} \leftarrow \hat{A}_t$, and *Linear.* Use Ridge regression in Equation (7)
end while
RECOMMENDED POLICY: Recommend $\boldsymbol{p}_t$, where $\boldsymbol{p}_t = \arg\max_{\boldsymbol{p}\in\Delta_K : \hat{A}_t\boldsymbol{p}\leq 0} \hat{\boldsymbol{\mu}}_t^\top \boldsymbol{p}$.

Update rule. We now define the FW update rule for the allocation policy as

$$\mathbf{x}_{t+1} \triangleq \arg\max_{\mathbf{x}\in\triangle_K} \min_{\mathbf{h}\in H_{F(\cdot)}(\boldsymbol{\omega}_{t-1}, r_t)} \langle \mathbf{x} - \boldsymbol{\omega}_t, \mathbf{h}\rangle,$$
$$\boldsymbol{\omega}_{t+1} \triangleq \frac{t}{t+1}\boldsymbol{\omega}_t + \frac{1}{t+1}\mathbf{x}_{t+1} . \tag{9}$$

Convergence of FW updates. For FW to converge, boundedness of the gradient and curvature constant is needed. For both multivariate Gaussian and linear bandits with positive definite covariances, we show that the gradient and curvatures of $F(\boldsymbol{\omega})$ are bounded helping FW to converge to the optimal allocation (ref. Appendix I). We observe that the convergence of FW here depends on $\min\{K, d\}$ instead of only K as in the unconstrained bandits (Wang et al., 2021a).

4.2. Stopping Time: Gaussian and Linear Bandits

Now, we derive the two stopping conditions that tests validity of a candidate policy candidate policy $\hat{\boldsymbol{\pi}}_t = \arg\max_{\hat{B}_t\boldsymbol{\pi}\leq 0} \hat{\boldsymbol{\mu}}^\top \boldsymbol{\pi}$ to be in SϵOP. These proofs extend the mixture of martingale technique of (Kaufmann & Koolen, 2021) and leverage rescaling of the multivariate Gaussian means with their covariance matrix.

Lemma 2 (Stopping criterion and Correctness)**.** *The Chernoff stopping criterion for any δ-correct ϵ-optimal pure exploration algorithm with d linear constraints for a candidate policy $\hat{\boldsymbol{\pi}}_t = \arg\max_{\hat{B}_t\boldsymbol{\pi}\leq 0} \hat{\boldsymbol{\mu}}^\top \boldsymbol{\pi}$, are as follows.*
(a) Multivariate Gaussian bandits:

$$\min\Bigg\{ \min_{\boldsymbol{\lambda}\in\Lambda_M^r(\hat{\boldsymbol{\nu}}_t, \hat{\boldsymbol{\pi}}_t, \epsilon)} \sum_{a=1}^K N_{a,t}\mathrm{KL}\left(\hat{\boldsymbol{\nu}}_{a,t} \parallel \boldsymbol{\lambda}_a\right), \min_{\boldsymbol{\lambda}\in\Lambda_M^c(\hat{\boldsymbol{\nu}}_t, \hat{\boldsymbol{\pi}}_t, \epsilon)} \sum_{a=1}^K N_{a,t}\mathrm{KL}\left(\hat{\boldsymbol{\nu}}_{a,t} \parallel \boldsymbol{\lambda}_a\right)\Bigg\} > \beta(t,\delta)$$

(b) Linear bandits: Given $V_t \triangleq \sum_{\mathbf{x}\in\mathcal{X}} N_{\mathbf{x},t}\mathbf{x}\mathbf{x}^\top$, we get

$$\min_{(\boldsymbol{\theta}', B')} \Bigg\{ \frac{\|\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}' - \Sigma_{rc}^\top\Sigma_c^{-1}(\hat{B}_t - B')\|_{\hat{V}_t}^2}{\sigma_r^2 - \Sigma_{rc}\Sigma_c^{-1}\Sigma_{rc}} + \sum_{i,j=1}^d e_i^\top\Sigma_c^{-1}(\hat{B}_t - B')V_t(\hat{B}_t - B')^\top\Sigma_c^{-1}e_j \Bigg\} > 2\beta(t,\delta),$$

where $\beta(t,\delta) = 2(d+1)\sum_{a=1}^K \log\left(4 + \log\left(N_{a,t}\right)\right) + K\mathcal{T}\left(\frac{\log\left(\frac{1}{\delta}\right)}{K}\right)$ and $\mathcal{T}(x) \approx x + \log x$.

Connection to sequential hypothesis testing and e-values: This stopping rule can be specialised to test feasibility of a linear optimisation problems with constraints over the observation distributions. Additionally, we can use the stopping rule for linear bandit to test optimality of a regressor while a part of it dictates the outcomes and the other part imposes a constraint. We find such examples in constrained regression literature. Finally, we observe that the GLRs used in the stopping times while normalised by the total runtime t yields e-values under constraints. (Clerico, 2024) states existence of an optimal e-class. Our GLR normalised by t provides a tractable and valid construction of it. Recently, (Agrawal & Ramdas, 2025) design power-one tests for hypothesis generated by a finite set of constraints. Our stopping rules can be used to generate two-sided sequential tests with multiple multivariate Gaussian and linear regression-generated hypothesis with linear constraints.

4.3. Asymptotic Optimality of PRUNE

Now, we show that PRUNE is an asymptotically optimal pure exploration algorithm for both Gaussian and linear bandits with multiple linear constraints.

Theorem 3 (Asymptotic Optimality)**.** *Under Assumption 1 and 2, as the error level $\delta \to 0$, the expected sample complexity of PRUNE matches up to constants the lower bound for both linear and unstructured Gaussian bandits, i.e.* $\lim_{\delta\to 0} \frac{\mathbb{E}[\tau_\delta]}{\log(1/\delta)} \leq T^*(\boldsymbol{\nu}, \epsilon)$ (or $T^*_{\mathrm{lin}}(\boldsymbol{\nu}, \epsilon)$).

We also prove correctness of the proposed algorithm and derive a non-asymptotic sample complexity bound (Appendix K), which we omit for brevity.

Connection to (Das & Basu, 2026): Instead of deriving the exact lower bound for identifying a policy in SϵOP, Das & Basu (2026) propose a Lagrangian relaxation of the lower bound that lend itself to efficient optimisation. They leverage this relaxed lower bound to design extensions of Track-and-Stop (Kaufmann et al., 2016) and Gamified Explorer (Degenne et al., 2019) to tackle multiple, unknown linear constraints, namely LaTS and LaGEx. While LaTS does not achieve asymptotic optimality, LaGEx does but its non-asymptotic optimality gets inflated by the shadow price of B, i.e. the ratio of its maximum and minimum eigenvalues. Due to our exact derivation of the lower bound and careful treatment of the optimisation with Frank-Wolfe, we do not observe any such factor appearing in our analysis making PRUNE statistically efficient than its existing counterparts.

Per-Round Computational Complexity. PRUNE incurs $\mathcal{O}(K \max\{K, d\})$ and $\mathcal{O}(d_{\text{lin}}^2 \max\{K, d_{\text{lin}}, d\})$ complexity per-round for unconstrained and linear bandits, respectively (Appendix L). Hence, the sample complexity of the proposed algorithm grows polynomially in problem dimensions, and improve on the conjecture of (Lardy et al., 2025) that resolving the pure exploration problem with d unknown constraints would incur exponential cost in d.

5. Discussions, Limitations, and Future Works

We study the problem of identifying the safely optimal policy with fixed confidence for bandits with correlated reward-cost observations and linear constraints on the cost. We show the statistical instability of exact identification, and derive the lower bounds on expected stopping times for identifying ϵ-optimal and safe policy under multiple constraints, and for both unstructured and linear bandits. We leverage the geometry of the lower bound to design a computationally efficient and asymptotically optimal algorithmic framework PRUNE. This resolves the open problem of handling unknown constraints in pure exploration (Carlsson et al., 2024), while refutes the conjecture of (Lardy et al., 2025) that handling multiple constraints would lead to exponential complexity in d.

Though our lower bounds are valid for any reward-cost distribution and the Frank-Wolfe allocation works efficiently for any distribution ensuring bounded curvature of the lower bound, the stopping rule hinders the proof of asymptotic optimality. This is due to lack of a mixture of martingales method available for multivariate distributions (Kaufmann & Koolen, 2021). In absence of such results, one can use the Lagrangian relaxation-driven stopping rule of (Das & Basu, 2026) along with our FW-based allocation strategy for non-Gaussian distributions. Another important direction is to explore the impacts of unknown nonlinear constraints in pure exploration with bandit feedback.

Acknowledgment

Authors acknowledge ANR JCJC project REPUBLIC (ANR-22-CE23-0003-01), PEPR project FOUNDRY (ANR23-PEIA-0003), and Inria-ISI Kolkata associate team SeRAI.

References

Agrawal, S. and Ramdas, A. On stopping times of power-one sequential tests: Tight lower and upper bounds. *arXiv preprint arXiv:2504.19952*, 2025.

Amani, S., Alizadeh, M., and Thrampoulidis, C. Linear stochastic bandits under safety constraints, 2019.

Ball, I. Bauer's maximum principle for quasiconvex functions. *arXiv preprint arXiv:2305.04893*, 2023.

Berge, C. *Topological spaces: Including a treatment of multi-valued functions, vector spaces and convexity*. Oliver & Boyd, 1877.

Bubeck, S., Munos, R., and Stoltz, G. Pure exploration in multi-armed bandits problems. In *Algorithmic Learning Theory: 20th International Conference, ALT 2009, Porto, Portugal, October 3-5, 2009. Proceedings 20*, pp. 23–37. Springer, 2009.

Bui, V.-H., Das, S., Hussain, A., Hollweg, G. V., and Su, W. A critical review of safe reinforcement learning techniques in smart grid applications. *arXiv preprint arXiv:2409.16256*, 2024.

Carlsson, E., Basu, D., Johansson, F., and Dubhashi, D. Pure exploration in bandits with linear constraints. In *International Conference on Artificial Intelligence and Statistics*, pp. 334–342. PMLR, 2024.

Clerico, E. Optimal e-value testing for properly constrained hypotheses. *arXiv preprint arXiv:2412.21125*, 2024.

Combes, R., Magureanu, S., and Proutiere, A. Minimal exploration in structured stochastic bandits, 2017. URL https://arxiv.org/abs/1711.00400.

Das, U. and Basu, D. Learning to explore with lagrangians for bandits under unknown linear constraints. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2026.

Das, U., Shukla, A., and Basu, D. Frappe: Fast and efficient preference-based pure exploration. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Degenne, R. and Koolen, W. M. Pure exploration with multiple correct answers. In *Neural Information Processing Systems*, 2019. URL https://api.semanticscholar.org/CorpusID:60440353.

Degenne, R., Koolen, W. M., and Ménard, P. Non-asymptotic pure exploration by solving games, 2019.

Even-Dar, E., Mannor, S., and Mansour, Y. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7(39):1079–1105, 2006. URL http://jmlr.org/papers/v7/evendar06a.html.

Feinberg, E. A., Kasyanov, P. O., and Voorneveld, M. Berge's maximum theorem for noncompact image sets. *Journal of Mathematical Analysis and Applications*, 413 (2):1040–1046, 2014.

Fiez, T., Jain, L., Jamieson, K. G., and Ratliff, L. Sequential experimental design for transductive linear bandits. *Advances in neural information processing systems*, 32, 2019.

Gangrade, A., Chen, T., and Saligrama, V. Safe linear bandits over unknown polytopes. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 1755–1795. PMLR, 2024a.

Gangrade, A., Gopalan, A., Saligrama, V., and Scott, C. Testing the feasibility of linear programs with bandit feedback. *arXiv preprint arXiv:2406.15648*, 2024b.

Garivier, A. and Kaufmann, E. Optimal best arm identification with fixed confidence. In Feldman, V., Rakhlin, A., and Shamir, O. (eds.), *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 998–1027, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR. URL https://proceedings.mlr.press/v49/garivier16a.html.

Garivier, A. and Kocák, T. Epsilon Best Arm Identification in Spectral Bandits. In *Thirtieth International Joint Conference on Artificial Intelligence {IJCAI-21}*, pp. 2636–2642, Montreal, Canada, August 2021. International Joint Conferences on Artificial Intelligence Organization. doi: 10.24963/ijcai.2021/363. URL https://hal.science/hal-03369546.

Jaggi, M. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *International conference on machine learning*, pp. 427–435. PMLR, 2013.

Kaufmann, E. and Koolen, W. M. Mixture martingales revisited with applications to sequential tests and confidence intervals. *Journal of Machine Learning Research*, 22(246):1–44, 2021. URL http://jmlr.org/papers/v22/18-798.html.

Kaufmann, E., Cappé, O., and Garivier, A. On the complexity of best arm identification in multi-armed bandit models, 2016.

Kumari, A. and Tanwar, S. A reinforcement-learning-based secure demand response scheme for smart grid system. *IEEE Internet of Things Journal*, 9(3):2180–2191, 2021.

Lardy, T. D., Katsimerou, C., and Koolen, W. M. Constrained best arm identification. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Lattimore, T. and Szepesvári, C. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401.

Li, Y., Du, N., Song, X., Yang, X., Cui, T., Xue, N., Farjudian, A., Ren, J., and Cheah, W. P. Cardinality and bounding constrained portfolio optimization using safe reinforcement learning. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2024.

Pacchiano, A., Ghavamzadeh, M., Bartlett, P., and Jiang, H. Stochastic bandits with linear constraints, 2020.

Pacchiano, A., Ghavamzadeh, M., and Bartlett, P. Contextual bandits with stage-wise constraints. *arXiv preprint arXiv:2401.08016*, 2024.

Soare, M. *Sequential resource allocation in linear stochastic bandits*. PhD thesis, Université Lille 1-Sciences et Technologies, 2015.

Thomas, P., Theocharous, G., and Ghavamzadeh, M. High-confidence off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Tikhonov, A. N. and Arsenin, V. Y. Solutions of ill-posed problems. vh winston & sons, 1977.

Wang, P.-A., Tzeng, R.-C., and Proutiere, A. Fast pure exploration via frank-wolfe. *Advances in Neural Information Processing Systems*, 34:5810–5821, 2021a.

Wang, X. and Liu, L. Risk-sensitive deep reinforcement learning for portfolio optimization. *Journal of Risk and Financial Management*, 18(7):347, 2025.

Wang, Z., Wagenmaker, A., and Jamieson, K. Best arm identification with safety constraints, 2021b.

Appendix

Table of Contents

A Notations **14**

B Proof of Lower Bounds for Exact Identification **15**
- B.1 Lower Bound . . . 15
- B.2 Impossibility of Learning under δ-correctness . . . 16

C Lower Bound for Identifying SϵOP Answer in Unstructured Bandits **19**
- C.1 Reward and Cost Model: Gaussian with Known and fixed Covariance . . . 22

D Lower Bound for Identifying SϵOP in Linear Bandits **27**
- D.1 Lower Bound . . . 27
- D.2 Analytical Form for Single Constraint Case . . . 30

E Structural Reduction of the Lower Bounds **35**

F Gradient and Curvature of Allocation Objective **39**
- F.1 Unstructured Bandits . . . 39
- F.2 Linear Bandits . . . 41

G Stopping Criterion **43**

H Concentration of Probability Bad Events **45**

I Convergence of Frank-Wolfe Optimiser for Allocations **47**
- I.1 Properties of the Allocation Function . . . 47
- I.2 Continuity of the Frank-Wolfe Iterates: Sub-Differentials . . . 48
- I.3 Contraction of the Frank-Wolfe Iterates . . . 49

J Tracking Results **51**

K Sample Complexity Analysis of PRUNE **52**

L Computational Complexity of PRUNE **53**
- L.1 Unstructured Bandits . . . 53
- L.2 Linear Bandits . . . 54
- L.3 Discussions . . . 54

M Auxiliary Definitions and Results **55**

A. Notations

Table 1. Summary of notation used throughout the paper.

Symbol	Description
General Notation	
$[n]$	Set $\{1, 2, \ldots, n\}$
$x, \mathbf{x}, \mathcal{X}$	Scalar, vector, and set respectively
$\|\cdot\|_p$	ℓ_p-norm
Δ_K	Probability simplex over K arms
$\mathrm{KL}(P\|Q)$	Kullback–Leibler divergence between distributions
Bandit Model	
K	Number of arms
A_t	Arm selected at time t
R_t, C_t	Reward and cost observed at time t
$\mathcal{H}_t$	History up to time t
$\boldsymbol{\nu}$	Bandit instance
Reward and Cost Distributions	
P_a, C_a	Reward and cost distributions of arm a
$\boldsymbol{\mu}$	Vector of mean rewards
B	Constraint matrix
d	Number of constraints
Policies and Optimization	
$\boldsymbol{p} \in \triangle_K$	Policy (distribution over arms)
$\boldsymbol{p}^\star(\boldsymbol{\nu})$	Safely optimal policy (SOP)
$\hat{\boldsymbol{p}}$	Recommended policy
$\Pi(\boldsymbol{\nu}, \epsilon)$	Set of ϵ-optimal safe policies
ϵ	Sub-optimality tolerance
Linear Bandit Setting	
$\mathbb{X}$	Feature matrix $[x_1, \ldots, x_K]$
$\mathbf{x}_t$	action chosen at time t
$\boldsymbol{\theta}$	Unknown reward parameter
B	Unknown constraint parameter
r_t	Observed reward at time t
c_t	Observed cost at time t
Σ	Noise covariance matrix
Sampling and Complexity	
$\boldsymbol{\omega} \in \triangle_K$	Sampling allocation
$\boldsymbol{\omega}^\star$	Optimal allocation
τ_δ	Stopping time
δ	Confidence level
$T^\star(\boldsymbol{\nu})$	Characteristic time
$T_P^\star(\boldsymbol{\nu})$	Characteristic time in unstructured bandit setting
$T_{\mathrm{lin}}^\star(\boldsymbol{\nu})$	Characteristic time in Linear bandit setting
$N_{a,t}$	Number of pulls of arm a up to time t
Alternative Instances	
$\boldsymbol{\lambda} \triangleq (P'_a, C'_a)_{a=1}^K$	Alternative bandit instance
$\Lambda(\boldsymbol{\nu})$	Set of alternative instances

B. Proof of Lower Bounds for Exact Identification

In this section, first we present the formal proof of the fundamental hardness of identification of (SOP), i.e., a lower bound on the expected number of interaction. Then we prove the existence of a statistical instability in Lemma 1. It shows that we can create indistinguishable alternative instances by only perturbing a single active constraint.

B.1. Lower Bound

Theorem 4 (Lower bound on expected stopping time under δ correctness)**.** *For any bandit model $\boldsymbol{\nu} = \{P_a, C_a\}_{a=1}^K$ and $\delta \in (0,1)$, any δ-correct pure exploration strategy satisfies,*

$$\mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta] \geq T^\star(\boldsymbol{\nu}) \log(1/(2.4\delta))$$

where

$$T^\star(\boldsymbol{\nu})^{-1} = \sup_{\boldsymbol{\omega} \in \triangle_K} \inf_{\boldsymbol{\lambda} \in \Lambda_M(\boldsymbol{\nu})} \sum_{a=1}^K \boldsymbol{\omega}_a \mathrm{KL}\left(P_a, C_a \parallel P'_a, C'_a\right)$$

Proof. Let $\boldsymbol{\pi}$ be a δ-correct policy. Let $\boldsymbol{\nu}$ be a bandit instance and $\boldsymbol{\lambda} \in \Lambda_M(\boldsymbol{\nu})$, where the alternative instances are all instances

$$\Lambda_M(\boldsymbol{\nu}) \triangleq \{\boldsymbol{\lambda} \triangleq (P'_a, C'_a)_{a \in [K]} \text{ s.t. } \boldsymbol{p}^\star(\lambda) \neq \boldsymbol{p}^\star(\boldsymbol{\nu})\}.$$

The data-processing inequality gives that

$$\begin{aligned} \mathrm{KL}\left(\mathbb{P}_{\boldsymbol{\nu},\boldsymbol{\pi}} \parallel \mathbb{P}_{\boldsymbol{\lambda},\boldsymbol{\pi}}\right) &\geq \mathrm{kl}\left(\mathbb{P}_{\boldsymbol{\nu},\boldsymbol{\pi}}\left(\hat{\boldsymbol{p}} = \boldsymbol{p}^\star(\boldsymbol{\nu})\right), \mathbb{P}_{\boldsymbol{\lambda},\boldsymbol{\pi}}\left(\hat{\boldsymbol{p}} = \boldsymbol{p}^\star(\boldsymbol{\nu})\right)\right) \\ &\geq \mathrm{kl}(1-\delta, \delta). \end{aligned} \tag{10}$$

where the second inequality is because $\boldsymbol{\pi}$ is δ-correct i.e. $\mathbb{P}_{\boldsymbol{\nu},\boldsymbol{\pi}}\left(\hat{\boldsymbol{p}} = \boldsymbol{p}^\star(\boldsymbol{\nu})\right) \geq 1-\delta$ and $\mathbb{P}_{\boldsymbol{\lambda},\boldsymbol{\pi}}\left(\hat{\boldsymbol{p}} = \boldsymbol{p}^\star(\boldsymbol{\nu})\right) \leq \delta$, and the monotonicity of the kl. Since this inequality holds for all $\boldsymbol{\lambda} \in \Lambda_M(\boldsymbol{\nu})$, we get

$$\begin{aligned} \mathrm{kl}(1-\delta, \delta) &\leq \inf_{\boldsymbol{\lambda} \in \Lambda_M(\boldsymbol{\nu})} \sum_{a=1}^K \mathbb{E}[N_a(\tau)] \mathrm{KL}\left(P_a, C_a \parallel P'_a, C'_a\right) \\ &\overset{(a)}{=} \mathbb{E}[\tau] \inf_{\boldsymbol{\lambda} \in \Lambda_M(\boldsymbol{\nu})} \sum_{a=1}^K \frac{\mathbb{E}\left[N_a(\tau)\right]}{\mathbb{E}[\tau]} \mathrm{KL}\left(P_a, C_a \parallel P'_a, C'_a\right) \\ &\overset{(b)}{\leq} \mathbb{E}[\tau] \left(\sup_{\boldsymbol{\omega} \in \Delta_K} \inf_{\boldsymbol{\lambda} \in \Lambda_M(\boldsymbol{\nu})} \sum_{a=1}^K \boldsymbol{\omega}_a \mathrm{KL}\left(P_a, C_a \parallel P'_a, C'_a\right) \right). \end{aligned}$$

(a) is due to the fact that $\mathbb{E}[\tau]$ does not depend on $\boldsymbol{\lambda}$. (b) is obtained by noting that the vector $(\boldsymbol{\omega}_a)_{a \in [K]} \triangleq \left(\frac{\mathbb{E}_{\boldsymbol{\nu},\boldsymbol{\pi}}[N_a(\tau)]}{\mathbb{E}_{\boldsymbol{\nu},\boldsymbol{\pi}}[\tau]}\right)_{a \in [K]}$ belongs to the simplex Δ_K.

By noting that for $\delta \in (0,1), \mathrm{kl}(1-\delta, \delta) \geq \log(1/2.4\delta)$, we get

$$\mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta] \geq T^\star(\boldsymbol{\nu}) \log(1/(2.4\delta))$$

with

$$T^\star(\boldsymbol{\nu})^{-1} \triangleq \sup_{\boldsymbol{\omega} \in \Delta_K} \inf_{\boldsymbol{\lambda} \in \Lambda_M(\boldsymbol{\nu})} \sum_{a=1}^K \boldsymbol{\omega}_a \mathrm{KL}\left(P_a, C_a \parallel P'_a, C'_a\right)$$

□

Theorem 5 (Lower bound on expected stopping time under δ' feasibility)**.** *For any bandit model $\boldsymbol{\nu} = \{P_a, C_a\}_{a=1}^K$ and $\delta' \in (0,1)$, any δ'-feasible pure exploration strategy satisfies,*

$$\mathbb{E}_{\boldsymbol{\nu}}[\tau_{\delta'}] \geq T_C^\star(\boldsymbol{\nu}) \log(1/(2.4\delta'))$$

where

$$T_C^\star(\boldsymbol{\nu})^{-1} \triangleq \sup_{\boldsymbol{\omega} \in \Delta_K} \inf_{\boldsymbol{\lambda} \in \Lambda_C(\boldsymbol{\nu})} \sum_{a \in [K]} \boldsymbol{\omega}_a \mathrm{KL}\left(C_a \parallel C'_a\right).$$

and $\Lambda_C(\boldsymbol{\nu}) \triangleq \{\boldsymbol{\lambda} \triangleq (P_a, C'_a)_{a \in [K]}$ s.t. $\exists \hat{\boldsymbol{p}} \in \Delta_K : \; B\hat{\boldsymbol{p}} \leq 0 \Rightarrow B\hat{\boldsymbol{p}} \not\leq 0\}$.

Proof. Let $\boldsymbol{\pi}$ be a δ'-feasible policy. Let $\boldsymbol{\nu}$ be a bandit instance and $\boldsymbol{\lambda} \in \Lambda_C(\boldsymbol{\nu})$, where the alternative instances are all instances

$$\Lambda_C(\boldsymbol{\nu}) \triangleq \{\boldsymbol{\lambda} \triangleq (P'_a, C'_a)_{a \in [K]} \text{ s.t. } \exists \hat{\boldsymbol{p}} \in \Delta_K : \; B\hat{\boldsymbol{p}} \leq 0 \Rightarrow B'\hat{\boldsymbol{p}} \not\leq 0\}$$

where $B' = (B'_a)_{a \in [K]}$, $B_a \triangleq \mathbb{E}_{X \sim C'_a}[X]$ and for $X \in \mathbb{R}^K$, $X \not\leq 0 \Leftrightarrow \exists j \in [1,K] : X_j > 0$, and $X \leq 0 \Leftrightarrow \forall j \in [1,K] : X_j \leq 0$.

The data-processing inequality gives that

$$\begin{aligned} \mathrm{KL}\left(\mathbb{P}_{\boldsymbol{\nu},\boldsymbol{\pi}} \parallel \mathbb{P}_{\boldsymbol{\lambda},\boldsymbol{\pi}}\right) &\geq \mathrm{kl}\left(\mathbb{P}_{\boldsymbol{\nu},\boldsymbol{\pi}}\left(B\hat{\boldsymbol{p}} \leq 0\right), \mathbb{P}_{\boldsymbol{\lambda},\boldsymbol{\pi}}\left(B\hat{\boldsymbol{p}} \leq 0\right)\right) \\ &\geq \mathrm{kl}(1-\delta', \delta'). \end{aligned} \tag{11}$$

where the second inequality is because $\boldsymbol{\pi}$ is δ'-feasible i.e. $\mathbb{P}_{\boldsymbol{\nu},\boldsymbol{\pi}}\left(B\hat{\boldsymbol{p}} \leq 0\right) \geq 1 - \delta'$ and $\mathbb{P}_{\boldsymbol{\lambda},\boldsymbol{\pi}}\left(B\hat{\boldsymbol{p}} \leq 0\right) \leq \mathbb{P}_{\boldsymbol{\lambda},\boldsymbol{\pi}}\left(B'\hat{\boldsymbol{p}} \not\leq 0\right) \leq \delta'$, and the monotonicity of the kl.

Since this inequality holds for all $\boldsymbol{\lambda} \in \Lambda_C(\boldsymbol{\nu})$, we get

$$\begin{aligned} \mathrm{kl}(1-\delta', \delta') &\leq \inf_{\boldsymbol{\lambda} \in \Lambda_C(\boldsymbol{\nu})} \sum_{a=1}^K \mathbb{E}[N_a(\tau)] \mathrm{KL}\left(P_a, C_a \parallel P'_a, C'_a\right) \\ &\overset{(a)}{=} \mathbb{E}[\tau] \inf_{\boldsymbol{\lambda} \in \Lambda_C(\boldsymbol{\nu})} \sum_{a=1}^K \frac{\mathbb{E}\left[N_a(\tau)\right]}{\mathbb{E}[\tau]} \mathrm{KL}\left(P_a, C_a \parallel P'_a, C'_a\right) \\ &\overset{(b)}{\leq} \mathbb{E}[\tau] \left(\sup_{\boldsymbol{\omega} \in \Delta_K} \inf_{\boldsymbol{\lambda} \in \Lambda_C(\boldsymbol{\nu})} \sum_{a=1}^K \boldsymbol{\omega}_a \mathrm{KL}\left(P_a, C_a \parallel P'_a, C'_a\right) \right). \end{aligned}$$

(a) is due to the fact that $\mathbb{E}[\tau]$ does not depend on $\boldsymbol{\lambda}$. (b) is obtained by noting that the vector $(\boldsymbol{\omega}_a)_{a \in [K]} \triangleq \left(\frac{\mathbb{E}_{\boldsymbol{\nu},\boldsymbol{\pi}}[N_a(\tau)]}{\mathbb{E}_{\boldsymbol{\nu},\boldsymbol{\pi}}[\tau]}\right)_{a \in [K]}$ belongs to the simplex Δ_K.

By noting that for $\delta' \in (0,1), \mathrm{kl}(1-\delta', \delta') \geq \log(1/2.4\delta')$, we get

$$\mathbb{E}_{\boldsymbol{\nu}}[\tau_{\delta'}] \geq T_C^\star(\boldsymbol{\nu}) \log(1/(2.4\delta'))$$

with

$$T_C^\star(\boldsymbol{\nu})^{-1} \triangleq \sup_{\boldsymbol{\omega} \in \Delta_K} \inf_{\boldsymbol{\lambda} \in \Lambda_C(\boldsymbol{\nu})} \sum_{a \in [K]} \boldsymbol{\omega}_a \mathrm{KL}\left(P_a, C_a \parallel P'_a, C'_a\right).$$

□

B.2. Impossibility of Learning under δ-correctness

Lemma 1 (Statistical instability of exact identification)**.** *Let there exists at least one tight constraint at the optimal solution $\boldsymbol{p}^\star(\boldsymbol{\nu})$,i.e., $\exists j \in [d]$ such that $\sum_{a=1}^K B_{j,a} \boldsymbol{p}_a^\star(\boldsymbol{\nu}) = 0$. Then given a bandit instance $\boldsymbol{\nu} \triangleq (\boldsymbol{\mu}_a, B_a)_{a=1}^K$, and $\epsilon > 0$, the expected stopping time $\mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta] \to \infty$.*

Proof. Let $\boldsymbol{\pi}$ be a δ-correct. Let $\boldsymbol{\nu}$ be a bandit instance and $\boldsymbol{\lambda} \in \Lambda_M(\boldsymbol{\nu})$, where the alternative instances are all instances

$$\Lambda_M(\boldsymbol{\nu}) \triangleq \{\boldsymbol{\lambda} \triangleq (P'_a, C'_a)_{a\in[K]} \text{ s.t. } \boldsymbol{p}^\star(\lambda) \neq \boldsymbol{p}^\star(\boldsymbol{\nu})\}.$$

The data-processing inequality gives that

$$\begin{aligned}\mathrm{KL}\left(\mathbb{P}_{\boldsymbol{\nu},\boldsymbol{\pi}} \parallel \mathbb{P}_{\boldsymbol{\lambda},\boldsymbol{\pi}}\right) &\geq \mathrm{kl}\left(\mathbb{P}_{\boldsymbol{\nu},\boldsymbol{\pi}}\left(\hat{\boldsymbol{p}} = \boldsymbol{p}^\star(\boldsymbol{\nu})\right), \mathbb{P}_{\boldsymbol{\lambda},\boldsymbol{\pi}}\left(\hat{\boldsymbol{p}} = \boldsymbol{p}^\star(\boldsymbol{\nu})\right)\right) \\ &\geq \mathrm{kl}(1-\delta, \delta). \end{aligned} \tag{12}$$

where the second inequality is because $\boldsymbol{\pi}$ is δ-correct i.e. $\mathbb{P}_{\boldsymbol{\nu},\boldsymbol{\pi}}\left(\hat{\boldsymbol{p}} = \boldsymbol{p}^\star(\boldsymbol{\nu})\right) \geq 1-\delta$ and $\mathbb{P}_{\boldsymbol{\lambda},\boldsymbol{\pi}}\left(\hat{\boldsymbol{p}} = \boldsymbol{p}^\star(\boldsymbol{\nu})\right) \leq \delta$, and the monotonicity of the kl. Since this inequality holds for all $\boldsymbol{\lambda} \in \Lambda_M(\boldsymbol{\nu})$, we get

$$\begin{aligned}\mathrm{kl}(1-\delta,\delta) &\leq \inf_{\boldsymbol{\lambda}\in\Lambda_M(\boldsymbol{\nu})} \sum_{a=1}^{K} \mathbb{E}[N_a(\tau)]\mathrm{KL}\left(P_a, C_a \parallel P'_a, C'_a\right) \\ \implies \mathrm{kl}(1-\delta,\delta) &\leq \mathbb{E}[\tau]\left(\inf_{\boldsymbol{\lambda}\in\Lambda_M(\boldsymbol{\nu})} \sum_{a=1}^{K} \boldsymbol{\omega}_a \mathrm{KL}\left(P_a, C_a \parallel P'_a, C'_a\right)\right).\end{aligned}$$

Defining the harder case(Only changing constraints). Let us define a specific alternative environment $(P'_a, C'_a)_{a=1}^K \in \Lambda_M(\boldsymbol{\nu})$ such that $\mathbb{E}_{R\sim P'_a}[R] = \boldsymbol{\mu}_a$ and $\mathbb{E}_{C\sim C'_a}[R] = B'_a$. Specifically, we define $B'_{j,a} \triangleq B_{j,a} + \epsilon \forall a \in [K]$ assuming $\exists j \in [d]$ s.t. $\sum_{a=1}^K B_{j,a}\boldsymbol{p}^\star_a(\boldsymbol{\nu}) = 0$, otherwise $B'_{i,a} \triangleq B_{i,a} \forall i \neq j$. Thus, we have

$$\mathrm{kl}(1-\delta,\delta) \leq \mathbb{E}[\tau]\left(\inf_{\boldsymbol{\lambda}\in\Lambda_M(\boldsymbol{\nu})} \sum_{a=1}^{K} \boldsymbol{\omega}_a \mathrm{KL}\left(P_a, C_a \parallel P'_a, C'_a\right)\right) \leq \mathbb{E}[\tau]\left(\sum_{a=1}^{K} \boldsymbol{\omega}_a \mathrm{KL}\left(C_a \parallel C'_a\right)\right).$$

Now, let us characterise this hardness for Multi-variate Gaussian cost vectors. Specifically, we assume $C_a \sim \mathcal{N}_d(B_a, \sigma^2\mathbf{I}_d)$ and $C'_a \sim \mathcal{N}_d(B'_a, \sigma^2\mathbf{I}_d)$. Thus,

$$\mathrm{kl}(1-\delta,\delta) \leq \mathbb{E}[\tau]\left(\sum_{a=1}^{K} \boldsymbol{\omega}_a \mathrm{KL}\left(B_a \parallel B'_a\right)\right) \tag{13}$$

Now, following the definition of B'_a, we have $\mathrm{KL}\left(B_a \parallel B'_a\right) = \frac{\epsilon^2}{2\sigma^2}$. Hence, Equation (13) yields,

$$\begin{aligned}\mathrm{kl}(1-\delta,\delta) &\leq \mathbb{E}[\tau]\left(\sum_{a=1}^{K} \boldsymbol{\omega}_a \frac{\epsilon^2}{2\sigma^2}\right) = \mathbb{E}[\tau]\frac{\epsilon^2}{2\sigma^2} \\ \implies \mathbb{E}[\tau] &\geq \frac{2\sigma^2 \mathrm{kl}(1-\delta,\delta)}{\epsilon^2}\end{aligned}$$

The choice of ϵ can be arbitrarily close to 0, since for any positive value of ϵ, $\boldsymbol{p}^\star(\boldsymbol{\nu})$ is infeasible with respect to the alternative constraint matrix B'. Thus, $\epsilon \to 0$ forces the expected number of interaction $\mathbb{E}[\tau] \to \infty$.

□

We extend Lemma 1 to prove even a stronger result. Under similar conditions as Lemma 1 we prove that any arm $a \in [K]$ has to be pulled infinitely many times in order to identify (SOP).

Lemma 3 (Learning Impossibility under δ-correctness (Stronger result))**.** *Let there exists at least one tight constraint at the optimal solution $\boldsymbol{p}^\star(\boldsymbol{\nu})$ i.e., $\exists$ at least one index $j \in [d]$ s.t. $\sum_{a=1}^K B_{j,a}\boldsymbol{p}^\star_a(\boldsymbol{\nu}) = 0$. Then given a bandit instance $\boldsymbol{\nu} \triangleq (\boldsymbol{\mu}_a, B_a)_{a=1}^K$, and $\epsilon > 0$, expected number of arm a till the stopping time τ i.e., $\mathbb{E}[N_a(\tau)] \to \infty \forall a \in [K]$.*

Proof. We start from the data-processing inequality:

$$\mathrm{kl}(1-\delta,\delta) \leq \inf_{\boldsymbol{\lambda}\in\Lambda_M(\boldsymbol{\nu})} \sum_{a=1}^{K} \mathbb{E}[N_a(\tau)]\mathrm{KL}\left(P_a, C_a \parallel P'_a, C'_a\right).$$

Now, define an alternative environment where $(P'_a, C'_a)_{a=1}^K \in \Lambda_M(\boldsymbol{\nu})$ such that $\mathbb{E}_{R\sim P'_a}[R] = \boldsymbol{\mu}_a$ and $\mathbb{E}_{C\sim C'_a}[R] = B'_a$. Specifically, we define $B'_{j,a} \triangleq B_{j,a} + \epsilon$ for some $a \in [K]$, where $\exists j \in [d]$ s.t. $\sum_{a=1}^K B_{j,a}\boldsymbol{p}^\star_a(\boldsymbol{\nu}) = 0$, otherwise $B'_{i,a} \triangleq B_{i,a} \forall i \neq j$. Thus, we have,

$$\mathrm{kl}(1-\delta,\delta) \leq \sum_{a=1}^K \mathbb{E}[N_a(\tau)]\mathrm{KL}\left(B_a \parallel B'_a\right) = \mathbb{E}[N_a(\tau)]\mathrm{KL}\left(B_a \parallel B'_a\right)$$

Now, let us characterise this hardness for Multi-variate Gaussian cost vectors. Specifically, we assume $C_a \sim \mathcal{N}_d(B_a, \sigma^2\mathbf{I}_d)$ and $C'_a \sim \mathcal{N}_d(B'_a, \sigma^2\mathbf{I}_d)$. Thus,

$$\begin{aligned} \mathrm{kl}(1-\delta,\delta) &\leq \mathbb{E}[N_a(\tau)]\mathrm{KL}\left(B_a \parallel B'_a\right) = \mathbb{E}[N_a(\tau)]\frac{\epsilon^2}{2\sigma^2} \\ \implies \mathbb{E}[N_a(\tau)] &\geq \frac{2\sigma^2\mathrm{kl}(1-\delta,\delta)}{\epsilon^2} \end{aligned}$$

Thus, $\epsilon \to 0$ forces the expected number of interaction $\mathbb{E}[N_a(\tau)] \to \infty$.

□

Then, the natural question arises:

What are the necessary conditions for a δ-correct policy to be learnable?

BNew Goal. Rather than going for exact optimality, we now relax the optimisation goal to find a ϵ-optimal policy, that is a policy that has reward inside an ϵ-ball around the reward of $\boldsymbol{p}^\star(\boldsymbol{\nu})$.

In the next section, we characterise the fundamental hardness of identifying such a ϵ-optimal safe policy.

C. Lower Bound for Identifying SϵOP Answer in Unstructured Bandits

We first remind the reader definition of the set of *correct answers*.

Definition 3 (SϵOP Set)**.** *The set of ϵ-optimal safe policies is defined as* $\Pi(\boldsymbol{\nu}, \epsilon) \triangleq \{\boldsymbol{p} \in \triangle_K \mid \boldsymbol{\mu}^\top \boldsymbol{p} + \epsilon \geq \boldsymbol{\mu}^\top \boldsymbol{p}^\star(\boldsymbol{\nu}) \wedge B\boldsymbol{p} \leq 0\}$ *for some* $\epsilon > 0$.

Goal. To recommend a policy $\hat{\boldsymbol{p}}$ such that $\hat{\boldsymbol{p}}$ *is ϵ-correct with confidence at least* $(1-\delta)$:

$$\mathbb{P}\left(\hat{\boldsymbol{p}} \in \Pi(\boldsymbol{\nu}, \epsilon)\right) \geq 1 - \delta$$

i.e., reward of the recommended policy is at most ϵ away that of the true optimal policy $\boldsymbol{p}^\star(\boldsymbol{\nu})$.

Theorem 1 (Lower bound on expected stopping time for unstructured bandits)**.** *For a bandit instance $\boldsymbol{\nu}$ and $\delta \in (0,1)$, any δ-correct pure explorer for identifying an ϵ-optimal safe policy satisfies*

$$\mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta] \geq T_P^\star\left(\boldsymbol{\nu}, \epsilon\right) \log(1/(2.4\delta))\,,$$

$$\text{s.t. } T_P^\star\left(\boldsymbol{\nu}, \epsilon\right)^{-1} = \max_{\boldsymbol{\omega} \in \triangle_K} \max_{\boldsymbol{p} \in \mathcal{E}(\Pi(\boldsymbol{\nu}, \epsilon))} \min \Bigg\{ \underbrace{\min_{\boldsymbol{q} \in \triangle_K} \min_{\substack{\boldsymbol{\mu}' : \boldsymbol{\mu}'^\top (\boldsymbol{q} - \boldsymbol{p}) = \epsilon \\ B' : B'\boldsymbol{q} = 0}} \sum_{a=1}^K \boldsymbol{\omega}_a \mathrm{KL}\left(\boldsymbol{\mu}_a, B_a \parallel \boldsymbol{\mu}'_a, B'_a\right)}_{\triangleq D^r(\boldsymbol{\nu}, \epsilon, \boldsymbol{\omega}, \boldsymbol{p})}, \underbrace{\min_{j \in [d]} \min_{B'_j : B'_j \boldsymbol{p} = 0} \sum_{a=1}^K \boldsymbol{\omega}_a \mathrm{KL}\left(B_a \parallel B'_a\right)}_{\triangleq D^c(B, \boldsymbol{\omega}, \boldsymbol{p})} \Bigg\}. \tag{14}$$

Proof. We can extend the lower bound proved in Lemma 4 here as a direct consequence of the lower bound argument for multiple correct answers (Degenne & Koolen, 2019), we have-

$$\mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta] \geq T_P^\star\left(\boldsymbol{\nu}, \epsilon\right) \log(1/(2.4\delta))$$

with

$$T_P^\star\left(\boldsymbol{\nu}, \epsilon\right)^{-1} \triangleq \sup_{\boldsymbol{\omega} \in \triangle_K} \sup_{\boldsymbol{p} \in \Pi(\boldsymbol{\nu}, \epsilon)} \inf_{\boldsymbol{\lambda} \in \Lambda_M(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon)} \sum_{a \in [K]} \boldsymbol{\omega}_a \mathrm{KL}\left(P_a, C_a \parallel P'_a, C'_a\right)\,,$$

since we can decompose the alt-set as following-

$$\begin{aligned}
\Lambda_M(\boldsymbol{\nu}, \epsilon) &\triangleq \left\{\boldsymbol{\lambda} \triangleq (\boldsymbol{\mu}'_a, B'_a)_{a \in [K]} : \text{ if } \boldsymbol{p} \in \Pi(\boldsymbol{\nu}, \epsilon) \text{ then } \boldsymbol{p} \notin \Pi(\boldsymbol{\lambda}, \epsilon)\right\} \\
&= \left\{\boldsymbol{\lambda} \triangleq (\boldsymbol{\mu}'_a, B'_a)_{a \in [K]} : \Pi(\boldsymbol{\lambda}, \epsilon) \cap \Pi(\boldsymbol{\nu}, \epsilon) = \emptyset\right\} \\
&= \bigcap_{\boldsymbol{p} \in \Pi(\boldsymbol{\nu}, \epsilon)} \left\{\boldsymbol{\lambda} \triangleq (\boldsymbol{\mu}'_a, B'_a)_{a \in [K]} : \boldsymbol{p} \notin \Pi(\boldsymbol{\lambda}, \epsilon)\right\} \triangleq \bigcap_{\boldsymbol{p} \in \Pi(\boldsymbol{\nu}, \epsilon)} \Lambda_M(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon)\,.
\end{aligned}$$

Lemma 4 (Structure of Alternative set under ϵ-Optimality)**.** *For a fixed "good answer" $\boldsymbol{p} \in \Pi(\boldsymbol{\nu}, \epsilon)$, any $\boldsymbol{\omega} \in \triangle_K$, the Alt-set satisfies–*

$$\Lambda_M(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon) = \left\{\bigcup_{\boldsymbol{q} \in \triangle_K} \Lambda_M^r(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon, \boldsymbol{q})\right\} \bigcup \left\{\bigcup_{j \in [d]} \Lambda_M^c(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon, j)\right\}\,,$$

where $\Lambda_M^r(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon) \triangleq \left\{\boldsymbol{\lambda} \in (\boldsymbol{\mu}'_a, B'_a)_{a=1}^K : \boldsymbol{\mu}'^\top \boldsymbol{p} < \boldsymbol{\mu}'^\top \boldsymbol{p}^\star(\boldsymbol{\lambda}) - \epsilon \bigwedge B'\boldsymbol{p} \leq 0\right\}$ *and,*
$\Lambda_M^c(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon, j) \triangleq \left\{\boldsymbol{\lambda} \in (\boldsymbol{\mu}'_a, B'_a)_{a=1}^K : B'_j \boldsymbol{p} > 0\right\}$.

Proof. We remind the definition of alternative instance set for a fixed correct answer $\boldsymbol{p} \in \Pi(\boldsymbol{\nu}, \epsilon)$,

$$\Lambda_M(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon) = \left\{\boldsymbol{\lambda} \in (P'_a, C'_a)_{a=1}^K : \boldsymbol{p} \notin \Pi(\boldsymbol{\lambda}, \epsilon)\right\}$$

Bcorrect answer $\boldsymbol{p}$ under $\boldsymbol{\nu}$ not being in the "good set" under the alternative environment $\boldsymbol{\lambda}$ either means $\boldsymbol{p}$ is not ϵ-optimal w.r.t $\boldsymbol{p}^\star(\boldsymbol{\lambda}) = \arg\max_{\boldsymbol{q} \in \triangle_K : B'\boldsymbol{q} \leq 0} \boldsymbol{\mu}'^\top \boldsymbol{q}$, or $\boldsymbol{p}$ is not feasible, or both.

$$\begin{aligned}
\Lambda_M(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon) &= \left\{\boldsymbol{\lambda} \in (\boldsymbol{\mu}'_a, B'_a)_{a=1}^K : \boldsymbol{p} \notin \Pi(\boldsymbol{\lambda}, \epsilon)\right\} \\
&= \left\{\boldsymbol{\lambda} \in (\boldsymbol{\mu}'_a, B'_a)_{a=1}^K : \boldsymbol{\mu}'^\top \boldsymbol{p} < \boldsymbol{\mu}'^\top \boldsymbol{p}^\star(\boldsymbol{\lambda}) - \epsilon\right\} \bigcup \left\{\boldsymbol{\lambda} \in (\boldsymbol{\mu}'_a, B'_a)_{a=1}^K : B'\boldsymbol{p} \not\leq 0\right\} \\
&\triangleq \Lambda_M^r(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon) \bigcup \Lambda_M^c(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon)
\end{aligned}$$

We further write $\Lambda_M^c(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon)$ as:

$$\begin{aligned}
\Lambda_M^c(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon) &= \left\{\boldsymbol{\lambda} \in (\boldsymbol{\mu}'_a, B'_a)_{a=1}^K : B'\boldsymbol{p} \not\leq 0\right\} \\
&= \left\{\boldsymbol{\lambda} \in (\boldsymbol{\mu}'_a, B'_a)_{a=1}^K, \exists j \in [d] : B'_j \boldsymbol{p} > 0\right\} \\
&= \bigcup_{j \in [d]} \left\{\boldsymbol{\lambda} \in (\boldsymbol{\mu}'_a, B'_a)_{a=1}^K : B'_j \boldsymbol{p} > 0\right\} \\
&\triangleq \bigcup_{j \in [d]} \Lambda_M^c(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon, j)
\end{aligned}$$

We also write $\Lambda_M^r(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon)$ as:

$$\begin{aligned}
\Lambda_M^r(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon) &= \left\{\boldsymbol{\lambda} \in (\boldsymbol{\mu}'_a, B'_a)_{a=1}^K : \boldsymbol{\mu}'^\top \boldsymbol{p} < \boldsymbol{\mu}'^\top \boldsymbol{p}^\star(\boldsymbol{\lambda}) - \epsilon\right\} \\
&= \left\{\boldsymbol{\lambda} \in (\boldsymbol{\mu}'_a, B'_a)_{a=1}^K, \exists \boldsymbol{q} \in \triangle_K : \boldsymbol{\mu}'^\top \boldsymbol{p} < \boldsymbol{\mu}'^\top \boldsymbol{q} - \epsilon \wedge B'\boldsymbol{q} \leq 0 \wedge \boldsymbol{q} \in \arg\max_{\boldsymbol{v} \in \triangle_K} \boldsymbol{\mu}'^\top \boldsymbol{v}\right\} \\
&= \bigcup_{\boldsymbol{q} \in \triangle_K} \left\{\boldsymbol{\lambda} \in (\boldsymbol{\mu}'_a, B'_a)_{a=1}^K : \boldsymbol{\mu}'^\top \boldsymbol{p} < \boldsymbol{\mu}'^\top \boldsymbol{q} - \epsilon \wedge B'\boldsymbol{q} \leq 0 \wedge \boldsymbol{q} \in \arg\max_{\boldsymbol{v} \in \triangle_K} \boldsymbol{\mu}'^\top \boldsymbol{v}\right\} \\
&\triangleq \bigcup_{\boldsymbol{q} \in \triangle_K} \Lambda_M^r(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon, \boldsymbol{q}),
\end{aligned}$$

Thus, the final expression for $\Lambda_M(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon)$ is given by:

$$\Lambda_M(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon) = \left\{\bigcup_{\boldsymbol{q} \in \triangle_K} \Lambda_M^r(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon, \boldsymbol{q})\right\} \bigcup \left\{\bigcup_{j \in [d]} \Lambda_M^c(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon, j)\right\}$$

□

Thus, using Lemma 4, Equation (3) yields, $T_P^\star(\boldsymbol{\nu})^{-1} =$

$$\sup_{\boldsymbol{\omega} \in \triangle_K} \sup_{\boldsymbol{p} \in \Pi(\boldsymbol{\nu}, \epsilon)} \min\left\{\min_{\boldsymbol{q} \in \mathrm{Vert}(\mathcal{F}')} \inf_{\boldsymbol{\lambda} \in \Lambda_M^r(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon, \boldsymbol{q})} \sum_{a=1}^K \boldsymbol{\omega}_a \mathrm{KL}\left(\boldsymbol{\nu}_a \parallel \boldsymbol{\lambda}_a\right), \min_{j \in [d]} \inf_{\boldsymbol{\lambda} \in \Lambda_M^c(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon, j)} \sum_{a=1}^K \boldsymbol{\omega}_a \mathrm{KL}\left(\boldsymbol{\nu}_a \parallel \boldsymbol{\lambda}_a\right)\right\}$$

$$= \sup_{\boldsymbol{\omega}\in\triangle_K} \sup_{\substack{\boldsymbol{p}:\boldsymbol{\mu}^\top(\boldsymbol{p}^\star-\boldsymbol{p})\le\epsilon \\ \wedge B\boldsymbol{p}\le 0}} \min\Bigg\{ \underbrace{\inf_{\boldsymbol{q}\in\triangle_K} \inf_{\substack{\boldsymbol{\mu}':\boldsymbol{\mu}'^\top(\boldsymbol{q}-\boldsymbol{p})>\epsilon \\ B':B'\boldsymbol{q}\le 0}} \sum_{a=1}^{K} \boldsymbol{\omega}_a \mathrm{KL}\left(\boldsymbol{\mu}_a, B_a \parallel \boldsymbol{\mu}'_a, B'_a\right)}_{\triangleq D^r(\boldsymbol{\nu},\epsilon,\boldsymbol{\omega},\boldsymbol{p})}, \min_{j\in[d]} \underbrace{\inf_{B'_j:B'_j\boldsymbol{p}>0} \sum_{a=1}^{K} \boldsymbol{\omega}_a \mathrm{KL}\left(B_a \parallel B'_a\right)}_{\triangleq D^c(B,\boldsymbol{\omega},\boldsymbol{p})} \Bigg\}. \tag{15}$$

Structural reduction of $D^r(\boldsymbol{\nu},\epsilon,\boldsymbol{\omega},\boldsymbol{p})$. By the virtue of closed and compactness of $\Lambda^r_M(\boldsymbol{\nu},\boldsymbol{p},\epsilon)$, in Corollary 3 we show that for a given allocation $\boldsymbol{\omega}\in\triangle_K$, and correct answer $\boldsymbol{p}\in\Pi(\boldsymbol{\nu},\epsilon)$, the inner $\inf-\inf$ optimisation can be reduced to a $\min-\min$ problem. In Proposition 1-2, we also show that the answer to these optimisation i.e., $(\boldsymbol{q}^*,\boldsymbol{\mu}'^*,B'^*)$ is unique.

Structural reduction of $D^c(B,\boldsymbol{\omega},\boldsymbol{p})$. By the same intuition for $\Lambda^c_M(\boldsymbol{\nu},\boldsymbol{p},\epsilon,j)$ for any $j\in[d]$, in Corollary 3 we show that for a given allocation $\boldsymbol{\omega}\in\triangle_K$, and correct answer $\boldsymbol{p}\in\Pi(\boldsymbol{\nu},\epsilon)$, the inner $\inf$ optimisation can be reduced to a $\min$ problem.

Outer optimisation. Using the fact that the objective with respect to $\boldsymbol{p}$ is quasi-convex, in Proposition 4 we show that the answer set to the outer $\sup_{\boldsymbol{p}}$ is finite. Further, using Bauer's max theorem (Theorem 12) this answer set is spanned by all the extremal points in $\Pi(\boldsymbol{\nu},\epsilon)$. Thus, for a given allocation $\boldsymbol{\omega}$, the optimisation $\sup_{\boldsymbol{p}}$ becomes a $\max_{\boldsymbol{p}}$ problem. Finally, we optimise the allocation in the simplex, which is a compact and convex set. Also, the objective $F(\boldsymbol{\omega})$ is strictly concave in $\boldsymbol{\omega}$. Thus, the outer most optimisation $\sup_{\boldsymbol{\omega}}$ also reduces to a $\max$ problem.

Hence, the characteristic time inverse takes the form of a $\max-\max-\min\{\min-\min,\min-\min\}$ nested optimisation problem as below:

$$T^\star_P(\boldsymbol{\nu},\epsilon)^{-1} = \max_{\boldsymbol{\omega}\in\triangle_K} \max_{\boldsymbol{p}\in\mathcal{E}(\Pi(\boldsymbol{\nu},\epsilon))} \min\Bigg\{ \underbrace{\min_{\boldsymbol{q}\in\triangle_K} \min_{\substack{\boldsymbol{\mu}':\boldsymbol{\mu}'^\top(\boldsymbol{q}-\boldsymbol{p})=\epsilon \\ B':B'\boldsymbol{q}=0}} \sum_{a=1}^{K} \boldsymbol{\omega}_a \mathrm{KL}\left(\boldsymbol{\mu}_a, B_a \parallel \boldsymbol{\mu}'_a, B'_a\right)}_{\triangleq D^r(\boldsymbol{\nu},\epsilon,\boldsymbol{\omega},\boldsymbol{p})}, \underbrace{\min_{j\in[d]} \min_{B'_j:B'_j\boldsymbol{p}=0} \sum_{a=1}^{K} \boldsymbol{\omega}_a \mathrm{KL}\left(B_a \parallel B'_a\right)}_{\triangleq D^c(B,\boldsymbol{\omega},\boldsymbol{p})} \Bigg\},$$

hence proved. □

C.1. Reward and Cost Model: Gaussian with Known and fixed Covariance

We assume for arm $a \in [K]$, the reward distribution $P_a \equiv \mathcal{N}(\boldsymbol{\mu}_a, \sigma_r^2)$ and cost distribution $C_a \equiv \mathcal{N}_d(B_a, \Sigma_c)$, where σ_r^2 and $\Sigma^{d\times d}$ is known and fixed across all $a \in [K]$.

Case I: Uncorrelated Reward and Cost We assume for any arm $a \in [K]$, the joint covariance between P_a and C_a is given by $\begin{bmatrix} \sigma_r^2 & \mathbf{0}^{1\times d} \\ \mathbf{0}^{d\times 1} & \Sigma_c^{d\times d} \end{bmatrix}$, i.e., reward and cost signals are uncorrelated. Thus, from the KL-divergence takes the form:

$$\mathrm{KL}\left(P_a, C_a \parallel P_a', C_a'\right) = \frac{(\boldsymbol{\mu}_a - \boldsymbol{\mu}_a')^2}{2\sigma_r^2} + \frac{1}{2}(B_a - B_a')^\top \Sigma_c^{-1}(B_a - B_a')$$

Now we focus on the Term $D^r(\boldsymbol{\nu}, \epsilon, \boldsymbol{\omega}, \boldsymbol{p})$ in Equation (15), i.e.,

$$\begin{aligned} D^r(\boldsymbol{\nu}, \epsilon, \boldsymbol{\omega}, \boldsymbol{p}) &= \inf_{\boldsymbol{q}\in\triangle_K} \inf_{\substack{\boldsymbol{\mu}':\boldsymbol{\mu}'^\top(\boldsymbol{q}-\boldsymbol{p})>\epsilon \\ B':B'\boldsymbol{q}\leq 0}} \sum_{a=1}^K \boldsymbol{\omega}_a \mathrm{KL}\left(\boldsymbol{\mu}_a, B_a \parallel \boldsymbol{\mu}_a', B_a'\right) \\ &= \inf_{\boldsymbol{q}\in\triangle_K} \inf_{\substack{\boldsymbol{\mu}':\boldsymbol{\mu}'^\top(\boldsymbol{q}-\boldsymbol{p})>\epsilon \\ B':B'\boldsymbol{q}\leq 0}} \sum_{a=1}^K \boldsymbol{\omega}_a \left(\frac{(\boldsymbol{\mu}_a - \boldsymbol{\mu}_a')^2}{2\sigma_r^2} + \frac{1}{2}(B_a - B_a')^\top \Sigma_c^{-1}(B_a - B_a') \right) \end{aligned} \tag{16}$$

We formulate the Lagrangian dual with respect to the constraints as:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\mu}', B'; \eta, \gamma) \triangleq \sum_{a=1}^K \boldsymbol{\omega}_a \frac{(\boldsymbol{\mu}_a - \boldsymbol{\mu}_a')^2}{2\sigma_r^2} + \frac{1}{2}\boldsymbol{\omega}_a (B_a - B_a')^\top \Sigma_c^{-1}(B_a - B_a') \\ + \sum_{a=1}^K \eta(\boldsymbol{\mu}_a'^\top(\boldsymbol{q}-\boldsymbol{p})_a - \epsilon) - \sum_{a=1}^K \gamma^\top B_a' \boldsymbol{q}_a \end{aligned}$$

We differentiate the Lagrangian dual w.r.t. $\boldsymbol{\mu}_a'$ to get:

$$\begin{aligned} &\nabla_{\boldsymbol{\mu}_a'} \mathcal{L}(\boldsymbol{\mu}', B'; \eta, \gamma) = 0 \\ \implies &\boldsymbol{\omega}_a \frac{(\boldsymbol{\mu}_a' - \boldsymbol{\mu}_a)}{\sigma_r^2} + \eta \boldsymbol{v}_a = 0 \\ \implies &\boldsymbol{\mu}_a' = \boldsymbol{\mu}_a - \frac{\eta\sigma_r^2 \boldsymbol{v}_a}{\boldsymbol{\omega}_a}, \end{aligned}$$

where $\boldsymbol{v}_a = (\boldsymbol{q} - \boldsymbol{p})_a$. Thus, putting the value of $\boldsymbol{\mu}_a'$ in the Lagrangian dual, we get the optimal value of the Lagrangian multiplier η as $\eta^\star = \frac{\sum_{a=1}^K \boldsymbol{\mu}_a \boldsymbol{v}_a - \epsilon}{\sigma_r^2 \sum_{a=1}^K \frac{\boldsymbol{v}_a^2}{\boldsymbol{\omega}_a}}$. Consequently, we get the exact form of the confusing instance $\boldsymbol{\mu}_a' = \boldsymbol{\mu}_a - \frac{\eta^\star \sigma_r^2 \boldsymbol{v}_a}{\boldsymbol{\omega}_a} = \boldsymbol{\mu}_a - \left(\frac{\sum_{a=1}^K \boldsymbol{\mu}_a \boldsymbol{v}_a - \epsilon}{\sum_{a=1}^K \frac{\boldsymbol{v}_a^2}{\boldsymbol{\omega}_a}} \right) \frac{\boldsymbol{v}_a}{\boldsymbol{\omega}_a}$.

Now, we again differentiate $\mathcal{L}(\boldsymbol{\mu}', B'; \eta, \gamma)$ w.r.t. B_a' to get:

$$\begin{aligned} &\nabla_{B_a'} \mathcal{L}(\boldsymbol{\omega}, \boldsymbol{p}, \boldsymbol{q}, \boldsymbol{\mu}', B') = 0 \\ \implies &\boldsymbol{\omega}_a \Sigma_c^{-1}(B_a' - B_a) - \gamma \boldsymbol{q}_a = 0 \\ \implies &B_a' = B_a + \frac{\boldsymbol{q}_a}{\boldsymbol{\omega}_a} \Sigma_c \gamma \end{aligned}$$

Putting the value of the confusing instance B_a in the Lagrangian dual, we get the optimal value of the Lagrangian multiplier γ as: $\gamma^\star = \frac{(B\boldsymbol{q})_+}{\sum_{a=1}^K \frac{\boldsymbol{q}_a^2}{\boldsymbol{\omega}_a}} \Sigma_c^{-1}$, where $(\mathbf{x})_+$ consists all the positive components of $\mathbf{x}$, 0 elsewhere. Thus, we get the exact form of the confusing instance as $B'_a = B_a + \frac{(B\boldsymbol{q})_+}{\sum_{a=1}^K \frac{\boldsymbol{q}_a^2}{\boldsymbol{\omega}_a}}$.

Hence, we get the final expression for the term $D^r(\boldsymbol{\nu}, \epsilon, \boldsymbol{\omega}, \boldsymbol{p})$ as:

$$D^r(\boldsymbol{\nu}, \epsilon, \boldsymbol{\omega}, \boldsymbol{p}) = \inf_{\boldsymbol{q} \in \triangle_K} \left(\frac{\left(\boldsymbol{\mu}^\top(\boldsymbol{q} - \boldsymbol{p}) - \epsilon\right)^2}{2\sigma_r^2 \|\boldsymbol{q} - \boldsymbol{p}\|_{W^{-1}}^2} + \frac{\|(B\boldsymbol{q})_+\|_{\Sigma_c^{-1}}^2}{2\|\boldsymbol{q}\|_{W^{-1}}^2} \right),$$

where $W\mathbb{R}^{K \times K}$ is a diagonal matrix with (a, a)-th entry as $\frac{1}{\boldsymbol{\omega}_a}$.

Bsimilar analysis for $D^c(\boldsymbol{\nu}, \boldsymbol{\omega}, \boldsymbol{p}) = \min_{j \in [d]} \inf_{A_j \boldsymbol{p} > 0} \sum_{a=1}^K \frac{\boldsymbol{\omega}_a}{2} (B_a - B'_a)^\top \Sigma_c^{-1} (B_a - B'_a)$ gives the final expression:

$$D^c(\boldsymbol{\nu}, \boldsymbol{\omega}, \boldsymbol{p}) = \min_{j \in [d]} \frac{(B\boldsymbol{p})_j^2}{2(\Sigma_c)_{j,j} \|\boldsymbol{p}\|_{W^{-1}}^2}$$

Thus, the final expression for the characteristic time inverse in Equation (15) as:

$$T_P^{\star -1}(\boldsymbol{\nu}) = \sup_{\boldsymbol{\omega} \in \triangle_K} \sup_{\boldsymbol{p} \in \Pi(\boldsymbol{\nu}, \epsilon)} \min \Bigg\{ \inf_{\boldsymbol{q} \in \triangle_K} \left(\frac{\left(\boldsymbol{\mu}^\top(\boldsymbol{q} - \boldsymbol{p}) - \epsilon\right)^2}{2\sigma_r^2 \|\boldsymbol{q} - \boldsymbol{p}\|_{W^{-1}}^2} + \frac{\|(B\boldsymbol{q})_+\|_{\Sigma_c^{-1}}^2}{2\|\boldsymbol{q}\|_{W^{-1}}^2} \right), \min_{j \in [d]} \frac{(B\boldsymbol{p})_j^2}{2(\Sigma_c)_{j,j} \|\boldsymbol{p}\|_{W^{-1}}^2} \Bigg\}$$

Case II: Correlated Reward and Cost

Corollary 1 (Analytical LB for multivariate Gaussians)**.**

$$T_P^{\star -1}(\boldsymbol{\nu}) = \max_{\boldsymbol{\omega} \in \triangle_K} \max_{\boldsymbol{p} \in \mathcal{E}(\Pi(\boldsymbol{\nu}, \epsilon))} \min \Bigg\{ \min_{\boldsymbol{q} \in \triangle_K} \beta_{cc} \frac{\left(\boldsymbol{\mu}^\top(\boldsymbol{q} - \boldsymbol{p}) - \epsilon\right)^2}{2\left(\sigma_r^2 \beta_{rr} \beta_{cc} - \beta_{rc}^2 \sigma_{rc}^2\right)} + \frac{1}{\beta_{cc}} \left(\frac{1}{2} \|A\boldsymbol{q}\|_{\Sigma_c^{-1}}^2 - \frac{\|A\boldsymbol{q}\|_{\Sigma_c^{-1} \Sigma_{rc} \Sigma_{rc}^\top \Sigma_c^{-1}}^2}{2\left(\sigma_r^2 \beta_{rr} \beta_{cc} - \beta_{rc}^2 \sigma_{rc}^2\right)} \right), \min_{j \in [d]} \frac{(A\boldsymbol{p})_j^2}{2(\Sigma_c)_{j,j} \|\boldsymbol{p}\|_{W^{-1}}^2} \Bigg\},$$

where $\beta_{rr} = \sum_{a=1}^K \frac{(\boldsymbol{q}_a - \boldsymbol{p}_a)^2}{\boldsymbol{\omega}_a}, \beta_{rc} = \sum_{a=1}^K \frac{\boldsymbol{q}_a(\boldsymbol{q}_a - \boldsymbol{p}_a)}{\boldsymbol{\omega}_a}, \beta_{cc} = \sum_{a=1}^K \frac{\boldsymbol{q}_a^2}{\boldsymbol{\omega}_a}$

Proof. We assume for any arm $a \in [K]$, the joint covariance between P_a and C_a is given by $\begin{bmatrix} \sigma_r^2 & \Sigma_{rc}^\top \\ \Sigma_{rc} & \Sigma_c^{d \times d} \end{bmatrix}$, where $\Sigma_{rc} = \mathbb{R}^d$ is a non-zero vector. Under Case II, the joint distribution of (P_a, C_a) is

$$\begin{pmatrix} P_a \\ C_a \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_a \\ B_a \end{pmatrix}, \Sigma \right), \qquad \Sigma = \begin{bmatrix} \sigma_r^2 & \Sigma_{rc}^\top \\ \Sigma_{rc} & \Sigma_c \end{bmatrix},$$

with $\Sigma_c \in \mathbb{R}^{d \times d}$ and $\Sigma_{rc} \in \mathbb{R}^d$ both known and fixed. Since the covariance Σ is common across instances, the KL divergence between the original and alternative arm a reduces to a Mahalanobis term:

$$\mathrm{KL}(\boldsymbol{\mu}_a, B_a \| \boldsymbol{\mu}'_a, B'_a) = \frac{1}{2} \begin{pmatrix} \delta\boldsymbol{\mu}_a \\ \delta B_a \end{pmatrix}^\top \Sigma^{-1} \begin{pmatrix} \delta\boldsymbol{\mu}_a \\ \delta B_a \end{pmatrix},$$

where $\delta\boldsymbol{\mu}_a \triangleq \boldsymbol{\mu}_a - \boldsymbol{\mu}'_a \in \mathbb{R}$ and $\delta B_a \triangleq B_a - B'_a \in \mathbb{R}^d$. Let $S \triangleq \mathrm{Var}(P_a \mid C_a) = \sigma_r^2 - \Sigma_{rc}^\top \Sigma_c^{-1} \Sigma_{rc} > 0$ be the conditional variance of the reward given the cost, and we define the *conditional mean shift*

$$\widetilde{\delta}_a \;:=\; \delta\boldsymbol{\mu}_a - \Sigma_{rc}^\top \Sigma_c^{-1} \delta B_a.$$

By the block-matrix inversion formula, the KL divergence decomposes as

$$\mathrm{KL}(\boldsymbol{\mu}_a, B_a \| \boldsymbol{\mu}'_a, B'_a) = \frac{\widetilde{\delta}_a^2}{2S} + \frac{1}{2}\, \delta B_a^\top \Sigma_c^{-1}\, \delta B_a. \tag{17}$$

For fixed $\boldsymbol{\omega}, \boldsymbol{q} \in \Delta_K$ and $\boldsymbol{p} \in \Pi(\boldsymbol{\nu}, \epsilon)$, The inner problem of $D^r(\boldsymbol{\nu}, \epsilon, \boldsymbol{\omega}, \boldsymbol{p})$ is

$$\inf_{\boldsymbol{\mu}', B'} \sum_{a=1}^{K} \boldsymbol{\omega}_a \,\mathrm{KL}(\boldsymbol{\mu}_a, B_a \| \boldsymbol{\mu}'_a, B'_a), \quad \text{subject to} \quad \begin{cases} \displaystyle\sum_{a=1}^{K} (\boldsymbol{q}_a - \boldsymbol{p}_a)\, \boldsymbol{\mu}'_a \;\geq\; \epsilon, \\ \displaystyle\sum_{a=1}^{K} \boldsymbol{q}_a\, B'_a \;\leq\; 0, \end{cases} \tag{18}$$

Thus, the Lagrangian relaxation of the inner optimization problem

$$\mathcal{L}(\boldsymbol{\mu}', B'; \eta, \gamma) = \sum_{a=1}^{K} \boldsymbol{\omega}_a \,\mathrm{KL}_a - \eta \left(\sum_a (\boldsymbol{q}_a - \boldsymbol{p}_a) \boldsymbol{\mu}'_a - \epsilon \right) + \gamma^\top \sum_a \boldsymbol{q}_a\, B'_a. \tag{19}$$

Using Equation (17) and $\partial \widetilde{\delta}_a / \partial \boldsymbol{\mu}'_a = -1$:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}'_a} = 0 \implies -\frac{\boldsymbol{\omega}_a\, \widetilde{\delta}_a}{S} - \eta(\boldsymbol{q}_a - \boldsymbol{p}_a) = 0.$$

Solving gives the conditional shift:

$$\widetilde{\delta}_a^* = -\frac{\eta S\, (\boldsymbol{q}_a - \boldsymbol{p}_a)}{\boldsymbol{\omega}_a}. \tag{20}$$

Using $\frac{\partial}{\partial B'_a}[\frac{\widetilde{\delta}_a^2}{2S}] = \frac{\widetilde{\delta}_a}{S} \Sigma_c^{-1} \Sigma_{rc}$ and $\frac{\partial}{\partial B'_a}[\delta B_a^\top \Sigma_c^{-1} \delta B_a / 2] = -\Sigma_c^{-1} \delta B_a$:

$$\frac{\partial \mathcal{L}}{\partial B'_a} = 0 \implies \boldsymbol{\omega}_a \Sigma_c^{-1} \left(\frac{\widetilde{\delta}_a}{S}\, \Sigma_{rc} - \delta B_a \right) + \boldsymbol{q}_a\, \gamma = 0.$$

Substituting Equation (20) and multiplying through by Σ_c yields the cost shift:

$$\delta B_a^* = -\frac{\eta(\boldsymbol{q}_a - \boldsymbol{p}_a)}{\boldsymbol{\omega}_a}\, \Sigma_{rc} + \frac{\boldsymbol{q}_a}{\boldsymbol{\omega}_a}\, \Sigma_c\, \gamma. \tag{21}$$

Recovering $\delta\boldsymbol{\mu}_a = \widetilde{\delta}_a + \Sigma_{rc}^\top \Sigma_c^{-1} \delta B_a$ and substituting Equation (20)–(21), and $S + \Sigma_{rc}^\top \Sigma_c^{-1} \Sigma_{rc} = \sigma_r^2$ gives:

$$\delta\boldsymbol{\mu}_a^* = -\frac{\eta \sigma_r^2 (\boldsymbol{q}_a - \boldsymbol{p}_a)}{\boldsymbol{\omega}_a} + \frac{\boldsymbol{q}_a\, \Sigma_{rc}^\top \gamma}{\boldsymbol{\omega}_a},$$

so the optimal alternative instances are given by:

$$\boxed{\boldsymbol{\mu}_a'^* = \boldsymbol{\mu}_a + \frac{\eta\, \sigma_r^2 (\boldsymbol{q}_a - \boldsymbol{p}_a)}{\boldsymbol{\omega}_a} - \frac{\boldsymbol{q}_a\, \Sigma_{rc}^\top \gamma}{\boldsymbol{\omega}_a}, \qquad B_a'^* = B_a + \frac{\eta(\boldsymbol{q}_a - \boldsymbol{p}_a)}{\boldsymbol{\omega}_a}\, \Sigma_{rc} - \frac{\boldsymbol{q}_a}{\boldsymbol{\omega}_a}\, \Sigma_c\, \gamma.} \tag{22}$$

We introduce some notations:

$$\beta_{rr} = \sum_{a=1}^{K} \frac{(\boldsymbol{q}_a - \boldsymbol{p}_a)^2}{\boldsymbol{\omega}_a}, \qquad \beta_{rc} = \sum_{a=1}^{K} \frac{\boldsymbol{q}_a(\boldsymbol{q}_a - \boldsymbol{p}_a)}{\boldsymbol{\omega}_a}, \qquad \beta_{cc} = \sum_{a=1}^{K} \frac{\boldsymbol{q}_a^2}{\boldsymbol{\omega}_a},$$

and the means $\bar{\boldsymbol{\mu}}_{\boldsymbol{q}-\boldsymbol{p}} = \boldsymbol{\mu}^\top(\boldsymbol{q}-\boldsymbol{p})$, $\bar{A}_{\boldsymbol{q}} = \sum_{a=1}^K \boldsymbol{q}_a B_a \in \mathbb{R}^d$.

At optimality $\sum_{a=1}^K (\boldsymbol{q}_a - \boldsymbol{p}_a)\boldsymbol{\mu}'^*_a = \epsilon$, thus substituting Equation (22):

$$\bar{\boldsymbol{\mu}}_{\boldsymbol{q}-\boldsymbol{p}} + \eta\,\sigma_r^2\,\beta_{rr} - \beta_{rc}\,\Sigma_{rc}^\top\gamma = \epsilon. \tag{23}$$

For the cost constraints, $\gamma \geq \mathbf{0}$ with $\gamma_j(B'^*\boldsymbol{q})_j = 0$. Computing $B'^*\boldsymbol{q} = \bar{A}_{\boldsymbol{q}} + \eta\,\beta_{rc}\,\Sigma_{rc} - \beta_{cc}\,\Sigma_c\,\gamma$, the *fully active* case ($\gamma > \mathbf{0}$, all constraints tight) sets $B'^*\boldsymbol{q} = \mathbf{0}$:

$$\gamma^* = \frac{1}{\beta_{cc}}\,\Sigma_c^{-1}\left(\bar{A}_{\boldsymbol{q}} + \eta\,\beta_{rc}\,\Sigma_{rc}\right). \tag{24}$$

Substituting Equation (24) into (23), and defining $\sigma_{rc}^2 = \Sigma_{rc}^\top\Sigma_c^{-1}\Sigma_{rc}$, yields the closed-form expression for the Lagrangian multiplier

$$\boxed{\eta^* = \frac{\beta_{cc}(\epsilon - \bar{\boldsymbol{\mu}}_{\boldsymbol{q}-\boldsymbol{p}}) + \beta_{rc}\,\Sigma_{rc}^\top\Sigma_c^{-1}\bar{A}_{\boldsymbol{q}}}{\sigma_r^2\,\beta_{rr}\beta_{cc} - \beta_{rc}^2\,\sigma_{rc}^2}} \tag{25}$$

The denominator equals $\sigma_r^2\beta_{rr} - \beta_{rc}^2\sigma_{rc}^2/\beta_{cc} > 0$ by the Cauchy-Schwarz inequality and the positive definiteness of Σ. Substituting (20)–(21) back into (17) and using definition of S:

$$\boldsymbol{\omega}_a\,\mathrm{KL}_a^* = \frac{1}{2\boldsymbol{\omega}_a}\Big[\eta^2\sigma_r^2(\boldsymbol{q}_a - \boldsymbol{p}_a)^2 - 2\eta(\boldsymbol{q}_a - \boldsymbol{p}_a)\boldsymbol{q}_a\,\Sigma_{rc}^\top\gamma + \boldsymbol{q}_a^2\,\gamma^\top\Sigma_c\,\gamma\Big].$$

Summing over arms:

$$\inf_{\boldsymbol{\mu}',B'}\sum_{a=1}^K \boldsymbol{\omega}_a\,\mathrm{KL}_a = \frac{(\eta^*)^2\sigma_r^2}{2}\,\beta_{rr} - \eta^*\,\Sigma_{rc}^\top\gamma^*\cdot\beta_{rc} + \frac{(\gamma^*)^\top\Sigma_c\,\gamma^*}{2}\,\beta_{cc}. \tag{26}$$

Here,

$$\frac{\beta_{rc}}{\beta_{cc}} = \frac{1}{\beta_{cc}}\sum_{a=1}^K \frac{\boldsymbol{q}_a(\boldsymbol{q}_a - \boldsymbol{p}_a)}{\boldsymbol{\omega}_a} = \frac{1}{\beta_{cc}}\sum_{a=1}^K \frac{\boldsymbol{q}_a^2 + (\boldsymbol{q}_a - \boldsymbol{p}_a)^2 - \boldsymbol{p}_a^2}{2\boldsymbol{\omega}_a} = \frac{1}{2}\left(1 + \frac{\beta_{rr}}{\beta_{cc}} - \frac{1}{\beta_{cc}}\sum_{a=1}^K \frac{\boldsymbol{p}_a^2}{\boldsymbol{\omega}_a}\right)$$

From Equation (24) and simplifying using ρ:

$$\begin{aligned}
&\inf_{\mu',B'}\sum_{a=1}^K \boldsymbol{\omega}_a\,\mathrm{KL}_a \\
=&\frac{\left(\epsilon - \boldsymbol{\mu}^\top(\boldsymbol{q}-\boldsymbol{p}) + \frac{\beta_{rc}}{\beta_{cc}}\Sigma_{rc}^\top\Sigma_c^{-1}B\boldsymbol{q}\right)^2}{2\left(\sigma_r^2\|\boldsymbol{q}-\boldsymbol{p}\|_{W^{-1}}^2 - \frac{\beta_{rc}^2}{\beta_{cc}}\Sigma_{rc}^\top\Sigma_c^{-1}\Sigma_{rc}\right)} - \frac{\eta^*\|\boldsymbol{q}^\top(\boldsymbol{q}-\boldsymbol{p})\|_{W^{-1}}}{\|\boldsymbol{q}\|_{W^{-1}}^2}\Sigma_{rc}^\top\Sigma_c^{-1}B\boldsymbol{q} + \frac{\|B\boldsymbol{q}\|_{\Sigma_c^{-1}}^2}{2\|\boldsymbol{q}\|_{W^{-1}}^2} \\
=&\frac{\left(\epsilon - \boldsymbol{\mu}^\top(\boldsymbol{q}-\boldsymbol{p})\right)^2 - \left(\frac{\beta_{rc}}{\beta_{cc}}\Sigma_{rc}^\top\Sigma_c^{-1}B\boldsymbol{q}\right)^2}{2\left(\sigma_r^2\|\boldsymbol{q}-\boldsymbol{p}\|_{W^{-1}}^2 - \frac{\beta_{rc}^2}{\beta_{cc}}\Sigma_{rc}^\top\Sigma_c^{-1}\Sigma_{rc}\right)} + \frac{\|B\boldsymbol{q}\|_{\Sigma_c^{-1}}^2}{2\|\boldsymbol{q}\|_{W^{-1}}^2} \\
=&\frac{\left(\boldsymbol{\mu}^\top(\boldsymbol{q}-\boldsymbol{p}) - \epsilon\right)^2}{2\left(\sigma_r^2\|\boldsymbol{q}-\boldsymbol{p}\|_{W^{-1}}^2 - \frac{\beta_{rc}^2}{\beta_{cc}}\Sigma_{rc}^\top\Sigma_c^{-1}\Sigma_{rc}\right)} \\
&\quad + \frac{\sigma_r^2\frac{\beta_{rr}}{\beta_{cc}}\|B\boldsymbol{q}\|_{\Sigma_c^{-1}}^2 - \frac{\beta_{rc}^2}{\beta_{cc}^2}\left(\Sigma_{rc}^\top\Sigma_c^{-1}\Sigma_{rc}\|B\boldsymbol{q}\|_{\Sigma_c^{-1}}^2 + \left(\Sigma_{rc}^\top\Sigma_c^{-1}B\boldsymbol{q}\right)^2\right)}{2\left(\sigma_r^2\|\boldsymbol{q}-\boldsymbol{p}\|_{W^{-1}}^2 - \frac{\beta_{rc}^2}{\beta_{cc}}\Sigma_{rc}^\top\Sigma_c^{-1}\Sigma_{rc}\right)} \\
=&\frac{\left(\boldsymbol{\mu}^\top(\boldsymbol{q}-\boldsymbol{p}) - \epsilon\right)^2}{2\left(\sigma_r^2\|\boldsymbol{q}-\boldsymbol{p}\|_{W^{-1}}^2 - \frac{\beta_{rc}^2}{\beta_{cc}}\sigma_{rc}^2\right)} + \frac{\sigma_r^2\frac{\beta_{rr}}{\beta_{cc}}\|B\boldsymbol{q}\|_{\Sigma_c^{-1}}^2 - \frac{\beta_{rc}^2}{\beta_{cc}^2}\left(\sigma_{rc}^2\|B\boldsymbol{q}\|_{\Sigma_c^{-1}}^2 + \left(\Sigma_{rc}^\top\Sigma_c^{-1}B\boldsymbol{q}\right)^2\right)}{2\left(\sigma_r^2\|\boldsymbol{q}-\boldsymbol{p}\|_{W^{-1}}^2 - \frac{\beta_{rc}^2}{\beta_{cc}}\sigma_{rc}^2\right)}
\end{aligned}$$

$$
\begin{aligned}
=&\beta_{cc}\frac{\left(\boldsymbol{\mu}^\top(\boldsymbol{q}-\boldsymbol{p})-\epsilon\right)^2}{2\left(\sigma_r^2\beta_{rr}\beta_{cc}-\beta_{rc}^2\sigma_{rc}^2\right)}\\
&\qquad+\frac{1}{\beta_{cc}}\frac{\sigma_r^2\beta_{rr}\beta_{cc}\|B\boldsymbol{q}\|^2_{\Sigma_c^{-1}}-\beta_{rc}^2\left(\sigma_{rc}^2\|B\boldsymbol{q}\|^2_{\Sigma_c^{-1}}+\|B\boldsymbol{q}\|^2_{\Sigma_c^{-1}\Sigma_{rc}\Sigma_{rc}^\top\Sigma_c^{-1}}\right)}{2\left(\sigma_r^2\beta_{rr}\beta_{cc}-\beta_{rc}^2\sigma_{rc}^2\right)}\\
=&\beta_{cc}\frac{\left(\boldsymbol{\mu}^\top(\boldsymbol{q}-\boldsymbol{p})-\epsilon\right)^2}{2\left(\sigma_r^2\beta_{rr}\beta_{cc}-\beta_{rc}^2\sigma_{rc}^2\right)}+\frac{1}{\beta_{cc}}\left(\frac{1}{2}\|B\boldsymbol{q}\|^2_{\Sigma_c^{-1}}-\frac{\|B\boldsymbol{q}\|^2_{\Sigma_c^{-1}\Sigma_{rc}\Sigma_{rc}^\top\Sigma_c^{-1}}}{2\left(\sigma_r^2\beta_{rr}\beta_{cc}-\beta_{rc}^2\sigma_{rc}^2\right)}\right)
\end{aligned}
$$

Let $e_j \in \mathbb{R}^d$ denote the j-th standard basis vector. For fixed j, the constraint $\sum_a \boldsymbol{p}_a A'_{a,j} = 0$ becomes

$$\sum_{a=1}^{K}\boldsymbol{p}_a\, e_j^\top \delta B_a \;=\; \bar{B}_{\boldsymbol{p},j}.$$

The inner problem for fixed j is therefore

$$\inf_{\delta B}\sum_{a=1}^{K}\frac{\boldsymbol{\omega}_a}{2}\,\delta B_a^\top\Sigma_c^{-1}\delta B_a,\quad \text{subject to}\quad \sum_{a=1}^{K}\boldsymbol{p}_a\, e_j^\top\delta B_a = \bar{B}_{\boldsymbol{p},j}. \tag{27}$$

This is a convex quadratic program with a single linear equality constraint.

The Lagrangian relaxation of inner optimization problem of $D^c(\boldsymbol{\nu},\boldsymbol{\omega},\boldsymbol{p})$ with multiplier $\eta\in\mathbb{R}$:

$$\mathcal{L}(\delta B;\,\eta) = \sum_{a=1}^{K}\frac{\boldsymbol{\omega}_a}{2}\,\delta B_a^\top\Sigma_c^{-1}\delta B_a - \eta\left(\sum_{a=1}^{K}\boldsymbol{p}_a\, e_j^\top\delta B_a - \bar{B}_{\boldsymbol{p},j}\right). \tag{28}$$

Taking the gradient with respect to δB_a and setting it to zero:

$$\nabla_{\delta B_a}\mathcal{L} = \boldsymbol{\omega}_a\,\Sigma_c^{-1}\delta B_a - \eta\,\boldsymbol{p}_a\, e_j = 0.$$

Solving gives the optimal cost shift:

$$\delta B_a^* = \frac{\eta^*\,\boldsymbol{p}_a}{\boldsymbol{\omega}_a}\,\Sigma_c\, e_j, \tag{29}$$

so that $B_a'^* = B_a - \frac{\eta^*\boldsymbol{p}_a}{\boldsymbol{\omega}_a}\,\Sigma_c e_j$. Geometrically, every arm is shifted in the same direction in cost space (along $\Sigma_c e_j$, the j-th column of Σ_c), proportionally to $\boldsymbol{p}_a/\boldsymbol{\omega}_a$.

Substitute Equation (29) into the constraint:

$$\sum_{a=1}^{K}\boldsymbol{p}_a\, e_j^\top\frac{\eta^*\boldsymbol{p}_a}{\boldsymbol{\omega}_a}\,\Sigma_c e_j = \bar{B}_{\boldsymbol{p},j}.$$

Letting $\Sigma_{c,jj} = e_j^\top\Sigma_c\, e_j > 0$ be the (j,j) entry of Σ_c : we obtain

$$\boxed{\eta^* = \frac{\bar{B}_{\boldsymbol{p},j}}{(\Sigma_c)_{j,j}\|\boldsymbol{p}\|^2_{W^{-1}}}.} \tag{30}$$

Since $\bar{B}_{\boldsymbol{p},j}\le 0$ and $(\Sigma_c)_{j,j}>0$, $\gamma>0$, we have $\eta^*\le 0$, confirming that the perturbation $\delta B_a^* = (\eta^*\boldsymbol{p}_a/\boldsymbol{\omega}_a)\Sigma_c e_j$ pulls $B'_{a,j}$ upward (toward violation).

Substituting Equation (29) into the objective and using $e_j^\top\Sigma_c^\top\Sigma_c^{-1}\Sigma_c\, e_j = e_j^\top\Sigma_c\, e_j = (\Sigma_c)_{j,j}$:

$$\sum_{a=1}^{K}\frac{\boldsymbol{\omega}_a}{2}\,(\delta B_a^*)^\top\Sigma_c^{-1}\delta B_a^* = \sum_{a=1}^{K}\frac{\boldsymbol{\omega}_a}{2}\cdot\frac{(\eta^*)^2\boldsymbol{p}_a^2}{\boldsymbol{\omega}_a^2}\, e_j^\top\Sigma_c\, e_j$$

$$= \frac{(\eta^*)^2 \, (\Sigma_c)_{j,j}}{2} \sum_{a=1}^{K} \frac{\boldsymbol{p}_a^2}{\boldsymbol{\omega}_a}$$
$$= \frac{(\eta^*)^2 \, (\Sigma_c)_{j,j} \|\boldsymbol{p}\|_{W^{-1}}^2}{2}.$$

Substituting Equation (30), we get:

$$\boxed{\inf_{\delta B} \sum_{a=1}^{K} \frac{\boldsymbol{\omega}_a}{2} \, \delta B_a^\top \Sigma_c^{-1} \delta B_a = \frac{\bar{B}_{\boldsymbol{p},j}^2}{2 \, (\Sigma_c)_{j,j} \|\boldsymbol{p}\|_{W^{-1}}^2} = \frac{(B\boldsymbol{p})_j^2}{2(\Sigma_c)_{j,j} \|\boldsymbol{p}\|_{W^{-1}}^2}.} \tag{31}$$

□

D. Lower Bound for Identifying SϵOP in Linear Bandits

D.1. Lower Bound

Theorem 2 (Lower Bound for Linear Bandits under Multiple Constraints)**.** *Under Assumption 2, the expected stopping time* $\frac{\mathbb{E}[\tau_\delta]}{2\log\left(\frac{1}{2\delta}\right)} \geq (T^*_{\text{lin}}(\boldsymbol{\nu},\epsilon))^{-1}$*, such that*

$$T^*_{\text{lin}}(\boldsymbol{\nu},\epsilon) = \max_{\boldsymbol{\omega}\in\triangle_K} \max_{\boldsymbol{p}\in\mathcal{E}(\Pi(\boldsymbol{\nu},\epsilon))} \min_{(\boldsymbol{\theta}',B')\in\Lambda(\boldsymbol{\nu},\boldsymbol{p},\epsilon)} \frac{\|\varepsilon_{\boldsymbol{\theta}} - \Sigma_{rc}^\top \Sigma_c^{-1/2}\varepsilon_B\|^2_{V(\boldsymbol{\omega})}}{S} + \sum_{i=1}^d \|\varepsilon_B^i\|^2_{V(\boldsymbol{\omega})} + \sum_{\substack{i,j\in[d]\\ i\neq j}} (\varepsilon_B^i)^\top V(\boldsymbol{\omega})\varepsilon_B^j\,.$$

The alternative set $\Lambda(\boldsymbol{\nu},\boldsymbol{p},\epsilon) = \{\tilde{\boldsymbol{\nu}} = (\boldsymbol{\theta}, B') : \boldsymbol{p} \notin \Pi(\tilde{\boldsymbol{\nu}},\epsilon)\} = \Lambda^c_M(\boldsymbol{\nu},\boldsymbol{p},\epsilon) \cup \Lambda^r_M(\boldsymbol{\nu},\boldsymbol{p},\epsilon)$*, i.e., all environments under which* $\boldsymbol{p}$ *is no longer a valid* ϵ*-correct answer, where* $\Lambda^c_M(\boldsymbol{\nu},\boldsymbol{p},\epsilon) = \bigcup_{j\in[d]} \left\{\boldsymbol{\lambda} \in (\boldsymbol{\mu}'_a, B'_a)_{a=1}^K : B'_j\mathbb{X}\boldsymbol{p} > 0\right\}$ *and* $\Lambda^r_M(\boldsymbol{\nu},\boldsymbol{p},\epsilon) = \bigcup_{\boldsymbol{q}\in\triangle_K} \left\{\boldsymbol{\lambda} \in (\boldsymbol{\mu}'_a, B'_a)_{a=1}^K : \boldsymbol{\theta}'^\top\mathbb{X}(\boldsymbol{q}-\boldsymbol{p}) > \epsilon \wedge B'\mathbb{X}\boldsymbol{q} \leq 0 \wedge \boldsymbol{q} \in \arg\max_{\boldsymbol{v}} \boldsymbol{\theta}'^\top\mathbb{X}\boldsymbol{v}\right\}$.

Proof. The arm set $\mathcal{X} \subset \mathbb{R}^{d_{\text{lin}}}$ is finite with $|\mathcal{X}| = K$ arms. The environment is a pair $\boldsymbol{\nu} = (\boldsymbol{\theta}, B)$ where $\boldsymbol{\theta} \in \mathcal{D}_{\boldsymbol{\theta}} \subseteq \mathbb{R}^{d_{\text{lin}}}$ governs rewards and $B \in \mathcal{D}_B \subseteq \mathbb{R}^{d\times d_{\text{lin}}}$ is the unknown constraint matrix. Pulling arm $\mathbf{x} \in \mathcal{X}$ at round t yields:

$$r_t = \langle \mathbf{x}, \theta\rangle + \eta_t, \quad \eta_t \sim \mathcal{N}(0,\sigma_r^2), \tag{32}$$

$$\mathbf{c}_t = B\mathbf{x} + \xi_t, \quad \xi_t \sim \mathcal{N}_d(\mathbf{0},\Sigma_c), \tag{33}$$

where σ_r^2 and $\Sigma_c \succ 0$ are known. The optimal mixture solves

$$\boldsymbol{p}^*(\boldsymbol{\theta}, B) \triangleq \operatorname*{arg\,max}_{\boldsymbol{p}\in\triangle_K:\, B\boldsymbol{p}_{\mathcal{X}}\leq 0} \boldsymbol{\mu}(\boldsymbol{p},\boldsymbol{\theta}) = \operatorname*{arg\,max}_{\boldsymbol{p}\in\triangle_K:\, B\mathbb{X}\boldsymbol{p}\leq 0} \boldsymbol{p}^\top\mathbb{X}^\top\boldsymbol{\theta},$$

We assume the joint distribution of $\gamma_t \triangleq (\eta_t, \xi_t) \sim \mathcal{N}(\mathbf{0}^{d+1}, \Sigma)$, where the covariance matrix $\Sigma = \begin{bmatrix} \sigma_r^2 & \Sigma_{rc}^\top \\ \Sigma_{rc} & \Sigma_c \end{bmatrix}$ is known, and $\Sigma_{rc} \in \mathbb{R}^d$. Bpolicy $\boldsymbol{\pi}$ is δ*-correct* over a class $\mathcal{E}$ if $\mathbb{P}_{\boldsymbol{\nu}}[\hat{\boldsymbol{p}} \neq \boldsymbol{p}^*(\boldsymbol{\theta}, B)] \leq \delta$ for all $\boldsymbol{\nu} \in \mathcal{E}$.

Information matrix. After pulling arms $\mathbf{x}_1,\ldots,\mathbf{x}_t \in \mathcal{X}$, the *design matrix* is

$$V_t = \sum_{s=1}^t \mathbf{x}_s\mathbf{x}_s^\top \;\in\; \mathbb{R}^{d_{\text{lin}}\times d_{\text{lin}}}. \tag{34}$$

For a fixed allocation $\boldsymbol{\omega} \in \triangle_K$ (i.e., arm $\mathbf{x}$ pulled $\boldsymbol{\omega}(\mathbf{x})t$ times), the design matrix satisfies

$$V_t = t\sum_{\mathbf{x}\in\mathbb{X}} \boldsymbol{\omega}(\mathbf{x})\,\mathbf{x}\mathbf{x}^\top \triangleq t\,V(\boldsymbol{\omega}),$$

where $G(\boldsymbol{\omega}) = \sum_{\mathbf{x}} \boldsymbol{\omega}(\mathbf{x})\mathbf{x}\mathbf{x}^\top$ is the *information matrix* of the design $\boldsymbol{\omega}$.

Case I: Uncorrelated reward and cost – $\Sigma_{rc} = \mathbf{0}^d$**. Step 1: Getting the expected LRT.** The Log-likelihood ratio for the observation till time step $t \in \mathbb{N}$:

$$\begin{aligned}
&\mathcal{L}_t(r_1,\mathbf{c}_1,\ldots,r_t,\mathbf{c}_t)\\
&= \log\left(\prod_{s=1}^t \frac{\mathbb{P}_{\boldsymbol{\nu}}(r_s,\mathbf{c}_s \mid \mathbf{x}_s)}{\mathbb{P}_{\tilde{\boldsymbol{\nu}}}(r_s,\mathbf{c}_s \mid \mathbf{x}_s)}\right)\\
&= \sum_{s=1}^t \log\left(\frac{\mathbb{P}_{\boldsymbol{\nu}}(r_s,\mathbf{c}_s \mid \mathbf{x}_s)}{\mathbb{P}_{\tilde{\boldsymbol{\nu}}}(r_s,\mathbf{c}_s \mid \mathbf{x}_s)}\right)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{s=1}^{t} \log\left(\frac{\mathbb{P}(\mathbf{x}_s^\top \boldsymbol{\theta} + \eta_s, B\mathbf{x} + \xi_s \mid \mathbf{x}_s)}{\mathbb{P}(\mathbf{x}_s^\top \boldsymbol{\theta}' + \eta_s', B'\mathbf{x} + \xi_s' \mid \mathbf{x}_s)}\right) \\
&= \sum_{s=1}^{t} \log\left(\frac{\mathbb{P}(\eta_s, \xi_s)}{\mathbb{P}(\eta_s', \xi_s')}\right) \\
&= \sum_{s=1}^{t} \log\left(\exp\left(\frac{1}{2}\|\gamma_s'\|^2_{\Sigma^{-1}} - \frac{1}{2}\|\gamma_s\|^2_{\Sigma^{-1}}\right)\right) \\
&= \sum_{s=1}^{t} \frac{1}{2}\left(\|\gamma_s'\|^2_{\Sigma^{-1}} - \|\gamma_s\|^2_{\Sigma^{-1}}\right) \\
&= \sum_{s=1}^{t} \Bigg(\frac{(r_s - \mathbf{x}_s^\top \boldsymbol{\theta}')^2 - (r_s - \mathbf{x}_s^\top \boldsymbol{\theta})^2}{2\sigma_r^2} \\
&\qquad + \frac{(\mathbf{c}_s - B'\mathbf{x}_s)^\top \Sigma_c^{-1}(\mathbf{c}_s - B'\mathbf{x}_s) - (\mathbf{c}_s - B\mathbf{x}_s)^\top \Sigma_c^{-1}(\mathbf{c}_s - B\mathbf{x}_s)}{2}\Bigg) \\
&= \sum_{s=1}^{t} \Bigg(\frac{\mathbf{x}_s^\top(\boldsymbol{\theta} - \boldsymbol{\theta}')(-2r_s + 2\mathbf{x}_s^\top \boldsymbol{\theta} + \mathbf{x}_s^\top(\boldsymbol{\theta} - \boldsymbol{\theta}'))}{2\sigma_r^2} \\
&\qquad + \frac{\mathbf{x}_s^\top (B' - B)^\top \Sigma_c^{-1}(B' - B)\mathbf{x}_s - 2(\mathbf{c}_s - B\mathbf{x}_s)^\top \Sigma_c^{-1}(B' - B)\mathbf{x}_s}{2}\Bigg) \\
&= \sum_{s=1}^{t} \Bigg(\frac{\mathbf{x}_s^\top(\boldsymbol{\theta} - \boldsymbol{\theta}')(-2\eta_s + \mathbf{x}_s^\top(\boldsymbol{\theta} - \boldsymbol{\theta}'))}{2\sigma_r^2} \\
&\qquad + \frac{\mathbf{x}_s^\top (B' - B)^\top \Sigma_c^{-1}(B' - B)\mathbf{x}_s - 2\gamma_s^\top \Sigma_c^{-1}(B' - B)\mathbf{x}_s}{2}\Bigg) \\
&= \sum_{s=1}^{t} \left(\frac{\mathbf{x}_s^\top \varepsilon_{\boldsymbol{\theta}}(-2\eta_s + \mathbf{x}_s^\top \varepsilon_{\boldsymbol{\theta}})}{2\sigma_r^2} + \frac{\mathbf{x}_s^\top \varepsilon_B^\top \varepsilon_B \mathbf{x}_s - 2\gamma_s^\top \Sigma_c^{-1/2} \varepsilon_B \mathbf{x}_s}{2}\right),
\end{aligned}
$$

where $\varepsilon_{\boldsymbol{\theta}} \triangleq (\boldsymbol{\theta} - \boldsymbol{\theta}')$, and $\varepsilon_B \triangleq \Sigma_c^{-1/2}(B' - B) = [(\varepsilon_B^1)^\top, \ldots, (\varepsilon_B^d)^\top]$.Taking expectation on the LRT, we get:

$$
\begin{aligned}
\mathbb{E}[\mathcal{L}_t(r_1, \mathbf{c}_1, \ldots, r_t, \mathbf{c}_t)] &= \mathbb{E}\left[\sum_{s=1}^{t}\left(\frac{(\mathbf{x}_s^\top \varepsilon_{\boldsymbol{\theta}})^2}{2\sigma_r^2} + \frac{\mathbf{x}_s^\top \varepsilon_B^\top \varepsilon_B \mathbf{x}_s}{2}\right)\right] \\
&= \left(\frac{\|\varepsilon_{\boldsymbol{\theta}}\|^2_{V_t}}{2\sigma_r^2} + \frac{1}{2}\sum_{i,j=1}^{d} (\varepsilon_B^i)^\top V_t \varepsilon_B^j\right) \\
&= \left(\frac{\|\varepsilon_{\boldsymbol{\theta}}\|^2_{V_t}}{2\sigma_r^2} + \frac{1}{2}\sum_{i=1}^{d} \|\varepsilon_B^i\|^2_{V_t} + \frac{1}{2}\sum_{i,j\in[d], i\neq j} (\varepsilon_B^i)^\top V_t \varepsilon_B^j\right)
\end{aligned}
$$

Using the definition of "soft allocation", i.e., $V_t = t\sum_{\mathbf{x}\in\mathbb{X}} \boldsymbol{\omega}(\mathbf{x})\mathbf{x}\mathbf{x}^\top \triangleq tV(\boldsymbol{\omega})$, where $\boldsymbol{\omega}(x) \triangleq \left[\frac{\mathbb{E}[N_{i,\tau_\delta}]}{\mathbb{E}[\tau_\delta]}\right]_{i=1}^{K}$ (N_{i,τ_δ} being the number of pulls of arm i till time τ_δ), till stopping time τ_δ in Equation (36) and Lemma 17, we get for any fixed Θ:

$$
\mathbb{E}[\tau_\delta] \geq 2\log\left(\frac{1}{2\delta}\right)\left(\frac{\|\varepsilon_{\boldsymbol{\theta}}\|^2_{V(\boldsymbol{\omega})}}{\sigma_r^2} + \sum_{i}^{d} \|\varepsilon_B^i\|^2_{V(\boldsymbol{\omega})} + \sum_{i,j\in[d], i\neq j} (\varepsilon_B^i)^\top V(\boldsymbol{\omega}) \varepsilon_B^j\right)^{-1} \tag{35}
$$

Case II: Correlated reward and cost – $\Sigma_{rc} \in \mathbb{R}^d$ is a non-zero vector **Step 1: Getting the expected LRT.** The Log-likelihood ratio for the observation till time step $t \in \mathbb{N}$:

$$\begin{aligned}
\mathcal{L}_t(r_1, \mathbf{c}_1, \ldots, r_t, \mathbf{c}_t) &= \log\left(\prod_{s=1}^{t} \frac{\mathbb{P}_{\boldsymbol{\nu}}(r_s, \mathbf{c}_s \mid \mathbf{x}_s)}{\mathbb{P}_{\tilde{\boldsymbol{\nu}}}(r_s, \mathbf{c}_s \mid \mathbf{x}_s)}\right) \\
&= \sum_{s=1}^{t} \log\left(\frac{\mathbb{P}_{\boldsymbol{\nu}}(r_s, \mathbf{c}_s \mid \mathbf{x}_s)}{\mathbb{P}_{\tilde{\boldsymbol{\nu}}}(r_s, \mathbf{c}_s \mid \mathbf{x}_s)}\right) \\
&= \sum_{s=1}^{t} \log\left(\frac{\mathbb{P}(\mathbf{x}_s^\top \boldsymbol{\theta} + \eta_s, B\mathbf{x} + \xi_s \mid \mathbf{x}_s)}{\mathbb{P}(\mathbf{x}_s^\top \boldsymbol{\theta}' + \eta'_s, B'\mathbf{x} + \xi'_s \mid \mathbf{x}_s)}\right) \\
&= \sum_{s=1}^{t} \log\left(\frac{\mathbb{P}(\eta_s, \xi_s)}{\mathbb{P}(\eta'_s, \xi'_s)}\right) \\
&= \sum_{s=1}^{t} \left(\log\left(\frac{\mathbb{P}(\eta_s \mid \xi_s)}{\mathbb{P}(\eta'_s \mid \xi'_s)}\right) + \log\left(\frac{\mathbb{P}(\xi_s)}{\mathbb{P}(\xi'_s)}\right)\right)
\end{aligned}$$

We know the conditional distribution of reward noise given the cost noise is $\eta_s \mid \xi_s \sim \mathcal{N}(\Sigma_{rc}^\top \Sigma_c^{-1} \xi_s, S)$, where the conditional variance $S = \sigma_r^2 - \Sigma_{rc}\Sigma_c^{-1}\Sigma_{rc}$. Thus, the log-likelihood ratio becomes:

$$\begin{aligned}
&\mathcal{L}_t(r_1, \mathbf{c}_1, \ldots, r_t, \mathbf{c}_t) \\
&= \sum_{s=1}^{t} \Bigg(\frac{(r_s - \mathbf{x}_s^\top \boldsymbol{\theta}' - \Sigma_{rc}^\top \Sigma_c^{-1}(\mathbf{c}_s - B'\mathbf{x}_s))^2 - (r_s - \mathbf{x}_s^\top \boldsymbol{\theta} - \Sigma_{rc}^\top \Sigma_c^{-1}(\mathbf{c}_s - B\mathbf{x}_s))^2}{2S} \\
&\qquad + \frac{1}{2}(\mathbf{c}_s - B'\mathbf{x}_s)^\top \Sigma_c^{-1}(\mathbf{c}_s - B'\mathbf{x}_s) - (\mathbf{c}_s - B\mathbf{x}_s)^\top \Sigma_c^{-1}(\mathbf{c}_s - B\mathbf{x}_s) \Bigg) \\
&= \sum_{s=1}^{t} \Bigg(\frac{\left(\mathbf{x}_s^\top(\boldsymbol{\theta} - \boldsymbol{\theta}') - \Sigma_{rc}^\top \Sigma_c^{-1}(B - B')\mathbf{x}_s\right)\left(2\eta_s + \mathbf{x}_s^\top(\boldsymbol{\theta} - \boldsymbol{\theta}') - \Sigma_{rc}^\top \Sigma_c^{-1}(B - B')\mathbf{x}_s - 2\Sigma_{rc}^\top \Sigma_c^{-1}\xi_s\right)}{2S} \\
&\qquad + \frac{\mathbf{x}_s^\top (B' - B)^\top \Sigma_c^{-1}(B' - B)\mathbf{x}_s - 2(\mathbf{c}_s - B\mathbf{x}_s)^\top \Sigma_c^{-1}(B' - B)\mathbf{x}_s}{2} \Bigg)
\end{aligned}$$

Thus, taking expectation over the noise factors both sides, we get:

$$\begin{aligned}
\mathbb{E}[\mathcal{L}_t(r_1, \mathbf{c}_1, \ldots, r_t, \mathbf{c}_t)] &= \mathbb{E}\Bigg[\sum_{s=1}^{t} \Bigg(\frac{\left(\mathbf{x}_s^\top(\boldsymbol{\theta} - \boldsymbol{\theta}') - \Sigma_{rc}^\top \Sigma_c^{-1}(B - B')\mathbf{x}_s\right)^2}{2S} \\
&\qquad + \frac{\mathbf{x}_s^\top (B' - B)^\top \Sigma_c^{-1}(B' - B)\mathbf{x}_s}{2} \Bigg)\Bigg] \qquad (36) \\
&= \mathbb{E}\left[\sum_{s=1}^{t} \left(\frac{\left(\mathbf{x}_s^\top \varepsilon_{\boldsymbol{\theta}} - \Sigma_{rc}^\top \Sigma_c^{-1/2} \varepsilon_B \mathbf{x}_s\right)^2}{2S} + \frac{\mathbf{x}_s^\top \varepsilon_B^\top \varepsilon_B \mathbf{x}_s}{2} \right)\right] \\
&= \frac{\|\varepsilon_{\boldsymbol{\theta}} - \Sigma_{rc}^\top \Sigma_c^{-1/2} \varepsilon_B\|_{V_t}^2}{2S} + \sum_{i}^{d} \|\varepsilon_B^i\|_{V_t}^2 + \sum_{i,j \in [d], i \neq j} (\varepsilon_B^i)^\top V_t \varepsilon_B^j\,, \qquad (37)
\end{aligned}$$

where $\varepsilon_{\boldsymbol{\theta}} \triangleq (\boldsymbol{\theta} - \boldsymbol{\theta}')$, and $\varepsilon_B \triangleq \Sigma_c^{-1/2}(B - B')$.

Using the definition of "soft allocation", i.e., $V_t = t\sum_{\mathbf{x} \in \mathbb{X}} \boldsymbol{\omega}(\mathbf{x})\mathbf{x}\mathbf{x}^\top \triangleq tV(\boldsymbol{\omega})$, where $\boldsymbol{\omega}(x) \triangleq \left[\frac{\mathbb{E}[N_{i,\tau_\delta}]}{\mathbb{E}[\tau_\delta]}\right]_{i=1}^{K}$ (N_{i,τ_δ} being the number of pulls of arm i till time τ_δ), till stopping time τ_δ in Equation (36) and Lemma 17, we get:

$$\mathbb{E}[\tau_\delta] \geq 2\log\left(\frac{1}{2\delta}\right)\left(\frac{\|\varepsilon_{\boldsymbol{\theta}} - \Sigma_{rc}^\top \Sigma_c^{-1/2} \varepsilon_B\|_{V(\boldsymbol{\omega})}^2}{S} + \sum_{i}^{d} \|\varepsilon_B^i\|_{V(\boldsymbol{\omega})}^2 + \sum_{i,j \in [d], i \neq j} (\varepsilon_B^i)^\top V(\boldsymbol{\omega}) \varepsilon_B^j\right)^{-1} \qquad (38)$$

Step 2: Specifying the LRT for the alternative set. Now to get the final lower bound, we need to choose the most confusing instance Θ' such that the R.H.S in Equation (39) is maximised.

Definition 4 (Alternative environment set)**.** *For a fixed environment $\boldsymbol{\nu} = (\boldsymbol{\theta}, B)$ and a fixed correct answer $\boldsymbol{p} \in \Pi(\boldsymbol{\nu}, \epsilon)$, define*

$$\Lambda(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon) = \{\tilde{\boldsymbol{\nu}} = (\boldsymbol{\theta}', B') : \boldsymbol{p} \notin \Pi(\tilde{\boldsymbol{\nu}}, \epsilon)\} = \Lambda^c_M(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon) \cup \Lambda^r_M(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon)\,,$$

i.e., all environments under which $\boldsymbol{p}$ is no longer a valid ϵ-correct answer, where

$$\Lambda^c_M(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon) = \bigcup_{j \in [d]} \left\{\boldsymbol{\lambda} \in (\boldsymbol{\mu}'_a, B'_a)_{a=1}^K : B'_j \mathbb{X}\boldsymbol{p} > 0\right\} \text{ and}$$

$$\Lambda^r_M(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon) = \bigcup_{\boldsymbol{q} \in \triangle_K} \left\{\boldsymbol{\lambda} \in (\boldsymbol{\mu}'_a, B'_a)_{a=1}^K : \boldsymbol{\theta}'^\top \mathbb{X}(\boldsymbol{q} - \boldsymbol{p}) > \epsilon \wedge B'\mathbb{X}\boldsymbol{q} \leq 0 \wedge \boldsymbol{q} \in \arg\max_{\boldsymbol{v}} \boldsymbol{\theta}'^\top \mathbb{X}\boldsymbol{v}\right\}.$$

Thus, we can define the lower bound as

$$\mathbb{E}[\tau_\delta] \geq 2\log\left(\frac{1}{2\delta}\right)\left(\sup_{\boldsymbol{\omega} \in \triangle_K} \sup_{\boldsymbol{p} \in \Pi(\boldsymbol{\nu}, \epsilon)} \inf_{(\boldsymbol{\theta}', B') \in \Lambda(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon)} \frac{\|\varepsilon_{\boldsymbol{\theta}} - \Sigma_{rc}^\top \Sigma_c^{-1/2} \varepsilon_B\|^2_{V(\boldsymbol{\omega})}}{S} + \sum_{i=1}^d \|\varepsilon^i_B\|^2_{V(\boldsymbol{\omega})} + \sum_{\substack{i,j \in [d] \\ i \neq j}} (\varepsilon^i_B)^\top V(\boldsymbol{\omega}) \varepsilon^j_B\right)^{-1}. \tag{39}$$

□

D.2. Analytical Form for Single Constraint Case

Corollary 2 (Lower Bound for Linear Bandits under Single Constraint)**.** *Given the reward variance σ_r^2, constraint variance σ_c^2, and correlation coefficient $\rho \in [-1, 1]$, we get $T^*_{\text{lin}}(\boldsymbol{\nu}, \epsilon) =$*

$$\max_{\boldsymbol{\omega} \in \triangle_K} \max_{\boldsymbol{p} \in \mathcal{E}(\Pi(\boldsymbol{\nu}, \epsilon))} \min\left\{\min_{\boldsymbol{q} \in \triangle_K} \frac{c^2(a - \rho b)}{(ac - \rho^2 b^2)^2}\left(\frac{\boldsymbol{\theta}_r^\top \mathbb{X}(\boldsymbol{q} - \boldsymbol{p}) - \epsilon}{\sigma_r}\right)^2 + \frac{a}{(ac - \rho^2 b^2)}\left(\frac{\boldsymbol{\theta}_c^\top \mathbb{X}\boldsymbol{q}}{\sigma_c}\right)^2 - \frac{2(\rho b)}{(ac - \rho^2 b^2)}\left(\frac{\boldsymbol{\theta}_c^\top \mathbb{X}\boldsymbol{q}}{\sigma_c}\right)\left(\frac{\boldsymbol{\theta}_r^\top \mathbb{X}(\boldsymbol{q} - \boldsymbol{p}) - \epsilon}{\sigma_r}\right), \frac{(\boldsymbol{\theta}_c^\top \mathbb{X}\boldsymbol{p})^2}{\sigma_c^2(1 - \rho^2)d}\right\}, \tag{40}$$

where $a \triangleq (\boldsymbol{q} - \boldsymbol{p})^\top \mathbb{X}^\top V(\boldsymbol{\omega})^{-1} \mathbb{X}(\boldsymbol{q} - \boldsymbol{p})$, $b \triangleq (\boldsymbol{q} - \boldsymbol{p})^\top \mathbb{X}^\top V(\boldsymbol{\omega})^{-1} \mathbb{X}\boldsymbol{q}$, $c \triangleq \boldsymbol{q}^\top \mathbb{X}^\top V(\boldsymbol{\omega})^{-1} \mathbb{X}\boldsymbol{q}$, and $d \triangleq \boldsymbol{p}^\top \mathbb{X}^\top V(\boldsymbol{\omega})^{-1} \mathbb{X}\boldsymbol{p}$.

Proof. Pulling arm $\mathbf{x} \in \mathcal{X}$ at round t yields:

$$r_t = \langle \mathbf{x}, \boldsymbol{\theta}_r \rangle + \eta_t, \quad \eta_t \sim \mathcal{N}(0, \sigma_r^2),$$
$$\mathbf{c}_t = \langle \mathbf{x}, \boldsymbol{\theta}_c \rangle + \xi_t, \quad \xi_t \sim \mathcal{N}(0, \sigma_c^2),$$

where σ_r^2 and σ_c^2 are known variances. We assume the joint distribution of $\gamma_t \triangleq (\eta_t, \xi_t) \sim \mathcal{N}(\mathbf{0}^2, \Sigma)$, where the covariance matrix $\Sigma = \begin{bmatrix} \sigma_r^2 & \rho\sigma_r\sigma_c \\ \rho\sigma_r\sigma_c & \sigma_c^2 \end{bmatrix}$ is known, and $-1 \leq \rho \leq 1$ is the Pearson correlation coefficient. Then, the lower bound in Equation (39) takes the form:

$$\mathbb{E}[\tau_\delta] \geq 2\log\left(\frac{1}{2\delta}\right)\left(\sup_{\boldsymbol{\omega} \in \triangle_K} \sup_{\boldsymbol{p} \in \Pi(\boldsymbol{\nu}, \epsilon)} \inf_{(\boldsymbol{\theta}'_r, \boldsymbol{\theta}'_c) \in \Lambda(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon)} \underbrace{\frac{\|\varepsilon_{\boldsymbol{\theta}_r} - \rho\sigma_r \varepsilon_{\boldsymbol{\theta}_c}\|^2_{V(\boldsymbol{\omega})}}{S} + \|\varepsilon_{\boldsymbol{\theta}_c}\|^2_{V(\boldsymbol{\omega})}}_{\triangleq f(\boldsymbol{\theta}'_r, \boldsymbol{\theta}'_c; \boldsymbol{\omega}, \boldsymbol{p})}\right)^{-1},$$

where the conditional variance of reward given cost noise $S = \mathrm{Var}(\eta_t \mid \xi_t) = \sigma_r^2(1-\rho^2)$

Thus, to get the final lower bound we peel the optimization problem in Equation (39) to get the first inner optimization for a fixed allocation $\boldsymbol{\omega}$, and correct answer $\boldsymbol{p}$ as:

$$
\begin{aligned}
F(\boldsymbol{\omega}, \boldsymbol{p}) &= \inf_{(\boldsymbol{\theta}'_r, \boldsymbol{\theta}'_c) \in \Lambda(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon)} f(\boldsymbol{\theta}'_r, \boldsymbol{\theta}'_c; \boldsymbol{\omega}, \boldsymbol{p}) \\
&= \min \left\{ \inf_{(\boldsymbol{\theta}'_r, \boldsymbol{\theta}'_c) \in \Lambda^r_M(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon)} f(\boldsymbol{\theta}'_r, \boldsymbol{\theta}'_c; \boldsymbol{\omega}, \boldsymbol{p}), \inf_{(\boldsymbol{\theta}'_r, \boldsymbol{\theta}'_c) \in \Lambda^c_M(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon)} f(\boldsymbol{\theta}'_r, \boldsymbol{\theta}'_c; \boldsymbol{\omega}, \boldsymbol{p}) \right\} \\
&= \min \left\{ \underbrace{\inf_{\boldsymbol{q} \in \triangle_K} \inf_{\substack{\boldsymbol{\theta'}_r^\top \mathbb{X}(\boldsymbol{q}-\boldsymbol{p}) > \epsilon \\ \boldsymbol{\theta'}_c^\top \mathbb{X}\boldsymbol{q} \le 0}} f(\boldsymbol{\theta}'_r, \boldsymbol{\theta}'_c; \boldsymbol{\omega}, \boldsymbol{p})}_{\triangleq D^r(\boldsymbol{\nu}, \epsilon, \boldsymbol{\omega}, \boldsymbol{p})}, \underbrace{\inf_{\boldsymbol{\theta'}_c^\top \mathbb{X}\boldsymbol{p} > 0} f(\boldsymbol{\theta}'_r, \boldsymbol{\theta}'_c; \boldsymbol{\omega}, \boldsymbol{p})}_{\triangleq D^c(\boldsymbol{\nu}, \boldsymbol{\omega}, \boldsymbol{p})} \right\}
\end{aligned}
$$

The objective function for a given allocation $\boldsymbol{\omega} \in \triangle_K$, a correct answer $\boldsymbol{p} \in \Pi(\boldsymbol{\nu}, \epsilon)$, and $\boldsymbol{q} \in \triangle_K$:

$$
\begin{aligned}
& f(\boldsymbol{\theta}'_r, \boldsymbol{\theta}'_c; \boldsymbol{\omega}, \boldsymbol{p}) \\
=& \frac{\|\varepsilon_{\boldsymbol{\theta}_r} - \rho\sigma_r \varepsilon_{\boldsymbol{\theta}_c}\|^2_{V(\boldsymbol{\omega})}}{S} + \|\varepsilon_{\boldsymbol{\theta}_c}\|^2_{V(\boldsymbol{\omega})} \\
=& \frac{\left(\boldsymbol{\theta}_r - \boldsymbol{\theta}'_r - \rho\frac{\sigma_r}{\sigma_c}(\boldsymbol{\theta}_c - \boldsymbol{\theta}'_c)\right)^\top V(\boldsymbol{\omega}) \left(\boldsymbol{\theta}_r - \boldsymbol{\theta}'_r - \rho\frac{\sigma_r}{\sigma_c}(\boldsymbol{\theta}_c - \boldsymbol{\theta}'_c)\right)}{S} + \frac{(\boldsymbol{\theta}_c - \boldsymbol{\theta}'_c)^\top V(\boldsymbol{\omega})(\boldsymbol{\theta}_c - \boldsymbol{\theta}'_c)}{\sigma_c^2}
\end{aligned}
$$

Now, we analyse the term $D^r(\boldsymbol{\nu}, \epsilon, \boldsymbol{\omega}, \boldsymbol{p})$ for a fixed $\boldsymbol{q} \in \triangle_K$ by writing the Lagrangian dual with Lagrangian multipliers $\gamma, \zeta \in \mathbb{R}_{\ge 0}$:

$$
\mathcal{L}(\gamma, \zeta; \boldsymbol{\theta}'_r, \boldsymbol{\theta}'_c) = f(\boldsymbol{\theta}'_r, \boldsymbol{\theta}'_c; \boldsymbol{\omega}, \boldsymbol{p}) + \gamma(\boldsymbol{\theta}_r{}'^\top \mathbb{X}(\boldsymbol{q} - \boldsymbol{p}) - \epsilon) - \zeta \boldsymbol{\theta'}_c^\top \mathbb{X}\boldsymbol{q} \tag{41}
$$

Differentiating w.r.t $\boldsymbol{\theta}'_r$, and equating it to 0, we get:

$$
\begin{aligned}
& \nabla_{\boldsymbol{\theta}'_r} \mathcal{L}(\gamma, \zeta; \boldsymbol{\theta}'_r, \boldsymbol{\theta}'_c) = 0 \\
\implies & -\frac{2}{S} V(\boldsymbol{\omega})(\boldsymbol{\theta}_r - \boldsymbol{\theta}'_r - \rho\frac{\sigma_r}{\sigma_c}(\boldsymbol{\theta}_c - \boldsymbol{\theta}'_c)) = -\gamma \mathbb{X}(\boldsymbol{q} - \boldsymbol{p}) \\
\implies & \boldsymbol{\theta}_r - \boldsymbol{\theta}'_r = \rho\frac{\sigma_r}{\sigma_c}(\boldsymbol{\theta}_c - \boldsymbol{\theta}'_c) + \frac{S}{2}\gamma V(\boldsymbol{\omega})^{-1}\mathbb{X}(\boldsymbol{q} - \boldsymbol{p})
\end{aligned}
$$

Putting the value of $\boldsymbol{\theta}'_r$ back into Equation (41) we get

$$
\begin{aligned}
\mathcal{L}(\gamma, \zeta; \boldsymbol{\theta}'_c) =& \frac{S\gamma^2}{4}(\boldsymbol{q} - \boldsymbol{p})^\top \mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}(\boldsymbol{q} - \boldsymbol{p}) + \frac{(\boldsymbol{\theta}_c - \boldsymbol{\theta}'_c)^\top V(\boldsymbol{\omega})(\boldsymbol{\theta}_c - \boldsymbol{\theta}'_c)}{\sigma_c^2} \\
&+ \gamma\left(\boldsymbol{\theta}_r - \rho\frac{\sigma_r}{\sigma_c}(\boldsymbol{\theta}_c - \boldsymbol{\theta}'_c) - \frac{S}{2}\gamma V(\boldsymbol{\omega})^{-1}\mathbb{X}(\boldsymbol{q} - \boldsymbol{p})\right)^\top \mathbb{X}(\boldsymbol{q} - \boldsymbol{p}) - \gamma\epsilon - \zeta \boldsymbol{\theta'}_c^\top \mathbb{X}\boldsymbol{q} \\
=& -\frac{S\gamma^2}{4}(\boldsymbol{q} - \boldsymbol{p})^\top \mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}(\boldsymbol{q} - \boldsymbol{p}) + \gamma\boldsymbol{\theta}_r^\top \mathbb{X}(\boldsymbol{q} - \boldsymbol{p}) - \gamma\rho\frac{\sigma_r}{\sigma_c}(\boldsymbol{\theta}_c - \boldsymbol{\theta}'_c)^\top \mathbb{X}(\boldsymbol{q} - \boldsymbol{p}) - \gamma\epsilon \\
&+ \frac{(\boldsymbol{\theta}_c - \boldsymbol{\theta}'_c)^\top V(\boldsymbol{\omega})(\boldsymbol{\theta}_c - \boldsymbol{\theta}'_c)}{\sigma_c^2} - \zeta \boldsymbol{\theta'}_c^\top \mathbb{X}\boldsymbol{q}
\end{aligned}
$$

Differentiating w.r.t $\boldsymbol{\theta}'_c$, and equating it to 0, we get:

$$
\begin{aligned}
&\nabla_{\boldsymbol{\theta}'_c}\mathcal{L}(\gamma,\zeta;\boldsymbol{\theta}'_c)=0\\
\implies &\gamma\rho\frac{\sigma_r}{\sigma_c}\mathbb{X}(\boldsymbol{q}-\boldsymbol{p})+2\frac{V(\boldsymbol{\omega})(\boldsymbol{\theta}'_c-\boldsymbol{\theta}_c)}{\sigma_c^2}=\zeta\mathbb{X}\boldsymbol{q}\\
\implies &\boldsymbol{\theta}'_c-\boldsymbol{\theta}_c=\frac{\sigma_c^2}{2}V(\boldsymbol{\omega})^{-1}\zeta\mathbb{X}\boldsymbol{q}-\gamma\rho\frac{\sigma_r\sigma_c}{2}V(\boldsymbol{\omega})^{-1}\mathbb{X}(\boldsymbol{q}-\boldsymbol{p})
\end{aligned}
$$

Thus, putting the value of $\boldsymbol{\theta}'_c$ back in $\mathcal{L}(\gamma,\zeta;\boldsymbol{\theta}'_c)$, we get:

$$
\begin{aligned}
&\mathcal{L}(\gamma,\zeta)\\
=&-\frac{S\gamma^2}{4}(\boldsymbol{q}-\boldsymbol{p})^\top\mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}(\boldsymbol{q}-\boldsymbol{p})+\gamma\boldsymbol{\theta}_r^\top\mathbb{X}(\boldsymbol{q}-\boldsymbol{p})-\gamma\epsilon\\
&-\gamma\rho\frac{\sigma_r}{\sigma_c}(-\frac{\sigma_c^2}{2}V(\boldsymbol{\omega})^{-1}\zeta\mathbb{X}\boldsymbol{q}+\gamma\rho\frac{\sigma_r\sigma_c}{2}V(\boldsymbol{\omega})^{-1}\mathbb{X}(\boldsymbol{q}-\boldsymbol{p}))^\top\mathbb{X}(\boldsymbol{q}-\boldsymbol{p})\\
&+\frac{(\frac{\sigma_c^2}{2}V(\boldsymbol{\omega})^{-1}\zeta\mathbb{X}\boldsymbol{q}-\gamma\rho\frac{\sigma_r\sigma_c}{2}V(\boldsymbol{\omega})^{-1}\mathbb{X}(\boldsymbol{q}-\boldsymbol{p}))^\top V(\boldsymbol{\omega})(\frac{\sigma_c^2}{2}V(\boldsymbol{\omega})^{-1}\zeta\mathbb{X}\boldsymbol{q}-\gamma\rho\frac{\sigma_r\sigma_c}{2}V(\boldsymbol{\omega})^{-1}\mathbb{X}(\boldsymbol{q}-\boldsymbol{p}))^\top}{\sigma_c^2}\\
&-\zeta(\boldsymbol{\theta}_c+\frac{\sigma_c^2}{2}V(\boldsymbol{\omega})^{-1}\zeta\mathbb{X}\boldsymbol{q}-\gamma\rho\frac{\sigma_r\sigma_c}{2}V(\boldsymbol{\omega})^{-1}\mathbb{X}(\boldsymbol{q}-\boldsymbol{p}))^\top\mathbb{X}\boldsymbol{q}\\
=&-\frac{\sigma_r^2(1-\rho^2)\gamma^2}{4}(\boldsymbol{q}-\boldsymbol{p})^\top\mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}(\boldsymbol{q}-\boldsymbol{p})+\gamma\left(\boldsymbol{\theta}_r^\top\mathbb{X}(\boldsymbol{q}-\boldsymbol{p})-\epsilon\right)\\
&+\gamma\zeta\rho\frac{\sigma_r\sigma_c}{2}\boldsymbol{q}^\top\mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}(\boldsymbol{q}-\boldsymbol{p})-\gamma^2\rho^2\frac{\sigma_r^2}{2}(\boldsymbol{q}-\boldsymbol{p})^\top\mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}(\boldsymbol{q}-\boldsymbol{p})\\
&+\frac{\sigma_c^2\zeta^2}{4}\boldsymbol{q}^\top\mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}\boldsymbol{q}+\frac{\rho^2\sigma_r^2\gamma^2}{4}(\boldsymbol{q}-\boldsymbol{p})^\top\mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}(\boldsymbol{q}-\boldsymbol{p})-\frac{\rho\sigma_r\sigma_c\zeta\gamma}{2}\boldsymbol{q}^\top\mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}(\boldsymbol{q}-\boldsymbol{p})\\
&-\zeta\boldsymbol{\theta}_c^\top\mathbb{X}\boldsymbol{q}-\zeta^2\frac{\sigma_c^2}{2}\boldsymbol{q}^\top\mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}\boldsymbol{q}+\gamma\zeta\rho\frac{\sigma_r\sigma_c}{2}(\boldsymbol{q}-\boldsymbol{p})^\top\mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}\boldsymbol{q}\\
=&-\frac{\sigma_r^2\gamma^2}{4}(\boldsymbol{q}-\boldsymbol{p})^\top\mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}(\boldsymbol{q}-\boldsymbol{p})+\gamma\left(\boldsymbol{\theta}_r^\top\mathbb{X}(\boldsymbol{q}-\boldsymbol{p})-\epsilon\right)\\
&+\gamma\zeta\rho\frac{\sigma_r\sigma_c}{2}(\boldsymbol{q}-\boldsymbol{p})^\top\mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}\boldsymbol{q}-\zeta^2\frac{\sigma_c^2}{4}\boldsymbol{q}^\top\mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}\boldsymbol{q}-\zeta\boldsymbol{\theta}_c^\top\mathbb{X}\boldsymbol{q}
\end{aligned}
$$

Differentiating w.r.t γ, and equating it to 0, we get

$$
\begin{aligned}
&0=\frac{\mathrm{d}\mathcal{L}(\gamma,\zeta)}{\mathrm{d}\gamma}\\
\implies\quad &0=-\frac{\sigma_r^2\gamma}{2}(\boldsymbol{q}-\boldsymbol{p})^\top\mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}(\boldsymbol{q}-\boldsymbol{p})+\left(\boldsymbol{\theta}_r^\top\mathbb{X}(\boldsymbol{q}-\boldsymbol{p})-\epsilon\right)\\
&\qquad+\zeta\rho\frac{\sigma_r\sigma_c}{2}(\boldsymbol{q}-\boldsymbol{p})^\top\mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}\boldsymbol{q}\\
\implies\quad &\gamma=\frac{2\left(\boldsymbol{\theta}_r^\top\mathbb{X}(\boldsymbol{q}-\boldsymbol{p})-\epsilon\right)+\zeta\rho\sigma_r\sigma_c(\boldsymbol{q}-\boldsymbol{p})^\top\mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}\boldsymbol{q}}{\sigma_r^2(\boldsymbol{q}-\boldsymbol{p})^\top\mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}(\boldsymbol{q}-\boldsymbol{p})}
\end{aligned}
$$

Now, substituting the value of γ, we get

$$
\begin{aligned}
\mathcal{L}(\zeta)=&\frac{\left(2\left(\boldsymbol{\theta}_r^\top\mathbb{X}(\boldsymbol{q}-\boldsymbol{p})-\epsilon\right)+\zeta\rho\sigma_r\sigma_c(\boldsymbol{q}-\boldsymbol{p})^\top\mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}\boldsymbol{q}\right)^2}{4\sigma_r^2(\boldsymbol{q}-\boldsymbol{p})^\top\mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}(\boldsymbol{q}-\boldsymbol{p})}\\
&-\zeta^2\frac{\sigma_c^2}{4}\boldsymbol{q}^\top\mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}\boldsymbol{q}-\zeta\boldsymbol{\theta}_c^\top\mathbb{X}\boldsymbol{q}
\end{aligned}
$$

Differentiating w.r.t ζ, and equating it to 0, we get

$$
0=\frac{\mathrm{d}\mathcal{L}(\gamma,\zeta)}{\mathrm{d}\zeta}
$$

$$
\begin{aligned}
\implies \quad 0 &= \frac{\rho\sigma_c}{\sigma_r}\frac{\left(\boldsymbol{\theta}_r^\top \mathbb{X}(\boldsymbol{q}-\boldsymbol{p})-\epsilon\right)}{(\boldsymbol{q}-\boldsymbol{p})^\top \mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}(\boldsymbol{q}-\boldsymbol{p})}(\boldsymbol{q}-\boldsymbol{p})^\top \mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}\boldsymbol{q} \\
&\quad + \zeta\frac{\rho^2\sigma_c^2}{2}\frac{\left((\boldsymbol{q}-\boldsymbol{p})^\top \mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}\boldsymbol{q}\right)^2}{(\boldsymbol{q}-\boldsymbol{p})^\top \mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}(\boldsymbol{q}-\boldsymbol{p})} \\
&\quad - \zeta\frac{\sigma_c^2}{2}\boldsymbol{q}^\top \mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}\boldsymbol{q} - \boldsymbol{\theta}_c^\top \mathbb{X}\boldsymbol{q} \\
\implies \quad \zeta &= \frac{-\boldsymbol{\theta}_c^\top \mathbb{X}\boldsymbol{q} + \frac{\rho\sigma_c}{\sigma_r}\frac{\left(\boldsymbol{\theta}_r^\top \mathbb{X}(\boldsymbol{q}-\boldsymbol{p})-\epsilon\right)}{(\boldsymbol{q}-\boldsymbol{p})^\top \mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}(\boldsymbol{q}-\boldsymbol{p})}(\boldsymbol{q}-\boldsymbol{p})^\top \mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}\boldsymbol{q}}{\frac{\sigma_c^2}{2}\boldsymbol{q}^\top \mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}\boldsymbol{q} - \frac{\rho^2\sigma_c^2}{2}\frac{\left((\boldsymbol{q}-\boldsymbol{p})^\top \mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}\boldsymbol{q}\right)^2}{(\boldsymbol{q}-\boldsymbol{p})^\top \mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}(\boldsymbol{q}-\boldsymbol{p})}} \\
&= \frac{-\boldsymbol{\theta}_c^\top \mathbb{X}\boldsymbol{q} + \frac{\rho\sigma_c}{\sigma_r}\frac{\left(\boldsymbol{\theta}_r^\top \mathbb{X}(\boldsymbol{q}-\boldsymbol{p})-\epsilon\right)}{a}b}{\frac{\sigma_c^2}{2}c - \frac{\rho^2\sigma_c^2}{2}\frac{b^2}{a}} = \frac{2}{\sigma_c^2}\frac{\frac{\rho\sigma_c b}{\sigma_r}\left(\boldsymbol{\theta}_r^\top \mathbb{X}(\boldsymbol{q}-\boldsymbol{p})-\epsilon\right) - a\boldsymbol{\theta}_c^\top \mathbb{X}\boldsymbol{q}}{ac-\rho^2 b^2}
\end{aligned}
$$

Here,

$$
a = (\boldsymbol{q}-\boldsymbol{p})^\top \mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}(\boldsymbol{q}-\boldsymbol{p}) \tag{42}
$$

$$
b = (\boldsymbol{q}-\boldsymbol{p})^\top \mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}\boldsymbol{q} \tag{43}
$$

$$
c = \boldsymbol{q}^\top \mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}\boldsymbol{q} \tag{44}
$$

Now substituting the value of ζ, we get

$$
\begin{aligned}
\mathcal{L} &= \frac{\left(2\left(\boldsymbol{\theta}_r^\top \mathbb{X}(\boldsymbol{q}-\boldsymbol{p})-\epsilon\right)+\zeta\rho\sigma_r\sigma_c b\right)^2}{4\sigma_r^2 a} - \zeta^2\frac{\sigma_c^2}{4}c - \zeta\boldsymbol{\theta}_c^\top \mathbb{X}\boldsymbol{q} \\
&= a\frac{\left(c\left(\boldsymbol{\theta}_r^\top \mathbb{X}(\boldsymbol{q}-\boldsymbol{p})-\epsilon\right) - \frac{\rho\sigma_r}{\sigma_c}b\boldsymbol{\theta}_c^\top \mathbb{X}\boldsymbol{q}\right)^2}{\sigma_r^2(ac-b^2\rho^2)^2} - c\rho^2\frac{\left(b\left(\boldsymbol{\theta}_r^\top \mathbb{X}(\boldsymbol{q}-\boldsymbol{p})-\epsilon\right) - \frac{\sigma_r}{\rho\sigma_c}a\boldsymbol{\theta}_c^\top \mathbb{X}\boldsymbol{q}\right)^2}{\sigma_r^2(ac-\rho^2 b^2)^2} \\
&\quad - \left(\frac{2\rho}{\sigma_r\sigma_c}\boldsymbol{\theta}_c^\top \mathbb{X}\boldsymbol{q}\right)\frac{b\left(\boldsymbol{\theta}_r^\top \mathbb{X}(\boldsymbol{q}-\boldsymbol{p})-\epsilon\right) - \frac{\sigma_r}{\rho\sigma_c}a\boldsymbol{\theta}_c^\top \mathbb{X}\boldsymbol{q}}{ac-\rho^2 b^2} \\
&= \frac{a\left(c\left(\boldsymbol{\theta}_r^\top \mathbb{X}(\boldsymbol{q}-\boldsymbol{p})-\epsilon\right) - (\rho b)\left(\frac{\sigma_r}{\sigma_c}\boldsymbol{\theta}_c^\top \mathbb{X}\boldsymbol{q}\right)\right)^2 - c\left((\rho b)\left(\boldsymbol{\theta}_r^\top \mathbb{X}(\boldsymbol{q}-\boldsymbol{p})-\epsilon\right) - a\left(\frac{\sigma_r}{\sigma_c}\boldsymbol{\theta}_c^\top \mathbb{X}\boldsymbol{q}\right)\right)^2}{\sigma_r^2(ac-\rho^2 b^2)^2} \\
&\quad - 2\frac{\left(\frac{\sigma_r}{\sigma_c}\boldsymbol{\theta}_c^\top \mathbb{X}\boldsymbol{q}\right)\left((\rho b)\left(\boldsymbol{\theta}_r^\top \mathbb{X}(\boldsymbol{q}-\boldsymbol{p})-\epsilon\right) - a\left(\frac{\sigma_r}{\sigma_c}\boldsymbol{\theta}_c^\top \mathbb{X}\boldsymbol{q}\right)\right)}{\sigma_r^2(ac-\rho^2 b^2)} \\
&= \frac{c^2(a-\rho b)}{\sigma_r^2(ac-\rho^2 b^2)^2}\left(\boldsymbol{\theta}_r^\top \mathbb{X}(\boldsymbol{q}-\boldsymbol{p})-\epsilon\right)^2 + \frac{a}{\sigma_r^2(ac-\rho^2 b^2)}\left(\frac{\sigma_r}{\sigma_c}\boldsymbol{\theta}_c^\top \mathbb{X}\boldsymbol{q}\right)^2 \\
&\qquad - \frac{2(\rho b)}{\sigma_r^2(ac-\rho^2 b^2)}\left(\frac{\sigma_r}{\sigma_c}\boldsymbol{\theta}_c^\top \mathbb{X}\boldsymbol{q}\right)\left(\boldsymbol{\theta}_r^\top \mathbb{X}(\boldsymbol{q}-\boldsymbol{p})-\epsilon\right) \\
&= \frac{c^2(a-\rho b)}{\sigma_r^2(ac-\rho^2 b^2)^2}\left(\boldsymbol{\theta}_r^\top \mathbb{X}(\boldsymbol{q}-\boldsymbol{p})-\epsilon\right)^2 \\
&\qquad + \frac{a}{\sigma_r^2(ac-\rho^2 b^2)}\left(\frac{\sigma_r}{\sigma_c}\boldsymbol{\theta}_c^\top \mathbb{X}\boldsymbol{q}\right)\left(a\left(\frac{\sigma_r}{\sigma_c}\boldsymbol{\theta}_c^\top \mathbb{X}\boldsymbol{q}\right) - 2(\rho b)\left(\left(\boldsymbol{\theta}_r^\top \mathbb{X}(\boldsymbol{q}-\boldsymbol{p})-\epsilon\right)\right)\right) \\
&= \frac{c^2(a-\rho b)}{(ac-\rho^2 b^2)^2}\left(\frac{\boldsymbol{\theta}_r^\top \mathbb{X}(\boldsymbol{q}-\boldsymbol{p})-\epsilon}{\sigma_r}\right)^2 + \frac{a}{(ac-\rho^2 b^2)}\left(\frac{\boldsymbol{\theta}_c^\top \mathbb{X}\boldsymbol{q}}{\sigma_c}\right)^2 \\
&\qquad - \frac{2(\rho b)}{(ac-\rho^2 b^2)}\left(\frac{\boldsymbol{\theta}_c^\top \mathbb{X}\boldsymbol{q}}{\sigma_c}\right)\left(\frac{\boldsymbol{\theta}_r^\top \mathbb{X}(\boldsymbol{q}-\boldsymbol{p})-\epsilon}{\sigma_r}\right)
\end{aligned}
$$

Now we analyse the term $D^c(\boldsymbol{\nu},\boldsymbol{\omega},\boldsymbol{p})$. We write the corresponding Lagrangian dual for $D^c(\boldsymbol{\nu},\boldsymbol{\omega},\boldsymbol{p})$ using Lagrangian

multiplier $\gamma \in \mathbb{R}_{\geq 0}$ as:

$$\mathcal{L}(\boldsymbol{\theta}'_c, \gamma; \boldsymbol{\omega}, \boldsymbol{p}) = \frac{(\boldsymbol{\theta}_c - \boldsymbol{\theta}'_c)^\top V(\boldsymbol{\omega})(\boldsymbol{\theta}_c - \boldsymbol{\theta}'_c)}{\sigma_c^2(1-\rho^2)} + \gamma {\boldsymbol{\theta}'}_c^\top \mathbb{X}\boldsymbol{p}$$

Differentiating $\mathcal{L}(\boldsymbol{\theta}'_c, \gamma; \boldsymbol{\omega}, \boldsymbol{p})$ with respect to $\boldsymbol{\theta}'_c$ and equating it to 0 we get:

$$\begin{aligned}
&\nabla_{\boldsymbol{\theta}'_c} \mathcal{L}(\boldsymbol{\theta}'_c, \gamma; \boldsymbol{\omega}, \boldsymbol{p}) = 0 \\
\implies & 2\frac{V(\boldsymbol{\omega})(\boldsymbol{\theta}'_c - \boldsymbol{\theta}_c)}{\sigma_c^2(1-\rho^2)} = -\gamma \mathbb{X}\boldsymbol{p} \\
\implies & \boldsymbol{\theta}'_c - \boldsymbol{\theta}_c = -\frac{\gamma\sigma_c^2(1-\rho^2)}{2} V(\boldsymbol{\omega})^{-1}\mathbb{X}\boldsymbol{p}
\end{aligned}$$

Putting the obtained value of $\boldsymbol{\theta}'_c$ in $\mathcal{L}(\boldsymbol{\theta}'_c; \boldsymbol{\omega}, \boldsymbol{p})$ yields:

$$\begin{aligned}
\mathcal{L}(\gamma; \boldsymbol{\omega}, \boldsymbol{p}) =& \frac{\gamma^2\sigma_c^2(1-\rho^2)}{4}\boldsymbol{p}^\top\mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}\boldsymbol{p} + \gamma\left(\boldsymbol{\theta}_c - \frac{\gamma\sigma_c^2(1-\rho^2)}{2}V(\boldsymbol{\omega})^{-1}\mathbb{X}\boldsymbol{p}\right)^\top \mathbb{X}\boldsymbol{p} \\
=& -\frac{\gamma^2\sigma_c^2(1-\rho^2)}{4}\boldsymbol{p}^\top\mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}\boldsymbol{p} + \gamma\boldsymbol{\theta}_c^\top\mathbb{X}\boldsymbol{p}
\end{aligned}$$

Differentiating $\mathcal{L}(\gamma; \boldsymbol{\omega}, \boldsymbol{p})$ with respect to $\boldsymbol{\theta}'_c$ and equating it to 0 we get:

$$\begin{aligned}
&\nabla_\gamma \mathcal{L}(\gamma; \boldsymbol{\omega}, \boldsymbol{p}) = 0 \\
\implies & -\frac{\gamma\sigma_c^2(1-\rho^2)}{2}\boldsymbol{p}^\top\mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}\boldsymbol{p} + \boldsymbol{\theta}_c^\top\mathbb{X}\boldsymbol{p} = 0 \\
\implies & \gamma = \frac{2}{\sigma_c^2(1-\rho^2)d}\boldsymbol{\theta}_c^\top\mathbb{X}\boldsymbol{p}\ ,
\end{aligned}$$

where $d \triangleq \boldsymbol{p}^\top\mathbb{X}^\top V(\boldsymbol{\omega})^{-1}\mathbb{X}\boldsymbol{p}$.

Thus, with the optimal value of the Lagrangian multiplier γ, we get the final expression of the dual:

$$\mathcal{L} = \frac{(\boldsymbol{\theta}_c^\top\mathbb{X}\boldsymbol{p})^2}{\sigma_c^2(1-\rho^2)d}$$

Putting these final dual values in place of $D^r(\boldsymbol{\nu}, \epsilon, \boldsymbol{\omega}, \boldsymbol{p})$ and $D^c(\boldsymbol{\nu}, \boldsymbol{\omega}, \boldsymbol{p})$ yields the Corollary statement. □

E. Structural Reduction of the Lower Bounds

In this section, we will prove some useful results on the structure of alt-set, the set of correct answers, and the optimization problems in general that will in turn, help us structurally reduce the lower bound formulation for (SOP).

Lemma 5 (Openness of $\Pi(\boldsymbol{\nu}, \epsilon)$ as a subset of $\boldsymbol{\nu}$)**.** *Under Assumption 1, the set of instances for which a fixed $\boldsymbol{p} \in \Pi(\boldsymbol{\nu}, \epsilon)$ remains ϵ-correct and feasible is relatively open in $\boldsymbol{\nu}$. That is, for any $\boldsymbol{\nu}_0 \triangleq (\boldsymbol{\mu}_0, B_0) \in (\mathcal{D}_{\boldsymbol{\mu}}, \mathcal{D}_B)$ such that $\boldsymbol{p} \in \Pi(\boldsymbol{\nu}_0, \epsilon)$, there exists $\eta > 0$ such that for all $\boldsymbol{\nu} \triangleq (\boldsymbol{\mu}, B)$ with $\|\boldsymbol{\nu} - \boldsymbol{\nu}_0\|_\infty \leq \eta$, we also have $\boldsymbol{p} \in \Pi(\boldsymbol{\nu}, \epsilon)$.*

Proof. Since $\boldsymbol{p} \in \Pi(\boldsymbol{\nu}_0, \epsilon)$, by definition:

(i) $\boldsymbol{\mu}_0^\top \boldsymbol{p} \geq \boldsymbol{\mu}_0^\top \boldsymbol{p}^\star(\boldsymbol{\nu}_0) - \epsilon$ ($\boldsymbol{p}$ is ϵ-optimal under $\boldsymbol{\nu}_0$),

(ii) $A_0 \boldsymbol{p} \leq 0$ component-wise ($\boldsymbol{p}$ is feasible under $\boldsymbol{\nu}_0$).

By Assumption 1, we have $\|A_{0,j}\|_2 \leq 1$ for all j, so the cost parameters are bounded, and the linear maps

$$\boldsymbol{\nu} \longmapsto \boldsymbol{\mu}^\top \boldsymbol{p} \quad \text{and} \quad \boldsymbol{\nu} \longmapsto (B\boldsymbol{p})_j, \ j \in [d]$$

are continuous in $\boldsymbol{\nu} = (\boldsymbol{\mu}, B)$.

For condition (ii): since $A_0 \boldsymbol{p} \leq 0$ (i.e., $(A_0 \boldsymbol{p})_j < 0$ for all constraint index j that are not tight), by continuity there exists $\eta_1 > 0$ such that $\|B - A_0\|_\infty \leq \eta_1 \Rightarrow B\boldsymbol{p} < 0$.

For condition (i): Since $\boldsymbol{p}^\star(\boldsymbol{\nu})$ varies continuously in $\boldsymbol{\nu}$ on the interior of the parameter space (by continuity of linear programs with strictly feasible solutions, under Assumption 1 which guarantees a strictly feasible point), there exists $\eta_2 > 0$ such that $\|\boldsymbol{\nu} - \boldsymbol{\nu}_0\|_\infty \leq \eta_2 \Rightarrow \boldsymbol{\mu}^\top \boldsymbol{p} \geq \boldsymbol{\mu}^\top \boldsymbol{p}^\star(\boldsymbol{\nu}) - \epsilon$.

Taking $\eta = \min\{\eta_1, \eta_2\}$ yields the open neighbourhood, proving that the set of instances under which $\boldsymbol{p}$ is a valid ϵ-correct answer is relatively open in $(\mathcal{D}_{\boldsymbol{\mu}}, \mathcal{D}_A)$.

Note that this result also holds trivially for linear bandits as well. □

Corollary 3. *For fixed $\boldsymbol{\omega} \in \triangle_K$, and $\boldsymbol{p} \in \Pi(\boldsymbol{\nu}, \epsilon)$, we can show that $\Lambda_M^r(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon)$ is a compact set, and thus:*

1. $$\inf_{\boldsymbol{q} \in \triangle_K} \inf_{\substack{\boldsymbol{\mu}' : \boldsymbol{\mu}'^\top (\boldsymbol{q} - \boldsymbol{p}) > \epsilon \\ B' : B'\boldsymbol{q} \leq 0}} \sum_{a=1}^K \boldsymbol{\omega}_a \mathrm{KL}\left(\boldsymbol{\mu}_a, B_a \parallel \boldsymbol{\mu}'_a, B'_a\right) = \min_{\boldsymbol{q} \in \triangle_K} \min_{\substack{\boldsymbol{\mu}' : \boldsymbol{\mu}'^\top (\boldsymbol{q} - \boldsymbol{p}) > \epsilon \\ B' : B'\boldsymbol{q} \leq 0}} \sum_{a=1}^K \boldsymbol{\omega}_a \mathrm{KL}\left(\boldsymbol{\mu}_a, B_a \parallel \boldsymbol{\mu}'_a, B'_a\right)$$

2. $\inf_{B'_j \boldsymbol{p} > 0} \sum_{a=1}^K \boldsymbol{\omega}_a \mathrm{KL}\left(B_a \parallel B'_a\right) = \min_{B'_j \boldsymbol{p} > 0} \sum_{a=1}^K \boldsymbol{\omega}_a \mathrm{KL}\left(B_a \parallel B'_a\right)$

Proof. **Step 1: Closedness of $\Lambda_M^r(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon)$.** From Lemma 5, we have shown that for a fixed allocation $\boldsymbol{\omega} \in \triangle_K$, the set of instances $\boldsymbol{\nu}$ for which a $\boldsymbol{p} \in \Pi(\boldsymbol{\nu}, \epsilon)$ is a "correct answer" is open in $(\mathcal{D}_{\boldsymbol{\mu}}, \mathcal{D}_A)$. As we know the set of alternative instance set is the complement of such $\boldsymbol{\nu}$, thus $\Lambda_M^r(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon) = \{\boldsymbol{\lambda} \in (\mathcal{D}_{\boldsymbol{\mu}}, \mathcal{D}_A) : \boldsymbol{p} \notin \Pi(\boldsymbol{\nu}, \epsilon)\}$ is a closed set.

Step 2: Boundedness of $\Lambda_M^r(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon)$. By Assumption 1, we know the parameter space $(\mathcal{D}_{\boldsymbol{\mu}}, \mathcal{D}_A)$ is bounded. Thus, combining Step 1–2, we conclude that $\Lambda_M^r(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon)$ is *compact*, and we can reduce the inner optimization from $\min \to \inf$. Hence, we conclude proof of Part (1). Proof of Part (2) follows similar argument. □

Theorem 6. *Consider the ϵ-optimal pure exploration problem under unknown linear constraints. Fix any correct answer $\boldsymbol{p} \in \Pi(\boldsymbol{\nu}, \epsilon)$. Then,*

1. *The set of instances $\boldsymbol{\nu}$ for which $\boldsymbol{p}$ remains ϵ-correct is relatively open in $(\mathcal{D}_{\boldsymbol{\mu}}, \mathcal{D}_A)$ (Lemma 5).*

2. *The alternative set $\Lambda_M^r(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon)$ decomposes as a finite union of convex sets:*

$$\Lambda_M^r(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon) = \underbrace{\Lambda_M^r(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon)}_{1 \text{ convex set}} \bigcup \underbrace{\bigcup_{j=1}^{d} \Lambda_M^c(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon, j)}_{d \text{ convex sets}},$$

a union of $d+1$ convex sets.

Hence (Wang et al., 2021a, Assumption 1) is satisfied with finite index family

$$\mathcal{J}_{\boldsymbol{p}} = \underbrace{[1]}_{\text{reward branch}} \cup \underbrace{[d]}_{\text{cost branch}}, \qquad |\mathcal{J}_{\boldsymbol{p}}| = 1 + d.$$

Proof. **Step 1: Decomposition.** By Lemma 4, we have the exact decomposition

$$\Lambda_M^r(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon) = \Big(\bigcup_{\boldsymbol{q} \in \triangle_K} \Lambda_M^r(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon, \boldsymbol{q}) \Big) \bigcup \Big(\bigcup_{j \in [d]} \Lambda_M^c(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon, j) \Big).$$

Step 2 Convexity of $\Lambda_M^r(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon)$. We first expand the definition of $\Lambda_M^r(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon)$ and define:

$$\Lambda_M^r(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon) = \underbrace{\Big\{ (\boldsymbol{\mu}', B') : \boldsymbol{\mu}'^{\top}(\boldsymbol{q} - \boldsymbol{p}) > \epsilon \Big\}}_{\triangleq H_{\text{reward}}} \cap \underbrace{\Big\{ (\boldsymbol{\mu}', B') : B'\boldsymbol{q} \leq 0 \Big\}}_{\triangleq H_{\text{cost}}}.$$

H_{reward} is convex. The constraint $\boldsymbol{\mu}'^{\top}(\boldsymbol{q} - \boldsymbol{p}) > \epsilon$ is a strict linear inequality in $\boldsymbol{\mu}'$ with fixed direction $\boldsymbol{q} - \boldsymbol{p} \in \mathbb{R}^K$. Its closure is a halfspace, hence the open version is the interior of a halfspace, which is convex.

H_{cost} is convex. The constraint $B'\boldsymbol{q} \leq 0$ means component-wise as $\sum_{a=1}^{K} \boldsymbol{q}_a B'_{a,j} \leq 0$ for each $j \in [d]$. Since $\boldsymbol{q} \in \triangle_K$ is a fixed non-negative vector, each inequality is linear in $(B'_a)_{a \in [K]}$, so H_{cost} is an intersection of d closed halfspaces in $(\boldsymbol{\mu}', B')$-space, hence convex.

The intersection of two convex sets is convex, thus $\Lambda_M^r(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon, \boldsymbol{q})$ is convex.

Step 3: Finiteness of cost branch. The cost branch $\bigcup_{j \in [d]} \Lambda_M^c(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon, j)$ is already a finite union over $j \in [d]$, giving d convex confusing sets.

Step 4: Openness of $\Pi(\boldsymbol{\nu}, \epsilon)$. By Lemma 5, the set of instances for which $\boldsymbol{p}$ is a valid ϵ-correct answer is relatively open in $(\mathcal{D}_{\boldsymbol{\mu}}, \mathcal{D}_A)$.

Combining Steps 1–4, the alternative set is a finite union of $d+1$ convex sets, and the set of "correct answers" $\Pi(\boldsymbol{\nu}, \epsilon)$ is open. Thus, both conditions of (Wang et al., 2021a, Assumption 1) are satisfied. □

We again write the expression of the characteristic time inverse for a δ-correct algorithm for pure exploration with unknown linear constraints using Corollary 3, i.e., $T_P^{\star}(\boldsymbol{\nu})^{-1} =$

$$\begin{aligned}
&\sup_{\boldsymbol{\omega} \in \triangle_K} \sup_{\boldsymbol{p} \in \Pi(\boldsymbol{\nu}, \epsilon)} \min \Bigg\{ \min_{\boldsymbol{q} \in \triangle_K} \min_{\substack{\boldsymbol{\mu}' : \boldsymbol{\mu}'^{\top}(\boldsymbol{q} - \boldsymbol{p}) > \epsilon \\ B' : B'\boldsymbol{q} \leq 0}} \sum_{a=1}^{K} \boldsymbol{\omega}_a \mathrm{KL}\left(\boldsymbol{\mu}_a, B_a \parallel \boldsymbol{\mu}'_a, B'_a\right), \\
&\qquad\qquad \min_{j \in [d]} \min_{B'_j : B'_j \boldsymbol{p} > 0} \sum_{a=1}^{K} \boldsymbol{\omega}_a \mathrm{KL}\left(B_a \parallel B'_a\right) \Bigg\} \\
&= \sup_{\boldsymbol{\omega} \in \triangle_K} \sup_{\boldsymbol{p} \in \Pi(\boldsymbol{\nu}, \epsilon)} F(\boldsymbol{\omega}, \boldsymbol{p}),
\end{aligned}$$

as we define $F(\boldsymbol{\omega}, \boldsymbol{p}) \triangleq \min\{F_1(\boldsymbol{\omega}, \boldsymbol{p}), F_2(\boldsymbol{\omega}, \boldsymbol{p})\}$, where

$$F_1(\boldsymbol{\omega}, \boldsymbol{p}) \triangleq \min_{\boldsymbol{q} \in \triangle_K} \underbrace{\min_{\substack{\boldsymbol{\mu}' : \boldsymbol{\mu}'^\top(\boldsymbol{q}-\boldsymbol{p}) > \epsilon \\ B' : B'\boldsymbol{q} \leq 0}} \sum_{a=1}^{K} \boldsymbol{\omega}_a \mathrm{KL}\left(\boldsymbol{\mu}_a, B_a \parallel \boldsymbol{\mu}'_a, B'_a\right)}_{\triangleq F_1(\boldsymbol{\omega}, \boldsymbol{p}, \boldsymbol{q})}, \text{ and}$$

$$F_2(\boldsymbol{\omega}, \boldsymbol{p}) \triangleq \min_{j \in [d]} \min_{B'_j : B'_j \boldsymbol{p} = 0} \sum_{a=1}^{K} \boldsymbol{\omega}_a \mathrm{KL}\left(B_a \parallel B'_a\right).$$

Proposition 1 (Uniqueness and continuity of inner minimisation)**.** *We define the set of bandit instances for which $\boldsymbol{p}$ is a fixed "correct answer" as $S_{\boldsymbol{p}} = \{\boldsymbol{\nu} \in (\mathcal{D}_{\boldsymbol{\mu}}, \mathcal{D}_A) : \boldsymbol{p} \in \Pi(\boldsymbol{\nu}, \epsilon)\}$. Then, for all $(\boldsymbol{\omega}, \boldsymbol{\nu}, \boldsymbol{q}) \in \mathring{\Delta}_K \times S_{\boldsymbol{p}} \times \triangle_K$,*

1. $(\boldsymbol{\mu}'^*, B'^*) = \underset{(\boldsymbol{\mu}', B') \in \Lambda^r_M(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon, \boldsymbol{q})}{\arg\min} \sum_{a=1}^{K} \boldsymbol{\omega}_a \mathrm{KL}\left(\boldsymbol{\mu}_a, B_a \parallel \boldsymbol{\mu}'_a, B'_a\right)$ *is unique, and continuous.*

2. $B'^* = \underset{j \in [d]}{\arg\min} \underset{B'_j \in \Lambda^c_M(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon, j)}{\arg\min} \sum_{a=1}^{K} \boldsymbol{\omega}_a \mathrm{KL}\left(B_a \parallel B'_a\right)$ *is unique, and continuous.*

Proof of Part (1). Existence. We define $\mathbb{X} \triangleq \triangle_K \times S_{\boldsymbol{p}} \times \triangle_K, \mathbb{Y} \triangleq \Lambda^r_M(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon, \boldsymbol{q}), \Phi \triangleq \Lambda^r_M(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon, \boldsymbol{q})$ and $u(\boldsymbol{\omega}, \boldsymbol{p}, \boldsymbol{q}) \triangleq \sum_{a=1}^{K} \boldsymbol{\omega}_a \mathrm{KL}\left(\boldsymbol{\mu}_a, B_a \parallel \boldsymbol{\mu}'_a, B'_a\right)$. Now, we observe that $\Phi : (\triangle_K \times S_{\boldsymbol{p}} \times \triangle_K) \rightrightarrows \Lambda^r_M(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon, \boldsymbol{q})$ is a *constant correspondence*. Thus, using Theorem 8 we can conclude that $(\boldsymbol{\mu}'^*, B'^*)$ is *upper-hemicontinuous* and $F_1(\boldsymbol{\omega}, \boldsymbol{p}, \boldsymbol{q})$ is *continuous* in $\triangle_K \times S_{\boldsymbol{p}} \times \triangle_K$.

Uniqueness. For each arm $a \in [K]$, (P_a, C_a) jointly follows a $d+1$-parameter exponential family (specifically, multivariate Gaussian in this case). It is evident that $\mathrm{KL}\left(\boldsymbol{\mu}_a, B_a \parallel \boldsymbol{\mu}'_a, B'_a\right)$ is *strongly convex* with respect to $(\boldsymbol{\mu}'_a, B'_a)$. Thus, the convex combination $\sum_{a=1}^{K} \boldsymbol{\omega}_a \mathrm{KL}\left(\boldsymbol{\mu}_a, B_a \parallel \boldsymbol{\mu}'_a, B'_a\right)$ for $\boldsymbol{\omega} \in \mathring{\Delta}_K$ also retains strong convexity in $(\boldsymbol{\mu}'_a, B'_a)$. Hence, we prove uniqueness of $(\boldsymbol{\mu}'^*, B'^*)$.

□

Proof of Part (2). This proof follows the similar steps as of Part (1). The existence is ensured using Theorem 8, and since C_a follows a d-parameter exponential family, $\mathrm{KL}\left(\boldsymbol{\mu}_a, B_a \parallel \boldsymbol{\mu}'_a, B'_a\right)$ is strongly convex in B'.

□

Proposition 2 (Existence and continuity of $\boldsymbol{q}^*$)**.** *For all $(\boldsymbol{\omega}, \boldsymbol{\nu}) \in \triangle_K \times S_{\boldsymbol{p}}$:*

$$\boldsymbol{q}^* = \underset{\boldsymbol{q} \in \triangle_K}{\arg\min} \sum_{a=1}^{K} \boldsymbol{\omega}_a \mathrm{KL}\left(\boldsymbol{\mu}_a, B_a \parallel \boldsymbol{\mu}'^*_a, B'^*_a\right)$$

is unique, and continuous in $\triangle_K \times S_{\boldsymbol{p}}$.

Proof. Existence. We define $\mathbb{X} \triangleq \triangle_K \times S_{\boldsymbol{p}}, \mathbb{Y} \triangleq \triangle_K, \Phi \triangleq \triangle_K$ and $u(\boldsymbol{\omega}, \boldsymbol{p}) \triangleq \sum_{a=1}^{K} \boldsymbol{\omega}_a \mathrm{KL}\left(\boldsymbol{\mu}_a, B_a \parallel \boldsymbol{\mu}'^*_a, B'^*_a\right)$. Now, we observe that $\Phi : (\triangle_K \times S_{\boldsymbol{p}}) \rightrightarrows \triangle_K$ is a *constant correspondence*. Thus, using Theorem 8 we can conclude that $\boldsymbol{q}^*$ is *upper-hemicontinuous* and $F_1(\boldsymbol{\omega}, \boldsymbol{p}, \boldsymbol{q})$ is *continuous* in $\triangle_K \times S_{\boldsymbol{p}}$.

Uniqueness. $\boldsymbol{q}$ contributes to the inner optimisation in term $T1$ only to define the linear constraints $\boldsymbol{\mu}'^\top(\boldsymbol{q} - \boldsymbol{p}) > \epsilon$, and $B'\boldsymbol{q} \leq 0$. Given unique $(\boldsymbol{\mu}'^*, B'^*)$ and continuity, it ensures the existence of an *unique* $\boldsymbol{q}^*$. Hence, proved. □

Proposition 3. 1. *For all $(\boldsymbol{\omega}, \boldsymbol{\nu}, \boldsymbol{q}) \in \mathring{\Delta}_K \times S_{\boldsymbol{p}} \times \triangle_K$, $F_1(\boldsymbol{\omega}, \boldsymbol{p}, \boldsymbol{q})$ is continuously differentiable in $\mathring{\Delta}_K \times S_{\boldsymbol{p}} \times \triangle_K$.*

2. *For all $(\boldsymbol{\omega}, \boldsymbol{\nu}) \in \mathring{\Delta}_K \times S_{\boldsymbol{p}}$, $F_1(\boldsymbol{\omega}, \boldsymbol{p}, \boldsymbol{q})$ is continuously differentiable in $\mathring{\Delta}_K \times S_{\boldsymbol{p}}$.*

Proof of Part (1). We define $\mathbb{X} \triangleq \Lambda^r_M(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon, \boldsymbol{q}), Y \triangleq \mathring{\Delta}_K \times S_{\boldsymbol{p}} \times \triangle_K$ □

Concavity of $F(\boldsymbol{\omega}, \boldsymbol{p})$ in $\boldsymbol{\omega}$. We observe that $\mathrm{KL_a}$ is convex in $(\boldsymbol{\mu}'_a, B'_a)$. It means $F_1(\boldsymbol{\omega}, \boldsymbol{p})$ is infimum over a sum of linear convex functions. Hence, $F_1(\boldsymbol{\omega}, \boldsymbol{p})$ is *concave* in $\boldsymbol{\omega}$. Following similar argument, $F_2(\boldsymbol{\omega}, \boldsymbol{p})$ is also *concave* in $\boldsymbol{\omega}$. We know that if $f_1(x), f_2(x), ..., f_m(x)$ are $m \in \mathbb{N}$ concave functions in x, then $f(x) = \min\{f_1(x), f_2(x), ..., f_m(x)\}$ is also concave in x. Thus, for a fixed "correct answer" $\boldsymbol{p} \in \Pi(\boldsymbol{\nu}, \epsilon)$, $F(\boldsymbol{\omega}, \boldsymbol{p})$ is also concave in $\boldsymbol{\omega}$.

Properties of $F(\boldsymbol{\omega}, \boldsymbol{p})$. Here, we explain all the properties of the objective function for a given allocation $\boldsymbol{\omega} \in \mathring{\Delta}_K$, and a "correct answer" $\boldsymbol{p} \in \Pi(\boldsymbol{\nu}, \epsilon)$ necessary for the convergence of the Frank-Wolfe sampler.

Corollary 4 (Lipschitz-ness of $F(\boldsymbol{\omega}, \boldsymbol{p})$)**.** *Using the fact that f_j is dD-Lipschitz and f_j's are continuous on $\mathring{\Delta}_K$, we have for any $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \mathring{\Delta}_K$, 1. $F(\cdot)$ is dD-Lipschitz for unstructured bandits, 2. $F(\cdot)$ is $dD_{\mathrm{lin}}\lambda_{\max}$-Lipschitz for unstructured bandits*

Proof. For any $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \mathring{\Delta}_K$:

$$\begin{aligned} F(\boldsymbol{\omega}, \boldsymbol{p}) = \min_{j \in \mathcal{J}_{\boldsymbol{p}}} f_j(\boldsymbol{\omega}, \boldsymbol{p}) &\geq \min_{j \in \mathcal{J}_{\boldsymbol{p}}} \left(f_j(\boldsymbol{\omega}') - dD\|\boldsymbol{\omega}' - \boldsymbol{\omega}\|_\infty\right) \\ &\geq \min_{j \in \mathcal{J}_{\boldsymbol{p}}} f_j(\boldsymbol{\omega}') - dD\|\boldsymbol{\omega}' - \boldsymbol{\omega}\|_\infty \\ &\geq F(\boldsymbol{\omega}', \boldsymbol{p}) - dD\,. \end{aligned}$$

Hence, proved. The result for linear bandits holds similarly. □

We further define $F(\boldsymbol{\omega}) = \sup_{\boldsymbol{p} \in \Pi(\boldsymbol{\nu}, \epsilon)} F(\boldsymbol{\omega}, \boldsymbol{p})$. Now, to prove convergence of the FW update in Equation (9),we need to ensure that there exists a finite set of minimizers $\mathcal{E}(\Pi(\boldsymbol{\nu}, \epsilon)) \triangleq \{\boldsymbol{p} \in \arg\sup_{\boldsymbol{p} \in \Pi(\boldsymbol{\nu}, \epsilon)} F(\boldsymbol{\omega}, \boldsymbol{p})\}$, given an allocation $\boldsymbol{\omega} \in \mathring{\Delta}_K$.

Proposition 4 (Properties of the outer maximisation over $\boldsymbol{p}$)**.** *For a fixed allocation $\boldsymbol{\omega} \in \mathring{\Delta}_K$:*

1. *$\Pi(\boldsymbol{\nu}, \epsilon)$ is compact.*
2. *$F(\boldsymbol{\omega}, \boldsymbol{p})$ is quasi-convex in $\boldsymbol{p}$.*
3. *$F(\boldsymbol{\omega}, \boldsymbol{p})$ is upper-semicontinuous in $\boldsymbol{p}$.*

Then by Theorem 12 we conclude that $\mathcal{E}(\Pi(\boldsymbol{\nu}, \epsilon))$ consists of the extremal point of the polytope $\Pi(\boldsymbol{\nu}, \epsilon)$, and has the finite combinatorial cardinality at most $\binom{K+d}{d}$.

Proof. **Compactness of $\Pi(\boldsymbol{\nu}, \epsilon)$.** We know that $\triangle_K$ is a closed and bounded, thus a compact set in $\mathbb{R}^K$. We further observe that $\Pi(\boldsymbol{\nu}, \epsilon)$ consists of $d+1$ linear inequalities defined over a bandit instance $\boldsymbol{\nu} \triangleq (\boldsymbol{\mu}, B)$, which makes it closed in $\boldsymbol{\nu}$. Then $\Pi(\boldsymbol{\nu}, \epsilon)$ as a proper subset of $\triangle_K$, is also compact, since any closed subset of a compact set is also compact.

Quasi-convexity of $F(\boldsymbol{\omega}, \boldsymbol{p})$. Note that $F(\boldsymbol{\omega}, \boldsymbol{p}) = \min_{j \in \mathcal{J}_{\boldsymbol{p}}} f_j(\boldsymbol{\omega}, \boldsymbol{p})$. We observe that for each $j \in \mathcal{J}_{\boldsymbol{p}}$, $f_j(\boldsymbol{\omega}, \boldsymbol{p})$ are ratio of two affine functions of $\boldsymbol{p}$. In general, these called *Rayleigh quotient* in optimisation theory literature, which are quasi-convex functions by definition. Also, quasi-convexity is preserved trivially if we take minimum over finite $j \in \mathcal{J}_{\boldsymbol{p}}$. Thus, $F(\boldsymbol{\omega}, \boldsymbol{p})$ is quasi-convex in $\boldsymbol{p}$.

Upper-semicontinuity of $F(\boldsymbol{\omega}, \boldsymbol{p})$. We define $\mathbb{X} \triangleq \triangle_K \times S_{\boldsymbol{p}}$, $\mathbb{Y} \triangleq \Pi(\boldsymbol{\nu}, \epsilon)$, $\Phi \triangleq \Pi(\boldsymbol{\nu}, \epsilon)$ and $u(\boldsymbol{\omega}, \boldsymbol{p}) \triangleq F(\boldsymbol{\omega}, \boldsymbol{p})$. Now, we observe that $\Phi : (\triangle_K \times S_{\boldsymbol{p}}) \rightrightarrows \triangle_K$ is a *constant correspondence*. Thus, using Theorem 8 we can conclude that $\boldsymbol{p}^*$ is *upper-hemicontinuous* and $F(\boldsymbol{\omega}, \boldsymbol{p},)$ is *continuous* in $\triangle_K \times S_{\boldsymbol{p}}$. Thus, $F(\boldsymbol{\omega}, \boldsymbol{p},)$ is also *upper semi-continuous* in $\boldsymbol{p}$.

Using these properties of $F(\boldsymbol{\omega}, \boldsymbol{p})$ with respect to $\boldsymbol{p}$, we conclude that the set of maximisers $\mathcal{E}(\Pi(\boldsymbol{\nu}, \epsilon))$ is finite, consisting of all the vertices of the polytope defined by $d+1$ linear constraints in $\Pi(\boldsymbol{\nu}, \epsilon)$. Also, we have $|\mathcal{E}(\Pi(\boldsymbol{\nu}, \epsilon))| \leq \binom{K}{d+1}$ following standard LP optimisation theory in a compact polytope. □

F. Gradient and Curvature of Allocation Objective

In this section we analyse the properties of the gradient norm and curvature of the optimization objective with respect to allocation $\boldsymbol{\omega} \in \mathring{\Delta}_K$.

Definition 5 (Interior of a simplex)**.** *We define a constant $c_{\min} > 0$ such that $\mathring{\Delta}_K = \{\boldsymbol{\omega}_a > c_{\min}, \forall a \in [K]\}$.*

F.1. Unstructured Bandits

Proposition 5 (Bound on Gradient norm and curvature of f_j for Unstructured Gaussian Bandits)**.** *Let $c_{\min} \in \left(0, \frac{1}{\min\{K,d\}}\right)$, $\boldsymbol{\omega} \in \mathring{\Delta}_K$, and $\boldsymbol{\omega}' \in \triangle_K$. Then, there exists a constant $D \geq 0$, we have for a fixed correct answer $\boldsymbol{p} \in \mathcal{E}(\Pi(\boldsymbol{\nu}, \epsilon))$:*

1. The upper bound on the L_∞-norm of the gradient is given by:

$$\boxed{\|\nabla_{\boldsymbol{\omega}} f_j(\boldsymbol{\omega}, \boldsymbol{p})\|_\infty \leq dD\,,}$$

2. Upper bound on the curvature constant is given by:

$$\boxed{C_{f_j(\boldsymbol{\omega},\boldsymbol{p})} \leq \sup_{\alpha \in (0,1]} \frac{8Dd}{\alpha c_{\min}}.}$$

Proof. Following Theorem 6 we recall that the objective $F(\boldsymbol{\omega}, \boldsymbol{p}) = \min_{j \in \mathcal{J}_{\boldsymbol{p}}} f_j(\boldsymbol{\omega}, \boldsymbol{p})$, where $|\mathcal{J}_{\boldsymbol{p}}| = d + 1$,

$$f_0(\boldsymbol{\omega}, \boldsymbol{p}) = \min_{\boldsymbol{q} \in \triangle_K} \min_{\substack{\boldsymbol{\mu}' : \boldsymbol{\mu}'^\top (\boldsymbol{q} - \boldsymbol{p}) > \epsilon \\ B' : B'\boldsymbol{q} \leq 0}} \sum_{a=1}^{K} \boldsymbol{\omega}_a \mathrm{KL}\left(\boldsymbol{\mu}_a, A_a \parallel \boldsymbol{\mu}'_a, B'_a\right),$$

and

$$f_j(\boldsymbol{\omega}, \boldsymbol{p}) = \min_{B'_j : B'_j \boldsymbol{p} = 0} \sum_{a=1}^{K} \boldsymbol{\omega}_a \mathrm{KL}\left(A_a \parallel B'_a\right), \forall j \in [d].$$

Upper bound on gradient norm. Leveraging Proposition 1-2 the derivative of $f_0(\boldsymbol{\omega}, \boldsymbol{p})$ with respect to any allocation $\boldsymbol{\omega} \in \mathring{\Delta}_K$ yields:

$$\nabla_{\boldsymbol{\omega}} f_0(\boldsymbol{\omega}, \boldsymbol{p}) = \sum_{a=1}^{K} \mathrm{KL}\left(\boldsymbol{\mu}_a, A_a \,\middle\|\, \boldsymbol{\mu}'^*_a, B'^*_a\right) e_a\,,$$

where e_a is a K-dimensional vector with entry 1 at a-th coordinate, and 0 elsewhere. Additionally, we have similar expression for the derivative for $f_j(\boldsymbol{\omega}, \boldsymbol{p})$ for all $j \in [d]$:

$$\nabla_{\boldsymbol{\omega}} f_j(\boldsymbol{\omega}, \boldsymbol{p}) = \sum_{a=1}^{K} \mathrm{KL}\left(A_a \,\middle\|\, B'^*_a\right) e_a$$

Thus, the L_∞ norm of the gradient

$$\|\nabla_{\boldsymbol{\omega}} f_0(\boldsymbol{\omega}, \boldsymbol{p})\|_\infty = \max_{a \in [K]} \mathrm{KL}\left(\boldsymbol{\mu}_a, A_a \,\middle\|\, \boldsymbol{\mu}'^*_a, B'^*_a\right), \text{ and } \|\nabla_{\boldsymbol{\omega}} f_j(\boldsymbol{\omega}, \boldsymbol{p})\|_\infty = \max_{a \in [K]} \mathrm{KL}\left(A_a \,\middle\|\, B'^*_a\right)$$

For Gaussian reward signal and multivariate Gaussian cost signal we know that $\mathrm{KL}\left(\boldsymbol{\mu}_a, A_a \,\middle\|\, \boldsymbol{\mu}'^*_a, B'^*_a\right) = \mathcal{O}(d)$. As the parameter domains $\mathcal{D}_{\boldsymbol{\mu}}$ and $\mathcal{D}_A$ are bounded as per Assumption 2, thus $\exists D_1, D_2 \geq 0$ such that

$$\|\nabla_{\boldsymbol{\omega}} f_0(\boldsymbol{\omega}, \boldsymbol{p})\|_\infty \leq dD_1, \text{ and } \|\nabla_{\boldsymbol{\omega}} f_j(\boldsymbol{\omega}, \boldsymbol{p})\|_\infty \leq dD_2$$

In general, we conclude for any $j \in \mathcal{J}_{\boldsymbol{p}}$:

$$\boxed{\|\nabla_{\boldsymbol{\omega}} f_j(\boldsymbol{\omega}, \boldsymbol{p})\|_\infty \leq dD\,,}$$

where $D \triangleq \max\{D_1, D_2\}$.

Upper bound on curvature. We apply chain rule on the expression for gradient:

$$\left(\nabla^2_{\boldsymbol{\omega}} f_0(\boldsymbol{\omega}, \boldsymbol{p})\right)_{a,a'} = \nabla_{(\boldsymbol{\mu'}^*_a, B'^*_a)} \mathrm{KL}\left(\boldsymbol{\mu}_a, A_a \,\|\, \boldsymbol{\mu'}^*_a, B'^*_a\right) \nabla_{\boldsymbol{\omega}_{a'}}(\boldsymbol{\mu'}^*_a, B'^*_a)$$
$$\left(\nabla^2_{\boldsymbol{\omega}} f_j(\boldsymbol{\omega}, \boldsymbol{p})\right)_{a,a'} = \nabla_{B'^*_a} \mathrm{KL}\left(A_a \,\|\, B'^*_a\right) \nabla_{\boldsymbol{\omega}_{a'}}(B'^*_a), \forall j \in [d].$$

Again, for Gaussian reward signal and multivariate Gaussian cost signal we observe that $\nabla_{\boldsymbol{\omega}_{a'}}(\boldsymbol{\mu'}^*_a, B'^*_a) = \mathcal{O}(\frac{1}{\min_{a\in[K]} \boldsymbol{\omega}_a})$. Then following Definition 5 there exists constants $M_1, M_2 > 0$ such that $\left(\nabla^2_{\boldsymbol{\omega}} f_0(\boldsymbol{\omega}, \boldsymbol{p})\right)_{a,a'} \leq \frac{dM_1}{c_{\min}}$, and $\left(\nabla^2_{\boldsymbol{\omega}} f_j(\boldsymbol{\omega}, \boldsymbol{p})\right)_{a,a'} \leq \frac{dM_2}{c_{\min}}, \forall j \in [d]$. Thus, finally we have for $M \triangleq \max\{M_1, M_2\}$:

$$\left(\nabla^2_{\boldsymbol{\omega}} f_j(\boldsymbol{\omega}, \boldsymbol{p})\right)_{a,a'} \leq \frac{d^2 M}{c_{\min}}$$
$$\implies \|\nabla^2_{\boldsymbol{\omega}} f_j(\boldsymbol{\omega}, \boldsymbol{p})\|_\infty \leq \sum_{a'=1}^{K} \frac{d^2 M}{c_{\min}} \leq \frac{MKd^2}{c_{\min}}, \forall j \in [d].$$

Bound on curvature constant. From the upper bound on the gradient norm we say that $f_j(\boldsymbol{\omega}, \boldsymbol{p})$ is dD-smooth. That means for any allocation pair $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \mathring{\Delta}_K$ we define $\boldsymbol{\omega}'' \triangleq \boldsymbol{\omega} + \alpha(\boldsymbol{\omega}' - \boldsymbol{\omega})$. Since $\mathring{\Delta}_K$ is a compact set, we write using Definition 7-

$$\begin{aligned}
C_{f_j(\boldsymbol{\omega},\boldsymbol{p})}(\mathring{\Delta}_K) &= \sup_{\alpha,\boldsymbol{\omega},\boldsymbol{\omega}'} \frac{2}{\alpha^2}\left[f_j(\boldsymbol{\omega}, \boldsymbol{p}) - f_j(\boldsymbol{\omega}'', \boldsymbol{p}) + \langle \boldsymbol{\omega} - \boldsymbol{\omega}'', \nabla_{\boldsymbol{\omega}} f_j(\boldsymbol{\omega}, \boldsymbol{p})\rangle\right] \\
&\leq \sup_{\alpha,\boldsymbol{\omega},\boldsymbol{\omega}'} \frac{2}{\alpha^2}\left[f_j(\boldsymbol{\omega}, \boldsymbol{p}) - (1-\alpha) f_j(\boldsymbol{\omega}, \boldsymbol{p}) + \alpha f_j(\boldsymbol{\omega}', \boldsymbol{p}) + \|\nabla_{\boldsymbol{\omega}} f_j(\boldsymbol{\omega}, \boldsymbol{p})\|_\infty \|\boldsymbol{\omega}'' - \boldsymbol{\omega}\|_1\right] \\
&\leq \sup_{\alpha,\boldsymbol{\omega},\boldsymbol{\omega}'} \frac{2}{\alpha^2}\left[\alpha \frac{Dd}{c_{\min}} \|\boldsymbol{\omega}' - \boldsymbol{\omega}\|_1 + \alpha \frac{Dd}{c_{\min}} \|\boldsymbol{\omega}' - \boldsymbol{\omega}\|_1\right] \\
&\leq \sup_{\alpha,\boldsymbol{\omega},\boldsymbol{\omega}'} \frac{4Dd}{\alpha c_{\min}} \|\boldsymbol{\omega}' - \boldsymbol{\omega}\|_1
\end{aligned}$$

We know as $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \mathring{\Delta}_K$, $\|\boldsymbol{\omega}' - \boldsymbol{\omega}\|_1 \leq 2$. Therefore, we have the final upper bound on the curvature constant as

$$\boxed{C_{f_j(\boldsymbol{\omega},\boldsymbol{p})} \leq \sup_{\alpha\in(0,1]} \frac{8Dd}{\alpha c_{\min}}.}$$

□

Corollary 5 (Curvature bound inside $\mathring{\Delta}_K$)**.** *Let $c_{\min} \in \left(0, \frac{1}{\min\{K,d\}}\right)$, $\boldsymbol{\omega} \in \mathring{\Delta}_K$, and $\boldsymbol{\omega}' \in \triangle_K$. Then we write for a given correct answer $\boldsymbol{p} \in \mathcal{E}(\Pi(\boldsymbol{\nu}, \epsilon))$:*

$$f_j(\boldsymbol{\omega}, \boldsymbol{p}) + \langle \boldsymbol{\omega}'' - \boldsymbol{\omega}, \nabla f_j(\boldsymbol{\omega}, \boldsymbol{p})\rangle - f_j(\boldsymbol{\omega}'', \boldsymbol{p}) \leq \frac{8Dd\alpha^2}{c_{\min}},$$

where $\boldsymbol{\omega}'' = \boldsymbol{\omega} + \alpha(\boldsymbol{\omega}' - \boldsymbol{\omega})$, and $\alpha = (0, \frac{1}{2}]$.

Proof. We first define a $\tilde{\boldsymbol{\omega}} \triangleq \frac{\boldsymbol{\omega}+\boldsymbol{\omega}'}{2}$. As all the allocation lies on the same line we have $\boldsymbol{\omega}'' = 2\alpha\tilde{\boldsymbol{\omega}} + (1-2\alpha)\boldsymbol{\omega}$. It is clear from the definition that $\boldsymbol{\omega}, \tilde{\boldsymbol{\omega}} \in \Delta_{\frac{c_{\min}}{2}}$. Thus, we have:

$$f_j(\boldsymbol{\omega}, \boldsymbol{p}) + \langle \boldsymbol{\omega}'' - \boldsymbol{\omega}, \nabla f_j(\boldsymbol{\omega}, \boldsymbol{p})\rangle - f_j(\boldsymbol{\omega}'', \boldsymbol{p}) \leq \frac{2 \times Dd(2\alpha)^2}{c_{\min}} \leq \frac{8Dd\alpha^2}{c_{\min}},$$

Hence, proved. □

F.2. Linear Bandits

We can retain similar results in case of linear bandits i.e., upper bounds on gradient norm and curvature constant.

Analogous to Theorem 6, for linear bandits we write–

Theorem 7. *Consider the ϵ-optimal pure exploration problem in linear bandits under unknown linear constraints. Fix any correct answer $\boldsymbol{p} \in \mathcal{E}(\Pi(\boldsymbol{\nu}, \epsilon))$. Then,*

1. *The set of instances $\boldsymbol{\nu}$ for which $\boldsymbol{p}$ remains ϵ-correct is relatively open in $(\mathcal{D}_{\boldsymbol{\mu}}, \mathcal{D}_A)$ (Lemma 5).*
2. *The alternative set $\Lambda_M^r(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon)$ in Definition 4 decomposes as a finite union of convex sets:*

$$\Lambda_M^r(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon) = \underbrace{\Lambda_M^r(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon)}_{1 \text{ convex set}} \bigcup \underbrace{\bigcup_{j=1}^{d} \Lambda_M^c(\boldsymbol{\nu}, \boldsymbol{p}, \epsilon, j)}_{d \text{ convex sets}},$$

a union of $d+1$ convex sets.

Proposition 6 (Bound on Gradient norm and curvature of f_j for Linear Bandits)**.** *Let $c_{\min} \in \left(0, \frac{1}{\min\{K,d\}}\right)$, $\boldsymbol{\omega} \in \mathring{\Delta}_K$, and $\boldsymbol{\omega}' \in \triangle_K$. Then, there exists a constant $D \geq 0$, we have for a fixed correct answer $\boldsymbol{p} \in \mathcal{E}(\Pi(\boldsymbol{\nu}, \epsilon))$:*

1. *The upper bound on the L_∞-norm of the gradient is given by:*

$$\boxed{\|\nabla_{\boldsymbol{\omega}} f_j(\boldsymbol{\omega}, \boldsymbol{p})\|_\infty \leq d D_{\text{lin}} \lambda_{\max}\,,}$$

2. *Upper bound on the curvature constant is given by:*

$$\boxed{C_{f_j(\boldsymbol{\omega}, \boldsymbol{p})} \leq \sup_{\alpha \in (0,1]} \frac{8 D_{\text{lin}} d \lambda_{\max}}{\alpha c_{\min}}.},$$

where $\lambda_{\max}$ is the largest Eigen value of the design matrix $V(\boldsymbol{\omega})$.

Proof. Following Theorem 6 we recall that the objective $F(\boldsymbol{\omega}, \boldsymbol{p}) = \min_{j \in \mathcal{J}_{\boldsymbol{p}}} f_j(\boldsymbol{\omega}, \boldsymbol{p})$, where $|\mathcal{J}_{\boldsymbol{p}}| = d+1$,

$$f_0(\boldsymbol{\omega}, \boldsymbol{p}) = \min_{\boldsymbol{q} \in \triangle_K} \min_{\substack{\boldsymbol{\theta}' : \boldsymbol{\theta}'^\top \mathbb{X}(\boldsymbol{q} - \boldsymbol{p}) > \epsilon \\ B' : B' \mathbb{X} \boldsymbol{q} \leq 0}} \frac{\|\varepsilon_{\boldsymbol{\theta}} - \Sigma_{rc}^\top \Sigma_c^{-1/2} \varepsilon_B\|_{V(\boldsymbol{\omega})}^2}{S} + \sum_{i=1}^{d} \|\varepsilon_B^i\|_{V(\boldsymbol{\omega})}^2 + \sum_{\substack{i,j \in [d] \\ i \neq j}} (\varepsilon_B^i)^\top V(\boldsymbol{\omega}) \varepsilon_B^j,$$

and

$$f_j(\boldsymbol{\omega}, \boldsymbol{p}) = \min_{B_j' : B_j'^\top \mathbb{X} \boldsymbol{p} > 0} \sum_{i=1}^{d} \frac{\|\varepsilon_B^i\|_{V(\boldsymbol{\omega})}^2}{S_B}, \forall j \in [d],$$

where $S_B \triangleq \text{Var}(c \mid r) = \Sigma_c - \frac{1}{\sigma_r^2} \Sigma_{rc} \Sigma_{rc}^\top$.

Upper bound on gradient norm. Following the similar arguments as unstructured bandits, we conclude that $\exists D_{\text{lin}} \geq 0$ such that

$$\boxed{\|\nabla_{\boldsymbol{\omega}} f_j(\boldsymbol{\omega}, \boldsymbol{p})\|_\infty \leq d D_{\text{lin}} \lambda_{\max}\,,}$$

where $\lambda_{\max}$ is the largest Eigen value of the design matrix $V(\boldsymbol{\omega})$.

Bound on Curvature. $\exists M_{\text{lin}} \geq 0$ such that

$$\left(\nabla^2_{\boldsymbol{\omega}} f_j(\boldsymbol{\omega}, \boldsymbol{p})\right)_{\mathbf{x},\mathbf{x}'} \leq \frac{d^2 M_{\text{lin}} \lambda^2_{\max}}{c_{\min}}$$
$$\implies \|\nabla^2_{\boldsymbol{\omega}} f_j(\boldsymbol{\omega}, \boldsymbol{p})\|_\infty \leq \sum_{\mathbf{x}' \in \mathcal{X}} \frac{M_{\text{lin}} d^2 \lambda^2_{\max}}{c_{\min}} \leq \frac{M_{\text{lin}} K d^2 \lambda^2_{\max}}{c_{\min}}, \forall j \in [d].$$

Bound on Curvature constant. From the upper bound on the gradient norm we say that $f_j(\boldsymbol{\omega}, \boldsymbol{p})$ is $dD_{\text{lin}}\lambda_{\max}$-smooth. That means for any allocation pair $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \mathring{\Delta}_K$ we define $\boldsymbol{\omega}'' \triangleq \boldsymbol{\omega} + \alpha(\boldsymbol{\omega}' - \boldsymbol{\omega})$. Since $\mathring{\Delta}_K$ is a compact set, we write using Definition 7-

$$\begin{aligned}
C_{f_j(\boldsymbol{\omega},\boldsymbol{p})}(\mathring{\Delta}_K) &= \sup_{\alpha,\boldsymbol{\omega},\boldsymbol{\omega}'} \frac{2}{\alpha^2} \left[f_j(\boldsymbol{\omega}, \boldsymbol{p}) - f_j(\boldsymbol{\omega}'', \boldsymbol{p}) + \langle \boldsymbol{\omega} - \boldsymbol{\omega}'', \nabla_{\boldsymbol{\omega}} f_j(\boldsymbol{\omega}, \boldsymbol{p})\rangle\right] \\
&\leq \sup_{\alpha,\boldsymbol{\omega},\boldsymbol{\omega}'} \frac{2}{\alpha^2} \left[f_j(\boldsymbol{\omega}, \boldsymbol{p}) - (1-\alpha) f_j(\boldsymbol{\omega}, \boldsymbol{p}) + \alpha f_j(\boldsymbol{\omega}', \boldsymbol{p}) + \|\nabla_{\boldsymbol{\omega}} f_j(\boldsymbol{\omega}, \boldsymbol{p})\|_\infty \|\boldsymbol{\omega}'' - \boldsymbol{\omega}\|_1\right] \\
&\leq \sup_{\alpha,\boldsymbol{\omega},\boldsymbol{\omega}'} \frac{2}{\alpha^2} \left[\alpha \frac{D_{\text{lin}} d \lambda_{\max}}{c_{\min}} \|\boldsymbol{\omega}' - \boldsymbol{\omega}\|_1 + \alpha \frac{D_{\text{lin}} d \lambda_{\max}}{c_{\min}} \|\boldsymbol{\omega}' - \boldsymbol{\omega}\|_1\right] \\
&\leq \sup_{\alpha,\boldsymbol{\omega},\boldsymbol{\omega}'} \frac{4 D_{\text{lin}} d \lambda_{\max}}{\alpha c_{\min}} \|\boldsymbol{\omega}' - \boldsymbol{\omega}\|_1
\end{aligned}$$

We know as $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \mathring{\Delta}_K$, $\|\boldsymbol{\omega}' - \boldsymbol{\omega}\|_1 \leq 2$. Therefore, we have the final upper bound on the curvature constant as

$$\boxed{C_{f_j(\boldsymbol{\omega},\boldsymbol{p})} \leq \sup_{\alpha \in (0,1]} \frac{8 D_{\text{lin}} d \lambda_{\max}}{\alpha c_{\min}}.}$$

□

Corollary 6 (Curvature bound inside $\mathring{\Delta}_K$ for Linear Bandits)**.** *Let* $c_{\min} \in \left(0, \frac{1}{\min\{K,d\}}\right)$, $\boldsymbol{\omega} \in \mathring{\Delta}_K$, *and* $\boldsymbol{\omega}' \in \triangle_K$. *Then we write for a given correct answer* $\boldsymbol{p} \in \mathcal{E}(\Pi(\boldsymbol{\nu}, \epsilon))$*:*

$$f_j(\boldsymbol{\omega}, \boldsymbol{p}) + \langle \boldsymbol{\omega}'' - \boldsymbol{\omega}, \nabla f_j(\boldsymbol{\omega}, \boldsymbol{p})\rangle - f_j(\boldsymbol{\omega}'', \boldsymbol{p}) \leq \frac{8 D_{\text{lin}} d \lambda_{\max} \alpha^2}{c_{\min}},$$

where $\boldsymbol{\omega}'' = \boldsymbol{\omega} + \alpha(\boldsymbol{\omega}' - \boldsymbol{\omega})$, *and* $\alpha = (0, \frac{1}{2}]$.

Proof. We first define a $\tilde{\boldsymbol{\omega}} \triangleq \frac{\boldsymbol{\omega}+\boldsymbol{\omega}'}{2}$. As all the allocation lies on the same line we have $\boldsymbol{\omega}'' = 2\alpha\tilde{\boldsymbol{\omega}} + (1-2\alpha)\boldsymbol{\omega}$. It is clear from the definition that $\boldsymbol{\omega}, \tilde{\boldsymbol{\omega}} \in \Delta_{\frac{c_{\min}}{2}}$. Thus, we have:

$$f_j(\boldsymbol{\omega}, \boldsymbol{p}) + \langle \boldsymbol{\omega}'' - \boldsymbol{\omega}, \nabla f_j(\boldsymbol{\omega}, \boldsymbol{p})\rangle - f_j(\boldsymbol{\omega}'', \boldsymbol{p}) \leq \frac{2 \times D_{\text{lin}} d \lambda_{\max} (2\alpha)^2}{c_{\min}} \leq \frac{8 D_{\text{lin}} d \lambda_{\max} \alpha^2}{c_{\min}},$$

Hence, proved. □

G. Stopping Criterion

Lemma 2 (Stopping and Correctness)**.** *The Chernoff stopping criterion for any δ-correct ϵ-optimal pure exploration algorithm with d linear constraints for a candidate policy $\hat{\boldsymbol{\pi}}_t = \arg\max_{\hat{B}_t\boldsymbol{\pi}\leq 0} \hat{\boldsymbol{\mu}}^\top \boldsymbol{\pi}$, is (a) Multivariate Gaussian bandits:*

$$\min\left\{\inf_{\boldsymbol{\lambda}\in\Lambda_M^r(\hat{\boldsymbol{\nu}}_t,\hat{\boldsymbol{\pi}}_t,\epsilon)}\sum_{a=1}^K N_{a,t}\mathrm{KL}\left(\hat{\boldsymbol{\nu}}_{a,t}\parallel\boldsymbol{\lambda}_a\right), \inf_{\boldsymbol{\lambda}\in\Lambda_M^c(\hat{\boldsymbol{\nu}}_t,\hat{\boldsymbol{\pi}}_t,\epsilon)}\sum_{a=1}^K N_{a,t}\mathrm{KL}\left(\hat{\boldsymbol{\nu}}_{a,t}\parallel\boldsymbol{\lambda}_a\right)\right\} > \beta(t,\delta)$$

(b) Linear bandits:

$$\inf_{(\boldsymbol{\theta}',B')\in\Lambda(\boldsymbol{\nu},\boldsymbol{p},\epsilon)} \frac{\|\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}'-\Sigma_{rc}^\top\Sigma_c^{-1}(\hat{B}_t-B')\|^2_{\sum_{\mathbf{x}\in\mathcal{X}}N_{\mathbf{x},t}\mathbf{x}\mathbf{x}^\top}}{\sigma_r^2-\Sigma_{rc}\Sigma_c^{-1}\Sigma_{rc}} + \sum_{i,j=1}^d e_i^\top\Sigma_c^{-1}(\hat{B}_t-B')\left(\sum_{\mathbf{x}\in\mathcal{X}}N_{\mathbf{x},t}\mathbf{x}\mathbf{x}^\top\right)(\hat{B}_t-B')^\top\Sigma_c^{-1}e_j > 2\beta(t,\delta),$$

where $\beta(t,\delta) = 2(d+1)\sum_{a=1}^K\log\left(4+\log\left(N_{a,t}\right)\right) + K\mathcal{T}\left(\frac{\log\left(\frac{1}{\delta}\right)}{K}\right)$ and $\mathcal{T}(x)\approx x+\log x$.

Proof of Part (a). **Step 1: Finiteness of τ_δ.** We claim that $\tau_\delta<\infty$, if the parameters in the model converges in finite time. Specifically, for PRUNE , we say it stops when the good events defined in Definition 6 holds with certainty, and additionally the allocation $\boldsymbol{\omega}_t$ has converged to the optimal allocation $\boldsymbol{\omega}^\star(\boldsymbol{\nu})$. While finiteness of the first event follows directly from tracking results in Appendix J (For $t>4K$, each arm has been played enough number of times for $\hat{\boldsymbol{\nu}}_t$ to $\boldsymbol{\nu}$; finiteness of the second event is due to convergence guaranties of the Frank-Wolfe iterates proven in Appendix I.

Step 2: Correctness of stopping threshold. We start by defining the 'incorrect' event when we stop at a finite time but do not yield a correct recommendation, i.e.

$$G_t = \left\{\tau_\delta<\infty\wedge\hat{\boldsymbol{\pi}}_{\tau_\delta}\notin\Pi(\boldsymbol{\nu},\epsilon)\right\}.$$

Thus,

$$\begin{aligned}
\mathbb{P}(G_t) = \quad & \mathbb{P}\Big(\tau_\delta<\infty\wedge\hat{\boldsymbol{\pi}}_{\tau_\delta}\notin\Pi(\boldsymbol{\nu},\epsilon)\Big)\\
&\leq \mathbb{P}\Big(\exists t\in\mathbb{N}:\mathrm{stat}_t>\beta(t,\delta)\wedge\hat{\boldsymbol{\pi}}_t\notin\Pi(\boldsymbol{\nu},\epsilon)\Big)\\
&= \mathbb{P}\Big(\exists t\in\mathbb{N},\boldsymbol{p}\notin\Pi(\boldsymbol{\nu},\epsilon):\mathrm{stat}_t>\beta(t,\delta)\wedge\hat{\boldsymbol{\pi}}_t=\boldsymbol{p}\Big)\\
&= \mathbb{P}\Big(\exists t\in\mathbb{N},\boldsymbol{p}\notin\Pi(\boldsymbol{\nu},\epsilon):\mathrm{stat}_t(\boldsymbol{p})>\beta(t,\delta)\Big)\\
&= \mathbb{P}\Big(\exists t\in\mathbb{N},\boldsymbol{p}\notin\Pi(\boldsymbol{\nu},\epsilon):\min\Big\{\inf_{\boldsymbol{\lambda}\in\Lambda_M^r(\hat{\boldsymbol{\nu}}_t,\boldsymbol{p},\epsilon)}\sum_{a=1}^K N_{a,t}\mathrm{KL}\left(\hat{\boldsymbol{\nu}}_{a,t}\parallel\boldsymbol{\lambda}_a\right),\\
&\qquad\qquad \inf_{\boldsymbol{\lambda}\in\Lambda_M^c(\hat{\boldsymbol{\nu}}_t,\boldsymbol{p},\epsilon)}\sum_{a=1}^K N_{a,t}\mathrm{KL}\left(\hat{\boldsymbol{\nu}}_{a,t}\parallel\boldsymbol{\lambda}_a\right)\Big\}>\beta(t,\delta)\Big)\\
&\leq \mathbb{P}\left(\exists t\in\mathbb{N},\boldsymbol{p}\notin\Pi(\boldsymbol{\nu},\epsilon):\sum_{a=1}^K N_{a,t}\mathrm{KL}\left(\hat{\boldsymbol{\nu}}_{a,t}\parallel\boldsymbol{\nu}_a\right)>\beta(t,\delta)\right) \qquad (45)\\
&= \mathbb{P}\left(\exists t\in\mathbb{N},\boldsymbol{p}\notin\Pi(\boldsymbol{\nu},\epsilon):\sum_{j=1}^{d+1}\sum_{a=1}^K N_{a,t}\mathrm{KL}\left(\bar{\boldsymbol{Y}}_{t,j}\,\middle\|\,(\Sigma^{-1/2}\boldsymbol{\nu}_a)_j\right)>\beta(t,\delta)\right) \qquad (46)
\end{aligned}$$

The last equality is true because of the following reasoning. We have collected $n=N_{a,t}$ i.i.d samples of a random variable $X_i\sim\mathcal{N}((\boldsymbol{\mu}_a,B_a),\Sigma)$ for any arm i. We define $Y_i=\Sigma^{-1/2}X_i$. Thus, $Y_i\sim\mathcal{N}(\Sigma^{-1/2}(\boldsymbol{\mu}_a,B_a),\mathbf{I}_d)$. We further define $Y_{i,j}\sim\mathcal{N}((\Sigma^{-1/2}\boldsymbol{\nu}_a)_j,1)$, for $j\in[d+1]$. The sample mean $\hat{Y}_n=\frac{1}{t}\sum_{i=1}^n Y_i$. Then, we can express the sum of KL-divergence as:

$$\sum_{j=1}^{d+1} \mathrm{KL}\left(\bar{\boldsymbol{Y}}_{n,j} \,\middle\|\, (\Sigma^{-1/2}\boldsymbol{\nu}_a)_j\right) = \frac{1}{2}\sum_{j=1}^{d+1} \|\bar{\boldsymbol{Y}}_n - (\Sigma^{-1/2}\boldsymbol{\nu}_a)_j\|_2^2 = \frac{1}{2}\|\hat{\boldsymbol{\nu}}_{a,n} - \boldsymbol{\nu}_a\|_{\Sigma^{-1}}^2$$

Now, we set

$$\beta(t,\delta) = 2(d+1)\sum_{a=1}^{K} \log\left(4 + \log\left(N_{a,t}\right)\right) + K\mathcal{T}\left(\frac{\log\left(\frac{1}{\delta}\right)}{K}\right)$$

Thus, by Lemma 16, we get

$$\mathbb{P}\Big(\tau_\delta < \infty \wedge \hat{\boldsymbol{\pi}}_{\tau_\delta} \notin \Pi(\boldsymbol{\nu},\epsilon)\Big) \leq \delta\,. \tag{47}$$

□

Proof of Part (b). **Step 1: Finiteness of τ_δ.** This follows similar argument as the unstructured bandit case.

Step 2: Correctness of stopping threshold. We start by defining the 'incorrect' event when we stop at a finite time but do not yield a correct recommendation, i.e.

$$G_t = \{\tau_\delta < \infty \wedge \hat{\boldsymbol{\pi}}_{\tau_\delta} \notin \Pi(\boldsymbol{\nu},\epsilon)\}\,.$$

Let us define $\hat{\boldsymbol{\nu}}_{\mathbf{x},t} = (\hat{\boldsymbol{\theta}}_t^\top, \hat{B}_t)\mathbf{x}$, $\hat{\varepsilon}_{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}'$, and $\hat{\varepsilon}_B = \Sigma_c^{-1/2}(\hat{B}_t - B')$.

Then, we get

$$\begin{aligned}
\mathbb{P}(G_t) = \quad & \mathbb{P}\Big(\tau_\delta < \infty \wedge \hat{\boldsymbol{\pi}}_{\tau_\delta} \notin \Pi(\boldsymbol{\nu},\epsilon)\Big) \\
&\leq \mathbb{P}\Big(\exists t \in \mathbb{N} : \mathrm{stat}_t > \beta(t,\delta) \wedge \hat{\boldsymbol{\pi}}_t \notin \Pi(\boldsymbol{\nu},\epsilon)\Big) \\
&= \mathbb{P}\left(\exists t \in \mathbb{N}, \boldsymbol{p} \notin \Pi(\boldsymbol{\nu},\epsilon) : \inf_{(\boldsymbol{\theta}'_r, B') \in \Lambda(\boldsymbol{\nu},\boldsymbol{p},\epsilon)} \frac{\|\hat{\varepsilon}_{\boldsymbol{\theta}} - \Sigma_{rc}^\top \Sigma_c^{-1/2}\hat{\varepsilon}_B\|^2_{\sum_{\mathbf{x}\in\mathcal{X}} N_{\mathbf{x},t}\mathbf{x}\mathbf{x}^\top}}{2(\sigma_r^2 - \Sigma_{rc}\Sigma_c^{-1}\Sigma_{rc})}\right. \\
&\qquad\qquad \left. + \frac{1}{2}\sum_{i,j=1}^{d} e_i^\top \hat{\varepsilon}_B \left(\sum_{\mathbf{x}\in\mathcal{X}} N_{\mathbf{x},t}\mathbf{x}\mathbf{x}^\top\right)\hat{\varepsilon}_B^\top e_j > \beta(t,\delta)\right) \\
&\leq \mathbb{P}\left(\exists t \in \mathbb{N}, \boldsymbol{p} \notin \Pi(\boldsymbol{\nu},\epsilon) : \sum_{\mathbf{x}\in\mathcal{X}} N_{\mathbf{x},t}\mathrm{KL}\left(\hat{\boldsymbol{\nu}}_{\mathbf{x},t} \,\|\, \boldsymbol{\nu}_{\mathbf{x}}\right) > \beta(t,\delta)\right) \\
&= \mathbb{P}\left(\exists t \in \mathbb{N}, \boldsymbol{p} \notin \Pi(\boldsymbol{\nu},\epsilon) : \sum_{j=1}^{d+1}\sum_{\mathbf{x}\in\mathcal{X}} N_{\mathbf{x},t}\mathrm{KL}\left(\bar{\boldsymbol{Y}}_{t,j} \,\middle\|\, (\Sigma^{-1/2}\boldsymbol{\nu}_{\mathbf{x}})_j\right) > \beta(t,\delta)\right)
\end{aligned} \tag{48}$$

The last equality is true because of the following reasoning. We have collected $n = N_{\mathbf{x},t}$ i.i.d samples of a random variable $X_i \sim \mathcal{N}((\boldsymbol{\theta}^\top\mathbf{x}, B\mathbf{x}), \Sigma)$ for any arm i. We define $Y_i = \Sigma_c^{-1/2}X_i$. Thus, $Y_i \sim \mathcal{N}(\Sigma^{-1/2}(\boldsymbol{\theta}^\top\mathbf{x}, B\mathbf{x}), \mathbf{I}_d)$. We further define $Y_{i,j} \sim \mathcal{N}((\Sigma^{-1/2}\boldsymbol{\nu}_{\mathbf{x}})_j, 1)$, for $j \in [d+1]$. The sample mean $\hat{Y}_n = \frac{1}{t}\sum_{i=1}^n Y_i$. Then, we can express the sum of KL-divergence as:

$$\sum_{j=1}^{d+1} \mathrm{KL}\left(\bar{\boldsymbol{Y}}_{n,j} \,\middle\|\, (\Sigma^{-1/2}\boldsymbol{\nu}_{\mathbf{x}})_j\right) = \frac{1}{2}\sum_{j=1}^{d+1} \|\bar{\boldsymbol{Y}}_n - (\Sigma^{-1/2}\boldsymbol{\nu}_{\mathbf{x}})_j\|_2^2 = \frac{1}{2}\|\hat{\boldsymbol{\nu}}_{\mathbf{x},n} - \boldsymbol{\nu}_{\mathbf{x}}\|_{\Sigma^{-1}}^2$$

Now, we set

$$\beta(t,\delta) = 2(d+1)\sum_{\mathbf{x}\in\mathcal{X}} \log\left(4 + \log\left(N_{\mathbf{x},t}\right)\right) + K\mathcal{T}\left(\frac{\log\left(\frac{1}{\delta}\right)}{K}\right)$$

Thus, by Lemma 16, we get

$$\mathbb{P}\Big(\tau_\delta < \infty \wedge \hat{\boldsymbol{\pi}}_{\tau_\delta} \notin \Pi(\boldsymbol{\nu},\epsilon)\Big) \leq \delta\,. \tag{49}$$

□

H. Concentration of Probability Bad Events

Lemma 6. *Under the good events defined in Definition 6,*

$$\mathbb{P}\left((\mathcal{G}_{1,\varepsilon}(T)\cap\mathcal{G}_{2,\varepsilon}(T))^c\right)\leq\infty$$

Proof. We first remind the definition of the good events for $\varepsilon\geq 0$:

$$\mathcal{G}_{1,\varepsilon}^t\triangleq\left\{\max_{\mathbf{x}\in\triangle_K}\min_{\mathbf{h}\in H_{F(\cdot)}(\boldsymbol{\omega}_{t-1},r_t)}\langle\mathbf{x}-\boldsymbol{\omega}_{t-1},\mathbf{h}\rangle-\varepsilon<\min_{\mathbf{h}\in H_{F(\cdot)}(\boldsymbol{\omega}_{t-1},r_t)}\langle\mathbf{x}_t-\boldsymbol{\omega}_{t-1},\mathbf{h}\rangle\right\}$$
$$\mathcal{G}_{2,\varepsilon}^t\triangleq\left\{\hat{\boldsymbol{\nu}}_t\in S_{\boldsymbol{p}}\bigwedge\mid F(\boldsymbol{\omega}\mid\hat{\boldsymbol{\nu}}_t)-F(\boldsymbol{\omega}\mid\boldsymbol{\nu})\mid<\varepsilon,\forall\boldsymbol{\omega}\in\mathring{\Delta}_K\right\}$$

We have $\mathcal{G}_1^t\subset\{\|\boldsymbol{\nu}-\hat{\boldsymbol{\nu}}_{t-1}\|_{\infty,\infty}\leq\xi_{1,\varepsilon}\}$ and also $\mathcal{G}_2^t\subset\{\|\boldsymbol{\nu}-\hat{\boldsymbol{\nu}}_t\|_{\infty,\infty}\leq\xi_{2,\varepsilon}\}$, if we apply Lemma 9 and Corollary 12 respectively.

Then we have

$$\mathbb{P}\left((\mathcal{G}_{1,\varepsilon}(T)\cap\mathcal{G}_{2,\varepsilon}(T))^c\right)\leq\sum_{\underline{\mathbf{h}}(T)}^{T}\mathbb{P}(\|\boldsymbol{\nu}-\hat{\boldsymbol{\nu}}_t\|_{\infty,\infty}>\xi_\varepsilon)$$

where $\xi_\varepsilon=\max(\xi_{1,\varepsilon},\xi_{2,\varepsilon})$. Now $\|\boldsymbol{\nu}-\hat{\boldsymbol{\nu}}_t\|_{\infty,\infty}=\max_{a\in[K]}\|\boldsymbol{\nu}_a-\hat{\boldsymbol{\nu}}_{a,t}\|_\infty$. Therefore

$$\begin{aligned}\mathbb{P}\Big((\mathcal{G}_{1,\varepsilon}(T)\cap\mathcal{G}_{2,\varepsilon}(T))^c\Big)&\leq\sum_{\underline{\mathbf{h}}(T)}^{T}\mathbb{P}(\max_{a\in[K]}\|\boldsymbol{\nu}_a-\hat{\boldsymbol{\nu}}_{a,t}\|_\infty>\xi_\varepsilon)\\&\leq\sum_{\underline{\mathbf{h}}(T)}^{T}\sum_{a=1}^{K}\mathbb{P}(\|\boldsymbol{\nu}_a-\hat{\boldsymbol{\nu}}_{a,t}\|_\infty>\xi_\varepsilon)\\&\leq\sum_{\underline{\mathbf{h}}(T)}^{T}\sum_{a=1}^{K}\sum_{j=1}^{d+1}\mathbb{P}(|\boldsymbol{\nu}_{a,j}-\hat{\boldsymbol{\nu}}_{a,j,t}|>\xi_\varepsilon)\end{aligned}$$

From tracking results (Lemma 12) we have at any time $t>4K$ and $\forall a\in[K]$, $N_{a,t}\geq\sqrt{\frac{t}{K}}-K$, ensuring that each arm, is played enough till time $t\in[T]$ to bring $\hat{\boldsymbol{\nu}}_{a,t}$ close to $\boldsymbol{\nu}_a$. Then applying Chernoff inequality on the probability on the R.H.S of the last inequality, we get

$$\mathbb{P}(|\boldsymbol{\nu}_{a,j}-\hat{\boldsymbol{\nu}}_{a,j,t}|>\xi_\varepsilon))\leq e^{Kd}\left[\exp(-\sqrt{t}A_{a,j}^-)+\exp(-\sqrt{t}A_{a,j}^+)\right]$$

where $A_{a,j}^-=\frac{\mathrm{KL}(\boldsymbol{\nu}_{a,j}-\xi_\varepsilon\parallel\boldsymbol{\nu}_{a,j})}{\sqrt{Kd}}$ and $A_k^+=\frac{\mathrm{KL}(\boldsymbol{\nu}_{a,j}+\xi_\epsilon\parallel\boldsymbol{\nu}_{a,j})}{\sqrt{Kd}}$. We leverage this upper bound the bad event to get

$$\begin{aligned}&\mathbb{P}\Big((\mathcal{G}_{1,\varepsilon}(T)\cap\mathcal{G}_{2,\varepsilon}(T))^c\Big)\leq e^{Kd}\sum_{\underline{\mathbf{h}}(T)}^{T}\sum_{a=1}^{K}\sum_{j=1}^{d+1}\left[\exp(-\sqrt{t}A_{a,j}^-)+\exp(-\sqrt{t}A_{a,j}^+)\right]<\infty\\\implies&\sum_{T=N}^{\infty}\mathbb{P}\Big((\mathcal{G}_{1,\varepsilon}(T)\cap\mathcal{G}_{2,\varepsilon}(T))^c\Big)<34e^{Kd}\sum_{a=1}^{K}\sum_{j=1}^{d+1}\frac{1}{\mathrm{KL}\left(\boldsymbol{\nu}_{a,j}-\xi_\varepsilon\parallel\boldsymbol{\nu}_{a,j}\right)^{19/4}}\\&\qquad+\frac{1}{\mathrm{KL}\left(\boldsymbol{\nu}_{a,j}+\xi_\varepsilon\parallel\boldsymbol{\nu}_{a,j}\right)^{19/4}}\end{aligned}\tag{50}$$

Where the last line is implied by leveraging Lemma 18 with $\alpha=\frac{8}{11}$, $\beta=\frac{1}{2}$ and $A=A_{a,j}^+=A_{a,j}^-$ respectively. □

Corollary 7. *For any $\boldsymbol{\nu} \in S_{\boldsymbol{p}}$, $\kappa \triangleq F(\boldsymbol{\omega} \mid \boldsymbol{\nu})$ 1. Unstructured bandits. $\varepsilon < (0, \kappa D d)$, if $\|\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}\|_{\infty,\infty} \leq \frac{\varepsilon}{2Dd}$, and 2. Linear bandits. $\varepsilon < (0, \kappa D_{\text{lin}} d \lambda_{\max})$, if $\|\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}\|_{\infty,\infty} \leq \frac{\varepsilon}{2D_{\text{lin}} d \lambda_{\max}}$, then:*

$$|F(\boldsymbol{\omega}|\boldsymbol{\nu}) - F(\boldsymbol{\omega}|\hat{\boldsymbol{\nu}})| < \varepsilon, \ \ \forall \boldsymbol{\omega} \in \mathring{\Delta}_K$$

Lemma 7. *Let $\boldsymbol{\nu} \in S_{\boldsymbol{p}}$ and $\varepsilon \in (0, \kappa D)$. For any $r \in (0,1)$, and $\boldsymbol{\omega} \in \mathring{\Delta}_K$, if $\hat{\boldsymbol{\nu}} \in S_{\boldsymbol{p}}$ that is*

1. Unstructured Bandits. $\|\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}\|_{\infty,\infty} \leq \frac{\varepsilon}{2Dd}$, and 2. Linear Bandits. $\|\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}\|_{\infty,\infty} \leq \frac{\varepsilon}{2D_{\text{lin}} d \lambda_{\max}}$, then we get

$$\left| \max_{\mathbf{x} \in \triangle_K} \min_{\mathbf{h} \in H_{F(\cdot)}(\boldsymbol{\omega}_{t-1}, r_t)} \langle \mathbf{x} - \boldsymbol{\omega}, \mathbf{h} \rangle - \max_{\mathbf{x} \in \triangle_K} \min_{\mathbf{h} \in H_{F(\cdot)}(\boldsymbol{\omega}, r)} \langle \mathbf{x} - \boldsymbol{\omega}, \mathbf{h} \rangle \right| < \frac{\varepsilon}{2}, \forall (\boldsymbol{\omega}, r) \in \mathring{\Delta}_K \times (0,1)$$

and

$$\left| \min_{\mathbf{h} \in H_{F(\cdot)}(\boldsymbol{\omega}_{t-1}, r_t)} \langle \mathbf{x} - \boldsymbol{\omega}, \mathbf{h} \rangle - \min_{\mathbf{h} \in H_{F(\cdot)}(\boldsymbol{\omega}, r)} \langle \mathbf{x} - \boldsymbol{\omega}, \mathbf{h} \rangle \right| < \frac{\varepsilon}{2}, \forall (\boldsymbol{\omega}, r, \mathbf{x}) \in \mathring{\Delta}_K \times (0,1) \times \triangle_K$$

Therefore putting $\xi_\varepsilon = \frac{\kappa}{2Dd}$ for unstructured bandits and $\xi_\varepsilon = \frac{\kappa}{2D_{\text{lin}} d \lambda_{\max}}$ for linear bandits in Equation (50) to get the final non-asymptotic bound on the bad event probabilities:

1. *Unstructured Bandits.*

$$\sum_{T=N}^{\infty} \mathbb{P}\Big((\mathcal{G}_{1,\varepsilon}(T) \cap \mathcal{G}_{2,\varepsilon}(T))^c \Big) < 34 e^{Kd} \sum_{a=1}^{K} \sum_{j=1}^{d+1} \frac{1}{\text{KL}\left(\boldsymbol{\nu}_{a,j} - \frac{\kappa}{2Dd} \,\middle\|\, \boldsymbol{\nu}_{a,j}\right)^{19/4}} + \frac{1}{\text{KL}\left(\boldsymbol{\nu}_{a,j} + \frac{\kappa}{2Dd} \,\middle\|\, \boldsymbol{\nu}_{a,j}\right)^{19/4}}$$

, and

2. *Linear Bandits.*

$$\sum_{T=N}^{\infty} \mathbb{P}\Big((\mathcal{G}_{1,\varepsilon}(T) \cap \mathcal{G}_{2,\varepsilon}(T))^c \Big) < 34 e^{Kd} \sum_{a=1}^{K} \sum_{j=1}^{d+1} \frac{1}{\text{KL}\left(\boldsymbol{\nu}_{a,j} - \frac{\kappa}{2D_{\text{lin}} d \lambda_{\max}} \,\middle\|\, \boldsymbol{\nu}_{a,j}\right)^{19/4}} + \frac{1}{\text{KL}\left(\boldsymbol{\nu}_{a,j} + \frac{\kappa}{2D_{\text{lin}} d \lambda_{\max}} \,\middle\|\, \boldsymbol{\nu}_{a,j}\right)^{19/4}}$$

I. Convergence of Frank-Wolfe Optimiser for Allocations

In this section, we present detailed proofs for convergence guaranty of Frank-Wolfe sampling algorithm. Note that the convergence of the FW iterates in Equation (9) is setting agnostic, meaning the proof of convergence does not assume any specific structure, but focuses only on the optimisation problem for identifying an SOP solution. The adaptation for linear bandits specifically comes in through the upper bounds on gradient norm and curvature constant.

I.1. Properties of the Allocation Function

Thus, using Proposition 4 we write

$$F(\boldsymbol{\omega}) = \sup_{\boldsymbol{p}\in\Pi(\boldsymbol{\nu},\epsilon)} F(\boldsymbol{\omega},\boldsymbol{p}) = \max_{\boldsymbol{p}\in\mathcal{P}(\boldsymbol{\omega})} F(\boldsymbol{\omega},\boldsymbol{p})$$

Sub-differential set. Thus, Theorem 6 and Lemma 5 yields $|\mathcal{J}_{\boldsymbol{p}}| = 1 + d$ sub-functions:

$$f_0(\boldsymbol{\omega},\boldsymbol{p}) \triangleq F_1(\boldsymbol{\omega},\boldsymbol{p}), \qquad f_j(\boldsymbol{\omega},\boldsymbol{p}) := F_{2,j}(\boldsymbol{\omega},\boldsymbol{p}), \quad j \in [d].$$

Hence $F(\boldsymbol{\omega},\boldsymbol{p}) = \min_{j\in\mathcal{J}_{\boldsymbol{p}}} f_j(\boldsymbol{\omega},\boldsymbol{p})$, and $F(\boldsymbol{\omega}) = \max_{\boldsymbol{p}\in\mathcal{P}(\boldsymbol{\omega})} F(\boldsymbol{\omega},\boldsymbol{p})$ using Proposition 4. Then, the r-subdifferential subspace used in the Frank–Wolfe update is defined as

$$H_{F(\cdot)}(\boldsymbol{\omega},r) = \mathrm{conv}\Big\{\nabla_{\boldsymbol{\omega}} f_j(\boldsymbol{\omega},\boldsymbol{p}) : \boldsymbol{p}\in\mathcal{P}(\boldsymbol{\omega}), j\in\mathcal{J}_{\boldsymbol{p}},\ f_j(\boldsymbol{\omega},\boldsymbol{p}) > F(\boldsymbol{\omega},\boldsymbol{p}) - r\Big\}.$$

Frank-Wolfe Update rule. The FW update rule for the allocation strategy is given by:

$$\mathbf{x}_{t+1} \triangleq \arg\max_{\mathbf{x}\in\triangle_K} \min_{\mathbf{h}\in H_{F(\cdot)}(\boldsymbol{\omega}_{t-1},r_t)} \langle \mathbf{x} - \boldsymbol{\omega}_t, \mathbf{h}\rangle, \quad \boldsymbol{\omega}_{t+1} \triangleq \frac{t}{t+1}\boldsymbol{\omega}_t + \frac{1}{t+1}\mathbf{x}_{t+1}\,.$$

Corollary 8 (Curvature of $F(\boldsymbol{\omega},\boldsymbol{p})$)**.** *Let $c_{\min} \in \left(0, \frac{1}{\min\{K,d\}}\right)$, $r \in (0,1)$, and $\boldsymbol{\omega} \in \mathring{\Delta}_K$, $\boldsymbol{\omega}' \in \triangle_K$. For any correct answer $\boldsymbol{p} \in \mathcal{E}(\Pi(\boldsymbol{\nu},\epsilon))$, 1. if $0 < \alpha < \min\left\{\frac{1}{2}, \frac{r}{dD}\right\}$, then for unstructured bandits with multivariate Gaussian reward-cost:*

$$F(\boldsymbol{\omega}'',\boldsymbol{p}) \geq F(\boldsymbol{\omega},\boldsymbol{p}) + \alpha \min_{\mathbf{h}\in H_{F(\cdot)}(\boldsymbol{\omega},r)} \langle \boldsymbol{\omega}' - \boldsymbol{\omega}, \mathbf{h}\rangle - \frac{8Dd\alpha^2}{c_{\min}}\,,$$

and 2. if $0 < \alpha < \min\left\{\frac{1}{2}, \frac{r}{dD_{\mathrm{lin}}\lambda_{\max}}\right\}$, then for linear bandits:

$$F(\boldsymbol{\omega}'',\boldsymbol{p}) \geq F(\boldsymbol{\omega},\boldsymbol{p}) + \alpha \min_{\mathbf{h}\in H_{F(\cdot)}(\boldsymbol{\omega},r)} \langle \boldsymbol{\omega}' - \boldsymbol{\omega}, \mathbf{h}\rangle - \frac{8D_{\mathrm{lin}}d\lambda_{\max}\alpha^2}{c_{\min}}$$

for $\boldsymbol{\omega}'' = (1-\alpha)\boldsymbol{\omega} + \alpha\boldsymbol{\omega}'$.

Proof. **Case I.** The result holds trivially in case $F(\boldsymbol{\omega},\boldsymbol{p}) = f_j(\boldsymbol{\omega},\boldsymbol{p}) = f_j(\boldsymbol{\omega}'',\boldsymbol{p}) = F(\boldsymbol{\omega}'',\boldsymbol{p})$i.e., both the functions $F(\boldsymbol{\omega},\boldsymbol{p})$ and $F(\boldsymbol{\omega}'',\boldsymbol{p})$ are minimised for the same $j \in \mathcal{J}_{\boldsymbol{p}}$ (using Proposition 5).

Case II. If there exists $j_1, j_2 \in \mathcal{J}_{\boldsymbol{p}}$ such that $F(\boldsymbol{\omega},\boldsymbol{p}) = f_{j_1}(\boldsymbol{\omega},\boldsymbol{p}) < f_{j_2}(\boldsymbol{\omega}',\boldsymbol{p})$ then we use the fact that $f_j(\cdot)$ is dD-Lipschitz in $\triangle_K$ as per Proposition 5. Thus, we write $\|\boldsymbol{\omega} - \boldsymbol{\omega}''\|_\infty \leq \alpha\|\boldsymbol{\omega}'\|_\infty < \min\left\{\frac{1}{2}, \frac{r}{dD}\right\}$. We also know that as from the definition of $H_{F(\cdot)}(\boldsymbol{\omega},r)$ that $f_{j_2}(\boldsymbol{\omega},\boldsymbol{p}) < F(\boldsymbol{\omega},\boldsymbol{p}) + r$. Thus, choosing $h = \nabla f_{j_2}(\boldsymbol{\omega})$, and using Corollary 5 we have:

$$\begin{aligned}
&F(\boldsymbol{\omega},\boldsymbol{p}) + \langle \boldsymbol{\omega}'' - \boldsymbol{\omega}, \mathbf{h}\rangle - F(\boldsymbol{\omega}'',\boldsymbol{p}) \leq f_{j_2}(\boldsymbol{\omega}) + \langle \boldsymbol{\omega}'' - \boldsymbol{\omega}, \mathbf{h}\rangle - f_{j_2}(\boldsymbol{\omega}'') \leq \frac{8Dd\alpha^2}{c_{\min}} \\
\implies &F(\boldsymbol{\omega}'',\boldsymbol{p}) \geq F(\boldsymbol{\omega},\boldsymbol{p}) - \langle \boldsymbol{\omega}'' - \boldsymbol{\omega}, \mathbf{h}\rangle + \frac{8Dd\alpha^2}{c_{\min}} \\
\implies &F(\boldsymbol{\omega}'',\boldsymbol{p}) \geq F(\boldsymbol{\omega},\boldsymbol{p}) - \alpha\langle \boldsymbol{\omega}' - \boldsymbol{\omega}, \mathbf{h}\rangle + \frac{8Dd\alpha^2}{c_{\min}}
\end{aligned}$$

$$\implies F(\boldsymbol{\omega}'', \boldsymbol{p}) \geq F(\boldsymbol{\omega}, \boldsymbol{p}) - \alpha \min_{\mathbf{h} \in H_{F(\cdot)}(\boldsymbol{\omega}, r)} \langle \boldsymbol{\omega}' - \boldsymbol{\omega}, \mathbf{h} \rangle + \frac{8Dd\alpha^2}{c_{\min}}$$

$$\implies F(\boldsymbol{\omega}'', \boldsymbol{p}) \geq F(\boldsymbol{\omega}, \boldsymbol{p}) - \min\left\{\frac{1}{2}, \frac{r}{dD}\right\} \min_{\mathbf{h} \in H_{F(\cdot)}(\boldsymbol{\omega}, r)} \langle \boldsymbol{\omega}' - \boldsymbol{\omega}, \mathbf{h} \rangle + \frac{8Dd\alpha^2}{c_{\min}}.$$

Hence, proved. The proof for the second part in linear bandits flows similarly using Proposition 6, and Corollary 6. □

Proposition 7 (Bound on Gradient norm and curvature of $F(\boldsymbol{\omega})$)**.** *1. $F(\boldsymbol{\omega})$ is dD-Lipschitz on $\triangle_K$.*

2. Let $c_{\min} \in \left(0, \frac{1}{\min\{K,d\}}\right)$, $r \in (0,1)$, and $\boldsymbol{\omega} \in \mathring{\Delta}_K$, $\boldsymbol{\omega}' \in \triangle_K$.

(a) If $0 < \alpha < \min\left\{\frac{1}{2}, \frac{r}{dD}\right\}$, then for unstructured bandits with multivariate Gaussian reward-cost:

$$F(\boldsymbol{\omega}'') \geq F(\boldsymbol{\omega}) + \alpha \min_{\mathbf{h} \in H_{F(\cdot)}(\boldsymbol{\omega}, r)} \langle \boldsymbol{\omega}' - \boldsymbol{\omega}, \mathbf{h} \rangle - \frac{8Dd}{\alpha c_{\min}},$$

and (b) if $0 < \alpha < \min\left\{\frac{1}{2}, \frac{r}{dD_{\text{lin}}\lambda_{\max}}\right\}$, then for Linear bandits:

$$F(\boldsymbol{\omega}'') \geq F(\boldsymbol{\omega}) + \alpha \min_{\mathbf{h} \in H_{F(\cdot)}(\boldsymbol{\omega}, r)} \langle \boldsymbol{\omega}' - \boldsymbol{\omega}, \mathbf{h} \rangle - \frac{8D_{\text{lin}}\lambda_{\max} d}{\alpha c_{\min}},$$

for $\boldsymbol{\omega}'' = (1-\alpha)\boldsymbol{\omega} + \alpha\boldsymbol{\omega}'$

Proof. Proof of this Proposition directly follows from the proof of Corollary 8 and 4 using monotonicity of $\max$ over all $\boldsymbol{p} \in \mathcal{E}(\Pi(\boldsymbol{\nu}, \epsilon))$. □

I.2. Continuity of the Frank-Wolfe Iterates: Sub-Differentials

Now we prove properties of the subdifferential set that handles the non-smoothness of $f_j(\boldsymbol{\omega} \mid \boldsymbol{\nu})$.

Corollary 9. *The mapping $\boldsymbol{\omega} \mapsto H_{F(\cdot)}(\boldsymbol{\omega}, r)$ is continuous.*

Proof. Let $\{\boldsymbol{\omega}_t, \hat{\boldsymbol{\nu}}_t, r_t\}_{t=1}^{\infty} \longrightarrow \{\boldsymbol{\omega}^\star(\boldsymbol{\nu}), \boldsymbol{\nu}, r\} \in \triangle_K \times S_{\boldsymbol{p}} \times (0,1)$ and $H_{F(\cdot)}(\boldsymbol{\omega}, r) = \text{ch}\left\{\nabla_{\boldsymbol{\omega}} f_j(\boldsymbol{\omega}, \boldsymbol{p} \mid \hat{\boldsymbol{\nu}})_{j=1}^d\right\}$ for some $\boldsymbol{p} \in \mathcal{P}(\boldsymbol{\omega})$. We can write for any $h \in H_{F(\cdot)}(\boldsymbol{\omega}, r), \exists$ a sequence of $\{\alpha_i \geq 0\}_{i=1}^d$ such that h can be expressed as a convex combination of $\{\nabla_{\boldsymbol{\omega}} f_j(\boldsymbol{\omega}, \boldsymbol{p} \mid \hat{\boldsymbol{\nu}})\}_{j=1}^d$. Further, leveraging continuity of $f_j(\cdot \mid \hat{\boldsymbol{\nu}})$ for all $j = [d]$(ref. Appendix E), we claim that $\nabla_{\boldsymbol{\omega}} f_j(\boldsymbol{\omega}_t, \boldsymbol{p} \mid \hat{\boldsymbol{\nu}}_t) \in H_{F(\cdot)}(\boldsymbol{\omega}_t, r_t \mid \hat{\boldsymbol{\nu}}_t)$ for some $t > N$.

Lower Hemicontinuity of subdifferential. Lower Hemicontinuity follows from the continuity result of $\nabla_{\boldsymbol{\omega}} f_j$ derived in Appendix I.2.

Upper Hemicontinuity of subdifferentials. We adapt similar proof structure as (Wang et al., 2021a) by adding a $\tilde{\epsilon}$-radius (Minkowski addition) to $H_{\hat{\boldsymbol{\nu}}_t}(\boldsymbol{\omega}_t, r_t)$ to show it is still contained by the open set containing $H_{\boldsymbol{\nu}}(\boldsymbol{\omega}^\star(\boldsymbol{\nu}), r)$. Thus, the subdifferential set is lower and upper hemicontinuous with respect to $\boldsymbol{\omega}$, i.e continuous. Hence, proved. □

Now, we prove the convergence of the Frank-Wolfe iterates. Let us remind the Frank-Wolfe steps once again for brevity,

$$\mathbf{x}_{t+1} \triangleq \arg\max_{\mathbf{x} \in \triangle_K} \min_{\mathbf{h} \in H_{F(\cdot)}(\boldsymbol{\omega}, r)} \langle \mathbf{x} - \boldsymbol{\omega}_t, \mathbf{h} \rangle \tag{51}$$

$$\boldsymbol{\omega}_{t+1} \triangleq \frac{t}{t+1}\boldsymbol{\omega}_t + \frac{1}{t+1}\mathbf{x}_{t+1} \tag{52}$$

First, we define the following maps

1. $\phi_1 : (\boldsymbol{\omega}, r, \hat{\boldsymbol{\nu}}, \mathbf{x}) \mapsto \min_{\mathbf{h} \in H_{F(\cdot)}(\boldsymbol{\omega}_{t-1}, r_t)} \langle \mathbf{x} - \boldsymbol{\omega}, \mathbf{h} \rangle$,

2. $\phi_2 : (\boldsymbol{\omega}, r, \hat{\boldsymbol{\nu}}) \mapsto \max_{\mathbf{x}\in\triangle_K} \phi_1$.

Corollary 10. *ϕ_1 is continuous on $\mathring{\Delta}_K \times (0,1) \times S_i \times \triangle_K$.*

Proof. We again leverage Theorem 10 with the following definitions, $\{X\} \triangleq \mathring{\Delta}_K \times (0,1) \times S_{\boldsymbol{p}} \times \triangle_K$, $\{Y\} \triangleq \mathbb{R}^K$, $\Phi(\boldsymbol{\omega}, r, \hat{\boldsymbol{\nu}}, \mathbf{x}) \triangleq H_{F(\cdot)}(\boldsymbol{\omega}_{t-1}, r_t)$ and $u(\boldsymbol{\omega}, r, \hat{\boldsymbol{\nu}}, \mathbf{x}, h) = \langle \mathbf{x} - \boldsymbol{\omega}, \mathbf{h}\rangle$. We are concerned only about continuity of the correspondence Φ, as u is a linear function, hence continuous. Though we have already proven continuity of Φ in Corollary 9. Thus, we conclude the proof. □

Corollary 11. *ϕ_2 is continuous on $\mathring{\Delta}_K \times (0,1) \times S_{\boldsymbol{p}}$.*

Proof. We again apply Theorem 10 with the definitions $\{X\} \triangleq \mathring{\Delta}_K \times (0,1) \times S_{\boldsymbol{p}}$, $\{Y\} \triangleq \triangle_K$, $\Phi(\boldsymbol{\omega}, r, \hat{\boldsymbol{\nu}}, \mathbf{x}) \triangleq \triangle_K$ and $u(\boldsymbol{\omega}, r, \hat{\boldsymbol{\nu}}, \mathbf{x}, h) = \phi_1(\boldsymbol{\omega}, r, \hat{\boldsymbol{\nu}}, \mathbf{x})$. As Φ is a constant corresponding, we claim continuity of ϕ_2 and conclude the proof. □

I.3. Contraction of the Frank-Wolfe Iterates

We start by defining two "good events":

Definition 6 (Good events)**.** *For $\varepsilon \geq 0$ we define:*

$$\begin{aligned}
\mathcal{G}^t_{1,\varepsilon} &\triangleq \left\{ \max_{\mathbf{x}\in\triangle_K} \min_{\mathbf{h}\in H_{F(\cdot)}(\boldsymbol{\omega}_{t-1}, r_t)} \langle \mathbf{x} - \boldsymbol{\omega}_{t-1}, \mathbf{h}\rangle - \varepsilon < \min_{\mathbf{h}\in H_{F(\cdot)}(\boldsymbol{\omega}_{t-1}, r_t)} \langle \mathbf{x}_t - \boldsymbol{\omega}_{t-1}, \mathbf{h}\rangle \right\} \\
\mathcal{G}^t_{2,\varepsilon} &\triangleq \left\{ \hat{\boldsymbol{\nu}}_t \in S_{\boldsymbol{p}} \bigwedge \mid F(\boldsymbol{\omega} \mid \hat{\boldsymbol{\nu}}_t) - F(\boldsymbol{\omega} \mid \boldsymbol{\nu}) \mid < \varepsilon, \forall \boldsymbol{\omega} \in \mathring{\Delta}_K \right\}
\end{aligned} \tag{53}$$

We further define $\mathcal{G}_{1,\varepsilon}(T) \triangleq \bigcup_{t=\underline{h}(T)}^T \mathcal{G}_{1,\varepsilon}$, and $\mathcal{G}_{2,\varepsilon}(T) \triangleq \bigcup_{t=\underline{h}(T)}^T \mathcal{G}_{2,\varepsilon}$

We further define the gap between current objective at $t \in \mathbb{N}$ and optimal objective as $\Delta_t \triangleq F(\boldsymbol{\omega}^*) - F(\boldsymbol{\omega}_t)$.

Lemma 8 (Contraction of Δ_t)**.** *Let $t \in \mathbb{N}$ such that $\lfloor \frac{t}{K} \rfloor \notin \mathbb{N}$, and $t > 4K$. Then, under the good events $\mathcal{G}_{1,\varepsilon}(T) \bigcap \mathcal{G}_{2,\varepsilon}(T)$, the FW iterates in Equation (9) with $D < \frac{r_t t}{d}$ satisfies the following contraction:*

$$\Delta_t \leq (1-\alpha)\Delta_{t-1} + \alpha(r+\varepsilon) + 16Dd\alpha^2\sqrt{tK}$$

Proof. From Lemma 11 we observe that when $t > 4K$, $\mathbf{x}_{t-1} \in \Delta_{\frac{1}{2\sqrt{tK}}}$. We leverage Proposition 7 for $c_{\min} = \frac{1}{2\sqrt{tK}}$ to write:

$$\begin{aligned}
F(\boldsymbol{\omega}'') \geq& F(\boldsymbol{\omega}) + \alpha \min_{\mathbf{h}\in H_{F(\cdot)}(\boldsymbol{\omega}, r)} \langle \boldsymbol{\omega}' - \boldsymbol{\omega}, \mathbf{h}\rangle - 16Dd\alpha^2\sqrt{tK} \\
\geq& F(\boldsymbol{\omega}) + \alpha \left(\max_{\boldsymbol{\pi}\in\triangle_K} \min_{\mathbf{h}\in H_{F(\cdot)}(\boldsymbol{\omega}, r)} \langle \boldsymbol{\pi} - \boldsymbol{\omega}, \mathbf{h}\rangle - \varepsilon \right) - 16Dd\alpha^2\sqrt{tK}
\end{aligned}$$

$$\begin{aligned}
\max_{\boldsymbol{\pi}\in\triangle_K} \min_{\mathbf{h}\in H_{F(\cdot)}(\boldsymbol{\omega}, r)} \langle \boldsymbol{\pi} - \boldsymbol{\omega}, \mathbf{h}\rangle - \varepsilon \geq& \max_{\boldsymbol{\pi}\in\triangle_K} \min_{\mathbf{h}\in\partial_r F(\boldsymbol{\omega})} \langle \boldsymbol{\pi} - \boldsymbol{\omega}, \mathbf{h}\rangle - \varepsilon \\
\geq& \Delta_{t-1} - r - \varepsilon
\end{aligned}$$

Thus, we have:

$$\begin{aligned}
& F(\boldsymbol{\omega}'') \geq F(\boldsymbol{\omega}) + \alpha\left(\Delta_{t-1} - r - \varepsilon\right) - 16Dd\alpha^2\sqrt{tK} \\
\implies & \Delta_t \leq (1-\alpha)\Delta_{t-1} + \alpha(r+\varepsilon) + 16Dd\alpha^2\sqrt{tK}\,,
\end{aligned}$$

where the last inequality is got by subtracting $F(\boldsymbol{\omega}^*)$ from both side. The final result follows by setting $\alpha = \frac{1}{t}$. □

Lemma 9. *For any $\varepsilon > 0$, $\exists$ a constant $\xi_{1,\epsilon} > 0$, such that if $\|\hat{\boldsymbol{\nu}}_t - \boldsymbol{\nu}\|_{\infty,\infty} < \xi_{1,\epsilon}$, then for all $\boldsymbol{\nu} \in S_{\boldsymbol{p}}$:*

$$\left| \max_{\mathbf{x}\in\triangle_K} \min_{\mathbf{h}\in H_{F(\cdot)}(\boldsymbol{\omega}_{t-1},r_t)} \langle \mathbf{x}-\boldsymbol{\omega}, \mathbf{h}\rangle - \max_{\mathbf{x}\in\triangle_K} \min_{\mathbf{h}\in H_{F(\cdot)}(\boldsymbol{\omega},r)} \langle \mathbf{x}-\boldsymbol{\omega}, \mathbf{h}\rangle \right| < \frac{\varepsilon}{2}, \forall(\boldsymbol{\omega}, r) \in \mathring{\Delta}_K \times (0,1) \tag{54}$$

and

$$\left| \min_{\mathbf{h}\in H_{F(\cdot)}(\boldsymbol{\omega}_{t-1},r_t)} \langle \mathbf{x}-\boldsymbol{\omega}, \mathbf{h}\rangle - \min_{\mathbf{h}\in H_{F(\cdot)}(\boldsymbol{\omega},r)} \langle \mathbf{x}-\boldsymbol{\omega}, \mathbf{h}\rangle \right| < \frac{\varepsilon}{2}, \forall(\boldsymbol{\omega}, r, \mathbf{x}) \in \mathring{\Delta}_K \times (0,1) \times \triangle_K \tag{55}$$

Proof. **Proof of Eq. (54).** We define for $\boldsymbol{\omega} \in \mathring{\Delta}_K$ and $r \in (0,1)$, $\psi(\hat{\boldsymbol{\nu}}) \triangleq \min\{- \mid \phi_2(\boldsymbol{\omega}, r, \hat{\boldsymbol{\nu}}) - \phi_2(\boldsymbol{\omega}, r, \boldsymbol{\nu}) \mid\}$. If we apply Theorem 11 with the following definitions $\{X\} = S_i$, $\{Y\} = \mathring{\Delta}_K \times (0,1)$, $\Phi(\hat{\boldsymbol{\nu}}) = \mathring{\Delta}_K \times (0,1)$ and finally $u(\hat{\boldsymbol{\nu}}, \boldsymbol{\omega}, r) = - \mid \phi_2(\boldsymbol{\omega}, r, \hat{\boldsymbol{\nu}}) - \phi_2(\boldsymbol{\omega}, r, \boldsymbol{\nu}) \mid$, we can claim that $\psi(\hat{\boldsymbol{\nu}})$ is continuous on $\mathring{\Delta}_K \times (0,1)$. So, by definition of ψ, $\exists \xi_{1,\varepsilon} > 0$, such that $\psi(\hat{\boldsymbol{\nu}}) > -\frac{\varepsilon}{2}$ for all $\|\hat{\boldsymbol{\nu}} - \boldsymbol{\nu}\|_{\infty,\infty} < \xi_{1,\varepsilon}$. Hence, proved.

Proof of Eq. (55). This proof is analogous by leveraging Corollary 10.

□

Corollary 12. *For any $\varepsilon > 0$, $\exists$ a constant $\xi_{2,\varepsilon} > 0$, such that if $\|\hat{\boldsymbol{\nu}} - \boldsymbol{\nu}\|_{\infty,\infty} < \xi_{2,\varepsilon}$, then*

$$\hat{\boldsymbol{\nu}} \in S_i \text{ and } \left|F(\boldsymbol{\omega}|\boldsymbol{\nu}) - F(\boldsymbol{\omega}|\hat{\boldsymbol{\nu}})\right| < \varepsilon, \forall \boldsymbol{\omega} \in \mathring{\Delta}_K.$$

Proof. We define $\{X\} = S_{\boldsymbol{p}}$, $\{Y\} = \mathring{\Delta}_K$, $\Phi(\hat{\boldsymbol{\nu}}) = \mathring{\Delta}_K$ and $u(\hat{\boldsymbol{\nu}}, \boldsymbol{\omega}) = - \mid F(\boldsymbol{\omega} \mid \hat{\boldsymbol{\nu}}) - F(\boldsymbol{\omega} \mid \boldsymbol{\nu}) \mid$. We again leverage Theorem 11 to claim that $\psi(\hat{\boldsymbol{\nu}}) \triangleq \min_{\boldsymbol{\omega}\in\mathring{\Delta}_K} - \mid F(\boldsymbol{\omega} \mid \hat{\boldsymbol{\nu}}) - F(\boldsymbol{\omega} \mid \boldsymbol{\nu}) \mid$ is continuous on the open set $S_{\boldsymbol{p}}$. Also, $u(\hat{\boldsymbol{\nu}})$ is continuous due to continuity results in Appendix E. Then, by definition of ψ, $\exists \xi_{2,\varepsilon} > 0$, such that $\psi(\hat{\boldsymbol{\nu}}) > -\frac{\varepsilon}{2}$ for all $\|\hat{\boldsymbol{\nu}} - \boldsymbol{\nu}\|_{\infty,\infty} < \xi_{2,\varepsilon}$. Hence, proved. □

We follow the recursion proof structure from (Wang et al., 2021a, Theorem 7) to prove convergence of FW iterates with modified upper bounds on gradient norm and curvature constant to get:

Lemma 10. *Let $\{r_t\}_{t\geq 1}$ be a sequence of positive numbers with the properties $\lim_{t\to\infty} \frac{1}{t}\sum_{s=1}^{t} r_s = 0$ and $\lim_{t\to\infty} tr_t = \infty$. Then*

1. for unstructured bandits with multivariate Gaussian reward-cost: $T \geq \max\left\{\left(\frac{35Dd}{\epsilon}\right)^{11}, T_{\epsilon,D}^{\frac{11}{8}}, (4K+1)^{\frac{11}{8}}\right\}$, and

2. for linear bandits: $T \geq \max\left\{\left(\frac{35D_{\text{lin}}d\lambda_{\max}}{\epsilon}\right)^{11}, T_{\epsilon,D}^{\frac{11}{8}}, (4K+1)^{\frac{11}{8}}\right\}$,

where $\exists T_{\epsilon,D} \in \mathbb{N}$ such that if $t \geq T_{\epsilon,D}$ then $\sum_{s=1}^{t} r_s < \epsilon t$ and $tr_t > D$ for any $\epsilon \in (0,1)$.Then under the good events defined in Definition 6, we have

$$F(\boldsymbol{\omega}^\star(\boldsymbol{\nu})) - F(\boldsymbol{\omega}_t) \leq 5\epsilon, \forall t = \bar{h}(T), \bar{h}(T)+1, \ldots, T\,,$$

where $\bar{h}(T) = \min\{t \in \mathbb{N} : t \geq T^{2/11}\underline{h}(T), \sqrt{\frac{t}{K}} \in \mathbb{N}\}$, and $\underline{h}(T) = \min\{t \in \mathbb{N} : t \geq T^{8/11} + 2, \sqrt{\frac{t}{K}} \in \mathbb{N}\}$.

J. Tracking Results

We use similar as (Wang et al., 2021a) tracking results here conforming with the notations used in this paper. These results help us prove convergence of the Frank-Wolfe iterates (Equation (9)) in Appendix I.

Lemma 11. *Let $\{\mathbf{x}_s\}_{s\in\mathbb{N}} \in \triangle_K$ be a sequence of vectors such that $\mathbf{x}_k$ is e_k for $k = [1, K]$. Then*

$$t = K, N_{k,t} = 1,$$
$$\forall t \geq K+1, A_t \in \arg\max_{k'} \frac{\sum_{s=1}^{t} \mathbf{x}_{k',s}}{N_{k',t-1}} \forall k \in [1, K], \text{ where } N_{k,t} = \sum_{s=1}^{t} \mathbb{1}\{A_s = k\}$$

where the $\arg\max$ *breaks ties arbitrarily. Then for all $t \geq K$, and all $k \in [1, K]$*

$$\sum_{s=1}^{t} \mathbf{x}_{k,s} - (K-1) \leq N_{k,t} \leq \sum_{s=1}^{t} \mathbf{x}_{k,s} + 1$$

Lemma 12. *At any $t \geq 4K$, the sampling rule of* PRUNE *satisfies $\boldsymbol{\omega}_s \in \Delta_{\frac{1}{2\sqrt{tK}}}$*

Proof. By forced exploration enforced in PRUNE, we can write

$$t\boldsymbol{\omega}_t = \sum_{s=1}^{t} \mathbf{x}_s \geq \frac{1}{K}\mathbb{1}\{\mathbf{x}_s = (1/K, 1/K, \ldots, 1/K)\} = \sqrt{\left\lfloor \frac{t}{K} \right\rfloor} \geq \sqrt{\frac{t}{K}} - 1 \geq \frac{1}{2}\sqrt{\frac{t}{K}}$$
$$\implies \boldsymbol{\omega}_t \geq \frac{1}{2\sqrt{tK}}$$

□

K. Sample Complexity Analysis of PRUNE

We follow the same proof structure for deriving non-asymptotic sample complexity for PRUNE as (Wang et al., 2021b).

Lemma 13 (Non-asymptotic Sample Complexity)**.** *1. Unstructured Bandits. For any $\boldsymbol{\nu} \in (D_{\boldsymbol{\mu}}, D_B)$, $\varepsilon, \tilde{\varepsilon} \in (0,1)$ and $\delta \in (0,1)$, and given preference cone $\mathcal{C}$ stopping time of the algorithm PRUNE satisfies*

$$\begin{aligned}\mathbb{E}[\tau_\delta] \leq\ & 34e^{Kd}\sum_{a=1}^{K}\sum_{j=1}^{d+1}\frac{1}{\mathrm{KL}\left(\boldsymbol{\nu}_{a,j}-\frac{\kappa}{2Dd}\,\|\,\boldsymbol{\nu}_{a,j}\right)^{19/4}}+\frac{1}{\mathrm{KL}\left(\boldsymbol{\nu}_{a,j}+\frac{\kappa}{2Dd}\,\|\,\boldsymbol{\nu}_{a,j}\right)^{19/4}}\\ &+\left(\frac{35Dd}{\varepsilon}\right)^{11}+(4K+1)^{\frac{11}{8}}+\max\left\{\left(\frac{2}{\tilde{\varepsilon}}\right)^{11},c_1(D_{\boldsymbol{\mu}},D_B)\right\}\\ &+\frac{1+\tilde{\varepsilon}}{F(\boldsymbol{\omega}^\star(\boldsymbol{\nu})\mid\boldsymbol{\nu})-6\varepsilon}\left[\log\left(\frac{(1+\tilde{\varepsilon})c_2(D_{\boldsymbol{\mu}},D_B)e}{\delta(F(\boldsymbol{\omega}^\star(\boldsymbol{\nu})\mid\boldsymbol{\nu})-6\varepsilon)}\right)+\log\log\left(\frac{(1+\tilde{\varepsilon})c_2(D_{\boldsymbol{\mu}},D_B)}{\delta(F(\boldsymbol{\omega}^\star(\boldsymbol{\nu})\mid\boldsymbol{\nu})-6\varepsilon)}\right)\right]\end{aligned}$$

2. Linear Bandits. For any $\boldsymbol{\nu} \in (D_{\boldsymbol{\theta}}, D_B)$, $\varepsilon, \tilde{\varepsilon} \in (0,1)$ and $\delta \in (0,1)$, and given preference cone $\mathcal{C}$ stopping time of the algorithm PRUNE satisfies

$$\begin{aligned}\mathbb{E}[\tau_\delta] \leq\ & 34e^{Kd}\sum_{a=1}^{K}\sum_{j=1}^{d+1}\frac{1}{\mathrm{KL}\left(\boldsymbol{\nu}_{a,j}-\frac{\kappa}{2D_{\mathrm{lin}}d\lambda_{\max}}\,\Big\|\,\boldsymbol{\nu}_{a,j}\right)^{19/4}}+\frac{1}{\mathrm{KL}\left(\boldsymbol{\nu}_{a,j}+\frac{\kappa}{2D_{\mathrm{lin}}d\lambda_{\max}}\,\Big\|\,\boldsymbol{\nu}_{a,j}\right)^{19/4}}\\ &+\left(\frac{35D_{\mathrm{lin}}d\lambda_{\max}}{\varepsilon}\right)^{11}+(4K+1)^{\frac{11}{8}}+\max\left\{\left(\frac{2}{\tilde{\varepsilon}}\right)^{11},c_1(D_{\boldsymbol{\mu}},D_B)\right\}\\ &+\frac{1+\tilde{\varepsilon}}{F(\boldsymbol{\omega}^\star(\boldsymbol{\nu})\mid\boldsymbol{\nu})-6\varepsilon}\left[\log\left(\frac{(1+\tilde{\varepsilon})c_2(D_{\boldsymbol{\theta}},D_B)e}{\delta(F(\boldsymbol{\omega}^\star(\boldsymbol{\nu})\mid\boldsymbol{\nu})-6\varepsilon)}\right)+\log\log\left(\frac{(1+\tilde{\varepsilon})c_2(D_{\boldsymbol{\theta}},D_B)}{\delta(F(\boldsymbol{\omega}^\star(\boldsymbol{\nu})\mid\boldsymbol{\nu})-6\varepsilon)}\right)\right]\end{aligned}$$

Thus, the asymptotic sample complexity is then given by,

$$\limsup_{\delta\to 0}\frac{\mathbb{E}[\tau_\delta]}{\log(\frac{1}{\delta})}\leq\frac{1+\tilde{\varepsilon}}{F(\boldsymbol{\omega}^\star(\boldsymbol{\nu})\mid\boldsymbol{\nu})-6\varepsilon}$$

We achieve optimality as ε and $\tilde{\varepsilon}$ are arbitrarily small, that is $\limsup_{\delta\to 0}\frac{\mathbb{E}_{[\tau_\delta]}}{\log(\frac{1}{\delta})}\leq\frac{1}{F(\boldsymbol{\omega}^\star(\boldsymbol{\nu})|\boldsymbol{\nu})}=T^*_P(\boldsymbol{\nu},\epsilon)/T^*_{\mathrm{lin}}(\boldsymbol{\nu},\epsilon)$.

L. Computational Complexity of PRUNE

In this section, we provide a detailed analysis of the computational complexity of PRUNE for both unstructured and linear bandit settings. We state the complexity as per-round computational cost. Our focus is on the dominant operations arising from the allocation step, the construction of alternative instances, and the stopping rule evaluation.

Each round of PRUNE consists of the following main steps:

1. Construction of a *confusing alternative instance* (the $\boldsymbol{\lambda}$-player).
2. Computation of a subgradient (or sub-differential element) of the allocation objective.
3. Frank–Wolfe (FW) update (Equation (9)) for the allocation $\boldsymbol{\omega}_t$.
4. Arm selection via tracking.
5. Update of empirical estimates.
6. Evaluation of the stopping rule.

We analyse the cost of each component.

L.1. Unstructured Bandits

Confusing Instance Computation. The inner optimization in the lower bound involves solving nested problems of the form:

$$\min_{\boldsymbol{q}\in\triangle_K} \min_{\boldsymbol{\mu}',B'} \sum_{a=1}^{K} \boldsymbol{\omega}_a \mathrm{KL}\left(\boldsymbol{\mu}_a, B_a \parallel \boldsymbol{\mu}'_a, B'_a\right),$$

subject to linear constraints.

For common parametric families such as Gaussian distributions with known covariance, this reduces to a *convex quadratic program (QP)*. The dominant cost arises from:

- Optimization over $\boldsymbol{q} \in \triangle_K$: $O(K)$ variables with simplex constraints.
- Solving small convex programs per constraint.

Using standard interior-point methods, this step has worst-case complexity: $\mathcal{O}(K^3 + dK^2)$, but can be reduced to $O(K^2)$ per iteration using first-order methods due to the simplex structure.

Subgradient Computation. The subgradient involves evaluating sensitivity of the objective with respect to $\boldsymbol{\omega}$. This requires:

- Computing KL divergences for each arm: $\mathcal{O}(Kd)$.
- Aggregation across arms: $\mathcal{O}(K)$.

Thus, the total cost is $\mathcal{O}(Kd)$.

Frank-Wolfe Update. Each FW step requires solving a linear minimization oracle like $\arg\min_{e_a\in\triangle_K}\langle g_t, e_a\rangle$, which reduces to selecting the minimum coordinate of the gradient: $\mathcal{O}(K)$

Tracking and Arm Selection. The tracking rule requires computing: $\arg\min_{a\in[K]}\left(N_{a,t} - \sum_{s=1}^{t}\boldsymbol{\omega}_{s,a}\right)$, which costs $\mathcal{O}(K)$.

Stopping Rule Evaluation. The stopping rule involves evaluating:

$$\min_{\lambda\in\Lambda} \sum_{a=1}^{K} N_{a,t}\,\mathrm{KL}(\hat{\boldsymbol{\nu}}_{a,t}\|\boldsymbol{\lambda}_a),$$

which resembles the confusing instance computation. Hence, it incurs a similar cost:

$$\mathcal{O}(K^3) \text{ (worst-case)}, \quad \text{or } \mathcal{O}(K^2) \text{ with first-order methods.}$$

Overall Per-Round Complexity. Combining all components, the per-round complexity is: $\mathcal{O}(K^2 + Kd) = \mathcal{O}(K \max\{K, d\})$, when using efficient first-order solvers. The cubic dependence appears only with generic second-order methods.

L.2. Linear Bandits

In the linear setting, the complexity is governed by operations involving the feature dimension d_{lin}.

Confusing Instance Computation. The lower bound involves optimization over $(\boldsymbol{\theta}', B')$:

$$\min_{(\boldsymbol{\theta}', B') \in \Lambda} \| \cdot \|^2_{V(\boldsymbol{\omega})},$$

where $V(\boldsymbol{\omega}) = \sum_{a=1}^{K} \boldsymbol{\omega}_a \mathbf{x}_a \mathbf{x}_a^\top$. The Key operations are:

- Construction of $V(\boldsymbol{\omega})$: $\mathcal{O}(K d_{\text{lin}}^2)$.
- Matrix inversion $V(\boldsymbol{\omega})^{-1}$: $\mathcal{O}(d_{\text{lin}}^3)$.
- Quadratic form evaluations: $\mathcal{O}(d_{\text{lin}}^2)$.

Thus, the dominant cost is: $\mathcal{O}(K d_{\text{lin}}^2 + d_{\text{lin}}^3)$

Subgradient Computation. Computing gradients requires evaluating quadratic forms $\mathbf{x}_a^\top V(\boldsymbol{\omega})^{-1} \mathbf{x}_a$, for each arm a, costing: $\mathcal{O}(K d_{\text{lin}}^2)$.

Frank-Wolfe Update. Same as in the unstructured case: $\mathcal{O}(K)$

Updating the parameters using ridge regressor. Using Sherman-Morrison rank-one update formula to incrementally update the inverse of $V(\boldsymbol{\omega})$, and then multiplying two matrices of order $(d+1) \times d_{\text{lin}}$ and $d_{\text{lin}} \times d_{\text{lin}}$ jointly shows complexity of $\mathcal{O}(d d_{\text{lin}}^2)$.

Stopping Rule Evaluation. The stopping rule involves evaluating norms of the form:

$$\|\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}' - \Sigma_{rc}^\top \Sigma_c^{-1} (\hat{B}_t - B')\|^2_{V_t},$$

which requires:

- Matrix-vector products: $\mathcal{O}(d_{\text{lin}}^2)$.
- Possibly recomputing or updating V_t^{-1}.

With incremental updates (e.g., Sherman–Morrison), the inverse can be updated in: $\mathcal{O}(d_{\text{lin}}^2)$ per round.

Overall Per-Round Complexity. Thus, the per-round complexity in linear bandits is: $\mathcal{O}(K d_{\text{lin}}^2 + d_{\text{lin}}^3 + d d_{\text{lin}}^2) = \mathcal{O}(d_{\text{lin}}^2 \max\{K, d, d_{\text{lin}}\})$.

L.3. Discussions

- **Polynomial complexity:** In both settings, the computational complexity scales polynomially in K, d, and d_{lin}, avoiding exponential dependence on the number of constraints.
- **Scalability:** The algorithm is practical for moderate K and d_{lin}, especially when leveraging first-order methods and incremental matrix updates.
- **Comparison with prior work:** Unlike approaches that require solving combinatorial optimization problems over policies, PRUNE operates over the simplex, leading to tractable updates via Frank–Wolfe.
- **Bottlenecks:** The main computational bottleneck lies in solving the inner optimization problems (confusing instances) and matrix inversions in the linear case.

M. Auxiliary Definitions and Results

Theorem 8 (Berge's Maximum Theorem)**.** *Let $\mathcal{U}$ and $\mathcal{V}$ be topological spaces, $f : \mathcal{U} \times \mathcal{V} \to \mathbb{R}$ and $C : \mathcal{U} \to \mathcal{V}$ be non-empty compact set for all $u \in \mathcal{U}$. Then, if C is continuous at u, $f^*(u) = \max_{v \in C(u)} f(u, v)$ is continuous and $C^*(u) = \{v \in C(u) : f^*(u) = f(u, v)\}$ is upper-hemicontinuous.*

Theorem 9 (Donsker-Vardhan Variational Formula)**.** *For mutual information $KL(P\|Q)$, we have that:*

$$\mathrm{KL}\left(P \parallel Q\right) = \sup_f \mathbb{E}_P\left[f\right] - \log \mathbb{E}_Q\left[\exp(f)\right]$$

lower

Lemma 14 (Peskun Ordering)**.** *For any two random variables X, Y on $\mathbb{R}^{KL}$ the following are equivalent:*

1. *$X \leq_s Y$*
2. *For all $x \in \mathbb{R}^{KL}$, $\mathbb{P}\left[X \geq x\right] \leq \mathbb{P}\left[Y \geq x\right]$*
3. *For all non-negative functions $f_1, f_2, \ldots, f_k$, we have that: $\Pi_{i=1}^K f_i \leq \Pi_{i=1}^K f_i$*

Theorem 10 ((Berge, 1877))**.** *Let $\mathbb{X}$ and $\mathbb{Y}$ be Hausdorff topological spaces. Assume that*

1. *$\Phi : \mathbb{X} \rightrightarrows \mathbb{K}(\mathbb{X})$ is continuous, where $\mathbb{K}(\mathbb{X}) = \{F \in \mathbb{S}(\mathbb{X}) : F$ is compact.$\}$ (i.e. both upper and lower hemiconituous),*
2. *$u : \mathbb{X} \times \mathbb{Y} \to \mathbb{R}$ is continuous.*

Then the function $v : \mathbb{X} \to \mathbb{R}$ is continuous and the solution multifunction $\Phi^\star : \mathbb{X} \to \mathbb{S}(\mathbb{Y})$ is upper hemicontinuous and compact valued, where $\mathbb{S}(\mathbb{Y})$ is the set of non-empty subsets of $\mathbb{Y}$.

Theorem 11 ((Feinberg et al., 2014))**.** *Assume that*

1. *$\mathbb{X}$ is compactly generated,*
2. *$\Phi : \mathbb{X} \rightrightarrows \mathbb{S}(\mathbb{Y})$ is lower hemicontinuous,*
3. *$u : \mathbb{X} \times \mathbb{Y} \to \mathbb{R}$ is $\mathbb{K}$-inf-compact and upper semi-continuous on $\mathrm{Gr}_{\mathbb{X}}(\Phi)$.*

Then the function $v : \mathbb{X} \to \mathbb{R}$ is continuous and the solution multifunction $\Phi^* : \mathbb{X} \rightrightarrows \mathbb{S}(\mathbb{Y})$ is upper hemicontinuous and compact valued.

Lemma 15 ((Combes et al., 2017))**.** *Let $\mathbb{X}$ be a metric space and Y be a nonempty open subset in $\mathbb{R}^K$. Let $u : \mathbb{X} \times Y \to \mathbb{R}$ and $\frac{\partial u}{\partial y}$ exists and is continuous in $\mathbb{X} \times Y$. For each $y \in Y$, let $x^\star(y)$ minimises $u(x, y)$ over $x \in \mathbb{X}$. Set*

$$v(y) = u(x^\star(y), y).$$

Assume that $x^\star : Y \to \mathbb{X}$ is a continuous function. Then v is continuously differentiable and

$$\frac{d}{dy} v(y) = \frac{\partial u}{\partial y}(x^\star(y), y).$$

Fact 1 (Existence and Continuity of Minimum)**.** *If X and Y are topological spaces and Y is compact. Then for any continuous $f : X \times Y \to \mathbb{R}$, the function $g(x) \triangleq \inf_{y \in Y} f(x, y)$ is well-defined and continuous. Additionally, $\inf_{y \in Y} f(x, y) = \min_{y \in Y} f(x, y)$.*

Fact 2 (inf and min over Union of Sets)**.** *Let us consider an ordered universe S and a set $A \subset S$ which is union of $|I|$ sets, i.e. $A = \cup_{i \in I} A_i$ and $A_i \subset S$. If the following statements are true:*

1. *$a \triangleq \inf A$ exists, and*
2. *$a_i \triangleq \inf A_i$ exists for each $i \in I$,*

they imply that $a = \inf\{a_i : i \in I\}$. *The same holds true for* $\min$.

Lemma 16 (Theorem 9 in (Kaufmann & Koolen, 2021))**.** *Let* $\delta > 0, \nu$ *be independent one-parameter exponential families with mean* $\boldsymbol{\mu}$ *and* $S \subset [d]$. *Then we have,*

$$\mathbb{P}_\nu \left[\exists t \in \mathbb{N} : \sum_{a \in S} N_{a,t} d_{KL}\left(\mu_{a,t}, \mu_a\right) \geq \sum_{a \in S} 2 \log\left(4 + \log\left(N_{a,t}\right)\right) + |S| \mathcal{T}\left(\frac{\log\left(\frac{1}{\delta}\right)}{|S|} \right) \right] \leq \delta \,.$$

Here, $\mathcal{T} : \mathbb{R}^+ \to \mathbb{R}^+$ *is such that* $\mathcal{T}(x) = 2\tilde{h}_{3/2}\left(\frac{h^{-1}(1+x) + \log\left(\frac{\pi^2}{3}\right)}{2} \right)$ *with*

$$\forall u \geq 1, \quad h(u) = u - \log(u)$$
$$\forall z \in [1, e], \forall x \geq 0, \quad \tilde{h}_z(x) = \begin{cases} \exp\left(\frac{1}{h^{-1}(x)}\right) h^{-1}(x) & \text{if } x \geq h^{-1}\left(\frac{1}{\log(z)}\right) \\ z(x - \log(\log(z))) & \text{else} \end{cases} .$$

Lemma 17 (Proposition 3.1 in (Soare, 2015))**.** *Let* $N_{x_i}(t)$ *denote the number of draws of arm* x_i *up to round* t *and suppose* t *is the stooping time of an algorithm* $\mathcal{A}$. *Also, let* $\boldsymbol{\nu}$ *and* $\boldsymbol{\nu}'$ *be two bandit models and* A *an event such that* $0 < \mathbb{P}[A] < 1$. *Then,*

$$\sum_{i=1}^{K} \mathbb{E}_{\boldsymbol{\nu}}\left[N_{x_i}(t)\right] \mathrm{KL}\left(\boldsymbol{\nu} \parallel \boldsymbol{\nu}'\right) \geq \mathrm{KL}\left(\mathbb{P}_{\boldsymbol{\nu}}[A] \parallel \mathbb{P}_{\boldsymbol{\nu}'}[A]\right)$$

Definition 7. *The curvature constant of a concave differentiable function* $\psi(\cdot) : \mathcal{K} \to \mathbb{R}$ *with respect to a compact set* $\mathcal{K}$ *is defined as:*

$$C_\psi(\mathcal{K}) \triangleq \sup_{\substack{\mathbf{x}, \mathbf{z} \in \mathcal{K} \\ \alpha \in (0,1] \\ \mathbf{y} = \mathbf{x} + \alpha(\mathbf{z} - \mathbf{x})}} \frac{1}{\alpha^2} \left[\psi(\mathbf{x}) - \psi(\mathbf{y}) + \langle \mathbf{y} - \mathbf{x}, \nabla\psi(\mathbf{x}) \rangle\right]$$

Theorem 12 (Quasiconvex maximum principle Theorem 1 in (Ball, 2023))**.** *Let* K *be a nonempty compact, convex subset of a Hausdorff locally convex space. Any upper semicontinuous, quasiconvex function from* K *to* $[-\infty, \infty)$ *achieves its maximum at some extreme point of* K

Lemma 18 ((Wang et al., 2021a))**.** *Let* $\alpha, \beta \in (0, 1)$ *and* $A > 0$.

$$\int_0^\infty \left(\int_{T^\alpha}^\infty \exp\left(-At^\beta\right) dt \right) dT = \frac{\Gamma\left(\frac{1}{\alpha\beta} + \frac{1}{\beta}\right)}{\beta A^{\frac{1}{\alpha\beta} + \frac{1}{\beta}}}$$