Sum-max Submodular Bandits

Anonymous Authors[1]

Abstract

Many online decision-making problems correspond to maximizing a sequence of submodular functions. In this work, we introduce sum-max functions, a subclass of monotone submodular functions capturing several interesting problems, including best-of-K-bandits, combinatorial bandits, and the bandit versions on M-medians and hitting sets. We show that all functions in this class satisfy a key property that we call pseudo-concavity. This allows us to prove $\left(1-\frac{1}{e}\right)$-regret bounds for bandit feedback in the nonstochastic setting of the order of $\sqrt{MKT}$ (ignoring log factors), where T is the time horizon and M is a cardinality constraint. This bound, attained by a simple and efficient algorithm, significantly improves on the $\widetilde{\mathcal{O}}\left(T^{2/3}\right)$ regret bound for online monotone submodular maximization with bandit feedback. We also extend our results to a bandit version of the facility location problem.

1. INTRODUCTION

In many concrete settings of sequential decision-making, decisions are subsets of a finite set $[K]$ (possibly with cardinality constraints) and utilities, or rewards, are non-linear set functions over $[K]$. Although we may know that utility functions have some specific structure, e.g., they are submodular, the feedback may not reveal anything beyond the utility of the current decision. For example, consider an advertising campaign over $[K]$ digital channels (e.g., web, apps, and social media). Due to budget constraints, the campaign can show ads only on a subset of M channels for every user. If a user ends up buying the advertised product, we observe that a sale occurred, but we may not know which of the M channels triggered the purchase. The advertiser's goal is to choose the subset of channels for each new user in order to maximize the number of sales.

The same problem was studied (with a different motivation) by Simchowitz et al. (2016) under stochastic assumptions on the generation of the Bernoulli random variables each indicating whether displaying an ad on a certain channel triggers a purchase for the current user. In this work, we study the nonstochastic variant of this problem, where the binary variables associated with the channels are chosen, for each user, by an oblivious adversary. Our main result is an efficient algorithm minimizing regret in a much larger class of problems containing the multichannel advertising problem as a special case. In particular, our regret analysis applies to any sequential decision-making problem where reward functions belong to a subclass of all monotone submodular functions called *sum-max*.

A sum-max function is defined by a nonnegative matrix with K columns and an arbitrary number of rows. The value of the function evaluated at a subset $\mathcal{S} \subset [K]$ of columns is the sum over the rows of the maximum row element over the subset $\mathcal{S}$ of columns. In the multichannel campaign example, the matrix is binary with a single row. The j-th entry indicates whether the current user would buy the product if advertised on channel j. If the matrix is square and symmetric, then we recover the non-metric facility location problem as a special case.

As we said earlier, our analysis of regret for sum-max functions assumes bandit feedback: at each time t we only observe the reward $r_t(\mathcal{A}_t)$ associated with our decision $\mathcal{A}_t$, where r_t is the sum-max function chosen by the adversary at time t. Hence, the reward $r_t(\mathcal{S})$ that we would have obtained by choosing any $\mathcal{S} \neq \mathcal{A}_t$ remains unknown. We also consider cardinality constraints, in the form of a parameter M requiring that the decision $\mathcal{A}_t$ at each time t satisfy $\left|\mathcal{A}_t\right| \leq M$. Note that when $M = 1$ we recover the adversarial K-armed bandit problem.

Our main result is an efficient algorithm, MSE3, achieving a $\widetilde{\mathcal{O}}\left(\sqrt{MKT}\right)$ bound on the γ_M-regret for $\gamma_M = 1-$

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the Workshop on Foundations of Reinforcement Learning and Control at the International Conference on Machine Learning (ICML). Do not distribute.

$\left(1-1/M\right)^M$. For comparison, for the class of all monotone submodular functions, Niazadeh et al. (2021) obtain a $\left(1-\frac{1}{e}\right)$-regret bound of $\mathcal{O}\left((\ln K)^{1/3}M(KT)^{2/3}\right)$. As $\gamma_M > 1-\frac{1}{e}$ for all $M > 1$, this bound is worse than ours in both approximation factor and regret.

When $M = 1$, algorithm MSE3 reduces to the standard Exp3 algorithm for K-armed bandits and our result specializes to the standard $\mathcal{O}\left(\sqrt{K(\ln K)T}\right)$ regret bound of Exp3. This implies that the $\sqrt{KT}$ dependence in the regret bound is not improvable, even disregarding efficiency. Moreover, we show that improving on the approximation factor γ_M with an efficient algorithm would give an efficient randomized algorithm for solving set cover on $[K]$ with an approximation ratio of $(1-\varepsilon)\ln K$, which is NP-hard for any $\varepsilon > 0$ (Dinur and Steurer, 2014).

In many real world problems, including an element i in the decision $\mathcal{A}_t$ at round t invokes a cost (i.e., a negative reward) $c_{t,i} \geq 0$. When this is the case we would like to maximize the cumulative reward:

$$\sum_{t\in[T]}\left(r_t(\mathcal{A}_t) - \sum_{i\in\mathcal{A}_t} c_{t,i}\right).$$

We show that MSE3 can handle this generalized problem if it receives, at the end of each round t, the values of $c_{t,i}$ for all $i \in \mathcal{A}_t$. We note that the bandit MSE3 without costs is a special case of MSE3 with costs.

The inclusion of costs creates a tension between including arms in $\mathcal{A}_t$ to increase the reward and, simultaneously, avoid including too many arms to control the costs. We address this trade-off in Section 5 by introducing and analyzing a variant of MSE3 for regret minimization with costs and bandit feedback where the rewards are sum-max functions without cardinality constraints. We call this setting the bandit facility location problem because it is a bandit version of the online facility location problem studied by Pasteris et al. (2021).

For $M > 1$ and arbitrary costs, MSE3 selects $\mathcal{A}_t$ by performing M independent draws $a_{t,1},\ldots,a_{t,M}$ from a distribution $\boldsymbol{p}_t = \left(p_{t,1},\ldots,p_{t,K}\right) \in \Delta_K$. Then, a reward estimate for each $i \in [K]$ is computed using

$$g_{t,i} = \frac{r_t(\mathcal{A}_t) - c_{t,i}}{p_{t,i}} \sum_{j\in[M]} [\![a_{t,j} = i]\!]\,, \qquad (1)$$

where, for any statement S, the Iverson bracket notation $[\![\cdot]\!]$ is defined as $[\![S]\!] = 1$ if S is true and $[\![S]\!] = 0$ otherwise. Note that for $M = 1$ and $c_{t,i} = 0$, the above reduces to the standard reward estimate of Exp3.

We now give an overview of how MSE3 works when we have no costs (i.e., $c_{t,i} = 0$). For all set functions r, we construct a function $\Phi^r : \mathbb{R}_+^K \to \mathbb{R}$ such that for all $\boldsymbol{q} \in \Delta_K$ we have that $\Phi^r(\boldsymbol{q})$ is the expected value of $r(\mathcal{A})$ when $\mathcal{A}$ is constructed by drawing M arms i.i.d. with replacement from $\boldsymbol{q}$. Specifically, we first show that there exists a function $\widetilde{r} : 2^{[K]} \to \mathbb{R}$ such that for all $\mathcal{Q} \subseteq [K]$ we have $r(\mathcal{Q}) = \sum_{\mathcal{S}\subseteq[K]} [\![\mathcal{Q} \subseteq \mathcal{S}]\!]\widetilde{r}(\mathcal{S})$. For all $\boldsymbol{q} \in \mathbb{R}_+^K$ we then define:

$$\Phi^r(\boldsymbol{q}) = \sum_{\mathcal{S}\subseteq[K]} \widetilde{r}(\mathcal{S}) \left(\sum_{i\in[K]} [\![i \in \mathcal{S}]\!] q_i\right)^M .$$

We learn via online exponentiated gradient ascent using the unbiased estimates (1) of the gradient of Φ^{r_t}. Clearly, for exponentiated gradient ascent to work we must have that, for all rounds t, our objective function Φ^{r_t} is concave over the simplex. We show that a sufficient condition for this to hold is that the function r_t is pseudo-concave, see Section 2 for a formal definition.

Next, we bound the regret with respect to any vector $\boldsymbol{p}^* \in \Delta_K$. Namely, we bound the expected reward of our algorithm relative to $\sum_{t\in[T]} \Phi^{r_t}(\boldsymbol{p}^*)$. By taking $\boldsymbol{p}^*$ such that $p_i^* = [\![i \in \mathcal{S}]\!]/|\mathcal{S}|$ for some set $\mathcal{S}$ we show that, because r_t is submodular, we have $\Phi^{r_t}(\mathcal{S}) \geq (1-\alpha^M)r_t(\mathcal{S})$ where $\alpha = (|\mathcal{S}|-1)/|\mathcal{S}|$. By bounding the variance of the gradient estimate we show that the regret term is $\widetilde{\mathcal{O}}(\sqrt{MKT})$.

We have provided an overview of how, when we have no costs, MSE3 works and why we require r_t to be pseudo-concave and submodular. We now describe how costs are incorporated. This is done by using, instead of Φ^{r_t}, the function Ψ_t defined by:

$$\Psi_t(\boldsymbol{q}) = \Phi^{r_t}(\boldsymbol{q}) - M \sum_{i\in[K]} q_i c_{t,i}\,,$$

so that $\Psi_t(\boldsymbol{p}_t)$ lower bounds the expected profit on trial t. Since Ψ_t differs from Φ^{r_t} by a linear function it is straightforward to extend the above methodology to this new objective function.

2. SUM-MAX FUNCTIONS

We now introduce sum-max functions and define the key property of this class that allows us to learn it with bandit feedback.

Definition 2.1. A set function $r : 2^{[K]} \to \mathbb{R}$ is *sum-max* if and only if there exists some $N \in \mathbb{N}$ and some matrix $\boldsymbol{V} \in \mathbb{R}^{N\times K}$ such that for all $\mathcal{S} \subseteq [K]$ with $\mathcal{S} \neq \emptyset$ we have:

$$r(\mathcal{S}) = \sum_{k\in[N]} \max_{i\in\mathcal{S}} V_{k,i} \quad \text{and} \quad r(\emptyset) \leq \sum_{k\in[N]} \min_{i\in[K]} V_{k,i}$$

For example, consider a marketplace with N buyers and K sellers. The value $V_{k,i}$ is the combined utility of buyer k going to seller i. The value $r(\mathcal{S})$ is the social welfare when a subset $\mathcal{S}$ of sellers participate in the marketplace, and buyers match up with sellers to optimize their combined utilities. When there is only one buyer ($N = 1$), $\boldsymbol{V}$ is a vector $(V_1, \dots, V_K)$ and we view each $i \in [K]$ as an arm with reward V_i. Then $r(\mathcal{S}) = \max_{i \in \mathcal{S}} V_i$, the maximum reward of an arm in the chosen set $\mathcal{S}$.

As sum-max functions are sums of monotone submodular functions, they are monotone submodular. We now list a number of sequential decision-making problems that can be expressed as regret minimization of specific sum-max functions under bandit feedback.

The multichannel campaign problem. This is our nonstochastic variant of the best-of-k bandit problem of Simchowitz et al. (2016). To view it as an instance of sum-max optimization, set $N = 1$ and let $V_i \in \{0, 1\}$ indicate whether a user makes a purchase when the ad is displayed on channel i. Then $(V_1, \dots, V_K)$ can be viewed as the incidence vector of a subset $\mathcal{D} \subseteq [K]$ of channels, and the reward is defined by $r(\mathcal{S}) = [\![\mathcal{S} \cap \mathcal{D} \neq \emptyset]\!]$. The feedback is bandit because we do not know what channel triggered the sale for that user.

Bandit hitting sets. This is a generalization of the previous example where $N \geq 1$ and $\boldsymbol{V}$ is a boolean matrix. Each row of $\boldsymbol{V}$ denotes a subset $\mathcal{C}_k$ of $[K]$ and $V_{k,i}$ indicates whether $i \in \mathcal{C}_k$. The value $r(\mathcal{S})$ then counts how many sets $\mathcal{C}_k$ have a non-empty intersection with $\mathcal{S}$. Bandit setting occurs when the sets remain unknown and each time we only observe the number of intersected sets.

Combinatorial bandits. Another important special case is when we receive the sum of the rewards r_i of the arms $i \in \mathcal{S}$. In this case $N = K$ and $V_{k,i} = [\![k = i]\!] r_i$. The problem is then equivalent to a combinatorial bandit (with full bandit feedback) over the class of M-sized subsets (Cesa-Bianchi and Lugosi, 2012).

Bandit k-medians. Given $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$ points in a metric space $(\mathcal{X}, d)$, consider the version of the k-medians problem (for $k = M$) where the M centroids have to be chosen in the given set of points. The value of the objective function at a candidate solution $\mathcal{S} \subset [K]$ with $|\mathcal{S}| \leq M$ can be written as

$$r(\mathcal{S}) = -\sum_{k \in [N]} \min_{i \in \mathcal{S}} d(\boldsymbol{x}_k, \boldsymbol{x}_i)\,.$$

Clearly, this is a sum-max function for $\boldsymbol{V}$ with elements $V_{k,i} := -d(\boldsymbol{x}_k, \boldsymbol{x}_i)$. The feedback is bandit when we do not know the metric, but we can observe the value of the objective function.

Next, we introduce an important property of sum-max functions.

Definition 2.2. Suppose we have a set function $r : 2^{[K]} \to \mathbb{R}$. For any $\mathcal{S} \subseteq [K]$ define the matrix $\boldsymbol{U}^{r,\mathcal{S}} \in \mathbb{R}^{K \times K}$ such that $U^{r,\mathcal{S}}_{i,j} = r(\mathcal{S} \cup \{i, j\})$ for all $i, j \in [K]$. We call the function r *pseudo-concave* if and only if $\boldsymbol{x}^\top \boldsymbol{U}^{r,\mathcal{S}} \boldsymbol{x} \leq 0$ for all $\mathcal{S} \subseteq [K]$ and all $\boldsymbol{x} \in \mathbb{R}^K$ with $\boldsymbol{x} \cdot \mathbf{1} = 0$.

In Appendix C, we show that there are monotone submodular functions that are not pseudo-concave. As a consequence, sum-max functions are indeed a proper subset of the class of monotone submodular functions. The following theorem confirms that all sum-max functions are pseudo-concave:

Theorem 2.3. *Any sum-max set function is pseudo-concave.*

Proof. Suppose we have some sum-max function $r : 2^{[K]} \to [0, 1]$. Let $\boldsymbol{V}$ be as in Definition 2.1. Without loss of generality, we will assume that all components of $\boldsymbol{V}$ are non-negative and $r(\emptyset) = 0$ (since any sum-max function can be transformed into this form by the addition of a constant).
Define, for any $\mathcal{Q} \subseteq [K]$, the set function $r^{\mathcal{Q}} : 2^{[K]} \to [0, 1]$ such that for all $\mathcal{S} \subseteq [K]$ we have

$$r^{\mathcal{Q}}(\mathcal{S}) := [\![\mathcal{S} \cap \mathcal{Q} \neq \emptyset]\!].$$

We shall now show that for all such $\mathcal{Q}$ we have that $r^{\mathcal{Q}}$ is pseudo-concave. Choose any $\boldsymbol{x} \in \mathbb{R}^K$ with $\boldsymbol{x} \cdot \mathbf{1} = 0$ and any $\mathcal{S} \subseteq [K]$. We have two cases:

1. If $\mathcal{S} \cap \mathcal{Q} \neq \emptyset$, for all $i, j \in [K]$ we have $r^{\mathcal{Q}}(\mathcal{S} \cup \{i, j\}) = 1$, this implies $\boldsymbol{U}^{r^{\mathcal{Q}},\mathcal{S}} = \mathbf{1}\mathbf{1}^\top$ and hence $\boldsymbol{x}^\top \boldsymbol{U}^{r^{\mathcal{Q}},\mathcal{S}} \boldsymbol{x} = 0$.
2. If $\mathcal{S} \cap \mathcal{Q} = \emptyset$ then for all $i, j \in [K]$ we have

$$r^{\mathcal{Q}}(\mathcal{S} \cup \{i, j\}) = [\![(i \in \mathcal{Q}) \vee (j \in \mathcal{Q})]\!].$$

Let $\boldsymbol{z} \in \{0, 1\}^K$ be such that for all $k \in [K]$ we have $z_k := [\![k \notin \mathcal{Q}]\!]$. Then for all $i, j \in [K]$ we have

$$[\![(i \in \mathcal{Q}) \vee (j \in \mathcal{Q})]\!] = 1 - z_i z_j\,,$$

so that, by above, we have $\boldsymbol{U}^{r^{\mathcal{Q}},\mathcal{S}} = \mathbf{1}\mathbf{1}^\top - \boldsymbol{z}\boldsymbol{z}^\top$ This implies that: $\boldsymbol{x}^\top \boldsymbol{U}^{r^{\mathcal{Q}},\mathcal{S}} \boldsymbol{x} = -(\boldsymbol{x} \cdot \boldsymbol{z})^2 \leq 0$.

And therefore, $r^{\mathcal{Q}}$ is pseudo-concave.

Now suppose we have a vector $\boldsymbol{v} \in \mathbb{R}^K_+$ and define the set function $r^{\boldsymbol{v}} : 2^{[K]} \to \mathbb{R}_+$ such that for all $\mathcal{S} \subseteq [K]$ we have

$$r^{\boldsymbol{v}}(\mathcal{S}) := \max_{i \in \mathcal{S}} v_i\,,$$

where the maximum of the empty set is defined as equal to zero. We can order the set $[K]$ into the sequence $\langle j_i \,|\, i \in [K] \rangle = [K]$ where $v_{j_{i+1}} \leq v_{j_i}$ for all $i \in [K-1]$. For all $i \in [K]$ we can define $\mathcal{Q}_i := \{j_k \,|\, k \leq i\}$. Now note then that for all $\mathcal{S} \subseteq [K]$ the set function $r^{\boldsymbol{v}}(\mathcal{S})$ can be expressed as

$$\sum_{i \in [K-1]} (v_{j_i} - v_{j_{i+1}}) [\![\mathcal{S} \cap \mathcal{Q}_i \neq \emptyset]\!] + v_{j_K} [\![\mathcal{S} \cap \mathcal{Q}_K \neq \emptyset]\!]$$

$$= \sum_{i \in [K-1]} (v_{j_i} - v_{j_{i+1}}) r^{\mathcal{Q}_i}(\mathcal{S}) + v_{j_K} r^{\mathcal{Q}_K}(\mathcal{S}) \,,$$

so, by above, $r^{\boldsymbol{v}}$ is a positive sum of pseudo-concave functions and is hence itself pseudo-concave. Note also that $r^{\boldsymbol{v}}$ is clearly submodular. Noting that r is a positive sum of functions of the form $r^{\boldsymbol{v}}$ we have now shown that it is both pseudo-concave and submodular as required. □

3. ADDITIONAL RELATED WORK

The work closest to ours is Pasteris et al. (2021), where they study online facility location with full information feedback. Our work improves on theirs in many respects: First, we solve the problem with bandit feedback, which requires designing an entirely different algorithm based on our discovery of an unbiased estimator for the gradient of our expected reward (we find it remarkable that such an estimator exists). As a consequence, our algorithm is also applicable to the full-information setting, where we obtain a per-trial running time of $\mathcal{O}(MK)$ when given an oracle for the reward function. When considering general sum-max functions, the methodology of Pasteris et al. (2021) would instead require a per-trial running time exponential in K[1]. Second, our algorithm can efficiently learn classes that are even more general than sum-max functions. Third, we obtain tighter approximation ratios and show optimality for the multichannel campaign problem (and thus optimality in general).

Sum-max functions are a special case of linear submodular functions (Yue and Guestrin, 2011), which are of the form $r(\mathcal{S}) = \sum_{i \in [N]} w_i F_i(\mathcal{S})$ for $F_1, \ldots, F_N$ monotone submodular functions and $w_1, \ldots, w_N$ non-negative coefficients. However, linear submodular functions have been only studied in stochastic settings, assuming preliminary knowledge of $F_1, \ldots, F_N$, and using a feedback model more informative than our bandit feedback.

Click-models (Lattimore and Szepesvári, 2020; Lattimore et al., 2018; Kveton et al., 2015) provide a different stochastic formalization of the best-of-k bandit problem. Here the user is presented with an ordered list of items, and the learner receives a positive reward if the user clicks on one of the presented items. The difference with our multichannel campaign problem is that the items are ordered, and the likelihood of clicking an item is also affected by the position of the item within the list.

[1] The work of Pasteris et al. (2021) only considered single-user cases, but it is straightforward to extend their methodology to general sum-max functions.

Algorithm 1 MSE3

Set $\eta := \ln(K)/R$ and $p_{1,i} := 1/K$ for $i \in [K]$

for $t = 1, 2, \ldots, T$ **do**:

1. For all $j \in [M]$ draw $a_{t,j} \in [K]$ from distribution $\boldsymbol{p}_t$
2. Define $\mathcal{A}_t := \{a_{t,j} \mid j \in [M]\}$
3. Receive $r_t(\mathcal{A}_t)$ and $\{c_{t,i} \,|\, i \in \mathcal{A}_t\}$
4. For all $i \in [K]$ set
$$g_{t,i} := \frac{r_t(\mathcal{A}_t) - c_{t,i}}{p_{t,i}} \sum_{j \in [M]} [\![a_{t,j} = i]\!]$$
5. For all $i \in [K]$ define $\widetilde{p}_{t,i} := p_{t,i} \exp(\eta g_{t,i})$
6. Define $\boldsymbol{p}_{t+1} := \widetilde{\boldsymbol{p}}_t / \|\widetilde{\boldsymbol{p}}_t\|_1$

4. MAIN RESULT

Our learning problem is formally defined as follows. The values $M, K \in \mathbb{N}$ and $C \in \mathbb{R}_+$ are all preliminarily known to the learner. Hidden from the learner, the adversary selects a sequence of set functions $\langle r_t \,|\, t \in [T] \rangle$, each with domain $2^{[K]}$ and a sequence of vectors $\langle \boldsymbol{c}_t \,|\, t \in [T] \rangle$ each in $[0, C]^K$. On each trial $t \in [T]$:

1. The learner chooses some $\mathcal{A}_t \subseteq [K]$ with $|\mathcal{A}_t| \leq M$.
2. The value $r_t(\mathcal{A}_t)$ is revealed.
3. For all $i \in \mathcal{A}_t$ the value $c_{t,i}$ is also revealed.

The learner maintains a probability vector $\boldsymbol{p}_t \in \Delta_K$, and behaves as described in Algorithm 1.

To aid our theorem statement we add the following definitions. For all $t \in [T]$ and $\mathcal{Q} \subseteq [K]$ we define $\hat{r}_t(\mathcal{Q}) := r_t(\mathcal{Q}) - r_t(\emptyset)$, which is the difference between the learner's profit on trial t and that which it would have obtained by selecting the empty set, $\psi_t := \hat{r}_t(\mathcal{A}_t) - \gamma_t(\mathcal{A}_t)$, and $\gamma_t(\mathcal{Q}) := \sum_{i \in \mathcal{Q}} c_{t,i}$. We note that by considering $\hat{r}_t$ instead of r_t our bounds do not change when r_t is shifted by an additive constant (which can be different for different trials t) as long as the range of r_t falls within the bounds described as follows.

We assume that the learner knows upper and lower bounds on the range r_t for all trials t. Hence, without loss of generality, assume that $r_t(\mathcal{Q}) \in [-1, 0]$ for all

$t \in [T]$ and $\mathcal{Q} \setminus [K]$ (otherwise scale and shift r_t and C). Let

$$R := (1+C)\sqrt{2\ln(K)M(K+M-1)T}\,.$$

Our results hold for a relaxed notion of submodularity, which we call *pseudo-submodularity*.

Definition 4.1. A set function $r : 2^{[K]} \to \mathbb{R}$ is *pseudo-submodular* if and only if for every set $\mathcal{S} \subseteq [K]$ with $\mathcal{S} \neq \emptyset$ there exists some $i \in \mathcal{S}$ such that for all $\mathcal{Q} \subseteq \mathcal{S} \setminus \{i\}$ we have $r(\mathcal{Q} \cup \{i\}) - r(\mathcal{Q}) \geq r(\mathcal{S}) - r(\mathcal{S} \setminus \{i\})$.

Note that all pseudo-submodular set functions are also submodular. We now present our main result.

Theorem 4.2. *Given r_t is pseudo-concave and pseudo-submodular for all $t \in [T]$, then for any set $\mathcal{S} \subseteq [K]$ with $\mathcal{S} \neq \emptyset$ we have*

$$\sum_{t\in[T]} \mathbb{E}[\psi_t] \geq \left(1-\alpha^M\right) \sum_{t\in[T]} \hat{r}_t(\mathcal{S}) - \frac{M}{|\mathcal{S}|} \sum_{t\in[T]} \gamma_t(\mathcal{S}) - R\,,$$

where

$$\alpha := 1 - \frac{1}{|\mathcal{S}|}.$$

Proof. See Section 6 □

We note that both the standard facility location and k-medians problems are often phrased as the minimization of a loss rather than a maximization of a profit. Our results easily capture this by considering the reward as a negative loss.

We now show that the approximation ratio $1-\alpha^M$ is not improvable in general in the class of sum-max functions. In particular, we show that obtaining an efficient online learning algorithm for the multichannel advertising problem with a sublinear γ-regret with $\gamma < 1-\alpha^M$ would give an efficient randomized algorithm for solving set cover on $[K]$ with an approximation better than $\ln K$. As shown in (Dinur and Steurer, 2014), obtaining an approximation of $(1-\varepsilon)\ln K$ for set cover is NP-hard for any $\varepsilon > 0$.

Recall that an instance of the multichannel campaign problem over K ads is defined by a sequence $\langle r_t \,|\, t \in [T]\rangle$ of set functions over $[K]$ such that for all $t \in [T]$ there exists some $\mathcal{D}_t \subseteq [K]$ with $r_t(\mathcal{Q}) = [\![\mathcal{Q} \cap \mathcal{D}_t \neq \emptyset]\!]$ for all $\mathcal{Q} \subseteq [K]$.

Theorem 4.3. *Suppose that there exists some $d \in \mathbb{N}$, $s \in (0,1)$, $\gamma > 1$, and a randomized polynomial time algorithm for the learner such that for all $K, M \in \mathbb{N}$ and for any instance of the multichannel advertising problem, it holds that $|\mathcal{A}_t| \leq M$ for all $t = 1, \ldots, T$ and, for any subset $\mathcal{S} \subseteq [K]$,*

$$\mathbb{E}\left[\sum_{t\in[T]} r_t(\mathcal{A}_t)\right] \geq \left(1-\alpha^{\gamma M}\right) \sum_{t\in[T]} r_t(\mathcal{S}) - R'\,,$$

where $R' \in \mathcal{O}(K^d T^s)$ and $\alpha := 1 - \frac{1}{|\mathcal{S}|}$. Then, for all $\varepsilon \in \left(0, 1-1/\gamma\right)$ and $B > 4^{1/((1-\varepsilon)\gamma-1)}$, there exists a randomized polynomial-time algorithm for the set cover problem on $[B]$ that, with probability at least $\frac{1}{2}$, achieves approximation ratio at least $(1-\varepsilon)\ln(B)$.

The proof can be found in Appendix B.

Algorithm 2 FLE3

Run MSE3 with $L = 2K$ arms and $M := \frac{K}{2}\ln(T/K^2)$.

On each trial $t \in [T]$:
1. Let $\mathcal{A}'_t$ be the output of MSE3
2. Output $\mathcal{A}_t := \mathcal{A}'_t \cap [K]$
3. Receive $r_t(\mathcal{A}_t)$ and $\{c_{t,i} \,|\, i \in [K]\}$
4. For all $i \in [L] \setminus [K]$ set $c_{t,i} := 0$
5. Feed $r_t(\mathcal{A}_t)$ and $\{c_{t,i} \,|\, i \in [L]\}$ back to MSE3

5. BANDIT FACILITY LOCATION

We can view this setting as a generalization of the marketplace example where sellers pay a known cost to enter the market. At each round, the platform admits a subset $\mathcal{A}_t$ of sellers and only observes the resulting social welfare (bandit feedback).

In this application, there are no restrictions on the set of arms $\mathcal{A}_t$ that we choose. We seek to maximize $r(\mathcal{A}_t) - \gamma(\mathcal{A}_t)$ where r is the sum-max reward function and γ is the linear and positive cost function.

For the facility location problem we must choose M, noting that although a high value of M increases the approximation ratio on the reward, it also increases that on the costs. To decrease the potentially large approximation ratio on the costs, we borrow from (Pasteris et al., 2021) the idea of *dummy arms* and the tuning of M. This leads to our algorithm FLE3 described in Algorithm 2. The bound on the total profit of FLE3 is given in the following theorem.

Theorem 5.1. *Given that $C = 1$ and $r_t : 2^K \to [-1, 0]$ is pseudo-concave and pseudo-submodular for all $t \in [T]$, we have that the algorithm $FLE3$ obtains the following bound for all $\mathcal{S} \subseteq [K]$ with $\mathcal{S} \neq \emptyset$:*

$$\sum_{t\in[T]} \mathbb{E}[\psi_t] \geq \sum_{t\in[T]} \hat{r}_t(\mathcal{S}) - \frac{1}{2}\ln\left(\frac{T}{K^2}\right) \sum_{t\in[T]} \gamma_t(\mathcal{S}) - R''\,,$$

where $R'' \in \widetilde{\mathcal{O}}(K\sqrt{T})$.

Proof. For all $t \in [T]$ define the set function $r'_t : 2^L \to [0,1]$ such that for all $\mathcal{Q} \subseteq [L]$, $r'_t(\mathcal{Q}) := r_t(\mathcal{Q} \cap [K])$ and, as consequence, $\hat{r}'_t(\mathcal{Q}) := r'_t(\mathcal{Q}) - r'_t(\emptyset)$. Now taking into consideration any possible comparator set $\mathcal{S} \subseteq [K]$, we define

$$\mathcal{S}' := \mathcal{S} \cup \{K + i \,|\, i \in [K - |\mathcal{S}|]\} ,$$

noting that $|\mathcal{S}'| = K$. Note that r'_t is sum-max and hence, by Theorem 2.3, is pseudo-concave and submodular for all $t \in [T]$. This allows us to apply Theorem 4.2, that gives us:

$$\sum_{t\in[T]} \mathbb{E}[\psi_t] \geq \left(1 - \alpha^M\right) \sum_{t\in[T]} \hat{r}'_t(\mathcal{S}') - \frac{M}{|\mathcal{S}'|} \sum_{t\in[T]} \gamma_t(\mathcal{S}') - R$$

$$= \left(1 - \alpha^M\right) \sum_{t\in[T]} \hat{r}_t(\mathcal{S}) - \frac{M}{|\mathcal{S}'|} \sum_{t\in[T]} \gamma_t(\mathcal{S}) - R \tag{2}$$

$$= \left(1 - \alpha^M\right) \sum_{t\in[T]} \hat{r}_t(\mathcal{S}) - \frac{1}{2} \ln \frac{T}{K^2} \sum_{t\in[T]} \gamma_t(\mathcal{S}) - R , \tag{3}$$

where equation (2) comes from the contribution of the dummy arms and equation (3) from the definition of M. Given that

$$\alpha := \frac{|\mathcal{S}'| - 1}{|\mathcal{S}'|} = \frac{K-1}{K} \leq \exp\left(-1/K\right) ,$$

we can therefore see that

$$\alpha^M \sum_{t\in[T]} \hat{r}_t(\mathcal{S}) \leq \alpha^M T = \exp(-M/K)T$$

$$= \frac{1}{\sqrt{T/K^2}} T = \sqrt{TK^2} , \tag{4}$$

where we used the definition of M given in Algorithm 2. Putting together (3) and (4) gives us the result, where $R'' = R + \sqrt{TK^2}$. □

6. ANALYSIS

We now give an overview of the proof of Theorem 4.2.

We first consider the case that we have no costs (i.e. $\boldsymbol{c}_t = \boldsymbol{0}$). MSE3 works by maintaining a probability distribution over the set of arms. Specifically, $\boldsymbol{p}_t \in \Delta_K$ is the vector whose components are the probabilities of drawing the actions on trial t. On trial t the algorithm constructs the set $\mathcal{A}_t$ by drawing a sequence $\langle a_{t,j} \,|\, j \in [M]\rangle$ of arms i.i.d. with replacement from $\boldsymbol{p}_t$ and then setting $\mathcal{A}_t := \{a_{t,j} \,|\, j \in [M]\}$.

This stochastic draw of a sequence and set from a probability vector will be represented by the following notation.

Definition 6.1. For all $\boldsymbol{q} \in \Delta_K$ let $\langle b_j(\boldsymbol{q}) \,|\, j \in [M]\rangle$ be a sequence of stochastic quantities drawn i.i.d. at random from (the probability distribution characterised by) $\boldsymbol{q}$. In addition, let $\mathcal{B}(\boldsymbol{q}) := \{b_j(\boldsymbol{q}) \,|\, j \in [M]\}$.

Note that our expected reward on trial t is $\mathbb{E}[r_t(\mathcal{B}(\boldsymbol{p}_t))]$ and hence, for all set functions r we shall construct a differentiable function $\Phi^r : \mathbb{R}^K \to \mathbb{R}$ such that for all $\boldsymbol{q} \in \Delta_K$ we have $\Phi^r(\boldsymbol{q}) = \mathbb{E}[r(\mathcal{B}(\boldsymbol{q}))]$. This construction is based on the following notion of a *subset decomposition*.

Definition 6.2. Given a function $r : 2^{[K]} \to \mathbb{R}$, we call a function $\widetilde{r} : 2^{[K]} \to \mathbb{R}$ a *subset decomposition* of r if and only if for all $\mathcal{Q} \subseteq [K]$ we have

$$r(\mathcal{Q}) = \sum_{\mathcal{S}\subseteq[K]} [\![\mathcal{Q} \subseteq \mathcal{S}]\!] \widetilde{r}(\mathcal{S}) .$$

The following lemma confirms that every set function has a unique subset decomposition.

Lemma 6.3. *Given a function $r : 2^{[K]} \to \mathbb{R}$ there exists a unique subset decomposition $\widetilde{r}$ of r.*

Proof. See Appendix A.1. □

Now we can define our function Φ^r.

Definition 6.4. For all $r : 2^K \to \mathbb{R}$ and all $\boldsymbol{q} \in \mathbb{R}^K$ define

$$\Phi^r(\boldsymbol{q}) := \sum_{\mathcal{S}\subseteq[K]} \widetilde{r}(\mathcal{S}) \left(\sum_{i\in[K]} [\![i \in \mathcal{S}]\!] q_i \right)^M ,$$

where, by Lemma 6.3, $\widetilde{r}$ is the unique subset decomposition of r.

The following lemma confirms that our function Φ^r indeed satisfies our condition.

Lemma 6.5. *For all $r : 2^K \to \mathbb{R}$ and all $\boldsymbol{q} \in \Delta_K$ we have $\Phi^r(\boldsymbol{q}) = \mathbb{E}[r(\mathcal{B}(\boldsymbol{q}))]$.*

Proof. See Appendix A.2. □

Drawing inspiration from (Auer et al., 2001) we will learn via online exponentiated gradient ascent with the functions Φ^{r_t} using unbiased gradient estimates. Of course, this means that we must be able to construct unbiased gradient estimates. Remarkably, we now show that we can use our sequence $\langle a_{t,j} \,|\, j \in [M]\rangle$ and the observed reward $r_t(\mathcal{A}_t)$ to construct an unbiased gradient estimate $\boldsymbol{g}_t$ defined in Algorithm 1 of the function Φ^{r_t} at $\boldsymbol{p}_t$.

Lemma 6.6. *For all $r : 2^K \to \mathbb{R}$, all $\boldsymbol{q} \in \Delta_K$ and all $i \in [K]$ we have*

$$\partial_i \Phi^r(\boldsymbol{q}) = \mathbb{E}\left[\frac{r(\mathcal{B}(\boldsymbol{q}))}{q_i} \sum_{j \in [M]} [\![b_j(\boldsymbol{q}) = i]\!]\right].$$

Proof. See Appendix A.3. □

For exponentiated gradient ascent to work, we must have that, for all trials t, our objective function Φ^{r_t} is concave over the simplex. We now show that a sufficient condition for this to hold is that the function r_t is pseudo-concave.

Lemma 6.7. *For all pseudo-concave set functions $r : 2^K \to \mathbb{R}$ we have that Φ^r is concave over the simplex Δ_K.*

Proof. See Appendix A.4. □

Now that we have all the underpinnings for exponentiated gradient ascent to function properly, we can establish a bound on the regret relative to any vector $\boldsymbol{p}^* \in \Delta_K$ via the following classic result.

Lemma 6.8. *For any vector $\boldsymbol{p}^* \in \Delta_K$ we have*

$$\sum_{t \in [T]} (\boldsymbol{p}^* - \boldsymbol{p}_t) \cdot \boldsymbol{g}_t \leq \frac{1}{\eta} \sum_{i \in [K]} p_i^* \ln(K p_i^*) + \eta \sum_{t \in [T]} \sum_{i \in [K]} p_{t,i} g_{t,i}^2 .$$

Proof. A classic result from the analysis of HEDGE. □

This lemma gives a bound on the regret since, because we have shown that $\boldsymbol{g}_t$ is an unbiased estimate of the gradient and the objective function is concave over the simplex, the term $(\boldsymbol{p}^* - \boldsymbol{p}_t) \cdot \boldsymbol{g}_t$ is bounded below by $\Phi^{r_t}(\boldsymbol{p}^*) - \Phi^{r_t}(\boldsymbol{p}_t)$. Note that we have shown above that $\Phi^{r_t}(\boldsymbol{p}_t)$ is equal to $\mathbb{E}[r_t(\mathcal{A}_t)]$.

We will later discuss the bounding of the regret itself, but first we shall show how to choose $\boldsymbol{p}^*$ such that we can bound $\Phi^{r_t}(\boldsymbol{p}^*)$ relative to $r_t(\mathcal{S})$ for some set $\mathcal{S} \subseteq [K]$. Specifically, we will choose $\boldsymbol{p}^*$ equal to $\boldsymbol{p}^{\mathcal{S}}$ in the following definition.

Definition 6.9. For all $\mathcal{S} \subseteq [K]$ with $\mathcal{S} \neq \emptyset$ define $\boldsymbol{p}^{\mathcal{S}} \in \Delta_K$ such that for all $i \in [K]$ we have

$$p_i^{\mathcal{S}} := \frac{[\![i \in \mathcal{S}]\!]}{|\mathcal{S}|} .$$

We use the following lemma will to bound $\Phi^{r_t}(\boldsymbol{p}^{\mathcal{S}})$, and it explains why we require r_t to be pseudo-submodular.

Lemma 6.10. *Let $\mathcal{S} \subseteq [K]$ with $\mathcal{S} \neq \emptyset$, $r : 2^{[K]} \to \mathbb{R}$ be a pseudo-submodular function, and $\mathcal{Z} \subseteq [K]$ be a set formed by drawing M elements uniformly at random (with replacement) from $\mathcal{S}$. Then we have*

$$\mathbb{E}[r(\mathcal{Z}) - r(\emptyset)] \geq \left(1 - \left(\frac{|\mathcal{S}| - 1}{|\mathcal{S}|}\right)^M\right)(r(\mathcal{S}) - r(\emptyset)) .$$

Proof. See Appendix A.5 □

With this lemma in hand, we can now bound $\Phi^{r_t}(\boldsymbol{p}^{\mathcal{S}})$.

Lemma 6.11. *Given any $\mathcal{S} \subseteq [K]$ and any pseudo-submodular set function $r : 2^{[K]} \to \mathbb{R}$ we have*

$$\Phi^r(\boldsymbol{p}^{\mathcal{S}}) \geq r(\emptyset) + \left(1 - \left(\frac{|\mathcal{S}| - 1}{|\mathcal{S}|}\right)^M\right)(r(\mathcal{S}) - r(\emptyset)) .$$

Proof. See Appendix A.6. □

Before we bound the regret term, we show how to incorporate the costs, so that $\boldsymbol{c}_t$ can be non-zero. This is done by choosing, instead of Φ^{r_t}, the objective function Ψ_t defined as follows.

Definition 6.12. For all trials $t \in [T]$ define $\Psi_t : \mathbb{R}^K \to \mathbb{R}$ such that for all $\boldsymbol{q} \in \mathbb{R}^K$ we have

$$\Psi_t(\boldsymbol{q}) := \Phi^{r_t}(\boldsymbol{q}) - M \boldsymbol{q} \cdot \boldsymbol{c}_t .$$

Note that by Lemma 6.5 we have that $\Psi_t(\boldsymbol{p}_t)$ is a lower bound on the expected profit and by Lemma 6.7 Ψ_t is concave over the simplex. It can hence serve as a surrogate concave objective function.

Lemma 6.6 leads to the following lemma, which confirms that $\boldsymbol{g}_t$ is an unbiased gradient estimate of Ψ_t at $\boldsymbol{p}_t$.

Lemma 6.13. *For all trials $t \in [T]$ and Ψ_t as defined in Definition 6.12, we have*

$$\nabla \Psi_t(\boldsymbol{p}_t) = \mathbb{E}[\boldsymbol{g}_t \,|\, \boldsymbol{p}_t] ,$$

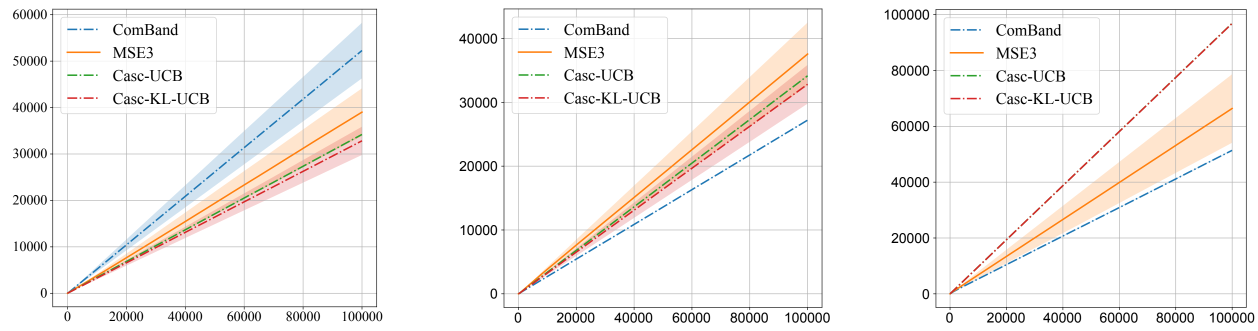
Proof. See Appendix A.7. □

Now we have shown that our results carry over to the case of non-zero costs, we can finally bound the regret via Lemma 6.8 and the following lemma.

Lemma 6.14. *For all trials $t \in [T]$ we have*

$$\mathbb{E}\left[\sum_{i \in [K]} p_{t,i} g_{t,i}^2\right] \leq (1 + C)^2 M (K + M - 1) .$$

Proof. See Appendix A.8. □



(a) Stochastic (b) Stochastic with adversarial corruptions (c) Worst-case stochastic

Figure 1. Cumulative reward over time in the three environment settings is described. The results also display the 95% confidence intervals over 35 runs with an Intel Xeon Gold 6312U, calculated using the standard error multiplied by the z-score of 1.96.

This completes the analysis. Although discussed here, Appendix A.9 formally shows how to piece the lemmas together in order to prove Theorem 4.2.

7. EXPERIMENTS

We experimentally evaluated the performance of our method by comparing it against two baselines: CascadeBandit from (Kveton et al., 2015) (in both the UCB and KL settings) and ComBand from (Cesa-Bianchi and Lugosi, 2012) for M-sized subsets, whose efficient implementation is described in Appendix D. We conducted our experiments in various synthetic settings. In each of these environments, a hidden vector $\boldsymbol{\theta} \in \mathbb{R}^K$ is maintained. For each $k \in [K]$, the entry θ_k represents the probability of obtaining a unit reward. These values can be viewed as *attraction probabilities*: the probability that a user clicks on the specific item. After presenting a subset of M elements, the learner gets a unit reward if any of the selected items returns a 1, and 0 otherwise. It is worth emphasizing that our model does not necessitate binary rewards; it offers the flexibility to accommodate any sum-max reward function (as discussed in Section 2). The use of a binary reward model is specifically required for comparisons with click models such as CascadeBandit.

Environments for the experiments. We experimentally evaluated our method in three different synthetic environments. We conducted experiments across a wide range of values for K, M, T, and for the probabilities associated with both optimal and suboptimal arms. In Figure 1, we display the cumulative reward over time obtained with $T = 10^5$, $K = 20$, $M = 3$ when the environments are set as follows:

1. **Stochastic** (Figure 1(a)): we randomly select M *good actions* to which we assign a reward probability of 0.3. The reward probabilities of the remaining $k - M$ arms are set to 0.1.

2. **Stochastic with adversarial corruptions** (Figure 1(b)): the rewards are generated as in the stochastic setting. However, in the first $\sqrt{T}$ rounds all good actions have a deterministic reward of 0.

3. **Worst-case stochastic** (Figure 1(c)): this setting is inspired by the lower bound of Cohen et al. (2017). Here the set $\mathcal{M} \subset [K]$ of M *good actions* is drawn uniformly at random. Then, for each $k \in [K]$, the probabilities are assigned as follows:

$$\theta_k = \begin{cases} X_k + \epsilon & \text{if } k \in \mathcal{M} \\ X_k & \text{otherwise}, \end{cases} \quad \text{where} \quad X_k \sim N\left(\frac{1}{2}, \sigma^2\right),$$

$$\sigma^2 = \frac{1}{192 + 96 \log T} \quad \text{and} \quad \epsilon = \sigma\sqrt{\frac{KM}{8T}}.$$

In Appendix E we present also results obtained varying the subset size M.

Results As expected, our most compelling results were achieved in the adversarial setting, where our approach demonstrated its superiority. In the two stochastic settings, we observed results that were on par with the established baseline methods, affirming the competitiveness of our proposed approach. These findings collectively underscore the effectiveness of our method, particularly in the challenging adversarial context, while also highlighting its versatility in stochastic scenarios. We emphasize that our method is the most efficient one, as each prediction only requires sampling M times from a probability distribution over the K available actions.

References

P. Auer, Y. Freund, and R. E. Schapire. The non-stochastic multi-armed bandit problem. 2001. URL https://api.semanticscholar.org/CorpusID:7732525.

N. Cesa-Bianchi and G. Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012.

A. Cohen, T. Hazan, and T. Koren. Tight bounds for bandit combinatorial optimization. In *Conference on Learning Theory*, pages 629–642. PMLR, 2017.

I. Dinur and D. Steurer. Analytical approach to parallel repetition. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 624–633, 2014.

D. Harvey and J. Van Der Hoeven. Integer multiplication in time o(nlog\,n). *Annals of Mathematics*, 193 (2):563–617, 2021.

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

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

T. Lattimore, B. Kveton, S. Li, and C. Szepesvari. Toprank: A practical algorithm for online stochastic ranking. *Advances in Neural Information Processing Systems*, 31, 2018.

R. Niazadeh, N. Golrezaei, J. R. Wang, F. Susan, and A. Badanidiyuru. Online learning via offline greedy algorithms: Applications in market design and optimization. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 737–738, 2021.

S. Pasteris, T. He, F. Vitale, S. Wang, and M. Herbster. Online learning of facility locations. In *Algorithmic Learning Theory*, pages 1002–1050. PMLR, 2021.

M. Simchowitz, K. Jamieson, and B. Recht. Best-of-k-bandits. In *Conference on Learning Theory*, pages 1440–1489. PMLR, 2016.

Y. Yue and C. Guestrin. Linear submodular bandits and their application to diversified retrieval. *Advances in Neural Information Processing Systems*, 24, 2011.

A. ANALYSIS PROOFS (PROOF OF THEOREM 4.2)

A.1. Lemma 6.3

Lemma A.1. *Given a function $r : 2^{[K]} \to \mathbb{R}$ there exists a unique subset decomposition $\widetilde{r}$ of r.*

Proof. For all $k \in [K] \cup \{0\}$ define $\mathcal{V}_k := \{\mathcal{S} \in 2^{[K]} \,|\, k \leq |\mathcal{S}|\}$. We take the inductive hypothesis that for all $k \in [K] \cup \{0\}$ there exists a unique function $\widetilde{r}_k : \mathcal{V}_k \to \mathbb{R}$ such that for all $\mathcal{Q} \in \mathcal{V}_k$ we have

$$r(\mathcal{Q}) = \sum_{\mathcal{S} \in \mathcal{V}_k} [\![\mathcal{Q} \subseteq \mathcal{S}]\!] \widetilde{r}_k(\mathcal{S}) \,.$$

We will prove the inductive hypothesis via reverse induction on k (i.e. from $k = K$ to $k = 0$).

The inductive hypothesis holds for $k = K$ since the only element of $\mathcal{V}_K$ is $[K]$ so we must have $\widetilde{r}_K([K]) := r([K])$

Now suppose, for some $i \in [K]$, the inductive hypothesis holds when $k = i$. Now consider the case that $k = i - 1$. Note that for all $\mathcal{Q} \in \mathcal{V}_i$ and $\mathcal{S} \in \mathcal{V}_{i-1} \setminus \mathcal{V}_i$ we must have that $\mathcal{Q} \not\subseteq \mathcal{S}$ and hence we must have that:

$$r(\mathcal{Q}) = \sum_{\mathcal{S} \in \mathcal{V}_i} [\![\mathcal{Q} \subseteq \mathcal{S}]\!] \widetilde{r}_{i-1}(\mathcal{S}) \,,$$

so, by the inductive hypothesis, the restriction of $\widetilde{r}_{i-1}$ to $\mathcal{V}_i$ is equal to $\widetilde{r}_i$. Now choose some arbitrary $\mathcal{Q} \in \mathcal{V}_{i-1} \setminus \mathcal{V}_i$ and define:

$$v(\mathcal{Q}) := \sum_{\mathcal{S} \in \mathcal{V}_i} [\![\mathcal{Q} \subseteq \mathcal{S}]\!] \widetilde{r}_{i-1}(\mathcal{S})$$

which, by above, is uniquely defined. Note that for all $\mathcal{S} \in \mathcal{V}_{i-1} \setminus \mathcal{V}_i$ we have that $\mathcal{Q} \subseteq \mathcal{S}$ if and only if $\mathcal{S} = \mathcal{Q}$ and hence we must have that:

$$r(\mathcal{Q}) = \sum_{\mathcal{S} \in \mathcal{V}_i} [\![\mathcal{Q} \subseteq \mathcal{S}]\!] \widetilde{r}_{i-1}(\mathcal{S}) + \widetilde{r}_{i-1}(\mathcal{Q}) = v(\mathcal{Q}) + \widetilde{r}_{i-1}(\mathcal{Q}) \,,$$

so that $\widetilde{r}_{i-1}(\mathcal{Q}) = r(\mathcal{Q}) - v(\mathcal{Q})$ which is unique.

We have hence shown that the inductive hypothesis holds for $k = i - 1$ and hence holds always. Noting that $\mathcal{V}_0 = 2^{[K]}$ we then get the result by necessarily setting $\widetilde{r} = \widetilde{r}_0$. □

A.2. Lemma 6.5

Lemma A.2. *For all $r : 2^K \to \mathbb{R}$ and all $\boldsymbol{q} \in \Delta_K$ we have $\Phi^r(\boldsymbol{q}) = \mathbb{E}[r(\mathcal{B}(\boldsymbol{q}))]$.*

Proof. Let $\widetilde{r}$ be a subset decomposition of r. We have

$$\begin{aligned}
\mathbb{E}[r(\mathcal{B}(\boldsymbol{q}))] &= \sum_{\mathcal{S} \subseteq [K]} \widetilde{r}(\mathcal{S}) \mathbb{P}[\mathcal{B}(\boldsymbol{q}) \subseteq \mathcal{S}] \\
&= \sum_{\mathcal{S} \subseteq [K]} \widetilde{r}(\mathcal{S}) \prod_{j \in [M]} \mathbb{P}[b_j(\boldsymbol{q}) \in \mathcal{S}] \\
&= \sum_{\mathcal{S} \subseteq [K]} \widetilde{r}(\mathcal{S}) \prod_{j \in [M]} \sum_{i \in [K]} [\![i \in \mathcal{S}]\!] q_i \\
&= \sum_{\mathcal{S} \subseteq [K]} \widetilde{r}(\mathcal{S}) \left(\sum_{i \in [K]} [\![i \in \mathcal{S}]\!] q_i \right)^M \\
&= \Phi^r(\boldsymbol{q})
\end{aligned}$$

as required. □

A.3. Lemma 6.6

Lemma A.3. *For all $r : 2^K \to \mathbb{R}$, all $\boldsymbol{q} \in \Delta_K$ and all $i \in [K]$ we have*

$$\partial_i \Phi^r(\boldsymbol{q}) = \mathbb{E}\left[\frac{r(\mathcal{B}(\boldsymbol{q}))}{q_i} \sum_{j \in [M]} [\![b_j(\boldsymbol{q}) = i]\!]\right].$$

Proof. Let $\widetilde{r}$ be a subset decomposition of r. For all $\boldsymbol{q}' \in \mathbb{R}^K$ and $\mathcal{S} \subseteq [K]$ define

$$\Lambda^{\mathcal{S}}(\boldsymbol{q}') := \left(\sum_{k \in [K]} [\![k \in \mathcal{S}]\!] q'_k\right)^M.$$

Fix some $j \in [M]$. Note that

$$\begin{aligned}
\partial_i \Lambda^{\mathcal{S}}(\boldsymbol{q}) &= M [\![i \in \mathcal{S}]\!] \left(\sum_{k \in [K]} [\![k \in \mathcal{S}]\!] q_k\right)^{M-1} \\
&= M [\![i \in \mathcal{S}]\!] \prod_{j' \in [M] \setminus \{j\}} \sum_{k \in [K]} [\![k \in \mathcal{S}]\!] q_k \\
&= M [\![i \in \mathcal{S}]\!] \prod_{j' \in [M] \setminus \{j\}} \mathbb{P}[b_{j'}(\boldsymbol{q}) \in \mathcal{S}] \\
&= \frac{M}{q_i} \mathbb{P}[b_j(\boldsymbol{q}) = i] [\![i \in \mathcal{S}]\!] \prod_{j' \in [M] \setminus \{j\}} \mathbb{P}[b_{j'}(\boldsymbol{q}) \in \mathcal{S}] \\
&= \frac{M}{q_i} \mathbb{P}[(b_j(\boldsymbol{q}) = i) \wedge (i \in \mathcal{S})] \prod_{j' \in [M] \setminus \{j\}} \mathbb{P}[b_{j'}(\boldsymbol{q}) \in \mathcal{S}] \\
&= \frac{M}{q_i} \mathbb{P}[(b_j(\boldsymbol{q}) = i) \wedge (b_j(\boldsymbol{q}) \in \mathcal{S})] \prod_{j' \in [M] \setminus \{j\}} \mathbb{P}[b_{j'}(\boldsymbol{q}) \in \mathcal{S}] \\
&= \frac{M}{q_i} \mathbb{P}[(b_j(\boldsymbol{q}) = i) \wedge (\forall j' \in [M], b_{j'}(\boldsymbol{q}) \in \mathcal{S})] \\
&= \frac{M}{q_i} \mathbb{P}[(b_j(\boldsymbol{q}) = i) \wedge (\mathcal{B}(\boldsymbol{q}) \subseteq \mathcal{S})] \\
&= \frac{M}{q_i} \mathbb{E}[[\![b_j(\boldsymbol{q}) = i]\!] [\![\mathcal{B}(\boldsymbol{q}) \subseteq \mathcal{S}]\!]],
\end{aligned}$$

so since:

$$\Phi^r(\boldsymbol{q}) = \sum_{\mathcal{S} \subseteq [K]} \widetilde{r}(\mathcal{S}) \Lambda^{\mathcal{S}}(\boldsymbol{q}),$$

we have

$$\begin{aligned}
\partial_i \Phi^r(\boldsymbol{q}) &= \sum_{\mathcal{S} \subseteq [K]} \widetilde{r}(\mathcal{S}) \partial_i \Lambda^{\mathcal{S}}(\boldsymbol{q}) \\
&= \frac{M}{q_i} \sum_{\mathcal{S} \subseteq [K]} \widetilde{r}(\mathcal{S}) \mathbb{E}[[\![b_j(\boldsymbol{q}) = i]\!] [\![\mathcal{B}(\boldsymbol{q}) \subseteq \mathcal{S}]\!]] \\
&= \frac{M}{q_i} \mathbb{E}\left[[\![b_j(\boldsymbol{q}) = i]\!] \sum_{\mathcal{S} \subseteq [K]} \widetilde{r}(\mathcal{S}) [\![\mathcal{B}(\boldsymbol{q}) \subseteq \mathcal{S}]\!]\right] \\
&= \frac{M}{q_i} \mathbb{E}[[\![b_j(\boldsymbol{q}) = i]\!] r(\mathcal{B}(\boldsymbol{q}))].
\end{aligned}$$

Summing over all $j \in [M]$ and dividing by M then gives us

$$\partial_i \Phi^r(\boldsymbol{q}) = \mathbb{E}\left[\frac{r(\mathcal{B}(\boldsymbol{q}))}{q_i} \sum_{j\in[M]} [\![b_j(\boldsymbol{q}) = i]\!]\right]$$

as required. □

A.4. Lemma 6.7

Lemma A.4. *For all pseudo-concave set functions $r : 2^K \to \mathbb{R}$ we have that Φ^r is concave over the simplex Δ_K.*

Proof. Choose any $\boldsymbol{q} \in \Delta_K$. Define $\langle b'_j(\boldsymbol{q}) \,|\, j \in [M-2]\rangle$ to be a sequence of stochastic quantities drawn i.i.d. at random from (the probability distribution characterised by) $\boldsymbol{q}$. In addition, let:

$$\mathcal{B}'(\boldsymbol{q}) := \{b'_j(\boldsymbol{q}) \,|\, j \in [M-2]\}\,.$$

Direct from the definition of Φ^r we have, for all $i, i' \in [K]$, that

$$\begin{aligned}
\partial_i \partial_{i'} \Phi^r &= \sum_{\mathcal{S}\subseteq[K]} \widetilde{r}(\mathcal{S}) [\![i \in \mathcal{S}]\!] [\![i' \in \mathcal{S}]\!] \left(\sum_{k\in[K]} [\![k \in \mathcal{S}]\!] q_k\right)^{M-2} \\
&= \sum_{\mathcal{S}\subseteq[K]} \widetilde{r}(\mathcal{S}) [\![i \in \mathcal{S}]\!] [\![i' \in \mathcal{S}]\!] \prod_{j\in[M-2]} \sum_{k\in[K]} [\![k \in \mathcal{S}]\!] q_k \\
&= \sum_{\mathcal{S}\subseteq[K]} \widetilde{r}(\mathcal{S}) [\![i \in \mathcal{S}]\!] [\![i' \in \mathcal{S}]\!] \prod_{j\in[M-2]} \mathbb{P}[b'_j(\boldsymbol{q}) \in \mathcal{S}] \\
&= \sum_{\mathcal{S}\subseteq[K]} \widetilde{r}(\mathcal{S}) [\![i \in \mathcal{S}]\!] [\![i' \in \mathcal{S}]\!] \mathbb{P}[\mathcal{B}'(\boldsymbol{q}) \subseteq \mathcal{S}] \\
&= \sum_{\mathcal{S}\subseteq[K]} \widetilde{r}(\mathcal{S}) \mathbb{P}[\mathcal{B}'(\boldsymbol{q}) \cup \{i, i'\} \subseteq \mathcal{S}] \\
&= \mathbb{E}\left[\sum_{\mathcal{S}\subseteq[K]} \widetilde{r}(\mathcal{S}) [\![\mathcal{B}'(\boldsymbol{q}) \cup \{i, i'\} \subseteq \mathcal{S}]\!]\right] \\
&= \mathbb{E}[r(\mathcal{B}'(\boldsymbol{q}) \cup \{i, i'\})] \\
&= \sum_{\mathcal{S}\subseteq[K]} \mathbb{P}[\mathcal{B}'(\boldsymbol{q}) = \mathcal{S}] r(\mathcal{S} \cup \{i, i'\}) \\
&= \sum_{\mathcal{S}\subseteq[K]} \mathbb{P}[\mathcal{B}'(\boldsymbol{q}) = \mathcal{S}] U^{r,\mathcal{S}}_{i,i'}\,.
\end{aligned}$$

So for all $\boldsymbol{x} \in \mathbb{R}^K$ with $\boldsymbol{x} \cdot \mathbf{1} = 0$ we have

$$\begin{aligned}
\boldsymbol{x}^\top (\nabla^2 \Phi^r(\boldsymbol{q})) \boldsymbol{x} &= \sum_{i,i'\in[K]} x_i (\partial_i \partial_{i'} \Phi^r) x_{i'} \\
&= \sum_{i,i'\in[K]} x_i x_{i'} \sum_{\mathcal{S}\subseteq[K]} \mathbb{P}[\mathcal{B}'(\boldsymbol{q}) = \mathcal{S}] U^{r,\mathcal{S}}_{i,i'} \\
&= \sum_{\mathcal{S}\subseteq[K]} \mathbb{P}[\mathcal{B}'(\boldsymbol{q}) = \mathcal{S}] \sum_{i,i'\in[K]} x_i U^{r,\mathcal{S}}_{i,i'} x_i \\
&= \sum_{\mathcal{S}\subseteq[K]} \mathbb{P}[\mathcal{B}'(\boldsymbol{q}) = \mathcal{S}] (\boldsymbol{x}^\top \boldsymbol{U}^{r,\mathcal{S}} \boldsymbol{x}) \\
&\leq 0\,,
\end{aligned}$$

which means that Φ^r is concave on Δ_K as required. □

A.5. Lemma 6.10

Lemma A.5. *Let $\mathcal{S} \subseteq [K]$ with $\mathcal{S} \neq \emptyset$, $r : 2^{[K]} \to \mathbb{R}$ be a pseudo-submodular function, and $\mathcal{Z} \subseteq [K]$ be a set formed by drawing M elements uniformly at random (with replacement) from $\mathcal{S}$. Then we have*

$$\mathbb{E}[r(\mathcal{Z}) - r(\emptyset)] \geq \left(1 - \left(\frac{|\mathcal{S}| - 1}{|\mathcal{S}|}\right)^M\right)(r(\mathcal{S}) - r(\emptyset))\,.$$

Proof. Without loss of generality assume that $r(\emptyset) = 0$.

We prove by induction on m that the lemma holds whenever $M \leq m$. In the case that $m = 0$ we have $\mathbb{E}[r(\mathcal{Z})] = r(\emptyset) = 0$ and $M = 0$ so the result holds. Now assume that it holds for all $M \leq m$ and consider the case that $M = m + 1$.

Since r is pseudo-submodular choose $i \in \mathcal{S}$ such that

$$r(\mathcal{Q} \cup \{i\}) - r(\mathcal{Q}) \geq r(\mathcal{S}) - r(\mathcal{S} \setminus \{i\}) \tag{5}$$

for all $\mathcal{Q} \subseteq \mathcal{S} \setminus \{i\}$. Define $\sigma := |\mathcal{S}|$ and

$$\phi := r(\mathcal{S}) - r(\mathcal{S} \setminus \{i\})\,. \tag{6}$$

Let $\langle z_s \,|\, s \in [M]\rangle$ be a sequence of M elements drawn uniformly at random from $\mathcal{S}$ such that $\mathcal{Z} = \{z_s \,|\, s \in [M]\}$. Define

$$\mu := \sum_{s \in [M]} [\![z_s \neq i]\!]\,.$$

For all $j \in [M] \cup \{0\}$ let $\mathcal{Z}_j$ be a set formed by sampling j actions independently and uniformly at random from $\mathcal{S} \setminus \{i\}$.

Note that by the inductive hypothesis, we have

$$\begin{aligned}
\mathbb{E}[[\![i \notin \mathcal{Z}]\!] r(\mathcal{Z})] &= \mathbb{P}[i \notin \mathcal{Z}]\mathbb{E}[r(\mathcal{Z}) \,|\, i \notin \mathcal{Z}] \\
&= \mathbb{P}[i \notin \mathcal{Z}]\left(1 - \left(\frac{|\mathcal{S} \setminus \{i\}| - 1}{|\mathcal{S} \setminus \{i\}|}\right)^M\right) r(\mathcal{S} \setminus \{i\}) \\
&= \mathbb{P}[i \notin \mathcal{Z}]\left(1 - \left(\frac{\sigma - 2}{\sigma - 1}\right)^M\right) r(\mathcal{S} \setminus \{i\}) \\
&= \mathbb{P}[\mu = M]\left(1 - \left(\frac{\sigma - 2}{\sigma - 1}\right)^M\right) r(\mathcal{S} \setminus \{i\})\,.
\end{aligned} \tag{7}$$

Note also that

$$\mathbb{E}[[\![i \in \mathcal{Z}]\!] r(\mathcal{Z})] = \sum_{j \in [m] \cup \{0\}} \mathbb{P}[\mu = j]\mathbb{E}[r(\mathcal{Z}_j \cup \{i\})]\,. \tag{8}$$

By equations (5) and (6) and the inductive hypothesis we have, for all $j \in [m] \cup \{0\}$, that

$$\begin{aligned}
\mathbb{E}[r(\mathcal{Z}_j \cup \{i\})] &\geq \mathbb{E}[\phi + r(\mathcal{Z}_j)] \\
&= \phi + \mathbb{E}[r(\mathcal{Z}_j)] \\
&\geq \phi + \left(1 - \left(\frac{|\mathcal{S} \setminus \{i\}| - 1}{|\mathcal{S} \setminus \{i\}|}\right)^j\right) r(\mathcal{S} \setminus \{i\}) \\
&= \phi + \left(1 - \left(\frac{\sigma - 2}{\sigma - 1}\right)^j\right) r(\mathcal{S} \setminus \{i\})\,.
\end{aligned} \tag{9}$$

We also have that

$$\sum_{j \in [m] \cup \{0\}} \mathbb{P}[\mu = j] = \mathbb{P}[i \in \mathcal{Z}]\,. \tag{10}$$

Substituting equations (9) and (10) into Equation (8) gives us

$$\mathbb{E}[\llbracket i \in \mathcal{Z} \rrbracket r(\mathcal{Z})] = \mathbb{P}[i \in \mathcal{Z}]\phi + \sum_{j \in [m] \cup \{0\}} \mathbb{P}[\mu = j] \left(1 - \left(\frac{\sigma - 2}{\sigma - 1}\right)^j\right) r(\mathcal{S} \setminus \{i\}) \,.$$

Adding this equation to Equation (7) gives us

$$\mathbb{E}[r(\mathcal{Z})] = \mathbb{P}[i \in \mathcal{Z}]\phi + \sum_{j \in [M] \cup \{0\}} \mathbb{P}[\mu = j] \left(1 - \left(\frac{\sigma - 2}{\sigma - 1}\right)^j\right) r(\mathcal{S} \setminus \{i\}) \,. \tag{11}$$

Take any $k \in \mathcal{S} \setminus \{i\}$. Note that

$$1 - \left(\frac{\sigma - 2}{\sigma - 1}\right)^j = 1 - (1 - 1/(\sigma - 1))^j = 1 - \mathbb{P}[k \notin \mathcal{Z}_j] = \mathbb{P}[k \in \mathcal{Z}_j] \,,$$

so that

$$\begin{aligned} \sum_{j \in [M] \cup \{0\}} \mathbb{P}[\mu = j] \left(1 - \left(\frac{\sigma - 2}{\sigma - 1}\right)^j\right) &= \sum_{j \in [M] \cup \{0\}} \mathbb{P}[\mu = j]\mathbb{P}[k \in \mathcal{Z}_j] \\ &= \sum_{j \in [M] \cup \{0\}} \mathbb{P}[\mu = j]\mathbb{P}[k \in \mathcal{Z} \setminus \{i\} \,|\, \mu = j] \\ &= \sum_{j \in [M] \cup \{0\}} \mathbb{P}[\mu = j]\mathbb{P}[k \in \mathcal{Z} \,|\, \mu = j] \\ &= \mathbb{P}[k \in \mathcal{Z}] \,. \end{aligned}$$

Substituting into Equation (11) gives us:

$$\begin{aligned} \mathbb{E}[r(\mathcal{Z})] &\geq \mathbb{P}[i \in \mathcal{Z}]\phi + \mathbb{P}[k \in \mathcal{Z}]r(\mathcal{S} \setminus \{i\}) \\ &= \mathbb{P}[i \in \mathcal{Z}](\phi + r(\mathcal{S} \setminus \{i\})) \\ &= \mathbb{P}[i \in \mathcal{Z}]r(\mathcal{S}) \\ &= (1 - \mathbb{P}[i \notin \mathcal{Z}])r(\mathcal{S}) \\ &= (1 - (1 - 1/\sigma)^M)r(\mathcal{S}) \\ &= \left(1 - \left(\frac{|\mathcal{S}| - 1}{|\mathcal{S}|}\right)^M\right) r(\mathcal{S}) \,. \end{aligned}$$

So the inductive hypothesis holds for all $M \in [m+1]$ and hence holds always. □

A.6. Lemma 6.11

Lemma A.6. *Given any $\mathcal{S} \subseteq [K]$ and any pseudo-submodular set function $r : 2^{[K]} \to \mathbb{R}$ we have*

$$\Phi^r(\boldsymbol{p}^{\mathcal{S}}) \geq r(\emptyset) + \left(1 - \left(\frac{|\mathcal{S}| - 1}{|\mathcal{S}|}\right)^M\right) (r(\mathcal{S}) - r(\emptyset)) \,.$$

Proof. Let $\mathcal{Z}$ be a set formed by drawing M elements i.i.d. with replacement from $\mathcal{S}$. Let z be an element drawn i.i.d. from $\mathcal{S}$. Let $\widetilde{r}$ be a subset-decomposition of r. Note that for all $i \in [K]$ we have

$$p_i^{\mathcal{S}} = \mathbb{P}[z = i] \,.$$

Hence, we have

$$\begin{aligned}
\Phi^r(\boldsymbol{p}^{\mathcal{S}}) &= \sum_{\mathcal{Q}\subseteq[K]} \widetilde{r}(\mathcal{Q}) \left(\sum_{i\in[K]} [\![i\in\mathcal{Q}]\!] p_i^{\mathcal{S}}\right)^M \\
&= \sum_{\mathcal{Q}\subseteq[K]} \widetilde{r}(\mathcal{Q}) \left(\sum_{i\in[K]} [\![i\in\mathcal{Q}]\!] \mathbb{P}[z=i]\right)^M \\
&= \sum_{\mathcal{Q}\subseteq[K]} \widetilde{r}(\mathcal{Q}) \mathbb{P}[z\in\mathcal{Q}]^M \\
&= \sum_{\mathcal{Q}\subseteq[K]} \widetilde{r}(\mathcal{Q}) \mathbb{P}[\mathcal{Z}\subseteq\mathcal{Q}] \\
&= \sum_{\mathcal{Q}\subseteq[K]} \widetilde{r}(\mathcal{Q}) \mathbb{E}[[\![\mathcal{Z}\subseteq\mathcal{Q}]\!]] \\
&= \mathbb{E}\left[\sum_{\mathcal{Q}\subseteq[K]} \widetilde{r}(\mathcal{Q}) [\![\mathcal{Z}\subseteq\mathcal{Q}]\!]\right] \\
&= \mathbb{E}[r(\mathcal{Z})]\,.
\end{aligned}$$

So

$$\Phi^r(\boldsymbol{p}^{\mathcal{S}}) - r(\emptyset) = \mathbb{E}[r(\mathcal{Z}) - r(\emptyset)]\,,$$

Lemma 6.10 then gives us the result. □

A.7. Lemma 6.13

Lemma A.7. *For all trials $t\in[T]$ and Ψ_t as defined in Definition 6.12, we have*

$$\nabla\Psi_t(\boldsymbol{p}_t) = \mathbb{E}[\boldsymbol{g}_t \,|\, \boldsymbol{p}_t]\,,$$

Proof. Take any $i\in[K]$. For any $j\in[M]$ we have

$$\begin{aligned}
c_{t,i} &= p_{t,i}c_{t,i}/p_{t,i} \\
&= \mathbb{P}[a_{t,j}=i \,|\, \boldsymbol{p}_t]c_{t,i}/p_{t,i} \\
&= \mathbb{E}[[\![a_{t,j}=i]\!]c_{t,i}/p_{t,i} \,|\, \boldsymbol{p}_t]\,.
\end{aligned}$$

So:

$$\begin{aligned}
Mc_{t,i} &= \sum_{j\in[M]} \mathbb{E}[[\![a_{t,j}=i]\!]c_{t,i}/p_{t,i} \,|\, \boldsymbol{p}_t] \\
&= \mathbb{E}\left[\frac{c_{t,j}}{p_{t,i}} \sum_{j\in[M]} [\![a_{t,j}=i]\!] \,\middle|\, \boldsymbol{p}_t\right]\,.
\end{aligned}$$

Hence, by Lemma 6.6, we have

$$\begin{aligned}
\partial_i\Psi_t(\boldsymbol{p}_t) &= \partial_i\Phi^{r_t}(\boldsymbol{p}_t) - Mc_{t,i} \\
&= \mathbb{E}\left[\frac{r_t(\mathcal{B}(\boldsymbol{p}_t))}{p_{t,i}} \sum_{j\in[M]} [\![b_j(\boldsymbol{p}_t)=i]\!]\right] - Mc_{t,i} \\
&= \mathbb{E}\left[\frac{r_t(\mathcal{A}_t)}{p_{t,i}} \sum_{j\in[M]} [\![a_{t,j}=i]\!] \,\middle|\, \boldsymbol{p}_t\right] - Mc_{t,i} \\
&= \mathbb{E}[g_{t,i} \,|\, \boldsymbol{p}_t]
\end{aligned}$$

as required. □

A.8. Lemma 6.14

Lemma A.8. *For all trials $t \in [T]$ we have*

$$\mathbb{E}\left[\sum_{i\in[K]} p_{t,i} g_{t,i}^2\right] \le (1+C)^2 M(K+M-1)\,.$$

Proof. Given $i \in [K]$ we have that

$$\begin{aligned}
\frac{\mathbb{E}[g_{t,i}^2]}{(1+C)^2} &= \frac{1}{(1+C)^2}\mathbb{E}\left[(r_t(\mathcal{A}_t) - c_{t,i})^2 \sum_{j,j'\in[M]} \frac{[\![a_{t,j}=i]\!][\![a_{t,j'}=i]\!]}{p_{t,i}^2}\right] \\
&\le \sum_{j,j'\in[M]} \mathbb{E}\left[\frac{[\![a_{t,j}=i]\!][\![a_{t,j'}=i]\!]}{p_{t,i}^2}\right] \\
&= \sum_{j\in[M]} \mathbb{E}\left[\frac{[\![a_{t,j}=i]\!]}{p_{t,i}^2}\right] + \sum_{j,j'\in[M]} [\![j\neq j']\!]\mathbb{E}\left[\frac{[\![a_{t,j}=i]\!][\![a_{t,j'}=i]\!]}{p_{t,i}^2}\right] \\
&= \sum_{j\in[M]} \frac{\mathbb{P}[a_{t,j}=i]}{p_{t,i}^2} + \sum_{j,j'\in[M]} [\![j\neq j']\!]\frac{\mathbb{P}[a_{t,j}=i]\mathbb{P}[a_{t,j'}=i]}{p_{t,i}^2} \\
&= \sum_{j\in[M]} \frac{1}{p_{t,i}} + \sum_{j,j'\in[M]} [\![j\neq j']\!] \\
&= \frac{M}{p_{t,i}} + M(M-1)\,,
\end{aligned}$$

and hence

$$\mathbb{E}\left[\sum_{i\in[K]} p_{t,i} g_{t,i}^2\right] = \sum_{i\in[K]} p_{t,i}\mathbb{E}[g_{t,i}^2] \le (1+C)^2 M(K+M-1)$$

as required. □

A.9. Theorem 4.2

Theorem 4.2. *Given r_t is pseudo-concave and pseudo-submodular for all $t \in [T]$, then for any set $\mathcal{S} \subseteq [K]$ with $\mathcal{S} \neq \emptyset$ we have*

$$\sum_{t\in[T]} \mathbb{E}[\psi_t] \ge \left(1-\alpha^M\right)\sum_{t\in[T]} \hat{r}_t(\mathcal{S}) - \frac{M}{|\mathcal{S}|}\sum_{t\in[T]} \gamma_t(\mathcal{S}) - R\,,$$

where

$$\alpha := 1 - \frac{1}{|\mathcal{S}|}.$$

Proof. Consider some trial $t \in [T]$. By Lemma 6.7 and the definition of Ψ_t we have that Ψ_t is concave over Δ_K. Hence, by Lemma 6.13, we have

$$\begin{aligned}
\mathbb{E}[(\boldsymbol{p}^{\mathcal{S}} - \boldsymbol{p}_t)\cdot \boldsymbol{g}_t \,|\, \boldsymbol{p}_t] &= (\boldsymbol{p}^{\mathcal{S}} - \boldsymbol{p}_t)\cdot \mathbb{E}[\boldsymbol{g}_t \,|\, \boldsymbol{p}_t] \\
&= (\boldsymbol{p}^{\mathcal{S}} - \boldsymbol{p}_t)\cdot \nabla\Psi_t(\boldsymbol{p}_t) \\
&\ge \Psi_t(\boldsymbol{p}^{\mathcal{S}}) - \Psi_t(\boldsymbol{p}_t)\,. \qquad (12)
\end{aligned}$$

Lemma 6.11 gives us:

$$\begin{aligned}\Psi_t(\boldsymbol{p}^{\mathcal{S}}) &= \Phi^{r_t}(\boldsymbol{p}^{\mathcal{S}}) - M \sum_{i\in[K]} p_i^{\mathcal{S}} c_{t,i} \\ &\geq r(\emptyset) + \left(1 - \left(\frac{|\mathcal{S}|-1}{|\mathcal{S}|}\right)^M\right)(r(\mathcal{S}) - r(\emptyset)) - \frac{M}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} c_{t,i} \end{aligned} \tag{13}$$

and Lemma 6.5 gives us:

$$\begin{aligned}\Psi_t(\boldsymbol{p}_t) &= \Phi^{r_t}(\boldsymbol{p}_t) - M \sum_{i\in[K]} p_{t,i} c_{t,i} \\ &= \mathbb{E}[r_t(\mathcal{B}(\boldsymbol{p}_t))] - M \sum_{i\in[K]} p_{t,i} c_{t,i} \\ &= \mathbb{E}[r_t(\mathcal{A}_t)\,|\,\boldsymbol{p}_t] - M \sum_{i\in[K]} p_{t,i} c_{t,i} \\ &= \mathbb{E}[r_t(\mathcal{A}_t)\,|\,\boldsymbol{p}_t] - \sum_{j\in[M]}\sum_{i\in[K]} \mathbb{P}[a_{t,j} = i\,|\,\boldsymbol{p}_t] c_{t,i} \\ &= \mathbb{E}[r_t(\mathcal{A}_t)\,|\,\boldsymbol{p}_t] - \sum_{j\in[M]} \mathbb{E}[c_{t,a_{t,j}}\,|\,\boldsymbol{p}_t] \\ &= \mathbb{E}[r_t(\mathcal{A}_t)\,|\,\boldsymbol{p}_t] - \mathbb{E}\left[\sum_{j\in[M]} c_{t,a_{t,j}} \,\middle|\, \boldsymbol{p}_t\right] \\ &\leq \mathbb{E}[r_t(\mathcal{A}_t)\,|\,\boldsymbol{p}_t] - \mathbb{E}\left[\sum_{i\in\mathcal{A}_t} c_{t,a_{t,j}} \,\middle|\, \boldsymbol{p}_t\right] \\ &= \mathbb{E}[\psi_t\,|\,\boldsymbol{p}_t] + r_t(\emptyset)\,. \end{aligned} \tag{14}$$

Substituting equations (13) and (14) into Equation (12) gives us:

$$\mathbb{E}[(\boldsymbol{p}^{\mathcal{S}} - \boldsymbol{p}_t)\cdot \boldsymbol{g}_t\,|\,\boldsymbol{p}_t] \geq -\mathbb{E}[\psi_t\,|\,\boldsymbol{p}_t] + \left(1 - \left(\frac{|\mathcal{S}|-1}{|\mathcal{S}|}\right)^M\right)\hat{r}_t(\mathcal{S}) - \frac{M}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} c_{t,i}$$

and hence:

$$\mathbb{E}[(\boldsymbol{p}^{\mathcal{S}} - \boldsymbol{p}_t)\cdot \boldsymbol{g}_t] \geq -\mathbb{E}[\psi_t] + \left(1 - \left(\frac{|\mathcal{S}|-1}{|\mathcal{S}|}\right)^M\right)\hat{r}_t(\mathcal{S}) - \frac{M}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} c_{t,i}\,. \tag{15}$$

Lemma 6.14 gives us:

$$\mathbb{E}\left[\sum_{i\in[K]} p_{t,i} g_{t,i}^2\right] \leq \frac{R^2}{T}\,, \tag{16}$$

Lemma 6.8 gives us:

$$\sum_{t\in[T]} \mathbb{E}[(\boldsymbol{p}^{\mathcal{S}} - \boldsymbol{p}_t)\cdot \boldsymbol{g}_t] \leq \frac{\ln(K)}{\eta} + \eta \sum_{t\in[T]} \mathbb{E}\left[\sum_{i\in[K]} p_{t,i} g_{t,i}^2\right]\,. \tag{17}$$

Substituting equations (15) and (16) into Equation (17) gives us:

$$-\sum_{t\in[T]} \mathbb{E}[\psi_t] + \left(1 - \left(\frac{|\mathcal{S}|-1}{|\mathcal{S}|}\right)^M\right)\sum_{t\in[T]} \hat{r}_t(\mathcal{S}) - \frac{M}{|\mathcal{S}|}\sum_{t\in[T]}\sum_{i\in\mathcal{S}} c_{t,i} \leq \frac{\ln(K)}{\eta} + \eta R^2\,.$$

Since $\eta = \ln(K)/R$ this implies the result. □

B. PROOF OF THEOREM 4.3

Theorem 4.3. *Suppose that there exists some $d \in \mathbb{N}$, $s \in (0,1)$, $\gamma > 1$, and a randomized polynomial time algorithm for the learner such that for all $K, M \in \mathbb{N}$ and for any instance of the multichannel advertising problem, it holds that $|\mathcal{A}_t| \leq M$ for all $t = 1, \ldots, T$ and, for any subset $\mathcal{S} \subseteq [K]$,*

$$\mathbb{E}\left[\sum_{t\in[T]} r_t(\mathcal{A}_t)\right] \geq \left(1 - \alpha^{\gamma M}\right) \sum_{t\in[T]} r_t(\mathcal{S}) - R',$$

where $R' \in \mathcal{O}(K^d T^s)$ and $\alpha := 1 - \frac{1}{|\mathcal{S}|}$. Then, for all $\varepsilon \in \left(0, 1 - 1/\gamma\right)$ and $B > 4^{1/((1-\varepsilon)\gamma - 1)}$, there exists a randomized polynomial-time algorithm for the set cover problem on $[B]$ that, with probability at least $\frac{1}{2}$, achieves approximation ratio at least $(1-\varepsilon)\ln(B)$.

Proof. Suppose we have such an algorithm. Let $c > 0$ and $\gamma > 1$ be such that

$$\mathbb{E}\left[\sum_{t\in[T]} r_t(\mathcal{A}_t)\right] \geq \left(1 - \left(\frac{|\mathcal{S}|-1}{|\mathcal{S}|}\right)^{\gamma M}\right) \sum_{t\in[T]} r_t(\mathcal{S}) - cK^d T^s. \tag{18}$$

Choose any $\rho \in (1/\gamma, 1)$ and then consider any $B \in \mathbb{N}$ such that $B > 4^{1/(\rho\gamma - 1)}$. Consider also any collection of sets $\{\mathcal{C}_k \mid k \in [K]\} \subseteq 2^{[B]}$ such that

$$\bigcup_{k\in[K]} \mathcal{C}_k = [B].$$

Let $\mathcal{S}$ be a subset of $[K]$ of minimum cardinality such that

$$\bigcup_{k\in\mathcal{S}} \mathcal{C}_k = [B].$$

Now choose

$$T := \left\lceil (4cK^d B)^{1/(1-s)} \right\rceil.$$

and choose any $M \in \mathbb{N}$ such that $M \geq \rho \ln(B)|\mathcal{S}|$. For all $t \in [T]$ draw $\mathcal{D}_t$ randomly as follows. First draw β_t uniformly at random from $[B]$ and then define

$$\mathcal{D}_t := \{k \in [K] \mid \beta_t \in \mathcal{C}_k\}.$$

It is a classic result that

$$\left(\frac{|\mathcal{S}|-1}{|\mathcal{S}|}\right)^{|\mathcal{S}|} \leq e^{-1}.$$

so by the conditions on B and M we have

$$\left(\frac{|\mathcal{S}|-1}{|\mathcal{S}|}\right)^{\gamma M} \leq \exp(-\gamma M/|\mathcal{S}|) = B^{-\rho\gamma} = \frac{B^{1-\rho\gamma}}{B} < \frac{1}{4B}. \tag{19}$$

By definition of $\mathcal{S}$ we have, for all $t \in [T]$, that there exists some $k \in \mathcal{S}$ such that $\beta_t \in \mathcal{C}_k$ so that $\mathcal{D}_t \cap \mathcal{S} \neq \emptyset$. This implies

$$\sum_{t\in[T]} r_t(\mathcal{S}) = T,$$

and hence, by (18) and (19), we have

$$\mathbb{E}\left[\sum_{t\in[T]} (1 - r_t(\mathcal{A}_t))\right] \leq T - T + \frac{T}{4B} + cK^d T^s \leq \frac{T}{4B} + \frac{TcK^d}{T^{1-s}} \leq \frac{T}{2B}. \tag{20}$$

Fix t and a realization of $\mathcal{A}_t$. If we have

$$\bigcup_{k\in\mathcal{A}_t} \mathcal{C}_k \neq [B]\,,$$

then we must also have that

$$\begin{aligned}\mathbb{E}[1 - r_t(\mathcal{A}_t) \mid \mathcal{A}_t] &= \mathbb{P}[\mathcal{A}_t \cap \mathcal{D}_t = \emptyset \mid \mathcal{A}_t] \\ &= \mathbb{P}[\forall\, k \in \mathcal{A}_t\,,\, \beta_t \notin \mathcal{C}_k \mid \mathcal{A}_t] \\ &= \mathbb{P}\left[\beta_t \notin \bigcup_{k\in\mathcal{A}_t} \mathcal{C}_k \,\middle|\, \mathcal{A}_t\right] \geq \frac{1}{B}\,.\end{aligned}$$

Hence, by taking the randomness of $\mathcal{A}_1, \ldots, \mathcal{A}_T$ into account,

$$\begin{aligned}&\mathbb{P}\left[\sum_{t\in[T]} [\![\bigcup_{k\in\mathcal{A}_t} \mathcal{C}_k \neq [B]]\!] = T\right] \\ &\leq \mathbb{P}\left[\mathbb{E}\left[\sum_{t\in[T]} \big(1 - r_t(\mathcal{A}_t)\big) \,\middle|\, \mathcal{A}_1, \ldots, \mathcal{A}_T\right] \geq \frac{T}{n}\right] \leq \frac{1}{2}\end{aligned}$$

by (20). Since T is polynomial in KB and $|\mathcal{A}_t| \leq M$, we have a randomized polynomial-time algorithm that, with probability at least $\frac{1}{2}$, solves the set cover problem on $[B]$ with approximation ratio $(1-\varepsilon)\ln(B)$ for $\varepsilon = 1 - \rho \in \big(0, 1 - 1/\gamma\big)$. □

C. SUBMODULAR MONOTONE NON-PSEUDOCONCAVE FUNCTIONS

We provide a function counterexample to show that there are submodular monotone functions which are not pseudoconcave.

Let $K = 8$, $\mathcal{P} = 2^{[K]}$, $\mathcal{S} = \{K\}$, and $\alpha > 0$. We define $U^{r,\mathcal{S}}$ as follows:

$$U^{r,\mathcal{S}} := \begin{pmatrix} 1 & 2 & 2 & 2 & 1+\alpha & 1+\alpha & 1+\alpha & 1 \\ 2 & 1 & 2 & 2 & 1+\alpha & 1+\alpha & 1+\alpha & 1 \\ 2 & 2 & 1 & 2 & 1+\alpha & 1+\alpha & 1+\alpha & 1 \\ 2 & 2 & 2 & 1 & 1+\alpha & 1+\alpha & 1+\alpha & 1 \\ 1+\alpha & 1+\alpha & 1+\alpha & 1+\alpha & 1 & 2 & 2 & 1 \\ 1+\alpha & 1+\alpha & 1+\alpha & 1+\alpha & 2 & 1 & 2 & 1 \\ 1+\alpha & 1+\alpha & 1+\alpha & 1+\alpha & 2 & 2 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{pmatrix}.$$

Now, let $\mathbf{x} = (1, 1, 1, 1, -1, -1, -1, -1)^\top$. Note that we have $\langle \mathbf{x}, \mathbf{1} \rangle = 0$ as required by the pseudoconcavity definition, and $\mathbf{x}^\top U^{r,\mathcal{S}} \mathbf{x} = 17 - 24\alpha$, which is positive for $\alpha \in \big(0, \frac{17}{24}\big)$, implying therefore the non-pseudoconcavity of r for such values of α.

We now show how to define r starting from $U^{r,\mathcal{S}}$ in such a way that it is both monotone and submodular while being therefore also non-pseudoconcave.

We have $|\mathcal{P}| = 2^K = 256$ possible subsets as the arguments of r, 29 of which are already defined by the above matrix $U^{r,\mathcal{S}}$:

- 1 subset ($\{K\}$) with cardinality 1,
- 7 subsets ($\{j, K\}_{j\in[K-1]}$) with cardinality 2,

- 21 subsets ($\{i, j, K\}_{1 \le j < i \le K-1}$) with cardinality 3.

For any $i \in [K]$, let δ_i and Δ_i be equal respectively to the minimum and the maximum difference (gain) over all values of r for subsets with cardinality i and all the ones for subsets with cardinality $i-1$. As previously anticipated, we construct function r starting from the above matrix $U^{r,\mathcal{S}}$ in such a way that for all $i \in [K-1]$, we have

$$\delta_i \ge \Delta_{i+1} \, ,$$

which is a sufficient condition for submodularity because, for all $i \in [K]$, *each* subset $S_i \in \mathcal{P}$ with cardinality i can be generated by adding one of its element *only* from a subset $S_{i-1} \subset S_i$ with cardinality $i-1$.

We set $\alpha = \frac{2}{3} < \frac{17}{24}$, which guarantees the non-pseudoconcavity of r. To ensure monotonicity and submodularity, we define

- $r(S_1) := 0$ for all subsets $S_1 \in \mathcal{P}$ with $|S_1| = 1$ (consistently with $U^{r,\mathcal{S}}_{K,K}$);
- $r(S_2) := 1$ for all subsets $S_2 \in \mathcal{P}$ with $|S_2| = 2$ (consistently with $U^{r,\mathcal{S}}_{K,j}, U^{r,\mathcal{S}}_{j,j}, U^{r,\mathcal{S}}_{j,1}$ for all $j \in [K-1]$);
- $r(S_3) := 1 + \frac{2}{3} = \frac{5}{3}$ for all subsets $S_3 \in \mathcal{P}$ with $|S_3| = 3$ that are not already defined by $U^{r,\mathcal{S}}$;
- $r(S_4) := r(S_3) + \frac{1}{2} = \frac{5}{3} + \frac{1}{2} = 2 + \frac{1}{6} > \max_{i,j} U^{r,\mathcal{S}}_{i,j} = 2$ for all subsets $S_4 \in \mathcal{P}$ with $|S_4| = 4$;
- $r(S_5) := r(S_4) + \frac{1}{6} = 2 + \frac{2}{6}$,
  $r(S_6) := r(S_5) + \frac{1}{6} = 2 + \frac{3}{6}$,
  $r(S_7) := r(S_6) + \frac{1}{6} = 2 + \frac{4}{6}$,
  $r(S_8) := r(S_7) + \frac{1}{6} = 2 + \frac{5}{6}$
  for all subsets $S_5, S_6, S_7, S_8 \in \mathcal{P}$ such that $|S_5| = 5$, $|S_6| = 6$, $|S_7| = 7$, $|S_8| = 8$.

Finally, we also set $r(\emptyset) = -1$. Note that, to ensure that submodularity is not violated, for each subset S_3^U with $|S_3^U| = 3$ defined by $U^{r,\mathcal{S}}$, we have that the difference $r(S_3^U) - r(S_2)$ for any subset $S_2 \subset S_3^U$ with $|S_2| = 2$ is either equal to $\alpha = \frac{2}{3}$ or 1, that is not smaller than the maximum difference $r(S_4) - r(S_3)$ over all $S_3, S_4 \in \mathcal{P}$ with $|S_3| = 3$ and $|S_4| = 4$, which in turn is equal to $\frac{1}{2} < \frac{2}{3}$. Furthermore, $r(S_4) = 2 + \frac{1}{6}$ is never smaller than any values of $r(S_3^U)$ for all subsets $S_3^U \in \mathcal{P}$ with $|S_3| = 3$ that are already defined by $U^{r,\mathcal{S}}$, because we have $r(S_3^U) \le 2$, thereby preserving monotonicity for all subsets in $\mathcal{P}$ with cardinality smaller or equal to 4.

Now, we recall that for any $i \in [K]$, δ_i and Δ_i are defined to be respectively equal to the minimum and the maximum difference (gain) over all values of r for subsets with cardinality i and all the ones for subsets with cardinality $i-1$. Since we have

- $\delta_1, \Delta_1, \delta_2, \Delta_2 = 1 \quad$ (which immediately implies $\Delta_2 \le \delta_1$),
- $\delta_3 = \frac{2}{3}; \quad \Delta_3 = 1 \le \delta_2$,
- $\delta_4 = \frac{1}{6}; \quad \Delta_4 = \frac{1}{2} \le \delta_3$,
- $\delta_5, \Delta_5, \delta_6, \Delta_6, \delta_7, \Delta_7, \delta_8, \Delta_8 = \frac{1}{6} \le \delta_4$,

then $\delta_i \ge \Delta_{i+1}$ for all $i \in [K-1]$ which guarantees the submodularity of r. Finally, it is immediate to verify that r is monotone also for all subsets in $\mathcal{P}$ with cardinality larger than 4. Hence, we conclude that r is monotone submodular and non-pseudoconcave.

□

D. EFFICIENT IMPLEMENTATION OF COMBAND

To implement the algorithm the ComBand presented in *(Cesa-Bianchi and Lugosi, 2012)*, it is necessary to devise an efficient method for sampling from a set whose size can be exponential in K. In fact, at each trial, given a set $\mathcal{S}$ of positive real numbers, we need to select any of the subsets S with a given size m from $\mathcal{S}$ with a probability proportional to the product of the elements contained in S itself.

To be consistent with the notation used in (Cesa-Bianchi and Lugosi, 2012), henceforth we use the symbol d in place of K.

Given a set $\mathcal{S} = \{q_1, q_2, \ldots, q_d\}$ of real positive numbers, we now show how to select a m-sized subset of indices $\{u_1, \ldots, u_m\}$ from $[d]$ with a probability proportional to $\Pi_{i=1}^m q_{u_i}$ by using dynamic programming. The running time of this sampling method is always linear[2] in $m \cdot d$.

For each sampling operation, consider the sequence of element indices $u_1, u_2, \ldots, u_m$ ordered according to the elements in $[d]$, i.e., $u_i < u_{i+1}$ for all $i \in [m-1]$.

The main idea of this method is to sample first u_m, and then $u_{m-1}, \ldots, u_1$ (i.e., in reverse order) having derived in a preliminary phase via dynamic programming all the probabilities that $u_m = j$ for all $m \le j \le d$, and the conditional probabilities that $u_{m'} = j$ given that $u_{m'+1} = j'$, for all $m' \in [m-1]$ and $m' \le j < j' \le d - m + m'$.

We denote the conditional probability that $u_{m'} = j$ given that $u_{m'+1} = j'$, where $m' \in [m-1]$ and $m' \le j < j' \le d - m + m'$ by

$$P_{m',j|j'} := \mathbb{P}(u_{m'} = j | u_{m'+1} = j')\,,$$

and, for the selection of u_m, we define for all $j \in [d]$

$$P_{m,j} := \mathbb{P}(u_m = j)\,,$$

because there is no element $u_{j'} > u_m$ (with $j' > m$) in the sequence of selected indices from $[d]$. We clearly have $\sum_{j=m'}^{j'-1} P_{m',j|j'} = 1$ and $\sum_{j=m}^{d} P_{m,j} = 1$.

For each $m' \in [m]$ and $m' \le j \le d - m + m'$ let $z_{m',j}$, be the the sum of the products of numbers of $\mathcal{S}$ with indices $u_1, u_2, \ldots, u_{m'}$ contained in each m'-sized subset of $[j]$ such that $u_{m'} = j$. We define $Z_{m',k} := \sum_{i=m'}^{k} z_{m',i}$ for any integer k such that $m' \le k \le d - m + m'$. Thus, for all $m' \in [m-1]$ and $m' \le j < j' \le d - m + m'$ we have

$$P_{m',j|j'} = \frac{z_{m',j}}{Z_{m',j'-1}}\,.$$

Analogously, for the selection of u_m, for all $m \le j \le d$ we can write

$$P_{m,j} = \frac{z_{m,j}}{Z_{m,d}}\,.$$

Hence, once we obtain $z_{m',j}$ and $Z_{m',j'-1}$ for all $m' \in [m-1]$ and $m' \le j < j' \le d - m + m'$, $z_{m,j}$ for all $m \le j \le d$, and $Z_{m,d}$, we can immediately compute the desired probabilities to sample $u_m, u_{m-1}, \ldots, u_1$ in this (reverse) order.

We now show how to calculate these values. To this goal, since $Z_{m',k} := \sum_{i=m'}^{k} z_{m',i}$, we only need to show how to compute the values appearing at the numerator in the above probability formulas.

The possibility to efficiently the above probabilities is given by the following observation:

[2] We assume that multiplying two numbers requires a constant time. Removing this assumption, since it is known that it is possible to multiply two numbers represented by at most m bits in time equal to $\widetilde{\mathcal{O}}(m)$ when $m \gg 1$ (Harvey and Van Der Hoeven, 2021), the total sampling time would be $\widetilde{\mathcal{O}}(m^2 d)$ instead of $\mathcal{O}(md)$.

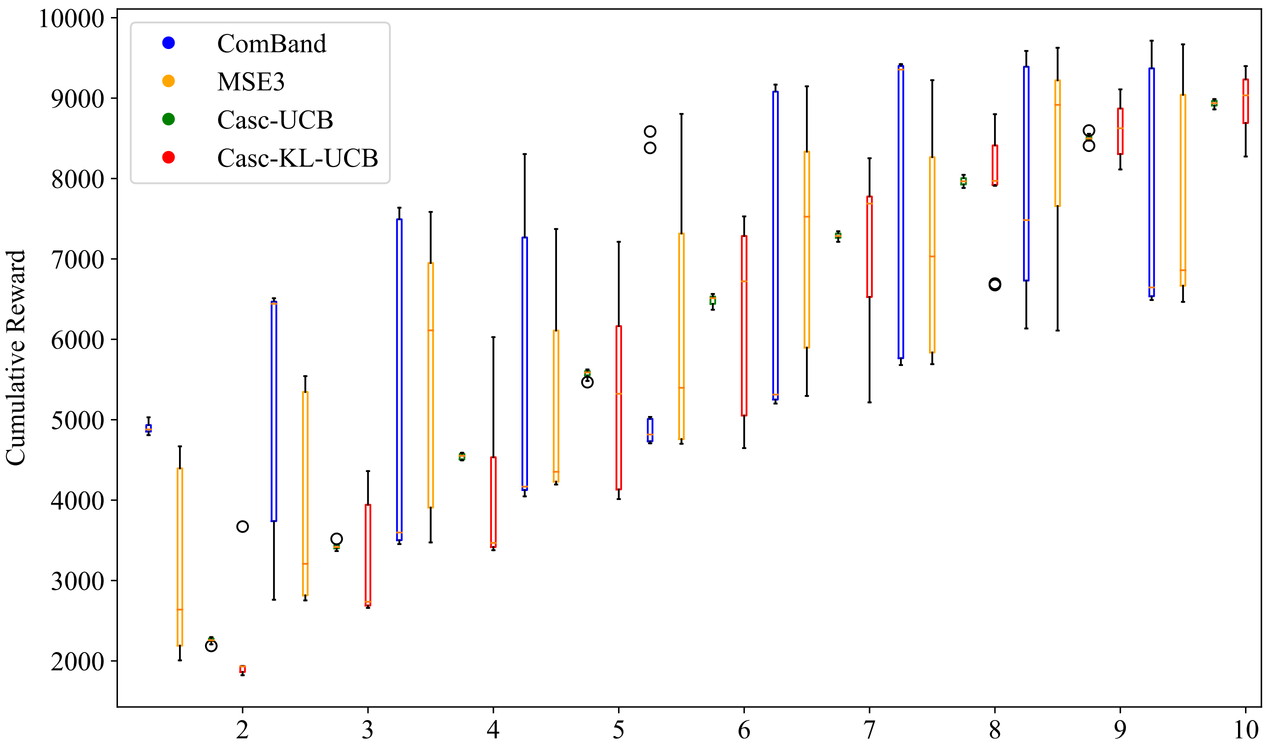


Figure 2. Stochastic environment, cumulative reward with respect to the M parameter

$$z_{m',j} = Z_{m'-1,j-1} \cdot q_j \, .$$

Note that $Z_{m'-1,j-1}$ can be in turn defined in terms of $z_{m'-1,m'-1}$, $z_{m'-1,m'}$, $z_{m'-1,m'+1}, \ldots, z_{m'-1,j-2}$, $z_{m'-1,j-1}$. This recurrence relation allows us to compute all these values once we know $z_{1,1}, z_{1,2}, \ldots, z_{1,d}$. Since we clearly have $z_{1,j} = q_j$ for all $j \in [d]$, we can therefore compute all these values and the above probabilities to efficiently accomplish this sampling operation by finding the indices $u_m, u_{m-1}, \ldots, u_1$ in this order. It is immediate to verify that both the number of sum and multiplication operations are equal to $\Theta(md)$.

E. ADDITIONAL EXPERIMENTAL RESULTS

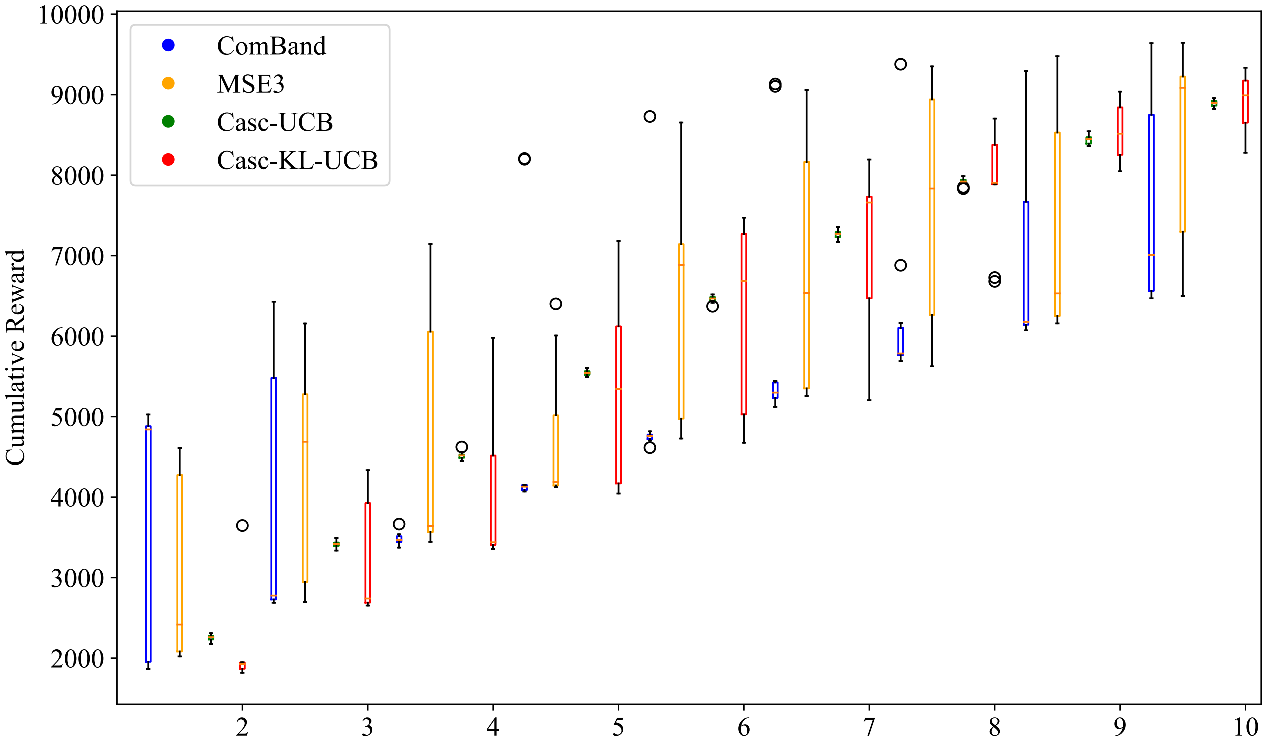


Figure 3. Stochastic with adversarial corruptions environment, cumulative reward with respect to the M parameter

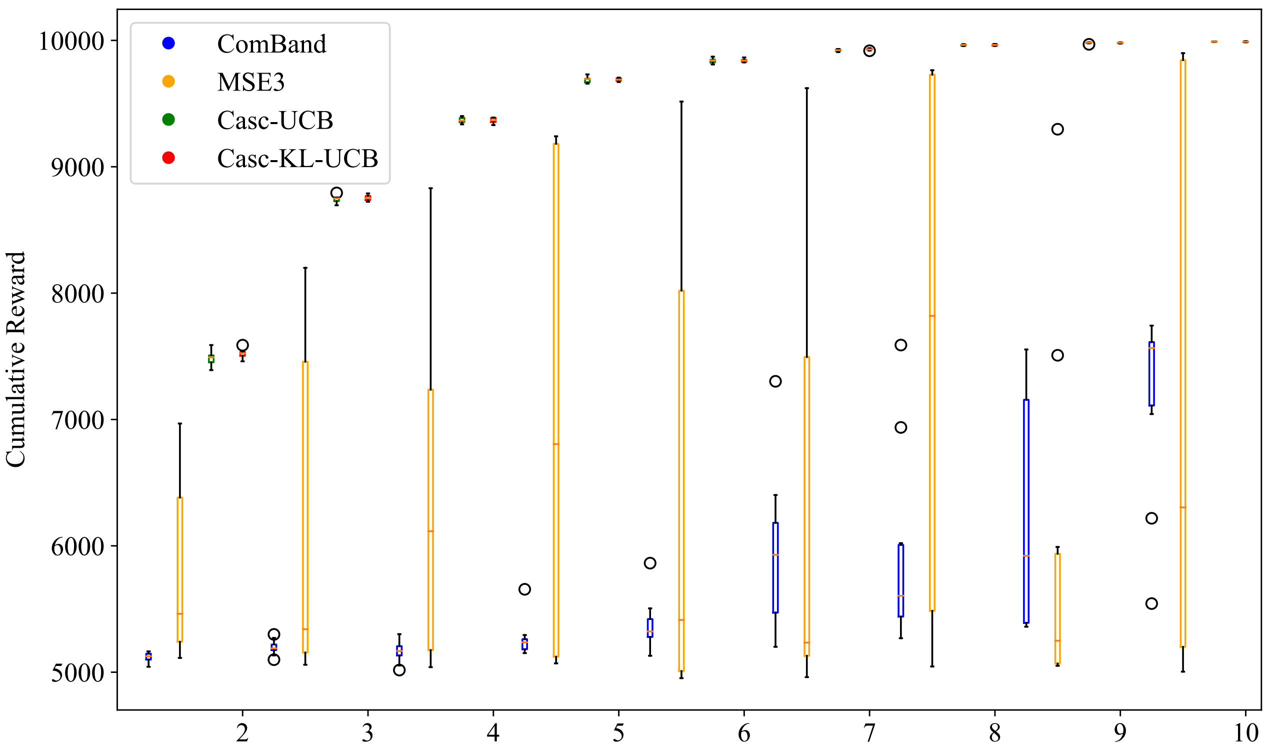


Figure 4. Worst case stochastic environment, cumulative reward with respect to the M parameter