Nested Bandits

Matthieu Martin [1] **Panayotis Mertikopoulos** [2 1] **Thibaud Rahier** [1] **Houssam Zenati** [1 3]

Abstract

In many online decision processes, the optimizing agent is called to choose between large numbers of alternatives with many inherent similarities; in turn, these similarities imply closely correlated losses that may confound standard discrete choice models and bandit algorithms. We study this question in the context of *nested bandits*, a class of adversarial multi-armed bandit problems where the learner seeks to minimize their regret in the presence of a large number of distinct alternatives with a hierarchy of embedded (non-combinatorial) similarities. In this setting, optimal algorithms based on the exponential weights blueprint (like Hedge, EXP3, and their variants) may incur significant regret because they tend to spend excessive amounts of time exploring irrelevant alternatives with similar, suboptimal costs. To account for this, we propose a *nested exponential weights* (NEW) algorithm that performs a layered exploration of the learner's set of alternatives based on a nested, step-by-step selection method. In so doing, we obtain a series of tight bounds for the learner's regret showing that online learning problems with a high degree of similarity between alternatives can be resolved efficiently, without a red bus / blue bus paradox occurring.

1. Introduction

Consider the following discrete choice problem (known as the "red bus / blue bus paradox" in the context of transportation economics). A commuter has a choice between taking a car or bus to work: commuting by car takes on average half an hour modulo random fluctuations, whereas commuting by bus takes an hour, again modulo random fluctuations (it's a long commute). Then, under the classical multinomial logit choice model for action selection [20, 21], the commuter's odds for selecting a car over a bus would be $\exp(-1/2)/\exp(-1) \approx 1.6 : 1$. This indicates a very clear preference for taking a car to work and is commensurate with the fact that, on average, commuting by bus takes twice as long.

Consider now the same model but with a twist. The company operating the bus network purchases a fleet of new buses that are otherwise completely identical to the existing ones, except for their color: old buses are red, the new buses are blue. This change has absolutely no effect on the travel time of the bus; however, since the new set of alternatives presented to the commuter is $\{\text{car}, \text{red bus}, \text{blue bus}\}$, the odds of selecting a car over a bus (red or blue, it doesn't matter) now drops to $\exp(-1/2)/[\exp(-1) + \exp(-1)] \approx 0.8 : 1$. Thus, by introducing an *irrelevant* feature (the color of the bus), the odds of selecting the alternative with the highest utility have dropped dramatically, to the extent that commuting by car is no longer the most probable outcome in this example.

Of course, the shift in choice probabilities may not always be that dramatic, but the point of this example is that the presence of an irrelevant alternative (the blue bus) would always induce such a shift – which is, of course, absurd. In fact, the red bus / blue bus paradox was originally proposed as a sharp criticism of the independence from irrelevant alternatives (IIA) axiom that underlies the multinomial logit choice model [20] and which makes it unsuitable for choice problems with inherent similarities between different alternatives. In turn, this has led to a vast corpus of literature in social choice and decision theory, with an extensive array of different axioms and models proposed to overcome the failures of the IIA assumption. For an introduction to the topic, we refer the reader to the masterful accounts of McFadden [21], Ben-Akiva & Lerman [7] and Anderson et al. [2].

Perhaps surprisingly, the implications of the red bus / blue bus paradox have not been explored in the context of online learning, despite the fact that similarities between alternatives are prevalent in the field's application domains – for example, in recommender systems with categorized product recommendation catalogues, in the economics of transport and product differentiation, etc. What makes this gap particularly pronounced is the fact that logit choice underlies some

All authors in alphabetical order. [1]Criteo AI Lab [2]Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000 Grenoble, France [3]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France. Correspondence to: Panayotis Mertikopoulos <panayotis.mertikopoulos@imag.fr>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

of the most widely used algorithmic schemes for learning in multi-armed bandit problems – namely the exponential weights algorithm for exploration and exploitation (EXP3) [4, 19, 29] as well as its variants, Hedge [5], EXP3.P [6], EXP3-IX [17], EXP4 [6] / EXP4-IX [23], etc. Thus, given the vulnerability of logit choice to irrelevant alternatives, it stands to reason that said algorithms may be suboptimal when faced with a set of alternatives with many inherent similarities.

Our contributions. Our paper examines this question in the context of repeated decision problems where a learner seeks to minimize their regret in the presence of a large number of distinct alternatives with a hierarchy of embedded (non-combinatorial) similarities. This similarity structure, which we formalize in Section 2, is defined in terms of a nested series of attributes – like "type" or "color" – and induces commensurate similarities to the losses of alternatives that lie in the same class (just as the red and blue buses have identical losses in the example described above).

Inspired by the nested logit choice model introduced by McFadden [21] to resolve the original red bus / blue bus paradox, we develop in Section 3 a *nested exponential weights* (NEW) algorithm for no-regret learning in decision problems of this type. Our main result is that the regret incurred by NEW is bounded as $\mathcal{O}(\sqrt{n_{\text{eff}} \log n \cdot T})$, where n is the total number of alternatives and n_{eff} is the "effective" number when taking similarities into account (for example, in the standard red bus / blue bus paradox, $n_{\text{eff}} = 2$, cf. Section 4). The gap between nested and non-nested algorithms can be quantified by the problem's *price of affinity* (PoAf), defined here as the ratio $\alpha = \sqrt{n/n_{\text{eff}}}$ measuring the worst-case ratio between the regret guarantees of the NEW and EXP3 algorithms (the latter scaling as $\mathcal{O}(\sqrt{n \log n \cdot T})$ in the problem at hand).

In practical applications (such as the type of recommendation problems that arise in online advertising), α can be exponential in the number of attributes, indicating that the NEW algorithm could lead to significant performance gains in this context. We verify that this is indeed the case in a range of synthetic experiments in Section 5.

Related Work. The problem of exploiting the structure of the loss model and/or any side information available to the learner is a staple of the bandit literature. More precisely, in the setting of contextual bandits, the learner is assumed to observe some "context-based" information and tries to learn the "context to reward" mapping underlying the model in order to make better predictions. Bandit algorithms of this type – like EXP4 – are often studied as "expert" models [6, 11] or attempt to model the agent's loss function with a semi-parametric contextual dependency in the stochastic setting to derive optimistic action selection rules [1]; for a survey, we refer the reader to [18] and references therein. While the nested bandit model we study assumes an additional layer of information relative to standard bandit models, there are no experts or a contextual mapping conditioning the action taken, so it is not comparable to the contextual setup.

The type of feedback we consider assumes that the learner observes the "intra-class" losses of their chosen alternative, similar to the semi-bandit in the study of combinatorial bandit algorithms [12, 15]. However, the similarity with combinatorial bandit models ends there: even though the categorization of alternatives gives rise to a tree structure with losses obtained at its leaves, there is no combinatorial structure defining these costs, and modeling this as a combinatorial bandit would lead to the same number of arms and ground elements, thus invalidating the concept.

Besides these major threads in the literature, [28] recently showed that the range of losses can be exploited with an additional free observation, while [13] improves the regret guarantees by using effective loss estimates. However, both works are susceptible to the advent of irrelevant alternatives and can incur significant regret when faced with such a problem. Finally, in the Lipschitz bandit setting, [14, 16] obtain order-optimal regret bounds by building a hierarchical covering model in the spirit of [10]; the correlations induced by a Lipschitz loss model cannot be compared to our model, so there is no overlap of techniques or results.

2. The general model

We begin in this section by defining our general nested choice model. Because the technical details involved can become cumbersome at times, it will help to keep in mind the running example of a music catalogue where songs are classified by, say, genre (classical music, jazz, rock,...), artist (Rachmaninov, Miles Davis, Led Zeppelin,...), and album. This is a simple – but not simplistic – use case which requires the full capacity of our model, so we will use it as our "go-to" example throughout.

2.1. Attributes, classes, and the relations between them

Let $\mathcal{A} = \{a_i : i = 1, \dots, n\}$ be a set of *alternatives* (or *atoms*) indexed by $i = 1, \dots, n$. A *similarity structure* (or *structure of attributes*) on $\mathcal{A}$ is defined as a tower of nested *similarity partitions* (or *attributes*) $\mathcal{S}_\ell$, $\ell = 0, \dots, L$, of $\mathcal{A}$ with $\{\mathcal{A}\} =: \mathcal{S}_0 \succcurlyeq \mathcal{S}_1 \succcurlyeq \cdots \succcurlyeq \mathcal{S}_L := \{\{a\} : a \in \mathcal{A}\}$. As a result of this definition, each partition $\mathcal{S}_\ell$ captures successively finer attributes of the elements of $\mathcal{A}$ (in our music catalogue example, these attributes would correspond to genre, artist, album, etc.).[1] Accordingly, each constituent

[1] The trivial partitions $\mathcal{S}_0 = \{\mathcal{A}\}$ and $\mathcal{S}_L = \{\{a\} : a \in \mathcal{A}\}$ do not carry much information in themselves, but they are included for completeness and notational convenience later on.

set A of a partition $\mathcal{S}_\ell$ will be referred to as a *similarity class* and we assume it collects all elements of $\mathcal{A}$ that share the attribute defining $\mathcal{S}_\ell$: for example, a similarity class for the attribute "artist" might consist of all Beethoven symphonies, all songs by Led Zeppelin, etc.

Collectively, a structure of attributes will be represented by the disjoint union

$$\mathcal{S} \coloneqq \coprod_{\ell=0}^{L} \mathcal{S}_\ell \equiv \bigcup_{\ell=0}^{L} \{(A,\ell) : A \in \mathcal{S}_\ell\} \qquad (1)$$

of all class/attribute pairs of the form (A,ℓ) for $A \in \mathcal{S}_\ell$. In a slight abuse of terminology (and when there is no danger of confusion), the pair $S = (A,\ell)$ will also be referred to as a "class", and we will write $S \in \mathcal{S}_\ell$ and $a \in S$ instead of $A \in \mathcal{S}_\ell$ and $a \in A$ respectively. By contrast, when we need to clearly distinguish between a class and its underlying set, we will write $A = \mathrm{elem}(S)$ for the set of atoms contained in S and $\ell = \mathrm{attr}(S)$ for the attached attribute label.

Remark 1. The reason for including the attribute label ℓ in the definition of $\mathcal{S}$ is that a set of alternatives may appear in different partitions of $\mathcal{A}$ in a different context. For example, if "IV" is the only album by Led Zeppelin in the catalogue, the album's track list represents both the set of "all songs in IV" as well as the set of "all Led Zeppelin songs". However, the focal attribute in each case is different – "artist" in the former versus "album" in the latter – and this additional information would be lost in the non-discriminating union $\bigcup_{\ell=0}^{L} \mathcal{S}_\ell$ (unless, of course, the partitions $\mathcal{S}_\ell$ happen to be mutually disjoint, in which case the distinction between "union" and "disjoint union" becomes set-theoretically superfluous). ¶

Moving forward, if a class $S \in \mathcal{S}_\ell$ contains the class $S' \in \mathcal{S}_k$ for some $k > \ell$, we will say that S' is a *descendant* of S (resp. S is an *ancestor* of S'), and we will write "$S' \prec S$" (resp. "$S \succ S'$").[2] As a special case of this relation, if $S' \prec S$ and $k = \ell + 1$, we will say that S' is a *child* of S (resp. S is *parent* of S') and we will write "$S' \lhd S$" (resp. "$S \rhd S'$"). For completeness, we will also say that S' and S'' are *siblings* if they are children of the same parent, and we will write $S' \sim S''$ in this case. Finally, when we wish to focus on descendants sharing a certain attribute, we will write "$S' \prec_\ell S$" as shorthand for the predicate "$S' \prec S$ and $\mathrm{attr}(S') = \ell$".

Building on this, a similarity structure on $\mathcal{A}$ can also be represented graphically as a rooted directed tree – an *arborescence* – by connecting two classes $S, S' \in \mathcal{S}$ with a directed edge $S \to S'$ whenever $S \rhd S'$. By construction, the root of this tree is $\mathcal{A}$ itself,[3] and the unique directed path $\mathcal{A} \equiv S_0 \rhd S_1 \rhd \cdots \rhd S_\ell \equiv S$ from $\mathcal{A}$ to any class $S \in \mathcal{S}$ will be referred to as the *lineage* of S. For notational simplicity, we will not distinguish between $\mathcal{S}$ and its graphical representation, and we will use the two interchangeably; for an illustration, see Fig. 1.

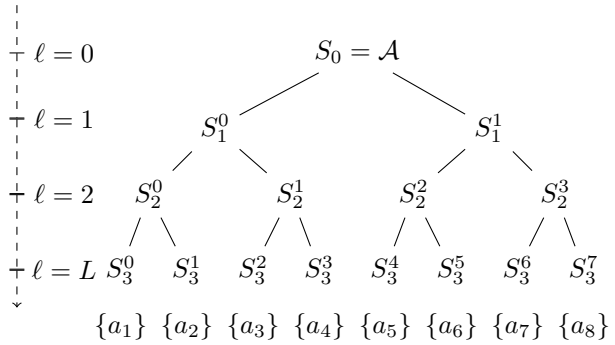

Figure 1: A structure with $L = 3$ attributes on the set $\mathcal{A} = \{a_1, \ldots, a_8\}$; for example, the class S_2^1 consists of $\{a_3, a_4\}$.

2.2. The loss model

Throughout what follows, we will consider loss models in which alternatives that share a common set of attributes incur similar costs, with the degree of similarity depending on the number of shared attributes. More precisely, given a similarity class $S \in \mathcal{S}$, we will assume that all its immediate subclasses S' share the same base cost c_S (determined by the parent class S) plus an idiosyncratic cost increment $r_{S'}$ (which is specific to the child $S' \lhd S$ in question). Formally, starting with $c_{\mathcal{A}} = 0$ (for the root class $\mathcal{A}$), this boils down to the recursive definition

$$c_{S'} = c_S + r_{S'} \quad \text{for all } S' \lhd S, \qquad (2)$$

which, when unrolled over the lineage $\mathcal{A} \equiv S_0 \rhd S_1 \rhd \cdots \rhd S_\ell \equiv S$ of a target class $S \in \mathcal{S}_\ell$, yields the expression

$$c_S = \sum\nolimits_{S' \succcurlyeq S} r_{S'} = r_{S_1} + \cdots + r_{S_\ell}. \qquad (3)$$

Thus, in particular, when $S \leftarrow a \in \mathcal{A}$, the cost assigned to an individual alternative $a \in \mathcal{A}$ will be given by

$$c_a = \sum\nolimits_{\ell=1}^{L} r_{S_\ell} = \sum\nolimits_{S \ni a} r_S \quad \text{for all } a \in \mathcal{A}. \qquad (4)$$

Finally, to quantify the "intra-class" variability of costs, we will assume throughout that the idiosyncratic cost increments within a given parent class S are bounded as

$$r_{S'} \in [0, R_S] \quad \text{for all } S' \lhd S. \qquad (5)$$

[2] More formally, we will write $S' \prec S$ when $\mathrm{elem}(S') \subseteq \mathrm{elem}(S)$ and $\mathrm{attr}(S') > \mathrm{attr}(S)$. The corresponding weak relation "$\preccurlyeq$" is defined in the standard way, i.e., allowing for the case $\mathrm{attr}(S') = \mathrm{attr}(S)$ which in turn implies that $S' = S$.

[3] Stricto sensu, the root of the tree is $(\mathcal{A}, 0)$, but since there is no danger of confusion, the attribute label "0" will be dropped.

This terminology is justified by the fact that, under the loss model (2), the costs $c_{S'}, c_{S''}$ to any two *sibling* classes $S', S'' \lhd S$ (i.e., any two classes parented by S) differ by at most R_S. Analogously, the costs to any two alternatives $a, a' \in \mathcal{A}$ that share a set of common attributes $S_1, \dots, S_\ell$ will differ by at most $\sum_{k=\ell+1}^{L} R_{S_k}$.

Example 1. To represent the original red bus / blue bus problem as an instance of the above framework, let $\mathcal{S}_1 = \{\{\text{red bus}, \text{blue bus}\}, \text{car}\}$ be the partition of the set $\mathcal{A} = \{\text{red bus}, \text{blue bus}, \text{car}\}$ by type ("`bus`" or "`car`"), and let $\mathcal{S}_2$ be the corresponding sub-partition by color ("`red`" or "`blue`" for elements of the class "`bus`"). The fact that color does not affect travel times may then be represented succinctly by taking $R_{\text{color}} = 0$ (representing the fact that color does not affect travel times). ¶

Remark 2. We make no distinction here between c_a and $c_{\{a\}}$, i.e., between an alternative a of $\mathcal{A}$ and the (unique) singleton class of $\{a\} \in \mathcal{S}_L$ containing it. This is done purely for reasons of notational convenience. ¶

Remark 3. For posterity, we also note that the optimizing agent is assumed to be aware of the cost decomposition (4) after selecting an alternative $a \in \mathcal{A}$. In the context of combinatorial bandits [12] this would correspond to the so-called "semi-bandit" setting. ¶

2.3. Sequence of events

With all this in hand, we will consider a generic online decision process that unfolds over a set of alternatives $\mathcal{A}$ endowed with a similarity structure $\mathcal{S} = \coprod_\ell \mathcal{S}_\ell$ as follows:

1. At each stage $t = 1, 2, \dots$, the learner selects an alternative $a_t \in \mathcal{A}$ by selecting attributes from $\mathcal{S}$ one-by-one.
2. Concurrently, nature sets the idiosyncratic, intra-class losses $r_{S,t}$ for each similarity class $S \in \mathcal{S}$.
3. The learner incurs $r_{S,t}$ for each chosen class $S \ni a_t$ for a total cost of $c_t = \sum_{S \ni a_t} r_{S,t}$, and the process repeats.

To align our presentation with standard bandit models with losses in $[0, 1]$, we will assume throughout that $\sum_{S \ni a} R_S \leq 1$ for all $a \in \mathcal{A}$, meaning in particular that the maximal cost incurred by any alternative $a \in \mathcal{A}$ is upper bounded by 1. Other than this normalization, the sequence of idiosyncratic loss vectors $r_t \in \mathbb{R}^{\mathcal{S}}$, $t = 1, 2, \dots$, is assumed arbitrary and unknown to the learner as per the standard adversarial setting [11, 26].

To avoid deterministic strategies that could be exploited by an adversary, we will assume that the learner selects an alternative a_t at time t based on a mixed strategy $x_t \in \Delta(\mathcal{A})$, i.e., $a_t \sim x_t$. The regret of a policy x_t, $t = 1, 2, \dots$, against a benchmark strategy $p \in \Delta(\mathcal{A})$ is then defined as the cumulative difference between the player's mean cost under p and x_t, that is

$$\text{Reg}_p(T) = \sum_{t=1}^{T} [\mathbb{E}_{x_t}[c_{a_t,t}] - \mathbb{E}_p[c_{a_t,t}]] = \sum_{t=1}^{T} \langle c_t, x_t - p \rangle \tag{6}$$

where $c_t = (c_{a,t})_{a \in \mathcal{A}} \in \mathbb{R}^{\mathcal{A}}$ denotes the vector of costs encountered by the learner at time t, i.e., $c_{a,t} = \sum_{S \ni a} r_{S,t}$ for all $a \in \mathcal{A}$. This definition will be our main figure of merit in the sequel.

3. The nested exponential weights algorithm

Our goal in what follows will be to design a learning policy capable of exploiting the type of similarity structures introduced in the previous section. The main ingredients of our method are a nested attribute selection and cost estimation rule, which we describe in detail in Sections 3.1 and 3.2 respectively; the proposed *nested exponential weights* (NEW) algorithm is then developed and discussed in Section 3.3.

3.1. Probabilities, propensities, and nested logit choice

We begin by introducing the attribute selection scheme that forms the backbone of our proposed policy. Our guiding principle in this is the *nested logit choice* (NLC) rule of McFadden [21] which selects an alternative $a \in \mathcal{A}$ by traversing $\mathcal{S}$ one attribute at a time and prescribing the corresponding conditional choice probabilities at each level of $\mathcal{S}$.

To set the stage for all this, if $x = (x_1, \dots, x_n) \in \Delta(\mathcal{A})$ is a mixed strategy on $\mathcal{A}$ we will write

$$x_S = \textstyle\sum_{a \in S} x_a \tag{7}$$

for the probability of choosing $S \in \mathcal{S}$ under x, and

$$x_{S'|S} = x_{S'} / x_S \tag{8}$$

for the conditional probability of choosing a descendant S' of S assuming that S has already been selected under x.[4] Then the NLC rule proceeds as follows: first, it prescribes choice probabilities x_{S_1} for all classes $S_1 \in \mathcal{S}_1$ (i.e., the coarsest ones); subsequently, once a class $S_1 \in \mathcal{S}_1$ has been selected, NLC prescribes the conditional choice probabilities $x_{S_2|S_1}$ for all children S_2 of S_1 and draws a class from $\mathcal{S}_2$ based on $x_{S_2|S_1}$. The process then continues downwards along $\mathcal{S}$ until reaching the finest partition $\mathcal{S}_L$ and selecting an atom $\{a\} \equiv S_L \lhd S_{L-1} \lhd \cdots \lhd S_0 \equiv \mathcal{A}$.

This step-by-step selection process captures the "nested" part of the nested logit choice rule; the "logit" part refers to the way that the conditional probabilities (8) are actually prescribed given the agent's predisposition towards each alternative $a \in \mathcal{A}$. To make this precise, suppose that the learner associates to each element $a \in \mathcal{A}$ a *propensity score*

[4] Note here that the joint probability of selecting *both* S and S' under x is simply $x_{S'}$ whenever $S' \preccurlyeq S$.

$y_a \in \mathbb{R}$ indicating their tendency – or *propensity* – to select it. The associated propensity score of a similarity class $S_{\ell-1} \in \mathcal{S}_{\ell-1}$, $\ell = 1, \dots, L$, is then defined inductively as

$$y_{S_{\ell-1}} = \mu_\ell \log \sum\nolimits_{S_\ell \lhd S_{\ell-1}} \exp(y_{S_\ell}/\mu_\ell) \tag{9}$$

where $\mu_\ell > 0$ is a tunable parameter that reflects the learner's *uncertainty level* regarding the ℓ-th attribute $\mathcal{S}_\ell$ of $\mathcal{S}$. In words, this means that the score of a class is the weighted softmax of the scores of its children; thus, starting with the individual alternatives of $\mathcal{A}$ – that is, the *leaves* of $\mathcal{S}$ – propensity scores are propagated backwards along $\mathcal{S}$, and this is repeated one attribute at a time until reaching the root of $\mathcal{S}$.

Remark 4. We should also note that Eq. (9) assigns a propensity score to *any* similarity class $S \in \mathcal{S}$. However, because the primitives of this assignment are the original scores assigned to each alternative $a \in \mathcal{A}$, we will reserve the notation $y = (y_1, \dots, y_n) \in \mathbb{R}^{\mathcal{A}}$ for the *profile* of propensity scores $(y_a)_{a \in \mathcal{A}}$ that comprises the basis of the recursive definition (9). ¶

With all this in hand, given a propensity score profile $y = (y_1, \dots, y_n) \in \mathbb{R}^{\mathcal{A}}$, the *nested logit choice* (NLC) rule is defined via the family of conditional selection probabilities

$$P_{S_\ell | S_{\ell-1}}(y) = \frac{\exp(y_{S_\ell}/\mu_\ell)}{\exp(y_{S_{\ell-1}}/\mu_\ell)} \tag{NLC}$$

where:

1. $S_\ell \in \mathcal{S}_\ell$ and $S_{\ell-1} \in \mathcal{S}_{\ell-1}$ is a child / parent pair of similarity classes of $\mathcal{S}$.
2. $\mu_1 \geq \cdots \geq \mu_L > 0$ is a nonincreasing sequence of uncertainty parameters (indicating a higher uncertainty level for coarser attributes; we discuss this later).

In more detail, the choice of an alternative $a \in \mathcal{A}$ under (NLC) proceeds as follows: given a propensity score $y_a \in \mathbb{R}$ for each $a \in \mathcal{A}$, every similarity class $S_{L-1} \in \mathcal{S}_{L-1}$ is assigned a propensity score via the recursive softmax expression (9), and the same procedure is applied inductively up to the root $\mathcal{A}$ of $\mathcal{S}$. Then, to select an alternative $a \in \mathcal{A}$, the conditional logit choice rule (NLC) proceeds in a top-down manner, first by selecting a similarity class $S_1 \lhd S_0 \equiv \mathcal{A}$, then by selecting a child $S_2 \lhd S_1$ of S_1, and so on until reaching a leaf $\{a\} \equiv S_L \lhd S_{L-1} \lhd \cdots \lhd S_0 \equiv \mathcal{A}$ of $\mathcal{S}$.

Equivalently, unrolling (NLC) over the lineage $\mathcal{A} \equiv S_0 \rhd S_1 \rhd \cdots \rhd S_\ell \equiv S$ of a target class $S \in \mathcal{S}_\ell$, we obtain the expression

$$P_S(y) = \prod\nolimits_{k=1}^{\ell} \frac{\exp(y_{S_k}/\mu_k)}{\exp(y_{S_{k-1}}/\mu_k)} \tag{10}$$

for the total probability of selecting class S under the propensity score profile $y \in \mathbb{R}^{\mathcal{A}}$. Clearly, (NLC) and (10) are mathematically equivalent, so we will refer to either one as the definition of the nested logit choice rule.

3.2. The nested importance weighted estimator

The second key ingredient of our method is how to estimate the costs of alternatives that were not chosen under (NLC). To that end, given a cost vector $c \in [0,1]^{\mathcal{A}}$ and a mixed strategy $x \in \Delta(\mathcal{A})$ with full support, a standard way to do this is via the importance-weighted estimator [9, 18]

$$\hat{c}_a = \frac{\mathbb{1}\{a = \hat{a}\}}{x_a} c_a \tag{IWE}$$

where $\hat{a} \sim x$ is the (random) element of $\mathcal{A}$ chosen under x.

This estimator enjoys the following important properties:

a) It is non-negative.

b) It is *unbiased*, i.e.,

$$\mathbb{E}[\hat{c}_a] = c_a \quad \text{for all } a \in \mathcal{A}. \tag{11}$$

c) Its *importance-weighted mean square* is bounded as

$$\mathbb{E}\Big[\sum\nolimits_{a \in \mathcal{A}} x_a \hat{c}_a^2\Big] \leq n \tag{12}$$

This trifecta of properties plays a key role in establishing the no-regret guarantees of the vanilla exponential weights algorithm [5, 19, 29]; at the same time however, (IWE) fails to take into account any side information provided by similarities between different elements of $\mathcal{A}$. This is perhaps most easily seen in the original red bus / blue bus paradox: if the commuter takes a red bus, the observed utility would be immediately translateable to the blue bus (and vice versa). However, (IWE) is treating the red and blue buses as unrelated, so $\hat{c}_{\text{blue bus}}$ is not updated under (IWE), even though $c_{\text{blue bus}} = c_{\text{red bus}}$ by default.

To exploit this type of similarities, we introduce below a layered estimator that shadows the step-by-step selection process of (NLC). To define it, let $x \in \Delta(\mathcal{A})$ be a mixed strategy on $\mathcal{A}$ with full support, and assume that an element $\hat{a} \in \mathcal{A}$ is selected progressively according to x as in the case of (NLC):[5] First, the learner chooses a similarity class $\hat{S}_1 \in \mathcal{S}_1$ with probability $\mathbb{P}(\hat{S}_1 = S_1) = x_{S_1}$; subsequently, conditioned on the choice of $\hat{S}_1$, a class $\hat{S}_2 \lhd \hat{S}_1$ is selected with probability $\mathbb{P}(\hat{S}_2 = S_2 | \hat{S}_1) = x_{S_2|\hat{S}_1}$, and the process repeats until reaching a leaf $\hat{S}_L = \{\hat{a}\}$ of $\mathcal{S}$ (at which point the selection procedure terminates and returns $\hat{a}$). Then, given a loss profile $r \in [0, +\infty)^{\mathcal{S}}$ and a mixed strategy $x \in \Delta(\mathcal{A})$, the *nested importance weighted estimator* (NIWE) is defined for all $\ell = 1, \dots, L$ as

[5] To clarify, this process adheres to the "nested" part of (NLC); the conditional probabilities $x_{S'|S}$ may of course differ.

$$\hat{r}_{S_\ell} = \frac{\mathbb{1}\{S_\ell = \hat{S}_\ell, \dots, S_1 = \hat{S}_1\}}{x_{S_\ell|S_{\ell-1}} \cdots x_{S_2|S_1} x_{S_1}} r_{S_\ell} \qquad \text{(NIWE)}$$

where the chain of categorical random variables $\mathcal{A} \equiv \hat{S}_0 \rhd \hat{S}_1 \rhd \cdots \rhd \hat{S}_L = \{\hat{a}\}$ is drawn according to $x \in \Delta(\mathcal{A})$ as outlined above.[6]

This estimator will play a central part in our analysis, so some remarks are in order. First and foremost, the non-nested estimator (IWE) is recovered as a special case of (NIWE) when there are no similarity attributes on $\mathcal{A}$ (i.e., $L = 1$). Second, in a bona fide nested model, we should note that $\hat{c}_{S_\ell}$ is $\hat{S}_\ell$-measurable but *not* $\hat{S}_{\ell-1}$-measurable: this property has no analogue in (IWE), and it is an intrinsic feature of the step-by-step selection process underlying (NIWE). Third, it is also important to note that (NIWE) concerns the idiosyncratic losses of each chosen class, *not* the base costs c_a of each alternative $a \in \mathcal{A}$. This distinction is again redundant in the non-nested case, but it leads to a distinct estimator for c_a in nested environments, namely

$$\hat{c}_a = \sum\nolimits_{S \ni a} \hat{r}_S \quad \text{for all } a \in \mathcal{A}. \qquad (13)$$

In particular, in the red bus / blue bus paradox, this means that an observation for the class "bus" automatically updates both $\hat{c}_{\text{red bus}}$ and $\hat{c}_{\text{blue bus}}$, thus overcoming one of the main drawbacks of (IWE) when facing irrelevant alternatives.

To complete the comparison with the non-nested setting, we summarize below the most important properties of the layered estimator (NIWE):

Proposition 1. *Let $\mathcal{S} = \coprod_{\ell=1}^{L} \mathcal{S}_\ell$ be a similarity structure on $\mathcal{A}$. Then, given a mixed strategy $x \in \Delta(\mathcal{A})$ and a vector of cost increments $r \in \mathbb{R}^{\mathcal{S}}$ as per (5), the estimator (NIWE) satisfies the following:*

1. It is unbiased:

$$\mathbb{E}[\hat{r}_S] = r_S \quad \textit{for all } S \in \mathcal{S}. \qquad (14)$$

2. It enjoys the importance-weighted mean-square bound

$$\mathbb{E}\big[x_S \hat{r}_S^2\big] \le R_S^2 \quad \textit{for all } S \in \mathcal{S}. \qquad (15)$$

Accordingly, the loss estimator (13) is itself unbiased and enjoys the bound

$$\mathbb{E}\Big[\sum\nolimits_{a \in \mathcal{A}} x_a \hat{c}_a^2\Big] \le n_{\text{eff}} \qquad (16)$$

where n_{eff} is defined as

$$\sqrt{n_{\text{eff}}} = \sum\nolimits_{\ell=1}^{L} \sqrt{n_\ell} \bar{R}_\ell \qquad (17)$$

with $n_\ell = |\mathcal{S}_\ell|$ denoting the number of classes of attribute $\mathcal{S}_\ell$, and

$$\bar{R}_\ell = \sqrt{\frac{1}{n_\ell} \sum\nolimits_{S_\ell \in \mathcal{S}_\ell} R_{S_\ell}^2} \qquad (18)$$

denoting the "root-mean-square" range of all classes in $\mathcal{S}_\ell$.

Of course, Proposition 1 yields the standard properties of (IWE) as a special case when $L = 1$ (in which case there are no similarities to exploit between alternatives). To streamline our presentation, we prove this result in Appendix B.

3.3. The nested exponential weights algorithm

We are finally in a position to present the *nested exponential weights* (NEW) algorithm in detail. Building on the original exponential weights blueprint [5, 19, 29], the main steps of the NEW algorithm can be summed up as follows:

1. For each stage $t = 1, 2, \dots$, the learner maintains and updates a propensity score profile $y_t \in \mathbb{R}^{\mathcal{A}}$.

2. The learner selects an action $a_t \in \mathcal{A}$ based on the nested logit choice rule $a_t \sim P(\eta_t y_t)$ where $\eta_t \ge 0$ is the method's *learning rate* and P is given by (NLC).

3. The learner incurs $r_{S,t}$ for each class $S \ni a_t$ and constructs a model $\hat{c}_t$ of the cost vector c_t of stage t via (NIWE).

4. The learner updates their propensity score profile based on $\hat{c}_t$ and the process repeats.

For a presentation of the algorithm in pseudocode form, see Algorithm 1; the tuning of the method's uncertainty parameters $\mu_1 \ge \dots \ge \mu_L > 0$ and the learning rate η_t is discussed in the next section, where we undertake the analysis of the NEW algorithm.

4. Analysis and results

We are now in a position to state and discuss our main regret guarantees for the NEW algorithm. These are as follows:

Theorem 1. *Suppose that Algorithm 1 is run with a non-increasing learning rate $\eta_t > 0$ and uncertainty parameters $\mu_1 \ge \cdots \ge \mu_L > 0$ against a sequence of cost vectors $c_t \in [0,1]^{\mathcal{A}}$, $t = 1, 2, \dots$, as per (4). Then, for all $p \in \Delta(\mathcal{A})$, the learner enjoys the regret bound*

$$\mathbb{E}[\mathrm{Reg}_p(T)] \le \frac{H}{\eta_{T+1}} + \frac{n_{\text{eff}}}{2\mu_L} \sum_{t=1}^{T} \eta_t \qquad (19)$$

with n_{eff} given by (17) and $H \equiv H(\mu_1, \dots, \mu_L)$ defined by

[6] The indicator in (NIWE) is assumed to take precedence over $x_{S_k|S_{k-1}}$, i.e., $\hat{c}_{S_\ell} = 0$ if $S_k \neq \hat{S}_k$ for some $k = 1, \dots, \ell$.

Algorithm 1: Nested exponential weights (NEW)

Require: set of alternatives $\mathcal{A}$; attribute structure $\mathcal{S} = \coprod_{\ell=1}^{L} \mathcal{S}_\ell$
Params: uncertainty levels $\mu_1, \dots, \mu_L > 0$; learning rate $\eta_t \geq 0$
Input: sequence of class costs $r_t \in [0,1]^{\mathcal{S}}$, $t = 1, 2, \dots$

initialize $y \leftarrow 0 \in \mathbb{R}^{\mathcal{A}}$, $S_0 = \mathcal{A}$ # *initialization*
for $t = 1, 2, \dots$ **do** # *scoring phase*
  for $\ell = L-1, \dots, 0$ and **for all** $S \in \mathcal{S}_\ell$ **do**
    $y_S \leftarrow \mu_{\ell+1} \log \sum_{S' \lhd S} \exp(y_{S'}/\mu_{\ell+1})$ # *as per* (9)
    set $\hat{r}_S \leftarrow 0$ # *baseline guess*
  end for
  for $\ell = 1, \dots, L$ **do** # *selection phase*
    select class $S_\ell \lhd S_{\ell-1}$ # *class choice*

$$S_\ell \sim x_{S_\ell | S_{\ell-1}} = \frac{\exp(\eta_t y_{S_\ell}/\mu_\ell)}{\exp(\eta_t y_{S_{\ell-1}}/\mu_\ell)} \qquad \text{\# (NLC)}$$

    get $r_{S_\ell, t}$ # *intra-class cost*
    set $\hat{r}_{S_\ell} \leftarrow \hat{r}_{S_\ell} + \dfrac{r_{S_\ell, t}}{x_{S_\ell|S_{\ell-1}} \cdots x_{S_1|S_0}}$ # (NIWE)
  end for
  set $\hat{c}_a \leftarrow \sum_{S \ni a} \hat{r}_S$ for all $a \in \mathcal{A}$ # *loss model*
  set $y \leftarrow y - \hat{c}$ # *update propensities*
end for

setting $y = 0$ in (9) and taking $H = y_{\mathcal{A}}$, i.e.,

$$H = \log \left[\sum_{S_1 \lhd S_0} \left[\sum_{S_2 \lhd S_1} \cdots \left[\sum_{S_L \lhd S_{L-1}} 1 \right]^{\frac{\mu_L}{\mu_{L-1}}} \cdots \right]^{\frac{\mu_2}{\mu_1}} \right]^{\mu_1} \tag{20}$$

In particular, if Algorithm 1 is run with $\mu_1 = \cdots = \mu_L = \sqrt{n_{\text{eff}}/2}$ and $\eta_t = \sqrt{\log n/(2t)}$, we have

$$\mathbb{E}[\mathrm{Reg}_p(T)] \leq 2\sqrt{n_{\text{eff}} \log n \cdot T}. \tag{21}$$

Theorem 1 is our main regret guarantee for NEW so, before discussing its proof (which we carry out in detail in Appendices A–C), some remarks are in order.

The first thing of note is the comparison to the corresponding bound for EXP3, namely

$$\mathbb{E}[\mathrm{Reg}_p(T)] \leq 2\sqrt{n \log n \cdot T}. \tag{22}$$

This shows that the guarantees of NEW and EXP3 differ by a factor of[7]

$$\alpha = \sqrt{n/n_{\text{eff}}}, \tag{23}$$

which, for reasons that become clear below, we call the *price of affinity* (PoAf).

[7] Depending on the source, the bound (22) may differ up to a factor of $\sqrt{2}$, compare for example [26, Corollary 4.2] and [18, Theorem 11.2]. This factor is due to the fact that (22) is usually stated for a known horizon T (which saves a factor of $\sqrt{2}$ relative to anytime algorithms). Ceteris paribus, the bound (21) can be sharpened by the same factor, but we omit the details.

Since the variabilities of the idiosyncratic losses within each attribute have been normalized to 1 (recall the relevant discussion in Section 2.3), Hölder's inequality trivially gives $n_{\text{eff}} \leq n$, no matter the underlying similarity structure. Of course, if there are no similarities to exploit ($L = 1$), we get $n_{\text{eff}} = n$, in which case the two bounds coincide ($\alpha = 1$).

At the other extreme, suppose we have a red bus / blue bus type of problem with, say, $n_1 = 2$ similarity classes, $n_2 = 100$ alternatives per class, and a negligible intra-class loss differential ($R_2 \approx 0$). In this case, EXP3 would have to wrestle with $n = n_1 n_2 = 200$ alternatives, while NEW would only need to discriminate between $n_{\text{eff}} \approx n_1 = 2$ alternatives, leading to an improvement by a factor of $\alpha \approx 10$ in terms of regret guarantee. Thus, even though the red bus / blue bus paradox could entangle EXP3 and cause the algorithm to accrue significant regret over time, this is no longer the case under the NEW method; we also explore this issue numerically in Section 5.

As another example, suppose that each non-terminal class in $\mathcal{S}$ has m children and the variability of the idiosyncratic losses likewise scales down by a factor of m per attribute. In this case, a straightforward calculation shows that n_{eff} scales as $\Theta(m)$, so the gain in efficiency would be of the order of $\alpha = \sqrt{n/n_{\text{eff}}} = \Theta(m^{(L-1)/2})$, i.e., polynomial in m and exponential in L. This gain in performance can become especially pronounced when there is a very large number of atlernatives organized in categories and subcategories of geometrically decreasing impact on the end cost of each alternative. We explore this issue in practical scenarios in Section 5 and Appendix D.

Finally, we should also note that the parameters of NEW have been tuned so as to facilitate the comparison with EXP3. This tuning is calibrated for the case where $\mathcal{S}$ is fully symmetric, i.e., all subcategories of a given attribute have the same number of children. Otherwise, in full generality, the tuning of the algorithm's uncertainty levels would boil down to a transcedental equation involving the nested term $H(\mu_1, \dots, \mu_L)$ of (19). This can be done efficiently offline via a line search, but since the result would be structure-dependent, we do not undertake this analysis here.

Proof outline of Theorem 1. The detailed proof of Theorem 1 is quite lengthy, so we defer it to Appendices A–C and only sketch here the main ideas.

The first basic step is to derive a suitable "potential function" that can be used to track the evolution of the NEW policy relative to the benchmark $p \in \Delta(\mathcal{A})$. The main ingredient of this potential is the "nested" entropy function

$$h(x) = \sum\nolimits_{k=0}^{L} \delta_k \sum\nolimits_{S_k \in \mathcal{S}_k} x_{S_k} \log x_{S_k}, \tag{24}$$

where $\delta_k = \mu_k - \mu_{k+1}$ for all $k = 1, \dots, L$ (with $\mu_{L+1} = 0$

by convention).[8] As we show in Proposition A.1 in Appendix A, the "tiers" of h can be unrolled to give the "non-tiered" recursive representation

$$h(x) = \sum\nolimits_{S \in \mathcal{S}} h(x|S) \tag{25}$$

where $h(x|S) = \mu_{\ell+1} \sum_{S' \lhd S} x_{S'} \log(x_{S'}/x_S)$ denotes the "conditional" entropy of x relative to class $S \in \mathcal{S}_\ell$. Then, by means of this decomposition and a delicate backwards induction argument, we show in Proposition A.2 that *a*) the recursively defined propensity score $y_{\mathcal{A}}$ of $\mathcal{A}$ can be expressed *non-recursively* as $y_{\mathcal{A}} = \arg\max_{x \in \Delta(\mathcal{A})} \{\langle y, x\rangle - h(x)\}$; and *b*) that the choice rule (NLC) can be expressed itself as

$$P_a(y) = \frac{\partial y_{\mathcal{A}}}{\partial y_a} \quad \text{for all } y \in \mathbb{R}^{\mathcal{A}},\, a \in \mathcal{A}. \tag{26}$$

This representation of (NLC) provides the first building block of our proof because, by Danskin's theorem [8], it allows us to rewrite Algorithm 1 in more concise form as

$$\begin{aligned} y_{t+1} &= y_t - \hat{c}_t \\ x_{t+1} &= \arg\max_{x \in \Delta(\mathcal{A})} \{\langle \eta_{t+1} y_{t+1}, x\rangle - h(x)\} \end{aligned} \tag{NEW}$$

with $\hat{c}_t$ given by (13) appplied to $x \leftarrow x_t$. Importantly, this shows that the NEW algorithm is an instance of the well-known *"follow the regularized leader"* (FTRL) algorithmic framework [26, 27]. Albeit interesting, this observation is not particularly helpful in itself because there is no universal, "regularizer-agnostic" analysis giving optimal (or near-optimal) regret rates for FTRL with bandit/partial information.[9] Nonetheless, by adapting a series of techniques that are used in the analysis of FTRL algorithms, we show in Appendix C that the iterates of (NEW) satisfy the "energy inequality"

$$\begin{aligned} \langle \hat{c}_t, x_t - p\rangle \leq E_t - E_{t+1} + \frac{1}{\eta_t} F(x_t, \eta_t y_{t+1}) \\ + (\eta_{t+1}^{-1} - \eta_t^{-1})[h(p) - \min h] \end{aligned} \tag{27}$$

where $\hat{c}_t$ is the nested importance weighted estimator (13) for the cost vector encountered c_t, and we have set

$$F(x, y) = h(x) + y_{\mathcal{A}} - \langle y, x\rangle \tag{28}$$

and $E_t = \eta_t^{-1} F(p, \eta_t y_t)$.

Then, by Proposition 1, we obtain:

[8] In the non-nested case, (24) boils down to the standard (negative) entropy $h(x) = \sum_a x_a \log x_a$. However, the inverse problem of deriving the "correct" form of h in a nested environment involves a technical leap of faith and a fair degree of trial-and-error.

[9] For the analysis of specific versions of FTRL with non-entropic regularizers, cf. [3, 30] and references therein.

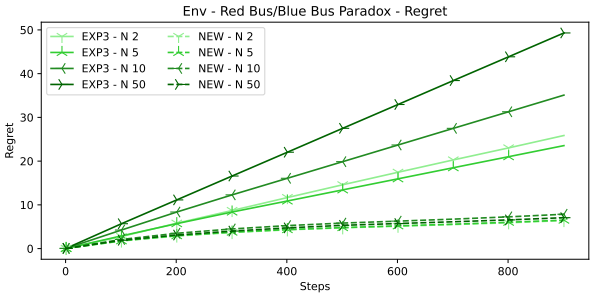

Figure 2: Regret of EXP3 and NEW in the red bus / blue bus problem with different numbers of buses.

Proposition 2. *The NEW algorithm enjoys the bound*

$$\mathbb{E}[\mathrm{Reg}_p(T)] \leq \frac{H}{\eta_{T+1}} + \sum_{t=1}^{T} \frac{\mathbb{E}[F(x_t, \eta_t y_{t+1})]}{\eta_t}. \tag{29}$$

Proposition 2 provides the first half of the bound (19), with the precise form of H derived in Lemma C.1. The second half of (19) revolves around the term $\mathbb{E}[F(x_t, \eta_t y_{t+1})]$ and boils down to estimating how propensity scores are back-propagated along $\mathcal{S}$. In particular, the main difficulty is to bound the difference $y_{\mathcal{A}}^+ - y_{\mathcal{A}}$ in the propensity score of the root node $\mathcal{A}$ of $\mathcal{S}$ when the underlying score profile $y \in \mathbb{R}^{\mathcal{A}}$ is incremented to $y^+ = y + w$ for some $w \in \mathbb{R}^{\mathcal{A}}$.

A first bound that can be obtained by convex analysis arguments is $|y_{\mathcal{A}}^+ - y_{\mathcal{A}}| \leq \langle y, P(y)\rangle + \|w\|_\infty^2$; however, because the increments of (NEW) are unbounded in norm, this global bound is far too lax for our puposes. A similar issue arises in the analysis of EXP3, and is circumvented by deriving a bound for the log-sum-exp function using the identity $\exp(x) \leq 1 + x + x^2/2$ for $x \leq 0$ and the fact that the estimator (IWE) is non-negative [11, 18, 26]. Extending this idea to nested environments is a very delicate affair, because each tier in $\mathcal{S}$ introduces an additional layer of error propagation in the increments $y_{t+1} - y_t$. However, by a series of inductive arguments that traverse $\mathcal{S}$ both forward and backward, we are able to show the bound

$$y_{\mathcal{A}}^+ - y_{\mathcal{A}} \leq \langle y, P(y)\rangle + \frac{1}{2\mu_L} \sum_{\ell=1}^{L} \sum_{S_\ell \in \mathcal{S}_\ell} P_{S_\ell}(y) r_{S_\ell}^2 \tag{30}$$

which, after taking expecations and using the bounds of Proposition 1, finally yields the pseudo-regret bound (19).

5. Numerical experiments

In this section we present a series of numerical experiments designed to test the efficiency of our method compared to EXP3. We use a synthetic environment where we simulate nested similarity partitions with trees. While NEW exploits the similarity structure by making forward/backward

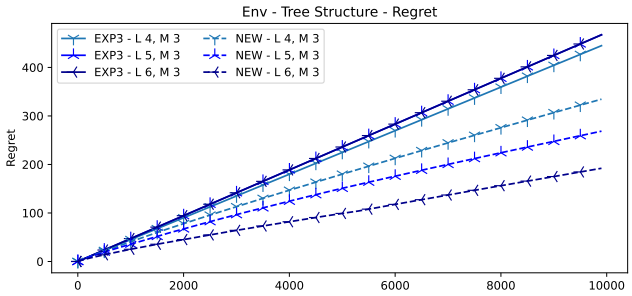

Figure 3: Regret of EXP3 and NEW in a tree environment with different values of levels L and classes per level M

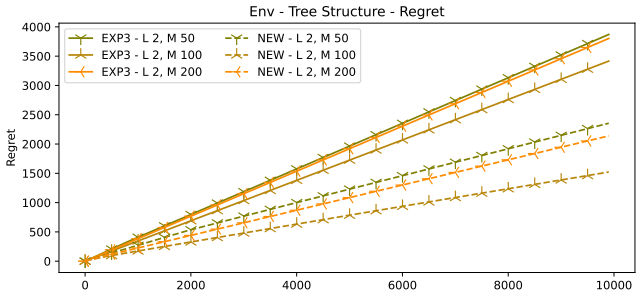

Figure 4: Regret of EXP3 and NEW in a tree environment with different values of levels L and classes per level M

passes through the associated tree with its logit choice rule (NLC), EXP3 is simply run over the leaves of the tree, i.e., $\mathcal{A}$. All experiment details (as well as additional results) are presented in Appendix D. For every setting, we report the results of our experiments by plotting the average regret of each algorithm for 20 seeds of randomly drawn losses. The code to reproduce the experiments can be found at `https://github.com/criteo-research/Nested-Exponential-Weights`.

Benefits in the red bus/blue bus problem. We consider here a variant of the red bus/blue bus problem with N different buses (the original paradox has $N = 2$). In this experiment (see illustration in Fig. 5, Appendix D.2) we allow each bus to have non-zero intrinsic losses and illustrate in Fig. 2 how both algorithms perform when N grows. We observe there that for all configurations NEW achieves better regret than EXP3. While both methods achieve sublinear regret, EXP3 requires far more steps to identify the best alternative as N grows and suffers overall from worse regret while NEW achieves similar regret and does not suffer as much from the number of irrelevant alternatives. We provide additional plots in Appendix D.2 which show that NEW performs consistently better than EXP3 when there exists a similarity structure allowing to efficiently update scores of classes that have very similar losses.

Performance in general nested structures. In this setting we generate symmetric trees and experiment with different values of number of levels L and number of child per nodes $M = |S_\ell|$ for $\ell = 1, \dots, L$. Specifically, in Fig. 3 with a fixed M, we see that NEW obtains better regret than EXP3 even when L increases. We provide variance plots for the experiments that generated the same performance on the plots in D.3 as well as additional visualisations. Finally, in Fig. 4, we can see that for a shallow tree ($L = 2$) NEW performs always better than EXP3, even for high values of M. Indeed, when the number of children per nodes M increases, the tree loses its "factorized" structure which also affects NEW due to the less "structured" tree. Thus, again, NEW performs consistently better than EXP3 when it is possible to efficiently handle classes with similar losses.

Overall, our experiments confirm that a learning algorithm based on nested logit choice can lead to significant benefits in problems with a high degree of similarity between alternatives. This leaves open the question of whether a similar approach can be applied to structures with *non-nested* attributes; we defer this question to future work.

6. Concluding remarks

One limitation of the current framework is that the nested estimator (13) requires knowledge of the intra-class cost increments r_S for every chosen similarity class $S \ni a_t$. This is akin to the difference between the "full bandit" and "semi-bandit" setting that arises in combinatorial bandits [12]. While relevant in a number of application domains (e.g., in path-planning or when layering a structured security, such as the tranches of a CDO), treating the fully unobservable case – possibly using an approach in the spirit of the hierarchical contextual analysis of Sen et al. [25] – is an important open question for future research.

Finally, it is also interesting to note that our analysis has been carried out in an arbitrarily changing "adversarial" environment. In a stochastic environment, it would be fruitful to consider other, contextual-based approaches such as LinUCB, KernelUCB and their variants [18]. Ideally, one would like to employ a nested variant of the "universal" algorithm of Zimmert & Seldin [30] that attains optimal regret guarantees in both stochastic and adversarial environments, but this question lies beyond the scope of our work.

Acknowledgements

P. Mertikopoulos is grateful for financial support by the French National Research Agency (ANR) in the framework of the "Investissements d'avenir" program (ANR-15-IDEX-02), the LabEx PERSYVAL (ANR-11-LABX-0025-01), MIAI@Grenoble Alpes (ANR-19-P3IA-0003), and the bilateral ANR-NRF grant ALIAS (ANR-19-CE48-0018-01).

References

[1] Abbasi-yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Adv. Neural Information Processing Systems (NIPS)*, 2011.

[2] Anderson, S. P., de Palma, A., and Thisse, J.-F. *Discrete Choice Theory of Product Differentiation*. MIT Press, Cambridge, MA, 1992.

[3] Audibert, J.-Y., Bubeck, S., and Lugosi, G. Minimax policies for combinatorial prediction games. In *COLT '11: Proceedings of the 24th Annual Conference on Learning Theory*, 2011.

[4] Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of the 36th Annual Symposium on Foundations of Computer Science*, 1995.

[5] Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.

[6] Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

[7] Ben-Akiva, M. and Lerman, S. R. *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT Press, Cambridge, 1985.

[8] Berge, C. *Topological Spaces*. Dover, New York, 1997.

[9] Bubeck, S. and Cesa-Bianchi, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

[10] Bubeck, S., Munos, R., Stoltz, G., and Szepesvári, C. $\mathcal{X}$-armed bandits. *Journal of Machine Learning Research*, 12: 1655–1695, 2011.

[11] Cesa-Bianchi, N. and Lugosi, G. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[12] Cesa-Bianchi, N. and Lugosi, G. Combinatorial bandits. *Journal of Computer and System Sciences*, 78:1404–1422, 2012.

[13] Cesa-Bianchi, N. and Shamir, O. Bandit regret scaling with the effective loss range. In *ALT '18: Proceedings of the 29th International Conference on Algorithmic Learning Theory*, 2018.

[14] Cesa-Bianchi, N., Gaillard, P., Gentile, C., and Gerchinovitz, S. Algorithmic chaining and the role of partial feedback in online nonparametric learning. In *COLT '17: Proceedings of the 30th Annual Conference on Learning Theory*, 2017.

[15] György, A., Linder, T., Lugosi, G., and Ottucsák, G. The online shortest path problem under partial monitoring. *Journal of Machine Learning Research*, 8:2369–2403, 2007.

[16] Héliou, A., Martin, M., Mertikopoulos, P., and Rahier, T. Zeroth-order non-convex learning via hierarchical dual averaging. In *ICML '21: Proceedings of the 38th International Conference on Machine Learning*, 2021.

[17] Kocák, T., Neu, G., Valko, M., and Munos, R. Efficient learning by implicit exploration in bandit problems with side observations. In *NIPS '14: Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2014.

[18] Lattimore, T. and Szepesvári, C. *Bandit Algorithms*. Cambridge University Press, Cambridge, UK, 2020.

[19] Littlestone, N. and Warmuth, M. K. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.

[20] Luce, R. D. *Individual Choice Behavior: A Theoretical Analysis*. Wiley, New York, 1959.

[21] McFadden, D. L. Conditional logit analysis of qualitative choice behavior. In Zarembka, P. (ed.), *Frontiers in Econometrics*, pp. 105–142. Academic Press, New York, NY, 1974.

[22] Nesterov, Y. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, 2009.

[23] Neu, G. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *NIPS '15: Proceedings of the 29th International Conference on Neural Information Processing Systems*, 2015.

[24] Rockafellar, R. T. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.

[25] Sen, R., Rakhlin, A., Ying, L., Kidambi, R., Foster, D., Hill, D., and Dhillon, I. Top-k eXtreme contextual bandits with arm hierarchy. In *ICML '21: Proceedings of the 38th International Conference on Machine Learning*, 2021.

[26] Shalev-Shwartz, S. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4 (2):107–194, 2011.

[27] Shalev-Shwartz, S. and Singer, Y. Convex repeated games and Fenchel duality. In *NIPS' 06: Proceedings of the 19th Annual Conference on Neural Information Processing Systems*, pp. 1265–1272. MIT Press, 2006.

[28] Thune, T. S. and Seldin, Y. Adaptation to easy data in prediction with limited advice. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[29] Vovk, V. G. Aggregating strategies. In *COLT '90: Proceedings of the 3rd Workshop on Computational Learning Theory*, pp. 371–383, 1990.

[30] Zimmert, J. and Seldin, Y. An optimal algorithm for stochastic and adversarial bandits. In *AISTATS '19: Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 2019.

A. The nested entropy and its properties

Our aim in this appendix is to prove the basic properties of the series of (negative) entropy functions that fuel the regret analysis of the nested exponential weights (NEW) algorithm.

To begin with, given a similarity structure $\mathcal{S}$ on $\mathcal{A}$ and a sequence of uncertainty parameters $\mu_1 \geq \cdots \geq \mu_L > 0$ (with $\mu_{L+1} = 0$ by convention), we define:

1. The *conditional entropy* of $x \in \Delta(\mathcal{A})$ relative to a target class $S \in \mathcal{S}_\ell$:

$$h(x|S) = \mu_{\ell+1} \sum_{S' \lhd S} x_{S'} \log \frac{x_{S'}}{x_S} = \mu_{\ell+1}\, x_S \sum_{S' \lhd S} x_{S'|S} \log x_{S'|S}. \tag{A.1}$$

2. The *nested entropy* of $x \in \Delta(\mathcal{A})$ relative to $S \in \mathcal{S}_\ell$:

$$h_S(x) = \sum_{k=\ell}^{L} \delta_k \sum_{S_k \preccurlyeq_k S} x_{S_k} \log x_{S_k} \tag{A.2}$$

   where $\delta_k = \mu_k - \mu_{k+1}$ for all $k = 1, \dots, L$.

3. The *restricted entropy* of $x \in \Delta(\mathcal{A})$ relative to $S \in \mathcal{S}_\ell$:

$$h_{|S}(x) = h_S(x) + \chi_{\Delta(S)}(x) = \begin{cases} h_S(x) & \text{if } x \in \Delta(S), \\ \infty & \text{otherwise,} \end{cases} \tag{A.3}$$

   where $\chi_{\Delta(S)}$ denotes the (convex) characteristic function of $\Delta(S)$, i.e., $\chi_{\Delta(S)}(x) = 0$ if $x \in \Delta(S)$ and $\chi_{\Delta(S)}(x) = \infty$ otherwise. [Obviously, $h_{|S}(x) = h_S(x)$ whenever $x \in \Delta(S)$.]

Remark 1. As per our standard conventions, we are treating S interchangeably as a subset of $\mathcal{A}$ or as an element of $\mathcal{S}$; by analogy, to avoid notational inflation, we are also viewing $\Delta(S)$ as a subset of $\Delta(\mathcal{A})$ – more precisely, a face thereof. Finally, in all cases, the functions $h(x|S)$, $h_S(x)$ and $h_{|S}(x)$ are assumed to take the value $+\infty$ for $x \in \mathbb{R}^{\mathcal{A}} \setminus \Delta(\mathcal{A})$. ¶

Remark 2. For posterity, we also note that the nested and restricted entropy functions ($h_S(x)$ and $h_{|S}(x)$ respectively) are both convex – though not necessarily *strictly* convex – over $\Delta(\mathcal{A})$. This is a consequence of the fact that each summand $x_S \log x_S$ in (A.2) is convex in x and that $\delta_k = \mu_k - \mu_{k+1} \geq 0$ for all $k = 1, \dots, L$. Of course, any two distributions $x, x' \in \Delta(\mathcal{A})$ that assign the same probabilities to elements of S but not otherwise have $h_S(x) = h_S(x')$, so h_S is *not* strictly convex over $\Delta(\mathcal{A})$ if $S \neq \mathcal{A}$. However, since the function $\sum_{a \in S} x_a \log x_a$ is strictly convex over $\Delta(S)$, it follows that h_S – and hence $h_{|S}$ – *is* strictly convex over $\Delta(S)$. ¶

Our main goal in the sequel will be to prove the following fundamental properties of the entropy functions defined above:

Proposition A.1. *For all $S \in \mathcal{S}_\ell$, $\ell = 1, \dots, L$, and for all $x \in \Delta(\mathcal{A})$, we have:*

$$h_S(x) = \sum_{S' \preccurlyeq S} h(x|S') + \mu_\ell\, x_S \log x_S. \tag{A.4}$$

Consequently, for all $x \in \Delta(S)$, we have:

$$h_{|S}(x) = \sum_{S' \preccurlyeq S} h(x|S'). \tag{A.5}$$

Proposition A.2. *For all $S \in \mathcal{S}$ and all $y \in \mathbb{R}^{\mathcal{A}}$, we have:*

1. The recursively defined propensity score y_S of S as given by (9) can be expressed as

$$y_S = \max_{x \in \Delta(S)} \{\langle y, x \rangle - h_{|S}(x)\} \tag{A.6}$$

2. *The conditional probability of choosing $a \in \mathcal{A}$ given that S has already been selected under* (NLC) *is given by*

$$P_{a|S}(y) = \frac{\partial y_S}{\partial y_a} \tag{A.7}$$

and the conditional probability vector $P_{|S}(y) = (P_{a|S}(y))_{a\in\mathcal{A}}$ solves the problem (A.6)*, viz.*

$$P_{|S}(y) = \operatorname*{arg\,max}_{x\in\Delta(S)}\{\langle y, x\rangle - h_{|S}(x)\} \tag{A.8}$$

These propositions will be the linchpin of the analysis to follow, so some remarks are in order:

Remark 3. Note here that the maximum in (A.6) is taken over the *restricted* entropy function $h_{|S}$, *not* the nested entropy h_S. This distinction will play a crucial role in the sequel; in particular, since $h_{|S}$ is strictly convex over $\Delta(S)$, it implies that the $\arg\max$ in (A.8) is a singleton. ¶

Remark 4. The first part of Proposition A.2 can be rephrased more concisely (but otherwise equivalently) as

$$y_S = h^*_{|S}(y) \tag{A.9}$$

where

$$h^*_{|S}(y) = \max_{x\in\Delta(\mathcal{A})}\{\langle y, x\rangle - h_{|S}(x)\} \tag{A.10}$$

denotes the convex conjugate of $h_{|S}$. This interpretation is conceptually important because it spells out the precise functional dependence between the (primitive) propensity score profile $y \in \mathbb{R}^{\mathcal{A}}$ and the propensity scores y_S that are propagated to higher-tier similarity classes $S \in \mathcal{S}$ via the recursive definition (9). In particular, this observation leads to the recursive rule

$$\exp\left(\frac{h^*_{|S}(y)}{\mu_{\ell+1}}\right) = \sum_{S'\lhd S}\exp\left(\frac{h^*_{|S'}(y)}{\mu_{\ell+1}}\right) \quad \text{for all } S\in\mathcal{S}_\ell,\ \ell = 0,1,\dots,L-1. \tag{A.11}$$

We will we use this representation freely in the sequel. ¶

Remark 5. It is also worth noting that the propensity scores y_{S_ℓ}, $S_\ell \in \mathcal{S}_\ell$, can also be seen as primitives for the arborescence $\mathcal{S}' = \coprod_{k=0}^{\ell}\mathcal{S}_k$ obtained from $\mathcal{S}$ by excising all (proper) descendants of $\mathcal{S}_\ell$. Under this interpretation, the second part of Proposition A.2 readily gives the more general expression

$$P_{S'|S}(y) = \frac{\partial y_S}{\partial y_{S'}} \quad \text{for all } S' \preccurlyeq S, \tag{A.12}$$

where, in the right-hand side, y_S is to be construed as a function of $y_{S'}$, defined recursively via (9) applied to the truncated arborescence $\mathcal{S}'$. Even though we will not need this specific result, it is instructive to keep it in mind for the sequel.

The rest of this appendix is devoted to the proofs of Propositions A.1 and A.2.

Proof of Proposition A.1. Let $\ell = \operatorname{attr}(S)$, and fix some attribute label $k > \ell$. We will proceed inductively by collecting all terms in (A.4) associated to the attribute $\mathcal{S}_k$ and then summing everything together. Indeed, we have:

$$\begin{aligned}
\mu_k \sum_{S'\preccurlyeq_k S} x_{S'}\log x_{S'} &= \mu_k \sum_{S_{k-1}\preccurlyeq_{k-1}S}\left[\sum_{S'\lhd S_{k-1}} x_{S'}\log x_{S'}\right] && \text{\# collect attributes}\\
&= \mu_k \sum_{S_{k-1}\preccurlyeq_{k-1}S}\left[\sum_{S'\lhd S_{k-1}} x_{S'|S_{k-1}}x_{S_{k-1}}\log(x_{S'|S_{k-1}}x_{S_{k-1}})\right] && \text{\# by definition}\\
&= \mu_k \sum_{S_{k-1}\preccurlyeq_{k-1}S}\left[\sum_{S'\lhd S_{k-1}} x_{S'|S_{k-1}}x_{S_{k-1}}\log x_{S'|S_{k-1}}\right] && \text{(A.13a)}\\
&\quad + \mu_k \sum_{S_{k-1}\preccurlyeq_{k-1}S}\left[\sum_{S'\lhd S_{k-1}} x_{S'|S_{k-1}}x_{S_{k-1}}\log x_{S_{k-1}}\right] && \text{(A.13b)}
\end{aligned}$$

with the tacit understanding that any empty sum that appears above is taken equal to zero.

Now, by the definition of the nested entropy, we readily obtain that

$$(\text{A.13a}) = \sum_{S_{k-1} \preccurlyeq_{k-1} S} h(x|S_{k-1}) \tag{A.14a}$$

whereas, by noting that $\sum_{S' \lhd S_{k-1}} x_{S'|S_{k-1}} = 1$ (by the definition of conditional class choice probabilities), Eq. (A.13b) becomes

$$(\text{A.13b}) = \mu_k \sum_{S_{k-1} \preccurlyeq_{k-1} S} x_{S_{k-1}} \log x_{S_{k-1}}. \tag{A.14b}$$

Hence, combining Eqs. (A.13), (A.14a) and (A.14b), we get:

$$\mu_k \sum_{S' \preccurlyeq_k S} x_{S'} \log x_{S'} = \sum_{S_{k-1} \preccurlyeq_{k-1} S} h(x|S_{k-1}) + \mu_k \sum_{S_{k-1} \preccurlyeq_{k-1} S} x_{S_{k-1}} \log x_{S_{k-1}}. \tag{A.15}$$

The above expression is our basic inductive step. Indeed, summing (A.15) over all $k = L, \dots, \ell = \operatorname{attr}(S)$, we obtain:

$$\begin{aligned}
h_S(x) &= \sum_{k=\ell}^{L} (\mu_k - \mu_{k+1}) \sum_{S' \preccurlyeq_k S} x_{S'} \log x_{S'} && \text{\# by definition} \\
&= \sum_{k=L}^{\ell+1} \left[\mu_k \sum_{S' \preccurlyeq_k S} x_{S'} \log x_{S'} - \mu_{k+1} \sum_{S' \preccurlyeq_k S} x_{S'} \log x_{S'} \right] + (\mu_\ell - \mu_{\ell+1})\, x_S \log x_S && \text{\# isolate } S \\
&= \sum_{k=L}^{\ell+1} \left[\sum_{S_{k-1} \preccurlyeq_{k-1} S} h(x|S_{k-1}) + \mu_k \sum_{S_{k-1} \preccurlyeq_{k-1} S} x_{S_{k-1}} \log x_{S_{k-1}} - \mu_{k+1} \sum_{S' \preccurlyeq_k S} x_{S'} \log x_{S'} \right] \\
&\quad + (\mu_\ell - \mu_{\ell+1})\, x_S \log x_S && \text{\# by (A.15)} \\
&= \sum_{k=\ell}^{L-1} \sum_{S' \preccurlyeq_k S} h(x|S') + \mu_\ell\, x_S \log x_S - \mu_{L+1} \sum_{S' \preccurlyeq_L S} x_{S'} \log x_{S'} && (\text{A.16})
\end{aligned}$$

with the last equality following by telescoping the terms involving μ_k. Now, given that $\mu_{L+1} = 0$ by convention, the third sum above is zero. Finally, since the conditional entropy of x relative to any childless class is zero by definition, the first sum in (A.16) can be rewritten as $\sum_{k=\ell}^{L-1} \sum_{S' \preccurlyeq_k S} h(x|S') = \sum_{S' \preccurlyeq S} h(x|S')$, and our claim follows.

Finally, (A.5) is a consequence of the fact that $x_S = 1$ whenever $x \in \Delta(S)$ – i.e., whenever $\operatorname{supp}(x) \subseteq S$. ■

Proof of Proposition A.2. We begin by noting that the optimization problem (A.6) can be written more explicitly as

$$\begin{aligned}
&\text{maximize} && \langle y, x \rangle - h_S(x), \\
&\text{subject to} && x \in \Delta(\mathcal{A}) \text{ and } \operatorname{supp}(x) \subseteq S.
\end{aligned} \tag{Opt_S}$$

We will proceed to show that the (unique) solution of (Opt_S) is given by the vector of conditional probabilities $(P_{a|S}(y))_{a \in \mathcal{A}}$. The expression (A.6) for the maximal value of (Opt_S) will then be derived from Proposition A.1, and the differential representation (A.8) will follow from Legendre's identity. We make all this precise in a series of individual steps below.

Step 1: Optimality conditions for (Opt_S). For all $a \in S$, the definition of the nested entropy gives

$$\begin{aligned}
\frac{\partial h_S}{\partial x_a} &= \sum_{k=\ell}^{L} \delta_k \sum_{S' \preccurlyeq_k S} \frac{\partial}{\partial x_a} (x_{S'} \log x_{S'}) = \sum_{k=\ell}^{L} \delta_k \sum_{S' \preccurlyeq_k S} (1 + \log x_{S'}) \frac{\partial x_{S'}}{\partial x_a} \\
&= \sum_{k=\ell}^{L} \delta_k \sum_{S' \preccurlyeq_k S} (1 + \log x_{S'})\, \mathbb{1}\{a \in S'\}
\end{aligned}$$

$$= \sum_{k=\ell}^{L} \delta_k (1 + \log x_{S_k})$$
$$= \mu_\ell + \sum_{k=\ell}^{L} \delta_k \log x_{S_k} \tag{A.17}$$

where $S \equiv S_\ell \rhd S_{\ell+1} \rhd \cdots \rhd S_L \equiv \{a\}$ denotes the lineage of a up to S (inclusive). This implies that $\partial_a h_S(x) \to -\infty$ whenever $x_a \to 0$, so any solution x of (Opt_S) must have $x_a > 0$ for all $a \in S$. In view of this, the first-order optimality conditions for (Opt_S) become

$$y_a - \frac{\partial h_S}{\partial x_a} = y_a - \mu_\ell - \sum_{k=\ell}^{L} \delta_k \log x_{S_k} = \lambda \quad \text{for all } a \in S, \tag{A.18}$$

where λ is the Lagrange multiplier for the equality constraint $\sum_{a \in \mathcal{A}} x_a = 1$.[10] Thus, after rearranging terms and exponentiating, we get

$$x_{S_L}^{\delta_L} \cdot x_{S_{L-1}}^{\delta_{L-1}} \cdots x_{S_\ell}^{\delta_\ell} = \frac{\exp(y_a)}{Z}, \tag{A.19}$$

for some proportionality constant $Z \equiv Z(y) > 0$.

Step 2: Solving (Opt_S)**.** The next step of our proof will focus on unrolling the chain (A.19), one attribute at a time. To start, recall that $\delta_L = \mu_L$, so (A.19) becomes

$$x_{S_L} \cdot x_{S_{L-1}}^{\delta_{L-1}/\mu_L} \cdots x_{S_\ell}^{\delta_\ell/\mu_L} = \frac{\exp(y_{S_L}/\mu_L)}{Z^{1/\mu_L}}, \tag{A.20}$$

where we used the fact that $S_L = a$ by definition. Now, since $S_{L-1} \preccurlyeq S_\ell = S$, it follows that all children of S_{L-1} are also desendants of S, so (A.20) applies to all siblings of S_L as well. Hence, summing (A.20) over $S_L \lhd S_{L-1}$, we get

$$x_{S_{L-1}} \cdot x_{S_{L-1}}^{\delta_{L-1}/\mu_L} \cdots x_{S_\ell}^{\delta_\ell/\mu_L} = \frac{\exp(y_{S_{L-1}}/\mu_L)}{Z^{1/\mu_L}}, \tag{A.21}$$

where we used the definition (7) of $x_{S_{L-1}} = \sum_{S_L \lhd S_{L-1}} x_{S_L}$ and the recursive definition (9) for $y_{S_{L-1}}$, i.e., the fact that $\exp(y_{S_{L-1}}/\mu_L) = \sum_{S_L \lhd S_{L-1}} \exp(y_{S_L}/\mu_L)$. Therefore, noting that

$$1 + \frac{\delta_{L-1}}{\mu_L} = 1 + \frac{\mu_{L-1} - \mu_L}{\mu_L} = \frac{\mu_{L-1}}{\mu_L} \tag{A.22}$$

the product (A.21) becomes

$$x_{S_{L-1}}^{\mu_{L-1}} \cdot x_{S_{L-2}}^{\delta_{L-2}} \cdots x_{S_\ell}^{\delta_\ell} = \frac{\exp(y_{S_{L-1}})}{Z} \tag{A.23}$$

or, equivalently

$$x_{S_{L-1}} \cdot x_{S_{L-2}}^{\delta_{L-2}/\mu_{L-1}} \cdots x_{S_\ell}^{\delta_\ell/\mu_{L-1}} = \frac{\exp(y_{S_{L-1}}/\mu_{L-1})}{Z^{1/\mu_{L-1}}}. \tag{A.24}$$

This last equation has the same form as (A.21) applied to the chain $S_\ell \rhd S_{\ell+1} \rhd \cdots \rhd S_{L-1}$ instead of $S_\ell \rhd S_{\ell+1} \rhd \cdots \rhd S_L$. Thus, proceeding inductively, we conclude that

$$x_{S_k}^{\mu_k} \prod_{j=k-1}^{\ell} x_{S_j}^{\delta_j} = \frac{\exp(y_{S_k})}{Z} \quad \text{for all } k = L, \ldots, \ell \tag{A.25}$$

with the empty product $\prod_{j \in \varnothing} x_{S_j}^{\delta_j}$ taken equal to 1 by standard convention.

Now, substituting $k \leftarrow k+1$ in (A.25), we readily get

$$x_{S_{k+1}}^{\mu_{k+1}} \cdot x_{S_k}^{\delta_k} \prod_{j=k-1}^{\ell} x_{S_j}^{\delta_j} = \frac{\exp(y_{S_{k+1}})}{Z} \quad \text{for all } k = L-1, \ldots, \ell. \tag{A.26}$$

[10] Since $x_a > 0$ for all $a \in S$, the multipliers for the corresponding inequality constraints all vanish by complementary slackness.

Consequently, recalling that $\delta_k = \mu_k - \mu_{k+1}$ and dividing (A.25) by (A.26), we get

$$\frac{x_{S_{k+1}}^{\mu_{k+1}}}{x_{S_k}^{\mu_{k+1}}} = \frac{\exp(y_{S_{k+1}})}{\exp(y_{S_k})}, \tag{A.27}$$

and hence

$$\frac{x_{S_{k+1}}}{x_{S_k}} = \frac{\exp(y_{S_{k+1}}/\mu_{k+1})}{\exp(y_{S_k}/\mu_{k+1})} = P_{S_{k+1}|S_k}(y) \tag{A.28}$$

by the definition of the conditional logit choice model (NLC). Therefore, by unrolling the chain

$$x_{a|S} = \frac{x_a}{x_S} = \frac{x_{S_L}}{x_{S_{L-1}}} \cdot \frac{x_{S_{L-1}}}{x_{S_{L-2}}} \cdots \frac{x_{S_{\ell+1}}}{x_{S_\ell}} = P_{S_L|S_{L-1}}(y) \times P_{S_{L-1}|S_{L-2}}(y) \times \cdots \times P_{S_{\ell+1}|S_\ell}(y) \tag{A.29}$$

we obtain the nested expression

$$x_a = x_S \prod_{k=\ell}^{L-1} P_{S_{k+1}|S_k}(y) \quad \text{for all } a \in S. \tag{A.30}$$

Thus, with $x_S = 1$ (by the fact that $\mathrm{supp}(x) = S$), we finally conclude that

$$x_a = \prod_{k=\ell}^{L-1} P_{S_{k+1}|S_k}(y) = P_{a|S}(y) \quad \text{for all } a \in S. \tag{A.31}$$

Step 3: The maximal value of (Opt_S)**.** To obtain the value of the maximization problem (Opt_S), we will proceed to substitute (A.31) in the expression (A.4) provided by Proposition A.1 for $h_S(x)$. To that end, for all $k = \ell, \dots, L-1$ and all $S_k \preccurlyeq_k S$, the definition (A.1) of the conditional entropy gives:

$$\begin{aligned}
h(x|S_k) &= \mu_{k+1}\, x_{S_k} \sum_{S_{k+1} \lhd S_k} x_{S_{k+1}|S_k} \log x_{S_{k+1}|S_k} && \text{\# by definition} \\
&= \mu_{k+1}\, x_{S_k} \sum_{S_{k+1} \lhd S_k} x_{S_{k+1}|S_k} \log \frac{\exp(y_{S_{k+1}}/\mu_{k+1})}{\exp(y_{S_k}/\mu_{k+1})} && \text{\# by (A.28)} \\
&= x_{S_k} \sum_{S_{k+1} \lhd S_k} x_{S_{k+1}|S_k} y_{S_{k+1}} - x_{S_k} y_{S_k} && \text{\# since } \textstyle\sum_{S_{k+1} \lhd S_k} x_{S_{k+1}|S_k} = 1 \\
&= \sum_{S_{k+1} \lhd S_k} x_{S_{k+1}} y_{S_{k+1}} - x_{S_k} y_{S_k} && \text{(A.32)}
\end{aligned}$$

and hence

$$\sum_{S_k \preccurlyeq_k S} h(x|S_k) = \sum_{S_k \preccurlyeq_k S} \left[\sum_{S_{k+1} \lhd S_k} x_{S_{k+1}} y_{S_{k+1}} - x_{S_k} y_{S_k} \right] = \sum_{S_{k+1} \preccurlyeq_{k+1} S} x_{S_{k+1}} y_{S_{k+1}} - \sum_{S_k \preccurlyeq_k S} x_{S_k} y_{S_k}. \tag{A.33}$$

Thus, telescoping this last releation over $k = \ell, \dots, L$ and invoking Proposition A.1, we obtain:

$$\begin{aligned}
h_S(x) &= \sum_{S' \preccurlyeq S} h(x|S') + \mu_k\, x_S \cancel{\log x_S} && \text{\# by \textit{Proposition A.1}} \\
&= \sum_{k=\ell}^{L-1} \sum_{S_k \preccurlyeq_k S} h(x|S_k) && \text{\# collect parent classes} \\
&= \sum_{k=\ell}^{L-1} \left[\sum_{S_{k+1} \preccurlyeq_{k+1} S} x_{S_{k+1}} y_{S_{k+1}} - \sum_{S_k \preccurlyeq_k S} x_{S_k} y_{S_k} \right] && \text{\# by (A.33)} \\
&= \langle y, x \rangle - x_S y_S && \text{(A.34)}
\end{aligned}$$

where, in the second line, we used the fact that the conditional entropy $h(x|S_L)$ relative to any childless class $S_L \in \mathcal{S}_L$ is zero by definition. Accordingly, substituting back to (Opt_S) we conclude that

$$\mathrm{val}\,(\mathrm{Opt}_S) = \langle y, x \rangle - h_S(x) = x_S y_S = y_S, \tag{A.35}$$

as claimed.

Step 4: Differential representation of conditional probabilities. To prove the second part of the proposition, recall that the restricted entropy function $h_{|S}$ is convex, and let

$$h^*_{|S}(y) = \max_{x \in \Delta(\mathcal{A})} \{\langle y, x\rangle - h_{|S}(x)\} \tag{A.36}$$

denote its convex conjugate.[11] By standard results in convex analysis [e.g., Theorem 23.5 in 24], $h^*_{|S}$ is differentiable in y and we have the Legendre identity:

$$x = \nabla h^*_{|S}(y) \iff y \in \partial h_{|S}(x) \iff x \in \operatorname*{arg\,max}_{x' \in \Delta(\mathcal{A})} \{\langle y, x'\rangle - h_{|S}(x')\} \tag{A.37}$$

Now, by (A.31), we have $x_a = P_{a|S}(y)$ whenever x solves (Opt_S) and hence, by Fermat's rule, whenever $y - \partial h_{|S}(x) \ni 0$. Our claim then follows by noting that $h^*_{|S}(y) = y_S$ and combining the first and third legs of the equivalence (A.37). ■

These properties of the nested entropy function (and its restricted variant) will play a key role in deriving a suitable energy function for the nested exponential weights algorithm. We make this precise in Appendix C below.

B. Auxiliary bounds and results

Throughout this appendix, we assume the following primitives:

- A fixed sequence of real numbers $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_L > 0$; all entropy-related objects will be defined relative to this sequence as per the previous section.
- A score vector $y \in \mathbb{R}^{\mathcal{A}}$ that defines inductively the score y_S of any class $S \in \mathcal{S}$ using (9), as well as the associated nested choice probability $P(y)$ as per (NLC).
- A vector of cost increments $r = (r_S)_{S \in \mathcal{S}} \in \mathbb{R}^{\mathcal{S}}$ that defines an associated *cost vector* $c \in \mathbb{R}^{\mathcal{A}}$ as per (4), viz.

$$c_a = \sum_{S \ni a} r_S \quad \text{for all } a \in \mathcal{A}. \tag{B.1}$$

Moreover, for all $c, y \in \mathbb{R}^{\mathcal{A}}$, we define the *nested power sum* function $\sigma_{c,y} \colon \mathcal{S}\backslash\mathcal{S}_L \to \mathbb{R}$ which, to any $S \in \mathcal{S}\backslash\mathcal{S}_L$, associates the real number

$$\sigma_{c,y}(S) = \begin{cases} \sum_{a \lhd S} P_{a|S}(y) \exp(-c_a/\mu_L) & \text{if } \operatorname{attr}(S) = L-1, \\ \sum_{S' \lhd S} P_{S'|S}(y) \sigma_{c,y}(S')^{\frac{\mu_{\ell+2}}{\mu_{\ell+1}}} & \text{if } \operatorname{attr}(S) = \ell < L-1. \end{cases} \tag{B.2}$$

The following lemma links the increments of the conjugate entropy h^* to the nested power sum defined above:

Lemma B.1. *For all $y \in \mathbb{R}^{\mathcal{A}}$, $c \in \mathbb{R}^{\mathcal{A}}$, we have*

$$h^*(y - c) = h^*(y) + \mu_1 \log(\sigma_{c,y}(\mathcal{A})). \tag{B.3}$$

Lemma B.1 will be proved as a corollary of the more general result below:

Lemma B.2. *Fix some $y \in \mathbb{R}^{\mathcal{A}}$ and $c \in \mathbb{R}^{\mathcal{A}}$. Then, for all $S_\ell \in \mathcal{S}_\ell$, $\ell < L$,we have*

$$\exp\left(\frac{h^*_{|S_\ell}(y-c)}{\mu_{\ell+1}}\right) = \exp\left(\frac{h^*_{|S_\ell}(y)}{\mu_{\ell+1}}\right) \sigma_{c,y}(S_\ell) \tag{B.4}$$

Proof of Lemma B.1. Simply invoke Lemma B.2 with $S \leftarrow \mathcal{A}$. ■

Proof of Lemma B.2. We proceed by descending induction on $\ell = \operatorname{attr}(S)$.

[11] Note here that $h^*_{|S}(y)$ is bounded from above by the convex conjugate $h^*_S(y)$ of $h_S(x)$ because the latter does not include the constraint $\operatorname{supp}(x) \subseteq S$.

Base step. Fix some $S \in \mathcal{S}$ with $\mathrm{attr}(S) = L - 1$. We then have:

$$\begin{aligned}
\exp\left(\frac{h^*_{|S}(y-c)}{\mu_L}\right) &= \sum_{a \lhd S} \exp\left(\frac{h^*_{|a}(y-c)}{\mu_L}\right) && \text{\# by Eq. (A.11)} \\
&= \sum_{a \lhd S} \exp\left(\frac{h^*_{|a}(y) - c_a}{\mu_L}\right) && \text{\# the } a\text{'s are leaves} \\
&= \sum_{a \lhd S} \exp\left(\frac{h^*_{|a}(y)}{\mu_L}\right) \exp\left(-\frac{c_a}{\mu_L}\right) \\
&= \exp\left(\frac{h^*_{|S}(y)}{\mu_L}\right) \underbrace{\sum_{a \lhd S} \left[\frac{\exp\left(\frac{h^*_{|a}(y)}{\mu_L}\right)}{\exp\left(\frac{h^*_{|S}(y)}{\mu_L}\right)}\right] \exp\left(-\frac{c_a}{\mu_L}\right)}_{=\sigma_{c,y}(S) \text{ by definition}} \\
&= \exp\left(\frac{h^*_{|S}(y)}{\mu_L}\right) \sigma_{c,y}(S) && \text{(B.5)}
\end{aligned}$$

with the last equality following from the definition of $P_{a|S}$ via (NLC) and by the definition of $\sigma_{c,y}(S)$. This concludes the start of the induction process.

Induction step. Fix some $S \in \mathcal{S}$ with $\mathrm{attr}(S) = \ell - 1$, $\ell < L$, and suppose that (B.4) holds at level ℓ. We then have:

$$\begin{aligned}
\exp\left(\frac{h^*_{|S}(y-c)}{\mu_\ell}\right) &= \sum_{S' \lhd S} \exp\left(\frac{h^*_{|S'}(y-c)}{\mu_\ell}\right) \\
&= \sum_{S' \lhd S} \exp\left(\frac{h^*_{|S'}(y-c)}{\mu_{\ell+1}}\right)^{\frac{\mu_{\ell+1}}{\mu_\ell}} \\
&= \sum_{S' \lhd S} \left[\exp\left(\frac{h^*_{|S'}(y)}{\mu_{\ell+1}}\right) \sigma_{c,y}(S')\right]^{\frac{\mu_{\ell+1}}{\mu_\ell}} && \text{\# inductive hypothesis} \\
&= \sum_{S' \lhd S} \exp\left(\frac{h^*_{|S'}(y)}{\mu_\ell}\right) \sigma_{c,y}(S')^{\frac{\mu_{\ell+1}}{\mu_\ell}} \\
&= \exp\left(\frac{h^*_{|S}(y)}{\mu_\ell}\right) \underbrace{\sum_{S' \lhd S} \left[\frac{\exp\left(\frac{h^*_{|S'}(y)}{\mu_\ell}\right)}{\exp\left(\frac{h^*_{|S}(y)}{\mu_\ell}\right)}\right] \sigma_{c,y}(S')^{\frac{\mu_{\ell+1}}{\mu_\ell}}}_{=\sigma_{c,y}(S) \text{ by definition}} \\
&= \exp\left(\frac{h^*_{|S}(y)}{\mu_L}\right) \sigma_{c,y}(S) && \text{(B.6)}
\end{aligned}$$

with the last equality following from the definition of $P_{S'|S}$ and $\sigma_{c,y}(S)$. This being true for all $S \in \mathcal{S}$ with $\mathrm{attr}(S) = \ell - 1$, the inductive step and – a fortiori – our proof are complete. ■

The next lemma provides an upper bound for $\sigma_{c,y}(\mathcal{A})$, which will in turn allow us to derive a bound for the increment of h^*.

Lemma B.3. *For $y \in \mathbb{R}^{\mathcal{A}}$ and $c \in [0, +\infty)^{\mathcal{A}}$, we have:*

$$\sigma_{c,y}(\mathcal{A}) \leq 1 - \frac{1}{\mu_1}\left[\sum_{a \in \mathcal{A}} P_a(y) c_a - \frac{1}{2\mu_L} \sum_{a \in \mathcal{A}} P_a(y) c_a^2\right]. \qquad \text{(B.7)}$$

As in the case of B.1, Lemma B.3 will follow as a special case of the more general, class-based result below:

Lemma B.4. *Fix some $y \in \mathbb{R}^{\mathcal{A}}$ and $c \in \mathbb{R}^{\mathcal{A}}_{+}$. Then, for all $S_\ell \in \mathcal{S}_\ell$, $\ell < L$,we have*

$$\sigma_{c,y}(S_\ell) \leq 1 - \frac{1}{\mu_{\ell+1}} \left[\sum_{a \in S_\ell} P_{a|S_\ell}(y) c_a - \frac{1}{2\mu_L} \sum_{a \in S_\ell} P_{a|S_\ell}(y) c_a^2 \right], \tag{B.8}$$

Proof of Lemma B.3. Simply invoke Lemma B.4 with $S \leftarrow \mathcal{A}$. ■

Proof of Lemma B.4. We proceed again by descending induction on $\ell = \mathrm{attr}(S)$.

Base step. Fix some $S \in \mathcal{S}$ with $\mathrm{attr}(S) = L - 1$. We then have:

$$\begin{aligned}
\sigma_{c,y}(S) &= \sum_{S' \lhd S} P_{S'|S}(y) \exp(-\frac{c_{S'}}{\mu_L}) \\
&\leq \sum_{S' \lhd S} P_{S'|S}(y)(1 - \frac{c_{S'}}{\mu_L} \frac{c_{S'}^2}{2\mu_L^2}) && \text{\# } e^{-x} \leq 1 - x + x^2/2 \text{ for } x \geq 0 \\
&= 1 - \frac{1}{\mu_L} \left[\sum_{S' \lhd S} P_{S'|S}(y) c_{S'} - \frac{1}{2\mu_L} \sum_{S' \lhd S} P_{S'|S}(y) c_{S'}^2 \right] \\
&= 1 - \frac{1}{\mu_{(L-1)+1}} \left[\sum_{a \lhd S} P_{a|S}(y) c_a - \frac{1}{2\mu_L} \sum_{a \lhd S} P_{a|S}(y) c_a^2 \right]
\end{aligned} \tag{B.9}$$

so the initialization of the induction process is complete.

Induction step. Fix some $S \in \mathcal{S}$ with $\mathrm{attr}(S) = \ell - 1$, $\ell < L$, and suppose that (B.8) holds at level ℓ. We then have:

$$\begin{aligned}
\sigma_{c,y}(S) &= \sum_{S' \lhd S} P_{S'|S}(y) \sigma_{c,y}(S')^{\frac{\mu_{\ell+1}}{\mu_\ell}} \\
&= \sum_{S' \lhd S} P_{S'|S}(y) \left[1 + \frac{1}{\mu_{\ell+1}} \left(- \sum_{a \lhd S'} P_{a|S'}(y) c_a + \frac{1}{2\mu_L} \sum_{a \lhd S'} P_{a|S'}(y) c_a^2 \right) \right]^{\frac{\mu_{\ell+1}}{\mu_\ell}} && \text{\# inductive hypothesis} \\
&\leq \sum_{S' \lhd S} P_{S'|S}(y) \left[1 + \frac{1}{\mu_\ell} \left(- \sum_{a \lhd S'} P_{a|S'}(y) c_a + \frac{1}{2\mu_L} \sum_{a \lhd S'} P_{a|S'}(y) c_a^2 \right) \right] && \text{\# } (1+x)^\beta \leq 1 + \beta x \text{ for } \beta \leq 1 \\
&= 1 + \frac{1}{\mu_\ell} \left[- \sum_{S' \lhd S} \sum_{a \lhd S'} P_{a|S'}(y) P_{S'|S}(y) c_a + \frac{1}{2\mu_L} \sum_{S' \lhd S} \sum_{a \lhd S'} P_{a|S'}(y) P_{S'|S}(y) c_a^2 \right] && \text{(B.10)} \\
&= 1 + \frac{1}{\mu_{(\ell-1)+1}} \left[\sum_{a \lhd S} P_{a|S}(y) c_a + \frac{1}{2\mu_L} \sum_{a \lhd S} P_{a|S}(y) c_a^2 \right] && \text{(B.11)}
\end{aligned}$$

This being true for all $S \in \mathcal{S}$ s.t. $\mathrm{attr}(S) = \ell - 1$, the induction step and the proof of our assertion are complete. ■

With all this in hand, we are now in a position to upper bound the increments of the conjugate nested entropy h^*.

Proposition B.1. *For $y \in \mathbb{R}^{\mathcal{A}}$ and $c \in [0, +\infty)^{\mathcal{A}}$, we have:*

$$h^*(y - c) - h^*(y) \leq -\langle P(y), c \rangle + \frac{1}{2\mu_L} \sum_{a \in \mathcal{A}} P_a(y) c_a^2. \tag{B.12}$$

Proof. Using Lemmas B.1 and B.3 and the concavity inequality $\log x \leq 1 + x$ directly delivers our assertion. ■

Remark 6. It is useful to note that, given a cost increment vector $r \in \mathbb{R}^{\mathcal{S}}$ with associated aggregate costs given by $c \in \mathbb{R}^{\mathcal{A}}$ we have:

$$\begin{aligned}\langle P(y), c\rangle &= \sum_{a\in\mathcal{A}} P_a(y)c_a \\ &= \sum_{a\in\mathcal{A}} P_a(y) \sum_{S\ni a} r_S \\ &= \sum_{a\in\mathcal{A}} P_a(y) \sum_{S\in\mathcal{S}} r_S \mathbb{1}_{a\in S} \\ &= \sum_{S\in\mathcal{S}} \left[\sum_{a\in\mathcal{A}} P_a(y) \mathbb{1}_{a\in S}\right] r_S \\ &= \sum_{S\in\mathcal{S}} P_S(y) r_S.\end{aligned}$$

We are finally in a position to prove the basic properties of the NIWE estimator, which we restate below for convenience:

Proposition 1. *Let $\mathcal{S} = \coprod_{\ell=1}^{L} \mathcal{S}_\ell$ be a similarity structure on $\mathcal{A}$. Then, given a mixed strategy $x \in \Delta(\mathcal{A})$ and a vector of cost increments $r \in \mathbb{R}^{\mathcal{S}}$ as per (5), the estimator (NIWE) satisfies the following:*

1. It is unbiased:

$$\mathbb{E}[\hat{r}_S] = r_S \quad \textit{for all } S \in \mathcal{S}. \tag{14}$$

2. It enjoys the importance-weighted mean-square bound

$$\mathbb{E}\big[x_S \hat{r}_S^2\big] \le R_S^2 \quad \textit{for all } S \in \mathcal{S}. \tag{15}$$

Accordingly, the loss estimator (13) is itself unbiased and enjoys the bound

$$\mathbb{E}\Big[\sum\nolimits_{a\in\mathcal{A}} x_a \hat{c}_a^2\Big] \le n_{\text{eff}} \tag{16}$$

where n_{eff} is defined as

$$\sqrt{n_{\text{eff}}} = \sum\nolimits_{\ell=1}^{L} \sqrt{n_\ell}\bar{R}_\ell \tag{17}$$

with $n_\ell = |\mathcal{S}_\ell|$ denoting the number of classes of attribute $\mathcal{S}_\ell$, and

$$\bar{R}_\ell = \sqrt{\frac{1}{n_\ell}\sum\nolimits_{S_\ell\in\mathcal{S}_\ell} R_{S_\ell}^2} \tag{18}$$

denoting the "root-mean-square" range of all classes in $\mathcal{S}_\ell$.

Proof. Fix some $S \in \mathcal{S}$ with $\operatorname{attr}(S) = \ell \in \{1, \dots, L\}$ and lineage $\mathcal{A} \equiv S_0 \rhd S_1 \rhd \cdots \rhd S_\ell \equiv S$. We will now prove both properties of the (NIWE) estimator.

Part 1. We begin by showing that the estimator (NIWE) is unbiased. Indeed, we have:

$$\begin{aligned}\mathbb{E}[\hat{r}_S] &= \mathbb{E}\left[\frac{\mathbb{1}\{S_\ell = \hat{S}_\ell, \dots, S_1 = \hat{S}_1\}}{x_{S_\ell|S_{\ell-1}}\cdots x_{S_2|S_1}x_{S_1}} r_{S_\ell}\right] = \mathbb{E}\left[\frac{\mathbb{1}\{S = \hat{S}\}}{x_S} r_S\right] && \textit{\# Rewriting (NIWE)} \\ &= \frac{r_S}{x_S}\underbrace{\mathbb{E}\Big[\mathbb{1}\{S = \hat{S}\}\Big]}_{x_S} = r_S. && \text{(B.13)}\end{aligned}$$

Part 2. We now turn to the proof of the importance-weighted mean-square bound of the estimator (NIWE). In this case, for any $S \in \mathcal{S}$, we have:

$$\mathbb{E}\big[x_S \hat{r}_S^2\big] = x_S\,\mathbb{E}\big[\hat{r}_S^2\big] = x_S\,\mathbb{E}\left[\left(\frac{\mathbb{1}\{S = \hat{S}\}}{x_S} r_{S_\ell}\right)^2\right]$$

$$= x_S \frac{r_{S_\ell}^2}{x_S^2} \mathbb{E}\Big[\mathbb{1}\big\{S = \hat{S}\big\}\Big] = r_{S_\ell}^2 \qquad \text{\# because } \mathbb{E}\big[\mathbb{1}\big\{S = \hat{S}\big\}\big] = x_S$$
$$\leq R_S^2. \tag{B.14}$$

We are left to derive the bound for the aggregate cost estimator (13), viz.

$$\hat{c}_a = \sum_{S \ni a} \hat{r}_S. \tag{B.15}$$

With this in mind, we can write:

$$\begin{aligned}
\sum_{a\in\mathcal{A}} x_a \hat{c}_a^2 &= \sum_{a\in\mathcal{A}} x_a \left(\sum_{S\ni a} \hat{r}_S\right)^2 \\
&= \sum_{a\in\mathcal{A}} x_a \left[\sum_{S\ni a} \hat{r}_S^2 + 2 \sum_{S'\ni a} \sum_{S\succ S'} \hat{r}_S \hat{r}_{S'}\right] \\
&= \sum_{a\in\mathcal{A}} \sum_{S\in\mathcal{S}} x_a \hat{r}_S^2 \, \mathbb{1}_{a\in S} + 2 \sum_{a\in\mathcal{A}} \sum_{S'\in\mathcal{S}} \sum_{S\succ S'} x_a \hat{r}_S \hat{r}_{S'} \, \mathbb{1}_{a\in S'} \\
&= \sum_{S\in\mathcal{S}} \hat{r}_S^2 \underbrace{\sum_{a\in\mathcal{A}} x_a \, \mathbb{1}_{a\in S}}_{x_S} + 2 \sum_{S'\in\mathcal{S}} \sum_{S\succ S'} \hat{r}_S \hat{r}_{S'} \underbrace{\sum_{a\in\mathcal{A}} x_a \, \mathbb{1}_{a\in S'}}_{x_{S'}} \\
&= \sum_{S\in\mathcal{S}} x_S \hat{r}_S^2 + 2 \sum_{S'\in\mathcal{S}} \sum_{S\succ S'} x_{S'} \hat{r}_S \hat{r}_{S'}.
\end{aligned} \tag{B.16}$$

Now, decomposing the above sums attribute-by-attribute and taking expectations in (B.16), we get:

$$\mathbb{E}\left[\sum_{a\in\mathcal{A}} x_a \hat{c}_a^2\right] = \sum_{\ell=1}^{L} \sum_{S_\ell\in\mathcal{S}_\ell} x_{S_\ell} \, \mathbb{E}\big[\hat{r}_{S_\ell}^2\big] + 2 \sum_{1\leq\ell<\ell'\leq L} \sum_{\substack{S_\ell\in\mathcal{S}_\ell \\ S_{\ell'}\prec_{\ell'}\mathcal{S}_\ell}} x_{S_{\ell'}} \, \mathbb{E}\big[\hat{r}_{S_\ell} \hat{r}_{S_{\ell'}}\big]. \tag{B.17}$$

The first term in (B.17) can simply be bounded using (B.14). Indeed:

$$\sum_{\ell=1}^{L} \sum_{S_\ell\in\mathcal{S}_\ell} x_{S_\ell} \, \mathbb{E}\big[\hat{r}_{S_\ell}^2\big] \leq \sum_{\ell=1}^{L} \sum_{S_\ell\in\mathcal{S}_\ell} R_{S_\ell}^2 = \sum_{\ell=1}^{L} n_\ell \bar{R}_\ell^2. \tag{B.18}$$

with $\bar{R}_\ell = \sqrt{\frac{1}{n_\ell} \sum_{S_\ell\in\mathcal{S}_\ell} R_{S_\ell}^2}$ for any $\ell = 1, \dots, L$.

We now turn to the second term in (B.17). Let $\{\epsilon_{\ell,\ell'}\}_{1\leq\ell'<\ell\leq L}$ be any fixed sequence of positive numbers. For any $\ell, \ell' \in \{1, \dots, L\}$ and any $S_\ell \in \mathcal{S}_\ell$ and $S_{\ell'} \in \mathcal{S}_{\ell'}$, the Peter-Paul inequality yields:

$$2 \hat{r}_{S_{\ell'}} \hat{r}_{S_\ell} \leq \frac{1}{\epsilon_{\ell,\ell'}} \hat{r}_{S_{\ell'}}^2 + \epsilon_{\ell,\ell'} \hat{r}_{S_\ell}^2 \tag{B.19}$$

Injecting (B.19) into the second term of (B.17) yields:

$$\begin{aligned}
&2 \sum_{1\leq\ell<\ell'\leq L} \sum_{\substack{S_\ell\in\mathcal{S}_\ell \\ S_{\ell'}\prec_{\ell'}\mathcal{S}_\ell}} x_{S_{\ell'}} \, \mathbb{E}\big[\hat{r}_{S_\ell} \hat{r}_{S_{\ell'}}\big] \\
&\qquad \leq \sum_{1\leq\ell<\ell'\leq L} \sum_{\substack{S_\ell\in\mathcal{S}_\ell \\ S_{\ell'}\prec_{\ell'}\mathcal{S}_\ell}} x_{S_{\ell'}} \left(\frac{1}{\epsilon_{\ell,\ell'}} \mathbb{E}\Big[\hat{r}_{S_{\ell'}}^2\Big] + \epsilon_{\ell,\ell'} \, \mathbb{E}\big[\hat{r}_{S_\ell}^2\big]\right) \\
&\qquad = \sum_{1\leq\ell<\ell'\leq L} \frac{1}{\epsilon_{\ell,\ell'}} \sum_{\substack{S_\ell\in\mathcal{S}_\ell \\ S_{\ell'}\prec_{\ell'}\mathcal{S}_\ell}} x_{S_{\ell'}} \, \mathbb{E}\Big[\hat{r}_{S_{\ell'}}^2\Big] + \sum_{1\leq\ell<\ell'\leq L} \epsilon_{\ell,\ell'} \sum_{\substack{S_\ell\in\mathcal{S}_\ell \\ S_{\ell'}\prec_{\ell'}\mathcal{S}_\ell}} x_{S_{\ell'}} \, \mathbb{E}\big[\hat{r}_{S_\ell}^2\big]
\end{aligned}$$

$$= \sum_{1\le\ell<\ell'\le L} \frac{1}{\epsilon_{\ell,\ell'}} \sum_{\substack{S_\ell\in\mathcal{S}_\ell \\ S_{\ell'}\prec_{\ell'}\mathcal{S}_\ell}} x_{S_{\ell'}}\, \mathbb{E}\Big[\hat{r}^2_{S_{\ell'}}\Big] + \sum_{1\le\ell<\ell'\le L} \epsilon_{\ell,\ell'} \sum_{S_\ell\in\mathcal{S}_\ell} \mathbb{E}\big[\hat{r}^2_{S_\ell}\big] \underbrace{\sum_{S_{\ell'}\prec_{\ell'}\mathcal{S}_\ell} x_{S_{\ell'}}}_{x_{S_\ell}}$$

$$= \sum_{1\le\ell<\ell'\le L} \frac{1}{\epsilon_{\ell,\ell'}} \sum_{S_{\ell'}\in\mathcal{S}_{\ell'}} x_{S_{\ell'}}\, \mathbb{E}\Big[\hat{r}^2_{S_{\ell'}}\Big] + \sum_{1\le\ell<\ell'\le L} \epsilon_{\ell,\ell'} \sum_{S_\ell\in\mathcal{S}_\ell} x_{S_\ell}\, \mathbb{E}\big[\hat{r}^2_{S_\ell}\big]$$

$$\le \sum_{1\le\ell<\ell'\le L} \frac{1}{\epsilon_{\ell,\ell'}} \sum_{S_{\ell'}\in\mathcal{S}_{\ell'}} R^2_{S_{\ell'}} + \sum_{1\le\ell<\ell'\le L} \epsilon_{\ell,\ell'} \sum_{S_\ell\in\mathcal{S}_\ell} R^2_{S_\ell} \qquad \text{\# by (B.14)}$$

$$\le \sum_{1\le\ell<\ell'\le L} \frac{1}{\epsilon_{\ell,\ell'}} n_{\ell'}\bar{R}^2_{\ell'} + \sum_{1\le\ell<\ell'\le L} \epsilon_{\ell,\ell'} n_\ell \bar{R}^2_\ell. \tag{B.20}$$

Injecting (B.18) and (B.20) into (B.17) ensures that:

$$\mathbb{E}\Big[\sum\nolimits_{a\in\mathcal{A}} x_a\hat{c}_a^2\Big] \le \sum\nolimits_{\ell=1}^{L} n_\ell\bar{R}^2_\ell + \sum\nolimits_{1\le\ell<\ell'\le L}\left(\frac{1}{\epsilon_{\ell,\ell'}} n_{\ell'}\bar{R}^2_{\ell'} + \epsilon_{\ell,\ell'} n_\ell\bar{R}^2_\ell\right)$$

holds for any sequence of positive numbers $\{\epsilon_{\ell,\ell'}\}_{1\le\ell'<\ell\le L}$. As a result, taking $\epsilon_{\ell,\ell'} = \sqrt{\frac{n_{\ell'}}{n_\ell}}\frac{\bar{R}_{\ell'}}{\bar{R}_\ell}$ yields the tight bound

$$\mathbb{E}\Big[\sum\nolimits_{a\in\mathcal{A}} x_a\hat{c}_a^2\Big] \le \sum\nolimits_{\ell=1}^{L} n_\ell\bar{R}^2_\ell + 2\sum\nolimits_{1\le\ell<\ell'\le L}\sqrt{n_{\ell'}}\bar{R}_{\ell'}\sqrt{n_\ell}\bar{R}^2_\ell = \left(\sum\nolimits_{\ell=1}^{L}\sqrt{n_\ell}\bar{R}_\ell\right)^2, \tag{B.21}$$

which proves our original assertion. ■

C. Regret analysis

As we mentioned in the main text, the principal component of our analysis is a recursive inequality which, when telescoped over $t = 1, 2, \ldots$, will yield the desired regret bound. To establish this "template inequality", we will first require an energy function measuring the disparity between a benchmark strategy $x \in \Delta(\mathcal{A})$ and a propensity score profile $y \in \mathbb{R}^{\mathcal{A}}$. To that end, building on the notions introduced in Appendix A, let $h\colon \Delta(\mathcal{A}) \to \mathbb{R}$ denote the total nested entropy function

$$h(x) = h_{\mathcal{A}}(x) = \sum_{k=0}^{L} \delta_k \sum_{S_k\in\mathcal{S}_k} x_{S_k}\log x_{S_k}, \quad x\in\Delta(\mathcal{A}), \tag{C.1}$$

and let

$$h^*(y) = \max_{x\in\Delta(\mathcal{A})}\{\langle y, x\rangle - h(x)\}, \qquad y\in\mathbb{R}^{\mathcal{A}}, \tag{C.2}$$

denote the convex conjugate of h so, by Proposition A.2, we have

$$h^*(y) = y_{\mathcal{A}} \quad \text{and} \quad P_a(y) = \frac{\partial h^*}{\partial y_a} \quad \text{for all } y\in\mathbb{R}^{\mathcal{A}}. \tag{C.3}$$

The *Fenchel coupling* between $x \in \Delta(\mathcal{A})$ and $y \in \mathbb{R}^{\mathcal{A}}$ is then defined as

$$F(x, y) = h(x) + h^*(y) - \langle y, x\rangle \quad \text{for all } x\in\Delta(\mathcal{A}),\, y\in\mathbb{R}^{\mathcal{A}}, \tag{C.4}$$

and we have the following key result:

Proposition C.1. *Let $\mathcal{S} = \coprod_{\ell=0}^{L}\mathcal{S}_\ell$ be a similarity structure on $\mathcal{A}$ with uncertainty parameters $\mu_1 \ge \cdots \ge \mu_L > 0$. Then:*

1. The Fenchel coupling (C.4) is positive-definite, i.e.,

$$F(x, y) \ge 0 \qquad \textit{for all } x\in\Delta(\mathcal{A}) \textit{ and all } y\in\mathbb{R}^{\mathcal{A}}, \tag{C.5}$$

with equality if and only if x is given by (NLC), i.e., if and only if $x = P(y)$.

2. *For all $x \in \mathcal{A}$, we have*

$$F(x, 0) = h(x) + h^*(0) = h(x) - \min h \tag{C.6}$$

where $\min h \equiv \min_{x' \in \Delta(\mathcal{A})} h(x')$ *denotes the minimum of* h *over* $\Delta(\mathcal{A})$.

Proof. Our first claim follows by setting $S \leftarrow \mathcal{A}$ in Propositions A.1 and A.2 and noting that $h_S = h_{|S}$ when $S = \mathcal{A}$: indeed, by Young's inequality, we have $h(x) + h^*(y) - \langle y, x\rangle \geq 0$ with equality if and only if $y \in \partial h(x)$, so the equality $x = P(y)$ follows from (A.37) applied to $S \leftarrow \mathcal{A}$ and the fact that $P_{a|\mathcal{A}}(y) = P_a(y)$. As for our second claim, simply note that $h^*(0) = \max_{x \in \Delta(\mathcal{A})}\{\langle 0, x\rangle - h(x)\} = -\min_{x \in \Delta(\mathcal{A})} h(x)$ and set $y \leftarrow 0$ in the definition (C.4) of the Fenchel coupling. ∎

With all this in hand, the specific energy function that we will use for our regret analysis is the "*rate-deflated*" Fenchel coupling

$$E_t = \frac{1}{\eta_t} F(p, \eta_t y_t) \tag{C.7}$$

where $p \in \Delta(\mathcal{A})$ is the regret comparator, η_t is the algorithm's learning rate at stage t, and y_t is the corrsponding propensity score estimate. In words, since the mixed strategy employed by the learner at stage t is $x_t = P(\eta_t y_t)$, the energy E_t essentially measures the disparity between x_t and the target strategy p (suitably rescaled by the method's learning rate). We then have the following fundamental estimate:

Proposition C.2. *For all $p \in \Delta(\mathcal{A})$ and all $t = 1, 2, \dots$, we have:*

$$E_{t+1} \leq E_t + \langle \hat{c}_t, x_t - p\rangle + (\eta_{t+1}^{-1} - \eta_t^{-1})[h(p) - \min h] + \frac{1}{\eta_t} F(x_t, \eta_t y_{t+1}). \tag{C.8}$$

Proof. By the definition of E_t, we have

$$E_{t+1} - E_t = \frac{1}{\eta_{t+1}} F(p, \eta_{t+1} y_{t+1}) - \frac{1}{\eta_t} F(p, \eta_t y_t) = \frac{1}{\eta_{t+1}} F(p, \eta_{t+1} y_{t+1}) - \frac{1}{\eta_t} F(p, \eta_t y_{t+1}) \tag{C.9a}$$

$$+ \frac{1}{\eta_t} F(p, \eta_t y_{t+1}) - \frac{1}{\eta_t} F(p, \eta_t y_t). \tag{C.9b}$$

We now proceed to upper-bound each of the two terms (C.9a) and (C.9b) separately.

For the term (C.9a), the definition of the Fenchel coupling (C.4) readily yields:

$$\text{(C.9a)} = \left[\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right] h(p) + \frac{1}{\eta_{t+1}} h^*(\eta_{t+1} y_{t+1}) - \frac{1}{\eta_t} h^*(\eta_t y_{t+1}). \tag{C.10}$$

Inspired by a trick of Nesterov [22], consider the function $\varphi(\eta) = \eta^{-1}[h^*(\eta y) + \min h]$. Then, by Proposition A.2, letting $x = P(\eta y)$ and differentiating φ with respect to η gives

$$\begin{aligned}
\varphi'(\eta) &= \frac{1}{\eta}\langle y, P(\eta y)\rangle - \frac{1}{\eta^2}[h^*(\eta y) + \min h] \\
&= \frac{1}{\eta^2}[\langle \eta y, x\rangle - h^*(\eta y) - \min h] \\
&= \frac{1}{\eta^2}[h(x) - \min h] \geq 0.
\end{aligned} \tag{C.11}$$

Since $\eta_{t+1} \leq \eta_t$, the above shows that $\varphi(\eta_t) \geq \varphi(\eta_{t+1})$. Accordingly, setting $y \leftarrow y_{t+1}$ in the definition of φ yields

$$\frac{1}{\eta_{t+1}} h^*(\eta_{t+1} y_{t+1}) - \frac{1}{\eta_t} h^*(\eta_t y_{t+1}) \leq \left[\frac{1}{\eta_t} - \frac{1}{\eta_{t+1}}\right] \min h \tag{C.12}$$

and hence

$$\text{(C.9a)} \leq (\eta_{t+1}^{-1} - \eta_t^{-1})[h(p) - \min h]. \tag{C.13}$$

Now, after a straightforward rearrangement, the second term of (C.9) becomes

$$\begin{aligned}
\text{(C.9b)} &= \frac{1}{\eta_t}[h(p) + h^*(\eta_t y_{t+1}) - \eta_t\langle y_{t+1}, p\rangle] - \frac{1}{\eta_t}[h(p) + h^*(\eta_t y_t) - \eta_t\langle y_t, p\rangle] \\
&= \frac{1}{\eta_t}[h^*(\eta_t y_{t+1}) - h^*(\eta_t y_t) - \eta_t\langle \hat{c}_t, p\rangle] && \text{\# by (NEW)} \\
&= \frac{1}{\eta_t}[h^*(\eta_t y_{t+1}) - h^*(\eta_t y_t) - \eta_t\langle \hat{c}_t, x_t\rangle] + \langle \hat{c}_t, x_t - p\rangle && \text{\# isolate benchmark} \\
&= \frac{1}{\eta_t}[h^*(\eta_t y_{t+1}) - \langle \eta_t y_t, x_t\rangle + h(x_t) - \eta_t\langle \hat{c}_t, x_t\rangle] + \langle \hat{c}_t, x_t - p\rangle && \text{\# by Proposition A.2} \\
&= \frac{1}{\eta_t}F(x_t, \eta_t y_{t+1}) + \langle \hat{c}_t, x_t - p\rangle && \text{(C.14)}
\end{aligned}$$

Thus, combining the above with (C.13), we finally obtain

$$\begin{aligned}
E_{t+1} &= E_t + \text{(C.9a)} + \text{(C.9b)} \\
&\leq E_t + (\eta_{t+1}^{-1} - \eta_t^{-1})[h(p) - \min h] + \langle \hat{c}_t, x_t - p\rangle + \frac{1}{\eta_t}F(x_t, \eta_t y_{t+1}) && \text{(C.15)}
\end{aligned}$$

and our proof is complete. ∎

We are now in a position to state and prove the template inequality that provides the scaffolding for our regret bounds:

Proposition 2. *The NEW algorithm enjoys the bound*

$$\mathbb{E}[\mathrm{Reg}_p(T)] \leq \frac{H}{\eta_{T+1}} + \sum_{t=1}^{T} \frac{\mathbb{E}[F(x_t, \eta_t y_{t+1})]}{\eta_t}. \tag{29}$$

Proof. Let $Z_t = \hat{c}_t - v_t$ denote the error in the learner's estimation of the t-th stage payoff vector v_t. Then, by substituting in Proposition C.2 and rearranging, we readily get:

$$\langle v_t, p - x_t\rangle \leq E_t - E_{t+1} + \langle Z_t, x_t - p\rangle + \big(\eta_{t+1}^{-1} - \eta_t^{-1}\big)[h(p) - \min h] + \eta_t F(p, \eta_t y_{t+1}) \tag{C.16}$$

Thus, telescoping over $t = 1, 2, \dots, T$, we have

$$\begin{aligned}
\mathrm{Reg}_p(T) &\leq E_1 - E_{T+1} + \left(\frac{1}{\eta_{T+1}} - \frac{1}{\eta_1}\right)[h(p) - \min h] + \sum_{t=1}^{T}\langle Z_t, x_t - p\rangle + \sum_{t=1}^{T}\frac{1}{\eta_t}F(x_t, \eta_t y_{t+1}) \\
&\leq \frac{h(p) - \min h}{\eta_{T+1}} + \sum_{t=1}^{T}\langle Z_t, x_t - p\rangle + \sum_{t=1}^{T}\frac{1}{\eta_t}F(x_t, \eta_t y_{t+1}) && \text{(C.17)}
\end{aligned}$$

where we used the fact that *a)* $E_t \geq 0$ for all t (a consequence of the first part of Proposition C.1); and that *b)* $E_1 = \eta_1^{-1}[h(p) + h^*(0)] = \eta_1^{-1}[h(p) - \min h]$ (from the second part of the same proposition). Our claim then follows by taking expectations in (C.17) and noting that $\mathbb{E}[Z_t \,|\, \mathcal{F}_t] = 0$ (by Proposition 1). ∎

In view of the above, our main regret bound follows by bounding the two terms in the template inequality (C.8). The second term is by far the most difficult one to bound, and is where Appendix B comes in; the first term is easier to handle, and it can be bounded as follows:

Lemma C.1. *Suppose that each class $S \in \mathcal{S}_{\ell-1}$ has at most m_ℓ children, $\ell = 1, \dots, L$. Then, for all $p \in \Delta(\mathcal{A})$, we have*

$$H \leq \sum_{\ell=1}^{L} \mu_\ell \log m_\ell \qquad \textit{with equality iff the tree is symmetric,} \tag{C.18}$$

$$H = \mu \log(n) \qquad \textit{if } \mu_1 = \mu_2 = \cdots = \mu_L = \mu. \tag{C.19}$$

Proof. Suppose that $y_a = 0$ for all $a \in \mathcal{A}$. Then, applying (9) inductively, we have:

$$\begin{aligned}
y_{S_L} &= 0 && \text{for all } S_L \in \mathcal{S}_L \\
y_{S_{L-1}} &= \mu_L \log \sum_{S_L \lhd S_{L-1}} \exp(0) \leq \mu_L \log m_L && \text{for all } S_{L-1} \in \mathcal{S}_{L-1} \\
y_{S_{L-2}} &= \mu_{L-1} \log \sum_{S_{L-1} \lhd S_{L-2}} \exp(y_{S_{L-1}}/\mu_{L-1}) \leq \mu_{L-1} \log m_{L-1} + \mu_L \log m_L && \text{for all } S_{L-2} \in \mathcal{S}_{L-2} \\
&\vdots && \vdots \\
y_{S_{\ell-1}} &= \mu_\ell \log \sum_{S_\ell \lhd S_{\ell-1}} \exp(y_{S_\ell}/\mu_\ell) \leq \sum_{k=\ell}^{L} \mu_k \log m_k && \text{for all } S_{\ell-1} \in \mathcal{S}_{\ell-1}
\end{aligned} \tag{C.20}$$

and hence $H = h^*(0) = y_{\mathcal{A}} \leq \sum_{\ell=1}^{L} \mu_\ell \log m_\ell$. Eq. (C.18) then follows from Proposition C.1.

Now, if $\mu_1 = \mu_2 = \cdots = \mu_L = \mu$, we have

$$\begin{aligned}
H &= \log \left[\sum_{S_1 \lhd S_0} \left[\sum_{S_2 \lhd S_1} \cdots \left[\sum_{S_L \lhd S_{L-1}} 1 \right]^{\frac{\mu_L}{\mu_{L-1}}} \cdots \right]^{\frac{\mu_2}{\mu_1}} \right]^{\mu_1} \\
&= \mu \log \sum_{S_1 \lhd S_0} \left[\sum_{S_2 \lhd S_1} \cdots \left[\sum_{S_L \lhd S_{L-1}} 1 \right] \cdots \right] \\
&= \mu \log \left[\sum_{S_L \lhd_L S_0} 1 \right] = \mu \log n,
\end{aligned} \tag{C.21}$$

which proves Eq. (C.19) and completes our proof. ∎

Proposition C.3. *For all $p \in \Delta(\mathcal{A})$ and all $t = \{1, 2, \dots\}$, we have:*

$$F(x_t, \mu_t y_{t+1}) + \eta_t \langle \hat{c}_t, x_t \rangle = h^*(\eta_t y_t + \eta_t \hat{c}_t) - h^*(\eta_t y_t). \tag{C.22}$$

Proof. Let $p \in \Delta(\mathcal{A})$ and $t \in 1, 2, \dots$, we simply write:

$$\begin{aligned}
F(x_t, \eta_t y_{t+1}) &= h(x_t) + h^*(\eta_t y_{t+1}) - \eta_t \langle y_{t+1}, x_t \rangle \\
&= \underbrace{h(x_t) + h^*(\eta_t y_t) - \langle \eta_t y_t, x_t \rangle}_{= F(x_t, \eta_t y_t)} + h^*(\eta_t y_{t+1}) - h^*(\eta_t y_t) - \eta_t \langle \hat{c}_t, x_t \rangle \\
&= h^*(\eta_t y_t + \eta_t \hat{c}_t) - h^*(y_t) - \eta_t \langle \hat{c}_t, x_t \rangle && \# \, F(x_t, \eta_t y_t) = 0
\end{aligned}$$

and our assertion follows. ∎

We are finally in a position to prove our main result (which we restate below for convenience):

Theorem 1. *Suppose that Algorithm 1 is run with a non-increasing learning rate $\eta_t > 0$ and uncertainty parameters $\mu_1 \geq \cdots \geq \mu_L > 0$ against a sequence of cost vectors $c_t \in [0,1]^{\mathcal{A}}$, $t = 1, 2, \dots$, as per (4). Then, for all $p \in \Delta(\mathcal{A})$, the learner enjoys the regret bound*

$$\mathbb{E}[\mathrm{Reg}_p(T)] \leq \frac{H}{\eta_{T+1}} + \frac{n_{\mathrm{eff}}}{2\mu_L} \sum_{t=1}^{T} \eta_t \tag{19}$$

with n_{eff} given by (17) and $H \equiv H(\mu_1, \dots, \mu_L)$ defined by setting $y = 0$ in (9) and taking $H = y_{\mathcal{A}}$, i.e.,

$$H = \log \left[\sum_{S_1 \lhd S_0} \left[\sum_{S_2 \lhd S_1} \cdots \left[\sum_{S_L \lhd S_{L-1}} 1 \right]^{\frac{\mu_L}{\mu_{L-1}}} \cdots \right]^{\frac{\mu_2}{\mu_1}} \right]^{\mu_1} \tag{20}$$

In particular, if Algorithm 1 is run with $\mu_1 = \cdots = \mu_L = \sqrt{n_{\text{eff}}/2}$ *and* $\eta_t = \sqrt{\log n/(2t)}$*, we have*

$$\mathbb{E}[\text{Reg}_p(T)] \leq 2\sqrt{n_{\text{eff}} \log n \cdot T}. \tag{21}$$

Proof. Injecting Eq. (C.22) in the result of Proposition 2 and using Proposition B.1 and Eq. (16) of Proposition 1 directly yields the pseudo-regret bound (19).

Finally, if we choose $\mu_1 = \cdots = \mu_L = \sqrt{n_{\text{eff}}/2}$, Lemma C.1 gives

$$H = \sqrt{n_{\text{eff}}/2} \log n. \tag{C.23}$$

Thus, taking $\eta_t = \sqrt{\log n/(2t)}$ and substituting in (19) along with (C.23) finally delivers

$$\mathbb{E}[\text{Reg}_p(T)] \leq 2\sqrt{n_{\text{eff}} \log n \cdot T}, \tag{C.24}$$

and our claim follows. ■

D. Additional Experiment Details and Discussions

In this appendix we provide additional details on the experiments as well as further discussions on the settings we presented. The code with the implementation of the algorithms as well as the code to reproduce the figures will be open-sourced and is provided along with the supplementary materials.

D.1. Experiment additional details

In the synthetic environment, at each level, the rewards are generated randomly according for each class nodes, through uniform distributions of randomly generated means and fixed bandwidth. From a level ℓ to the next $\ell + 1$, the rewards range are divided by a multiplicative factor $R_\ell / R_{\ell+1} = 10$. The implemented method of NEW uses the reward based IW. Moreover, no model selection was used in this experiment as no hyperparameter was tuned. Indeed, a decaying rate of $\frac{1}{\sqrt{t}}$ was used for the score updates for all methods, as is common in the bandit litterature [18].

D.2. Blue Bus/ Red Bus environment

We detail in Figure 5 a graphical representation of such blue bus/red bus environment, where many colors of the bus item build irrelevant alternatives. In this setting, with few arms, we run the methods up to the horizon $T = 1000$. We provide in Figure 6 the average reward of the two methods NEW and EXP3 with varying number of subclasses of the "bus".

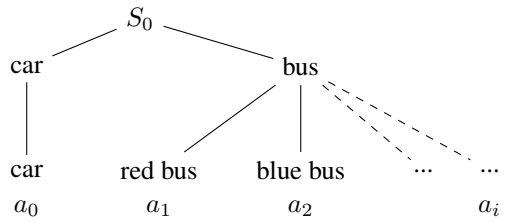

Figure 5: Diagram of the blue Bus/Red Bus environment.

While the NEW method ends up selecting the best alternative and having the lowest regret, the EXP3 seems to pick wrong alternative in some experiments, and ends up having higher regret and requiring more iterations to converge to higher average reward. In some of our experiments over the multiple random runs, alternatives of very low sampling probability that were sampled changed the score vector too brutally in the IPS estimator which seemed to hurt the EXP3 method much more than the NEW algorithm.

D.3. Tree structures

In this appendix we show additional results and visualisations for the second setting presented in the main paper. We start with discussions on the depth parameter L and follow with the breadth parameter related to the number of child per class $M = |S|$.

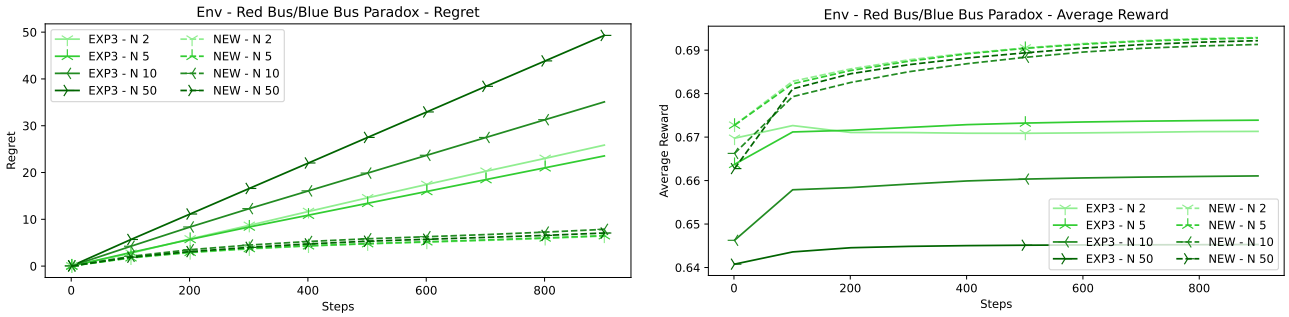

Figure 6: Regret and Average Reward of NEW and EXP3 on the Blue Bus/ Red Bus environment.

Influence of the depth parameter L In Figure 7 we show the influence of the depth parameter with a fixed number of child per class. By making the tree deeper, we illustrate the effect of knowing the nested structure compared to running the logit choice to the whole alternative set. As shown in both the regret and average reward plots, the NEW method outperforms the EXP3 algorithm. While the NEW method also use an IPS estimator, it is less prone to variance issues than the EXP3 method. Indeed, due to the nested structure and the reward decay related to the ratio $R_{\ell+1}/R_{\ell}$, the NEW estimator end up not hurting the regret by still selecting "right" parent classes.

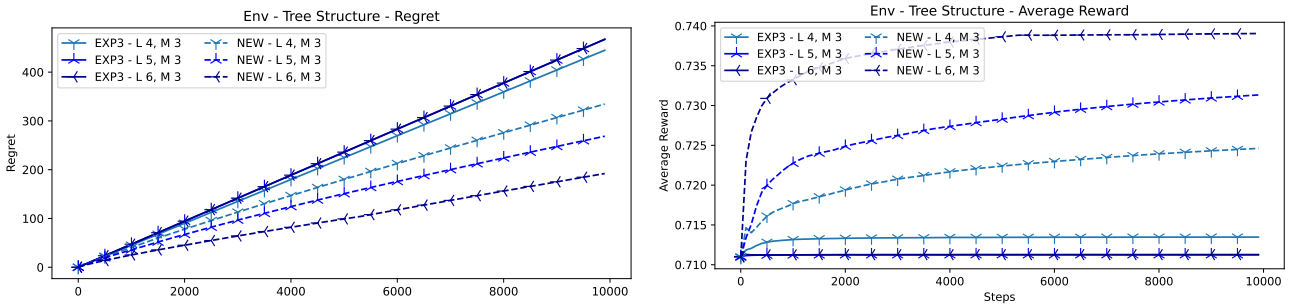

Figure 7: Regret and Average Reward of NEW and EXP3 on the synthetic environment with varying number of levels L.

Influence of the number of child per class (wideness) $M = |S|$ In this setting we fix the number of levels L and vary the number of child per classes M. In Figure 8 we can see that the NEW method outperforms the EXP3 in terms of regret and average reward. Interestingly, we see that the gap between the two methods shrinks when the number of child per class augments. This is because when the size of a class increase, the NEW method also end up having less knowledge locally and end up having a large number of alternatives to choose among.

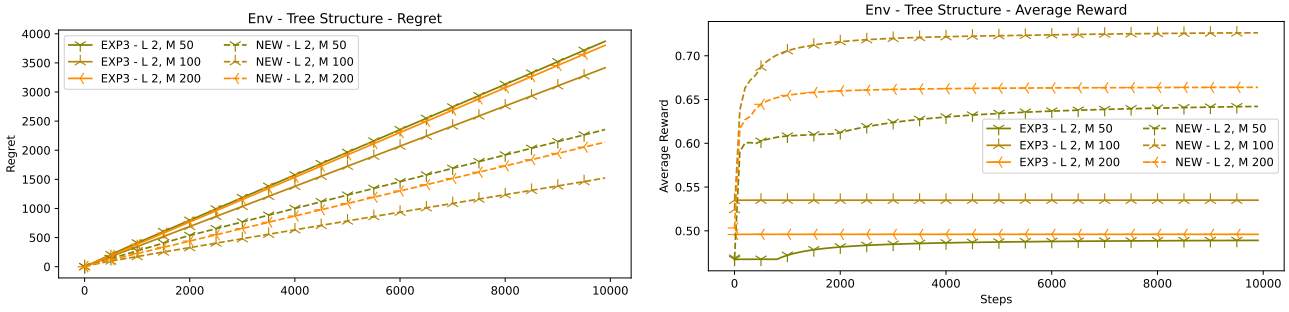

Figure 8: Regret and Average Reward of NEW and EXP3 on the synthetic environment with varying number of child per class $M = |S|$.

D.4. A visualisation of the effects of NEW

In this appendix we want to show the effects of NEW through the simple setting where we assume a nested structure with $L = 4$ and $M = |S| = 3$. We illustrate in Figure 9 the score vectors of the NEW method along the optimal path in the tree (path which nodes have the highest cumulated mean, i.e which generates the highest reward) along with the oracle means of the child nodes. We can see that the algorithm takes advantage of the nested structure and updates the scores vectors optimally with regards to the oracle means of all the nodes. The NEW algorithm therefore estimates correctly the rewards of the environment.

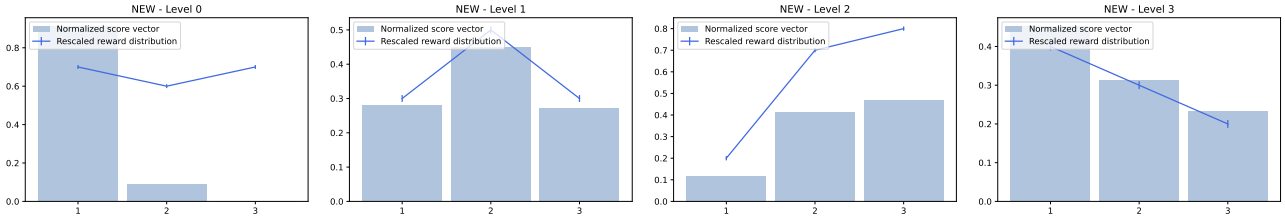

Figure 9: Histograms of the score vectors along the optimal path in the nested structure, with visualisation of the mean value of the node.

Inversely we see in Figure 10 that the EXP3 method has suffered from variance issue and selected a suboptimal alternative among the $|S|^L = 81$ possible ones. The EXP3 did not take advantage of the nested structure and therefore did not learn as correctly as the NEW algorithm the reward values.

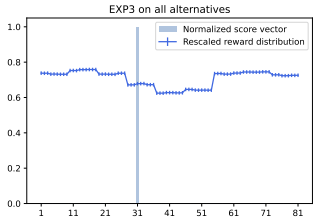

Figure 10: Histogram of the score vector of the all alternatives, with a visualisation of the mean value of all nodes.

D.5. Cases where both algorithms perform identically

In this appendix we merely show that the implementation of the NEW and EXP3 algorithm match exactly and observe the same behavior when the number of levels L is set to 1. This setting is where we have no knowledge of any nested structure, therefore both algorithms perform identically in Figure 11.

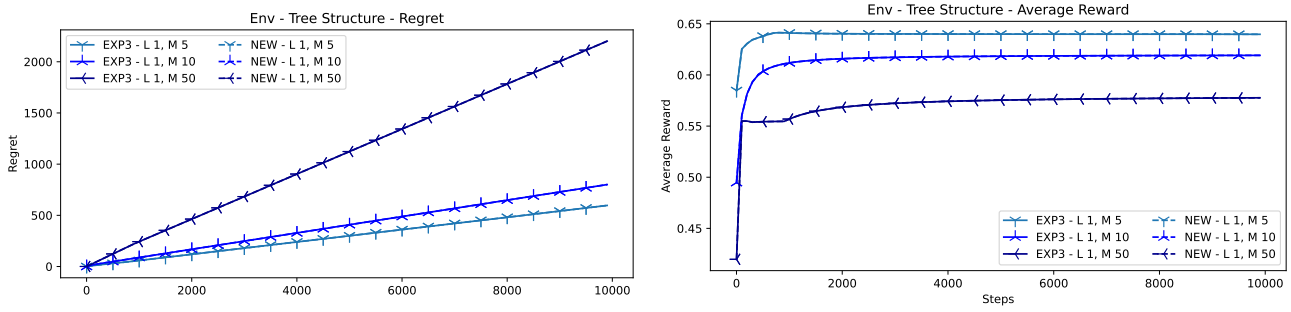

Figure 11: Regret and Average Reward of NEW and EXP3 on the synthetic environment where $L = 1$.

D.6. Variance plots for the synthetic experiments

We discuss here the variance of the regret at the final timestep $T = 10000$. Indeed, as shown on Figure 6 for the NEW algorithm , on Figure 7 for both algorithms EXP3 and NEW, and on Figure 8 for EXP3, some of the plots do no exhibit the

monotonicity one would have expected when increasing the number of arms through L or M, and are even overlapping on the regret plot. This can be explained on Figures 12 for the Red Bus/Blue Bus environment, and in Figures 13 and 14 respectively for depth and wideness tree experiments. Those plots show the variances (across the 20 random seeds) of the final regret for both methods at the final step-size. In Figure 13 we see that the EXP3 arms have similar mean values with large variances, which explains why they are overlapping on the plot in Figure 3. In Figure 14 when varying M we can also have a closer look on how NEW outperforms EXP3 and how the close values of NEW regrets through different M can be explained by their high variance.

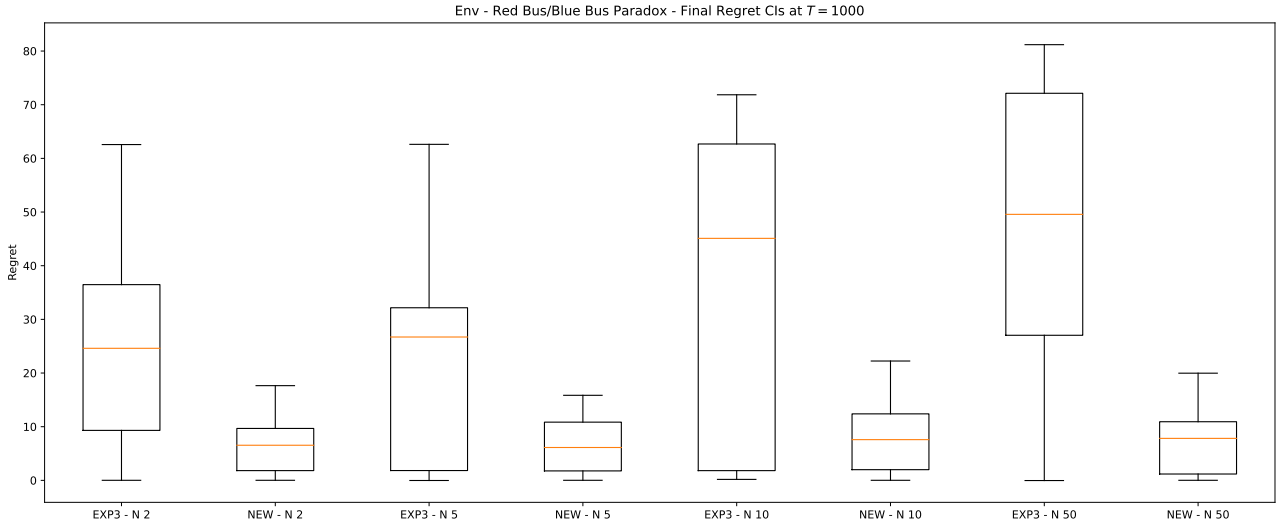

Figure 12: Regret distribution at the final stepsize $T = 1000$ for the Red Bus/Blue Bus environment.

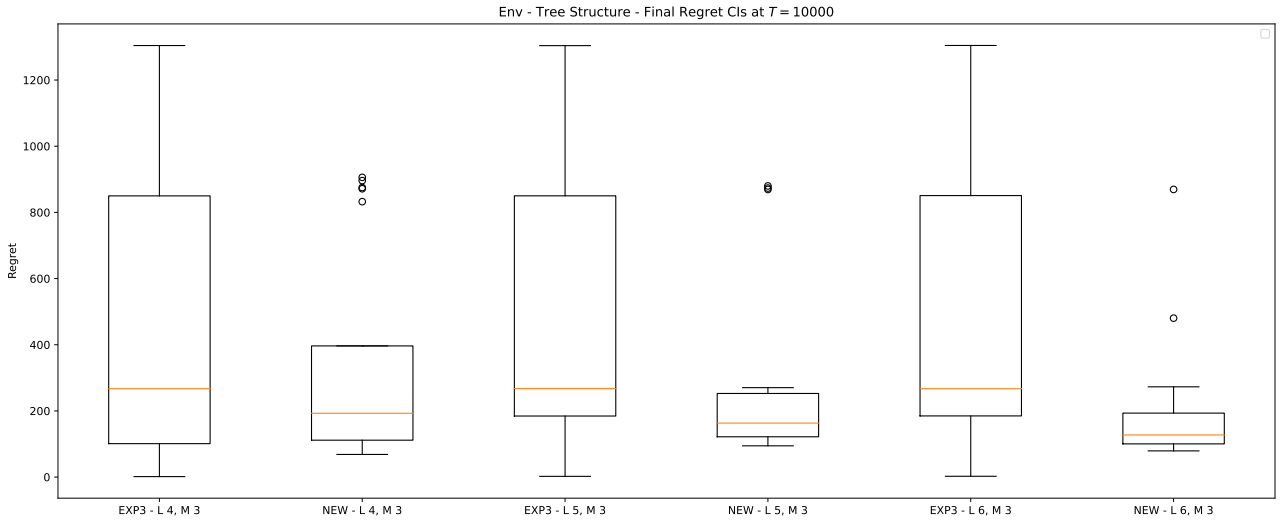

Figure 13: Regret distribution at the final stepsize $T = 10000$ when varying the depth parameter L.

D.7. Reproducibility

We provide code for reproducibility of our experiments and plots, in addition to a more general implementation of both the NEW algorithm and EXP3 baseline. All experiments were run on a Mac book pro laptop, with 1 processor of 6 cores @2.6GHz (6-Core Intel Core i7). The code and all experiments can be found in the attached .zip.

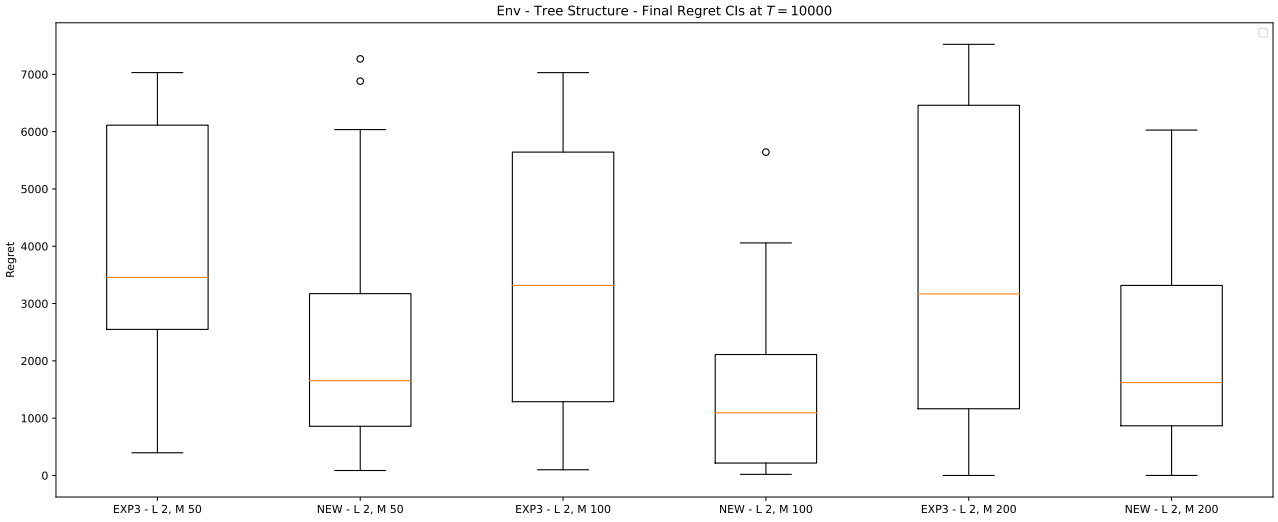

Figure 14: Regret distribution at the final stepsize $T = 10000$ when varying the wideness parameter M.