Online Multitask Learning with Long-Term Memory

Mark Herbster, Stephen Pasteris, Lisa Tse
Department of Computer Science
University College London
London
United Kingdom
(m.herbster|s.pasteris|l.tse)@cs.ucl.ac.uk

Abstract

We introduce a novel online multitask setting. In this setting each task is partitioned into a sequence of segments that is unknown to the learner. Associated with each segment is a hypothesis from some hypothesis class. We give algorithms that are designed to exploit the scenario where there are many such segments but significantly fewer associated hypotheses. We prove regret bounds that hold for any segmentation of the tasks and any association of hypotheses to the segments. In the single-task setting this is equivalent to *switching with long-term memory* in the sense of [1]. We provide an algorithm that predicts on each trial in time linear in the number of hypotheses when the hypothesis class is finite. We also consider infinite hypothesis classes from reproducing kernel Hilbert spaces for which we give an algorithm whose per trial time complexity is cubic in the number of cumulative trials. In the single-task special case this is the first example of an efficient regret-bounded switching algorithm with long-term memory for a non-parametric hypothesis class.

1 Introduction

We consider a model of online prediction in a non-stationary environment with multiple interrelated tasks. Associated with each task is an asynchronous data stream. As an example, consider a scenario where a team of drones may need to decontaminate an area of toxic waste. In this example, the tasks correspond to drones. Each drone is receiving a data stream from its sensors. The data streams are non-stationary but interdependent as the drones are travelling within a common site. At any point in time, a drone receives an instance x and is required to predict its label y. The aim is to minimize mispredictions. As is standard in regret-bounded learning we have no statistical assumptions on the data-generation process. Instead, we aim to predict well relative to some hypothesis class of predictors. Unlike a standard regret model, where we aim to predict well in comparison to a single hypothesis, we instead aim to predict well relative to a completely unknown sequence of hypotheses in each task's data stream, as illustrated by the "*coloring*" in Figure 1. Each *mode* (color) corresponds to a distinct hypothesis from the hypothesis class. A *switch* is said to have occurred whenever we move between modes temporally within the same task.

Thus in task 1, there are three modes and four switches. We are particularly motivated by the case that a mode once present will possibly recur multiple times even within different tasks, i.e., "modes" $\ll$ "switches." We will give algorithms and regret bounds for finite hypothesis classes (the "experts" model [2, 3, 4]) and for infinite non-parametric Reproducing Kernel Hilbert Space (RKHS) [5] hypothesis classes.

The paper is organized as follows. In the next section, we introduce our formal model for online switching multitask learning. In doing so we provide a brief review of some related online learning results which enable us to provide a prospectus for attainable regret bounds. This is done by

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

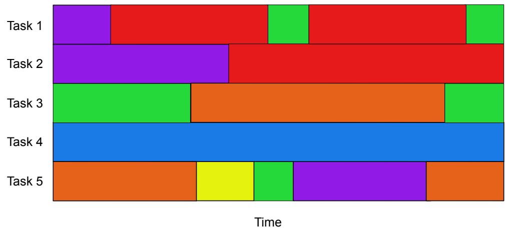

Figure 1: A Coloring of Data Streams (5 tasks, 6 modes, and 11 switches).

For $\tau = 1$ to T do
  Receive task $\ell^\tau \in [s]$.
  Set $i \leftarrow \ell^\tau$; $t \leftarrow \boldsymbol{\sigma}(\tau)$.
  Receive instance $x^\tau \equiv x_t^i \in \mathcal{X}$.
  Predict $\hat{y}^\tau \equiv \hat{y}_t^i \in \{-1, 1\}$.
  Receive label $y^\tau \equiv y_t^i \in \{-1, 1\}$.
  Incur Loss $\mathcal{L}_{01}(y^\tau, \hat{y}^\tau)$.

Figure 2: The Switching Multitask Model

considering the bounds achievable by non-polynomial time algorithms. We then provide a brief survey of related work as well as our notational conventions. In Sections 3 and 4 we provide algorithms and bounds for finite hypothesis classes and RKHS hypothesis classes, respectively. Finally, we provide a few concluding remarks in Section 5. The supplementary appendices contain our proofs.

2 Online Learning with Switching, Memory, and Multiple Tasks

We review the models and regret bounds for online learning in the single-task, switching, and switching with memory models as background for our multitask switching model with memory.

In the single-task online model a *learner* receives data sequentially so that on a trial $t = 1, \ldots, T$: 1) the learner receives an instance $x_t \in \mathcal{X}$ from the *environment*, then 2) predicts a label $\hat{y}_t \in \{-1, 1\}$, then 3) receives a label from the environment $y_t \in \{-1, 1\}$ and then 4) incurs a *zero-one* loss $\mathcal{L}_{01}(y_t, \hat{y}_t) := [y_t \neq \hat{y}_t]$. There are no probabilistic assumptions on how the environment generates its instances or their labels; it is an arbitrary process which in fact may be adversarial. The only restriction on the environment is that it does not "see" the learner's $\hat{y}_t$ until after it reveals y_t. The learner's aim will be to compete with a hypothesis class of predictors $\mathcal{H} \subseteq \{-1, 1\}^{\mathcal{X}}$ so as to minimize its *expected regret*, $R_T(h) := \sum_{t=1}^{T} \mathbb{E}[\mathcal{L}_{01}(y_t, \hat{y}_t)] - \mathcal{L}_{01}(y_t, h(x_t))$ for every hypothesis $h \in \mathcal{H}$, where the expectation is with respect to the learner's internal randomization.

In this paper we will consider two types of hypothesis classes: a finite set of hypotheses $\mathcal{H}_{\text{fin}}$, and a set $\mathcal{H}_K$ induced by a kernel K. A "multiplicative weight" (MW) algorithm [6] that achieves a regret bound[1] of the form

$$R_T(h) \in \mathcal{O}\left(\sqrt{\log(|\mathcal{H}_{\text{fin}}|)T}\right) \quad (\forall h \in \mathcal{H}_{\text{fin}}) \tag{1}$$

was given in [7] for finite hypothesis classes. This is a special case of the framework of "prediction with expert advice" introduced in [2, 3]. Given a reproducing kernel $K : \mathcal{X} \times \mathcal{X} \to \Re$ we denote the induced norm of the reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ as $\|\cdot\|_K$ (for details on RKHS see [5] also Appendix C.2). Given an instance sequence $\boldsymbol{x} := (x_1, \ldots, x_T)$, we let $\mathcal{H}_K^{(\boldsymbol{x})} := \{h \in \mathcal{H}_K : h(x_t) \in \{-1, 1\}, \forall t \in [T]\}$ denote the functions in $\mathcal{H}_K$ that are binary-valued on the sequence. An analysis of online gradient descent (OGD_K) with the hinge loss, kernel K and randomized prediction [8, see e.g., Ch. 2 & 3] (proof included in Appendix C.3 for completeness) gives an expected regret bound of

$$R_T(h) \in \mathcal{O}\left(\sqrt{\|h\|_K^2 X_K^2 T}\right) \quad (\forall h \in \mathcal{H}_K^{(\boldsymbol{x})}), \tag{2}$$

where $X_K^2 \geq \max_{t \in [T]} K(x_t, x_t)$.

In the switching single-task model the hypothesis becomes a sequence of hypotheses $\boldsymbol{h} = (h_1, h_2, \ldots, h_T) \in \mathcal{H}^T$ and the regret is $R_T(\boldsymbol{h}) := \sum_{t=1}^{T} \mathbb{E}[\mathcal{L}_{01}(y_t, \hat{y}_t)] - \mathcal{L}_{01}(y_t, h_t(x_t))$. Two parameters of interest are the number of *switches* $k := \sum_{t=1}^{T-1} [h_t \neq h_{t+1}]$ and the number of *modes* $m := |\cup_{t=1}^{T} \{h_t\}|$, i.e., the number of the distinct hypotheses in the sequence. In this work we are

[1]Technically, when we say that an algorithm *achieves a bound*, it may be that the algorithm depends on a small set of parameters which we have then assumed are "tuned" optimally.

interested in *long-term memory*, that is, algorithms and bounds that are designed to exploit the case of $m \ll k$.

The methodology of [9] may be used to derive an expected regret bound for $\mathcal{H}_{\text{fin}}$ in the switching single-task model of the form $R_T(\boldsymbol{h}) \in \mathcal{O}(\sqrt{(k\log(|\mathcal{H}_{\text{fin}}|)+k\log(T/k))T})$. Freund in [10] posed an open problem to improve the results of [9] in the case of long-term memory ($m \ll k$). Freund gave counting arguments that led to an exponential-time algorithm with a regret bound of $R_T(\boldsymbol{h}) \in \mathcal{O}(\sqrt{(m\log(|\mathcal{H}_{\text{fin}}|)+k\log m+k\log(T/k))T})$. In [1] an efficient algorithm was given with nearly this bound, except for a small additional additive "$T\log\log T$" term under the square root. For the hypothesis class $\mathcal{H}_K^{(\boldsymbol{x})}$ we may give non-memory bounds of the form $R_T(\boldsymbol{h}) \in \mathcal{O}(\sqrt{k\max_t \|h_t\|_K^2 X_K^2 T})$ by using a simple modification [11] of OGD_K (see Appendix C.3). To the best of our knowledge there are no previous long-term memory bounds for $\mathcal{H}_K^{(\boldsymbol{x})}$ (however see the discussion of [12] in Section 2.2); these will be a special case of our multitask model, to be introduced next.

2.1 Switching Multitask Model

In Figure 2 we illustrate the protocol for our multitask model. The model is essentially the same as the switching single-task model, except that we now have s tasks. On each (global) trial τ the environment reveals the active task $\ell^\tau \in [s]$. The ordering of tasks chosen by the environment is arbitrary, and therefore the active task may change on every (global) trial τ. We use the following notational convention: (global time) $\tau \equiv {}^i_t$ (local time) where $i=\ell^\tau$, $t=\boldsymbol{\sigma}(\tau)$ and $\boldsymbol{\sigma}(\tau) := \sum_{j=1}^{\tau}[\ell^j=\ell^\tau]$. Thus $x^\tau \equiv x_t^i$, $y^\tau \equiv y_t^i$, etc., where the mapping is determined implicitly by the task vector $\boldsymbol{\ell} \in [s]^T$. Each task $i \in [s]$ has its own data pair (instance, label) sequence $(x_1^i, y_1^i), \ldots, (x_{T^i}^i, y_{T^i}^i)$ where $T = T^1 + \ldots + T^s$. The *multitask hypotheses multiset* is denoted as $\boldsymbol{h}^* = (h^1, \ldots, h^T) \equiv (h_1^1, \ldots, h_{T^1}^1, \ldots, h_1^s, \ldots, h_{T^s}^s) \in \mathcal{H}^T$. In the multitask model we denote the number of switches as $k(\boldsymbol{h}^*) := \sum_{i=1}^s \sum_{t=1}^{T^i-1}[h_t^i \neq h_{t+1}^i]$, the set of modes as $m(\boldsymbol{h}^*) := \cup_{i=1}^s \cup_{t=1}^{T^i} \{h_t^i\}$ and the multitask regret as $R_T(\boldsymbol{h}^*) := \sum_{i=1}^s \sum_{t=1}^{T^i} \mathbb{E}[\mathcal{L}_{01}(y_t^i, \hat{y}_t^i)] - \mathcal{L}_{01}(y_t^i, h_t^i(x_t^i))$. In the following, we give motivating upper bounds based on exponential-time algorithms induced by "meta-experts." We provide a lower bound with respect to $\mathcal{H}_K^{(\boldsymbol{x})}$ in Proposition 4.

The idea of "meta-experts" is to take the base class of hypotheses and to construct a class of "meta-hypotheses" by combining the original hypotheses to form new ones, and then apply an MW algorithm to the constructed class; in other words, we reduce the "meta-model" to the "base-model." In our setting, the base class is $\mathcal{H}_{\text{fin}} \subseteq \{-1,1\}^{\mathcal{X}}$ and our meta-hypothesis class will be some $\mathcal{H}' \subseteq \{-1,1\}^{\mathcal{X}'}$ where $\mathcal{X}' := \{(x,t,i) : x \in \mathcal{X}, t \in [T^i], i \in [s]\}$. To construct this set we define $\bar{\mathcal{H}}(k,m,s,\mathcal{H}_{\text{fin}},T^1,\ldots,T^s) := \{(h_1^1,\ldots,h_{T^s}^s) = \bar{\boldsymbol{h}} \in \mathcal{H}_{\text{fin}}^T : k = k(\bar{\boldsymbol{h}}), m = |m(\bar{\boldsymbol{h}})|\}$ and then observe that for each $\bar{\boldsymbol{h}} \in \bar{\mathcal{H}}$ we may define an $h' : \mathcal{X}' \to \{-1,1\}$ via $h'((x,t,i)) := h_t^i(x)$, where h_t^i is an element of $\bar{\boldsymbol{h}}$. We thus construct $\mathcal{H}'$ by converting each $\bar{\boldsymbol{h}} \in \bar{\mathcal{H}}$ to an $h' \in \mathcal{H}'$. Hence we have reduced the switching multitask model to the single-task model with respect to $\mathcal{H}'$. We proceed to obtain a bound by observing that the cardinality of $\bar{\mathcal{H}}$ is bounded above by $\binom{T-s}{k}\binom{n}{m}m^s(m-1)^k$ where $n = |\mathcal{H}_{\text{fin}}|$. If we then substitute into (1) and then further upper bound we have

$$R_T(\boldsymbol{h}^*) \in \mathcal{O}\left(\sqrt{(m\log(n/m)+s\log m+k\log m+k\log((T-s)/k))T}\right), \tag{3}$$

for any $\boldsymbol{h}^* \in \mathcal{H}_{\text{fin}}^T$ such that $k = k(\boldsymbol{h}^*)$ and $m = |m(\boldsymbol{h}^*)|$. The drawback is that the algorithm requires exponential time. In Section 3 we will give an algorithm whose time to predict per trial is $\mathcal{O}(|\mathcal{H}_{\text{fin}}|)$ and whose bound is equivalent up to constant factors.

We cannot directly adapt the above argument to obtain an algorithm and bound for $\mathcal{H}_K^{(\boldsymbol{x})}$ since the cardinality, in general, is infinite, and additionally we do not know $\boldsymbol{x}$ in advance. However, the structure of the argument is the same. Instead of using hypotheses from $\mathcal{H}_K^{(\boldsymbol{x})}$ as building blocks to construct meta-hypotheses, we use multiple instantiations of an online algorithm for $\mathcal{H}_K^{(\boldsymbol{x})}$ as our building blocks. We let $\mathcal{A}_K := \{a[1],\ldots,a[m]\}$ denote our set of m instantiations that will act as a surrogate for the hypothesis class $\mathcal{H}_K^{(\boldsymbol{x})}$. We then construct the set, $\bar{\mathcal{A}}_K(k,m,s,T^1,\ldots,T^s) := \{\bar{\boldsymbol{a}} \in \mathcal{A}_K^T : k = k(\bar{\boldsymbol{a}}), m = |m(\bar{\boldsymbol{a}})|\}$. Each $\bar{\boldsymbol{a}} \in \bar{\mathcal{A}}_K$ now defines a meta-algorithm for the multitask setting. That is, given an online multitask data sequence $(x^i_1, y^i_1), \ldots, (x^j_{T^j}, y^j_{T^j})$, each element of $\bar{\boldsymbol{a}}$ will "color" the corresponding data pair with one of the m instantiations (we will use the function $\alpha : \{(t,i) : t \in [T^i], i \in [s]\} \to [m]$ to denote this mapping with respect to $\bar{\boldsymbol{a}}$). Each instantiation will *receive as inputs only the online sequence of the data pairs corresponding to its "color"*; likewise, the prediction of meta-algorithm $\bar{\boldsymbol{a}}$ will be that of the instantiation active on that trial. We will use as our base algorithm OGD_K. Thus for the meta-algorithm $\bar{\boldsymbol{a}}$ we have from (2),

$$\sum_{i=1}^{s}\sum_{t=1}^{T^i} \mathbb{E}[\mathcal{L}_{01}(y^i_t, \hat{y}^i_t)] \leq \sum_{i=1}^{s}\sum_{t=1}^{T^i} \mathcal{L}_{01}(y^i_t, h[\alpha(^i_t)](x^i_t)) + \sum_{j=1}^{m} \mathcal{O}\left(\sqrt{\|h[j]\|^2_K X^2 T^j}\right) \tag{4}$$

for any received instance sequence $\boldsymbol{x} \in \mathcal{X}^T$ and for any $h[1], \ldots, h[m] \in \mathcal{H}^{(\boldsymbol{x})}_K$. The MW algorithm [3, 2, 4] does not work just for hypothesis classes; more generally, it works for collections of algorithms. Hence we may run the MW as a meta-meta-algorithm to combine all of the meta-algorithms $\bar{\boldsymbol{a}} \in \bar{\mathcal{A}}_K$. Thus by substituting the loss for each meta-algorithm $\bar{\boldsymbol{a}}$ (the R.H.S. of (4)) into (1) and using the upper bound $\binom{T-s}{k} m^s (m-1)^k$ for the cardinality of $\bar{\mathcal{A}}_K$, we obtain (using upper bounds for binomial coefficients and the inequality $\sum_i \sqrt{p_i q_i} \leq \sqrt{(\sum_i p_i)(\sum_i q_i)}$),

$$R_T(\boldsymbol{h}^*) \in \mathcal{O}\left(\sqrt{(\textstyle\sum_{h \in m(\boldsymbol{h}^*)} \|h\|^2_K X^2_K + s \log m + k \log m + k \log((T-s)/k))T}\right), \tag{5}$$

for any received instance sequence $\boldsymbol{x} \in \mathcal{X}^T$ and for any $\boldsymbol{h}^* \in \mathcal{H}^{(\boldsymbol{x})^T}_K$ such that $k = k(\boldsymbol{h}^*)$ and $m = |m(\boldsymbol{h}^*)|$.

The terms $m \log(n/m)$ (assuming $m \ll n$) and $\sum_{h \in m(\boldsymbol{h}^*)} \|h\|^2_K X^2_K$ may be viewed as *learner complexities*, i.e., the price we "pay" for identifying the hypotheses that fit the modes. A salient feature of long-term memory bounds is that although the data pairs associated with each hypothesis are intermixed in the multitask sequence, we pay the learner complexity only modestly in terms of potentially leading multiplicative constants. A switching algorithm without long-term memory "forgets" and pays the full price for a mode on every switch or new task. We gave exponential-time algorithms for $\mathcal{H}_{\mathrm{fin}}$ and $\mathcal{H}^{(\boldsymbol{x})}_K$ with $\mathcal{O}(1)$ leading multiplicative constants in the discussion leading to (3) and (5). We give efficient algorithms for finite hypothesis classes and RKHS hypothesis classes in Sections 3 and 4, with time complexities of $\mathcal{O}(n)$ and $\mathcal{O}(T^3)$ per trial, and in terms of learner complexities they gain only leading multiplicative constants of $\mathcal{O}(1)$ and $\mathcal{O}(\log T)$.

2.2 Related Work

In this section we briefly describe other related work in the online setting that considers either *switching* or *multitask* models.

The first result for switching in the experts model was the WML algorithm [3] which was generalized in [9]. There is an extensive literature building on those papers, with some prominent results including [1, 13, 14, 15, 16, 17, 15, 18, 19, 20, 21, 22]. Relevant for our model are those papers [1, 14, 17, 15, 20, 21, 22] that address the problem of long-term memory ($m \ll k$), in particular [1, 14, 17].

Analogous to the problem of long-term memory in online learning is the problem of catastrophic forgetting in artificial neural network research [23, 24]. That is the problem of how a system can adapt to new information without forgetting the old. In online learning that is the problem of how an algorithm can both quickly adapt its prediction hypothesis and recall a previously successful prediction hypothesis when needed. In the experts model this problem was first addressed by [1], which gave an algorithm that stores each of its past state vectors, and then at each update mixes these vectors into the current state vector. In [14], an algorithm and bounds were given that extended the base comparision class of experts to include Bernoulli models. An improved algorithm with a Bayesian intepretation based on the idea of "circadian specialists" was given for this setting in [17]. Our construction of Algorithm 1 was based on this methodology.

The problem of linear regression with long term memory was posed as an open problem in [17, Sec. 5]. Algorithm 2 gives an algorithm for linear interpolation in a RKHS with a regret bound that reflects long-term memory. Switching linear prediction has been considered in [11, 25, 26, 12]. Only [12] addresses the issue of long-term memory. The methodology of [12] is a direct inspiration for Algorithm 2. We significantly extend the result of [12, Eq. (1)]. Their result was i) restricted to a mistake as opposed to a regret bound, ii) restricted to finite positive definite matrices and iii) in their mistake bound the term analogous to $\sum_{h\in m(\boldsymbol{h}^*)} \|h\|_K^2 X_K^2$ was increased by a multiplicative factor of $\tilde{\mathcal{O}}(|m(\boldsymbol{h}^*)|)$, a significantly weaker result.

Multitask learning has been considered extensively in the batch setting, with some prominent early results including [27, 28, 29]. In the online multitask *expert* setting [30, 31, 32, 17] considered a model which may be seen as a special case of ours where each task is associated only with a single hypothesis, i.e., no internal switching within a task. Also in the expert setting [33, 34] considered models where the prediction was made for all tasks simultaneously. In [34] the aim was to predict well relative to a set of possible predefined task interrelationships and in [33] the interrelationships were to be discovered algorithmically. The online multitask *linear* prediction setting was considered in [35, 36, 37]. The models of [36, 37] are similar to ours, but like previous work in the expert setting, these models are limited to one "hypothesis" per task. In the work of [35], the predictions were made for all tasks simultaneously through a joint loss function.

2.3 Preliminaries

For any positive integer m, we define $[m] := \{1, 2, \ldots, m\}$. For any predicate $[\text{PRED}] := 1$ if PRED is true and equals 0 otherwise, and for any $x \in \Re$, $[x]_+ := x[x > 0]$. We denote the inner product of vectors as both $\boldsymbol{x}, \boldsymbol{w} \in \Re^n$ as $\langle \boldsymbol{x}, \boldsymbol{w} \rangle = \boldsymbol{x} \cdot \boldsymbol{w} = \sum_{i=1}^n x_i w_i$, component-wise multiplication $\boldsymbol{x} \odot \boldsymbol{w} := (x_1 w_1, \ldots, x_n w_n)$ and the norm as $\|\boldsymbol{w}\| = \sqrt{\langle \boldsymbol{w}, \boldsymbol{w} \rangle}$. If $f : \Re \to \Re$ and $\boldsymbol{x} \in \Re^n$ then $f(\boldsymbol{x}) := (f(x_1), \ldots, f(x_n))$. The xth-coordinate vector is denoted $\boldsymbol{e}_X^x := ([x = z])_{z \in X}$; we commonly abbreviate this to $\boldsymbol{e}^x$. We denote the probability simplex as $\Delta_{\mathcal{H}} := \{h \in [0,1]^{\mathcal{H}}\} \cap \{h : \sum_{h\in\mathcal{H}} = 1\}$ and set $\Delta_n := \Delta_{[n]}$. We denote the binary entropy as $H(p) := p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p}$. If $\boldsymbol{v} \in \Delta_{\mathcal{H}}$ then $h \sim \boldsymbol{v}$ denotes that h is a random sample from the probability vector $\boldsymbol{v}$ over the set $\mathcal{H}$. For vectors $\boldsymbol{p} \in \Re^m$ and $\boldsymbol{q} \in \Re^n$ we define $[\boldsymbol{p}; \boldsymbol{q}] \in \Re^{m+n}$ to be the concatenation of $\boldsymbol{p}$ and $\boldsymbol{q}$, which we regard as a column vector. Hence $[\boldsymbol{p}; \boldsymbol{q}]^\top [\bar{\boldsymbol{p}}; \bar{\boldsymbol{q}}] = \boldsymbol{p}^\top \bar{\boldsymbol{p}} + \boldsymbol{q}^\top \bar{\boldsymbol{q}}$.

The notation $\boldsymbol{M}^+$ and $\sqrt{\boldsymbol{M}}$ denotes the pseudo-inverse and the unique positive square root, respectively, of a positive semi-definite matrix $\boldsymbol{M}$. The trace of a square matrix is denoted by $\operatorname{tr}(\boldsymbol{Y}) := \sum_{i=1}^n Y_{ii}$ for $\boldsymbol{Y} \in \Re^{n\times n}$. The $m \times m$ identity matrix is denoted $\boldsymbol{I}^m$. A function $K : \mathcal{X} \times \mathcal{X} \to \Re$ is a strictly positive definite (SPD) kernel iff for every finite $X \subseteq \mathcal{X}$ the matrix $K(x, x')_{x,x' \in X}$ is symmetric and strictly positive definite, for example, the Gaussian kernel. In addition, we define $\boldsymbol{S}^m$ to be the set of $m \times m$ symmetric matrices and let $\boldsymbol{S}_+^m$ and $\boldsymbol{S}_{++}^m$ be the subset of positive semidefinite and strictly positive definite matrices, respectively. We define the squared radius of $\boldsymbol{M} \in \boldsymbol{S}_+^m$ as $\mathcal{R}_{\boldsymbol{M}} := \max_{i\in[m]} M_{ii}^+$. The (undirected) graph Laplacian matrix is defined by $\boldsymbol{L} := \boldsymbol{D} - \boldsymbol{A}$ where $\boldsymbol{D}$ is the degree matrix and $\boldsymbol{A}$ is the adjacency matrix. The corresponding (strictly) positive definite *PDLaplacian* of an m-vertex connected graph is $\boldsymbol{L}^\circ := \boldsymbol{L} + \mathcal{R}_L^{-1} \left(\frac{\boldsymbol{1}}{m}\right) \left(\frac{\boldsymbol{1}}{m}\right)^\top$.

3 Finite Hypothesis Classes

In this section we present the algorithm and the regret bound for finite hypothesis classes, with proofs given in Appendix A. The design and analysis of the algorithm is inspired by [17], which considers a Bayesian setting where, on each trial, each hypothesis h gives an estimated probability $P(y^\tau = \hat{y}|h)$ of the outcome y^τ. The idea is for the learner to predict a probability $\hat{P}(y^\tau = \hat{y})$ and the loss incurred is the log loss, $\log(1/\hat{P}(y^\tau))$. Our algorithm, on the other hand, is framed in the well known "Allocation" setting [38] where the learner must play, on trial τ, a vector $\boldsymbol{v}^\tau \in \Delta_n$ and incurs a loss of $\boldsymbol{c}^\tau \cdot \boldsymbol{v}^\tau$ where all components of $\boldsymbol{c}^\tau$ are in $[0, 1]$.

To gain some intuition about the algorithm we observe the following. The algorithm maintains and updates the following vectors: a "global" probability vector $\boldsymbol{\pi}^\tau \in \Delta_{\mathcal{H}_{\text{fin}}}$ and the "local" task weight vectors $\boldsymbol{w}_t^1, \ldots, \boldsymbol{w}_t^s \in [0,1]^{\mathcal{H}_{\text{fin}}}$. Given an hypothesis $h \in \mathcal{H}_{\text{fin}}$, the scalar π_h^τ represents our "confidence", on trial τ, that hypothesis h is in $m(\boldsymbol{h}^*)$. For a given task i, hypothesis $h \in \mathcal{H}_{\text{fin}}$, and local time t, the scalar $w_{t,h}^i$ represents our confidence that $h = h_t^i$ if we knew that h was in $m(\boldsymbol{h}^*)$. Putting together, $\pi_h^\tau w_{\boldsymbol{\sigma}(\tau),h}^i$ represents our confidence, on trial τ, that $h = h_t^i$. The weights $\boldsymbol{\pi}^\tau$ and $\boldsymbol{w}_t^i$ (for tasks i) are designed in such a way that, not only do they store all the information required by the algorithm, but also on each trial τ we need only update $\boldsymbol{\pi}^\tau$ and $\boldsymbol{w}_t^{\ell^\tau}$. Thus the algorithm

Algorithm 1 Predicting $\mathcal{H}_{\text{fin}}$ in a switching multitask setting.

Parameters: $\mathcal{H}_{\text{fin}} \subseteq \{-1,1\}^{\mathcal{X}}$; $s, m, k, T \in \mathbb{N}$

Initialization: $n := |\mathcal{H}_{\text{fin}}|$; $\boldsymbol{\pi}^1 \leftarrow \frac{\mathbf{1}}{n}$; $\mu := \frac{1}{m}$; $\boldsymbol{w}_1^1 = \cdots = \boldsymbol{w}_1^s \leftarrow \mu\mathbf{1}$; $\theta := 1 - \frac{k}{T-s}$; $\phi := \frac{k}{(m-1)(T-s)}$ and

$$\eta := \sqrt{\left(m \log\left(\tfrac{n}{m}\right) + smH\left(\tfrac{1}{m}\right) + (T-s)H\left(\tfrac{k}{T-s}\right) + (m-1)(T-s)H\left(\tfrac{k}{(m-1)(T-s)}\right)\right)\tfrac{2}{T}}$$

For $\tau = 1, \ldots, T$

- Receive task $\ell^\tau \in [s]$.
- Receive $x^\tau \in \mathcal{X}$.
- Set $i \leftarrow \ell^\tau$; $t \leftarrow \boldsymbol{\sigma}(\tau)$.
- Predict

$$\boldsymbol{v}^\tau \leftarrow \frac{\boldsymbol{\pi}^\tau \odot \boldsymbol{w}_t^i}{\boldsymbol{\pi}^\tau \cdot \boldsymbol{w}_t^i}, \qquad \hat{h}^\tau \sim \boldsymbol{v}^\tau, \qquad \hat{y}^\tau \leftarrow \hat{h}^\tau(x^\tau)\,.$$

- Receive $y^\tau \in \{-1,1\}$.
- Update:

$i)\ \ \forall h \in \mathcal{H}_{\text{fin}},\ c_h^\tau = \mathcal{L}_{01}(h(x^\tau), y^\tau)$ $\qquad ii)\ \boldsymbol{\delta} \leftarrow \boldsymbol{w}_t^i \odot \exp(-\eta \boldsymbol{c}^\tau)$

$iii)\ \beta \leftarrow (\boldsymbol{\pi}^\tau \cdot \boldsymbol{w}_t^i)/(\boldsymbol{\pi}^\tau \cdot \boldsymbol{\delta})$ $\qquad iv)\ \boldsymbol{\epsilon} \leftarrow \mathbf{1} - \boldsymbol{w}_t^i + \beta\boldsymbol{\delta}$

$v)\ \ \boldsymbol{\pi}^{\tau+1} \leftarrow \boldsymbol{\pi}^\tau \odot \boldsymbol{\epsilon}$ $\qquad vi)\ \boldsymbol{w}_{t+1}^i \leftarrow (\phi(\mathbf{1} - \boldsymbol{w}_t^i) + \theta\beta\boldsymbol{\delta}) \odot \boldsymbol{\epsilon}^{-1}$

---

predicts in $\mathcal{O}(n)$ time per trial and requires $\mathcal{O}(sn)$ space. We bound the regret of the algorithm in the following theorem.

Theorem 1. *The expected regret of Algorithm 1 with parameters $\mathcal{H}_{\text{fin}} \subseteq \{-1,1\}^{\mathcal{X}}$; $s, m, k, T \in \mathbb{N}$ and*

$$C := m \log\left(\frac{n}{m}\right) + smH\left(\frac{1}{m}\right) + (T-s)H\left(\frac{k}{T-s}\right) + (m-1)(T-s)H\left(\frac{k}{(m-1)(T-s)}\right)$$

is bounded above by

$$\sum_{i=1}^{s}\sum_{t=1}^{T^i} \mathbb{E}[\mathcal{L}_{01}(y_t^i, \hat{y}_t^i)] - \mathcal{L}_{01}(y_t^i, h_t^i(x_t^i)) \leq \sqrt{2CT}$$

for any $\boldsymbol{h}^* \in \mathcal{H}_{\text{fin}}^T$ such that $k = k(\boldsymbol{h}^*)$, $m \geq |m(\boldsymbol{h}^*)|$, $m > 1$. Furthermore,

$$C \leq m \log\left(\frac{n}{m}\right) + s(\log(m) + 1) + k\left(\log(m-1) + 2\log\left(\frac{T-s}{k}\right) + 2\right)\,.$$

In further comparison to [17] we observe that we can obtain bounds for the log loss with our algorithm by defining $\hat{P}(y^\tau = \hat{y}) := \sum_h v_h^\tau P(y^\tau = \hat{y}|h)$ and redefining $c_h^\tau := -\frac{1}{\eta}\log(P(y^\tau = \hat{y}|h))$ in the update. The resultant theorem then matches the bound of [17, Thm. 4] for single-task learning with long-term memory ($s = 1$) and the bound of [17, Thm. 6] for multitask learning with no switching ($k = 0$).

4 RKHS Hypothesis Classes

Our algorithm and its analysis builds on the algorithm for online inductive matrix completion with side-information (IMCSI) from [39, Theorem 1, Algorithm 2 and Proposition 4]. IMCSI is an example of a matrix multiplicative weight algorithm [40, 6]. We give notation and background from [39] to provide insight.

The max-norm (or γ_2 norm [41]) of a matrix $\boldsymbol{U} \in \Re^{m \times n}$ is defined by

$$\|\boldsymbol{U}\|_{\max} := \min_{\boldsymbol{P}\boldsymbol{Q}^\top = \boldsymbol{U}} \left\{ \max_{1 \leq i \leq m} \|\boldsymbol{P}_i\| \times \max_{1 \leq j \leq n} \|\boldsymbol{Q}_j\| \right\}, \tag{6}$$

where the minimum is over all matrices $\boldsymbol{P} \in \Re^{m \times d}$ and $\boldsymbol{Q} \in \Re^{n \times d}$ and every integer d. We denote the class of $m \times d$ *row-normalized* matrices as $\mathcal{N}^{m,d} := \{\hat{\boldsymbol{P}} \subset \Re^{m \times d} : \left\|\hat{\boldsymbol{P}}_i\right\| = 1,\, i \in [m]\}$. The quasi-dimension of a matrix is defined as follows.

Definition 2 ([39, Equation (3)]). *The* quasi-dimension *of a matrix* $\boldsymbol{U} \in \Re^{m \times n}$ *with respect to* $\boldsymbol{M} \in \boldsymbol{S}^m_{++}$, $\boldsymbol{N} \in \boldsymbol{S}^n_{++}$ *at* γ *as*

$$\mathcal{D}^{\gamma}_{\boldsymbol{M},\boldsymbol{N}}(\boldsymbol{U}) := \min_{\hat{\boldsymbol{P}}\hat{\boldsymbol{Q}}^{\top}=\gamma\boldsymbol{U}} \operatorname{tr}\left(\hat{\boldsymbol{P}}^{\top}\boldsymbol{M}\hat{\boldsymbol{P}}\right)\mathcal{R}_{\boldsymbol{M}} + \operatorname{tr}\left(\hat{\boldsymbol{Q}}^{\top}\boldsymbol{N}\hat{\boldsymbol{Q}}\right)\mathcal{R}_{\boldsymbol{N}}\,, \tag{7}$$

where the infimum is over all row-normalized matrices $\hat{\boldsymbol{P}} \in \mathcal{N}^{m,d}$ *and* $\hat{\boldsymbol{Q}} \in \mathcal{N}^{n,d}$ *and every integer d. If the infimum does not exist then* $\mathcal{D}^{\gamma}_{\boldsymbol{M},\boldsymbol{N}}(\boldsymbol{U}) := +\infty$ *(The infimum exists iff* $\|\boldsymbol{U}\|_{max} \leq 1/\gamma$*).*

The algorithm IMCSI addresses the problem of the online prediction of a binary comparator matrix $\boldsymbol{U}$ with side information. The side information is supplied as a pair of kernels over the row indices and the column indices. In [39, Theorem 1] a regret bound $\tilde{\mathcal{O}}(\sqrt{(\widehat{\mathcal{D}}/\gamma^2)T})$ is given, where $1/\gamma^2 \geq \|\boldsymbol{U}\|^2_{\max}$ and $\widehat{\mathcal{D}} \geq \mathcal{D}^{\gamma}_{\boldsymbol{M},\boldsymbol{N}}(\boldsymbol{U})$ are parameters of the algorithm that serve as upper estimates on $\|\boldsymbol{U}\|^2_{\max}$ and $\mathcal{D}^{\gamma}_{\boldsymbol{M},\boldsymbol{N}}(\boldsymbol{U})$. The first estimate $1/\gamma^2$ is an upper bound on the squared max-norm (Eq. (6)) which like the trace-norm may be seen as a proxy for the rank of the matrix [42]. The second estimate $\widehat{\mathcal{D}}$ is an upper bound of the *quasi-dimension* (Eq. (7)) which measures the quality of the side-information. The quasi-dimension depends upon the "best" factorization $(1/\gamma)\hat{\boldsymbol{P}}\hat{\boldsymbol{Q}}^{\top} = \boldsymbol{U}$, which will be smaller when the row $\hat{\boldsymbol{P}}$ (column $\hat{\boldsymbol{Q}}$) factors are in congruence with the row (column) kernel. We bound the quasi-dimension in Theorem 47 in Appendix B as a key step to proving Theorem 3.

In the reduction of our problem to a matrix completion problem with side information, the row indices correspond to the domain of the learner-supplied kernel K and the column indices correspond to the temporal dimension. On each trial we receive an x^{τ} (a.k.a. x^i_t). Thus the column of the comparator matrix (now $\boldsymbol{H}$) corresponding to time τ will contain the entries $H^{\tau} = (h^{\tau}(x^{\upsilon}))_{\upsilon\in[T]}$. Although we are predicting functions that are changing over time, the underlying assumption is that the change is sporadic; otherwise it is infeasible to prove a non-vacuous bound. Thus we expect $H^i_t \approx H^i_{t+1}$ and as such our column side-information kernel should reflect this expectation. Topologically we would therefore expect a kernel to present as s separate time *paths*, where nearness in time is nearness on the path. In the following we introduce the *path-tree-kernel* (the essence of the construction was first introduced in [43]), which satisfies this expectation in the single-task case. We then adapt this construction to the multitask setting.

A *path-tree* kernel $P : [T] \times [T] \to \Re$, is formed via the Laplacian of a fully complete binary *tree* with $N := 2^{\lceil \log_2 T\rceil+1} - 1$ vertices. The *path* corresponds to the first T leaves of the tree, numbered sequentially from the leftmost to the rightmost leaf of the first T leaves. Denote this Laplacian as $\boldsymbol{L}$ where the path is identified with $[T]$ and the remaining vertices are identified with $[N] \setminus [T]$. Then using the definition $\boldsymbol{L}^{\circ} := \boldsymbol{L} + \left(\frac{\boldsymbol{1}}{N}\right)\left(\frac{\boldsymbol{1}}{N}\right)^{\top}\mathcal{R}^{-1}_{\boldsymbol{L}}$ we define $P(\tau,\upsilon) := (\boldsymbol{L}^{\circ})^{+}_{\tau\upsilon}$ where $\tau,\, \upsilon \in [T]$. We extend the path-tree kernel to a *multitask-path-tree* kernel by dividing the path into s contiguous segments, where segment i is a path of length T^i, and the task vector $\boldsymbol{\ell} \in [s]^T$ determines the mapping from global trial τ to task ℓ^{τ} and local trial $\boldsymbol{\sigma}(\tau)$. We define $\tilde{P}^{\boldsymbol{\ell},T^1,\ldots,T^s} : [T]\times[T] \to \Re$ as $\tilde{P}^{\boldsymbol{\ell},T^1,\ldots,T^s}(\tau,\upsilon) := P\left(\sum_{i=1}^{\ell^{\tau}-1} T^i + \boldsymbol{\sigma}(\tau), \sum_{i=1}^{\ell^{\upsilon}-1} T^i + \boldsymbol{\sigma}(\upsilon)\right)$. Observe we do not need to know the task vector $\boldsymbol{\ell}$ in advance; we only require upper bounds on the lengths of the tasks to be able to use this kernel. Finally, we note that it is perhaps surprising that we use a tree rather than a path directly. We discuss this issue following Lemma 49 in Appendix B.

Algorithm 2 requires $\mathcal{O}(t^3)$ time per trial t since we need to compute the eigendecomposition of three $\mathcal{O}(t) \times \mathcal{O}(t)$ matrices as well as sum $\mathcal{O}(t) \times \mathcal{O}(t)$ matrices up to t times. We bound the regret of the algorithm as follows.

Theorem 3. *The expected regret of Algorithm 2 with upper estimates,* $k \geq k(\boldsymbol{h}^*)$, $m \geq |m(\boldsymbol{h}^*)|$,

$$\hat{C} \geq C(\boldsymbol{h}^*) := \left(\sum_{h\in m(\boldsymbol{h}^*)} \|h\|^2_K X^2_K + 2(s+k-1)m\lceil \log_2 T\rceil^2 + 2m^2\right),$$

$\hat{X}^2_K \geq \max_{\tau\in[T]} K(x^{\tau},x^{\tau})$*, and learning rate* $\eta = \sqrt{\frac{\hat{C}\log(2T)}{2Tm}}$ *is bounded by*

$$\sum_{i=1}^{s}\sum_{t=1}^{T^i} \mathbb{E}[\mathcal{L}_{01}(y^i_t,\hat{y}^i_t)] - \mathcal{L}_{01}(y^i_t,h^i_t(x^i_t)) \leq 4\sqrt{2\hat{C}\,T\log(2T)} \tag{8}$$

with received instance sequence $\boldsymbol{x} \in \mathcal{X}^T$ *and for any* $\boldsymbol{h}^* \in \mathcal{H}^{(\boldsymbol{x})^T}_K$.

Algorithm 2 Predicting $\mathcal{H}_K^{(\boldsymbol{x})}$ in a switching multitask setting.

Parameters: Tasks $s \in \mathbb{N}$, task lengths $T^1, \ldots, T^s \in \mathbb{N}$, $T := \sum_{i=1}^{s} T^i$, learning rate: $\eta > 0$, complexity estimate: $\hat{C} > 0$, modes: $m \in [T]$, SPD Kernel $K : \mathcal{X} \times \mathcal{X} \to \Re$, $\tilde{P} := \tilde{P}^{\boldsymbol{\ell}, T^1, \ldots, T^s} : [T] \times [T] \to \Re$, with $\max_{\tau \in [T]} K(x^\tau, x^\tau) \leq \hat{X}_K^2$, and $\hat{X}_P^2 := 2\lceil \log_2 T \rceil$.

Initialization: $\mathbb{U} \leftarrow \emptyset\,, \mathcal{X}^1 \leftarrow \emptyset\,, \mathcal{T}^1 \leftarrow \emptyset\,.$

For $\tau = 1, \ldots, T$

- Receive task $\ell^\tau \in [s]\,.$
- Receive $x^\tau \in \mathcal{X}\,.$
- Set $i \leftarrow \ell^\tau$; $t \leftarrow \boldsymbol{\sigma}(\tau)$; $x_t^i \equiv x^\tau\,.$
- Define

$$\boldsymbol{K}^\tau := (K(x,z))_{x,z \in \mathcal{X}^\tau \cup \{x^\tau\}}\,; \quad \boldsymbol{P}^\tau := (\tilde{P}(\tau, \upsilon))_{\tau, \upsilon \in \mathcal{T}^\tau \cup \{\tau\}}\,,$$

$$\tilde{\boldsymbol{X}}^\tau(\upsilon) := \left[\frac{\sqrt{\boldsymbol{K}^\tau}\boldsymbol{e}^{x^\upsilon}}{\sqrt{2\hat{X}_K^2}}; \frac{\sqrt{\boldsymbol{P}^\tau}\boldsymbol{e}^{\upsilon}}{\sqrt{2\hat{X}_P^2}}\right]\left[\frac{\sqrt{\boldsymbol{K}^\tau}\boldsymbol{e}^{x^\upsilon}}{\sqrt{2\hat{X}_K^2}}; \frac{\sqrt{\boldsymbol{P}^\tau}\boldsymbol{e}^{\upsilon}}{\sqrt{2\hat{X}_P^2}}\right]^\top,$$

$$\tilde{\boldsymbol{W}}^\tau \leftarrow \exp\left(\log\left(\frac{\hat{C}}{2Tm}\right)\boldsymbol{I}^{|\mathcal{X}^\tau| + |\mathcal{T}^\tau| + 2} + \sum_{\upsilon \in \mathbb{U}} \eta y_\upsilon \tilde{\boldsymbol{X}}^\tau(\upsilon)\right).$$

- Predict

$$Y^\tau \sim \textsc{Uniform}(-\gamma, \gamma)\,; \quad \bar{y}^\tau \leftarrow \operatorname{tr}\left(\tilde{\boldsymbol{W}}^\tau \tilde{\boldsymbol{X}}^\tau\right) - 1\,; \quad \hat{y}_t^i := \hat{y}^\tau \leftarrow \operatorname{sign}(\bar{y}^\tau - Y^\tau)\,.$$

- Receive label $y_t^i := y^\tau \in \{-1, 1\}\,.$
- If $y^\tau \bar{y}^\tau \leq \frac{1}{\sqrt{m}}$ then

$$\mathbb{U} \leftarrow \mathbb{U} \cup \{t\}\,, \ \ \mathcal{X}^{\tau+1} \leftarrow \mathcal{X}^\tau \cup \{x^\tau\}, \text{ and } \mathcal{T}^{\tau+1} \leftarrow \mathcal{T}^\tau \cup \{\tau\}\,.$$

- Else $\mathcal{X}^{\tau+1} \leftarrow \mathcal{X}^\tau$ and $\mathcal{T}^{\tau+1} \leftarrow \mathcal{T}^\tau\,.$

Comparing roughly to the bound of the exponential-time algorithm (see (5)), we see that the $\log m$ term has been replaced by an m term and we have gained a multiplicative factor of $\log 2T$. From the perspective of long-term memory, we note that the potentially dominant learner complexity term $\sum_{h \in m(\boldsymbol{h}^*)} \|h\|_K^2 X_K^2$ has only increased by a slight $\log 2T$ term. To gain more insight into the problem we also have the following simple lower bound.

Proposition 4. *For any (randomized) algorithm and any $s, k, m, \Gamma \in \mathbb{N}$, with $k + s \geq m > 1$ and $\Gamma \geq m \log_2 m$, there exists a kernel K and a $T_0 \in \mathbb{N}$ such that for every $T \geq T_0$:*

$$\sum_{\tau=1}^{T} \mathbb{E}[\mathcal{L}_{01}(y^\tau, \hat{y}^\tau)] - \mathcal{L}_{01}(y^\tau, h^\tau(x^\tau)) \in \Omega\left(\sqrt{(\Gamma + s\log m + k \log m)\,T}\right),$$

for some multitask sequence $(x^1, y^1), \ldots, (x^T, y^T) \in (\mathcal{X} \times \{-1, 1\})^T$ and some $\boldsymbol{h}^* \in [\mathcal{H}_K^{(\boldsymbol{x})}]^T$ such that $m \geq |m(\boldsymbol{h}^*)|$, $k \geq k(\boldsymbol{h}^*)$, $\sum_{h \in m(\boldsymbol{h}^*)} \|h\|_K^2 X_K^2 \geq |m(\boldsymbol{h}^*)| \log_2 m$, where $X_K^2 = \max_{\tau \in [T]} K(x^\tau, x^\tau)$.

Comparing the above proposition to the bound of the exponential-time algorithm (see (5)), the most striking difference is the absence of the $\log T$ terms. We conjecture that these terms are not necessary for the 0-1 loss. A proof of Theorem 3 and a proof sketch of Proposition 4 are given in Appendix B.

5 Discussion

We have presented a novel multitask setting which generalizes single-task switching under the long-term memory setting. We gave algorithms for finite hypothesis classes and for RKHS hypothesis classes with per trial prediction times of $\mathcal{O}(n)$ and $\mathcal{O}(T^3)$. We proved upper bounds on the regret for both cases as well as a lower bound in the RKHS case. An open problem is to resolve the gap in the RKHS case. On the algorithmic side, both algorithms depend on a number of parameters. There is extensive research in online learning methods to design parameter-free methods. Can some of these methods be applied here (see e.g., [44])? For a non-parametric hypothesis class, intuitively it seems we must expect some time complexity dependence on T. However can we perhaps utilize decay methods such as [45, 46] or sketching methods [47] that have had success in simpler models to improve running times? More broadly, for what other infinite hypothesis classes can we give efficient regret-bounded algorithms in this switching multitask setting with long-term memory?

6 Acknowledgements

This research was supported by the U.S. Army Research Laboratory and the U.K. Ministry of Defence under Agreement Number W911NF-16-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon. This research was further supported by the Engineering and Physical Sciences Research Council grant number EP/L015242/1.

Broader Impact

In general this work does not present any specific foreseeable societal consequence in the authors' joint opinion.

This is foundational research in *regret-bounded online learning*. As such it is not targeted towards any particular application area. Although this research may have societal impact for good or for ill in the future, we cannot foresee the shape and the extent.

Funding Transparency Statement

Funding

The authors were supported by the U.S. Army Research Laboratory and the U.K. Ministry of Defence under agreement number W911NF-16-3-0001 and by the Engineering and Physical Sciences Research Council grant number EP/L015242/1.

Competing Interests

The authors assert no competing interests.

References

[1] O. Bousquet and M.K. Warmuth. Tracking a small set of experts by mixing past posteriors. *Journal of Machine Learning Research*, 3:363–396, 2003.

[2] Volodimir G. Vovk. Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, COLT '90, pages 371–386, 1990.

[3] Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Inf. Comput.*, 108(2):212–261, February 1994.

[4] Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006.

[5] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

[6] Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(6):121–164, 2012.

[7] Nicolò Cesa-Bianchi, Yoav Freund, David Haussler, David P. Helmbold, Robert E. Schapire, and Manfred K. Warmuth. How to use expert advice. *J. ACM*, 44(3):427–485, May 1997.

[8] S. Shalev-Shwartz. Online Learning and Online Convex Optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2011.

[9] M. Herbster and M.K. Warmuth. Tracking the best expert. *Machine Learning*, 32(2):151–178, 1998.

[10] Y. Freund. Private communication, 2000. Also posted on http://www.learning-theory.org.

[11] M. Herbster and M.K. Warmuth. Tracking the best linear predictor. *Journal of Machine Learning Research*, 1:281–309, 2001.

[12] M. Herbster, S. Pasteris, and S. Pontil. Predicting a switching sequence of graph labelings. *Journal of Machine Learning Research*, 16:2003–2022, 2015.

[13] A. György, T. Linder, and G. Lugosi. Tracking the best of many experts. In *Proceedings 18th Annual Conference on Learning Theory*, pages 204–216, 2005.

[14] Wouter M. Koolen and Tim van Erven. Freezing and sleeping: Tracking experts that learn by evolving past posteriors, 2010.

[15] N. Cesa-Bianchi, P. Gaillard, G. Lugosi, and G. Stoltz. Mirror descent meets fixed share (and feels no regret). In *Advances in Neural Information Processing Systems 24*, pages 989–997, 2012.

[16] A. György, T. Linder, and G. Lugosi. Efficient tracking of large classes of experts. *IEEE Transactions on Information Theory*, 58(11):6709–6725, Nov 2012.

[17] Wouter M. Koolen, Dmitry Adamskiy, and Manfred K. Warmuth. Putting bayes to sleep. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS 12, pages 135–143, Red Hook, NY, USA, 2012. Curran Associates Inc.

[18] D. Adamskiy, W. M. Koolen, A. Chernov, and V. Vovk. A closer look at adaptive regret. In *Proceedings of the 23rd International Conference on Algorithmic Learning Theory*, ALT'12, pages 290–304, 2012.

[19] A. Daniely, A. Gonen, and S. Shalev-Shwartz. Strongly adaptive online learning. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 1405–1411, 2015.

[20] Jaouad Mourtada and Odalric-Ambrym Maillard. Efficient tracking of a growing number of experts. In *Proceedings of the 28th International Conference on Algorithmic Learning Theory (ALT)*, volume 76 of *Proceedings of Machine Learning Research*, pages 517–539, 2017.

[21] Maria-Florina Balcan, Travis Dick, and Dravyansh Sharma. Online optimization of piecewise lipschitz functions in changing environments. *CoRR*, abs/1907.09137, 2019.

[22] Kai Zheng, Haipeng Luo, Ilias Diakonikolas, and Liwei Wang. Equipping experts/bandits with long-term memory. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d. Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 5929–5939. Curran Associates, Inc., 2019.

[23] Michael Mccloskey and Neil J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *The Psychology of Learning and Motivation*, 24:104–169, 1989.

[24] Robert M. French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128 – 135, 1999.

[25] J. Kivinen, A. J. Smola, and R. C. Williamson. Online learning with kernels. *IEEE Transactions on Signal Processing*, 52:2165–2176, 2004.

[26] N. Cesa-Bianchi and C. Gentile. Tracking the best hyperplane with a simple budget perceptron. In *Proceedings of the 18th Conference on Learning Theory*, pages 483–498, 2006.

[27] Jonathan Baxter. Learning internal representations. In *Proceedings of the Eighth Annual Conference on Computational Learning Theory*, COLT ?95, page 311?320, New York, NY, USA, 1995. Association for Computing Machinery.

[28] Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

[29] Theodoros Evgeniou and Massimiliano Pontil. Regularized multi–task learning. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 109–117, New York, NY, USA, 2004. Association for Computing Machinery.

[30] J. Abernethy, P. Bartlett, and A. Rakhlin. Multitask learning with expert advice. In *Proceedings 20th Annual Conference on Learning Theory*, pages 484–498, 2007.

[31] Alekh Agarwal, Alexander Rakhlin, and Peter Bartlett. Matrix regularization techniques for online multitask learning. Technical Report UCB/EECS-2008-138, EECS Department, University of California, Berkeley, Oct 2008.

[32] S. Avishek, R. Piyush, H. Daumé III, and S. Venkatasubramanian. Online learning of multiple tasks and their relationships. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pages 643–651, 2011.

[33] Alexander Rakhlin, Jacob D. Abernethy, and Peter L. Bartlett. Online discovery of similarity mappings. In Zoubin Ghahramani, editor, *Machine Learning, Proceedings of the Twenty-Fourth International Conference (ICML 2007), Corvallis, Oregon, USA, June 20-24, 2007*, volume 227 of *ACM International Conference Proceeding Series*, pages 767–774. ACM, 2007.

[34] Gábor Lugosi, Omiros Papaspiliopoulos, and Gilles Stoltz. Online multi-task learning with hard constraints. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009.

[35] O. Dekel, P.M. Long, and Y. Singer. Online learning of multiple tasks with a shared loss. *Journal of Machine Learning Research*, 8(10):2233–2264, 2007.

[36] G. Cavallanti, N. Cesa-Bianchi, and C. Gentile. Linear algorithms for online multitask classification. *Journal of Machine Learning Research*, 1:2901–2934, 2010.

[37] Christoph Hirnschall, Adish Singla, Sebastian Tschiatschek, and Andreas Krause. Coordinated online learning with applications to learning user preferences, 2017.

[38] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119?139, August 1997.

[39] Mark Herbster, Stephen Pasteris, and Lisa Tse. Online matrix completion with side information. In *Advances in Neural Information Processing Systems 33*. 2020.

[40] K. Tsuda, G. Rätsch, and M.K. Warmuth. Matrix exponentiated gradient updates for on-line learning and bregman projection. *Journal of Machine Learning Research*, 6:995–1018, 2005.

[41] N. Linial, S. Mendelson, G. Schechtman, and A. Shraibman. Complexity measures of sign matrices. *Combinatorica*, 27(4):439–463, 2007.

[42] Jason D Lee, Ben Recht, Nathan Srebro, Joel Tropp, and Russ R Salakhutdinov. Practical large-scale optimization for max-norm regularization. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1297–1305. Curran Associates, Inc., 2010.

[43] M. Herbster, G. Lever, and M. Pontil. Online prediction on large diameter graphs. In *Advances in Neural Information Processing Systems 21*, pages 649–656, 2008.

[44] Francesco Orabona and Dávid Pál. Coin betting and parameter-free online learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS16, pages 577–585, Red Hook, NY, USA, 2016. Curran Associates Inc.

[45] Ofer Dekel, Shai Shalev-shwartz, and Yoram Singer. The forgetron: A kernel-based perceptron on a fixed budget. In Y. Weiss, B. Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 259–266. MIT Press, 2006.

[46] Giovanni Cavallanti, Nicolò Cesa-Bianchi, and Claudio Gentile. Tracking the best hyperplane with a simple budget perceptron. *Mach. Learn.*, 69(2-3):143–167, 2007.

[47] Yair Carmon, John C. Duchi, Aaron Sidford, and Kevin Tian. A rank-1 sketch for matrix multiplicative weights. In Alina Beygelzimer and Daniel Hsu, editors, *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, volume 99 of *Proceedings of Machine Learning Research*, pages 589–623. PMLR, 2019.

[48] Yoav Freund, Robert E. Schapire, Yoram Singer, and Manfred K. Warmuth. Using and combining predictors that specialize. In *Proceedings of the Twenty-Ninth Annual ACM Symposium on Theory of Computing*, STOC 97, pages 334–343, New York, NY, USA, 1997. Association for Computing Machinery.

[49] M. Herbster and M. Pontil. Prediction on a graph with a perceptron. In *Advances in Neural Information Processing Systems 19*, pages 577–584, 2006.

[50] Douglas Klein and Milan Randic. Resistance distance. *Journal of Mathematical Chemistry*, 12:81–95, 12 1993.

[51] M. Herbster, M. Pontil, and S. Rojas-Galeano. Fast prediction on a tree. In *Advances in Neural Information Processing Systems*, pages 657–664, 2009.

[52] A.B. Novikoff. On convergence proofs on perceptrons. In *Proceedings of the Symposium on the Mathematical Theory of Automata*, pages 615–622, 1962.

[53] J. Forster, N. Schmitt, and H.U. Simon. Estimating the optimal margins of embeddings in euclidean half spaces. In *Proceedings Computational Learning Theory*, pages 402–415, 2001.

[54] N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285–318, April 1988.

[55] Shai Ben-David, Dávid Pál, and Shai Shalev-Shwartz. Agnostic online learning. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009.

[56] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

[57] Martin Zinkevich. Online Convex Programming and Generalized Infinitesimal Gradient Ascent. In *Proceedings, Twentieth International Conference on Machine Learning*, volume 2, pages 928–935, 2003.

[58] N. Cesa-Bianchi, P. M. Long, and M. K. Warmuth. Worst-case quadratic loss bounds for on-line prediction of linear functions by gradient descent. *IEEE Transactions on Neural Networks*, 7(3):604–619, 1996. Earlier version in 6th COLT, 1993.

A Analysis of Algorithm 1

This section is divided into two subsections: in Section A.1 we outline the analysis of Algorithm 1 as a sequence of theorems. These theorems are then proved in Section A.2. We start with the following definitions:

Given a finite set X and functions $f, g : X \to [0, 1]$ with $\sum_{x\in X} f(x) = \sum_{x\in X} g(x) = 1$ we define:

$$\mathrm{KL}(f, g) := \sum_{x\in X} f(x) \ln\left(\frac{f(x)}{g(x)}\right) .$$

Given a finite set X and functions $f, g : X \to [0, 1]$ we define:

$$\langle f, g\rangle := \sum_{x\in X} f(x)g(x) .$$

A.1 Overview of Analysis

Our general problem is to play the "(Multitask) Allocation game[2]" defined as follows. We have n "experts" and s "tasks". The game proceeds over $\tau = 1, \ldots, T$ trials. On trial τ the following happens:

1. Nature chooses a task $\ell^\tau \in [s]$.
2. Learner chooses a prediction vector $\boldsymbol{v}^\tau \in \Delta_n$.
3. Nature chooses a loss vector $\boldsymbol{c}^\tau \in [0, 1]^n$.
4. Learner suffers loss $c^\tau_* := \boldsymbol{v}^\tau \cdot \boldsymbol{c}^\tau$.

The goal of learner is to minimise the cumulative loss $\sum_{\tau\in[T]} c^\tau_*$.

Note that Algorithm 1 is applicable to the Allocation game. We will compare the performance of the algorithm against an arbitrary sequence $z_1, z_2, \ldots, z_T \in [n]$. We define:

- $\mathbb{M} := \{z_\tau \ : \ \tau \in [T]\}$
- $m := |\mathbb{M}|$
- $k := |\{\tau \in [T] : [U_\tau \neq \emptyset] \wedge [z_\tau \neq z_{(\min U_\tau)}]\}|$ where $U_\tau := \{\tau' \in [T] : [\tau' > \tau] \wedge [\ell^{\tau'} = \ell^\tau]\}$ for all $\tau \in [T]$.

In the Allocation game we refer to the elements of $[n]$ as "experts". At the end of the analysis we will reduce our main problem to the Allocation game which involves enumerating the set of hypotheses $\mathcal{H}_{\text{fin}}$ as $[n]$, i.e., each hypothesis $h \in \mathcal{H}_{\text{fin}}$ corresponds to an expert in $[n]$. The sequence $z_1, z_2, \ldots, z_T$ then corresponds to the hypothesis sequence $\boldsymbol{h}^*$. Hence we have that $k = k(\boldsymbol{h}^*)$ and $m = |m(\boldsymbol{h}^*)|$.

We now begin the analysis of Algorithm 1. We will analyse the algorithm via a reduction of the Allocation game to another game called the "Specialist allocation game". We will now define this game and introduce an algorithm called "Specialist hedge" which is applicable to it.

A.1.1 The Specialist Allocation Game

The Specialist allocation game is the following game between Nature and Learner. We have a finite set $\mathcal{E}$ of "specialists". On trial τ:

1. Nature chooses a non-empty set $\mathcal{W}^\tau \subseteq \mathcal{E}$ and reveals it to Learner.
2. Learner chooses a function $\tilde{v}^\tau : \mathcal{W}^\tau \to [0, 1]$ with $\sum_{\xi\in\mathcal{W}^\tau} \tilde{v}^\tau(\xi) = 1$ and reveals it to Nature.
3. Nature chooses a function $\tilde{c}^\tau : \mathcal{W}^\tau \to [0, 1]$ and reveals it to Learner.
4. Learner incurs loss $c^\tau_* := \langle \tilde{v}^\tau, \tilde{c}^\tau\rangle$

[2] We mean "Game" in an informal sense.

This game is a generalisation of the Single-task allocation game where on every trial only a subset $\mathcal{W}^\tau$ of experts (now called "specialists") are awake. Note that we have changed the notation from the Allocation game in that what was represented by vectors in the definition of the Allocation game have now become, equivalently, represented by functions. Instead of selecting, on trial τ, a probability distribution (represented by $\boldsymbol{v}^\tau$) over the set of experts (as in the Allocation game) Learner now plays a probability distribution $\tilde{v}^\tau$ over only the set $\mathcal{W}^\tau$.

A.1.2 The Specialist Hedge Algorithm

Algorithm 3 Specialist Hedge

Parameters:
$\eta \in \mathbb{R}^+$ and $\rho^* : \mathcal{E} \to [0,1]$ with $\sum_{\xi \in \mathcal{E}} \rho^*(\xi) = 1$.

Initialization:
- $\rho^1 \leftarrow \rho^*$

Prediction on trial τ:
- Receive $\mathcal{W}^\tau \subseteq \mathcal{E}$
- $Y^\tau \leftarrow \sum_{\xi \in \mathcal{W}^\tau} \rho^\tau(\xi)$
- $\forall \xi \in \mathcal{W}^\tau,\ \tilde{v}^\tau(\xi) \leftarrow \rho^\tau(\xi)/Y^\tau$

Update on trial τ:
- Receive $\tilde{c}^\tau$
- $Z^\tau \leftarrow \sum_{\xi \in \mathcal{W}^\tau} \rho^\tau(\xi) \exp(-\eta \tilde{c}^\tau(\xi))$
- $\forall \xi \in \mathcal{W}^\tau,\ \rho^{\tau+1}(\xi) \leftarrow \frac{Y^\tau}{Z^\tau} \rho^\tau(\xi) \exp(-\eta \tilde{c}^\tau(\xi))$
- $\forall \xi \in \mathcal{E} \setminus \mathcal{W}^\tau,\ \rho^{\tau+1}(\xi) \leftarrow \rho^\tau(\xi)$

We now introduce Specialist hedge as an algorithm for Learner in the Specialist allocation game. Specialist hedge generalizes the classic Hedge algorithm for the Single-task allocation game so that it now may be applied to the Specialist allocation game. The algorithm is seeded with a probability distribution ρ^* over the specialists: i.e. $\rho^* : \mathcal{E} \to [0,1]$ with $\sum_{\xi \in \mathcal{E}} \rho^*(\xi) = 1$. For a specialist $\xi \in \mathcal{E}$ the quantity $\rho^*(\xi)$ represents our a-priori confidence that the specialist ξ has a low cumulative loss over the trials in which it is awake. Specialist hedge maintains, over trials τ, a probability distribution ρ^τ over specialists. The quantity $\rho^\tau(\xi)$ represents our confidence, on trial τ, that the specialist ξ has a low cumulative loss over the trials in which it is awake.

A.1.3 The Specialist Hedge Bound

We will now give an inequality for Specialist hedge. The inequality is relative to any arbitrary probability distribution u over the set of all specialists.

Theorem 5. *Let $u : \mathcal{E} \to [0,1]$ be any function such that $\sum_\xi u(\xi) = 1$. On any trial τ we define*

$$u(\mathcal{W}^\tau) := \sum_{\xi \in \mathcal{W}^\tau} u(\xi)$$

and for all $\xi \in \mathcal{W}^\tau$ we define

$$\bar{u}^\tau(\xi) := u(\xi)/u(\mathcal{W}^\tau).$$

Then we have:

$$\sum_{\tau \in [T]} u(\mathcal{W}^\tau) \langle \tilde{v}^\tau - \bar{u}^\tau, \tilde{c}^\tau \rangle \leq \frac{1}{\eta} \mathrm{KL}(u, \rho^*) + \frac{\eta}{2} \sum_{\tau \in [T]} u(\mathcal{W}^\tau)$$

A.1.4 Definition of a Markov Circadian

We now generalize the proof methodology found in [17] by defining specialists via what we call "(Generalized) Markov circadians". (Generalized) Markov circadians are a generalisation of the 0/1 (wake/sleep) Markov circadians found in [17]. We will drop the word "generalized" in the following.

A "Markov circadian" is defined as a Markov chain over trials along with, on every trial, a subset of states which are denoted "awake". Specifically a Markov circadian is defined by the following:

- A finite set Φ of states,
- A map $\lambda' : \Phi \to [0,1]$ with $\sum_{\varphi\in\Phi} \lambda'(\varphi) = 1$,
- For every trial $\tau \in [T-1]$ a map $\lambda^\tau : \Phi^2 \to [0,1]$ with $\sum_{\varphi'\in\Phi} \lambda^\tau(\varphi,\varphi') = 1$ for all $\varphi \in \Phi$.
- On every trial τ a set $\widetilde{\mathcal{W}}^\tau \subseteq \Phi$.

We define $\mathcal{A} := \Phi^T$. We call the elements of $\mathcal{A}$ "Circadian Instances". Each Circadian Instance $\alpha \in \mathcal{A}$ has an associated weight $q(\alpha)$ defined as:

$$q(\alpha) = \lambda'(\alpha(1)) \prod_{t\in[T-1]} \lambda^\tau(\alpha(\tau), \alpha(\tau+1))$$

A.1.5 Reduction to the Specialist Allocation Game and Specialist Hedge

We now show how a Markov circadian converts the Allocation game to the Specialist allocation game. Specifically we choose each specialist to be a circadian instance paired with an expert. A specialist is awake whenever its associated circadian instance is in an awake state and the losses of the specialist equal the losses of its associated expert.

Formally, given a Markov circadian, we define the set of specialists as:

$$\mathcal{E} := [n] \times \mathcal{A} \,.$$

We then define, for all $\tau \in [T]$:

$$\mathcal{W}^\tau := \left\{(i,\alpha) \in \mathcal{E} \mid \alpha(\tau) \in \widetilde{\mathcal{W}}^\tau\right\}$$

$$\tilde{c}^\tau(i,\alpha) := c_i^\tau \quad \forall (i,\alpha) \in \mathcal{W}^\tau$$

We will implicitly run Specialist hedge on this set of specialists. We seed the specialist hedge algorithm with ρ^* defined as:

$$\rho^*(i,\alpha) := \frac{1}{n} q(\alpha) \,.$$

The choice, $\boldsymbol{v}^\tau$, of Learner in the Allocation game is then defined from the choice, $\tilde{v}^\tau$ of Specialist Hedge as follows:

$$v_i^\tau := \sum_{\alpha\in\mathcal{A}:(i,\alpha)\in\mathcal{W}^\tau} \tilde{v}^\tau(i,\alpha)$$

for all experts $i \in [n]$.

The following theorem asserts the equivalence of $\boldsymbol{v}^\tau$ and $\tilde{v}^\tau$:

Theorem 6. *For all $\tau \in [T]$, the loss $\boldsymbol{v}^\tau \cdot \boldsymbol{c}^\tau$ of Learner in the allocation game (when $\boldsymbol{v}^\tau$ is defined as above) is equal to the loss $\langle \tilde{v}^\tau, \tilde{c}^\tau \rangle$ of Specialist hedge when using the above specialists and seeding distribution.*

A.1.6 Implicit Specialist Hedge Updates and Choices

For all trials $\tau \in [T]$, experts $i \in [n]$ and states $\varphi \in \Phi$ we define:

$$\gamma_i^\tau(\varphi) := \sum_{\alpha\in\mathcal{A}:\alpha(\tau)=\varphi} \rho^\tau(i,\alpha)$$

where ρ^τ is defined as in the Specialist hedge algorithm run with specialists and seed defined in Section A.1.5. We have the following theorem:

Theorem 7. *For all experts $i \in [n]$ and states $\varphi \in \Phi$ we have:*

$$\gamma_i^1(\varphi) = \frac{1}{n}\lambda'(\varphi)$$

and, in addition, for all trials $\tau \in [T-1]$ we have:

$$\gamma_i^{\tau+1}(\varphi) = \frac{Y^\tau}{Z^\tau} \sum_{\varphi' \in \widetilde{\mathcal{W}}^\tau} \lambda^\tau(\varphi', \varphi)\gamma_i^\tau(\varphi') \exp(-\eta c_i^\tau) + \sum_{\varphi' \in \Phi \setminus \widetilde{\mathcal{W}}^\tau} \lambda^\tau(\varphi', \varphi)\gamma_i^\tau(\varphi')$$

where:

$$Y^\tau = \sum_{i \in [n]} \sum_{\varphi \in \widetilde{\mathcal{W}}^\tau} \gamma_i^\tau(\varphi)$$

and:

$$Z^\tau = \sum_{i \in [n]} \exp(-\eta c_i^\tau) \sum_{\varphi \in \widetilde{\mathcal{W}}^\tau} \gamma_i^\tau(\varphi)$$

Also, when $\boldsymbol{v}^\tau$ is defined as in Section A.1.5 we have:

$$v_i^\tau = \frac{1}{Y^\tau} \sum_{\varphi \in \widetilde{\mathcal{W}}^\tau} \gamma_i^\tau(\varphi)$$

for all experts $i \in [n]$.

A.1.7 A bound for Markov circadians

By Theorem 6 we can bound Learner's loss when playing the allocation game with a Markov circadian (i.e. when using the strategy of A.1.5) by the loss of Specialist hedge with a Markov circadian. Hence, in this section, we will bound the latter.

In this section we assume (in retrospect) that we have, for all $i \in \mathbb{M}$, a Circadian Instance $\hat{\alpha}^i$ with:

$$\{\tau \in [T] \mid \hat{\alpha}^i(\tau) \in \widetilde{\mathcal{W}}^\tau\} = \{\tau \in [T] \mid z_\tau = i\}$$

We will now bound the loss of Specialist hedge via the following intermediate theorem:

Theorem 8. *If $u : \mathcal{E} \to [0,1]$ is defined as:*

$$u(i, \alpha) = \frac{1}{m} \mathbb{I}\left(i \in \mathbb{M} \, \wedge \, \alpha = \hat{\alpha}^i\right)$$

then we have:

$$m \, \mathrm{KL}(u, \rho^*) = m \ln\left(\frac{n}{m}\right) - \sum_{i \in \mathbb{M}} \ln(\lambda'(\hat{\alpha}^i(1))) - \sum_{i \in \mathbb{M}} \sum_{\tau \in [T-1]} \ln(\lambda^\tau(\hat{\alpha}^i(\tau), \hat{\alpha}^i(\tau+1)))$$

By combining Theorem 8 and Theorem 5 (with choice of u as in Theorem 8) we obtain the following bound on the Specialist hedge algorithm with a Markov circadian.

Theorem 9. *The regret of Specialist hedge with a Markov circadian is bounded by:*

$$\sum_{\tau \in [T]} (c_*^\tau - c_{z_\tau}^\tau) \leq \frac{\eta}{2} T + \frac{1}{\eta} C$$

where

$$C = m \ln\left(\frac{n}{m}\right) - \sum_{i \in \mathbb{M}} \ln(\lambda'(\hat{\alpha}^i(1))) - \sum_{i \in \mathbb{M}} \sum_{\tau \in [T-1]} \ln(\lambda^\tau(\hat{\alpha}^i(\tau), \hat{\alpha}^i(\tau+1)))$$

A.1.8 The Multitask Markov Circadian

Algorithm 1 implicitly follows the reduction given in Section A.1.5, using the following Markov Circadian.

Our state set is defined as $\Phi := \{0, 1, \psi\}^s$ where ψ is a symbol that denotes that a particular task will not be seen again. On each trial τ we have $\widetilde{\mathcal{W}}^\tau := \{\varphi \in \Phi \; : \; \varphi_{\ell^\tau} = 1\}$. Since Φ is now a set of vectors we now use vector notation for its members. Our transition matrices are defined as follows.

We define $\bar{\Omega} := \{\tau \in [T] \; : \; \forall \tau' > \tau, \; \ell^{\tau'} \neq \ell^\tau\}$. The (modified) Multitask Markov circadian is defined as follows:

For all states $\boldsymbol{\varphi} \in \Phi$ we define:

$$\lambda'(\boldsymbol{\varphi}) := \prod_{j \in [s]} \left(\mu \mathbb{I}\left(\varphi_j = 1\right) + (1-\mu)\, \mathbb{I}\left(\varphi_j = 0\right)\right)$$

and for all trials $\tau \in [T-1]$ and pairs of states $\boldsymbol{\varphi}, \boldsymbol{\varphi}' \in \Phi$ we have:

- If $\tau \notin \bar{\Omega}$ then
  - If $\varphi'_{\ell^\tau} = 0$ then:
    - If $\varphi_{\ell^\tau} = 0$ then $\lambda^\tau(\boldsymbol{\varphi}', \boldsymbol{\varphi}) := (1-\phi)\, \mathbb{I}\left(\forall j \in [s] \setminus \{\ell^\tau\},\ \varphi_j = \varphi'_j\right)$
    - If $\varphi_{\ell^\tau} = 1$ then $\lambda^\tau(\boldsymbol{\varphi}', \boldsymbol{\varphi}) := \phi\, \mathbb{I}\left(\forall j \in [s] \setminus \{\ell^\tau\},\ \varphi_j = \varphi'_j\right)$
    - If $\varphi_{\ell^\tau} = \psi$ then $\lambda^\tau(\boldsymbol{\varphi}', \boldsymbol{\varphi}) := 0$
  - If $\varphi'_{\ell^\tau} = 1$ then:
    - If $\varphi_{\ell^\tau} = 0$ then $\lambda^\tau(\boldsymbol{\varphi}', \boldsymbol{\varphi}) := (1-\theta)\, \mathbb{I}\left(\forall j \in [s] \setminus \{\ell^\tau\},\ \varphi_j = \varphi'_j\right)$
    - If $\varphi_{\ell^\tau} = 1$ then $\lambda^\tau(\boldsymbol{\varphi}', \boldsymbol{\varphi}) := \theta\, \mathbb{I}\left(\forall j \in [s] \setminus \{\ell^\tau\},\ \varphi_j = \varphi'_j\right)$
    - If $\varphi_{\ell^\tau} = \psi$ then $\lambda^\tau(\boldsymbol{\varphi}', \boldsymbol{\varphi}) := 0$
- If $\tau \in \bar{\Omega}$ then $\lambda^\tau(\boldsymbol{\varphi}', \boldsymbol{\varphi}) := \mathbb{I}\left(\varphi_{\ell^\tau} = \psi\right) \mathbb{I}\left(\forall j \in [s] \setminus \{\ell^\tau\},\ \varphi_j = \varphi'_j\right)$

where μ, ϕ and θ are as in Algorithm 1.

A.1.9 Derivation of Algorithm 1

Here we will show that the choice $\boldsymbol{v}^\tau$ of Algorithm 1 on trial τ is equal to that in Section A.1.5 when using the Multitask Markov circadian. For all $j \in [s]$ we first define:

$$\Omega_j := \{\tau \in [T] \mid \forall \tau' \geq \tau,\ \ell^{\tau'} \neq j\}$$

Also, for all $\tau \in [T]$, all $j \in [s]$ and all $b \in \{0, 1\}$ we define $\varpi_j^\tau(b)$ by:

- If $\tau \notin \Omega_j$ then $\varpi_j^\tau(b) = b$
- If $\tau \in \Omega_j$ then $\varpi_j^\tau(b) = \psi$

In addition, for all $\boldsymbol{\varphi} \in \Phi$ and $b \in \{0, 1\}$ we define $\boldsymbol{\varphi}^{\tau|b}$ by:

- $\varphi_j^{\tau|b} := \varphi_j \quad \forall j \in [s] \setminus \{\ell^\tau\}$
- $\varphi_{\ell^\tau}^{\tau|b} := b$

In this section we take the vectors $\boldsymbol{\pi}$ and $\boldsymbol{w}$ as in Algorithm 1 (with subscript and superscript removed). For every task $j \in [s]$ and every trial $\tau \in [T]$ we define the vector $\bar{\boldsymbol{w}}^{\tau,j}$ as follows:

- $\bar{\boldsymbol{w}}^{1,j} = \mu$
- $\bar{\boldsymbol{w}}^{\tau+1,\ell^\tau} := \boldsymbol{w}_{\boldsymbol{\sigma}(\tau)+1}^{\ell^\tau}$
- If $j \neq \ell^\tau$ then $\bar{\boldsymbol{w}}^{\tau+1,j} := \bar{\boldsymbol{w}}^{\tau,j}$

We now analyse the values $\gamma_i^\tau(\boldsymbol{\varphi})$ for all $\tau \in [T]$, $i \in [n]$, and $\boldsymbol{\varphi} \in \Phi$. Theorem 7 leads to the following theorem:

Theorem 10. *For all trials $\tau \in [T]$ and experts $i \in [n]$*

- *If $\tau \in \bar{\Omega}$ we have, for all $\boldsymbol{\varphi} \in \Phi$:*

$$\gamma_i^{\tau+1}(\boldsymbol{\varphi}) = \mathbb{I}\left(\varphi_{\ell^\tau} = \psi\right) \left(\gamma_i^\tau(\boldsymbol{\varphi}^{\tau|0}) + \gamma_i^\tau(\boldsymbol{\varphi}^{\tau|1}) \exp(-\eta c_i^\tau) \frac{Y^\tau}{Z^\tau}\right)$$

- *If $\tau \notin \bar{\Omega}$ we have the following:*

– *For all $\boldsymbol{\varphi} \in \Phi$ with $\varphi_{\ell^\tau} = 0$ we have:*

$$\gamma_i^{\tau+1}(\boldsymbol{\varphi}) = (1-\phi)\gamma_i^\tau(\boldsymbol{\varphi}^{\tau|0}) + (1-\theta)\gamma_i^\tau(\boldsymbol{\varphi}^{\tau|1})\exp(-\eta c_i^\tau)\frac{Y^\tau}{Z^\tau}$$

– *For all $\boldsymbol{\varphi} \in \Phi$ with $\varphi_{\ell^\tau} = 1$ we have:*

$$\gamma_i^{\tau+1}(\boldsymbol{\varphi}) = \phi\gamma_i^\tau(\boldsymbol{\varphi}^{\tau|0}) + \theta\gamma_i^\tau(\boldsymbol{\varphi}^{\tau|1})\exp(-\eta c_i^\tau)\frac{Y^\tau}{Z^\tau}$$

– *For all $\boldsymbol{\varphi} \in \Phi$ with $\varphi_{\ell^\tau} = \psi$ we have:*

$$\gamma_i^{\tau+1}(\boldsymbol{\varphi}) = 0$$

Induction with Theorem 10 leads to the following theorem:

Theorem 11. *For all $\tau \in [T]$ and for all $\boldsymbol{\varphi} \in \Phi$ we have:*

$$\gamma_i^\tau(\boldsymbol{\varphi}) = \pi_i^\tau \prod_{j\in[s]} \left(\bar{w}_i^{\tau,j}\mathbb{I}\left(\varphi_j = \varpi_j^\tau(1)\right) + (1-\bar{w}_i^{\tau,j})\mathbb{I}\left(\varphi_j = \varpi_j^\tau(0)\right)\right)$$

which implies:

$$\sum_{\boldsymbol{\varphi}\in\widetilde{\mathcal{W}}^\tau} \gamma_i^\tau(\boldsymbol{\varphi}) = \pi_i^\tau \bar{w}_i^{\tau,\ell^\tau} .$$

Theorems 11 and 7 then lead to the following theorem:

Theorem 12. *The prediction $\boldsymbol{v}^\tau$ of Algorithm 1 is equal to that defined in Section A.1.5 when using the Markov circadian defined in Section A.1.8.*

A.1.10 Participating Circadian Instances

In this section we define the Circadian Instances $\{\hat{\alpha}^i \mid i \in \mathbb{M}\}$ used in Section A.1.7 to bound the loss of Algorithm 1. Firstly, for all $j \in [n]$ and $\tau \in [T] \setminus \Omega_j$ we define $\bar{\nu}_j(\tau)$ as:

$$\bar{\nu}_j(\tau) := \min\left\{\tau' \in [T] \mid \tau' \geq \tau \ \wedge \ \ell^{\tau'} = j\right\}$$

For all $i \in \mathbb{M}$, we define the Circadian Instance $\hat{\alpha}^i$ by, for all $j \in [s]$:

- For all $\tau \in [T] \setminus \Omega_j$ we have:

$$\hat{\alpha}^i(\tau)_j := \mathbb{I}\left(i = z_{\bar{\nu}_j(\tau)}\right)$$

- For all $\tau \in \Omega_j$ we have:

$$\hat{\alpha}^i(\tau)_j := \psi$$

A.1.11 The Bound

We now bound the performance of Algorithm 1. Firstly, for all $\tau \in [T] \setminus \bar{\Omega}$ we define $\nu(\tau)$ as:

$$\nu(\tau) := \min\{\tau' \in [T] \mid \tau' > \tau \ \wedge \ \ell^{\tau'} = j\}$$

We start with the following theorem:

Theorem 13. *For all $i \in \mathbb{M}$ we have:*

$$\begin{aligned}
&\ln(\lambda'(\hat{\alpha}^i(1))) + \sum_{\tau\in[T-1]} \ln(\lambda^\tau(\hat{\alpha}^i(\tau), \hat{\alpha}^i(\tau+1))) \\
&= \sum_{j\in[s]} (\ln(\mu)\mathbb{I}\left(z_{\bar{\nu}_j(1)} = i\right) + \ln(1-\mu)\mathbb{I}\left(z_{\bar{\nu}_j(1)} \neq i\right)) \\
&\quad + \ln(\theta) \sum_{\tau\in[T]\setminus\bar{\Omega}} \mathbb{I}\left(z_\tau = i \ \wedge \ z_{\nu(\tau)} = i\right) \\
&\quad + \ln(1-\phi) \sum_{\tau\in[T]\setminus\bar{\Omega}} \mathbb{I}\left(z_\tau \neq i \ \wedge \ z_{\nu(\tau)} \neq i\right)
\end{aligned}$$

$$+\ln(1-\theta)\sum_{\tau\in[T]\setminus\bar{\Omega}}\mathbb{I}\left(z_\tau=i\ \wedge\ z_{\nu(\tau)}\neq i\right)$$
$$+\ln(\phi)\sum_{\tau\in[T]\setminus\bar{\Omega}}\mathbb{I}\left(z_\tau\neq i\ \wedge\ z_{\nu(\tau)}=i\right)$$

Substituting the inequality of Theorem 13 into that of Theorem 9 gives us the following theorem:

Theorem 14. *The regret of the Algorithm 1 is bounded as:*

$$\sum_{\tau\in[T]}\left(c_*^\tau-c_{z_\tau}^\tau\right)\leq\frac{\eta}{2}T+\frac{1}{\eta}C$$

where:

$$C:=m\ln\left(\frac{n}{m}\right)+s\left(\ln\left(\mu\right)+(m-1)\ln\left(1-\mu\right)\right)$$
$$+\ln(\theta)(T-s-k)+\ln(1-\phi)((T-s)(m-1)-k)+\ln(1-\theta)k+\ln(\phi)k$$

Tuning the parameters then leads to the following theorem:

Theorem 15. *Setting $\mu:=1/m$, $\theta:=1-k/(T-s)$ and $\phi:=k/((m-1)(T-s))$ we have that the regret, $\sum_{\tau\in[T]}\left(c_*^\tau-c_{z_\tau}^\tau\right)$ of Algorithm 1 is bounded above by:*

$$\sqrt{2T\left(m\log\left(\frac{n}{m}\right)+s(\log(m)+1)+k\left(\log(m-1)+2\log\left(\frac{T-s}{k}\right)+2\right)\right)}\qquad(9)$$

and also, more tightly, bounded above by:

$$\sqrt{2T}\sqrt{m\log\left(\frac{n}{m}\right)+smH\left(\frac{1}{m}\right)+(T-s)H\left(\frac{k}{T-s}\right)+(m-1)(T-s)H\left(\frac{k}{(m-1)(T-s)}\right)}.\qquad(10)$$

We now reduce the finite hypothesis class setting to the Multitask Allocation game, which will prove Theorem 1. First, we let $n:=|\mathcal{H}_{\text{fin}}|$, define a bijection $\kappa:[n]\to\mathcal{H}_{\text{fin}}$, and for all $\tau\in[T]$ define $z_\tau=\kappa^{-1}(h_\tau)$. On trial τ the Learner randomly draws $i^\tau\in[n]$ with probability $v_{i^\tau}^t$ and predicts with $\hat{y}^\tau:=[\kappa(i^\tau)](x^\tau)$. We then define $\boldsymbol{c}^\tau$ by $c_i^\tau:=\mathcal{L}_{01}(y^\tau,[\kappa(i)](x^\tau))$.

Theorem 16. *We have the following equivalence,*

$$R_T=\sum_{i\in[s]}\sum_{t\in[T^i]}\mathbb{E}[\mathcal{L}_{01}(y_t^i,\hat{y}_t^i)]-\mathcal{L}_{01}(y_t^i,h_t^i(x_t^i))=\sum_{\tau\in[T]}\left(c_*^\tau-c_{z_\tau}^\tau\right)\ .$$

Theorem 1 follows from theorems 15 and 16, noting that $k=k(\boldsymbol{h}^*)$ and $m=|m(\boldsymbol{h}^*)|$. ■

A.2 Proofs

We now prove the theorems of Section A.1 in order.

A.2.1 Theorem 5

Proof. The proof is similar to that in [48], and utilizes the allocation model introduced in [38].

We first show, by induction on τ, that for all $\tau\in[T]$ we have $\sum_{\xi\in\mathcal{E}}\rho^\tau(\xi)=1$. This is clearly true for $\tau=1$ since $\sum_{\xi\in\mathcal{E}}\rho^1(\xi)=\sum_{\xi\in\mathcal{E}}\rho^*(\xi)=1$. Now suppose that it is true for $\tau=\tau'$ (for some $\tau'\in[T]$). We now show that it is true for $\tau=\tau'+1$.

$$\sum_{\xi\in\mathcal{E}}\rho^{\tau'+1}(\xi)=\sum_{\xi\in\mathcal{E}\setminus\mathcal{W}^{\tau'}}\rho^{\tau'+1}(\xi)+\sum_{\xi\in\mathcal{W}^{\tau'}}\rho^{\tau'+1}(\xi)\qquad(11)$$
$$=\sum_{\xi\in\mathcal{E}\setminus\mathcal{W}^{\tau'}}\rho^{\tau'}(\xi)+\sum_{\xi\in\mathcal{W}^{\tau'}}\rho^{\tau'+1}(\xi)\qquad(12)$$

$$
\begin{aligned}
&= \sum_{\xi\in\mathcal{E}\setminus\mathcal{W}^{\tau'}} \rho^{\tau'}(\xi) + \sum_{\xi\in\mathcal{W}^{\tau'}} \frac{Y^{\tau'}}{Z^{\tau'}} \rho^{\tau'}(\xi)\exp(-\eta\tilde{c}^{\tau'}(\xi)) && (13)\\
&= \sum_{\xi\in\mathcal{E}\setminus\mathcal{W}^{\tau'}} \rho^{\tau'}(\xi) + \sum_{\xi\in\mathcal{W}^{\tau'}} \frac{\rho^{\tau'}(\xi)\exp(-\eta\bar{c}^{\tau'}(\xi))\sum_{\xi'\in\mathcal{W}^{\tau'}}\rho^{\tau'}(\xi')}{\sum_{\xi'\in\mathcal{W}^{\tau'}}\rho^{\tau'}(\xi')\exp(-\eta\bar{c}^{\tau'}(\xi'))} && (14)\\
&= \sum_{\xi\in\mathcal{E}\setminus\mathcal{W}^{\tau'}} \rho^{\tau'}(\xi) + \frac{\sum_{\xi\in\mathcal{W}^{\tau'}}\rho^{\tau'}(\xi)\exp(-\eta\bar{c}^{\tau'}(\xi))\sum_{\xi'\in\mathcal{W}^{\tau'}}\rho^{\tau}(\xi')}{\sum_{\xi'\in\mathcal{W}^{\tau'}}\rho^{\tau'}(\xi')\exp(-\eta\bar{c}^{\tau'}(\xi'))} && (15)\\
&= \sum_{\xi\in\mathcal{E}\setminus\mathcal{W}^{\tau'}} \rho^{\tau'}(\xi) + \sum_{\xi'\in\mathcal{W}^{\tau'}} \rho^{\tau'}(\xi') && (16)\\
&= \sum_{\xi\in\mathcal{E}} \rho^{\tau'}(\xi) && (17)\\
&= 1 && (18)
\end{aligned}
$$

With this in hand, we now have that $\mathrm{KL}(u,\rho^\tau)$ is defined and positive for all $\tau\in[T]$.

Let $A^\tau := \sum_{\xi'\in\mathcal{W}^\tau}\bar{v}^\tau(\xi')\exp(-\eta\bar{c}^\tau(\xi'))$. By definition of $\bar{v}^\tau$ we have:

$$
A^\tau = \frac{\sum_{\xi'\in\mathcal{W}^\tau}\rho^\tau(\xi')\exp(-\eta\bar{c}^\tau(\xi'))}{\sum_{\xi'\in\mathcal{W}^\tau}\rho^\tau(\xi')} \tag{19}
$$

so:

$$
\begin{aligned}
&\mathrm{KL}(u,\rho^\tau) - \mathrm{KL}(u,\rho^{\tau+1})\\
=&\sum_{\xi\in\mathcal{E}} u(\xi)\left(\log\left(\frac{u(\xi)}{\rho^\tau(\xi)}\right) - \log\left(\frac{u(\xi)}{\rho^{\tau+1}(\xi)}\right)\right)\\
=&\sum_{\xi\in\mathcal{E}} u(\xi)\log\left(\frac{\rho^{\tau+1}(\xi)}{\rho^\tau(\xi)}\right)\\
=&\sum_{\xi\in\mathcal{W}^\tau} u(\xi)\log\left(\frac{\rho^{\tau+1}(\xi)}{\rho^\tau(\xi)}\right) && (20)\\
=&u(\mathcal{W}^\tau)\sum_{\xi\in\mathcal{W}^\tau} \frac{u(\xi)}{u(\mathcal{W}^\tau)}\log\left(\frac{\rho^{\tau+1}(\xi)}{\rho^\tau(\xi)}\right)\\
=&u(\mathcal{W}^\tau)\sum_{\xi\in\mathcal{W}^\tau} \bar{u}^\tau(\xi)\log\left(\frac{\rho^{\tau+1}(\xi)}{\rho^\tau(\xi)}\right)\\
=&u(\mathcal{W}^\tau)\sum_{\xi\in\mathcal{W}^\tau} \bar{u}^\tau(\xi)\log\left(\frac{Y^{\tau'}}{Z^{\tau'}}\exp(-\eta\tilde{c}^{\tau'}(\xi))\right) && (21)\\
=&u(\mathcal{W}^\tau)\sum_{\xi\in\mathcal{W}^\tau} \bar{u}^\tau(\xi)\log\left(\frac{\exp(-\eta\bar{c}^\tau(\xi))\sum_{\xi'\in\mathcal{W}^\tau}\rho^\tau(\xi')}{\sum_{\xi'\in\mathcal{W}^\tau}\rho^\tau(\xi')\exp(-\eta\bar{c}^\tau(\xi'))}\right) && (22)\\
=&u(\mathcal{W}^\tau)\sum_{\xi\in\mathcal{W}^\tau} \bar{u}^\tau(\xi)\log\left(\frac{\exp(-\eta\bar{c}^\tau(\xi))}{A^\tau}\right) && (23)\\
=&-u(\mathcal{W}^\tau)\eta\left(\sum_{\xi\in\mathcal{W}^\tau}\bar{u}^\tau(\xi)\bar{c}^\tau(\xi)\right) - u(\mathcal{W}^\tau)\log(A^\tau)\left(\sum_{\xi\in\mathcal{W}^\tau}\bar{u}^\tau(\xi)\right)\\
=&-u(\mathcal{W}^\tau)\eta\langle\bar{u}^\tau,\bar{c}^\tau\rangle - u(\mathcal{W}^\tau)\log(A^\tau)\\
=&-u(\mathcal{W}^\tau)\eta\langle\bar{u}^\tau,\bar{c}^\tau\rangle - u(\mathcal{W}^\tau)\log\left(\sum_{\xi\in\mathcal{W}^\tau}\bar{v}^\tau(\xi)\exp(-\eta\bar{c}^\tau(\xi))\right)
\end{aligned}
$$

$$\geq - u(\mathcal{W}^\tau)\eta\langle \bar{u}^\tau, \bar{c}^\tau\rangle - u(\mathcal{W}^\tau)\log\left(\sum_{\xi\in\mathcal{W}^\tau}\bar{v}^\tau(\xi)\left(1-\eta\bar{c}^\tau(\xi)+\frac{1}{2}\eta^2\left(\bar{c}^\tau(\xi)\right)^2\right)\right) \tag{24}$$

$$= - u(\mathcal{W}^\tau)\eta\langle \bar{u}^\tau, \bar{c}^\tau\rangle - u(\mathcal{W}^\tau)\log\left(1-\eta\langle\bar{v}^\tau,\bar{c}^\tau\rangle+\frac{1}{2}\eta^2\sum_{\xi\in\mathcal{W}^\tau}\bar{v}^\tau(\xi)\left(\bar{c}^\tau(\xi)\right)^2\right)$$

$$\geq - u(\mathcal{W}^\tau)\eta\langle \bar{u}^\tau, \bar{c}^\tau\rangle - u(\mathcal{W}^\tau)\log\left(1-\eta\langle\bar{v}^\tau,\bar{c}^\tau\rangle+\frac{1}{2}\eta^2\sum_{\xi\in\mathcal{W}^\tau}\bar{v}^\tau(\xi)\right)$$

$$= - u(\mathcal{W}^\tau)\eta\langle \bar{u}^\tau, \bar{c}^\tau\rangle - u(\mathcal{W}^\tau)\log\left(1-\eta\langle\bar{v}^\tau,\bar{c}^\tau\rangle+\frac{1}{2}\eta^2\right)$$

$$\geq - u(\mathcal{W}^\tau)\eta\langle \bar{u}^\tau, \bar{c}^\tau\rangle + u(\mathcal{W}^\tau)\eta\langle\bar{v}^\tau,\bar{c}^\tau\rangle - u(\mathcal{W}^\tau)\frac{1}{2}\eta^2 \tag{25}$$

$$= u(\mathcal{W}^\tau)\eta\langle\bar{v}^\tau-\bar{u}^\tau,\bar{c}^\tau\rangle - u(\mathcal{W}^\tau)\frac{1}{2}\eta^2$$

where Equation (20) comes from the fact that if $\xi \in \mathcal{E}\setminus\mathcal{W}^\tau$ then $\rho^{\tau+1}(\xi) = \rho^\tau(\xi)$ so $\log(\rho^{\tau+1}(\xi)/\rho^\tau(\xi)) = 0$, Equation (21) comes from the update of $\rho^\tau(\xi)$ to $\rho^{\tau+1}(\xi)$ when $\xi \in \mathcal{W}^\tau$, Equation (23) comes from Equation (19), Equation (24) comes from the inequality $\exp(x) \leq 1 - x + x^2/2$ for $x \geq 0$, and Equation (25) comes from the inequality $\log(1+x) \leq x$.

A telescoping sum then gives us:

$$\begin{aligned}
\mathrm{KL}(u^1,\rho^*) &= \mathrm{KL}(u^1,\rho^1)\\
&\geq \mathrm{KL}(u^1,\rho^1) - \mathrm{KL}(u,\rho^{T+1})\\
&= \sum_{\tau\in[T]}\mathrm{KL}(u,\rho^\tau) - \mathrm{KL}(u,\rho^{\tau+1})\\
&= \sum_{\tau\in[T]}\left(u(\mathcal{W}^\tau)\eta\langle\bar{v}^\tau-\bar{u}^\tau,\bar{c}^\tau\rangle - u(\mathcal{W}^\tau)\frac{1}{2}\eta^2\right).
\end{aligned}$$

Dividing by η and rearranging then gives us the result. □

A.2.2 Theorem 6

We have:

$$\begin{aligned}
\langle\tilde{v}^\tau,\tilde{c}^\tau\rangle &= \sum_{\xi\in\mathcal{W}^\tau}\tilde{v}^\tau(\xi)\boldsymbol{c}^\tau(\xi)\\
&= \sum_{(i,\alpha)\in\mathcal{W}^\tau}\tilde{v}^\tau(i,\alpha)\boldsymbol{c}^\tau(i,\alpha)\\
&= \sum_{(i,\alpha)\in\mathcal{W}^\tau}\tilde{v}^\tau(i,\alpha)c_i^\tau\\
&= \sum_{i\in[n]}c_i^\tau\sum_{\alpha\in\mathcal{A}:(i,\alpha)\in\mathcal{W}^\tau}\tilde{v}^\tau(i,\alpha)\\
&= \sum_{i\in[n]}c_i^\tau v_i^\tau\\
&= \boldsymbol{v}^\tau\cdot\boldsymbol{c}^\tau
\end{aligned}$$

■

A.2.3 Theorem 7

We start with the following definition:

Definition 17. *Given some $\tau \in T$ and some $\alpha \in \Phi^\tau$ we define:*

$$\mathcal{B}(\alpha) := \{\alpha' \in \mathcal{A} : \forall \tau' \in [\tau], \alpha'(\tau) = \alpha(\tau)\}$$

and define:

$$q(\alpha) := \lambda'(\alpha(1)) \prod_{\tau' \in [\tau-1]} \lambda^{\tau'}(\alpha(\tau'), \alpha(\tau'+1))$$

.

Lemma 18. *Given some $\tau \in T$ and some $\alpha \in \Phi^\tau$ we have:*

$$\sum_{\alpha' \in \mathcal{B}(\alpha)} q(\alpha') = q(\alpha)$$

Proof. We prove by reverse induction of τ. i.e. from $\tau = T$ to $\tau = 1$. It is clear that it holds for $\tau = T$ as if $\alpha \in \Phi^T$ we have $\mathcal{B}(\alpha) = \{\alpha\}$. Now suppose it holds for $\tau = \bar{\tau}$ (for some $\bar{\tau} \in [T]$ with $\bar{\tau} > 1$) and suppose we have some $\alpha \in \Phi^{\bar{\tau}-1}$. For $\varphi \in \Phi$ define $\alpha^\varphi \in \Phi^{\bar{\tau}}$ by $\alpha^\varphi(\bar{\tau}) := \varphi$ and for all $\tau' < \bar{\tau}$, $\alpha^\varphi(\tau') := \alpha(\tau')$. The sets $\{\mathcal{B}(\alpha^\varphi) \mid \varphi \in \Phi\}$ are pairwise disjoint with union $\mathcal{B}(\alpha)$ and hence:

$$\sum_{\alpha' \in \mathcal{B}(\alpha)} q(\alpha') = \sum_{\varphi \in \Phi} \sum_{\alpha' \in \mathcal{B}(\alpha^\varphi)} q(\alpha')$$

so by the inductive hypothesis, and noting. that $q(\alpha^\varphi) = q(\alpha)\lambda^{\tau-1}(\alpha(\tau-1), \varphi)$ we have:

$$\begin{aligned}
&\sum_{\alpha' \in \mathcal{B}(\alpha)} q(\alpha')) \\
=&\sum_{\varphi \in \Phi} \sum_{\alpha' \in \mathcal{B}(\alpha^\varphi)} q(\alpha') \\
=&\sum_{\varphi \in \Phi} q(\alpha^\varphi) \\
=&\sum_{\varphi \in \Phi} q(\alpha)\lambda^{\tau-1}(\alpha(\tau-1), \varphi) \\
=&q(\alpha) \sum_{\varphi \in \Phi} \lambda^{\tau-1}(\alpha(\tau-1), \varphi) \\
=&q(\alpha)
\end{aligned}$$

□

Lemma 19. *For all $\tau \in [T]$, all $i \in [n]$ and all $\alpha \in \Phi^\tau$ there exists some value $f^\tau(i, \alpha) \in \mathbb{R}$ such that for all $\alpha' \in \mathcal{B}(\alpha)$ we have $\rho^{\tau+1}(i, \alpha') = f^\tau(i, \alpha)q(\alpha')$. In addition, if we define $\bar{\alpha} \in \Phi^{\tau-1}$ by $\bar{\alpha}(\bar{\tau}) = \alpha(\bar{\tau})$ for all $\bar{\tau} \in [\tau-1]$ we have:*

- *If $\alpha(\tau) \in \widetilde{\mathcal{W}}^\tau$ then $f^\tau(i, \alpha) := f^{\tau-1}(i, \bar{\alpha})\frac{Y^\tau}{Z^\tau}\exp(-\eta c_i^\tau)$*
- *If $\alpha(\tau) \notin \widetilde{\mathcal{W}}^\tau$ then $f^\tau(i, \alpha) := f^{\tau-1}(i, \bar{\alpha})$*

Proof. We prove by induction on τ. For $\tau = 0$ we can define $f^\tau(i, \alpha) = 1$ so for all $\alpha' \in \mathcal{B}(\alpha)$ we have $\rho^{\tau+1}(i, \alpha') = \rho^1(i, \alpha') = q(\alpha') = f^\tau(i, \alpha)q(\alpha')$. Now suppose it holds for $\tau := \tau'$ (for some $\tau' \in [T]$). We shall now show it holds for $\tau := \tau' + 1$. Define $\bar{\alpha} \in \Phi^{\tau'}$ by $\bar{\alpha}(\bar{\tau}) = \alpha(\bar{\tau})$ for all $\bar{\tau} \in [\tau']$. Choose any $\alpha' \in \mathcal{B}(\alpha)$. Then we have $\alpha' \in \mathcal{B}(\bar{\alpha})$ and $\alpha'(\tau'+1) = \alpha(\tau'+1)$. Since $\alpha' \in \mathcal{B}(\bar{\alpha})$ we have, from the inductive hypothesis, that $\rho^{\tau'+1}(i, \alpha') = f^{\tau'}(i, \bar{\alpha})q(\alpha')$. We have two cases:

- If $\alpha(\tau) \in \widetilde{\mathcal{W}}^\tau$ then $\alpha'(\tau'+1) = \alpha(\tau'+1) = \alpha(\tau) \in \widetilde{\mathcal{W}}^{\tau'+1}$ so $(i, \alpha') \in \mathcal{W}^{\tau'+1}$ and hence:

$$\rho^{\tau+1}(i, \alpha') = \rho^{\tau'+2}(i, \alpha')$$

$$
\begin{aligned}
&= \frac{Y^{\tau'+1}}{Z^{\tau'+1}} \rho^{\tau'+1}(i,\alpha') \exp(-\eta \tilde{c}^{(\tau'+1)}(i,\alpha')) \\
&= \frac{Y^{\tau'+1}}{Z^{\tau'+1}} \rho^{\tau'+1}(i,\alpha') \exp(-\eta c_i^{\tau'+1}) \\
&= \frac{Y^{\tau'+1}}{Z^{\tau'+1}} f^{\tau'}(i,\bar{\alpha}) q(\alpha') \exp(-\eta c_i^{\tau'+1}) \\
&= \frac{Y^{\tau}}{Z^{\tau}} f^{\tau-1}(i,\bar{\alpha}) q(\alpha') \exp(-\eta c_i^{\tau})
\end{aligned}
$$

and hence we have the result with $f^\tau(i,\alpha) := f^{\tau-1}(i,\bar{\alpha})\frac{Y^\tau}{Z^\tau}\exp(-\eta c_i^\tau)$

- If $\alpha(\tau) \notin \widetilde{\mathcal{W}}^\tau$ then $\alpha'(\tau'+1) = \alpha(\tau'+1) = \alpha(\tau) \notin \widetilde{\mathcal{W}}^{\tau'+1}$ so $(i,\alpha') \notin \mathcal{W}^{\tau'+1}$ and hence:

$$
\begin{aligned}
\rho^{\tau+1}(i,\alpha') &= \rho^{\tau'+2}(i,\alpha') \\
&= \rho^{\tau'+1}(i,\alpha') \\
&= f^{\tau'}(i,\bar{\alpha}) q(\alpha') \\
&= f^{\tau-1}(i,\bar{\alpha}) q(\alpha')
\end{aligned}
$$

and hence we have the result with $f^\tau(i,\alpha) := f^{\tau-1}(i,\bar{\alpha})$

□

Definition 20. *For all $\tau \in [T]$, all $i \in [n]$ and all $\alpha \in \Phi^\tau$ let $f^\tau(i,\alpha)$ be defined as in Lemma 19. For all $\varphi \in \Phi$ and $i \in [n]$ we define:*

$$
g^\tau(i,\varphi) := \sum_{\alpha \in \Phi^\tau : \alpha(\tau) = \varphi} f^\tau(i,\alpha) q(\alpha)
$$

In addition, given $\alpha \in \Phi^\tau$, define $\bar{\alpha} \in \Phi^{\tau-1}$ by $\bar{\alpha}(\bar{\tau}) = \alpha(\bar{\tau})$ for all $\bar{\tau} \in [\tau-1]$

Lemma 21. *Given $\tau \in [T]$ with $\tau \geq 2$ and some $\varphi \in \Phi$ we have:*

- *If $\varphi \in \widetilde{\mathcal{W}}^\tau$ then $g^\tau(i,\varphi) = \frac{Y^\tau}{Z^\tau}\exp(-\eta c_i^\tau)\sum_{\varphi' \in \Phi} \lambda^{\tau-1}(\varphi',\varphi) g^{\tau-1}(i,\varphi')$*
- *If $\varphi \in \widetilde{\mathcal{W}}^\tau$ then $g^\tau(i,\varphi) = \sum_{\varphi' \in \Phi} \lambda^{\tau-1}(\varphi',\varphi) g^{\tau-1}(i,\varphi')$*

Proof. Given $\alpha \in \Phi\tau$, define $\bar{\alpha} \in \Phi^{\tau-1}$ by $\bar{\alpha}(\bar{\tau}) = \alpha(\bar{\tau})$ for all $\bar{\tau} \in [\tau-1]$. Let x be defined as follows:

- If $\varphi \in \widetilde{\mathcal{W}}^\tau$ then $x := \frac{Y^\tau}{Z^\tau}\exp(-\eta c_i^\tau)$
- If $\varphi \notin \widetilde{\mathcal{W}}^\tau$ then $x := 1$

From Lemma 19 we have, for all $\alpha \in \Phi\tau$, that If $\varphi \in \widetilde{\mathcal{W}}^\tau$ we have, from Lemma 19 that $f^\tau(i,\alpha) = x f^{\tau-1}(i,\bar{\alpha})$. Hence, we have:

$$
\begin{aligned}
g^\tau(i,\varphi) &= \sum_{\alpha \in \Phi^\tau : \alpha(\tau) = \varphi} f^\tau(i,\alpha) q(\alpha) \\
&= \sum_{\alpha \in \Phi^\tau : \alpha(\tau) = \varphi} f^\tau(i,\alpha) q(\bar{\alpha}) \lambda^{\tau-1}(\alpha(\tau-1), \alpha(\tau)) \\
&= \sum_{\alpha \in \Phi^\tau : \alpha(\tau) = \varphi} f^\tau(i,\alpha) q(\bar{\alpha}) \lambda^{\tau-1}(\bar{\alpha}(\tau-1), \varphi) \\
&= \sum_{\alpha \in \Phi^\tau : \alpha(\tau) = \varphi} x f^{\tau-1}(i,\bar{\alpha}) q(\bar{\alpha}) \lambda^{\tau-1}(\bar{\alpha}(\tau-1), \varphi)
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{\bar{\alpha}\in\Phi^{\tau-1}} x f^{\tau-1}(i,\bar{\alpha})q(\bar{\alpha})\lambda^{\tau-1}(\bar{\alpha}(\tau-1),\varphi)\\
&= \sum_{\alpha\in\Phi^{\tau-1}} x f^{\tau-1}(i,\alpha)q(\alpha)\lambda^{\tau-1}(\alpha(\tau-1),\varphi)\\
&= \sum_{\varphi'\in\Phi}\sum_{\alpha\in\Phi^{\tau-1}:\alpha(\tau-1)=\varphi'} x f^{\tau-1}(i,\alpha)q(\alpha)\lambda^{\tau-1}(\alpha(\tau-1),\varphi)\\
&= \sum_{\varphi'\in\Phi}\sum_{\alpha\in\Phi^{\tau-1}:\alpha(\tau-1)=\varphi'} x f^{\tau-1}(i,\alpha)q(\alpha)\lambda^{\tau-1}(\varphi',\varphi)\\
&= x\sum_{\varphi'\in\Phi}\lambda^{\tau-1}(\varphi',\varphi)\sum_{\alpha\in\Phi^{\tau-1}:\alpha(\tau-1)=\varphi'} f^{\tau-1}(i,\alpha)q(\alpha)\\
&= x\sum_{\varphi'\in\Phi}\lambda^{\tau-1}(\varphi',\varphi)g^{\tau-1}(i,\varphi')
\end{aligned}
$$

□

Lemma 22. *For all trials $\tau\in[T]$ with $\tau>1$, experts $i\in[n]$ and states $\varphi\in\Phi$ we have:*

$$\gamma_i^\tau(\varphi)=\sum_{\varphi'\in\Phi}\lambda^{\tau-1}(\varphi',\varphi)g^{\tau-1}(i,\varphi')$$

Proof. Let x be defined as follows:

- If $\varphi\in\widetilde{\mathcal{W}}^\tau$ then $x:=\frac{Y^\tau}{Z^\tau}\exp(-\eta c_i^\tau)$
- If $\varphi\notin\widetilde{\mathcal{W}}^\tau$ then $x:=1$

Note first that $\{\alpha'\in\mathcal{A}:\alpha'(\tau)=\varphi\}=\bigcup_{\alpha\in\Phi^\tau:\alpha(\tau)=\varphi}\mathcal{B}(\alpha)$ where the sets in the union are disjoint. So from Lemma 18 we have:

$$
\begin{aligned}
&\gamma_i^\tau(\varphi)\\
&= \sum_{\alpha'\in\mathcal{A}:\alpha'(\tau)=\varphi}\rho^\tau(i,\alpha')\\
&= \sum_{\alpha\in\Phi^\tau:\alpha(\tau)=\varphi}\sum_{\alpha'\in\mathcal{B}(\alpha)}\rho^\tau(i,\alpha')\\
&= \sum_{\alpha\in\Phi^\tau:\alpha(\tau)=\varphi}\sum_{\alpha'\in\mathcal{B}(\alpha)} f^{\tau-1}(i,\bar{\alpha})q(\alpha')\\
&= \sum_{\alpha\in\Phi^\tau:\alpha(\tau)=\varphi} f^{\tau-1}(i,\bar{\alpha})\sum_{\alpha'\in\mathcal{B}(\alpha)} q(\alpha')\\
&= \sum_{\alpha\in\Phi^\tau:\alpha(\tau)=\varphi} f^{\tau-1}(i,\bar{\alpha})q(\alpha)\\
&= \sum_{\alpha\in\Phi^\tau:\alpha(\tau)=\varphi} f^{\tau-1}(i,\bar{\alpha})\lambda^{\tau-1}(\alpha(\tau-1),\alpha(\tau))q(\bar{\alpha})\\
&= \sum_{\alpha\in\Phi^\tau:\alpha(\tau)=\varphi} f^{\tau-1}(i,\bar{\alpha})\lambda^{\tau-1}(\bar{\alpha}(\tau-1),\varphi)q(\bar{\alpha})\\
&= \sum_{\bar{\alpha}\in\Phi^{\tau-1}} f^{\tau-1}(i,\bar{\alpha})\lambda^{\tau-1}(\bar{\alpha}(\tau-1),\varphi)q(\bar{\alpha})\\
&= \sum_{\alpha\in\Phi^{\tau-1}} f^{\tau-1}(i,\alpha)\lambda^{\tau-1}(\alpha(\tau-1),\varphi)q(\alpha)\\
&= \sum_{\varphi'\in\Phi}\sum_{\alpha\in\Phi^{\tau-1}:\alpha(\tau-1)=\varphi'} f^{\tau-1}(i,\alpha)\lambda^{\tau-1}(\alpha(\tau-1),\varphi)q(\alpha)
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{\varphi' \in \Phi} \sum_{\alpha \in \Phi^{\tau-1} : \alpha(\tau-1) = \varphi'} f^{\tau-1}(i, \alpha) \lambda^{\tau-1}(\varphi', \varphi) q(\alpha) \\
&= \sum_{\varphi' \in \Phi} \lambda^{\tau-1}(\varphi', \varphi) \sum_{\alpha \in \Phi^{\tau-1} : \alpha(\tau-1) = \varphi'} f^{\tau-1}(i, \alpha) q(\alpha) \\
&= \sum_{\varphi' \in \Phi} \lambda^{\tau-1}(\varphi', \varphi) g^{\tau-1}(i, \varphi')
\end{aligned}
$$

□

Lemma 23. *We have* $\gamma_i^1(\varphi) = \lambda'(\varphi)/n$

Proof. Let $\alpha \in \Phi^1$ be such that $\alpha(1) := \varphi$. We have, from Lemma 18:

$$
\begin{aligned}
\gamma_i^1(\varphi) &= \sum_{\alpha' \in \mathcal{A} : \alpha'(1) = \varphi} \rho^1(i, \alpha') \\
&= \sum_{\alpha' \in \mathcal{A} : \alpha'(1) = \varphi} \rho^*(i, \alpha') \\
&= \sum_{\alpha' \in \mathcal{A} : \alpha'(1) = \varphi} \frac{1}{n} q(\alpha') \\
&= \sum_{\alpha' \in \mathcal{B}(\alpha)} \frac{1}{n} q(\alpha') \\
&= \frac{1}{n} \sum_{\alpha' \in \mathcal{B}(\alpha)} q(\alpha') \\
&= \frac{1}{n} q(\alpha) \\
&= \frac{1}{n} \lambda'(\alpha(1)) \\
&= \frac{1}{n} \lambda'(\varphi)
\end{aligned}
$$

□

Lemma 24. *For all* $\tau \in [T]$ *we have:*

$$
\begin{aligned}
\gamma_i^{\tau+1}(\varphi) =& \frac{Y^\tau}{Z^\tau} \sum_{\varphi' \in \widetilde{\mathcal{W}}^\tau} \lambda^\tau(\varphi', \varphi) \gamma_i^\tau(\varphi') \exp(-\eta c_i^\tau) \\
&+ \sum_{\varphi' \in \Phi \setminus \widetilde{\mathcal{W}}^\tau} \lambda^\tau(\varphi', \varphi) \gamma_i^\tau(\varphi')
\end{aligned}
$$

Proof. From Lemma 22 we have:

$$
\begin{aligned}
&\gamma_i^{\tau+1}(\varphi) \\
&= \sum_{\varphi' \in \Phi} \lambda^\tau(\varphi', \varphi) g^\tau(i, \varphi') \\
&= \sum_{\varphi' \in \widetilde{\mathcal{W}}^\tau} \lambda^\tau(\varphi', \varphi) g^\tau(i, \varphi') + \sum_{\varphi' \in \Phi \setminus \widetilde{\mathcal{W}}^\tau} \lambda^\tau(\varphi', \varphi) g^\tau(i, \varphi')
\end{aligned}
$$

So by Lemma 21 we have:

$$
\begin{aligned}
&\gamma_i^{\tau+1}(\varphi) \\
&= \sum_{\varphi' \in \widetilde{\mathcal{W}}^\tau} \lambda^\tau(\varphi', \varphi) \frac{Y^\tau}{Z^\tau} \exp(-\eta c_i^\tau) \sum_{\varphi'' \in \Phi} \lambda^{\tau-1}(\varphi'', \varphi') g^{\tau-1}(i, \varphi'')
\end{aligned}
$$

$$+\sum_{\varphi'\in\Phi\setminus\widetilde{\mathcal{W}}^\tau}\lambda^\tau(\varphi',\varphi)\sum_{\varphi''\in\Phi}\lambda^{\tau-1}(\varphi'',\varphi')g^{\tau-1}(i,\varphi'')$$

on which the application of Lemma 22 gives us the result □

Lemmas 23 and 24 give us the first part of the Theorem. The rest is proved as follows:
For all trials $\tau\in[T]$ we have, from the Specialist hedge algorithm, that:

$$\begin{aligned}
Y^\tau &= \sum_{\xi\in\mathcal{W}^\tau}\rho^\tau(\xi)\\
&= \sum_{(i,\alpha)\in\mathcal{W}^\tau}\rho^\tau(i,\alpha)\\
&= \sum_{(i,\alpha)\in[n]\times\mathcal{A}:\alpha(\tau)\in\widetilde{\mathcal{W}}^\tau}\rho^\tau(i,\alpha)\\
&= \sum_{i\in[n]}\sum_{\alpha\in\mathcal{A}:\alpha(\tau)\in\widetilde{\mathcal{W}}^\tau}\rho^\tau(i,\alpha)\\
&= \sum_{i\in[n]}\sum_{\varphi\in\widetilde{\mathcal{W}}^\tau}\sum_{\alpha\in\mathcal{A}:\alpha(\tau)=\varphi}\rho^\tau(i,\alpha)\\
&= \sum_{i\in[n]}\sum_{\varphi\in\widetilde{\mathcal{W}}^\tau}\gamma_i^\tau(\varphi)
\end{aligned}$$

and that:

$$\begin{aligned}
Z^\tau &= \sum_{\xi\in\mathcal{W}^\tau}\rho^\tau(\xi)\exp(-\eta\tilde{c}^\tau(\xi))\\
&= \sum_{(i,\alpha)\in\mathcal{W}^\tau}\rho^\tau(i,\alpha)\exp(-\eta\tilde{c}^\tau(i,\alpha))\\
&= \sum_{(i,\alpha)\in\mathcal{W}^\tau}\rho^\tau(i,\alpha)\exp(-\eta c_i^\tau)\\
&= \sum_{(i,\alpha)\in[n]\times\mathcal{A}:\alpha(\tau)\in\widetilde{\mathcal{W}}^\tau}\rho^\tau(i,\alpha)\exp(-\eta c_i^\tau)\\
&= \sum_{i\in[n]}\sum_{\alpha\in\mathcal{A}:\alpha(\tau)\in\widetilde{\mathcal{W}}^\tau}\rho^\tau(i,\alpha)\exp(-\eta c_i^\tau)\\
&= \sum_{i\in[n]}\exp(-\eta c_i^\tau)\sum_{\alpha\in\mathcal{A}:\alpha(\tau)\in\widetilde{\mathcal{W}}^\tau}\rho^\tau(i,\alpha)\\
&= \sum_{i\in[n]}\exp(-\eta c_i^\tau)\sum_{\varphi\in\widetilde{\mathcal{W}}^\tau}\sum_{\alpha\in\mathcal{A}:\alpha(\tau)=\varphi}\rho^\tau(i,\alpha)\\
&= \sum_{i\in[n]}\exp(-\eta c_i^\tau)\sum_{\varphi\in\widetilde{\mathcal{W}}^\tau}\gamma_i^\tau(\varphi)
\end{aligned}$$

and for all $i\in[n]$:

$$\begin{aligned}
v_i^\tau &= \sum_{\alpha\in\mathcal{A}:(i,\alpha)\in\mathcal{W}^\tau}\tilde{v}^\tau(i,\alpha)\\
&= \sum_{\alpha\in\mathcal{A}:(i,\alpha)\in\mathcal{W}^\tau}\rho^\tau(i,\alpha)/Y^\tau\\
&= \sum_{\alpha\in\mathcal{A}:\alpha(\tau)\in\widetilde{\mathcal{W}}^\tau}\rho^\tau(i,\alpha)/Y^\tau
\end{aligned}$$

$$= \frac{1}{Y^\tau} \sum_{\alpha \in \mathcal{A}:\alpha(\tau) \in \widetilde{\mathcal{W}}^\tau} \rho^\tau(i,\alpha)$$

$$= \frac{1}{Y^\tau} \sum_{\varphi \in \widetilde{\mathcal{W}}^\tau} \sum_{\alpha \in \mathcal{A}:\alpha(\tau)=\varphi} \rho^\tau(i,\alpha)$$

$$= \frac{1}{Y^\tau} \sum_{\varphi \in \widetilde{\mathcal{W}}^\tau} \gamma_i^\tau(\varphi)$$

■

A.2.4 Theorem 8

We have, from the definition of u:

$$
\begin{aligned}
& m\,\mathrm{KL}(u,\rho^*) \\
=& m \sum_{\xi \in \mathcal{E}} u(\xi) \ln\left(\frac{u(\xi)}{\rho^*(\xi)}\right) \\
=& m \sum_{(i,\alpha) \in [n]\times\mathcal{A}} u(i,\alpha) \ln\left(\frac{u(i,\alpha)}{\rho^*(i,\alpha)}\right) \\
=& m \sum_{i \in [n]} \sum_{\alpha \in \mathcal{A}} u(i,\alpha) \ln\left(\frac{u(i,\alpha)}{\rho^*(i,\alpha)}\right) \\
=& m \sum_{i \in \mathbb{M}} u(i,\hat{\alpha}^i) \ln\left(\frac{u(i,\hat{\alpha}^i)}{\rho^*(i,\hat{\alpha}^i)}\right) \\
=& m \sum_{i \in \mathbb{M}} \frac{1}{m} \ln\left(\frac{1/m}{\rho^*(i,\hat{\alpha}^i)}\right) \\
=& \sum_{i \in \mathbb{M}} \ln\left(\frac{1/m}{\rho^*(i,\hat{\alpha}^i)}\right) \\
=& \sum_{i \in \mathbb{M}} \ln\left(\frac{1/m}{q(\hat{\alpha}^i)/n}\right) \\
=& \sum_{i \in \mathbb{M}} \ln\left(\frac{n/m}{q(\hat{\alpha}^i)}\right) \\
=& \sum_{i \in \mathbb{M}} \left(\ln\left(\frac{n}{m}\right) + \ln\left(\frac{1}{q(\hat{\alpha}^i)}\right)\right) \\
=& m \ln\left(\frac{n}{m}\right) + \sum_{i \in \mathbb{M}} \ln\left(\frac{1}{q(\hat{\alpha}^i)}\right) \\
=& m \ln\left(\frac{n}{m}\right) - \sum_{i \in \mathbb{M}} \ln(q(\hat{\alpha}^i)) \\
=& m \ln\left(\frac{n}{m}\right) - \sum_{i \in \mathbb{M}} \ln\left(\lambda'(\alpha(1)) \prod_{t \in [T-1]} \lambda^\tau(\alpha(\tau),\alpha(\tau+1))\right) \\
=& m \ln\left(\frac{n}{m}\right) - \sum_{i \in \mathbb{M}} \ln(\lambda'(\hat{\alpha}^i(1))) - \sum_{i \in \mathbb{M}} \sum_{\tau \in [T-1]} \ln(\lambda^\tau(\hat{\alpha}^i(\tau),\hat{\alpha}^i(\tau+1)))
\end{aligned}
$$

A.2.5 Theorem 9

Choose u as in Theorem 8 and for all $\tau \in [T]$ let $\bar{u}^\tau$ and $u(\mathcal{W}^\tau)$ be defined from u as in Theorem 5. We start with the following lemma:

Lemma 25. *For all $\tau \in [T]$ we have $u(\mathcal{W}^\tau) = 1/m$*

Proof. From the definition of u and $\hat{\alpha}^i$ we have, for all $\tau \in [T]$:

$$
\begin{aligned}
&u(\mathcal{W}^\tau) \\
&= \sum_{\xi \in \mathcal{W}^\tau} u(\xi) \\
&= \sum_{(i,\alpha) \in [n] \times \mathcal{A} : (i,\alpha) \in \mathcal{W}^\tau} u(i, \alpha) \\
&= \sum_{(i,\alpha) \in [n] \times \mathcal{A} : \alpha(\tau) \in \widetilde{\mathcal{W}}^\tau} u(i, \alpha) \\
&= \sum_{i \in [n]} \sum_{\alpha \in \mathcal{A} : \alpha(\tau) \in \widetilde{\mathcal{W}}^\tau} u(i, \alpha) \\
&= \sum_{i \in [n]} \sum_{\alpha \in \mathcal{A} : \alpha(\tau) \in \widetilde{\mathcal{W}}^\tau} \frac{1}{m} \mathbb{I}\left(i \in \mathbb{M} \,\wedge\, \alpha = \hat{\alpha}^i\right) \\
&= \sum_{i \in [n]} \sum_{\alpha \in \mathcal{A} : \alpha(\tau) \in \widetilde{\mathcal{W}}^\tau} \frac{1}{m} \mathbb{I}\left(i \in \mathbb{M}\right) \mathbb{I}\left(\alpha = \hat{\alpha}^i\right) \\
&= \frac{1}{m} \sum_{i \in [n]} \mathbb{I}\left(i \in \mathbb{M}\right) \sum_{\alpha \in \mathcal{A} : \alpha(\tau) \in \widetilde{\mathcal{W}}^\tau} \mathbb{I}\left(\alpha = \hat{\alpha}^i\right) \\
&= \frac{1}{m} \sum_{i \in [n]} \mathbb{I}\left(i \in \mathbb{M}\right) \sum_{\alpha \in \mathcal{A}} \mathbb{I}\left(\alpha = \hat{\alpha}^i\right) \mathbb{I}\left(\alpha(\tau) \in \widetilde{\mathcal{W}}^\tau\right) \\
&= \frac{1}{m} \sum_{i \in \mathbb{M}} \sum_{\alpha \in \mathcal{A}} \mathbb{I}\left(\alpha = \hat{\alpha}^i\right) \mathbb{I}\left(\alpha(\tau) \in \widetilde{\mathcal{W}}^\tau\right) \\
&= \frac{1}{m} \sum_{i \in \mathbb{M}} \mathbb{I}\left(\hat{\alpha}^i(\tau) \in \widetilde{\mathcal{W}}^\tau\right) \\
&= \frac{1}{m} \sum_{i \in \mathbb{M}} \mathbb{I}\left(z_\tau = i\right) \\
&= \frac{1}{m} \sum_{i \in \mathbb{M} : i = z_\tau} 1 \\
&= \frac{1}{m}
\end{aligned}
$$

□

Lemma 26. *We have:*

$$
\sum_{\tau \in [T]} u(\mathcal{W}^\tau) \langle \tilde{v}^\tau - \bar{u}^\tau, \tilde{c}^\tau \rangle = \frac{1}{m} \sum_{\tau \in [T]} (c_*^\tau - c_{z_\tau}^\tau)
$$

Proof. Utilising Lemma 25 we have, for all $\tau \in [T]$,

$$
\begin{aligned}
&\langle \bar{u}^\tau, \tilde{c}^\tau \rangle \\
&= \sum_{\xi \in \mathcal{W}^\tau} \bar{u}^\tau(\xi) \tilde{c}^\tau(\xi) \\
&= \sum_{\xi \in \mathcal{E}} \mathbb{I}\left(\xi \in \mathcal{W}^\tau\right) \bar{u}^\tau(\xi) \tilde{c}^\tau(\xi)
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{(i,\alpha)\in[n]\times\mathcal{A}} \mathbb{I}\left((i,\alpha)\in\mathcal{W}^\tau\right)\bar{u}^\tau(i,\alpha)\tilde{c}^\tau(i,\alpha)\\
&= \sum_{(i,\alpha)\in[n]\times\mathcal{A}} \mathbb{I}\left((i,\alpha)\in\mathcal{W}^\tau\right)\bar{u}^\tau(i,\alpha)c_i^\tau\\
&= \sum_{(i,\alpha)\in[n]\times\mathcal{A}} \mathbb{I}\left((i,\alpha)\in\mathcal{W}^\tau\right)^2(u(i,\alpha)/u(\mathcal{W}^\tau))c_i^\tau\\
&= \sum_{(i,\alpha)\in[n]\times\mathcal{A}} \mathbb{I}\left((i,\alpha)\in\mathcal{W}^\tau\right)(u(i,\alpha)/u(\mathcal{W}^\tau))c_i^\tau\\
&= m\sum_{(i,\alpha)\in[n]\times\mathcal{A}} \mathbb{I}\left((i,\alpha)\in\mathcal{W}^\tau\right)u(i,\alpha)c_i^\tau\\
&= m\sum_{(i,\alpha)\in\mathcal{W}^\tau} u(i,\alpha)c_i^\tau\\
&= m\sum_{(i,\alpha)\in[n]\times\mathcal{A}:\alpha\in\widetilde{\mathcal{W}}^\tau} u(i,\alpha)c_i^\tau\\
&= m\sum_{i\in[n]}\sum_{\alpha\in\widetilde{\mathcal{W}}^\tau} u(i,\alpha)c_i^\tau\\
&= m\sum_{i\in[n]}\sum_{\alpha\in\widetilde{\mathcal{W}}^\tau} \frac{1}{m}\mathbb{I}\left(i\in\mathbb{M}\ \wedge\ \alpha=\hat{\alpha}^i\right)c_i^\tau\\
&= \sum_{i\in[n]}\sum_{\alpha\in\widetilde{\mathcal{W}}^\tau} \mathbb{I}\left(i\in\mathbb{M}\ \wedge\ \alpha=\hat{\alpha}^i\right)c_i^\tau\\
&= \sum_{i\in[n]}\sum_{\alpha\in\widetilde{\mathcal{W}}^\tau} \mathbb{I}\left(i\in\mathbb{M}\right)\mathbb{I}\left(\alpha=\hat{\alpha}^i\right)c_i^\tau\\
&= \sum_{i\in[n]}\mathbb{I}\left(i\in\mathbb{M}\right)\sum_{\alpha\in\widetilde{\mathcal{W}}^\tau} \mathbb{I}\left(\alpha=\hat{\alpha}^i\right)c_i^\tau\\
&= \sum_{i\in\mathbb{M}}\sum_{\alpha\in\widetilde{\mathcal{W}}^\tau} \mathbb{I}\left(\alpha=\hat{\alpha}^i\right)c_i^\tau\\
&= \sum_{i\in\mathbb{M}}\mathbb{I}\left(\hat{\alpha}^i\in\widetilde{\mathcal{W}}^\tau\right)c_i^\tau\\
&= \sum_{i\in\mathbb{M}}\mathbb{I}\left(i=z_\tau\right)c_i^\tau\\
&= c_{z_\tau}^\tau
\end{aligned}
$$

We also have that:

$$
\begin{aligned}
&\langle\tilde{v}^\tau,\tilde{c}^\tau\rangle\\
&= \sum_{\xi\in\mathcal{W}^\tau}\tilde{v}^\tau(\xi)\tilde{c}^\tau(\xi)\\
&= \sum_{(i,\alpha)\in[n]\times\mathcal{A}:(i,\alpha)\in\mathcal{W}^\tau}\tilde{v}^\tau(i,\alpha)\tilde{c}^\tau(i,\alpha)\\
&= \sum_{(i,\alpha)\in[n]\times\mathcal{A}:(i,\alpha)\in\mathcal{W}^\tau}\tilde{v}^\tau(i,\alpha)c_i^\tau\\
&= \sum_{i\in[n]}\sum_{\alpha\in\mathcal{A}:(i,\alpha)\in\mathcal{W}^\tau}\tilde{v}^\tau(i,\alpha)c_i^\tau
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{i\in[n]} c_i^\tau \sum_{\alpha\in\mathcal{A}:(i,\alpha)\in\mathcal{W}^\tau} \tilde{v}^\tau(i,\alpha) \\
&= \sum_{i\in[n]} c_i^\tau v_i^\tau \\
&= \boldsymbol{c}^\tau \cdot \boldsymbol{v}^\tau \\
&= c_*^\tau
\end{aligned}
$$

Utilising these two equations, along with that of Lemma 25 gives us:

$$
\begin{aligned}
\sum_{\tau\in[T]} u(\mathcal{W}^\tau)\langle \tilde{v}^\tau - \bar{u}^\tau, \tilde{c}^\tau\rangle &= \sum_{\tau\in[T]} \frac{1}{m}\langle \tilde{v}^\tau - \bar{u}^\tau, \tilde{c}^\tau\rangle \\
&= \frac{1}{m}\sum_{\tau\in[T]} (\langle \tilde{v}^\tau, \tilde{c}^\tau\rangle - \langle \bar{u}^\tau, \tilde{c}^\tau\rangle) \\
&= \frac{1}{m}\sum_{\tau\in[T]} (c_*^\tau - c_{z_\tau}^\tau)
\end{aligned}
$$

□

From Lemma 26 and Theorem 5 we have:

$$
\begin{aligned}
\frac{1}{m}\sum_{\tau\in[T]} (c_*^\tau - c_{z_\tau}^\tau) &= \sum_{\tau\in[T]} u(\mathcal{W}^\tau)\langle \tilde{v}^\tau - \bar{u}^\tau, \tilde{c}^\tau\rangle \\
&\leq \frac{1}{\eta}\,\mathrm{KL}(u,\rho^*) + \frac{\eta}{2}\sum_{\tau\in[T]} u(\mathcal{W}^\tau)
\end{aligned}
$$

Applying Lemma 25 and multiplying through by m then gives us:

$$
\sum_{\tau\in[T]} (c_*^\tau - c_{z_\tau}^\tau) \leq \frac{m}{\eta}\,\mathrm{KL}(u,\rho^*) + \frac{\eta}{2}T
$$

Applying Theorem 8 then gives us the result: ■

A.2.6 Theorem 10

We start with the following lemma:

Lemma 27. *For all $\tau \in [T]$ and $\boldsymbol{\varphi} \in \Phi$ with $\varphi_{\ell^\tau} = \psi$ we have $\gamma_i^\tau(\boldsymbol{\varphi}) = 0$*

Proof. Suppose, for contradiction, that we have a circadian instance $\alpha \in \mathcal{A}$ with $\alpha(\tau) = \boldsymbol{\varphi}$ and $q(\alpha) > 0$. Then since $\varphi_{\ell^\tau} = \psi$ and, as $\lambda'(\alpha) \neq 0$, $\alpha(1)_{\ell^\tau} \neq \psi$ there exists $\tau' \in [\tau-1]$ with $\alpha(\tau')_{\ell^\tau} \neq \psi$ and $\alpha(\tau'+1)_{\ell^\tau} = \psi$, so choose such a τ'. Since $q(\alpha) > 0$ we must have $\lambda^{\tau'}(\alpha(\tau'), \alpha(\tau'+1)) > 0$. We have two cases:

- If $\ell^{\tau'} \neq \ell^\tau$ then since $\alpha(\tau')_{\ell^\tau} \neq \alpha(\tau'+1)_{\ell^\tau}$ we have, from the definition of the Multitask Markov circadian, that $\lambda^{\tau'}(\alpha(\tau'), \alpha(\tau'+1) = 0$ which is a contradiction.
- If $\ell^{\tau'} = \ell^\tau$ then since $\tau > \tau'$ we have $\tau' \notin \bar{\Omega}$ and hence, by the definition of the Multitask Markov circadian, we have, since $\lambda^{\tau'}(\alpha(\tau'), \alpha(\tau'+1)) > 0$, that $\alpha(\tau'+1)_{\ell^{\tau'}} \neq \psi$ and hence that $\alpha(\tau'+1)_{\ell^\tau} = \alpha(\tau'+1)_{\ell^{\tau'}} \neq \psi$ which is a contradiction.

So either way is a contradiction and hence we have shown that for all $\alpha \in \mathcal{A}$ with $\alpha(\tau) = \boldsymbol{\varphi}$, it is the case that $q(\alpha) = 0$ and hence that $\rho^1(i,\alpha) = \rho^*(i,\alpha) = 0$. By a simple induction on τ'' we then have, from the Specialist hedge algorithm, that $\rho^{\tau''}(i,\alpha) = 0$ for all $\tau'' \in [T]$. This implies:

$$
\gamma_i^\tau(\boldsymbol{\varphi}) := \sum_{\alpha\in\mathcal{A}:\alpha(\tau)=\boldsymbol{\varphi}} \rho^\tau(i,\alpha) = 0
$$

□

We first consider the case that $\tau \in \bar{\Omega}$. From Theorem 7 and Lemma 27 we have:

$$
\begin{aligned}
\gamma_i^{\tau+1}(\boldsymbol{\varphi}) &= \frac{Y^\tau}{Z^\tau} \sum_{\boldsymbol{\varphi}' \in \widetilde{\mathcal{W}}^\tau} \lambda^\tau(\boldsymbol{\varphi}', \boldsymbol{\varphi}) \gamma_i^\tau(\boldsymbol{\varphi}') \exp(-\eta c_i^\tau) + \sum_{\boldsymbol{\varphi}' \in \Phi \setminus \widetilde{\mathcal{W}}^\tau} \lambda^\tau(\boldsymbol{\varphi}', \boldsymbol{\varphi}) \gamma_i^\tau(\boldsymbol{\varphi}') \\
&= \frac{Y^\tau}{Z^\tau} \sum_{\boldsymbol{\varphi}' \in \Phi :\ \varphi'_{\ell^\tau} = 1} \lambda^\tau(\boldsymbol{\varphi}', \boldsymbol{\varphi}) \gamma_i^\tau(\boldsymbol{\varphi}') \exp(-\eta c_i^\tau) + \sum_{\boldsymbol{\varphi}' \in \Phi :\ \varphi'_{\ell^\tau} \in \{0, \psi\}} \lambda^\tau(\boldsymbol{\varphi}', \boldsymbol{\varphi}) \gamma_i^\tau(\boldsymbol{\varphi}') \\
&= \frac{Y^\tau}{Z^\tau} \sum_{\boldsymbol{\varphi}' \in \Phi :\ \varphi'_{\ell^\tau} = 1} \lambda^\tau(\boldsymbol{\varphi}', \boldsymbol{\varphi}) \gamma_i^\tau(\boldsymbol{\varphi}') \exp(-\eta c_i^\tau) + \sum_{\boldsymbol{\varphi}' \in \Phi :\ \varphi'_{\ell^\tau} = 0} \lambda^\tau(\boldsymbol{\varphi}', \boldsymbol{\varphi}) \gamma_i^\tau(\boldsymbol{\varphi}') \\
&= \frac{Y^\tau}{Z^\tau} \sum_{\boldsymbol{\varphi}' \in \Phi :\ \varphi'_{\ell^\tau} = 1} \mathbb{I}\left(\varphi_{\ell^\tau} = \psi\right)\ \mathbb{I}\left(\forall j \in [s] \setminus \{\ell^\tau\},\ \varphi_j = \varphi'_j\right) \gamma_i^\tau(\boldsymbol{\varphi}') \exp(-\eta c_i^\tau) \\
&\quad + \sum_{\boldsymbol{\varphi}' \in \Phi :\ \varphi'_{\ell^\tau} = 0} \mathbb{I}\left(\varphi_{\ell^\tau} = \psi\right)\ \mathbb{I}\left(\forall j \in [s] \setminus \{\ell^\tau\},\ \varphi_j = \varphi'_j\right) \gamma_i^\tau(\boldsymbol{\varphi}') \\
&= \frac{Y^\tau}{Z^\tau} \sum_{\boldsymbol{\varphi}' \in \Phi} \mathbb{I}\left(\varphi_{\ell^\tau} = \psi\right)\ \mathbb{I}\left(\varphi'_{\ell^\tau} = 1\ \wedge\ \forall j \in [s] \setminus \{\ell^\tau\},\ \varphi_j = \varphi'_j\right) \gamma_i^\tau(\boldsymbol{\varphi}') \exp(-\eta c_i^\tau) \\
&\quad + \sum_{\boldsymbol{\varphi}' \in \Phi} \mathbb{I}\left(\varphi_{\ell^\tau} = \psi\right)\ \mathbb{I}\left(\varphi'_{\ell^\tau} = 0\ \wedge\ \forall j \in [s] \setminus \{\ell^\tau\},\ \varphi_j = \varphi'_j\right) \gamma_i^\tau(\boldsymbol{\varphi}') \\
&= \frac{Y^\tau}{Z^\tau} \sum_{\boldsymbol{\varphi}' \in \Phi} \mathbb{I}\left(\boldsymbol{\varphi}' = \boldsymbol{\varphi}^{\tau|1}\right) \gamma_i^\tau(\boldsymbol{\varphi}') \exp(-\eta c_i^\tau) + \sum_{\boldsymbol{\varphi}' \in \Phi} \mathbb{I}\left(\varphi_{\ell^\tau} = \psi\right)\ \mathbb{I}\left(\boldsymbol{\varphi}' = \boldsymbol{\varphi}^{\tau|0}\right) \gamma_i^\tau(\boldsymbol{\varphi}') \\
&= \frac{Y^\tau}{Z^\tau} \mathbb{I}\left(\varphi_{\ell^\tau} = \psi\right) \gamma_i^\tau(\boldsymbol{\varphi}^{\tau|1}) \exp(-\eta c_i^\tau) + \mathbb{I}\left(\varphi_{\ell^\tau} = \psi\right) \gamma_i^\tau(\boldsymbol{\varphi}^{\tau|0}) \\
&= \mathbb{I}\left(\varphi_{\ell^\tau} = \psi\right) \left(\gamma_i^\tau(\boldsymbol{\varphi}^{\tau|0}) + \gamma_i^\tau(\boldsymbol{\varphi}^{\tau|1}) \exp(-\eta c_i^\tau) \frac{Y^\tau}{Z^\tau} \right)
\end{aligned}
$$

We now consider the case that $\tau \notin \bar{\Omega}$. Define x and y as follows:

- If $\varphi_{\ell^\tau} = 0$ then $x := 1 - \theta$ and $y := 1 - \phi$.
- If $\varphi_{\ell^\tau} = 1$ then $x := \theta$ and $y := \phi$.
- If $\varphi_{\ell^\tau} = \psi$ that $x := 0$ and $y := 0$

From Theorem 7 and Lemma 27 we have:

$$
\begin{aligned}
\gamma_i^{\tau+1}(\boldsymbol{\varphi}) &= \frac{Y^\tau}{Z^\tau} \sum_{\boldsymbol{\varphi}' \in \widetilde{\mathcal{W}}^\tau} \lambda^\tau(\boldsymbol{\varphi}', \boldsymbol{\varphi}) \gamma_i^\tau(\boldsymbol{\varphi}') \exp(-\eta c_i^\tau) + \sum_{\boldsymbol{\varphi}' \in \Phi \setminus \widetilde{\mathcal{W}}^\tau} \lambda^\tau(\boldsymbol{\varphi}', \boldsymbol{\varphi}) \gamma_i^\tau(\boldsymbol{\varphi}') \\
&= \frac{Y^\tau}{Z^\tau} \sum_{\boldsymbol{\varphi}' \in \Phi :\ \varphi'_{\ell^\tau} = 1} \lambda^\tau(\boldsymbol{\varphi}', \boldsymbol{\varphi}) \gamma_i^\tau(\boldsymbol{\varphi}') \exp(-\eta c_i^\tau) + \sum_{\boldsymbol{\varphi}' \in \Phi :\ \varphi'_{\ell^\tau} \in \{0, 1\}} \lambda^\tau(\boldsymbol{\varphi}', \boldsymbol{\varphi}) \gamma_i^\tau(\boldsymbol{\varphi}') \\
&= \frac{Y^\tau}{Z^\tau} \sum_{\boldsymbol{\varphi}' \in \Phi :\ \varphi'_{\ell^\tau} = 1} \lambda^\tau(\boldsymbol{\varphi}', \boldsymbol{\varphi}) \gamma_i^\tau(\boldsymbol{\varphi}') \exp(-\eta c_i^\tau) + \sum_{\boldsymbol{\varphi}' \in \Phi :\ \varphi'_{\ell^\tau} = 0} \lambda^\tau(\boldsymbol{\varphi}', \boldsymbol{\varphi}) \gamma_i^\tau(\boldsymbol{\varphi}') \\
&= \frac{Y^\tau}{Z^\tau} \sum_{\boldsymbol{\varphi}' \in \Phi :\ \varphi'_{\ell^\tau} = 1} x\, \mathbb{I}\left(\forall j \in [s] \setminus \{\ell^\tau\},\ \varphi_j = \varphi'_j\right) \gamma_i^\tau(\boldsymbol{\varphi}') \exp(-\eta c_i^\tau) \\
&\quad + \sum_{\boldsymbol{\varphi}' \in \Phi :\ \varphi'_{\ell^\tau} = 0} y\, \mathbb{I}\left(\forall j \in [s] \setminus \{\ell^\tau\},\ \varphi_j = \varphi'_j\right) \gamma_i^\tau(\boldsymbol{\varphi}') \\
&= \frac{Y^\tau}{Z^\tau} \sum_{\boldsymbol{\varphi}' \in \Phi} x\, \mathbb{I}\left(\varphi'_{\ell^\tau} = 1\ \wedge\ \forall j \in [s] \setminus \{\ell^\tau\},\ \varphi_j = \varphi'_j\right) \gamma_i^\tau(\boldsymbol{\varphi}') \exp(-\eta c_i^\tau)
\end{aligned}
$$

$$+\sum_{\boldsymbol{\varphi}'\in\Phi} y\,\mathbb{I}\left(\varphi'_{\ell^\tau}=0\ \wedge\ \forall j\in[s]\setminus\{\ell^\tau\},\ \varphi_j=\varphi'_j\right)\gamma_i^\tau(\boldsymbol{\varphi}')$$
$$=\frac{Y^\tau}{Z^\tau}\sum_{\boldsymbol{\varphi}'\in\Phi} x\,\mathbb{I}\left(\boldsymbol{\varphi}'=\boldsymbol{\varphi}^{\tau|1}\right)\gamma_i^\tau(\boldsymbol{\varphi}')\exp(-\eta c_i^\tau)+\sum_{\boldsymbol{\varphi}'\in\Phi} y\,\mathbb{I}\left(\boldsymbol{\varphi}'=\boldsymbol{\varphi}^{\tau|0}\right)\gamma_i^\tau(\boldsymbol{\varphi}')$$
$$=\frac{Y^\tau}{Z^\tau}x\gamma_i^\tau(\boldsymbol{\varphi}^{\tau|1})\exp(-\eta c_i^\tau)+y\gamma_i^\tau(\boldsymbol{\varphi}^{\tau|0})$$
$$=y\gamma_i^\tau(\boldsymbol{\varphi}^{\tau|0})+x\gamma_i^\tau(\boldsymbol{\varphi}^{\tau|1})\exp(-\eta c_i^\tau)\frac{Y^\tau}{Z^\tau}$$

■

A.2.7 Theorem 11

We take the inductive hypothesis:

$$\gamma_i^\tau(\boldsymbol{\varphi})=\pi_i^\tau\prod_{j\in[s]}\left(\bar{w}_i^{\tau,j}\mathbb{I}\left(\varphi_j=\varpi_j^\tau(1)\right)+(1-\bar{w}_i^{\tau,j})\mathbb{I}\left(\varphi_j=\varpi_j^\tau(0)\right)\right)$$

and prove it by induction on τ. We first consider the case that $\tau=1$. In this case we have, from Theorem 7 and definition of λ', that:

$$\gamma_i^1(\boldsymbol{\varphi})=\frac{1}{n}\lambda'(\boldsymbol{\varphi})$$
$$=\frac{1}{n}\prod_{j\in[s]}\left(\mu\mathbb{I}\left(\varphi_j=1\right)+(1-\mu)\,\mathbb{I}\left(\varphi_j=0\right)\right)$$
$$=\pi_i^1\prod_{j\in[s]}\left(\mu\mathbb{I}\left(\varphi_j=1\right)+(1-\mu)\,\mathbb{I}\left(\varphi_j=0\right)\right)$$
$$=\pi_i^1\prod_{j\in[s]}\left(\bar{w}_i^{1,j}\mathbb{I}\left(\varphi_j=1\right)+(1-\bar{w}_i^{1,j})\mathbb{I}\left(\varphi_j=0\right)\right)$$
$$=\pi_i^1\prod_{j\in[s]}\left(\bar{w}_i^{1,j}\mathbb{I}\left(\varphi_j=\varpi_j^\tau(1)\right)+(1-\bar{w}_i^{1,j})\mathbb{I}\left(\varphi_j=\varpi_j^\tau(0)\right)\right)$$

So the inductive hypothesis holds for $\tau=1$. Now suppose the inductive hypothesis holds for $\tau=\tau'$ (for some $\tau\in[T-1]$). We now show it holds for $\tau=\tau'+1$. To do this suppose that $\boldsymbol{\delta}$, $\boldsymbol{\epsilon}$ and β are as created in Algorithm 1 during the update of trial τ' and let $\boldsymbol{w}:=\boldsymbol{w}_{\boldsymbol{\sigma}(\tau')}^{\ell^{\tau'}}$, noting that this is equal to $\bar{\boldsymbol{w}}^{\tau',\ell^{\tau'}}$ We start with the following lemmas:

Lemma 28. *For all $\tau\in[T]$ and all $j\in[s]$ we have:*

$$\sum_{\varphi_j\in\{0,1,\psi\}}\left(\bar{w}_i^{\tau',j}\mathbb{I}\left(\varphi_j=\varpi_j^\tau(1)\right)+(1-\bar{w}_i^{\tau',j+1})\mathbb{I}\left(\varphi_j=\varpi_j^\tau(0)\right)\right)=1$$

Proof. If $\tau\in\Omega_j$ then $\varpi_j^\tau(0)=\varpi_j^\tau(1)=\psi$ and if $\tau\notin\Omega_j$ then $\varpi_j^\tau(0)=0$ and $\varpi_j^\tau(0)=1$. In either case we then have that:

$$\sum_{\varphi_j\in\{0,1,\psi\}}\left(\bar{w}_i^{\tau',j}\mathbb{I}\left(\varphi_j=\varpi_j^\tau(1)\right)+(1-\bar{w}_i^{\tau',j+1})\mathbb{I}\left(\varphi_j=\varpi_j^\tau(0)\right)\right)=\bar{w}_i^{\tau',j}+(1-\bar{w}_i^{\tau',j+1})=1$$

□

Lemma 29. *We have:*

$$\sum_{\boldsymbol{\varphi}\in\widetilde{\mathcal{W}}^{\tau'}}\gamma_i^{\tau'}(\boldsymbol{\varphi})=\pi_i^{\tau'}\bar{w}_i^{\tau',\ell^{\tau'}}$$

Proof. Without loss of generality assume $\ell^{\tau'}=1$. For all $\chi\in[s]$ we take the inductive hypothesis:

$$\sum_{\boldsymbol{\varphi}\in\{0,1,\psi\}^\chi:\ \varphi_1=1}\pi_i^{\tau'}\prod_{j\in[\chi]}\left(\bar{w}_i^{\tau',j}\mathbb{I}\left(\varphi_j=\varpi_j^{\tau'}(1)\right)+(1-\bar{w}_i^{\tau',j})\mathbb{I}\left(\varphi_j=\varpi_j^{\tau'}(0)\right)\right)=\pi_i^{\tau'}\bar{w}_i^{\tau',1}$$

This is since:

$$
\begin{aligned}
\sum_{\boldsymbol{\varphi}\in\widetilde{\mathcal{W}}^{\tau'}} \gamma_i^{\tau'}(\boldsymbol{\varphi}) &= \sum_{\boldsymbol{\varphi}\in\Phi:\ \varphi_{\ell^{\tau'}}=1} \gamma_i^{\tau'}(\boldsymbol{\varphi})\\
&= \sum_{\boldsymbol{\varphi}\in\Phi:\ \varphi_1=1} \gamma_i^{\tau'}(\boldsymbol{\varphi})\\
&= \sum_{\boldsymbol{\varphi}\in\Phi:\ \varphi_1=1} \pi_i^{\tau'} \prod_{j\in[s]} \left(\bar{w}_i^{\tau',j}\mathbb{I}\left(\varphi_j=\varpi_j^{\tau'}(1)\right)+(1-\bar{w}_i^{\tau',j})\mathbb{I}\left(\varphi_j=\varpi_j^{\tau'}(0)\right)\right)\\
&= \sum_{\boldsymbol{\varphi}\in\{0,1,\psi\}^s:\ \varphi_1=1} \pi_i^{\tau'} \prod_{j\in[s]} \left(\bar{w}_i^{\tau',j}\mathbb{I}\left(\varphi_j=\varpi_j^{\tau'}(1)\right)+(1-\bar{w}_i^{\tau',j})\mathbb{I}\left(\varphi_j=\varpi_j^{\tau'}(0)\right)\right)
\end{aligned}
$$

For $\chi=1$ we have $\ell^{\tau'}=1$ so $1=\tau'\notin\Omega_1$ and hence $\varpi_j^{\tau'}(0)=0$ and $\varpi_j^{\tau'}(1)=1$. Hence, for $\chi=1$ we have:

$$
\begin{aligned}
&\sum_{\boldsymbol{\varphi}\in\{0,1,\psi\}^\chi:\ \varphi_1=1} \pi_i^{\tau'} \prod_{j\in[\chi]} \left(\bar{w}_i^{\tau',j}\mathbb{I}\left(\varphi_j=\varpi_j^{\tau'}(1)\right)+(1-\bar{w}_i^{\tau',j})\mathbb{I}\left(\varphi_j=\varpi_j^{\tau'}(0)\right)\right)\\
&= \sum_{\boldsymbol{\varphi}\in\{0,1,\psi\}^\chi:\ \varphi_1=1} \pi_i^{\tau'} \prod_{j\in[\chi]} \left(\bar{w}_i^{\tau',j}\mathbb{I}\left(\varphi_j=1\right)+(1-\bar{w}_i^{\tau',j})\mathbb{I}\left(\varphi_j=0\right)\right)\\
&= \sum_{\boldsymbol{\varphi}\in\{0,1,\psi\}^1:\ \varphi_1=1} \pi_i^{\tau'} \prod_{j\in[1]} \left(\bar{w}_i^{\tau',j}\mathbb{I}\left(\varphi_j=1\right)+(1-\bar{w}_i^{\tau',j})\mathbb{I}\left(\varphi_j=0\right)\right)\\
&= \sum_{\boldsymbol{\varphi}\in\{0,1,\psi\}^1:\ \varphi_1=1} \pi_i^{\tau'} \left(\bar{w}_i^{\tau',1}\mathbb{I}\left(\varphi_1=1\right)+(1-\bar{w}_i^{\tau',1})\mathbb{I}\left(\varphi_1=0\right)\right)\\
&= \pi_i^{\tau'}\left(\bar{w}_i^{\tau',1}1+(1-\bar{w}_i^{\tau',1})0\right)\\
&= \pi_i^{\tau'}\bar{w}_i^{\tau',1}
\end{aligned}
$$

so the inductive hypothesis holds for $\chi=1$. Now suppose it holds for $\chi=\chi'$ (for some $\chi'\in[s]$). We will show that it also holds for $\chi=\chi'+1$. Specifically we have, by Lemma 28 and the inductive hypothesis:

$$
\begin{aligned}
&\sum_{\boldsymbol{\varphi}\in\{0,1,\psi\}^{\chi'+1}:\ \varphi_1=1} \pi_i^{\tau'} \prod_{j\in[\chi'+1]} \left(\bar{w}_i^{\tau',j}\mathbb{I}\left(\varphi_j=\varpi_j^{\tau'}(1)\right)+(1-\bar{w}_i^{\tau',j})\mathbb{I}\left(\varphi_j=\varpi_j^{\tau'}(0)\right)\right)\\
&= \sum_{\varphi_{\chi'+1}\in\{0,1,\psi\}} \sum_{\boldsymbol{\varphi}\in\{0,1\}^{\chi'}:\ \varphi_1=1} \pi_i^{\tau'} \prod_{j\in[\chi'+1]} \left(\bar{w}_i^{\tau',j}\mathbb{I}\left(\varphi_j=\varpi_j^{\tau'}(1)\right)+(1-\bar{w}_i^{\tau',j})\mathbb{I}\left(\varphi_j=\varpi_j^{\tau'}(0)\right)\right)\\
&= \left(\sum_{\varphi_{\chi'+1}\in\{0,1,\psi\}} \left(\bar{w}_i^{\tau',\chi'+1}\mathbb{I}\left(\varphi_{\chi'+1}=\varpi_j^{\tau'}(1)\right)+(1-\bar{w}_i^{\tau',\chi'+1})\mathbb{I}\left(\varphi_{\chi'+1}=\varpi_j^{\tau'}(0)\right)\right)\right)\\
&\quad\times \pi_i^{\tau'} \prod_{j\in[\chi']} \left(\bar{w}_i^{\tau',j}\mathbb{I}\left(\varphi_j=\varpi_j^{\tau'}(1)\right)+(1-\bar{w}_i^{\tau',j})\mathbb{I}\left(\varphi_j=\varpi_j^{\tau'}(0)\right)\right)\\
&= \left(\sum_{\varphi_{\chi'+1}\in\{0,1,\psi\}} \left(\bar{w}_i^{\tau',\chi'+1}\mathbb{I}\left(\varphi_{\chi'+1}=\varpi_j^{\tau'}(1)\right)+(1-\bar{w}_i^{\tau',\chi'+1})\mathbb{I}\left(\varphi_{\chi'+1}=\varpi_j^{\tau'}(0)\right)\right)\right)\pi_i^{\tau'}\bar{w}_i^{\tau',1}\\
&= \pi_i^{\tau'}\bar{w}_i^{\tau',1}
\end{aligned}
$$

were the last equality is due to Lemma 28. This proves the inductive hypothesis. □

Lemma 30. *We have:*

$$
\frac{Y^{\tau'}}{Z^{\tau'}}=\beta
$$

Proof. From Theorem 7 and Lemma 29 we have:

$$
\begin{aligned}
Y^{\tau'} &= \sum_{i\in[n]} \sum_{\boldsymbol{\varphi}\in\widetilde{\mathcal{W}}^{\tau'}} \gamma_i^{\tau'}(\boldsymbol{\varphi}) \\
&= \sum_{i\in[n]} \pi_i^{\tau} \bar{w}_i^{\tau',\ell^{\tau'}} \\
&= \sum_{i\in[n]} \pi_i^{\tau} w_i \\
&= \boldsymbol{\pi}^{\tau} \cdot \boldsymbol{w}
\end{aligned}
$$

and we have:

$$
\begin{aligned}
Z^{\tau'} &= \sum_{i\in[n]} \exp(-\eta c_i^{\tau'}) \sum_{\varphi\in\widetilde{\mathcal{W}}^{\tau}} \gamma_i^{\tau}(\varphi) \\
&= \sum_{i\in[n]} \exp(-\eta c_i^{\tau'})\pi_i^{\tau'} \bar{w}_i^{\tau',\ell^{\tau'}} \\
&= \sum_{i\in[n]} \exp(-\eta c_i^{\tau'})\pi_i^{\tau'} w_i \\
&= \sum_{i\in[n]} \pi_i^{\tau'} \left(\exp(-\eta c_i^{\tau'}) w_i\right) \\
&= \sum_{i\in[n]} \pi_i^{\tau'} \delta_i \\
&= \boldsymbol{\pi}^{\tau'} \cdot \boldsymbol{\delta}
\end{aligned}
$$

so:

$$
\begin{aligned}
\frac{Y^{\tau'}}{Z^{\tau'}} &= \frac{\boldsymbol{\pi}^{\tau'} \cdot \boldsymbol{w}}{\boldsymbol{\pi}^{\tau'} \cdot \boldsymbol{\delta}} \\
&= \beta
\end{aligned}
$$

□

Definition 31. *Let us define $\bar{\gamma}$ as:*

$$
\bar{\gamma} := \pi_i^{\tau'} \prod_{j\in[s]\setminus\ell^{\tau'}} \left(\bar{w}_i^{\tau',j}\mathbb{I}\left(\varphi_j = \varpi_j^{\tau'}(1)\right) + (1-\bar{w}_i^{\tau',j})\mathbb{I}\left(\varphi_j = \varpi_j^{\tau'}(0)\right)\right)
$$

Lemma 32. *We have:*

$$
\gamma_i^{\tau'}(\boldsymbol{\varphi}^{\tau'|0}) = \bar{\gamma}(1-w_i)
$$

and:

$$
\gamma_i^{\tau'}(\boldsymbol{\varphi}^{\tau'|1}) = \bar{\gamma}w_i
$$

Proof. Take any $b \in \{0,1\}$. Since $\tau' \notin \Omega_{\ell^{\tau'}}$ we have $\varpi_{\ell^{\tau'}}^{\tau'}(1) = 1$ and $\varpi_{\ell^{\tau'}}^{\tau'}(0) = 0$. From this result and the inductive hypothesis we have:

$$
\begin{aligned}
\gamma_i^{\tau'}(\boldsymbol{\varphi}^{\tau'|b}) &= \pi_i^{\tau'} \prod_{j\in[s]} \left(\bar{w}_i^{\tau',j}\mathbb{I}\left(\varphi_j^{\tau'|b} = \varpi_j^{\tau'}(1)\right) + (1-\bar{w}_i^{\tau',j})\mathbb{I}\left(\varphi_j^{\tau'|b} = \varpi_j^{\tau'}(0)\right)\right) \\
&= \pi_i^{\tau'} \left(\bar{w}_i^{\tau',\ell^{\tau'}}\mathbb{I}\left(\varphi_{\ell^{\tau'}}^{\tau'|b} = \varpi_{\ell^{\tau'}}^{\tau'}(1)\right) + (1-\bar{w}_i^{\tau',\ell^{\tau'}})\mathbb{I}\left(\varphi_{\ell^{\tau'}}^{\tau'|b} = \varpi_{\ell^{\tau}}^{\tau'}(0)\right)\right) \\
&\quad \times \prod_{j\in[s]\setminus\{\ell^{\tau}\}} \left(\bar{w}_i^{\tau',j}\mathbb{I}\left(\varphi_j^{\tau'|b} = \varpi_j^{\tau'}(1)\right) + (1-\bar{w}_i^{\tau',j})\mathbb{I}\left(\varphi_j^{\tau'|b} = \varpi_j^{\tau'}(0)\right)\right) \\
&= \pi_i^{\tau'} \left(\bar{w}_i^{\tau',\ell^{\tau'}}\mathbb{I}\left(\varphi_{\ell^{\tau'}}^{\tau'|b} = 1\right) + (1-\bar{w}_i^{\tau',\ell^{\tau'}})\mathbb{I}\left(\varphi_{\ell^{\tau'}}^{\tau'|b} = 0\right)\right)
\end{aligned}
$$

$$
\begin{aligned}
&\times \prod_{j\in[s]\setminus\{\ell^\tau\}} \left(\bar{w}_i^{\tau',j}\mathbb{I}\left(\varphi_j^{\tau'|b}=\varpi_j^{\tau'}(1)\right)+(1-\bar{w}_i^{\tau',j})\mathbb{I}\left(\varphi_j^{\tau'|b}=\varpi_j^{\tau'}(0)\right)\right)\\
&=\pi_i^{\tau'}\left(\bar{w}_i^{\tau',\ell^{\tau'}}\mathbb{I}\left(b=1\right)+(1-\bar{w}_i^{\tau',\ell^{\tau'}})\mathbb{I}\left(b=0\right)\right)\\
&\times \prod_{j\in[s]\setminus\{\ell^\tau\}} \left(\bar{w}_i^{\tau',j}\mathbb{I}\left(\varphi_j^{\tau'|b}=\varpi_j^{\tau'}(1)\right)+(1-\bar{w}_i^{\tau',j})\mathbb{I}\left(\varphi_j^{\tau'|b}=\varpi_j^{\tau'}(0)\right)\right)\\
&=\pi_i^{\tau'}\left(\bar{w}_i^{\tau',\ell^{\tau'}}\mathbb{I}\left(b=1\right)+(1-\bar{w}_i^{\tau',\ell^{\tau'}})\mathbb{I}\left(b=0\right)\right)\\
&\times \prod_{j\in[s]\setminus\{\ell^\tau\}} \left(\bar{w}_i^{\tau',j}\mathbb{I}\left(\varphi_j=\varpi_j^{\tau'}(1)\right)+(1-\bar{w}_i^{\tau',j})\mathbb{I}\left(\varphi_j=\varpi_j^{\tau'}(0)\right)\right)\\
&=\bar{\gamma}\left(\bar{w}_i^{\tau',\ell^{\tau'}}\mathbb{I}\left(b=1\right)+(1-\bar{w}_i^{\tau',\ell^{\tau'}})\mathbb{I}\left(b=0\right)\right)\\
&=\bar{\gamma}\left(w_i\mathbb{I}\left(b=1\right)+(1-w_i)\mathbb{I}\left(b=0\right)\right)
\end{aligned}
$$

□

Lemma 33. *We have:*

$$\bar{w}_i^{\tau'+1,\ell^{\tau'}}=(\phi(1-w_i)+\theta\beta\delta_i)/\epsilon_i$$

and for all $j\in[s]\setminus\{\ell^{\tau'}\}$ we have:

$$\bar{w}_i^{\tau'+1,j}=\bar{w}_i^{\tau',j}$$

Proof. Direct from Algorithm 1 and the definition of $\bar{w}_i^{\tau'+1,j}$ (for all $j\in[s]$). □

Lemma 34. *We have:*

$$\gamma_i^{\tau'+1}(\boldsymbol{\varphi})=\bar{\gamma}\epsilon_i\left(\bar{w}_i^{\tau'+1,\ell^{\tau'}}\mathbb{I}\left(\varphi_{\ell^\tau}=\varpi_j^{\tau'}(1)\right)+(1-\bar{w}_i^{\tau'+1,\ell^{\tau'}})\mathbb{I}\left(\varphi_{\ell^\tau}=\varpi_j^{\tau'}(0)\right)\right)$$

Proof. We first consider the case that $\tau'\in\bar{\Omega}$. Note that since $\tau'\in\bar{\Omega}$ we have $\ell^{\tau''}\neq\ell^{\tau'}$ for all $\tau''>\tau'$ and hence $\tau'\in\Omega_{\ell^{\tau'}}$ so we have $\varpi_{\ell^{\tau'}}^{\tau'}(1)=\varpi_{\ell^{\tau'}}^{\tau'}(0)=\psi$. Combing this result with Theorem 10, Lemma 30, Lemma 32 and Lemma 33 we have:

$$
\begin{aligned}
\gamma_i^{\tau'+1}(\boldsymbol{\varphi})&=\mathbb{I}\left(\varphi_{\ell^\tau}=\psi\right)\left(\gamma_i^\tau(\boldsymbol{\varphi}^{\tau|0})+\gamma_i^\tau(\boldsymbol{\varphi}^{\tau|1})\exp(-\eta c_i^\tau)\frac{Y^\tau}{Z^\tau}\right)\\
&=\mathbb{I}\left(\varphi_{\ell^\tau}=\psi\right)\left(\gamma_i^\tau(\boldsymbol{\varphi}^{\tau|0})+\gamma_i^\tau(\boldsymbol{\varphi}^{\tau|1})\exp(-\eta c_i^\tau)\beta\right)\\
&=\mathbb{I}\left(\varphi_{\ell^\tau}=\psi\right)\left(\bar{\gamma}(1-w_i)+\bar{\gamma}w_i\exp(-\eta c_i^\tau)\beta\right)\\
&=\left(\bar{w}_i^{\tau'+1,\ell^{\tau'}}+\left(1-\bar{w}_i^{\tau'+1,\ell^{\tau'}}\right)\right)\mathbb{I}\left(\varphi_{\ell^\tau}=\psi\right)\left(\bar{\gamma}(1-w_i)+\bar{\gamma}w_i\exp(-\eta c_i^\tau)\beta\right)\\
&=\left(\bar{w}_i^{\tau'+1,\ell^{\tau'}}\mathbb{I}\left(\varphi_{\ell^\tau}=\psi\right)+(1-\bar{w}_i^{\tau'+1,\ell^{\tau'}})\mathbb{I}\left(\varphi_{\ell^\tau}=\psi\right)\right)\bar{\gamma}\left((1-w_i)+w_i\exp(-\eta c_i^\tau)\beta\right)\\
&=\left(\bar{w}_i^{\tau'+1,\ell^{\tau'}}\mathbb{I}\left(\varphi_{\ell^\tau}=\varpi_{\ell^{\tau'}}^{\tau'}(1)\right)+(1-\bar{w}_i^{\tau'+1,\ell^{\tau'}})\mathbb{I}\left(\varphi_{\ell^\tau}=\varpi_{\ell^{\tau'}}^{\tau'}(0)\right)\right)\bar{\gamma}\left((1-w_i)+w_i\exp(-\eta c_i^\tau)\beta\right)\\
&=\left(\bar{w}_i^{\tau'+1,\ell^{\tau'}}\mathbb{I}\left(\varphi_{\ell^\tau}=\varpi_{\ell^{\tau'}}^{\tau'}(1)\right)+(1-\bar{w}_i^{\tau'+1,\ell^{\tau'}})\mathbb{I}\left(\varphi_{\ell^\tau}=\varpi_{\ell^{\tau'}}^{\tau'}(0)\right)\right)\bar{\gamma}\left((1-w_i)+\delta_i\right)\\
&=\bar{\gamma}\epsilon_i\left(\bar{w}_i^{\tau'+1,\ell^{\tau'}}\mathbb{I}\left(\varphi_{\ell^\tau}=\varpi_{\ell^{\tau'}}^{\tau'}(1)\right)+(1-\bar{w}_i^{\tau'+1,\ell^{\tau'}})\mathbb{I}\left(\varphi_{\ell^\tau}=\varpi_{\ell^{\tau'}}^{\tau'}(0)\right)\right)
\end{aligned}
$$

We next consider the case that $\tau'\notin\bar{\Omega}$. We have two subcases. We first consider the subcase that $\varphi_{\ell^{\tau'}}=0$. From Theorem 10, Lemma 30, Lemma 32 and Lemma 33 we have:

$$
\begin{aligned}
\gamma_i^{\tau'+1}(\boldsymbol{\varphi}) &= (1-\phi)\gamma_i^{\tau'}(\boldsymbol{\varphi}^{\tau|0}) + (1-\theta)\gamma_i^{\tau'}(\boldsymbol{\varphi}^{\tau'|1})\exp(-\eta c_i^{\tau'})\frac{Y^{\tau'}}{Z^{\tau'}} \\
&= (1-\phi)\gamma_i^{\tau'}(\boldsymbol{\varphi}^{\tau|0}) + (1-\theta)\gamma_i^{\tau'}(\boldsymbol{\varphi}^{\tau'|1})\exp(-\eta c_i^{\tau'})\beta \\
&= (1-\phi)\bar{\gamma}(1-w_i) + (1-\theta)\bar{\gamma}w_i\exp(-\eta c_i^{\tau'})\beta \\
&= \bar{\gamma}\left((1-\phi)(1-w_i) + (1-\theta)w_i\exp(-\eta c_i^{\tau'})\beta\right) \\
&= \bar{\gamma}\left((1-\phi)(1-w_i) + (1-\theta)\delta_i\beta\right) \\
&= \bar{\gamma}\left((1-w_i) + \delta_i\beta - (\phi(1-w_i) + \theta\delta_i\beta)\right) \\
&= \bar{\gamma}\left((1-w_i+\delta_i\beta) - \bar{w}_i^{\tau'+1,\ell^{\tau'}}\epsilon_i\right) \\
&= \bar{\gamma}(\epsilon_i - \bar{w}_i^{\tau'+1,\ell^{\tau'}}\epsilon_i) \\
&= \bar{\gamma}\epsilon_i(1-\bar{w}_i^{\tau'+1,\ell^{\tau'}})
\end{aligned}
$$

We next consider the subcase that $\varphi_{\ell^{\tau'}} = 1$. From Theorem 10, Lemma 30, Lemma 32 and Lemma 33 we have:

$$
\begin{aligned}
\gamma_i^{\tau'+1}(\boldsymbol{\varphi}) &= \phi\gamma_i^{\tau'}(\boldsymbol{\varphi}^{\tau|0}) + \theta\gamma_i^{\tau'}(\boldsymbol{\varphi}^{\tau'|1})\exp(-\eta c_i^{\tau'})\frac{Y^{\tau'}}{Z^{\tau'}} \\
&= \phi\gamma_i^{\tau'}(\boldsymbol{\varphi}^{\tau|0}) + \theta\gamma_i^{\tau'}(\boldsymbol{\varphi}^{\tau'|1})\exp(-\eta c_i^{\tau'})\beta \\
&= \phi\bar{\gamma}(1-w_i) + \theta\bar{\gamma}w_i\exp(-\eta c_i^{\tau'})\beta \\
&= \bar{\gamma}\left(\phi(1-w_i) + \theta w_i\exp(-\eta c_i^{\tau'})\beta\right) \\
&= \bar{\gamma}\left(\phi(1-w_i) + \theta\delta_i\beta\right) \\
&= \bar{\gamma}\bar{w}_i^{\tau'+1,\ell^{\tau'}}\epsilon_i
\end{aligned}
$$

Putting together gives us:

$$
\gamma_i^{\tau'+1}(\boldsymbol{\varphi}) = \bar{\gamma}\epsilon_i\left(\bar{w}_i^{\tau'+1,\ell^{\tau'}}\mathbb{I}(\varphi_{\ell^{\tau'}}=1) + (1-\bar{w}_i^{\tau'+1,\ell^{\tau'}})\mathbb{I}(\varphi_{\ell^{\tau'}}=0)\right)
$$

Since $\tau' \notin \bar{\Omega}$ there exists $\tau'' > \tau'$ with $\ell^{\tau''} = \ell^{\tau'}$ and hence we have $\tau' \notin \Omega_{\ell^{\tau'}}$ so $1 = \varpi_{\ell^{\tau'}}^{\tau'}(1)$ and $0 = \varpi_{\ell^{\tau'}}^{\tau'}(0)$. The result follows.

□

By Lemma 33 and the update of $\pi_i^{\cdot}$ we have:

$$
\begin{aligned}
\epsilon_i\bar{\gamma} &= (\epsilon_i\pi_i^{\tau'})\prod_{j\in[s]\setminus\ell^{\tau'}}\left(\bar{w}_i^{\tau',j}\mathbb{I}\left(\varphi_j = \varpi_j^{\tau'}(1)\right) + (1-\bar{w}_i^{\tau',j})\mathbb{I}\left(\varphi_j = \varpi_j^{\tau'}(0)\right)\right) \\
&= \pi_i^{\tau'+1}\prod_{j\in[s]\setminus\ell^{\tau'}}\left(\bar{w}_i^{\tau',j}\mathbb{I}\left(\varphi_j = \varpi_j^{\tau'}(1)\right) + (1-\bar{w}_i^{\tau',j})\mathbb{I}\left(\varphi_j = \varpi_j^{\tau'}(0)\right)\right) \\
&= \pi_i^{\tau'+1}\prod_{j\in[s]\setminus\ell^{\tau'}}\left(\bar{w}_i^{\tau'+1,j}\mathbb{I}\left(\varphi_j = \varpi_j^{\tau'}(1)\right) + (1-\bar{w}_i^{\tau'+1,j})\mathbb{I}\left(\varphi_j = \varpi_j^{\tau'}(0)\right)\right)
\end{aligned}
$$

and hence, by Lemma 34:

$$\begin{aligned}
\gamma_i^{\tau'+1}(\boldsymbol{\varphi}) &= \bar{\gamma}\epsilon_i \left(\bar{w}_i^{\tau'+1,\ell^{\tau'}} \mathbb{I}\left(\varphi_{\ell^{\tau'}} = \varpi_j^{\tau'}(1)\right) + (1-\bar{w}_i^{\tau'+1,\ell^{\tau'}})\mathbb{I}\left(\varphi_{\ell^{\tau'}} = \varpi_j^{\tau'}(0)\right)\right) \\
&= \pi_i^{\tau'+1} \prod_{j\in[s]\setminus\ell^{\tau'}} \left(\bar{w}_i^{\tau'+1,j}\mathbb{I}\left(\varphi_j = \varpi_j^{\tau'}(1)\right) + (1-\bar{w}_i^{\tau'+1,j})\mathbb{I}\left(\varphi_j = \varpi_j^{\tau'}(0)\right)\right) \\
&\quad \times \left(\bar{w}_i^{\tau'+1,\ell^{\tau'}} \mathbb{I}\left(\varphi_{\ell^{\tau'}} = \varpi_j^{\tau'}(1)\right) + (1-\bar{w}_i^{\tau'+1,\ell^{\tau'}})\mathbb{I}\left(\varphi_{\ell^{\tau'}} = \varpi_j^{\tau'}(0)\right)\right) \\
&= \pi_i^{\tau'+1} \prod_{j\in[s]} \left(\bar{w}_i^{\tau'+1,j}\mathbb{I}\left(\varphi_j = \varpi_j^{\tau'}(1)\right) + (1-\bar{w}_i^{\tau'+1,j})\mathbb{I}\left(\varphi_j = \varpi_j^{\tau'}(0)\right)\right)
\end{aligned}$$

which proves the inductive hypothesis. Lemma 29 completes the proof.

A.2.8 Theorem 12

By Theorem 11 we have that:

$$\sum_{\boldsymbol{\varphi}\in\widetilde{\mathcal{W}}^\tau} \gamma_i^\tau(\boldsymbol{\varphi}) = \pi_i^\tau \bar{w}_i^{\tau,\boldsymbol{\sigma}(\tau)}$$

so, by Theorem 7, we have:

$$Y^\tau = \sum_{i\in[n]} \sum_{\boldsymbol{\varphi}\in\widetilde{\mathcal{W}}^\tau} \gamma_i^\tau(\boldsymbol{\varphi}) = \sum_{i\in[n]} \pi_i^\tau \bar{w}_i^{\tau,\boldsymbol{\sigma}(\tau)} = \boldsymbol{\pi}^\tau \cdot \bar{\boldsymbol{w}}^{\tau,\boldsymbol{\sigma}(\tau)}$$

which, by Theorem 7, implies that the vector component v_i^τ defined in Section A.1.5 is equal to:

$$v_i^\tau = \frac{1}{Y^\tau} \sum_{\boldsymbol{\varphi}\in\widetilde{\mathcal{W}}^\tau} \gamma_i^\tau(\boldsymbol{\varphi}) = (\pi_i^\tau \bar{w}_i^{\tau,\boldsymbol{\sigma}(\tau)})/(\boldsymbol{\pi}^\tau \cdot \bar{\boldsymbol{w}}^{\tau,\boldsymbol{\sigma}(\tau)})$$

Since $\bar{\boldsymbol{w}}^{\tau,\boldsymbol{\sigma}(\tau)} = \boldsymbol{w}_{\boldsymbol{\sigma}(\tau)}^{\ell^\tau}$ this is also the vector component played by Algorithm 1 on trial τ.

■

A.2.9 Theorem 13

Since $\hat{\alpha}^i(1)_j = \mathbb{I}\left(i = z_{\bar{\nu}_j(1)}\right)$ we have:

$$\begin{aligned}
\ln(\lambda'(\hat{\alpha}^i(1))) &= \ln\left(\prod_{j\in[s]} \left(\mu\mathbb{I}\left(\hat{\alpha}^i(1)_j = 1\right) + (1-\mu)\,\mathbb{I}\left(\hat{\alpha}^i(1)_j = 0\right)\right)\right) \\
&= \ln\left(\prod_{j\in[s]} \left(\mu\mathbb{I}\left(\mathbb{I}\left(i = z_{\bar{\nu}_j(1)}\right) = 1\right) + (1-\mu)\,\mathbb{I}\left(\mathbb{I}\left(i = z_{\bar{\nu}_j(1)}\right) = 0\right)\right)\right) \\
&= \ln\left(\prod_{j\in[s]} \left(\mu\mathbb{I}\left(i = z_{\bar{\nu}_j(1)}\right) + (1-\mu)\,\mathbb{I}\left(i \neq z_{\bar{\nu}_j(1)}\right)\right)\right) \\
&= \sum_{j\in[s]} \ln\left(\mu\mathbb{I}\left(i = z_{\bar{\nu}_j(1)}\right) + (1-\mu)\,\mathbb{I}\left(i \neq z_{\bar{\nu}_j(1)}\right)\right) \\
&= \sum_{j\in[s]} \left(\mathbb{I}\left(i = z_{\bar{\nu}_j(1)}\right)\ln(\mu) + \mathbb{I}\left(i \neq z_{\bar{\nu}_j(1)}\right)\ln(1-\mu)\right)
\end{aligned}$$

We have the following lemmas:

Lemma 35. *For all $\tau \in [T] \setminus \bar{\Omega}$ we have:*

$$\hat{\alpha}^i(\tau)_{\ell^\tau} = \mathbb{I}\left(i = z_\tau\right)$$

Proof. By definition of $\hat{\alpha}^i(\tau)_{\ell^\tau}$ and since $\bar{\nu}_{\ell^\tau}(\tau) = \tau$ we have:

$$\hat{\alpha}^i(\tau)_{\ell^\tau} = \mathbb{I}\left(z_{\bar{\nu}_{\ell^\tau}(\tau)} = i\right) = \mathbb{I}\left(i = z_\tau\right)$$

□

Lemma 36. *For all $\tau \in [T] \setminus \bar{\Omega}$ we have:*

$$\hat{\alpha}^i(\tau+1)_{\ell^\tau} = \mathbb{I}\left(i = z_{\nu(\tau)}\right)$$

Proof. Note that since $\tau \notin \bar{\Omega}$ we have that there exists $\tau' \in [T]$ with $\tau' > \tau$ and $\ell^{\tau'} = \ell^\tau$ so $\tau + 1 \notin \Omega_{\ell^\tau}$ and hence:

$$\hat{\alpha}^i(\tau+1)_{\ell^\tau} = \mathbb{I}\left(z_{\bar{\nu}_{\ell^\tau}(\tau+1)} = i\right)$$

By definition of $\bar{\nu}_{\ell^\tau}(\tau+1)$ we have

$$\begin{aligned}
\bar{\nu}_{\ell^\tau}(\tau+1) &= \min\{\tau' \in [T] \mid \tau' \geq \tau + 1 \ \wedge\ \ell^{\tau'} = \ell^\tau\} \\
&= \min\{\tau' \in [T] \mid \tau' > \tau \ \wedge\ \ell^{\tau'} = \ell^\tau\} \\
&= \nu(\tau)
\end{aligned}$$

Putting together we have:

$$\hat{\alpha}^i(\tau+1)_{\ell^\tau} = \mathbb{I}\left(z_{\bar{\nu}_{\ell^\tau}(\tau+1)} = i\right) = \mathbb{I}\left(z_{\nu(\tau)} = i\right)$$

□

Lemma 37. *For all $\tau \in [T-1]$ we have:*

$$\mathbb{I}\left(\forall j \in [s] \setminus \{\ell^\tau\},\ \hat{\alpha}^i(\tau)_j = \hat{\alpha}^i(\tau+1)_j\right) = 1$$

Proof. Suppose we have some $j \in [s] \setminus \{\ell^\tau\}$. We have two cases. First consider the case that $\tau \notin \Omega_j$. Then since $\ell^\tau \neq j$ we have:

$$\begin{aligned}
\bar{\nu}_j(\tau) &= \min\{\tau' \in [T] \mid \tau' \geq \tau \ \wedge\ \ell^{\tau'} = j\} \\
&= \min\{\tau' \in [T] \mid \tau' \geq \tau + 1 \ \wedge\ \ell^{\tau'} = j\} \\
&= \bar{\nu}_j(\tau+1)
\end{aligned}$$

Hence we have

$$\hat{\alpha}^i(\tau)_j = \mathbb{I}\left(z_{\bar{\nu}_j(\tau)} = i\right) = \mathbb{I}\left(z_{\bar{\nu}_j(\tau+1)} = i\right) = \hat{\alpha}^i(\tau+1)_j$$

Next consider the case that $\tau \in \Omega_j$. Then we immediately have, by definition of Ω_j, that $\tau + 1 \in \Omega_j$. Therefore we have:

$$\hat{\alpha}^i(\tau)_j = \psi = \hat{\alpha}^i(\tau+1)_j$$

So in either case we have:

$$\hat{\alpha}^i(\tau)_j = \hat{\alpha}^i(\tau+1)_j$$

The result follows. □

Lemma 38. *For all $\tau \in \bar{\Omega}$ we have:*

$$\ln\left(\lambda^\tau(\hat{\alpha}^i(\tau), \hat{\alpha}^i(\tau+1))\right) = 1$$

Proof. Since $\tau \in \bar{\Omega}$ there does not exist $\tau' \in [T]$ with $\tau' \geq \tau + 1$ and $\ell^{\tau'} = \ell^\tau$. Hence we have that $\tau + 1 \in \Omega_{\ell^\tau}$ so $\hat{\alpha}^i(\tau+1)_{\ell^\tau} = \psi$. By the definition of the Multitask circadian and Lemma 37 we then have:

$$\begin{aligned}
\lambda^\tau(\hat{\alpha}^i(\tau), \hat{\alpha}^i(\tau+1)) &= \mathbb{I}\left(\hat{\alpha}^i(\tau+1)_{\ell^\tau} = \psi\right) \mathbb{I}\left(\forall j \in [s] \setminus \{\ell^\tau\},\ \hat{\alpha}^i(\tau+1)_j = \hat{\alpha}^i(\tau)_j\right) \\
&= \mathbb{I}\left(\forall j \in [s] \setminus \{\ell^\tau\},\ \hat{\alpha}^i(\tau+1)_j = \hat{\alpha}^i(\tau)_j\right) \\
&= 1
\end{aligned}$$

□

Lemma 39. *For all $\tau \in [T] \setminus \bar{\Omega}$ we have the following:*

- *If $z_\tau = i \wedge z_{\nu(\tau)} = i$ then $\lambda^\tau(\hat{\alpha}^i(\tau), \hat{\alpha}^i(\tau+1)) = \theta$*
- *If $z_\tau \neq i \wedge z_{\nu(\tau)} \neq i$ then $\lambda^\tau(\hat{\alpha}^i(\tau), \hat{\alpha}^i(\tau+1)) = 1 - \phi$*
- *If $z_\tau \neq i \wedge z_{\nu(\tau)} = i$ then $\lambda^\tau(\hat{\alpha}^i(\tau), \hat{\alpha}^i(\tau+1)) = \phi$*
- *If $z_\tau = i \wedge z_{\nu(\tau)} \neq i$ then $\lambda^\tau(\hat{\alpha}^i(\tau), \hat{\alpha}^i(\tau+1)) = 1 - \theta$*

Proof. We only explicitly prove the first item - the rest are proved in exactly the same way. If $z_\tau = i \wedge z_{\nu(\tau)} = i$ then we have, by Lemma 35, that $\hat{\alpha}^i(\tau)_{\ell^\tau} = 1$ and, by Lemma 36, that $\hat{\alpha}^i(\tau+1)_{\ell^\tau} = 1$. By definition of the Multitask Markov circadian this implies that $\lambda^\tau(\hat{\alpha}^i(\tau), \hat{\alpha}^i(\tau+1)) = \theta \mathbb{I}\left(\forall j \in [s] \setminus \{\ell^\tau\},\ \hat{\alpha}^i(\tau)_j = \hat{\alpha}^i(\tau+1)_j\right)$ which, by Lemma 37 is equal to θ □

Combing Lemma 38 and Lemma 39 with the equation derived at the start of the proof we have the result.

■

A.2.10 Theorem 14

By Theorem 9 we have:

$$\sum_{\tau \in [T]} (c_*^\tau - c_{z_\tau}^\tau) \leq \frac{\eta}{2} T + \frac{1}{\eta} C$$

where

$$C = m \ln \left(\frac{n}{m}\right) - \sum_{i \in \mathbb{M}} \ln(\lambda'(\hat{\alpha}^i(1))) - \sum_{i \in \mathbb{M}} \sum_{\tau \in [T-1]} \ln(\lambda^\tau(\hat{\alpha}^i(\tau), \hat{\alpha}^i(\tau+1)))$$

We now compute the value of C. We start with the following lemmas:

Lemma 40. *We have:*

$$\sum_{i \in \mathbb{M}} \sum_{j \in [s]} \left(\mathbb{I}\left(i = z_{\bar{\nu}_j(1)}\right) \ln(\mu) + \mathbb{I}\left(i \neq z_{\bar{\nu}_j(1)}\right) \ln(1-\mu)\right) = s\left(\ln(\mu) + (m-1)\ln(1-\mu)\right)$$

Proof. Given $j \in [s]$ we have:

$$\begin{aligned}\sum_{i \in \mathbb{M}} \mathbb{I}\left(i = z_{\bar{\nu}_j(1)}\right) \ln(\mu) &= \sum_{i = z_{\bar{\nu}_j(1)}} \mathbb{I}\left(i = z_{\bar{\nu}_j(1)}\right) \ln(\mu) + \sum_{i \in \mathbb{M} \setminus \{z_{\bar{\nu}_j(1)}\}} \mathbb{I}\left(i = z_{\bar{\nu}_j(1)}\right) \ln(\mu) \\ &= \sum_{i = z_{\bar{\nu}_j(1)}} \ln(\mu) + \sum_{i \in \mathbb{M} \setminus \{z_{\bar{\nu}_j(1)}\}} 0 \\ &= \ln(\mu)\end{aligned}$$

and:

$$\begin{aligned}\sum_{i \in \mathbb{M}} \mathbb{I}\left(i \neq z_{\bar{\nu}_j(1)}\right) \ln(1-\mu) &= \sum_{i = z_{\bar{\nu}_j(1)}} \mathbb{I}\left(i \neq z_{\bar{\nu}_j(1)}\right) \ln(1-\mu) + \sum_{i \in \mathbb{M} \setminus \{z_{\bar{\nu}_j(1)}\}} \mathbb{I}\left(i \neq z_{\bar{\nu}_j(1)}\right) \ln(1-\mu) \\ &= \sum_{i = z_{\bar{\nu}_j(1)}} 0 + \sum_{i \in \mathbb{M} \setminus \{z_{\bar{\nu}_j(1)}\}} \ln(1-\mu) \\ &= (m-1)\ln(1-\mu)\end{aligned}$$

Hence, we have:

$$\sum_{i \in \mathbb{M}} \sum_{j \in [s]} \left(\mathbb{I}\left(i = z_{\bar{\nu}_j(1)}\right) \ln(\mu) + \mathbb{I}\left(i \neq z_{\bar{\nu}_j(1)}\right) \ln(1-\mu)\right)$$

$$
\begin{aligned}
&= \sum_{j\in[s]} \sum_{i\in\mathbb{M}} \left(\mathbb{I}\left(i = z_{\bar{\nu}_j(1)}\right) \ln(\mu) + \mathbb{I}\left(i \neq z_{\bar{\nu}_j(1)}\right) \ln(1-\mu)\right) \\
&= \sum_{j\in[s]} \left(\ln(\mu) + (m-1)\ln(1-\mu)\right) \\
&= s\left(\ln(\mu) + (m-1)\ln(1-\mu)\right)
\end{aligned}
$$

□

Lemma 41. *Given $\tau \in [T] \setminus \bar{\Omega}$ we have:*

- $\sum_{i\in\mathbb{M}} \mathbb{I}\left(z_\tau = i \ \wedge \ z_{\nu(\tau)} = i\right) = \mathbb{I}\left(z_{\nu(\tau)} = z_\tau\right)$
- $\sum_{i\in\mathbb{M}} \mathbb{I}\left(z_\tau = i \ \wedge \ z_{\nu(\tau)} \neq i\right) = \mathbb{I}\left(z_{\nu(\tau)} \neq z_\tau\right)$
- $\sum_{i\in\mathbb{M}} \mathbb{I}\left(z_\tau \neq i \ \wedge \ z_{\nu(\tau)} = i\right) = \mathbb{I}\left(z_{\nu(\tau)} \neq z_\tau\right)$
- $\sum_{i\in\mathbb{M}} \mathbb{I}\left(z_\tau \neq i \ \wedge \ z_{\nu(\tau)} \neq i\right) = n - 1 - \mathbb{I}\left(z_{\nu(\tau)} \neq z_\tau\right)$

Proof. We have:

$$
\begin{aligned}
\sum_{i\in\mathbb{M}} \mathbb{I}\left(z_\tau = i \ \wedge \ z_{\nu(\tau)} = i\right) &= \sum_{i=z_\tau} \mathbb{I}\left(z_\tau = i \ \wedge \ z_{\nu(\tau)} = i\right) + \sum_{i\in\mathbb{M}\setminus\{z_\tau\}} \mathbb{I}\left(z_\tau = i \ \wedge \ z_{\nu(\tau)} = i\right) \\
&= \sum_{i=z_\tau} \mathbb{I}\left(z_\tau = i \ \wedge \ z_{\nu(\tau)} = i\right) + \sum_{i\in\mathbb{M}\setminus\{z_\tau\}} 0 \\
&= \mathbb{I}\left(z_\tau = z_\tau \ \wedge \ z_{\nu(\tau)} = z_\tau\right) \\
&= \mathbb{I}\left(z_{\nu(\tau)} = z_\tau\right)
\end{aligned}
$$

and:

$$
\begin{aligned}
\sum_{i\in\mathbb{M}} \mathbb{I}\left(z_\tau = i \ \wedge \ z_{\nu(\tau)} \neq i\right) &= \sum_{i=z_\tau} \mathbb{I}\left(z_\tau = i \ \wedge \ z_{\nu(\tau)} \neq i\right) + \sum_{i\in\mathbb{M}\setminus\{z_\tau\}} \mathbb{I}\left(z_\tau = i \ \wedge \ z_{\nu(\tau)} \neq i\right) \\
&= \sum_{i=z_\tau} \mathbb{I}\left(z_\tau = i \ \wedge \ z_{\nu(\tau)} \neq i\right) + \sum_{i\in\mathbb{M}\setminus\{z_\tau\}} 0 \\
&= \mathbb{I}\left(z_\tau = z_\tau \ \wedge \ z_{\nu(\tau)} \neq z_\tau\right) \\
&= \mathbb{I}\left(z_{\nu(\tau)} \neq z_\tau\right)
\end{aligned}
$$

and:

$$
\begin{aligned}
\sum_{i\in\mathbb{M}} \mathbb{I}\left(z_\tau \neq i \ \wedge \ z_{\nu(\tau)} = i\right) &= \sum_{i=z_{\nu(\tau)}} \mathbb{I}\left(z_\tau \neq i \ \wedge \ z_{\nu(\tau)} = i\right) + \sum_{i\in\mathbb{M}\setminus\{z_{\nu(\tau)}\}} \mathbb{I}\left(z_\tau \neq i \ \wedge \ z_{\nu(\tau)} = i\right) \\
&= \sum_{i=z_{\nu(\tau)}} \mathbb{I}\left(z_\tau \neq i \ \wedge \ z_{\nu(\tau)} = i\right) + \sum_{i\in\mathbb{M}\setminus\{z_{\nu(\tau)}\}} 0 \\
&= \mathbb{I}\left(z_\tau \neq z_{\nu(\tau)} \ \wedge \ z_{\nu(\tau)} = z_{\nu(\tau)}\right) \\
&= \mathbb{I}\left(z_\tau \neq z_{\nu(\tau)}\right)
\end{aligned}
$$

and:

$$
\begin{aligned}
\sum_{i\in\mathbb{M}} \mathbb{I}\left(z_\tau \neq i \ \wedge \ z_{\nu(\tau)} \neq i\right) &= \sum_{i\in\{z_\tau, z_{\nu(\tau)}\}} \mathbb{I}\left(z_\tau \neq i \ \wedge \ z_{\nu(\tau)} \neq i\right) + \sum_{i\in\mathbb{M}\setminus\{z_\tau, z_{\nu(\tau)}\}} \mathbb{I}\left(z_\tau \neq i \ \wedge \ z_{\nu(\tau)} \neq i\right) \\
&= \sum_{i\in\{z_\tau, z_{\nu(\tau)}\}} 0 + \sum_{i\in\mathbb{M}\setminus\{z_\tau, z_{\nu(\tau)}\}} 1 \\
&= |\mathbb{M} \setminus \{z_\tau, z_{\nu(\tau)}\}| \\
&= m - |\{z_\tau, z_{\nu(\tau)}\}| \\
&= m - 1 - \mathbb{I}\left(z_\tau \neq z_{\nu(\tau)}\right)
\end{aligned}
$$

□

Lemma 42. *We have:*

$$\sum_{\tau \in [T] \setminus \bar{\Omega}} \mathbb{I}\left(z_{\nu(\tau)} \neq z_\tau\right) = k$$

Proof. Direct from the definition of k and $\nu(\cdot)$. □

Lemma 43. *We have* $|\bar{\Omega}| = s$

Proof. For every task $j \in [s]$, the number $\max\{\tau \in [T] \mid \ell^\tau = j\}$ is the unique $\tau \in [T]$ such that $\ell^\tau = j$ and for all $\tau' > \tau$ we have $\ell^{\tau'} = j$. The result follows. □

Lemma 44. *We have:*

- $\sum_{i \in \mathbb{M}} \sum_{\tau \in [T] \setminus \bar{\Omega}} \mathbb{I}\left(z_\tau = i \ \wedge \ z_{\nu(\tau)} = i\right) = T - s - k$
- $\sum_{i \in \mathbb{M}} \sum_{\tau \in [T] \setminus \bar{\Omega}} \mathbb{I}\left(z_\tau = i \ \wedge \ z_{\nu(\tau)} \neq i\right) = k$
- $\sum_{i \in \mathbb{M}} \sum_{\tau \in [T] \setminus \bar{\Omega}} \mathbb{I}\left(z_\tau \neq i \ \wedge \ z_{\nu(\tau)} = i\right) = k$
- $\sum_{i \in \mathbb{M}} \sum_{\tau \in [T] \setminus \bar{\Omega}} \mathbb{I}\left(z_\tau \neq i \ \wedge \ z_{\nu(\tau)} \neq i\right) = (T - s)(n - 1) - k$

Proof. From Lemma 41, Lemma 42 and Lemma 43 we have:

$$\begin{aligned}
\sum_{i \in \mathbb{M}} \sum_{\tau \in [T] \setminus \bar{\Omega}} \mathbb{I}\left(z_\tau = i \ \wedge \ z_{\nu(\tau)} = i\right) &= \sum_{\tau \in [T] \setminus \bar{\Omega}} \mathbb{I}\left(z_{\nu(\tau)} = z_\tau\right) \\
&= \sum_{\tau \in [T] \setminus \bar{\Omega}} \left(1 - \mathbb{I}\left(z_{\nu(\tau)} \neq z_\tau\right)\right) \\
&= (T - |\bar{\Omega}|) - \sum_{\tau \in [T] \setminus \bar{\Omega}} \mathbb{I}\left(z_{\nu(\tau)} \neq z_\tau\right) \\
&= (T - |\bar{\Omega}|) - k \\
&= (T - s) - k
\end{aligned}$$

and:

$$\sum_{i \in \mathbb{M}} \sum_{\tau \in [T] \setminus \bar{\Omega}} \mathbb{I}\left(z_\tau = i \ \wedge \ z_{\nu(\tau)} \neq i\right) = \sum_{\tau \in [T] \setminus \bar{\Omega}} \mathbb{I}\left(z_{\nu(\tau)} \neq z_\tau\right) = k$$

and:

$$\sum_{i \in \mathbb{M}} \sum_{\tau \in [T] \setminus \bar{\Omega}} \mathbb{I}\left(z_\tau \neq i \ \wedge \ z_{\nu(\tau)} = i\right) = \sum_{\tau \in [T] \setminus \bar{\Omega}} \mathbb{I}\left(z_{\nu(\tau)} \neq z_\tau\right) = k$$

and:

$$\begin{aligned}
\sum_{i \in \mathbb{M}} \sum_{\tau \in [T] \setminus \bar{\Omega}} \mathbb{I}\left(z_\tau \neq i \ \wedge \ z_{\nu(\tau)} \neq i\right) &= \sum_{\tau \in [T] \setminus \bar{\Omega}} \left(m - 1 - \mathbb{I}\left(z_{\nu(\tau)} \neq z_\tau\right)\right) \\
&= (T - |\bar{\Omega}|)(m - 1) - \sum_{\tau \in [T] \setminus \bar{\Omega}} \mathbb{I}\left(z_{\nu(\tau)} \neq z_\tau\right)) \\
&= (T - |\bar{\Omega}|)(m - 1) - k \\
&= (T - s)(m - 1) - k
\end{aligned}$$

□

From Theorem 13 we have:

$$\sum_{i \in \mathbb{M}} \ln(\lambda'(\hat{\alpha}^i(1))) + \sum_{i \in \mathbb{M}} \sum_{\tau \in [T-1]} \ln(\lambda^\tau(\hat{\alpha}^i(\tau), \hat{\alpha}^i(\tau + 1)))$$

$$
\begin{aligned}
&= \sum_{i \in \mathbb{M}} \left(\ln(\lambda'(\hat{\alpha}^i(1))) + \sum_{\tau \in [T-1]} \ln(\lambda^\tau(\hat{\alpha}^i(\tau), \hat{\alpha}^i(\tau+1))) \right) \\
&= \sum_{i \in \mathbb{M}} \sum_{j \in [s]} \left(\mathbb{I}\left(i = z_{\bar{\nu}_j(1)}\right) \ln(\mu) + \mathbb{I}\left(i \neq z_{\bar{\nu}_j(1)}\right) \ln(1-\mu) \right) \\
&\quad + \ln(\theta) \sum_{i \in \mathbb{M}} \sum_{\tau \in [T] \setminus \bar{\Omega}} \mathbb{I}\left(z_\tau = i \ \wedge \ z_{\nu(\tau)} = i\right) \\
&\quad + \ln(1-\phi) \sum_{i \in \mathbb{M}} \sum_{\tau \in [T] \setminus \bar{\Omega}} \mathbb{I}\left(z_\tau \neq i \ \wedge \ z_{\nu(\tau)} \neq i\right) \\
&\quad + \ln(1-\theta) \sum_{i \in \mathbb{M}} \sum_{\tau \in [T] \setminus \bar{\Omega}} \mathbb{I}\left(z_\tau = i \ \wedge \ z_{\nu(\tau)} \neq i\right) \\
&\quad + \ln(\phi) \sum_{i \in \mathbb{M}} \sum_{\tau \in [T] \setminus \bar{\Omega}} \mathbb{I}\left(z_\tau \neq i \ \wedge \ z_{\nu(\tau)} = i\right)
\end{aligned}
$$

So by Lemma 40 and Lemma 44 we then have:

$$
\begin{aligned}
&\sum_{i \in \mathbb{M}} \ln(\lambda'(\hat{\alpha}^i(1))) + \sum_{i \in \mathbb{M}} \sum_{\tau \in [T-1]} \ln(\lambda^\tau(\hat{\alpha}^i(\tau), \hat{\alpha}^i(\tau+1))) \\
=& s\left(\ln(\mu) + (m-1)\ln(1-\mu)\right) + \ln(\theta)(T-s-k) + \ln(1-\phi)((T-s)(m-1)-k) + \ln(1-\theta)k + \ln(\phi)k
\end{aligned}
$$

Substituting into the definition of C in Theorem 9 gives us the result

A.2.11 Theorem 15

We start with the following lemma:

Lemma 45. *Given $x, y \in \mathbb{R}^+$, if we set $z := \frac{x}{x+y}$ then:*

$$
-x\log(z) - y\log(1-z) = (x+y)H\left(\frac{x}{x+y}\right)
$$

and

$$
-x\log(z) - y\log(1-z) \leq x\log\left(\frac{x+y}{x}\right) + x.
$$

Proof. We have $x = (x+y)z$ and $y = (x+y)(1-z)$ so

$$
\begin{aligned}
-x\log(z) - y\log(1-z) &= (x+y)(-z\log(z) - (1-z)\log(1-z)) \\
&= (x+y)H(z) \qquad (26) \\
&= (x+y)H\left(\frac{x}{x+y}\right).
\end{aligned}
$$

We recall the standard inequality:

$$
\frac{1}{z}H(z) \leq \log(1/z) + 1.
$$

Plugging this into Equation (26) gives us:

$$
-x\log(z) - y\log(1-z) \leq (x+y)(z\log(1/z) + z).
$$

Substituting in the value of z then gives us the result. ◻

The next lemma utilizes the inequality in Lemma 45 to give us the values of certain quantities when μ, θ and ϕ are tuned.

Lemma 46. *Setting $\mu := 1/m$, $\theta := 1 - k/(T-s)$ and $\phi := k/((m-1)(T-s))$ we have:*

1. $$-\log(\mu) - (m-1)\log(1-\mu) = mH\left(\frac{1}{m}\right) \leq \log(m) + 1$$

2. $$-(T-s-k)\log(\theta) - k\log(1-\theta) = (T-s)H\left(\frac{k}{T-s}\right) \leq k\log\left(\frac{T-s}{k}\right) + k$$

3. $$-k\log(\phi) + ((m-1)(T-s)-k)\log(1-\phi) = (m-1)(T-s)H\left(\frac{k}{(m-1)(T-s)}\right) \leq k\log\left(\frac{(m-1)(T-s)}{k}\right) + k$$

Proof. All three items follow from Lemma 45 with x, y and z defined, for each item, as follows:

1. $z := \mu = 1/m$, $x := 1$ and $y := m - 1$
2. $z := 1 - \theta = k/(T-s)$, $x := k$ and $y := T - s - k$
3. $z := \phi = k/((m-1)(T-s))$, $x := k$ and $y := (m-1)(T-s) - k$.

□

Substituting the inequalities of Lemma 46 into the definition of C in Theorem 14 gives us the result.

■

A.2.12 Theorem 16

We have:

$$\begin{aligned} c_*^\tau &= \boldsymbol{v}^\tau \cdot \boldsymbol{c}^\tau \\ &= \sum_{i\in[n]} v_i^\tau c_i^\tau \\ &= \sum_{i\in[n]} \mathbb{P}(i^\tau = i) c_i^\tau \\ &= \sum_{i\in[n]} \mathbb{P}(i^\tau = i) \mathcal{L}_{01}(y^\tau, [\kappa(i)](x_t)) \\ &= \mathbb{E}\left(\mathcal{L}_{01}(y^\tau, [\kappa(i^\tau)](x_t))\right) \\ &= \mathbb{E}\left(\mathcal{L}_{01}(y^\tau, \hat{y}^\tau)\right) \end{aligned}$$

Noting that $\mathcal{L}_{01}(y_t^i, h_t^i(x_t^i)) = \mathcal{L}_{01}(y_t^i, [\kappa(z_t^i)](x_t^i)) = c_{t,z_t^i}^i$ then gives us the result.

■

B Proofs for Section 4

We prove Theorem 3 and give a proof sketch of Proposition 4 in this section. We first provide a brief overview of the proof of Theorem 3 and a discussion of Theorem 47, which is a key result in the proof of the theorem.

Sketch of Theorem 3 and Proof of Theorem 47

In the proof, we give a reduction of Algorithm 2 to [39, Alg. 2] (IMCSI). Two necessary additional results that we need include Theorem 47 and Corollary 50. In Corollary 50, we bound a normalized margin-like quantity of the multitask-path-tree kernel used in the algorithm. Then in Theorem 47, we bound the quasi-dimension which indicates how to set the parameters of Algorithm 2 as well as determines the value of $C(\boldsymbol{h}^*)$ in the main theorem. As this bound of the quasi-dimension is a key element of our proof, we contrast it to a parallel result proved in [39, Thm. 3].

We recall that the regret (see [39, Thm. 1]) of IMCSI is $\tilde{\mathcal{O}}(\sqrt{(\widehat{\mathcal{D}}/\gamma^2)T})$ where $1/\gamma^2 \geq \|\boldsymbol{U}\|^2_{\max}$ and $\widehat{\mathcal{D}} \geq \mathcal{D}^{\gamma}_{\boldsymbol{M},\boldsymbol{N}}(\boldsymbol{U})$. We will prove in our setting $1/\gamma^2 = m \geq \|\boldsymbol{U}\|^2_{\max}$ in the discussion following (44).

We now contrast our bound on the quasi-dimension (Theorem 47) to the bound of [39, Thm. 3].

The quasi-dimension depends on γ so that if $\gamma \leq \gamma'$, then $\mathcal{D}^{\gamma}_{\boldsymbol{M},\boldsymbol{N}}(\boldsymbol{U}) \leq \mathcal{D}^{\gamma'}_{\boldsymbol{M},\boldsymbol{N}}(\boldsymbol{U})$. In [39, Thm. 3] the given bound on quasi-dimension is independent of the value of γ. Thus to minimize the regret bound of $\tilde{\mathcal{O}}(\sqrt{(\widehat{\mathcal{D}}/\gamma^2)T})$, it is sensible in [39] to select the smallest possible $1/\gamma^2 = \|\boldsymbol{U}\|^2_{\max}$. The situation in this paper is essentially reversed. In the following theorem, it is required that $1/\gamma^2 = m \geq \|\boldsymbol{U}\|^2_{\max}$. In fact, m is the maximum possible value of the squared max norm in the case that $m = |m(\boldsymbol{h}^*)|$ with respect to all possible comparators $\boldsymbol{h}^*$ (see (44)). Thus in contrast to [39, Thm. 3], our result trade-offs a potentially larger value in $1/\gamma^2$ for a smaller possible $\widehat{\mathcal{D}}$. If we were instead to use the bound of [39, Thm 3], then the term in this paper $\sum_{h\in m(\boldsymbol{h}^*)} \|h\|^2_K X^2_K$ would gain a leading multiplicative factor of m^2 (terrible!).

We introduce the following notation. We recall the class of $m \times d$ *row-normalized* as $\mathcal{N}^{m,d} := \{\hat{\boldsymbol{P}} \subset \Re^{m\times d} : \left\|\hat{\boldsymbol{P}}_i\right\| = 1, i \in [m]\}$ and denote the class of *block expansion* matrices as $\mathcal{B}^{m,d} := \{\boldsymbol{R} \subset \{0,1\}^{m\times d} : \|\boldsymbol{R}_i\| = 1 \text{ for } i \in [m], \operatorname{rank}(\boldsymbol{R}) = d\}$. Block expansion matrices may be seen as a generalization of permutation matrices, additionally duplicating rows (columns) by left (right) multiplication. The class of (k,ℓ)-binary-biclustered matrices is defined as

$$\mathbb{B}^{m,n}_{k,\ell} = \{\boldsymbol{U} = \boldsymbol{R}\boldsymbol{U}^*\boldsymbol{C}^\top \in \{-1,1\}^{m\times n} : \boldsymbol{U}^* \in \{-1,1\}^{k\times\ell}, \boldsymbol{R} \in \mathcal{B}^{m,k}, \boldsymbol{C} \in \mathcal{B}^{n,\ell}\}.$$

Theorem 47. *If* $\boldsymbol{U} \in \mathbb{B}^{p,T}_{p,m}$, $\gamma = 1/\sqrt{m}$ *and if*

$$\mathcal{D}^{\star}_{\boldsymbol{M},\boldsymbol{N}}(\boldsymbol{U}) := \gamma^2 \operatorname{tr}((\boldsymbol{U}^*)^\top \boldsymbol{M}\boldsymbol{U}^*)\mathcal{R}_{\boldsymbol{M}} + \operatorname{tr}(\boldsymbol{C}^\top\boldsymbol{N}\boldsymbol{C})\mathcal{R}_{\boldsymbol{N}} \tag{27}$$

is defined as the minimum over all decompositions of $\boldsymbol{U} = \boldsymbol{U}^*\boldsymbol{C}^\top$ *for* $\boldsymbol{U}^* \in \{-1,1\}^{p\times m}$ *and* $\boldsymbol{C} \in \mathcal{B}^{T,m}$ *then*

$$\mathcal{D}^{\gamma}_{\boldsymbol{M},\boldsymbol{N}}(\boldsymbol{U}) \leq \mathcal{D}^{\star}_{\boldsymbol{M},\boldsymbol{N}}(\boldsymbol{U}) \qquad (\gamma = 1/\sqrt{m}).$$

Proof. Recall by supposition $\gamma = 1/\sqrt{m}$. Set $\hat{\boldsymbol{P}}' := \gamma\boldsymbol{U}^*$ and $\hat{\boldsymbol{Q}}' := \boldsymbol{C}$ hence $\hat{\boldsymbol{P}}' \in \mathcal{N}^{p,m}$, $\hat{\boldsymbol{Q}}' \in \mathcal{N}^{T,m}$ and $\hat{\boldsymbol{P}}'\hat{\boldsymbol{Q}}'^\top = \gamma\boldsymbol{U}$.

Recall (7),

$$\mathcal{D}^{\gamma}_{\boldsymbol{M},\boldsymbol{N}}(\boldsymbol{U}) := \min_{\hat{\boldsymbol{P}}\hat{\boldsymbol{Q}}^\top = \gamma\boldsymbol{U}} \operatorname{tr}\left(\hat{\boldsymbol{P}}^\top\boldsymbol{M}\hat{\boldsymbol{P}}\right)\mathcal{R}_{\boldsymbol{M}} + \operatorname{tr}\left(\hat{\boldsymbol{Q}}^\top\boldsymbol{N}\hat{\boldsymbol{Q}}\right)\mathcal{R}_{\boldsymbol{N}}. \tag{28}$$

Observe that $(\hat{\boldsymbol{P}}', \hat{\boldsymbol{Q}}')$ is in the feasible set of the above optimization. Hence

$$\begin{aligned}\mathcal{D}^{\gamma}_{\boldsymbol{M},\boldsymbol{N}}(\boldsymbol{U}) &\leq \operatorname{tr}\left(\hat{\boldsymbol{P}}'^\top\boldsymbol{M}\hat{\boldsymbol{P}}'\right)\mathcal{R}_{\boldsymbol{M}} + \operatorname{tr}\left(\hat{\boldsymbol{Q}}'^\top\boldsymbol{N}\hat{\boldsymbol{Q}}'\right)\mathcal{R}_{\boldsymbol{N}} \\ &= \gamma^2 \operatorname{tr}((\boldsymbol{U}^*)^\top\boldsymbol{M}\boldsymbol{U}^*)\mathcal{R}_{\boldsymbol{M}} + \operatorname{tr}(\boldsymbol{C}^\top\boldsymbol{N}\boldsymbol{C})\mathcal{R}_{\boldsymbol{N}}.\end{aligned}$$

□

Proof of Corollary 50

In this section, we prove Corollary 50, which is utilized in the proof of Theorem 3.

We recall the notions of *effective resistance* between vertices in a graph and the *resistance diameter* of a graph. A graph may naturally interpreted as an resistive network where each edge in the graph is viewed as a unit resistor. Thus the *effective resistance* between two vertices is the potential difference needed to induce a unit current flow between them and the *resistance diameter* is the maximum effective resistance between all pairs of vertices.

To prove the corollary, we will need to bound the diagonal element of the Laplacian pseudo-inverse by the resistance diameter. In the following Lemma, we will improve upon [49, Eq. (9)] by a factor of $\frac{1}{2}$ for the special case of fully complete trees.

Lemma 48. *For the graph Laplacian $\boldsymbol{L} \in \Re^{N\times N}$ of a fully complete tree graph,*

$$\mathcal{R}_{\boldsymbol{L}} = \max_{i\in[N]} L^+_{ii} \leq \frac{1}{2}\mathcal{R}_{diam}(\boldsymbol{L}),$$

where $R_{diam}(\boldsymbol{L})$ is the resistance diameter of the graph described by $\boldsymbol{L}$.

Proof. Before proving the result, we shall recall 4 general facts about graphs, trees and Laplacians. We also denote the set of vertices at a given depth a *level*. The root is at level 0.

1. The effective resistance between vertices i and j is given by (see [50]),
$$\mathcal{R}(i,j) = L^+_{ii} + L^+_{jj} - 2L^+_{ij}. \tag{29}$$
2. The diagonal element of L^+ is given by (see eg. [51])
$$L^+_{ii} = \frac{\mathcal{R}(i)}{N} - \frac{\mathcal{R}_{tot}}{N^2}, \tag{30}$$
where $\mathcal{R}(i) = \sum_{j=1}^N \mathcal{R}(i,j)$ and $\mathcal{R}_{tot} = \sum_{i,j<i} \mathcal{R}(i,j)$.
3. For fully complete trees, we have that $\mathcal{R}(i) = \mathcal{R}(j)$ and $L^+_{ii} = L^+_{jj}$ if i and j are in the same level due to symmetry.
4. For trees, the effective resistance between vertices i and j is given by the geodesic distance (path length) between the two vertices.

Next, we prove the following intermediate result.

Lemma: For a given vertex i, the vertex j that minimizes L^+_{ij} is the leaf vertex with the largest geodesic distance from i.

Proof. Define h to be the height of the tree. We take vertex i' to be at level $k \in [h-1]$ and vertex j' at level $k+1$. Recalling that $\mathcal{R}(i) = \sum_{j=1}^N \mathcal{R}(i,j)$, we will consider the individual summands that compose $\mathcal{R}(j')$ and $\mathcal{R}(i')$, given by the geodesic distances between i' and j' respectively and the other vertices due to fact 4. From fact 3 (with respect to the summands), we can assume without loss of generality that vertex j' is the child of i'. Going from the summation of $\mathcal{R}(j')$ to the summation of $\mathcal{R}(i')$, there are 3 possible changes to the geodesic distances in the summation:

1. the descendants of j' will have a geodesic distance reduced by 1
2. the geodesic distance between i' and j' remains constant
3. all the other vertices will have a geodesic distance increased by 1.

Hence, defining $\mathcal{D}_{j'}$ to be the set of descendants of node j',

$$\begin{aligned}\mathcal{R}(j') &= \mathcal{R}(i') - \sum_{i\in\mathcal{D}_{j'}} 1 + \sum_{i'\in[N]\setminus\mathcal{D}_{j'}\cup\{j'\}} 1\\ &= \mathcal{R}(i') - |\mathcal{D}_{j'}| + N - (|\mathcal{D}_{j'}| + 1)\\ &= \mathcal{R}(i') + N - 2|\mathcal{D}_{j'}| - 1.\end{aligned}$$

This gives that $\mathcal{R}(j') - \mathcal{R}(i') \leq N$, and

$$\frac{\mathcal{R}(j') - \mathcal{R}(i')}{N} \leq 1. \tag{31}$$

We show that vertex j that minimizes L^+_{ij} must be a leaf vertex by contradiction. Suppose j is not a leaf vertex then there exists a child of $j \neq i$. Call the child j' which thus satisfies $\mathcal{R}(i, j') - \mathcal{R}(i, j) = 1$. Hence, Equations (29) and (30) give

$$L^+_{ij'} - L^+_{ij} = \frac{1}{2}\left(\frac{\mathcal{R}(j')}{N} - \frac{\mathcal{R}(j)}{N} - \mathcal{R}(i, j') + \mathcal{R}(i, j)\right) \tag{32}$$

$$\leq 0, \tag{33}$$

where the inequality is due to (31) for which we let $i' = j$. Hence, we have that $L^+_{ij} \geq L^+_{ij'}$ which is a contradiction.

Then, using Equations (29) and (30), we have

$$\operatorname*{argmin}_j L^+_{ij} = \operatorname*{argmin}_j \frac{\mathcal{R}(j)}{N} - \mathcal{R}(i, j).$$

Since all leaf vertices have the same $\mathcal{R}(i)$, the leaf vertex that minimizes must be the one with the largest geodesic distance from i. □

Recall that for a tree, the resistance diameter is equal to its geodesic diameter, and hence the vertices that maximize the effective resistance are given by the two leaf vertices with the largest geodesic distance. We therefore proceed by considering i and j to be any of the vertices that maximize the effective resistance, giving the resistance diameter. Due to fact 3, we have $L^+_{ii} = L^+_{jj}$. Then, from (29), we obtain,

$$\begin{aligned}\frac{1}{2}\mathcal{R}_{diam}(\boldsymbol{L}) &= L^+_{ii} - L^+_{ij} \\ &= L^+_{ii} - \min_k L^+_{ik} &(34) \\ &\geq L^+_{ii} - \frac{1}{N}\sum_{k=1}^{N} L^+_{ik} \\ &\geq L^+_{ii} &(35)\end{aligned}$$

where (34) comes from the intermediate lemma, and (35) comes from the fact that $\sum_{j=1}^N L^+_{ij} = 0$ for all $i \in [N]$ since $\boldsymbol{L}\mathbf{1} = \mathbf{0}$ for connected graphs. □

The following Lemma is essentially a simplification of the argument in [43, Section 6] for Laplacians,

Lemma 49. *(See [43, Section 6].) If $f \in \mathcal{H}_P \cap \{0, 1\}^T$ then*

$$\max_{\tau \in [T]} P(t, t) \leq 2\lceil \log_2 T \rceil$$

$$\|f\|^2_P \max_{t \in [T]} P(t, t) \leq k(f)\lceil \log_2 T \rceil^2 + 2, , \tag{36}$$

where $k(f) := \sum_{t=1}^{T-1}[f(t) \neq f(t + 1)]$.

Proof. First we recall the following standard fact about the graph Laplacian $\boldsymbol{L}$ of an unweighted graph $\mathcal{G} = (V, E)$,

$$\boldsymbol{u}^\top \boldsymbol{L}\boldsymbol{u} = \sum_{(i,j)\in E} (u_i - u_j)^2 ,$$

where V is the set of vertices and E is the set of edges in the graph. Call this quantity the *cut* of the labeling $\boldsymbol{u}$. Consider a fully complete binary tree with a depth of $\lceil \log_2 T \rceil + 1$. For simplicity now assume that there are exactly T leaf nodes, i.e., $\log_2 T \in \mathbb{N}$. Assume some natural linear ordering[3] of

[3] Given every three vertices in ordering (a, b, c) we have that $d(a, b) \leq d(a, c)$ where $d(p, q)$ is the path length between p and q.

the leaves. This ordering then defines our *path*. We call each set of vertices at a given depth a "level" and they inherit a natural linear ordering from their children. Suppose that there are n vertices at a given level ℓ, and define $w_i^\ell := u_{v_i^\ell}$, where v_i^ℓ is the ith vertex on level ℓ. The *path-cut* at this level is given by $\sum_{i=1}^{n-1} |w_i^\ell - w_{i+1}^\ell|$.

We now proceed to argue that for a given binary labeling of a path with associated path-cut $k(f)$, we can identify a (real-numbered) labeling of the tree, such that: a. the labeling of the tree leaves is binary and consistent with that of the path and b. the tree has a cut of no more than $\frac{1}{2}k(f)\lceil \log T \rceil$. The construction is as follows: *each parent inherits the average of the labels of its children.* We make two observations about the constructed labeling:

1. The path-cut at a higher level cannot be more than the level below. Consider two adjacent levels with the lower level ℓ having n vertices. Denote the set of odd numbers that is a subset of $[n-1]$ as I_{odd}, and the set of even number that is a subset of $[n-2]$ as I_{even}. Recall that the path-cut of the lower level is $\sum_{i=1}^{n-1} |w_i^\ell - w_{i+1}^\ell|$. This can be upper bounded as follows:

$$
\begin{aligned}
&\sum_{i=1}^{n-1} |w_i^\ell - w_{i+1}^\ell| \\
&= \sum_{i \in I_{odd}} \left| w_i^\ell - \frac{w_i^\ell + w_{i+1}^\ell}{2} \right| + \left| \frac{w_i^\ell + w_{i+1}^\ell}{2} - w_{i+1}^\ell \right| + \sum_{i \in I_{even}} |w_i^\ell - w_{i+1}^\ell| \\
&= \sum_{i \in I_{odd}} \left| w_i^\ell - \frac{w_i^\ell + w_{i+1}^\ell}{2} \right| + \left| \frac{w_i^\ell + w_{i+1}^\ell}{2} - w_{i+1}^\ell \right| + \sum_{i \in I_{odd} \setminus \{n-1\}} |w_{i+1}^\ell - w_{i+2}^\ell| \\
&\geq \sum_{i \in I_{odd}} \left| w_i^\ell - \frac{w_i^\ell + w_{i+1}^\ell}{2} \right| + \sum_{i \in I_{odd} \setminus \{n-1\}} \left| \frac{w_i^\ell + w_{i+1}^\ell}{2} - w_{i+1}^\ell \right| + |w_{i+1}^\ell - w_{i+2}^\ell| \\
&\geq \sum_{i \in I_{odd}} \left| w_i^\ell - \frac{w_i^\ell + w_{i+1}^\ell}{2} \right| + \sum_{i \in I_{odd} \setminus \{n-1\}} \left| \frac{w_i^\ell + w_{i+1}^\ell}{2} - w_{i+2}^\ell \right| \qquad (37) \\
&\geq \sum_{i \in I_{odd} \setminus \{1\}} \left| u_i - \frac{u_i + w_{i+1}^\ell}{2} \right| + \sum_{i \in I_{odd} \setminus \{n-1\}} \left| \frac{w_i^\ell + w_{i+1}^\ell}{2} - w_{i+2}^\ell \right| \\
&= \sum_{i \in I_{odd} \setminus \{n-1\}} \left| w_{i+2}^\ell - \frac{w_{i+2}^\ell + w_{i+3}^\ell}{2} \right| + \left| \frac{w_i^\ell + w_{i+1}^\ell}{2} - w_{i+2}^\ell \right| \\
&\geq \sum_{i \in I_{odd} \setminus \{n-1\}} \left| \frac{w_i^\ell + w_{i+1}^\ell}{2} - \frac{w_{i+2}^\ell + w_{i+3}^\ell}{2} \right| \qquad (38)
\end{aligned}
$$

   where (37) and (38) follow from $|a-b| + |b-c| \geq |a-c|$ (triangle inequality). Observing that the R.H.S. of (38) is the path cut of the upper level, we are then done.

2. If we denote the set of edges between two adjacent levels by $\tilde{E}$, we have that $\sum_{(i,j)\in\tilde{E}}(u_i - u_j)^2$ is at most half the path-cut of the lower level. This can be seen by considering the edges between a given parent i and its two children j and j'. Let us define x as half the path cut due to the children, i.e. $\frac{1}{2}|u_j - u_{j'}|$. Since all labelings are in $[0,1]$, we have that $x \in [0, 1/2]$. The cut made due to the parent and the children, i.e. $(u_i - u_j)^2 + (u_i - u_{j'})^2$ is then given by $2x^2$. Using the inequality $x - 2x^2 \geq 0$ for $x \in [0, 1/2]$, and applying this to all the parents on the same level as vertex i, we then prove the statement.

Hence, combining the two above observations and recalling that there are $\lceil \log_2 T \rceil + 1$ levels (and therefore $\lceil \log_2 T \rceil$ transitions between the levels), we have that

$$
\sum_{(i,j)\in E} (u_i - u_j)^2 \leq \sum_{\ell=1}^{\lceil \log_2 T \rceil} \frac{1}{2} p(\ell) \leq \sum_{k=1}^{\lceil \log_2 T \rceil} \frac{1}{2} k(f),
$$

where $p(\ell)$ is the path-cut of the tree at level ℓ, the first inequality is due to observation 2, and the second inequality is due to observation 1. Hence we have shown our premise that the cut is upper bounded by $\frac{1}{2}k(f)\lceil\log_2 T\rceil$. Observe that our premise still holds if there are more than T leaf nodes, as we can treat any additional leaves on the bottom level as being labeled with the last label on their level; thus the cut will not increase. Hence we have shown the following inequality where $\boldsymbol{L}$ is the Laplacian of a fully complete binary tree with N vertices and a path of T leaves labeled by an $f \in \{0,1\}^{[T]}$.

$$\min_{\boldsymbol{u}\in\boldsymbol{R}^{[N]}:u_i=f(i),i\in[T]} \boldsymbol{u}^\top \boldsymbol{L}\boldsymbol{u} \leq \frac{1}{2}k(f)\lceil\log_2 T\rceil\,. \tag{39}$$

We next observe that

$$\mathcal{R}_{\boldsymbol{L}} \leq \lceil\log_2 T\rceil\,. \tag{40}$$

This follows from Lemma 48, where $\mathcal{R}_{\boldsymbol{L}} = \max_{i\in[N]} L^+_{ii}$ is bounded by half the resistance diameter, which is then just bounded by half the geodesic diameter. Furthermore if $\boldsymbol{L}$ is the Laplacian of a connected graph and $\boldsymbol{L}^\circ := \boldsymbol{L} + \left(\frac{\boldsymbol{1}}{m}\right)\left(\frac{\boldsymbol{1}}{m}\right)^\top \mathcal{R}_{\boldsymbol{L}}^{-1}$ then if $\boldsymbol{u}\in[-1,1]^m$ we have

$$\mathcal{R}_{\boldsymbol{L}^\circ} = 2\mathcal{R}_{\boldsymbol{L}}\,,$$
$$\boldsymbol{u}^\top\boldsymbol{L}^\circ\boldsymbol{u} \leq \boldsymbol{u}^\top\boldsymbol{L}\boldsymbol{u} + \frac{1}{\mathcal{R}_{\boldsymbol{L}}}\,.$$

Thus combining the above with (39) and (40), we have,

$$\mathcal{R}_{\boldsymbol{L}^\circ} \leq 2\lceil\log_2 T\rceil\,,$$
$$\min_{\boldsymbol{u}\in\boldsymbol{R}^{[N]}:u_i=f(i),i\in[T]} (\boldsymbol{u}^\top\boldsymbol{L}^\circ\boldsymbol{u})\mathcal{R}_{\boldsymbol{L}^\circ} \leq k(f)\lceil\log_2 T\rceil^2 + 2\,,$$

which proves the Lemma. □

Observe that the left hand side in (36) is up to constant factors, the normalized margin of f in the sense of Novikoff's Theorem [52]. The construction is somewhat counterintuitive as one may expect that one can use a path graph directly in the construction of the kernel. However, then $\max_{t\in[T]} P(t,t) \in \Theta(T)$ which would lead to a vacuous regret bound. Also one may wonder if one can reduce the term $(\log T)^2$ while maintaining a linear factor in $k(f)$. In fact the term $(\log T)^2$ is known [53, Theorem 6.1] to be required when $k(f)=1$.

As a straightforward corollary to Lemma 49, we have

Corollary 50. *If $f \in \mathcal{H}_{\tilde{P}} \cap \{0,1\}^T$ then*

$$\|f\|^2_{\tilde{P}} \max_{\tau\in[T]} \tilde{P}(\tau,\tau) \leq (k(f)+s(f))\lceil\log_2 T\rceil^2 + 2\,, \tag{41}$$

where $\tilde{P} = \tilde{P}^{\boldsymbol{\ell},T^1,\ldots,T^s}$, $k(f) := \sum_{i=1}^{s}\sum_{t=1}^{T^i-1}[f(^i_t) \neq f(^{\;i}_{t+1})]$ and $s(f) := \sum_{i=1}^{s-1}[f(^{\;i}_{T^i}) \neq f(^{i+1}_{1})]$.

Proof. Since each task is laid out contiguously along the bottom layer, we pay the path-cut for each task individually and we pay $s(f)$ for the intertask boundaries. □

Proof of Theorem 3

We first recall some of the notation introduced earlier in the section. The *block expansion* matrices are defined as $\mathcal{B}^{m,d} := \{\boldsymbol{R}\subset\{0,1\}^{m\times d} : \|\boldsymbol{R}_i\| = 1 \text{ for } i\in[m], \operatorname{rank}(\boldsymbol{R}) = d\}$. The class of (k,ℓ)-binary-biclustered matrices is defined as $\mathbb{B}^{m,n}_{k,\ell} = \{\boldsymbol{U} = \boldsymbol{R}\boldsymbol{U}^*\boldsymbol{C}^\top \in \{-1,1\}^{m\times n} : \boldsymbol{U}^* \in \{-1,1\}^{k\times\ell}, \boldsymbol{R}\in\mathcal{B}^{m,k}, \boldsymbol{C}\in\mathcal{B}^{n,\ell}\}$. Next, we recall Theorem 3 and provide a proof.

Theorem 3. *The expected regret of Algorithm 2 with upper estimates, $k \geq k(\boldsymbol{h}^*)$, $m \geq |m(\boldsymbol{h}^*)|$,*

$$\hat{C} \geq C(\boldsymbol{h}^*) := \left(\sum_{h\in m(\boldsymbol{h}^*)} \|h\|^2_K X^2_K + 2(s+k-1)m\lceil\log_2 T\rceil^2 + 2m^2\right)\,,$$

$\hat{X}_K^2 \geq \max_{\tau \in [T]} K(x^\tau, x^\tau)$, *and learning rate* $\eta = \sqrt{\frac{\hat{C}\log(2T)}{2Tm}}$ *is bounded by*

$$\sum_{i=1}^{s}\sum_{t=1}^{T^i} \mathbb{E}[\mathcal{L}_{01}(y_t^i, \hat{y}_t^i)] - \mathcal{L}_{01}(y_t^i, h_t^i(x_t^i)) \leq 4\sqrt{2\hat{C}\,T\log(2T)} \tag{42}$$

with received instance sequence $\boldsymbol{x} \in \mathcal{X}^T$ *for any* $\boldsymbol{h}^* \in \mathcal{H}_K^{(\boldsymbol{x})^T}$.

Proof. Algorithm 2 is the same as [39, Algorithm 2] (which we call IMCSI) except for some redefinitions in the notation. For convenience we recall IMCSI below[4]

Algorithm 2 ([39]).

Parameters: Learning rate: $0 < \eta$ quasi-dimension estimate: $1 \leq \widehat{\mathcal{D}}$, margin estimate: $0 < \gamma \leq 1$ and side-information kernels $\mathcal{M}^+ : \mathcal{I} \times \mathcal{I} \to \Re$, $\mathcal{N}^+ : \mathcal{J} \times \mathcal{J} \to \Re$, with $\mathcal{R}_{\mathcal{M}} := \max_{i \in \mathcal{I}} \mathcal{M}^+(i,i)$ and $\mathcal{R}_{\mathcal{N}} := \max_{j \in \mathcal{J}} \mathcal{N}^+(j,j)$, and maximum distinct rows m and columns n, where $m + n \geq 3$.

Initialization: $\mathbb{M} \leftarrow \emptyset\,, \mathbb{U} \leftarrow \emptyset\,, \mathcal{I}^1 \leftarrow \emptyset\,,\ \mathcal{J}^1 \leftarrow \emptyset$.

For $t = 1, \ldots, T$

- Receive pair $(i_t, j_t) \in \mathcal{I} \times \mathcal{J}$.
- Define

$$(\boldsymbol{M}^t)^+ := (\mathcal{M}^+(i_r, i_s))_{r,s \in \mathcal{I}^t \cup \{i_t\}}\,; \quad (\boldsymbol{N}^t)^+ := (\mathcal{N}^+(j_r, j_s))_{r,s \in \mathcal{J}^t \cup \{j_t\}}\,,$$

$$\tilde{\boldsymbol{X}}^t(s) := \left[\frac{(\sqrt{(\boldsymbol{M}^t)^+})\boldsymbol{e}^{i_s}}{\sqrt{2\mathcal{R}_{\mathcal{M}}}}; \frac{(\sqrt{(\boldsymbol{N}^t)^+})\boldsymbol{e}^{j_s}}{\sqrt{2\mathcal{R}_{\mathcal{N}}}}\right]\left[\frac{(\sqrt{(\boldsymbol{M}^t)^+})\boldsymbol{e}^{i_s}}{\sqrt{2\mathcal{R}_{\mathcal{M}}}}; \frac{(\sqrt{(\boldsymbol{N}^t)^+})\boldsymbol{e}^{j_s}}{\sqrt{2\mathcal{R}_{\mathcal{N}}}}\right]^\top,$$

$$\tilde{\boldsymbol{W}}^t \leftarrow \exp\left(\log\left(\frac{\widehat{\mathcal{D}}}{m+n}\right)\boldsymbol{I}^{|\mathcal{I}^t| + |\mathcal{J}^t| + 2} + \sum_{s \in \mathbb{U}} \eta y_s \tilde{\boldsymbol{X}}^t(s)\right).$$

- Predict

$$Y_t \sim \textsc{Uniform}(-\gamma, \gamma)\,; \bar{y}_t \leftarrow \mathrm{tr}\left(\tilde{\boldsymbol{W}}^t \tilde{\boldsymbol{X}}^t\right) - 1\,; \quad \hat{y}_t \leftarrow \mathrm{sign}(\bar{y}_t - Y_t)\,.$$

- Receive label $y_t \in \{-1, 1\}$.
- If $y_t \bar{y}_t \leq \gamma$ then

$$\mathbb{U} \leftarrow \mathbb{U} \cup \{t\}\,,\ \mathcal{I}^{t+1} \leftarrow \mathcal{I}^t \cup \{i_t\},\ \text{and } \mathcal{J}^{t+1} \leftarrow \mathcal{J}^t \cup \{j_t\}\,.$$

- Else $\mathcal{I}^{t+1} \leftarrow \mathcal{I}^t$ and $\mathcal{J}^{t+1} \leftarrow \mathcal{J}^t$.

The following table summarizes the notational changes between the two algorithms.

Description	IMCSI	Algorithm 2
Row space	$\mathcal{I}$	$\mathcal{X}$
Column space	$\mathcal{J}$	$[T]$
Row kernel	$\mathcal{M}^+$	K
Column kernel	$\mathcal{N}^+$	$P := \tilde{P}^{\boldsymbol{\ell}, T^1, \ldots, T^s}$
Row squared radius	$\mathcal{R}_{\mathcal{M}}$	$\hat{X}_K^2$
Column squared radius	$\mathcal{R}_{\mathcal{N}}$	$\hat{X}_P^2$
Margin estimate	γ^{-2}	m
Complexity Estimate	$\widehat{\mathcal{D}}\gamma^{-2}$	$\hat{C}$
Dimensions[5]	m, n	T, T
Time	t	τ
Instance	(i_t, j_t)	(x^τ, τ)

We now recall the following regret bound for IMCSI.

[4]Since we are only concerned with regret bound we have set the parameter **NON-CONSERVATIVE** $= 1$ in our restating of the algorithm.

Theorem 1 ([39, Theorem 1/Proposition 4]) *The expected regret of [39, Algorithm 2] with parameters* $\gamma \in (0,1]$, $\widehat{\mathcal{D}} \geq \mathcal{D}^{\gamma}_{\boldsymbol{M},\boldsymbol{N}}(\boldsymbol{U})$, $\eta = \sqrt{\frac{\widehat{\mathcal{D}}\log(m+n)}{2T}}$*, p.d. matrices* $\boldsymbol{M} \in \boldsymbol{S}^m_{++}$ *and* $\boldsymbol{N} \in \boldsymbol{S}^n_{++}$ *is bounded by*

$$\sum_{t\in[T]} \mathbb{E}[\mathcal{L}_{01}(y_t,\hat{y}_t)] - \mathcal{L}_{01}(y_t, U_{i_t j_t}) \leq 4\sqrt{2\frac{\widehat{\mathcal{D}}}{\gamma^2}\log(m+n)T} \tag{43}$$

for all $\boldsymbol{U} \in \{-1,1\}^{m\times n}$ *with* $\|\boldsymbol{U}\|_{max} \leq 1/\gamma$.

We introduce the following notation: the matrix $\boldsymbol{H} := (h^{\tau}(x) : x \in \mathcal{X}^{\text{fin}}, \tau \in [T])$, the set $\mathcal{X}^{\text{fin}} := \cup_{\tau\in[T]}\{x^{\tau}\}$, the matrices $\bar{\boldsymbol{K}} = [(K(x,x') : x,x' \in \mathcal{X}^{\text{fin}})]^{-1}$, and $\bar{\boldsymbol{P}} = [(\tilde{P}(\tau,\upsilon) : \tau,\upsilon \in [T])]^{-1}$.

Initially we note that we very minorly extend the algorithm and thus its analysis [39, Theorem 1] in so far as we use the upper bounds and $\hat{X}^2_K \geq \mathcal{R}_{\bar{\boldsymbol{K}}}$ and $\hat{X}^2_P \geq \mathcal{R}_{\bar{\boldsymbol{P}}}$.

It now remains that in order to complete the reduction of Theorem 3 to [39, Theorem 1] we need to demonstrate the following two inequalities,

$$\|\boldsymbol{H}\|_{\max} \leq \sqrt{m} \tag{44}$$

$$\mathcal{D}^{1/\sqrt{m}}_{\bar{\boldsymbol{K}},\bar{\boldsymbol{P}}}(\boldsymbol{H}) \leq \frac{1}{m}C(\boldsymbol{h}^*)\,. \tag{45}$$

First we show (44). Initially we derive the following simple inequality,

$$\|\boldsymbol{U}\|_{\max} \leq \min(\sqrt{m},\sqrt{n})\,, \tag{46}$$

which follows since we may decompose $\boldsymbol{U} = \boldsymbol{U}\boldsymbol{I}^n$ or as $\boldsymbol{U} = \boldsymbol{I}^m\boldsymbol{U}$. Let $p = |\mathcal{X}^{\text{fin}}|$. Recall by definition $m \geq |m(\boldsymbol{h}^*)|$ and thus there are only at most m distinct columns which implies $\boldsymbol{H} = \boldsymbol{I}^p\boldsymbol{H}^*\boldsymbol{C}^{\top}$ where $\boldsymbol{H}^* \in \{-1,1\}^{p\times m}$ and $\boldsymbol{C} \in \mathcal{B}^{T,m}$ hence $\boldsymbol{H} \in \mathbb{B}^{p,T}_{p,m}$. We now show,

$$\|\boldsymbol{H}^*\|_{\max} \geq \|\boldsymbol{H}\|_{\max}\,. \tag{47}$$

For every factorization $\boldsymbol{H}^* = \boldsymbol{P}^*\boldsymbol{Q}^{*\top}$ there exists a factorization $\boldsymbol{H} = \boldsymbol{P}^*(\boldsymbol{Q}^{*\top}\boldsymbol{C}^{\top})$ for some $\boldsymbol{C} \in \mathcal{B}^{T,m}$ such that

$$\max_{1\leq i\leq p}\|\boldsymbol{P}^*_i\| \max_{1\leq j\leq m}\left\|\boldsymbol{Q}^*_j\right\| = \max_{1\leq i'\leq p}\|\boldsymbol{P}^*_{i'}\| \max_{1\leq j'\leq T}\|(\boldsymbol{C}\boldsymbol{Q}^*)_{j'}\|\,,$$

since for every row vector $\boldsymbol{Q}^*_j$ there exists a row vector $(\boldsymbol{C}\boldsymbol{Q}^*)_{j'}$ such that $\boldsymbol{Q}^*_j = (\boldsymbol{C}\boldsymbol{Q}^*)_{j'}$ and vice versa. Therefore since the max norm (see (6)) is the minimum over all factorizations we have shown (47). Since $\boldsymbol{H}^* \in \{-1,1\}^{p\times m}$ we have $\|\boldsymbol{H}\|_{\max} \leq \|\boldsymbol{H}^*\|_{\max} \leq \min(\sqrt{p},\sqrt{m})$ by (46) and thus we have demonstrated (44).

We now show (45). We recall the following useful equality,

$$\boldsymbol{u}^{\top}\boldsymbol{K}^{-1}\boldsymbol{u} = \operatorname*{argmin}_{f\in\mathcal{H}_K : f(x)=u_x : x\in X} \|f\|^2_K\,. \tag{48}$$

where $\boldsymbol{K} = (K(x,x'))_{x,x'\in X}$, $\boldsymbol{u} \in \Re^X$ and $\boldsymbol{K}$ is invertible and K is a kernel. By Theorem 47 we have

$$\mathcal{D}^{1/\sqrt{m}}_{\bar{\boldsymbol{K}},\bar{\boldsymbol{P}}}(\boldsymbol{H}) \leq \frac{1}{m}\operatorname{tr}((\boldsymbol{H}^*)^{\top}\bar{\boldsymbol{K}}\boldsymbol{H}^*)\hat{X}^2_K + \operatorname{tr}(\boldsymbol{C}^{\top}\bar{\boldsymbol{P}}\boldsymbol{C})\hat{X}^2_P$$

where $\boldsymbol{H} = \boldsymbol{H}^*\boldsymbol{C}^{\top}$ with $\boldsymbol{H}^* := (h(x))_{x\in\mathcal{X}^{\text{fin}},h\in m(\boldsymbol{h}^*)}$ and $\boldsymbol{C} := ([h^{\tau} = h])_{\tau\in[T],h\in m(\boldsymbol{h}^*)}$ (note $\boldsymbol{C} \in \mathcal{B}^{T,m}$).

Simplifying and using (48) we have,

$$\mathcal{D}^{1/\sqrt{m}}_{\bar{\boldsymbol{K}},\bar{\boldsymbol{P}}}(\boldsymbol{H}) \leq \frac{1}{m}\sum_{h\in m(\boldsymbol{h}^*)}\|h\|^2_K\,\hat{X}^2_K + \operatorname{tr}(\boldsymbol{C}^{\top}\bar{\boldsymbol{P}}\boldsymbol{C})\hat{X}^2_P\,. \tag{49}$$

From (48) we have,

$$\operatorname{tr}(\boldsymbol{C}^{\top}\bar{\boldsymbol{P}}\,\boldsymbol{C}) = \sum_{h\in m(\boldsymbol{h}^*)}\boldsymbol{c}^{\top}_h\bar{\boldsymbol{P}}\boldsymbol{c}_h = \sum_{h\in m(\boldsymbol{h}^*)}\|f_h\|^2_{\tilde{P}} \tag{50}$$

where $\boldsymbol{c}_h$ is the column vector formed by taking the h^{th} column of C. The vector $\boldsymbol{c}_h \in \{0,1\}^T$ indicates if hypothesis h is "active" on trial τ, i.e., $c_h^\tau = [h^\tau = h]$. Next we define $f_h(\tau) := c_h^\tau$ for $\tau = 1, \cdots, T$. Recalling $\tau \equiv {\ell^\tau \atop \sigma(\tau)}$, we also have $f_h(\tau) \equiv f_h\,({\ell^\tau \atop \sigma(\tau)})$.

From (50) and Corollary 50 we have,

$$
\begin{aligned}
\operatorname{tr}(\boldsymbol{C}^\top \bar{\boldsymbol{P}} \boldsymbol{C}) \hat{X}_P^2 &= \sum_{h \in m(\boldsymbol{h}^*)} \|f_h\|_{\tilde{P}}\, \hat{X}_P^2 && (51)\\
&\leq \sum_{h \in m(\boldsymbol{h}^*)} \Big((k(f_h) + s(f_h)) \lceil \log_2 T \rceil^2 + 2 \Big) \\
&\leq \sum_{h \in m(\boldsymbol{h}^*)} (k(f_h) + s(f_h)) \lceil \log_2 T \rceil^2 + 2m(\boldsymbol{h}^*) \\
&\leq \sum_{h \in m(\boldsymbol{h}^*)} (k(f_h) + s(f_h)) \lceil \log_2 T \rceil^2 + 2m \\
&\leq 2(s + k - 1) \lceil \log_2 T \rceil^2 + 2m && (52)
\end{aligned}
$$

where

$$
k(f) = \sum_{i=1}^{s} \sum_{t=1}^{T^i - 1} [f({i \atop t}) \neq f({i \atop t+1})]\,,\; s(f) = \sum_{i=1}^{s-1} [f({i \atop T^i}) \neq f({i+1 \atop 1})]\,,
$$

and where (52) comes from using $\sum_{h \in m(\boldsymbol{h}^*)} k(f_h) = k(\boldsymbol{h}^*) \leq 2k$ and $\sum_{h \in m(\boldsymbol{h}^*)} s(h) \leq 2(s-1)$, where the factors of two are due to each switch of f_h on successive time steps as well as intertask boundaries being counted twice.

Substituting (52) into (49), we have

$$
\mathcal{D}^{1/\sqrt{m}}_{\bar{\boldsymbol{K}}, \bar{\boldsymbol{P}}}(\boldsymbol{H}) \leq \frac{1}{m} \sum_{h \in m(\boldsymbol{h}^*)} \|h\|_K^2\, \hat{X}_K^2 + 2(s + k - 1) \lceil \log_2 T \rceil^2 + 2m\,,
$$

This demonstrates (45) thus completing the reduction. □

Proof Sketch of Proposition 4

First we recall and then give a proof sketch of Proposition 4.

Proposition 4. *For any (randomized) algorithm and any $s, k, m, \Gamma \in \mathbb{N}$, with $k + s \geq m > 1$ and $\Gamma \geq m \log_2 m$, there exists a kernel K and a $T_0 \in \mathbb{N}$ such that for every $T \geq T_0$:*

$$
\sum_{\tau=1}^{T} \mathbb{E}[\mathcal{L}_{01}(y^\tau, \hat{y}^\tau)] - \mathcal{L}_{01}(y^\tau, h^\tau(x^\tau)) \in \Omega\left(\sqrt{(\Gamma + s \log m + k \log m)\, T} \right)\,,
$$

for some multitask sequence $(x^1, y^1), \ldots, (x^T, y^T) \in (\mathcal{X} \times \{-1, 1\})^T$ and some $\boldsymbol{h}^* \in [\mathcal{H}_K^{(\boldsymbol{x})}]^T$ such that $m \geq |m(\boldsymbol{h}^*)|$, $k \geq k(\boldsymbol{h}^*)$, $\sum_{h \in m(\boldsymbol{h}^*)} \|h\|_K^2 X_K^2 \geq |m(\boldsymbol{h}^*)| \log_2 m$, where $X_K^2 = \max_{\tau \in [T]} K(x^\tau, x^\tau)$.

Proof Sketch. We recall the following online learning terminology. A sequence of examples $(x_1, y_1), \ldots, (x_T, y_T)$ is *realizable* with respect to a hypothesis class $\mathcal{H}$ if there exists an $h \in \mathcal{H}$, such that $\sum_{t=1}^{T} \mathcal{L}_{01}(y_t, h(x_t)) = 0$. The *optimal mistake bound* (Ldim($\mathcal{H}$)) with respect to a hypothesis class $\mathcal{H}$ also known as the *Littlestone dimension* [54, 55] is, informally speaking, the minimum over all deterministic learning algorithms, of the maximum over all realizable example sequences of the number of mistaken predictions.

We will apply the following useful result [55, Lemma 14] which we quote below for convenience,

> **Lemma 14** (Lower Bound). *Let $\mathcal{H}$ be any hypothesis class with a finite Ldim($\mathcal{H}$). For any (possibly randomized) algorithm, exists a sequence $(x_1, y_1), \ldots, (x_T, y_T)$ such that*
>
> $$\mathbb{E} \sum_{t=1}^{T} \mathcal{L}_{01}(y_t, \hat{y}_t) - \min_{h \in \mathcal{H}} \mathcal{L}_{01}(y_t, h(x_t)) \geq \sqrt{\frac{\text{Ldim}(\mathcal{H}) T}{8}}\,.$$

In essence, this allows one to go from a lower bound on mistakes in the *realizable* case to a lower bound in the non-realizable case. However, Lemma 14 only applies directly to the standard single-task model. To circumvent this, we recall as discussed in Section 2, that the switching multitask model may be reduced to the single-task model with a domain $\mathcal{X}' = \mathcal{X} \times [T] \times [s]$ and hypothesis class $\mathcal{H}'$. Therefore a lower bound in the switching multitask model with respect to $\mathcal{H}$ implies a lower bound in the single-task non-switching case for $\mathcal{H}'$ via the reduction. There are some slight technical issues over the fact that "time" is now part of the domain $\mathcal{X}'$ and thus e.g., valid example sequences cannot be permuted. We gloss over these issues in this proof sketch noting that they do not in fact impact our arguments. The argument proceeds by demonstrating that there exists for any $s, k, m, \Gamma \in \mathbb{N}$ a kernel K and a realizable multitask sequence $(x^1, y^1), \ldots, (x^T, y^T)$ for which

$$\sum_{\tau=1}^{T} \mathcal{L}_{01}(y^\tau, \hat{y}^\tau) \in \Omega\left(\Gamma + s \log m + k \log m\right) , \tag{53}$$

where $\boldsymbol{x} \in \mathcal{X}^T$, $X_K^2 = \max_{\tau \in [T]} K(x^\tau, x^\tau)$, $\Gamma \geq \sum_{h \in m(\boldsymbol{h}^*)} \|h\|_K^2 X_K^2 \geq m \log_2 m$, $k \geq k(\boldsymbol{h}^*)$, $m \geq |m(\boldsymbol{h}^*)|$ and $k + s \geq m > 1$. After demonstrating that there exists such an example sequence we can apply [55, Lemma 14] to demonstrate the proposition. Since the lower bound is in the form $\Omega(P+Q+R)$ which is equivalent to $\Omega(\max(P,Q,R))$, we may treat P, Q and R, independently to prove the bound. Before we treat the individual cases, we give a straightforward result for a simplistic hypothesis class.

Define $\mathcal{X}_d := [d]$ and $\mathcal{H}_d := \{-1,1\}^d$ (i.e., the set of functions that map $[d] \to \{-1,1\}$). Observe that $\text{Ldim}(\mathcal{H}_d) = d$, as an algorithm can force a mistake for every component and then no more. Also, observe that if we define a kernel $K_d(x,x') := [x = x']$ over the domain $\mathcal{X}_d$ that $\mathcal{H}_d = \mathcal{H}_{K_d}^{([d])}$, $\max_{x \in [d]} K_d(x,x) = 1$ and that $\|h\|_{K_d}^2 = d$ for all $h \in \mathcal{H}_d$. Finally, note that if $m = |\mathcal{H}_d|$ then $\sum_{h \in \mathcal{H}_d} \|h\|_{K_d}^2 X_{K_d}^2 = m \log_2 m$.

We proceed by sketching an adversary for each of the three cases.

Case Γ is the max.

To force Γ mistakes, we choose $K = K_d$ and set $d = \Gamma/m$ and without loss of generality assume that d is an integer and recall that $k + s \geq m$. Since $\text{Ldim}(\mathcal{H}_d) = d$, an adversary may force d mistakes within a single task in the first d trials. This strategy may repeated k more times within a single task thus forcing $(k+1)d$ mistakes. If $k+1 = m$, we are done. Otherwise, the constraint $k + s \geq m$ implies that we may force d mistakes per task in $m - (k+1)$ other tasks. Thus after md trials, $md = \Gamma$ mistakes have been forced while maintaining the condition $m \geq |m(\boldsymbol{h}^*)|$.

Case $k \log_2 m$ is the max.

Set $d = \log_2 m$ and without loss of generality assume d is positive integer. Using $\mathcal{H}_d$ we force kd mistakes by first forcing d mistakes within a single task then "switching" $k-1$ times forcing $kd = k \log_2 m$ mistakes, while maintaining the conditions $m \geq |m(\boldsymbol{h}^*)|$ and $k \geq k(\boldsymbol{h}^*)$.

Case $s \log_2 m$ is the max. Same instance as the above case, except we force d mistakes per task. $\square$

C Proofs and Details for Section 2

For the reader's convenience, we collect some standard well-known online learning results or minor extensions thereof in this appendix.

C.1 Proof the MW Bound

The algorithm and analysis corresponds essentially to the classic weighted majority algorithm introduced in [3]. In the following, we will denote $|\mathcal{H}_{\text{fin}}|$ as n. We introduce the MW algorithm and give the corresponding regret.

Theorem 51. *For Algorithm 4, setting* $\eta = \sqrt{(2 \log n)/T}$

$$\sum_{t=1}^{T} \mathbb{E}[\mathcal{L}_{01}(y_t, \hat{y}_t)] - \mathcal{L}_{01}(y_t, h(x_t)) \leq \sqrt{2 \log(n) T} \tag{54}$$

Algorithm 4 MW Algorithm

Parameters: Learning rate η; finite hypothesis set $\{h^1, \dots, h^n\} = \mathcal{H}_{\text{fin}} \subset \{-1,1\}^{\mathcal{X}}$

Initialization: Initialize $\boldsymbol{v}_1 = \frac{1}{n}\mathbf{1}^n$

For $t = 1, \dots, T$

- Receive instance $\boldsymbol{x}_t \in \mathcal{X}$.
- Set $\boldsymbol{h}_t = (h^1(\boldsymbol{x}_t) \dots h^n(\boldsymbol{x}_t)) \in \{-1,1\}^n$.
- Predict

$$i_t \sim v_t \; ; \; \hat{y}_t \leftarrow h_t^{i_t} \, .$$

- Receive label $y_t \in \{-1,1\}$.
- Update

$$\begin{aligned}
\boldsymbol{\ell}_t &\leftarrow \frac{1}{2}|\boldsymbol{h}_t - \hat{y}_t \mathbf{1}| \\
\boldsymbol{w}_{t+1} &\leftarrow \boldsymbol{w}_t \odot \exp\left(-\eta \boldsymbol{\ell}_t\right) \\
\boldsymbol{v}_{t+1} &\leftarrow \frac{\boldsymbol{w}_{t+1}}{\sum_{i=1}^n w_{t+1,i}}
\end{aligned}$$

for any $h \in \mathcal{H}_{\text{fin}}$.

Proof. Recalling that $\boldsymbol{\ell}_t = \frac{|\boldsymbol{h}_t - \hat{y}_t \mathbf{1}|}{2}$, we have that $\boldsymbol{v}_t \cdot \boldsymbol{\ell}_t = \mathbb{E}[\mathcal{L}_{01}(y_t, \hat{y}_t)]$ and that $\boldsymbol{e}^i \cdot \boldsymbol{\ell}_t = \mathcal{L}_{01}(y_t, h^i(x_t))$. In what follows, we will therefore bound $\boldsymbol{v}_t \cdot \boldsymbol{\ell}_t - \boldsymbol{e}^i \cdot \boldsymbol{\ell}_t$. We first prove the following "progress versus regret" inequality.

$$\boldsymbol{v}_t \cdot \boldsymbol{\ell}_t - \boldsymbol{e}^i \cdot \boldsymbol{\ell}_t \leq \frac{1}{\eta}\left(d(\boldsymbol{e}^i, \boldsymbol{v}_t) - d(\boldsymbol{e}^i, \boldsymbol{v}_{t+1})\right) + \frac{\eta}{2}\sum_{i=1}^n v_{t,i}\ell_{t,i}^2. \tag{55}$$

Let $Z_t := \sum_{i=1}^n v_{t,i} \exp(-\eta \ell_{t,i})$. Defining $d(\boldsymbol{u}, \boldsymbol{v})$ as the relative entropy between $\boldsymbol{u}$ and $\boldsymbol{v} \in \Delta_n$, observe that from the algorithm

$$\begin{aligned}
d(\boldsymbol{e}^i, \boldsymbol{v}_t) - d(\boldsymbol{e}^i, \boldsymbol{v}_{t+1}) &= \sum_{j=1}^n e_j^i \log \frac{v_{t+1,j}}{v_{t,j}} \\
&= -\eta \sum_{j=1}^n e_j^i \ell_{t,j} - \log Z_t \\
&= -\eta \boldsymbol{e}^i \cdot \boldsymbol{\ell}_t - \log \sum_{i=1}^n v_{t,i} \exp(-\eta \ell_{t,i}) \\
&\geq -\eta \boldsymbol{e}^i \cdot \boldsymbol{\ell}_t - \log \sum_{i=1}^n v_{t,i}(1 - \eta \ell_{t,i} + \frac{1}{2}\eta^2 \ell_{t,i}^2) \qquad (56) \\
&= -\eta \boldsymbol{e}^i \cdot \boldsymbol{\ell}_t - \log(1 - \eta \boldsymbol{v}_t \cdot \boldsymbol{\ell}_t + \frac{1}{2}\eta^2 \sum_{i=1}^n v_{t,i}\ell_{t,i}^2) \\
&\geq \eta(\boldsymbol{v}_t \cdot \boldsymbol{\ell}_t - \boldsymbol{e}^i \cdot \boldsymbol{\ell}_t) - \frac{1}{2}\eta^2 \sum_{i=1}^n v_{t,i}\ell_{t,i}^2 \qquad (57)
\end{aligned}$$

using inequalities $e^{-x} \leq 1 - x + \frac{x^2}{2}$ for $x \geq 0$ and $\log(1+x) \leq x$ for (56) and (57) respectively.

Summing over t and rearranging we have

$$\begin{aligned}
\sum_{t=1}^m \left(\boldsymbol{v}_t \cdot \boldsymbol{\ell}_t - \boldsymbol{e}^i \cdot \boldsymbol{\ell}_t\right) &\leq \frac{1}{\eta}\left(d(\boldsymbol{e}^i, \boldsymbol{v}_1) - d(\boldsymbol{e}^i, \boldsymbol{v}_{m+1})\right) + \frac{\eta}{2}\sum_{t=1}^T \sum_{i=1}^n v_{t,i}\ell_{t,i}^2 \\
&\leq \frac{\log n}{\eta} + \frac{\eta}{2}\sum_{t=1}^T \sum_{i=1}^n v_{t,i}\ell_{t,i}^2 \qquad (58)
\end{aligned}$$

where (58) comes from noting that $d(\boldsymbol{u}, \boldsymbol{v}_1) \leq \log n$, $-d(\boldsymbol{u}, \boldsymbol{v}_{m+1}) \leq 0$, and $\sum_{t=1}^{T} \sum_{i=1}^{n} v_{t,i} \ell_{t,i}^2 \leq T$. Finally we substitute the value of η and obtain the theorem. □

C.2 Review of Reproducing Kernel Hilbert Spaces

For convenience we provide a minimal review of RKHS see [5, 56] for more details.

A real RKHS $\mathcal{H}_K$ is induced by a kernel $K : \mathcal{X} \times \mathcal{X} \to \Re$. Where K is a symmetric and positive definite function. A function K is (strictly) positive definite iff the matrix $(K(x', x''))_{x',x'' \in X}$ is (strictly) positive definite for every finite cardinality $X \subseteq \mathcal{X}$. In this paper we are only concerned with strictly positive definite kernels. The pre-Hilbert space induced by kernel K is the set $H_K := \operatorname{span}(\{K(x, \cdot)\}_{\forall x \in \mathcal{X}})$ with the inner product of $f = \sum_{i=1}^{m} \alpha_i K(x_i, \cdot)$ and $g = \sum_{j=1}^{n} \alpha'_j k(x'_j, \cdot)$ defined as $\langle f, g \rangle_K := \sum_{i=1}^{m} \sum_{j=1}^{n} \alpha_i \alpha'_j K(x_i, x'_j)$. The completion of H_K is denoted $\mathcal{H}_K$. Finally the fact that K is positive definite implies the *reproducing property:* if $f \in \mathcal{H}_K$ and $x \in \mathcal{X}$ then $f(x) = \langle f(\cdot), K(x, \cdot) \rangle_K$.

C.3 Proof of the Online Gradient Descent Regret Bound

In this section, we will prove expected regret bounds for Online Gradient Descent for both the switching and non-switching case. The proofs are adapted from the material in [8, 11, 57] (see [58] for the seminal work on worst case bounds for online gradient descent with the square loss). Recall that we wish to proof the following for the non-switching case:

$$\sum_{t=1}^{T} \mathbb{E}[\mathcal{L}_{01}(y_t, \hat{y}_t)] - \mathcal{L}_{01}(y_t, h(x_t)) \in \mathcal{O}\left(\sqrt{\|h\|_K^2 \, X_K^2 T}\right) \quad (\forall h \in \mathcal{H}_K^{(\boldsymbol{x})}) \tag{59}$$

where $X_K^2 := \max_{t \in [T]} K(x_t, x_t)$. For the switching case, we wish to prove

$$\sum_{t=1}^{T} \mathbb{E}[\mathcal{L}_{01}(y_t, \hat{y}_t)] - \mathcal{L}_{01}(y_t, h(x_t)) \in \mathcal{O}\left(\sqrt{k \max_t \|h_t\|_K^2 \, X_K^2 T}\right). \tag{60}$$

For simplicity, we prove for an arbitrary inner product $\langle \cdot, \cdot \rangle$ space with induced norm $\|\cdot\|$. The RKHS setting reduces to this setting by identifying $\boldsymbol{x} := K(x, \cdot)$, $\boldsymbol{u} := h$, and $\langle \boldsymbol{u}, \boldsymbol{x} \rangle := h(x)$.

Algorithm 5 Randomized Constrained Online Gradient Descent Algorithm

Parameters: Learning rate η, radius γ

Initialization: Initialize $\boldsymbol{w}_1 = \boldsymbol{0}$

For $t = 1, \ldots, T$

- Receive vector $\boldsymbol{x}_t$.
- Predict
$$Y_t \sim \text{UNIFORM}(-1, 1)\,;\ \bar{y}_t \leftarrow \langle \boldsymbol{w}_t, \boldsymbol{x}_t \rangle\,;\ \hat{y}_t \leftarrow \operatorname{sign}(\bar{y}_t - Y_t)\,.$$
- Receive label $y_t \in \{-1, 1\}$.
- If $\bar{y}_t y_t \leq 1$ then
$$\boldsymbol{w}_t^m \leftarrow \boldsymbol{w}_t + \eta y_t \boldsymbol{x}_t$$
$$\boldsymbol{w}_{t+1} \leftarrow P_\gamma(\boldsymbol{w}_t^m)$$
- Else $\boldsymbol{w}_t^m \leftarrow \boldsymbol{w}_t\,;\ \ \boldsymbol{w}_{t+1} \leftarrow \boldsymbol{w}_t$

In the following, we define the hinge loss $\mathcal{L}_{\text{hi}}(y', y'') = [1 - y'y'']_+$ for $y', y'' \in \Re$. We define $\boldsymbol{z}_t := -y_t \boldsymbol{x}_t [1 - y_t \langle \boldsymbol{w}_t, \boldsymbol{x}_t \rangle \geq 0] \in \nabla_{\boldsymbol{w}} \mathcal{L}_{\text{hi}}(y_t, \langle \boldsymbol{w}, \boldsymbol{x}_t \rangle)$, where $\boldsymbol{w}_t$, $\boldsymbol{x}_t$ and y_t are as defined in Algorithm 5. We denote $P_\gamma(\boldsymbol{w})$ to be the projection into the closed origin-centered ball with radius γ, so that

$$P_\gamma(\boldsymbol{w}) = \begin{cases} \boldsymbol{w} & \text{if } \|\boldsymbol{w}\| \leq \gamma \\ \gamma \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} & \text{otherwise}\,. \end{cases}$$

We also present a lemma, used as a starting point for both the switching and non-switching proofs.

Lemma 52. *For Algorithm 5 and any $\boldsymbol{u}$ lying in the convex set $\{\boldsymbol{w} : \|\boldsymbol{w}\| \leq \gamma\}$,*

$$\langle \boldsymbol{w}_t - \boldsymbol{u}, \boldsymbol{z}_t \rangle \leq \frac{1}{2\eta} \left(\|\boldsymbol{w}_t - \boldsymbol{u}\|^2 - \|\boldsymbol{w}_{t+1} - \boldsymbol{u}\|^2 + \eta^2 \|\boldsymbol{z}_t\|^2 \right)$$

Proof. Using the update rule of the algorithm, we have

$$\begin{aligned} \|\boldsymbol{w}_t^m - \boldsymbol{u}\|^2 &= \|\boldsymbol{w}_t - \eta \boldsymbol{z}_t - \boldsymbol{u}\|^2 \\ &= \|\boldsymbol{w}_t - \boldsymbol{u}\|^2 - 2\eta \langle \boldsymbol{w}_t - \boldsymbol{u}, \boldsymbol{z}_t \rangle + \eta^2 \|\boldsymbol{z}_t\|^2 \end{aligned}$$

Next note that

$$\|\boldsymbol{w}_{t+1} - \boldsymbol{u}\|^2 \leq \|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t^m\|^2 + \|\boldsymbol{w}_{t+1} - \boldsymbol{u}\|^2 \leq \|\boldsymbol{w}_t^m - \boldsymbol{u}\|^2$$

where the rightmost inequality is the Pythogorean inequality for projection onto a convex set where $\boldsymbol{w}_{t+1}$ is the projection of $\boldsymbol{w}_t^m$ on to the convex set $\{\boldsymbol{w} : \|\boldsymbol{w}\| \leq \gamma\}$ which contains $\boldsymbol{u}$. Thus,

$$\|\boldsymbol{w}_{t+1} - \boldsymbol{u}\|^2 \leq \|\boldsymbol{w}_t - \boldsymbol{u}\|^2 - 2\eta \langle \boldsymbol{w}_t - \boldsymbol{u}, \boldsymbol{z}_t \rangle + \eta^2 \|\boldsymbol{z}_t\|^2 .$$

Rearranging then results in the lemma. □

We will use the following lemma to upper bound the zero-one loss of our randomized prediction by the hinge loss.

Lemma 53. *For $y \in \{-1, 1\}$, $\bar{y} \in \Re$, $Y \sim \text{UNIFORM}(-1, 1)$, and $\hat{y} := \operatorname{sign}(\bar{y} - Y)$,*

$$2\mathbb{E}[\mathcal{L}_{01}(y, \hat{y})] \leq \mathcal{L}_{hi}(y, \bar{y}).$$

Proof. We have

$$p(\hat{y} = 1) = \begin{cases} 0 & \text{if } \bar{y} \leq -1 \\ \frac{1}{2}(1 + \bar{y}) & \text{if } -1 < \bar{y} \leq 1 \\ 1 & \text{if } \bar{y}_t > 1 \end{cases}$$

and

$$p(\hat{y} = -1) = \begin{cases} 1 & \text{if } \bar{y} \leq -1 \\ \frac{1}{2}(1 - \bar{y}) & \text{if } -1 < \bar{y} \leq 1 \\ 0 & \text{if } \bar{y} > 1. \end{cases}$$

The possible cases are as follows.

1. If $|\bar{y}| < 1$ then $2\mathbb{E}[\mathcal{L}_{01}(y, \hat{y})] = \mathcal{L}_{\text{hi}}(y, \bar{y})$. This is since if $y = 1$ then $\mathbb{E}[\mathcal{L}_{01}(y, \hat{y})] = \frac{1}{2}(1 - \bar{y})$ and $\mathcal{L}_{\text{hi}}(y, \bar{y}) = 1 - \bar{y}$. Similarly if $y = -1$ then $\mathbb{E}[\mathcal{L}_{01}(y, \hat{y})] = \frac{1}{2}(1 + \bar{y})$ and $\mathcal{L}_{\text{hi}}(y, \bar{y}) = (1 + \bar{y})$.
2. If $|\bar{y}| \geq 1$ and $\mathbb{E}[\mathcal{L}_{01}(y, \hat{y})] = 0$, then $\mathcal{L}_{\text{hi}}(y, \bar{y}) = [1 - |\bar{y}|]_+ = 0$.
3. If $|\bar{y}| \geq 1$ and $\mathbb{E}[\mathcal{L}_{01}(y, \hat{y})] = 1$ then, $\mathcal{L}_{\text{hi}}(y, \bar{y}) = [1 + |\bar{y}|]_+ \geq 2 = 2\mathbb{E}[\mathcal{L}_{01}(y, \hat{y})]$.

□

C.3.1 Non-switching bound

Lemma 54. *For Algorithm 5, given $X = \max_t \|\boldsymbol{x}_t\|$, $\|\boldsymbol{u}\| \leq U$ and $\eta = \frac{U}{X\sqrt{T}}$ we have that*

$$\sum_{t=1}^{T} \mathcal{L}_{hi}(y_t, \bar{y}_t) - \mathcal{L}_{hi}(y_t, \langle \boldsymbol{u}, \boldsymbol{x}_t \rangle) \leq \sqrt{U^2 X^2 T}\,, \tag{61}$$

for any vector $\boldsymbol{u}$.

Proof. Using the convexity of the hinge loss (with respect to its second argument), we have

$$\mathcal{L}_{\text{hi}}(y_t, \bar{y}_t) - \mathcal{L}_{\text{hi}}(y_t, \langle \boldsymbol{u}, \boldsymbol{x}_t \rangle) \leq \langle \boldsymbol{w}_t - \boldsymbol{u}, \boldsymbol{z}_t \rangle.$$

We may therefore proceed by bounding $\sum_{t=1}^{T} \langle \boldsymbol{w}_t - \boldsymbol{u}, \boldsymbol{z}_t \rangle$. Starting with Lemma 52 and summing over t, we have

$$\begin{aligned}
\sum_{t=1}^{T} \langle \boldsymbol{w}_t - \boldsymbol{u}, \boldsymbol{z}_t \rangle &\leq \frac{1}{2\eta} \left(\|\boldsymbol{w}_1 - \boldsymbol{u}\|^2 - \|\boldsymbol{w}_{T+1} - \boldsymbol{u}\|^2 + \eta^2 \sum_{t=1}^{T} \|\boldsymbol{z}_t\|^2 \right) \\
&\leq \frac{1}{2\eta} \left(\|\boldsymbol{u}\|^2 + \eta^2 \sum_{t=1}^{T} \|\boldsymbol{z}_t\|^2 \right) \qquad (62) \\
&= \frac{1}{2\eta} \|\boldsymbol{u}\|^2 + \frac{\eta}{2} \sum_{t=1}^{T} \|\boldsymbol{x}_t\|^2 \left[1 - y_t \langle \boldsymbol{w}_t, \boldsymbol{x}_t \rangle \geq 0\right] \\
&\leq \frac{1}{2\eta} \|\boldsymbol{u}\|^2 + \frac{\eta}{2} \sum_{t=1}^{T} \|\boldsymbol{x}_t\|^2 \\
&\leq \frac{1}{2\eta} U^2 + \frac{\eta}{2} X^2 T \\
&= \sqrt{U^2 X^2 T}
\end{aligned}$$

where Equation (62) results from the fact that $\boldsymbol{w}_1 = 0$. ☐

Theorem 55. *For Algorithm 5, given* $X = \max_t \|\boldsymbol{x}_t\|$, $\|\boldsymbol{u}\| \leq U$, $\eta = \frac{U}{X\sqrt{T}}$,

$$\sum_{t=1}^{T} \mathbb{E}[\mathcal{L}_{01}(y_t, \hat{y}_t)] - \mathcal{L}_{01}(y_t, \langle \boldsymbol{u}, \boldsymbol{x}_t \rangle) \leq \frac{1}{2} \sqrt{U^2 X^2 T}\,,$$

for any vector $\boldsymbol{u}$ *such that* $|\langle \boldsymbol{u}, \boldsymbol{x}_t \rangle| = 1$ *for* $t = 1, \dots, T$.

Proof. Applying the lower bound on the hinge loss from Lemma 53 to (61) gives

$$2 \sum_{t=1}^{T} \mathbb{E}[\mathcal{L}_{01}(y_t, \hat{y}_t)] \leq \sum_{t=1}^{T} \mathcal{L}_{\text{hi}}(y_t, \langle \boldsymbol{u}, \boldsymbol{x}_t \rangle) + \sqrt{U^2 X^2 T}\,,$$

observe that $\mathcal{L}_{\text{hi}}(y_t, \langle \boldsymbol{u}, \boldsymbol{x}_t \rangle) = 2\mathcal{L}_{01}(y_t, \langle \boldsymbol{u}, \boldsymbol{x}_t \rangle)$ since we have the condition $|\langle \boldsymbol{u}, \boldsymbol{x}_t \rangle| = 1$ for $t = 1, \dots, T$ dividing both sides by 2 proves the theorem. ☐

The bound for the non-switching case in (59) then follows by setting $U = ||\boldsymbol{u}||$.

C.3.2 Switching bound

Lemma 56. *For Algorithm 5, given* $X = \max_t \|\boldsymbol{x}_t\|$, $\{\boldsymbol{u}_1, \dots \boldsymbol{u}_T\} \subset \{\boldsymbol{u} : \|\boldsymbol{u}\| \leq \gamma\}$, $\eta = \frac{U}{X\sqrt{T}}$ *and* $\sqrt{\|\boldsymbol{u}_T\|^2 + 2\gamma \sum_{t=1}^{T-1} \|\boldsymbol{u}_{t+1} - \boldsymbol{u}_t\|} \leq U$, *we have that*

$$\sum_{t=1}^{T} \mathcal{L}_{hi}(y_t, \bar{y}_t) - \mathcal{L}_{hi}(y_t, \langle \boldsymbol{u}_t, \boldsymbol{x}_t \rangle) \leq \sqrt{U^2 X^2 T}.$$

Proof. Using the convexity of the hinge loss (with respect to its second argument), we have

$$\mathcal{L}_{\text{hi}}(y_t, \bar{y}_t) - \mathcal{L}_{\text{hi}}(y_t, \langle \boldsymbol{u}_t, \boldsymbol{x}_t \rangle) \leq \langle \boldsymbol{w}_t - \boldsymbol{u}_t, \boldsymbol{z}_t \rangle.$$

We may therefore proceed by bounding $\sum_{t=1}^{T} \langle \boldsymbol{w}_t - \boldsymbol{u}_t, \boldsymbol{z}_t \rangle$. Starting with Lemma 52 and summing over t, we have

$$\sum_{t=1}^{T} \langle \boldsymbol{w}_t - \boldsymbol{u}_t, \boldsymbol{z}_t \rangle \leq \frac{1}{2\eta} \sum_{t=1}^{T} \left(\|\boldsymbol{w}_t - \boldsymbol{u}_t\|^2 - \|\boldsymbol{w}_{t+1} - \boldsymbol{u}_t\|^2 + \eta^2 \|\boldsymbol{z}_t\|^2 \right) \qquad (63)$$

To transform the right hand side of the above equation into a telescoping sum, we add and subtract the term $A_t = \|\boldsymbol{w}_{t+1} - \boldsymbol{u}_t\|^2 - \|\boldsymbol{w}_{t+1} - \boldsymbol{u}_{t+1}\|^2$, giving

$$\begin{aligned}
\sum_{t=1}^{T} \|\boldsymbol{w}_t - \boldsymbol{u}_t\|^2 - \|\boldsymbol{w}_{t+1} - \boldsymbol{u}_t\|^2 &= \sum_{t=1}^{T} \|\boldsymbol{w}_t - \boldsymbol{u}_t\|^2 - \|\boldsymbol{w}_{t+1} - \boldsymbol{u}_{t+1}\|^2 - A_t \\
&= \|\boldsymbol{u}_1\|^2 - \|\boldsymbol{w}_{T+1} - \boldsymbol{u}_{T+1}\|^2 - \sum_{t=1}^{T}(\|\boldsymbol{w}_{t+1} - \boldsymbol{u}_t\|^2 - \|\boldsymbol{w}_{t+1} - \boldsymbol{u}_{t+1}\|^2) \\
&= \|\boldsymbol{u}_1\|^2 - \|\boldsymbol{w}_{T+1} - \boldsymbol{u}_T\|^2 - \sum_{t=1}^{T-1}(\|\boldsymbol{w}_{t+1} - \boldsymbol{u}_t\|^2 - \|\boldsymbol{w}_{t+1} - \boldsymbol{u}_{t+1}\|^2) \qquad (64) \\
&\le \|\boldsymbol{u}_1\|^2 - \sum_{t=1}^{T-1}(\|\boldsymbol{w}_{t+1} - \boldsymbol{u}_t\|^2 - \|\boldsymbol{w}_{t+1} - \boldsymbol{u}_{t+1}\|^2), \qquad (65)
\end{aligned}$$

where Equation (64) comes from evaluating $t = T$ in the summation.

Computing the sum, we obtain

$$\begin{aligned}
\sum_{t=1}^{T-1} \|\boldsymbol{w}_{t+1} - \boldsymbol{u}_t\|^2 - \|\boldsymbol{w}_{t+1} - \boldsymbol{u}_{t+1}\|^2 &= \sum_{t=1}^{T-1} \|\boldsymbol{u}_t\|^2 - \|\boldsymbol{u}_{t+1}\|^2 - 2\langle \boldsymbol{w}_{t+1}, (\boldsymbol{u}_t - \boldsymbol{u}_{t+1})\rangle \\
&\ge \sum_{t=1}^{T-1} \|\boldsymbol{u}_t\|^2 - \|\boldsymbol{u}_{t+1}\|^2 - 2\,\|\boldsymbol{w}_{t+1}\|\,\|\boldsymbol{u}_t - \boldsymbol{u}_{t+1}\| \\
&\ge \|\boldsymbol{u}_1\|^2 - \|\boldsymbol{u}_T\|^2 - 2\gamma \sum_{t=1}^{T-1} \|\boldsymbol{u}_t - \boldsymbol{u}_{t+1}\| \qquad (66)
\end{aligned}$$

where Equation (66) comes from $\|\boldsymbol{w}_{t+1}\| \le \gamma$, a consequence of the projection step. Substituting this back into Equations (63) and (65), we then obtain

$$\begin{aligned}
\sum_{t=1}^{T} \langle \boldsymbol{w}_t - \boldsymbol{u}_t, \boldsymbol{z}_t\rangle &\le \frac{1}{2\eta}\left(\|\boldsymbol{u}_T\|^2 + 2\gamma \sum_{t=1}^{T-1} \|\boldsymbol{u}_t - \boldsymbol{u}_{t+1}\| + \sum_{t=1}^{T} \eta^2 \|\boldsymbol{z}_t\|^2 \right) \\
&\le \frac{1}{2\eta} U^2 + \frac{\eta}{2} X^2 T. \\
&= \sqrt{U^2 X^2 T},
\end{aligned}$$

where the second inequality comes from the definitions of $\boldsymbol{z}_t$, U and X, and the equality comes from the definition of η. □

Theorem 57. *For Algorithm 5, given* $X = \max_t \|\boldsymbol{x}_t\|$, $\{\boldsymbol{u}_1, \dots \boldsymbol{u}_T\} \subset \{\boldsymbol{u} : \|\boldsymbol{u}\| \le \gamma\}$, *and* $\sqrt{\|\boldsymbol{u}_T\|^2 + 2\gamma \sum_{t=1}^{T-1} \|\boldsymbol{u}_{t+1} - \boldsymbol{u}_t\|} \le U$, *and* $\eta = \frac{U}{X\sqrt{T}}$ *we have that*

$$\sum_{t=1}^{T} \mathbb{E}[\mathcal{L}_{01}(y_t, \hat{y}_t)] - \mathcal{L}_{01}(y_t, \langle \boldsymbol{u}_t, \boldsymbol{x}_t\rangle) \le \frac{1}{2}\sqrt{U^2 X^2 T}\,,$$

for any sequence of vectors $\boldsymbol{u}_1, \dots \boldsymbol{u}_T$ such that $|\langle \boldsymbol{u}_t, \boldsymbol{x}_t\rangle| = 1$ for $t = 1, \dots, T$.

Proof. The proof follows the structure of the proof of Theorem 55 except that Lemma 56 is the base inequality. □

The bound for the switching case then follows from Theorem 57 by setting $\gamma = \max_t ||\boldsymbol{u}_t||$, and $U = \sqrt{(4k+1)\max_t ||\boldsymbol{u}_t||^2}$, noting that

$$\|\boldsymbol{u}_T\|^2 + 2\gamma \sum_{t=1}^{T-1} \|\boldsymbol{u}_{t+1} - \boldsymbol{u}_t\| \le \|\boldsymbol{u}_T\|^2 + 2\max_t ||\boldsymbol{u}_t||\,(2k\,\max_t ||\boldsymbol{u}_t||)$$

$$
\begin{aligned}
&= \|\boldsymbol{u}_T\|^2 + 4k \max_t ||\boldsymbol{u}_t||^2 \\
&\leq (4k+1) \max_t ||\boldsymbol{u}_t||^2 \\
&= U^2.
\end{aligned}
$$

This gives us a regret bound of

$$
\sum_{t=1}^{T} \mathbb{E}[\mathcal{L}_{01}(y_t, \hat{y}_t)] - \mathcal{L}_{01}(y_t, \langle \boldsymbol{u}_t, \boldsymbol{x}_t \rangle) \leq \frac{1}{2}\sqrt{(4k+1) \max_t ||\boldsymbol{u}_t||^2 X^2 T},
$$

as desired.