Adapting to non-stationarity with growing expert ensembles

Abigail Jacobs

Cosma Shalizi & Aaron Clauset

August 13, 2010

Abstract

Forecasting sequences by expert ensembles generally assumes stationary or near-stationary processes; however, in complex systems and many real-world applications, we are frequently confronted with non-stationarities, including both abrupt and gradual changes in distribution.

We present an algorithm for forecasting non-stationary time series by combining predictions of a growing ensemble of models, adding new models as new data becomes available. Our approach modifies the exponentially weighted average forecaster [7] and fixed shares forecaster [3] to let the ensemble models grow, but retains the property of performing almost as well as an oracle which knows the optimal sequence of models to use. Additionally, we relate this to recent work in sequential anomaly detection using exponential-family models [9] and to the larger context of universal prediction.

1 Introduction

Although we observe non-stationarity in times series from many contexts, from social and biological processes, in macroeconomic data, in climate systems, and in engineering (including but certainly not limited to anomaly detection), among many other settings, our tools to adapt with non-stationarity are limited. Fortunately, if we are forecasting time series, there are some cases in which we may be able to deal with non-stationarities, or changes in the underlying distribution, directly; for example, we may be able to estimate the underlying trends parametrically for detrending. Many cases resist this capability, however, and we observe this frequently in complex systems, where parametric models tend to serve as poor representations of changes in the system.

Furthermore, we are interested in approaching the prediction of time series in an online learning setting, in which we sequentially predict and observe elements of the sequence. This has gained particular attention for its application to data streams, in which we might be only able access data sequentially. Anomaly detection and coping with concept drift would then have to be dealt with in this online setting, without access to previous data, and so problems of decision-theoretic online learning have gained much recent interest.

However, rather than attempting to improve prediction of a single model, we define our work within the context of *expert ensembles*: we define group of 'expert' forecasters, and we make predictions using expert advice by aggregating our beliefs about the system and about the experts.

Standard statistical decision theory generally assumes the sequence has been generated by an easily removed and easily extrapolated deterministic trend, with stationary fluctuation around the trend. Usual coping methods for non-stationarity then involve some kind of detrending: if we can, for example, estimate a trend parametrically, we would be able to evaluate our model more easily. Instead, we abandon these assumptions, and allow for the possibility of a potentially deterministic, stochastic, or even adversarially adaptive mechanism generating our sequence. Therefore, we can't hope to guarantee a small expected error (risk) in the future, or reason about that future risk, because that would require even the existence of a trend for us to depart from.

Instead, the best we can hope to do is abandon trying to minimize risk, and instead minimize *regret*: the difference between our (the forecaster's) loss to the experts'. The goal then becomes to predict (nearly) as well as the best expert for each time; however, that would require knowing in advance which expert that was —and in that case, we would not need to consider the other experts at that time. This does, however, frame our goal: we would like our forecaster to predict nearly as well as an oracle who knows in advance the best expert for each moment in time.

In this work, we seek to address this problem, with particular attention to the problems of forecasting non-stationary sequences. We refer to non-stationarity as concept drift almost interchangeably here, largely due to the contributions of the machine learning literature to prediction with expert advice.

1.1 Prediction with expert advice

By design, work on prediction with expert advice draws from ideas in (computational) learning theory, game theory, statistics, machine learning, and information theory. At the intersection, we find different ways of framing our problem: for example, prediction with expert advice fits naturally with online learning problems as well as a repeated game. Alternatively, just as we seek to forecast nearly as well as an oracle following the best expert, we may relate this to information theory through the problem of universal prediction (see [2] for a thorough overview), where our prediction seek to follow an unknown, independent underlying distribution.

These ideas are united by their emphasis on regret, not loss, and their ability to facilitate our loosened assumptions. For example, we are able to black box our experts' predictions by referring to them as $f_{i,t}$: we need not make any assumptions as to how they came to those predictions (or

if they used outside information). Furthermore, we are able to extend past traditional settings —where the general online learning setting would likely look at the loss of one model, we are able to generalize across models and time.

By the repeated, sequential nature of the problem of prediction with expert advice, we may frame the problem as a repeated game; our assumptions fit naturally into how we define the rules. We may think of this as an interaction between two players, a forecaster and the environment, which controls both the experts' forecasts and the outcomes, and could be deterministic, stochastic, or even adversarial. Formally, we may then introduce the problem:

Game theoretic setup [1]
Players Forecaster, (potentially Adversarial) Environment
Initialize N experts; decision space D; outcome space Y; loss function ℓ
For $t = 1, 2, \ldots$

1. **Environment** chooses next outcome $y_t \in Y$ and expert advice $\{f_{i,t} \in D\}$, and reveals the expert advice to the **Forecaster**.
2. **Forecaster** chooses prediction $\hat{p}_t \in D$
3. **Environment** reveals outcome y_t
4. **Forecaster** incurs loss $\ell(\hat{p}_t, y_t)$, and each expert i incurs loss $\ell(f_{i,t}, y_t)$

This allows us to formalize our notion of regret compared to any one expert [1]: define *instantaneous regret* $r_{i,t} = \ell(\hat{p}_t, y_t) - \ell(f_{i,t}, y_t)$, and let *cumulative regret* be defined:

$$R(i_1, \ldots, i_t) = \sum_{s=1}^{t} \ell(\hat{p}_s, y_s) - \sum_{s=1}^{t} \ell(f_{i,s}, y_s),$$

where we consider the cumulative loss of the expert i to the forecaster's performance.

Our goal has been to consider the *best* expert. A simple solution would be to compare against the best-performing expert, fitting with this notion of regret. We can then compare across all experts, and redefine regret with respect to the best overall expert

$$R(i_1, \ldots, i_t) = \sum_{s=1}^{t} \ell(\hat{p}_s, y_s) - \min_{i=1,\ldots,N} \sum_{s=1}^{t} \ell(f_{i,s}, y_s).$$

However, it is not unreasonable to assume that some experts may perform better during some lengths of time than others, and that, in fact, the best expert for a given time period may be different than for other time periods. We would most expect this in cases of non-stationarity, where it would be easy to imagine that perhaps some experts adapt more quickly to the new

paradigm, they have side information that indicates a regime change, or simply that the old 'best' model has been broken.

To handle this, we would want to consider the best *sequence* of experts: we may define this as an action sequence $i_1, \ldots, i_n$ which we would want to track. Furthermore, we may generalize this to consider randomized actions drawn by the forecaster, and so we can examine expected regret across the best sequence of actions:

$$\bar{R}(i_1, \ldots, i_n) = \sum_{t=1}^{n} \bar{\ell}(\mathbf{p_t}, Y_t) - \sum_{t=1}^{n} \ell(i_t, Y_t)$$

where $\mathbf{p_t} = (p_{1,t}, \ldots, p_{N,t})$ gives distribution for random actions I_t and expected loss $\hat{\ell}(\mathbf{p_t}, Y_t) = \sum_{i=1}^{N} p_{i,t}\ell(I_t, Y_t)$.

This best sequence of actions – in one sense, known by the oracle, for us to compare performance – becomes a very useful construct, known as *tracking regret* [3]. However, even if we limit the number m of switches of expert (*size*), we still end up with combinatorially many action sequences. We will see that some forecasters can cope with these options more efficiently than others.

Up until now, we have remained vague about how the forecaster creates their predictions $\hat{p}_t$. Ideally, we want to incorporate their regret (maybe both recently or over time) into their beliefs about the experts, and aggregate these beliefs with the forecasts in order to do as well as possible, with respect to the experts. Better yet, this should be with respect to the best expert —although what the forecaster believes to be the best expert, and the truly optimal expert, need not be the same. This is especially true in the case of non-stationarity, where it would be surprising if the best expert for all time was the same expert in the ensemble. Here we present an early expert, the exponentially weighted average forecaster, developed by Littlestone and Warmuth (1994) and Vovk (1990), based on the exponential potential. Later, we prove regret bounds for this forecaster, and present a modified version based on growing ensembles.

The exponentially weighted average forecaster The exponentially weighted average forecaster [7] is derived from the exponential potential and is convenient to work with. At each time step, the weights are updated with respect to its loss in the previous round. We define the *learning rate* $\eta > 0$.

We define expert weights $w_{i,t}$ for all experts $i = 1, \ldots, N$ and instances $t = 1, \ldots, n$:

$$\begin{aligned} w_{i,t} &= \frac{e^{\eta R_{i,t-1}}}{\sum_{j=1}^{N} e^{\eta R_{j,t-1}}} \\ &= \frac{w_{i,t-1} e^{-\eta \ell(f_{i,t}, y_t)}}{\sum_{j=1}^{N} w_{j,t-1} e^{-\eta \ell(f_{i,t}, y_t)}}. \end{aligned}$$

We define the forecaster's prediction as a function of the expert predictions and the weights assigned

to those experts:

$$\begin{aligned}\hat{p}_t &= \frac{\sum_{i=1}^{N} e^{-\eta L_{i,t-1}} f_{i,t}}{\sum_{i=1}^{N} e^{-\eta L_{j,t-1}}} \\ &= \sum_{i=1}^{N} w_{i,t-1} f_{i,t}\end{aligned}$$

where the assignment and updating of expert weights corresponds to the notion of sequential belief assignment.

The fixed shares forecaster The fixed shares forecaster [3] defines a version of the exponentially weighted average forecaster, using the same conditional distribution of actions, but that avoids following a combinatorial number of experts for tracking the best expert. Additionally, we use expected regret and so evaluate across the distribution $\hat{\mathbf{p}}_\mathbf{t}$ of random actions, as we consider our sequence of actions $I_1, \ldots, I_n$ as a *randomized* sequence of action.

We define initial expert weights $w_{i,0} = \frac{1}{N}$ for all experts, and define the expert weights for all $i = 1, \ldots, N$:

$$w_{i,t} = \alpha \frac{\sum_{i=1}^{N} v_{i,t}}{N} + (1 - \alpha) v_{i,t}$$

where

$$v_{i,t} = w_{i,t-1} e^{-\eta \ell(i, Y_t)}$$

We define the forecaster's selected action I_t as drawn from the distribution, for $i = 1, \ldots, N$,

$$p_{i,t} = \frac{w_{i,t-1}}{\sum_{j=1}^{N} w_{j,t-1}}$$

This setup is taken from [1], which draws it from [3]. We next draw on related bounds for these forecasters from the literature.

1.2 Related results

Theorem (Cesa-Bianchi & Lugosi, 2006, Theorem 2.2, p. 16) *Assume that the loss function ℓ is convex in its first argument and that it takes values in $[0, 1]$. For any n and $\eta > 0$, and for all $y_1, \ldots, y_n \in Y$, the regret of the exponentially weighted average forecaster satisfies*

$$R(i_1, \ldots, i_n) = \hat{L}_n - \min_{i=1,\ldots,N} L_{i,n} \leq \frac{\ln N}{\eta} + \frac{n}{8}\eta,$$

In particular, with $\eta = \sqrt{\frac{8}{n} \ln N}$, the upper bound becomes $\sqrt{\frac{n}{2} \ln N}$.

Recall that we define η to be the learning rate and Y the space of outcomes.

In addition, these results may be easily generalized to include unbounded loss functions.

It is interesting to note that our regret bounds appear to grow with respect to N, the number of experts. Given that η is our learning rate, and that we could potentially have one expert who correctly predicted the true outcome for all time, we would be measuring our loss against the best expert; in this case, the best expert would always have zero loss. This would mean that while this expert's weights would increase as the others' decreased, but meanwhile the forecaster would not necessarily have identified the correctness of that expert, and so their prediction $\hat{p}_t$ could be heavily swayed by very diverse, if also very incorrect, expert predictions $f_{i,t}$.

Recall that in the problem of tracking the best expert, we want to choose from the best action sequence with at most m changes of expert. Herbster and Warmuth (1998) presented these results, using the fixed shares forecaster, which eliminates the computational inefficiency of keeping track of combinatorially many experts. For equivalence of the conditional distributions assigned to the action sequences in the exponentially weighted average forecaster to the fixed shares forecaster, and therefore equivalence in the following result, consider, e.g., Theorem 5.1 [1].

Theorem (as in Cesa-Bianchi & Lugosi, 2006, Theorem 5.2, pg. 104) Tracking the Best Expert *For all $n \geq 1$, the tracking regret of the fixed share forecaster satisfies*

$$\bar{R}(i_1, \ldots, i_n) \leq \frac{m+1}{\eta} \ln N + \frac{1}{\eta} \ln \frac{1}{(\alpha/N)^m (1-\alpha)^{n-m-1}} + \frac{\eta}{8} n$$

for all action sequences $i_1, \ldots, i_n$, where $m = size(i_1, \ldots, i_n)$.

Corollary (as in Cesa-Bianchi & Lugosi, 2006, Corollary 5.1, pg. 105)
For all n, m such that $0 \leq m \leq n$, if the fixed share forecaster is run with parameters $\alpha = \frac{m}{n-1}$, where for $m = 0$ we let $\alpha = 0$, and

$$\eta = \sqrt{\frac{8}{n} \left((m+1) \ln N + (n-1) H \left(\frac{m}{n-1} \right) \right)}$$

then

$$\bar{R}(i_1, \ldots, i_n) \leq \sqrt{\frac{n}{2} \left((m+1) \ln N + (n-1) H \left(\frac{m}{n-1} \right) \right)}.$$

2 Growing the Ensemble

We propose a new type of expert ensemble that grows with time, in which we introduce new experts every τ instances, initialized by training them only on the most recent epoch. Ideally, this group of experts would be more responsive to changes in the underlying distribution by capturing different subsets of the past data.

We generally consider the sequence of outcomes $y_1, y_2, \ldots, y_n$, throughout which we could expect any number of changes in underlying distribution. To grow the ensemble, we consider the partition of the sequence into epochs, each of length τ, into $y_1, y_2, \ldots, y_{q\tau}, y_{q\tau+1}, \ldots, y_{(q+1)\tau}, \ldots, y_n$. We begin with an initial number of experts (we use one, but this can easily be extended to more), and introduce some number (again, we use one, but this is can easily be extended) of experts at the beginning of each epoch. Then, by the nth round, we would expect to have one expert trained only on the most recent epoch, one expert trained on the most recent two epochs, and so on, through one expert trained over all time.

By introducing *temporally specialized* experts, we seek to gain a sensitivity to concept drift without falling entirely subject to noise. Better yet, we may be able to develop models otherwise inaccessible based on other (larger, different) subsets of data; with new data, ideally, we produce new models that could not have otherwise evolved. If there was an abrupt shift, we would be effectively guaranteed to 'catch' it within τ rounds, where the most recent experts could be most valuable, having only observed data near and after the shift; similarly, for gradual shifts, our more recent experts would have different subsets of the data, and should then adapt with the shift well. On the other hand, if we were forecasting an entirely (or mostly) stationary sequence, our oldest forecaster(s) would be well prepared, having had exposure to the most examples from that distribution.

Previous attempts have been made to develop temporally specialized experts: Kolter and Maloof (2005) introduced additive expert ensembles via the algorithm **AddExp**, through which new experts are introduced in response to errors above a certain threshold. However, this threshold is set in advance, and therefore is susceptible in particular to both adversarial and highly noisy settings, and is possibly weak in cases of non-stationarity. Cases of an adversarial mechanism or high noise could potentially lead the algorithm to produce a new expert every, or nearly every, time step, especially if the threshold was relatively low with respect to the system. Alternatively, if the threshold is sufficiently high with respect to the system, their algorithm risks underfitting the model and not catching more significant changes to the underlying distribution; similarly, by this resistance to introducing a new model, they become unable to train the new models on data that may have occurred after the break in stationarity.

To handle the efficiency problems that would come from this sudden influx of experts, Kolter and Maloof bound the number of experts available [6]. Their pruning methods, aptly named "Oldest First" and "Weakest First," however, are particularly susceptible given our assumptions: Oldest First expert removal would likely degrade performance in cases of stationarity, or other long term behavior. For maximum K experts, the most severe case could eliminate all but the experts trained on the last K instances. Similarly, Weakest First removal would be highly susceptible to attack by an adversary, who would need only to 'weaken' the best experts available.

The SEA algorithm [11] introduces an "incremental learning" algorithm: new experts are introduced with each batch, but only included in the ensemble if they perform sufficiently well. It preserves a fixed number of experts in the ensemble, and considers data in *batches*: while a batch of data (for example, a set of τ data points) is similar to our notion of epochs, there is a tradeoff

in either memory (storing and training experts over that batch set) or redefining our framework to consider subsets, rather than individual instances, of the data. This small issue aside, SEA uses unweighted expert majority voting, and is based on classifying two-class problems. Their algorithm trains experts only on single batches, so once introduced, they cannot learn (but only be evaluated against) new data. This creates weakness both by inhibiting learning new models and because they require initially strong performance, they potentially eliminate future optimal models that are not yet relevant or strong. Both of these weaknesses are of special concern in cases of non-stationarity.

Another related notion to our ensemble is algorithms that use an adaptive time window (e.g. [10]), using an adaptive amount of recent data. While these algorithms generally don't use a growing ensemble, they do take advantage of temporally-specialized experts. They maintain the size of the ensemble, and train and replace the ensemble only if the resulting ensemble would perform better. While this addresses some of the difficulties created by non-stationarity (i.e. in theory, new ensembles would be introduced during times of concept drift, and not during stationarity), this method remains susceptible in similar ways to **AddExp**. It would be susceptible to adversarial choices of outcome that rapidly and frequently replaced the ensemble, creating both computational issues and potentially eliminating optimal models (or even optimal ensembles).

While these methods and results are valuable, this is by no means exhaustive of previous applications of expert ensembles to concept drift, or even attempts towards growing, temporally specialized ensembles, nor do they fully address the problem at hand. (For example, even Kolter and Maloof's own Dynamic Weighted Majority algorithm address some of these issues, abandoning Oldest First pruning methods and developing the Weakest First method further [5]). However, they serve as instructive examples for constructing algorithms for coping with non-stationarity, and at this time we have not found an algorithm constructed as ours; additionally, our current emphasis is on theoretical bounds, although we intend to find numerical results in the future.

Furthermore, while we do not present pruning methods, we seek modest efficiency by both limiting the number of experts added in the short and long term –capping the number of total experts, bound by the horizon (number of instances) n and our choice of epoch length, but also by beginning with a small number of initial experts and increasing that slowly, subject only to the choice of epoch length.

In the context of other attempts, our framework currently assigns a fixed initial weight to new experts. Kolter and Maloof (2005) use an adaptive initial weight, giving greater weight to new experts added after large errors; pursuing adaptive initial weights could be a valuable direction to pursue. It is possible, however, that our weights would vary quickly enough to adapt to changes, such that an adaptive initial weight would not be necessary.

We present modified algorithms of the exponentially weighted average forecaster [7] and fixed shares forecaster [3] by introducing the framework of the growing ensemble. We prove that they still yield the same conditional distribution of actions, and we then prove regret bounds for those forecasters.

3 Modifying the Forecasters

To incorporate growing ensembles into previously existing frameworks, we first must introduce several new parameters:

- β, the initial weight for new experts
- N_t, the number of experts at time t
- τ, the number of instances (time steps) in an epoch

3.1 The modified exponentially weighted average forecaster

We modify the exponentially weighted average forecaster [7] to accommodate growing ensembles.

Expert weights We begin with $N_0 = 1$ initial experts.

We define the initial weights at time $t = 0$, and let initial weights $w_{i,0} = w_{1,0} = 1$. (Recall that we may generalize this to any number of initial experts).

Within epoch We assign the weights to all experts at t within an epoch (i.e., where t mod $\tau \neq 0$) as:

$$w_{i,t} = \frac{e^{\eta R_{i,t-1}}}{\sum_{j=1}^{N_{t-1}} e^{\eta R_{j,t-1}}} = \frac{w_{i,t-1}e^{-\eta \ell(f_{i,t},y_t)}}{\sum_{j=1}^{N_{t-1}} w_{j,t-1}e^{-\eta \ell(f_{j,t-1},y_t)}} \tag{1}$$

for all $i \in \{1, \ldots, N_t\}$. Observe that within epochs, as there is no new expert introduced, we have $N_t = N_{t-1}$. This formulation of the weights is very similar to the regular exponentially weighted average forecaster, but we substitute N_t for N.

Between epochs: introducing new experts At the beginning of a new epoch, we reweight the pre-existing experts and introduce a new expert with parameter β.

Since $N_t = N_{t-1} + 1$, the new expert is the N_tth expert in the ensemble.
For $i \in \{1, \ldots, N_{t-1}\}$, we define

$$w_{i,t} = \frac{(1-\beta)w_{i,t-1}e^{-\eta \ell(f_{i,t},y_t)}}{\beta + (1-\beta)\sum_{j=1}^{N_{t-1}} w_{j,t-1}e^{-\eta \ell(f_{j,t-1},y_t)}} \tag{2}$$

and for $i = N_t$, we define

$$w_{N_t,t} = \frac{\beta}{\beta + (1-\beta)\sum_{j=1}^{N_{t-1}} w_{j,t-1}e^{-\eta \ell(f_{j,t-1},y_t)}}. \tag{3}$$

Notice that we preserve the property that $\sum_{i=1}^{N_t} w_{i,t} = 1$.

Our new parameter β captures the new expert's weight, and we assign the new expert zero loss ($\ell(f_{N_t,t-1}, y_{t-1}) = 0$) for the rounds before it was introduced. Zero loss means that our equation for $w_{N_t,t}$ can hide the alternate formulation:

$$\beta = \beta e^{-\eta \ell(f_{N_t,t-1}, y_{t-1})}$$

which is useful for proving regret bounds.

The weighted average forecaster We define the prediction at time t, given by the modified exponentially weighted average forecaster, to be:

$$\hat{p}_t = \sum_{i=1}^{N_t} w_{i,t-1} f_{i,t}$$

for $t = 1, \ldots, n$. During the change of epoch, when we introduce a new expert, we capture this change in our choice of β in the initial weight $w_{N_{t-1},t-1}$.

3.2 The modified fixed shares forecaster

Recall that implementing the exponentially weighted average forecaster requires us to keep track of a combinatorial number of experts; the usual fixed shares algorithm [3] allows us to sidestep this issue.

The modified exponentially weighted average forecaster helps reduce the number of possible action sequences: there is only one possible action sequence for the first epoch, and only two experts to possibly switch between in the second epoch, and so on and so forth. However, to improve upon working with this smaller (but still combinatorial) set of action sequences, we introduce the modified fixed shares algorithm. It is worthwhile to note that, while both forecasters included sums across all N experts, the fixed shares algorithm also explicitly incorporated the number of experts in its $\frac{\alpha}{N}$ term in its calculation of expert weights.

Construction of the modified fixed shares forecaster We begin with $N_0 = 1$ initial experts.

We define the initial weights at time $t = 0$, and let initial weights $w_{i,0} = w_{1,0} = \frac{1}{N_0} = 1$.

For $\mathbf{t = 1, 2, \ldots, n}$

1. If within new epoch ($t \bmod \tau \neq 0$)
   - set $N_t = N_{t-1}$
   - draw action from distribution:

$$p_{i,t} = \frac{w_{i,t-1}}{\sum_{j=1}^{N_t} w_{j,t-1}}, i = 1, \ldots, N_t$$

If beginning of new epoch ($t \bmod \tau = 0$)

- set $N_t = N_{t-1} + 1$
- draw action from distribution:

$$p_{i,t} = \frac{(1-\beta) w_{i,t-1}}{\beta + (1-\beta) \sum_{j=1}^{N_{t-1}} w_{j,t-1}}, i = 1, \ldots, N_{t-1},$$

and

$$p_{i,t} = \frac{\beta}{\beta + (1-\beta) \sum_{j=1}^{N_{t-1}} w_{j,t-1}}, i = N_t.$$

2. obtain Y_t and compute for each $i = 1, \ldots, N_t$,

$$v_{i,t} = w_{i,t} e^{-\eta \ell(i_t, Y_t)}$$

3. for each $i = 1, \ldots, N_t$, let

$$w_{i,t} = \alpha \frac{\sum_{i=1}^{N_t} v_{i,t}}{N_t} + (1-\alpha) v_{i,t}.$$

3.3 Static regret bounds for the growing ensemble

As with the previous forecasters, our simplest notion of regret is defined using a single expert. For analogous bounds (with proof), and why they misrepresent bounds and our ensemble (with discussion), see Appendix A.

3.4 Conditional distributions of the modified forecasters

Previous results [3] allowed us to prove regret bounds identical for the exponentially weighted average forecaster and the fixed shares forecaster. This requires proof that the conditional, given the past, distributions of actions I_t are the same for both forecasters.

Theorem (adapted from Cesa-Bianchi & Lugosi, 2006, pg. 103, Theorem 5.1) *For all $\alpha \in [0, 1]$, for any sequence of n outcomes, and for all $t = 1, 2, \ldots, n$, the conditional (given the past) distribution of the action I_t, drawn at time t by the modified fixed shares forecaster with input parameter α, is the same as the conditional distribution of actions I'_t drawn at time t by the exponentially weighted forecaster run over the compound actions $(i_{1,\ldots,i_n})$ using initial weights $w_0(i_{1,\ldots,i_n})$ set with the same value of α.*

Preliminaries To prove this, we take advantage of the construction of the weights of exponentially weighted forecaster (from Cesa-Bianchi & Lugosi, 2006, pg. 103):
Define $w'_{i,t}$ to be the weight of expert i at time t for the (modified) exponentially weighted average forecaster, and $w_{i,t}$ for the modified fixed shares forecaster.
A weight at time t follows

$$w'_t(i_1, \ldots, i_n) = w'_0(i_1, \ldots, i_n) \exp\left(-\eta \sum_{s=1}^{t} \ell(i_s, Y_s)\right) \tag{4}$$

where, considering across possible future compound actions,

$$w'_{i,t} = \sum_{i_1,\ldots,i_t,i_{t+2},\ldots,i_n} w'_t(i_1, \ldots, i_t, i, i_{t+2}, \ldots, i_n)$$

for t ≥ 1 and $w'_{i,0} = \frac{1}{N}$, although recall that in our case, $w'_{i,0}$ is the weight for the single expert available, such that $w'_{1,0} = \frac{1}{N_0} = 1$.
This also lends us the form:

$$w'_t(i_1, \ldots, i_n) = w'_{t-1}(i_1, \ldots, i_n) \exp(-\eta \ell(i_s, Y_s))$$

which is trivial, but used in the proof below.

Additionally, we have the recursive definition

$$\frac{w'_0(i_1, \ldots, i_{t+1})}{w'_0(i_1, \ldots, i_t)} = \frac{\alpha}{N_{t+1}} + (1 - \alpha)\mathbf{1}_{\{i_{t+1}=i_t\}}, \tag{5}$$

because we take the conditional distribution dependent on t, we need not be concerned whether or not we are within or between epochs: we choose N_t and have either $N_{t+1} = N_t$ or $N_{t+1} = N_t + 1$ in our formulation of the distribution.

Proof To prove this, we need only show that $w'_{i,t} = w_{i,t}$ for all i and t by induction. We proceed by adapting Cesa-Bianchi & Lugosi, 2006, Theorem 5.1.

Recall that we let $w'_{i,t}$ denote the weight on the ith expert at time t using the exponentially weighted forecaster, and let $w_{i,t}$ denote the weight for that expert at time t using the fixed shares forecaster, both within the modified frameworks for growing expert ensembles.

To begin, it is obvious that $\mathbf{w}_0 = \mathbf{w}'_0$, where $\mathbf{w}'_0 = w'_{1,0} = w_{1,0} = \frac{1}{N_0} = 1$.

To proceed by induction, we assume $w_{i,s} = w'_{i,s}$ for all i and $s < t$.

From there, we may incorporate our framework throughout the proof:

$$
\begin{aligned}
w'_{i,t} &= \sum_{i_1,\ldots,i_t,i_{t+2},\ldots,i_n} w'_t(i_1,\ldots,i_t,i,i_{t+2},\ldots,i_n) \qquad (6)\\
&= \sum_{i_1,\ldots,i_t} \exp\left(-\eta\sum_{s=1}^{t}\ell(i_s,Y_s)\right) w'_0(i_1,\ldots,i_t,i)\\
&= \sum_{i_1,\ldots,i_t} \exp\left(-\eta\sum_{s=1}^{t}\ell(i_s,Y_s)\right) w'_0(i_1,\ldots,i_t)\frac{w'_0(i_1,\ldots,i_t,i)}{w'_0(i_1,\ldots,i_t)}\\
&= \sum_{i_1,\ldots,i_t} \exp\left(-\eta\sum_{s=1}^{t}\ell(i_s,Y_s)\right) w'_0(i_1,\ldots,i_t)\left(\frac{\alpha}{N_t}+(1-\alpha)\mathbf{1}_{\{i_{t-1}=i_t\}}\right)\\
&= \sum_{i_1,\ldots,i_t} w'_{t-1}(i_1,\ldots,i_t)\exp(-\eta\ell(i_t,Y_t))\left(\frac{\alpha}{N_t}+(1-\alpha)\mathbf{1}_{\{i_{t-1}=i_t\}}\right)\\
&= w'_{i_t,t-1}\exp(-\eta\ell(i_t,Y_t))\left(\frac{\alpha}{N_t}+(1-\alpha)\mathbf{1}_{\{i_{t-1}=i_t\}}\right)\\
&= w_{i_t,t-1}\exp(-\eta\ell(i_t,Y_t))\left(\frac{\alpha}{N_t}+(1-\alpha)\mathbf{1}_{\{i_{t-1}=i_t\}}\right) \qquad (7)
\end{aligned}
$$

replacing $w'_{i_t,t-1}$ with $w_{i_t,t-1}$, by $t-1<t$ and the induction assumption,

$$
\begin{aligned}
&= v_{i,t}\left(\frac{\alpha}{N_t}+(1-\alpha)\mathbf{1}_{\{i_{t-1}=i_t\}}\right) \qquad (8)\\
&= w_{i,t}
\end{aligned}
$$

by steps 2 and 3 of the modified fixed shares algorithm, which completes the proof.

3.5 Regret bounds for the modified forecasters

To define the regret bounds within our framework, we define the following:

- 1st epoch: $t = 0,\ldots,\tau-1$
- qth epoch: $t = (q-1)\tau,\ldots,(q-1)\tau+\tau-1$
- $q_{max} = \lceil\frac{n}{\tau}\rceil$, the total number of epochs
- $m_q = size(q\text{th epoch})$

  where $m_q = sizep(q)$; define $sizep(q)$ analogously to $size(i_1,\ldots,i_n)$,

  let m_q representing the number of switches of expert $(i_t \neq i_{t-1})$ within that epoch

- $\sum_{q=1}^{q_{max}} m_q = m$, as we defined previously

Recall t goes up to horizon n; where convenient, we assume n perfectly divisible by τ; this may be easily modified. This framework also lets us reason about intermediate values: for example, whether or not we allow much switching during the first epoch, if we only began with one expert, then there is no way to switch between experts during the first epoch.

Theorem (adapted from Cesa-Bianchi & Lugosi, 2006, pg. 103, Theorem 5.1) *For all $n \geq 1$, the tracking regret of the modified fixed shares forecaster satisfies*

$$R(i_1, \ldots, i_n) \leq \frac{1}{\eta} \ln \frac{1}{\alpha^m (1-\alpha)^{n-m}} + \frac{1}{\eta} \sum_{q=1}^{q_{max}} m_q \ln q + \frac{\eta}{8} n$$

for all action sequences $i_1, \ldots, i_n$, where $m = \sum_{q=1}^{q_{max}} m_q = \sum_{q=1}^{q_{max}} sizep(q)$

Preliminaries From our construction of the generic weights for the exponentially weighted forecaster, we have

$$w'_t(i_1, \ldots, i_n) = w'_0(i_1, \ldots, i_n) \exp\left(-\eta \sum_{s=1}^{t} \ell(i_s, Y_s)\right) \tag{9}$$

where action i_t is drawn with probability $\frac{w_{i,t}}{\sum_{j=1}^{N_t} w_{j,t}}$.
Additionally, we build from the following facts:

$$\ln w'_n(i_1, \ldots, i_n) = \ln w'_0(i_1, \ldots, i_n) - \eta \sum_{t=1}^{n} \ell(i_t, Y_t), \tag{10}$$

and allowing switching, let $m_q = sizep(q)$ and let $m = \sum_{q=1}^{q_{max}} m_q$. Then we have:

$$w'_0(i_1, \ldots, i_n) = \prod_{q=1}^{q_{max}} (\frac{\alpha}{q})^{m_q} (\frac{\alpha}{q} + 1 - \alpha)^{\tau - m_q} \geq \prod_{q=1}^{q_{max}} (\frac{\alpha}{q})^{m_q} (1-\alpha)^{\tau - m_q}. \tag{11}$$

Proof Adapted from Cesa-Bianchi & Lugosi, 2006, Theorem 5.1.
By construction:

$$\begin{aligned} \sum_{t=1}^{n} \ell(\mathbf{p_t}, Y_t) &\leq \frac{1}{\eta} \ln \frac{1}{W'_n} + \frac{\eta}{8} n \qquad (12) \\ &\leq \sum_{t=1}^{n} \ell(i_t, Y_t) + \frac{1}{\eta} \ln \frac{1}{w'_n(i_1, \ldots, i_n)} + \frac{\eta}{8} n \\ &\leq \sum_{t=1}^{n} \ell(i_t, Y_t) + \frac{1}{\eta} \ln \frac{1}{\prod_{q=1}^{q_{max}} (\frac{\alpha}{q})^{m_q} (1-\alpha)^{\tau - m_q}} + \frac{\eta}{8} n \end{aligned}$$

using the facts established above. Then, from the fact

$$\frac{1}{\eta}\ln\frac{1}{w'_n(i_1,\ldots,i_n)} = -\frac{1}{\eta}\left(\ln(w'_0(i_1,\ldots,i_n)) - \eta\sum_{t=1}^{n}\ell(i_t,Y_t)\right) \tag{13}$$

this lends itself to the regret bound:

$$\begin{aligned}
\sum_{t=1}^{n}\ell(\mathbf{p_t},Y_t) &\leq -\frac{1}{\eta}\left(\ln(w'_0(i_1,\ldots,i_n)) - \eta\sum_{t=1}^{n}\ell(i_t,Y_t)\right) + \frac{\eta}{8}n \qquad (14)\\
&= \sum_{t=1}^{n}\ell(i_t,Y_t) - \frac{1}{\eta}\left(\ln(\prod_{q=1}^{q_{max}}(\frac{\alpha}{q})^{m_q}(\frac{\alpha}{q}+1-\alpha)^{\tau-m_q})\right) + \frac{\eta}{8}n\\
&\leq \sum_{t=1}^{n}\ell(i_t,Y_t) - \frac{1}{\eta}\left(\ln(\prod_{q=1}^{q_{max}}(\frac{\alpha}{q})^{m_q}(1-\alpha)^{\tau-m_q})\right) + \frac{\eta}{8}n\\
&= \sum_{t=1}^{n}\ell(i_t,Y_t) - \frac{1}{\eta}\sum_{q=1}^{q_{max}}\left(m_q\ln\frac{\alpha}{q} + (\tau-m_q)\ln(1-\alpha)\right) + \frac{\eta}{8}n\\
&= \sum_{t=1}^{n}\ell(i_t,Y_t) - \frac{1}{\eta}\sum_{q=1}^{q_{max}}m_q\ln\frac{\alpha}{q} - \frac{\ln(1-\alpha)}{\eta}\sum_{q=1}^{q_{max}}(\tau-m_q) + \frac{\eta}{8}n\\
&= \sum_{t=1}^{n}\ell(i_t,Y_t) - \frac{1}{\eta}(\sum_{q=1}^{q_{max}}m_q\ln\frac{\alpha}{q}) - \frac{(n-m)(\ln(1-\alpha))}{\eta} + \frac{\eta}{8}n\\
&= \sum_{t=1}^{n}\ell(i_t,Y_t) - \frac{1}{\eta}\left(\sum_{q=1}^{q_{max}}m_q\ln\alpha - \sum_{q=1}^{q_{max}}m_q\ln q + (n-m)\ln(1-\alpha)\right) + \frac{\eta}{8}n\\
&= \sum_{t=1}^{n}\ell(i_t,Y_t) - \frac{1}{\eta}(m\ln\alpha + (n-m)\ln(1-\alpha)) + \frac{1}{\eta}\sum_{q=1}^{q_{max}}m_q\ln q + \frac{\eta}{8}n\\
&= \sum_{t=1}^{n}\ell(i_t,Y_t) + \frac{1}{\eta}\ln\frac{1}{\alpha^m(1-\alpha)^{n-m}} + \frac{1}{\eta}\sum_{q=1}^{q_{max}}m_q\ln q + \frac{\eta}{8}n \qquad (15)
\end{aligned}$$

which is analogous to the regret bounds for the original fixed shares forecaster: here we find $\frac{1}{\eta}\sum_{q=1}^{q_{max}}m_q\ln q$ term in lieu of $\frac{m+1}{\eta}\ln N$. In terms of comparison, it is useful to observe that N is analogous to q_{max}, $m=\sum_{q=1}^{q_{max}}m_q$, and:

for $\mu \geq max_{q=1,\ldots,q_{max}} m_q$,

$$\begin{aligned} \frac{1}{\eta} \sum_{q=1}^{q_{max}} m_q \ln q &\leq \frac{1}{\eta} \sum_{q=1}^{q_{max}} \mu \ln q & (16) \\ &= \frac{\mu}{\eta} \ln(q_{max}!) & (17) \end{aligned}$$

Appropriately, for $m = 0$ (no switching allowed) this reduces to a regret bound of $\frac{\eta}{8}n$, which is as we found in Appendix A, static regret bounds for the modified exponentially weighted average forecaster.

We are able to set α and η to create a desirable upper bound. We establish this in the following corollary.

3.6 Corollary: Setting η and α

Corollary (Adapted from Cesa-Bianchi & Lugosi, 2006, Corollary 5.1, pg. 105.) *For all n, m such that $0 \leq m < n$, if the modified fixed share forecaster is run with parameters $\alpha = \frac{m}{n-1}$, where for $m = 0$ we let $\alpha = 0$, and*

$$\eta = \sqrt{\frac{8}{n}\left((n-1)H(\frac{m}{n-1}) - \ln(\frac{n-m-1}{n-1}) + \sum_{q=1}^{q_{max}} m_q \ln q\right)},$$

then

$$R(i_1, \ldots, i_n) \leq \sqrt{\frac{n}{2}\left((n-1)H(\frac{m}{n-1}) - \ln(\frac{n-m-1}{n-1}) + \sum_{q=1}^{q_{max}} m_q \ln q\right)}$$

for all action sequences $i_1, \ldots, i_n$ such that $size(i_1, \ldots, i_n) \leq m$.

Preliminaries Let $\alpha = \frac{m}{n-1}$. Then:

$$\begin{aligned} \ln\left(\frac{1}{\alpha^m(1-\alpha)^{n-m}}\right) &= -m \ln \alpha - (n-m)\ln(1-\alpha) & (18) \\ &= -m \ln \frac{m}{n-1} - (n-m)\ln\left(\frac{n-m-1}{n-1}\right) \\ &\geq -m \ln \frac{m}{n-1} - (n-m-1)\ln\left(\frac{n-m-1}{n-1}\right) \\ &= (n-1)H(\frac{m}{n-1}) & (19) \end{aligned}$$

where $H(x)$ is the binary entropy function for $x \in [0,1]$, $H(x) = -x \ln x - (1-x)\ln(1-x)$. We can also keep this at equality, which we use in the proof:

$$\begin{aligned}
\ln\left(\frac{1}{\alpha^m(1-\alpha)^{n-m}}\right) &= -m\ln\alpha - (n-m)\ln(1-\alpha) \\
&= -m\ln\frac{m}{n-1} - (n-m)\ln\left(\frac{n-m-1}{n-1}\right) \\
&= -m\ln\frac{m}{n-1} - (n-m-1)\ln\left(\frac{n-m-1}{n-1}\right) - \ln\left(\frac{n-m-1}{n-1}\right) \\
&= (n-1)H(\frac{m}{n-1}) - \ln\left(\frac{n-m-1}{n-1}\right) \qquad (20)
\end{aligned}$$

Proof Substituting η into the regret bound, and using the bounds from the preliminaries, we are done.

4 Conclusion

This paper introduced a new algorithm for adapting to non-stationarity in prediction of individual sequences. We proposed a framework for growing expert ensembles, in which we add an expert every τ time steps, therefore allowing experts to become temporally specialized. We examine related algorithms, and motivate this framework in part by potential weaknesses of these other algorithms in cases of non-stationarity or an adversarial opponent. By using the idea of epochs, we were able to extend results from the exponentially weighted average forecaster and the fixed shares forecaster to find regret bounds for our modified forecasters. In particular, we were able to find regret bounds analogous to Tracking the Best Expert [3], which compares our performance to an oracle that is able to choose the best sequence of experts.

5 Future work

This problem still remains of theoretical interest. In the future, we would like to explore the regret bounds more deeply and see if they may be improved. Additionally, we would like to investigate the properties of initial weights for new experts, and consider implementing responsive initial weights (i.e. [6]). Extending our framework to include an adaptive switching rate α [8], an unknown horizon n and lower bounds over all n [1] could be a beneficial way to test our proposed algorithm.

Our next steps will include implementation, seeking numerical results from actual data, and comparing performance across non-stationary sequences (including with both gradual and abrupt changes) and stationary sequences. We want to compare this work to other models, including linear autoregressive models, and tie this work in with anomaly detection (as in Raginsky, et al. 2010).

Acknowledgements

The author would like to thank the Santa Fe Institute, the NSF, and her wonderful mentors, Cosma (especially for his patience via email) and Aaron (especially for his open door).

References

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

[2] M. Feder and N. Merhav. Universal prediction. *IEEE Transactions on Information Theory*, 1998.

[3] Mark Herbster and Manfred Warmuth. Tracking the best expert. *Machine Learning*, 32(2):151–178, 1998.

[4] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[5] J. Zico Kolter and Marcus A. Maloof. Dynamic weighted majority: An ensemble method for drifting concepts. *Journal of Machine Learning Research*, 8:2755–2790, 2007.

[6] Jeremy Z. Kolter and Marcus A. Maloof. Using additive expert ensembles to cope with concept drift. In *In Proceedings of the 22nd International Conference on Machine Learning*, pages 449–456. ACM Press, 2005.

[7] N. Littlestone and Manfred Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994.

[8] C. Monteleoni and T. Jaakkola. Online learning of non-stationary sequences. *Advances in Neural Information Processing Systems (NIPS)*, 16, 2003.

[9] Maxim Raginsky, Rebecca Willett, Corinne Horn, Jorge Silva, and Roummel Marcia. Sequential anomaly detection in the presence of noise and limited feedback. *IEEE Transactions on Information Theory*, Submitted 2010.

[10] Martin Scholz and Ralf Klinkenberg. An ensemble classifier for drifting concepts. In Porto ECML, editor, *Proceedings of the Second International Workshop on Knowledge Discovery and Data Mining (KDD)*, pages 53–64, 2005.

[11] W. N. Street and Y. Kim. A streaming ensemble algorithm (sea) for large-scale classification. In *Proceedings of the Second International Workshop on Knowledge Discovery and Data Mining (KDD)*, pages 377–382. ACM Press, 2001.

[12] Volodimir Vovk. Aggregating strategies. In *Annual Workshop on Computational Learning Theory: Proceedings of the third annual workshop on Computational learning theory*, August 1990.

Appendix A

Modifying the static expert regret bounds

By blindly extending our modified exponentially weighted average forecaster, we may solve for static regret bounds within our framework.

Theorem *Static regret bounds for the modified exponentially weighted average forecaster* (Adapted from Cesa-Bianchi & Lugosi, 2006, Theorem 2.2, p.16)

Assume that the loss function ℓ is convex in its first argument and that it takes values in $[0, 1]$. For any n and $\eta > 0$, and for all $y_1, \ldots, y_n \in Y$, the outcome space, the regret of the modified exponentially weighted average forecaster satisfies

$$\hat{L}_n - \min_{i=1,\ldots,N_n} L_{i,n} \leq \frac{\eta^2}{8} n \tag{21}$$

with respect to the single best expert.

Recall also that we originally required $\eta > 0$; however, we can still set our regret bounds to be arbitrarily small by our choice of η.

Observe that this result becomes trivial: we can always let regret be arbitrarily close to zero, by simply predicting what that one best expert is predicting. In all other contexts, we do not know what that best expert is: in this setting, however, we are comparing to our single initial expert. (Again, this can be generalized for any number N_0 of initial experts, in which case we need only add an $\ln N_0$ term). We then see that in this and the general case, we are working with an equivalent version of the standard static regret bounds, but consider only $N = N_0$, experts, the number of initial experts, over time.

This is a direct consequence of the setup of the *growing* ensemble. We cannot compare the cumulative loss of recently added experts, because it is effectively undefined for all t prior to their existence:

$$L_{i,n} = \sum_{t=1}^{n} \ell(f_{i,t}, y_t)$$

Alternatively, we may think of this as setting $\ell(f_{i,t}, y_t) = 0$ for all t prior to the introduction of the expert. Then taking the minimum across all experts is similarly meaningless.

Lemma (Hoeffding's Inequality) *Let X be a random variable with $a \leq X \leq b$. Then for any $c \in \mathbf{R}$,*

$$\ln \mathbf{E}[e^{cX}] \leq c\mathbf{E}X + \frac{c^2(b-a)^2}{8}$$

(From Cesa-Bianchi & Lugosi, 2006, Lemma 2.2, p.16, originally from [4])

Proof *Modified static regret bounds*
We define $W_t = \sum_{i=1}^{N_t} w_{i,t} = \sum_{i=1}^{N_t} e^{-\eta L_{i,t}}$ for $t \geq 1$ and $W_0 = 1$; recall that $L_{i,t}$ refers to cumulative loss, given loss function ℓ.

First, consider

$$\begin{aligned} \ln \frac{W_n}{W_0} &= \ln W_n & (22) \\ &= \ln \left(\sum_{i=1}^{N_n} e^{-\eta L_{i,n}} \right) \\ &\geq \ln \left(\max_{i=1,\ldots,N_n} e^{-\eta L_{i,n}} \right) \\ &= -\eta \min_{i=1,\ldots,N_n} L_{i,n} & (23) \end{aligned}$$

N.B. This is comparable to $\ln \frac{W_n}{W_0} \geq -\eta \min_{i=1,\ldots,N} L_{i,n} - \ln N$ for the usual forecaster. Consider what happens to $\ln W_0 = \ln N$ term; here that term disappears as $N_0 = 1$.

Next, we consider $\ln \frac{W_t}{W_{t-1}}$, but separate the cases where we are within and between epochs, and proceed with each.

Within epoch For $t \in \{1, \ldots, n\}$ such that t is in the same epoch as $t-1$, we have $N_t = N_{t-1}$, such that any sum for $i = 1, \ldots, N_t$ is equivalent to that as for $i = 1, \ldots, N_{t-1}$. We proceed as in Theorem 2.2:

$$\begin{aligned} \ln \frac{W_t}{W_{t-1}} &= \ln \frac{\sum_{i=1}^{N_t} e^{-\eta \ell(f_{i,t}, y_t)} e^{-\eta L_{i,t-1}}}{e^{\sum_{j=1}^{N_{t-1}} \eta L_{j,t-1}}} & (24) \\ &= \ln \frac{\sum_{i=1}^{N_t} w_{i,t-1} e^{-\eta \ell(f_{i,t}, y_t)}}{\sum_{j=1}^{N_{t-1}} w_{j,t-1}} \end{aligned}$$

and, using Hoeffding's inequality, the above quantity is bounded by

$$\begin{aligned} -\eta \frac{\sum_{i=1}^{N_t} w_{i,t-1} \ell(f_{i,t}, y_t)}{\sum_{j=1}^{N_{t-1}} w_{j,t-1}} + \frac{\eta^2}{8} &\leq -\eta \ell \left(\frac{\sum_{i=1}^{N_t} w_{i,t-1} f_{i,t}}{\sum_{j=1}^{N_{t-1}} w_{j,t-1}}, y_t \right) + \frac{\eta^2}{8} & (25) \\ &= -\eta \ell \left(\sum_{i=1}^{N_t} w_{i,t-1} f_{i,t}, y_t \right) + \frac{\eta^2}{8} \\ &= -\eta \ell(\hat{p}_t, y_t) + \frac{\eta^2}{8} & (26) \end{aligned}$$

using convexity of the loss function in the first argument and the definition of the modified exponentially weighted average forecaster.

Between epochs For $t \in \{1, \ldots, n\}$ such that t begins a new epoch, we have $N_t = N_{t-1} + 1$, and by our choice of initial weights for the introduction of new experts, we must modify the proof.

First observe how our forecaster's prediction looks at the beginning of the new epoch:

$$\begin{aligned} \hat{p}_t &= \sum_{i=1}^{N_t} w_{i,t-1} f_{i,t} \\ &= \beta f_{N_t,t} + (1-\beta) \sum_{i=1}^{N_{t-1}} w_{i,t-1} f_{i,t} \end{aligned}$$

although I'm afraid I have some wrong subscripts with N_t and t that need to be rechecked.

We return to the comparison of weights between time steps, but alter the numerator to include the newly added expert.

$$\begin{aligned} \ln \frac{W_t}{W_{t-1}} &= \frac{\beta + (1-\beta) \sum_{i=1}^{N_{t-1}} w_{i,t-1} e^{-\eta \ell(f_{i,t}, y_t)}}{\sum_{j=1}^{N_{t-1}} w_{j,t-1}} \\ &= \frac{\beta e^{-\eta \ell(f_{N_t}, y_t)} + (1-\beta) \sum_{i=1}^{N_{t-1}} w_{i,t-1} e^{-\eta \ell(f_{i,t}, y_t)}}{\sum_{j=1}^{N_{t-1}} w_{j,t-1}} \end{aligned} \tag{27}$$

letting $\ell(f_{N_t,t}, y_t) = 0$. We may then apply Hoeffding's inequality, so this is bounded by:

$$-\eta \frac{\beta \ell(f_{N_t,t}) + (1-\beta) \sum_{i=1}^{N_{t-1}} w_{i,t-1} \ell(f_{i,t}, y_t)}{\sum_{j=1}^{N_{t-1}} w_{j,t-1}} + \frac{\eta^2}{8} \tag{28}$$

$$\begin{aligned} &\leq -\eta \ell \left(\frac{\beta f_{N_t,t} + (1-\beta) \sum_{i=1}^{N_{t-1}} w_{i,t-1} f_{i,t}}{\sum_{j=1}^{N_{t-1}} w_{j,t-1}}, y_t \right) + \frac{\eta^2}{8} \\ &= -\eta \ell \left(\beta f_{N_t,t} + (1-\beta) \sum_{i=1}^{N_{t-1}} w_{i,t-1} f_{i,t}, y_t \right) + \frac{\eta^2}{8} \\ &= -\eta \ell(\hat{p}_t, y_t) + \frac{\eta^2}{8} \end{aligned} \tag{29}$$

which is equivalent to the bounds found before.

Combining over all t Summing over all $t = 1, \ldots, n$ using these bounds, we get

$$\ln \frac{W_n}{W_0} \leq -\eta \hat{L}_n + \frac{\eta^2}{8} n,$$

but we may substitute our bounds from before, such that

$$-\eta \min_{i=1,\dots,N_n} L_{i,n} \leq -\eta \hat{L}_n + \frac{\eta^2}{8} n$$

yielding the regret bound

$$\hat{L}_n - \min_{i=1,\dots,N_n} L_{i,n} \leq \frac{\eta^2}{8} n. \tag{30}$$