Notation	Meaning
$\mathcal{F}_h$	Set of all h^{th}-order tree experts $\mathbf{f} : \mathcal{X}^h \to \mathcal{X}$
$l^{(\mathsf{tree})}_{t,\mathbf{f}} \Big\vert L^{(\mathsf{tree})}_{t,\mathbf{f}}$	Instantaneous $\vert$ cumulative loss suffered by tree expert $\mathbf{f}$
$\mathbf{l}^{(\mathsf{tree})}_{t,\mathbf{f}} = [l^{(\mathsf{tree})}_{t,\mathbf{f}}]_{\mathbf{f}\in\mathcal{F}_D} \Big\vert \mathbf{L}^{(\mathsf{tree})}_{t,\mathbf{f}} = [L^{(\mathsf{tree})}_{t,\mathbf{f}}]_{\mathbf{f}\in\mathcal{F}_D}$	Vector of instantaneous $\vert$ cumulative losses suffered by tree experts in $\mathcal{F}_D$
$l_{t,y}$	Instantaneous loss at time t suffered by predicting $y \in \mathcal{X}$
$L_{x,t,y}$	Cumulative loss obtained by predicting $y \in \mathcal{X}$ after seeing $x \in \mathcal{X}^h$
$\widehat{F}_h(t) := \arg\min_{\mathbf{f}\in\mathcal{F}_h} L_{t,\mathbf{f}}$	Best h^{th}-order tree expert at time t
$\widehat{L}_{t,h} := L_{t,\widehat{F}_h(t)}$	Cumulative loss suffered by tree expert $\widehat{F}_h(t)$
$R_{T,h}$	Regret suffered with respect to best h^{th}-order tree expert

Table 1: Basic notation for regret minimization under contextual experts framework.

Notation	Meaning
$\eta_1^T = \{\eta_t\}_{t=1}^T$	Sequence of learning-rates used in exponential weights updates
$g : \{0, 1, \dots, D\} \to \mathbb{R}_+$	Function for prior on tree experts function of order.
$\mathbf{w}_1(g) \Big\vert \mathbf{w}_1^{(\mathsf{tree})}(g)$	Initial distribution on prediction $\vert$ choice of tree expert
$\mathbf{w}_t(\eta_t; g) \Big\vert \mathbf{w}_t^{(\mathsf{tree})}(\eta_t; g)$	Distribution at round t on prediction $\vert$ choice of tree expert
$Z(g)$	Normalizing factor for initial distribution on tree experts
$h_t(\eta_t; g) \Big\vert H_t(\eta_1^t; g)$	Instantaneous $\vert$ cumulative expected loss incurred by algorithm at time t
$\delta_t(\eta_t; g) \Big\vert \Delta_t(\eta_1^t; g)$	Instantaneous $\vert$ cumulative mixability gap of algorithm at time t
$v_t(\eta_t; g) \Big\vert V_t(\eta_1^t; g)$	Instantaneous $\vert$ cumulative variance of loss incurred by algorithm at time t

Table 2: Notation specific to algorithm ContextTreeAdaHedge.

A Main proofs of ContextTreeAdaHedge(D)

A.1 Second-order regret bound and adversarial result

We first obtain our second-order-regret bound, stated generally for a prior function $g : \{0, 1, \dots, D\} \to \mathbb{R}$. Tables 1 and 2 recap the basic notation for regret minimization and important algorithmic notation, and are useful to look at while reading the proof of the second-order bound.

Recall the expression for the computationally naive update in Equation (5):

$$w^{(\mathsf{tree})}_{t,\mathbf{f}}(\eta_t; g) = \frac{\left(\sum_{h=\mathrm{order}(\mathbf{f})}^{D} g(h)\right) e^{-\eta_t L_{t,\mathbf{f}}}}{\sum_{\mathbf{f}\in\mathcal{F}_D} \left(\sum_{h=\mathrm{order}(\mathbf{f})}^{D} g(h)\right) e^{-\eta_t L_{t,\mathbf{f}}}}.$$

and the expression for the initial distribution on tree experts based on Definition 1:

$$w^{(\mathsf{tree})}_{1,\mathbf{f}}(g) = \frac{\sum_{h=\mathrm{order}(f)}^{D} g(h)}{Z(g)}$$

where $Z(g) > 0$ is the initial normalizing factor. The explicit expression for the normalizing factor is $Z(g) = \sum_{h=0}^{D} 2^{2^h} g(h)$.

Lemma 1. ContextTreeAdaHedge(D) *with prior function* $g(\cdot)$ *obtains regret*

$$R_{T,d} \leq \left(\sqrt{V_T \ln 2} + \frac{2}{3}\ln 2 + 1\right)\left(1 + \frac{\ln\left(\frac{Z(g)}{g(d)}\right)}{\ln 2}\right)$$

for every $d \in \{0, 1, \dots, D\}$.

Proof. Recall that $\widehat{F}_d(T)$ denotes the best d^{th}-order tree expert at round T for the given loss sequence. We denote $\widehat{L}_{T,d} := L_{t,\widehat{F}_d(t)}$ as the actual loss incurred by this expert. We start with the computationally naive update in probability distribution over tree experts as in Equation (5), and the proof proceeds in a very similar manner to the variance-based regret bound for vanilla AdaHedge [DRVEGK14]. We denote

$$\begin{aligned}
h_t(\eta_t; g) &:= \langle \mathbf{w}_t(\eta_t; g),\, \mathbf{l}_t \rangle = \langle \mathbf{w}_t^{(\mathsf{tree})}(\eta_t; g),\, \mathbf{l}_t^{(\mathsf{tree})} \rangle \\
H_T(\eta_1^T; g) &:= \sum_{t=1}^{T} h_t(\eta_t; g) \\
m_t(\eta_t; g) &:= \frac{1}{\eta_t} \ln \langle \mathbf{w}_t(\eta_t; g),\, e^{-\eta_t \mathbf{l}_t} \rangle = \frac{1}{\eta_t} \ln \langle \mathbf{w}_t^{(\mathsf{tree})}(\eta_t; g),\, e^{-\eta_t \mathbf{l}_t^{(\mathsf{tree})}} \rangle \\
M_T(\eta_1^T; g) &:= \sum_{t=1}^{T} m_t(\eta_t; g).
\end{aligned}$$

Recall that the mixability gap $\delta_t(\eta_t; g) = h_t(\eta_t; g) - m_t(\eta_t; g)$ and $\Delta_T(\eta_1^T; g) = \sum_{t=1}^{T} \delta_t(\eta_t; g)$. Since the instantaneous losses are bounded between 0 and 1, it is easy to show that $0 \leq \delta_t(\eta_t; g) \leq 1$.

A standard argument tells us that

$$\begin{aligned}
R_{T,d} &= H_T(\eta_1^T; g) - L^*_{T,d} \\
&= H_T(\eta_1^T; g) - M_T(\eta_1^T; g) + M_T(\eta_1^T; g) - L^*_{T,d} \\
&= M_T(\eta_1^T; g) - L^*_{T,d} + \Delta_T(\eta_1^T; g).
\end{aligned}$$

Recall that the sequence η_1^T is decreasing as an automatic consequence of the update in Equation (2), and non-negativity of δ_t. Handling a time-varying, data-dependent learning rate is well known to be challenging [EKRG11, DRVEGK14]. We invoke a simple lemma from the original proof of AdaHedge [DRVEGK14] that helps us effectively subsitute the final learning rate.

Lemma 2 ([DRVEGK14])**.** *For any exponential-weights update with a decreasing learning rate η_1^T and prior function $g(\cdot)$, we have $M_T(\eta_1^T; g) \leq M_T(\{\eta_T\}_{t=1}^T; g)$.*

Thus, we get

$$R_{T,d} \leq M_T(\{\eta_T\}_{t=1}^T; g) - L^*_{T,d} + \Delta_T(\eta_1^T; g). \tag{15}$$

We also have the following simple intermediate result for $M_T(\{\eta_T\}_{t=1}^T; g)$, which is simply a slightly more general version of the lemma in [DRVEGK14] that can apply to non-uniform priors.

Lemma 3.

$$M_T(\{\eta_T\}_{t=1}^T; g) \leq L^*_{T,d} + \frac{1}{\eta_T} \ln\left(\frac{Z(g)}{g(d)}\right).$$

Proof. We note that

$$\langle \mathbf{w}_1^{(\mathsf{tree})}(g),\, e^{-\eta_T \mathbf{L}_T^{(\mathsf{tree})}} \rangle \geq w^{(\mathsf{tree})}_{1,f^*_{T,d}}(g) e^{-\eta_T L^*_{T,d}}.$$

Because the initial distribution $\mathbf{w}_1^{(\mathsf{tree})}$ is normalized to sum to 1, a simple telescoping argument can be used to give $M_T(\{\eta_T\}_{t=1}^T; g) = \sum_{t=1}^{T} m_t(\{\eta_T\}_{t=1}^T; g) = -\frac{1}{\eta_T} \ln\left(\langle \mathbf{w}_1^{(\mathsf{tree})}(g),\, e^{-\eta_T \mathbf{L}_T^{(\mathsf{tree})}} \rangle\right)$.

This automatically tells us that

$$\begin{aligned}
M_T(\{\eta_T\}_{t=1}^T; g) &= -\frac{1}{\eta_T}\ln\left(\langle \mathbf{w}_1^{(\mathsf{tree})}(g),\, e^{-\eta_T \mathbf{L}_T^{(\mathsf{tree})}}\rangle\right) \\
&\le -\frac{1}{\eta_T}\ln(w_{1,f_{T,d}^*}^{(\mathsf{tree})}(g)) + L_{T,d}^* \\
&= L_{T,d}^* + \frac{1}{\eta_T}\ln\left(\frac{1}{w_{1,f_{T,d}^*}^{(\mathsf{tree})}(g)}\right) \\
&= L_{T,d}^* + \frac{1}{\eta_T}\ln\left(\frac{Z(g)}{\sum_{h=d}^D g(h)}\right) \\
&\le L_{T,d}^* + \frac{1}{\eta_T}\ln\left(\frac{Z(g)}{g(d)}\right)
\end{aligned}$$

thus proving the lemma. □

Now, Equation (15) and Lemma 3 together with the definition of η_t in Equation (2) give us

$$\begin{aligned}
R_{T,d} &\le \frac{1}{\eta_T}\ln\left(\frac{Z(g)}{g(d)}\right) + \Delta_T(\eta_1^T; g) \\
&= \frac{\ln\left(\frac{Z(g)}{g(d)}\right)}{\ln 2}\Delta_{T-1}(\eta_1^{T-1}; g) + \Delta_T(\eta_1^T; g).
\end{aligned}$$

From non-negativity of δ_t, we have $\Delta_{T-1}(\eta_1^T; g) \le \Delta_T(\eta_1^T; g)$ and so

$$R_{T,d} \le \Delta_T(\eta_1^T; g)(1 + \frac{\ln\left(\frac{Z(g)}{g(d)}\right)}{\ln 2}). \tag{16}$$

It now remains to bound the quantity Δ_T in terms of variance. In fact, it will be useful to define slightly more generic quantities

$$\begin{aligned}
\Delta_{T_0}^T(\eta_{T_0}^T; g) &:= \sum_{t=T_0}^T \delta_t(\eta_t; g) \\
V_{T_0}^T(\eta_{T_0}^T; g) &:= \sum_{t=T_0}^T v_t(\eta_t; g) \text{ where} \\
v_t(\eta_t; g) &:= \mathrm{var}_{K_t \sim \mathbf{w}_t(\eta_t; g)}\left[l_{t,K_t}\right].
\end{aligned}$$

The bound is described below.

Lemma 4. *We have*

$$\Delta_{T_0}^T(\eta_{T_0}^T; g) \le \sqrt{V_{T_0}^T(\eta_{T_0}^T; g)\ln 2} + \left(\frac{2}{3}\ln 2 + 1\right).$$

Proof. The argument is similar to the original AdaHedge proof [DRVEGK14] and proceeds below. We use a telescoping sum to get

$$\begin{aligned}
\Big(\Delta_{T_0}^{T}\big(\eta_{T_0}^{T};g\big)\Big)^2 &= \sum_{t=T_0+1}^{T} \Big(\Delta_{T_0}^{t}\big(\eta_{T_0}^{t};g\big)\Big)^2 - \Big(\Delta_{T_0}^{t-1}\big(\eta_{T_0}^{t-1};g\big)\Big)^2 \\
&= \sum_{t=T_0}^{T} \Big(\Delta_{T_0}^{t-1}\big(\eta_{T_0}^{t-1};g\big) + \delta_t\big(\eta_t;g\big)\Big)^2 - \Big(\Delta_{T_0}^{t-1}\big(\eta_{T_0}^{t-1};g\big)\Big)^2 \\
&= \sum_{t=T_0}^{T} 2\delta_t\big(\eta_t;g\big)\Delta_{T_0}^{t-1}\big(\eta_{T_0}^{t-1};g\big) + \Big(\delta_t\big(\eta_t;g\big)\Big)^2 \\
&\leq \sum_{t=T_0}^{T} 2\delta_t\big(\eta_t;g\big)\Delta_{t-1}\big(\eta_1^{t-1};g\big) + \Big(\delta_t\big(\eta_t;g\big)\Big)^2 \\
&= \sum_{t=T_0}^{T} 2\delta_t\big(\eta_t;g\big)\frac{\ln 2}{\eta_t} + \Big(\delta_t\big(\eta_t;g\big)\Big)^2) \\
&\leq \sum_{t=T_0}^{T} 2\delta_t\big(\eta_t;g\big)\frac{\ln 2}{\eta_t} + \delta_t\big(\eta_t;g\big) \text{ since } \delta_t(\eta_t;g) \leq 1 \\
&\leq (2\ln 2)\sum_{t=T_0}^{T} \frac{\delta_t(\eta_t;g)}{\eta_t} + \Delta_{T_0}^{T}\big(\eta_{T_0}^{T};g\big).
\end{aligned}$$

We also recall the following lemma from the original proof of AdaHedge [DRVEGK14]. The proof of this lemma involves a Bernstein tail bounding argument.

Lemma 5 ([DRVEGK14])**.** *We have*

$$\frac{\delta_t\big(\eta_t;g\big)}{\eta_t} \leq \frac{1}{2}v_t\big(\eta_t;g\big) + \frac{1}{3}\delta_t\big(\eta_t;g\big).$$

Using Lemma 5, we then get

$$\Big(\Delta_{T_0}^{T}\big(\eta_{T_0}^{T};g\big)\Big)^2 \leq V_{T_0}^{T}\big(\eta_{T_0}^{T};g\big)\ln 2 + \left(\frac{2}{3}\ln 2 + 1\right)\Delta_{T_0}^{T}\big(\eta_{T_0}^{T};g\big) \tag{17}$$

which is an inequality for the quantity $\Delta_{T_0}^{T}\big(\eta_{T_0}^{T};g\big)$ in quadratic form. We now solve Equation (17), and use Fact 2 from Appendix D to get

$$\Delta_{T_0}^{T}\big(\eta_{T_0}^{T};g\big) \leq \sqrt{V_{T_0}^{T}\big(\eta_{T_0}^{T};g\big)\ln 2} + \frac{2}{3}\ln 2 + 1. \tag{18}$$

□

Now we complete the proof of Lemma 1 by combining Equations (16) and (18) for the special case of $T_0 = 1$. □

Now, noting that $V_T(\eta_1^T;g) \leq \frac{T}{4}$ and substituting the expression for $g = g_{\mathsf{prop}}$ from Equation (11) directly proves Equation (12) from Lemma 1. To see this, we substitute $g = g_{\mathsf{prop}}$ into the statement of

Lemma 1 to get

$$
\begin{aligned}
R_{T,d} &\leq \left(\sqrt{V_T(\eta_1^T; g)\ln 2} + \frac{2}{3}\ln 2 + 1\right)\left(1 + \frac{\ln\left(\frac{Z(g_{\mathsf{prop}})}{g_{\mathsf{prop}}(d)}\right)}{\ln 2}\right) \\
&= \left(\sqrt{V_T(\eta_1^T; g)\ln 2} + \frac{2}{3}\ln 2 + 1\right)\left(1 + \frac{\ln\left(\frac{\sum_{h=0}^{D} 2^{2^h} 2^{-2^{h+1}}}{2^{-2^{d+1}}}\right)}{\ln 2}\right) \\
&= \left(\sqrt{V_T(\eta_1^T; g)\ln 2} + \frac{2}{3}\ln 2 + 1\right)\left(1 + \frac{\ln\left(\frac{\sum_{h=0}^{D} 2^{-2^h}}{2^{-2^{d+1}}}\right)}{\ln 2}\right) \\
&\leq \left(\sqrt{V_T(\eta_1^T; g)\ln 2} + \frac{2}{3}\ln 2 + 1\right)\left(1 + \frac{\ln\left(2 \cdot 2^{2^{d+1}}\right)}{\ln 2}\right) \\
&= \left(\sqrt{V_T(\eta_1^T; g)\ln 2} + \frac{2}{3}\ln 2 + 1\right)\left(2 + 2^{d+1}\right) \\
&\leq \left(\frac{1}{2}\sqrt{T\ln 2} + \frac{2}{3}\ln 2 + 1\right)\left(2 + 2^{d+1}\right)
\end{aligned}
$$

which is precisely Equation (12) when expressed in big-$\mathcal{O}$ notation.

A.2 Exploiting stochasticity

To effectively bound regret for the “easier" stochastic instances, we need finer control on the cumulative mixability gap term $\Delta_T(\eta_1^T; g)$. Our starting point is the following thresholding lemma.

Lemma 6. *Fix $t_0 > 0$. Let $T_0 := \max\{0 < t \leq T : \eta_t > \frac{\ln 2}{t_0}\}$. Then, we have*

$$
\Delta_T(\eta_1^T; g) \leq t_0 + 1 + \sqrt{V_{T_0}^T(\eta_{T_0}^T; g)\ln 2} + \frac{2}{3}\ln 2 + 1. \tag{19}
$$

Proof. From the definition of T_0, we observe that

$$
\begin{aligned}
\eta_{T_0} &= \frac{\ln 2}{\Delta_{T_0-1}(\eta_1^{T_0-1}; g)} > \frac{\ln 2}{t_0} \\
\implies \Delta_{T_0-1}(\eta_1^{T_0-1}; g) &< t_0 \\
\implies \Delta_{T_0}(\eta_1^{T_0}; g) &< t_0 + 1.
\end{aligned}
$$

Then, using $\Delta_T(\eta_1^T; g) = \Delta_{T_0}(\eta_1^{T_0}; g) + \Delta_{T_0}^T(\eta_{T_0}^T; g)$ and Lemma 4 directly gives us the statement in Equation (19) and completes the proof. □

We observe that the threshold T_0 depends on the choice of t_0 as well as the data (in fact, it is a random variable when the process $\{(X_t, Y_t)\}_{t=1}^T$ is stochastic). We have the freedom to choose $t_0 > 0$ for our analysis. Conceptually, in the stochastic regime, the choice of t_0 thresholds the number of rounds T_0 below which we can make few, if any, statistical guarantees, and will become clear in subsequent sections. Effectively, Lemma 6 uses the elegant inverse relationship between learning rate and mixability (in Equation (2)) to show that a minimal amount of regret, precisely, in terms of t_0, is accumulated *even before we can make high-probability statistical guarantees.*

Notation	Meaning/Interpretation
$N_t(x(h))$	Appearance frequency of a sub-context $x(h) \in \mathcal{X}^h$
$\widehat{P}_t(h\|x(h))$	Fraction of times that we observed $X_t(h) = x(h), Y_t = y$
$S_{t,h}$	Number-of-seen sub-contexts of length h at time t
$\widehat{\pi}_h(t)$	Estimated unpredictability based on h^{th}-order tree expert predictors
$D_t(h)$	Gap between correct and incorrect predictors at time t
$\mathbf{w}_t^{(h)}$	Probability distribution on predictions
$v_t^{(h)}$	Variance of loss of $\textsc{ContextTreeAdaHedge}(h)$ with uniform prior at time t
$q_t(h) \propto Q_t(h)$	Posterior probability that the h^{th}-order model is the right model
d	True model order of data $(X_t, Y_t)_{t=1}^T$
$Q_h^*(\cdot), h \leq d$	Marginal distribution on $X_t(h), h \leq d$
$P^*(\cdot\|x(h))$	Conditional distribution on Y_t given $X_t = x(h)$
$\beta(x(d)), \beta^*$	Average prediction accuracy with conteext $x(d)$
$\pi_h^*, h \leq D$	Asymptotic unpredictability under h^{th}-order model.
$t_{\mathsf{high}}(h), h > d$	Number of epochs of $x(h) \in \mathcal{X}^h$ after which we can guarantee a unique best predictor
$t_{\mathsf{low}}(h), h \leq d$	Number of rounds after which we can conclusively rule out lower h^{th}-order model

Table 3: Notation for analysis.

A.2.1 Notation for contextual prediction

First, we define a couple of convenient counts for the number of appearances of a particular context, and the number of contexts that have so far appeared.

Definition 8. *The* ***appearance frequency*** *of a particular context $x(h) \in \mathcal{X}^h$ at time t is given by*

$$N_t(x(h)) := \sum_{s=1}^{t-1} \mathbb{I}[X_s(h) = x(h)],$$

The ***fraction of times the value*** *$y \in \mathcal{X}$ seen after a particular context is given by*

$$\widehat{P}_t(y|x(h)) := \frac{\sum_{s=1}^{t-1} \mathbb{I}[X_s(h) = x(h), Y_s = y]}{\sum_{s=h}^{t-1} \mathbb{I}[X_s(h) = x(h)]}$$
$$\Big(= 1 - \frac{L_{x(h),t-1,y}}{N_t(x(h))} \Big)$$

The ***number-of-seen-contexts*** *is given by*

$$S_{t,h} := \sum_{x(h) \in \mathcal{X}^h} \mathbb{I}[N_t(x(h)) > 0].$$

Next, we define our estimates for unpredictability, effectively an estimate for the approximation error, under various model orders.

Definition 9 ([FMG92])**.** *For every value of $h \geq 0$ and a sequence $\{(X_t, Y_t)\}_{t \geq 1}$, we define its estimated unpredictability*

$$\begin{aligned}\widehat{\pi}_h(t) &:= \sum_{x(h) \in \mathcal{X}^h} \frac{N_t(x(h))}{t} \left(1 - \max_{y \in \mathcal{X}} \{\widehat{P}_t(y|x(h))\}\right) \\ &= \sum_{x(h) \in \mathcal{X}^h} \frac{1}{t} \min_{y \in \mathcal{X}} \{L_{x(h),t,y}\}.\end{aligned}$$

This definition is inspired by the information-theoretic perspective on universal sequence prediction [FMG92]. In this line of work, the quantity $\widehat{\pi}_h(t)$ represents the estimated unpredictability of a binary sequence under a h-memory Markov model. This is the natural estimate of *approximation error of the h^{th}-order model* that is used to carry out data-driven model selection.

Finally, we denote the *true prediction* (the one we would make if we had oracle knowledge of the best predictor $f^*(\cdot)$) as

$$Y_t^* := f^*(X_t(d)).$$

Then, for every $h \geq d$ we define

$$D_t(h) := L_{X_t(h),t,1-Y_t^*} - L_{X_t(h),t,Y_t^*} \tag{20}$$

represents the "gap" between the correct predictor Y_t^* and the worse predictor $1 - Y_t^*$ at time t, and pertaining to the current context $X_t(h)$.

A.2.2 Explicit model selection

We have stated the problem of wanting to exploit the structure of a d^{th}-order stochastic sequence $\{(X_t, Y_t)\}_{t\geq 1}$ in an online fashion, as a model selection problem. This has been implicitly clear in the choice of prior function in Equation (11): more complex experts are downweighted. Now, we make the connection clear.

As a reminder, we evaluate the performance of the algorithm ContextTreeAdaHedge(D) with prior function $g_{\mathsf{prop}}(\cdot)$), and using Equation (18) as a jumping point, we are concerned with bounding the cumulative variance $V_{T_0}^T(\eta_{T_0}^T; g)$.

First, we observe that

$$\begin{aligned}
V_{T_0}^T(\eta_{T_0}^T; g_{\mathsf{prop}}) &= \sum_{t=T_0}^{T} v_t(\eta_t; g_{\mathsf{prop}}) \\
&= \sum_{t=T_0}^{T} w_{t,Y_t^*}(\eta_t; g_{\mathsf{prop}}) \left(1 - w_{t,1-Y_t^*}(\eta_t; g_{\mathsf{prop}})\right) \text{ since } l_{t,K_t} \text{ i.i.d } \sim \mathrm{Ber}(w_{t,1}) \\
&\leq \sum_{t=T_0}^{T} w_{t,1-Y_t^*}(\eta_t; g_{\mathsf{prop}})
\end{aligned}$$

and thus, it is sufficient to control the evolution of the term $w_{t,1-Y_t^*}(\eta_t; g_{\mathsf{prop}})$ with t. This is the probability with which we select the prediction $1 - Y_t^*$ that is more likely to be wrong under the stochastic model for the data.

The first step is to express the update in this probability in terms of a posterior probability on the effective *order of the model* the algorithm is selecting. Explicitly, we can re-write Equation (6a) as

$$w_{t,1-Y_t^*}(\eta_t; g_{\mathsf{prop}}) = \sum_{h=0}^{D} q_t(h; \eta_t, g_{\mathsf{prop}}) w_{t,1-Y_t^*}^{(h)}(\eta_t)$$

where we have defined the shorthand notation for the update used by ContextTreeAdaHedge(h) with uniform prior,

$$w_{t,1-Y_t^*}^{(h)}(\eta_t) := w_{t,1-Y_t^*}(\eta_t; g_{\mathsf{unif}}) = \frac{e^{-\eta_t D_t(h)}}{1 + e^{-\eta_t D_t(h)}},$$

where $D_t(h)$ is according to Equation (20) and the quantities $\{q_t(h; \eta_t, g_{\mathsf{prop}})\}$ are explicitly written as

$$q_t(h; \eta_t, g_{\mathsf{prop}}) \propto Q_t(h; \eta_t, g_{\mathsf{prop}}) := g_{\mathsf{prop}}(h) \prod_{x(h)\in\mathcal{X}^h} \left(\sum_{y\in\mathcal{X}} e^{-\eta_t L_{x(h),t,y}} \right) \tag{21}$$

where the proportionality constant is set such that $\sum_{h'=0}^{D} q_t(h; \eta_t, g_{\mathsf{prop}}) = 1$. The quantity $q_t(h; \eta_t, g_{\mathsf{prop}})$ is exactly the *posterior probability* that the algorithm ContextTreeAdaHedge(D) selects a h^{th}-order model. We will see that controlling the posterior on model order selection is crucial to bounding the variance in our desired manner.

First, we state a simple lemma that bounds Equation (21) in terms of more intuitive quantities.

Lemma 7. *We have*

$$\exp\{-\eta_t \widehat{\pi}_h(t)t + \ln g_{\mathsf{prop}}(h)\} \leq Q_t(h; \eta_t, g_{\mathsf{prop}}) \leq \exp\{-\eta_t \widehat{\pi}_h(t)t + 2^h \ln 2 + \ln g_{\mathsf{prop}}(h)\}. \tag{22}$$

Proof. For the upper bound, we have

$$\begin{aligned}
Q_t(h; \eta_t, g_{\mathsf{prop}}) &:= g_{\mathsf{prop}}(h) \prod_{x(h) \in \mathcal{X}^h} \left(\sum_{y \in \mathcal{X}} e^{-\eta_t L_{x(h),t,y}} \right) \\
&= \exp\left\{ \sum_{x(h) \in \mathcal{X}^h} \ln \left(\sum_{y \in \mathcal{X}} e^{-\eta_t L_{x(h),t,y}} \right) + \ln g_{\mathsf{prop}}(h) \right\} \\
&\leq \exp\left\{ \sum_{x(h) \in \mathcal{X}^h} \ln \left(2 e^{-\eta_t \min_{y \in \mathcal{X}} \{L_{x(h),t,y}\}} \right) + \ln g_{\mathsf{prop}}(h) \right\} \\
&= \exp\left\{ - \sum_{x(h) \in \mathcal{X}^h} \eta_t \min_{y \in \mathcal{X}} \{L_{x(h),t,y}\} + 2^h \ln 2 + \ln g_{\mathsf{prop}}(h) \right\} \\
&= \exp\left\{ -\eta_t \widehat{\pi}_h(t)t + 2^h \ln 2 + \ln g_{\mathsf{prop}}(h) \right\}
\end{aligned}$$

and for the lower bound, we have

$$\begin{aligned}
Q_t(h; \eta_t, g_{\mathsf{prop}}) &:= \exp\left\{ \sum_{x(h) \in \mathcal{X}^h} \ln \left(\sum_{y \in \mathcal{X}} e^{-\eta_t L_{x(h),t,y}} \right) + \ln g_{\mathsf{prop}}(h) \right\} \\
&\geq \exp\left\{ \sum_{x(h) \in \mathcal{X}^h} \ln \left(e^{-\eta_t \min_{y \in \mathcal{X}} \{L_{x(h),t,y}\}} \right) + \ln g_{\mathsf{prop}}(h) \right\} \\
&= \exp\left\{ - \sum_{x(h) \in \mathcal{X}^h} \eta_t \min_{y \in \mathcal{X}} \{L_{x(h),t,y}\} + \ln g_{\mathsf{prop}}(h) \right\} \\
&= \exp\left\{ -\eta_t \widehat{\pi}_h(t)t + \ln g_{\mathsf{prop}}(h) \right\}
\end{aligned}$$

□

Substituting $\ln g_{\mathsf{prop}}(h) = -2^{h+1} \ln 2 = -2 \cdot 2^h \ln 2$, we get

$$\exp\{-\eta_t \widehat{\pi}_h(t)t - 2 \cdot 2^h \ln 2\} \leq Q_t(h; \eta_t, g_{\mathsf{prop}}) \leq \exp\{-\eta_t \widehat{\pi}_h(t)t - 2^h \ln 2\}. \tag{23}$$

Equation (23) effectively makes the tradeoff between approximation error (reflected by the quantity $\widehat{\pi}_h(t)$) and model complexity (reflected by the quantity $2^h \ln 2$ clear in the model-order selection problem. We can think of the model orders as “meta-experts" that are being randomized over. Note that the learning rate that is being used to randomize their selection is still η_t!

A.2.3 Analysis for a higher-than-needed model order

Here, we analyze the contribution of a specific selected model order to the variance, an important intermediate step. Formally, we consider the algorithm ContextTreeAdaHedge(h) equipped with the uniform prior function $g_{\mathsf{unif}}(h') = \mathbb{I}[h' = h]$. The regret guarantee is given by the following proposition.

Proposition 1. 1. *For any sequence $\{X_t, Y_t\}_{t=1}^T$ the algorithm* ContextTreeAdaHedge(h) *with uniform prior gives us regret rate*

$$R_{T,d} = \mathcal{O}\left(\sqrt{T}2^h\right) \tag{24}$$

with respect to the best d^{th}-order tree expert in hindsight, and for every $d \leq h$.

2. ContextTreeAdaHedge(h) *with uniform prior gives regret with probability greater than $(1-\epsilon)$:*

$$R_{T,d} = \mathcal{O}\Big(\frac{2^{2h}}{(2\beta^* - 1)^2}\left(h + \ln\left(\frac{1}{\epsilon(2\beta^* - 1)}\right)\right)\Big).$$

on a sequence $(X_t, Y_t)_{t \geq 1}$ that satisfies the d^{th}-order stochastic condition with parameter β^*.

Observe the suboptimal scaling in terms of 2^{2h} in the regret bound for the case where $d < h$. We now proceed to prove Proposition 1.

Formally, the algorithm ContextTreeAdaHedge(h) equipped with the uniform prior function $g_{\mathsf{unif}}(h') = \mathbb{I}[h' = h]$ gives us $q_t(h'; \eta_t, g_{\mathsf{unif}}) = \mathbb{I}[h' = h]$, and we would get

$$\begin{aligned}
\sum_{t=1}^{T}\sum_{h'=0}^{D} q_t(h'; \eta_t, g_{\mathsf{unif}}) w^{(h)}_{t,1-Y_t^*}(\eta_t) &= \sum_{t=1}^{T} w^{(h)}_{t,1-Y_t^*} \\
&= \sum_{t=1}^{T} \frac{e^{-\eta_t D_t(h)}}{1 + e^{-\eta_t D_t(h)}} \\
&\leq \sum_{t=1}^{T} \min\{e^{-\eta_t D_t(h)}, 1\} \\
&\leq \sum_{t=1}^{T} \min\{e^{-\eta_T D_t(h)}, 1\}
\end{aligned}$$

where $D_t(h)$ is the gap between predictions as in Equation (20), and the last inequality is because η_1^T is a decreasing sequence according to the update in Equation (2).

Therefore, we have

$$V_T(\eta_1^T; g_{\mathsf{unif}}) \leq \sum_{t=1}^{T} \min\{e^{-\eta_t D_t(h)}, 1\}. \tag{25}$$

We observe that Equation (25) can be effectively unraveled to get a closed-form variance bound for particular evolutions of $\{D_t(h)\}_{t \geq 1}$. Particularly, we care about $D_t(h)$ as a function of $N_t(X_t(h))$, the number of appearances so far of the current context. We show this result in the following lemma.

Lemma 8. *Let the following condition hold for some $t_0(h) > 0$ and $\alpha > 0$.*

$$D_t(h) \geq \alpha N_t(X_t(h)) \text{ for all } t \text{ such that } N_t(X_t(h)) \geq t_0(h) \tag{26}$$

for some $\alpha > 0$.
Then, we have

$$\sum_{t=1}^{\infty} w^{(h)}_{t,1-Y_t^*}(\eta_t) \leq 2^h \left(t_0(h) + \frac{1}{\eta_T \alpha}\right). \tag{27}$$

Proof. We can directly use the condition in Equation (26). For values of t such that $N_t(X_t(h)) < t_0(h)$, we apply $w^{(h)}_{t,1-Y_t^*}(\eta_t) \leq 1$. Otherwise, we use $w^{(h)}_{t,1-Y_t^*}(\eta_t) \leq e^{-\eta_T \alpha N_t(X_t(h))}$.

Combining the two gives us

$$
\begin{aligned}
\sum_{t=1}^{\infty} w_{t,1-Y_t^*}^{(h)}(\eta_t) &\leq \sum_{x(h)\in\mathcal{X}^h} \left(t_0 + \sum_{s=t_0(h)}^{N_T(x(h))} e^{-\eta_T \alpha s} \right) \\
&\leq 2^h t_0(h) + \sum_{x(h)\in\mathcal{X}^h} \sum_{s=t_0(h)}^{\infty} e^{-\eta_T \alpha s} \\
&\leq 2^h \left(t_0(h) + \sum_{s=t_0(h)}^{\infty} e^{-\eta_T \alpha s} \right) \\
&\leq 2^h \left(t_0(h) + \frac{e^{-\eta_T \alpha}}{1 - e^{-\eta_T \alpha}} \right).
\end{aligned}
$$

Now, we have

$$
\begin{aligned}
\frac{e^{-\eta_T \alpha}}{1 - e^{-\eta_T \alpha}} &= \frac{1}{e^{\eta_T \alpha} - 1} \\
&\leq \frac{1}{\eta_T \alpha}
\end{aligned}
$$

by the inequality $e^a \geq 1 + a$ for $a \geq 0$. Substituting this above gives us our required result. □

It remains to show that the condition in Equation (26) is met with high probability for $(X_t, Y_t)_{t\geq 1}$ satisfying the d^{th}-order *realizability condition* with paramter β^*, and for any $d \leq h$. We use a standard Hoeffding-bounding technique to show this.

Lemma 9. *Let $\epsilon \in (0, 1]$. For a process $\{(X_t, Y_t)\}_{t\geq 1}$ satisfying the d^{th}-order realizability condition with parameter $\beta^* > 1/2$, the condition in Equation (26) holds for all $h \geq d$ for parameter values*

$$
\alpha := \frac{2\beta^* - 1}{2} \tag{28}
$$

$$
t_0(h) = t_{\mathsf{high}}(h) := \frac{2}{\alpha^2} \ln \left(\frac{4(D-d)\cdot 2^{h+1}}{\alpha^2 \epsilon} \right) \tag{29}
$$

with probability greater than or equal to $(1 - \epsilon/2)$.

Proof. Essentially, we need to obtain to bound properties of the gap sequence $\{D_{t,(h)}\}_{t=1}^T$ so defined in Equation (20) – we use the Hoeffding bound for this. This proof is a simple adaptation of the proof in the original AdaHedge paper [EKRG11] to the case of contextual prediction.

We denote the p^{th} epoch of arrival of context $x(h) \in \mathcal{X}^h$ by $T_p(x(h))$. Showing that the condition in Equation (26) holds with probability greater than or equal to $(1 - \epsilon/2)$ is exactly equivalent to showing that the probability of the following bad event

$$
\left\{ \cup_{h=d}^{D} \cup_{\mathbf{x}_{(h)}\in\mathcal{X}^h} \cup_{p=t_0(h)}^{N_T(x(h))} \left\{ D_{T_p(x(h))}(h) < \alpha p \right\} \right\} \tag{30}
$$

is less than or equal to $\frac{\epsilon}{2}$. We proceed by showing exactly this.

From the definition of a d^{th}-order stochastic process, we have $Y_t | \{X_t, (X_s, Y_s)_{s=1}^{t-1}\}$ i.i.d $\sim P^*(\cdot | X_t(d))$. This means that Y_t is independent of $(X_t(D, \ldots, D_d), X_s, Y_s)_{s=1}^{t-1}$ conditioned on $X_t(d)$, and we can write

$$
D_{T_p(x(h))}(h) = \sum_{s'=1}^{p} 2Z_{s'}
$$

where

$$
\{Z_s'\}_{s'\geq 1} \text{ i.i.d } \sim \begin{cases} 1 \text{ w. p. } \beta(x(d)) \\ -1 \text{ otherwise }. \end{cases}
$$

Denote $\alpha := \frac{2\beta^*-1}{2}$. We have $\mathbb{E}[Z_s] = 2\beta(x(d))-1 \geq 2\beta^* - 1 = 2\alpha$ and so we have $\mathbb{E}[D_{T_p(x(h))}(h)] \geq 2\alpha p$. Noting that $Z_s \in \{-1, 1\}$, we can directly use the Hoeffding bound to get

$$\begin{aligned}
\Pr\left[D_{T_p(x(h))}(h) < \alpha p\right] &\leq \Pr\left[D_{T_p(x(h))}(h) < \left(\frac{2\beta(x(d))-1}{2}\right) p\right] \\
&\leq \exp\{-\frac{(2\beta(x(d))-1)^2 p}{8}\} \\
&\leq \exp\{-\frac{\alpha^2 p}{2}\},
\end{aligned}$$

and so, for any $t_0(h) \geq 1$ and $x(h) \in \mathcal{X}^h$, we can use the union bound to get

$$\begin{aligned}
\Pr\left[\cup_{p=t_0(h)}^{N_T(x(h))}\left\{D_{T_p(x(h))}(h) < \alpha p\right\}\right] &\leq \sum_{p=t_0(h)}^{N_T(x(h))} \exp\{-\frac{\alpha^2 p}{2}\} \\
&\leq \sum_{p=t_0(h)}^{\infty} \exp\{-\frac{\alpha^2 p}{2}\} \\
&\leq \int_{u=t_0(h)}^{\infty} \exp\{-\frac{\alpha^2 u}{2}\} du \\
&= \frac{2e^{-\frac{\alpha^2 t_0(h)}{2}}}{\alpha^2}.
\end{aligned}$$

We need to bound the probability that the above bad event happens *for any* context $x(h) \in \mathcal{X}^h$ and model order $h \geq d$. To do this, we apply the union bound twice more, to get

$$\begin{aligned}
\Pr\left[\cup_{h=d}^{D} \cup_{x(h)\in\mathcal{X}^h} \cup_{p=t_0(h)}^{N_T(x(h))}\left\{D_{T_p(x(h))}(h) < \alpha p\right\}\right] &\leq \sum_{h=d}^{D} \sum_{x(h)\in\mathcal{X}^h} \frac{2e^{-\frac{\alpha^2 t_0(h)}{2}}}{\alpha^2} \\
&= \left(\sum_{h=d}^{D} \frac{2 \cdot 2^h \cdot e^{-\frac{\alpha^2 t_0(h)}{2}}}{\alpha^2}\right) \\
&\leq \epsilon/2
\end{aligned}$$

if $t_0(h) \geq t_{\mathsf{high}}(h) = \frac{2}{(\alpha)^2} \ln\left(\frac{4(D-d)\cdot 2^h}{\epsilon(\alpha)^2}\right)$.

Setting $t_0(h) = t_{\mathsf{high}}(h)$ bounds the probability of the bad event as defined in Equation (30), and completes our proof.

□

A.2.4 Completing proof of Proposition 1

Now, the proof of Proposition 1 follows directly from Lemmas 1 and 8. We denote as shorthand the following:

$$\begin{aligned}
\Delta_T^{(h)} &= \Delta_T(\eta_1^T; g_{\mathsf{unif}}) \\
V_T^{(h)} &= V_T(\eta_1^T; g_{\mathsf{unif}})
\end{aligned}$$

Substituting $g(\cdot) = g_{\mathsf{unif}}(\cdot)$ into Lemma 1, we have

$$\begin{aligned}
R_{T,d} \leq R_{T,h} &\leq \left(\sqrt{V_T^{(h)} \ln 2} + \frac{2}{3}\ln 2 + 1\right)\left(1 + \frac{\ln\left(\frac{Z}{g_{\mathsf{unif}}(h)}\right)}{\ln 2}\right) \\
&\leq \left(\sqrt{V_T^{(h)} \ln 2} + \frac{2}{3}\ln 2 + 1\right)(1 + 2^h)
\end{aligned}$$

Thus, it remains to bound the variance term $V_T^{(h)}$. We denote the final learning rate as

$$\eta_T^{(h)} = \frac{\ln 2}{\Delta_{T-1}^{(h)}} \geq \frac{\ln 2}{\Delta_T^{(h)}}$$

and from [DRVEGK14] that

$$\begin{aligned}
\Delta_T^{(h)} &\leq \sqrt{V_T^{(h)} \ln 2} + \frac{2}{3} \ln 2 + 1 \\
&\leq \sqrt{V_T^{(h)}} \Big(\sqrt{\ln 2} + \frac{4}{3} \ln 2 + 2 \Big) \Big(\text{ as } \sqrt{V_T^{(h)}} \geq \sqrt{v_1^{(h)}} = \frac{1}{2} \Big) \\
&\leq 6 \sqrt{V_T^{(h)}} \ln 2.
\end{aligned}$$

Together, these give us

$$\eta_T^{(h)} \geq \frac{1}{6\sqrt{V_T^{(h)}}}$$

and therefore, we have with probability greater than or equal to $(1-\epsilon)$,

$$\begin{aligned}
V_T^{(h)} &\leq \sum_{t=1}^{T} w_{t,1-X_t^*}^{(h)} \\
&\leq 2^h \left(t_{\mathsf{high}}(h) + \frac{1}{\eta_T^{(h)}(2\beta^* - 1)} \right) \\
&\leq 2^h \left(t_{\mathsf{high}}(h) + \frac{6\sqrt{V_T^{(h)}}}{(2\beta^* - 1)} \right) \\
&\leq 2^h \left(\frac{8}{(2\beta^* - 1)^2} \ln \left(\frac{8 \cdot 2^h}{\epsilon(2\beta^* - 1)^2} \right) + \frac{6\sqrt{V_T^{(h)}}}{(2\beta^* - 1)} \right)
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\sqrt{V_T^{(h)}} &\leq \frac{8 \cdot 2^h}{(2\beta^* - 1)^2} \ln \left(\frac{8 \cdot 2^h}{\epsilon(2\beta^* - 1)^2} \right) + \frac{6 \cdot 2^h}{(2\beta^* - 1)} \\
&\leq \frac{14 \cdot 2^h}{(2\beta^* - 1)^2} \ln \left(\frac{8 \cdot 2^h}{\epsilon(2\beta^* - 1)^2} \right)
\end{aligned}$$

This gives us

$$R_{T,d} = \mathcal{O}\Big(\frac{2^{2h}}{(2\beta^* - 1)^2} \left(h + \ln \left(\frac{1}{\epsilon(2\beta^* - 1)} \right) \right) \Big).$$

with probability greater than or equal to $(1-\epsilon)$. This completes the proof.

A.2.5 Ruling out higher-order models

We can make two clear inferences from Lemma 8:

1. ContextTreeAdaHedge(d) gives us the true regret scaling in terms of $\mathcal{O}(2^{2d} \left(d + \ln \left(\frac{1}{\epsilon} \right) \right))$.
2. For $h > d$, ContextTreeAdaHedge(h) gives us suboptimal scaling $Oh(2^{2h} \left(h + \ln \left(\frac{1}{\epsilon} \right) \right))$. The reason for suboptimality is because of sample splitting: for every true context $x(d) \in \mathcal{X}^d$, we are unnecessarily splitting the data into 2^{d-h} extra contexts and treating the best predictors for these contexts as independent.

It is clear, particularly from the second inference, that we would like to control the posterior probability with which we select overly complex models. This quantity is expressed as $q_t(h;\eta_t,g_{\mathsf{prop}})$ for all $h>d$. Now, we consider an explicit upper bound on $q_t(h;\eta_t,g_{\mathsf{prop}})$ and show how it decreases with t.

Using Equation (23), it is convenient to consider the following upper bound on the quantity $q_t(h;\eta_t,g_{\mathsf{prop}})$ for $h>d$:

$$\begin{aligned}
q_t(h;\eta_t,g_{\mathsf{prop}}) &= \frac{Q_t(h;\eta_t,g_{\mathsf{prop}})}{\sum_{h'=0}^{D} Q_t(h';\eta_t,g_{\mathsf{prop}})} \\
&\leq \frac{Q_t(h;\eta_t,g_{\mathsf{prop}})}{Q_t(d;\eta_t,g_{\mathsf{prop}})} \\
&\leq \exp\{\eta_t(\widehat{\pi}_d(t)-\widehat{\pi}_h(t))t - 2^h\ln 2 + 2\cdot 2^d \ln 2\}
\end{aligned}$$

We should expect that as t becomes large the difference in estimated approximation errors is negligible, i.e. we will observe that $\widehat{\pi}_h(t)=\widehat{\pi}_d(t)$ with high probability. We would then get a scaling of $q_t(h;\eta_t,g_{\mathsf{prop}})\leq \exp\{-2^h\ln 2\}$. However, we can say $\widehat{\pi}_h(t)=\widehat{\pi}_d(t)$ with high probability only after $\mathcal{O}(2^h)$ rounds. Before this, and particularly for times between $\mathcal{O}(2^d)$ and $\mathcal{O}(2^h)$, we have to worry about the difference in approximation errors, $\eta_t(\widehat{\pi}_h(t)-\widehat{\pi}_d(t))t$. This is the *overfitting regime* in which the hth order model may look deceptively better. Luckily, we can cap this quantity as well owing to already established statistical guarantees on the sequence $\{X_t\}_{t\geq 1}$. The following lemma expresses this.

Lemma 10. *The process $\{(X_t,Y_t)\}_{t\geq 1}$ satfisfying Equation (26) for all $h\geq d$ and for*

$$t_0(h)=t_{\mathsf{high}}(h)$$

directly implies

$$(\widehat{\pi}_d(t)-\widehat{\pi}_h(t))t \leq \min\{\frac{t}{2}, 2^{h-1}t_{\mathsf{high}}(h)\}. \tag{31}$$

The two quantities on the right hand side of Equation (31) have different operational meaning. The bound in terms of $\frac{t}{2}$ will be used to show that for a small number of rounds, the doubly exponential prior on model order h will weigh this model order down and prevent it from being selected prematurely *even if it could be leveraged for more accurate prediction in later rounds, as would be the case when the data is out-of-model.* On the other hand, the bound in terms of $2^{h-1}t_{\mathsf{high}}(h)$ is useful to conclusively rule out the h^{th}-order model even in later rounds *for the case where data is realized from a d^{th}-order model*, by which time it is clear that the higher-order model does not lead to any improvement in approximability.

Proof. It suffices to prove the following two inequalities separately:

$$\begin{aligned}
(\widehat{\pi}_d(t)-\widehat{\pi}_h(t))t &\leq \frac{t}{2} \\
(\widehat{\pi}_d(t)-\widehat{\pi}_h(t))t &\leq 2^{h-1}t_{\mathsf{high}}(h).
\end{aligned}$$

Recall the notation we defined for the best d^{th}-order tree expert at time t, $\widehat{F}_d(t)$, as well as the number of appearances of context $x(h)$ at time t, denoted by $N_t(x(h))$.

From Definition 9, we have

$$\begin{aligned}
&(\widehat{\pi}_d(t)-\widehat{\pi}_h(t))t \\
&= \sum_{x(d)\in\mathcal{X}^d} N_t(x(d))\left(1-\max_{y\in\mathcal{X}}\{\widehat{P}_t(y|x(d))\}\right) - \sum_{x(h)\in\mathcal{X}^h} N_t(x(h))\left(1-\max_{y\in\mathcal{X}}\{\widehat{P}_t(y|x(h))\}\right) \\
&= \sum_{x(d)\in\mathcal{X}^d} \underbrace{\left(\sum_{x(h):x(d)\subset x(h)} N_t(x(h))\left(\max_{y\in\mathcal{X}}\{\widehat{P}_t(y|x(h))\} - \widehat{P}_t(\widehat{F}_d(t)(x(d))|x(d))\right)\right)}_{T_1}
\end{aligned}$$

Let T_1 be the quantity under the brace (for shorthand). We also define the number of *super-contexts* of length h that contain $x(d)$,

$$S_{t,h-d}(x(d)) := \sum_{x(h):x(d) \subset x(h)} \mathbb{I}[N_t(x(h)) > 0].$$

Now, we have one of two cases:

1. We have $N_t(x(d)) \leq t_{\mathsf{high}}$. In this case, we have $T_1 \leq \frac{t_{\mathsf{high}}}{2}$.
2. $N_t(x(d)) > t_{\mathsf{high}}$. In this case, we have $\widehat{F}_d(t)(x(d)) = f^*(x(d))$ from Equation (26), and we directly get

   $$T_1 = \sum_{x(h):x(d) \subset x(h) \text{ and } \arg\max\{\widehat{P}_t(y|x(h))\} \neq f^*(x(d))} N_t(x(h)) \left(\max_{y \in \mathcal{X}} \{\widehat{P}_t(y|x(h))\} - \widehat{P}_t(f_d^*(x(d))|x(d)) \right)$$

   Clearly, the overfitting effect is created *only* by the set of contexts $x(h)$ for which the best predictor does not match $f^*(x(d))$. From Lemma 9, Equation (26) is satisfied for all $h \geq d$ and for $N_t(x(h)) \geq t_{\mathsf{high}}(h)$. It is easy to see that Equation (26) implies a non-negative separation between the truly correct predictor $f^*(x(d))$ and its alternative, and so we have

   $$\arg\max_{y \in \mathcal{X}} \{\widehat{P}_t(y|x(h))\} = f^*(x(d)) \text{ if } N_t(x(h)) \geq t_{\mathsf{high}}(h).$$

   Substituting this directly, and noting that

   $$\max_{y \in \mathcal{X}} \{\widehat{P}_t(y|x(h))\} - \widehat{P}_t(f_d^*(x(d))|x(d)) \leq 1/2$$

   gives us

   $$\begin{aligned} T_1 &\leq \sum_{x(h):x(d) \subset x(h) \text{ and } N_t(x(h)) \leq t_{\mathsf{high}}(h)} \frac{\min\{N_t(x(h), t_{\mathsf{high}}(h)\}}{2}\} \\ &\leq \sum_{x(h):x(d) \subset x(h) \text{ and } N_t(x(h)) \leq t_{\mathsf{high}}(h)} \frac{t_{\mathsf{high}}(h)}{2} \\ &\leq S_{t,h-d}(x(d)) \frac{t_{\mathsf{high}}(h)}{2}. \end{aligned}$$

Noting that $1 \leq 2^{h-d}$ and $S_{t,h-d}(x(d)) \leq 2^{h-d}$ gives us

$$T_1 \leq 2^{h-d} \frac{t_{\mathsf{high}}(h)}{2},$$

and substituting back this expression yields

$$\begin{aligned} (\widehat{\pi}_d(t) - \widehat{\pi}_h(t))t &\leq \sum_{x(d) \in \mathcal{X}^d} T_1 \\ &\leq 2^{h-1} t_{\mathsf{high}}(h). \end{aligned}$$

This completes our proof.

□

Recall that for all $t > T_0(h)$ where $T_0(h)$ is as defined in Lemma 6 with respect to $t_0(h) = t_{\mathsf{high}}(h)$, we have $\eta_t < \frac{\ln 2}{t_0}$. Under this condition, the explicit cap on the overfitting effect as defined in Lemma 10, together with the adaptive regularization of AdaHedge, ensures that we can sufficiently restrict the contribution of higher-order models.

We use Equation (31) to get

$$\begin{aligned} q_t(h;\eta_t,g_{\mathsf{prop}}) &\leq \exp\{\eta_t(\widehat{\pi}_d(t)-\widehat{\pi}_h(t))t - 2^h \ln 2 + 2\cdot 2^d \ln 2\} \\ &\leq \exp\{\frac{2^{h-1}t_{\mathsf{high}}(h)\ln 2}{t_{\mathsf{high}}(h)} - 2^h \ln 2 + 2^{d+1}\ln 2\} \\ &\leq \exp\{-2^{h-1}\ln 2 + 2^{d+1}\ln 2\} \\ &= 2^{-2^{h-1}+2^{d+1}}. \end{aligned}$$

Therefore, we can apply Lemma 8 to get

$$\begin{aligned} \sum_{t=T_0}^{T} q_t(h;\eta_t,g_{\mathsf{prop}}) w^{(h)}_{t,1-Y_t^*}(\eta_t) &\leq 2^{-2^{h-1}+2^{d+1}} \sum_{t=T_0}^{T} w^{(h)}_{t,1-Y_t^*} \\ &\leq 2^{h-2^{h-1}+2^{d+1}} \left(t_{\mathsf{high}}(h) + \frac{1}{\eta_T \alpha}\right). \end{aligned}$$

It is now easy to check that

$$\begin{aligned} & 2h \leq 2^{h-1} - 2^{d+1} \text{ for all } h \geq d+4 \text{ and } d \geq 0 \\ \implies & h - 2^{h-1} + 2^{d+1} \leq -h \\ \implies & 2^{h-2^{h-1}+2^{d+1}} \leq 2^{-h}. \end{aligned}$$

Therefore, for $h \geq d+4$, we get

$$\sum_{t=T_0}^{T} q_t(h;\eta_t,g_{\mathsf{prop}}) w^{(h)}_{t,1-Y_t^*}(\eta_t) \leq 2^{-h}\left(t_{\mathsf{high}}(h) + \frac{1}{\eta_T\alpha}\right).$$

For $h < d+4$, we do not try to non-trivially bound $q_t(h;\eta_t,g_{\mathsf{prop}})$. We directly use Lemma 8 to get

$$\sum_{t=T_0}^{T} q_t(h) w^{(h)}_{t,1-Y_t^*}(\eta_t) \leq 2^{h}\left(t_{\mathsf{high}}(h) + \frac{1}{\eta_T\alpha}\right).$$

We have thus guaranteed that the contribution from the higher-order models (particularly for $h \geq d+4$) not only has no exponential dependence on h, but is in fact exponentially decaying in h! Ultimately, we will see that we get a very weak linear dependence on D, the maximum model order, in our regret bound.

A.2.6 Ruling out *bad* lower-order models

Using Equation (23), it is convenient to consider the following upper bound on the quantity $q_t(h)$ for $h < d$:

$$\begin{aligned} q_t(h;\eta_t,g_{\mathsf{prop}}) &\leq \frac{Q_t(h;\eta_t,g_{\mathsf{prop}})}{Q_t(d;\eta_t,g_{\mathsf{prop}})} && \text{(32a)} \\ &\leq \exp\{-\eta_t(\widehat{\pi}_h(t)-\widehat{\pi}_d(t))t + 2\cdot 2^d\ln 2 - 2^h \ln 2\} && \text{(32b)} \end{aligned}$$

Ruling out lower-order models actually stems from the fact that we can make concrete statements about the sequence's unpredictability (poor approximability) under these models.

The kind of concrete statement that we would like is detailed in the lemma below.

Lemma 11. *Let $h < d$. Consider a sequence $\{x_t\}_{t \geq 1}$ such that we have*

$$(\widehat{\pi}_h(t) - \widehat{\pi}_d(t))t \geq \alpha_{h,d} t \text{ for all } t \geq t_0(h) > 0 \tag{33}$$

for some $\alpha_{h,d} > 0$.
Then, we have

$$\sum_{t=1}^{T} q_t(h; \eta_t, g_{\mathsf{prop}}) w^{(h)}_{t,1-Y_t^*}(\eta_t) \leq t'_{\mathsf{low}}(h) + \frac{1}{\eta_T \alpha_{h,d}} \tag{34}$$

where

$$t'_{\mathsf{low}}(h) = \max\{t_0(h), \frac{2 \cdot 2^d \ln 2}{\eta_T \alpha_{h,d}}\}. \tag{35}$$

Proof. The condition in Equation (33) is essentially the same as the condition on gaps between losses in the original AdaHedge paper [EKRG11] used to prove constant regret bounds. We use a similar argument here.

First, we subsitute the condition in Equation (33) into Equation (32b) to get the upper bound

$$\begin{aligned} q_t(h; \eta_t, g_{\mathsf{prop}}) &\leq \exp\{-\eta_t \alpha_{h,d} t + 2 \cdot 2^d \ln 2 - 2^h \ln 2\} \\ &\leq \exp\{-\eta_t \alpha_{h,d} t + 2 \cdot 2^d\} \\ &= \exp\{2 \cdot 2^d \ln 2 - \eta_t \alpha_{h,d} t\} \\ &\leq \exp\{2 \cdot 2^d \ln 2 - \eta_T \alpha_{h,d} t\}. \end{aligned}$$

where the last inequality applies because η_1^T is a decreasing sequence. Putting this together with the trivial bound $q_t(h; \eta_t, g_{\mathsf{prop}}) \leq 1$ gives us

$$q_t(h; \eta_t, g_{\mathsf{prop}}) \leq \begin{cases} 1 \text{ for } t \leq t'_{\mathsf{low}}(h) \\ \exp\{2 \cdot 2^d \ln 2 - \eta_T \alpha_{h,d} t\} \text{ for } t > t'_{\mathsf{low}}(h). \end{cases}$$

where we have

$$t'_{\mathsf{low}} = \max\{t_0(h), \frac{2 \cdot 2^d \ln 2}{\eta_T \alpha_{h,d}}\}.$$

From this, using the trivial bound $w_{t,1-Y_t^*}(\eta_t) \leq 1$ we get

$$\begin{aligned} \sum_{t=1}^{T} q_t(h; \eta_t, g_{\mathsf{prop}}) w_{t,1-Y_t^*}(\eta_t) &\leq t'_{\mathsf{low}}(h) + \sum_{t=t'_{\mathsf{low}}(h)+1}^{\infty} \exp\{2 \cdot 2^d \ln 2 - \eta_T \alpha_{h,d} t\} \\ &\leq t'_{\mathsf{low}}(h) + \exp\{2 \cdot 2^d \ln 2 - \eta_T \alpha_{h,d} t'_{\mathsf{low}}(h)\} \left(\sum_{t=1}^{\infty} e^{-\eta_T \alpha_{h,d} t} \right) \\ &= t'_{\mathsf{low}}(h) + \sum_{t=1}^{\infty} e^{-\eta_T \alpha_{h,d} t} \\ &\leq t'_{\mathsf{low}}(h) + \int_{u=0}^{\infty} e^{-\eta_T \alpha_{h,d} u} du \\ &= t'_{\mathsf{low}}(h) + \frac{1}{\eta_T \alpha_{h,d}} \int_{v=0}^{\infty} e^{-v} dv \\ &= t'_{\mathsf{low}}(h) + \frac{1}{\eta_T \alpha_{h,d}}, \end{aligned}$$

This completes the proof. □

From Lemma 11, we can clearly bound the contribution of lower-order models to cumulative variance by a constant term. This is because the difference in estimated unpredictability between the right model and the bad lower-order model remains as the number of rounds increase – leading to an exponentially decaying likelihood of selecting the lower-order model. (We do not even need to use any information about whether the online learning algorithm would ensure low regret when selecting a lower-order model, although this is sometimes the case in practice[13].)

It is of interest to characterize when the condition in Equation (33) holds. We show that this holds in a variety of settings under both the stronger *realizability* and the weaker *approximability* condition. The informal statement is stated below; for a formal statement and proof see Appendix B.

Lemma 12 (Informal.). *The condition in Equation* (33) *holds for* $\alpha_{h,d} = \frac{\pi^*_h - \pi^*_d}{2}$*, some constant* $c > 0$*, and*

$$t_0(h) = t_{\mathsf{low}}(h) := \frac{32}{\alpha^2_{h,d}} \left(d \cdot 2^h \ln 2 + \ln \left(\frac{64d}{\epsilon \alpha^2_{h,d}} \right) \right). \tag{36}$$

with probability greater than equal to $(1 - \epsilon)$ *for the following cases:*

1. $(X_s, Y_s)_{s \geq 1}$ *are stochastic and iid (contextual prediction).*
2. $(Y_s)_{s \geq 1}$ *is a* d^{th}*-memory mixing Markov process. (Here,* $X_s = Y^{s-1}_{s-D}$*.)*
3. $(Y_s)_{s \geq 1}$ *is a mixing hidden Markov models. (Here,* $X_s = Y^{s-1}_{s-D}$*.)*

A.2.7 Putting the pieces together: Proof of Theorem 1

In Section A.2.3, we determined the overall contribution to the cumulative variance coming from the vicinity of the true model orders, $h \in \{d, d+1, d+2, d+3\}$. Then, in Section A.2.5 + A.2.6, we appropriately limited the contribution of lower-order and higher-order models to the cumulative variance. Now, we put together the pieces and characterize cumulative regret to complete the proof of Theorem 1.

First, we apply Lemma 6 setting $t_0 = t_{\mathsf{high}}(D)$. Recall that $t_{\mathsf{high}}(D)$ represents the number of appearances of a full context before which we cannot necessarily make statistical guarantees about the predictor. This gives us[14]

$$\Delta_T \leq t_{\mathsf{high}}(D) + \sqrt{V^T_{T_0(D)} \ln 2} + \frac{2}{3} \ln 2 + 2. \tag{37}$$

We now proceed to bound the quantity $V^T_{T_0(D)}$. Recall that

$$\begin{aligned} V^T_{T_0(D)} &\leq \sum_{h=0}^{D} q_t(h) \sum_{t=T_0(D)}^{T} w^{(h)}_{t,1-X^*_t} \\ &\leq \underbrace{\sum_{h=0}^{d-1} q_t(h) \sum_{t=T_0(D)}^{T} w^{(h)}_{t,1-X^*_t}}_{T_1} + \underbrace{\sum_{h=d}^{d+3} \sum_{t=T_0(D)}^{T} w^{(d)}_{t,1-X^*_t}}_{T_2} + \underbrace{\sum_{h=d+4}^{D} q_t(h) \sum_{t=T_0(D)}^{T} w^{(h)}_{t,1-X^*_t}}_{T_3} \end{aligned}$$

[13] In fact, models that are close in approximability to the true model will suffer less regret. Ideally, our analysis should consider this nuance, but doing so is likely to be technically challenging because of the data-dependent learning rate.

[14] Equation (37) exposes new conceptual beauty in the umbrella of approaches to varying the learning rate inversely proportional to accumulated regret so far. The only reason a high learning rate does not affect us is because it means that very little regret has been accumulated up to that point. Effectively, $t_0 = t_{\mathsf{high}}(D)$ represents the extent of cumulative mixability the algorithm is willing to tolerate in this regime before carrying out probabilistic stochastic model selection, and is the natural statistical quantity to reflect this.

We start with summarizing the lower-order model contribution T_1. From Lemmas 11 and 13, we have

$$\begin{aligned} T_1 &\leq \sum_{h=0}^{d-1} t'_{\mathsf{low}}(h) + \frac{1}{\eta_T}\left(\sum_{h=0}^{d-1}\frac{1}{\alpha_{h,d}}\right) \\ &\leq d t'_{\mathsf{low}}(d-1) + \frac{1}{\eta_T}\left(\sum_{h=0}^{d-1}\frac{1}{\alpha_{h,d}}\right). \end{aligned}$$

Notice that T_1 is a constant independent of the horizon T as long as η_T does not decay with T.

Next, we move on to the vicinity of the true model order contribution, represented by model orders $\{d, d+1, d+2, d+3\}$. From Lemmas 8 and 9, we get

$$\begin{aligned} T_2 &\leq \sum_{h=d}^{d+3} 2^h \left(t_{\mathsf{high}}(h) + \frac{1}{\eta_T(2\beta^* - 1)}\right) \\ &\leq 15 \cdot 2^d \left(t_{\mathsf{high}}(d+3) + \frac{1}{\eta_T(2\beta^* - 1)}\right). \end{aligned}$$

Notice that T_2 is roughly what we should expect (upto constant factors) if we knew the model order exactly.

Finally, we summarize the higher-order-model contribution T_3. From Lemma 10 and the analysis in Section A.2.5, we have

$$\begin{aligned} T_3 &\leq \sum_{h=d+4}^{D} 2^{-h} \left(t_{\mathsf{high}}(h) + \frac{1}{\eta_T(2\beta^* - 1)}\right) \\ &= \sum_{h=d+4}^{D} 2^{-h} t_{\mathsf{high}}(h) + \frac{2}{\eta_T(2\beta^* - 1)}. \end{aligned}$$

Recall from Equation (28) that

$$\begin{aligned} t_{\mathsf{high}}(h) &= \frac{2}{(2\beta^* - 1)^2} \ln\left(\frac{(D-d)\cdot 2^h}{(2\beta^* - 1)^2 \epsilon}\right) \\ &= \frac{2h}{(2\beta^* - 1)^2} \ln 2 + \frac{2}{(2\beta^* - 1)^2} \ln\left(\frac{(D-d)}{(2\beta^* - 1)^2 \epsilon}\right) \end{aligned}$$

and since $\sum_{h=0}^{\infty} 2^{-h} \leq \sum_{h=0}^{\infty} h \cdot 2^{-h} = 4$, we get

$$T_3 \leq \frac{8}{(2\beta^* - 1)^2} \ln 2 + \frac{8}{(2\beta^* - 1)^2} \ln\left(\frac{(D-d)}{(2\beta^* - 1)^2 \epsilon}\right) + \frac{2}{\eta_T(2\beta^* - 1)} = 8 t_{\mathsf{high}}(1) + \frac{2}{\eta_T(2\beta^* - 1)}.$$

Notice that T_3 is a constant that scales only logarithmically in the maximum model order D!

Now combining the three equations for T_1,T_2 and T_3, we get

$$V^T_{T_0(D)} \leq d t'_{\mathsf{low}}(d-1) + 15 \cdot 2^d t_{\mathsf{high}}(d+3) + 8 t_{\mathsf{high}}(1) + \frac{(d+1)\cdot 2^d}{\eta_T \overline{\gamma}},$$

where

$$\frac{1}{\overline{\gamma}} := \frac{1}{d+1}\Big(\sum_{h=0}^{d-1}\frac{1}{\alpha_{h,d}} + \frac{15}{(2\beta^* - 1)}\Big)$$

Next, recall from Equation (35) that

$$t'_{\mathsf{low}}(d-1) = \max\{t_{\mathsf{low}}(d-1), \frac{2 \cdot 2^d}{\eta_T \alpha_{d-1,d}}\} \leq t_{\mathsf{low}}(d-1) + \frac{2\cdot 2^d}{\eta_T \alpha_{d-1,d}}$$

using Fact 1. Substituting this expression gives us

$$V^T_{T_0(D)} \leq d \cdot t_{\mathsf{low}}(d-1) + 15 \cdot 2^d \cdot t_{\mathsf{high}}(d+3) + 8t_{\mathsf{high}}(1) + \frac{(d+2) \cdot 2^d}{\eta_T \overline{\gamma}}.$$

Next, we use the connection between learning rate and mixability gap from Equation (2) to get

$$\begin{aligned} \eta_T &= \frac{\ln 2}{\Delta_{T-1}} \geq \frac{\ln 2}{\Delta_T} \\ \implies \frac{1}{\eta_T} &\leq \frac{\Delta_T}{\ln 2} \\ &\leq \frac{t_{\mathsf{high}}(D)}{\ln 2} + \frac{1}{\ln 2}\left(\sqrt{V^T_{T_0(D)} \ln 2} + \frac{2}{3}\ln 2 + 1\right) \end{aligned}$$

where in the last step we applied Equation (37).

Ultimately, we get the following inequality for $V^T_{T_0(D)}$:

$$V^T_{T_0(D)} \leq d \cdot t_{\mathsf{low}}(d-1) + 15 \cdot 2^d \cdot t_{\mathsf{high}}(d+3) + 8t_{\mathsf{high}}(1) + \frac{(d+2) \cdot 2^d}{\overline{\gamma}}\left(\frac{t_{\mathsf{high}}(D)}{\ln 2} + \frac{1}{\ln 2}\left(\sqrt{V^T_{T_0(D)} \ln 2} + \frac{2}{3}\ln 2 + 1\right)\right).$$

Now, we have two cases:

1. $V^T_{T_0(D)} < \frac{1}{4}$.
2. $V^T_{T_0(D)} \geq \frac{1}{4}$, in which case, we get

$$\begin{aligned} V^T_{T_0(D)} &\leq \sqrt{V^T_{T_0(D)}}\Big(2d \cdot t_{\mathsf{low}}(d-1) + 30 \cdot 2^d \cdot t_{\mathsf{high}}(d+3) + 16 \cdot t_{\mathsf{high}}(1) \\ &\quad + \frac{2 \cdot (d+2) \cdot 2^d \cdot t_{\mathsf{high}}(D)}{\overline{\gamma}\ln 2} + \frac{1}{\sqrt{\ln 2}} + \frac{2}{3} + \frac{1}{\ln 2}\Big) \\ \implies \sqrt{V^T_{T_0(D)}} &\leq 2d \cdot t_{\mathsf{low}}(d-1) + 30 \cdot 2^d \cdot t_{\mathsf{high}}(d+3) + 16 \cdot t_{\mathsf{high}}(1) + \frac{2 \cdot (d+2) \cdot 2^d \cdot t_{\mathsf{high}}(D)}{\overline{\gamma}\ln 2} \\ &\quad + \frac{1}{\sqrt{\ln 2}} + \frac{2}{3} + \frac{1}{\ln 2}. \end{aligned}$$

So, we have bounded the cumulative variance term $V^T_{T_0(D)}$. We now substitute back into Equation (37) to get

$$\begin{aligned} \Delta_T &\leq t_{\mathsf{high}}(D) + \Big(2d \cdot t_{\mathsf{low}}(d-1) + 30 \cdot 2^d \cdot t_{\mathsf{high}}(d+3) + 16 \cdot t_{\mathsf{high}}(1) + \frac{2 \cdot (d+2) \cdot 2^d \cdot t_{\mathsf{high}}(D)}{\overline{\gamma}\ln 2} \\ &\quad + \frac{1}{\sqrt{\ln 2}} + \frac{2}{3} + \frac{1}{\ln 2}\Big)\sqrt{\ln 2} + \frac{2}{3}\ln 2 + 2. \end{aligned}$$

Observe, from this inequality, that the cumulative mixability gap Δ_T is dominated by three intuitive quantities (other than the constant additive term):

1. $t_{\mathsf{low}}(d-1)$, which represents the number of rounds after which all lower-order models can be conclusively ruled out. The dependence on $t_{\mathsf{low}}(d-1)$ is saying that this much mixability could have accumulated (due to poor approximability) before then.
2. $t_{\mathsf{high}}(D)$, which represents the amount of mixability the algorithm has to accumulate before performing effective higher-order model selection to rule out the overfitting models[15].

[15] It is also possible that the algorithm would not have accumulated even this mixability, and the model selection phase is never reached – however, we never observed this case empirically.

3. $2^d \cdot t_{\mathsf{high}}(d)$, which represents the amount of mixability accumulated by the algorithm *at the right model order*. This is the term in analysis that corresponds to standard best-of-both-worlds analysis over a fixed model order.

Now, we know from Equation (28) that $t_{\mathsf{high}}(h) = \frac{2}{(2\beta^*-1)^2} \ln\left(\frac{(D-d)\cdot 2^h}{(2\beta^*-1)^2\epsilon}\right)$ and from Equation (36) that $t_{\mathsf{low}}(d-1) = \frac{32d}{\alpha^2_{d-1,d}}\left(d \cdot 2^{d-1}\ln 2 + \ln\left(\frac{64d}{\epsilon\alpha^2_{d-1,d}}\right)\right)$. Substituting these in, we get

$$\Delta_T = \mathcal{O}\left(2^d\left(\frac{d^2}{\alpha^2_{d-1,d}}\ln\left(\frac{d}{\alpha^2_{d-1,d}\epsilon}\right) + \frac{D(d+2)}{\overline{\gamma}(2\beta^*-1)^2}\ln\left(\frac{D}{(2\beta^*-1)^2\epsilon}\right)\right)\right) \tag{38}$$

and substituting this into Lemma 1 gives

$$R_{T,d} = \mathcal{O}\left(2^{2d}\left(\frac{d^2}{\alpha^2_{d-1,d}}\ln\left(\frac{d}{\alpha^2_{d-1,d}\epsilon}\right) + \frac{D(d+2)}{\overline{\gamma}(2\beta^*-1)^2}\ln\left(\frac{D}{(2\beta^*-1)^2\epsilon}\right)\right)\right), \tag{39}$$

completing the proof. To highlight the dependence on true model order d and maximum model order D (as is expressed in the informal statement of Theorem 1), we can hide the constants in terms of parameters and write

$$\begin{aligned} R_{T,d} &= \Delta_T\left(1 + 2^d\right) && (40)\\ &= \mathcal{O}\left(2^{2d}\left(D \cdot d \cdot \ln\left(\frac{D}{\epsilon}\right)\right)\right). && (41)\end{aligned}$$

B Stochastic model selection guarantees

Whether the sequence was actually *realized* from a finite-order model, or whether it was merely *well-approximated*, we required the estimates of approximation error to concentrate sufficiently quickly – in particular, we required that the difference in approximability between a higher-order model and lower-order model not look too small – in order to rule out lower-order models when appropriate. This was encapsulated in Lemma 11. It is therefore of interest to understand when the condition in Equation (33) holds, and in particular, characterize $t'_{\mathsf{low}}(h)$. Recall the definition of asymptotic unpredictability

$$\pi^*_h := \sum_{x(h)\in\mathcal{X}^h} Q^*(x(h))\left[1 - \max_{y\in\mathcal{X}}\{P^*(y|x(h))\}\right] \tag{42}$$

Also recall that for $h > d$, we have $\pi^*_h = \pi^*_d$; and for $h < d$, we have $\pi^*_h > \pi^*_d$. It is also well-known [FMG92] that

$$\widehat{\pi}_h(t) \xrightarrow{\mathsf{prob.}} \pi^*_h \text{ for all } h \in \{0, 1, \ldots, D\}.$$

So the intuition is that for a large enough value of t, we should also start to see a *strict* decaying in the estimated unpredictability as h increases to d – and we should be able to rule out the poorly performing hth order models when $h < d$. That is,

$$\widehat{\pi}_h(t) > \widehat{\pi}_d(t) \text{ for all } h < d.$$

In this section, we show that this condition holds for a broad class of stationary, stochastic, predictable processes. One clear case is that of iid context-response pairs, for which proving concentration bounds is standard. The second case considers sequences generated from a finite-memory process or a hidden Markov model – to prove concentration bounds here, we invoke results from the information theory community on transportation-cost inequalities, used to establish concentration of measure for weakly dependent random variables.

B.1 Sufficient condition for concentration of estimate of approximability

We start by expressing our estimate for approximability for the h^{th}-order model, $\widehat{\pi}_h(t)$, as a minimum of $|\mathcal{F}_h|$ Lipschitz functions as below:

$$
\begin{aligned}
t\widehat{\pi}_h(t) &= \min_{\mathbf{f}\in\mathcal{F}_h} \left\{ f_{(h)}(\{(X_s, Y_s)\}_{s=1}^t; \mathbf{f}) \right\} \text{ where} \\
f_{(h)}(\{(X_s, Y_s)\}_{s=1}^t; \mathbf{f}) &:= \sum_{s=1}^t \mathbb{I}[Y_s \neq \mathbf{f}(X_s(h))] \\
&= \sum_{s=1}^t Z_s
\end{aligned}
$$

where $Z_s = \mathbb{I}[Y_s \neq \mathbf{f}(X_s(h))]$. Note that $\{Z_s\}_{s=1}^t$ are independent variables taking values in $\{0,1\}$. We now state the following technical lemma:

Lemma 13. *Let the following condition hold for every $f \in \mathcal{F}^h$, $t \geq h+1$ and $\delta > 0$:*

$$
\Pr\left[|f_{(h)}((X_s, Y_s)_{s=1}^t; \mathbf{f}) - \mathbb{E}[f_{(h)}((X_s, Y_s)_{s=1}^t; \mathbf{f})|] > t\delta\right] \leq 2\exp\{-ct\delta^2\} \tag{43}
$$

for some constant $c > 0$ (that can depend linearly on d as well as h).
Then, the condition in Equation (33) holds for $\alpha_{h,d} = \frac{\pi_h^* - \pi_d^*}{2}$ and

$$
t_0(h) = t_{\mathsf{low}}(h) := \frac{32}{c\cdot\alpha_{h,d}^2}\left(d\cdot 2^h \ln 2 + \ln\left(\frac{64d}{c\cdot\epsilon\alpha_{h,d}^2}\right)\right).
$$

with probability greater than equal to $(1-\epsilon)$.

Proof. Observe that $\widehat{\pi}_h(t)$ itself is not an unbiased estimate of π_h^*. But we know that

$$
\mathbb{E}[t\widehat{\pi}_h(t)] = \mathbb{E}\left[\min_{\mathbf{f}\in\mathcal{F}_h} f_{(h)}(\{(X_s, Y_s)\}_{s=1}^t; \mathbf{f})\right] \leq \mathbb{E}[f_{(h)}(\{(X_s, Y_s)\}_{s=1}^t; \mathbf{f}_h^*)] = t\pi_h^*
$$

for all $\mathbf{f} \in \mathcal{F}_h$. The upper tail bound therefore follows easily – from Equation (43), we have

$$
\begin{aligned}
\Pr\left[t\widehat{\pi}_h(t) - t\pi_h^* > \delta t\right] &\leq \Pr\left[f_{(h)}((X_s, Y_s)_{s=1}^t; \mathbf{f}_h^*) - \mathbb{E}[f_{(h)}((X_s, Y_s)_{s=1}^t; \mathbf{f}_h^*)]\right] \\
&\leq \exp\{-ct\delta^2\}.
\end{aligned}
$$

To get the lower tail bound, we need to use the union bound.

$$
\begin{aligned}
\Pr\left[t\pi_h^* - t\widehat{\pi}_h(t) > \delta t\right] &= \Pr\left[t\widehat{\pi}_h(t) < t\pi_h^* - \delta t\right] \\
&\leq \sum_{\mathbf{f}\in\mathcal{F}^h} \Pr\left[f_{(h)}((X_s, Y_s)_{s=1}^t; \mathbf{f}) < t\pi_h^* - \delta t\right] \\
&= \sum_{\mathbf{f}\in\mathcal{F}^h} \Pr\Big[f_{(h)}(\{(X_s, Y_s)\}_{s=1}^t; \mathbf{f}) - \mathbb{E}[f_{(h)}(\{(X_s, Y_s)\}_{s=1}^t; \mathbf{f})] < \\
&\qquad t\pi_h^* - \mathbb{E}[f_{(h)}(\{(X_s, Y_s)\}_{s=1}^t; \mathbf{f})] - \delta t\Big] \\
&\leq \sum_{\mathbf{f}\in\mathcal{F}^h} \Pr\left[f_{(h)}(\{(X_s, Y_s)\}_{s=1}^t; \mathbf{f}) - \mathbb{E}[f_{(h)}(\{(X_s, Y_s)\}_{s=1}^t; \mathbf{f})] < -\delta t\right] \\
&\leq 2^{2^h}\exp\{-ct\delta^2\}.
\end{aligned}
$$

Next, we plug in $\delta = \frac{\alpha_{h,d}}{2} = \frac{\pi_h^* - \pi_d^*}{4}$ and re-apply the union bound to get

$$\begin{aligned}
&\Pr\left[\cup_{h=0}^{d-1}\{(\widehat{\pi}_h(t) - \widehat{\pi}_d(t)) \le \alpha_{h,d} \text{ for some } t \ge t_0(h)\}\right] \\
&\le \Pr\left[\cup_{h=0}^{d-1}\{\pi_h^* - \widehat{\pi}_h(t) \le \frac{\alpha_{h,d}}{2} \text{ for some } t \ge t_0(h)\} \cup \{\widehat{\pi}_d(t) - \pi_d^* \le \frac{\alpha_{h,d}}{2} \text{ for some } t \ge t_0(h)\}\right] \\
&\le \sum_{h=0}^{d-1} \Pr\left[\pi_h^* - \widehat{\pi}_h(t) \le \frac{\alpha_{h,d}}{2} \text{ for some } t \ge t_0(h)\right] + \Pr\left[\widehat{\pi}_d(t) - \pi_d^* \le \frac{\alpha_{h,d}}{2} \text{ for some } t \ge t_0(h)\right] \\
&\le \sum_{h=0}^{d-1} \frac{32 \cdot 2^{2^h}}{c \cdot \alpha_{h,d}^2} e^{-\frac{c \cdot \alpha_{h,d}^2 t_0(h)}{32}} + \frac{32}{c \cdot \alpha_{h,d}^2} e^{-\frac{c \cdot \alpha_{h,d}^2 t_0(h)}{32}} \\
&\le \epsilon/2 \text{ when}
\end{aligned}$$

$$t_0(h) \ge t_{\mathsf{low}}(h) := \frac{32}{c \cdot \alpha_{h,d}^2}\left(d \cdot 2^h \ln 2 + \ln\left(\frac{64d}{c \cdot \epsilon \alpha_{h,d}^2}\right)\right).$$

This completes our proof. □

Clearly, Lemma 13 holds for the case where $(X_s, Y_s)_{s=1}^t$ are iid, as the Z_s's are iid, by the Hoeffding bound. We proceed to show that it also holds for finite-memory Markov models and hidden Markov models.

B.2 Concentration for finite-memory Markov models

The concentration of sums of random variables to its mean is a classical topic in statistics and probability theory. The special case when the random variables are iid is well-understood. Intuitively, a Markov process that is well-approximated by an iid process should follow similar concentration laws – the transportation cost argument uses this to prove concentration bounds on sums of random variables following a Markov process.

Formally, this notion of approximability by a product distribution is captured by the *contractivity* of a Markov process which we define below.

Definition 10. *1. For a d^{th}-memory Markov process on $Y_1, \dots, Y_t$ on state space $\mathcal{X}$, we define the **aggregated state** at time s by*

$$W_s = (W_{s,1}, \dots, W_{s,d}) = (Y_{(s-1)d+1}, Y_{(s-1)d+2}, \dots, Y_{sd}). \tag{44}$$

Then, clearly, for any $s \ge 1$, we have $W_s \perp W_{s-2} | W_{s-1}$ and so that the states $\{W_s \in \mathcal{X}^d\}_{s \ge 1}$ satisfy the 1-memory Markov property. Explicitly, we have for any $w, w' \in \mathcal{X}^d$,

$$\begin{aligned}
\mathbb{P}(w) &= P_{(d)}(w) \\
\mathbb{P}_{-1}(w'|w) &= \prod_{i=1}^{d} P(w'_i | (w_i, \dots, w_d, \dots, w'_{i-1})).
\end{aligned}$$

*2. A d^{th}-memory Markov process on $Y_1, \dots, Y_t$ is γ-**contractive** if for every $w, w' \in \mathcal{X}^d$, we have*

$$\|\mathbb{P}_{-1}(.|w) - \mathbb{P}_{-1}(.|w')\|_{TV} \le \gamma < 1. \tag{45}$$

The *transportation cost method* can then be easily leveraged to show that Equation (43) holds, as we show in the following minor lemma.

Lemma 14. *Equation* (43) *holds for a d^{th}-memory γ-contractive Markov process with constant $c = \frac{(1-\gamma)}{d}$.*

Proof. We invoke Marton's concentration theorem for γ-contractive Markov processes as described in Theorem 3 (details about the transportation cost method are provided in Section B.4).

We recall the definition of functions

$$f_{(h)}((X_s, Y_s)_{s=1}^t; \mathbf{f}) = f_{(h)}((Y_s)_{s=1}^t; \mathbf{f}) := \sum_{s=1}^{t} Z_s$$

where $Z_s = \mathbb{I}[Y_s \neq \mathbf{f}(Y_s(h))]$. To obtain concentration bounds from Theorem 3, it remains to rewrite $f((Y_s)_{s=1}^t; \mathbf{f})$ as a sum of indicator functions on $\{W_s\}_{s \geq 1}$ and show the Lipschitz property.

Let $t = \lfloor t/d \rfloor + k$ for some $k \in \{0, \ldots, d-1\}$. Then, we can write (with a slight abuse of notation) for any $h \in \{0, 1, \ldots, d\}$,

$$\begin{aligned} f_{(h)}(Y^t; \mathbf{f}) := df_{(h)}(\{W_s\}_{s=1}^{\lceil t/d \rceil}; \mathbf{f}) = \Big(& \sum_{s=1}^{\lfloor t/d \rfloor} \sum_{i=h+1}^{d} \mathbb{I}[W_{s,i} \neq f(W_{s,i-h}, \ldots, W_{s,i-1})] \\ & + \sum_{i=1}^{h} \mathbb{I}[W_{s+1,i} \neq f(W_{s,d-(h-i)}, \ldots, W_{s,d}, W_{s+1,1}, \ldots, W_{s+1,i-1})] \Big) \end{aligned}$$

Now, it's easy to verify that that a change in W_s will only affect two terms in the sum over $\lceil t/d \rceil$ terms, and some simple algebra tell us that the Lipschitz constant of function $f_{(h)}$ is at most 2. We now apply Theorem 3 directly to get

$$\Pr\left[|f_{(h)}(W^{\lceil \frac{t}{d} \rceil}; \mathbf{f}_d) - \mathbb{E}[f_{(h}(W^{\lceil \frac{t}{d} \rceil}; \mathbf{f}_d)]| > \delta \lceil \frac{t}{d} \rceil \right] \leq 2 \exp\{-\frac{2\delta^2 (1-\gamma)^2 t}{4d}\} \tag{46}$$

and so, we finally get

$$\Pr\left[f_{(h)}(Y^t; \mathbf{f}_d) - \mathbb{E}[f_{(h)}(Y^t; \mathbf{f}_d)] > (t-h-1)\delta \right] \leq \exp\{-\frac{\delta^2 (1-\gamma)^2 (t-h-1)}{2d}\} \text{ and}$$
$$\Pr\left[\mathbb{E}[f_{(h)}(Y^t; \mathbf{f}_d)] - f_{(h)}(Y^t; f) > (t-h-1)\delta \right] \leq \exp\{-\frac{\delta^2 (1-\gamma)^2 (t-h-1)}{2d}\},$$

completing the proof of Lemma 15.

□

B.3 Concentration for hidden Markov models

Now, we consider the hidden states $(W_s)_{s=1}^t$ which form a Markov chain.

The definition of a hidden Markov model directly implies that Y_s is independent of $(W_{s'}, Y_{s'})_{s'=1}^{s-2}$ and Y_{s-1} given W_{s-1}. Effectively, the hidden Markov model is a special case of a 1-memory Markov chain on $(W_s, Y_s)_{s \geq 1}$.

Like in the case of d-memory Markov processes, we consider a condition of contractivity. Let $\mathbb{P}$ denote the stationary distribution on the tuple (W_s, Y_s), and $\mathbb{P}_{-1}(\cdot, \cdot | x, y)$ denote the transition probability matrix. (Note that this only depends on the value of x; we have $\mathbb{P}_{-1}(\cdot, \cdot | x, y) = \mathbb{P}_{-1}(\cdot, \cdot | x, y')$ for any $y' \neq y$.)

Definition 11. *Consider a hidden Markov model with transition probability matrix $\mathbb{P}_{-1}(\cdot, \cdot | x, y)$ on the joint state (W_s, Y_s). The model is γ-**contractive** if for every $(x, y), (x', y') \in \mathcal{X} \times \mathcal{X}$, we have*

$$||\mathbb{P}_{-1}(.|x) - \mathbb{P}_{-1}(.|x')||_{TV} \leq \gamma < 1.$$

The *transportation cost method* can then be easily leveraged to show that Equation (43) holds, as we show in the following minor lemma.

Lemma 15. *Equation* (43) *holds for a γ-contractive Markov process with constant $c = \frac{(1-\gamma)}{h^2}$ (on the function used to expressed h^{th}-order unpredictability).*

Proof. As before, we invoke Theorem 3 on the Markov process $(W_s, Y_s)_{s=1}^t$. For the estimate of h^{th}-order approximability, we consider the function (overloading notation)

$$f_{(h)}((W_s, Y_s)_{s=1}^t; \mathbf{f}) := f_{(h)}(Y^t; \mathbf{f}) = \sum_{s=h}^{t} \mathbb{I}\left[Y_s \neq \mathbf{f}(Y_{s-h}, \ldots, Y_{s-1})\right]. \tag{47}$$

Clearly, a change in Y_s can affect at most h terms in the sum on the right hand side of Equation (47), and therefore, the function $f_{(h)}(\cdot; \mathbf{f})$ is h-Lipschitz. Thus, we apply the theorem to get the following concentration rate:

$$\Pr\left[f_{(h)}(Y^t; \mathbf{f}) - \mathbb{E}[f_{(h)}(Y^t; \mathbf{f})] > (t-h-1)\delta\right] \leq \exp\{-\frac{\delta^2(1-\gamma)^2(t-h-1)}{2h^2}\} \text{ and}$$
$$\Pr\left[\mathbb{E}[f_{(h)}(Y^t; \mathbf{f})] - f_{(h)}(Y^t; \mathbf{f}) > (t-h-1)\delta\right] \leq \exp\{-\frac{\delta^2(1-\gamma)^2(t-h-1)}{2h^2}\},$$

□

B.4 Technical details about the transportation cost method

Let $t > 0$. Consider a metric space $\mathcal{X}^t$ with metric ρ.

We will consider functions of the form $f : \mathcal{X}^t \to \mathbb{R}$ that are Lipschitz with respect to metric ρ; that is, there exists some $L > 0$ such that

$$|f(X_1^t) - f(X_2^t)| \leq L\rho(X_1^t, X_2^t).$$

We denote the Lipschitz constant of the function by $\|f\|_{\mathsf{Lip}}$.

Now we define a useful notion of distance called the Wasserstein distance.

Definition 12. *The Wasserstein distance between distributions $\mathbb{P}$ and $\mathbb{Q}$ on $\mathcal{X}^t$ with respect to metric ρ is defined as*

$$W_\rho(\mathbb{P}, \mathbb{Q}) = \sup_{f:\|f\|_{\mathsf{Lip}} \leq 1} \int f d\mathbb{P} - f d\mathbb{Q} = \inf_{\mathbb{M} \text{ couples } \mathbb{P} \text{ and } \mathbb{Q} \text{ on } (X_1^t, X_2^t)} \mathbb{E}\left[\rho(X_1^t, X_2^t)\right]$$

We will consider $X^t \in \mathcal{X}^t$ be distributed according to $\mathbb{P}$. For a function f such that $\|f\|_{\mathsf{Lip}} = L$, we care about the concentration of the quantity $f(X^t)$ around its mean, $\mathbb{E}[f(X^t)]$, as a function of t.

In our case, $\mathcal{X}^t = \{0, 1\}^t$ is finite. We consider the *additive Hamming metric*

$$\rho(X_1^t, X_2^t) := \sum_{s=1}^{t} \mathbb{I}[X_{1,s} \neq X_{2,s}]. \tag{48}$$

(For the special case of $t = 1$ the Wasserstein distance between $\mathbb{P}$ and $\mathbb{Q}$ corresponding to this metric is the total variation distance, denoted by $\|\mathbb{P} - \mathbb{Q}\|_{TV}$.

Our basic ingredient is a transportation cost inequality, which we define below.

Definition 13. *We say that the distribution $\mathbb{P}$ satisfies a transportation cost inequality if, for every distribution $\mathbb{Q}$, we have*

$$W_\rho(\mathbb{P}, \mathbb{Q}) \leq \sqrt{2\gamma D(\mathbb{Q} \parallel \mathbb{P})} \tag{49}$$

Marton showed [M+96] that a transportation inequality on the underlying distribution $\mathbb{P}$ on X^t implied nice concentration bounds on $f(X^t)$ around its mean, when $f(\cdot)$ is Lipshitz with respect to the metric ρ. This technique is powerful because we can establish transportation cost inequalities for a much broader class of distributions $\mathbb{P}$ than just product distributions; in particular, we can handle weak dependencies. In the special case of the Wasserstein metric corresponding to *total variation distance*, the classical Pinsker's inequality is a special case of the transportation cost inequality (49). It turns out we can adapt Pinsker's inequality together with the chain rule on KL-divergence to prove a transportation cost inequality on the additive Hamming distance for product distributions [Mar86]. We can also do this more generally for the case where $\mathbb{P}$ is a Markov distribution on $\mathcal{X}^t$, provided the Markov chain satisfies an important *contractivity condition.* Consider the Markov process with stationary distribution $\mathbb{P}_1(.)$, and transition probabilities $\mathbb{P}_{-1}(.|x)$ for all $x \in \mathcal{X}$. We again define γ-contractivity for a general-state-space Markov process below.

Definition 14. *A Markov chain is γ-contractive if for every two states $x, x' \in \mathcal{X}$, we have*

$$\|\mathbb{P}_{-1}(.|x) - \mathbb{P}_{-1}(.|x')\|_{TV} \leq \gamma < 1. \tag{50}$$

Under this condition, the Markov distribution satisfies a transportation cost inequality, as shown by the following theorem.

Theorem 3 ([M$^+$96])**.** *Let $\mathbb{P}$ be a Markov distribution on $\mathcal{X}^t$ that satisfies Equation* (50) *with parameter $\gamma < 1$. Then, we have*

$$W_\rho(\mathbb{P}, \mathbb{Q}) \leq \frac{1}{1-\gamma}\sqrt{\frac{t}{2} D(\mathbb{Q} \parallel \mathbb{P})} \tag{51}$$

This directly implies a concentration bound of the form

$$\Pr[|f(X^t) - \mathbb{E}[f(X^t)]| > \delta t] \leq 2\exp\{-\frac{2\delta^2(1-\gamma)^2 t}{L^2}\} \tag{52}$$

C Algorithmic benefits of ContextTreeAdaHedge(D)

In this section, we expound on the algorithmic benefits of ContextTreeAdaHedge(D) equipped with prior function $g(\cdot)$: in particular, we formally show the reduced computational complexity of the algorithm, and the equivalence of the computationally efficient update in Equation (6a) and the computationally naive update in Equation (5). The equivalence was originally proved for the multiplicative weights algorithm with a fixed learning rate [HS97]: here, we generalize the argument to include the family of exponential-weights updates with a time-varying, data-dependent learning rate.

Proposition 2. *The runtime of* ContextTreeAdaHedge(D) *per prediction round is $\mathcal{O}(2^D)$.*

Proof. Consider round t of prediction. To carry out the efficient update in Equation (6a), we need to visit every node in the path of the context X_t. Since the full context is of length D, the update runs in $\mathcal{O}(D)$. To perform the prediction, we must calculate the probability distribution $\mathbf{w}_t$, which has 2 entries. To calculate $\mathbf{w}_t$, we must visit every node in the single complete height D tree to access the cumulative loss vectors $\{\mathbf{L}_{x(D),t}\}_{x(D) \in \mathcal{X}^D}$.

Since there are 2^D such loss vectors (i.e. 2^D nodes to visit), this operation takes $\mathcal{O}(2^D)$ time. For a general prior, these cumulative contextual losses are accessed for every value of $h \in \{0, 1, \ldots, D\}$. Thus, the total computational complexity of performing an update is

$$\sum_{h=0}^{D} 2^h = 2^{D+1} - 1 \in \mathcal{O}(2^D).$$

After performing prediction and receiving loss feedback, we need to access all these nodes again and update the cumulative losses. By a similar argument as above, this is also a $\mathcal{O}(2^D)$ operation. Therefore, the total computational compelexity per round is $\mathcal{O}(2^D)$. □

Computational complexity reduction: equivalence of updates Here, we state and prove the following proposition which shows equivalence of the naive update in Equation (5) and the computationally efficient update in Equation (6a).

Proposition 3. *For any prior function $g : \{0, 1, \ldots, D\} \to \mathbb{R}_+$, the updates in Equation* (6a) *and Equation* (5) *are equivalent.*

Proof. It is convenient, for the purposes of this proof, to consider the overcounted set of tree experts ranging from orders 0 to D. In particular, any d^{th}-order tree expert is described by a function $f' : \mathcal{X}^d \to \mathcal{X}$ and there are 2^{2^d} such experts. Corresponding to prior function $g(\cdot)$, we set the initial distribution on tree experts:

$$w_{1,\mathbf{f}}^{(\mathsf{tree})} = \frac{\sum_{h=\mathrm{order}(\mathbf{f})}^{D} g(h)}{Z(g)}$$

where $Z(g)$ is the initial normalizing factor, i.e. $Z(g) = \sum_{h=0}^{D} 2^{2^h} g(h)$.

Recall Equation (5) for the probability of choosing tree expert $\mathbf{f}$ at time t:

$$w_{t,\mathbf{f}}^{(\mathsf{tree})} = \frac{\left(\sum_{h=\mathrm{order}(\mathbf{f})}^{D} g(h)\right) e^{-\eta_t L_{t,\mathbf{f}}}}{Z_t(g)}$$

where

$$Z_t(g) := \sum_{\mathbf{f} \in \mathcal{F}_D} \left(\sum_{h=\mathrm{order}(\mathbf{f})}^{D} g(h) \right) e^{-\eta_t L_{t,\mathbf{f}}}.$$

Also recall Equation (6a) for the probability of $y \in \mathcal{X}$ at time t:

$$w_{t,y} = \frac{\sum_{h=0}^{D} g'(h;\eta_t) e^{-\eta_t L_{X_t(h),t,y}}}{\sum_{h=0}^{D} g'(h;\eta_t) \left(\sum_{y \in \mathcal{X}} e^{-\eta_t L_{X_t(h),t,y}}\right)} \text{ where}$$

$$g'(h;\eta_t) = g(h) \prod_{x(h) \neq X_t(h)} \left(\sum_{y \in \mathcal{X}} e^{-\eta_t L_{x(h),t,y}} \right)$$

To show equivalence, it clearly suffices to show for every $y \in \mathcal{X}$ that

$$\sum_{\mathbf{f} \in \mathcal{F}_D : \mathbf{f}(X_t) = y} w_{t,\mathbf{f}}^{(\mathsf{tree})} = w_{t,y}. \tag{53}$$

We have

$$\begin{aligned}
\sum_{\mathbf{f} \in \mathcal{F}_D : f(X_t) = j} w_{t,\mathbf{f}}^{(\mathsf{tree})} &= \sum_{h=0}^{D} \sum_{\substack{f:\mathrm{order}(f)=h \\ f:f(X_t(h))=y}} w_{t,\mathbf{f}}^{(\mathsf{tree})} \\
&= \sum_{h=0}^{D} \sum_{\substack{f:\mathrm{order}(f)=h \\ f:f(X_t(h))=y}} \frac{g(h)}{Z_t(g)} \prod_{x(h) \in \mathcal{X}^h} e^{-\eta_t L_{x(h),t,f(x(h))}} \\
&= \sum_{h=0}^{D} \frac{g(h)}{Z_t(g)} e^{-\eta_t L_{X_t(h),t,y}} \prod_{x(h) \neq X_t(h)} \left(\sum_{y' \in \mathcal{X}} e^{-\eta_t L_{x(h),t,y}} \right) \\
&= \frac{\sum_{h=0}^{D} g'(h;\eta_t) e^{-\eta_t L_{X_t(h),t,y}}}{Z_t(g)}
\end{aligned}$$

where we have used the distributive law of multiplication over addition, and substituted the definition of $g'(h;\eta_t)$. To complete the proof of equivalence, it remains to show that

$$Z_t(g) = \sum_{h=0}^{D} g'(h;\eta_t) \left(\sum_{y \in \mathcal{X}} e^{-\eta_t L_{X_t(h),t,y}} \right). \tag{54}$$

We use the distributive law to get

$$
\begin{aligned}
Z_t(g) &:= \sum_{f \in \mathcal{F}_D} \left(\sum_{h=\text{order}(f)}^{D} g(h) \right) \prod_{x(h) \in \mathcal{X}^h} e^{-\eta_t L_{x(h),t,f(x(h))}} \\
&= \sum_{h=0}^{D} g(h) \prod_{x(h) \in \mathcal{X}^h} \left(\sum_{y \in \mathcal{X}} e^{-\eta_t L_{x(h),t,y}} \right).
\end{aligned}
$$

We also substitute the expression for $g'(h;\eta_t)$ to get

$$
\begin{aligned}
\sum_{h=0}^{D} g'(h;\eta_t) \left(\sum_{y \in \mathcal{X}} e^{-\eta_t L_{X_t(h),t,y}} \right) &= \sum_{h=0}^{D} g(h) \left(\prod_{x(h) \neq X_t(h)} \left(\sum_{y \in \mathcal{X}} e^{-\eta_t L_{x(h),t,y}} \right) \right) \left(\sum_{y \in \mathcal{X}} e^{-\eta_t L_{X_t(h),t,y}} \right) \\
&= \sum_{h=0}^{D} g(h) \prod_{x(h) \in \mathcal{X}^h} \left(\sum_{y \in \mathcal{X}} e^{-\eta_t L_{x(h),t,y}} \right).
\end{aligned}
$$

Thus, Equation (54) holds. This completes the proof of equivalence of algorithms.

□

D Supplementary algebra

In this section, we state a couple of supplementary algebraic statements (and prove them when necessary).

Fact 1. *For two quantities $B, C \geq 0$, we have* $\max\{B, C\} \leq B + C$.

Fact 2. *For two numbers $B, C \geq 0$,*

$$
x^2 - Bx - C \leq 0 \implies x \leq \sqrt{C} + B.
$$

This results from the quadratic formula, which gives us

$$
\begin{aligned}
x &\leq \frac{B + \sqrt{B^2 + 4C}}{2} \\
&\leq \frac{B + B + 2\sqrt{C}}{2} = \sqrt{C} + B
\end{aligned}
$$

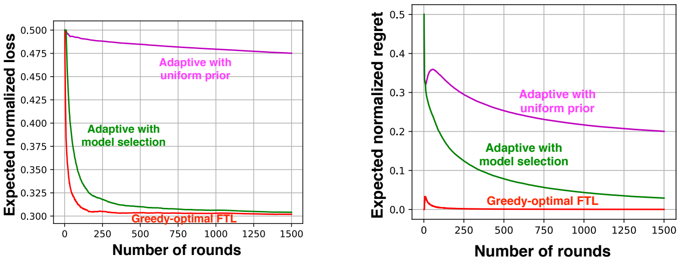
where the last inequality is a consequence of

$$
a, b \geq 0 \implies \sqrt{a+b} \leq \sqrt{a} + \sqrt{b}.
$$

E Extra simulations to illustrate model adaptivity

In this section, we provide a couple of supplementary simulation to the ones in Figure 2 to show the maximal extent of advantage that adaptivity to the model order can give us. First, we examine the 0^{th}-order stochastic model on $\{(X_t, Y_t)\}_{t \geq 1}$, that is, Y_t i.i.d Ber(0.7) and Y_t is independent of X_t, and again compare three algorithms: the *optimal online algorithm with oracle knowledge of this structure* (the greedy FOLLOW-THE-LEADER); uniform-prior CONTEXTTREEADAHEDGE(D), which adapts to stochasticity but not model order; and our two-fold adaptive algorithm, CONTEXTTREEADAHEDGE(D) with the prior function $g_{\mathsf{prop}}(\cdot)$.

Figure 3 shows the evolution of regret and cumulative loss of all three algorithms. The advantage of adaptivity is even more stark in the simple iid case: CONTEXTTREEADAHEDGE(D) with prior function $g_{\mathsf{prop}}(\cdot)$ is very close in its performance to the greedy optimal Follow-the-Leader algorithm. The



(a) Total loss as a function of T.

(b) $R_{T,0}$ as a function of T.

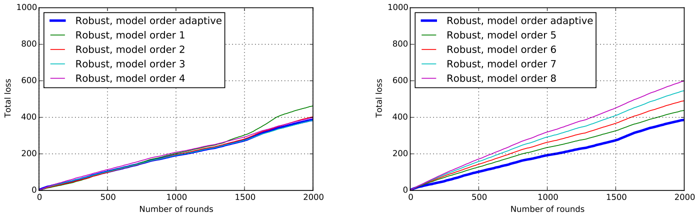
Figure 3. Comparison of optimal greedy FTL, ContextTreeAdaHedge(D) with uniform prior and prior function $g_{\mathsf{prop}}(\cdot)$ (where $D = 8$); against iid structure, upto $T = 1500$ rounds.

disadvantage of adaptivity is also very clearly illustrated: uniform-prior ContextTreeAdaHedge(D) is hugely overfitting for this simple iid example.

Second, we explore an additional example of an HMM with the following parameters:

$$\text{Hidden state evolution } W_{t+1} \sim \text{Ber}\left(|W_t - 0.001|\right)$$
$$Y_t|W_t = 0 \sim \text{Ber}(0.2) \text{ and } Y_t|W_t = 1 \sim \text{Ber}(0.9).$$

This is an interesting example of a HMM with very slowly transitioning hidden states – we expect there to be longer-range dependencies here than in the HMM with quickly transitioning hidden states that we considered in Section 5. From the simulation results in Figure 4, it appears that the best model fit is of order 3 or 4; we observe that our adaptive algorithm naturally tracks the performance of such a model fit in this example as well. If we do not select models of roughly this order, we either overfit or underfit as seen in the simulations. It is worth noting that depending on the parameters of the HMM, different model orders could be considered as optimal fits for increasing numbers of round; it is notable that ContextTreeAdaHedge(D) adapts to a suitable model order for different choices of parameters.



(a) Total loss as a function of T compared to lower-model orders.

(b) Total loss as a function of T compared to higher-model orders.

Figure 4. Comparison of model-adaptive ContextTreeAdaHedge(D) with uniform-prior ContextTreeAdaHedge(d) for fixed model orders on a HMM with slowly transitioning states.