

Prediction with Expert Advice under Discounted Loss

Alexey Chernov and Fedor Zhdanov

Computer Learning Research Centre, Department of Computer Science
Royal Holloway, University of London, Egham, Surrey TW20 0EX, UK
`{chernov,fedor}@cs.rhul.ac.uk`

June 7, 2010

Abstract

We study prediction with expert advice in the setting where the losses are accumulated with some discounting and the impact of old losses can gradually vanish. We generalize the Aggregating Algorithm and the Aggregating Algorithm for Regression, propose a new variant of exponentially weighted average algorithm, and prove bounds on the cumulative discounted loss.

1 Introduction

Prediction with expert advice is a framework for online sequence prediction. Predictions are made step by step. The quality of each prediction (the discrepancy between the prediction and the actual outcome) is evaluated by a real number called loss. The losses are accumulated over time. In the standard framework for prediction with expert advice (see the monograph [2] for a comprehensive review), the losses from all steps are just summed. In this paper, we consider a generalization where older losses can be devalued; in other words, we use discounted cumulative loss.

Predictions are made by Experts and Learner according to Protocol 1. In

Protocol 1 Prediction with expert advice under general discounting

$\mathcal{L}_0 := 0$.
 $\mathcal{L}_0^\theta := 0, \theta \in \Theta$.
for $t = 1, 2, \dots$ **do**
 Accountant announces $\alpha_{t-1} \in (0, 1]$.
 Experts announce $\gamma_t^\theta \in \Gamma, \theta \in \Theta$.
 Learner announces $\gamma_t \in \Gamma$.
 Reality announces $\omega_t \in \Omega$.
 $\mathcal{L}_t^\theta := \alpha_{t-1} \mathcal{L}_{t-1}^\theta + \lambda(\gamma_t^\theta, \omega_t), \theta \in \Theta$.
 $\mathcal{L}_t := \alpha_{t-1} \mathcal{L}_{t-1} + \lambda(\gamma_t, \omega_t)$.
end for

this protocol, Ω is the set of possible outcomes and $\omega_1, \omega_2, \omega_3 \dots$ is the sequence

to predict; Γ is the set of admissible predictions, and $\lambda: \Gamma \times \Omega \rightarrow [0, \infty]$ is the loss function. The triple $(\Omega, \Gamma, \lambda)$ specifies the game of prediction. The most common examples are the binary square loss, log loss, and absolute loss games. They have $\Omega = \{0, 1\}$ and $\Gamma = [0, 1]$, and their loss functions are $\lambda^{\text{sq}}(\gamma, \omega) = (\gamma - \omega)^2$, $\lambda^{\text{log}}(\gamma, 0) = -\log(1 - \gamma)$ and $\lambda^{\text{log}}(\gamma, 1) = -\log \gamma$, $\lambda^{\text{abs}}(\gamma, \omega) = |\gamma - \omega|$, respectively.

The players in the game of prediction are Experts θ from some pool Θ , Learner, and also Accountant and Reality. We are interested in (worst-case optimal) strategies for Learner, and thus the game can be regarded as a two-player game, where Learner opposes the other players. The aim of Learner is to keep his total loss \mathcal{L}_t small as compared to the total losses \mathcal{L}_t^θ of all experts $\theta \in \Theta$.

The standard protocol of prediction with expert advice (as described in [19, 20]) is a special case of Protocol 1 where Accountant always announces $\alpha_t = 1$, $t = 0, 1, 2, \dots$. The new setting gives some more freedom to Learner's opponents.

Another important special case is the exponential (geometric) discounting $\alpha_t = \alpha \in (0, 1)$. Exponential discounting is widely used in finance and economics (see, e.g., [16]), time series analysis (see, e.g., [8]), reinforcement learning [18], and other applications. In the context of prediction with expert advice, Freund and Hsu [6] noted that the discounted loss provides an alternative to “tracking the best expert” framework [11]. Indeed, an exponentially discounted sum depends almost exclusively on the last $O(\log(1/\alpha))$ terms. If the expert with the best one-step performance changes at this rate, then Learner observing the α -discounted losses will mostly follow predictions of the current best expert. Under our more general discounting, more subtle properties of best expert changes may be specified by varying the discount factor. In particular, one can cause Learner to “restart mildly” giving $\alpha_t = 1$ (or $\alpha_t \approx 1$) most of the time and $\alpha_t \ll 1$ at crucial moments. (We prohibit $\alpha_t = 0$ in the protocol, since this is exactly the same as the stopping the current game and starting a new, independent game; on the other hand, the assumption $\alpha_t \neq 0$ simplifies some statements.)

Cesa-Bianchi and Lugosi [2, § 2.11] discuss another kind of discounting

$$L_T = \sum_{t=1}^T \beta_{T-t} l_t, \quad (1)$$

where l_t are one-step losses and β_t are some decreasing discount factors. To see the difference, let us rewrite our definition in the same style:

$$\begin{aligned} L_T &= \alpha_{T-1} L_{T-1} + l_T = \alpha_{T-2} \alpha_{T-1} L_{T-2} + \alpha_{T-1} l_{T-1} + l_T = \dots \\ &= \sum_{t=1}^T \alpha_t \cdots \alpha_{T-1} l_t = \frac{1}{\beta_T} \sum_{t=1}^T \beta_t l_t, \end{aligned} \quad (2)$$

where $\beta_t = 1/\alpha_1 \cdots \alpha_{t-1}$, $\beta_1 = 1$. The sequence β_t is *non-decreasing*, $\beta_1 \leq \beta_2 \leq \beta_3 \leq \dots$; but it is applied “in the reverse order” compared to (1). So, in both definitions, the older losses are the less weight they are ascribed. However, according to (1), the losses l_t have different relative weights in L_T , L_{T+1} and so on, whereas (2) fixes the relative weight of l_t with respect to all previous losses forever starting from the moment t . The latter property allows us to get uniform

algorithms for Learner with loss guarantees that hold for all $T = 1, 2, \dots$; in contrast, Theorem 2.8 in [2] gives a guarantee only at one moment T chosen in advance. The only kind of discounting that can be expressed both as (1) and as (2) is the exponential discounting $\sum_{t=1}^T \alpha^{T-t} l_t$. Under this discounting, NormalHedge algorithm is analysed in [6]; we briefly compare the obtained bounds in Section 3.

Let us say a few words about “economical” interpretation of discounting. Recall that $\alpha_t \leq 1$ in Protocol 1, in other words, the previous cumulative loss cannot become more important at later steps. If the losses are interpreted as the lost money, it is more natural to assume that the old losses must be multiplied by something greater than 1. Indeed, the money could have been invested and have brought some interest, so the current value of an ancient small loss can be considerably large. Nevertheless, there is a not so artificial interpretation for our discounting model as well. Assume that the loss at each step is expressed as a quantity of some goods, and we pay for them in cash; say, we pay for apples damaged because of our incorrect weather prediction. The price of apples can increase but never decreases. Then β_t in (2) is the current price, $\sum_{t=1}^T \beta_t l_t$ is the total sum of money we lost, and L_T is the quantity of apples that we could have bought now if we had not lost so much money. (We must also assume that we cannot hedge our risk by buying a lot of cheap apples in advance—the apples will rot—and that the bank interest is zero.)

We need the condition $\alpha_t \leq 1$ for our algorithms and loss bounds. However, the case of $\alpha_t \geq 1$ is no less interesting. We cannot say anything about it and leave it as an open problem, as well as the general case of arbitrary positive α_t .

The rest of the paper is organized as follows. In Section 2, we propose a generalization of the Aggregating Algorithm [20] and prove the same bound as in [20] but for the discounted loss. In Section 3, we consider convex loss functions and propose an algorithm similar to the Weak Aggregating Algorithm [14] and the exponentially weighted average forecaster with time-varying learning rate [2, § 2.3], with a similar loss bound. In Section 4, we consider the use of prediction with expert advice for the regression problem and adapt the Aggregating Algorithm for Regression [22] (applied to spaces of linear functions and to reproducing kernel Hilbert spaces) to the discounted square loss. All our algorithms are inspired by the methodology of defensive forecasting [4]. We do not explicitly use or refer to this technique in the main text. However, to illustrate these ideas we provide an alternative treatment of the regression task with the help of defensive forecasting in Appendix A.2.

2 Linear Bounds for Learner’s Loss

In this section, we assume that the set of experts is finite, $\Theta = \{1, \dots, K\}$, and show how Learner can achieve a bound of the form $\mathcal{L}_t \leq c\mathcal{L}_t^k + (c \ln K)/\eta$ for all Experts k , where $c \geq 1$ and $\eta > 0$ are constants. Bounds of this kind were obtained in [19]. Loosely speaking, such a bound holds for certain c and η if and only if the game $(\Omega, \Gamma, \lambda)$ has the following property:

$$\exists \gamma \in \Gamma \forall \omega \in \Omega \quad \lambda(\gamma, \omega) \leq -\frac{c}{\eta} \ln \left(\sum_{i \in I} p_i e^{-\eta \lambda(\gamma_i, \omega)} \right) \quad (3)$$

for any finite index set I , for any $\gamma_i \in \Gamma$, $i \in I$, and for any $p_i \in [0, 1]$ such that $\sum_{i \in I} p_i = 1$. It turns out that this property is sufficient for the discounted case as well.

Theorem 1. *Suppose that the game $(\Omega, \Gamma, \lambda)$ satisfies condition (3) for certain $c \geq 1$ and $\eta > 0$. In the game played according to Protocol 1, Learner has a strategy guaranteeing that, for any T and for any $k \in \{1, \dots, K\}$, it holds*

$$\mathcal{L}_T \leq c\mathcal{L}_T^k + \frac{c \ln K}{\eta}. \quad (4)$$

We formulate the strategy for Learner in Subsection 2.1 and prove the theorem in Subsection 2.2.

For the standard undiscounted case (Accountant announces $\alpha_t = 1$ at each step t), this theorem was proved by Vovk in [19] with the help of the Aggregating Algorithm (AA) as Learner's strategy. It is known ([10, 20]) that this bound is asymptotically optimal for large pools of Experts (for games satisfying some assumptions): if the game does not satisfy (3) for some $c \geq 1$ and $\eta > 0$, then, for sufficiently large K , there is a strategy for Experts and Reality (recall that Accountant always says $\alpha_t = 1$) such that Learner cannot secure (4). For the special case of $c = 1$, bound (4) is tight for any fixed K as well [21]. These results imply optimality of Theorem 1 in the new setting with general discounting (when we allow arbitrary behaviour of Accountant with the only requirement $\alpha_t \in (0, 1]$). However, they leave open the question of lower bounds under different discounting assumptions (that is, when Accountant moves are fixed); a particularly interesting case is the exponential discounting $\alpha_t = \alpha \in (0, 1)$.

2.1 Learner's Strategy

To prove Theorem 1, we will exploit the AA with a minor modification.

Algorithm 1 The Aggregating Algorithm

- 1: Initialize weights of Experts $w_0^k := 1/K$, $k = 1, \dots, K$.
 - 2: **for** $t = 1, 2, \dots$ **do**
 - 3: Get Experts' predictions $\gamma_t^k \in \Gamma$, $k = 1, \dots, K$.
 - 4: Calculate $g_t(\omega) = -\frac{c}{\eta} \ln \left(\sum_{k=1}^K w_{t-1}^k e^{-\eta \lambda(\gamma_t^k, \omega)} \right)$, for all $\omega \in \Omega$.
 - 5: Output $\gamma_t := \sigma(g_t) \in \Gamma$.
 - 6: Get $\omega_t \in \Omega$.
 - 7: Update the weights $\tilde{w}_t^k := w_{t-1}^k e^{-\eta \lambda(\gamma_t^k, \omega_t)}$, $k = 1, \dots, K$,
 - 8: and normalize them $w_t^k := \tilde{w}_t^k / \sum_{k=1}^K \tilde{w}_t^k$, $k = 1, \dots, K$.
 - 9: **end for**.
-

The pseudocode of the AA is given as Algorithm 1. The algorithm has three parameters, which depend on the game $(\Omega, \Gamma, \lambda)$: $c \geq 1$, $\eta > 0$, and a function $\sigma: \mathbb{R}^\Omega \rightarrow \Gamma$. The function σ is called a *substitution function* and must have the following property: $\lambda(\sigma(g), \omega) \leq g(\omega)$ for all $\omega \in \Omega$ if for $g \in \mathbb{R}^\Omega$ there exists any $\gamma \in \Gamma$ such that $\lambda(\gamma, \omega) \leq g(\omega)$ for all $\omega \in \Omega$. A natural example of substitution function is given by

$$\sigma(g) = \arg \min_{\gamma \in \Gamma} (\lambda(\gamma, \omega) - g(\omega)) \quad (5)$$

(if the minimum is attained at several points, one can take any of them). An advantage of this σ is that the normalization step in line 8 is not necessary and one can take $w_t^k = \tilde{w}_t^k$. Indeed, multiplying all w_t^k by a constant (independent of k) we add to all $g_t(\omega)$ a constant (independent of ω), and $\sigma(g_t)$ does not change.

The Aggregating Algorithm with Discounting (AAD) differs only by the use of the weights in the computation of g_t and the update of the weights.

The pseudocode of the AAD is given as Algorithm 2.

Algorithm 2 The Aggregating Algorithm with Discounting

- 1: Initialize weights of Experts $w_0^k := 1, k = 1, \dots, K$.
 - 2: **for** $t = 1, 2, \dots$ **do**
 - 3: Get discount $\alpha_{t-1} \in (0, 1]$.
 - 4: Get Experts' predictions $\gamma_t^k \in \Gamma, k = 1, \dots, K$.
 - 5: Calculate $g_t(\omega) = -\frac{c}{\eta} \left(\ln \sum_{k=1}^K \frac{1}{K} (w_{t-1}^k)^{\alpha_{t-1}} e^{-\eta \lambda(\gamma_t^k, \omega)} \right)$, for all $\omega \in \Omega$.
 - 6: Output $\gamma_t := \sigma(g_t) \in \Gamma$.
 - 7: Get $\omega_t \in \Omega$.
 - 8: Update the weights $w_t^k := (w_{t-1}^k)^{\alpha_{t-1}} e^{\eta \lambda(\gamma_t, \omega_t) / c - \eta \lambda(\gamma_t^k, \omega_t)}$, $k = 1, \dots, K$,
 - 9: **end for**.
-

For a substitution function satisfying (5), one can use in line 8 the update rule $w_t^k := (w_{t-1}^k)^{\alpha_{t-1}} e^{-\eta \lambda(\gamma_t^k, \omega_t)}$, which does not contain Learner's losses, in the same manner as the normalization in Algorithm 1 can be omitted.

2.2 Proof of the Bound

Assume that c and η are such that condition (3) holds for the game. Let us show that Algorithm 2 preserves the following condition:

$$\sum_{k=1}^K \frac{1}{K} w_t^k \leq 1. \quad (6)$$

Condition (6) trivially holds for $t = 0$. Assume that (6) holds for $t - 1$, that is, $\sum_{k=1}^K w_{t-1}^k / K \leq 1$. Thus, we have

$$\sum_{k=1}^K \frac{1}{K} (w_{t-1}^k)^{\alpha_{t-1}} \leq \left(\sum_{k=1}^K \frac{1}{K} w_{t-1}^k \right)^{\alpha_{t-1}} \leq 1,$$

since the function $x \mapsto x^\alpha$ is concave for $\alpha \in (0, 1]$, $x \geq 0$, and since $x \leq 1$ implies $x^\alpha \leq 1$ for $\alpha \geq 0$ and $x \geq 0$.

Let \tilde{w}^k be any reals such that $\tilde{w}^k \geq (w_{t-1}^k)^{\alpha_{t-1}} / K$ and $\sum_{k=1}^K \tilde{w}^k = 1$. Due to condition (3) there exists $\gamma \in \Gamma$ such that for all $\omega \in \Omega$

$$\begin{aligned} \lambda(\gamma, \omega) &\leq -\frac{c}{\eta} \ln \left(\sum_{k=1}^K \tilde{w}^k e^{-\eta \lambda(\gamma^k, \omega)} \right) \\ &\leq -\frac{c}{\eta} \ln \left(\sum_{k=1}^K \frac{1}{K} (w_{t-1}^k)^{\alpha_{t-1}} e^{-\eta \lambda(\gamma^k, \omega)} \right) = g_t(\omega) \end{aligned}$$

(the second inequality holds due to our choice of \tilde{w}^k). Thus, due to the property of σ , we have $\lambda(\gamma_t, \omega) \leq g_t(\omega)$ for all $\omega \in \Omega$. In particular, this holds for $\omega = \omega_t$, and we get

$$\lambda(\gamma_t, \omega_t) \leq -\frac{c}{\eta} \ln \left(\sum_{k=1}^K \frac{1}{K} (w_{t-1}^k)^{\alpha_{t-1}} e^{-\eta \lambda(\gamma_t^k, \omega_t)} \right),$$

which is equivalent to (6).

To get the loss bound (4), it remains to note that

$$\ln w_t^k = \eta (\mathcal{L}_t/c - \mathcal{L}_t^k).$$

Indeed, for $t = 0$, this is trivial. If this holds for w_{t-1}^k , then

$$\begin{aligned} \ln w_t^k &= \alpha_{t-1} \ln(w_{t-1}^k) + \eta \lambda(\gamma_t, \omega_t)/c - \eta \lambda(\gamma_t^k, \omega_t) \\ &= \alpha_{t-1} \eta (\mathcal{L}_{t-1}/c - \mathcal{L}_{t-1}^k) + \eta \lambda(\gamma_t, \omega_t)/c - \eta \lambda(\gamma_t^k, \omega_t) \\ &= \eta ((\alpha_{t-1} \mathcal{L}_{t-1} + \lambda(\gamma_t, \omega_t))/c - (\alpha_{t-1} \mathcal{L}_{t-1}^k + \lambda(\gamma_t^k, \omega_t))) = \eta (\mathcal{L}_t/c - \mathcal{L}_t^k) \end{aligned}$$

and we get the equality for w_t^k . Thus, condition (6) means that

$$\sum_{k=1}^K \frac{1}{K} e^{\eta (\mathcal{L}_t/c - \mathcal{L}_t^k)} \leq 1, \quad (7)$$

and (4) follows by lower-bounding the sum by any of its terms.

Remark. Everything in this section remains valid, if we replace the equal initial Experts' weights $1/K$ by arbitrary non-negative weights w^k , $\sum_{k=1}^K w^k = 1$. This leads to a variant of (4), where the last additive term is replaced by $\frac{c}{\eta} \ln \frac{1}{w^k}$. Additionally, we can consider any measurable space Θ of Experts and a non-negative weight function $w(\theta)$, and replace sums over K by integrals over Θ . Then the algorithm and its analysis remain valid (if we impose natural integrability conditions on Experts' predictions γ_t^θ ; see [22] for more detailed discussion)—this will be used in Section 4.

3 Learner's Loss in Bounded Convex Games

The linear bounds of the form (4) are perfect when $c = 1$. However, for many games (for example, the absolute loss game), condition (3) does not hold for $c = 1$ (with any $\eta > 0$), and one cannot get a bound of the form $\mathcal{L}_t \leq \mathcal{L}_t^k + O(1)$. Since Experts' losses \mathcal{L}_T^θ may grow as T in the worst case, any bound with $c > 1$ only guarantees that Learner's loss may exceed an Expert's loss by at most $O(T)$. However, for a large class of interesting games (including the absolute loss game), one can obtain guarantees of the form $\mathcal{L}_T \leq \mathcal{L}_T^k + O(\sqrt{T})$ in the undiscounted case. In this section, we prove an analogous result for the discounted setting.

A game $(\Omega, \Gamma, \lambda)$ is non-empty if Ω and Γ are non-empty. The game is called *bounded* if $L = \max_{\omega, \gamma} \lambda(\gamma, \omega) < \infty$. One may assume that $L = 1$ (if not, consider the scaled loss function λ/L). The game is called *convex* if

for any predictions $\gamma_1, \dots, \gamma_M \in \Gamma$ and for any weights $p_1, \dots, p_M \in [0, 1]$, $\sum_{m=1}^M p_m = 1$,

$$\exists \gamma \in \Gamma \forall \omega \in \Omega \quad \lambda(\gamma, \omega) \leq \sum_{m=1}^M p_m \lambda(\gamma_m, \omega). \quad (8)$$

Note that if Γ is a convex set (e.g., $\Gamma = [0, 1]$) and $\lambda(\gamma, \omega)$ is convex in γ (e.g., λ^{abs}), then the game is convex.

Theorem 2. *Suppose that $(\Omega, \Gamma, \lambda)$ is a non-empty convex game, and $\lambda(\gamma, \omega) \in [0, 1]$ for all $\gamma \in \Gamma$ and $\omega \in \Omega$. In the game played according to Protocol 1, Learner has a strategy guaranteeing that, for any T and for any $k \in \{1, \dots, K\}$, it holds*

$$\mathcal{L}_T \leq \mathcal{L}_T^k + \sqrt{\ln K} \sqrt{\frac{B_T}{\beta_T}}, \quad (9)$$

where $\beta_t = 1/(\alpha_1 \cdots \alpha_{t-1})$ and $B_T = \sum_{t=1}^T \beta_t$.

Note that B_T/β_T is the maximal predictors' loss, which incurs when the predictor suffers the maximal possible loss $l_t = 1$ at each step.

In the undiscounted case, $\alpha_t = 1$, thus $\beta_t = 1$, $B_T = T$, and (9) becomes

$$\mathcal{L}_T \leq \mathcal{L}_T^k + \sqrt{T \ln K}.$$

A similar bound (but with worse constant $\sqrt{2}$ instead of 1 before $\sqrt{T \ln K}$) is obtained in [2, Theorem 2.3]:

$$\mathcal{L}_T \leq \mathcal{L}_T^k + \sqrt{2T \ln K} + \sqrt{\frac{\ln K}{8}}.$$

For the exponential discounting $\alpha_t = \alpha$, we have $\beta_t = \alpha^{-t+1}$ and $B_T = (1 - \alpha^{-T})/(1 - 1/\alpha)$, and (9) transforms into

$$\mathcal{L}_T \leq \mathcal{L}_T^k + \sqrt{\ln K} \sqrt{\frac{1 - \alpha^T}{1 - \alpha}} \leq \mathcal{L}_T^k + \sqrt{\frac{\ln K}{1 - \alpha}}.$$

A similar bound (with worse constants) is obtained in [6] for NormalHedge:

$$\mathcal{L}_T \leq \mathcal{L}_T^k + \sqrt{\frac{8 \ln 2.32K}{1 - \alpha}}.$$

The NormalHedge algorithm has an important advantage: it can guarantee the last bound without knowledge of the number of experts K (see [3] for a precise definition). We can achieve the same with the help of a more complicated algorithm but at the price of a worse bound (Theorem 3).

3.1 Learner's Strategy for Theorem 2

The pseudocode of Learner's strategy is given as Algorithm 3. It contains a constant $a > 0$, which we will choose later in the proof.

The algorithm is not fully specified, since lines 6–7 of Algorithm 3 allow arbitrary choice of γ satisfying the inequality. The algorithm can be completed

with the help of a substitution function σ as in Algorithm 2, so that lines 6–8 are replaced by

$$g_t(\omega) = -\frac{1}{\eta_t} \ln \left(\sum_{k=1}^K \frac{1}{K} (w_{t-1}^k)^{\alpha_{t-1}\eta_t/\eta_{t-1}} e^{-\eta_t \lambda(\gamma_t^k, \omega) - \eta_t^2/8} \right)$$

and $\gamma_t = \sigma(g_t)$. However, the current form of Algorithm 3 emphasizes the similarity to the Algorithm 5, which is described later (Subsection 3.3) but actually inspired our analysis.

Algorithm 3 Learner’s Strategy for Convex Games

- 1: Initialize weights of Experts $w_0^k := 1$, $k = 1, \dots, K$.
Set $\beta_1 = 1$, $B_0 = 0$.
 - 2: **for** $t = 1, 2, \dots$ **do**
 - 3: Get discount $\alpha_{t-1} \in (0, 1]$; update $\beta_t = \beta_{t-1}/\alpha_{t-1}$, $B_t = B_{t-1} + \beta_t$.
 - 4: Compute $\eta_t = a\sqrt{\beta_t/B_t}$.
 - 5: Get Experts’ predictions $\gamma_t^k \in \Gamma$, $k = 1, \dots, K$.
 - 6: Find $\gamma \in \Gamma$ s.t. for all $\omega \in \Omega$
 - 7: $\lambda(\gamma, \omega) \leq -\frac{1}{\eta_t} \ln \left(\sum_{k=1}^K \frac{1}{K} (w_{t-1}^k)^{\alpha_{t-1}\eta_t/\eta_{t-1}} e^{-\eta_t \lambda(\gamma_t^k, \omega) - \eta_t^2/8} \right)$
 - 8: Output $\gamma_t := \gamma$.
 - 9: Get $\omega_t \in \Omega$.
 - 10: Update the weights $w_t^k := (w_{t-1}^k)^{\alpha_{t-1}\eta_t/\eta_{t-1}} e^{\eta_t (\lambda(\gamma_t, \omega_t) - \lambda(\gamma_t^k, \omega_t)) - \eta_t^2/8}$,
 - 11: $k = 1, \dots, K$,
 - 12: **end for**.
-

Let us explain the relation of Algorithm 3 to the Weak Aggregating Algorithm [14] and the exponentially weighted average forecaster with time-varying learning rate [2, § 2.3]. To this end, consider Algorithm 4.

Algorithm 4 Weak Aggregating Algorithm with Discounting

- 1: Initialize Experts’ cumulative losses $\mathcal{L}_0^k := 0$, $k = 1, \dots, K$.
Set $\beta_1 = 1$, $B_0 = 0$.
 - 2: **for** $t = 1, 2, \dots$ **do**
 - 3: Get discount $\alpha_{t-1} \in (0, 1]$; update $\beta_t = \beta_{t-1}/\alpha_{t-1}$, $B_t = B_{t-1} + \beta_t$.
 - 4: Compute $\eta_t = a\sqrt{\beta_t/B_t}$.
 - 5: Compute the weights $q_t^k = e^{-\alpha_{t-1}\eta_t \mathcal{L}_{t-1}^k}$, $k = 1, \dots, K$.
 - 6: Compute the normalized weights $\tilde{w}_t^k = q_t^k / \sum_{j=1}^K q_t^j$.
 - 7: Get Experts’ predictions $\gamma_t^k \in \Gamma$, $k = 1, \dots, K$.
 - 8: Find $\gamma \in \Gamma$ s.t. for all $\omega \in \Omega$ $\lambda(\gamma, \omega) \leq \sum_{k=1}^K \tilde{w}_t^k \lambda(\gamma_t^k, \omega)$.
 - 9: Output $\gamma_t := \gamma$.
 - 10: Get $\omega_t \in \Omega$.
 - 11: Update $\mathcal{L}_t^k := \alpha_{t-1} \mathcal{L}_{t-1}^k + \lambda(\gamma_t^k, \omega_t)$, $k = 1, \dots, K$.
 - 12: **end for**.
-

The proof of Theorem 2 implies that Algorithm 4 is a special case of Algorithm 3. Indeed, (15) implies that $w_{t-1}^k = e^{-\eta_{t-1} \mathcal{L}_{t-1}^k + C}$, where C does not depend on k and w_{t-1}^k are the weights from Algorithm 3. Therefore $q_t^k =$

$C'(w_{t-1}^k)^{\alpha_{t-1}\eta_t/\eta_{t-1}}$, where C' does not depend on k , and one can take \tilde{w}_t^k for \tilde{w}^k in the proof of Theorem 2. Thus, if Algorithm 4 output some γ_t then Algorithm 3 can output this γ_t as well.

Recall that if $\alpha_t = 1$ for all t (the undiscounted case), $\beta_t = 1$ and $B_t = t$, hence $\eta_t = a/\sqrt{t}$. In this case, Algorithm 4 is just the Weak Aggregating Algorithm as described in [14].

Consider now the case when Γ is a convex set and $\lambda(\gamma, \omega)$ is convex in γ . Then one can take $\gamma_t = \sum_{k=1}^K \tilde{w}_t^k \gamma_t^k$ in Algorithm 4. For $\alpha_t = 1$, we get exactly the exponentially weighted average forecaster with time-varying learning rate [2, § 2.3].

3.2 Proof of Theorem 2

Similarly to the case of the AAD, let us show that Algorithm 3 always can find γ in lines 6–7 and preserves the following condition:

$$\sum_{k=1}^K \frac{1}{K} w_t^k \leq 1. \quad (10)$$

First check that $\alpha_{t-1}\eta_t/\eta_{t-1} \leq 1$. Indeed, $\alpha_{t-1} = \beta_{t-1}/\beta_t$, and thus

$$\alpha_{t-1} \frac{\eta_t}{\eta_{t-1}} = \frac{\beta_{t-1}}{\beta_t} \frac{a\sqrt{\beta_t/B_t}}{a\sqrt{\beta_{t-1}/B_{t-1}}} = \sqrt{\frac{\beta_{t-1}}{\beta_t} \frac{B_{t-1}}{B_t}} = \sqrt{\alpha_{t-1}} \sqrt{\frac{B_{t-1}}{B_{t-1} + \beta_t}} \leq 1. \quad (11)$$

Condition (10) trivially holds for $t = 0$. Assume that (10) holds for $t - 1$, that is, $\sum_{k=1}^K w_{t-1}^k/K \leq 1$. Thus, we have

$$\sum_{k=1}^K \frac{1}{K} (w_{t-1}^k)^{\alpha_{t-1}\eta_t/\eta_{t-1}} \leq \left(\sum_{k=1}^K \frac{1}{K} w_{t-1}^k \right)^{\alpha_{t-1}\eta_t/\eta_{t-1}} \leq 1, \quad (12)$$

since the function $x \mapsto x^\alpha$ is concave for $\alpha \in (0, 1]$, $x \geq 0$, and since $x \leq 1$ implies $x^\alpha \leq 1$ for $\alpha \geq 0$ and $x \geq 0$.

Let \tilde{w}^k be any reals such that $\tilde{w}^k \geq (w_{t-1}^k)^{\alpha_{t-1}\eta_t/\eta_{t-1}}/K$ and $\sum_{k=1}^K \tilde{w}^k = 1$. (For example, $\tilde{w}^k = (w_{t-1}^k)^{\alpha_{t-1}\eta_t/\eta_{t-1}} / \sum_{j=1}^K (w_{t-1}^j)^{\alpha_{t-1}\eta_t/\eta_{t-1}}$.) By the Höfdding inequality (see, e. g., [2, Lemma 2.2]), we have

$$\ln \sum_{k=1}^K \tilde{w}^k e^{-\eta_t \lambda(\gamma_t^k, \omega)} \leq -\eta_t \sum_{k=1}^K \tilde{w}^k \lambda(\gamma_t^k, \omega) + \frac{\eta_t^2}{8}, \quad (13)$$

since $\lambda(\gamma, \omega) \in [0, 1]$ for any $\gamma \in \Gamma$ and $\omega \in \Omega$. Since the game is convex, there exists $\gamma \in \Gamma$ such that $\lambda(\gamma, \omega) \leq \sum_{k=1}^K \tilde{w}^k \lambda(\gamma_t^k, \omega)$ for all $\omega \in \Omega$. For this γ and for all $\omega \in \Omega$ we have

$$\begin{aligned} \lambda(\gamma, \omega) &\leq \sum_{k=1}^K \tilde{w}^k \lambda(\gamma_t^k, \omega) \leq -\frac{1}{\eta_t} \ln \left(\sum_{k=1}^K \tilde{w}^k e^{-\eta_t \lambda(\gamma_t^k, \omega) - \eta_t^2/8} \right) \\ &\leq -\frac{1}{\eta_t} \ln \left(\sum_{k=1}^K \frac{1}{K} (w_{t-1}^k)^{\alpha_{t-1}\eta_t/\eta_{t-1}} e^{-\eta_t \lambda(\gamma_t^k, \omega) - \eta_t^2/8} \right) \end{aligned} \quad (14)$$

(the second inequality follows from (13), and the third inequality holds due to our choice of \tilde{w}^k). Thus, one can always find γ in lines 6–7 of Algorithm 3. It remains to note that the inequality in line 7 with γ_t substituted for γ and ω_t substituted for ω is equivalent to

$$1 \geq \sum \frac{1}{K} (w_{t-1}^k)^{\alpha_{t-1}\eta_t/\eta_{t-1}} e^{\eta_t \lambda(\gamma_t, \omega_t) - \eta_t \lambda(\gamma_t^k, \omega_t) - \eta_t^2/8} = \sum \frac{1}{K} w_t^k.$$

Now let us check that

$$\ln w_t^k = \eta_t (\mathcal{L}_t - \mathcal{L}_t^k) - \frac{\eta_t}{8\beta_t} \sum_{\tau=1}^t \beta_\tau \eta_\tau. \quad (15)$$

Indeed, for $t = 0$, this is trivial. Assume that it holds for w_{t-1}^k . Then, taking the logarithm of the update expression in line 10 of Algorithm 3 and substituting $\ln w_{t-1}^k$, we get

$$\begin{aligned} \ln w_t^k &= \frac{\alpha_{t-1}\eta_t}{\eta_{t-1}} \ln w_{t-1}^k + \eta_t (\lambda(\gamma_t, \omega_t) - \lambda(\gamma_t^k, \omega_t)) - \frac{\eta_t^2}{8} \\ &= \frac{\alpha_{t-1}\eta_t}{\eta_{t-1}} \left(\eta_{t-1} (\mathcal{L}_{t-1} - \mathcal{L}_{t-1}^k) - \frac{\eta_{t-1}}{8\beta_{t-1}} \sum_{\tau=1}^{t-1} \beta_\tau \eta_\tau \right) + \eta_t (\lambda(\gamma_t, \omega_t) - \lambda(\gamma_t^k, \omega_t)) - \frac{\eta_t^2}{8} \\ &= \eta_t (\alpha_{t-1} \mathcal{L}_{t-1} + \lambda(\gamma_t, \omega_t) - \alpha_{t-1} \mathcal{L}_{t-1}^k - \lambda(\gamma_t^k, \omega_t)) - \frac{\eta_t}{8\beta_t} \sum_{\tau=1}^{t-1} \beta_\tau \eta_\tau - \frac{\eta_t^2}{8} \\ &= \eta_t (\mathcal{L}_t - \mathcal{L}_t^k) - \frac{\eta_t}{8\beta_t} \sum_{\tau=1}^t \beta_\tau \eta_\tau. \end{aligned}$$

Condition (10) implies that $w_T^k \leq K$ for all k and T , hence we get a loss bound

$$\mathcal{L}_T \leq \mathcal{L}_T^k + \frac{\ln K}{\eta_T} + \frac{1}{8\beta_T} \sum_{t=1}^T \beta_t \eta_t. \quad (16)$$

Recall that $\eta_t = a\sqrt{\beta_t/B_t}$. To estimate $\sum_{t=1}^T \beta_t \eta_t$, we use the following inequality (see Appendix A.1 for the proof).

Lemma 1. *Let β_t be any reals such that $1 \leq \beta_1 \leq \beta_2 \leq \dots$. Let $B_T = \sum_{t=1}^T \beta_t$. Then, for any T , it holds*

$$\frac{1}{\beta_T} \sum_{t=1}^T \beta_t \sqrt{\frac{\beta_t}{B_t}} \leq 2\sqrt{\frac{B_T}{\beta_T}}.$$

Then (16) implies

$$\mathcal{L}_T \leq \mathcal{L}_T^k + \frac{\ln K}{a} \sqrt{\frac{B_T}{\beta_T}} + \frac{2a}{8} \sqrt{\frac{B_T}{\beta_T}} = \mathcal{L}_T^k + \left(\frac{\ln K}{a} + \frac{a}{4} \right) \sqrt{\frac{B_T}{\beta_T}}.$$

Choosing $a = 2\sqrt{\ln K}$, we finally get

$$\mathcal{L}_T \leq \mathcal{L}_T^k + \sqrt{\ln K} \sqrt{\frac{B_T}{\beta_T}}.$$

3.3 A Bound with respect to ϵ -Best Expert

Algorithm 3 originates in the “Fake Defensive Forecasting” (FDF) algorithm from [5, Theorem 9]. That algorithm is based on the ideas of defensive forecasting [4], in particular, Hoeffding supermartingales [24], combined with the ideas from an early version of the Weak Aggregating Algorithm [13]. However, our analysis in Theorem 2 is completely different from [5], following the lines of [2, Theorem 2.2] and [13].

In this subsection, we consider a direct extension of the FDF algorithm from [5, Theorem 9] to the discounted case. Algorithm 5 becomes the FDF algorithm when $\alpha_t = 1$.

Algorithm 5 Fake Defensive Forecasting Algorithm with Discounting

- 1: Initialize cumulative losses $\mathcal{L}_0 = 0$, $\mathcal{L}_0^k := 0$, $k = 1, \dots, K$.
Set $\beta_1 = 1$, $B_0 = 0$.
 - 2: **for** $t = 1, 2, \dots$ **do**
 - 3: Get discount $\alpha_{t-1} \in (0, 1]$; update $\beta_t = \beta_{t-1}/\alpha_{t-1}$, $B_t = B_{t-1} + \beta_t$.
 - 4: Compute $\eta_t = \sqrt{\beta_t/B_t}$.
 - 5: Get Experts’ predictions $\gamma_t^k \in \Gamma$, $k = 1, \dots, K$.
 - 6: Find $\gamma \in \Gamma$ s.t. for all $\omega \in \Omega$ $f_t(\gamma, \omega) \leq C_t$,
where f_t and C_t are defined by (17) and (18), respectively.
 - 7: Output $\gamma_t := \gamma$.
 - 8: Get $\omega_t \in \Omega$.
 - 9: Update $\mathcal{L}_t := \alpha_{t-1}\mathcal{L}_{t-1} + \lambda(\gamma_t, \omega_t)$.
 - 10: Update $\mathcal{L}_t^k := \alpha_{t-1}\mathcal{L}_{t-1}^k + \lambda(\gamma_t^k, \omega_t)$, $k = 1, \dots, K$.
 - 11: **end for**.
-

Algorithm 5 in line 6 uses the function

$$f_t(\gamma, \omega) = \sum_{k=1}^K \frac{1}{K} \sum_{j=1}^{\infty} \frac{c}{j^2} \exp \left(j\alpha_{t-1}\eta_t(\mathcal{L}_{t-1} - \mathcal{L}_{t-1}^k) - \frac{j^2\eta_t}{2\beta_t} \sum_{\tau=1}^{t-1} \beta_\tau \eta_\tau \right) \\ \times \exp \left(j\eta_t(\lambda(\gamma, \omega) - \lambda(\gamma_t^k, \omega)) - \frac{j^2\eta_t^2}{2} \right) \quad (17)$$

and the constant

$$C_t = \sum_{k=1}^K \frac{1}{K} \sum_{j=1}^{\infty} \frac{c}{j^2} \exp \left(j\alpha_{t-1}\eta_t(\mathcal{L}_{t-1} - \mathcal{L}_{t-1}^k) - \frac{j^2\eta_t}{2\beta_t} \sum_{\tau=1}^{t-1} \beta_\tau \eta_\tau \right), \quad (18)$$

where $1/c = \sum_{j=1}^{\infty} \frac{1}{j^2}$.

Algorithm 5 is more complicated than Algorithm 3, and the loss bound we get is weaker and holds for a narrower class of games. However, this bound can be stated as a bound for ϵ -quantile regret introduced in [3]. Namely, let \mathcal{L}_t^ϵ be any value such that for at least ϵK Experts their loss \mathcal{L}_t^k after step t is not greater than \mathcal{L}_t^ϵ . The ϵ -quantile regret is the difference between \mathcal{L}_t and \mathcal{L}_t^ϵ . For $\epsilon = 1/K$, we can choose $\mathcal{L}_t^\epsilon = \min_k \mathcal{L}_t^k \leq \mathcal{L}_t^k$ for all $k = 1, \dots, K$, and thus a bound in terms of the ϵ -quantile regret implies a bound in terms of \mathcal{L}_t^k . The value $1/\epsilon$ plays the role of the “effective” number of experts. Algorithm 5 guarantees a bound in terms of \mathcal{L}_t^ϵ for any $\epsilon > 0$, without the prior knowledge

of ϵ , and in this sense the algorithm works for the unknown number of Experts (see [5] for a more detailed discussion).

For Algorithm 5 we need to restrict the class of games we consider. The game is called *compact* if the set $\Lambda = \{\lambda(\gamma, \cdot) \in \mathbb{R}^\Omega \mid \gamma \in \Gamma\}$ is compact in the standard topology of \mathbb{R}^Ω .

Theorem 3. *Suppose that $(\Omega, \Gamma, \lambda)$ is a non-empty convex compact game, Ω is finite, and $\lambda(\gamma, \omega) \in [0, 1]$ for all $\gamma \in \Gamma$ and $\omega \in \Omega$. In the game played according to Protocol 1, Learner has a strategy guaranteeing that, for any T and for any $\epsilon > 0$, it holds*

$$\mathcal{L}_T \leq \mathcal{L}_T^\epsilon + 2\sqrt{\frac{B_T}{\beta_T} \ln \frac{1}{\epsilon}} + 7\sqrt{\frac{B_T}{\beta_T}}, \quad (19)$$

where $\beta_t = 1/(\alpha_1 \cdots \alpha_{t-1})$ and $B_T = \sum_{t=1}^T \beta_t$.

Proof. The most difficult part of the proof is to show that one can find γ in line 6 of Algorithm 5. We do not do this here, but refer to [5]; the proof is literally the same as in [5, Theorem 9] and is based on the supermartingale property of f_t . (The rest of the proof below also follows [5, Theorem 9]; the only difference is in the definition of f_t and C_t .)

Let us check that $C_t \leq 1$ for all t . Clearly, $C_1 = 1$. Assume that we have $C_t \leq 1$. This implies $f_t(\gamma_t, \omega_t) \leq 1$ due to the choice of γ_t , and thus $(f_t(\gamma_t, \omega_t))^{\alpha_t \eta_{t+1}/\eta_t} \leq 1$. Similarly to (11), we have $\alpha_t \eta_{t+1}/\eta_t \leq 1$. Since the function $x \mapsto x^\alpha$ is concave for $\alpha \in (0, 1]$, $x \geq 0$, we get

$$\begin{aligned} 1 &\geq (f_t(\gamma_t, \omega_t))^{\alpha_t \eta_{t+1}/\eta_t} \\ &= \left(\sum_{k=1}^K \frac{1}{K} \sum_{j=1}^{\infty} \frac{c}{j^2} \exp \left(j\eta_t(\mathcal{L}_t - \mathcal{L}_t^k) - \frac{j^2 \eta_t}{2\beta_t} \sum_{\tau=1}^t \beta_\tau \eta_\tau \right) \right)^{\alpha_t \eta_{t+1}/\eta_t} \\ &\geq \sum_{k=1}^K \frac{1}{K} \sum_{j=1}^{\infty} \frac{c}{j^2} \left(\exp \left(j\eta_t(\mathcal{L}_t - \mathcal{L}_t^k) - \frac{j^2 \eta_t}{2\beta_t} \sum_{\tau=1}^t \beta_\tau \eta_\tau \right) \right)^{\alpha_t \eta_{t+1}/\eta_t} \\ &= \sum_{k=1}^K \frac{1}{K} \sum_{j=1}^{\infty} \frac{c}{j^2} \exp \left(j\alpha_t \eta_{t+1}(\mathcal{L}_t - \mathcal{L}_t^k) - \frac{j^2 \eta_{t+1}}{2\beta_{t+1}} \sum_{\tau=1}^t \beta_\tau \eta_\tau \right) = C_{t+1}. \end{aligned}$$

Thus, for each t we have $f_t(\gamma_t, \omega_t) \leq 1$, that is,

$$\sum_{k=1}^K \frac{1}{K} \sum_{j=1}^{\infty} \frac{c}{j^2} \exp \left(j\eta_t(\mathcal{L}_t - \mathcal{L}_t^k) - \frac{j^2 \eta_t}{2\beta_t} \sum_{\tau=1}^t \beta_\tau \eta_\tau \right) \leq 1.$$

For any $\epsilon > 0$, let us take any \mathcal{L}_T^ϵ such that for at least ϵK Experts their losses \mathcal{L}_T^k are smaller than or equal to \mathcal{L}_T^ϵ . Then we have

$$1 \geq \sum_{k=1}^K \frac{1}{K} \sum_{j=1}^{\infty} \frac{c}{j^2} \exp \left(j\eta_t(\mathcal{L}_t - \mathcal{L}_t^k) - \frac{j^2 \eta_t}{2\beta_t} \sum_{\tau=1}^t \beta_\tau \eta_\tau \right)$$

$$\begin{aligned}
&\geq \epsilon \sum_{j=1}^{\infty} \frac{c}{j^2} \exp \left(j\eta_t(\mathcal{L}_t - \mathcal{L}_t^\epsilon) - \frac{j^2\eta_t}{2\beta_t} \sum_{\tau=1}^t \beta_\tau \eta_\tau \right) \\
&\geq \frac{c\epsilon}{j^2} \exp \left(j\eta_t(\mathcal{L}_t - \mathcal{L}_t^\epsilon) - \frac{j^2\eta_t}{2\beta_t} \sum_{\tau=1}^t \beta_\tau \eta_\tau \right)
\end{aligned}$$

for any natural j . Taking the logarithm and rearranging, we get

$$\mathcal{L}_t \leq \mathcal{L}_t^\epsilon + \frac{j}{2\beta_t} \sum_{\tau=1}^t \beta_\tau \eta_\tau + \frac{1}{j\eta_t} \ln \frac{j^2}{c\epsilon}.$$

Substituting $\eta_t = \sqrt{\beta_t/B_t}$ and using Lemma 1, we get

$$\mathcal{L}_t \leq \mathcal{L}_t^\epsilon + \left(j + \frac{2}{j} \ln j + \frac{1}{j} \ln \frac{1}{\epsilon} + \frac{1}{j} \ln \frac{1}{c} \right) \sqrt{\frac{B_t}{\beta_t}}.$$

Letting $j = \left\lceil \sqrt{\ln(1/\epsilon)} \right\rceil + 1$ and using the estimates $j \leq \sqrt{\ln(1/\epsilon)} + 2$, $(\ln j)/j \leq 2$, $(\ln(1/\epsilon))/j \leq \sqrt{\ln(1/\epsilon)}$, $1/j \leq 1$, and $\ln(1/c) = \ln(\pi^2/6) \leq 1$, we obtain the final bound. \square

4 Regression with Discounted Loss

In this section we consider a task of regression, where Learner must predict “labels” $y_t \in \mathbb{R}$ for input instances $x_t \in \mathbf{X} \subseteq \mathbb{R}^n$. The predictions proceed according to Protocol 2. This task can be embedded into prediction with expert

Protocol 2 Competitive online regression

for $t = 1, 2, \dots$ **do**
 Reality announces $x_t \in \mathbf{X}$.
 Learner announces $\gamma_t \in \Gamma$.
 Reality announces $y_t \in \Omega$.
end for

advice if Learner competes with all functions $x \rightarrow y$ from some large class serving as a pool of (imaginary) Experts.

4.1 The Framework and Linear Functions as Experts

Let the input space be $\mathbf{X} \subseteq \mathbb{R}^n$, the set of predictions be $\Gamma = \mathbb{R}$, and the set of outcomes be $\Omega = [Y_1, Y_2]$. In this section we consider the square loss $\lambda^{\text{sq}}(\gamma, y) = (\gamma - y)^2$. Learner competes with a pool of experts $\Theta = \mathbb{R}^n$ (treated as linear functionals on \mathbb{R}^n). Each individual expert is denoted by $\theta \in \Theta$ and predicts $\theta'x_t$ at step t .

Let us take any distribution over the experts $P(d\theta)$. It is known from [19] that (3) holds for the square loss with $c = 1$, $\eta = \frac{2}{(Y_2 - Y_1)^2}$:

$$\exists \gamma \in \Gamma \forall y \in \Omega = [Y_1, Y_2] \quad (\gamma - y)^2 \leq -\frac{1}{\eta} \ln \left(\int_{\Theta} e^{-\eta(\theta'x_t - y)^2} P(d\theta) \right). \quad (20)$$

Denote by X the matrix of size $T \times n$ consisting of the rows of the input vectors x'_1, \dots, x'_T . Let also $W_T = \text{diag}(\beta_1/\beta_T, \beta_2/\beta_T, \dots, \beta_T/\beta_T)$, i.e., W_T is a diagonal matrix $T \times T$. In a manner similar to [22], we prove the following upper bound for Learner's loss.

Theorem 4. *For any $a > 0$, there exists a prediction strategy for Learner in Protocol 2 achieving, for every T and for any linear predictor $\theta \in \mathbb{R}^n$,*

$$\sum_{t=1}^T \frac{\beta_t}{\beta_T} (\gamma_t - y_t)^2 \leq \sum_{t=1}^T \frac{\beta_t}{\beta_T} (\theta' x_t - y_t)^2 + a \|\theta\|^2 + \frac{(Y_2 - Y_1)^2}{4} \ln \det \left(\frac{X' W_T X}{a} + I \right). \quad (21)$$

If, in addition, $\|x_t\|_\infty \leq Z$ for all t , then

$$\sum_{t=1}^T \frac{\beta_t}{\beta_T} (\gamma_t - y_t)^2 \leq \sum_{t=1}^T \frac{\beta_t}{\beta_T} (\theta' x_t - y_t)^2 + a \|\theta\|^2 + \frac{n(Y_2 - Y_1)^2}{4} \ln \left(\frac{Z^2 \sum_{t=1}^T \beta_t}{a \beta_T} + 1 \right). \quad (22)$$

In the undiscounted case ($\alpha_t = 1$ for all t), the bounds in the theorem coincide with the bounds for the Aggregating Algorithm for Regression [22, Theorem 1] with $Y_2 = Y$ and $Y_1 = -Y$, since, as remarked after Theorem 2, $\beta_t = 1$ and $(\sum_{t=1}^T \beta_t) / \beta_T = T$ in the undiscounted case. Recall also that in the case of the exponential discounting ($\alpha_t = \alpha \in (0, 1)$) we have $\beta_t = \alpha^{-t+1}$ and $(\sum_{t=1}^T \beta_t) / \beta_T = (1 - \alpha^{T-1}) / (1 - \alpha) \leq 1 / (1 - \alpha)$. Thus, for the exponential discounting bound (22) becomes

$$\sum_{t=1}^T \alpha^{T-t} (\gamma_t - y_t)^2 \leq \sum_{t=1}^T \alpha^{T-t} (\theta' x_t - y_t)^2 + a \|\theta\|^2 + \frac{n(Y_2 - Y_1)^2}{4} \ln \left(\frac{Z^2 (1 - \alpha^{T-1})}{a(1 - \alpha)} + 1 \right). \quad (23)$$

4.2 Functions from an RKHS as Experts

In this section we apply the kernel trick to the linear method to compete with wider sets of experts. Each expert $f \in \mathcal{F}$ predicts $f(x_t)$. Here \mathcal{F} is a reproducing kernel Hilbert space (RKHS) with a positive definite kernel $k: \mathbf{X} \times \mathbf{X} \rightarrow \mathbb{R}$. For the definition of RKHS and its connection to kernels see [17]. Each kernel defines a unique RKHS. We use the notation $\mathbf{K}_T = \{k(x_i, x_j)\}_{i,j=1,\dots,T}$ for the kernel matrix for the input vectors at step T . In a manner similar to [7], we prove the following upper bound on the discounted square loss of Learner.

Theorem 5. *For any $a > 0$, there exists a strategy for Learner in Protocol 2*

achieving, for every positive integer T and any predictor $f \in \mathcal{F}$,

$$\sum_{t=1}^T \frac{\beta_t}{\beta_T} (\gamma_t - y_t)^2 \leq \sum_{t=1}^T \frac{\beta_t}{\beta_T} (f(x_t) - y_t)^2 + a \|f\|^2 + \frac{(Y_2 - Y_1)^2}{4} \ln \det \left(\frac{\sqrt{W_T} \mathbf{K}_T \sqrt{W_T}}{a} + I \right). \quad (24)$$

Corollary 1. Assume that $c_{\mathcal{F}}^2 = \sup_{x \in \mathbf{X}} k(x, x) < \infty$ for the RKHS \mathcal{F} . Under the conditions of Theorem 5, given in advance any constant \mathcal{T} such that $(\sum_{t=1}^T \beta_t) / \beta_T \leq \mathcal{T}$, one can choose parameter a such that the strategy in Theorem 5 achieves for any $f \in \mathcal{F}$

$$\sum_{t=1}^T \frac{\beta_t}{\beta_T} (\gamma_t - y_t)^2 \leq \sum_{t=1}^T \frac{\beta_t}{\beta_T} (f(x_t) - y_t)^2 + \left(\frac{(Y_2 - Y_1)^2}{4} + \|f\|^2 \right) c_{\mathcal{F}} \sqrt{\mathcal{T}}. \quad (25)$$

where $c_{\mathcal{F}}^2 = \sup_{x \in \mathbf{X}} k(x, x) < \infty$ characterizes the RKHS \mathcal{F} .

Proof. The determinant of a symmetric positive definite matrix is upper bounded by the product of its diagonal elements (see Chapter 2, Theorem 7 in [1]), and thus we have

$$\begin{aligned} \ln \det \left(I + \frac{\sqrt{W_T} \mathbf{K}_T \sqrt{W_T}}{a} \right) &\leq T \ln \left(1 + \frac{c_{\mathcal{F}}^2 \left(\prod_{t=1}^T \frac{\beta_t}{\beta_T} \right)^{1/T}}{a} \right) \\ &\leq T \frac{c_{\mathcal{F}}^2}{a} \left(\prod_{t=1}^T \frac{\beta_t}{\beta_T} \right)^{1/T} \leq T \frac{c_{\mathcal{F}}^2}{a \beta_T} \frac{\sum_{t=1}^T \beta_t}{T} \leq \frac{c_{\mathcal{F}}^2 \mathcal{T}}{a} \end{aligned}$$

(we use $\ln(1+x) \leq x$ and the inequality between the geometric and arithmetic means). Choosing $a = c_{\mathcal{F}} \sqrt{\mathcal{T}}$, we get bound (25) from (24). \square

Recall again that $(\sum_{t=1}^T \beta_t) / \beta_T = (1 - \alpha^{T-1}) / (1 - \alpha) \leq 1 / (1 - \alpha)$ in the case of the exponential discounting ($\alpha_t = \alpha \in (0, 1)$), and we can take $\mathcal{T} = 1 / (1 - \alpha)$.

In the undiscounted case ($\alpha_t = 1$), we have $(\sum_{t=1}^T \beta_t) / \beta_T = T$, so we need to know the number of steps in advance. Then, bound (25) matches the bound obtained in [23, the displayed formula after (33)]. If we do not know an upper bound \mathcal{T} in advance, it is still possible to achieve a bound similar to (25) using the Aggregating Algorithm with Discounting to merge Learner's strategies from Theorem 5 with different values of parameter a , in the same manner as in [23, Theorem 3].

Corollary 2. Assume that $c_{\mathcal{F}}^2 = \sup_{x \in \mathbf{X}} k(x, x) < \infty$ for the RKHS \mathcal{F} . Under the conditions of Theorem 5, there exists a strategy for Learner in Protocol 2 achieving, for every positive integer T and any predictor $f \in \mathcal{F}$,

$$\begin{aligned} \sum_{t=1}^T \frac{\beta_t}{\beta_T} (\gamma_t - y_t)^2 &\leq \sum_{t=1}^T \frac{\beta_t}{\beta_T} (f(x_t) - y_t)^2 + c_{\mathcal{F}} \|f\| (Y_2 - Y_1) \sqrt{\frac{\sum_{t=1}^T \beta_t}{\beta_T}} \\ &\quad + \frac{(Y_2 - Y_1)^2}{2} \ln \frac{\sum_{t=1}^T \beta_t}{\beta_T} + \|f\|^2 + (Y_2 - Y_1)^2 \ln \left(\frac{c_{\mathcal{F}} (Y_2 - Y_1)}{\|f\|} + 2 \right). \quad (26) \end{aligned}$$

Proof. Let us take the strategies from Theorem 5 for $a = 1, 2, 3, \dots$ and provide them as Experts to the Aggregating Algorithm with Discounting, with the square loss function, $\eta = 2/(Y_2 - Y_1)^2$ and initial Experts' weights proportional to $1/a^2$. Then Theorem 1 (extended as described in Remark at the end of Section 2) guarantees that the extra loss of the aggregated strategy (compared to the strategy from Theorem 5 with parameter a) is not greater than $\frac{(Y_2 - Y_1)^2}{2} \ln \frac{a^2}{c}$, where $c = \sum_{k=1}^K 1/k^2$. On the other hand, for the strategy from Theorem 5 with parameter a similarly to the proof of Corollary 1 we get

$$\sum_{t=1}^T \frac{\beta_t}{\beta_T} (\gamma_t - y_t)^2 \leq \sum_{t=1}^T \frac{\beta_t}{\beta_T} (f(x_t) - y_t)^2 + a \|f\|^2 + \frac{c_{\mathcal{F}}^2 (Y_2 - Y_1)^2}{4a} \frac{\sum_{t=1}^T \beta_t}{\beta_T}.$$

Adding $\frac{(Y_2 - Y_1)^2}{2} \ln \frac{a^2}{c}$ to the right-hand side and choosing

$$a = \left\lceil \frac{c_{\mathcal{F}} (Y_2 - Y_1)}{2 \|f\|} \sqrt{\frac{\sum_{t=1}^T \beta_t}{\beta_T}} \right\rceil,$$

we get the statement after simple estimations. \square

4.3 Proofs of Theorems 4 and 5

Let us begin with several technical lemmas from linear algebra. The proofs of some of these lemmas are moved to Appendix A.1.

Lemma 2. *Let A be a symmetric positive definite matrix of size $n \times n$. Let $\theta, b \in \mathbb{R}^n$, c be a real number, and $Q(\theta) = \theta' A \theta + b' \theta + c$. Then*

$$\int_{\mathbb{R}^n} e^{-Q(\theta)} d\theta = e^{-Q_0} \frac{\pi^{n/2}}{\sqrt{\det A}},$$

where $Q_0 = \min_{\theta \in \mathbb{R}^n} Q(\theta)$.

The proof of this lemma can be found in [9, Theorem 15.12.1].

Lemma 3. *Let A be a symmetric positive definite matrix of size $n \times n$. Let $b, z \in \mathbb{R}^n$, and*

$$F(A, b, z) = \min_{\theta \in \mathbb{R}^n} (\theta' A \theta + b' \theta + z' \theta) - \min_{\theta \in \mathbb{R}^n} (\theta' A \theta + b' \theta - z' \theta).$$

Then $F(A, b, z) = -b' A^{-1} z$.

Lemma 4. *Let A be a symmetric positive definite matrix of size $n \times n$. Let $\theta, b_1, b_2 \in \mathbb{R}^n$, c_1, c_2 be real numbers, and $Q_1(\theta) = \theta' A \theta + b_1' \theta + c_1$, $Q_2(\theta) = \theta' A \theta + b_2' \theta + c_2$. Then*

$$\frac{\int_{\mathbb{R}^n} e^{-Q_1(\theta)} d\theta}{\int_{\mathbb{R}^n} e^{-Q_2(\theta)} d\theta} = e^{c_2 - c_1 - \frac{1}{4} (b_2 + b_1)' A^{-1} (b_2 - b_1)}.$$

The previous three lemmas were implicitly used in [22] to derive a bound on the cumulative undiscounted square loss of the algorithm competing with linear experts.

Lemma 5. For any matrix B of size $n \times m$, any matrix C of size $m \times n$, and any real number a such that the matrices $aI_m + CB$ and $aI_n + BC$ are nonsingular, it holds

$$B(aI_m + CB)^{-1} = (aI_n + BC)^{-1}B, \quad (27)$$

where I_n, I_m are the unit matrices of sizes $n \times n$ and $m \times m$, respectively.

Proof. Note that this is equivalent to $(aI_n + BC)B = B(aI_m + CB)$. \square

Lemma 6. For matrix B of size $n \times m$, any matrix C of size $m \times n$, and any real number a , it holds

$$\det(aI_n + BC) = \det(aI_m + CB), \quad (28)$$

where I_n, I_m are the unit matrices of sizes $n \times n$ and $m \times m$, respectively.

4.3.1 Proof of Theorem 4.

We take the Gaussian initial distribution over the experts with a parameter $a > 0$:

$$P_0(d\theta) = \left(\frac{a\eta}{\pi}\right)^{n/2} e^{-a\eta\|\theta\|^2} d\theta.$$

and use “Algorithm 2 with infinitely many Experts”. Repeating the derivations from Subsection 2.2, we obtain the following analogue of (7):

$$\left(\frac{a\eta}{\pi}\right)^{n/2} \int_{\Theta} e^{\eta\left(\sum_{t=1}^T \frac{\beta_t}{\beta_T} (\gamma_t - y_t)^2 - \sum_{t=1}^T \frac{\beta_t}{\beta_T} (\theta' x_t - y_t)^2\right)} e^{-a\eta\|\theta\|^2} d\theta \leq 1.$$

The simple equality

$$\sum_{t=1}^T \frac{\beta_t}{\beta_T} (\theta' x_t - y_t)^2 + a\|\theta\|^2 = \theta'(aI + X'W_T X)\theta - 2 \sum_{t=1}^T \frac{\beta_t}{\beta_T} y_t \theta' x_t + \sum_{t=1}^T \frac{\beta_t}{\beta_T} y_t^2 \quad (29)$$

shows that the integral can be evaluated with the help of Lemma 2:

$$\begin{aligned} & \left(\frac{a\eta}{\pi}\right)^{n/2} \int_{\Theta} e^{-\eta\left(\sum_{t=1}^T \frac{\beta_t}{\beta_T} (\theta' x_t - y_t)^2 + a\|\theta\|^2\right)} d\theta \\ &= \left(\frac{a}{\pi}\right)^{n/2} e^{-\eta \min_{\theta} \left(\sum_{t=1}^T \frac{\beta_t}{\beta_T} (\theta' x_t - y_t)^2 + a\|\theta\|^2\right)} \frac{\pi^{n/2}}{\sqrt{\det(aI + X'W_T X)}}. \end{aligned}$$

We take the natural logarithms of both parts of the bound and using the value $\eta = \frac{2}{(Y_2 - Y_1)^2}$ obtain (21). The determinant of a symmetric positive definite matrix is upper bounded by the product of its diagonal elements (see Chapter 2, Theorem 7 in [1]):

$$\det\left(\frac{X'W_T X}{a} + I\right) \leq \left(\frac{Z^2 \sum_{t=1}^T \beta_t}{a\beta_T} + 1\right)^n,$$

and thus we obtain (22).

4.3.2 Proof of Theorem 5.

We must prove that for each T and each sequence $(x_1, y_1, \dots, x_T, y_T) \in (\mathbf{X} \times \mathbb{R})^T$ the guarantee (24) is satisfied. Fix T and $(x_1, y_1, \dots, x_T, y_T)$. Fix an isomorphism between the linear span of k_{x_1}, \dots, k_{x_T} obtained for the Riesz Representation theorem and $\mathbb{R}^{\tilde{T}}$, where $\tilde{T} \leq T$ is the dimension of the linear span of k_{x_1}, \dots, k_{x_T} . Let $\tilde{x}_1, \dots, \tilde{x}_T \in \mathbb{R}^{\tilde{T}}$ be the images of k_{x_1}, \dots, k_{x_T} , respectively, under this isomorphism. We have then $k(\cdot, x_i) = \langle \cdot, \tilde{x}_i \rangle$ for any x_i .

We apply the strategy from Theorem 4 to $\tilde{x}_1, \dots, \tilde{x}_T$. The predictions of the strategies are the same due to Proposition 1 below. Any expert $\theta \in \mathbb{R}^{\tilde{T}}$ in bound (21) can be represented as

$$\theta = \sum_{i=1}^T c_i \tilde{x}_i = \sum_{i=1}^T c_i k(\cdot, x_i)$$

for some $c_i \in \mathbb{R}$. Thus the experts' predictions are $\theta' \tilde{x}_t = \sum_{i=1}^T c_i k(x_t, x_i)$, and the norm is $\|\theta\|^2 = \sum_{i,j=1}^T c_i c_j k(x_i, x_j)$.

Denote by \tilde{X} the $T \times \tilde{T}$ matrix consisting of the rows of the vectors $\tilde{x}'_1, \dots, \tilde{x}'_T$. From Lemma 6 we have

$$\det \left(\frac{\tilde{X}' W_T \tilde{X}}{a} + I \right) = \det \left(\frac{\sqrt{W_T} \tilde{X} \tilde{X}' \sqrt{W_T}}{a} + I \right).$$

Thus using $\mathbf{K}_T = \tilde{X} \tilde{X}'$ we obtain the upper bound

$$\begin{aligned} \sum_{t=1}^T \frac{\beta_t}{\beta_T} (\gamma_t - y_t)^2 &\leq \sum_{t=1}^T \frac{\beta_t}{\beta_T} \left(\sum_{i=1}^T c_i k(x_t, x_i) - y_t \right)^2 \\ &\quad + a \sum_{i,j=1}^T c_i c_j k(x_i, x_j) + \frac{(Y_2 - Y_1)^2}{4} \ln \det \left(\frac{\sqrt{W_T} \mathbf{K}_T \sqrt{W_T}}{a} + I \right) \end{aligned}$$

for any $c_i \in \mathbb{R}$, $i = 1, \dots, T$. By the Representer theorem (see Theorem 4.2 in [17]) the minimum of $\sum_{t=1}^T \frac{\beta_t}{\beta_T} (f(x_t) - y_t)^2 + a \|f\|^2$ over all $f \in \mathcal{F}$ is achieved on one of the linear combinations from the bound obtained above. This concludes the proof.

4.4 Regression Algorithms

In this subsection we derive explicit form of the prediction strategies for Learner used in Theorems 4 and 5.

4.4.1 Strategy for Theorem 4.

In [22] Vovk suggests for the square loss the following substitution function satisfying (5):

$$\gamma_T = \frac{Y_2 + Y_1}{2} - \frac{g_T(Y_2) - g_T(Y_1)}{2(Y_2 - Y_1)}. \quad (30)$$

It allows us to calculate g_T with unnormalized weights:

$$g_T(y) = -\frac{1}{\eta} \left(\ln \int_{\Theta} e^{-\eta(\theta' A_T \theta - 2\theta'(b_{T-1} + y x_T) + (\sum_{t=1}^{T-1} \frac{\beta_t}{\beta_T} y_t^2 + y^2))} d\theta \right)$$

for any $y \in \Omega = [Y_1, Y_2]$ (here we use the expansion (29)), where

$$A_T = aI + \sum_{t=1}^{T-1} \frac{\beta_t}{\beta_T} x_t x_t' + x_T x_T' = aI + X' W_T X,$$

and $b_{T-1} = \sum_{t=1}^{T-1} \frac{\beta_t}{\beta_T} y_t x_t$. The direct calculation of g_T is inefficient: it requires numerical integration. Instead, we notice that

$$\begin{aligned} \gamma_T &= \frac{Y_2 + Y_1}{2} - \frac{g_T(Y_2) - g_T(Y_1)}{2(Y_2 - Y_1)} \\ &= \frac{Y_2 + Y_1}{2} - \frac{1}{2(Y_2 - Y_1)\eta} \ln \frac{\int_{\Theta} e^{-\eta(\theta' A_T \theta - 2\theta'(b_{T-1} + Y_1 x_T) + (\sum_{t=1}^{T-1} \frac{\beta_t}{\beta_T} y_t^2 + Y_1^2))} d\theta}{\int_{\Theta} e^{-\eta(\theta' A_T \theta - 2\theta'(b_{T-1} + Y_2 x_T) + (\sum_{t=1}^{T-1} \frac{\beta_t}{\beta_T} y_t^2 + Y_2^2))} d\theta} \\ &= \frac{Y_2 + Y_1}{2} - \frac{1}{2(Y_2 - Y_1)\eta} \ln e^{\eta(Y_2^2 - Y_1^2 - (b_{T-1} + (\frac{Y_2 + Y_1}{2})x_T)' A_T^{-1} (\frac{Y_2 - Y_1}{2} x_T))} \\ &= \left(b_{T-1} + \left(\frac{Y_2 + Y_1}{2} \right) x_T \right)' A_T^{-1} x_T, \quad (31) \end{aligned}$$

where the third equality follows from Lemma 4.

The strategy which predicts according to (31) requires $O(n^3)$ operations per step. The most time-consuming operation is the inverse of the matrix A_T . Note that for the undiscounted case the inverse could be computed incrementally using the Sherman-Morrison formula, which leads to $O(n^2)$ operations per step.

4.4.2 Strategy for Theorem 5.

We use following notation. Let

$$\begin{aligned} \mathbf{k}_T &\text{ be the last column of the matrix } \mathbf{K}_T, \mathbf{k}_T = \{k(x_i, x_T)\}_{i=1}^T, \\ \mathbf{Y}_T &\text{ be the column vector of the outcomes } \mathbf{Y}_T = (y_1, \dots, y_T)'. \end{aligned} \quad (32)$$

When we write $\mathbf{Z} = (\mathbf{V}; \mathbf{Y})$ or $\mathbf{Z} = (\mathbf{V}'; \mathbf{Y}')'$ we mean that the column vector \mathbf{Z} is obtained by concatenating two column vectors \mathbf{V}, \mathbf{Y} vertically or \mathbf{V}', \mathbf{Y}' horizontally.

As it is clear from the proof of Theorem 5, we need to prove that the strategy for this theorem is the same as the strategy for Theorem 4 in the case when the kernel is the scalar product.

Proposition 1. *The predictions (31) can be represented as*

$$\gamma_T = \left(\mathbf{Y}_{T-1}; \frac{Y_2 + Y_1}{2} \right)' \sqrt{W_T} \left(aI + \sqrt{W_T} \mathbf{K}_T \sqrt{W_T} \right)^{-1} \sqrt{W_T} \mathbf{k}_T \quad (33)$$

for the scalar product kernel $k(x, y) = \langle x, y \rangle$, the unit $T \times T$ matrix I , and $a > 0$.

Proof. For the scalar product kernel we have we have $\mathbf{K}_T = X'X$ and $\sqrt{W_T} \mathbf{k}_T = \sqrt{W_T} X x_T$. By Lemma 5 we obtain

$$\left(aI + \sqrt{W_T} X X' \sqrt{W_T} \right)^{-1} \sqrt{W_T} X x_T = \sqrt{W_T} X (aI + X' W_T X)^{-1} x_T.$$

It is easy to see that

$$\left(\mathbf{Y}_{T-1}; \frac{Y_2 + Y_1}{2}\right)' W_T X = \left(\sum_{t=1}^{T-1} \frac{\beta_t}{\beta_T} y_t x_t + \left(\frac{Y_2 + Y_1}{2}\right) x_T\right)'$$

and

$$X' W_T X = \sum_{t=1}^{T-1} \frac{\beta_t}{\beta_T} x_t x_t' + x_T x_T'.$$

Thus we obtain the formula (31) from (33). \square

Acknowledgements

We are grateful to Yura Kalnishkan and Volodya Vovk for numerous illuminating discussions. This work was supported by EPSRC (grant EP/F002998/1).

A Appendix

A.1 Proofs of Technical Lemmas

Proof of Lemma 1. For $T = 1$ the inequality is trivial. Assume it for $T - 1$. Then

$$\begin{aligned} \frac{1}{\beta_T} \sum_{t=1}^T \beta_t \sqrt{\frac{\beta_t}{B_t}} &= \frac{\beta_{T-1}}{\beta_T} \left(\frac{1}{\beta_{T-1}} \sum_{t=1}^{T-1} \beta_t \sqrt{\frac{\beta_t}{B_t}} \right) + \sqrt{\frac{\beta_T}{B_T}} \\ &\leq 2 \frac{\beta_{T-1}}{\beta_T} \sqrt{\frac{B_{T-1}}{\beta_{T-1}}} + \sqrt{\frac{\beta_T}{B_T}} = 2 \sqrt{\frac{\beta_{T-1}}{\beta_T}} \sqrt{\frac{B_{T-1}}{\beta_T}} + \sqrt{\frac{\beta_T}{B_T}} \\ &\leq 2 \sqrt{\frac{B_{T-1}}{\beta_T}} + \sqrt{\frac{\beta_T}{B_{T-1} + \beta_T}} \leq 2 \sqrt{\frac{B_{T-1} + \beta_T}{\beta_T}} = 2 \sqrt{\frac{B_T}{\beta_T}}. \end{aligned}$$

The first inequality is by the induction assumption, and the second inequality holds since $\beta_{T-1} \leq \beta_T$. The last inequality is $2\sqrt{x}/\sqrt{y} + \sqrt{y}/\sqrt{x+y} \leq 2\sqrt{x+y}/\sqrt{y}$, which holds for any positive x and y . (Indeed, it is equivalent to $2\sqrt{x}\sqrt{x+y} + y \leq 2(x+y)$ and $2\sqrt{x}\sqrt{x+y} \leq x+y+x$.) \square

Proof of Lemma 3. This lemma is proven by taking the derivative of the quadratic forms in F by θ and calculating the minimum: $\min_{\theta \in \mathbb{R}^n} (\theta' A \theta + c' \theta) = -\frac{(A^{-1}c)'}{4} c$ for any $c \in \mathbb{R}^n$ (see Theorem 19.1.1 in [9]). \square

Proof of Lemma 4. After evaluating each of the integrals using Lemma 2 the ratio is represented as follows:

$$\frac{\int_{\mathbb{R}^n} e^{-Q_1(\theta)} d\theta}{\int_{\mathbb{R}^n} e^{-Q_2(\theta)} d\theta} = e^{\min_{\theta \in \mathbb{R}^n} Q_2(\theta) - \min_{\theta \in \mathbb{R}^n} Q_1(\theta)}.$$

The difference of minimums can be calculated using Lemma 3 with $b = \frac{b_2 + b_1}{2}$ and $z = \frac{b_2 - b_1}{2}$:

$$\min_{\theta \in \mathbb{R}^n} Q_2(\theta) - \min_{\theta \in \mathbb{R}^n} Q_1(\theta) = c_2 - c_1 - \frac{1}{4} (b_2 + b_1)' A^{-1} (b_2 - b_1).$$

\square

Proof of Lemma 6. Consider the product of block matrices:

$$\begin{pmatrix} I_n & B \\ 0 & I_m \end{pmatrix} \begin{pmatrix} aI_n + BC & 0 \\ -C & aI_m \end{pmatrix} = \begin{pmatrix} aI_n & aB \\ -C & aI_m \end{pmatrix} = \begin{pmatrix} aI_n & 0 \\ -C & aI_m + CB \end{pmatrix} \begin{pmatrix} I_n & B \\ 0 & I_m \end{pmatrix}$$

Taking the determinant of both sides, and using formulas for the determinant of a block matrix, we get the statement of the lemma. \square

A.2 An Alternative Derivation of Regression Algorithms Using Defensive Forecasting

In this section we derive the upper bound and the algorithms using a different technique, the defensive forecasting [4].

A.2.1 Description of the Proof Technique

We denote the predictions of any expert θ (from a finite set or following strategies from Section 4) by ξ_t^θ . For each step T and each expert θ we define the function

$$\begin{aligned} Q_t^\theta &: \Gamma \times \Omega \rightarrow [0, \infty) \\ Q_t^\theta(\gamma, y) &:= e^{\eta(\lambda(\gamma, y) - \lambda(\xi_t^\theta, y))}. \end{aligned} \tag{34}$$

We also define the mixture function

$$Q_T := \int_{\Theta} \prod_{t=1}^{T-1} (Q_t^\theta)^{\Pi_{i=t}^{T-1} \alpha_i} Q_T^\theta P_0(d\theta)$$

with some initial weights distribution $P_0(d\theta)$ on the experts. Here η is a learning rate coefficient; it will be defined later in the section. We define the correspondence

$$\gamma^p = p(Y_2 - Y_1) + Y_1, \quad p \in [0, 1], \tag{35}$$

between $[0, 1]$ and Learner's predictions $\gamma^p \in \Gamma$.

Let us introduce the notion of a defensive property. We use the notation $\delta\Omega := \{Y_1, Y_2\}$. Assume that there is a fixed bijection between the space $\mathcal{P}(\delta\Omega)$ of all probability measures on $\delta\Omega$ and the set $[0, 1]$. Each $p^\pi \in [0, 1]$ corresponds to some unique $\pi \in \mathcal{P}(\delta\Omega)$.

Definition 1. A sequence R of functions R_1, R_2, \dots such that $R_t : \Gamma \times \Omega \rightarrow (-\infty, \infty]$ is said to have *the defensive property* if, for any T and any $\pi_T \in \mathcal{P}(\delta\Omega)$, it holds that

$$\mathbb{E}_{\pi_T} R_T(\gamma^{p^{\pi_T}}, y) \leq 1, \tag{36}$$

where \mathbb{E}_π is the expectation with respect to a measure π .

A sequence R is called *forecast-continuous* if, for all T and all $y \in \Omega$, all the functions $R_T(\gamma, y)$ are continuous in γ .

We now prove that Q_t^θ has the defensive property.

Lemma 7. For $\eta \in \left(0, \frac{2}{(Y_2 - Y_1)^2}\right]$

$$Q_t^\theta = e^{\eta((\gamma_t - y_t)^2 - (\xi_t^\theta - y_t)^2)}$$

is a forecast-continuous sequence having the defensive property.

Proof. The continuity is obvious. We need to prove that

$$pe^{\eta((\gamma-Y_2)^2-(\xi_t^\theta-Y_2)^2)} + (1-p)e^{\eta((\gamma-Y_1)^2-(\xi_t^\theta-Y_1)^2)} \leq 1 \quad (37)$$

holds for all $\gamma \in [Y_1, Y_2]$ and $\eta \in \left(0, \frac{2}{(Y_2-Y_1)^2}\right]$. Indeed, for any $\gamma \in \mathbb{R} \setminus [Y_1, Y_2]$ there exists $\tilde{\gamma} \in \{Y_1, Y_2\}$ such that $(\tilde{\gamma} - y)^2 \leq (\gamma - y)^2$ for any $y \in \Omega$. Since the exponent function is increasing, the inequality (37) for any $\gamma \in \mathbb{R}$ will follow.

We use the correspondence (35), $\xi_t^\theta = q(Y_2 - Y_1) + Y_1$ for some $q \in \mathbb{R}$, and $\mu = \eta(Y_2 - Y_1)^2$. Then we have to show that for all $p \in [0, 1]$, $q \in \mathbb{R}$ and $\eta \in \left(0, \frac{2}{(Y_2-Y_1)^2}\right]$

$$pe^{\mu((1-p)^2-(1-q)^2)} + (1-p)e^{\mu(p^2-q^2)} \leq 1.$$

If we substitute $q = p + x$, the last inequality will reduce to

$$pe^{2\mu(1-p)x} + (1-p)e^{-2\mu px} \leq e^{\mu x^2}, \quad \forall x \in \mathbb{R}.$$

Applying Hoeffding's inequality (see [12]) to the random variable X that is equal to 1 with probability p and to 0 with probability $(1-p)$, we obtain

$$pe^{h(1-p)} + (1-p)e^{-hp} \leq e^{h^2/8}$$

for any $h \in \mathbb{R}$. With the substitution $h := 2\mu x$ it reduces to

$$pe^{2\mu(1-p)x} + (1-p)e^{-2\mu px} \leq e^{\mu^2 x^2/2} \leq e^{\mu x^2},$$

where the last inequality holds if $\mu \leq 2$. The last inequality is equivalent to $\eta \leq \frac{2}{(Y_2-Y_1)^2}$, which we assumed. \square

We will further use the maximum value for η , $\eta = \frac{2}{(Y_2-Y_1)^2}$.

The following lemma states the most important for us property of the sequences having the defensive property originally proven in [15].

Lemma 8. *Let R be a forecast-continuous sequence having the defensive property. For any T there exists $p \in [0, 1]$ such that for all $y \in \delta\Omega$*

$$R_T(\gamma^p, y) \leq 1.$$

Proof. Define a function $f_t : \delta\Omega \times [0, 1] \rightarrow (-\infty, \infty]$ by

$$f_t(p, y) = R_t(\gamma^p, y) - 1.$$

Since R is forecast-continuous and the correspondence (35) is continuous, $f_t(y, p)$ is continuous in p . Since R has the defensive property, we have

$$pf(p, Y_2) + (1-p)f(1-p, Y_1) \leq 0 \quad (38)$$

for all $p \in [0, 1]$. In particular, $f(0, Y_1) \leq 0$ and $f(1, Y_2) \leq 0$.

Our goal is to show that for some $p \in [0, 1]$ we have $f(p, Y_1) \leq 0$ and $f(p, Y_2) \leq 0$. If $f(0, Y_2) \leq 0$, we can take $p = 0$. If $f(1, Y_1) \leq 0$, we can take $p = 1$. Assume that $f(0, Y_2) > 0$ and $f(1, Y_1) > 0$. Then the difference

$$f(p) := f(p, Y_2) - f(p, Y_1)$$

is positive for $p = 0$ and negative for $p = 1$. By the intermediate value theorem, $f(p) = 0$ for some $p \in (0, 1)$. By (38) we have $f(p, Y_2) = f(p, Y_1) \leq 0$. \square

This lemma shows that at each step there is a probability measure (corresponding to $p \in [0, 1]$) such that the sequence having the defensive property remains less than one for any outcome.

The proof of the upper bounds for Defensive Forecasting is based on the following argument.

Lemma 9. *Assume that the sequence of functions Q_t^θ is forecast-continuous and has the defensive property. Then the mixtures Q_t as functions of two variables y, γ at the step t form a forecast-continuous sequence having the defensive property.*

Proof. The continuity easily follows from the continuity of Q_t^θ and the integration functional. We proceed by induction in T . For $T = 0$ we have $E_\pi Q_0 = E_\pi 1 \leq 1$. For $T > 0$ assume that for any $y_1, \dots, y_{T-2} \in \delta\Omega$ and any $\gamma_1, \dots, \gamma_{T-2} \in \Gamma$

$$E_\pi Q_{T-1}(y_1, \gamma_1, \dots, y_{T-2}, \gamma_{T-2}, y, \gamma^{p^\pi}) \leq 1$$

for any $\pi \in \mathcal{P}(\delta\Omega)$. Then by Lemma 8 there exists $\pi_{T-1} \in \mathcal{P}(\delta\Omega)$ such that

$$Q_{T-1}(y_1, \gamma_1, \dots, y_{T-2}, \gamma_{T-2}, y, \gamma^{p^{\pi_{T-1}}}) = \int_{\Theta} \prod_{t=1}^{T-2} (Q_t^\theta)^{\prod_{i=t}^{T-2} \alpha_i} Q_{T-1}^\theta P_0(d\theta) \leq 1 \quad (39)$$

for any $y \in \delta\Omega$. We denote $\gamma_{T-1} = \gamma^{p^{\pi_{T-1}}}$ and fix any $y_{T-1} \in \Omega$. We obtain

$$\begin{aligned} & E_\pi Q_T(y_1, \gamma_1, \dots, y_{T-1}, \gamma_{T-1}, y, \gamma^{p^\pi}) \\ &= E_\pi \int_{\Theta} \prod_{t=1}^{T-1} (Q_t^\theta(\gamma_t, y_t))^{\prod_{i=t}^{T-1} \alpha_i} Q_T^\theta(\gamma^{p^\pi}, y) P_0(d\theta) \\ &= \int_{\Theta} \prod_{t=1}^{T-1} (Q_t^\theta(\gamma_t, y_t))^{\prod_{i=t}^{T-1} \alpha_i} \left(E_\pi Q_T^\theta(\gamma^{p^\pi}, y) \right) P_0(d\theta) \\ &\leq \int_{\Theta} \prod_{t=1}^{T-1} (Q_t^\theta(\gamma_t, y_t))^{\prod_{i=t}^{T-1} \alpha_i} P_0(d\theta) \\ &= \int_{\Theta} \left(\prod_{t=1}^{T-2} (Q_t^\theta)^{\prod_{i=t}^{T-2} \alpha_i} Q_{T-1}^\theta \right)^{\alpha_{T-1}} P_0(d\theta) \\ &\leq \left(\int_{\Theta} \prod_{t=1}^{T-2} (Q_t^\theta)^{\prod_{i=t}^{T-2} \alpha_i} Q_{T-1}^\theta P_0(d\theta) \right)^{\alpha_{T-1}} \leq 1. \end{aligned}$$

The first inequality holds because $E_\pi Q_T^\theta(\gamma^{p^\pi}, y) \leq 1$ for any $\pi \in \mathcal{P}(\delta\Omega)$. The penultimate inequality holds due to the concavity of the function x^α with $x > 0$, $\alpha \in [0, 1]$. The last inequality holds due to (39). This completes the proof. \square

By Lemma 8 at each step t there exists a prediction γ_t such that Q_t is less than one. Now we only need to generalize Lemma 8 for the case when the outcome set is the full interval: $\Omega = [Y_1, Y_2]$.

Lemma 10. *If γ_T is such that $Q_T(y_1, \gamma_1, \dots, y_{T-1}, \gamma_{T-1}, y, \gamma_T) \leq 1$ for all $y \in \{Y_1, Y_2\}$, then $Q_T(y_1, \gamma_1, \dots, y_{T-1}, \gamma_{T-1}, y, \gamma_T) \leq 1$ for all $y \in [Y_1, Y_2]$.*

Proof. Note that any $y \in [Y_1, Y_2]$ can be represented as $y = uY_{T,2} + (1-u)Y_{T,1}$ for some $u \in [0, 1]$. Thus

$$\begin{aligned} (\zeta_1 - y)^2 - (\zeta_2 - y)^2 &= \zeta_1^2 - \zeta_2^2 - 2y(\zeta_1 - \zeta_2) \\ &= u[(\zeta_1 - Y_2)^2 - (\zeta_2 - Y_2)^2] + (1-u)[(\zeta_1 - Y_1)^2 - (\zeta_2 - Y_1)^2] \end{aligned}$$

for any $\zeta_1, \zeta_2 \in \mathbb{R}$. Due to the convexity of the exponent function we have for any $\eta \geq 0$

$$e^{\eta[(\zeta_1 - y)^2 - (\zeta_2 - y)^2]} \leq ue^{\eta[(\zeta_1 - Y_2)^2 - (\zeta_2 - Y_2)^2]} + (1-u)e^{\eta[(\zeta_1 - Y_1)^2 - (\zeta_2 - Y_1)^2]}.$$

Thus

$$Q_T^\theta(\gamma_T, y) \leq uQ_T^\theta(\gamma_T, Y_2) + (1-u)Q_T^\theta(\gamma_T, Y_1)$$

and therefore

$$\begin{aligned} Q_T(y_1, \gamma_1, \dots, y_{T-1}, \gamma_{T-1}, y, \gamma_T) &\leq uQ_T(y_1, \gamma_1, \dots, y_{T-1}, \gamma_{T-1}, Y_2, \gamma_T) \\ &\quad + (1-u)Q_T(y_1, \gamma_1, \dots, y_{T-1}, \gamma_{T-1}, Y_1, \gamma_T) \leq 1 \end{aligned}$$

where the second inequality follows from the condition of the lemma. \square

Finally we obtain

$$\int_{\Theta} \prod_{t=1}^{T-1} e^{\eta \sum_{i=t}^{T-1} \alpha_i (\lambda(\gamma_t, y_t) - \lambda(\xi_t^\theta, y_t))} e^{\eta (\lambda(\gamma_T, y_T) - \lambda(\xi_T^\theta, y_T))} P_0(d\theta) \leq 1. \quad (40)$$

A.2.2 Derivation of the Prediction Strategies Using Defensive Forecasting

Lemma 8 describes an explicit strategy of making predictions. This strategy relies on the search for a fixed point and may become very inefficient especially for the cases of infinite number of experts. Therefore we develop a more efficient strategies for each of our problems.

We first note that the strategy in Lemma 8 solves

$$\begin{aligned} \int_{\Theta} \prod_{t=1}^{T-1} e^{\eta \sum_{i=t}^{T-1} \alpha_i (\lambda(\gamma_t, y_t) - \lambda(\xi_t^\theta, y_t))} e^{\eta (\lambda(\gamma_T, Y_2) - \lambda(\xi_T^\theta, Y_2))} P_0(d\theta) \\ - \int_{\Theta} \prod_{t=1}^{T-1} e^{\eta \sum_{i=t}^{T-1} \alpha_i (\lambda(\gamma_t, y_t) - \lambda(\xi_t^\theta, y_t))} e^{\eta_T (\lambda(\gamma_T, Y_1) - \lambda(\xi_T^\theta, Y_1))} P_0(d\theta) = 0 \end{aligned}$$

in $\gamma \in [Y_1, Y_2]$ if the trivial predictions are not satisfactory (the integral becomes a sum in the case of finite number of experts). We define

$$g_T(y) := -\frac{1}{\eta} \ln \int_{\Theta} e^{-\eta \lambda(\xi_T^\theta, y)} \prod_{t=1}^{T-1} e^{-\eta \sum_{i=t}^{T-1} \alpha_i \lambda(\xi_t^\theta, y_t)} P_0(d\theta) \quad (41)$$

for any $y \in \Omega$. Rewriting the equation for the root we have

$$e^{\eta (\lambda_T(\gamma, Y_2) - g_T(Y_2))} - e^{\eta (\lambda_T(\gamma, Y_1) - g_T(Y_1))} = 0$$

Moving the second exponent to the right-hand side and taking \log_η of both sides we obtain

$$\lambda(\gamma, Y_2) - g_T(Y_2) = \lambda(\gamma, Y_1) - g_T(Y_1). \quad (42)$$

For the square loss we can solve (42) in γ :

$$\gamma = \frac{Y_2 + Y_1}{2} - \frac{g(Y_2) - g(Y_1)}{2(Y_2 - Y_1)}. \quad (43)$$

This formula for predictions is equivalent to (30).

References

- [1] Beckenbach, E.F., Bellman, R.: Inequalities. Springer, Berlin (1961)
- [2] Cesa-Bianchi, N., Lugosi, G.: Prediction, Learning, and Games. Cambridge University Press, Cambridge, England (2006)
- [3] Chaudhuri, K., Freund, Y., Hsu, D.: A parameter-free hedging algorithm. In: Advances in Neural Information Processing Systems 22, 297–305 (2009)
- [4] Chernov, A., Kalnishkan, Y., Zhdanov, F., Vovk, V.: Supermartingales in prediction with expert advice. Theoretical Computer Science, 411, pp. 2647–2669 (2010). See also: arXiv:1003.2218 [cs.LG]
- [5] Chernov, A., Vovk, V.: Prediction with Advice of Unknown Number of Experts Technical report, arXiv:1006.0475 [cs.LG], arXiv.org e-Print archive (2010)
- [6] Freund, Y., Hsu, D.: A new hedging algorithm and its application to inferring latent random variables. Technical report, arXiv:0806.4802v1 [cs.GT], arXiv.org e-Print archive (2008)
- [7] Gammerman, A., Kalnishkan, Y., Vovk, V.: On-line prediction with kernels and the complexity approximation principle. In: Uncertainty in Artificial Intelligence, Proc. of 20th Conf., pp. 170–176 (2004)
- [8] Gardner, E.S.: Exponential smoothing: The state of the art – part II. International Journal of Forecasting 22, 637–666 (2006)
- [9] Harville, D.A.: Matrix algebra from a statistician’s perspective. Springer, New York (1997)
- [10] Haussler, D., Kivinen, J., Warmuth, M.: Sequential prediction of individual sequences under general loss functions. IEEE Transactions on Information Theory, 44:1906–1925 (1998).
- [11] Herbster, M., Warmuth, M.K.: Tracking the best expert. Machine Learning 32, 151–178 (1998)
- [12] Hoeffding, W.: Probability inequalities for sums of bounded random variables. Journal of the American Statistical Association 58, 13–30 (1963)

- [13] Kalnishkan, Y., Vyugin, M.: The weak aggregating algorithm and weak mixability. Technical report, CLRC-TR-03-01, Computer Learning Research Centre, Royal Holloway, University of London (2003). <http://www.clrc.rhul.ac.uk/publications/files/tr0301.ps>
- [14] Kalnishkan, Y., Vyugin, M.: The weak aggregating algorithm and weak mixability. *Journal of Computer and System Sciences*, 74(8), 1228–1244 (2008)
- [15] Levin, L.: Uniform tests of randomness. *Soviet Mathematics Doklady* 17, 337–340 (1976)
- [16] Muth, J.F.: Optimal properties of exponentially weighted forecasts. *Journal of the American Statistical Association* 55, 299–306 (1960)
- [17] Schölkopf, B., Smola, A.J.: *Learning with kernels: Support Vector Machines, regularization, optimization, and beyond*. MIT Press, Cambridge, MA, USA (2002)
- [18] Sutton, R., Barto, A.: *Reinforcement learning: An introduction*. Cambridge, MA, MIT Press (1998)
- [19] Vovk, V.: Aggregating strategies. In: *Proceedings of the Third Annual Workshop on Computational Learning Theory*. pp. 371–383. Morgan Kaufmann, San Mateo, CA (1990)
- [20] Vovk, V.: A Game of Prediction with Expert Advice. *Journal of Computer and System Sciences*, 56:153–173 (1998)
- [21] Vovk V.: Derandomizing stochastic prediction strategies. *Machine Learning*, 35:247–282 (1999)
- [22] Vovk, V.: Competitive on-line statistics. *Int. Stat. Review* 69, 213–248 (2001)
- [23] Vovk, V.: On-line regression competitive with reproducing kernel Hilbert spaces. Technical report, arXiv:cs/0511058 [cs.LG], arXiv.org e-Print archive (2005)
- [24] Vovk, V.: Hoeffding’s inequality in game-theoretic probability. Technical Report, arXiv:0708.2502 [math.PR], arXiv.org e-Print archive (2007)