Iterative Ridge Regression using the Aggregating Algorithm⋆

Waqas Jamil[1,*], Abdelhamid Bouchachia

Bournemouth University, Department of Computing and Informatics, Poole, BH12 5BB, UK

ABSTRACT

In this paper, regularised regression for sequential data is investigated and new ridge regression algorithm is proposed. It uses the Aggregating Algorithm (AA) to devise an iterative version of ridge regression (IRR). This algorithm is called AAIRR. A competitive analysis is conducted to show that the guarantee on the performance of AAIRR is better than that of the known online ridge regression algorithms. Moreover, an empirical study is carried out on real-world datasets to demonstrate the superior performance over those state-of-the-art algorithms.

1. Introduction

The problem of online regularised regression aims to predict the outcome, lying on a real number line, for a given sequence of data examples. In such a setting, the algorithm receives the sequence example by example and attempts to predict the outcome for each element before seeing the ground truth (actual outcome). If there is a discrepancy between the predicted outcome and the true one, the algorithm suffers a loss. This loss adds up over the whole sequence to obtain the total loss. The exact description of a learning environment can be thought of a game defined by triple $(\Omega, \Gamma, \lambda)$ indicating a set of possible outcomes, a set of allowed predictions and a function measuring the loss respectively.

The protocol of online learning assumes that at each step t, the learner receives a data example $x_t \in \mathbb{R}^n$ which is processed by a decision pool (i.e., set of experts) $w \in \Theta$, whose prediction is denoted by $\gamma_t^w = w' x_t$. Ridge regression in this online setting was studied by (Vovk, 2001; Azoury and Warmuth, 2001) leading to the following upper bound on the cumulative square loss:

$$L_T(AAR) \leq L_T^* + aW^2 + nY^2 \ln\left(1 + \frac{TR^2}{a}\right) \tag{1}$$

where the data examples are taken from $\ell_\infty-$ball $\{x \in \mathbb{R}^n : \|x\|_\infty \leq R\}$, the decision pool $\Theta = \{w \in \mathbb{R}^n : \|w\|_1 \leq W\}$ and $y \in [-Y, Y]$

⋆This work was funded by the EU Commission through the H2020 PROTEUS project (Ref: 687691).

*Corresponding author

e-mail: wjamil@bournemouth.ac.uk (Waqas Jamil), abouchachia@bournemouth.ac.uk (Abdelhamid Bouchachia)

such that $Y \geq 0$, $a > 0$, and L_T^* is the best linear forecaster in hindsight, given by:

$$L_T^* = \inf_w \sum_{t=1}^{T} (w' x_t - y_t)^2$$

The algorithm introduced by Vovk (2001) is derived using a Bayesian strategy, while the algorithm proposed by Azoury and Warmuth (2001) exploits the duality properties of the exponential family distributions. A simplification of the analysis is presented in (Forster, 1999) using min-max optimisation. Essentially the learner's prediction can be obtained by solving the following optimisation problem:

$$\operatorname{argmin} \sup_{y \in [-Y,Y]} \left(L_T(Learner) - \inf_{w \in \mathbb{R}^n} \left(a\|w\|_2^2 + \sum_{t=1}^{T} (y_t - \gamma_t^w)^2 \right) \right) \tag{2}$$

Later Cesa-Bianchi and Lugosi (2006) obtained a similar bound as (1) by using a gradient-based forecaster with time varying elliptical potentials.

To derive the proposed learning algorithm AAIRR, we will rely on the approach described by Forster (1999). We will then follow the approach by Vovk (2001) to obtain a guarantee for AAIRR and thus providing a connection with Laplace prior which is well understood within the statistical literature on similar matter.

Most of the existing literature proves the performance guarantee for ℓ_2 norm. However, the Iterative Ridge Regression (IRR) has not been studied in the online setting. For instance, Schmidt (2005); Tibshirani (1996), and Fan and Li (2001) proposed some algorithms for offline ridge regression by considering $w^{k+1} \in \mathbb{R}^n$ where k denotes the number of passes with the condition $w_i \neq 0$ for $i = 1, 2, \cdots, n$:

$$\|w^{k+1}\|_1 \approx \sum_{i=1}^{n} \frac{\left(w_i^{k+1}\right)^2}{|w_i^k|} = \|D_{w^k}^{-\frac{1}{2}} w^{k+1}\|_2^2 \tag{3}$$

such that $D_{w^k}^{-\frac{1}{2}} = \text{diag}(1/\sqrt{|w_1^k|}, ..., 1/\sqrt{|w_n^k|})$. In (Fan and Li, 2001), it is argued that (3) is a good approximation to ℓ_1 norm due to its similarity with the Newton's method; see for example (Kelley, 2003).

In the present work, it is shown that by scaling the ridge penalty, one can obtain a better regret than (1) under certain circumstances. For the sake of comparison we bound the input, output and the decision pool. The proposed AAIRR is compared against the Aggregating Algorithm for Regression (AAR) theoretically and empirically.

In summary, the major contributions of this work are as follows:

1. Derivation of AAIRR.
2. Provision of a competitive analysis for AAIRR to show the circumstances under which it is better than the algorithm proposed in (Vovk, 2001) and (Azoury and Warmuth, 2001).
3. Carrying out an empirical study and comparing AAIRR against the state-of-the-art algorithms.

The organisation of the rest of this paper is as follows. Section 2 describes the AAIRR algorithm. Section 3 and 4 presents mathematical and empirical analysis of AAIRR before concluding in Section 5.

2. Problem formulation and Derivation of AAIRR

Given a sequence of instances and their corresponding outcomes i.e. $(x_1, y_1), \cdots, (x_t, y_t)$. Let $\gamma_t^w : \Theta \to \Gamma$ denote the prediction given by the decision strategy/expert at time t. Let $w_{t,i}$ $(i = 1, \cdots, n)$ denotes the $i-$th component of the decision vector w_t at time t and γ_t is the prediction given by the learner. Then the operational cycle of the proposed AAIRR follows Protocol 1. The

Protocol 1. Online Regression

```
FOR t = 1, 2, ...
   (1) Read input x_t ∈ R^n
   (2) Learner outputs γ_t ∈ Γ
   (3) Receive outcome y_t ∈ [−Y, Y]
   (4) Update weights w ∈ Θ
END FOR
```

overarching goal is to ensure that the loss of the learner:

$$L_t(\text{Learner}) = \sum_{s=1}^{t} (y_s - \gamma_s)^2 \tag{4}$$

is almost as good as the loss of the best expert w (optimal weight vector):

$$L_t(w) := \sum_{s=1}^{t} (y_s - \gamma_s^w)^2 \tag{5}$$

Assuming that the input is taken from the $\ell_\infty-$ ball of radius R: $\{x_t \in \mathbb{R}^n : \|x\|_\infty \leq R\}$ and the vector w is indexed by $\Theta = \{w \in \mathbb{R}^n : \|w\|_1 \leq W\}$. Let us define the following quantities:

$$b_t := \sum_{s=1}^{t} y_s x_s \in \mathbb{R}^n \tag{6}$$

$$A_t := \left(aD^{-1} + \sum_{s=1}^{t} x_s \otimes x_s \right) \in \mathbb{R}^{n \times n},\ a > 0 \tag{7}$$

and

$$D^{-1} = \text{diag}(1/C, ..., 1/C) \tag{8}$$

where $\|w\|_1 \geq C \neq 0$ and let w to be initialised in $\mathbb{R}^n$ uniformly. Let also $\triangledown f(w)$ denote the first derivative of f and $H \triangledown f(w)$ the second derivative with respect to w and H is the Hessian matrix. The aim is to compete against the iterative ridge regression algorithm (IRR), which was suggested as an approximate solution for the following problem:

$$\inf_{w \in \mathbb{R}^n} (L_t(w) + a\|w\|_1) \tag{9}$$

where $a > 0$. The problem (9) is very difficult to bound because ℓ_1 norm is not differentiable, but it is convex. Hence, one may use sub-differentiation. Unfortunately, the problem is that the sub-differentiation of ℓ_1 norm does not lead to a unique dual vector. Thus, given the training data $\mathbf{X} \in \mathbb{R}^{p \times n}$ and the corresponding target output $\mathbf{Y} \in \mathbb{R}^p$, substituting (3) into (9)[1] gives an expression similar to that of ridge regression, which is as follows (see Equation (22) in (Schmidt, 2005) and Equation (7) in (Rajaratnam et al., 2016)):

$$w^{k+1} = \left(\mathbf{X}'\mathbf{X} + aD_{w^k}^{-1}\right)^{-1} \mathbf{X}'\mathbf{Y} \tag{10}$$

where $D_{w^k}^{-\frac{1}{2}} = \text{diag}(1/\sqrt{|w_1^k|}, ..., 1/\sqrt{|w_n^k|})$. Notice that this formulation corresponds to the offline learning setting. The online setting requires solving the following optimisation problem:

$$\inf_{w \in \mathbb{R}^n} \left(L_t(w) + a\|D^{-\frac{1}{2}}w\|_2^2\right) \tag{11}$$

For the sake of comparison and interpretation, we use Cauchy-Schwartz inequality to obtain following:

$$\inf_{w \in \mathbb{R}^n} \left(L_t(w) + a\|D^{-\frac{1}{2}}w\|_2^2\right) \leq \inf_{w \in \mathbb{R}^n} \left(L_t(w) + \frac{a}{C}\|w\|_2^2\right) \tag{12}$$

This is inequality will be proven later. Like AAR, we consider the exponential discounting of the predictions:

$$P_t(dw) = e^{-\frac{1}{2Y^2}(y_t - \gamma_t^w)^2} P_{t-1}(dw) \tag{13}$$

for all measurable set $E \in \mathbb{R}^n$:

$$P_t(E) = \int_E e^{-\frac{1}{2Y^2}(y_t - \gamma_t^w)^2} P_{t-1}(dw)$$

We set the prior for all t to be:

$$P_0 = \left(\frac{a\frac{1}{2Y^2}}{2}\right)^n \exp\left(-a\frac{1}{2Y^2}\frac{W^2}{C}\right) \tag{14}$$

such that $C \leq \|w\|_1 \leq W$, $C \neq 0$ and w is initialised with the vector $\mathbf{1}$. Essentially, we replace $e^{-a\frac{1}{2Y^2}\|w\|_2^2}$ in the Gaussian prior by $e^{-a\frac{1}{2Y^2}w'D^{-1}w}$ and $\left(\frac{a\frac{1}{2Y^2}}{\pi}\right)^{\frac{1}{2}}$ by $\frac{a\frac{1}{2Y^2}}{2}$. While in AAR only the initial distribution, P_0, is set to be Gaussian prior, in AAIRR the selected distribution over the weights is inspired by the Laplace distribution.

The Laplace distribution (Tibshirani, 1996) is written as:

$$P_0 = \frac{1}{2\tau}e^{\|w\|_1|/\tau}$$

where $\tau = \frac{1}{\lambda}$ and $\lambda > 0$. In this paper, $\tau = \frac{1}{a\eta}$ ($a > 0$) and $\eta = \frac{1}{2Y^2}$. This leads to the following Lemma:

[1] For details on the derivation of the offline IRR algorithm, see Section 4.4.2 in (van Wieringen, 2018)

Lemma 1. *For a prior (14), denoted by P_0, and $0 < C \leq \|w\|_1 \leq W$, $w_0 = \mathbf{1}$ on the topology of $\Gamma \in \mathbb{R}$ and $t = 1, 2, ...$, the cumulative loss of the Modified-Aggregating-Pseudo-Algorithm (MAPA) is:*

$$L_t(MAPA) \leq \log_\beta \int_\Theta \beta^{L_t(w)} P_0(dw)$$

where $\beta = e^{-\frac{1}{2Y^2}}$

Proof. We use induction to prove the Lemma. The pseudo-prediction is defined as:

$$g_t(y) = \log_\beta \int_\Theta \beta^{(y_t - \gamma_t^w)^2} P^*_{t-1}(dw)$$

where $P^*_{t-1}(dw) = \frac{P_{t-1}(dw)}{P_{t-1}(\Theta)}$ such that $P_t(\Theta) = \int_\Theta P_t(dw)$. For $t = 1$, then $L_t(MAPA) = g_1(y)$ (assuming this holds for $t-1$). We consider $L_t(MAPA) = g_t(y) + L_{t-1}(MAPA)$, the following holds:

$$L_t(MAPA) = \log_\beta \frac{\int_\Theta \beta^{(y_t - \gamma_t^w)^2} P_{t-1}(dw)}{P_{t-1}(\Theta)} + \log_\beta \int_\Theta \beta^{L_{t-1}(w)} P_0(dw) \tag{15}$$

For $0 < C \leq \|w\|_1 \leq W$, eq. (13) can be written as:

$$\begin{aligned} P_{t-1}(dw) &= \beta^{(y_{t-1} - \gamma_{t-1}^w)^2 + ... + (y_1 - \gamma_1^w)^2} \left(\frac{a \frac{1}{2Y^2}}{2} \right)^n \exp\left(-a \frac{1}{2Y^2} \frac{W^2}{C} \right) \\ &= \beta^{L_{t-1}(w)} P_0(dw) \end{aligned} \tag{16}$$

It follows that:

$$L_t(MAPA) \leq \log_\beta \frac{\int_\Theta \beta^{(y_t - \gamma_t^w)^2 + L_{t-1}(w)} P_0(dw)}{P_{t-1}(\Theta)} + \log_\beta \int_\Theta \frac{\beta^{L_{t-1}(w)} P_{t-1}(dw)}{\beta^{L_{t-1}(w)}} \tag{17}$$

$$L_t(MAPA) \leq \log_\beta \frac{\int_\Theta \beta^{(y_t - \gamma_t^w)^2 + L_{t-1}(w)} P_0(dw)}{P_{t-1}(\Theta)} + \log_\beta P_{t-1}(\Theta) \tag{18}$$

$$= \log_\beta \frac{P_{t-1}(\Theta) \int_\Theta \beta^{(y_t - \gamma_t^w)^2 + L_{t-1}(w)} P_0(dw)}{P_{t-1}(\Theta)} = \log_\beta \int_\Theta \beta^{L_t(w)} P_0(dw) \tag{19}$$

Therefore, the statement holds $\forall t \geq 1$. □

In the previous lemma, we confirmed that the foundation on which the prediction stand is correct. Now we optimise the weights, that is, we choose the best expert (strategy) from the decision pool using the following Lemma.

Lemma 2. *For all $t \geq 0$, $f(w) := a\|D^{-\frac{1}{2}} w\|_2^2 + L_t(w)$ is minimal at a unique point w and the function $f(w)$ is given as follows:*

$$w = A_t^{-1} b_t \quad \textit{and} \quad f(w) = \sum_{s=1}^{t} y_s^2 - b_t' A_t^{-1} b_t$$

such that none of the elements of the weight vector has its absolute value at any step equal to zero. The definition of b_t, A_t, $D^{-\frac{1}{2}}$ and $L_t(w)$ is given in (6), (7), (8) and (5) respectively.

Proof. Please see Appendix A.1. □

Theorem 1. *Let the distribution on the weights of the decision pool be (14). The prediction γ_t given by AAIRR is $b'_{t-1}A_t^{-1}x_t$, where b_t and A_t are as defined in (6) and (7) respectively.*

Proof. Please see Appendix A.2. □

The following Lemma can be used to lift the condition of $C \neq 0$ in (8), to obtain line 4 in Protocol 2 for formulating the AAIRR protocol.

Lemma 3. *For all $s = 1, 2, ..., t$, $a > 0$*

$$\left(aD^{-1} + \sum_{s=1}^{t} x_s \otimes x_s\right)^{-1} = D^{\frac{1}{2}}\left(a\mathbf{I} + D^{\frac{1}{2}}\left(\sum_{s=1}^{t} x_s \otimes x_s\right)D^{\frac{1}{2}}\right)^{-1} D^{\frac{1}{2}}$$

Proof. From the properties of a diagonal matrix, it follows that:

$$\left(aD^{-1} + \sum_{s=1}^{t} x_s \otimes x_s\right)^{-1} = \left(aD^{-\frac{1}{2}}D^{-\frac{1}{2}} + \sum_{s=1}^{t} x_s \otimes x_s\right)^{-1}$$

$$= D^{\frac{1}{2}}\left(a\mathbf{I} + D^{\frac{1}{2}}\left(\sum_{s=1}^{t} x_s \otimes x_s\right)D^{\frac{1}{2}}\right)^{-1} D^{\frac{1}{2}}$$

□

Protocol 2. AAIRR

Initialise: $a > 0, A = \mathbf{0}^{n\times n}$, $b = \mathbf{0}^{n\times 1}$ and $w = \mathbf{1} \in \mathbb{R}^{n\times 1}$.
FOR $t = 1, 2, ...,$
(1) Read $x_t \in \mathbb{R}^n$
(2) $D = \text{diag}(\sqrt{\text{abs}(w)})$ (Regularisation)
(3) $A = A + x_t \otimes x_t$ (Covariance matrix)
(4) $A^{-1} = D\,(a\mathbf{I} + DAD)^{-1}\,D$ (Lemma 3)
(5) $\gamma_t = b'A^{-1}x_t$ (Corollary 1)
(6) Read $y_t \in \mathbb{R}$
(7) $b = b + y_t x_t$ (convention)
(8) $w = A^{-1}b$ (Lemma 2)
END FOR

---

3. Analysis

The following corollary presents the limiting behaviour of AAIRR. It shows that as $\|x_t\| \to \infty$, $\gamma_t \to 0$, thus making AAIRR less likely to overestimate predictions in comparison to the usual convex optimisation methods that predict by multiplying the optimal decision strategy from the decision pool by x_t (Cesa-Bianchi and Lugosi, 2006).

Corollary 1. *For all $s = 1, 2, ..., t$, the AAIRR's prediction is is follows:*

$$\gamma_t = \frac{s_t}{1 + x_t' D^{\frac{1}{2}} \left(a\mathbf{I} + D^{\frac{1}{2}} \left(\sum_{s=1}^{t-1} x_s \otimes x_s \right) D^{\frac{1}{2}} \right)^{-1} D^{\frac{1}{2}} x_t}$$

where

$$s_t = \left(\sum_{s=1}^{t-1} y_s x_s \right)' D^{\frac{1}{2}} \left(a\mathbf{I} + D^{\frac{1}{2}} \left(\sum_{s=1}^{t-1} x_s \otimes x_s \right) D^{\frac{1}{2}} \right)^{-1} D^{\frac{1}{2}} x_t$$

Proof. Please see Appendix A.3. □

The rest of this section provides the upper bounds on the cumulative square loss for AAIRR. The main objective is to deduce the circumstances under which AAIRR has a better regret than AAR (i.e., it has better upper bound on the cumulative square loss in the online setting). To achieve this goal first the performance guarantee of AAIRR is obtained. Then, the input and weights are bounded to simplify the comparison. Finally, the regret of AAIR and AAIRR is compared.

Lemma 4. *The following upper bound on the cumulative square loss holds:*

$$L_t(AAIRR) \leq \log_\beta \int_\Theta \beta^{L_t(w)} P_0(dw)$$

Proof. The square loss function is $\eta-$mixable. For details on mixability of the loss functions see (Haussler et al., 1994; Vovk, 1990). □

Lemma 5. *For $D \in \mathbb{R}^{m \times n}$ with entries a_{ij} and $w \in \mathbb{R}^n$ with entries w_j*

$$||Dw||_2^2 \leq ||D||_F^2 ||w||_2^2$$

Proof. From Cauchy-Schwartz inequality:

$$\left(\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2 \right) \sum_{k=1}^{n} w_k$$

$$= \sum_{i=1}^{m} \left(\sum_{j=1}^{n} a_{ij}^2 \sum_{k=1}^{n} (w_k)^2 \right) \geq \sum_{i=1}^{m} \left(\sum_{j=1}^{n} a_{ij} w_j \right)^2$$

□

Remark 1. *For $n = m$ in Lemma 5*

$$\left(\sum_{i=1}^{m} \sum_{j=1}^{m} a_{ij}^2 \right) \sum_{k=1}^{m} w_k \geq \sum_{i=1}^{m} \left(\sum_{j=1}^{m} a_{ij} w_j \right)^2$$

By definition $||D||_F^2 = Tr(DD^H)$, where Tr denotes the trace of a matrix and D^H is the conjugate transpose. In other words, $||D||_F^2$ is the Sum of Squares (SS) of the absolute value of the entries of D. Also, if D is a diagonal matrix, then $||D||_F^2$ is the sum of squares of diagonal elements. This justifies the inequality (12).

Bounding $\|x_t\|_\infty \leq R$ and $\|w\|_1 \leq W$ for $s = 1, 2, ..., t$, then from Lemma 5, we have:

$$wD^{-1}w \leq \frac{\|w\|_2^2}{C} \leq \frac{\|w\|_1^2}{C} \leq \frac{W^2}{C} \tag{20}$$

We also need to upper bound the following:

$$\ln \det A_t = \ln \det \left(aD^{-1} + \sum_{s=1}^{t} x_s \otimes x_s \right)$$

To do that, we use (Beckenbach and Bellman, 1961), Theorem 7 in Chapter 2, to obtain:

$$\ln \det A_t \leq n \ln \left(aC^{-1} + tR^2 \right) = n \ln \frac{a + CtR^2}{C} \tag{21}$$

We now bound the loss of AAIRR, by using Lemma 4 and Remark 1.

Theorem 2. *For any point in time $s = 1, 2, ..., t$ and any $a > 0$ such that $\|x_t\|_\infty \leq R$ and $C \leq \|w\|_1 \leq W$, the following holds:*

$$L_t(AAIRR) \leq L_t^* + aW^2C^{-1} + nY^2 \ln \left(\frac{8Y^2(a + CtR^2)}{a^2C\pi} \right)$$

such that $C \neq 0$.

Proof. From Lemma 4, $L_t(AAIRR) \leq L_t(MAPA)$ and the rest of the proof is shown in Appendix A.4. □

Remark 2. *$L_t(IRR) = \inf_w \left(L_t(w) + a\|D^{-\frac{1}{2}}w\|_2^2 \right)$ can be written as $\sum_{t=1}^{T} y_t^2 - b'A^{-1}b$ (see Lemma 2), where A^{-1} is defined as in Lemma 3, and (21) becomes $n \ln(aW + tW^2R^2)$. Also, the upper bound on the determinant of AAIRR is $\ln \frac{16Y^4}{a^2\pi}(W(a + tWR^2))$ compare to AAR's one which is $\ln \frac{a+tR^2}{a}$. When $W \leq 1$, then $\inf_w \left(L_t(w) + a\|w\|_2^2 \right) = L_t(RR) \leq L_t(IRR)$. By setting $a \geq \frac{16Y^4}{\pi}$ one can ensure that $\ln \frac{16Y^4}{a^2\pi}(W(a + tWR^2)) \leq \ln \frac{a+tR^2}{a}$, because $\ln \frac{16Y^4(W(a+tWR^2))}{a\pi(a+tR^2)} \leq 0$. Nevertheless, this way of analysis does not provide a clean comparison of AAR and AAIRR. It however indicates that AAIRR has a better bound when $\|w\|_1 \leq 1$ and the noise term has a greater influence on the prediction accuracy than the true regression function.*

The following Theorem presents circumstances under which the regret of AAIRR is better than AAR's.

Theorem 3. *Let $\mathcal{R}_t = L_t(Learner) - L_t(w)$ (see eqs. (4) and (5)) $\|x_t\|_\infty \leq R$, $C \leq \|w\|_1 \leq W$ and n be some positive integer. Then $\forall t$, $\mathcal{R}_t(AAIR) \leq \mathcal{R}_t(AAR)$ when $C \geq 1$ and $a \geq \frac{8Y^2}{\pi}$.*

Proof. To prove this Theorem, it is sufficient to show that $\mathcal{R}_t(AAIR) - \mathcal{R}_t(AAR) \leq 0$. From (1) and Theorem 2, we have the following:

$$aW^2 \left(\frac{1}{C} - 1 \right) + nY^2 \ln \left(\frac{8Y^2(a + CtR^2)}{a^2C\pi} \right) - nY^2 \ln \left(\frac{a + tR^2}{a} \right) \leq 0$$

Thus,

$$aW^2 \left(\frac{1}{C} - 1 \right) + nY^2 \ln \frac{8Y^2(a + CtR^2)}{aC\pi(a + tR^2)} \leq 0$$

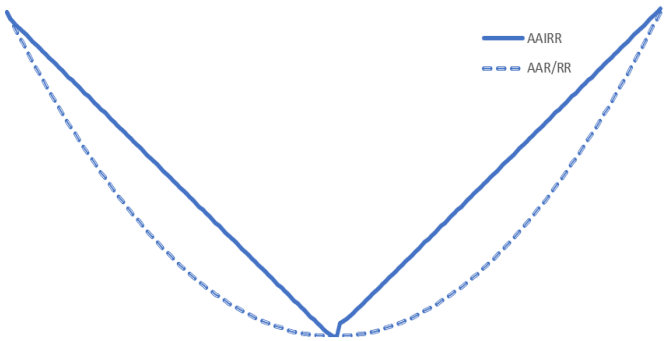

Fig. 1. AAR's penalty vs. AAIRR's penalty.

$C \geq 1$, $aW^2\left(\frac{1}{C}-1\right) \leq 0$. Also, from Lemma 5, it is clear that $\|w\|_2^2 \geq \|D^{-\frac{1}{2}}w\|_2^2$ for $C \geq 1$. The condition $a \geq \frac{8Y^2}{\pi}$ ensures that $\pi aC(a+tR^2) \geq 8Y^2(a+CtR^2)$. This concludes the proof. □

Remark 3. *Figure 1 shows that AAIRR penalty resembles ℓ_1-norm (also known as LASSO) in contrast to AAR's ℓ_2-norm (also known as ridge penalty).*

Table 1. Statistical properties of the datasets: Cook distance, mean and variance

dataset	max.cook.dist	min.cook.dist	med.cooks.dist	label mean	label variance	lr.model variance
Gaze	1.90×10^{-1}	1.35×10^{-8}	7.18×10^{-4}	5.44×10^{2}	6.31×10^{4}	3.29×10^{3}
ISE	1.37×10^{-1}	7.28×10^{-10}	4.23×10^{-4}	1.55×10^{-3}	4.46×10^{-4}	3.23×10^{-5}
NO_2	4.25×10^{-2}	3.11×10^{-8}	7.52×10^{-4}	2.18×10^{-6}	1.00×10^{0}	4.98×10^{-1}
$F-16$	5.10×10^{-2}	1.50×10^{-6}	2.30×10^{-5}	-8.68×10^{-4}	1.69×10^{-7}	3.01×10^{-8}
Weather	9.18×10^{-6}	1.61×10^{-15}	9.83×10^{-4}	1.09×10	1.14×10^{2}	1.15×10^{0}

4. Empirical study

In the following we show the empirical performance of AAIRR through a set of experiments. Specifically, we will compare it against state-of-the-art algorithms: RLS (Hayes, 1996), AROWR(Crammer et al., 2009), AAR/ORR(Vovk, 2001), ONS (Orabona et al., 2012) and the optimal offline solution. To achieve a fair comparison, five (5) datasets are considered differing from each other in terms of amount of outliers, noise, complexity (dimensionality) and volume (size). In the following, a brief description of the datasets:

- The Istanbul stock exchange (ISE) dataset (Akbilgic et al., 2014) - 536 observations with 8 attributes that are: S&P 500 Index, Deutscher Aktien Index, FTSE 100 Index, Nikkel Index, Bovespa Index, Bovespa Index, MSCI Europe Index and MSCU Emerging Markets Index. This dataset is chosen due to its simplicity. There is no noise or outlier(s).

Table 2. Performance of the algorithms on 5 real-world dataset

Algorithm	RMSE	R^2	MAE	LQE	MQE	UQE
dataset: Gaze						
AROWR	4.88×10^{14}	5.91×10^{-5}	3.21×10^{13}	-3.31×10^{12}	-1.20×10^{12}	-3.69×10^{11}
RLS	2.19×10^{17}	1.19×10^{-4}	1.35×10^{16}	-6.26×10^{14}	-6.23×10^{12}	-7.61×10^{11}
ORR	2.19×10^{17}	1.19×10^{-4}	1.35×10^{16}	-6.26×10^{14}	-1.55×10^{10}	-1.77×10^{9}
AAR	1.48×10^{5}	7.63×10^{-3}	1.26×10^{5}	-1.84×10^{4}	-1.26×10^{5}	-6.23×10^{4}
ONS	5.33×10^{3}	9.91×10^{-4}	1.06×10^{3}	-5.52×10^{2}	-5.84×10	6.69×10^{2}
AAIRR	1.61×10^{2}	6.65×10^{-1}	1.03×10^{2}	-2.04×10	4.37×10	1.13×10^{2}
Naive	3.66×10^{2}	3.44×10^{-3}	2.99×10^{2}	-2.70×10^{2}	1.95×10	2.73×10^{2}
$\mathbf{X}w^*$	$\mathbf{5.65 \times 10}$	$\mathbf{9.49 \times 10^{-1}}$	$\mathbf{4.48 \times 10}$	$\mathbf{-3.94 \times 10}$	$\mathbf{-2.25 \times 10^{0}}$	$\mathbf{3.51 \times 10^{0}}$
dataset: F−16						
AROWR	1.29×10^{11}	1.22×10^{-4}	1.15×10^{10}	-1.22×10^{8}	1.21×10^{7}	4.66×10^{8}
RLS	1.25×10^{11}	2.70×10^{-4}	1.10×10^{10}	-1.37×10^{8}	1.44×10^{7}	5.09×10^{8}
ORR	1.75×10^{7}	2.83×10^{-4}	1.60×10^{6}	-2.30×10^{4}	3.17×10^{3}	8.50×10^{4}
AAR	4.62×10^{-1}	1.64×10^{-4}	1.41×10^{-1}	-4.70×10^{-2}	8.49×10^{-4}	4.84×10^{-2}
ONS	2.30×10^{4}	1.11×10^{-2}	1.79×10^{4}	-1.23×10^{4}	1.29×10^{3}	1.72×10^{4}
AAIRR	2.08×10^{-4}	7.82×10^{-1}	1.51×10^{-4}	-7.32×10^{-5}	4.21×10^{-5}	1.39×10^{-4}
Naive	2.75×10^{-4}	6.05×10^{-1}	2.09×10^{-3}	-1.00×10^{-4}	-1.00×10^{-4}	-1.00×10^{-4}
$\mathbf{X}w^*$	$\mathbf{1.73 \times 10^{-4}}$	$\mathbf{8.24 \times 10^{-1}}$	$\mathbf{1.27 \times 10^{-4}}$	$\mathbf{-9.15 \times 10^{-5}}$	$\mathbf{3.36 \times 10^{-6}}$	$\mathbf{9.98 \times 10^{-5}}$
dataset: NO_2						
AROWR	3.11×10^{5}	1.09×10^{-1}	1.40×10^{5}	-5.02×10^{4}	-4.29×10^{3}	3.81×10^{4}
RLS	3.15×10^{5}	1.14×10^{-1}	1.46×10^{5}	-5.90×10^{4}	-5.63×10^{3}	4.27×10^{4}
ORR	8.90×10^{2}	1.59×10^{-1}	4.78×10^{2}	-2.38×10^{2}	-2.51×10	1.69×10^{2}
AAR	4.35×10	1.95×10^{-1}	3.24×10	-3.16×10	5.71×10^{0}	1.37×10
ONS	8.25×10^{-1}	4.04×10^{-1}	6.23×10^{-1}	-4.78×10^{-1}	2.07×10^{-2}	5.11×10^{-1}
AAIRR	7.31×10^{-1}	4.69×10^{-1}	5.72×10^{-1}	-3.56×10^{-1}	1.48×10^{-1}	5.58×10^{-1}
Naive	1.09×10^{0}	1.58×10^{-1}	8.19×10^{-1}	-6.04×10^{-1}	-2.74×10^{-2}	5.99×10^{-1}
$\mathbf{X}w^*$	$\mathbf{7.01 \times 10^{-1}}$	$\mathbf{5.07 \times 10^{-1}}$	$\mathbf{5.47 \times 10^{-1}}$	$\mathbf{-4.13 \times 10^{-1}}$	$\mathbf{3.65 \times 10^{-2}}$	$\mathbf{4.62 \times 10^{-1}}$
dataset: ISE						
AROWR	1.80×10^{-2}	3.00×10^{-1}	1.30×10^{-2}	-8.62×10^{-3}	9.20×10^{-4}	1.01×10^{-2}
RLS	1.01×10^{-1}	5.94×10^{-1}	7.17×10^{-2}	-5.72×10^{-2}	-1.42×10^{-2}	1.28×10^{-2}
ORR	2.79×10^{-2}	4.85×10^{-1}	1.98×10^{-2}	-1.58×10^{-2}	-4.04×10^{-4}	1.23×10^{-2}
AAR	2.00×10^{-2}	3.77×10^{-1}	1.48×10^{-2}	-1.19×10^{-3}	2.04×10^{-3}	1.22×10^{-2}
ONS	2.08×10^{-2}	5.50×10^{-1}	1.56×10^{-2}	-9.54×10^{-3}	2.57×10^{-3}	1.34×10^{-2}
AAIRR	7.61×10^{-3}	8.77×10^{-1}	5.07×10^{-3}	-4.25×10^{-3}	-1.47×10^{-4}	3.21×10^{-3}
Naive	2.87×10^{-2}	5.22×10^{-3}	2.14×10^{-2}	-1.77×10^{-2}	-1.38×10^{-3}	1.61×10^{-2}
$\mathbf{X}w^*$	$\mathbf{5.64 \times 10^{-3}}$	$\mathbf{9.29 \times 10^{-1}}$	$\mathbf{4.30 \times 10^{-3}}$	$\mathbf{-3.351 \times 10^{-3}}$	$\mathbf{3.02 \times 10^{-4}}$	$\mathbf{3.24 \times 10^{-3}}$
dataset: Weather						
AROWR	–	–	–	–	–	–
RLS	–	–	–	–	–	–
ORR	5.38×10^{15}	1.55×10^{-5}	1.34×10^{14}	-9.16×10^{10}	-3.60×10^{8}	4.46×10^{9}
AAR	3.90×10^{7}	3.16×10^{-4}	1.53×10^{6}	-7.72×10^{5}	5.06×10^{5}	-2.56×10^{5}
ONS	5.73×10^{5}	5.17×10^{-1}	5.51×10^{5}	-6.63×10^{5}	-5.58×10^{5}	4.50×10^{5}
AAIRR	1.09×10^{0}	9.89×10^{-1}	8.49×10^{-1}	-7.33×10^{-1}	-1.13×10^{-1}	6.56×10^{-1}
Naive	1.81×10^{0}	9.71×10^{-1}	1.21×10^{-1}	-9.00×10^{-1}	-2.22×10^{-2}	9.22×10^{-1}
$\mathbf{X}w^*$	$\mathbf{1.07 \times 10^{0}}$	$\mathbf{9.89 \times 10^{-1}}$	$\mathbf{8.43 \times 10^{-1}}$	$\mathbf{-7.29 \times 10^{-1}}$	$\mathbf{-1.05 \times 10^{-1}}$	$\mathbf{6.61 \times 10^{-1}}$

- Gaze dataset (Quinonero-Candela et al., 2006) consists of 450 observations of 12 features related to measurements obtained from head-mounted cameras for eye tracking, estimating the positions of the eyes of the subject when the subject is looking at the monitor. This dataset is chosen due to the presence of outlier(s).

- The NO_2 dataset (Vlachos and Meyer, 2005) consists of 500 observations from a road air pollution study collected by the Norwegian Public Roads Administration, measured at Alnabru in Oslo, Norway, between October 2001 and August 2003. There are 7 predictor variables: the logarithm of the number of cars per hour, temperature (×2), wind speed and direction,

hour of the day and the date when the observations were taken. This dataset is chosen because it shows non-linearity.

- Ailerons ($F-16$) dataset (Van Rijn et al., 2013) consists of 13750 observations with a total of 40 attributes that describe the status of the $F-16$. This dataset is chosen due to its complexity; it has the highest number of features and illustrates algorithms shrinkage ability.

- Weather dataset (Budincsevity, 2016) has historical weather around Szeged, Hungary, from 2006 to 2016 with 9 features namely: temperature, apparent temperature, humidity, wind speed, wind bearing, visibility, cloud cover, precipitation type and summary. In total there are 96453 observations. This dataset is chosen due to its considerable size; it has the highest number of observations among all datasets.

Table 1 shows the statistical properties of the datasets.

.

To run the experiments, we observed the following:

- For all algorithms setting tuning parameter or the learning rate as $\frac{1}{T}$, where T denotes the length/size of the dataset. Clearly, it is assumed that the length of the dataset is known in advance.

- The naive baseline (using y_{t-1} as prediction for y_t) is also reported.

- We consider a solution optimal after exhausting the whole dataset, that is: $\mathbf{X}w^*$, where $\mathbf{X} \in \mathbb{R}^{T\times n}$ since it has direct link to the theoretical results (see Lemma 2). The bounds given are compared against $L_T^* = \inf_w \|\mathbf{Y}-\mathbf{X}w\|_2^2$, which is the optimal loss considered and $w^* = argmin_w \|\mathbf{Y}-\mathbf{X}w\|_2^2$. This means the baseline uses the optimal weights, where the optimal loss is achieved.

Table 2 reports the root mean square error (RMSE), coefficient of determination (R^2), mean absolute error (MAE) and error quantiles: lower quantile error (LQE (25%)), mean quantile error (MQE (50%)) and upper quantile error (UQE (75%)). The main outcomes of the comparison are:

- AAIRR is overall the best algorithm in terms of $RMSE$, R^2 and MAE among the algorithms (AROWR and RLS fail to give a sensible result on the weather dataset).

- None of the algorithms is able to outperform Xw^* on any of the datasets. However, on the weather dataset, AAIRR is very close to the optimal solution in terms of RMSE and MAE. AAIRR achieves the optimal solution in terms of R^2 outperforming the naive baseline on all datasets.

5. Conclusion

In this paper, we proposed a new algorithm, AAIRR, and showed its performance guarantees. The theoretical analysis indicates that AAIRR has a better guarantee than AAR by setting $C > 1$ - see Theorem 3. The empirical study on number of real-world datasets shows the superiority of AAIRR.

In the future, the presented analysis will be extended to the stochastic setting and to study the algorithm using different loss functions (i.e., the logarithmic loss). Also, it is worth noting that tightness of AAR and AAIR bound is still an open problem.

References

Akbilgic, O., Bozdogan, H., Balaban, M.E., 2014. A novel hybrid rbf neural networks model as a forecaster. Statistics and Computing 24, 365–375.

Azoury, K.S., Warmuth, M.K., 2001. Relative loss bounds for on-line density estimation with the exponential family of distributions. Machine Learning. 43.

Beckenbach, E.F., Bellman, R.E., 1961. Inequalities. Springer.

Budincsevity, N., 2016. Weather in szeged 2006-2016. `https://www.kaggle.com/budincsevity/szeged-weather`.

Cesa-Bianchi, N., Lugosi, G., 2006. Prediction, learning, and games. Cambridge university press.

Crammer, K., Kulesza, A., Dredze, M., 2009. Adaptive regularization of weight vectors. Advances in neural information processing systems 22.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association 96, 1348–1360.

Forster, J., 1999. On relative loss bounds in generalized linear regression, in: Fundamentals of Computation Theory, Springer. pp. 829–829.

Haussler, D., Littlestone, N., Warmuth, M.K., 1994. Predicting {0, 1}-functions on randomly drawn points. Information and Computation 115, 248–292.

Hayes, M., 1996. 9.4: Recursive least squares. Statistical Digital Signal Processing and Modeling , 541.

Kelley, C.T., 2003. Solving nonlinear equations with Newton's method. SIAM.

Orabona, F., Cesa-Bianchi, N., Gentile, C., 2012. Beyond logarithmic bounds in online learning, in: Artificial intelligence and statistics, PMLR. pp. 823–831.

Quinonero-Candela, J., Dagan, I., Magnini, B., d'Alché Buc, F., 2006. Machine Learning Challenges: Evaluating Predictive Uncertainty, Visual Object Classification, and Recognizing Textual Entailment, First Pascal Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers. volume 3944. Springer.

Rajaratnam, B., Roberts, S., Sparks, D., Dalal, O., 2016. Lasso regression: estimation and shrinkage via the limit of gibbs sampling. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78, 153–174.

Schmidt, M., 2005. Least squares optimization with ℓ_1-norm regularization.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) , 267–288.

Van Rijn, J.N., Bischl, B., Torgo, L., Gao, B., Umaashankar, V., Fischer, S., Winter, P., Wiswedel, B., Berthold, M.R., Vanschoren, J., 2013. Openml: A collaborative science platform, in: Joint european conference on machine learning and knowledge discovery in databases, Springer. pp. 645–649.

Vlachos, P., Meyer, M., 2005. Statlib datasets archive. URL http://lib. stat. cmu. edu/datasets .

Vovk, V., 1990. Aggregating strategies, in: Proc. Third Workshop on Computational Learning Theory, Morgan Kaufmann. pp. 371–383.

Vovk, V., 2001. Competitive on-line statistics. International Statistical Review/Revue Internationale de Statistique , 213–248.

van Wieringen, W.N., 2018. Lecture notes on ridge regression. arXiv preprint arXiv:1509.09169 .

Appendix A. Proofs

Appendix A.1. Proof of Lemma 2

By definition, we have:

$$f(w) = a\|D^{-\frac{1}{2}}w\|_2^2 + \sum_{s=1}^{t}(y_s - w'x_s)^2$$

$$= aw'D^{-1}w + \sum_{s=1}^{t}(y_s^2 - 2y_s w'x_s + w'(x_s \otimes x_s)w)$$

$$= \sum_{s=1}^{t} y_s^2 - 2w' \sum_{s=1}^{t} y_s x_s + w'\left(aD^{-1} + \sum_{s=1}^{t} x_s \otimes x_s\right)w$$

$$= \sum_{s=1}^{t} y_s^2 - 2w'b_t + w'A_t$$

$$= \sum_{s=1}^{t} y_s^2 - \left(\sum_{s=1}^{t} 2y_s w'x_s\right) + w'\left(aD^{-1} + \sum_{s=1}^{t} x_s \otimes x_s\right)w$$

Differentiating $f(w)$ with respect to w (treating w_{t-1} as a constant), we obtain:

$$\nabla f(w) = 2\sum_{s=1}^{t} y_s x_s + 2w'\left(aD^{-1} + \sum_{s=1}^{t} x_s \otimes x_s\right)$$

$$\implies H\nabla f(w) = 2aD^{-1} + 2\sum_{s=1}^{t} x_s \otimes x_s$$

Having $\nabla f(w) = 0 - 2b_t + 2A_t w_t$ and $H\nabla f(w) = 2A_t$ indicates that f is convex and to attain its minimum, we set $\nabla f(w) = 0$ which gives $w = b_t' A_t^{-1}$. Thus,

$$f(w) = f(b_t' A_t^{-1}) = \sum_{s=1}^{t} y_s^2 - 2b_t' A_t^{-1} b_t + b_t' A_t^{-1} A_t A_t^{-1} b_t$$

$$= \sum_{s=1}^{t} y_s^2 - b_t' A_t^{-1} b_t$$

Appendix A.2. Proof of Theorem 1

In relation to Protocol 1, we use Lemma 2 to write:

$$\arg\inf_{\gamma_t \in \mathbb{R}} \sup_{y_t \in [-Y,Y]} \left(\sum_{s=1}^{t} (y_s - \gamma_s)^2 - \sum_{s=1}^{t} y_s^2 + b_t' A_t^{-1} b_t\right)$$

$$= \arg\inf_{\gamma_t \in \mathbb{R}} \sup_{y_t \in [-Y,Y]} \left(\sum_{s=1}^{t} (y_s - \gamma_s)^2 - \sum_{s=1}^{t} y_s^2 + b_{t-1} A_t^{-1} b_{t-1} + 2y_t b_{t-1}' A_t^{-1} x_t + y_t^2 x_t' A_t^{-1} x_t\right) \tag{A.1}$$

$$\implies \arg\inf_{\gamma_t \in \mathbb{R}} \sup_{y_t \in [-Y,Y]} \left(-2y_t\gamma_t + \gamma_t^2 + 2y_t b_{t-1}' A_t^{-1} x_t + y_t^2 x_t' A_t^{-1} x_t\right)$$

$$= \arg\inf_{\gamma_t \in \mathbb{R}} \sup_{y_t \in [-Y,Y]} \left(2y_t(b_{t-1}' A_t^{-1} x_t - \gamma_t) + y_t^2 (x_t' A_t^{-1} x_t) + \gamma_t^2\right) \tag{A.2}$$

Given that $y_t \in [-Y, Y]$ and A_t is positive definite, γ_t should be chosen in a way that:

$$2Y\left(b_{t-1} A_t^{-1} x_t - \gamma_t\right) + \gamma_t^2 \tag{A.3}$$

(A.3) is minimised according to the following cases:

- **Case 1**: $b_{t-1} A_t^{-1} x_t \in [-Y, Y]$. If $b_{t-1} A_t^{-1} x_t \geq Y$, then (A.3) decreases when $\gamma_t \leq Y$ and increases when $\gamma_t \geq Y$. Similar argument holds for the case when $b_{t-1} A_t^{-1} x_t \geq -Y$. Thus, (A.3) is attained at Y.
- **Case 2**: $\gamma_t \leq b_{t-1} A_t^{-1} x_t$ attains its minimum on the domain $\min(Y, b_{t-1} A_t^{-1} x_t)$.
- **Case 3**: $\gamma_t \geq b_{t-1} A_t^{-1} x_t$ attains the minimum on the domain $\max(-Y, b_{t-1} A_t^{-1} x_t)$.

Therefore, (A.1) attains the minimum for $\gamma_t = b_{t-1} A_t^{-1} x_t$.

Appendix A.3. Proof of Corollary 1

The learner's prediction is:

$$\gamma_t = \left(\sum_{s=1}^{t-1} y_s x_s\right)' D^{\frac{1}{2}} \left(a\mathbf{I} + D^{\frac{1}{2}} \left(\sum_{s=1}^{t} x_s \otimes x_s\right) D^{\frac{1}{2}}\right)^{-1} D^{\frac{1}{2}} x_t$$

$$= \left(\sum_{s=1}^{t-1} y_s x_s\right)' D^{\frac{1}{2}} \left(a\mathbf{I} + D^{\frac{1}{2}} \left(\sum_{s=1}^{t-1} x_s \otimes x_s\right) D^{\frac{1}{2}}\right)^{-1} D^{\frac{1}{2}} x_t$$

$$- \left(\sum_{s=1}^{t-1} y_s x_s\right)' \frac{\left(D^{\frac{1}{2}} \left(a\mathbf{I} + D^{\frac{1}{2}} \left(\sum_{s=1}^{t-1} x_s \otimes x_s\right) D^{\frac{1}{2}}\right)^{-1} D^{\frac{1}{2}} x_t\right)}{1 + x_t' D^{\frac{1}{2}} \left(a\mathbf{I} + D^{\frac{1}{2}} \left(\sum_{s=1}^{t-1} x_s \otimes x_s\right) D^{\frac{1}{2}}\right)^{-1} D^{\frac{1}{2}} x_t}$$

$$\times \frac{\left(D^{\frac{1}{2}} \left(a\mathbf{I} + D^{\frac{1}{2}} \left(\sum_{s=1}^{t-1} x_s \otimes x_s\right) D^{\frac{1}{2}}\right)^{-1} D^{\frac{1}{2}} x_t\right)'}{1 + x_t' D^{\frac{1}{2}} \left(a\mathbf{I} + D^{\frac{1}{2}} \left(\sum_{s=1}^{t-1} x_s \otimes x_s\right) D^{\frac{1}{2}}\right)^{-1} D^{\frac{1}{2}} x_t} x_t$$

After some algebraic manipulation, we obtain:

$$\gamma_t = \frac{s_t}{1 + x_t' D^{\frac{1}{2}} \left(a\mathbf{I} + D^{\frac{1}{2}} \left(\sum_{s=1}^{t-1} x_s \otimes x_s\right) D^{\frac{1}{2}}\right)^{-1} D^{\frac{1}{2}} x_t} \tag{A.4}$$

Appendix A.4. Proof of Theorem 2

The bound on MAPA's loss is given as follows:

$$L_t(MAPA) \leq \log_\beta \int_{\mathbb{R}^n} dw \left(\frac{a^{\frac{1}{2Y^2}}}{2}\right)^n \times \exp\left(-\frac{1}{2Y^2} w' \left(\sum_{s=1}^{t} x_s \otimes x_s + aD^{-1}\right) w + 2\frac{1}{2Y^2}\left(\sum_{s=1}^{t} y_s x_s\right) w - \frac{1}{2Y^2} \sum_{s=1}^{t} y_t^2\right) \tag{A.5}$$

Let $Q(w) = w' A_t w + b_{t-1} w + x_t' w$, where A_t is symmetric positive definite matrix and $x_t, w, b_t \in \mathbb{R}^n$. Using Theorem 3 in (Beckenbach and Bellman, 1961), we have:

$$\int_{\mathbb{R}^n} e^{Q(w)} dw = e^{-Q_0} \frac{\pi^{n/2}}{\sqrt{\det A_t}} \tag{A.6}$$

where $Q_0 = \inf_w Q(w)$. Using (A.5) and (A.6), we obtain:

$$L_t(MAPA) \leq \inf_w \left(L_t(w) + a\|D^{-\frac{1}{2}} w\|_2^2\right) + \log_\beta \left(\left(\frac{a^{\frac{1}{2Y^2}}}{2}\right)^n \frac{\pi^{n/2}}{\sqrt{\det \frac{1}{2Y^2} A_t}}\right)$$

$$= \inf_w \left(L_t(w) + a||D^{-\frac{1}{2}}w||_2^2\right) + \log_\beta \left(\left(\frac{a\frac{1}{2Y^2}}{2}\right)^{\frac{2n}{2}} \frac{\pi^{n/2}}{\sqrt{\det \frac{1}{2Y^2}A_t}}\right)$$

$$= \inf_w \left(L_t(w) + a||D^{-\frac{1}{2}}w||_2^2\right) - \frac{1}{2}\log_\beta \left(\left(\frac{2}{a\frac{1}{2Y^2}}\right)^{2n} \frac{\det \frac{1}{2Y^2}A_t}{\pi^n}\right)$$

$$= \inf_w \left(L_t(w) + a||D^{-\frac{1}{2}}w||_2^2\right) - \frac{1}{2}\log_\beta \left(\left(\frac{4}{a^2\frac{1}{2Y^2}^2\pi}\right)^{n} \det \frac{1}{2Y^2}A_t\right)$$

$$= \inf_w \left(L_t(w) + a||D^{-\frac{1}{2}}w||_2^2\right) - \frac{1}{2}\frac{\ln\left(\left(\frac{4}{a^2\frac{1}{2Y^2}^2\pi}\right)^{n} \det \frac{1}{2Y^2}A_t\right)}{\ln\beta}$$

$$= \inf_w \left(L_t(w) + a||D^{-\frac{1}{2}}w||_2^2\right) - \frac{1}{2}\frac{\ln\left(\left(\frac{16Y^4}{a^2\pi}\right)^{n} \det \frac{A_t}{2Y^2}\right)}{-\frac{1}{2Y^2}}$$

$$= \inf_w \left(L_t(w) + a||D^{-\frac{1}{2}}w||_2^2\right) + Y^2\ln\left(\left(\frac{16Y^4}{a^2\pi}\right)^{n} \det \frac{A_T}{2Y^2}\right)$$

$$= \inf_w \left(L_t(w) + a||D^{-\frac{1}{2}}w||_2^2\right) + Y^2\left(n\ln\left(\frac{16Y^4}{a^2\pi}\right) + \ln\left(\det \frac{A_t}{2Y^2}\right)\right)$$

$$L_t(AAIRR) \leq \inf_w \left(L_t(w) + a||D^{-\frac{1}{2}}w||_2^2\right)$$
$$+Y^2\left(2n\ln\left(\frac{4Y^2}{a\sqrt{\pi}}\right) + \ln\det\frac{A_t}{2Y^2}\right)$$

Finally from (21), we obtain:

$$L_t(AAIRR) \leq \inf_w \left(L_t(w) + a||D^{-\frac{1}{2}}w||_2^2\right)$$
$$+Y^2\left(2n\ln\frac{4Y^2}{a\sqrt{\pi}} + n\ln\frac{a+CtR^2}{2Y^2C}\right)$$

$$= \inf_w \left(L_t(w) + a||D^{-\frac{1}{2}}w||_2^2\right) + Y^2\left(n\ln\frac{16Y^4}{a^2\pi} + n\ln\frac{a+CtR^2}{2Y^2C}\right)$$

$$= \inf_w \left(L_t(w) + a||D^{-\frac{1}{2}}w||_2^2\right) + Y^2\left(n\ln\left(\frac{16Y^4(a+CtR^2)}{2a^2\pi Y^2C}\right)\right)$$

$$= \inf_w \left(L_t(w) + a||D^{-\frac{1}{2}}w||_2^2\right) + Y^2\left(n\ln\left(\frac{8Y^2(a+CtR^2)}{a^2C\pi}\right)\right)$$

$$= \inf_w \left(L_t(w) + a||D^{-\frac{1}{2}}w||_2^2\right) + nY^2\ln\left(\frac{8Y^2(a+CtR^2)}{a^2C\pi}\right)$$

(20) proves the statement.