

Online Bayesian Shrinkage Regression

Waqas Jamil · Abdelhamid Bouchachia

Received: date / Accepted: date

Abstract The present work introduces an original and new online regression method that extends the Shrinkage via Limit of Gibbs sampler (SLOG) in the context of online learning. In particular, we theoretically show how the proposed Online SLOG (OSLOG) is obtained using the Bayesian framework without resorting to the Gibbs sampler or considering a hierarchical representation. Moreover, in order to define the performance guarantee of OSLOG, we derive an upper bound on the cumulative squared loss. It is the only online regression algorithm with sparsity that gives logarithmic regret. Furthermore, we do an empirical comparison with two state-of-the-art algorithms to illustrate the performance of OSLOG relying on three aspects: normality, sparsity and multi-collinearity showing an excellent achievement of trade-off between these properties.

Keywords Regression · Regularisation · Online learning · Competitive analysis

1 Introduction

Offline L_1 -regularised regression Tibshirani [1996], known as lasso, has been studied well in the past. In batch setting the goal is to find the regression model weights, w , by solving:

$$w^{\text{lasso}} = \underset{w \in \mathbb{R}^n}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{X}w\|_2^2 + \lambda \|w\|_1 \quad (1)$$

given training data \mathbf{X} , labels vector \mathbf{Y} and a hyper-parameter λ . A Bayesian solution for lasso weights estimation using Gibbs Sampler was proposed in Park and Casella [2008] and later developed further in Rajaratnam et al. [2016] resulting

A shorter version of the paper was presented at the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning 2019.

W. Jamil · A. Bouchachia
Department of Computing & Informatics
Machine Intelligence Group
Bournemouth University
E-mail: {wjamil, abouchachia}@bournemouth.ac.uk

in the deterministic Bayesian lasso or better known as SLOG. By multiplying w^{lasso} with test data one can obtain predictions in batch setting.

On the other hand, in online learning predictions are made sequentially. Online learning is useful when the application lends itself continuous learning (*concept drift*) [Sambasivan et al., 2018] or there is too much data that can't fit into memory at once. Most of the work related to online L_1 -regularised regression relies on gradient descent methods (e.g., sub-gradient, coordinate descent and other proximal algorithms) to compute the estimates of the model weights see for example [Langford et al., 2009, Gerchinovitz, 2013, Duchi and Singer, 2009, Shalev-Shwartz and Tewari, 2011].

In contrast, the proposed algorithm learns by updating covariance matrix. At each trial $T = 1, 2, \dots$, our learning algorithm receives input $x_T \in \mathbb{R}^n$, makes prediction $\gamma_T \in \mathbb{R}$ and then receives the actual output $y_T \in \mathbb{R}$. Arguably the proposed method might not retain the sparsity properties when implemented with only one pass over the data. Nevertheless, it will have some degree of sparsity, we leave this matter for latter part of the paper (please see Remark 2 and Fig 2.). The fundamental advantage of using covariance-based approach is that one can obtain logarithmic regret, which is so far not possible when using gradient and sub-gradient descent approaches to solve the least squares regression problem. In [Zinkevich, 2003], it is shown that for an arbitrary convex loss function, online gradient descent has the regret growth rate of \sqrt{T} . Moreover in general, for arbitrary convex loss function, this can't be improved. However, it is possible to obtain logarithmic regret using the online Newton step [Hazan et al., 2007]; but such approach gives no advantage in terms of time complexity over the covariance-based approach for regression [Orabona et al., 2012].

It is worth noting that SLOG assumes that the entries of the regressor matrix are drawn from a distribution that is absolutely continuous with respect to Lebesgue measure [Tibshirani et al., 2013, Rajaratnam et al., 2016]. We will make no such assumption for OSLOG.

The SLOG algorithm proposed by Rajaratnam et al. [2016] maximises the posterior distribution $w \in \mathbb{R}^n$ given the response $\mathbf{y} \in \mathbb{R}^n$ i.e., $\text{argmax}_w p(w|\mathbf{y})$. It is assumed that $\mathbf{y}|w$ follows the normal distribution and w follows the Laplace or double exponential distribution. To derive SLOG, Rajaratnam et al. [2016] tweaks the approach mentioned by Park and Casella [2008] for Bayesian lasso algorithm. Both SLOG and the Bayesian lasso consider a hierarchical model by writing the Laplace distribution as a scale mixture of the Gaussian distribution [Andrews and Mallows, 1974]. The weight updating rule of the Bayesian lasso is the joint posterior obtained through the hierarchical model. Then, it is shown that by using the Gibbs sampler on the joint posterior converges to the L_1 -regularisation regression solution. SLOG uses the same approach as the Bayesian lasso with a different tuning parameter. SLOG replaces the tuning parameter $\lambda > 0$ in (1) by $a\sqrt{\sigma^2}$ with known variance σ^2 . Consequently, as the limit $\sigma^2 \rightarrow 0$ of the Gibbs sampler, it reduces to a deterministic sequence, giving the weight updating rule of SLOG. In this work, for OSLOG same weight updating equation as SLOG is obtained but without the use of Gibbs Sampler. Also, a performance guarantee for OSLOG is given. So, the major contributions of this paper are:

1. derivation of an algorithm for OSLOG without considering any hierarchical representation.

2. formulation of an upper bound on the cumulative square loss of the OSLOG algorithm.
3. empirical comparison with state-of-the-art.

The organisation of the paper is as follows. The next section introduces the derivation of OSLOG. Section 3 analyses the performance guarantee followed by the empirical study. Section 5 concludes the paper.

2 Derivation of OSLOG

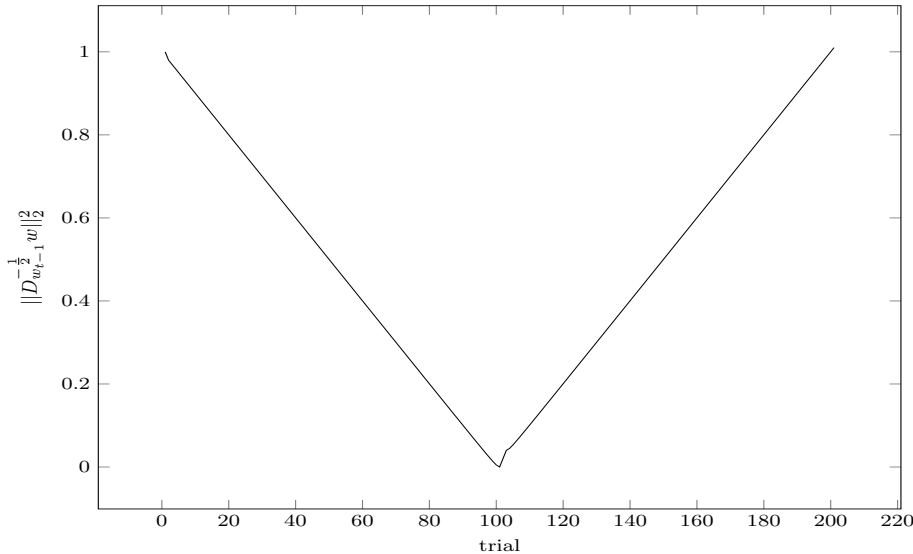


Fig. 1 L_1 -norm approximation done by OSLOG.

We consider the online protocol which assumes that at each trial the input arrives. Then, the algorithm predicts the outcome before the actual outcome is revealed and adjustment of the weights is conducted. We assume the following prior on weights:

$$p(w) = \left(\frac{a\eta}{2}\right)^n \exp\left(-a\eta w' D_{w_{t-1}}^{-1} w\right) \quad (2)$$

where $D_{w_{t-1}}$ denotes the diagonal matrix such that the diagonal vector contains the absolute value of each element of the weight vector obtained at the previous trial. The selected prior distribution on weights is inspired by the Laplace distribution which is written as Tibshirani [1996]:

$$\frac{1}{2\tau} e^{-\|w\|_1/\tau}, \quad \tau = \frac{1}{\lambda}, \quad \lambda > 0$$

In this paper, we consider: $\tau = \frac{1}{a\eta}$, where scalar $\eta = \frac{1}{2\sigma^2}$ such that $a, \eta > 0$. Also, we replace $\|w\|_1$ by $\|D_{w_{t-1}}^{-1/2} w\|_2$. Clearly in the expression $\|D_{w_{t-1}}^{-1/2} w\|_2^2$ we need

a restriction on weights. So, at trial $T - 1$ absolute value of each element of the weight vector should not be zero in (2). Despite this restriction Figure 1 shows reasonable similarity to $\|w\|_1$. A visible difference is near the kink point (100, 0). To overcome the issue of the situation where $\frac{\mathbb{R}}{0}$, we present the following Lemma:

Lemma 1. For all $t = 1, 2, \dots$

$$\left(aD_{w_{t-1}}^{-1} + \sum_{s=1}^t x_s x_s' \right)^{-1} = D_{w_{t-1}}^{\frac{1}{2}} \left(a\mathbf{I} + D_{w_{t-1}}^{\frac{1}{2}} \left(\sum_{s=1}^t x_s x_s' \right) D_{w_{t-1}}^{\frac{1}{2}} \right)^{-1} D_{w_{t-1}}^{\frac{1}{2}}$$

Proof.

$$\begin{aligned} \left(aD_{w_{t-1}}^{-1} + \sum_{s=1}^t x_s x_s' \right)^{-1} &= \left(aD_{w_{t-1}}^{-\frac{1}{2}} D_{w_{t-1}}^{-\frac{1}{2}} + \sum_{s=1}^t x_s x_s' \right)^{-1} \\ &= D_{w_{t-1}}^{\frac{1}{2}} \left(a\mathbf{I} + D_{w_{t-1}}^{\frac{1}{2}} \left(\sum_{s=1}^t x_s x_s' \right) D_{w_{t-1}}^{\frac{1}{2}} \right)^{-1} D_{w_{t-1}}^{\frac{1}{2}} \end{aligned}$$

□

Lemma 2. For any $x, b \in \mathbb{R}^n$ and a symmetric positive definite matrix A :

$$x'Ax - 2b'x = (x - A^{-1}b)'A(x - A^{-1}b) - b'A^{-1}b$$

Proof. Expanding quadratic form:

$$\begin{aligned} (x - A^{-1}b)'A(x - A^{-1}b) &= x'Ax - 2b'A^{-1}Ax + b'A^{-1}AA^{-1}b \\ &= x'Ax - 2b'x + b'A^{-1}b \end{aligned}$$

□

Remark 1. From Lemma 2 immediately follows:

$$\begin{aligned} w' \left(\sum_{t=1}^{T-1} x_t x_t' + aD_{w_{t-1}}^{-1} \right) w - 2w' \left(\sum_{t=1}^{T-1} x_t y_t \right) &= \\ \left(w - \left(\sum_{t=1}^{T-1} x_t y_t \right)' \left(\sum_{t=1}^{T-1} x_t x_t' + aD_{w_{t-1}}^{-1} \right)^{-1} \right)' \left(\sum_{t=1}^{T-1} x_t x_t' + aD_{w_{t-1}}^{-1} \right) & \\ \left(w - \left(\sum_{t=1}^{T-1} x_t y_t \right)' \left(\sum_{t=1}^{T-1} x_t x_t' + aD_{w_{t-1}}^{-1} \right)^{-1} \right) - & \\ \left(\sum_{t=1}^{T-1} x_t y_t \right)' \left(\sum_{t=1}^{T-1} x_t x_t' + aD_{w_{t-1}}^{-1} \right)^{-1} \left(\sum_{t=1}^{T-1} x_t y_t \right) & \quad (3) \end{aligned}$$

Lemma 3. If an algorithm follows Bayesian strategy with Gaussian likelihood and prior (2) such that absolute value of the each element of the weight vector is not zero, w_0 is initialised uniformly and $a > 0$, then the posterior distribution is:

$$\mathcal{N} \left(\left(\left(\sum_{t=1}^{T-1} x_t y_t \right)' \left(\sum_{t=1}^{T-1} x_t x_t' + aD_{w_{t-1}}^{-1} \right)^{-1} \right), \frac{1}{2\sigma^2} \left(\sum_{t=1}^{T-1} x_t x_t' + aD_{w_{t-1}}^{-1} \right)^{-1} \right)$$

Proof. Expanding posterior (6), by using (2) and ignoring the normalising constant we get:

$$\begin{aligned}
p(w|S_{T-1}) &\propto \exp\left(-\eta \sum_{t=1}^{T-1} (y_t - w'x_t)^2 - a\eta w' D_{w_{t-1}}^{-1} w\right) \\
&= \exp\left(-\eta \left(w' \left(\sum_{t=1}^{T-1} x_t x_t' + aD_{w_{t-1}}^{-1}\right) w - 2w' \sum_{t=1}^{T-1} x_t y_t + \sum_{t=1}^{T-1} y_t^2\right)\right) \\
&= \exp\left(-\eta \left(w - \left(\sum_{t=1}^{T-1} x_t y_t\right)'\left(\sum_{t=1}^{T-1} x_t x_t' + aD_{w_{t-1}}^{-1}\right)^{-1}\right)'\right. \\
&\quad \left.\left(\sum_{t=1}^{T-1} x_t x_t' + aD_{w_{t-1}}^{-1}\right) \left(w - \left(\sum_{t=1}^{T-1} x_t y_t\right)'\left(\sum_{t=1}^{T-1} x_t x_t' + aD_{w_{t-1}}^{-1}\right)^{-1}\right)\right. \\
&\quad \left.- \left(\sum_{t=1}^{T-1} x_t y_t\right)'\left(\sum_{t=1}^{T-1} x_t x_t' + aD_{w_{t-1}}^{-1}\right)^{-1} \left(\sum_{t=1}^{T-1} x_t y_t\right)\right. \\
&\quad \left.+ \left(\sum_{t=1}^{T-1} x_t y_t\right)'\left(\sum_{t=1}^{T-1} x_t x_t' + aD_{w_{t-1}}^{-1}\right)^{-1} \left(\sum_{t=1}^{T-1} x_t y_t\right) + \sum_{t=1}^{T-1} y_t^2\right) \\
&\propto \exp\left(-\eta \left(w - \left(\sum_{t=1}^{T-1} x_t y_t\right)'\left(\sum_{t=1}^{T-1} x_t x_t' + aD_{w_{t-1}}^{-1}\right)^{-1}\right)'\right. \\
&\quad \left.\left(\sum_{t=1}^{T-1} x_t x_t' + aD_{w_{t-1}}^{-1}\right) \left(w - \left(\sum_{t=1}^{T-1} x_t y_t\right)'\left(\sum_{t=1}^{T-1} x_t x_t' + aD_{w_{t-1}}^{-1}\right)^{-1}\right)\right) \quad (4)
\end{aligned}$$

The last and the second last equality follows from (8) and (3) respectively. The last proportionality (4) can be recognised as probability density function of the multivariate Normal distribution. \square

Theorem 1. *If an algorithm follows a Bayesian strategy with Gaussian likelihood and prior (2) such that weights at trial $T-1$ are not null, w_0 is initialised uniformly and $a > 0$, then the predictive distribution is expressed as:*

$$\mathcal{N}\left(\left(\sum_{t=1}^{T-1} x_t y_t\right)'\left(\sum_{t=1}^{T-1} x_t x_t' + aD_{w_{t-1}}^{-1}\right)^{-1} x_T, \frac{1}{2\sigma^2} x_T \left(\sum_{t=1}^{T-1} x_t x_t' + aD_{w_{t-1}}^{-1}\right)^{-1} x_T\right)$$

Proof. To obtain the predictive distribution for Normal/Gaussian likelihood with sequence S we need to solve the following:

$$p(y|x_T, S_{T-1}) = \int_{\mathbb{R}^n} p(y|x_T, w) p(w|S_{T-1}) dw \quad (5)$$

with the prior distribution (2) and the posterior is:

$$p(w|S_{T-1}) = \frac{\left(\prod_{t=1}^{T-1} p(y_t|x_t, w)\right) p(w)}{\int_{\mathbb{R}^n} \left(\prod_{t=1}^{T-1} p(y_t|x_t, w)\right) p(w) dw} \quad (6)$$

Thus, the predictive distribution at time T for y given the sequence $S_{T-1} = x_1, y_1, \dots, x_{T-1}, y_{T-1}$ requires evaluation of the following integral:

$$\frac{\int_{\mathbb{R}^n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(w'x_T - y)^2}{2\sigma^2}} \prod_{t=1}^{T-1} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(w'x_t - y_t)^2}{2\sigma^2}} \exp\left(-\frac{a}{2\sigma^2} w' D_{w_{t-1}}^{-1} w\right) dw}{\int_{\mathbb{R}^n} \prod_{t=1}^{T-1} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(w'x_t - y_t)^2}{2\sigma^2}} \exp\left(-\frac{a}{2\sigma^2} w' D_{w_{t-1}}^{-1} w\right) dw} \quad (7)$$

Let $\eta = \frac{1}{2\sigma^2}$ and,

$$L_T^w = \sum_{t=1}^T (y_t - w'x_t)^2 = \sum_{t=1}^T y_t^2 - 2w' \left(\sum_{t=1}^T x_t y_t\right) + w' \left(\sum_{t=1}^T x_t x_t'\right) w \quad (8)$$

The posterior distribution Lemma 3 can be thought of online variant of the posterior obtained by Park and Casella [2008]. Since the posterior predictive distribution is a weighted average over parameter space where each parameter is weighted by its posterior probability (see (5) and for further details see for example [Murphy, 2014]).

□

By applying Lemma 1 we lift the condition on weights and get the following explicit algorithm for OSLOG. We place the absolute value of each element of the weight vector on the diagonal of a matrix that has all off diagonal entries zero and in the algorithm we denote it as: $\text{diag}(|w_{t-1,1}|, \dots, |w_{t-1,n}|) = \text{diag}(\text{abs}(w))$.

Algorithm 1 OSLOG

Initialise: $a > 0, M = \mathbf{0}^{n \times n}, b = \mathbf{0}^{n \times 1}$ and $w = \mathbf{1} \in \mathbb{R}^{n \times 1}$

FOR $t = 1, 2, \dots$

- (1) **Read** $x_t \in \mathbb{R}^n$
- (2) $D_{w_{t-1}} = \text{diag}(\text{abs}(w))$
- (3) $\gamma = w'x_t$
- (4) $M = M + x_t x_t'$
- (5) $A^{-1} = \sqrt{D_{w_{t-1}}} (a\mathbf{I} + \sqrt{D_{w_{t-1}}} M \sqrt{D_{w_{t-1}}})^{-1} \sqrt{D_{w_{t-1}}}$
- (6) **Read** $y_t \in \mathbb{R}$
- (7) $b = b + y_t x_t$
- (8) $w = A^{-1}b$

END FOR

Remark 2. In Algorithm 1 line 8 can be allowed to make passes until convergence to have higher level of sparsity. We know from the sequential compactness theorem (see for example [Kotowicz, 1990]) that any closed and bounded sequence in Euclidean space converges. Further details can be found in [Abbott, 2001, Rudin et al., 1976, Tao, 2011]. Theorem 8 in [Rajaratnam et al., 2016] shows that SLOG converges to the lasso solution under some regularity conditions.

In Algorithm 1, the matrix A^{-1} is symmetric and positive definite, so its inverse exists at each trial. At each trial, the system of equations solved is unique without making stochastic assumptions. However, calculating the posterior predictive distribution involves measures and integrals. Therefore for measure, we assume consistency with the topological space. It is also assumed that the prediction space is a topological space equipped with σ -algebra, and the set of parameter $w \in \Theta = \mathbb{R}^n$ is equipped with σ -algebra¹.

3 Analysis of the performance guarantee

The goal is to formulate the upper bound on the cumulative squared loss. Theorem 1 implies that the prediction of Algorithm 1 corresponds to the mean of the posterior predictive parameter w weighted by the posterior probability [Murphy, 2014]. Interestingly, Kivinen and Warmuth [1999] showed that the likelihood of the weighted average can be interpreted as the loss of the Online Bayesian Strategy.

In the following, we denote the cumulative squared loss $\sum_{t=1}^T (y_t - w'x_t)^2$ by L_T^w and set A_T to be $(\sum_{t=1}^T x_t x_t' + aD_{w_{t-1}}^{-1})$.

Theorem 2. *For any trial $t = 1, 2, \dots, T$, any $a > 0$ the following holds:*

$$L_T(OSLOG) \leq \inf_w \left(L_T^w + a \|D_{w_{t-1}}^{-\frac{1}{2}} w\|_2^2 \right) + Y^2 \left(2n \ln \left(\frac{16Y^2}{a\sqrt{\pi}} \right) + \ln \det \frac{A_T}{8Y^2} \right) \quad (9)$$

where $y_t \in [-Y, Y]$ such that $Y \geq 0$ and absolute value of each element of the weight vector at $T - 1$ is not zero.

Proof. To prove the theorem considering following Lemma and the Remark:

Lemma 4. *For prior (2) at time $t = 1, 2, \dots$ the cumulative loss of OSLOG is:*

$$L_t(OSLOG) = \log_{\beta} \int_{\mathbb{R}^n} \beta^{L_t^w} p(w) dw$$

where $\beta = e^{-\eta}$.

Proof. One could proof the statement by noticing that Bayesian Strategy Q such that $\{Q_w | w \in \mathbb{R}^n\}$ with prior $p(w)$ is defined by:

$$Q = \int_{\mathbb{R}^n} Q_w p(w) dw$$

So, the main statement of the Lemma is the definition of $\log_{\beta} Q$. Hence, it holds by the definition of the Bayesian decision rule. This is a popular approach for Online Bayesian algorithms, see for example [Kakade and Ng, 2005]. \square

Remark 3. *From [Kivinen and Warmuth, 1999] we know the equality “ = ” in the above Lemma is replaced by the inequality “ \leq ” for $\eta = \frac{1}{8Y^2}$ such that L_T^w is (8) and the outcomes are bounded in $[-Y, Y]$. In other words for any value of $\eta > \frac{1}{8Y^2}$, $\beta^{(y_t - w'x_t)^2}$ will not be concave for $w'x_t$.*

¹ This is a mild assumption which is always satisfied in practice. Not making such assumption will lead to counter intuitive results such as Banach-Tarski paradox. For details see, for example, [Tao, 2011]

The problem is reduced to evaluating the integral of Lemma 4. For direct evaluation of the integral see Theorem 3 of Chapter 2 in [Beckenbach and Bellman, 2012].

$$\log_{\beta} \int_{\mathbb{R}^n} dw \left(\frac{a\eta}{2} \right)^n \times \exp \left(-\eta w' \left(\sum_{s=1}^t x_s x'_s + aD_{w_{t-1}}^{-1} \right) w + 2\eta \left(\sum_{s=1}^t y_s x_s \right) w - \eta \sum_{s=1}^t y_s^2 \right) \quad (10)$$

Remark 4. *The integral to be calculated is of the form:*

$$\int_{\mathbb{R}^n} e^{-f(w)} dw = e^{-f_0} \frac{\pi^{n/2}}{\sqrt{\det A}}$$

where $f_0 = \inf_w f(w)$. Notice,

$$f(w) = - \left(\sum_{s=1}^t 2y_s (w' x_s) \right) + w' \left(aD_{w_{t-1}}^{-1} + \sum_{s=1}^t x_s x'_s \right) w + \sum_{s=1}^t y_s^2$$

We proceed by differentiating with respect to w :

$$\nabla f(w) = - \left(\sum_{s=1}^t 2y_s x_s \right) + 2w' \left(aD_{w_{t-1}}^{-1} + \sum_{s=1}^t x_s x'_s \right)$$

clearly the second differential is negative implying the infimum is attained and by substitution the result is obtained.

From (10) and as per the Remark 3:

$$\begin{aligned} L_T(OSLOG) &= \log_{\beta} \int_{\mathbb{R}^n} dw \left(\frac{a\eta}{2} \right)^n \\ &\times \exp \left(-\eta w' \left(\sum_{t=1}^T x_t x'_t + aD_{w_{t-1}}^{-1} \right) w + 2\eta \left(\sum_{t=1}^T y_t x_t \right) w - \eta \sum_{t=1}^T y_t^2 \right) \\ &= \log_{\beta} e^{-\eta \inf \left(L_T^w + a \| D_{w_{t-1}}^{-\frac{1}{2}} w \|_2^2 \right)} \frac{\pi^{n/2}}{\det \eta \left(\sum_{t=1}^T x_t x'_t + aD_{w_{t-1}}^{-1} \right)} \\ &= \inf_w \left(L_T^w + a \| D_{w_{t-1}}^{-\frac{1}{2}} w \|_2^2 \right) + \log_{\beta} \left(\left(\frac{a\eta}{2} \right)^n \frac{\pi^{n/2}}{\sqrt{\det \eta A_T}} \right) \\ &= \inf_w \left(L_T^w + a \| D_{w_{t-1}}^{-\frac{1}{2}} w \|_2^2 \right) + \log_{\beta} \left(\left(\frac{a\eta}{2} \right)^{\frac{2n}{2}} \frac{\pi^{n/2}}{\sqrt{\det \eta A_T}} \right) \\ &= \inf_w \left(L_T^w + a \| D_{w_{t-1}}^{-\frac{1}{2}} w \|_2^2 \right) - \frac{1}{2} \log_{\beta} \left(\left(\frac{2}{a\eta} \right)^{2n} \frac{\det \eta A_T}{\pi^n} \right) \\ &= \inf_w \left(L_T^w + a \| D_{w_{t-1}}^{-\frac{1}{2}} w \|_2^2 \right) - \frac{1}{2} \log_{\beta} \left(\left(\frac{4}{a^2 \eta^2 \pi} \right)^n \det \eta A_T \right) \end{aligned}$$

$$\begin{aligned}
&= \inf_w \left(L_T^w + a \|D_{w_{t-1}}^{-\frac{1}{2}} w\|_2^2 \right) - \frac{1}{2} \frac{\ln \left(\left(\frac{4}{a^2 \eta^2 \pi} \right)^n \det \eta A_T \right)}{\ln \beta} \\
&\leq \inf_w \left(L_T^w + a \|D_{w_{t-1}}^{-\frac{1}{2}} w\|_2^2 \right) - \frac{1}{2} \frac{\ln \left(\left(\frac{16Y^4}{a^2 \pi} \right)^n \det \frac{A_T}{8Y^2} \right)}{-\frac{1}{8Y^2}} \\
&= \inf_w \left(L_T^w + a \|D_{w_{t-1}}^{-\frac{1}{2}} w\|_2^2 \right) + Y^2 \ln \left(\left(\frac{256Y^4}{a^2 \pi} \right)^n \det \frac{A_T}{8Y^2} \right) \\
&= \inf_w \left(L_T^w + a \|D_{w_{t-1}}^{-\frac{1}{2}} w\|_2^2 \right) + Y^2 n \ln \left(\frac{256Y^4}{a^2 \pi} \right) + Y^2 \ln \det \frac{A_T}{8Y^2} \\
&= \inf_w \left(L_T^w + a \|D_{w_{t-1}}^{-\frac{1}{2}} w\|_2^2 \right) + Y^2 \left(2n \ln \left(\frac{16Y^2}{a\sqrt{\pi}} \right) + \ln \det \frac{A_T}{8Y^2} \right) \quad (11)
\end{aligned}$$

□

Bounding $\|x_t\|_\infty \leq R$ and $C \leq \|w\|_1 \leq P$ for $t = 1, 2, \dots, T$ and denoting elements of diagonal matrix $D_{w_{t-1}}$ by d_{ij} . Now we upper bound the following expression:

$$\ln \det A_T = \ln \det \left(aD_{w_{t-1}}^{-1} + \sum_{t=1}^T x_t x_t' \right)$$

we use Beckenbach and Bellman [2012] Theorem 7 (in Chapter 2) to bound the determinant i.e.:

$$\begin{aligned}
\ln \det A_T &\leq \ln \prod_{i=1}^n \left(\frac{a}{d_{ii}} + \sum_{t=1}^T (x_{t,i})^2 \right) \leq \sum_{i=1}^n \ln (aC^{-1} + TR^2) \\
\ln \det A_T &\leq n \ln (aC^{-1} + TR^2) = n \ln \frac{a + CTR^2}{C} \quad (12)
\end{aligned}$$

Corollary 1. For any trial $t = 1, 2, \dots, T$ and any $a > 0$ such that $\|x_t\|_\infty \leq R$ and $C \leq \|w\|_1 \leq P$, the following holds:

$$L_T(OSLOG) \leq \inf_w \left(L_T^w + a \|D_{w_{t-1}}^{-\frac{1}{2}} w\|_2^2 \right) + nY^2 \ln \left(\frac{32Y^2(a + CTR^2)}{a^2 C \pi} \right)$$

for $y_t \in [-Y, Y]$, such that $Y \geq 0$ and $C \neq 0$.

Proof. From Theorem 2 and (12), we write:

$$\begin{aligned}
L_T(OSLOG) &\leq \inf_w \left(L_T^w + a \|D_{w_{t-1}}^{-\frac{1}{2}} w\|_2^2 \right) + Y^2 \left(2n \ln \frac{16Y^2}{a\sqrt{\pi}} + n \ln \frac{a + CTR^2}{8Y^2 C} \right) \\
&= \inf_w \left(L_T^w + a \|D_{w_{t-1}}^{-\frac{1}{2}} w\|_2^2 \right) + Y^2 \left(n \ln \frac{256Y^4}{a^2 \pi} + n \ln \frac{a + CTR^2}{8Y^2 C} \right) \\
&= \inf_w \left(L_T^w + a \|D_{w_{t-1}}^{-\frac{1}{2}} w\|_2^2 \right) + Y^2 \left(n \ln \left(\frac{256Y^4(a + CTR^2)}{8a^2 \pi Y^2 C} \right) \right) \\
&= \inf_w \left(L_T^w + a \|D_{w_{t-1}}^{-\frac{1}{2}} w\|_2^2 \right) + Y^2 \left(n \ln \left(\frac{32Y^2(a + CTR^2)}{a^2 C \pi} \right) \right)
\end{aligned}$$

We may write the above expression as follows:

$$L_T(OSLOG) \leq L_T^w + aP^2C^{-1} + nY^2 \ln \left(\frac{32Y^2(a + CTR^2)}{a^2C\pi} \right)$$

□

AAR mentioned in [Vovk, 2001] has the following guarantee:

$$L_T(AAR) \leq L_T^w + aP^2 + nY^2 \ln \left(1 + \frac{TR^2}{a} \right) \quad (13)$$

and the guarantee of OSLOG is as follows:

$$L_T(OSLOG) \leq L_T^w + aP^2C^{-1} + nY^2 \ln \left(\frac{32Y^2(a + CTR^2)}{a^2C\pi} \right)$$

The following theorem shows that under certain conditions OSLOG has a better guarantee:

Theorem 3. *If $\|x_t\|_\infty \leq R$ and $C \leq \|w\|_1 \leq P$ such that $C \geq 1$, $a \geq \frac{32Y^2}{\pi}$, and n is some positive integer, then $\forall t$, the following holds:*

$$L_T^U(OSLOG) \leq L_T^U(AAR)$$

where $L_T^U(\cdot)$ denotes the upper bound on the cumulative squared loss.

Proof. We show that $L_T^U(OSLOG) - L_T^U(AAR) \leq 0$. From (13) and Corollary 1 we write:

$$\begin{aligned} aP^2 \left(\frac{1}{C} - 1 \right) + nY^2 \ln \left(\frac{32Y^2(a + CTR^2)}{a^2C\pi} \right) - nY^2 \ln \left(\frac{a + TR^2}{a} \right) &\leq 0 \\ aP^2 \left(\frac{1}{C} - 1 \right) + nY^2 \ln \frac{32Y^2(a + CTR^2)}{aC\pi(a + TR^2)} &\leq 0 \end{aligned}$$

For $C \geq 1$, $aP^2 \left(\frac{1}{C} - 1 \right) \leq 0$. It is clear that $\|w\| \geq \|D_w^{-\frac{1}{2}}w\|$ for $C \geq 1$. The condition $a \geq \frac{32Y^2}{\pi}$ ensures that $\pi aC(a + TR^2) \geq 32Y^2(a + CTR^2)$. This concludes the proof. □

4 Empirical study

To show² the usefulness of our suggested algorithm compared to the baselines, Aggregation Algorithm for Regression (AAR) and Online Ridge Regression (ORR) [Vovk, 2001], two real-world datasets, *Gaze Data* and *Istanbul Stock Exchange Data* are used.

Gaze data [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. The dependent variable is the position of the marker

² All algorithms are available from SOLMA library: <https://github.com/proteus-h2020/proteus-solma>

Table 1 Performance comparison.

Algorithm	Mean	Variance	CSL	R^2
<i>Gaze data</i>				
AAR	504.26	46851.78	7901991	0.747
ORR	507.78	940718.40	406403726	0.042
OSLOG	544.79	41697.51	35829520	0.059
<i>Istanbul Exchange Stock data</i>				
AAR	0.002	0.0003	0.032	0.873
ORR	0.002	0.0004	0.0232	0.903
OSLOG	0.002	0.0004	0.0210	0.912

Table 2 Computational efficiency comparison in milliseconds.

Alg.	Min.	LQ.	Mean	Median	UQ.	Max.
<i>Gaze data</i>						
AAR	115.30	116.73	120.73	119.80	122.05	150.11
ORR	119.12	121.90	126.06	124.11	1126.55	203.90
OSLOG	65.21	70.58	72.62	72.78	74.03	88.21
<i>Istanbul Exchange Stock data</i>						
AAR	111.58	116.21	119.14	118.17	120.32	174.21
ORR	110.95	116.17	119.22	117.60	120.59	160.95
OSLOG	74.74	80.00	82.86	82.59	84.75	127.25

displayed on a computer monitor. We expect cameras to lose their calibration occasionally (high variance).

Istanbul stock exchange (ISE) Akbilgic et al. [2014] data³ has 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. Day and time sort all the attributes. The goal is to make the prediction of ISE in USD.

We evaluate their accuracy and efficiency. We use 20% of the data to find the best tuning parameter $a > 0$ and then we fit all the algorithms on the data in an online mode.

For the sake of analysis of the performance, we report the mean, variance, Cumulative Squared Loss (CSL) and the R^2 statistic of the predicted outcomes for each dataset. Table 1 show that in the case of gaze data, AAR outperforms all the algorithms, ORR being the worst. On Istanbul Stock Exchange data OSLOG outperforms all the algorithms, AAR being the worst.

The empirical study shows that when the statistical assumptions of normality is violated, AAR is likely to perform better than OSLOG. However, when statistical assumptions are satisfied OSLOG is likely to outperform AAR.

Figure 2 studies the effect of sparsity and multicollinearity on OSLOG. The true model: $\mathbf{y} = \mathbf{X}w + \epsilon$ is considered. To study sparsity, simulation is conducted using 1000 observations and 100 predictors. The sparsity plot is generated by varying the number of predictors in the true model from 2 to 100. The plot illustrates that as the sparsity decreases, the RMSE increases for both AAR and OSLOG. The aim of the second plot is to study multicollinearity. It shows no clear pattern, which indicates that multicollinearity sometimes helps OSLOG to estimate

³ <https://archive.ics.uci.edu/ml/datasets/ISTANBUL+STOCK+EXCHANGE>

the error term. However, this is not the case for AAR. There as multicollinearity increases, RMSE also increases. On the other hand, OSLOG handles multicollinearity and sparsity better, mainly because at each trial OSLOG weights are updated. This is not done for AAR (there is no explicit update of weights in the AAR algorithm). The simulation is done by considering correlation in predictors, i.e., $\text{Cov}(\mathbf{X})_{ij} = m^{|i-j|}$, where $m = 0.1, 0.2, \dots, 0.9$.

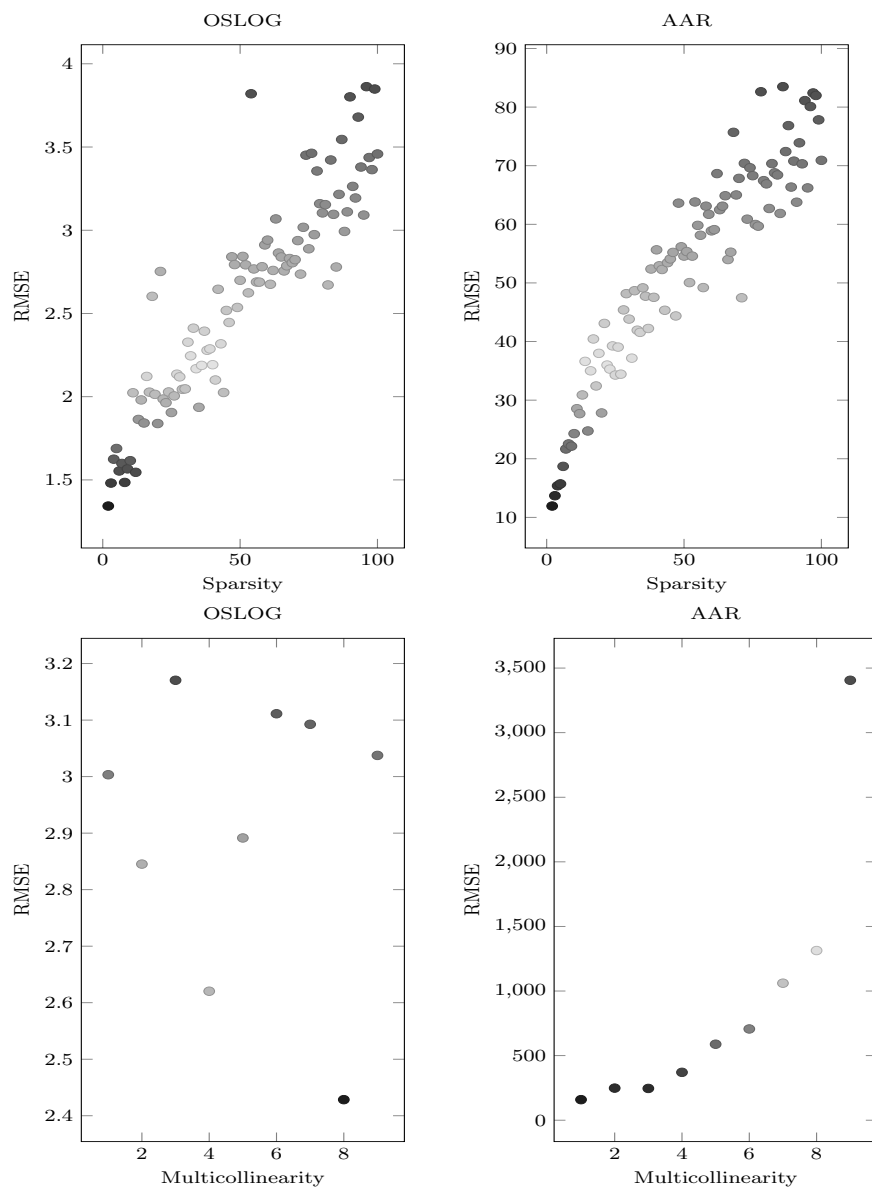


Fig. 2 Effect of sparsity and multicollinearity on AAR and OSLOG.

5 Conclusion

We proposed an online algorithm for SLOG regression and presented its performance guarantee (without making any distributional assumptions) with regret bounded by a logarithmic function of T . Our online formulation of SLOG does not require a hierarchical structure. Another fundamental difference in SLOG and OSLOG is that SLOG requires $\sigma^2 \rightarrow 0$, while OSLOG requires $\sigma^2 = 4Y^2$. In this sense, OSLOG could be considered as an online variant of the Bayesian lasso with known fixed σ^2 .

The empirical study shows that when the assumptions of multicollinearity and sparsity are violated, OSLOG is much better compared to the other algorithms. But, when the assumption of normality is violated, AAR performs a little better compared to OSLOG. Thus, if the underlying statistical properties are unknown, OSLOG is a better choice as a tradeoff between normality, multicollinearity and sparsity.

One of the interesting future research direction as a follow-up of this study to investigate the tightness of the given guarantee. Also as a natural extension, it is quite appealing to explore other loss functions besides the squared loss function.

Disclosure

Authors have no conflict of interest.

References

- Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.
- Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.
- Bala Rajaratnam, Steven Roberts, Doug Sparks, and Onkar Dalal. Lasso regression: estimation and shrinkage via the limit of gibbs sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(1):153–174, 2016.
- Rajiv Sambasivan, Sourish Das, and Sujit K Saha. A bayesian perspective of statistical machine learning for big data. *arXiv preprint arXiv:1811.04788*, 2018.
- John Langford, Lihong Li, and Tong Zhang. Sparse online learning via truncated gradient. *Journal of Machine Learning Research*, 10(Mar):777–801, 2009.
- Sébastien Gerchinovitz. Sparsity regret bounds for individual sequences in online linear regression. *Journal of Machine Learning Research*, 14(Mar):729–769, 2013.
- John Duchi and Yoram Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10(Dec):2899–2934, 2009.
- Shai Shalev-Shwartz and Ambuj Tewari. Stochastic methods for l1-regularized loss minimization. *Journal of Machine Learning Research*, 12(Jun):1865–1892, 2011.

- Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. Technical Report CMU-CS-03-110, School of Computer Science, Carnegie Mellon University, 2003.
- Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.
- Francesco Orabona, Nicolo Cesa-Bianchi, and Claudio Gentile. Beyond logarithmic bounds in online learning. In *Artificial Intelligence and Statistics*, pages 823–831, 2012.
- Ryan J Tibshirani et al. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490, 2013.
- David F Andrews and Colin L Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 99–102, 1974.
- Kevin Murphy. *Machine learning, a probabilistic perspective*. Taylor & Francis, 2014.
- Jarosław Kotowicz. Convergent real sequences. upper and lower bound of sets of real numbers. *Formalized Mathematics*, 1(3):477–481, 1990.
- Stephen Abbott. *Understanding analysis*. Springer, 2001.
- Walter Rudin et al. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1976.
- Terence Tao. *An introduction to measure theory*. American Mathematical Society Providence, RI, 2011.
- Jyrki Kivinen and Manfred Warmuth. Averaging expert predictions. In *Computational Learning Theory*, pages 638–638. Springer, 1999.
- Sham M Kakade and Andrew Y Ng. Online bounds for bayesian algorithms. In *Advances in neural information processing systems*, pages 641–648, 2005.
- Edwin F Beckenbach and Richard Bellman. *Inequalities*, volume 30. Springer Science & Business Media, 2012.
- Vladimir Vovk. Competitive on-line statistics. *International Statistical Review/Revue Internationale de Statistique*, pages 213–248, 2001.
- Joaquin Quinero-Candela, Ido Dagan, Bernardo Magnini, and Florence d’Alché Buc. *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, 2006.
- Oguz Akbilgic, Hamparsum Bozdogan, and M Erdal Balaban. A novel hybrid rbf neural networks model as a forecaster. *Statistics and Computing*, 24(3):365–375, 2014.