ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

Online Bayesian Shrinkage Regression

Waqas Jamil[1*] and Abdelhamid Bouchachia[2 *]

[1,2]Department of Computing and Informatics - Machine Intelligence Group
Bournemouth University - Poole, UK

Abstract. The present work introduces a new online regression method that extends the Shrinkage via Limit of Gibbs sampler (SLOG) in the context of online learning. In particular, we theoretically demonstrate that the proposed Online SLOG (OSLOG) is derived using the Bayesian framework without resorting to the Gibbs sampler. We also state the performance guarantee of OSLOG.

1 Introduction

Offline L_1-regularised regression [18], 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:

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

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 [11] and later developed further in [12] resulting 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*) [15] 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 [9, 5, 4, 16].

In contrast, the proposed algorithm learns by updating covariance matrix. At each trial $T = 1, 2, ...$, our learning algorithm receives input $x_T \in \mathbb{R}^n$, makes prediction $\gamma_T \in \mathbb{R}$ and than 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. 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 [20], 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

[*] This work has been supported by the European Commission under the Horizon 2020 Grant 687691 related to the project *PROTEUS: Scalable Online Machine Learning for Predictive Analytic and Real-Time Interactive Visualization.*

can't be improved. However, it is possible to obtain logarithmic regret using the online Newton step [6]; but such approach gives no advantage in terms of time complexity over the covariance-based approach for regression [10].

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 [19, 12]. We will make no such assumption for OSLOG.

The SLOG algorithm proposed by Rajaratnam et al. [12] maximises the posterior distribution $w \in \mathbb{R}^n$ given the response $\mathbf{y} \in \mathbb{R}^n$ i.e., $\operatorname{argmax}_{w \in \mathbb{R}^n} p(w|\mathbf{y})$. 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. [12] tweaks the approach mentioned by Park and Casella [11] 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 [2]. 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 $\sigma^2 \to 0$ 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 by using an iterated prior and without considering any hierarchical representation.

2. formulation of an upper bound on the cumulative square loss of the Online SLOG algorithm.

The organisation of the paper is as follows. The next section introduces the derivation of OSLOG. Section 3 analyses the performance guarantee. Section 4 concludes the paper.

2 Derivation of 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:

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

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

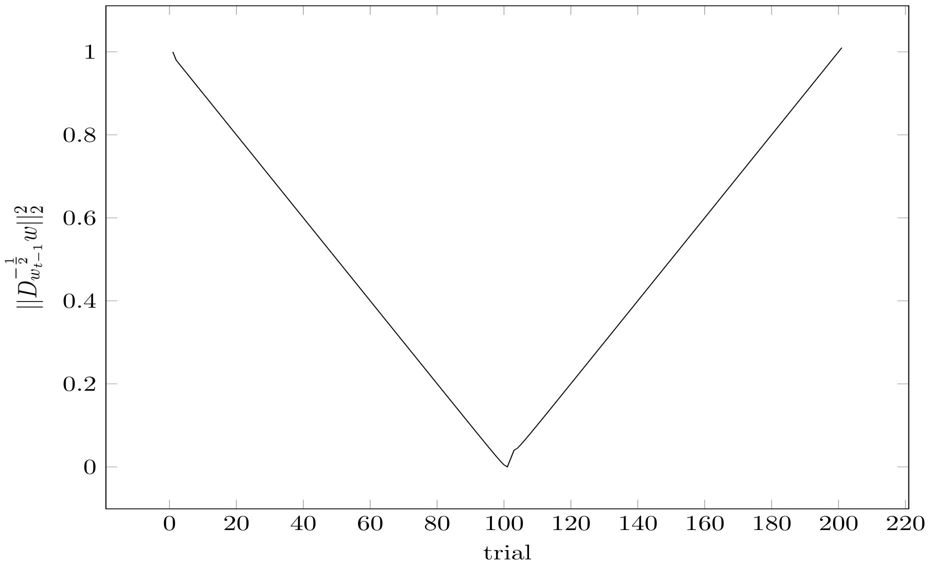

Fig. 1: L_1-norm approximation.

distribution which is written as [18]:

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

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 $||D^{-\frac{1}{2}}_{w_{t-1}}w||_2^2$. Clearly in the expression $||D^{-\frac{1}{2}}_{w_{t-1}}w||_2^2$ we need a restriction on weights. So, at trial $T-1$ absolute value of each element of the weight vector should not to 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* $t = 1, 2, ...$

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

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:*

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

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: $\text{diag}(|w_{t-1,1}|, ..., |w_{t-1,n}|) = \text{diag}(\text{abs}(w))$.

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

Remark 1. *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 [8]) that any closed and bounded sequence in Euclidean space converges. Further details can be found in [1, 14, 17]. Theorem 8 in [12] 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 $\sigma-$algebra, and the set of parameter $w \in \Theta = \mathbb{R}^n$ is equipped with $\sigma-$ algebra[1].

3 Analysis of the performance guarantee

The goal is to formulate the upper bound on the cumulative square 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 [13]. Interestingly, Kivinen and Warmuth [7] showed that the likelihood of the weighted average can be interpreted as the loss of the Online Bayesian Strategy.

[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, [17]

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

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

$$L_T(OSLOG) \leq \inf_w \left(L_T^w + a||D_{w_{t-1}}^{-\frac{1}{2}} w||_2^2\right) + Y^2\left(2n \ln\left(\frac{16Y^2}{a\sqrt{\pi}}\right) + \ln\det\frac{A_T}{8Y^2}\right) \tag{3}$$

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.

We assume that $||x_t||_\infty \leq R$ and $C \leq ||w||_1 \leq P$ for $t = 1, 2, ..., T$ and denote 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 [3] Theorem 7 (in Chapter 2) to bound the determinant i.e.:

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

$$\ln\det A_T \leq n\ln\left(aC^{-1} + TR^2\right) = n\ln\frac{a + CTR^2}{C} \tag{4}$$

Corollary 1. *For any trial $t = 1, 2, ..., T$ 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 $C \neq 0$.

4 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 $\sigma^2 \to 0$, 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.

In the future[2], we will carry the empirical evaluation of OSLOG. Also, we will explore other loss functions. A more interesting direction is to inspect the tightness of the given guarantee.

[2] Proofs will be given in the extended version of the paper.

References

[1] Stephen Abbott. *Understanding analysis.* Springer, 2001.

[2] 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.

[3] Edwin F Beckenbach and Richard Bellman. *Inequalities*, volume 30. Springer Science & Business Media, 2012.

[4] John Duchi and Yoram Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10(Dec):2899–2934, 2009.

[5] Sébastien Gerchinovitz. Sparsity regret bounds for individual sequences in online linear regression. *Journal of Machine Learning Research*, 14(Mar):729–769, 2013.

[6] Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

[7] Jyrki Kivinen and Manfred Warmuth. Averaging expert predictions. In *Computational Learning Theory*, pages 638–638. Springer, 1999.

[8] Jarosław Kotowicz. Convergent real sequences. upper and lower bound of sets of real numbers. *Formalized Mathematics*, 1(3):477–481, 1990.

[9] John Langford, Lihong Li, and Tong Zhang. Sparse online learning via truncated gradient. *Journal of Machine Learning Research*, 10(Mar):777–801, 2009.

[10] Francesco Orabona, Nicolo Cesa-Bianchi, and Claudio Gentile. Beyond logarithmic bounds in online learning. In *Artificial Intelligence and Statistics*, pages 823–831, 2012.

[11] Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[12] 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.

[13] Christian Robert. *Machine learning, a probabilistic perspective.* Taylor & Francis, 2014.

[14] Walter Rudin et al. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1976.

[15] Rajiv Sambasivan, Sourish Das, and Sujit K Saha. A bayesian perspective of statistical machine learning for big data. *arXiv preprint arXiv:1811.04788*, 2018.

[16] Shai Shalev-Shwartz and Ambuj Tewari. Stochastic methods for l1-regularized loss minimization. *Journal of Machine Learning Research*, 12(Jun):1865–1892, 2011.

[17] Terence Tao. *An introduction to measure theory.* American Mathematical Society Providence, RI, 2011.

[18] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[19] Ryan J Tibshirani et al. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490, 2013.

[20] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. Technical Report CMU-CS-03-110, School of Computer Science, Carnegie Mellon University, 2003.