On regret bounds for continual single-index learning *

The Tien Mai[1][0000−0002−3514−9636]

Oslo Centre for Biostatistics and Epidemiology,
Department of Biostatistics,
University of Oslo, Norway
t.t.mai@medisin.uio.no

Abstract. In this paper, we generalize the problem of single-index model to the context of continual learning in which a learner is challenged with a sequence of tasks one by one and the dataset of each task is revealed in an online fashion. We propose a randomized strategy that is able to learn a common single-index for all tasks and a specific link function for each task. The common single-index allows to transfer the information gained from the previous tasks to a new one. We provide a rigorous theoretical analysis of our proposed strategy by proving some regret bounds under different assumption on the loss function.

Keywords: Continual learning · Single-index model · Regret bounds · Exponentially weighted aggregation · Online learning

1 introduction

Recently, studying of learning algorithms in the setting in which the tasks are presented sequentially has received a lot of attention, see e.g. [17,7,2,10,16,9,22] among others. This setting is often referred to as contunual learning, also called as learning-to-learn or incremental learning [20,4,2]. Clearly, using information gained from previously learned tasks is useful and important for learning a new similar task. This is motivated from that human are able to learn a new task quite accurately by ultilizing knowledge from previous learned tasks.

In order to reuse the information from previous tasks, the new task must share some commonalities with previous ones. In this work, we consider that different tasks share a common feature representation space. This direction has been explored by various works, e.g. [19,18,2,22] and is natural for classification and regression problem. More precisely, different predictor for each task is built on top of a common representation in order to make predictions.

In this paper, we extend the single-index model [15] to the learning-to-learn setting. More specifically, we assume that the tasks share a common single-index in this problem. The predictor is constructed on top of this common single-index

* The research of T.T.M. is supported by the European Research Council (SCARABEE, no. 742158)

through a task-specific link functions. This grants the learner to reuse/transfer the knowlegde (the commonality) learned from previous tasks to a new task through the common single-index. Moreover, the learner still has the ability to deal with the differency between tasks through a task-specific link function.

Continual learning can be casted as a generalization of online learning and a standard way to provide theoretical guarantees for online algorithms is a regret bound. This bound measures the discrepancy between the prediction error of the forecaster and the error of an ideal predictor. We extend the EWA-LL meta-procedure in [2] to our continual single-index learning problem. Through this procedure, we provide the regret bounds for continual learning single-index. These theoretical analysis show that it is possible to learn such model in a continual context.

Interestingly, as a by-product from our work that is to provide an example of a within-task algorithm, we develop an online algorithm for learning single-index model in an online setting. More specifically, it is based on the exponentially weighted aggregation (EWA) procedure for online learning, see e.g. [6] and references therein. We also provide a regret bound for this algorithm which is also novel in the context of online single-index learning.

The paper is structured as follow. In Section 2 we introduce the continual learning context and then extend the single-index model to this context. After that, we present a meta algorithm for learning the continual single-index model based on EWA-LL procedure. The regret bound analysis is given in Section 3. A within-task online algorithm for single-index model and its regret bound is presented in Section 4. Some discussion and conclusion are given in Section 5. All technical proofs are given in Section A.

2 Continual single-index setting

2.1 Setting

We assume that, at each time step $t \in \{1, \dots, T\}$, the learner is challenged with a task sequentially, corresponding to a dataset

$$\mathcal{S}_t = \{(x_{t,1}, y_{t,1}), \dots, (x_{t,n_t}, y_{t,n_t})\} \in (\mathcal{X} \times \mathcal{Y})^{n_t}, n_t \in \mathbb{N}.$$

Furthermore, we assume that the dataset $\mathcal{S}_t$ is itself revealed sequentially, that is, at each inner step $i \in \{1, \dots, n_t\}$:

- the object $x_{t,i}$ is revealed and the learner has to predict $y_{t,i}$ by $\hat{y}_{t,i}$;
- then $y_{t,i}$ is revealed and the learner incurs the loss $\hat{\ell}_{t,i} := \ell(\hat{y}_{t,i}, y_{t,i})$.

Let $f : \mathcal{X} \to \mathcal{Y}$ be a predictor, where $\mathcal{Y} = \mathbb{R}$ for regression and $\mathcal{Y} = \{-1, 1\}$ for binary classification. Put $\hat{y}_{t,i} := f(x_{t,i})$ denote the prediction for $y_{t,i}$.

As we want to transfer the information (a common data representation) gained from the previous tasks to a new one. Formally, we let $\mathcal{Z}$ be a set and prescribe a set $\mathcal{G}$ of feature maps (also called *representations*) $g : \mathcal{X} \to \mathcal{Z}$, and a set $\mathcal{H}$ of functions $h : \mathcal{Z} \to \mathbb{R}$. We shall design an algorithm that is useful when there is a function $g \in \mathcal{G}$, common to all the tasks, and task-specific functions $h_1, \ldots, h_T$ such that $f_t = h_t \circ g$ is a good predictor for task t, in the sense that the corresponding prediction error (see below) is small.

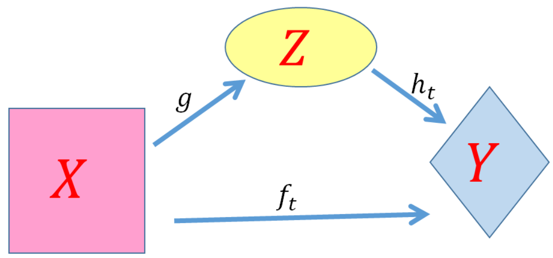

Fig. 1. The predictor f_t is built on top of a representation g and a task-specific function h_t.

In the single index model, let the set $\mathcal{X} = \mathcal{Z} = \mathbb{R}^d$, and we define $\mathcal{G} = \{x \mapsto \theta^\top x, \theta \in \mathbb{R}^d\}$ linear functions on $\mathcal{X}$. Furthermore, let $\mathcal{H}$ be a set of L_2-Lipschitz univariate measurable functions on $\mathbb{R}$. Recall that our predictor here is of the form

$$f_t(x_{t,i}) = h_t(\theta^\top x_{t,i}).$$

The goal is to learn the common single-index vector θ for all tasks and the link function h_t for each task t.

Remark 1. The predictor can be interpreted as: The predictor changes only in the direction θ (single-index), and the way it changes in this direction is defined by the link function h_t.

Remark 2. The single-index model [15] is known as a particularly useful variation of the linear model. This model has been applied to a variety of fields, such as discrete choice analysis in econometrics and dose-response models in biometrics, where high-dimensional regression models are often employed. See for example [11,14,12,13].

Noted that the task t ends at time n_t and the average prediction error at that point is $\frac{1}{n_t}\sum_{i=1}^{n_t} \hat{\ell}_{t,i}$. This process is repeated for each task t, so that at the end of all the tasks, the average error is $\frac{1}{T}\sum_{t=1}^{\top} \frac{1}{n_t}\sum_{i=1}^{n_t} \hat{\ell}_{t,i}$. Our principal objective is to design a procedure (meta-algorithm) that is able to learn the common single-index vector θ for all tasks and the link function h_t for each task t and control the *(compound) regret* of our procedure

$$\mathcal{R} := \frac{1}{T}\sum_{t=1}^{\top} \frac{1}{n_t}\sum_{i=1}^{n_t} \hat{\ell}_{t,i} - \inf_{g\in\mathcal{G}} \frac{1}{T}\sum_{t=1}^{\top} \inf_{h_t\in\mathcal{H}} \frac{1}{n_t}\sum_{i=1}^{n_t} \ell\big(h_t(\theta^\top x_{t,i}), y_{t,i}\big).$$

2.2 A randomized strategy for continual single-index learning

The EWA-LL meta-algorithm proposed in [2] based on the exponentially weighted aggregation (EWA) is a general procedure in lifelong learning. Here, we propose an application of this algorithm to the context of single-index learning. The details of our proposal algorithm is outlined in Algorithm 1.

Algorithm 1 EWA-LL for continual single-index learning

Data A sequence of datasets $\mathcal{S}_t = \big((x_{t,1}, y_{t,1}), \ldots, (x_{t,n_t}, y_{t,n_t})\big), 1 \leq t \leq T$; the points within each dataset are also given sequentially.

Input A prior π_1, a learning parameter $\eta > 0$ and a learning algorithm for each task t which, for any single-index θ returns a sequence of predictions $\hat{y}^{\theta}_{t,i}$ and suffers a loss

$$\hat{L}_t(\theta) := \frac{1}{n_t} \sum_{i=1}^{n_t} \ell\left(\hat{y}^{\theta}_{t,i}, y_{t,i}\right).$$

Loop For $t = 1, \ldots, T$

i Draw $\hat{\theta}_t \sim \pi_t$.

ii Run the within-task learning algorithm on $\mathcal{S}_t$ and suffer loss $\hat{L}_t(\hat{\theta}_t)$.

iii Update

$$\pi_{t+1}(\mathrm{d}\theta) := \frac{\exp(-\eta \hat{L}_t(\theta))\pi_t(\mathrm{d}\theta)}{\int \exp(-\eta \hat{L}_t(\gamma))\pi_t(\mathrm{d}\gamma)}.$$

More specifically, the algorithm 1 is based on the exponentially weighted aggregation (EWA) procedure, see e.g. [6,3] and references therein. It updates a probability distribution π_t on the set of single-index representation $\mathcal{G}$ before the encounter of task t. It is noticed that this procedure allows the user to freely choose the within-task algorithm (step **ii**) to learn the task-specific link function h_t, which does not even need to be the same for each task.

Furthermore, the step **i** is crucial during the learning procedure, because to draw $\hat{\theta}_t$ from π_t is not straightforward and varies in different specific situation. While the effect of Step **iii** is that any single-index θ which does not perform well on task t, is less likely to be reused on the next task.

3 Regret bounds

3.1 Bound with expectation

We make the following assumptions on our model.

Assumption 1. *We assume that* $\|\theta\|_1 = 1$ *and* $\|x_{t,i}\|_2 \leq M < +\infty$.

Assumption 2. *We assume that the loss* ℓ *is* L_1*-Lipschitz with respect to its first component, i.e, there exists* $L_1 > 0$ *such that*

$$|\ell(a_1, \cdot) - \ell(a_2, \cdot)| \leq L_1 |a_1 - a_2|.$$

We further assume that $\ell(x,\cdot) \in [0,C], \forall x$.

Assume that we have some within-task algorithms that learn h_t at each time t. And

$$\beta(n_t) := \sup_{\|\theta\|_1=1} \beta(n_t,\theta) < +\infty,$$

$\beta(n_t)$ being an upper bound of the within-task algorithm that learns h_t. We will detail one possible such algorithm in Section 4.

Let π_1 be uniform on the unit ℓ_1-ball. We note that as Algortihm 1 is a randomized algorithm, we first provide a bound on the expected regret. A simple result for continual single-index learning is given in the following theorem.

Theorem 1. *Under the Assumptions 1 and 2, we have*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\hat\theta\sim\pi_t}\left[\frac{1}{n_t}\sum_{i=1}^{n_t}\hat\ell_{t,i}\right] - \inf_{\|\theta\|_1=1}\frac{1}{T}\sum_{t=1}^{\top}\inf_{h_t\in\mathcal{H}}\frac{1}{n_t}\sum_{i=1}^{n_t}\ell(h_t(\theta^\top x_{t,i}),y_{t,i})$$
$$\leq \frac{c_{(L_1,L_2,C,M)}d\log(T)+2d\log(d)}{\sqrt{T}} + \frac{1}{T}\sum_{t=1}^{T}\beta(n_t).$$

where $c_{(L_1,L_2,C,M)}$ is a universal constant that depends only on L_1, L_2, M and C.

The proof relies on an application of Theorem 3.1 in [2]. We postpone the proof to Section A.

3.2 Uniform bound

Now, under additional assumption that the loss function is convex with respect to (w.r.t.) its first component, it is possible to obtain a uniform regret bound. However, rather than using a random draw that $\hat\theta_t \sim \pi_t$ as in Step **i** of Algorithm 1, we need to consider an aggregation step for predicting that is

$$\hat y_{t,i} = \int h_t(\theta^\top x_{t,i})\pi_t(\mathrm{d}\theta). \qquad (1)$$

The unifrom regret bound is presented in the following theorem.

Theorem 2. *Under the assumptions of Theorem 1 and the loss function is convex w.r.t its first argument, we have*

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{n_t}\sum_{i=1}^{n_t}\ell(\hat y_{t,i},y_{t,i}) - \inf_{\|\theta\|_1=1}\frac{1}{T}\sum_{t=1}^{T}\inf_{h_t\in\mathcal{H}}\frac{1}{n_t}\sum_{i=1}^{n_t}\ell(h_t(\theta^\top x_{t,i}),y_{t,i})$$
$$\leq \frac{c_{(L_1,L_2,C,M)}d\log(T)+2d\log(d)}{\sqrt{T}} + \frac{1}{T}\sum_{t=1}^{T}\beta(n_t).$$

where $c_{(L_1,L_2,C,M)}$ is a universal constant that depends only on L_1, L_2, M and C.

Proof. We have that

$$\frac{1}{n_t}\sum_{i=1}^{n_t}\ell(\hat{y}_{t,i},y_{t,i}) = \frac{1}{n_t}\sum_{i=1}^{n_t}\ell\left(\int h_t(\theta^\top x_{t,i})\pi_t(d\theta), y_{t,i}\right).$$

As the loss is convex w.r.t its first component, Jensen's inequality leads to

$$\frac{1}{n_t}\sum_{i=1}^{n_t}\ell\left(\int h_t(\theta^\top x_{t,i})\pi_t(d\theta), y_{t,i}\right) \leq \int \frac{1}{n_t}\sum_{i=1}^{n_t}\ell\left(h_t(\theta^\top x_{t,i}), y_{t,i}\right)\pi_t(d\theta).$$

The proof completes by applying Theorem 1.

Remark 3. Our regret bound is typically at $\log(T)/\sqrt{T}$ order, which tends to 0 as the number of tasks, T increase.

Remark 4. Noted that if all the tasks have the same sample size, that is $n_t = n$ for all t, then $\frac{1}{T}\sum_{t=1}^{T}\beta(n_t) = \beta(n)$ and thus the analysis will not be changed. Here after, to ease our presentation, we assume that all the tasks have the same sample size, that is $n_t = n, \forall t$.

In practice, for an infinite set $\mathcal{G}$ we are not able to run simultaneously the within-task algorithm for all single-index θ. So, we cannot compute the prediction (1) exactly. A possible strategy is to draw N elements i.i.d. from π_t, say $\hat{\theta}_t^{(1)},\dots,\hat{\theta}_t^{(N)}$, and to replace (1) by its Monte Carlo approximation

$$\hat{y}_{t,i}^{(N)} = \frac{1}{N}\sum_{j=1}^{N} h_t(\hat{\theta}_t^{(j)\top} x_{t,i}).$$

Let's call MC-EWA this new version.

Algorithm 2 MC-EWA for continual single-index learning with convex loss

Data and Input as in Algorithm 1.

Loop For $t = 1,\dots,T$

- **i** Draw $\hat{\theta}_t^{(1)},\dots,\hat{\theta}_t^{(N)}$ i.i.d from π_t.
- **ii** Run the within-task learning algorithm $\mathcal{S}_t$ for each $\hat{\theta}_t^{(j)}$ and return as predictions:

$$\hat{y}_{t,i}^{(N)} = \frac{1}{N}\sum_{j=1}^{N} h_t(\hat{\theta}_t^{(j)\top} x_{t,i}).$$

- **iii** Update $\pi_{t+1}(\mathrm{d}\theta) := \frac{\exp(-\eta \hat{L}_t(\theta))\pi_t(\mathrm{d}\theta)}{\int \exp(-\eta \hat{L}_t(\gamma))\pi_t(\mathrm{d}\gamma)}$.

In order to analyze the performance of this algorithm, we can directly use Theorem 2. We only have to control the discrepancy between the theoretical integral with respect to π_t and the corresponding empirical mean. Hoeffding's inequality leads to

$$\frac{1}{N}\sum_{j=1}^{N}\hat{L}_t(\hat{\theta}_t^{(j)}) \leq \mathbb{E}_{\theta\sim\pi_t}[\hat{L}_t(\theta)] + C\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2N}}$$

with probability at least $1-\delta$. A union bound over the T tasks leads to the following result directly.

Corollary 1. *Assuming that the assumptions of Theorem 2 are hold. Then, with probability at least $1-\delta$ over the drawing of all the $\hat{\theta}_t^{(j)}$'s,*

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{m_t}\sum_{i=1}^{m_t}\ell\left(\hat{y}_{t,i}^{(N)}, y_{t,i}\right) - \inf_{\|\theta\|_1=1}\frac{1}{T}\sum_{t=1}^{T}\inf_{h_t\in\mathcal{H}}\frac{1}{n_t}\sum_{i=1}^{n_t}\ell(h_t(\theta^\top x_{t,i}), y_{t,i})$$
$$\leq \frac{c_{(L_1,L_2,C,M)}d\log(T) + 2d\log(d)}{\sqrt{T}} + \frac{1}{T}\sum_{t=1}^{T}\beta(n_t) + C\sqrt{\frac{\log\left(\frac{T}{\delta}\right)}{2N}}.$$

In the next Section, we provide an example of a within task online algorithm and derive its regret bound.

4 A within-task algorithm

4.1 EWA for online single-index learning

Here, we propose an online algorithm for learning within each task, detailed in Algorithm 3. The algorithm is based on the EWA procedure on the space $\mathcal{H}\circ g$ for a prescribed single-index representation $g\in\mathcal{G}$, with $g(x)=\theta x$.

Algorithm 3 EWA for online single-index learning

Data A task $\mathcal{S}_t = \big((x_{t,1}, y_{t,1}), \ldots, (x_{t,n_t}, y_{t,n_t})\big)$ where the data points are given sequentially.

Input A learning rate $\zeta > 0$;
a prior distribution μ_1 on $\mathcal{H}$.

Loop For $i = 1, \ldots, n_t$,

i Predict $\hat{y}_{t,i}^{\theta} = \int_{\mathcal{H}} h(\theta x_{t,i})\mu_i(\mathrm{d}h)$,

ii $y_{t,i}$ is revealed, update

$$\mu_{i+1}(\mathrm{d}h) = \frac{\exp(-\zeta\ell(\hat{y}_{t,i}^{\theta}, y_{t,i}))\mu_i(\mathrm{d}h)}{\int\exp(-\zeta\ell(u(\theta x_{t,i}), y_{t,i}))\mu_i(\mathrm{d}u)}.$$

To learn h_t, we use Algorithm 3 and consider a structure for $\mathcal{H}$. We consider, for a positive interger S, the link function

$$h_t \in \mathcal{H}_{S,C_2+1} := \{h \in \mathcal{H} : h = \sum_{j=1}^{S} \beta_j \phi_j, \sum_{j=1}^{S} j|\beta_j| \leq C_2 + 1\},$$

where $\{\phi_j\}_{j=1}^{\infty}$ is a dictionary of measurable functions, each ϕ_j is assumed to be defined on $[-1, 1]$ and to take values in $[-1, 1]$. The trigonometric system [21] is an example for this kind of dictionary, that is $\phi_1(z) = 1, \phi_{2j}(z) = \cos(\pi j z), \phi_{2j+1}(z) = \sin(\pi j z)$ with $j = 1, 2, \dots$ and $z \in [-1, 1]$.

Let

$$\mathcal{B}_S(C_2 + 1) := \{(\beta_1, \dots, \beta_S) \in \mathbb{R}^S : \sum_{j=1}^{S} j|\beta_j| \leq C_2 + 1\}.$$

We define $\mu_1(dh)$ on the set $\mathcal{H}_{S,C_2+1}$ as the image of the uniform measure on $\mathcal{B}_S(C_2 + 1)$ induced by the map $(\beta_1, \dots, \beta_S) \mapsto \sum_{j=1}^{S} \beta_j \phi_j$.

Remark 5. The choice of $C_2 + 1$ instead of C_2 in the definition of the prior support is just convenient for technical proofs. This ensures that when the target h_t belongs to $\mathcal{H}_{S,C_2}$, then a small ball around it is contained in $\mathcal{H}_{S,C_2+1}$.

Remark 6. The integer S should be understood as a measure of the "dimension" of the link function h_t; the larger S, the more complex the function.

Now, we are ready to provide a regret bound for Algorithm 3. Remind that we assume that $n_t = n, \forall t$.

Proposition 1. *By choosing* $\zeta = \sqrt{\frac{8S}{C^2 n}}$ *, we have*

$$\frac{1}{n}\sum_{i=1}^{n} \hat{\ell}_{t,i} - \inf_{h_t \in \mathcal{H}_{S,C_2+1}} \frac{1}{n}\sum_{i=1}^{n} \ell(h_t(\theta^\top x_{t,i}), y_{t,i}) \leq a_{(L_1,S,C,C_2)} \frac{\log(n)}{\sqrt{n}},$$

where $a_{(L_1,S,C,C_2)}$ *is a universal constant that depends only on* L_1, S, C, C_2.

As the proof of the Proposition 1 is not straightforward, we postpone the proof to Section A.

4.2 A detailed regret bound

We are ready to provide a full regret bound for continual single-index learning. The following result is obtained by plug in Proposition 1 into Theorem 2.

Corollary 2. *Under the assumptions of Theorem 2 and Proposition 1, we have*

$$\frac{1}{T}\sum_{t=1}^{\top} \frac{1}{n}\sum_{i=1}^{n} \hat{\ell}_{t,i} - \inf_{\|\theta\|_1=1} \frac{1}{T}\sum_{t=1}^{\top} \inf_{h_t \in \mathcal{H}_{S,C_2+1}} \frac{1}{n}\sum_{i=1}^{n} \ell(h_t(\theta^\top x_{t,i}), y_{t,i})$$

$$\leq a_{(L_1,S,C,C_2)} \frac{\log(n)}{\sqrt{n}} + \frac{c_{(L_1,L_2,C)} d \log(T) + 2d \log(d)}{\sqrt{T}},$$

where $c_{(L_1,L_2,C,M)}$ *is a universal constant that depends only on* L_1, L_2, C, M *and* $a_{(L_1,S,C,C_2)}$ *is a universal constant that depends only on* L_1, S, C, C_2.

Typically, we obtain a regret bound at the order of $\log(n)/\sqrt{n} + \log(T)/\sqrt{T}$.

5 Discussion and Conclusion

We presented a meta-algorithm for continual single-index learning and provided for the first time a fully online analysis of its regret. We also provided an online algortihm for learning within task and proved its regret bound. This is novel to our knowledge.

A fundamental question is to provide a more computationally efficient algorithm, such as approximations of EWA [8], or fully gradient based algorithms as in [19,16].

A Proofs

First, we state the following result that is useful for our proofs.

Theorem 3. [2, Theorem 3.1] *If, for any* $g \in \mathcal{G}, \ell(x) \in [0, C]$ *and the within-task algorithm has a regret bound* $\beta(g, m_t)$*, then*

$$\frac{1}{T}\sum_{t=1}^{\top} \mathbb{E}_{\hat{g}_t \sim \pi_t}\left[\frac{1}{m_t}\sum_{i=1}^{m_t} \hat{\ell}_{t,i}\right] \leq \inf_{\rho}\Bigg\{\mathbb{E}_{g\sim\rho}\Bigg[\frac{1}{T}\sum_{t=1}^{\top}\inf_{h_t\in\mathcal{H}}\frac{1}{m_t}\sum_{i=1}^{m_t}\ell\big(h_t\circ g(x_{t,i}), y_{t,i}\big) + \frac{1}{T}\sum_{t=1}^{\top}\beta(g, m_t)\Bigg] + \frac{\eta C^2}{8} + \frac{\mathcal{K}(\rho,\pi_1)}{\eta T}\Bigg\},$$

where the infimum is taken over all probability measures ρ *and* $\mathcal{K}(\rho, \pi_1)$ *is the Kullback-Leibler divergence between* ρ *and* π_1.

A.1 Proof of Theorem 1

Proof. Let θ^* denote a minimizer of the optimization problem

$$\min_{\|\theta\|_1=1} \frac{1}{T}\sum_{t=1}^{\top}\inf_{h_t\in\mathcal{H}}\frac{1}{n_t}\sum_{i=1}^{n_t}\ell(h_t(\theta^\top x_{t,i}), y_{t,i}).$$

We apply Theorem 3.1 in [2] and upper bound the infimum w.r.t any ρ by an infimum with respect to ρ_ϵ in the following parametric family

$$\rho_\epsilon(\mathrm{d}\theta) \propto \mathbf{1}\{\|\theta - \theta^*\|_2 \leq \epsilon\}\pi_1(d\theta).$$

where ϵ is a positive parameter. Note that when ϵ is small, ρ_ϵ highly concentrates around θ^*, but we will show this is at a price of an increase in $\mathcal{K}(\rho_\epsilon, \pi_1)$. The proof then proceeds in optimizing with respect to ϵ.

More specifically, Theorem 3.1 in [2] becomes

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{\top}\mathbb{E}_{\hat{g}_t\sim\pi_t}\left[\frac{1}{n_t}\sum_{i=1}^{n_t}\hat{\ell}_{t,i}\right]\\ &\leq \inf_{\epsilon}\left\{\mathbb{E}_{\theta\sim\rho_\epsilon}\left[\frac{1}{T}\sum_{t=1}^{\top}\inf_{h_t\in\mathcal{H}}\frac{1}{n_t}\sum_{i=1}^{n_t}\ell(h_t(\theta^\top x_{t,i}),y_{t,i})+\beta(n_t)\right]+\frac{\eta C^2}{8}+\frac{\mathcal{K}(\rho_\epsilon,\pi_1)}{\eta T}\right\}.\end{aligned}$$

Furthermore, using the notation

$$h_t^* := \arg\inf_{h_t\in\mathcal{H}}\frac{1}{n_t}\sum_{i=1}^{n_t}\ell(h_t(\theta^\top x_{t,i}),y_{t,i}),$$

we get

$$\begin{aligned}&\inf_{h_t\in\mathcal{H}}\frac{1}{n_t}\sum_{i=1}^{n_t}\ell(h_t(\theta^\top x_{t,i}),y_{t,i})-\frac{1}{n_t}\sum_{i=1}^{n_t}\ell(h_t^*(\theta^{*\top}x_{t,i}),y_{t,i})\\ &\qquad\leq\frac{1}{n_t}\sum_{i=1}^{n_t}\ell(h_t^*(\theta^\top x_{t,i}),y_{t,i})-\frac{1}{n_t}\sum_{i=1}^{n_t}\ell(h_t^*(\theta^{*\top}x_{t,i}),y_{t,i}).\end{aligned}$$

Under the condition on the loss, we have

$$\begin{aligned}\left|\ell(h_t^*(\theta^\top x_{t,i}),y_{t,i})-\ell(h_t^*(\theta^{*\top}x_{t,i}),y_{t,i})\right| &\leq L\left|h_t^*(\theta^\top x_{t,i})-h_t^*(\theta^{*\top}x_{t,i})\right|\\ &\leq L_1L_2|(\theta-\theta^*)^\top x_{t,i}|\\ &\leq \epsilon L_1L_2\|x_{t,i}\|_2.\end{aligned}$$

We obtain an upper-bound

$$\begin{aligned}&\mathbb{E}_{\theta\sim\rho_\epsilon}\frac{1}{T}\sum_{t=1}^{\top}\inf_{h_t\in\mathcal{H}}\frac{1}{n_t}\sum_{i=1}^{n_t}\ell(h_t(\theta^\top x_{t,i}),y_{t,i})\\ &\leq\inf_{\|\theta\|_1=1}\left\{\frac{1}{T}\sum_{t=1}^{\top}\inf_{h_t\in\mathcal{H}}\frac{1}{n_t}\sum_{i=1}^{n_t}\ell(h_t(\theta^\top x_{t,i}),y_{t,i})+\epsilon\,L_1L_2\frac{1}{T}\sum_{t=1}^{\top}\frac{1}{n_t}\sum_{i=1}^{n_t}\|x_{t,i}\|_2\right\}.\end{aligned}$$

Now, dealing with the Kullback-Leibler, we have

$$\mathcal{K}(\rho_\epsilon,\pi_1)=-\log\pi_1(\{\|\theta-\theta^*\|_2\leq\epsilon\}),$$

and

$$\pi_1(\{\|\theta-\theta^*\|_2\leq\epsilon\})\geq\frac{\pi^{(d-1)/2}}{\Gamma((d+1)/2)}\epsilon^{(d-1)}\Bigg/\frac{2^d}{d!}$$

$$\geq \frac{\epsilon^{(d-1)}}{\sqrt{\pi(d-1)}}\left(\frac{2\pi e}{d-1}\right)^{(d-1)/2} \Bigg/ \frac{2^d}{d!}$$
$$\geq \epsilon^{d-1}2^{d-2}\frac{d!}{(d-1)^{d/2}}.$$

Note that the first inequality follows by observing that, since π_1 is the uniform distribution on the unit ℓ_1 ball, the probability to be calculated is greater or equal to the ration between the volume of the $(d-1)-$ball radius ϵ over the volume of the unit ℓ_1 ball. The second inequality is just using the Stirling formula.

Consequently, we obtain

$$\mathcal{K}(\rho_\epsilon, \pi_1) \leq (d-1)\log(1/\epsilon) + \log\left(\frac{2^{d-2}d!}{(d-1)^{d/2}}\right).$$

Therefore, Theorem 3.1 in [2] leads to

$$\frac{1}{T}\sum_{t=1}^{\top}\mathbb{E}_{\theta\sim\pi_t}\left[\frac{1}{n_t}\sum_{i=1}^{n_t}\hat{\ell}_{t,i}\right] - \inf_{\|\theta\|_1=1}\frac{1}{T}\sum_{t=1}^{\top}\inf_{h_t\in\mathcal{H}}\frac{1}{n_t}\sum_{i=1}^{n_t}\ell(h_t(\theta^\top x_{t,i}), y_{t,i})$$
$$\leq \inf_\epsilon\left\{\epsilon\, L_1L_2M + \frac{(d-1)\log(1/\epsilon)}{\eta T}\right\} + \frac{\log\left(\frac{2^{d-2}d!}{(d-1)^{d/2}}\right)}{2\eta T} + \beta(n_t) + \frac{\eta C^2}{8}.$$

The choices $\epsilon = \sqrt{\frac{1}{T}}$ and $\eta = \frac{2}{C}\sqrt{\frac{1}{T}}$ make the right-hand side becomes

$$\frac{L_1L_2M}{\sqrt{T}} + \frac{Cd\log(T) + \log\left(\frac{2^{d-2}d!}{(d-1)^{d/2}}\right) + C}{4\sqrt{T}} + \beta(n_t).$$

The proof is completed by using the Stirling's approximation that $\log(d!) \sim d\log(d)$.

A.2 Proof of Proposition 1

Proof. We follows the same steps as in the proof of Theorem 1 in [3]. First, we have that

$$\mu_i(\mathrm{d}h) = \frac{\exp(-\zeta\ell(h(\theta x_{t,i}), y_{t,i}))\mu_i(\mathrm{d}h)}{\int\exp(-\zeta\ell(u(\theta x_{t,i}), y_{t,i}))\mu_i(\mathrm{d}u)}. = \frac{\exp(-\zeta\ell(h(\theta x_{t,i}), y_{t,i}))\mu_i(\mathrm{d}h)}{W_i}. \tag{2}$$

where we introduce the notation W_i for the sake of shortness. We denote $\ell(h) := \ell(h(\theta x_{t,i}), y_{t,i}))$ and put

$$E_i = \int \ell(h)\mu_i(\mathrm{d}h) = \mathbb{E}_{h_i\sim\mu_i}[\ell(h_i)].$$

Using Hoeffding's inequality on the bounded random variable $\ell(h) \in [0, C]$ we have, for any i, that

$$\mathbb{E}_{h\sim\mu_i}\left[\exp\left\{\zeta(E_i - \ell(h)\right\}\right] \leq \exp\left\{\frac{C^2\zeta^2}{8}\right\}$$

which can be rewritten as

$$\exp\left\{-\zeta\mathbb{E}_{h\sim\mu_i}[\ell(h)]\right\} \geq \exp\left(-\frac{C^2\zeta^2}{8}\right)\mathbb{E}_{h\sim\mu_i}\left\{\exp\left[-\zeta\ell(h)\right]\right\}. \tag{3}$$

Next, we note that

$$\begin{aligned}
&\exp\left\{-\zeta\sum_{i=1}^{n}\mathbb{E}_{h\sim\mu_i}[\ell(h)]\right\} \\
&= \prod_{i=1}^{n}\exp\left\{-\zeta\mathbb{E}_{h\sim\mu_i}[\ell(h)]\right\} \\
&\geq \exp\left(-\frac{nC^2\zeta^2}{8}\right)\prod_{i=1}^{n}\mathbb{E}_{h\sim\mu_i}\left\{\exp\left[-\zeta\ell(h)\right]\right\}, \text{ using (3)} \\
&= \exp\left(-\frac{nC^2\zeta^2}{8}\right)\prod_{i=1}^{n}\int\left\{\exp\left[-\zeta\ell(h)\right]\right\}\mu_i(\mathrm{d}h) \\
&= \exp\left(-\frac{nC^2\zeta^2}{8}\right)\prod_{i=1}^{n}\int\frac{\exp\left\{-\zeta\sum_{u=1}^{n}\ell(h_u)\right\}}{W_i}\mu_1(\mathrm{d}h), \text{ using (2)} \\
&= \exp\left(-\frac{nC^2\zeta^2}{8}\right)\prod_{i=1}^{n}\frac{W_{i+1}}{W_i} = \exp\left\{\frac{nC^2\zeta^2}{8}\right\}W_{n+1}.
\end{aligned}$$

So

$$\begin{aligned}
\sum_{i=1}^{n}\mathbb{E}_{h\sim\mu_i}[\ell(h)] &\leq -\frac{\log W_{n+1}}{\zeta} + \frac{nC^2\zeta}{8} \\
&= -\frac{\log\int\exp\left[-\zeta\sum_{i=1}^{n}\ell(h_i)\right]\mu_1(\mathrm{d}h)}{\zeta} + \frac{nC^2\zeta}{8}
\end{aligned}$$

and finally we use [5, Equation (5.2.1)] which states that

$$-\frac{\log\int\exp\left[-\zeta\sum_{i=1}^{n}\ell(h_i)\right]\mu_1(\mathrm{d}h)}{\zeta} = \inf_{\nu}\left\{\mathbb{E}_{h_i\sim\nu}\left[\sum_{i=1}^{n}\ell(h_i)\right] + \frac{\mathcal{K}(\nu,\mu_1)}{\zeta}\right\}.$$

Therefore, for each t and given a θ, we obtain a general bound

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\ell}_{t,i} \quad\leq\quad \inf_{\nu}\mathbb{E}_{h_t\sim\nu}\left\{\frac{1}{n}\sum_{i=1}^{n}\ell(h_t(\theta^\top x_{t,i}), y_{t,i}) \;+\; \frac{\zeta C^2}{8} \;+\; \frac{\mathcal{K}(\nu,\mu_1)}{\zeta n}\right\}. \tag{4}$$

Put

$$h_t^* := \arg\inf_{h_t\in\mathcal{H}_{S,C_2+1}}\frac{1}{n}\sum_{i=1}^{n}\ell(h_t(\theta^\top x_{t,i}), y_{t,i}).$$

We define

$$\|h\|_S = \sum_{j=1}^{S} j|\beta_j|, \forall h \in \mathcal{H}_{S,C_2+1}.$$

and let

$$\nu_\gamma = \mathbf{1}(\|h - h_t^*\|_S \le \gamma)\mu_1(dh).$$

The bound in (4) becomes

$$\frac{1}{n}\sum_{i=1}^{n} \hat{\ell}_{t,i} \quad \le \quad \inf_{\gamma} \mathbb{E}_{h_t\sim\nu_\gamma}\left\{\frac{1}{n}\sum_{i=1}^{n} \ell(h_t(\theta^\top x_{t,i}), y_{t,i}) \ + \ \frac{\zeta C^2}{8} \ + \ \frac{\mathcal{K}(\nu_\gamma, \mu_1)}{\zeta n}\right\}.$$

Under the condition on the loss, we have

$$\begin{aligned}\left|\ell(h_t^*(\theta^\top x_{t,i}), y_{t,i}) - \ell(h_t(\theta^\top x_{t,i}), y_{t,i})\right| &\le L_1 \left|h_t^*(\theta^\top x_{t,i}) - h_t(\theta^\top x_{t,i})\right| \\ &\le L_1 \sup_z |h_t^*(z) - h_t(z)| \\ &\le L_1\gamma.\end{aligned}$$

Using the Lemma 10 in [1], we have

$$\mathcal{K}(\nu_\gamma, \mu_1) \le S \log \frac{(C_2+1)}{\gamma}.$$

Thus we obtain

$$\frac{1}{n}\sum_{i=1}^{n} \hat{\ell}_{t,i} - \inf_{h_t\in\mathcal{H}_{S,C_2+1}} \frac{1}{n}\sum_{i=1}^{n} \ell(h_t(\theta^\top x_{t,i}), y_{t,i}) \le \inf_{\gamma}\left\{L_1\gamma + \frac{\zeta C^2}{8} + \frac{S\log\frac{(C_2+1)}{\gamma}}{\zeta n}\right\}.$$

By choosing $\gamma = 1/\sqrt{n}$ and then the optimum is reached at $\zeta = \sqrt{\frac{8S}{C^2 n}}$

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n} \hat{\ell}_{t,i} - \inf_{h_t\in\mathcal{H}_{S,C_2+1}} \frac{1}{n}\sum_{i=1}^{n} \ell(h_t(\theta^\top x_{t,i}), y_{t,i}) \\ &\qquad\qquad \le \frac{L_1}{\sqrt{n}} + \frac{C\sqrt{S}}{2\sqrt{2n}} + \frac{C\sqrt{S}\log[(C_2+1)\sqrt{n}]}{2\sqrt{2n}}.\end{aligned}$$

This completes the proof.

References

1. Alquier, P., Biau, G.: Sparse single-index model. J. Mach. Learn. Res. **14**, 243–280 (2013)

2. Alquier, P., Mai, T.T., Pontil, M.: Regret bounds for lifelong learning. In: PMLR (AISTATS). vol. 54, pp. 261–269 (2017)
3. Audibert, J.Y.: A randomized online learning algorithm for better variance control. In: Proc. 19th Annual Conference on Learning Theory, pp. 392–407. Springer (2006)
4. Baxter, J.: A model of inductive bias learning. Journal of Artificial Intelligence Research **12**, 149–198 (2000)
5. Catoni, O.: Statistical learning theory and stochastic optimization, Saint-Flour Summer School on Probability Theory 2001 (Jean Picard ed.), Lecture Notes in Mathematics, vol. 1851. Springer-Verlag, Berlin (2004). https://doi.org/10.1007/b99352, http://dx.doi.org/10.1007/b99352
6. Cesa-Bianchi, N., Lugosi, G.: Prediction, learning, and games. Cambridge University Press (2006)
7. Chen, Z., Liu, B.: Lifelong machine learning. Synthesis Lectures on Artificial Intelligence and Machine Learning **12**(3), 1–207 (2018)
8. Chérief-Abdellatif, B.E., Alquier, P., Khan, M.E.: A generalization bound for online variational inference. In: Asian Conference on Machine Learning. pp. 662–677. PMLR (2019)
9. Denevi, G., Ciliberto, C., Stamos, D., Pontil, M.: Learning to learn around a common mean. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 31 (NIPS 2018). vol. 31. NIPS Proceedings (2018)
10. Finn, C., Rajeswaran, A., Kakade, S., Levine, S.: Online meta-learning. In: International Conference on Machine Learning. pp. 1920–1930. PMLR (2019)
11. Horowitz, J.L.: Semiparametric methods in econometrics, vol. 131. Springer Science & Business Media (2012)
12. Ichimura, H.: Semiparametric least squares (sls) and weighted sls estimation of single-index models. Journal of Econometrics **58**(1-2), 71–120 (1993)
13. Lopez, O.: Single-index regression models with right-censored responses. Journal of Statistical Planning and Inference **139**(3), 1082–1097 (2009)
14. McAleer, M., Da Veiga, B.: Single-index and portfolio models for forecasting value-at-risk thresholds. Journal of Forecasting **27**(3), 217–235 (2008)
15. McCullagh, P., Nelder, J.A.: Generalized Linear Models, vol. 37. CRC Press (1989)
16. Nguyen, C.V., Li, Y., Bui, T.D., Turner, R.E.: Variational continual learning. In: International Conference on Learning Representations (2018)
17. Pentina, A., Lampert, C.: A pac-bayesian bound for lifelong learning. In: Proc. 31st International Conference on Machine Learning. pp. 991–999 (2014)
18. Pentina, A., Ben-David, S.: Multi-task and lifelong learning of kernels. In: Proc. 26th International Conference on Algorithmic Learning Theory. pp. 194–208 (2015)
19. Ruvolo, P., Eaton, E.: Ella: An efficient lifelong learning algorithm. In: Proc. 30th International Conference on Machine Learning, pp. 507–515 (2013)
20. Thrun, S., Pratt, L.: Learning to Learn. Kluwer Academic Publishers (1998)
21. Tsybakov, A.B.: Introduction to nonparametric estimation. Springer Science & Business Media (2008)
22. Wu, Y.S., Wang, P.A., Lu, C.J.: Lifelong optimization with low regret. In: The 22nd International Conference on Artificial Intelligence and Statistics. pp. 448–456. PMLR (2019)