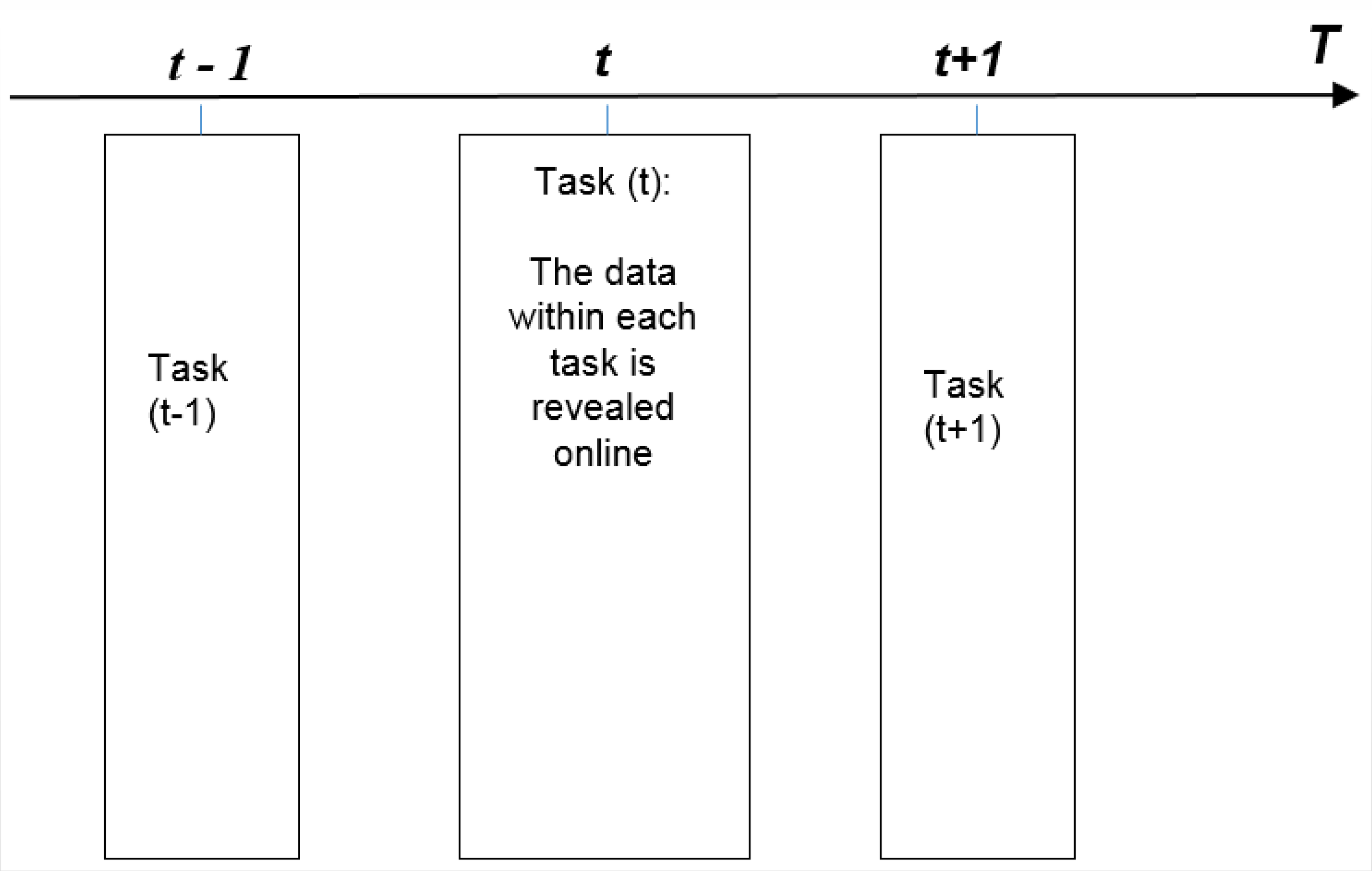
Regret bounds for ℓifelong ℓearning

Pierre Alquier and **T.Tien Mai** CREST, ENSAE, Université Paris Saclay
Massimiliano Pontil University College London & Istituto Italiano di Tecnologia

Abstract

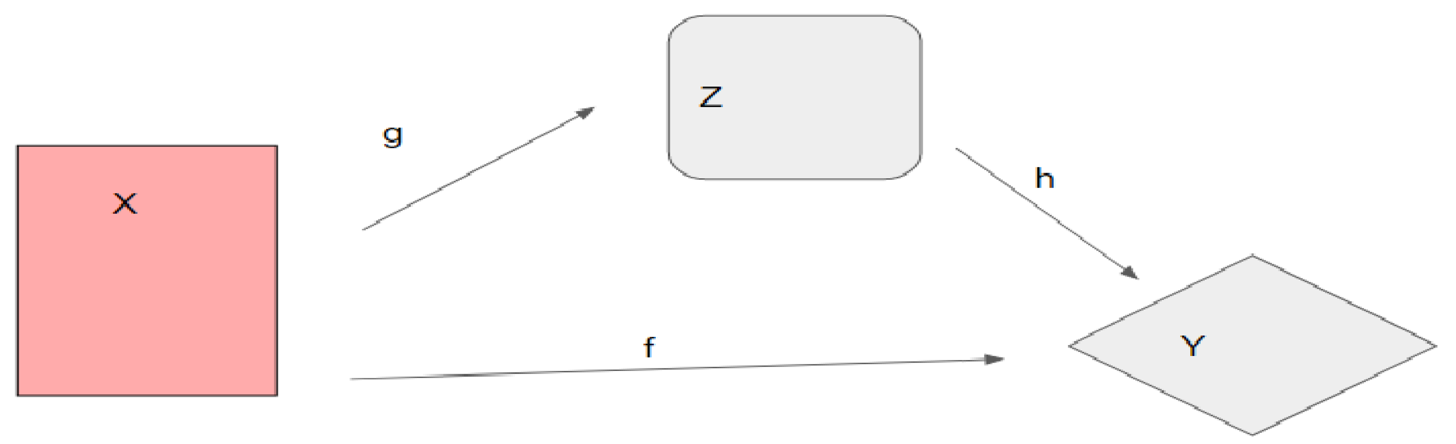
We consider the problem of transfer learning in an online setting. Different tasks are presented sequentially and processed by a within-task algorithm. We propose a lifelong learning strategy which refines the underlying data representation used by the within-task algorithm, thereby transferring information from one task to the next. We show that when the within-task algorithm comes with some regret bound, our strategy inherits this good property. Our bounds are in expectation for a general loss function, and uniform for a convex loss. We discuss applications to dictionary learning and finite set of predictors. In the latter case, we improve previous $O(1/\sqrt{m})$ bounds to $O(1/m)$, where m is the per task sample size.

Problem setting



Objective We wish to design a procedure (meta-algorithm) that,

- transfer the learned information from previous tasks to the next,



Let's define that $g \in \mathcal{G}$ is a feature map (common data representation) and $h_t \in \mathcal{H}$ is a task-specific function. Such that

$$f_t = h_t \circ g$$

is a good predictor for task t.

- control the *compound regret* of our procedure

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{m_t}\sum_{i=1}^{m_t}\hat{\ell}_{t,i} - \inf_{g\in\mathcal{G}}\frac{1}{T}\sum_{t=1}^{T}\inf_{h_t\in\mathcal{H}}\frac{1}{m_t}\sum_{i=1}^{m_t}\ell\left(h_t\circ g(x_{t,i}), y_{t,i}\right).$$

Examples of Within Task Algorithms Given an online task (data) $\mathcal{S}_t = \left((x_{t,1}, y_{t,1}), \ldots, (x_{t,m_t}, y_{t,m_t})\right)$ and a prescribed representation g.

Online Gradient Algortihm OGA

Given a step-size $\zeta > 0$ and $\theta_1 = 0$. Loop for $i = 1, \ldots, m_t$,

1. Predict $\hat{y}^g_{t,i} = h_{\theta_i} \circ g(x_{t,i})$,
2. $y_{t,i}$ is revealed, update $\theta_{i+1} = \theta_i - \zeta\nabla_\theta \ell\left(h_\theta \circ g(x_{t,i}), y_{t,i}\right)\big|_{\theta=\theta_i}$.

- A regret bound for OGA is $\beta(g, m_t) = \mathcal{O}(1/\sqrt{m_t})$ (convex, Lipschitz).
- [3] provides bounds for $\beta(g, m_t)$ in $\mathcal{O}(\log(m_t)/m_t)$ under additional assumptions.

Exponentially Weighted Aggregation EWA

Given a learning rate $\zeta > 0$; a prior distribution μ_1 on $\mathcal{H}$. Loop for $i = 1, \ldots, m_t$,

1. Predict $\hat{y}^g_{t,i} = \int_{\mathcal{H}} h \circ g(x_{t,i})\mu_i(\mathrm{d}h)$,
2. $y_{t,i}$ is revealed, update $\mu_{i+1}(\mathrm{d}h) = \frac{\exp(-\zeta\ell(h\circ g(x_{t,i}), y_{t,i}))\mu_i(\mathrm{d}h)}{\int\exp(-\zeta\ell(u\circ g(x_{t,i}), y_{t,i}))\mu_i(\mathrm{d}u)}$.

- A regret bound for EWA is $\beta(g, m_t) = \mathcal{O}(\sqrt{\log(|\mathcal{H}|)/m_t})$.
- better bound for EWA is $\beta(g, m_t) = \mathcal{O}(\log|\mathcal{H}|/m_t)$, under exp-concavity [2].

Lifelong learning procedure

EWA-LL Algorithm

1. **Input:** datasets $\mathcal{S}_t = \left((x_{t,1}, y_{t,1}), \ldots, (x_{t,m_t}, y_{t,m_t})\right)$ are given in sequence for different learning tasks $t = 1, \ldots, T$; the points within each dataset are also given sequentially. A prior π_1, a learning rate $\eta > 0$.
2. A learning algorithm for each task t which, for any representation g returns a sequence of predictions $\hat{y}^g_{t,i}$ and suffers a loss $\hat{L}_t(g) := \frac{1}{m_t}\sum_{i=1}^{m_t}\ell\left(\hat{y}^g_{t,i}, y_{t,i}\right)$.
3. **Loop:** For $t = 1, \ldots, T$

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

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

   iii Update

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

Theorem

If, for any $g \in \mathcal{G}$, $\hat{L}_t(g) \in [0, C]$ and the within-task algorithm has a regret bound $\mathcal{R}_t(g) \leq \beta(g, m_t)$, then

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

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.

Finite Subset of Relevant Predictors

$\mathcal{G}$ is a set of K functions and $\mathcal{H}$ is finite. Assume: $\ell(\cdot, y)$ is ζ_0-exp-concave and upper bounded by a constant C.
Then the EWA-LL algorithm with $\eta = (2/C)\sqrt{2\log(K)/T}$ using the EWA within task with $\zeta = \zeta_0$ satisfies

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\hat{g}_t\sim\pi_t}\left[\frac{1}{m}\sum_{i=1}^{m}\hat{\ell}_{t,i}\right] \leq \min_{1\leq k\leq K}\frac{1}{T}\sum_{t=1}^{T}\min_{h_t\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\ell(h_t\circ g_k(x_{t,i}), y_{t,i}) + \frac{\zeta_0\log|\mathcal{H}|}{m} + C\sqrt{\frac{\log K}{2T}}.$$

In particular, our $\mathcal{O}(1/m)$ bound improves upon [4] who derived an $\mathcal{O}(1/\sqrt{m})$ bound.

Lifelong dictionary learning

$\mathcal{X} = \mathbb{R}^d$. $\mathcal{D}_K = \{D_{d\times K} : \|D_{\cdot,j}\|_2 = 1, j = 1, \ldots, K\}$, let $\mathcal{G} = \{x \mapsto Dx : D \in \mathcal{D}_K\}$.
Assume: $\|x_{t,i}\| \leq 1$ and ℓ is convex and Φ-Lipschitz w.r.t its 1^{st} component.
The prior π_1: the columns of D are i.i.d. uniformly distributed on the d-dimensional unit sphere.
Algorithm EWA-LL for dictionary learning, with $\eta = (2/C)\sqrt{Kd/T}$, and using the OGA algorithm within tasks, with step $\zeta = B/(\Phi\sqrt{2mK})$, satisfies

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\hat{\ell}_{t,i} \leq \inf_{D\in\mathcal{D}_K}\frac{1}{T}\sum_{t=1}^{T}\inf_{h_t\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\ell(\langle h_t, Dx_{t,i}\rangle, y_{t,i}) + \frac{C}{4}\sqrt{\frac{Kd}{T}}(\log(T) + 7) + \frac{B\Phi}{\sqrt{T}} + \frac{\Phi B\sqrt{2K}}{\sqrt{m}}.$$

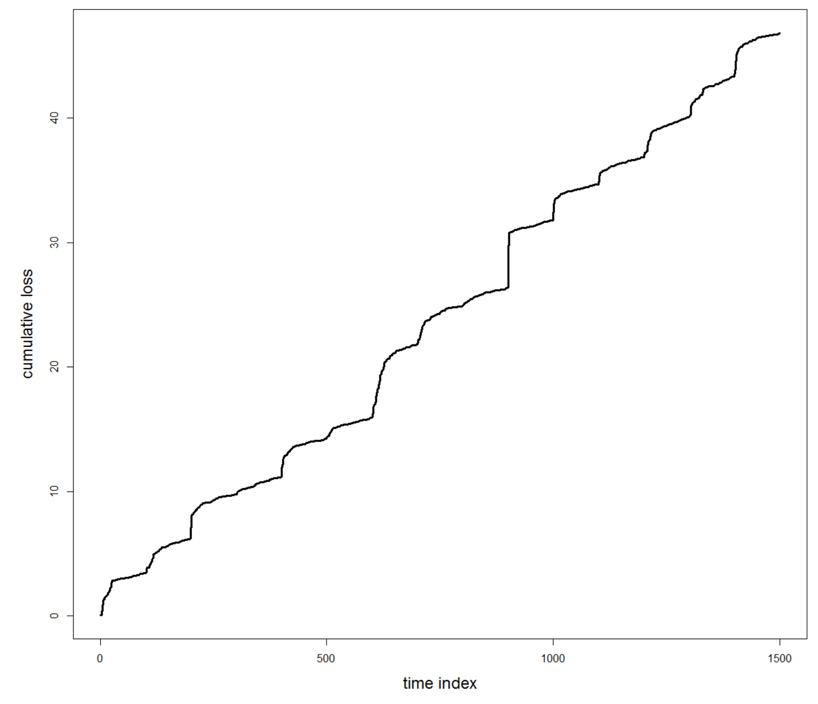


Figure 1: The cumulative loss of the oracle for the first 15 tasks.

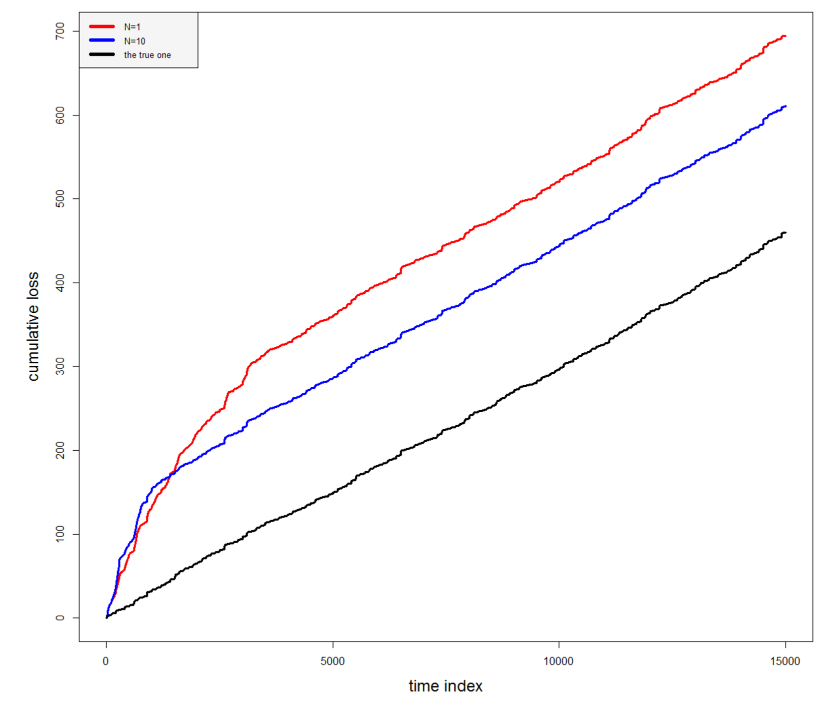


Figure 2: Cumulative loss of EWA-LL ($N = 1$ in red and $N = 10$ in blue) and cumulative loss of the oracle.

Acknowledgements: We are grateful to Mark Herbster for useful comments. This work was supported by the research programme *New Challenges for New Data* from LCL and GENES, hosted by the *Fondation du Risque*, from Labex ECODEC (ANR - 11-LABEX-0047) and EPSRC grant EP/P009069/1.

[1] Audibert. A randomized online learning algorithm for better variance control. In *the 19th COLT*. Springer, 2006.

[2] Gerchinovitz. *Prediction of individual sequences and prediction in the statistical framework: some links around sparse regression and aggregation techniques.* PhD thesis, U. Paris 11, 2011.

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

[4] Pentina and Lampert. A pac-bayesian bound for lifelong learning. In *Proc. of the 31st ICML*, 2014.

[5] Thrun and Pratt (Eds.). (2012). Learning to learn. Springer Science & Business Media.