Learning in actionable universes

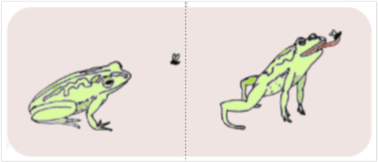
Emmanuel Daucé

Centrale Marseille, Marseille, France
Aix Marseille Univ, Inserm, INS, Institut de Neurosciences des Systèmes, Marseille, France

1. Introduction
2. Contributions
   - Daucé, E, Proix, T, Ralaivola, L ; proc. of IJCNN 2015
   - Zhong, H and Daucé, E, hal-01345825, submitted
3. Numerical experiments
   - Zhong, H and Daucé, E, hal-01345825, submitted
   - Daucé, E, Proix, T, Ralaivola, L ; proc. of IJCNN 2015
4. Conclusion

Recommender systems vs. Biological control systems

Vs.



- Online learning
- Large response set
- Context adaptation
- Frequent renewal

Main intuition

Labels are actions !

Labels vs. Actions

Category learning :

- Label = Latent variable
- The label y is the *cause* of observation x
- Generative models, Predictive coding, ...

Action-based learning :

- Label = Action
- The observation x is the *cause* of action y
- (and y is possibly the cause of *next* observation x')
- Bandit models, POMDP, Active inference, ...

Assumption : natural systems control is grounded on action-based learning.

Actionable universe (Bandit problem)

- Repertoire of K possible *actions* : 1, ..., K
- Learn the effect-of-action through a sequence of trial : $\tilde{y}_1$, ..., $\tilde{y}_t$, ...
- Every action $\tilde{y}_t$ brings an *information* (feedback) f_t (which in turn provides a *loss* l_t)
- Exploration/exploitation dilemma :
  - (either) Better predict the effect-of-action $W_t \simeq P(f|\tilde{y})$
  - (or) minimize regret

$$\sum_{t \in 1,..,T} l_t - l_t^*$$

Contextual actionable universe (Online learning)

A sequential update :

For all $t \in 1, \dots T$:

1. Read a *context* $x_t \in \mathbb{R}^d$
2. Choose a *label* $\tilde{y}_t \in \{1, \dots, K\}$
3. Read feedback f_t
4. Update model W_t

Related problems :

- Contextual bandits [Lai and Robbins, 1985, Auer et al., 2002]
- Supervised online learning [Rosenblatt, 1958, Duda et al., 1973]

Role of feedback

The feedback/loss function interplay is at the core of learning.

- Bandit case :
  - quantitative feedback $f_t \in \mathbb{R}$
  - $f_t = -l_t$ relates to the direct outcome of action
- Category learning :
  - A qualitative feedback :
    - either explicit : $f_t = y_t$
    - or implicit : "good"/"bad" (reward)
  - The loss l_t is *derived* from f_t.

Binary guiding in multi-class classification

- Binary feedback :
  - "all-or-nothing" feedback
  - clic, like, visit, follow, retweet ...
  - very common in man-machine interaction
- Unary coding / Binary guiding :
  - every x_t (context) promotes a unique expected label y_t (any other response is a miss)
  - The proposed label $\tilde{y}_t$
  - $f_t = \delta(\tilde{y}_t, y_t) \in \{0, 1\}$
- Methods :
  - "Banditron" [Kakade et al., 2008]
  - Contextual bandits

Linear classifiers

Similarity measure $\langle .,. \rangle$ on observation space

- Multiclass :
  - Task : character recognition, face recognition etc.
  - read $X_t = x_t \in \mathbb{R}^d$, find its category y_t
  - Labels : $1, \ldots, K$

$$\left.\begin{array}{l} W = (w_1, .., w_K) \in \mathbb{R}^{Kd} \\ X_t^k \triangleq (\vec{0}, \ldots, x_t, \ldots, \vec{0}) \in \mathbb{R}^{Kd} \end{array}\right\} \text{ so that : } \langle W, X_t^k \rangle = \langle w_k, x_t \rangle$$

- Scoring/ranking :
  - read $X_t = (x_t^1, \ldots, x_t^K) \in \mathbb{R}^{Kd}$, identify best match y to target

$$\left.\begin{array}{l} W = w \in \mathbb{R}^d \\ X_t^k \triangleq x_t^k \in \mathbb{R}^d \end{array}\right\} \text{ so that : } \langle W, X_t^k \rangle = \langle w, x_t^k \rangle$$

The Banditron [Kakade et al., 2008]

Inspired by the multiclass perceptron [Rosenblatt, 1958, Duda et al., 1973]

$\forall t \in 1, \ldots, T$:

1. Read X_t
2. Choose : $\hat{y}_t = \underset{k \in \{1,..,K\}}{\text{argmax}} \langle W_{t-1}, X_t^k \rangle$

   $\tilde{y}_t \sim (1-\varepsilon)\delta(k, \hat{y}_t) + \frac{\varepsilon}{K}$
3. Read $f_t = \delta(\tilde{y}_t, y_t)$
4. Update : $W_t = W_{t-1} + \frac{f_t \cdot X_t^{\tilde{y}_t}}{P(\tilde{Y}_t = \tilde{y}_t)} - X_t^{\hat{y}_t}$

Regret : $O(T^{2/3})$; Non-sparse.

Banditron alternatives

- Contextual bandits
- Choice based on :
  - Confidence intervals (UCB)[Lai and Robbins, 1985], LinUCB [Auer, 2002]
  - 2nd order Perceptron [Cesa-Bianchi et al., 2005]
  - Examples :
    - [Li et al., 2010]
    - [Hazan and Kale, 2011]
    - [Crammer and Gentile, 2013]
    - [Ngo et al., 2013]
- Regret $O(\sqrt{T})$ in the stationary case.
- The update relies on an estimator of $\mathbb{E}\left[{X_t^{\hat{y}_t}}^T X_t^{\hat{y}_t}\right]$:
  - $O(d^2)$ cost in space
  - Non sparse

Main question

- Binary feedback : $f_t \in \{0, 1\}$
  - "all-or-nothing" feedback
  - clic, like, visit, follow, retweet ...
- Unary coding ("one-hot")/ Binary guiding :
  - every x_t (context) promotes a unique expected label y_t (any other response is a miss)

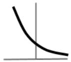
$\rightarrow$ The uniqueness of the expected label y_t needs to be expressed in a loss function.

Classical label-aware online learning loss functions

Observation : X ; expected label : y ;

- Logistic loss :

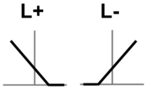
$$l = -\log \frac{\exp\langle W, X^y \rangle}{\sum_{k=1}^{K} \exp\langle W, X^k \rangle}$$



- Hinge loss :
  - Relative loss (Kessler) :

$$l = \left[1 - \langle W, X^y \rangle + \max_{k \neq y} \langle W, X^k \rangle \right]_+$$

  - Absolute (One-Versus-All) :



$$l = \sum_{k=1}^{K} \left[1 + (1 - 2\delta(y, k)) \langle W, X^k \rangle \right]_+$$

Policy gradient [Williams, 1992]

Model-free RL, POMDP, temporal credit assignment etc...

- Read X
- Stochastic policy : $\tilde{y} \sim \text{Multinomial}(\boldsymbol{\pi})$ with :

$$\pi(k) = \frac{\exp\langle W, X^k\rangle}{\sum_\ell \exp\langle W, X^\ell\rangle}$$

$$\pi(k) = (1-\varepsilon)\delta(k,\hat{y}) + \frac{\varepsilon}{K}$$

etc.

- Each response $\tilde{y} \sim \text{Multinomial}(\boldsymbol{\pi})$ provides a *reward* r

$\rightarrow$ find W^* that maximizes the reward expectation :

$$W^* = \text{argmax}_W \mathbb{E}(r)|_W$$

using gradient ascent :

$$W \leftarrow W + \eta \nabla_W \mathbb{E}(r)|_W$$

Policy gradient with binary guiding

$\forall t \in 1, ..., T$:

1. Read $\quad X_t$
2. Choose : $\quad \forall k, \pi(k|X_t) = \frac{\exp\langle W_{t-1}, X_t^k \rangle}{\sum_\ell \exp\langle W_{t-1}, X_t^\ell \rangle}$

   $\tilde{y}_t \sim \text{Multinomial}(\boldsymbol{\pi})$
3. Read $\quad f_t = \delta(\tilde{y}_t, y_t)$

   $r_t = f_t r^+ + (1 - f_t) r^-$
4. Update : $\quad W_t = W_{t-1} + \eta r_t \left(X_t^{\tilde{y}_t} - \sum_k \pi(k|X_t) X_t^k \right)$

Main result

Let :

- $\tilde{y}$ be the current response (random variable) ;
- y be the actual label associated with X
- $\boldsymbol{g}(X, \tilde{y})) = r_t \left(X^{\tilde{y}_t} - \sum_k \pi(k|X)X^k\right)$ be the policy gradient

Then :

$$\mathbb{E}_{\tilde{y}|\mathcal{X}=X}(\boldsymbol{g}(X, \tilde{y})) = \underbrace{(r^+ - r^-)\pi(y|X)}_{=1\,?} \underbrace{\left(X^y - \sum_k \pi(k|X)X^k\right)}_{\text{Logistic gradient}}$$

Multiclass PG with binary guiding recipe

$$\mathrm{cov}_{\mathcal{X},\mathcal{Y}}(\boldsymbol{g}(X,\tilde{y})) = \mathbb{E}_{\mathcal{X}}\left[\left(r^- + (1-\pi(y|X)(r^+ - r^-)\right)^2 \frac{\pi(y|X)}{(1-\pi(y|X))}\boldsymbol{g}(X)^T\boldsymbol{g}(X)\right]$$
$$+ {r^-}^2\mathbb{E}_{\mathcal{X}}\left[\tilde{\boldsymbol{\Sigma}}(X)\right] + (r^+ - r^-)^2\mathrm{cov}_{\mathcal{X}}\left[\frac{\pi(y|X)}{(1-\pi(y|X))}\boldsymbol{g}(X)\right]$$

Consider

- $a = r^+ - r^-$ the reward "amplitude"
- $b \in [0,1]$ the reward "baseline" : $\left|\frac{r^+}{r^-}\right| = \frac{1-b}{b}$

Then with $\pi^+ = \mathbb{E}_{\mathcal{X}}(\pi(y|X))$:

- Take $a = \frac{1}{\pi^+}$ (logistic gradient "speed")
- Take $b = \frac{\pi^+(1-\pi^+)}{1+\pi^+-\frac{2}{K}}$ (variance reduction)

Learning and forgetting : Regularized Policy gradient update

- Regularized optimization (λ hyperparameter) :

$$\max_{\boldsymbol{W}} \mathcal{H} = \max_{\boldsymbol{W}} E(r) - \frac{\lambda}{2}||\boldsymbol{W}||^2$$

- Regularized gradient ascent : $\nabla_W \mathcal{H} = E(r \nabla_W \ln \pi(\tilde{y}|X)) - \lambda W$
  - Gradient estimator (stochastic gradient) :

$$\langle r_t \nabla_W \ln \pi(\tilde{y}_t|X_t)\rangle_{1..T}$$

  - Online update (learning rate $\eta << 1$) :

$$\begin{aligned} W &\leftarrow W + \eta(r\nabla_W \ln \pi(\tilde{y}|X) - \lambda W) \\ &= (1-\eta\lambda)W + \eta r \nabla_W \ln \pi(\tilde{y}|X) \end{aligned}$$

- The old examples “fade away” as time passes → tracking algorithm and novelty detection (Kivinen et al, 2010)

Kernel extension

4. Update : $W_t(.) = (1 - \eta\lambda) W_{t-1}(.) + \eta r_t \left(\mathcal{K}(X_t^{\tilde{y}_t}, .) - \sum_k \pi(k|X_t) \mathcal{K}(X_t^k, .) \right)$

Or :

$$W_t(.) = \sum_{u=1}^{t} \sum_{k=1}^{K} \alpha_{k,u} \beta_{t-u} \mathcal{K}(X_t^k, .)$$

with :

$$\alpha_{k,u} = \eta r_u (\delta(\tilde{y}_u, k) - \pi(k|X_u; W_u))$$

and

$$\beta_v = (1 - \eta\lambda)^v$$

1. Introduction
2. Contributions
   - Daucé, E, Proix, T, Ralaivola, L ; proc. of IJCNN 2015
   - Zhong, H and Daucé, E, hal-01345825, submitted
3. Numerical experiments
   - Zhong, H and Daucé, E, hal-01345825, submitted
   - Daucé, E, Proix, T, Ralaivola, L ; proc. of IJCNN 2015
4. Conclusion

Online empirical risk minimization

- "Passive-agressive" approach [Crammer et al., 2006]
- "Kessler" Hinge loss : $l_t = \left[1 - \langle W_{t-1}, X^{y}{}_t \rangle + \max_{k \neq y} \langle W_{t-1}, X_t^k \rangle \right]_+$
- Local quadratic optimization : $\forall t$, solve :

$$W_t = \arg\min_W \frac{1}{2} \|W - W_{t-1}\|^2 + C\xi^2 \text{ s.t. } l_t \leq \xi$$

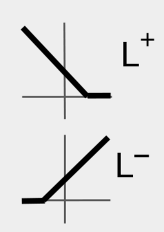
- Update :

$$W_t = W_{t-1} + \frac{l_t}{2\|X_t\|^2 + \frac{1}{2C}} (X_t^{y_t} - \max_{k \neq y_t} X_t^k)$$

- In the linearly separable case :

$$\sum_{t=1}^{T} l_t^2 \leqslant 4R^2 \parallel U \parallel^2$$

Hinge loss : "Bandit" reduction



- Reduction (sample) of the OVA loss :

$$l_t = \left[1 + (1 - 2\delta(y_t, \tilde{y}_t))\langle W_{t-1}, X_t^{\tilde{y}_t}\rangle\right]_+$$

- Online empirical risk minimization :

$$W_t = W_{t-1} + \frac{l_t}{\|X_t\|^2 + \frac{1}{2C}}(2\delta(y_t, \tilde{y}_t) - 1)X_t^{\tilde{y}_t}$$

- Aggressiveness / conservatism ($C \to \infty$) :
  - "one-shot" update
  - label-error sensitivity

Bandit "Passive-Aggressive" (BPA)

$\forall t \in 1, ..., T$:

1. Read $\quad x_t$
2. Choose : $\quad \hat{y}_t = \underset{k \in \{1,..,K\}}{\text{argmax}} \langle W_{t-1}, X_t^k \rangle$

   $\tilde{y}_t \sim (1-\varepsilon)\delta(k, \hat{y}_t) + \frac{\varepsilon}{K}$
3. Read $\quad f_t = \delta(\tilde{y}_t, y_t)$
4. Update : $\quad W_t = W_{t-1} + \frac{l_t}{\|X_t\|^2 + \frac{1}{2C}}(2f_t - 1) \cdot X_t^{\tilde{y}_t}$

Linearly separable case

Theorem

Let $(x_1, y_1), ..., (x_T, y_T)$ be a sequence of separable examples where $x_t \in \mathbb{R}^d$, $y_t \in \{1, ..., K\}$ and $\| x_t \| \leqslant R$ for all t, let $\tilde{y}_1, ..., \tilde{y}_T$ be a sequence of responses, with $\tilde{y}_t \in \{1, ..., K\}$, and let $U \in \mathbb{R}^{Kd}$ be such that $\forall t, l_t^* = 0$. Then, assuming $C \to \infty$, the cumulative squared loss of BPA is bounded by :

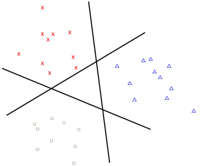
$$\sum_{t=1}^{T} l_t^2 \leqslant R^2 \| U \|^2 \quad (1)$$

Warning ! !

The squared loss sum is not here an upper bound of the number of classification errors.

Additional conditions

Consider the greedy choice $\tilde{y}_t = \hat{y}_t = \underset{k \in \{1,..,K\}}{\text{argmax}} \langle W_{t-1}, X_t^k \rangle$:



- if all X^k's k belong to a convex set $\mathcal{C}_k \subset \mathbb{R}^d$
- If $\exists t^*$ so that $\forall t \geq t^*, l_t = 0$
- If $\exists t \geq t^*$ so that $\tilde{y}_t = y_t = k$

then *all* examples from C_k are correctly classified for $t \geq t^*$.

Moreover, this can be assumed almost surely if :

$$\tilde{y}_t \sim (1-\varepsilon)\delta(k, \hat{y}_t) + \frac{\varepsilon}{K}$$

(ε-greedy)

Non-separable stationary case

Theorem

Let $(x_1, y_1), \ldots, (x_T, y_T)$ *be a sequence of examples where* $x_t \in \mathbb{R}^d$, $y_t \in \{1, \ldots, K\}$ *and* $\| x_t \| \leqslant R$ *for all t, let* $\tilde{y}_1, \ldots, \tilde{y}_T$ *be a sequence of responses, with* $\tilde{y}_t \in \{1, \ldots, K\}$. *Then for any* $U \in \mathbb{R}^{Kd}$, *and assuming* $C \to \infty$, *the cumulative squared loss of BPA is bounded by :*

$$\sum_{t=1}^{T} l_t^2 \leqslant \left(R \| U \| + 2 \sqrt{\sum_{t=1}^{T} (l_t^*)^2} \right)^2$$

- For large T, the loss is at worst twice of the optimal loss
- Needs a finite C to reach an $O(\sqrt{T})$ regret

Kernel extension

4. Update : $W_t(.) = W_{t-1}(.) + \frac{l_t}{\mathcal{K}(X_t, X_t) + \frac{1}{2C}}(2f_t - 1) \cdot \mathcal{K}(X_t^{\tilde{y}_t}, .)$

Or :

$$W_t(.) = \sum_{u=1}^{t} \alpha_u \mathcal{K}(X_t^{\tilde{y}_u}, .)$$

with :

$$\alpha_u = \frac{l_u}{\mathcal{K}(X_u, X_u) + \frac{1}{2C}}(2f_u - 1)$$

1. Introduction
2. Contributions
   - Daucé, E, Proix, T, Ralaivola, L ; proc. of IJCNN 2015
   - Zhong, H and Daucé, E, hal-01345825, submitted
3. Numerical experiments
   - Zhong, H and Daucé, E, hal-01345825, submitted
   - Daucé, E, Proix, T, Ralaivola, L ; proc. of IJCNN 2015
4. Conclusion

Datasets

Table: Five datasets considered, with n the number of instances, d the vectors dimension and K the number of labels.

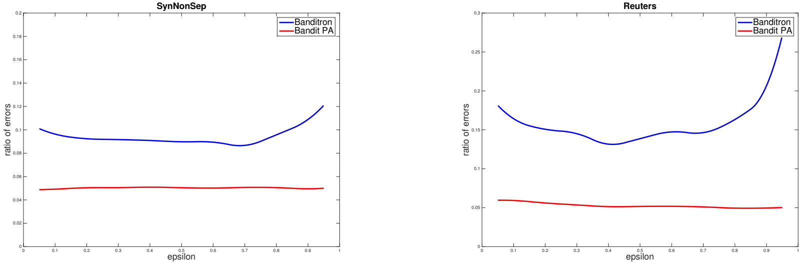
Dataset	n	d	K
SynSep	10^5	400	9
SynNonSep	10^5	400	9
RCV1-v2	10^5	47236	53
Segment	2310	19	7
Pendigits	7494	16	10

Parameters

Table: Parameters setting for different algorithms and different datasets.

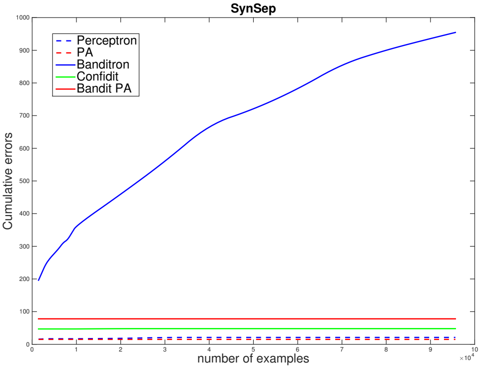
Dataset	P	PA	B	C	BPA
Synsep	$\varnothing$	$C \to \infty$	$\varepsilon = 0.014$	$\eta = 10^3$	$\varepsilon = 0.4$ $C \to \infty$
SynNonSep	$\varnothing$	$C = 10^{-2}$	$\varepsilon = 0.65$	$\eta = 10^3$	$\varepsilon = 0.8$ $C = 10^{-2}$
RCV1-v2	$\varnothing$	$C = 10^{-2}$	$\varepsilon = 0.4$	$\eta = 10^2$	$\varepsilon = 0.2$ $C = 10^{-2}$
	K-B	**BPA**	**K-BPA**	**K-SGD**	
Segment	$\sigma = 1$ $\varepsilon = 0.1$	$\varepsilon = 0.3$	$\sigma = 1$ $\varepsilon = 0.3$	$\sigma = 1$ $H = 200$	
Pendigits	$\sigma = 10$ $\varepsilon = 0.1$	$\varepsilon = 0.3$	$\sigma = 10$ $\varepsilon = 0.3$	$\sigma = 10$ $H = 500$	

Exploration rate



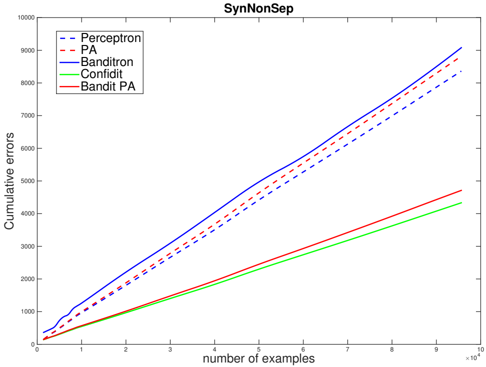
- Final error rate in function of ε
- SynNonSep (left) and Reuters (right) datasets.

SynSep Cumulative errors



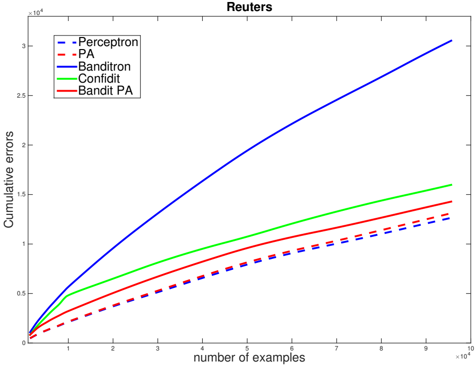
- Perceptron, PA, Banditron, Confidit and BPA
- 9 classes, $d = 400$

SynNonSep Cumulative errors



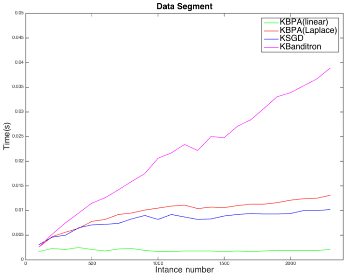
- Perceptron, PA, Banditron, Confidit and BPA
- 9 classes, $d = 400$

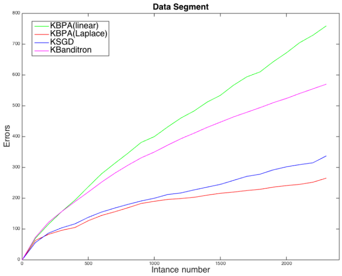
Reuters Cumulative errors



- Perceptron, PA, Banditron, Confidit and BPA
- 53 classes, $d = 47236$

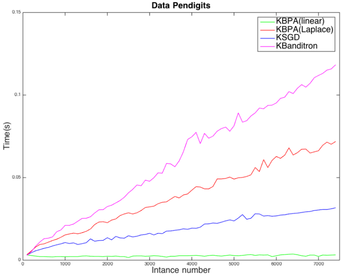
Segment (with Kernels)

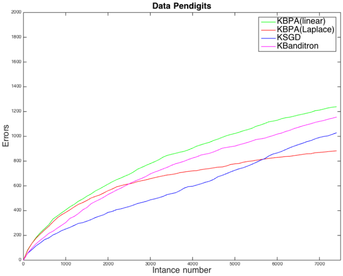




- BPA, K-BPA, K-SGD, K-Banditron
- 7 classes, $d = 19$

Pendigits (with Kernels)



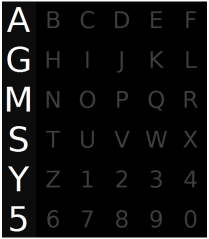


- BPA, K-BPA, K-SGD, K-Banditron
- 10 classes, $d = 16$

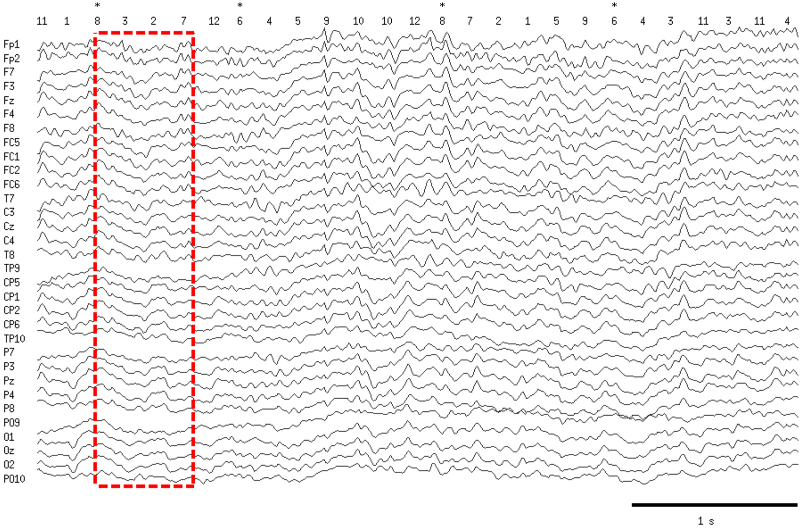
1. Introduction
2. Contributions
   - Daucé, E, Proix, T, Ralaivola, L ; proc. of IJCNN 2015
   - Zhong, H and Daucé, E, hal-01345825, submitted
3. Numerical experiments
   - Zhong, H and Daucé, E, hal-01345825, submitted
   - Daucé, E, Proix, T, Ralaivola, L ; proc. of IJCNN 2015
4. Conclusion

P300 speller

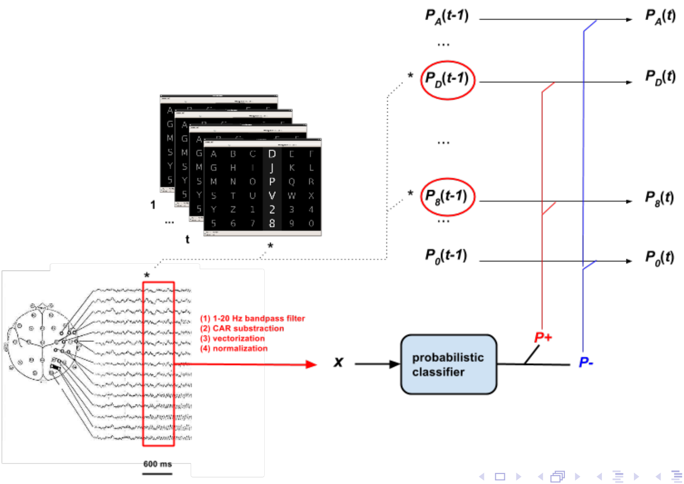
- EEG :
  - 10 - 60 channels (surface electric potential - H Berger, 1929)
  - high temporal resolution / low spatial resolution
  - noisy, non-reliable,... "Evoqued potentials" technique
  - the "P300" ERP is "surprise" effect ("oddball" paradigm)
- P300-speller (Farwell and Donchin, 1988) :
  - based on the "oddball" paradigm
  - 6 x 6 letters grid
  - random row/column magnification (every 150-300 ms)
  - row/column evidence build-up + argmax choice
  - low SNR / low bit rate (many flashes for one letter)
  - spelling accuracy tends to decrease in the long run



EEG data (from Inserm U1028, Lyon, France)

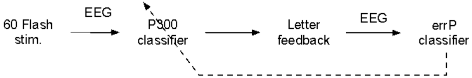


Data processing pipeline



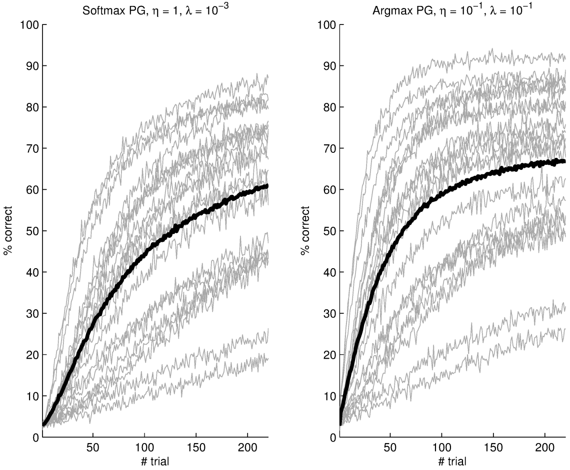
Rewards in classifiction

- Online stochastic classifier :
  - read input observations set : $X = (\boldsymbol{x}_1, ..., \boldsymbol{x}_K)$
  - give a score to every class : $\forall k, \pi(k|X;W) = \frac{\exp\langle W, X^k \rangle}{\sum_\ell \exp\langle W, X^\ell \rangle}$
  - choose the response at random (Softmax choice)
  - read the *reward* r
  - update W
- Which reward ?
  - "error" potential after the classifier's response :

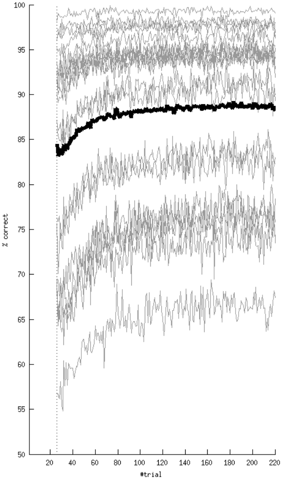


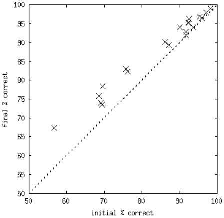
  - "BACKSPACE" key on the virtual keyboard

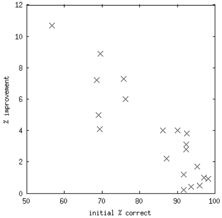
Softmax/Argmax spelling improvement



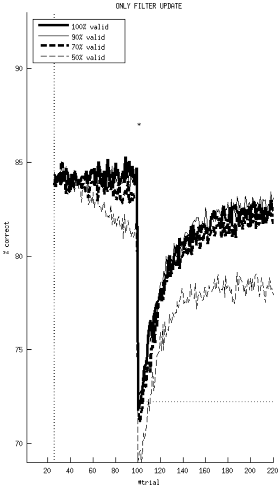
Classification improvement after a 25-trials training session



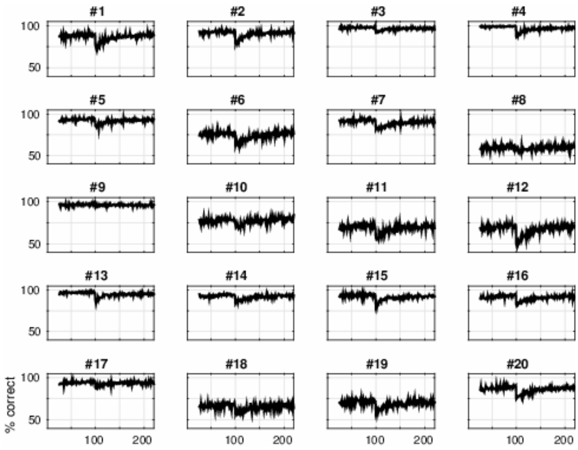




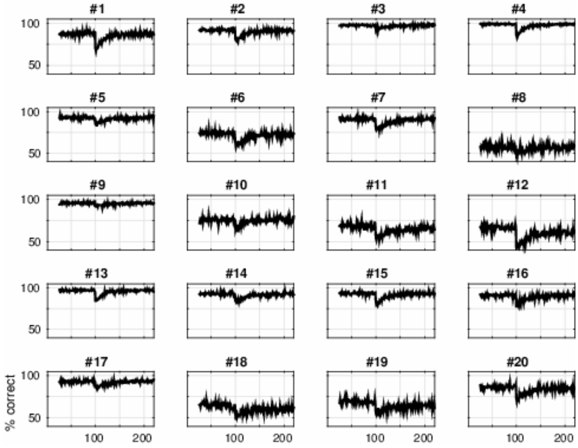
Global recovery after electrode break



Individual recovery, label noise = 10%



Individual recovery, label noise = 30%



1. Introduction
2. Contributions
   - Daucé, E, Proix, T, Ralaivola, L ; proc. of IJCNN 2015
   - Zhong, H and Daucé, E, hal-01345825, submitted
3. Numerical experiments
   - Zhong, H and Daucé, E, hal-01345825, submitted
   - Daucé, E, Proix, T, Ralaivola, L ; proc. of IJCNN 2015
4. Conclusion

Conclusion

Policy gradient

Pros

- Cheap (linear cost)
- Neural networks / Multinomial generative friendly (logistic gradient)
- Label noise resistance
- 2nd order expandable

Cons

- Non-sparse
- η, λ parameter fitting

OVA online quadratic optimization

Pros

- Cheap (linear cost)
- Sparse
- Upper bound when linearly separable

Cons

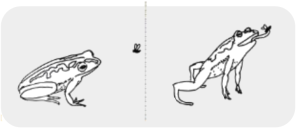
- Aggressive update $\rightarrow$ label noise sensitivity
- Needs an optimal "stiffness" C hyperparameter (cross-validated)
- Non deterministic : needs an ε (possibly decreasing)

Open questions

- Unary coding + binary guiding :
  - a more structured/constrained bandit problem
  - multiclass gradient, multiclass bounds
- Adversarial case :
  - the more robust, the less sparse ?
  - learning and forgetting (tracking)

More "challenging" open questions

- Learning in :
  - embedded controllers
  - real time
  - many decisions in limited time/limited resources
  - non-stationary environments
- Binary guiding in nature :



  - Label = actions ?
  - Many actions = many labels
  - Complex motor realization space (many DOFs)
  - All or nothing

Auer, P. (2002).
Using confidence bounds for exploitation-exploration trade-offs.
Journal of Machine Learning Research, 3(Nov) :397–422.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002).
Finite-time analysis of the multiarmed bandit problem.
Machine learning, 47(2-3) :235–256.

Cesa-Bianchi, N., Conconi, A., and Gentile, C. (2005).
A second-order perceptron algorithm.
SIAM Journal on Computing, 34(3) :640–668.

Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., and Singer, Y. (2006).
Online passive-aggressive algorithms.
The Journal of Machine Learning Research, 7 :551–585.

Crammer, K. and Gentile, C. (2013).
Multiclass classification with bandit feedback using adaptive regularization.
Machine learning, 90(3) :347–383.

Duda, R. O., Hart, P. E., et al. (1973).
Pattern classification and scene analysis, volume 3.
Wiley New York.

Hazan, E. and Kale, S. (2011).
Newtron : an efficient bandit algorithm for online multiclass prediction.
In *Advances in Neural Information Processing Systems*, pages 891–899.

Kakade, S. M., Shalev-Shwartz, S., and Tewari, A. (2008).
Efficient bandit algorithms for online multiclass prediction.

In *Proceedings of the 25th international conference on Machine learning*, pages 440–447. ACM.

Kivinen, J., Smola, A. J., and Williamson, R. C. (2004).
Online learning with kernels.
Signal Processing, IEEE Transactions on, 52(8) :2165–2176.

Lai, T. L. and Robbins, H. (1985).
Asymptotically efficient adaptive allocation rules.
Advances in applied mathematics, 6(1) :4–22.

Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010).
A contextual-bandit approach to personalized news article recommendation.
In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM.

Ngo, H. Q., Luciw, M. D., Vien, N. A., and Schmidhuber, J. (2013).

Upper confidence weighted learning for efficient exploration in multiclass prediction with binary feedback.
In *IJCAI*.

Rosenblatt, F. (1958).
The perceptron : a probabilistic model for information storage and organization in the brain.
Psychological review, 65(6) :386.

Sutton, R. S. and Barto, A. G. (1998).
Reinforcement learning : An introduction, volume 1.
MIT press Cambridge.

Williams, R. (1992).
Simple statistical gradient following algorithms for connectionnist reinforcement learning.
Machine Learning, 8 :229–256.