Robust and adaptive online learning with BOA algorithm

Olivier Wintenberger, olivier.wintenberger@upmc.fr

February 26, 2019

Data and Analytics for Short-Term Operations, INI, Cambridge

The online learning setting

Online learning in a stochastic environment

From Gerchinovitz (2013): observe random $(Y_t, X_t)_{t \geqslant 1}$ recursively (online).

Aim of Online Learning

Predict $Y_{t+1} \in \mathbb{R}$ given $X_{t+1} \in \mathcal{X}$ thanks to $\widehat{f}_t(X_{t+1})$ with a learner

$$\widehat{f}_t : \qquad \mathcal{X} \to \mathbb{R} \qquad \text{depending only on} \qquad (Y_s, X_s)_{1 \leqslant s \leqslant t}.$$

Remarks

- Both processes $(\widehat{f}_t)$ and (Y_t, X_t) are adapted to the natural filtration $(\mathcal{F}_t)$ with $\mathcal{F}_t = \sigma((Y_s, X_s)_{1 \leqslant s \leqslant t})$.
- Adversarial, iid, auto-regressive, ... are included in that general setting.

Definition (Experts)

A dictionary $\{f_1, \dots f_M\}$ of experts $f_j = (f_{j,t})$ with $f_{j,t} : \mathcal{X} \to \mathbb{R}$ is given, $f_t = (f_{j,t})_{1 \leqslant j \leqslant M}$ are learners adapted to $(\mathcal{F}_t)$.

Examples: deterministic experts, outputs of sequential statistical algorithms (Kalman filters, OGD, ONS,..., Hazan, 2015). In all cases they are black boxes for us.

Aggregation $\widehat{f}_t = \sum_{j=1}^M \pi_{j,t} f_{j,t} = \mathbb{E}_{\pi_t}[f_t]$ for some weights $\pi_{j,t} \geqslant 0$, $\sum_{j=1}^M \pi_{j,t} = 1$ in the simplex Λ_M.

Weights $(\pi_{j,t})$ are adapted to the filtration $\mathcal{F}_t$. Both the experts f_t and their aggregation $\widehat{f}_t$ are learners.

Aim $\Rightarrow$ Solution

Find an online aggregation procedure that is adaptive and robust in that general framework

$\Rightarrow$ Bernstein Online Aggregation (BOA) algorithm provides a reasonable answer.

The regret bound and the predictive risk

Regret bound

Let $\ell(y, x)$ be a convex loss in x. Consider the adversarial setting with expert advice.

Definition

The regret in the expert aggregation setting is defined as

$$Regret_T = \sum_{t=1}^{T} \ell(Y_t, \widehat{f}_{t-1}(X_t)) - \min_{1 \leqslant j \leqslant M} \sum_{t=1}^{T} \ell(Y_t, f_{j,t-1}(X_t)).$$

Remark

The regret is non necessarily positive.

Regularity of the loss

Optimal rate of convergence in $\sqrt{T}$ achieved by OGD algorithm, Zinkevich (2003).

Definition (Exp-concavity)

The loss ℓ is $\delta > 0$ exp-concave when

$$\ell(z, x) - \ell(z, y) \geqslant \ell'(z, y)(x - y) + \delta(\ell'(z, y)(x - y))^2, \qquad x, y, z \in \mathcal{K}\,.$$

Theorem (Vovk, 1998)

When ℓ is exp-concave then the optimal fast rates are achieved by the Exponentially Weighted Averaging with $\text{Regret}_T = \mathcal{O}(\log M)$.

Definition (EWA)

EWA algorithm computes weights, with fixed learning rate η, as

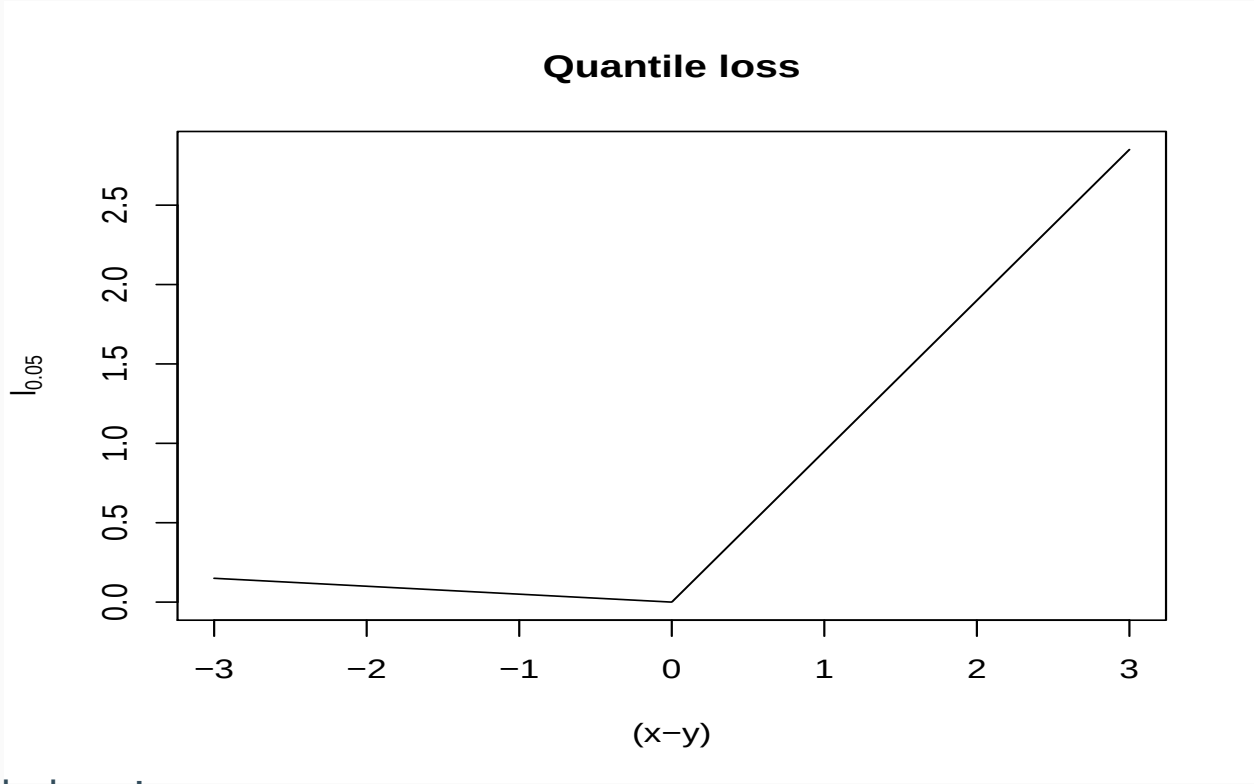
$$\pi_{j,t} \propto \exp(-\eta \ell(Y_t, f_{j,t}(X_t)))\pi_{j,t-1}.$$

Quantile or pinball loss

Definition (Koenker, 2005)

The quantile loss of rate $\tau \in (0, 1)$ is defined as

$$\ell_\tau(x, y) = \begin{cases} \tau(x - y), & x - y > 0, \\ -(1 - \tau)(x - y), & \text{else.} \end{cases}$$



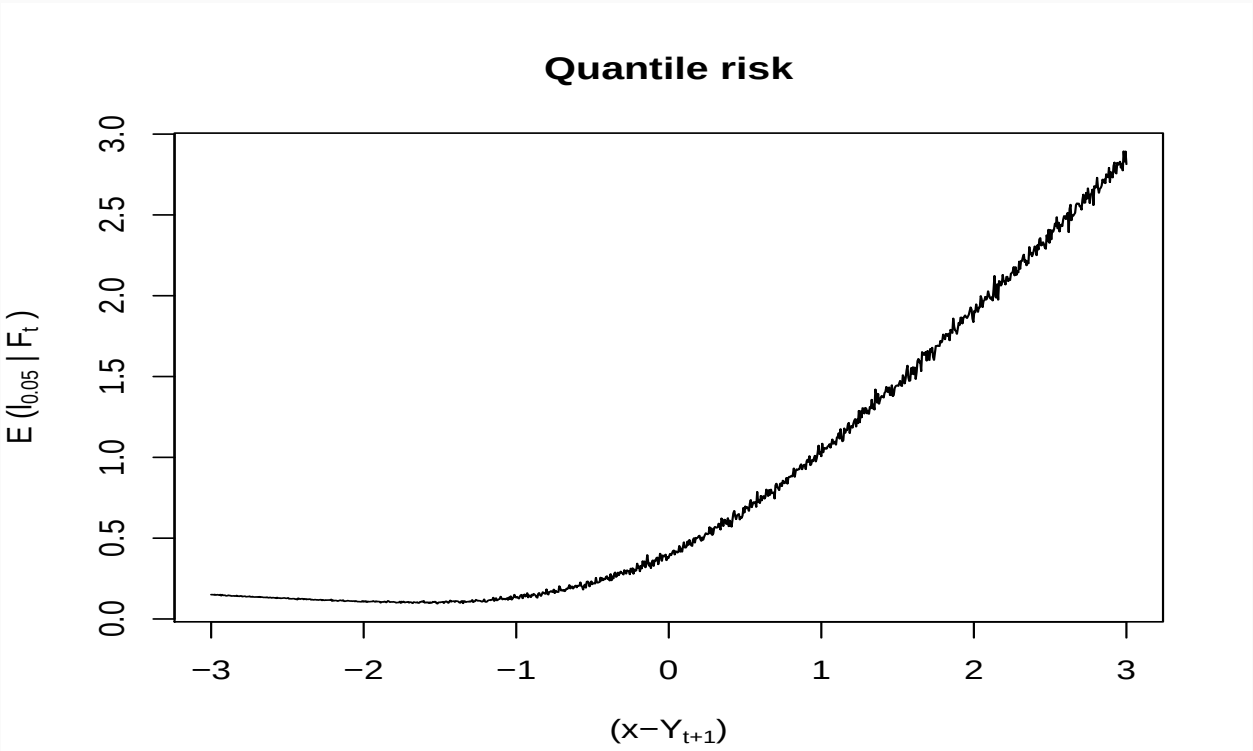
The quantile loss is convex.

The conditional quantile

Lemma (Koenker, 2005, Biau & Patra, 2011)

The conditional quantile satisfies

$$F^{-1}_{Y_t|\mathcal{F}_{t-1}}(\tau) = \arg\min_{q\in\mathbb{R}} \mathbb{E}[\ell_\tau(Y_t, q) \mid \mathcal{F}_{t-1}], \qquad a.s.$$



The quantile predictive risk is exp-concave and even strongly convex. The constant of exp-concavity depends on the conditional distribution $Y_t \mid \mathcal{F}_{t-1}$.

The quantile quantile risk as objective

Minimizing the regret, hopefully

$$\min_{1 \leqslant j \leqslant M} \sum_{t=1}^{T} \ell_\tau(Y_t, f_{j,t-1}(X_t)) \approx 0\,.$$

For $T = 1$, the minimizer may achieve $f_{j^*,0}(X_1) \approx Y_1$ independent of τ!

Definition

The predictive risk at time t of any learner f is defined as

$$\mathbb{E}[\ell_\tau(Y_t, f(X_t)) \mid \mathcal{F}_{t-1}]$$

What we gain:

- optimum close to our objective the conditional quantile,
- regularization of the objective function.

The cumulative predictive risk

Definition (Cumulative predictive risk, W., 2017)

In the general setting, for any loss ℓ and any sequential learners $\widehat{f} = (\widehat{f}_t)$ we define

$$R(\widehat{f}) = \sum_{t=1}^{T} \mathbb{E}[\ell(Y_t, \widehat{f}_{t-1}(X_t)) \mid \mathcal{F}_{t-1}] - \min_{1 \leqslant j \leqslant M} \sum_{t=1}^{T} \mathbb{E}[\ell(Y_t, f_{j,t-1}(X_t)) \mid \mathcal{F}_{t-1}].$$

Remark

In the iid setting, we consider $\bar{f} = T^{-1} \sum_{t=1}^{T} \widehat{f}_{t-1}$ and by convexity

$$R(\bar{f}) = T\mathbb{E}[\ell(Y, \bar{f}(X)) \mid \bar{f}] - \min_{1 \leqslant j \leqslant M} \sum_{t=1}^{T} \mathbb{E}[\ell(Y, f_{j,t-1}(X))] \leqslant R(\widehat{f})$$

for some copy (Y, X).

For constants experts $f_j = f_{j,t}$,$t \geqslant 1$, we recover usual oracle bound from the batch setting.

Back to the adversarial setting considering the Dirac masses as conditional probabilities

Quantitative bounds and BOA

Fast regret bounds

Theorem (Fast regret bound, Cesa-Bianchi & Lugosi, 2005)

If ℓ is δ exp-concave then EWA for a well chosen η satisfies,

$$Regret_T \leqslant C \log(M).$$

Remark

The quantile loss is not exp-concave.

It does not exist any aggregation procedure with less than a $\mathcal{O}(\sqrt{T})$ for such loss.

And for the cumulative predictive risk?

Fast rate in expectation

EWA satisfies, for a well chosen η, the cumulative risk bound in expectation

$$\mathbb{E}[R(\widehat{f})] \leqslant C \log(M).$$

Theorem (Slow rate in probability, Audibert, 2007)

For any learning rate, in the iid setting, EWA satisfies

$$\mathbb{P}\left(R(\bar{f}) \leqslant C\sqrt{T(\log(M)+x)}\right) \geqslant 1 - e^{-x}, \qquad x > 0,$$

and the rate cannot be improved.

We say that EWA is not robust in the stochastic setting.

Robustness

Open question from Audibert (2009)

Can we modify EWA to obtain a robust and adaptive procedure in the stochastic setting?

Fast rate in probability

The aim is to build an algorithm so that for strongly convex predictive risk $\mathbb{E}[\ell(Y_t, \cdot) \mid \mathcal{F}_{t-1}]$ we have

$$\mathbb{P}\Big(R(\bar{f}) \leqslant C(\log(M) + x)\Big) \geqslant 1 - e^{-x}, \qquad x > 0.$$

Usually the strong convexity constant depends on the conditional distribution $Y_t \mid \mathcal{F}_{t-1}$ and is unknown. The algorithm may not depend on it.

Regret $\Rightarrow$ Cumulative risk in probability

Key idea: online to batch conversion

The sequence of differences $(M_T = R(\widehat{f}) - Regret_T)$ constitutes a martingale (Zhang, 2005, Kakade & Tewari, 2008, Audibert, 2009).

Theorem (Empirical Bernstein inequalities for martingales, W. 2017)

For a martingale (M_t) such that $\Delta M_t \geqslant -1/2$ with quadratic variation $[M]_n = \sum_{t=1}^n \Delta M_t^2$ then for any $n \geqslant 1$

$$\mathbb{P}(M_n \leqslant [M]_n + x) \geqslant 1 - e^{-x}, \qquad x > 0.$$

New adversarial to stochastic conversion

Applied to $(R(\widehat{f}) - Regret_T)$ we control the adversarial to stochastic conversion with an observable second order term

$$\sum_{t=1}^{T} (\ell(Y_t, \widehat{f}_{t-1}(X_t)) - \ell(Y_t, f_{j,t-1}(X_t)))^2.$$

Bernstein Online Aggregation (BOA)

Idea: Include the second order term to regularize EWA

Modify the EWA by adding a second order term correction.

Definition (Bernstein Online Aggregation algorithm)

The Bernstein Online Aggregation procedure defines recursively weights as

$$\pi_{j,t} \propto \exp(-\eta \ell(Y_t, \widehat{f}_{j,t-1}(X_t)) - \eta^2(\ell(Y_t, \widehat{f}_{j,t-1}(X_t)) - \mathbb{E}_{\pi_t}[\ell(Y_t, \widehat{f}_{t-1}(X_t)])^2)\pi_{j,t-1}$$

with some learning rate $\eta > 0$.

Theorem (Fast rate, exp-concave losses)

For exp-concave losses and for $\eta > 0$ well chosen, then BOA achieves fast rates in probability for well chosen learning rate.

Problem: the quantile loss ℓ_τ is not exp-concave....

Second order regret bound for BOA

Idea: the gradient trick

Use the linearized loss $\tilde{\ell}_t(f) = \ell'(Y_t, \widehat{f}_{t-1}(X_t))(\widehat{f}_{t-1}(X_t) - f(X_t))$ for any learner f.

Theorem (Second order regret bound, W. 2017)

For $\eta > 0$ well chosen, as $\eta\tilde{\ell}_t(f_j)$ is a centered r.v. under the distribution π_{t-1}, we have

$$\tilde{Regret}_T(f_j) \leqslant 2\sqrt{\sum_{t=1}^{T}(\ell'(Y_t, \widehat{f}_{t-1}(X_t))(f_j(X_t) - \widehat{f}_{t-1}(X_t))^2 \log(M)}\,,$$

where $\tilde{Regret}_T(f_j) = \ell'(Y_t, \widehat{f}_{t-1}(X_t))(\widehat{f}_{t-1}(X_t) - f_j(X_t))$.

Remark

- $Regret_T \leqslant \max_{1\leqslant j\leqslant M} \tilde{Regret}_T(f_j)$ by convexity,
- Second order regret bound similar than for MLProd in Gaillard *et al.* (2014).

Oracle inequality in probability for BOA

We apply the new adversarial to stochastic conversion

Theorem (Second order cumulative risk bound, W. 2017)

For η well chosen we obtain, w. p. $1 - e^{-x}$,

$$R(\widehat{f}) \leqslant \max_{1 \leqslant j \leqslant M} 2\sqrt{\sum_{t=1}^{T} (\ell'(Y_t, \widehat{f}_{t-1}(X_t))(f(X_t) - f_{t-1}(X_t))^2 (\log(M) + x)}.$$

No assumption on the stochastic environment (Y_t, X_t).

Optimality of the second order term

For ℓ a loss satisfying the following Bernstein condition, Koolen et al. (2015),

$$\mathbb{E}[\ell(Y_t, x) - \ell(Y_t, y) \mid \mathcal{F}_{t-1}] \geqslant \mathbb{E}\Big[\ell'(Y_t, y)(x-y) \mid \mathcal{F}_{t-1}\Big] + \alpha\, \mathbb{E}\Big[\big(\ell'(Y_t, y)(x-y)\big)^2 \mid \mathcal{F}_{t-1}\Big], \quad a.s.$$

Theorem (Fast rate in probability, Gaillard and W., 2018)

When the loss satisfies the Bernstein condition and $\eta > 0$ is well chosen, then

$$\mathbb{P}\Big(R(\widehat{f}) \leqslant R(f_j) + C(\log(M) + x)\Big) \geqslant 1 - e^{-x}.$$

Remarks

- Under the Bernstein condition then $\mathbb{E}[\ell(Y_t, \cdot) \mid \mathcal{F}_{t-1}]$ is δ exp-concave with $\delta \geqslant \alpha$,
- The quantile predictive risk is strongly convex and satisfies the Bernstein condition with α depending on the conditional distribution of Y_t.

Adaptivity of BOA

Conditions on the learning rate

From the empirical Bernstein inequality we assumed $\Delta M_t = \eta \tilde{\ell}_t(f_j) \geqslant -1/2$ but the linearized losses are difficult to control.

Idea: multiple learning rates

The result extends to the difference of martingale $\Delta M_t = \eta_j \tilde{\ell}_t(f_j) \geqslant -1/2$ for multiple learning rates as in Gaillard et al., 2014.

The requirement ΔM_t centered under π_{t-1} is not satisfied anymore....

Lemma (Centering under tilted weights, W. 2017)

The difference of martingale $\Delta M_t = \eta_j \tilde{\ell}_t(f_j)$ is centered under the BOA weights

$$\pi_{j,t} \propto \eta_j \exp(-\eta_j \ell(Y_t, \widehat{f}_{j,t-1}(X_t)) - \eta_j^2 (\ell(Y_t, \widehat{f}_{j,t-1}(X_t)) - \mathbb{E}_{\pi_t}[\ell(Y_t, \widehat{f}_{t-1}(X_t)])^2) \pi_{j,t-1}.$$

Adaptive learning rates for BOA

Using the doubling trick in Cesa-Bianchi et al. (2007), we define the adaptive BOA procedure as

Initialization: Set $L_{j,0} = 0$, $\eta_{j,0} = 0$, $\pi_{j,0} = M^{-1}$.

For: each time round $t \geqslant 1$,

- Compute recursively

$$L_{j,t} = L_{j,t-1} + \tilde{\ell}_t(f_j)(1 + \eta_{j,t-1}\tilde{\ell}_t(f_j)),$$

- Estimate the ranges $E_{j,t} = 2^{k+1}$ where k is the smallest integer such that $\max_{1 \leqslant s \leqslant t} |\ell_{j,t}| \leqslant 2^k$, $1 \leqslant j \leqslant M$,
- Compute the adaptive learning rate

$$\eta_{j,t} = \min\left\{\frac{1}{E_{j,t}}, \sqrt{\frac{\log M}{\sum_{s=1}^{t} \tilde{\ell}_s(f_j)^2}}\right\}, \qquad 1 \leqslant j \leqslant M,$$

- Compute the weights vector $\tilde{\pi}_t = (\tilde{\pi}_{j,t})_{1 \leqslant j \leqslant M}$:

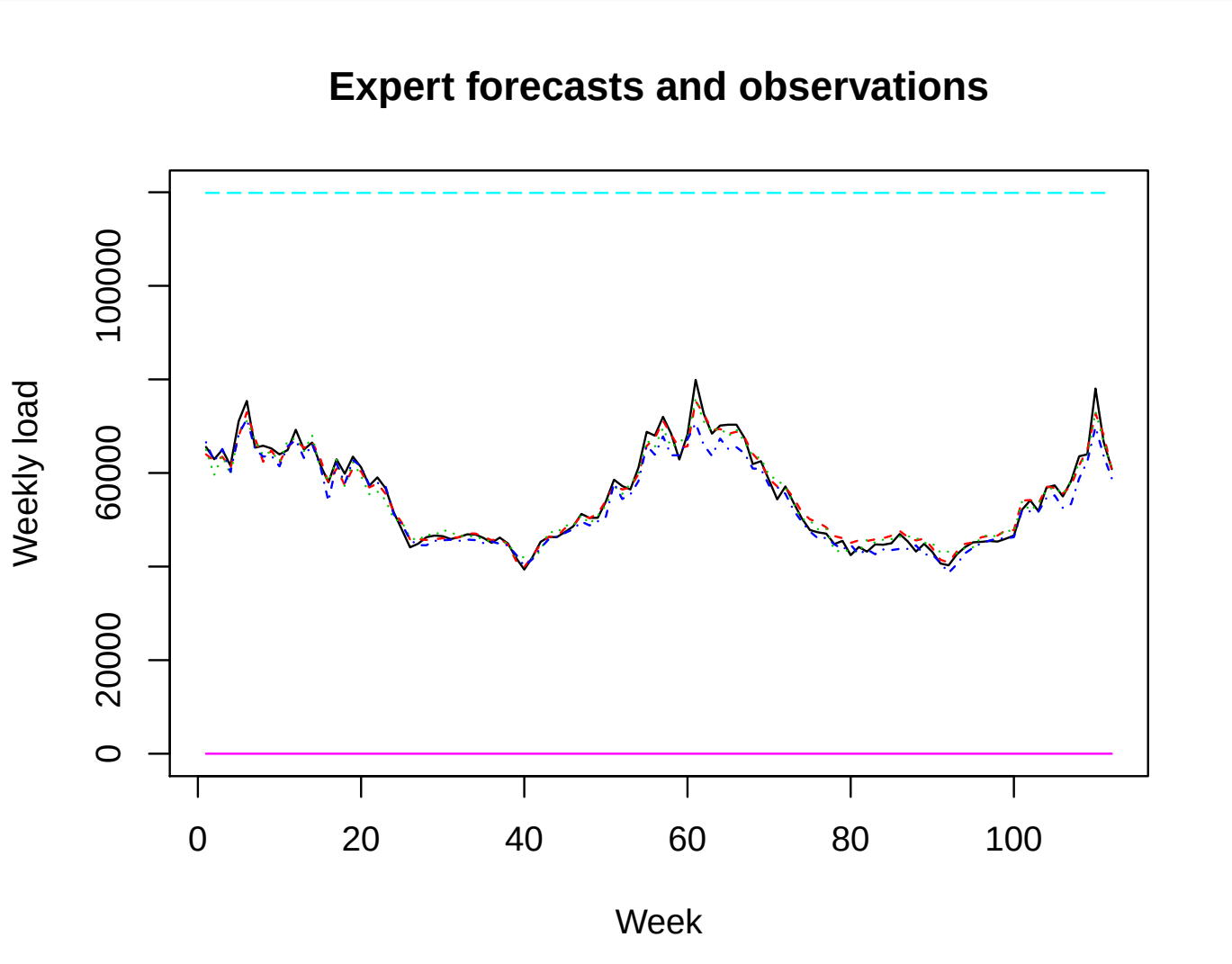
$$\pi_{j,t} = \frac{\eta_{j,t} \exp(-\eta_{j,t} L_{j,t})}{\sum_{j=1}^{M} \eta_{j,t} \exp(-\eta_{j,t} L_{j,t})}.$$

Some application

Illustrative example

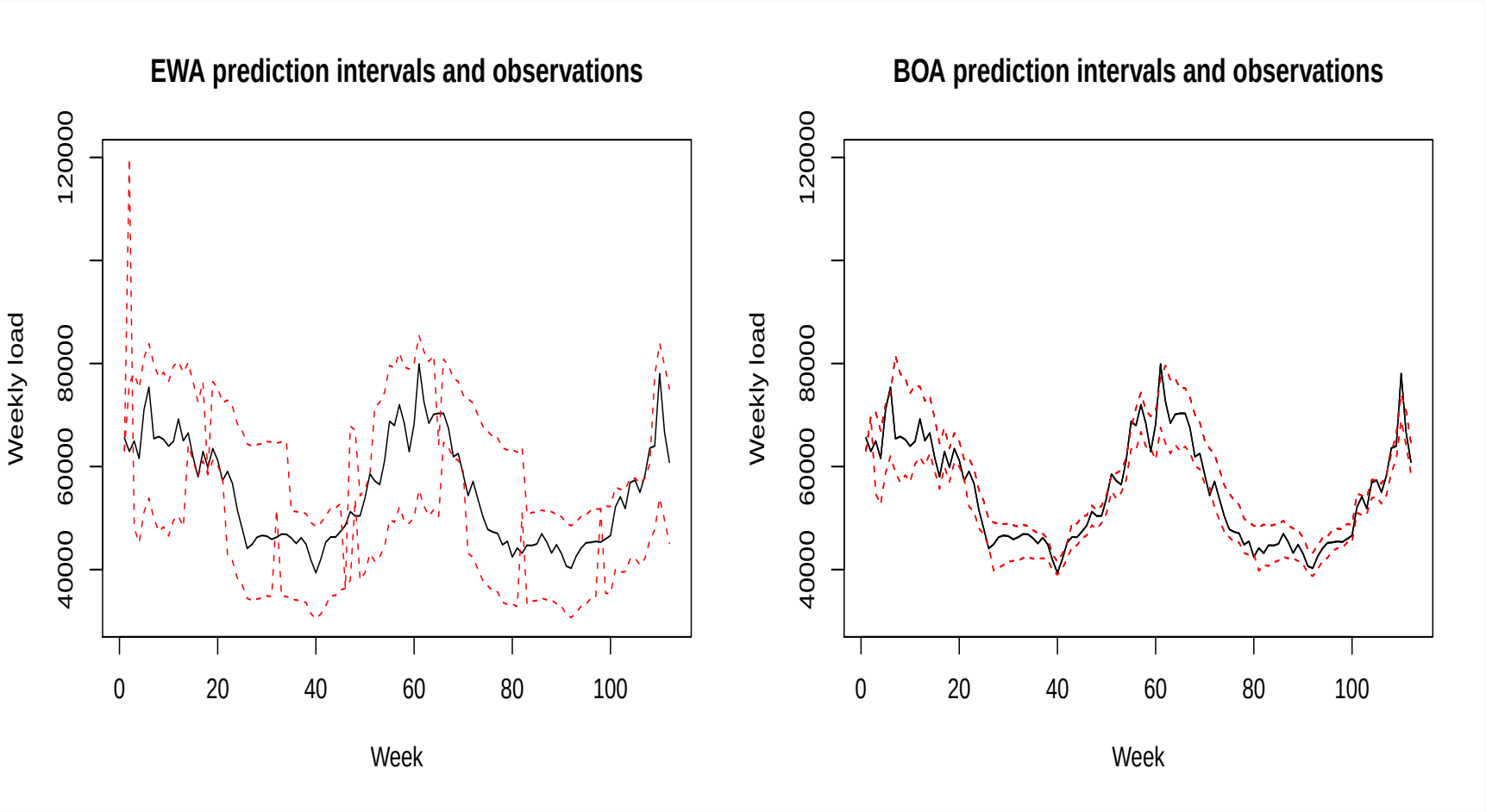
Prediction of the weekly electricity consumption (Opera's vignette, Pierre Gaillard, EDF).

5 experts: GAM, Autoregressive models, GBM, Upper bound $1.5 * \max(\text{observations})$, Lower bound 0.



Prediction intervals

Apply a mixture of the 5 experts with the quantile loss ℓ_τ with $\tau = .05$ and $\tau = .95$ in order to obtain a prediction interval of level 90%.



Conclusion

- Prediction intervals can be built using quantile losses in the expert advice setting,
- a general stochastic setting is useful because conditional quantile risk are more regular then quantile losses,
- The associated aggregation algorithms as BOA are robust and adaptive,
- It works in practice (applications at EDF, Meteo France, Advestis based on package Opera of Pierre Gaillard, EDF).

Perspective (ongoing project with E. Adjakossa)

Adapt the aggregation procedure to the statistical predictors thanks to feedbacks. Example: MetaGrad (Koolen and Van Erven, 2016) aggregates OGD with different learning rates.

Sparse rates in the iid setting (Gaillard & W., 2018)

Sparse rates

Theorem (Kolchintsky et al., 2011)

When ℓ is the square loss, under Restricted Eigenvalue condition on i.i.d. (Y_t, X_t), the LASSO achieves, for $\pi^* \in \mathbb{R}^M$ the minimizer of $\mathbb{E}[\ell(Y, \sum_{j=1}^M \pi_j f_j(X))]$

$$\mathbb{P}\left(R(\widehat{f}^{Lasso}) \leqslant C \sum_{j=1}^{M} 1_{\pi_j^* \neq 0} (\log M + x) \right) \geqslant 1 - e^{-x},$$

where $f_\pi = \sum_{j=1}^M \pi_j X_j$ and $\widehat{f}^{Lasso}(X) = \sum_{j=1}^M \pi_j^{Lasso} X_j$.

With aggregation, seems difficult to compete efficiently with LASSO in $\mathbb{R}^M$ because any discretization grid is exponentially complex in the dimension, Gerchinovitz (2013).

Any efficient and optimal algorithm may require some Restricted Eigenvalue condition, Zhang et al. (2014).

Accelerable rates

Crucial second order properties of the LASSO from Giraud (2015):

For any $\pi \in \mathbb{R}^M$ we have

$$R(\widehat{f}^{\text{Lasso}}) \leqslant R(\pi[f]) + \sum_{j=1}^{M} |\pi_j^{\text{Lasso}} - \pi_j| C\sqrt{\frac{\log M}{T}}.$$

Theorem (Gaillard & W.)

BOA on experts $f_k(X) = \sum_{j=1}^M \pi_j^{(k)} X_j$, $1 \leqslant k \leqslant M$ with $\|\pi^{(k)}\|_1 \leqslant 1$ and the corners of the ℓ_1-ball B_1, satisfies,

$$R(\bar{f}) \leqslant R(\widehat{f}) \leqslant \min_{1 \leqslant k \leqslant M} \sum_{j=1}^{M} |\pi_j^{(k)} - \pi_j| C\sqrt{\frac{\log M}{T}}, \qquad \pi \in B_1,$$

only if Supp$(\pi^{(k)}) \subseteq$ Supp(π) when $\|\pi\|_1 = 1$.

Sparse acceleration of boa

SABOA algorithm:

- Use the doubling trick to run BOA on exponentially growing long sessions.
- At each session, apply BOA on sparse versions of the averaging of the last session and the corners of the ℓ_1-ball B_1.

Theorem (Gaillard & W., 2017)

Assume the Łojasiewicz's condition on the ℓ_1-ball, there exist $\beta > 0$ and $\mu > 0$ such that for all $\pi \in B_1$, it exists a minimizer $\pi^* \in B_1$ of the risk satisfying

$$\mu \left\| \pi - \pi^* \right\|_2^2 \leqslant R(f_\pi) - R(f_{\pi^*}) \,.$$

Assume the set of minimizers π^* is included in $\subseteq B_{1-\gamma}$, $\gamma \geqslant 0$, then with probability at least $1 - e^{-x}$

$$R(\widehat{f}) \leqslant \sup_{\pi^*} (\log M + x) \left(\frac{1}{\alpha} + \frac{1}{\mu} \left(\left(\sum_{j=1}^{M} 1_{\pi_j^* \neq 0} \right)^2 \wedge \frac{\sum_{j=1}^{M} 1_{\pi_j^* \neq 0}}{\gamma^2} \right) \right) ,$$

Conclusion on sparse optimisation

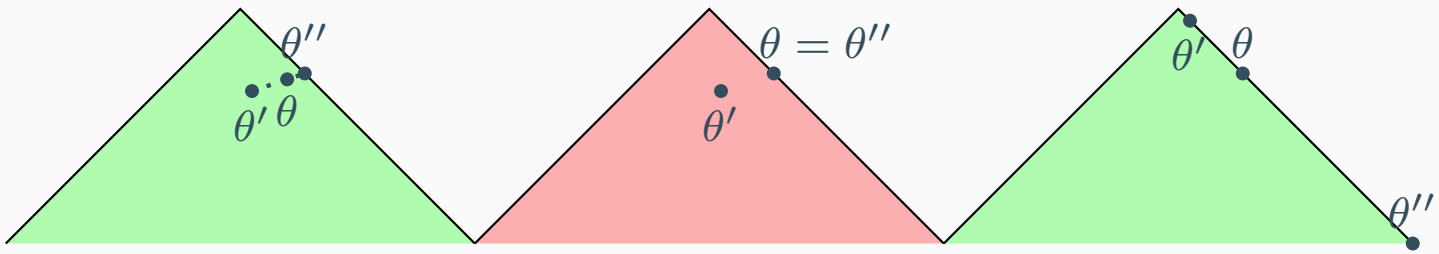


Figure 1: Averaging accelerability for 3 different configurations.

- Versions of BOA may be used for the optimisation problem in some ℓ_1-ball,
- It is an effective online algorithm, robust to the design,
- It is not as fast as LASSO in the favorable cases,
- Optimality of the new rates of convergence?
- How to tune the radius of the ℓ_1-ball?

Thank you for your attention!