Introduction to sequential prediction problems

Pierre Alquier

RIKEN

AIP Center for Advanced Intelligence Project

Online Summer School of ML, Moscow Skoltech, Aug. 19, 2020

Pierre Alquier

- PhD in statistics, Université Paris 6 (2006)
- Lecturer, Université Paris Diderot (2007-2012)
- Lecturer, UCD Dublin (2012-2014)
- Professor, ENSAE Paris (2014-2019)
- Research scientist, RIKEN (2019-...)

For my research, please visit my page :
https ://pierrealquier.github.io/
In case you have any question, please send an e-mail :
pierrealain.alquier@riken.jp

Sequential Prediction

Sequential classification problem - $y_t \in \{0, 1\}$

Sequential Prediction

Sequential classification problem - $y_t \in \{0,1\}$

1. 1. x_1 given

Sequential Prediction

Sequential classification problem - $y_t \in \{0, 1\}$

1. 1. x_1 given
   2. predict y_1 : $\hat{y}_1$

Sequential Prediction

Sequential classification problem - $y_t \in \{0, 1\}$

1.
   1. x_1 given
   2. predict y_1 : $\hat{y}_1$
   3. y_1 is revealed

Sequential Prediction

Sequential classification problem - $y_t \in \{0, 1\}$

1. 1. x_1 given
   2. predict y_1 : $\hat{y}_1$
   3. y_1 is revealed
2. 1. x_2 given

Sequential Prediction

Sequential classification problem - $y_t \in \{0, 1\}$

1. 1. x_1 given
   2. predict y_1 : $\hat{y}_1$
   3. y_1 is revealed
2. 1. x_2 given
   2. predict y_2 : $\hat{y}_2$

Sequential Prediction

Sequential classification problem - $y_t \in \{0, 1\}$

1.
   1. x_1 given
   2. predict y_1 : $\hat{y}_1$
   3. y_1 is revealed
2.
   1. x_2 given
   2. predict y_2 : $\hat{y}_2$
   3. y_2 revealed

Sequential Prediction

Sequential classification problem - $y_t \in \{0,1\}$

1. 1. x_1 given
   2. predict y_1 : $\hat{y}_1$
   3. y_1 is revealed
2. 1. x_2 given
   2. predict y_2 : $\hat{y}_2$
   3. y_2 revealed
3. 1. x_3 given

Sequential Prediction

Sequential classification problem - $y_t \in \{0, 1\}$

1. 1. x_1 given
   2. predict y_1 : $\hat{y}_1$
   3. y_1 is revealed
2. 1. x_2 given
   2. predict y_2 : $\hat{y}_2$
   3. y_2 revealed
3. 1. x_3 given
   2. predict y_3 : $\hat{y}_3$

Sequential Prediction

Sequential classification problem - $y_t \in \{0, 1\}$

1.
    1. x_1 given
    2. predict y_1 : $\hat{y}_1$
    3. y_1 is revealed
2.
    1. x_2 given
    2. predict y_2 : $\hat{y}_2$
    3. y_2 revealed
3.
    1. x_3 given
    2. predict y_3 : $\hat{y}_3$
    3. y_3 revealed
4. . . .

Sequential Prediction

Sequential classification problem - $y_t \in \{0, 1\}$

1.
   1. x_1 given
   2. predict y_1 : $\hat{y}_1$
   3. y_1 is revealed
2.
   1. x_2 given
   2. predict y_2 : $\hat{y}_2$
   3. y_2 revealed
3.
   1. x_3 given
   2. predict y_3 : $\hat{y}_3$
   3. y_3 revealed
4. ...

Objective :

Sequential Prediction

Sequential classification problem - $y_t \in \{0, 1\}$

1. 1. x_1 given
   2. predict y_1 : $\hat{y}_1$
   3. y_1 is revealed
2. 1. x_2 given
   2. predict y_2 : $\hat{y}_2$
   3. y_2 revealed
3. 1. x_3 given
   2. predict y_3 : $\hat{y}_3$
   3. y_3 revealed
4. ...

Objective : make sure that we learn to predict well **as soon as possible**.

Sequential Prediction

Sequential classification problem - $y_t \in \{0,1\}$

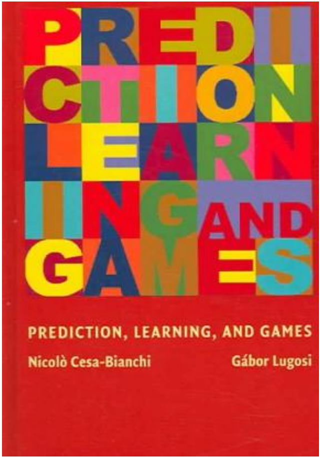
1. 1. x_1 given
   2. predict y_1 : $\hat{y}_1$
   3. y_1 is revealed
2. 1. x_2 given
   2. predict y_2 : $\hat{y}_2$
   3. y_2 revealed
3. 1. x_3 given
   2. predict y_3 : $\hat{y}_3$
   3. y_3 revealed
4. ...

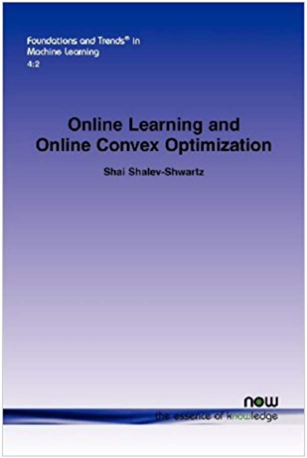
Objective : make sure that we learn to predict well **as soon as possible**. Keep

$$\sum_{t=1}^{T} 1(\hat{y}_t \neq y_t)$$

as small as possible for any T, **without unrealistic assumptions on the data.**

References


PREDICTION LEARNING AND GAMES
PREDICTION, LEARNING, AND GAMES
Nicolò Cesa-Bianchi
Gábor Lugosi


Foundations and Trends® in
Machine Learning
4:2
Online Learning and
Online Convex Optimization
Shai Shalev-Shwartz
now
the essence of knowledge

Outline of the talk

Setting of the problem

Notations : loss function

General notations

Notations : loss function

General notations

- $x_t \in \mathcal{X}$.

Notations : loss function

General notations

- $x_t \in \mathcal{X}$.
- $y_t \in \mathbb{R}$ (regression...) or $y_t \in \{0, 1\}$ (classification).

Notations : loss function

General notations

- $x_t \in \mathcal{X}$.
- $y_t \in \mathbb{R}$ (regression...) or $y_t \in \{0, 1\}$ (classification).
- $\hat{y}_t$ prediction.

Notations : loss function

General notations

- $x_t \in \mathcal{X}$.
- $y_t \in \mathbb{R}$ (regression...) or $y_t \in \{0, 1\}$ (classification).
- $\hat{y}_t$ prediction.
- loss incurred at time t : $\ell(\hat{y}_t, y_t)$ for some real-valued, convex loss function ℓ.

Notations : loss function

General notations

- $x_t \in \mathcal{X}$.
- $y_t \in \mathbb{R}$ (regression...) or $y_t \in \{0,1\}$ (classification).
- $\hat{y}_t$ prediction.
- loss incurred at time t : $\ell(\hat{y}_t, y_t)$ for some real-valued, convex loss function ℓ.

Classical examples : $\ell(y) = |y - y'|$ or $\ell(y) = |y - y'|^2$...

The data

We want to avoid assumptions on the data (x_t, y_t), in order to include situations like :

The data

We want to avoid assumptions on the data (x_t, y_t), in order to include situations like :

- $y_t = F(x_t, \varepsilon_t)$ and the noise variables ε_t are i.i.d.

The data

We want to avoid assumptions on the data (x_t, y_t), in order to include situations like :

- $y_t = F(x_t, \varepsilon_t)$ and the noise variables ε_t are i.i.d.
- $y_t = G(x_{t-1}, y_{t-1}, x_t, \varepsilon_t)$.

The data

We want to avoid assumptions on the data (x_t, y_t), in order to include situations like :

- $y_t = F(x_t, \varepsilon_t)$ and the noise variables ε_t are i.i.d.
- $y_t = G(x_{t-1}, y_{t-1}, x_t, \varepsilon_t)$.
- $y_t = H(x_t, z_t, \varepsilon_t)$ where z_t : omitted variables.

The data

We want to avoid assumptions on the data (x_t, y_t), in order to include situations like :

- $y_t = F(x_t, \varepsilon_t)$ and the noise variables ε_t are i.i.d.
- $y_t = G(x_{t-1}, y_{t-1}, x_t, \varepsilon_t)$.
- $y_t = H(x_t, z_t, \varepsilon_t)$ where z_t : omitted variables.
- $y_t = I(t, x_t, \varepsilon_t)$.

The data

We want to avoid assumptions on the data (x_t, y_t), in order to include situations like :

- $y_t = F(x_t, \varepsilon_t)$ and the noise variables ε_t are i.i.d.
- $y_t = G(x_{t-1}, y_{t-1}, x_t, \varepsilon_t)$.
- $y_t = H(x_t, z_t, \varepsilon_t)$ where z_t : omitted variables.
- $y_t = I(t, x_t, \varepsilon_t)$.
- $y_t = J(\hat{y}_t)$.

The data

We want to avoid assumptions on the data (x_t, y_t), in order to include situations like :

- $y_t = F(x_t, \varepsilon_t)$ and the noise variables ε_t are i.i.d.
- $y_t = G(x_{t-1}, y_{t-1}, x_t, \varepsilon_t)$.
- $y_t = H(x_t, z_t, \varepsilon_t)$ where z_t : omitted variables.
- $y_t = I(t, x_t, \varepsilon_t)$.
- $y_t = J(\hat{y}_t)$.
- $y_t = K(t, (\hat{y}_1, \dots, \hat{y}_t), (x_1, \dots, x_t), (y_1, \dots, y_{t-1}), \varepsilon_t, z_t)$.

Prediction strategy

On the other hand, a realistic prediction cannot be completely arbitrary.

Prediction strategy

On the other hand, a realistic prediction cannot be completely arbitrary.

- We have to be able to compute $\hat{y}_t$ it can depend on $(x_1, \dots, x_t)$ and $(y_1, \dots, y_{t-1})$.

Prediction strategy

On the other hand, a realistic prediction cannot be completely arbitrary.

- We have to be able to compute $\hat{y}_t$ it can depend on $(x_1, \dots, x_t)$ and $(y_1, \dots, y_{t-1})$.
- It must be computationnally feasible.

Prediction strategy

On the other hand, a realistic prediction cannot be completely arbitrary.

- We have to be able to compute $\hat{y}_t$ it can depend on $(x_1, \dots, x_t)$ and $(y_1, \dots, y_{t-1})$.
- It must be computationnally feasible.
- We can use expert advice.

What performance can we achieve in this setting ?

Consider binary classification with $\ell(y, y') = 1(y \neq y')$, as we allowed $y_t = J(\hat{y}_t)$, the opponent can always chose $y_t = 1 - \hat{y}_t$ which leads to

$$\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) = T.$$

What performance can we achieve in this setting ?

Consider binary classification with $\ell(y, y') = 1(y \neq y')$, as we allowed $y_t = J(\hat{y}_t)$, the opponent can always chose $y_t = 1 - \hat{y}_t$ which leads to

$$\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) = T.$$

On the other hand, many real world phenomena can be "quite well" described by models. These models allow to do "sensible" predictions.

What performance can we achieve in this setting ?

Consider binary classification with $\ell(y, y') = 1(y \neq y')$, as we allowed $y_t = J(\hat{y}_t)$, the opponent can always chose $y_t = 1 - \hat{y}_t$ which leads to

$$\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) = T.$$

On the other hand, many real world phenomena can be "quite well" described by models. These models allow to do "sensible" predictions.

The extreme case would be the constraint $y_t = f(x_t)$, where $f \in \mathcal{F}$ for a known class $\mathcal{F}$. This is called the *realizable case*. Let's study it as a toy example when $\mathcal{F}$ is finite.

A naive strategy

Here $y_t = f_{i^*}(x_t)$ where $i^* \in \{1, \dots, M\}$ is unknown.

Naive strategy

A naive strategy

Here $y_t = f_{i^*}(x_t)$ where $i^* \in \{1, \dots, M\}$ is unknown.

Naive strategy

Start with $i(1) = 1$ and $C(1) = \{1, \dots, M\}$. At step t,

A naive strategy

Here $y_t = f_{i^*}(x_t)$ where $i^* \in \{1, \dots, M\}$ is unknown.

Naive strategy

Start with $i(1) = 1$ and $C(1) = \{1, \dots, M\}$. At step t,

1. predict $\hat{y}_t = f_{i(t)}(x_t)$, observe y_t,

A naive strategy

Here $y_t = f_{i^*}(x_t)$ where $i^* \in \{1, \dots, M\}$ is unknown.

Naive strategy

Start with $i(1) = 1$ and $C(1) = \{1, \dots, M\}$. At step t,

1. predict $\hat{y}_t = f_{i(t)}(x_t)$, observe y_t,
2. update $\begin{cases} C(t+1) = \{i \in C(t) : f_i(x_t) = y_t\}, \\ i(t+1) = \min C(t+1). \end{cases}$

A naive strategy

Here $y_t = f_{i^*}(x_t)$ where $i^* \in \{1, \dots, M\}$ is unknown.

Naive strategy

Start with $i(1) = 1$ and $C(1) = \{1, \dots, M\}$. At step t,

1. predict $\hat{y}_t = f_{i(t)}(x_t)$, observe y_t,
2. update $\begin{cases} C(t+1) = \{i \in C(t) : f_i(x_t) = y_t\}, \\ i(t+1) = \min C(t+1). \end{cases}$

Theorem

$$\forall T, \sum_{t=1}^{T} \ell(\hat{y}_t, y_t) \leq M - 1.$$

Naive strategy : an example

Naive strategy : an example

Naive strategy : an example

$2-1$

Naive strategy : an example

$$2-1$$

Naive strategy : an example

$2-1$ $\quad$ $1-1$

Naive strategy : an example

$2-1$ $1-1$

Naive strategy : an example

$2-1$ $\quad$ $1-1$ $\quad$ $2-0$

Naive strategy : an example

$2-1$ $1-1$ $2-0$

Naive strategy : an example

$2-1$ $1-1$ $2-0$ $3-0$

Naive strategy : an example

$2-1$ $\quad$ $1-1$ $\quad$ $2-0$ $\quad$ $3-0$

Naive strategy : an example

$2-1$

$1-1$

$2-0$

$3-0$

$0-1$

Naive strategy : an example





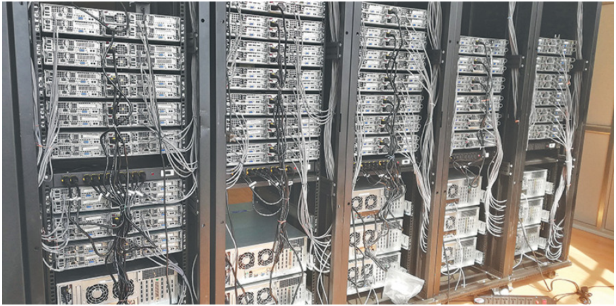
W	DL	W	W	DL

Naive strategy : an example

Remind that we have to assume that one of the experts is never wrong. Is there such an expert ?

Naive strategy : an example

Remind that we have to assume that one of the experts is never wrong. Is there such an expert ?



This is not a realistic example, so let us imagine a data scientist built a perfect IA.

Naive strategy : an example

Naive strategy : an example

$t = 1$ W DL W W DL W

Naive strategy : an example

						
$t = 1$	W	<u>DL</u>	W	W	DL	W

Naive strategy : an example

$t = 1$	W	DL	W	W	DL	W

Naive strategy : an example

$t = 1$	W	DL	W	W	DL	W
$t = 2$	DL		DL	W		DL

Naive strategy : an example

						
$t = 1$	W	DL	W	W	DL	W
$t = 2$	DL		DL	W		DL

Naive strategy : an example






$t = 1$	W	DL	W	W	DL	W
$t = 2$	DL		DL	W		DL

Naive strategy : an example






$t = 1$	W	DL	W	W	DL	W
$t = 2$	DL		DL	W		DL
$t = 3$	W		DL			W

Naive strategy : an example

						
$t = 1$	W	DL	W	W	DL	W
$t = 2$	DL		DL	W		DL
$t = 3$	W		DL			W

Naive strategy : an example

$t = 1$	W	DL	W	W	DL	W
$t = 2$	DL		DL	W		DL
$t = 3$	W		DL			W

Naive strategy : an example






$t = 1$	W	DL	W	W	DL	W
$t = 2$	DL		DL	W		DL
$t = 3$	W		DL			W
$t = 4$	DL					W

Naive strategy : an example

						
$t = 1$	W	DL	W	W	DL	W
$t = 2$	DL		DL	W		DL
$t = 3$	W		DL			W
$t = 4$	DL					W

Naive strategy : an example

$t=1$	W	<u>DL</u>	W	W	DL	W
$t=2$	DL		DL	<u>W</u>		DL
$t=3$	W		<u>DL</u>			W
$t=4$	<u>DL</u>					W

Naive strategy : an example






$t = 1$	W	DL	W	W	DL	W
$t = 2$	DL		DL	W		DL
$t = 3$	W		DL			W
$t = 4$	DL					W
$t = 5$						W

The halving algorithm

(Still $y_t = f_{i^*}(x_t)$ where $i^* \in \{1, \dots, M\}$ is unknown).

The halving algorithm

The halving algorithm

(Still $y_t = f_{i^*}(x_t)$ where $i^* \in \{1, \dots, M\}$ is unknown).

The halving algorithm

Start with $i(1) = 1$ and $C(1) = \{1, \dots, M\}$. At step t,

The halving algorithm

(Still $y_t = f_{i^*}(x_t)$ where $i^* \in \{1, \dots, M\}$ is unknown).

The halving algorithm

Start with $i(1) = 1$ and $C(1) = \{1, \dots, M\}$. At step t,

1. predict $\hat{y}_t =$ "majority vote in $C(t)$", observe y_t,

The halving algorithm

(Still $y_t = f_{i^*}(x_t)$ where $i^* \in \{1, \dots, M\}$ is unknown).

The halving algorithm

Start with $i(1) = 1$ and $C(1) = \{1, \dots, M\}$. At step t,

1. predict $\hat{y}_t =$ "majority vote in $C(t)$", observe y_t,
2. update $C(t+1) = \{i \in C(t) : f_i(x_t) = y_t\}$.

The halving algorithm

(Still $y_t = f_{i^*}(x_t)$ where $i^* \in \{1, \dots, M\}$ is unknown).

The halving algorithm

Start with $i(1) = 1$ and $C(1) = \{1, \dots, M\}$. At step t,

1. predict $\hat{y}_t =$ "majority vote in $C(t)$", observe y_t,
2. update $C(t+1) = \{i \in C(t) : f_i(x_t) = y_t\}$.

Theorem

$$\forall T, \ \sum_{t=1}^{T} \ell(\hat{y}_t, y_t) \leq \log_2(M).$$

Halving algorithm : an example






Halving algorithm : an example






$t = 1$	W	DL	W	W	DL	W

Halving algorithm : an example

$t = 1$	W	DL	W	W	DL	W

Halving algorithm : an example

$t = 1$	W	DL	W	W	DL	W

Halving algorithm : an example

						
$t = 1$	W	DL	W	W	DL	W
$t = 2$	DL		DL	W		DL

Halving algorithm : an example

						
$t = 1$	W	DL	W	W	DL	W
$t = 2$	DL		DL	W		DL

Halving algorithm : an example

$t = 1$	W	DL	W	W	DL	W
$t = 2$	DL		DL	W		DL

Halving algorithm : an example

						
$t = 1$	W	DL	W	W	DL	W
$t = 2$	DL		DL	W		DL
$t = 3$	W		DL			W

Halving algorithm : an example






$t = 1$	W	DL	W	W	DL	W
$t = 2$	DL		DL	W		DL
$t = 3$	W		DL			W

Halving algorithm : an example

$t = 1$	W	DL	W	W	DL	W
$t = 2$	DL		DL	W		DL
$t = 3$	W		DL			W

Halving algorithm : an example






$t = 1$	W	DL	W	W	DL	W
$t = 2$	DL		DL	W		DL
$t = 3$	W		DL			W
$t = 4$	DL					W

Halving algorithm : an example

						
$t = 1$	W	DL	W	W	DL	W
$t = 2$	DL		DL	W		DL
$t = 3$	W		DL			W
$t = 4$	DL					W

Halving algorithm : an example






$t = 1$	W	DL	W	W	DL	W
$t = 2$	DL		DL	W		DL
$t = 3$	W		DL			W
$t = 4$	DL					W

Halving algorithm : an example

						
$t = 1$	W	DL	W	W	DL	W
$t = 2$	DL		DL	W		DL
$t = 3$	W		DL			W
$t = 4$	DL					W
$t = 5$						W

Halving algorithm : another example

Halving algorithm : another example

$t = 1$	W	DL	DL	DL	DL	W

Halving algorithm : another example

						
$t = 1$	W	<u>DL</u>	<u>DL</u>	<u>DL</u>	<u>DL</u>	W

Halving algorithm : another example

$t = 1$	W	DL	DL	DL	DL	W

Halving algorithm : another example






$t = 1$	W	DL	DL	DL	DL	W
$t = 2$	DL					DL

Halving algorithm : another example

$t = 1$	W	DL	DL	DL	DL	W
$t = 2$	DL					DL

Halving algorithm : another example

$t = 1$	W	DL	DL	DL	DL	W
$t = 2$	DL					DL

Halving algorithm : another example






$t = 1$	W	DL	DL	DL	DL	W
$t = 2$	DL					DL
$t = 3$	DL					W

Halving algorithm : another example

$t = 1$	W	DL	DL	DL	DL	W
$t = 2$	DL					DL
$t = 3$	DL					W

Halving algorithm : another example

$t = 1$	W	DL	DL	DL	DL	W
$t = 2$	DL					DL
$t = 3$	DL					W

Halving algorithm : another example

						
$t = 1$	W	DL	DL	DL	DL	W
$t = 2$	DL					DL
$t = 3$	DL					W
$t = 4$						W

Halving algorithm : another example






$t = 1$	W	DL	DL	DL	DL	W
$t = 2$	DL					DL
$t = 3$	DL					W
$t = 4$						W
$t = 5$						W

A feasible objective

Two extremes :

- playing against the devil $y_t = 1 - \hat{y}_t$,
- assuming a true, exact model $\mathcal{F}$.

A feasible objective

Two extremes :

- playing against the devil $y_t = 1 - \hat{y}_t$,
- assuming a true, exact model $\mathcal{F}$.

Real-life is somewhere in between !

A feasible objective

Two extremes :

- playing against the devil $y_t = 1 - \hat{y}_t$,
- assuming a true, exact model $\mathcal{F}$.

Real-life is somewhere in between !

Objective

Strategy such that

$$\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) \leq \underbrace{\inf_{f\in\mathcal{F}} \sum_{t=1}^{T} \ell(f(x_t), y_t)}_{\substack{= T \text{ in the worst case (devil),} \\ = 0 \text{ in the ideal case (true model),} \\ \text{almost always in between.}}} + \underbrace{B(T)}_{\text{as small as possible !!}} .$$

The regret

$$\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) \leq \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(f(x_t), y_t) + B(T)$$

The regret

$$\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(f(x_t), y_t) \leq B(T)$$

The regret

$$\mathrm{Regret}(T) = \sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{T} \ell(f(x_t), y_t) \leq B(T)$$

The regret

$$\mathrm{Regret}(T) = \sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(f(x_t), y_t) \leq B(T)$$

Objective

Strategy such that $\mathrm{Regret}(T) \leq B(T)$ as small as possible, at least $B(T) = o(T)$.

The regret

$$\mathrm{Regret}(T) = \sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(f(x_t), y_t) \leq B(T)$$

Objective

Strategy such that $\mathrm{Regret}(T) \leq B(T)$ as small as possible, at least $B(T) = o(T)$.

We'll see that

- for a bounded ℓ, $B(T) = \mathcal{O}(\sqrt{T})$ always feasible with a randomized strategy.
- deterministic results, and $B(T) = \mathcal{O}(\log(T))$ or even $B(T) = \mathcal{O}(1)$, possible under more assumptions.

Important remarks

1. Common misunderstanding :
   machine learning $\simeq$ prediction, opposed to modelization.

Important remarks

1. Common misunderstanding :
   machine learning $\simeq$ prediction, opposed to modelization.
2. However ! modelization (economics, physics, epidemiology) is required to build $\mathcal{F}$:

$$\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) \leq \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(f(x_t), y_t) + B(T).$$

Important remarks

1. Common misunderstanding :
   machine learning $\simeq$ prediction, opposed to modelization.
2. However ! modelization (economics, physics, epidemiology) is required to build $\mathcal{F}$:

$$\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) \leq \inf_{f\in\mathcal{F}} \sum_{t=1}^{T} \ell(f(x_t), y_t) + B(T).$$

3. Common mistake : machine learning provides good predictions in practice, but has no theoretical ground.

Important remarks

1. Common misunderstanding :
   machine learning $\simeq$ prediction, opposed to modelization.
2. However ! modelization (economics, physics, epidemiology) is required to build $\mathcal{F}$:

$$\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) \leq \inf_{f\in\mathcal{F}} \sum_{t=1}^{T} \ell(f(x_t), y_t) + B(T).$$

3. Common mistake : machine learning provides good predictions in practice, but has no theoretical ground.
4. Wrong ! We'll see some theoretical results below.

Proposition

My own view is that machine learning theory is itself a model for “the performance of a scientist who uses a model for prediction in an environment where the model might not be exactly correct”.

Exponentially Weighted Aggregation (EWA)

Finite number of predictors

Let us start with the case of a finite set of M predictors :

$$\mathcal{F} = (f_1, \dots, f_M).$$

Finite number of predictors

Let us start with the case of a finite set of M predictors :

$$\mathcal{F} = (f_1, \dots, f_M).$$

What should the f_i's be ?

Finite number of predictors

Let us start with the case of a finite set of M predictors :

$$\mathcal{F} = (f_1, \dots, f_M).$$

What should the f_i's be ? By including side information in $\tilde{x}_t$ such as the past $\tilde{x}_t = (x_1, y_1, \dots, x_{t-1}, y_{t-1}, x_t)$, we can have rich predictors. For example :

$$f_1(\tilde{x}_t) = \hat{\beta}_t^T x_t$$

where

$$\hat{\beta}_t = \arg\min_{\beta} \sum_{i=1}^{t-1} (y_i - \beta^T x_i)^2.$$

Expert advice

More importantly, we can use "expert advice" : an expert e proposes at each time t a forecast $\hat{y}_t^{\mathrm{e}}$, why not using it ?

Expert advice

More importantly, we can use "expert advice" : an expert e proposes at each time t a forecast $\hat{y}_t^{\mathrm{e}}$, why not using it ?

For a while, we forget about the x_t's. At each time t, M different forecasts are proposed :

$$(\hat{y}_t^{(1)}, \dots, \hat{y}_t^{(M)}).$$

Some come from models, others from experts. For short we refer to all of them as "experts advice". I have to make my own prediction $\hat{y}_t$ based on this.

Expert advice

More importantly, we can use "expert advice" : an expert e proposes at each time t a forecast $\hat{y}_t^{\mathrm{e}}$, why not using it ?

For a while, we forget about the x_t's. At each time t, M different forecasts are proposed :

$$(\hat{y}_t^{(1)}, \dots, \hat{y}_t^{(M)}).$$

Some come from models, others from experts. For short we refer to all of them as "experts advice". I have to make my own prediction $\hat{y}_t$ based on this.

$$\mathrm{Regret}(T) = \sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \min_{i=1,\dots,M} \sum_{t=1}^{T} \ell(\hat{y}_t^{(i)}, y_t) \leq ?$$

Randomized EWA strategy

EWA : Exponentially Weighted Aggregation. Input :

- learning rate $\eta > 0$,
- initial weights $p_1(1), \dots, p_1(M) \geq 0$ with $\sum_{i=1}^{M} p_1(i) = 1$.

Randomized EWA strategy

EWA : Exponentially Weighted Aggregation. Input :

- learning rate $\eta > 0$,
- initial weights $p_1(1), \dots, p_1(M) \geq 0$ with $\sum_{i=1}^{M} p_1(i) = 1$.

Algorithm 1 EWA (Randomized version)

1: **for** $i = 1, 2, \dots$ **do**
2: Draw I_t with $\mathbb{P}(I_t = i) = p_t(i)$
3: Predict $\hat{y}_t = \hat{y}_t^{(I_t)}$,
4: y_t revealed, update $p_{t+1}(i) = \frac{p_t(i) \exp[-\eta \ell(\hat{y}_t^{(i)}, y_t)]}{\sum_{j=1}^{M} p_t(j) \exp[-\eta \ell(\hat{y}_t^{(j)}, y_t)]}$
5: **end for**

Guarantees (in expectation)

Theorem

Assume that $\ell(\cdot,\cdot) \in [0, C]$ (e.g. classification). Then

$$\mathbb{E}\left(\mathrm{Regret}(T)\right) \leq \frac{\eta C^2 T}{8} + \frac{\log(M)}{\eta}$$

Guarantees (in expectation)

Theorem

Assume that $\ell(\cdot,\cdot) \in [0, C]$ (e.g. classification). Then

$$\mathbb{E}\left(\mathrm{Regret}(T)\right) \leq \frac{\eta C^2 T}{8} + \frac{\log(M)}{\eta}$$

$$\eta = \frac{1}{C}\sqrt{\frac{8\log(M)}{T}} \Rightarrow \mathbb{E}\left(\mathrm{Regret}(T)\right) \leq C\sqrt{\frac{T\log(M)}{2}}.$$

Guarantees (in expectation)

Theorem

Assume that $\ell(\cdot,\cdot) \in [0, C]$ (e.g. classification). Then

$$\mathbb{E}\left(\mathrm{Regret}(T)\right) \leq \frac{\eta C^2 T}{8} + \frac{\log(M)}{\eta}$$

$$\eta = \frac{1}{C}\sqrt{\frac{8\log(M)}{T}} \Rightarrow \mathbb{E}\left(\mathrm{Regret}(T)\right) \leq C\sqrt{\frac{T\log(M)}{2}}.$$

- the expectation is only w.r.t the algorithm. No assumption on the data.

Guarantees (in expectation)

Theorem

Assume that $\ell(\cdot,\cdot) \in [0, C]$ (e.g. classification). Then

$$\mathbb{E}\left(\mathrm{Regret}(T)\right) \leq \frac{\eta C^2 T}{8} + \frac{\log(M)}{\eta}$$

$$\eta = \frac{1}{C}\sqrt{\frac{8\log(M)}{T}} \Rightarrow \mathbb{E}\left(\mathrm{Regret}(T)\right) \leq C\sqrt{\frac{T\log(M)}{2}}.$$

- the expectation is only w.r.t the algorithm. No assumption on the data.
- possible to take $\eta_t \sim 1/\sqrt{t}$.

Guarantees (in expectation)

Theorem

Assume that $\ell(\cdot,\cdot) \in [0, C]$ (e.g. classification). Then

$$\mathbb{E}\left(\mathrm{Regret}(T)\right) \leq \frac{\eta C^2 T}{8} + \frac{\log(M)}{\eta}$$

$$\eta = \frac{1}{C}\sqrt{\frac{8\log(M)}{T}} \Rightarrow \mathbb{E}\left(\mathrm{Regret}(T)\right) \leq C\sqrt{\frac{T\log(M)}{2}}.$$

- the expectation is only w.r.t the algorithm. No assumption on the data.
- possible to take $\eta_t \sim 1/\sqrt{t}$.
- what about deterministic prediction ?

EWA strategy

Input :

- learning rate $\eta > 0$,
- weights $p_1(1), \dots, p_1(M)$.

EWA strategy

Input :

- learning rate $\eta > 0$,
- weights $p_1(1), \dots, p_1(M)$.

Algorithm 2 EWA

1: **for** $i = 1, 2, \dots$ **do**
2: Predict $\hat{y}_t = \sum_{i=1}^M p_t(i) \hat{y}_t^{(i)}$,
3: y_t revealed, update $p_{t+1}(i) = \frac{p_t(i) \exp[-\eta \ell(\hat{y}_t^{(i)}, y_t)]}{\sum_{j=1}^M p_t(j) \exp[-\eta \ell(\hat{y}_t^{(j)}, y_t)]}$
4: **end for**

EWA - theorem

Theorem

Assume that $0 \leq \ell \leq C$ and ℓ is convex. Then

$$\mathrm{Regret}(T) \leq \frac{\eta C^2 T}{8} + \frac{\log(M)}{\eta}.$$

EWA - theorem

Theorem

Assume that $0 \leq \ell \leq C$ and ℓ is convex. Then

$$\mathrm{Regret}(T) \leq \frac{\eta C^2 T}{8} + \frac{\log(M)}{\eta}.$$

In other words, without any assumption on the data, with $\eta = \frac{1}{C}\sqrt{\frac{8\log(M)}{T}}$,

$$\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) \leq \min_{i=1,\dots,M} \sum_{t=1}^{T} \ell\left(\hat{y}_t^{(i)}, y_t\right) + C\sqrt{\frac{T\log(M)}{2}}.$$

EWA - theorem

Theorem

Assume that $0 \leq \ell \leq C$ and ℓ is convex. Then, with

$$\eta_t = \frac{1}{C}\sqrt{\frac{8\log(M)}{t}}.$$

Then

$$\mathrm{Regret}(T) \leq 2C\sqrt{\frac{T\log(M)}{2}} + \sqrt{\log\left(\frac{M}{8}\right)}.$$

Example 1 : air quality prediction

Journal de la Société Française de Statistique
Vol. 151 No. 2 (2010)

Agrégation séquentielle de prédicteurs : méthodologie générale et applications à la prévision de la qualité de l'air et à celle de la consommation électrique

Title: Sequential aggregation of predictors: General methodology and application to air-quality forecasting and to the prediction of electricity consumption

Gilles Stoltz *

Résumé : Cet article fait suite à la conférence que j'ai eu l'honneur de donner lors de la réception du prix Marie-Jeanne Laurent-Duhamel, dans le cadre des XLᵉ Journées de Statistique à Ottawa, en 2008. Il passe en revue les résultats fondamentaux, ainsi que quelques résultats récents, en prévision séquentielle de suites arbitraires par agrégation d'experts. Il décline ensuite la méthodologie ainsi décrite sur deux jeux de données, l'un pour un problème de prévision de qualité de l'air, l'autre pour une question de prévision de consommation électrique. La plupart des résultats mentionnés dans cet article reposent sur des travaux en collaboration avec Yannig Goude (EDF R&D) et Vivien Mallet (INRIA), ainsi qu'avec les stagiaires de master que nous avons co-encadrés : Marie Devaine, Sébastien Gerchinovitz et Boris Mauricette.

Abstract: This paper is an extended written version of the talk I delivered at the "XLᵉ Journées de Statistique" in Ottawa, 2004, when being awarded the Marie-Jeanne Laurent-Duhamel prize. It is devoted to surveying some fundamental as well as some more recent results in the field of sequential prediction of individual sequences with expert advice. It then performs two empirical studies following the stated general methodology: the first one to air-quality forecasting and the second one to the prediction of electricity consumption. Most results mentioned in the paper are based on joint works with Yannig Goude (EDF R&D) and Vivien Mallet (INRIA), together with some students whom we co-supervised for their M.Sc. theses: Marie Devaine, Sébastien Gerchinovitz and Boris Mauricette.

Classification AMS 2000 : primaire 62-02, 62L99, 62P12, 62P30

Mots-clés : Agrégation séquentielle, prévision avec experts, suites individuelles, prévision de la qualité de l'air, prévision de la consommation électrique

Keywords: Sequential aggregation of predictors, prediction with expert advice, individual sequences, air-quality forecasting, prediction of electricity consumption

Ecole normale supérieure, CNRS, 45 rue d'Ulm, 75005 Paris
& HEC Paris, CNRS, 1 rue de la Libération, 78350 Jouy-en-Josas
E-mail : gilles.stoltz@ens.fr
URL : http://www.math.ens.fr/∼stoltz

* L'auteur remercie l'Agence nationale de la recherche pour son soutien à travers le projet JCJC06-137444 ATLAS ("From applications to theory in learning and adaptive statistics").
† Ces recherches ont été menées dans le cadre du projet CLASSIC de l'INRIA, hébergé par l'Ecole normale supérieure et le CNRS.

Journal de la Société Française de Statistique, **Vol. 151** No. 2 66-106
http://www.sfds.asso.fr/journal
 ISSN: 2102-6238

The data and the problem

- 126 days during summer 2001. 241 stations in France and Germany.

The data and the problem

- 126 days during summer 2001. 241 stations in France and Germany.
- one-day ahead prediction, quadratic loss.

The data and the problem

- 126 days during summer 2001. 241 stations in France and Germany.
- one-day ahead prediction, quadratic loss.
- typical ozone concentrations between $40\mu\mathrm{gm}^{-3}$ and $150\mu\mathrm{gm}^{-3}$, a few extreme values up to $240\mu\mathrm{gm}^{-3}$.

The data and the problem

- 126 days during summer 2001. 241 stations in France and Germany.
- one-day ahead prediction, quadratic loss.
- typical ozone concentrations between $40\mu\mathrm{gm}^{-3}$ and $150\mu\mathrm{gm}^{-3}$, a few extreme values up to $240\mu\mathrm{gm}^{-3}$.
- $M = 48$ experts taken from a paper in geophysics by choosing a physical and chemical formulation, a numerical approximation scheme to solve the involved PDEs, and a set of input data.

Prediction by the experts

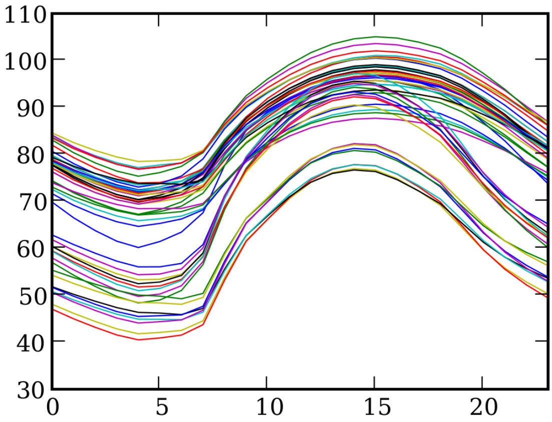


Figure – Predictions by the 48 experts for one day at one station.

Numerical performances

	RMSE
Best expert	22.43
Uniform mean	24.41
EWA	21.47

Figure – Numerical performances (RMSE).

Weights

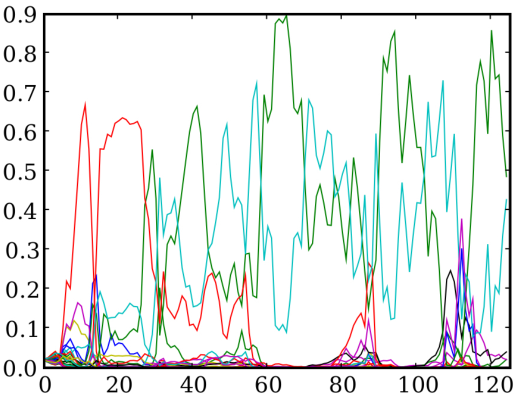


Figure – Evolution of the weights $p_i(t)$ w.r.t t.

Example 2 : GDP growth in France

Prediction of Quantiles by Statistical Learning and Application to GDP Forecasting

Pierre Alquier[1,3] and Xiaoyin Li[2]

[1] LPMA (Université Paris 7)
175, rue du Chevaleret
75013 Paris, France
alquier@math.jussieu.fr
http://alquier.ensae.net/
[2] Laboratoire de Mathématiques (Université de Cergy-Pontoise)
UCP site Saint-Martin, 2 boulevard Adolphe Chauvin
95000 Cergy-Pontoise, France
xiaoyin.li@u-cergy.fr
[3] CREST (ENSAE)

Abstract. In this paper, we tackle the problem of prediction and confidence intervals for time series using a statistical learning approach and quantile loss functions. In a first time, we show that the Gibbs estimator is able to predict as well as the best predictor in a given family for a wide set of loss functions. In particular, using the quantile loss function of [1], this allows to build confidence intervals. We apply these results to the problem of prediction and confidence regions for the French Gross Domestic Product (GDP) growth, with promising results.

Keywords: Statistical learning theory, time series, quantile regression, GDP forecasting, PAC-Bayesian bounds, oracle inequalities, weak dependence, confidence intervals, business surveys.

1 Introduction

Motivated by economics problems, the prediction of time series is one of the most emblematic problems of statistics. Various methodologies are used that come from such various fields as parametric statistics, statistical learning, computer science or game theory.

In the parametric approach, one assumes that the time series is generated from a parametric model, e.g. ARMA or ARIMA, see [2,3]. It is then possible to estimate the parameters of the model and to build confidence intervals on the prevision. However, such an assumption is unrealistic in most applications.

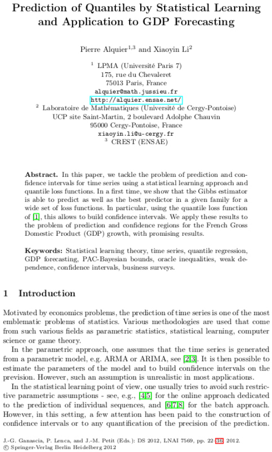
In the statistical learning point of view, one usually tries to avoid such restrictive parametric assumptions - see, e.g., [4,5] for the online approach dedicated to the prediction of individual sequences, and [6,7,8] for the batch approach. However, in this setting, a few attention has been paid to the construction of confidence intervals or to any quantification of the precision of the prediction.

J.-G. Ganascia, P. Lenca, and J.-M. Petit (Eds.): DS 2012, LNAI 7569, pp. 22–36, 2012.

Data from :

INSEE
INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES

Example 2 : GDP growth in France



Prediction of Quantiles by Statistical Learning and Application to GDP Forecasting

Pierre Alquier[1,3] and Xiaoyin Li[2]

[1] LPMA (Université Paris 7)
175, rue du Chevaleret
75013 Paris, France
alquier@math.jussieu.fr
http://alquier.ensae.net/
[2] Laboratoire de Mathématiques (Université de Cergy-Pontoise)
UCP site Saint-Martin, 2 boulevard Adolphe Chauvin
95000 Cergy-Pontoise, France
xiaoyin.li@u-cergy.fr
[3] CREST (ENSAE)

Abstract. In this paper, we tackle the problem of prediction and confidence intervals for time series using a statistical learning approach and quantile loss functions. In a first time, we show that the Gibbs estimator is able to predict as well as the best predictor in a given family for a wide set of loss functions. In particular, using the quantile loss function of [1], this allows to build confidence intervals. We apply these results to the problem of prediction and confidence regions for the French Gross Domestic Product (GDP) growth, with promising results.

Keywords: Statistical learning theory, time series, quantile regression, GDP forecasting, PAC-Bayesian bounds, oracle inequalities, weak dependence, confidence intervals, business surveys.

1 Introduction

Motivated by economics problems, the prediction of time series is one of the most emblematic problems of statistics. Various methodologies are used that come from such various fields as parametric statistics, statistical learning, computer science or game theory.

In the parametric approach, one assumes that the time series is generated from a parametric model, e.g. ARMA or ARIMA, see [2,3]. It is then possible to estimate the parameters of the model and to build confidence intervals on the prevision. However, such an assumption is unrealistic in most applications.

In the statistical learning point of view, one usually tries to avoid such restrictive parametric assumptions - see, e.g., [4,5] for the online approach dedicated to the prediction of individual sequences, and [6,7,8] for the batch approach. However, in this setting, a few attention has been paid to the construction of confidence intervals or to any quantification of the precision of the prediction.

J.-G. Ganascia, P. Lenca, and J.-M. Petit (Eds.): DS 2012, LNAI 7569, pp. 22–36, 2012.
© Springer-Verlag Berlin Heidelberg 2012

Data from :

INSEE
INSTITUT NATIONAL
DE LA STATISTIQUE
ET DES ÉTUDES
ÉCONOMIQUES

It is a French institution similar to :

Федеральная служба
государственной статистики

GDP growth forecasting

Objective : during the 3rd month of quarter t, predict what will be the GDP growth during the quarter : $\Delta\mathrm{GDP}_t$.

GDP growth forecasting

Objective : during the 3rd month of quarter t, predict what will be the GDP growth during the quarter : $\Delta\mathrm{GDP}_t$.

Available from INSEE :

GDP growth forecasting

Objective : during the 3rd month of quarter t, predict what will be the GDP growth during the quarter : $\Delta\mathrm{GDP}_t$.

Available from INSEE :

1. the past : $\Delta\mathrm{GDP}_{t-1}$, ..., $\Delta\mathrm{GDP}_1$, with $t = 1$: 1988-T1.

GDP growth forecasting

Objective : during the 3rd month of quarter t, predict what will be the GDP growth during the quarter : $\Delta\mathrm{GDP}_t$.

Available from INSEE :

1. the past : $\Delta\mathrm{GDP}_{t-1}$, ..., $\Delta\mathrm{GDP}_1$, with $t = 1$: 1988-T1.
2. French business surveys.

GDP growth forecasting

Objective : during the 3rd month of quarter t, predict what will be the GDP growth during the quarter : $\Delta\mathrm{GDP}_t$.

INSEE
INSTITUT NATIONAL
DE LA STATISTIQUE
ET DES ÉTUDES
ÉCONOMIQUES

Available from INSEE :

1. the past : $\Delta\mathrm{GDP}_{t-1}$, ..., $\Delta\mathrm{GDP}_1$, with $t = 1$: 1988-T1.
2. French business surveys.
3. much more...

Business surveys

Business surveys : forms sent monthly to big companies, and to a sample of small companies. These data are to be taken into account because

Business surveys

Business surveys : forms sent monthly to big companies, and to a sample of small companies. These data are to be taken into account because

1. they don't come from economists, but from economic agents.

Business surveys

Business surveys : forms sent monthly to big companies, and to a sample of small companies. These data are to be taken into account because

1. they don't come from economists, but from economic agents.
2. they are available almost immediately. During the 3rd month of quarter t, the analysis of the forms for the 1st and the 2nd months are already known.

Business surveys

Business surveys : forms sent monthly to big companies, and to a sample of small companies. These data are to be taken into account because

1. they don't come from economists, but from economic agents.
2. they are available almost immediately. During the 3rd month of quarter t, the analysis of the forms for the 1st and the 2nd months are already known.

$\rightarrow$ this information is summarized in the *business climate indicator* I_{t-1}.

The experts

30th CIRET Conference, New York, October 2010

Constructing a conditional GDP fan chart with an application to French business survey data

Matthieu CORNEC
INSEE Business Surveys Unit

Abstract

Among economic forecasters, it has become a more common practice to provide point projection with a density forecast. This realistic view acknowledges that nobody can predict future evolution of the economic outlook with absolute certainty. Interval confidence and density forecasts have thus become useful tools to describe in probability terms the uncertainty inherent to any point forecast (for a review see Tay and Wallis 2000). Since 1996, the Central Bank of England (CBE) has published a density forecast of inflation in its quarterly *Inflation Report*, so called "fan chart". More recently, INSEE has also published a fan chart of its Gross Domestic Production (GDP) prediction in the *Note de Conjoncture*. Both methodologies estimate parameters of exponential families on the sample of past errors. They thus suffer from some drawbacks. First, INSEE fan chart is *unconditional* which means that whatever the economic outlook is, the magnitude of the displayed uncertainty is the same. On the contrary, it is common belief among practitioners that the forecasting exercise highly depends on the state of the economy, especially during crisis. A second limitation is that CBE fan chart is not reproducible as it introduces subjectivity. Eventually, another inadequacy is the parametric shape of the ditribution. In this paper, we tackle those issues to provide a reproducible conditional and non-parametric fan chart. For this, following Taylor 1999, we combine quantile regression approach together with regularization techniques to display a density forecast conditional on the available information. In the same time, we build a Forecasting Risk Index associated to this fan chart to measure the intrinsic difficulty of the forecasting exercise. The proposed methodology is applied to the French economy. Using balances of different business surveys, the GDP fan chart captures efficiently the growth stall during the crisis on an real-time basis. Moreover, our Forecasting Risk Index increased substantially in this period of turbulence, showing signs of growing uncertainty.

Key Words: density forecast, quantile regression, business tendency surveys, fan chart.

JEL Classification: E32, E37, E66, C22

In this paper, M. Cornec proposed simple econometric models that can be seen as experts :

The experts

30th CIRET Conference, New York, October 2010

Constructing a conditional GDP fan chart with an application to French business survey data

Matthieu CORNEC
INSEE Business Surveys Unit

Abstract

Among economic forecasters, it has become a more common practice to provide point projection with a density forecast. This realistic view acknowledges that nobody can predict future evolution of the economic outlook with absolute certainty. Interval confidence and density forecasts have thus become useful tools to describe in probability terms the uncertainty inherent to any point forecast (for a review see Tay and Wallis 2000). Since 1996, the Central Bank of England (CBE) has published a density forecast of inflation in its quarterly *Inflation Report*, so called "fan chart". More recently, INSEE has also published a fan chart of its Gross Domestic Production (GDP) prediction in the *Note de Conjoncture*. Both methodologies estimate parameters of exponential families on the sample of past errors. They thus suffer from some drawbacks. First, INSEE fan chart is *unconditional* which means that whatever the economic outlook is, the magnitude of the displayed uncertainty is the same. On the contrary, it is common belief among practitioners that the forecasting exercise highly depends on the state of the economy, especially during crisis. A second limitation is that CBE fan chart is not reproducible as it introduces subjectivity. Eventually, another inadequacy is the parametric shape of the ditribution. In this paper, we tackle those issues to provide a reproducible conditional and non-parametric fan chart. For this, following Taylor 1999, we combine quantile regression approach together with regularization techniques to display a density forecast conditional on the available information. In the same time, we build a Forecasting Risk Index associated to this fan chart to measure the intrinsic difficulty of the forecasting exercise. The proposed methodology is applied to the French economy. Using balances of different business surveys, the GDP fan chart captures efficiently the growth stall during the crisis on an real-time basis. Moreover, our Forecasting Risk Index increased substantially in this period of turbulence, showing signs of growing uncertainty.

Key Words: density forecast, quantile regression, business tendency surveys, fan chart.

JEL Classification: E32, E37, E66, C22

In this paper, M. Cornec proposed simple econometric models that can be seen as experts :

1. forecasts similars to the ones by the most complex models used by INSEE.

The experts

30[th] CIRET Conference, New York, October 2010

Constructing a conditional GDP fan chart with an application to French business survey data

Matthieu CORNEC
INSEE Business Surveys Unit

Abstract

Among economic forecasters, it has become a more common practice to provide point projection with a density forecast. This realistic view acknowledges that nobody can predict future evolution of the economic outlook with absolute certainty. Interval confidence and density forecasts have thus become useful tools to describe in probability terms the uncertainty inherent to any point forecast (for a review see Tay and Wallis 2000). Since 1996, the Central Bank of England (CBE) has published a density forecast of inflation in its quarterly *Inflation Report*, so called "fan chart". More recently, INSEE has also published a fan chart of its Gross Domestic Production (GDP) prediction in the *Note de Conjoncture*. Both methodologies estimate parameters of exponential families on the sample of past errors. They thus suffer from some drawbacks. First, INSEE fan chart is *unconditional* which means that whatever the economic outlook is, the magnitude of the displayed uncertainty is the same. On the contrary, it is common belief among practitioners that the forecasting exercise highly depends on the state of the economy, especially during crisis. A second limitation is that CBE fan chart is not reproducible as it introduces subjectivity. Eventually, another inadequacy is the parametric shape of the ditribution. In this paper, we tackle those issues to provide a reproducible conditional and non-parametric fan chart. For this, following Taylor 1999, we combine quantile regression approach together with regularization techniques to display a density forecast conditional on the available information. In the same time, we build a Forecasting Risk Index associated to this fan chart to measure the intrinsic difficulty of the forecasting exercise. The proposed methodology is applied to the French economy. Using balances of different business surveys, the GDP fan chart captures efficiently the growth stall during the crisis on an real-time basis. Moreover, our Forecasting Risk Index increased substantially in this period of turbulence, showing signs of growing uncertainty.

Key Words: density forecast, quantile regression, business tendency surveys, fan chart.

JEL Classification: E32, E37, E66, C22

In this paper, M. Cornec proposed simple econometric models that can be seen as experts :

1. forecasts similars to the ones by the most complex models used by INSEE.
2. when $\widehat{\Delta \mathrm{GDP}}_t^f$ is small, the accuracy deteriorates.

Forecastings

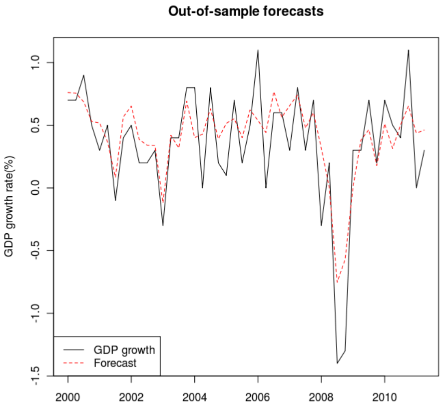


Figure – Using M. Cornec's predictor and the absolute loss function $\ell(x, x') = |x - x'|$.

Confidence intervals

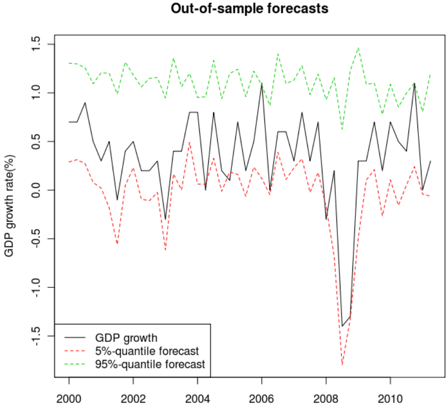


Figure – Using quantile loss $\ell(x, x') = (x - x')(\tau - 1(x - x' < 0))$.

The infinite case

Infinite family of predictors $f_\theta : \mathcal{X} \rightarrow \mathbb{R}$, $\theta \in \Theta$.

The infinite case

Infinite family of predictors $f_\theta : \mathcal{X} \to \mathbb{R}$, $\theta \in \Theta$.

- learning rate $\eta > 0$.

The infinite case

Infinite family of predictors $f_\theta : \mathcal{X} \to \mathbb{R}$, $\theta \in \Theta$.

- learning rate $\eta > 0$.
- prior distribution on Θ, $p_1 = \pi$.

The infinite case

Infinite family of predictors $f_\theta : \mathcal{X} \to \mathbb{R}$, $\theta \in \Theta$.

- learning rate $\eta > 0$.
- prior distribution on Θ, $p_1 = \pi$.

Algorithm 3 EWA (general case)

1: **for** $i = 1, 2, \dots$ **do**
2: $\hat{y}_t = \int f_\theta(x_t) p_t(\mathrm{d}\theta)$,
3: y_t revealed, update $p_{t+1}(\mathrm{d}\theta) = \frac{\exp[-\eta \ell(f_\theta(x_t), y_t)] p_t(\mathrm{d}\theta)}{\int \exp[-\eta \ell(f_\vartheta(x_t), y_t)] p_t(\mathrm{d}\vartheta)}$.
4: **end for**

Regret bound in the general case

Theorem

Assume that $0 \leq \ell \leq C$. Then

$$\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) \leq \inf_{p} \left[\int \sum_{t=1}^{T} \ell(f_\vartheta(x_t), y_t) p(\mathrm{d}\vartheta) + \frac{\eta C^2 T}{8} + \frac{KL(p||\pi)}{\eta} \right].$$

Regret bound in the general case

Theorem

Assume that $0 \leq \ell \leq C$. Then

$$\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) \leq \inf_{p} \left[\int \sum_{t=1}^{T} \ell(f_\vartheta(x_t), y_t) p(\mathrm{d}\vartheta) + \frac{\eta C^2 T}{8} + \frac{KL(p||\pi)}{\eta} \right].$$

- the inf. is with respect to any probability distribution p.

Regret bound in the general case

Theorem

Assume that $0 \leq \ell \leq C$. Then

$$\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) \leq \inf_{p} \left[\int \sum_{t=1}^{T} \ell(f_\vartheta(x_t), y_t) p(\mathrm{d}\vartheta) + \frac{\eta C^2 T}{8} + \frac{KL(p||\pi)}{\eta} \right].$$

- the inf. is with respect to any probability distribution p.
- $KL(p||\pi)$ is the Kullback divergence.

Reminder

The Kullback divergence, or relative entropy :

$$KL(p||\pi) = \begin{cases} \int \log \left[\frac{\mathrm{d}p}{\mathrm{d}\pi}(\vartheta) \right] p(\mathrm{d}\vartheta) \text{ if } p \ll \pi, \\ +\infty \text{ otherwise.} \end{cases}$$

Reminder

The Kullback divergence, or relative entropy :

$$KL(p||\pi) = \begin{cases} \int \log\left[\frac{\mathrm{d}p}{\mathrm{d}\pi}(\vartheta)\right] p(\mathrm{d}\vartheta) \text{ if } p \ll \pi, \\ +\infty \text{ otherwise.} \end{cases}$$

When π is uniform on $\{1, \dots, M\}$ and when p is the Dirac mass on $i \in \{1, \dots, M\}$ then

$$KL(p||\pi) = \log(M)$$

so the result in the finite case is indeed a corollary of the general result.

Link with Bayesian statistics

$$
\begin{aligned}
p_{t+1}(\mathrm{d}\theta) & \propto \exp[-\eta\ell(f_\theta(x_t), y_t)] p_t(\mathrm{d}\theta \\
& \propto \left\{ \prod_{i=1}^{t} \exp[-\eta\ell(f_\theta(x_i), y_i)] \right\} \pi(\mathrm{d}\theta).
\end{aligned}
$$

Link with Bayesian statistics

$$
\begin{aligned}
p_{t+1}(\mathrm{d}\theta) & \propto \exp[-\eta\ell(f_\theta(x_t), y_t)] p_t(\mathrm{d}\theta \\
& \propto \left\{ \prod_{i=1}^{t} \exp[-\eta\ell(f_\theta(x_i), y_i)] \right\} \pi(\mathrm{d}\theta).
\end{aligned}
$$

Assume x_t deterministic, $y_t \sim \mathcal{N}(f_{\theta^*}(x_t), \sigma^2)$, take $\eta = 1$ and $\ell(y, y') = \frac{(y-y')^2}{2\sigma^2}$. Then the likelihood is given by

$$
\mathcal{L}(\theta, y_1, \dots, y_t) = \prod_{i=1}^{t} \exp[-\eta\ell(f_\theta(x_i), y_i)]
$$

Link with Bayesian statistics

$$
\begin{aligned}
p_{t+1}(\mathrm{d}\theta) & \propto \exp[-\eta \ell(f_\theta(x_t), y_t)] p_t(\mathrm{d}\theta \\
& \propto \left\{ \prod_{i=1}^{t} \exp[-\eta \ell(f_\theta(x_i), y_i)] \right\} \pi(\mathrm{d}\theta).
\end{aligned}
$$

Assume x_t deterministic, $y_t \sim \mathcal{N}(f_{\theta^*}(x_t), \sigma^2)$, take $\eta = 1$ and $\ell(y, y') = \frac{(y-y')^2}{2\sigma^2}$. Then the likelihood is given by

$$
\mathcal{L}(\theta, y_1, \dots, y_t) = \prod_{i=1}^{t} \exp[-\eta \ell(f_\theta(x_i), y_i)]
$$

$$
\Rightarrow p_{t+1}(\mathrm{d}\theta) \propto \mathcal{L}(\theta, y_1, \dots, y_t) \pi(\mathrm{d}\theta) \propto \pi(\theta | y_1, \dots, y_t).
$$

Online gradient and online variational inference

Context

1. 1. x_1 given
   2. predict y_1 : $\hat{y}_1$
   3. y_1 is revealed, we suffer loss $\ell(\hat{y}_1, y_1)$.
2. 1. x_2 given
   2. predict y_2 : $\hat{y}_2$
   3. y_2 revealed, we suffer loss $\ell(\hat{y}_2, y_2)$.
3. ...

and this time, we would like to use a model like

$$\hat{y}_t = \langle \theta_t, x_t \rangle .$$

Context

1. 1. x_1 given
   2. predict y_1 : $\hat{y}_1$
   3. y_1 is revealed, we suffer loss $\ell(\hat{y}_1, y_1)$.
2. 1. x_2 given
   2. predict y_2 : $\hat{y}_2$
   3. y_2 revealed, we suffer loss $\ell(\hat{y}_2, y_2)$.
3. ...

and this time, we would like to use a model like

$$\hat{y}_t = \langle \theta_t, x_t \rangle .$$

More generally, we study $\hat{y}_t = g(\theta_t, x_t)$.

OGA

Input :

- learning rate $\eta > 0$,
- starting value θ_1, often 0,
- link function $g(\theta, x)$ convex in θ, loss $\ell(y, y')$ convex in y.

OGA

Input :

- learning rate $\eta > 0$,
- starting value θ_1, often 0,
- link function $g(\theta, x)$ convex in θ, loss $\ell(y, y')$ convex in y.

Algorithm 3 OGA

1: **for** $i = 1, 2, \ldots$ **do**
2: Predict $\hat{y}_t = g(\theta_t, x_t)$,
3: y_t revealed, update

$$\theta_{t+1} = \theta_t - \eta \frac{\partial}{\partial \theta} \ell(g(\theta_t, x_t), y_t).$$

4: **end for**

OGA - theorem

Theorem

Assume that $\theta \mapsto \ell(g(\theta, x), y)$ is convex and L-Lispchitz with respect to θ, then :

$$\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) \leq \inf_{\theta \in \mathbb{R}^d} \left[\sum_{t=1}^{T} \ell(g(\theta, x_t), y_t) + \frac{\|\theta\|^2}{2\eta} + \eta T L^2. \right]$$

OGA - theorem

Theorem

Assume that $\theta \mapsto \ell(g(\theta, x), y)$ is convex and L-Lispchitz with respect to θ, then :

$$\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) \leq \inf_{\theta \in \mathbb{R}^d} \left[\sum_{t=1}^{T} \ell(g(\theta, x_t), y_t) + \frac{\|\theta\|^2}{2\eta} + \eta T L^2. \right]$$

Choose $B > 0$ and $\eta = \frac{B}{L\sqrt{2T}}$ to obtain :

$$\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) \leq \inf_{\|\theta\| \leq B} \sum_{t=1}^{T} \ell(g(\theta, x_t), y_t) + BL\sqrt{2T}.$$

OGA - theorem

Theorem

Assume that $\theta \mapsto \ell(g(\theta, x), y)$ is convex and L-Lispchitz with respect to θ, then :

$$\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) \leq \inf_{\theta \in \mathbb{R}^d} \left[\sum_{t=1}^{T} \ell(g(\theta, x_t), y_t) + \frac{\|\theta\|^2}{2\eta} + \eta T L^2 . \right]$$

Choose $B > 0$ and $\eta = \frac{B}{L\sqrt{2T}}$ to obtain :

$$\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) \leq \inf_{\|\theta\| \leq B} \sum_{t=1}^{T} \ell(g(\theta, x_t), y_t) + BL\sqrt{2T}.$$

The choice $\eta_t = \frac{B}{L\sqrt{2t}}$ leads to similar rate with worse constant.

Example : glass identification

Example : glass identification

Glass identification

Dataset from the Machine Learning Repository :
https ://archive.ics.uci.edu/ml/index.php
$y_t = 1$ (window) or $y_t = -1$ (non window) ; attributes : x_t (chemical composition).

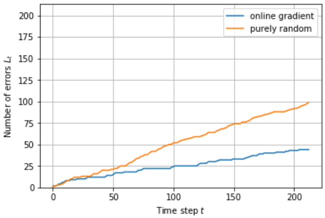
Glass identification

Dataset from the Machine Learning Repository :
https ://archive.ics.uci.edu/ml/index.php
$y_t = 1$ (window) or $y_t = -1$ (non window) ; attributes : x_t (chemical composition).
Prediction : $\hat{y}_t = \langle \theta_t, x_t \rangle$. Loss : $\ell(\hat{y}, y) = (1 - \hat{y}y)_+$.

Glass identification

Dataset from the Machine Learning Repository :
https ://archive.ics.uci.edu/ml/index.php
$y_t = 1$ (window) or $y_t = -1$ (non window) ; attributes : x_t (chemical composition).
Prediction : $\hat{y}_t = \langle \theta_t, x_t \rangle$. Loss : $\ell(\hat{y}, y) = (1 - \hat{y}y)_+$.

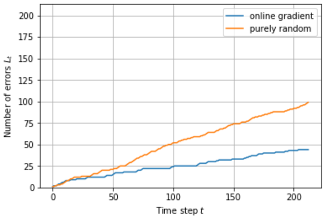
$$\frac{\partial}{\partial \theta} \ell(< \theta_t, x_t >, y_t) = \begin{cases} -y_t x_t \text{ if } \mathrm{sign}(< \theta_t, x_t >) \neq y_t \\ 0 \text{ otherwise,} \end{cases}$$

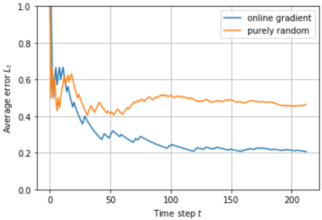


Glass identification

Dataset from the Machine Learning Repository :
https ://archive.ics.uci.edu/ml/index.php
$y_t = 1$ (window) or $y_t = -1$ (non window) ; attributes : x_t (chemical composition).
Prediction : $\hat{y}_t = \langle \theta_t, x_t \rangle$. Loss : $\ell(\hat{y}, y) = (1 - \hat{y}y)_+$.

$$\frac{\partial}{\partial \theta} \ell(< \theta_t, x_t >, y_t) = \begin{cases} -y_t x_t \text{ if } \mathrm{sign}(< \theta_t, x_t >) \neq y_t \\ 0 \text{ otherwise,} \end{cases}$$

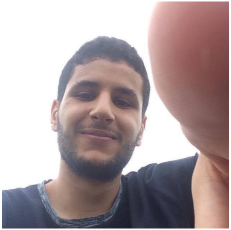




Online gradient and online variational inference

1. Setting of the problem
   - Definitions
   - Toy examples
   - The regret
2. Exponentially Weighted Aggregation (EWA)
   - Prediction with expert advice
   - Examples : air quality / GDP growth
   - The infinite case
3. Online gradient and online variational inference
   - Online gradient algorithm
   - Example : glass identification
   - Online variational inference

Co-authors



ENSAE
IP PARIS

RIKEN
AIP Center for Advanced Intelligence Project

Approximate Bayesian Inference team

https : //emtiyaz.github.io/

Chérief-Abdellatif, B.-E., Alquier, P. and Khan, M. E. (2019). A Generalization Bound for Online Variational Inference. *ACML*.

Bayes and computational efficiency

(Generalized) Bayes rule non feasible in complex models

$$\exp\left[-\eta \sum_{t=1}^{T-1} \ell(f_\theta(x_t), y_t)\right] \pi(\theta)$$

Bayes and computational efficiency

(Generalized) Bayes rule non feasible in complex models

$$\exp\left[-\eta \sum_{t=1}^{T-1} \ell(f_\theta(x_t), y_t)\right] \pi(\theta) \leftarrow \min_{p} \sum_{t=1}^{T-1} \mathbb{E}_{\theta\sim p}\left[\ell(f_\theta(x_t), y_t)\right] + \frac{KL(p\|\pi)}{\eta}$$

Bayes and computational efficiency

(Generalized) Bayes rule non feasible in complex models

$$\exp\left[-\eta \sum_{t=1}^{T-1} \ell(f_\theta(x_t), y_t)\right] \pi(\theta) \leftarrow \min_{p} \sum_{t=1}^{T-1} \mathbb{E}_{\theta\sim p}\left[\ell(f_\theta(x_t), y_t)\right] + \frac{KL(p\|\pi)}{\eta}$$

We will propose a feasible version thanks to the ideas in :

Online variational inference

- propose a parametric family of probability distributions : (q_μ).
- approximate p_t by some q_{μ_t}.

Online variational inference

- propose a parametric family of probability distributions : (q_μ).
- approximate p_t by some q_{μ_t}.

How ? We could for example perform an online gradient algorithm on μ for :

$$\mathbb{E}_{\theta \sim q_\mu}[\ell(f_\theta(x_t), y_t)].$$

Online variational inference

- propose a parametric family of probability distributions : (q_μ).
- approximate p_t by some q_{μ_t}.

How ? We could for example perform an online gradient algorithm on μ for :

$$\mathbb{E}_{\theta\sim q_\mu}[\ell(f_\theta(x_t),y_t)].$$

But there is a better thing to do...

Reminder on online gradient

$$\hat{y}_t = f_{\theta_t}(x_t) \quad \text{and} \quad \theta_{t+1} = \theta_t - \eta \nabla_\theta \ell(f_{\theta_t}(x_t), y_t).$$

Reminder on online gradient

$$\hat{y}_t = f_{\theta_t}(x_t) \quad \text{and} \quad \theta_{t+1} = \theta_t - \eta \nabla_\theta \ell_t(\theta_t).$$

Reminder on online gradient

$$\hat{y}_t = f_{\theta_t}(x_t) \quad \text{and} \quad \theta_{t+1} = \theta_t - \eta \nabla_\theta \ell_t(\theta_t).$$

Note that θ_{t+1} can be obtained by :

1. $$\min_\theta \left\{ \left\langle \theta, \sum_{s=1}^{t} \nabla_\theta \ell_s(\theta_s) \right\rangle + \frac{\|\theta - \theta_1\|^2}{2\eta} \right\},$$
2. $$\min_\theta \left\{ \left\langle \theta, \nabla_\theta \ell_t(\theta_t) \right\rangle + \frac{\|\theta - \theta_t\|^2}{2\eta} \right\}.$$

Two options for online VI

Two options for online VI

1. Sequential Variational Approximation (SVA) :

$$\theta_{t+1} = \arg\min_{\theta} \left\{ \left\langle \theta, \sum_{s=1}^{t} \nabla_{\theta} \ell_s(\theta_s) \right\rangle + \frac{\|\theta - \theta_1\|^2}{2\eta} \right\},$$

2. Streaming Variational Bayes (SVB) :

$$\theta_{t+1} = \arg\min_{\theta} \left\{ \left\langle \theta, \nabla_{\theta} \ell_t(\theta_t) \right\rangle + \frac{\|\theta - \theta_t\|^2}{2\eta} \right\},$$

Two options for online VI

1. Sequential Variational Approximation (SVA) :

$$\theta_{t+1} = \arg\min_{\theta} \left\{ \left\langle \theta, \sum_{s=1}^{t} \nabla_{\theta} \ell_s(\theta_s) \right\rangle + \frac{\|\theta - \theta_1\|^2}{2\eta} \right\},$$

$$\mu_{t+1} = \arg\min_{\mu} \left\{ \left\langle \mu, \sum_{s=1}^{t} \nabla_{\mu} \mathbb{E}_{\theta \sim q_{\mu_s}} [\ell_s(\theta)] \right\rangle + \frac{KL(q_{\mu}, \pi)}{\eta} \right\}.$$

2. Streaming Variational Bayes (SVB) :

$$\theta_{t+1} = \arg\min_{\theta} \left\{ \left\langle \theta, \nabla_{\theta} \ell_t(\theta_t) \right\rangle + \frac{\|\theta - \theta_t\|^2}{2\eta} \right\},$$

Two options for online VI

1. Sequential Variational Approximation (SVA) :

$$\theta_{t+1} = \arg\min_{\theta} \left\{ \left\langle \theta, \sum_{s=1}^{t} \nabla_\theta \ell_s(\theta_s) \right\rangle + \frac{\|\theta - \theta_1\|^2}{2\eta} \right\},$$

$$\mu_{t+1} = \arg\min_{\mu} \left\{ \left\langle \mu, \sum_{s=1}^{t} \nabla_\mu \mathbb{E}_{\theta \sim q_{\mu_s}} [\ell_s(\theta)] \right\rangle + \frac{KL(q_\mu, \pi)}{\eta} \right\}.$$

2. Streaming Variational Bayes (SVB) :

$$\theta_{t+1} = \arg\min_{\theta} \left\{ \left\langle \theta, \nabla_\theta \ell_t(\theta_t) \right\rangle + \frac{\|\theta - \theta_t\|^2}{2\eta} \right\},$$

$$\mu_{t+1} = \arg\min_{\mu} \left\{ \left\langle \mu, \nabla_\mu \mathbb{E}_{\theta \sim q_{\mu_t}} [\ell_t(\theta)] \right\rangle + \frac{KL(q_\mu, q_{\mu_t})}{\eta} \right\}.$$

SVA & SVB are tractable, and not equivalent

Example : Gaussian prior $\theta \sim \pi = \mathcal{N}(0, s^2 I)$ and mean-field Gaussian approximation, $\mu = (m, \sigma)$.

$$\mathrm{SVA}: m_{t+1} \leftarrow m_t - \eta s^2 \bar{g}_{m_t}, \qquad g_{t+1} \leftarrow g_t + \bar{g}_{\sigma_t},$$
$$\sigma_{t+1} \leftarrow h\left(\eta s g_{t+1}\right) s,$$
$$\mathrm{SVB}: m_{t+1} \leftarrow m_t - \eta \sigma_t^2 \bar{g}_{m_t},$$
$$\sigma_{t+1} \leftarrow \sigma_t h\left(\eta \sigma_t \bar{g}_{\sigma_t}\right)$$

where $h(x) := \sqrt{1 + x^2} - x$ is applied componentwise, as well as the multiplication of two vectors, and

$$\bar{g}_{m_t} = \frac{\partial}{\partial m} \mathbb{E}_{\theta \sim \pi_{m_t, \sigma_t}}[\ell_t(\theta)],$$
$$\bar{g}_{\sigma_t} = \frac{\partial}{\partial \sigma} \mathbb{E}_{\theta \sim \pi_{m_t, \sigma_t}}[\ell_t(\theta)].$$

Theoretical analysis of SVA

Theorem 1

Under convexity and L-Lipschitz assumption on the loss, under α-strong convexity assumption on the KL term, SVA leads to

$$\sum_{t=1}^{T} \mathbb{E}_{\theta\sim q_{\mu_t}}[\ell_t(\theta)] \leq \inf_{\mu\in M} \left\{ \sum_{t=1}^{T} \mathbb{E}_{\theta\sim q_{\mu}}[\ell_t(\theta)] + \frac{\eta L^2 T}{\alpha} + \frac{KL(q_\mu,\pi)}{\eta} \right\}.$$

Theoretical analysis of SVA

Theorem 1

Under convexity and L-Lipschitz assumption on the loss, under α-strong convexity assumption on the KL term, SVA leads to

$$\sum_{t=1}^{T} \mathbb{E}_{\theta\sim q_{\mu_t}}[\ell_t(\theta)] \leq \inf_{\mu\in M} \left\{ \sum_{t=1}^{T} \mathbb{E}_{\theta\sim q_{\mu}}[\ell_t(\theta)] + \frac{\eta L^2 T}{\alpha} + \frac{KL(q_\mu,\pi)}{\eta} \right\}.$$

Application to Gaussian approximation leads to

$$\sum_{t=1}^{T} \mathbb{E}_{\theta\sim q_{\mu_t}}[\ell_t(\theta)] \leq \inf_{\theta} \sum_{t=1}^{T} \ell_t(\theta) + (1+o(1))\frac{2L}{\alpha}\sqrt{dT\log(T)}.$$

Theoretical analysis of SVB

Theorem 2

Using Gaussian approximations, assuming the loss is convex, L-Lipschitz and the parameter space bounded (diameter $= D$), SVB with adequate η leads to

$$\sum_{t=1}^{T} \ell_t\left(\mathbb{E}_{\theta\sim q_{\mu_t}}(\theta)\right) \leq \inf_{\theta} \sum_{t=1}^{T} \ell_t(\theta) + DL\sqrt{2T}.$$

Theoretical analysis of SVB

Theorem 2

Using Gaussian approximations, assuming the loss is convex, L-Lipschitz and the parameter space bounded (diameter $= D$), SVB with adequate η leads to

$$\sum_{t=1}^{T} \ell_t\Big(\mathbb{E}_{\theta\sim q_{\mu_t}}(\theta)\Big) \leq \inf_{\theta} \sum_{t=1}^{T} \ell_t(\theta) + DL\sqrt{2T}.$$

If, moreover, the loss is H-strongly convex,

$$\sum_{t=1}^{T} \ell_t\Big(\mathbb{E}_{\theta\sim q_{\mu_t}}(\theta)\Big) \leq \inf_{\theta} \sum_{t=1}^{T} \ell_t(\theta) + \frac{L^2(1+\log(T))}{H}.$$

Test on a simulated dataset

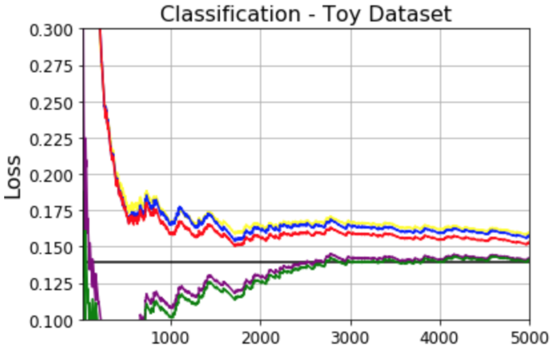


Figure – Average cumulative losses on different datasets for classification and regression tasks with OGA (yellow), OGA-EL (red), SVA (blue), SVB (purple) and NGVI (green).

Test on the Breast dataset

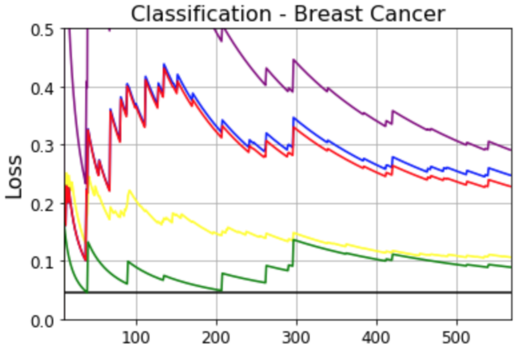


Figure – Average cumulative losses on different datasets for classification and regression tasks with OGA (yellow), OGA-EL (red), SVA (blue), SVB (purple) and NGVI (green).

Open questions

Open questions

1. Analysis of SVB in the general case.

Open questions

1. Analysis of SVB in the general case.
2. Analysis of the uncertainty quantification.

Open questions

1. Analysis of SVB in the general case.
2. Analysis of the uncertainty quantification.
3. NGVI is the next step in going closer to algorithms used to train Neural Networks with Bayesian principles. But being based on a different parametrization, it does not satisfy our convexity assumption...

Open questions

1. Analysis of SVB in the general case.
2. Analysis of the uncertainty quantification.
3. NGVI is the next step in going closer to algorithms used to train Neural Networks with Bayesian principles. But being based on a different parametrization, it does not satisfy our convexity assumption...

   Uses exponential family approximations $\{q_\mu, \mu \in M\}$ where m is the mean parameter. Denoting λ the natural parameter (with $\lambda = F(\mu)$),

$$\lambda_{t+1} = (1-\rho)\lambda_t + \rho \nabla_\mu \mathbb{E}_{\theta \sim q_{\mu_t}} \left[\ell_t(\theta)\right],$$

M. E. Khan, D. Nielsen (2018). *Fast yet Simple Natural-Gradient Descent for Variational Inference in Complex Models.* ISITA.

Thank you !