Introduction to Sequential Prediction

Pierre Alquier

École nationale de la statistique et de l'administration économique
ENSAE ParisTech
université PARIS-SACLAY

London Business School, Nov. 14, 2018

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 - $y_t \in \{0, 1\}$

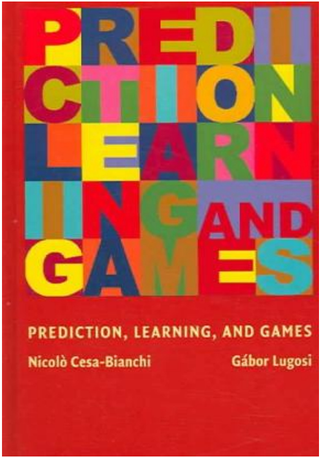
1.
   1. x_1 given
   2. predict y_1 : $\hat{y}_1$
   3. y_1 is revealed
2.
   1. x_2 given
   2. predict y_2 : $\hat{y}_2$
   3. y_2 revealed
3.
   1. x_3 given
   2. predict y_3 : $\hat{y}_3$
   3. y_3 revealed
4. ...

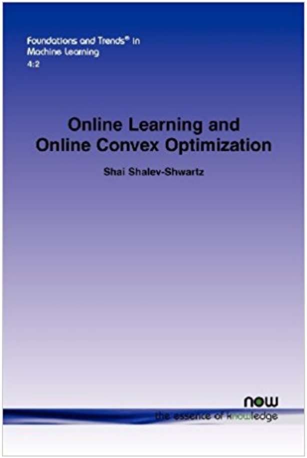
Objective : make sure that we learn to predict well **as soon as possible**. Keep

$$\sum_{t=1}^{T} \mathbf{1}(\hat{Y}_t \neq Y_t)$$

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

1. Setting of the problem
   - Definitions
   - Toy examples
   - The regret
2. Exponentially Weighted Aggregation (EWA)
   - Prediction with expert advice
   - Further topics
   - The infinite case
3. Open questions
   - Confidence intervals
   - Fast algorithms
   - More open questions

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 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 loss function ℓ.

Classical examples :

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 loss function ℓ.

Classical examples :

- $\ell(y, y') = a\mathbf{1}(y = 1, y' = 0) + b\mathbf{1}(y = 0, y' = 1)$ for 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.
- loss incurred at time t : $\ell(\hat{y}_t, y_t)$ for some real-valued loss function ℓ.

Classical examples :

- $\ell(y, y') = a\mathbf{1}(y = 1, y' = 0) + b\mathbf{1}(y = 0, y' = 1)$ for classification,
- $\ell(y, y') = (y - y')^2$ or $\ell(y, y') = |y - y'|$ for regression,

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 loss function ℓ.

Classical examples :

- $\ell(y, y') = a\mathbf{1}(y = 1, y' = 0) + b\mathbf{1}(y = 0, y' = 1)$ for classification,
- $\ell(y, y') = (y - y')^2$ or $\ell(y, y') = |y - y'|$ for regression,
- $\ldots$

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 $(y_1, \dots, y_{t-1})$. We can also use randomization if necessary.

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 $(y_1, \dots, y_{t-1})$. We can also use randomization if necessary.
- 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 $(y_1, \dots, y_{t-1})$. We can also use randomization if necessary.
- It must be computationnally feasible.
- We can use expert advice.

What performance can we achieve in this setting ?

Consider binary classification with $\ell(y, y') = \mathbf{1}(y \neq y')$, 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 $\ell(y, y') = \mathbf{1}(y \neq y')$, 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 $\ell(y, y') = \mathbf{1}(y \neq y')$, 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

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

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

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

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

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

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

- 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

Assume that $\ell(\cdot, y)$ is convex. Input :

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

EWA strategy

Assume that $\ell(\cdot, y)$ is convex. 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 - convex case

Theorem

Assume that $\ell(\cdot,\cdot) \in [0, C]$ and $\ell(\cdot, y)$ is convex. Then

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

EWA - convex case

Theorem

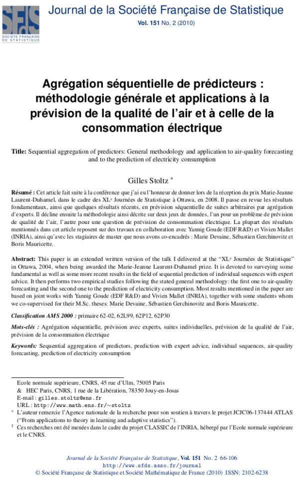
Assume that $\ell(\cdot,\cdot) \in [0, C]$ and $\ell(\cdot, y)$ is convex. Then

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

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

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

An example : 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

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

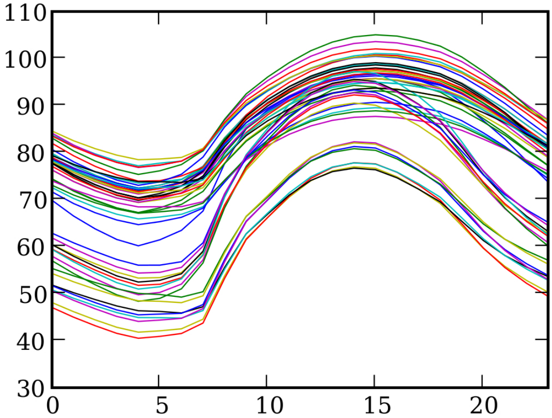


Figure – Predictions by the 48 experts for one day at one station.

Numerical performances

	RMSE
Best expert	22.43
Uniform mean	24.41
EWA	21.47

Figure – Numerical performances (RMSE).

Weights

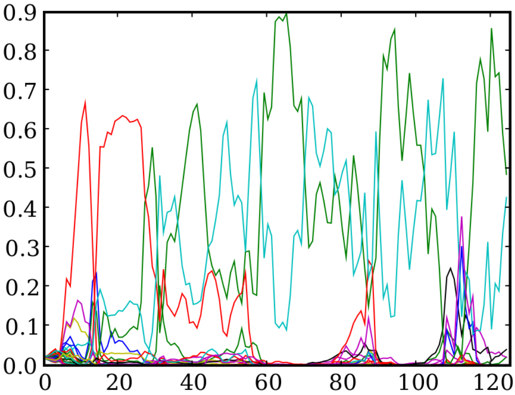


Figure – Evolution of the weights $p_i(t)$ w.r.t t.

Further topics

Better regret bounds

We obtained

$$\mathrm{Regret}(T) = \mathcal{O}(\sqrt{T \log(M)})$$

for EWA.

Further topics

Better regret bounds

We obtained

$$\mathrm{Regret}(T) = \mathcal{O}(\sqrt{T \log(M)})$$

for EWA. Under a stronger assumption (exp-concave loss ℓ),

$$\mathrm{Regret}(T) = \mathcal{O}(\log(M)).$$

Further topics

Better regret bounds

We obtained

$$\mathrm{Regret}(T) = \mathcal{O}(\sqrt{T \log(M)})$$

for EWA. Under a stronger assumption (exp-concave loss ℓ),

$$\mathrm{Regret}(T) = \mathcal{O}(\log(M)).$$

Other strategies

See the introduction by Shalev-Shwartz :

- online ridge regression,

Further topics

Better regret bounds

We obtained

$$\mathrm{Regret}(T) = \mathcal{O}(\sqrt{T \log(M)})$$

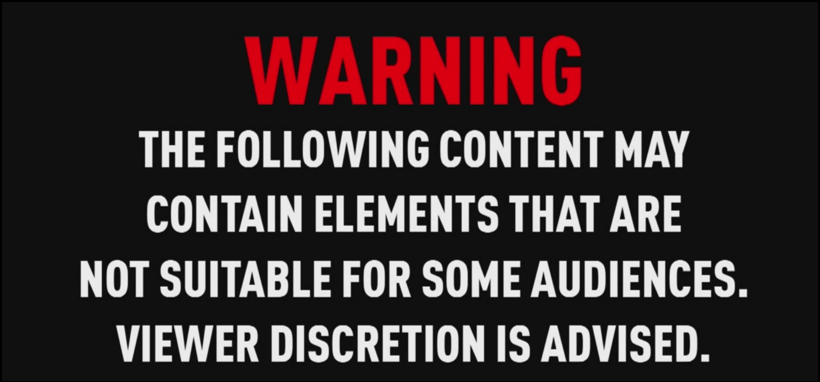
for EWA. Under a stronger assumption (exp-concave loss ℓ),

$$\mathrm{Regret}(T) = \mathcal{O}(\log(M)).$$

Other strategies

See the introduction by Shalev-Shwartz :

- online ridge regression, that is itself a special case of
- online gradient descent...


WARNING
THE FOLLOWING CONTENT MAY
CONTAIN ELEMENTS THAT ARE
NOT SUITABLE FOR SOME AUDIENCES.
VIEWER DISCRETION IS ADVISED.

The infinite case

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

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 Randomized EWA (general case)

1: **for** $i = 1, 2, \ldots$ **do**
2: Draw $\theta_t \sim p_t$, predict $\hat{y}_t = f_{\theta_t}(x_t)$,
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 $\ell(\cdot,\cdot) \in [0, C]$ (e.g. classification). Then

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

Regret bound in the general case

Theorem

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

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

- the expectation is w.r.t the algorithm. Convex case : possible to replace randomization by averaging.

Regret bound in the general case

Theorem

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

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

- the expectation is w.r.t the algorithm. Convex case : possible to replace randomization by averaging.
- the inf. is with respect to any probability distribution p.

Regret bound in the general case

Theorem

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

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

- the expectation is w.r.t the algorithm. Convex case : possible to replace randomization by averaging.
- the inf. is with respect to any probability distribution p.
- $\mathcal{K}(p,\pi)$ is the Kullback divergence.

Reminder

The Kullback divergence, or relative entropy :

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

Reminder

The Kullback divergence, or relative entropy :

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

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

$$\mathcal{K}(p,\pi) = \log(M)$$

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

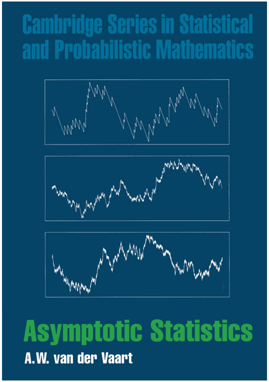
$$
\begin{aligned}
p_{t+1}(\mathrm{d}\theta) &\propto \exp[-\eta\ell(f_\theta(x_t), y_t)]p_t(\mathrm{d}\theta \\
&\propto \left\{\prod_{i=1}^{t} \exp[-\eta\ell(f_\theta(x_i), y_i)]\right\} \pi(\mathrm{d}\theta).
\end{aligned}
$$

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

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

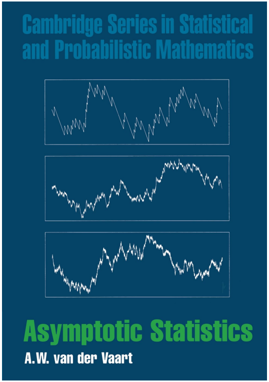
$$
\Rightarrow p_{t+1}(\mathrm{d}\theta) \propto \mathcal{L}(\theta, y_1, \dots, y_t)\pi(\mathrm{d}\theta) \propto \pi(\theta | y_1, \dots, y_t).
$$

Concentration of the posterior in Bayesian statistics



The asymptotic concentration of $\pi(\theta|y_1, \ldots, y_t)$ is a well-known topic. Requires :

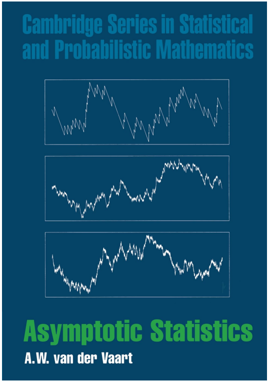
Concentration of the posterior in Bayesian statistics



The asymptotic concentration of $\pi(\theta|y_1, \ldots, y_t)$ is a well-known topic. Requires :

1. model well specified,

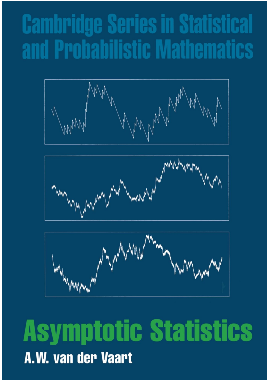
Concentration of the posterior in Bayesian statistics



The asymptotic concentration of $\pi(\theta|y_1, \dots, y_t)$ is a well-known topic. Requires :

1. model well specified,
2. a technical "test" condition,

Concentration of the posterior in Bayesian statistics



The asymptotic concentration of $\pi(\theta|y_1, \dots, y_t)$ is a well-known topic. Requires :

1. model well specified,
2. a technical "test" condition,
3. the prior mass condition : find r such that

$$\pi\{B(\theta^*, \varepsilon)\} \geq \mathrm{e}^{-r(\varepsilon)},$$

$$B(\theta, x) = \{\theta' : \|\theta - \theta'\| \leq x\}.$$

Explicit regret bound

Here, we did not assume the model is well speficied, nor the test condition, nor $\eta = 1$. Put $\pi_{\theta,\varepsilon}$ as π restricted to $B(\theta, \varepsilon)$.

$$\mathbb{E}\left(\sum_{t=1}^{T} \ell(\hat{y}_t, y_t)\right)$$

Explicit regret bound

Here, we did not assume the model is well speficied, nor the test condition, nor $\eta = 1$. Put $\pi_{\theta,\varepsilon}$ as π restricted to $B(\theta,\varepsilon)$.

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

Explicit regret bound

Here, we did not assume the model is well speficied, nor the test condition, nor $\eta = 1$. Put $\pi_{\theta,\varepsilon}$ as π restricted to $B(\theta, \varepsilon)$.

$$\mathbb{E}\left(\sum_{t=1}^{T} \ell(\hat{y}_t, y_t)\right)$$

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

$$\leq \inf_{\theta,\varepsilon} \left[\int \sum_{t=1}^{T} \ell(f_\vartheta(x_t), y_t) \pi_{\theta,\varepsilon}(\mathrm{d}\vartheta) + \frac{\eta C^2 T}{8} + \frac{\mathcal{K}(\pi_{\theta,\varepsilon},\pi)}{\eta}\right]$$

Explicit regret bound

Here, we did not assume the model is well speficied, nor the test condition, nor $\eta = 1$. Put $\pi_{\theta,\varepsilon}$ as π restricted to $B(\theta, \varepsilon)$.

$$\mathbb{E}\left(\sum_{t=1}^{T} \ell(\hat{y}_t, y_t)\right) \quad \text{(assume } \theta \mapsto \ell(f_\vartheta(x_t), y_t) \text{ is } L\text{-Lipschitz)}$$

$$\leq \inf_p \left[\int \sum_{t=1}^{T} \ell(f_\vartheta(x_t), y_t) p(\mathrm{d}\vartheta) + \frac{\eta C^2 T}{8} + \frac{\mathcal{K}(p,\pi)}{\eta}\right]$$

$$\leq \inf_{\theta,\varepsilon} \left[\int \sum_{t=1}^{T} \ell(f_\vartheta(x_t), y_t) \pi_{\theta,\varepsilon}(\mathrm{d}\vartheta) + \frac{\eta C^2 T}{8} + \frac{\mathcal{K}(\pi_{\theta,\varepsilon},\pi)}{\eta}\right]$$

$$\leq \inf_\theta \sum_{t=1}^{T} \ell(f_\theta(x_t), y_t) + \inf_\varepsilon \left(TL\varepsilon + \frac{\eta C^2 T}{8} + \frac{r(\varepsilon)}{\eta}\right)$$

Explicit regret bound

$$\mathbb{E}\left[\mathrm{Regret}(T)\right] = \inf_{\varepsilon>0}\left(T(\eta B^2 + L\varepsilon) + \frac{d \log\left(\frac{1}{\varepsilon}\right)}{\eta} \right).$$

Explicit regret bound

$$\mathbb{E}\left[\mathrm{Regret}(T)\right] = \inf_{\varepsilon>0}\left(T(\eta B^2 + L\varepsilon) + \frac{d\log\left(\frac{1}{\varepsilon}\right)}{\eta}\right).$$

The choice $\varepsilon = d/(TL\eta)$ and $\eta = \sqrt{d/(TB^2)}$ leads to the regret bound

$$\mathbb{E}\left[\mathrm{Regret}(T)\right] = \leq B\sqrt{dT\left[2+\log\left(\frac{LT}{Bd}\right)\right]}.$$

Open questions

Example - 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.

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 : $\Delta \mathrm{GDP}_t$.

INSEE
INSTITUT NATIONAL
DE LA STATISTIQUE
ET DES ÉTUDES
ÉCONOMIQUES

Available from INSEE :

GDP growth forecasting

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

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.

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 : $\Delta\mathrm{GDP}_t$.

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}.

M. Cornec's predictors

$$\widehat{\Delta\mathrm{GDP}}_t^f = \alpha + \beta \Delta\mathrm{GDP}_{t-1} + \gamma I_{t-1} + \delta (I_{t-1} - I_{t-2}) |I_{t-1} - I_{t-2}|$$

M. Cornec's predictors

$$\widehat{\Delta \mathrm{GDP}}_t^f = \alpha + \beta \Delta \mathrm{GDP}_{t-1} + \gamma I_{t-1} + \delta (I_{t-1} - I_{t-2}) |I_{t-1} - I_{t-2}|$$

proposed by

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

M. Cornec's predictors

$$\widehat{\Delta \mathrm{GDP}}_t^f = \alpha + \beta \Delta \mathrm{GDP}_{t-1} + \gamma I_{t-1} + \delta (I_{t-1} - I_{t-2}) |I_{t-1} - I_{t-2}|$$

proposed by

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

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

M. Cornec's predictors

$$\widehat{\Delta\mathrm{GDP}}_t^f = \alpha + \beta \Delta\mathrm{GDP}_{t-1} + \gamma I_{t-1} + \delta (I_{t-1} - I_{t-2}) | I_{t-1} - I_{t-2} |$$

proposed by

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

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

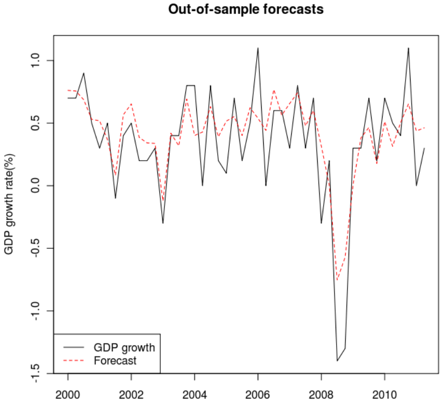


Figure – Using M. Cornec's predictor and the absolute loss function $\ell(x, x') = |x - x'|$.

Confidence intervals

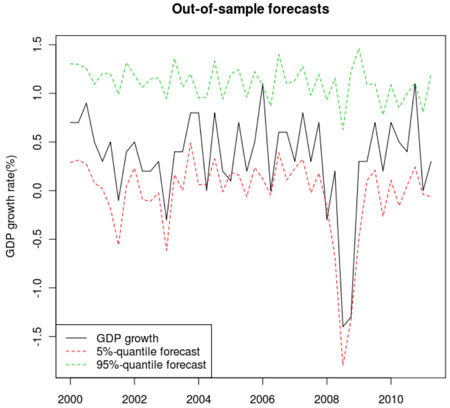


Figure – Using quantile loss $\ell(x, x') = (x - x')(\tau - \mathbf{1}(x - x' < 0))$.

Matthieu Cornec - Xiaoyin Li

R. Deswarte's algorithm

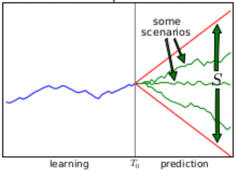
Algorithm 15 Methodology

Preliminaries:
Observe $(y_0, \ldots, y_{T_0-1})$

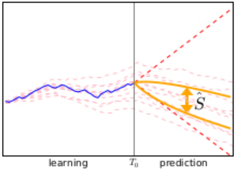
for $t = T_0, \ldots, T$:
I. Building $\widehat{S}_t$:
Initialize $\widehat{S}_t = \emptyset$
for each $(z_{T_0}, \ldots, z_T) \in S$:
1. Feed any classical learning algorithm with $(y_0, \ldots, y_{T_0-1}, z_{T_0}, \ldots, z_{t-1})$ and $(f_{k,\tau})_{1 \leqslant k \leqslant K, 1 \leqslant \tau \leqslant t}$
2. Predict $\widehat{z}_t$
3. Update $\widehat{S}_t \leftarrow \widehat{S}_t \cup \{\widehat{z}_t\}$

II. Output:
Output the forecast interval $[\widehat{y}_t^{\min}, \widehat{y}_t^{\max}]$ defined as the smallest interval containing $\widehat{S}_t$

The set S of the possible scenarios



The set $\hat{S}$ of the forecast intervals



Application : oil prediction forecasting

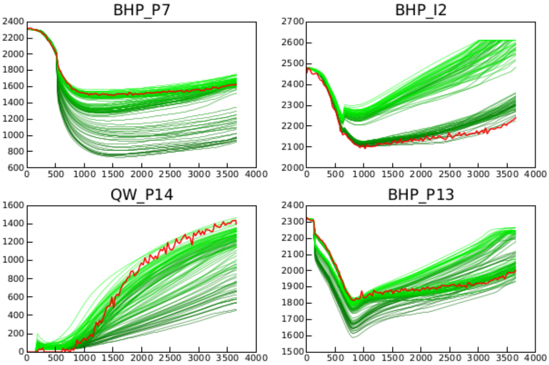


Figure – 104 physical models build to predict oil production in various wells.

Results

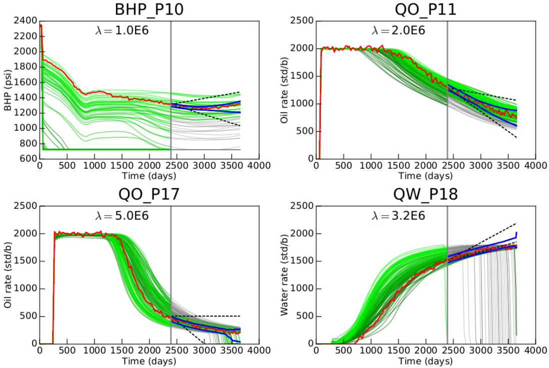


Figure – Confidence intervals by R. Deswarte's algorithm.

Results

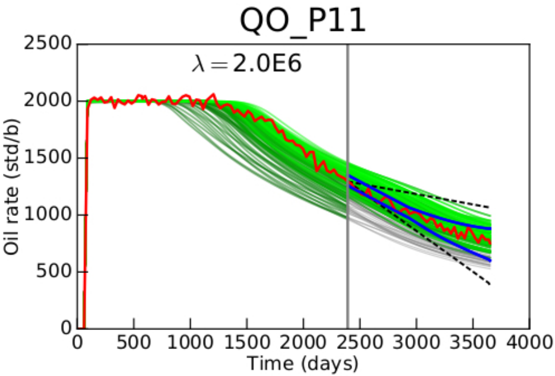
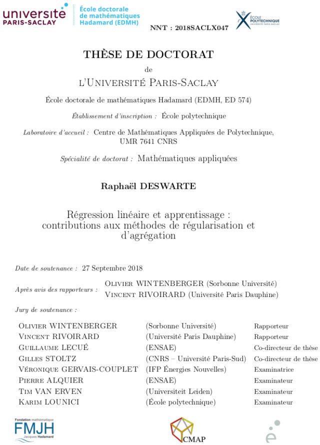


Figure – Confidence intervals by R. Deswarte's algorithm.

Raphaël Deswarte


université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

NNT : 2018SACLX047

ÉCOLE POLYTECHNIQUE

THÈSE DE DOCTORAT

de

L'UNIVERSITÉ PARIS-SACLAY

École doctorale de mathématiques Hadamard (EDMH, ED 574)

Établissement d'inscription : École polytechnique

Laboratoire d'accueil : Centre de Mathématiques Appliquées de Polytechnique, UMR 7641 CNRS

Spécialité de doctorat : Mathématiques appliquées

Raphaël DESWARTE

Régression linéaire et apprentissage : contributions aux méthodes de régularisation et d'agrégation

Date de soutenance : 27 Septembre 2018

Après avis des rapporteurs : Olivier WINTENBERGER (Sorbonne Université)
Vincent RIVOIRARD (Université Paris Dauphine)

Jury de soutenance :

Olivier WINTENBERGER	(Sorbonne Université)	Rapporteur
Vincent RIVOIRARD	(Université Paris Dauphine)	Rapporteur
Guillaume LECUÉ	(ENSAE)	Co-directeur de thèse
Gilles STOLTZ	(CNRS – Université Paris-Sud)	Co-directeur de thèse
Véronique GERVAIS-COUPLET	(IFP Énergies Nouvelles)	Examinatrice
Pierre ALQUIER	(ENSAE)	Examinateur
Tim VAN ERVEN	(Universiteit Leiden)	Examinateur
Karim LOUNICI	(École polytechnique)	Examinateur

FMJH CMAP

Raphaël Deswarte

université PARIS-SACLAY — École doctorale de mathématiques Hadamard (EDMH) — NNT : 2018SACLX047 — ÉCOLE POLYTECHNIQUE

THÈSE DE DOCTORAT

de

L'UNIVERSITÉ PARIS-SACLAY

École doctorale de mathématiques Hadamard (EDMH, ED 574)

Établissement d'inscription : École polytechnique

Laboratoire d'accueil : Centre de Mathématiques Appliquées de Polytechnique, UMR 7641 CNRS

Spécialité de doctorat : Mathématiques appliquées

Raphaël DESWARTE

Régression linéaire et apprentissage : contributions aux méthodes de régularisation et d'agrégation

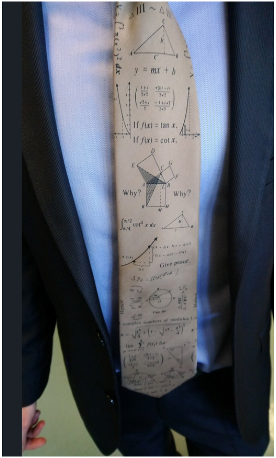
Date de soutenance : 27 Septembre 2018

Après avis des rapporteurs : OLIVIER WINTENBERGER (Sorbonne Université)
VINCENT RIVOIRARD (Université Paris Dauphine)

Jury de soutenance :

OLIVIER WINTENBERGER	(Sorbonne Université)	Rapporteur
VINCENT RIVOIRARD	(Université Paris Dauphine)	Rapporteur
GUILLAUME LECUÉ	(ENSAE)	Co-directeur de thèse
GILLES STOLTZ	(CNRS – Université Paris-Sud)	Co-directeur de thèse
VÉRONIQUE GERVAIS-COUPLET	(IFP Énergies Nouvelles)	Examinatrice
PIERRE ALQUIER	(ENSAE)	Examinateur
TIM VAN ERVEN	(Universiteit Leiden)	Examinateur
KARIM LOUNICI	(École polytechnique)	Examinateur

FMJH — CMAP



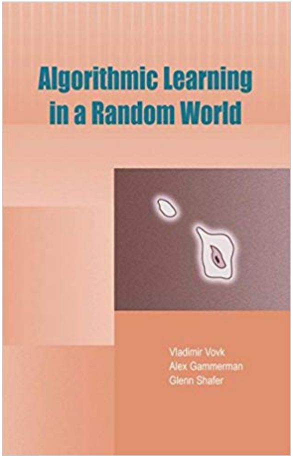
Conformal prediction

Another approach was proposed by Vovk and co-authors.

Conformal prediction

Another approach was proposed by Vovk and co-authors.

It is extremely nice, flexible and theoretically grounded. But requires stochastic assumptions on the data. Also, very different from the previous approaches, so would be too long to explain here... so read :



Fast algorithms ?

In the infinite case, the computation of EWA might be infeasible or very slow...

Fast algorithms ?

In the infinite case, the computation of EWA might be infeasible or very slow...

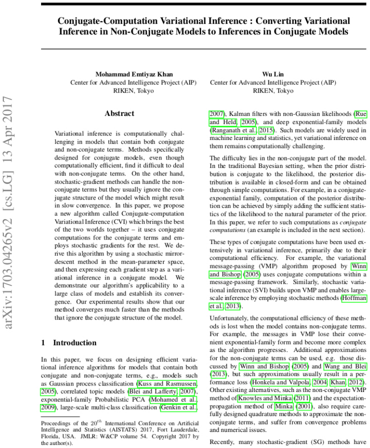
In Bayesian statistics, fast approximations of $\pi(\theta|y_1, \ldots, y_t)$ available via "variational inference".

Fast algorithms ?

In the infinite case, the computation of EWA might be infeasible or very slow...

In Bayesian statistics, fast approximations of $\pi(\theta|y_1, \ldots, y_t)$ available via "variational inference".

Similar approaches are currently being developped in the online (sequential prediction) framework for p_t...



arXiv:1703.04265v2 [cs.LG] 13 Apr 2017

Conjugate-Computation Variational Inference : Converting Variational Inference in Non-Conjugate Models to Inferences in Conjugate Models

Mohammad Emtiyaz Khan
Center for Advanced Intelligence Project (AIP)
RIKEN, Tokyo

Wu Lin
Center for Advanced Intelligence Project (AIP)
RIKEN, Tokyo

Abstract

Variational inference is computationally challenging in models that contain both conjugate and non-conjugate terms. Methods specifically designed for conjugate models, even though computationally efficient, find it difficult to deal with non-conjugate terms. On the other hand, stochastic-gradient methods can handle the non-conjugate terms but they usually ignore the conjugate structure of the model which might result in slow convergence. In this paper, we propose a new algorithm called Conjugate-computation Variational Inference (CVI) which brings the best of the two worlds together – it uses conjugate computations for the conjugate terms and employs stochastic gradients for the rest. We derive this algorithm by using a stochastic mirror-descent method in the mean-parameter space, and then expressing each gradient step as a variational inference in a conjugate model. We demonstrate our algorithm's applicability to a large class of models and establish its convergence. Our experimental results show that our method converges much faster than the methods that ignore the conjugate structure of the model.

1 Introduction

In this paper, we focus on designing efficient variational inference algorithms for models that contain both conjugate and non-conjugate terms, e.g., models such as Gaussian process classification (Kuss and Rasmussen, 2005), correlated topic models (Blei and Lafferty, 2007), exponential-family Probabilistic PCA (Mohamed et al., 2009), large-scale multi-class classification (Genkin et al., 2007), Kalman filters with non-Gaussian likelihoods (Rue and Held, 2005), and deep exponential-family models (Ranganath et al., 2015). Such models are widely used in machine learning and statistics, yet variational inference on them remains computationally challenging.

The difficulty lies in the non-conjugate part of the model. In the traditional Bayesian setting, when the prior distribution is conjugate to the likelihood, the posterior distribution is available in closed-form and can be obtained through simple computations. For example, in a conjugate-exponential family, computation of the posterior distribution can be achieved by simply adding the sufficient statistics of the likelihood to the natural parameter of the prior. In this paper, we refer to such computations as *conjugate computations* (an example is included in the next section).

These types of conjugate computations have been used extensively in variational inference, primarily due to their computational efficiency. For example, the variational message-passing (VMP) algorithm proposed by Winn and Bishop (2005) uses conjugate computations within a message-passing framework. Similarly, stochastic variational inference (SVI) builds upon VMP and enables large-scale inference by employing stochastic methods (Hoffman et al., 2013).

Unfortunately, the computational efficiency of these methods is lost when the model contains non-conjugate terms. For example, the messages in VMP lose their convenient exponential-family form and become more complex as the algorithm progresses. Additional approximations for the non-conjugate terms can be used, e.g. those discussed by Winn and Bishop (2005) and Wang and Blei (2013), but such approximations usually result in a performance loss (Honkela and Valpola, 2004; Khan, 2012). Other existing alternatives, such as the non-conjugate VMP method of Knowles and Minka (2011) and the expectation-propagation method of Minka (2001), also require carefully designed quadrature methods to approximate the non-conjugate terms, and suffer from convergence problems and numerical issues.

Recently, many stochastic-gradient (SG) methods have

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54. Copyright 2017 by the author(s).

Other (fast) approximations

Stochastic Particle Gradient Descent for Infinite Ensembles

Atsushi Nitanda [*1,2] and Taiji Suzuki [†1,2,3]

[1]Graduate School of Information Science and Technology, The University of Tokyo
[2]Center for Advanced Intelligence Project, RIKEN
[3]PRESTO, Japan Science and Technology Agency

Abstract

The superior performance of ensemble methods with infinite models are well known. Most of these methods are based on optimization problems in infinite-dimensional spaces with some regularization, for instance, boosting methods and convex neural networks use L^1-regularization with the non-negative constraint. However, due to the difficulty of handling L^1-regularization, these problems require early stopping or a rough approximation to solve it inexactly. In this paper, we propose a new ensemble learning method that performs in a space of probability measures, that is, our method can handle the L^1-constraint and the non-negative constraint in a rigorous way. Such an optimization is realized by proposing a general purpose stochastic optimization method for learning probability measures via parameterization using transport maps on base models. As a result of running the method, a transport map to output an *infinite ensemble* is obtained, which forms a residual-type network. From the perspective of functional gradient methods, we give a convergence rate as fast as that of a stochastic optimization method for finite dimensional nonconvex problems. Moreover, we show an *interior optimality property* of a local optimality condition used in our analysis.

1 Introduction

The goal of the binary classification problem is to find a measurable function, called a classifier, from the feature space to the range $[-1, 1]$, which is required to minimize the expected classification error. The ensemble, including boosting and bagging, is one method used to solve this problem, by constructing a complex classifier by combining base classifiers. It is well-known empirically that such a classifier attains good generalization performance in experiments and applications [3, 12, 51].

*atsushi_nitanda@mist.i.u-tokyo.ac.jp
†taiji@mist.i.u-tokyo.ac.jp

1

Perturbed Bayesian Inference for Online Parameter Estimation

Mathieu Gerber* Kari Heine†

In this paper we introduce perturbed Bayesian inference, a new Bayesian based approach for online parameter inference. Given a sequence of stationary observations $(Y_t)_{t\geq 1}$, a parametric model $\{f_\theta, \theta \in \mathbb{R}^d\}$ and $\theta_\star := \mathrm{argmax}_{\theta\in\mathbb{R}^d}\mathbb{E}[\log f_\theta(Y_1)]$, the sequence $(\tilde{\pi}_t^N)_{t\geq 1}$ of *perturbed posterior distributions* has the following properties: (i) $\tilde{\pi}_t^N$ does not depend on $(Y_s)_{s>t}$, (ii) the time and space complexity of computing $\tilde{\pi}_t^N$ from $\tilde{\pi}_{t-1}^N$ and Y_t is at most cN, where $c < +\infty$ is independent of t, and (iii) for N large enough $\tilde{\pi}_t^N$ converges almost surely as $t \to +\infty$ to $\theta_\star$ at rate $\log(t)^{(1+\varepsilon)/2}t^{-1/2}$, with $\varepsilon > 0$ arbitrary and under classical conditions that can be found in the literature on maximum likelihood estimation and on Bayesian asymptotics.

Keywords: Bayesian inference, online inference, streaming data

1 Introduction

In many modern applications a large number of observations arrive continuously and need to be processed in real time, either because it is impracticable to store the data or because a decision should be made and/or revised as soon as possible as you data arrives. This is for instance the case with digital financial transactions data, where the number of observations per day frequently exceeds the million and where online fraud detection is of obvious importance (Zhang et al., 2018). In this context, the notion of data stream is more appropriate than that of a dataset, which supposes infrequent updates (O'callaghan et al., 2002). Following Henzinger et al. (1998), we informally refer to a data stream as a sequence of observations that can be read only once and in the order in which they arrive. The data stream model is also relevant for large datasets, where the number of observations is such that each of them can only be read a small number of times for practical considerations (O'callaghan et al., 2002).

Beyond computations of simple descriptive statistics, statistical inference from data streams is a challenging task. This is particularly true for parameter estimation in parametric models, the focus of this paper. Indeed, current approaches to online parameter

*School of Mathematics, University of Bristol, UK.
†Department of Mathematical Sciences, University of Bath, UK.

1

More open questions

More open questions

- theoretical study of the confidence intervals.

More open questions

- theoretical study of the confidence intervals.
- theoretical study of the fast algorithms.

More open questions

- theoretical study of the confidence intervals.
- theoretical study of the fast algorithms.
- causal inference ?

More open questions

- theoretical study of the confidence intervals.
- theoretical study of the fast algorithms.
- causal inference ?
- tests ?

More open questions

- theoretical study of the confidence intervals.
- theoretical study of the fast algorithms.
- causal inference ?
- tests ?
- ...

Thank you!!