Prediction with expert advice under budget constraints

Gilles Blanchard

Université Paris-Saclay, Inria

Spring School “Data Assimilation”, 20-24 March 2023

Based on joint work with El Mehdi Saad (INRAE Montpellier)

1. Lecture 1: Prediction with expert advice: basics

2. Lecture 2: Fast rates on a budget I (stochastic + simple regret)

3. Lecture 3: Fast rates on a budget II (fixed sequence prediction, cumulative regret)

Course Plan

1. Lecture 1: Prediction with expert advice: basics
2. Lecture 2: Fast rates on a budget I (stochastic + simple regret)
3. Lecture 3: Fast rates on a budget II (fixed sequence prediction, cumulative regret)

Prediction with the help of experts

A large panel of possibly diverse **experts**...

Prediction with the help of experts

A large panel of possibly diverse **experts**...

Prediction with the help of experts

A large panel of possibly diverse **experts**...

- Each expert is giving predictions
- Experts may be more or less skilled
- How can one make the best use of expert advice?

Prediction with the help of experts

- In many applications we want to predict an output Y as accurately as possible and have access to a panel of “expert” predictions $F_1, \ldots, F_K$.
- Concretely, experts might be human experts (e.g. **finance, sports, crowdsourcing**) but more often statistical or numerical models.
- Models might differ by architecture, assumptions they are built on, or tuning parameters.
- Often the case in industrial applications: R&D teams will want to try out and compare many existing models (+ in-house developed models).
- Models/experts are treated as “black boxes” and a loose general goal is to find a way to find a prediction “not much worse than the best expert”

Problem 1: "batch" learning

- We have a large (but finite) family of prediction models (e.g. weather forecast, electricity consumption) $F_1, \ldots, F_K$.
- Each model can be run under different initial conditions $X \in \mathcal{X}$, giving rise to predictions $F_1(X), \ldots, F_K(X)$.
- We can compare the output of the different models on the same input to an observed "truth" $Y \in \mathcal{Y}$.
- Quality of a **single prediction** is measured through a **loss function** (e.g. squared loss)

$$\ell(F_k(X), Y).$$

Problem 1: "batch" learning (continued)

- The quality (**risk**) of a prediction model is measured through an average according to a probability distribution P:

$$L_P(F) := \mathbb{E}_{(X,Y)\sim P}[\ell(F(X), Y)].$$

- We have access to data $(X_t, Y_t), t = 1, \ldots, N$ generated i.i.d from P.

Separate "training" and prediction: observe all the data and the model predictions, **then** based on this decide of a "final" prediction strategy F.

- **Goal:** have a small **regret**

$$\mathcal{R}(F) := L_P(F) - \min_{k \in [\![K]\!]} L_P(F_k).$$

- **Note:** due to randomness in the training data, and possible internal randomization of F, we can ask for this guarantee in expectation or with high probability (**whp**) with respect to $(X_t, Y_t)_{t \in [\![N]\!]}$ (and internal randomization).

Some notation (setting 1)

- We will "forget" about the covariate X and identify $F_k = F_k(X)$ (random variable)
- The training sequence will be denoted

$$S_N = (\boldsymbol{F}_t, Y_t)_{t \in [\![N]\!]},$$

- where $\boldsymbol{F}_t = (F_{1,t}, \dots, F_{K,t})$ is the vector of expert predictions at time t (with $F_{k,t} = F_k(X_t)$).
- The loss of expert k at training round t is denoted

$$\ell_{k,t} = \ell(F_{k,t}, Y_t).$$

- The population expected loss of expert k is

$$L_k = L_P(F_k) = \mathbb{E}[\ell(F_k, Y)].$$

Problem 2: sequential learning/prediction

- We want to predict sequentially outputs $Y_1, \ldots, Y_N$. The generating mechanism for the outputs is unknown (not assumed i.i.d. – "**adversarial**")

- We have access to a large but finite family of "expert predictions" $F_1, \ldots, F_K$ (Each expert might have access to some privileged information that we don't see.)

- **Examples:** time series, recommender systems...

- Each expert is identified with a **sequence of predictions**: $F_k \equiv (F_{k,1}, \ldots, F_{k,N})$.

- As before we measure quality of a single prediction F for output Y via a loss function $\ell(F, Y)$.

Problem 2: sequential setting (cont'd)

- The quality of a prediction sequence $F = (F_1, \dots, F_N)$ is measured through its averaged **cumulative loss**

$$L_{\mathrm{seq}}^{(N)}(F) := \frac{1}{N} \sum_{t=1}^{N} \ell(F_t, Y_t).$$

Constraint: a valid prediction sequence $(F_1, \dots, F_N)$ is such that prediction F_t may only depend on **past outputs** $(Y_{t'})_{t'<t}$ and **past and present expert predictions** $(F_{i,t'})_{i \in [\![K]\!], t' \le t}$. We will talk of a **prediction strategy.**

- In this scenario "learning" and "prediction" are **intertwined.**
- **Goal:** guarantee a small **regret**

$$\mathcal{R}(F) := L_{\mathrm{seq}}(F) - \min_{k \in [\![K]\!]} L_{\mathrm{seq}}(F_k).$$

- **Note:** due to possible randomization in the prediction strategy, we can ask for this guarantee in expectation or with high probability (**whp**).

How to combine experts?

- We assume that the loss $\ell(.,.)$ is **convex** in its first variable (the prediction, assumed to take values in a vector space).

We will only consider "combination of experts" strategies that are **convex combinations:**

$$F_w := \sum_{i=1}^{K} w_i F_i, \qquad \text{where } \boldsymbol{w} \in \Delta,$$

where Δ is the $(K-1)$-dimensional simplex.

- Instead of exactly combining experts we might also use w as a **probability distribution** on $[\![K]\!]$ and draw one expert at random.

Two scenarios: overview

Stochastic + simple regret

- $S_N = (Y_t, F_t)_{1 \le t \le N}$ are **i.i.d.** wrt. t, with joint distribution P.
- Observe **all** the above, **then** pick a combination $w \in \Delta$.
- **Goal:** small average "**simple**" regret on future predictions:

$$\mathcal{R}(w) = \mathbb{E}[\ell(F_w, Y)] - \min_i \mathbb{E}[\ell(F_i, Y)].$$

Fixed seq. + cumulative regret

- $(Y_t, F_t)_{1 \le t \le N}$ are **a fixed sequence**
- Observe the above up to $t-1$, then pick a combination $w_t \in \Delta$, sequentially for $t = 1, \ldots, N$.
- **Goal:** small **cumulative** regret

$$\mathcal{R}((w_t)_{t \le N}) = \frac{1}{N}\left(\sum_{t=1}^N \ell(F_{w_t}, Y_t) - \min_i \sum_{t=1}^N \ell(F_{i,t}, Y_t) \right)$$

From fixed sequence to stochastic

Proposition : Adversarial Regret > Expected Stochastic Regret

Assume $\widehat{\boldsymbol{w}} = (\widehat{w}_t)_{t \le N}$ is an expert combination strategy in the fixed sequence scenario, such that for some deterministic number $\mathcal{B}$:

$$L_{\mathrm{seq}}(F_{\widehat{\boldsymbol{w}}}) \le \min_i L_{\mathrm{seq}}(F_i) + \mathcal{B}.$$

Then if the sequence $(\boldsymbol{F}_t, Y_t)_{t \in [\![N]\!]}$ is actually i.i.d. from a distribution P, then

$$\mathbb{E}\left[L_{\mathrm{seq}}(F_{\widehat{\boldsymbol{w}}})\right] \le \min_i L_P(F_i) + \mathcal{B}.$$

Proof:

$$\mathbb{E}\left[\min_i L_{\mathrm{seq}}(F_i)\right] \le \min_i \mathbb{E}\left[L_{\mathrm{seq}}(F_i)\right] = \min_i \frac{1}{N}\sum_{t=1}^{N} \mathbb{E}\left[\ell(F_{i,t}, Y_t)\right] = \min_i L_P(F_i).$$

From cumulative regret to simple regret

Proposition : Online to batch conversion (Progressive mixture)

Assume $\widehat{\boldsymbol{w}} = (\widehat{w}_t)_{t \geq 0}$ is an expert combination strategy in the fixed sequence scenario.
In the "batch" learning scenario, with batch training sample $S_N = (\boldsymbol{F}_t, Y_t)_{t \in [\![N]\!]} \overset{i.i.d.}{\sim} P$, let $\widehat{\boldsymbol{w}}$ be the result of the above strategy applied to the sample considered as a sequence, and consider

$$\widetilde{w} := \frac{1}{N+1} \sum_{t=1}^{N+1} \widehat{w}_t \in \Delta.$$

(Recall $\widehat{w}_t$ only depends on data observed for $t' < t$.)
Then the **simple risk** of the above aggregate is bounded as

$$\mathbb{E}_{S_N}\left[L_P(F_{\widetilde{w}})\right] \leq \mathbb{E}_{S_{N+1}}\left[L_{\mathrm{seq}}^{(N+1)}(\widehat{\boldsymbol{w}})\right].$$

A "universal" strategy

Exponential Weights Averaging (EWA) (Vovk, 1998)

- Define the **cumulative loss** of each expert:

$$\widehat{L}_{k,t} = \sum_{u=1}^{t} \ell(F_{k,u}, Y_k) = \sum_{u=1}^{t} \ell_{k,u}.$$

- And the combination weights (for some $\lambda > 0$)

$$w_{k,t}^{EWA} \propto \exp(-\lambda \widehat{L}_{k,t}).$$

- (**Note:** in the stochastic scenario, $\lambda = \infty$ is the "empirical risk minimization" (ERM).)

Pseudo-Bayesian interpretation of EWA

- Interpret the loss $\ell(F_i, Y)$ as a **"pseudo-log-likelihood"** for expert i

- The EWA weights
$$w_{k,t}^{EWA} \propto \exp\left(-\lambda \widehat{L}_{k,t}\right)$$
can then be interpreted as a **pseudo-posterior** in the **Bayesian** sense (up to the rescaling λ).

- Alternative **"thermodynamic"** interpretation: the reweighting of experts follows a **"Gibbsian"** distribution where the losses play the role of the **minus** energy, and λ the inverse temperature.

A "universal" strategy

Exponential Weights Averaging (EWA)

Theorem :

- In the stochastic+simple regret scenario, if $\lambda \gtrsim \sqrt{\log K / N}$:

$$\mathbb{E}_{S_N}\left[\mathcal{R}(\boldsymbol{w}_N^{EWA})\right] \lesssim \sqrt{\frac{\log K}{N}},$$

- And in the sequence+cumulative regret scenario, if $\lambda \simeq \sqrt{\log K / N}$:

$$\mathcal{R}\left((\boldsymbol{w}_t^{EWA})_{1 \le t \le N}\right) \lesssim \sqrt{\frac{\log K}{N}}.$$

A “universal” strategy
Exponential Weights Averaging (EWA)

Theorem :

- In the stochastic+simple regret scenario, if $\lambda \gtrsim \sqrt{\log K / N}$:

$$\mathbb{E}_{S_N}\left[\mathcal{R}(\boldsymbol{w}_N^{EWA})\right] \lesssim \sqrt{\frac{\log K}{N}},$$

  also holds **with high probability wrt. observations S_N.** ☺

- And in the sequence+cumulative regret scenario, if $\lambda \simeq \sqrt{\log K / N}$:

$$\mathcal{R}((\boldsymbol{w}_t^{EWA})_{1 \leq t \leq N}) \lesssim \sqrt{\frac{\log K}{N}}.$$

- In both scenarios: also holds for **randomized** version
  (Pick 1 random expert using weights as probability)
  (In expectation or with high probability wrt. randomization). ☺

Fast rates

- **Improved bounds** if we assume some form of **strong convexity** of the loss.
- In what follows we will assume

Assumption (BSL)

Predictions and target belong to $[0, 1]$ and loss is squared loss.

(can be generalized to **bounded, exp-concave** losses)

- Then for $\lambda \simeq 1$:
  - In the **sequence+cumulative regret** scenario,

$$\mathcal{R}((\boldsymbol{w}_t^{EWA})_{1 \leq t \leq N}) \lesssim \frac{\log K}{N}.$$

  - In the **stochastic+simple regret** scenario, combined with "online-to-batch/progressive mixture"

$$\mathbb{E}_{S_N}\left[\mathcal{R}(\boldsymbol{w}_N^{EWA})\right] \lesssim \frac{\log K}{N},$$

Fast rates

- **Improved bounds** if we assume some form of **strong convexity** of the loss.
- In what follows we will assume

  Assumption (BSL)

  Predictions and target belong to $[0, 1]$ and loss is squared loss.

  (can be generalized to **bounded, exp-concave** losses)
- Then for $\lambda \simeq 1$:
  - In the **sequence+cumulative regret** scenario,

$$\mathcal{R}\left((\boldsymbol{w}_t^{EWA})_{1 \le t \le N}\right) \lesssim \frac{\log K}{N}.$$

  - In the **stochastic+simple regret** scenario, combined with "online-to-batch/progressive mixture"

$$\mathbb{E}_{S_N}\left[\mathcal{R}(\boldsymbol{w}_N^{EWA})\right] \lesssim \frac{\log K}{N},$$

    But: **not true** with high probability wrt. observations S_n! ☹
  - In either scenario: **not true** for randomized version (even in expectation). ☹

Course Plan

1. Lecture 1: Prediction with expert advice: basics
2. Lecture 2: Fast rates on a budget I (stochastic + simple regret)
3. Lecture 3: Fast rates on a budget II (fixed sequence prediction, cumulative regret)

Prediction with costly expert advice?

- **Asking for expert advice** is **costly!**

- **“Monetary” cost:**
  - Consulting an expert **before** the event (**a priori**) is expensive
  - Observing an expert’s individual loss **after** the event (**a posteriori**) may be cheaper

- **Time/Computation cost:**
  - computing/consulting all individual prediction models **a priori** might be subject to strong constraints due to time, communication or computation constraints
  - constraints for observing losses **a posteriori** might me looser

- **“Frugal” learning:** Integrate such constraints into the mathematical setup

Prediction with budgetary constraints

- **Constraint for prediction (a priori):** use only up to p expert queries (i.e. combination weights must belong to Δ_p)

- **Constraint for observation (a posteriori):** several settings:

  - **Global** budget constraint: (**simple regret** scenario only)
    Limitation of total number Q observed expert losses during training.
    (No limitation on number of training rounds.)

  - **Local** budget constraint: (both scenarios)
    Limitation to m observed expert losses in each round.
    (Simple regret scenario: still limited to N training rounds.)

What is known? The slow rate setting, $p = 1$

- **"Slow rate"** setting:
  - **Stochastic + Simple regret scenario**:
    Spread out equally training observations so that each expert is observed Q/K times ($Q = Nm$ for local budget constraint)
    Then use **randomized** EWA strategy for prediction.
    Simple regret: $\mathcal{O}(\sqrt{(\log K)K/Q})$.
  - **Remark**: equivalently if one aims at a guaranteed regret less than $\varepsilon > 0$, then $Q_\varepsilon = \mathcal{O}(K \log(K)\varepsilon^{-2})$ queries are necessary.

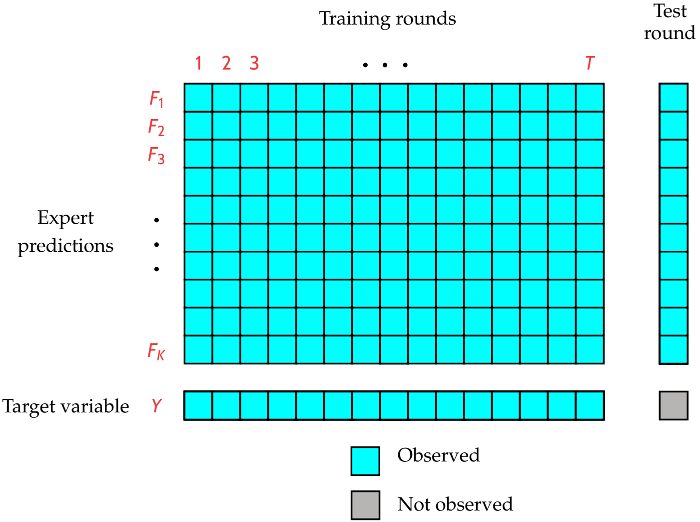
What are our aims?

Assumption (BSL)

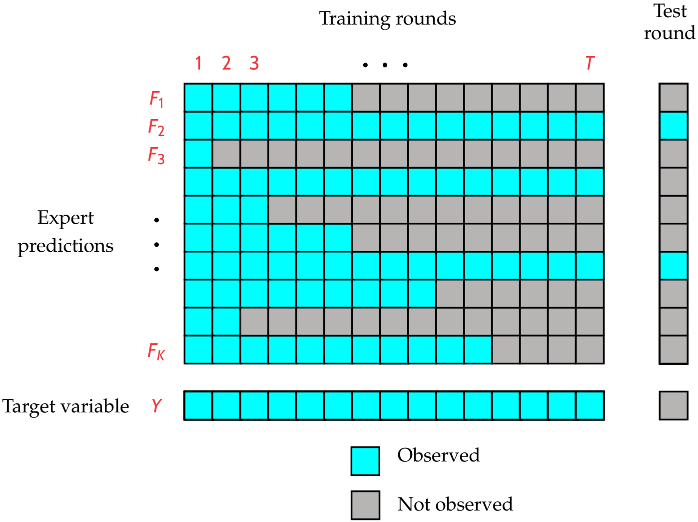
Predictions and target belong to $[0, 1]$ and loss is squared loss.

- When are **fast rates** possible, impossible under budget constraints?
- What is the influence of the **budgetary constraints** on the regret?
- Are fast rates bounds **with high probability** possible?
- In the stochastic scenario, is it possible to obtain **fast context dependent bounds** (i.e. faster than worst case if many experts are largely sub-optimal)

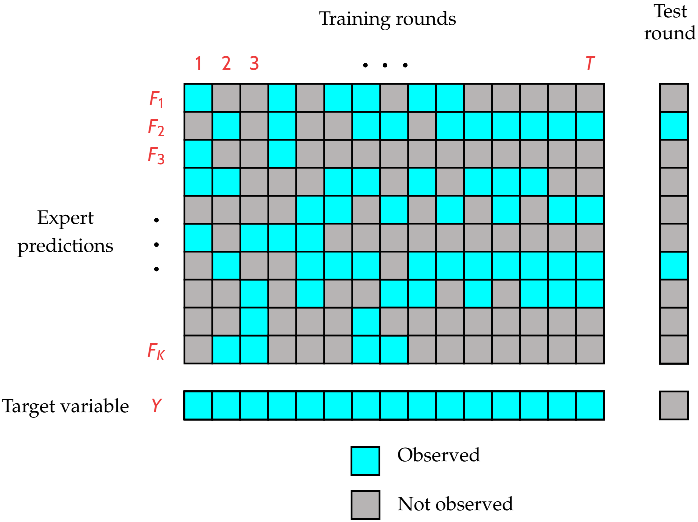
Unconstrained (=full information) setting



“Global budget” setting



"Local budget" setting



The full information/unconstrained case ($m = p = K$)

A.k.a. "aggregation for model selection" problem in batch setting

Assumption (BSL)

Predictions and target belong to $[0, 1]$ and loss is squared loss.

- **Audibert**(2008): Although progressive EWA is has fast regret in expectation, it is **deviation suboptimal** i.e. excess risk is $\Omega(1/\sqrt{N})$ with constant prob.
- **Lecué-Mendelson**(2009): ERM on **convex combinations** of experts is suboptimal
- Both **Audibert**(2008) and **Lecué-Mendelson**(2009) propose specific strategies with optimal fast rate ($O(1/N)$) excess risk deviations with high probability
- **Fact: proper** decision rules selecting **one** expert for prediction (e.g. ERM) cannot attain fast rates in general – at best $O(1/\sqrt{N})$.
- **Audibert**'s "empirical star" algorithm outputs a combination of **only 2 experts**.

Revisiting the unconstrained case

Notation:

- $\widehat{L}_i$ empirical average loss of expert i;
- $\widehat{d}_{ij}$ empirical mean of $(\ell(F_i, Y) - \ell(F_j, Y))^2$.

Test statistic for expert i vs expert j:

$$\widehat{\Delta}_{ij} := \widehat{L}_j - \widehat{L}_i - \alpha\widehat{d}_{ij} - \alpha^2. \qquad \left(\alpha \simeq \sqrt{\log(K\delta^{-1})/N}\right)$$

Fact: (from empirical Bernstein's inequality)

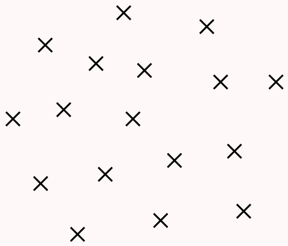
$\widehat{\Delta}_{ij} > 0$ implies $L_j > L_i$ w.p. $(1-\delta)$ uniformly over i, j

A simple algorithm – unconstrained case

Full information algorithm

Set of candidates:
non-rejected experts

$$S := \left\{ j \in [\![K]\!] : \sup_{i \in [\![K]\!]} \widehat{\Delta}_{ij} \leq 0. \right\}$$

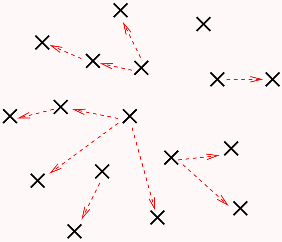


A simple algorithm – unconstrained case

Full information algorithm

Set of candidates:
non-rejected experts

$$S := \left\{ j \in [\![K]\!] : \sup_{i \in [\![K]\!]} \widehat{\Delta}_{ij} \leq 0. \right\}$$

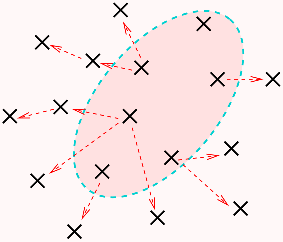


A simple algorithm – unconstrained case

Full information algorithm

Set of candidates:
non-rejected experts

$$S := \left\{ j \in [\![K]\!] : \sup_{i \in [\![K]\!]} \widehat{\Delta}_{ij} \leq 0. \right\}$$



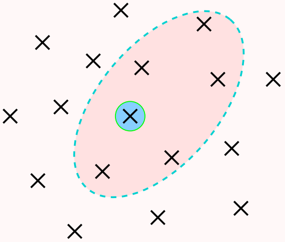
A simple algorithm – unconstrained case

Full information algorithm

Set of candidates:
non-rejected experts

$$S := \left\{ j \in [\![K]\!] : \sup_{i \in [\![K]\!]} \widehat{\Delta}_{ij} \leq 0. \right\}$$

- Choose $\bar{k} \in S$ arbitrarily ;

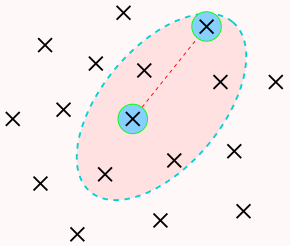


A simple algorithm – unconstrained case

Full information algorithm

Set of candidates:
non-rejected experts

$$S := \left\{ j \in [\![K]\!] : \sup_{i \in [\![K]\!]} \widehat{\Delta}_{ij} \leq 0. \right\}$$



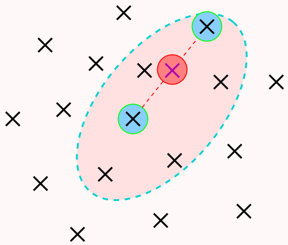
- Choose $\bar{k} \in S$ arbitrarily ;
- Pick $\bar{j} \in \mathrm{Arg\,Max}_{j \in S} \widehat{d}_{\bar{k}j}$;

A simple algorithm – unconstrained case

Full information algorithm

Set of candidates:
non-rejected experts

$$S := \left\{ j \in [\![K]\!] : \sup_{i \in [\![K]\!]} \widehat{\Delta}_{ij} \le 0. \right\}$$



- Choose $\bar{k} \in S$ arbitrarily ;
- Pick $\bar{j} \in \operatorname{Arg\,Max}_{j \in S} \widehat{d}_{\bar{k}j}$;
- Predict $\widehat{F} := \frac{1}{2}(F_{\bar{k}} + F_{\bar{j}})$.

Fast rate in full information case

Theorem

Under (BSL), for the predictor $\widehat{F}$ previously defined for the full information case, with probability $1-\delta$ over the training phase:

$$\mathcal{R}(\widehat{F}) \lesssim \frac{\log(K\delta^{-1})}{N}.$$

- Same type of result as Audibert (2008) and Lecué and Mendelson (2009) but with simpler algorithm & proof.

The global budget setting

- $\widehat{R}_i$, $\widehat{d}_{ij}$ and $\widehat{\Delta}_{ij}$ are defined as before but are updated on-line
- κ is a numerical constant

Budgeted setting algorithm

Input δ.
Initialization: $S \leftarrow [\![K]\!]$.
for $t = 1, 2, \ldots$ **do**
  Remove experts marked for time t from S.
  Observe losses of all the experts in S at time t.
  Update $\widehat{\Delta}_{ij}, \widehat{L}_{ij}, \widehat{d}_{ij}$ for all $i, j \in S$.
  For all $i, j \in [\![K]\!]$, if $\widehat{\Delta}_{ij} > 0$, mark j for deletion from S at time κt.
  if the budget is consumed **then**
    let $\bar{k} \in S$, and $\bar{l} \leftarrow \operatorname{argmax}_{j \in S} \widehat{d}_{\bar{k}j}$.
    return $\widehat{F} = (F_{\bar{k}} + F_{\bar{l}})/2$.
  end if
end for

Result for the global budget setting

- Introduce $\Delta_{ij} = L_i - L_j$ and $d_{ij} = \mathbb{E}\left[\left(\ell(F_i, Y) - \ell(F_j, Y)\right)^2\right]$.

$$T_{ij} := \frac{1}{\Delta_{ij}} \max\left(\frac{d_{ij}^2}{\Delta_{ij}}; 1\right).$$

  If $L_j < L_i$: T_{ij} is the number of joint queries to (i, j) so that i is eliminated by j (w.h.p.)

- Let $\mathcal{S}^*$ denote the set of optimal experts and let

$$T_i^* := \min_{j: L_j < L_i} T_{ij}; \qquad \bar{T}^* := \max_{i \neq \mathcal{S}^*} T_i$$

- T_i^* is the minimum of joint queries for i to be eliminated by **any other (better)** expert.

Result for the global budget setting

- For $\varepsilon \geq 0$ let

$$C_\varepsilon := \sum_{i \in [\![K]\!]} \min\left(T_i^*, \bar{T}^*, \frac{1}{\varepsilon}\right),$$

Theorem : Instance dependent-bound, global budget setting

Assume (BSL).
For the predictor $\widehat{F}$ output by the algorithm in the **global budget** setting, if the budget Q is such that

$$Q \gtrsim C_\varepsilon \log(K\delta^{-1}C_\varepsilon),$$

then with probability $1 - \delta$ over the training phase it holds

$$\mathcal{R}(\widehat{F}) \lesssim \varepsilon.$$

Comparison to unconstrained setting

- It holds
$$C_\varepsilon = \sum_{i \in [\![K]\!]} \min\left(T_i, \bar{T}^*, \frac{1}{\varepsilon}\right) \leq \frac{K}{\varepsilon},$$
- So that a sufficient budget constraint is
$$Q \gtrsim \frac{K}{\varepsilon} \log \frac{K\delta^{-1}}{\varepsilon} \geq C_\varepsilon \log(K\delta^{-1} C_\varepsilon)$$
- In full observation model, to reach the same precision, need number of expert observations
$$Q_\varepsilon \eqsim \frac{K}{\varepsilon} \log(K\delta^{-1})$$
- Hence, at worst additional logarithmic factor w.r.t. full information (and potentially much more efficient)

Local budget setting, $p = 2$, $m \geq 2$ arbitrary

Similar algorithm as before, but sample at each training round m experts uniformly from the set of remaining candidates S, observe their losses and update corresponding quantities.

Theorem : Instance independent-bound, m-queries setting ($m \geq 2$)

Under (BSL), for the predictor $\widehat{F}$ output by the algorithm in the m-queries setting. Then with probability $1 - \delta$ over the training phase it holds

$$\mathcal{R}(\widehat{F}) \lesssim \frac{(K/m)^2 \log\left(NK\delta^{-1}\right)}{N}.$$

Lower bounds

Under (BSL):

Proposition : ($p = 1$)

For $K = m = 2$ and $p = 1$, for any N, and for any output $\hat{F} = F_{\hat{k}}$ after N training rounds, there exists a joint probability distribution for experts $\{F_1, F_2\}$ and target variable Y (all bounded by 1) s.t., with probability at least 0.1,

$$\mathcal{R}(\widehat{F}) \gtrsim \frac{1}{\sqrt{N}}.$$

Proposition : ($m = 1$)

For $K = p = 2$, and $m = 1$, for any N, for any training observation strategy and convex combination output $\widehat{F}$ following the game protocol for N training rounds, there exists a joint probability distribution for experts $\{F_1, F_2\}$ and target variable Y (all bounded by 1) such that with probability at least 0.1,

$$R(\widehat{F}) \gtrsim \frac{1}{\sqrt{N}}.$$

Course Plan

1. Lecture 1: Prediction with expert advice: basics
2. Lecture 2: Fast rates on a budget I (stochastic + simple regret)
3. Lecture 3: Fast rates on a budget II (fixed sequence prediction, cumulative regret)

Fixed sequence scenario under limited advice

- At each round:
  - Predict a convex combination of p experts
  - Observe **a posteriori** the loss of m experts

- “Inclusion Condition” **(IC)** in effect if the set of m **a posteriori** observed experts must include the set of p experts used for prediction

- The case $m = p = K$ is the full information setting.

- The case $m = p = 1$ **(IC)**= true is the bandit setting.

Previous results (fixed sequence scenario)

<table>
<tr><th></th><th colspan="2">$p = 1$</th><th>$p \geq 2$</th><th>$p = K$</th></tr>
<tr><td>$m = 1$</td><td colspan="2">$\sqrt{\frac{K}{N}}$
(Bandit setting) [1,2]</td><td rowspan="4">Fast rates?</td><td rowspan="3"></td></tr>
<tr><td rowspan="2">$m \geq 2$</td><td>Lower bound</td><td>Upper bound</td></tr>
<tr><td rowspan="2">$\sqrt{\frac{K}{m}\frac{1}{N}}$
[3]</td><td rowspan="2">$\sqrt{\frac{K}{m}\frac{\log(K)}{N}}$
[3]</td></tr>
<tr><td>$m = K$</td><td>$(\log K)/N$
[4]</td></tr>
</table>

Bounds up to absolute numerical factors.

[1]:Auer et al., 2002; [2]: Audibert and Bubeck, 2010; [3]: Seldin et al., 2014; [4]: Vovk, 1990

The slow rate setting, $p = 1$, $m \geq 1$

Seldin et al. 2014

- Exponential combination weights (for some $\lambda > 0$) using **pseudo-losses** $\widehat{\ell}_{i,t}$:

$$\widehat{w}_{i,t}^{EWA} \propto \exp\left(-\lambda \sum_{k=1}^{t} \widehat{\ell}_{i,k}\right),$$

- Draw expert I_t at random according to $\widehat{w}^{EWA}$. Use their prediction.
- If $m > 2$ observe additional $m - 1$ expert losses drawn uniformly at random. Denote $\mathcal{O}_t$ the total set of observed experts (including I_t).
- Define pseudo-losses

$$\widehat{\ell}_{i,t} = \frac{\mathbf{1}\{i \in \mathcal{O}_t\}}{\mathbb{P}[i \in \mathcal{O}_t | \mathcal{F}_t]} \ell_{i,t}.$$

- Note that $\mathbb{E}\left[\widehat{\ell}_{i,t} | \mathcal{F}_t\right] = \ell_{i,t}$ (unbiased estimate).

The slow rate setting, $p = 1$, $m \geq 1$

Seldin et al. 2014

Theorem :

In the sequence+cumulative regret scenario, if $\lambda \simeq \sqrt{m \log K / N}$:

$$\mathcal{R}\left((\boldsymbol{w}_t^{EWA})_{1 \leq t \leq N}\right) \lesssim \sqrt{\frac{K}{m} \frac{\log K}{N}}.$$

What is known? The fast rate setting / full information, $m = p = K$

Theorem

Under (BSL), for any input parameter: $\lambda \in \left(0, \frac{1}{4}\right)$, the regret of the (vanilla) EWA $\widehat{w}^{EWA}$ satisfies for any sequence of target variables and expert predictions:

$$\mathcal{R}_T \lesssim \frac{\log(K\delta^{-1})}{\lambda N}.$$

Previous results (fixed sequence scenario)

	$p = 1$		$p \geq 2$	$p = K$
$m = 1$	$\sqrt{\frac{K}{N}}$ (Bandit setting) [1,2]		Fast rates?	
	Lower bound	Upper bound		
$m \geq 2$	$\sqrt{\frac{K}{m}\frac{1}{N}}$ [3]	$\sqrt{\frac{K}{m}\frac{\log(K)}{N}}$ [3]		
$m = K$				$\left(\log K\right)/N$ [4]

Bounds up to absolute numerical factors.

[1]:Auer et al., 2002; [2]: Audibert and Bubeck, 2010; [3]: Seldin et al., 2014; [4]: Vovk, 1990

Modified EWA strategy

- Exponential combination weights (for some $\lambda > 0$) using **pseudo-losses** $\widehat{\ell}_{i,t}$:

$$\widehat{w}_{i,t}^{EWA} \propto \exp\left(-\lambda \sum_{k=1}^{t} \widehat{\ell}_{i,k}\right),$$

- **Modification 1**: $p = 2$ **sufficient.** Draw at random 2 independent experts I_t, J_t from $\widehat{w}^{EWA}$ and predict their midpoint

$$\frac{F_{I_t} + F_{J_t}}{2}.$$

- **Modification 2**: If $m > 2$ observe loss of I_t and additional $m - 2$ expert losses in set $\mathcal{O}_t$ drawn uniformly at random. Estimate pseudo-losses $\widehat{\ell}_{i,t}$ from observed losses only.

Modified EWA strategy: pseudo-loss

- **Unbiased** loss estimation using **"smart centering"** on one expert picked by EWA:

$$\widehat{\ell}_{i,t} = \ell_{I_t,t} + \mathbf{1}\{i \in \mathcal{O}_t\} \frac{K}{m-2} (\ell_{i,t} - \ell_{I_t,t})$$

- **Modification 3: Second-order adjustment**:

$$\widetilde{\ell}_{i,t} = \widehat{\ell}_{i,t} - \lambda \mathbf{1}\{i \in \mathcal{O}_t\} \frac{K}{m-2} (\ell_{i,t} - \ell_{I_t,t})^2 .$$

$\longrightarrow$ corresponding EWA weights denoted as $\widetilde{w}^{EWA}$

Note: it is an "anti-penalty" on estimated losses: **optimism in the face of uncertainty**.

Algorithmic complexity considerations

- The pseudo-losses take the form

$$\widetilde{\ell}_{i,t} = \ell_{I_t,t} + \mathbf{1}\{i \in \mathcal{O}_t\}\Psi(\ell_{i,t} - \ell_{I_t,t}).$$

- Because of exponential weight normalization, the weights are **unchanged** if we **shift** all pseudo-losses by the **same quantity** (for all experts).
- Thus, we can use instead the shifted pseudo-losses

$$\check{\ell}_{i,t} = \widetilde{\ell}_{i,t} - \ell_{I_t,t} = \mathbf{1}\{i \in \mathcal{O}_t\}\Psi(\ell_{i,t} - \ell_{I_t,t}).$$

- **Only** need to update for **observed** experts!
- Using binary tree storage of weights, total complexity per round (weight update + random draw of expert indices) is only $\mathcal{O}(m \log K)$.

Limited feedback I ($m \geq 3, p = 2$)

Theorem

Under (BSL), for $\lambda \simeq \frac{m}{K}$, the regret of the modified EWA $\widehat{\boldsymbol{w}}^{EWA}$ algorithm satisfies for any sequence of target variables and expert predictions:

$$\mathbb{E}[\mathcal{R}_T] \lesssim \frac{K}{m} \frac{\log(K)}{N},$$

where the expectation is with respect to the strategy randomization.

Theorem

Under (BSL), for $\lambda \simeq \frac{m}{K}$, the regret of the second-order modified EWA $\widetilde{\boldsymbol{w}}^{EWA}$ satisfies for any sequence of target variables and expert predictions, with probability $1 - \delta$ wrt the strategy randomization:

$$\mathcal{R}_T \lesssim \frac{K}{m} \frac{\log(K\delta^{-1})}{N}.$$

Fast rates results (seq. prediction, cumul. regret)

Additional results in green

<table>
<tr><td></td><td colspan="2">$p = 1$</td><td colspan="2">$p \geq 2$</td></tr>
<tr><td rowspan="2">$m = 1$</td><td colspan="2" rowspan="3">$\sqrt{\frac{K}{N}}$</td><td>Lower Bound</td><td>Upper Bound ($p = 2$)</td></tr>
<tr><td>$\sqrt{\frac{K}{N}}$</td><td>$\sqrt{\frac{K}{N}}$</td></tr>
<tr><td>$m = 2$</td><td>$\frac{K}{N}$</td><td>IC $=$ True : $\frac{K^2 \log(K)}{N}$
IC $=$ False : $\frac{K \log(K)}{N}$</td></tr>
<tr><td rowspan="2">$m \geq 3$</td><td>Lower bound</td><td>Upper bound</td><td rowspan="2">$\frac{K}{mN}$</td><td rowspan="2">$\frac{K \log(K)}{mN}$</td></tr>
<tr><td>$\sqrt{\frac{K}{mN}}$</td><td>$\sqrt{\frac{K \log(K)}{mN}}$</td></tr>
</table>

Bounds up to absolute numerical factors. Upper bounds also hold w.h.p $(1 - \delta)$ with factor $\log \delta^{-1}$.

Lower bounds

The distinction between fast and slow rates in the upper bounds is not an artifact but is also supported by (worst case) lower bounds.

Theorem

Under (BSL), if either $p = 1$ or $m = 1$, it holds

$$\inf \sup \mathbb{E}[\mathcal{R}_T] \gtrsim \sqrt{\frac{K}{mN}}.$$

and for $m \geq 2$, $p \geq 2$ it holds

$$\inf \sup \mathbb{E}[\mathcal{R}_T] \gtrsim \frac{K}{mN},$$

where the inf is over convex aggregation strategies and the sup over sequences (the expectation is over possible randomization of the strategy).

Take home messages

- Scenarios for "**frugal learning**" under budget limitations for expert access
  Suitable assumption on loss allows fast rates.

- In all scenarios, in order to attain **fast rates** $\mathcal{O}(1/Q)$ (vs. $\mathcal{O}(1/\sqrt{Q})$) for regret as a function of the number of queries Q, it is **necessary** and **sufficient** to:
  - be able to predict a combination of $p = 2$ experts;
  - be able to observe at least $m \geq 2$ experts' losses per round.

- Results in expectation and with high probability.

- The natural regret bound to aim for appears to be K/Q. Some loose ends remaining:
  - Extra logarithmic factors everywhere
  - For stochastic + simple regret, local budget scenario: extra factor K/m
  - For fixed seq + cumul. regret, $m = p = 2$, **(IC)=true** (the "**bi-bandit**"): extra factor K

Thank you for your attention

J.-Y. Audibert. Progressive mixture rules are deviation suboptimal.
NeurIPS 2008.

G. Lecué and S. Mendelson. Aggregation via empirical risk minimization.
Probability theory and related fields, 145(3-4):591–613, 2009.

Y. Seldin, P. Bartlett, K. Crammer, and Y. Abbasi-Yadkori. Prediction with limited advice and multiarmed bandits with paid observations.
ICML 2014.

E.M. Saad, G. Blanchard. Fast rates for prediction with limited expert advice.
arXiv:2110.14485 + NeurIPS 2021

E.M. Saad, G. Blanchard. Constant regret for sequence prediction with limited advice.
arXiv:2210.02256

E.M. Saad. Contributions to frugal learning.
Ph.D. Thesis, 2022.

General assumption on loss function

- A function $f : E \to \mathbb{R}$, where E is a convex set, is in the class $\mathcal{E}(c)$ if:

$$\forall x, y \in E : \qquad f\left(\frac{x+y}{2}\right) \leq \frac{1}{2}\Big(f(x) + f(y) - c^{-1}(f(x) - f(y))^2\Big).$$

- Our "fast rates" results hold if predictions take values in a convex set E and for all y, $\ell(\cdot, y)$ is in the class $\mathcal{E}(c)$ (the constant c comes into the bounds).

- Exp-concave, range bounded functions belong to $\mathcal{E}(c)$ for a suitable c.

- Conversely, $f \in \mathcal{E}(c)$ and continuous implies range-bounded by c and f is $(4/c)$-exp-concave.

- Strongly convex, Lipschitz functions also belong to $\mathcal{E}(c)$ for a suitable c.