Statistical and Sequential Learning for Time Series Forecasting

Expert online aggregation

Margaux Brégère

Introduction

Framework

Let $Y = (Y_t)_{t \in \mathbb{N}^\star}$ be a time series

Assumption: at a time step $t = 1, 2, 3, \ldots$

- Observe the data with a delay d: Y_{t-d}
- Receive K predictions $f_{1t}, \ldots, f_{Kt}$ from expert advice / (deterministic or statistic) models

Aim

Providing the best possible forecast $\hat{Y}_t$ of the future realization of Y by mixing the predictions

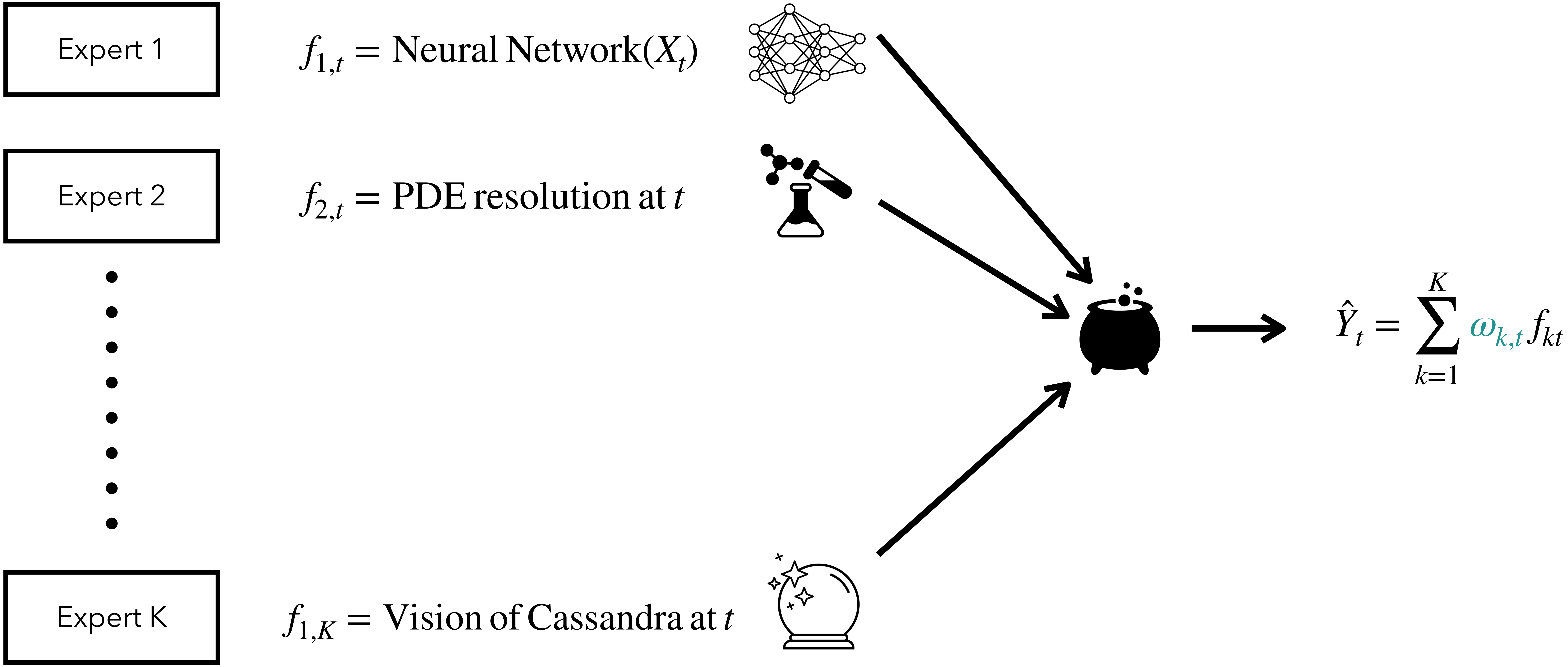
☞ Aggregation $\hat{Y}_t = \hat{f}(f_{1t}, \ldots, f_{Kt}) = \sum_{k=1}^{K} \omega_{k,t} f_{kt}$

Forecast evaluation:

On a testing dataset $\{Y_t, f_{1t}, \ldots, f_{Kt}\}_{t=1,\ldots,T}$ and a loss function ℓ, we aim to minimise

$$\frac{1}{T} \sum_{t=1}^{T} \ell(Y_t, \hat{Y}_t)$$

Illustration



References

- Hannan(1957) and Blackwell et al. (1956) in a game theory framework
- Littlestone and Warmuth (1994) and Vovk (1990) in a statistical learning framework
- Cesa-Bianchi et al. (1997), Freund et al. (1997) and Vovk (1998) for theoretical results
- Cesa-Bianchi and Lugosi (2006) for a review
- Goude (2008) and Gaillard (2015) PhDs for an application to electricity consumption forecasting and the development of the « opera » package (in R and Python)

Regret

To assess the quality of the final forecast, a benchmark is needed!

We could look directly at the performance of $\hat{Y}_t$, but that wouldn't make much sense:

- if all the experts are bad, the mixture of forecasts will has poor performance, whereas it's possible that the aggregation performs well (that the mixture is better than each forecast)
- Conversely, if all the forecasts are good, it is highly likely that whatever the mix, it will be good

We need:

- a set for the weights ω_{kt} (the simplex of K-dimension for exemple)
- a set S of strategies to compare ourselves (the set of constant strategies for example)

$$\text{Regret: } R_T = \sum_{t=1}^{T} \ell\left(Y_t, \sum_{k=1}^{K} \omega_{kt} f_{kt}\right) - \min_{s \in S} \sum_{t=1}^{T} \ell\left(Y_t, \sum_{k=1}^{K} \omega_{kt}(s) f_{kt}\right)$$

Examples

Regret regarding the best expert:

$$R_T = \sum_{t=1}^{T} \ell\left(Y_t, \sum_{k=1}^{K} \omega_{kt} f_{kt}\right) - \min_{k=1,\ldots K} \sum_{t=1}^{T} \ell\left(Y_t, f_{kt}\right)$$

Regret regarding the best constant convex combination of experts:

$$R_T = \sum_{t=1}^{T} \ell\left(Y_t, \sum_{k=1}^{K} \omega_{kt} f_{kt}\right) - \min_{\omega_1,\ldots,\omega_K} \sum_{t=1}^{T} \ell\left(Y_t, \sum_{k=1}^{K} \omega_k f_{kt}\right)$$

with $\sum_{k=1}^{K} \omega_k = 1$ and $\forall k = 1,\ldots,K, \quad \omega_k \in [0,1]$

Question: What kind of regret should our strategy have?

Clue: What is the regret of a dumb strategy?

Regret bounds

If the loss function is bounded (true as soon as Y_t is too), the regret is at most proportional to T

→ our strategy should satisfy

$$\lim_{T\to\infty} \sup_{f_{1,1},\ldots,f_{k,t},\ldots,f_{K,T}} \frac{R_T}{T} \to 0$$

So as time goes by, we get closer to the strategy we're comparing ourselves to, or even better: we beat it!

Algorithms

Exponentially Weighted Aggregation (EWA)

Parameter: $\eta > 0$

Initialization:

- $\forall k = 1,\ldots,K, \quad \omega_{k,1} = \dfrac{1}{K}$ (uniform weights)
- Prediction: $\hat{Y}_1 = \dfrac{1}{K}\displaystyle\sum_{k=1}^{K} f_{k1}$ (empirical mean)

For $t = 2,\ldots,T$

- Weight updates: $\forall k = 1,\ldots,K, \quad \omega_{k,t} = \dfrac{\exp\big(-\eta \sum_{s=1}^{t-1} \ell(Y_s, f_{ks})\big)}{\sum_{j=1}^{K} \exp\big(-\eta \sum_{s=1}^{t-1} \ell(Y_s, f_{js})\big)}$
- Prediction: $\hat{Y}_t = \displaystyle\sum_{k=1}^{K} \omega_{k,t} f_{k,t}$ (empirical mean)

EWA regret bound (Stoltz, 2010)

Assumptions:

- Loss function $\ell : \mathbb{R} \times \mathbb{R} \to [0,M]$ is bounded
- $\forall Y$, $\ell(Y, \cdot\,)$ is convex

Then, for any $\eta > 0$

$$\sup_{f_{1,1},\ldots,f_{k,t},\ldots,f_{K,T}} \left(\sum_{t=1}^{T} \ell\left(Y_t, \hat{Y}_t^{\mathrm{EWA}}\right) - \min_{k=1,\ldots,K} \sum_{t=1}^{T} \ell(Y_t, f_{k,t}) \right) \leq \frac{\ln K}{\eta} + \frac{\eta M^2}{8} T$$

How to choose η?

EWA regret bound (Stoltz, 2010)

Assumptions:

- Loss function $\ell : \mathbb{R} \times \mathbb{R} \to [0, M]$ is bounded
- $\forall Y$, $\ell(Y, \cdot\,)$ is convex

Then, for any $\eta > 0$

$$\sup_{f_{1,1},\ldots,f_{k,t},\ldots,f_{K,T}} \left(\sum_{t=1}^{T} \ell\left(Y_t, \hat{Y}_t^{\mathrm{EWA}}\right) - \min_{k=1,\ldots,K} \ell(Y_t, f_{k,t}) \right) \leq \frac{\ln K}{\eta} + \frac{\eta M^2}{8} T$$

With $\eta = \dfrac{2}{M}\sqrt{\dfrac{2\ln K}{T}}$, we get $R_T = \mathcal{O}\left(M \sqrt{\dfrac{T}{2\ln K}} \right)$

EWA with Gradient Trick = Exponential Gradient

Parameter: $\eta > 0$

Initialization:

- $\forall k = 1,\ldots,K, \quad \omega_{k1} = \dfrac{1}{K}$ (uniform weights)
- Prediction: $\hat{Y}_1 = \dfrac{1}{K}\displaystyle\sum_{k=1}^{K} f_{k1}$ (empirical mean)

For $t = 2,\ldots,T$

- Weight updates: $\forall k = 1,\ldots,K, \quad \omega_k = \dfrac{\exp\big(-\eta\sum_{s=1}^{t-1}\partial\ell(Y_s,\hat{Y}_s)\cdot f_{ks}\big)}{\sum_{j=1}^{K}\exp\big(-\eta\sum_{s=1}^{t-1}\partial\ell(Y_s,\hat{Y}_s)\cdot f_{js}\big)}$
- Prediction: $\hat{Y}_t = \displaystyle\sum_{k=1}^{K}\omega_{kt}f_{kt}$ (empirical mean)

Exponential Gradient (EG) - L2

$$\forall k = 1,\ldots,K, \quad \omega_k = \frac{\exp\big(-\eta \sum_{s=1}^{t-1} 2(\hat{Y}_s - Y_s) f_{ks}\big)}{\sum_{j=1}^{K} \exp\big(-\eta \sum_{s=1}^{t-1} 2(\hat{Y}_s - Y_s) f_{js}\big)}$$

Intuition:

- If $\hat{Y}_s > Y_s$, experts who forecast the lowest values are at an advantage
- If $\hat{Y}_s < Y_s$, experts who forecast the highest values are at an advantage

In practice

How do you choose experts?

How do you choose experts?

Encouraging diversity!

Encouraging diversity!

→ Train models using a variety of data:
- Estimation periods
- Input variables / features
- Spatial / Temporal resolution

→ Consider various methods:
- Linear models
- Ensemble models
- Neural networks models
- Deliberately biased models, …

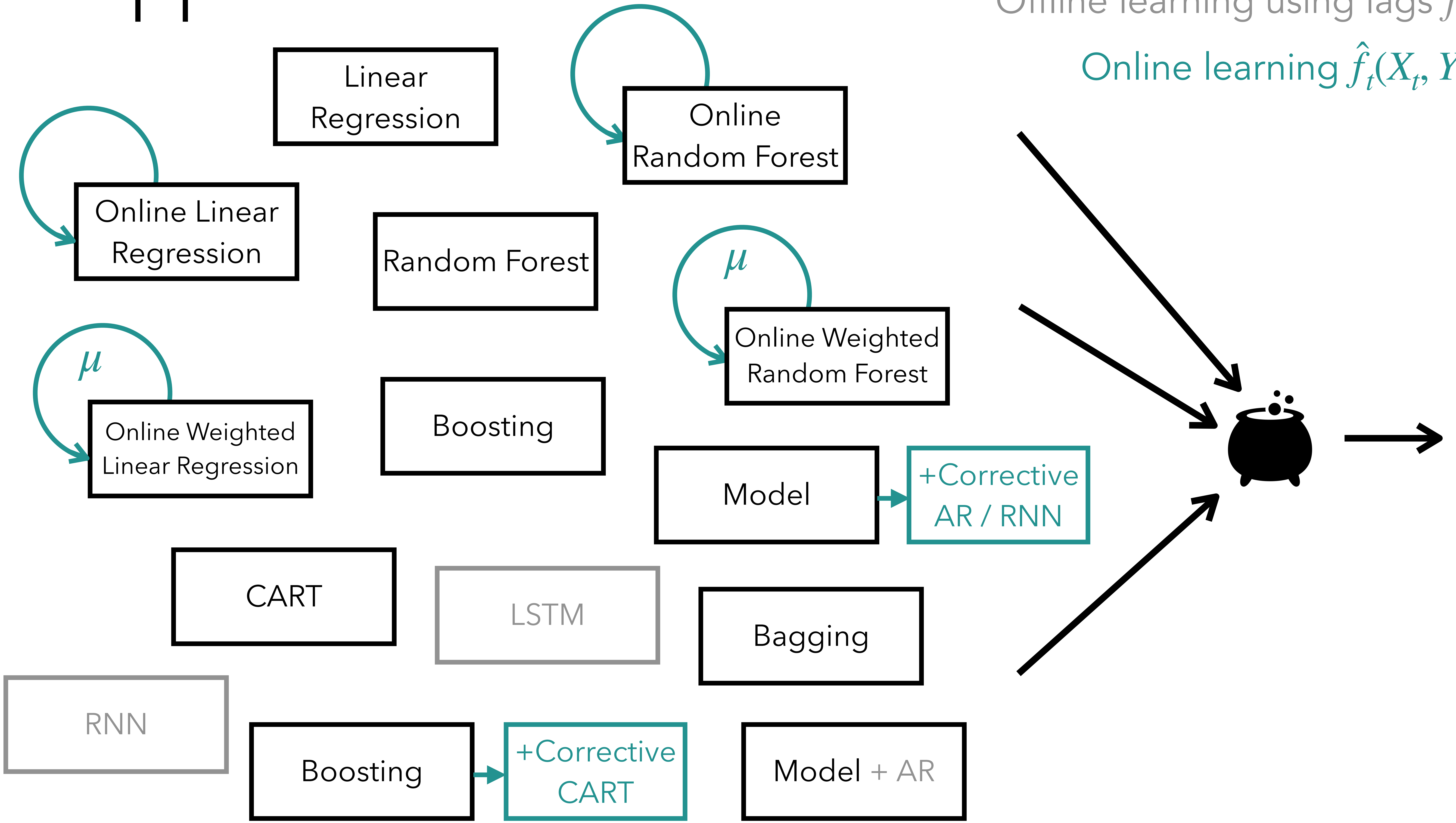
→ Consider various loss functions:
- L2
- L1
- Multiple quantile loss (so The variable to be forecast is in the convex envelope of the experts' forecasts)…

Application

Offline learning $\hat{f}(X_t)$

Offline learning using lags $\hat{f}(X_t, Y_{t-1}, Y_{t-2}, \ldots)$

Online learning $\hat{f}_t(X_t, Y_{t-1}, Y_{t-2}, \ldots)$



$$\hat{Y}_t = \sum_{k=1}^{K} \omega_{k,t} f_{kt}$$

That’s all folks!