Multiobjective Prediction with Expert Advice

Alexey Chernov

Computer Learning Research Centre and Department of Computer Science
Royal Holloway University of London

GTP Workshop, June 2010

Example: Prediction of Sport Match Outcome

V. Vovk, F. Zhdanov. Predictions with Expert Advice for Brier Game. ICML'08

Bookmakers data:
4 bookmakers, odds for $\sim$ 10000 tennis matches (2 outcomes)
8 bookmakers, odds for $\sim$ 9000 football matches (3 outcomes)

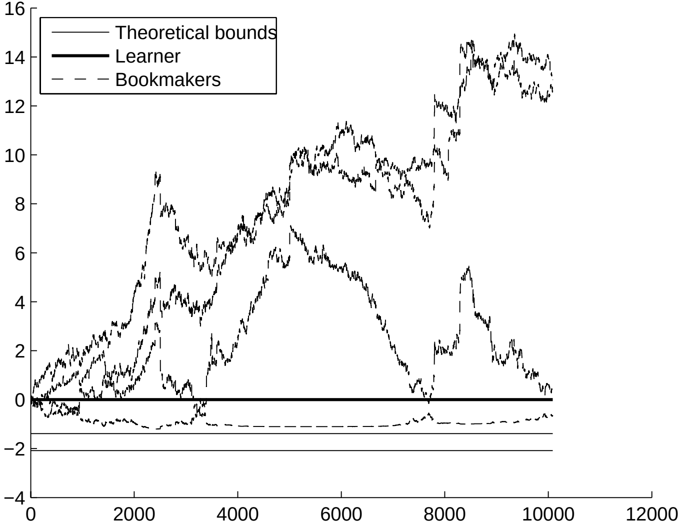
Odds a_i can be transformed to probabilities $Prob[i]$:

$$Prob[i] = \frac{1/a_i}{\sum_j 1/a_j}$$

The loss is measured by the square (Brier) loss function.

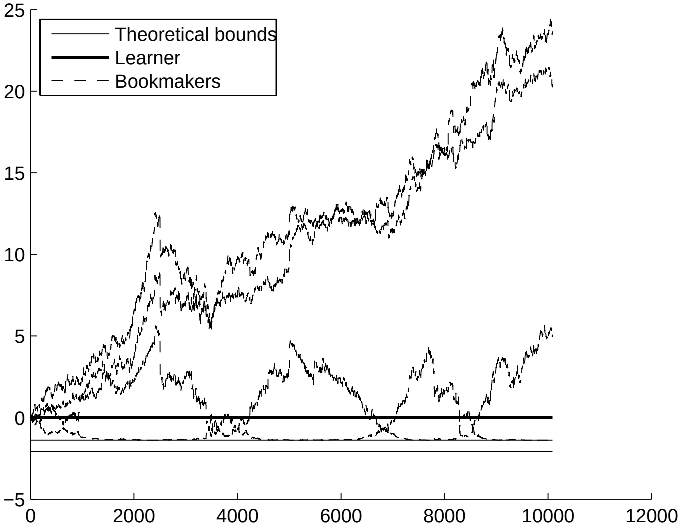
Learner's strategy is the Aggregating Algorithm.

Tennis Prediction, Square Loss



Graph of the negative regret $\mathrm{Loss}_{\mathcal{E}_k}(T) - \mathrm{Loss}(T)$, 4 Experts
Learner is the AA for the square loss

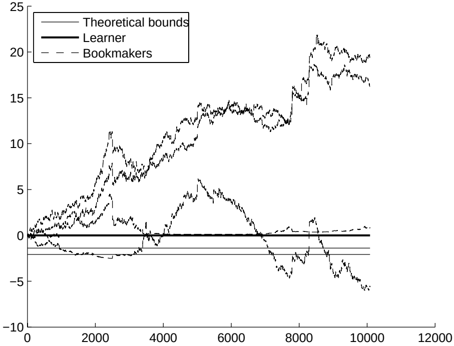
Tennis Prediction, Log Loss



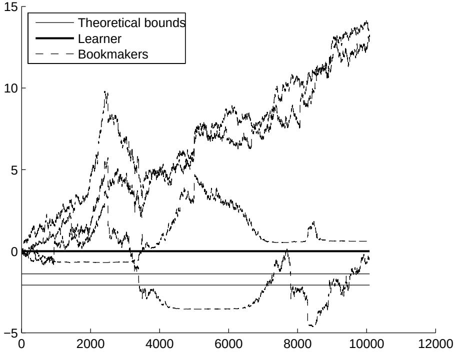
Graph of the negative regret $\mathrm{Loss}_{\mathcal{E}_k}(T) - \mathrm{Loss}(T)$, 4 Experts
Learner is the AA for the log loss (Bayes mixture)

Tennis Predictions: "Wrong" Losses

Graphs of the negative regret $\mathrm{Loss}_{\mathcal{E}_k}(T) - \mathrm{Loss}(T)$



log loss
the AA for the square loss



square loss
the AA for the log loss

Learner optimizes for a "wrong" loss function

Aggregating Algorithm with Wrong Losses

Fact

For the game with 2 outcomes, one can construct a sequence of predictions of 2 Experts and a sequence of outcomes with the following property. If Learner's predictions are generated by the Aggregating Algorithm for the log loss then for almost all T

$$\mathrm{Loss}(T) \geq \mathrm{Loss}_{\mathcal{E}_1}(T) + T/10\,,$$

where $\mathrm{Loss}(T)$ *and* $\mathrm{Loss}_{\mathcal{E}_1}(T)$ *are the square losses of Learner and Expert 1.*

A similar statement holds for the Aggregating Algorithm for the square loss evaluated by the log loss.

New Settings

Experts: $\gamma_t^{(1)}, \dots, \gamma_t^{(k)}$
Learner: γ_t
Reality: ω_t

Many loss functions

$$\mathrm{Loss}_{\mathcal{E}_k}^{(m)}(T) = \sum_{t=1}^{T} \lambda^{(m)}(\gamma_t^{(k)}, \omega_t)$$

$$\mathrm{Loss}^{(m)}(T) = \sum_{t=1}^{T} \lambda^{(m)}(\gamma_t, \omega_t)$$

New Settings

Experts: $\gamma_t^{(1)}, \dots, \gamma_t^{(k)}$
Learner: γ_t
Reality: ω_t

Many loss functions

$$\mathrm{Loss}_{\mathcal{E}_k}^{(m)}(T) = \sum_{t=1}^{T} \lambda^{(m)}(\gamma_t^{(k)}, \omega_t)$$

$$\mathrm{Loss}^{(m)}(T) = \sum_{t=1}^{T} \lambda^{(m)}(\gamma_t, \omega_t)$$

Expert Evaluator's advice

$$\mathrm{Loss}_{\mathcal{E}_k}^{(k)}(T) = \sum_{t=1}^{T} \lambda^{(k)}(\gamma_t^{(k)}, \omega_t)$$

$$\mathrm{Loss}^{(k)}(T) = \sum_{t=1}^{T} \lambda^{(k)}(\gamma_t, \omega_t)$$

Bound for New Settings

Theorem

If $\lambda^{(k)}$ are $\eta^{(k)}$-mixable proper loss functions, $k = 1, \ldots, K$, Learner has a strategy (e. g. the Defensive Forecasting algorithm) that guarantees, for all T and for all k, that

$$\mathrm{Loss}^{(k)}(T) \leq \mathrm{Loss}^{(k)}_{\mathcal{E}_k}(T) + \frac{1}{\eta^{(k)}} \ln K \,.$$

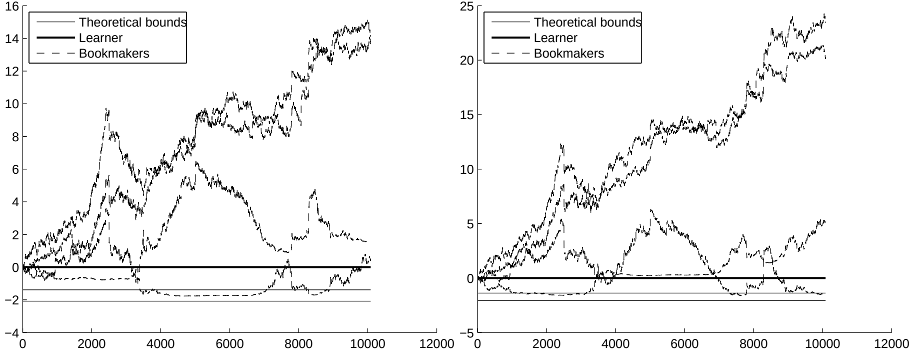
Corollary

If $\lambda^{(m)}$ are $\eta^{(m)}$-mixable proper loss functions, $m = 1, \ldots, M$, Learner has a strategy that guarantees, for all T, for all k and for all m, that

$$\mathrm{Loss}^{(m)}(T) \leq \mathrm{Loss}^{(m)}_{\mathcal{E}_k}(T) + \frac{1}{\eta^{(m)}} (\ln K + \ln M) \,.$$

Tennis Predictions, Two Losses

Graphs of the negative regret $\mathrm{Loss}_{\mathcal{E}_k}^{(m)}(T) - \mathrm{Loss}^{(m)}(T)$



square loss — log loss

Learner optimizes for both loss functions, using the DF algorithm.

Defensive Forecasting Algorithm

$$\exists\pi\ \forall\omega \quad \sum_{k=1}^{K} p_{t-1}^{(k)} \mathrm{e}^{\eta(\lambda(\pi,\omega)-\lambda(\pi_t^{(k)},\omega))} \leq 1\,,$$

where $p_{t-1}^{(k)} = p_0^{(k)} \mathrm{e}^{\eta(\mathrm{Loss}(t-1)-\mathrm{Loss}_{\mathcal{E}_k}(t-1))}$

To get this (from Levin's Lemma) we need that $\lambda(\pi,\omega)$ is continuous and for all π,π'

$$\mathbf{E}_\pi \mathrm{e}^{\eta(\lambda(\pi,\cdot)-\lambda(\pi',\cdot))} = \sum_{\omega\in\Omega} \pi(\omega) \mathrm{e}^{\eta(\lambda(\pi,\omega)-\lambda(\pi',\omega))} \leq 1$$

Defensive Forecasting Algorithm

$$\exists \pi \, \forall \omega \quad \sum_{k=1}^{K} p_{t-1}^{(k)} \mathrm{e}^{\eta^{(k)}(\lambda^{(k)}(\pi,\omega) - \lambda^{(k)}(\pi_t^{(k)},\omega))} \leq 1 \, ,$$

where $p_{t-1}^{(k)} = p_0^{(k)} \mathrm{e}^{\eta^{(k)}(\mathrm{Loss}^{(k)}(t-1) - \mathrm{Loss}_{\mathcal{E}_k}^{(k)}(t-1))}$

To get this (from Levin's Lemma) we need that $\lambda^{(k)}(\pi, \omega)$ is continuous and for all π, π'

$$\mathbf{E}_\pi \mathrm{e}^{\eta^{(k)}(\lambda^{(k)}(\pi,\cdot) - \lambda^{(k)}(\pi',\cdot))} = \sum_{\omega \in \Omega} \pi(\omega) \mathrm{e}^{\eta^{(k)}(\lambda^{(k)}(\pi,\omega) - \lambda^{(k)}(\pi',\omega))} \leq 1$$

The DFA and the AA

λ is continuous and $\forall \pi, \pi'$ $\mathbf{E}_\pi e^{\eta(\lambda(\pi,\cdot)-\lambda(\pi',\cdot))} \le 1 \quad \Rightarrow \quad \lambda$ is η-mixable

λ is η-mixable and ? $\quad \Rightarrow \quad \forall \pi, \pi'$ $\mathbf{E}_\pi e^{\eta(\lambda(\pi,\cdot)-\lambda(\pi',\cdot))} \le 1$

The DFA and the AA

λ is continuous and $\forall \pi, \pi'$ $\mathbf{E}_\pi e^{\eta(\lambda(\pi,\cdot)-\lambda(\pi',\cdot))} \le 1 \quad \Rightarrow \quad \lambda$ is η-mixable

$$\forall \pi, \pi' \; \mathbf{E}_\pi e^{\eta(\lambda(\pi,\cdot)-\lambda(\pi',\cdot))} \le 1 \quad \Rightarrow \quad \lambda \text{ is proper}$$

λ is η-mixable and proper $\quad \Rightarrow \quad \forall \pi, \pi'$ $\mathbf{E}_\pi e^{\eta(\lambda(\pi,\cdot)-\lambda(\pi',\cdot))} \le 1$

Proper Loss Functions

λ is proper if for any $\pi, \pi' \in \mathcal{P}(\Omega)$

$$\mathbf{E}_\pi \lambda(\pi, \cdot) \le \mathbf{E}_\pi \lambda(\pi', \cdot)$$

If $\omega \sim \pi$ then $\mathbf{E}_\pi \lambda(\pi', \omega)$ is the expected loss for prediction π'.

The expected loss is minimal for the true distribution
$\Rightarrow$ the forecaster is encouraged to give the true probabilities

The square loss and the log loss are proper.

Example: Hellinger Loss

$$\lambda^{Hellinger}(\gamma,\omega) = \frac{1}{2}\sum_{j=1}^{r}\left(\sqrt{\gamma(j)} - \sqrt{\mathbb{I}_{\{\omega=j\}}}\right)^2$$

The Hellinger loss is $\sqrt{2}$-mixable

The Hellinger loss is not proper

Proper Mixable Loss Functions

Each mixable loss function $\lambda(\gamma, \omega)$ has a proper analogue $\lambda^{\textit{proper}}(\pi, \omega)$ such that

1. $\forall\pi\exists\gamma\forall\omega\ \lambda^{\textit{proper}}(\pi, \omega) = \lambda(\gamma, \omega)$
2. $\forall\pi\forall\gamma \quad \mathbf{E}_\pi\lambda^{\textit{proper}}(\pi, \cdot) \leq \mathbf{E}_\pi\lambda(\gamma, \cdot)$

Proper Mixable Loss Functions

Each mixable loss function $\lambda(\gamma, \omega)$ has a proper analogue $\lambda^{proper}(\pi, \omega)$ such that

1. $\forall\pi\exists\gamma\forall\omega\ \lambda^{proper}(\pi, \omega) = \lambda(\gamma, \omega)$
2. $\forall\pi\forall\gamma \quad \mathbf{E}_\pi \lambda^{proper}(\pi, \cdot) \le \mathbf{E}_\pi \lambda(\gamma, \cdot)$

For the Hellinger loss, the proper analogue is the spherical loss

$$\lambda^{spherical}(\pi, \omega) = 1 - \frac{\pi(\omega)}{\sqrt{\sum_{j=1}^{r} (\pi(j))^2}}$$

$\lambda^{spherical}(\pi, \omega) = \lambda^{Hellinger}(\gamma, \omega)$ for $\gamma(\omega) = \frac{(\pi(\omega))^2}{\sum_{j=1}^{r} (\pi(j))^2}$

Example: Mixable and Non-Mixable Losses

Experts $1, \ldots, K$ predict $\pi^{(k)} \in \mathcal{P}(\{0,1\})$.
Experts $1, \ldots, N$ predict $\gamma^{(n)} \in \{0,1\}$.
Learner predicts $(\pi, \tilde{\pi}) \in \mathcal{P}(\{0,1\}) \times \mathcal{P}(\{0,1\})$ such that if $\pi(0) > 1/2$ then $\tilde{\pi}(0) = 1$ and if $\pi(1) > 1/2$ then $\tilde{\pi}(1) = 1$.

There exists a strategy for Learner that guarantees for any k

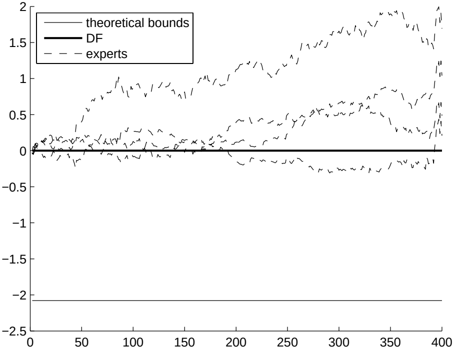
$$\sum_{t=1}^{T} \lambda^{square}(\pi_t, \omega_t) \le \sum_{t=1}^{T} \lambda^{square}(\pi_t^{(k)}, \omega_t) + \ln(K+N)$$

and for any m

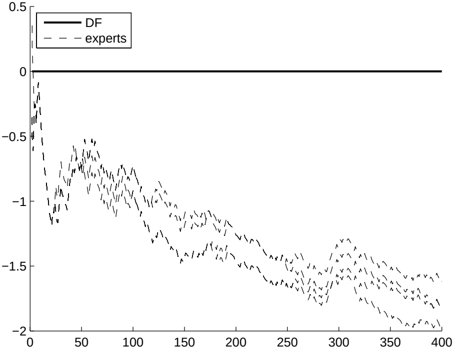
$$\sum_{t=1}^{T} \lambda^{abs}(\tilde{\pi}_t, \omega_t) \le \sum_{t=1}^{T} \lambda^{simple}(\gamma_t^{(n)}, \omega_t) + O(\sqrt{T \ln(K+N) + T \ln \ln T})$$

Tennis Predictions, Square and Absolute Losses

Graphs of the negative regret $\mathrm{Loss}_{\mathcal{E}_k}^{(m)}(T) - \mathrm{Loss}^{(m)}(T)$



square loss



absolute loss

Learner optimizes for both loss functions, using the DF algorithm with "mixability" and Hoeffding supermartingales.

The Mixed Supermartingale

$$\frac{1}{K+N}\sum_{k=1}^{K} e^{2\sum_{t=1}^{T-1}((p_t-\omega_t)^2-(p_t^{(k)}-\omega_t)^2)} \times e^{2((p-\omega)^2-(p_T^{(k)}-\omega)^2)}$$
$$+\frac{1}{K+N}\sum_{n=1}^{N}\int_0^{1/e}\frac{d\eta}{\eta\left(\ln\frac{1}{\eta}\right)^2} e^{\eta\sum_{t=1}^{T-1}(|\tilde{p}_t-\omega_t|-[\gamma_t^{(n)}\neq\omega_t])-\eta^2/2}$$
$$\times e^{\eta(|\tilde{p}-\omega|-[\gamma_T^{(n)}\neq\omega])-\eta^2/2}$$

where $p_t=\pi_t(1)$, $p_t^{(k)}=\pi_t^{(k)}(1)$, $\tilde{p}_t=\tilde{\pi}_t(1)$.
$[x\neq y]=1$ if $x\neq y$ and $[x\neq y]=0$ if $x=y$.

References

V. Vovk, F. Zhdanov. Predictions with Expert Advice for Brier Game. ICML 2008. `http://arxiv.org/abs/0710.0485`

A. Chernov, Y. Kalnishkan, F. Zhdanov, V. Vovk. Supermartingales in Prediction with Expert Advice. ALT 2008 and TCS. `http://arxiv.org/abs/1003.2218`

A. Chernov, V. Vovk. Prediction with expert evaluators' advice. ALT 2009. `http://arxiv.org/abs/0902.4127`

A. Chernov, V. Vovk. Prediction with Advice of Unknown Number of Experts. UAI 2010. `http://arxiv.org/abs/1006.0475`

`http://onlineprediction.net/`