Prediction with Expert Advice and Game-Theoretic Supermartingales

Alexey Chernov

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

22 March 2010, Durham

Outline

1. Framework of Prediction with Expert Advice
2. Motivation: Minimal Expected Loss, Calibration, Martingales
3. Defensive Forecasting

Sequence Prediction

Sequence of events

$$\omega_1, \omega_2, \omega_3, \ldots$$

Outcomes $\omega_t \in \Omega$

Sequence Prediction

Sequence of events

$$\omega_1, \omega_2, \omega_3, \ldots$$

Outcomes $\omega_t \in \Omega$

We try to predict the outcomes

$$\gamma_1, \omega_1, \gamma_2, \omega_2, \gamma_3, \omega_3, \ldots$$

Predictions $\gamma_t \in \Gamma$

Sequence Prediction

Sequence of events

$$\omega_1, \omega_2, \omega_3, \ldots$$

Outcomes $\omega_t \in \Omega$

We try to predict the outcomes

$$\gamma_1, \omega_1, \gamma_2, \omega_2, \gamma_3, \omega_3, \ldots$$

Predictions $\gamma_t \in \Gamma$

The quality of each prediction is measured by a loss function:

$$(\gamma, \omega) \mapsto \lambda(\gamma, \omega) \in \mathbb{R}$$

The quality of the first T predictions: $L_T = \sum_{t=1}^{T} \lambda(\gamma_t, \omega_t)$

Goal: $L_T \to \min$

Simple Loss

Two outcomes, two possible predictions
$\Gamma = \Omega = \{0, 1\}$

$$\lambda^{\text{simple}}(\gamma, \omega) = 1 - \mathbb{I}_{\{\gamma=\omega\}} = \begin{cases} 0 & \text{if } \gamma = \omega, \\ 1 & \text{if } \gamma \neq \omega \end{cases}$$

$\sum_{t=1}^{T} \lambda^{\text{simple}}(\gamma_t, \omega_t)$ is the number of errors

Absolute Loss

Two outcomes: $\Omega = \{0, 1\}$

Probabilistic predictions: $\Gamma = \{(\gamma(0), \gamma(1)) \in [0, 1]^2 \mid \gamma(0) + \gamma(1) = 1\}$

$$\lambda^{\text{abs}}(\gamma, \omega) = |\gamma(1) - \omega| = \gamma(0)\lambda^{\text{simple}}(0, \omega) + \gamma(1)\lambda^{\text{simple}}(1, \omega)$$

$\sum_{t=1}^{T} \lambda^{\text{abs}}(\gamma_t, \omega_t)$ is the expected number of errors

Brier Loss

G. Brier. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, 1950.

Finitely many outcomes: $\Omega = \{1, \ldots, r\}$

Probabilistic predictions:
$\Gamma = \{\gamma = (\gamma(1), \ldots, \gamma(r)) \in [0,1]^r \mid \sum_{j=1}^{r} \gamma(j) = 1\}$

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

$L_T^{Brier} \to \min$ encourages unbiased estimates of the true probabilities

Logarithmic Loss

Finitely many outcomes: $\Omega = \{1, \ldots, r\}$

Probabilistic predictions:
$\Gamma = \{\gamma = (\gamma(1), \ldots, \gamma(r)) \in [0,1]^r \mid \sum_{j=1}^{r} \gamma(j) = 1\}$

$$\lambda^{\log}(\gamma, \omega) = -\ln \gamma(\omega)$$

Measures the “quantity of information”.

Logarithmic Loss

P is a probability measure on all sequences $\omega_1\omega_2\omega_3\ldots \in \Omega^\infty$

Prediction strategy:

$$\gamma_{t+1} = P(\cdot \mid \omega_1 \ldots \omega_t)$$

that is $\gamma_{t+1}(\omega) = \frac{P(\omega_1 \ldots \omega_t \omega)}{P(\omega_1 \ldots \omega_t)}$

$$L_T = \sum_{t=1}^{T} \lambda^{\log}(\gamma_t, \omega_t) = -\ln \prod_{t=1}^{T} \frac{P(\omega_1 \ldots \omega_{t-1}\omega_t)}{P(\omega_1 \ldots \omega_{t-1})} = -\ln P(\omega_1 \ldots \omega_T)$$

$L_T \to \min \quad \Leftrightarrow \quad$ the likelihood $P(\omega_1 \ldots \omega_T) \to \max$.

Prediction with Expert Advice

At step t	Expert 1	$\ldots$	Expert K	Learner
Prediction	γ_t^1	$\cdots$	γ_t^K	
Outcome				

Prediction with Expert Advice

At step t	Expert 1	$\ldots$	Expert K	Learner
Prediction	γ_t^1	$\cdots$	γ_t^K	γ_t
Outcome				

Prediction with Expert Advice

At step t	Expert 1	...	Expert K	Learner
Prediction	γ_t^1	$\cdots$	γ_t^K	γ_t
Outcome		ω_t		

Prediction with Expert Advice

At step t	Expert 1	...	Expert K	Learner
Prediction	γ_t^1	$\cdots$	γ_t^K	γ_t
Outcome		ω_t		
Loss	$\lambda(\gamma_t^1, \omega_t)$	$\cdots$	$\lambda(\gamma_t^K, \omega_t)$	$\lambda(\gamma_t, \omega_t)$

Prediction with Expert Advice

At step t	Expert 1	$\ldots$	Expert K	Learner
Prediction	γ_t^1	$\cdots$	γ_t^K	γ_t
Outcome	ω_t			
Loss	$\lambda(\gamma_t^1, \omega_t)$	$\cdots$	$\lambda(\gamma_t^K, \omega_t)$	$\lambda(\gamma_t, \omega_t)$

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

Goal: after each step T, for any Expert k,

$$L_T \leq L_T^k + \text{something small}$$

Loss Bound

Theorem

If λ is an η-mixable loss function, Learner has strategy that guarantees

$$\sum_{t=1}^{T} \lambda(\gamma_t, \omega_t) \leq \sum_{t=1}^{T} \lambda(\gamma_t^k, \omega_t) + \frac{\ln K}{\eta}.$$

If λ is a convex loss function, Learner has strategy that guarantees

$$\sum_{t=1}^{T} \lambda(\gamma_t, \omega_t) \leq \sum_{t=1}^{T} \lambda(\gamma_t^k, \omega_t) + O(\sqrt{T \ln K}).$$

(Both bounds hold uniformly for all T and for all k.)

Log loss and Brier loss are 1-mixable.
Absolute loss is convex but not mixable. Simple loss is not convex.

λ is η-mixable if $\forall K \, \forall \gamma^k \in \Gamma \, \forall w^k \quad \exists \gamma \in \Gamma \, \forall \omega \quad e^{-\eta\lambda(\gamma,\omega)} \geq \sum_{k=1}^{K} w^k e^{-\eta\lambda(\gamma^k,\omega)}$.

Example: Bayesian Prediction (1)

Logarithmic loss $\lambda^{\log}(\gamma, \omega) = -\ln \gamma(\omega)$
Experts are probability measures $P^1, \ldots, P^K$:

$$\gamma_T^k(\omega) = P^k(\omega \mid \omega_1 \ldots \omega_{T-1})$$

Example: Bayesian Prediction (1)

Logarithmic loss $\lambda^{\log}(\gamma, \omega) = -\ln \gamma(\omega)$
Experts are probability measures $P^1, \ldots, P^K$:

$$\gamma_T^k(\omega) = P^k(\omega \mid \omega_1 \ldots \omega_{T-1})$$

Learner's strategy is a mixture:

$$P = \sum_{k=1}^{K} w^k P^k$$

Example: Bayesian Prediction (1)

Logarithmic loss $\lambda^{\log}(\gamma, \omega) = -\ln \gamma(\omega)$
Experts are probability measures $P^1, \ldots, P^K$:

$$\gamma_T^k(\omega) = P^k(\omega \mid \omega_1 \ldots \omega_{T-1})$$

Learner's strategy is a mixture:

$$P = \sum_{k=1}^{K} w^k P^k$$

$$\gamma_T(\omega) = P(\omega \mid \omega_1 \ldots \omega_{T-1}) = \frac{\sum_{k=1}^K w^k P^k(\omega_1 \ldots \omega_{T-1}\omega)}{\sum_{k=1}^K w^k P^k(\omega_1 \ldots \omega_{T-1})}$$

$$= \sum_{k=1}^{K} \frac{w^k \prod_{t=1}^{T-1} \gamma_t^k(\omega_t)}{\sum_{i=1}^K w^i \prod_{t=1}^{T-1} \gamma_t^i(\omega_t)} \gamma_t^k(\omega)$$

Example: Bayesian Prediction (2)

Logarithmic loss $\lambda^{\log}(\gamma, \omega) = -\ln \gamma(\omega)$
Experts are probability measures $P^1, \ldots, P^K$
Learner's strategy:

$$P = \sum_{k=1}^{K} \frac{1}{K} P^k$$

Example: Bayesian Prediction (2)

Logarithmic loss $\lambda^{\log}(\gamma, \omega) = -\ln \gamma(\omega)$
Experts are probability measures $P^1, \ldots, P^K$
Learner's strategy:

$$P = \sum_{k=1}^{K} \frac{1}{K} P^k$$

Then for any $\omega_1 \ldots \omega_T$

$$P(\omega_1 \ldots \omega_T) \geq \frac{1}{K} P_k(\omega_1 \ldots \omega_T)$$

$$L_T = -\ln P(\omega_1 \ldots \omega_T) \leq -\ln P_k(\omega_1 \ldots \omega_T) + \ln K = L_T^k + \ln K$$

Counterexample: Simple Game of Prediction

$\omega, \gamma \in \{0, 1\}$

$$\lambda^{\text{simple}}(\gamma, \omega) = 1 - \mathbb{I}_{\{\gamma=\omega\}} = \begin{cases} 0 & \text{if } \gamma = \omega, \\ 1 & \text{if } \gamma \neq \omega \end{cases}$$

Experts:

$$\gamma_t^1 = 0,\ \gamma_t^2 = 1 \quad \forall t$$

Outcome:

$$\omega_t = 1 - \gamma_t$$

$$L_T = T,\ L_T^1 + L_T^2 = T \quad \Rightarrow \quad L_T \geq \min_k L_T^k + T/2$$

Outline

Minimal Expected Loss

At step t, ω_t is sampled from a distribution P_t
and Learner knows the distributions P_t

Learner's prediction: $\gamma_t = \arg\min_{\gamma \in \Gamma} \mathbf{E}_t \lambda(\gamma, \omega_t)$

Then

$$\sum_{t=1}^{T} \mathbf{E}_t \lambda(\gamma_t, \omega_t) \leq \sum_{t=1}^{T} \mathbf{E}_t \lambda(\gamma_t^k, \omega_t)$$

Minimal Expected Loss

At step t, ω_t is sampled from a distribution P_t
and Learner knows the distributions P_t

Learner's prediction: $$\gamma_t = \arg\min_{\gamma \in \Gamma} \mathbf{E}_t \lambda(\gamma, \omega_t)$$

Then with high probability

$$\sum_{t=1}^{T} \lambda(\gamma_t, \omega_t) + O(\sqrt{T})$$

$$\|$$

$$\sum_{t=1}^{T} \mathbf{E}_t \lambda(\gamma_t, \omega_t) \quad \leq \quad \sum_{t=1}^{T} \mathbf{E}_t \lambda(\gamma_t^k, \omega_t)$$

$$\|$$

$$\sum_{t=1}^{T} \lambda(\gamma_t^k, \omega_t) + O(\sqrt{T})$$

Calibration

Dawid, 1982

Sequence of outcomes $\omega_t \in \{0,1\}$:

$$\omega_1, \omega_2, \omega_3, \ldots$$

We consider probability forecasts $p_t \in [0,1]$:

$$p_1, \omega_1, p_2, \omega_2, p_3, \omega_3, \ldots$$

Forecasts are well-calibrated if for any $p \in [0,1]$

$$\frac{\sum_{t\colon p_t=p} \omega_t}{\#\{t\colon p_t = p\}} \quad \to \quad p$$

“Ignorant” Calibration

Theorem (Foster, Vohra, 1998)

There is a randomised strategy constructing p_t given $\omega_1 \ldots \omega_{t-1}$ s.t. for any $\omega_1\omega_2\ldots$ the forecasts p_t are well-calibrated with high probability

“Ignorant” Calibration

Theorem (Foster, Vohra, 1998)

There is a randomised strategy constructing p_t given $\omega_1 \dots \omega_{t-1}$ s.t. for any $\omega_1\omega_2 \dots$ the forecasts p_t are well-calibrated with high probability

Generally:
P is a distribution on $\vec{\omega} \in \Omega^\infty$, $\mathit{Test}(P, \vec{\omega}) \in \{\mathit{accept}, \mathit{reject}\}$

Theorem (Sandroni, 2003)

If Test accepts $\vec{\omega}$ sampled from P with P-probability $1-\epsilon$ for any P then there is a randomised strategy that constructs P on-line given $\vec{\omega}$ s.t. $\mathit{Test}(P, \vec{\omega})$ accepts with probability $1-\epsilon$.

Informal Idea: "Ignorant" Expected Loss

Given $\omega_1, \omega_2, \ldots$ and Expert's γ^k
we want to construct a distribution P s.t.

$$\mathbf{E} \sum_{t=1}^{T} \lambda(\gamma_t^P, \omega_t) = \sum_{t=1}^{T} \lambda(\gamma_t^P, \omega_t) + O(\sqrt{T})$$

and

$$\mathbf{E} \sum_{t=1}^{T} \lambda(\gamma_t^k, \omega_t) = \sum_{t=1}^{T} \lambda(\gamma_t^k, \omega_t) + O(\sqrt{T})$$

where

$$\gamma_t^P = \arg\min_{\gamma \in \Gamma} \mathbf{E}_t \lambda(\gamma, \omega_t)$$

Martingales

$\omega_1\omega_2\ldots$ sampled from some distribution P
$S_t = S(\omega_1, \ldots, \omega_t)$

S is a martingale if

$$\mathbf{E}[S_t \mid \omega_1, \ldots, \omega_{t-1}] = S_{t-1}$$

Martingales

$\omega_1\omega_2\ldots$ sampled from some distribution P
$S_t = S(\omega_1, \ldots, \omega_t)$

S is a martingale if

$$\mathbf{E}[S_t \mid \omega_1, \ldots, \omega_{t-1}] = S_{t-1}$$

Theorem (Ville, 1939)

If $P(A) < \epsilon$ then a supermartingale S exists s.t. $\lim_{t\to\infty} S(\omega_1, \ldots, \omega_t) \geq 1/\epsilon$ for $\vec{\omega} \in A$.

Martingales

$\omega_1\omega_2\ldots$ sampled from some distribution P
$S_t = S(\omega_1, \ldots, \omega_t)$

S is a martingale if

$$\mathbf{E}[S_t \mid \omega_1, \ldots, \omega_{t-1}] = S_{t-1}$$

Theorem (Ville, 1939)

If $P(A) < \epsilon$ then a supermartingale S exists s.t. $\lim_{t\to\infty} S(\omega_1, \ldots, \omega_t) \geq 1/\epsilon$ for $\vec{\omega} \in A$.

Sandroni theorem test: $P\{\vec{\omega} \mid \mathit{Test}(P, \omega) = \mathit{reject}\} < \epsilon$
(i.e., uniformly $P(A_P) \leq \epsilon$)

Outline

Game-Theoretic Supermartingales

Informally:
S_t is player's capital after round t
ω_t is outcome of round t
distribution P is the rules of the game

If player has a uniform strategy for all P then S_t is a function of P, $\vec{\omega}$ and also player's additional knowledge

Game-Theoretic Supermartingales

Informally:
S_t is player's capital after round t
ω_t is outcome of round t
distribution P is the rules of the game

If player has a uniform strategy for all P then S_t is a function of P, $\vec{\omega}$ and also player's additional knowledge

$\omega_1, \omega_2, \ldots \in \Omega$
$\pi_1, \pi_2, \ldots$ are distributions on Ω
S is a game-theoretic supermartingale if for any π

$$\int_\Omega S(e_1, \pi_1, \omega_1, \ldots, e_T, \pi, \omega)\pi(d\omega) \leq S(e_1, \pi_1, \omega_1, \ldots, e_{T-1}, \pi_{T-1}, \omega_{T-1})$$

Levin's Lemma

Lemma (Levin, 1976)

If $s(\pi, \omega)$ is continuous in π and for some C

$$\forall \pi \quad \int_{\Omega} s(\pi, \omega) \pi(d\omega) \leq C$$

then there exists π s.t.

$$\forall \omega \quad s(\pi, \omega) \leq C$$

Levin's Lemma

Lemma (Levin, 1976)

If $s(\pi, \omega)$ is continuous in π and for some C

$$\forall \pi \quad \int_\Omega s(\pi, \omega)\pi(d\omega) \le C$$

then there exists π s.t.

$$\forall \omega \quad s(\pi, \omega) \le C$$

Proof idea: Consider $\phi(\pi', \pi) = \int_\Omega s(\pi, \omega)\pi'(d\omega)$

$$\begin{aligned} s(\pi, \omega_0) = \int_\Omega s(\pi, \omega)\delta_{\omega_0}(d\omega) &= \phi(\delta_{\omega_0}, \pi) \\ \le \max_{\pi'} \phi(\pi', \pi) = \min_\pi \max_{\pi'} \phi(\pi', \pi) &= \max_{\pi'} \min_\pi \phi(\pi', \pi) \\ &\le \max_{\pi'} \phi(\pi', \pi') \le C \end{aligned}$$

Supermartingales for PEA: Mixable Games

If $\lambda(\gamma,\omega)$ is η-mixable then for any distribution π and for any $\gamma \in \Gamma$

$$\int_{\Omega} e^{\eta(\lambda(\pi,\omega)-\lambda(\gamma,\omega))}\pi(d\omega) \le 1$$

where $\lambda(\pi,\omega)$ is a proper loss function: for any π and any $\gamma \in \Gamma$

$$\int_{\Omega} \lambda(\pi,\omega)\pi(d\omega) \le \int_{\Omega} \lambda(\gamma,\omega)\pi(d\omega)$$

Supermartingales for PEA: Mixable Games

If $\lambda(\gamma,\omega)$ is η-mixable then for any distribution π and for any $\gamma \in \Gamma$

$$\int_\Omega e^{\eta(\lambda(\pi,\omega)-\lambda(\gamma,\omega))}\pi(d\omega) \le 1$$

where $\lambda(\pi,\omega)$ is a proper loss function: for any π and any $\gamma \in \Gamma$

$$\int_\Omega \lambda(\pi,\omega)\pi(d\omega) \le \int_\Omega \lambda(\gamma,\omega)\pi(d\omega)$$

$$S_T = \sum_{k=1}^{K}\left(\frac{1}{K}\prod_{t=1}^{T} e^{\eta(\lambda(\pi_t,\omega_t)-\lambda(\gamma_t^k,\omega_t))}\right)$$

is a supermartingale.
Choosing π_t by Levin's lemma, we can guarantee that $S_T \le 1$ for all T.

Supermartingales for PEA: Logarithmic Loss

Consider $\lambda^{\log}(\gamma, \omega) = -\ln \gamma(\omega)$ (which is 1-mixable)
For any distribution π and for any $\gamma \in \Gamma$

$$\int_\Omega e^{\lambda^{\log}(\pi,\omega) - \lambda^{\log}(\gamma,\omega)} \pi(d\omega)$$

$$= \sum_{\omega \in \Omega} e^{-\ln \pi(\omega) + \ln \gamma(\omega)} \pi(\omega) = \sum_{\omega \in \Omega} \frac{\gamma(\omega)}{\pi(\omega)} \pi(\omega) = 1$$

$$S_T = \sum_{k=1}^{K} \frac{1}{K} \prod_{t=1}^{T} \frac{\gamma_t^k(\omega_t)}{\pi_t(\omega_t)} \le 1$$

Supermartingales for PEA: Convex Games

If $\lambda(\gamma, \omega)$ is convex then
for any distribution π, for any $\gamma \in \Gamma$, for any $\eta > 0$,

$$\int_\Omega e^{\eta(\lambda(\pi,\omega)-\lambda(\gamma,\omega))-\eta^2/2}\pi(d\omega) \le 1$$

where $\lambda(\pi, \omega)$ is a proper loss (multi-)function: for any π and any $\gamma \in \Gamma$

$$\int_\Omega \lambda(\pi, \omega)\pi(d\omega) \le \int_\Omega \lambda(\gamma, \omega)\pi(d\omega)$$

$$S_T = \sum_{k=1}^{K} \left(\frac{1}{K} \prod_{t=1}^{T} e^{\eta(\lambda(\pi_t,\omega_t)-\lambda(\gamma_t^k,\omega_t))-\eta^2/2} \right)$$

is a supermartingale.
Letting $\eta = O(1/\sqrt{T})$ and choosing π_t by Levin's lemma, we can guarantee that $S_T \le 1$ for all T.

Laws of Probability (1)

Probability law: $P(A_P)$ is small for any P
Game-theoretic supermartingales correspond to probability laws

Laws of Probability (1)

Probability law: $P(A_P)$ is small for any P
Game-theoretic supermartingales correspond to probability laws

Supermartingale for convex games: Hoeffding inequality
If $X \in [-1, 1]$ then

$$\mathbf{E}\mathrm{e}^{\eta X} \leq \mathrm{e}^{\eta \mathbf{E} X + \eta^2/2}$$

For independent $X_1, \ldots X_N \in [-1, 1]$

$$P\left[\frac{1}{N}\left|\sum_{n=1}^{N}(X_n - \mathbf{E}X_n)\right| > \epsilon\right] \leq 2\mathrm{e}^{-\epsilon^2 N/2}$$

Laws of Probability (2)

Supermartingale for mixable games:

λ is proper η-mixable loss function,
P is any distribution, $\pi_t = P(\omega \mid \omega_1 \dots \omega_{t-1})$,
$P^1, \dots, P^K$ are any distributions and $\pi_t^k = P^k(\omega \mid \omega_1 \dots \omega_{t-1})$

$$P\left\{\vec{\omega} \,\middle|\, \forall T \forall k = 1, \dots, K \ \sum_{t=1}^{T} \lambda(\pi_t, \omega_t) \geq \sum_{t=1}^{T} \lambda(\pi_t^k, \omega_t) + \frac{1}{\eta} \ln \frac{K}{\delta}\right\} \leq \delta$$

Special case: $\lambda^{\log}(\pi, \omega) = -\ln \pi(\omega)$

$$P\left\{\vec{\omega} \,\middle|\, \forall T \, \forall k = 1, \dots, K \quad \frac{P^k(\omega_1 \dots \omega_t)}{P(\omega_1 \dots \omega_t)} \geq \frac{\delta}{K}\right\} \leq \delta$$

References

N. Cesa-Bianchi, G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, Cambridge, England, 2006.

G. Shafer, V. Vovk. *Probability and Finance: It's Only a Game*! Wiley, New York, 2001.

V. Vovk. Predictions as statements and decisions. `http://arxiv.org/abs/cs/0606093`

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

`http://vovk.net/df/index.html`

`http://onlineprediction.net/`

These slides:
`pareto.cs.rhul.ac.uk/~chernov/PEAmartingales.pdf`