From Competitive Investment to Aggregating Algorithm and Defensive Forecasting

Yuri Kalnishkan

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

June 2010
Outline

1. Laissez-Faire Investment
2. Aggregating Algorithm
3. Defensive Forecasting
1. Laissez-Faire Investment

2. Aggregating Algorithm

3. Defensive Forecasting
Sequential Investment (1)

- there are $M$ stocks ($0, 1, \ldots, M - 1$) we can invest into
  — no cash or deposit (or one of the stocks is the deposit)
  — no inflation
- time is discrete, $t = 0, 1, 2, \ldots$
- an investment decision is a vector

$$\gamma_t = (\gamma_{t,0}, \gamma_{t,1}, \ldots, \gamma_{t,M-1})$$

such that $\gamma_{t,i} \in [0, 1]$ and $\sum_{j=0}^{M-1} \gamma_{t,i} = 1$
— it shows the distribution of our capital among the stocks
— on step $t - 1$ we spend the fraction $\gamma_{t,j}$ of our capital to buy stock $j$, $j = 0, 1, 2, \ldots, M - 1$
Sequential Investment (2)

- vector $x_t$ represents the market shift
  - let $x_{t,j}$ be the ratio of the prices of stock $i$ at the moments $t$ and $t-1$
  - assume $x_{t,j} \geq 0$
- between steps $t-1$ and $t$ our capital changes from $W_{t-1}$ to

$$W_t = \sum_{j=0}^{M-1} W_{t-1} \gamma_{t,j} x_{t,j} = W_{t-1} \langle \gamma_t, x_t \rangle$$

- if we start from $W_0 = 1$, then after $T$ steps we get

$$W_T = \prod_{t=1}^{T} \langle \gamma_t, x_t \rangle$$
Experts

• suppose that there are $N$ experts $E_1, E_2, \ldots, E_N$ that suggest investment decisions to us — before deciding on $\gamma_t$, we can observe decisions $\gamma_t^{(i)}$, $i = 1, 2, \ldots, N$, output by the experts

• we want to merge experts decisions in such a way so that our capital $W_t$ is not much less than the capital of any expert $W_t^{(i)}$, $i = 1, 2, \ldots, N$

— we want an inequality of the type $W_t \gtrsim W_t^{(i)}$ to hold uniformly for all $i$ and possibly for all $t$

• no assumptions are made as to how the experts arrive at their decisions
Laissez-Faire Merging

• we can think of a merging strategy as follows: on step \( t - 1 \) we invest the share \( p_t^{(i)} \) of our wealth as suggested by expert \( \mathcal{E}_i \)

• we need to decide how much money to entrust to expert \( \mathcal{E}_i \)

• let the experts be self-financing!
  — initially we split the capital among the experts so that \( \mathcal{E}_i \) gets \( p_0^{(i)} \) and then the experts operate on their share of the wealth
  — we do not redistribute the wealth

• then at every moment in time

\[
W_t = \sum_{i=1}^{N} p_0^{(i)} W_t^{(i)} \text{ and } W_t \geq p_0^{(i)} W_t^{(i)}
\]
Weights for the Merging

- the actual weights $p_t^{(i)}$ are given by

$$p_t^{(i)} = p_0^{(i)} W_t^{(i)}/W_t$$

$$= p_0^{(i)} W_t^{(i)} / \sum_{k=1}^{N} p_0^{(k)} W_t^{(i)}.$$  

— take the values $p_0^{(i)} W_t^{(i)}$ and normalise them to sum up to 1

- our investment decision is the weighted sum

$$\gamma_t = \sum_{i=1}^{N} \gamma_t^{(i)} p_t^{(i)}$$
1. Laissez-Faire Investment

2. Aggregating Algorithm

3. Defensive Forecasting
Sequential Prediction

- outcomes $\omega_t \in \Omega$ occur sequentially in discrete time: $\omega_1, \omega_2, \ldots$
- we need to output a prediction $\gamma_t \in \Gamma$ before seeing $\omega_t$
- the quality is measured by the loss function $\lambda : \Gamma \times \Omega \rightarrow [0, +\infty]$
  
  — our performance is measured by the cumulative loss $\text{Loss}(T) = \sum_{t=1}^{T} \lambda(\gamma_t, \omega_t)$
- a triple $\langle \Omega, \Gamma, \lambda \rangle$ (outcome space / prediction space / loss function) is called a game
Cover’s Game

- let
  - $\Omega = [0, +\infty)^M$ (the positive orthant),
  - $\Gamma = \{ \gamma = (\gamma_0, \gamma_1, \ldots, \gamma_M \in \mathbb{R}^M | \sum_{i=0}^{M-1} \gamma_j = 1 \}$ (the $(M - 1)$-dimensional simplex) and
  - $\lambda(\gamma, \omega) = -\ln \langle \gamma, \omega \rangle$

- this game is called Cover’s game and it describes the sequential investment scenario:

$$\text{Loss}(T) = -\ln W_T$$
Binary Games

• binary games have $\Omega = \{0, 1\}$ and $\Gamma = [0, 1]$
• square-loss game has $\lambda(\gamma, \omega) = (\omega - \gamma)^2$
• absolute-loss game has $\lambda(\gamma, \omega) = |\omega - \gamma|$
• logarithmic game has
  \[
  \lambda(\gamma, \omega) = \begin{cases} 
  -\log_2 \gamma, & \text{if } \omega = 1; \\
  -\log_2(1 - \gamma), & \text{if } \omega = 0
  \end{cases}
  \]
• simple prediction game has
  \[
  \lambda(\gamma, \omega) = \begin{cases} 
  0, & \text{if } \omega = \gamma; \\
  1, & \text{otherwise}
  \end{cases}
  \]
Prediction with Expert Advice

• suppose that we have access to predictions of $N$ experts $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_N$

  (1) FOR $t = 1, 2, \ldots$
  (2) we read $\gamma_t^{(1)}, \gamma_t^{(2)}, \ldots, \gamma_t^{(N)} \in \Gamma$
  (3) we output $\gamma_t \in \Gamma$
  (4) we observe $\omega_t \in \Omega$
  (5) END FOR

• we want to suffer loss little worse than the best expert — i.e., we want an inequality of the type $\text{Loss}(t) \lesssim \text{Loss}_{\mathcal{E}_i}(t)$ to hold uniformly for all $i$ and possibly for all $t$
Cover’s Game

- for Cover’s game laissez-faire investment yields

\[ W_t \geq p_0^{(i)} W_t^{(i)} \]

— by taking the logarithms we get

\[ \text{Loss}(t) \leq \text{Loss}_{\varepsilon_i}(t) + \ln(1/p_0^{(i)}) \]

- can we extend the algorithm to other games?
Extension of Laissez-Faire Investment (1)

- the wealth is the exponent of negative loss
  — let us introduce a parameter $\eta > 0$ (learning rate) to consider different bases
  — so $W_t^{(i)} = e^{-\eta \text{Loss}_E(t)}$

- consider the pseudo-algorithm that follows the laissez-faire investment strategy: it divides up its initial ‘wealth’ among the experts and lets them invest their shares
  — it has the notional ‘wealth’

$$\tilde{W}_t = \sum_{i=1}^{N} p_0^{(i)} W_t^{(i)}$$
Extension of Laissez-Faire Investment (2)

- but we need to produce actual predictions not ‘investment decisions’
- we can calculate the weights

\[ p_t^{(i)} = p_0^{(i)} W_t^{(i)} / \sum_{i=1}^{N} p_0^{(i)} W_t^{(i)} \]

\[ = p_0^{(i)} e^{-\eta \text{Loss} \varepsilon_i(t)} / \sum_{k=1}^{N} p_0^{(i)} e^{-\eta \text{Loss} \varepsilon_k(t)} \]

- but we cannot take \( \gamma_t = \sum_{i=1}^{N} \gamma_t^{(i)} p_t^{(i)} \) because there is no linearity in the general case
The Capital Inequality (1)

- we would like to achieve $e^{-\eta \text{Loss}(t)} = W_t \geq \tilde{W}_t$
- we can do it by making sure that

$$\frac{W_t}{W_{t-1}} \geq \frac{\tilde{W}_t}{\tilde{W}_{t-1}}$$

on every step
— we shall call this the capital inequality
- we have

$$\frac{W_t}{W_{t-1}} = e^{-\eta \lambda(\gamma_t, \omega_t)}$$

and

$$\frac{\tilde{W}_t}{\tilde{W}_{t-1}} = \sum_{i=1}^{N} p^{(i)}_t \frac{\tilde{W}_{t-1}}{\tilde{W}_{t-1}} e^{-\eta \lambda(\gamma^{(i)}_t, \omega_t)} = \sum_{i=1}^{N} p^{(i)}_t e^{-\eta \lambda(\gamma^{(i)}_t, \omega_t)}$$
The Capital Inequality (2)

- the capital inequality thus amounts to

\[ e^{-\eta \lambda(\gamma_t, \omega_t)} \geq \sum_{i=1}^{N} p_t^{(i)} e^{-\eta \lambda(\gamma_t^{(i)}, \omega_t)} \]

- we need to choose \( \gamma_t \) before we have seen \( \omega_t \) so we need to make sure this holds for all \( \omega_t \)

- so we need to look for \( \gamma \) such that

\[ \lambda(\gamma, \omega) \leq -\frac{1}{\eta} \ln \sum_{i=1}^{N} p_0^{(i)} e^{-\eta \lambda(\gamma^{(i)}, \omega)} \]

for all \( \omega \in \Omega \)

- are there any such \( \gamma \in \Gamma \)?
Superpredictions

- each prediction $\gamma$ can be thought of as a function $g : \Omega \rightarrow [0, +\infty]$ acting as follows: $g(\omega) = \lambda(\gamma, \omega)$
- a superprediction is a function $g : \Omega \rightarrow [0, +\infty]$ such that there is $\gamma \in \Gamma$ such that $g(\omega) \geq \lambda(\gamma, \omega)$ for all $\omega \in \Omega$ — we will denote the set of superpredictions by $\Sigma_\Gamma$
- if $\Omega$ is finite and $\Omega = \{\omega_0, \omega_1, \ldots, \omega_{K-1}\}$, the set of superpredictions can be identified with the set of points $s = (s_0, s_1, \ldots, s_{K-1}) \in \mathbb{R}^K$ such that
  
  $s_0 \geq \lambda(\gamma, \omega_0)$
  
  $s_1 \geq \lambda(\gamma, \omega_1)$
  
  $\ldots$
  
  $s_{K-1} \geq \lambda(\gamma, \omega_{K-1})$

for some $\gamma$
Sets of Superpredictions for Binary Games (1)

square-loss game
\[ \lambda(\gamma, \omega) = (\omega - \gamma)^2 \]

absolute-loss game
\[ \lambda(\gamma, \omega) = |\omega - \gamma| \]
Sets of Superpredictions for Binary Games (2)

logarithmic game

$$\lambda(\gamma, \omega) = \begin{cases} -\log_2(1 - \gamma), & \omega = 0 \\ -\log_2 \gamma, & \omega = 1 \end{cases}$$

simple prediction game

$$\lambda(\gamma, \omega) = \begin{cases} 0, & \omega = \gamma \\ 1, & \omega \neq \gamma \end{cases}$$
Mixability (1)

- Consider the transformation $\mathcal{B}_\eta : [0, +\infty]^{\Omega} \rightarrow [0, 1]^{\Omega}$ that transforms $g(\omega)$ into $e^{-\eta g(\omega)}$.

- If $\mathcal{B}_\eta(\Sigma_\Gamma)$, the image of the set of superpredictions under $\mathcal{B}_\eta$ is convex, we call the game $\eta$-mixable — for an $\eta$-mixable game we can always find $\gamma$ to satisfy the capital inequality.
Mixability (2)

- mixability implies convexity of the set of superpredictions, but it is a bit stronger than it — the logarithmic and square-loss game are mixable (for some $\eta$) — the absolute-loss and simple prediction games are not mixable
- for an $\eta$-mixable game we can make sure that the capital inequality holds and therefore $W_t \geq p_0^{(i)} W_t^{(i)}$, i.e.,

$$\text{Loss}(t) \leq \text{Loss}_{\varepsilon_i}(t) + \frac{1}{\eta} \ln\left(\frac{1}{p_0^{(i)}}\right)$$
\((c, \eta)\)-realisability (1)

- if the game is not \(\eta\)-mixable, the convex hull \(\mathcal{H}(\mathcal{B}_\eta(\Sigma_\Gamma))\) is not a subset of \(\mathcal{B}_\eta(\Sigma_\Gamma)\) and \(\mathcal{B}_\eta^{-1}(\mathcal{H}(\mathcal{B}_\eta(\Sigma_\Gamma))) \not\subset \Sigma_\Gamma\)
- the set of superpredictions is defined in such a way that for \(c \geq 1\) we have \(\Sigma_\eta \subseteq \frac{1}{c} \Sigma_\eta\), i.e., \(\frac{1}{c} \Sigma_\eta\) is ‘larger’
- so it may still be possible that \(\mathcal{B}_\eta^{-1}(\mathcal{H}(\mathcal{B}_\eta(\Sigma_\Gamma))) \subseteq \frac{1}{c} \Sigma_\Gamma\) — in this case we say that the aggregating algorithm is \((c, \eta)\)-realisable
(c, η)-realisability (2)

- if the aggregating algorithm is (c, η)-realisable, we can find γ such that

\[ \lambda(\gamma, \omega) \leq -\frac{c}{\eta} \ln \sum_{i=1}^{N} p_0^{(i)} e^{-\eta \lambda(\gamma^{(i)}, \omega)} \]

for all \( \omega \in \Omega \)

- instead of the capital inequality, we will be able to achieve

\[ \left( \frac{W_t}{W_{t-1}} \right)^{1/c} \geq \frac{\tilde{W}_t}{\tilde{W}_{t-1}} \]

— let us call this the generalised capital inequality

- we get therefore \( (W_t)^{1/c} \geq p_0^{(i)} W_t^{(i)} \), i.e.,

\[ \text{Loss}(t) \leq c \text{Loss}_i(t) + \frac{c}{\eta} \ln \left( \frac{1}{p_0^{(i)}} \right) \]
Optimality of the Aggregating Algorithm

- let us fix the uniform initial distribution $\rho_0^{(i)} = 1/N$
- if the aggregating algorithm is $(c, \eta)$-realisable, than for all $t$, all outcomes and all experts’ predictions it achieves

$$\text{Loss}(t) \leq c \text{Loss}_{E_i}(t) + \frac{c}{\eta} \ln N$$

for each expert $E_1, E_2, \ldots, E_N$

- suppose that some merging strategy for all $N$, all $t$, all outcomes and all experts’ predictions achieves

$$\text{Loss}(t) \leq a \text{Loss}_{E_i}(t) + b \ln N$$

for each expert $E_1, E_2, \ldots, E_N$ and some constants $a$ and $b$ — then the aggregating algorithm can do the same
1. Laissez-Faire Investment

2. Aggregating Algorithm

3. Defensive Forecasting
Capital Inequality Revisited (1)

- the generalised capital inequality

\[
\left( \frac{W_t}{W_{t-1}} \right)^{1/c} \geq \frac{\tilde{W}_t}{\tilde{W}_{t-1}}
\]

can be rewritten as

\[
\frac{\tilde{W}_t}{W_t^{1/c}} \leq \frac{\tilde{W}_{t-1}}{W_{t-1}^{1/c}}
\]
Capital Inequality Revisited (2)

- we have

\[ \frac{\widetilde{W}_t}{W_t^{1/c}} = \sum_{i=1}^{N} p_0^{(i)} \frac{W_t^{(i)}}{W_t^{1/c}} = \sum_{i=1}^{N} p_0^{(i)} e^{\eta \left( \frac{\text{Loss}(t)}{c} - \text{Loss}_i(t) \right)} \]

- if \( Q_t^{(i)} = e^{\eta \left( \frac{\text{Loss}(t)}{c} - \text{Loss}_i(t) \right)} \), the generalised capital inequality is equivalent to

\[ \sum_{i=1}^{N} p_0^{(i)} Q_t^{(i)} \leq \sum_{i=1}^{N} p_0^{(i)} Q_{t-1}^{(i)} \]

- this has important consequences...
Supermartingales

- let $\mathcal{P}(\Omega)$ be the set of probability distributions on $\Omega$ — from now on we consider finite spaces $\Omega$, so $\mathcal{P}(\Omega)$ is the $(|\Omega| - 1)$-simplex in $\mathbb{R}^{|\Omega|}$
- let $E$ be some set of parameters
- the function $S : (E \times \mathcal{P}(\Omega) \times \Omega)^* \rightarrow \mathbb{R}$ is a supermartingale if

$$\sum_{\omega \in \Omega} \pi_t(\omega)S(e_1, \pi_1, \omega_1, \ldots, e_{t-1}, \pi_{t-1}, \omega_{t-1}, e_t, \pi_t, \omega) \leq S(e_1, \pi_1, \omega_1, \ldots, e_{t-1}, \pi_{t-1}, \omega_{t-1})$$

for all $e_1, \pi_1, \omega_1, \ldots, e_{t-1}, \pi_{t-1}, \omega_{t-1}, e_t, \pi_t$
— the left-hand side is the expectation of $S$ given $e_1, \pi_1, \omega_1, \ldots, e_{t-1}, \pi_{t-1}, \omega_{t-1}, e_t, \pi_t$
Levin’s Lemma

- let $S$ be a supermartingale and let $S$ be forecast-continuous
  — i.e., $S(e_1, \pi_1, \omega_1, \ldots, e_{t-1}, \pi_{t-1}, \omega_{t-1}, e_t, \pi, \omega_t)$ is continuous over $\pi$ for all values of other parameters including $t$
- then for all $e_1, \pi_1, \omega_1, \ldots, e_{t-1}, \pi_{t-1}, \omega_{t-1}, e_t$ there is $\pi$ such that

$$S(e_1, \pi_1, \omega_1, \ldots, e_{t-1}, \pi_{t-1}, \omega_{t-1}, e_t, \pi, \omega) \leq S(e_1, \pi_1, \omega_1, \ldots, e_{t-1}, \pi_{t-1}, \omega_{t-1})$$

for all $\omega$
- that is, we can always choose $\pi \in \mathcal{P}(\Omega)$ such that $S$ does not grow no matter what $\omega$ the nature chooses
Applying Levin’s Lemma (1)

- Levin’s lemma guarantees the existence of a distribution — and we need a prediction
- if two games $\langle \Omega, \Gamma_1, \lambda_1 \rangle$ and $\langle \Omega, \Gamma_2, \lambda_2 \rangle$ specify the same set of superpredictions $\Sigma$, they are essentially equivalent — different $\Gamma$’s and $\lambda$’s can be thought of as different parametrisations of $\Sigma$
- for many games with finite $\Omega$ we can construct an equivalent parametrisation with the prediction set $\mathcal{P}(\Omega)$ — if $\Omega = \{0, 1\}$, the set $\mathcal{P}(\Omega)$ can be identified with $[0, 1]$ (cf. our definition of a binary game)
- consider $Q(i) = e^{\eta \left( \frac{\text{Loss}(t)}{c} - \text{Loss} \epsilon_i(t) \right)}$ and $Q = \sum_{i=1}^{N} p_0^{(i)} Q^{(i)}$
  — the array of experts’ predictions $\pi^{(1)}_t, \pi^{(2)}_t, \ldots, \pi^{(N)}_t$ makes $e_t$
Applying Levin’s Lemma (2)

- If $Q$ is a forecast-continuous supermartingale for some $c$ and $\eta$, on each step we can find $\pi_t$ ensuring that $Q$ does not grow — i.e., the generalised wealth equation is satisfied and we achieve
  \[
  \text{Loss}(t) \leq c \text{Loss}_i(t) + \frac{c}{\eta} \ln \left( \frac{1}{p_0^{(i)}} \right)
  \]
  — we shall call this method defensive forecasting

- If $Q$ is a forecast-continuous supermartingale for some $c$ and $\eta$, the aggregating algorithm must be $(c, \eta)$-realisable and output the same predictions
  — the inverse statement is a bit trickier...
Proper Loss Function

- a function $\lambda : \mathcal{P}(\Omega) \times \Omega \rightarrow [0, +\infty]$ is proper if for all $\pi, \pi' \in \mathcal{P}(\Omega)$ we have

$$E_{\omega \sim \pi} \lambda(\pi, \omega) \leq E_{\omega \sim \pi} \lambda(\pi', \omega)$$

- the expectation of $\lambda(\pi', \omega)$ when $\omega$ is distributed according to $\pi$ is the smallest when $\pi' = \pi$

- a proper and continuous loss function exists given some mild technical restrictions on the game — the square and logarithmic loss functions are already proper
Geometric Interpretation

\[(1 - p)x + py = H(p)\]

- A distribution \( \pi \) is a family of parallel hyperplanes.
- For a proper loss function \( \lambda \) this family touches \( S \) at \( (\lambda(\omega_0, \pi), \lambda(\omega_1, \pi), \ldots, \lambda(\omega_{K-1}, \pi)) \).
The Inverse Statement

- if AA is $(c, \eta)$-realisable and there is a proper continuous loss function, then $Q$ is a supermartingale
- and defensive forecasting achieves the same loss bound as the aggregating algorithm
Bibliography

- N. Cesa-Bianchi and G. Lugosi, Prediction, learning, and games, CUP, 2006
  — the monograph on prediction with expert advice
  — AA and its optimality
  — connections between the AA and defensive forecasting