On-line learning of individual sequences

Yoav Freund, AT&T Research

A review talk, some results from collaborations with:
Peter Auer
Nicoló Cesa-Bianchi
David Haussler
David Helmbold
Rob Schapire
Manfred Warmuth

Papers and transparencies in:
http://www.research.att.com/orgs/ssr/people/yoav

A simple prediction problem

- On-line prediction of a binary sequence.
- N experts provide predictions of the sequence.
- *Assumption:* one of the experts makes no mistakes.
- *Goal:* Predict to make the minimal number of mistakes.

Example:

$$Sequence: \; 0, \; 1, \; 1, \; 0, \; \ldots$$

$$\begin{array}{ll} expert_1: & 1, \; 1, \; 0, \; 1, \; \ldots \\ expert_2: & 0, \; 1, \; 1, \; 1, \; \ldots \\ expert_3: & 0, \; 1, \; 1, \; 0, \; \ldots \\ expert_4: & 0, \; 1, \; 1, \; 1, \; \ldots \end{array}$$

Solution of simple problem

- Predict according to the majority of experts that have made no mistake.
- At each mistake the pool of active experts is *halved*.

$\Rightarrow$ At most $\log_2 N$ mistakes can be made.

Analysis characteristics:

- *Online* (iterative) setup.
- *No* probabilistic assumptions.
- Bounds hold for *worst-case* sequences of experts, predictions and outcomes.

Can we generalize to when no expert is perfect?

Preview

1. A simple and general (but not optimal) on-line prediction algorithm.
2. An even more general (but non-constructive) meta-algorithm.
3. Detailed results for specific cases.
4. Continuous sets of experts.

Predicting when no expert is perfect

Prediction game: algorithm vs. adversary

for t=1,2,...,T

1. **Adversary:** Each expert suggest an action.
   A loss $\ell_i^t \in [0, 1]$ is associated with each action.
2. **Algorithm:** Chooses an expert i_t (can randomize).
3. All losses $\ell_1^t, \ldots, \ell_N^t$ are revealed to player.

Quantities of interest

Total loss of expert i:

$$L_i \doteq \sum_{t=1}^{T} \ell_i^t$$

Total loss of best expert:

$$L_{\min} \doteq \min_{i=1,\ldots,N} L_i$$

Expected total loss of algorithm:

$$L_A \doteq E_{i_1,\ldots,i_t}\left[\sum_{t=1}^{T} \ell_{i_t}^t\right]$$

Goal: Minimize $L_A - L_{\min}$ in the worst case.

Algorithm Hedge (Weighted Majority)

[Littlestone, Warmuth 89],[Freund, Schapire 95]

"Weight" $W_i^t > 0$ associated with expert i at time t.

One parameter: $\eta > 0$.

Initial weights: $\forall i \in \{1 \dots N\}\ \ W_i^1 = 1/N$

1. Prob. of choosing expert i on iteration t:

$$\frac{W_i^t}{\Sigma_{j=1}^N W_j^t}$$

2. Weights update after iteration t:

$$W_i^{t+1} = W_i^t \, e^{-\eta \ell_i^t}$$

Theorem regarding the performance of Hedge

For *any* sequence of predictions and outcomes of length T

$$L_A \leq \frac{\eta L_{\min} + \ln N}{1 - e^{-\eta}}$$

Setting η to minimize bound:

- Decreasing η decreases numerator.
- If $\eta \to 0$ denominator becomes zero.

Corollary For any T

$$\text{If } \eta = \ln\left(1 + \sqrt{\frac{2 \ln N}{T}}\right)$$

then

$$L_A - L_{\min} \leq \sqrt{2T \ln N} + \ln N \ .$$

In other words:

$$\frac{L_A}{T} \leq \frac{L_{\min}}{T} + \sqrt{\frac{2 \ln N}{T}} + \frac{\ln N}{T} \ .$$

Proof sketch, Main argument

Lemma I: If expected loss is large, final total weight is small:

$$\ln \sum_i W_i^{T+1} \leq -(1 - e^{-\eta}) \sum_{t=1}^{T} E_{i_t}[\ell_{i_t}^t]$$

Lemma II: If the best expert is good, the final weight cannot be too small:

$$\sum_i W_i^{T+1} \geq \frac{1}{N} \exp(-\eta L_{\min})$$

Combining the bounds we get

$$-(1 - e^{-\eta}) \sum_{t=1}^{T} E_{i_t}[\ell_{i_t}^t] \geq \ln \left(\frac{1}{N} \exp(-\eta L_{\min}) \right)$$

$$\sum_{t=1}^{T} E_{i_t}[\ell_{i_t}^t] \leq \frac{\ln N + \eta L_{\min}}{1 - e^{-\eta}}$$

Proof of lemma II

For *any* $j \in \{1 \ldots N\}$:

$$\sum_i W_i^{T+1} \geq W_j^{T+1} = \frac{1}{N} \exp(-\eta L_j)$$

Proof of lemma I

$$\begin{aligned}\sum_i W_i^{t+1} &= \sum_i W_i^t \quad \exp(-\eta \ell_i^t) \\ &\leq \sum_i W_i^t \left(1 - (1 - e^{-\eta})\ell_i^t\right) \\ &= \sum_i W_i^t - (1 - e^{-\eta}) \sum_i W_i^t \ell_i^t\end{aligned}$$

As

$$E_{i_t}[\ell_{i_t}^t] = \sum_i \frac{W_i^t}{\sum_j W_j^t} \ell_i^t$$

We get

$$\frac{\sum_i W_i^{t+1}}{\sum_i W_i^t} \leq 1 - (1 - e^{-\eta}) E_{i_t}[\ell_{i_t}^t] \; .$$

Taking logs and Combining for $t = 1 \ldots T$:

$$\begin{aligned}\ln \frac{\sum_i W_i^{T+1}}{\sum_i W_i^1} &\leq \sum_{t=1}^{T} \ln\left(1 - (1 - e^{-\eta}) E_{i_t}[\ell_{i_t}^t]\right) \\ &\leq -(1 - e^{-\eta}) \sum_{t=1}^{T} E_{i_t}[\ell_{i_t}^t]\end{aligned}$$

But $\sum_i W_i^1 = N \frac{1}{N} = 1.$

What can be done beyond this?

- Unbounded loss per trial.
- Better bounds for special losses
  - Define loss as a function of *prediction* and *outcome*.
  - Combine predictions of experts instead of randomly choosing a single expert.
- Better bounds for special model classes
  - Models parameterized by continuous parameters.
  - Neighborhood structure of models.

Some useful loss functions

Outcomes: binary $x^1, x^2, \ldots$
Predictions: $p^1, p^2, \ldots\ p^t \in [0,1]$

- **Absolute loss (Prediction error)**

$$\ell^t = |x^t - p^t|$$

  Probability of making a mistake if predicting 0 or 1 using a biased coin
  If $P[x^t = 1] = q$, then the optimal prediction is

$$p^t = \begin{cases} 1, & \text{if } q > 1/2 \\ 0, & \text{otherwise} \end{cases} .$$

- **Log loss (Entropy loss)**

$$\ell^t = -x^t \ln p^t - (1 - x^t) \ln(1 - p^t)$$

  Cumulative log loss = coding length ± 1
  If $P[x^t = 1] = q$, optimal prediction $p^t = q$.

- **Square loss (Breier Loss)**

$$\ell^t = (x^t - p^t)^2$$

  If $P[x^t = 1] = q$, optimal prediction $p^t = q$.
  Loss is bounded.
  Example: predicting forehand/backhand in ping-pong.

Vovk's general result [Vovk 89,95]

A game, between Nature and a Learner:

For $t = 1, 2, \ldots$:

1. Each expert $i \in \{1 \ldots N\}$ makes a prediction $\gamma_i^t \in \Gamma$.
2. The learner, after observing $\langle \gamma_1^t \ldots \gamma_N^T \rangle$, makes its own prediction γ^t.
3. Nature chooses an outcome $\omega^t \in \Omega$.
4. Each expert incurs loss $\ell_i^t = \lambda(\omega^t, \gamma_i^t)$. The learner incurs loss $\ell^t = \lambda(\omega^t, \gamma^t)$.

Goal: guarantee that

$$L_A \leq a L_{\min} + c \ln N$$

for any sequence and for the smallest possible pairs $(a, c) \in [0, \infty)^2$.

The loss function should be well-behaved

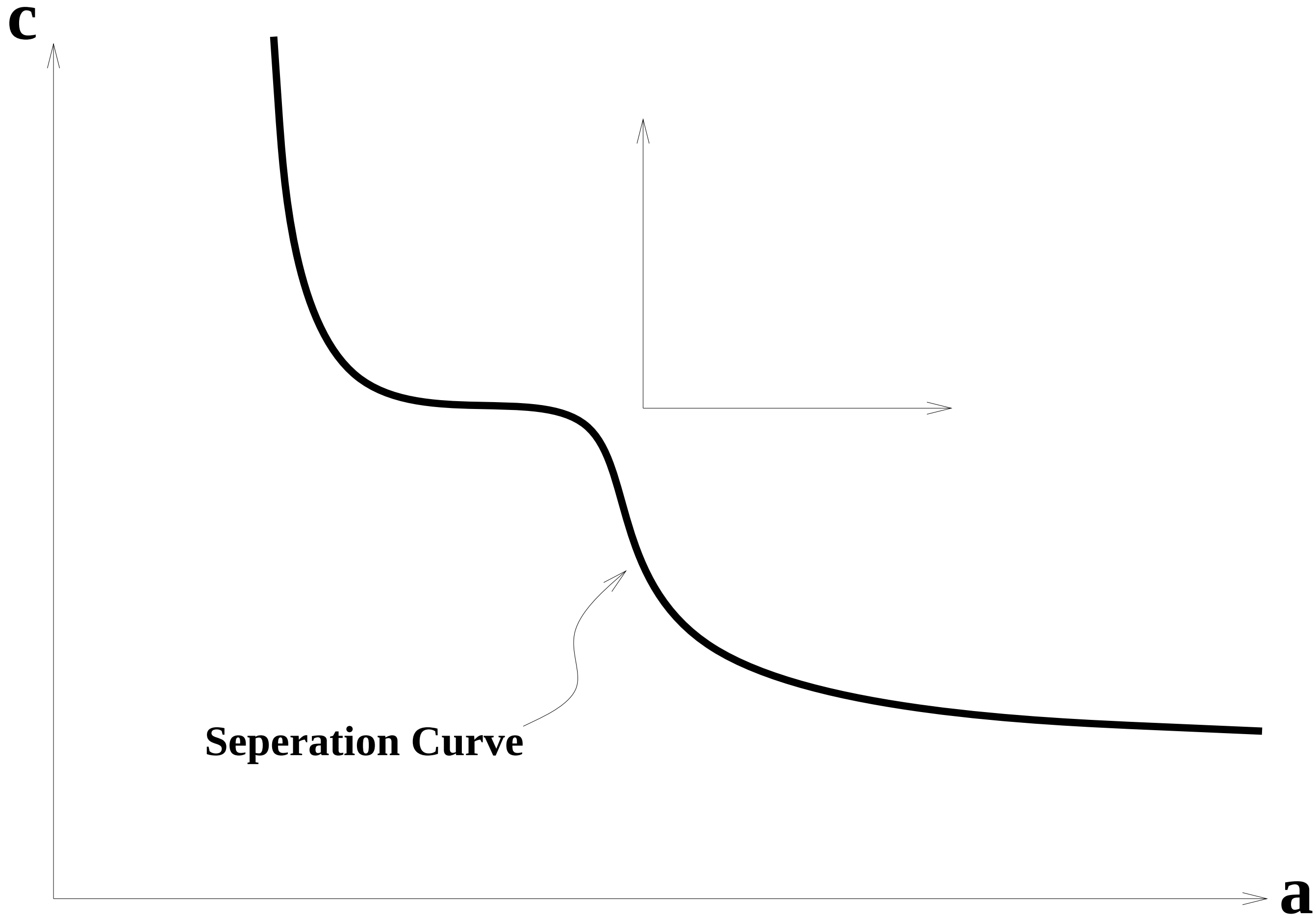
There should be a topology on the set Γ such that

- Γ is compact.
- $\forall \omega \in \Omega$, the function $\gamma \to \lambda(\omega, \gamma)$ is continuous.
- There is a universally reasonable prediction $\exists \gamma \in \Gamma$, $\forall \omega \in \Omega$, $\lambda(\omega, \gamma) < \infty$.
- There is no universally optimal prediction $\neg \exists \gamma \in \Gamma$, $\forall \omega \in \Omega$, $\lambda(\omega, \gamma) = 0$.

The set of achievable bounds

The pair (a, c) is *achievable* if there exists *some* prediction algorithm such that for *any* $N > 0$, *any* set of N prediction sequences and *any* sequence of outcomes

$$L_A \leq a L_{\min} + c \ln N$$



Vovk's meta-algorithm

Fix an achievable pair (a, c).

Set $\eta = a/c$.

1. Define
$$W_i^t = \frac{1}{N}\, e^{-\eta L_i^t}$$
2. Choose γ_t so that, for all $\omega^t \in \Omega$:
$$\lambda(\omega^t, \gamma^t) - c \ln \sum_i W_i^t \leq -c \ln \left(\sum_i W_i^t e^{-\eta\lambda(\omega^t, \gamma_i^t)} \right)$$

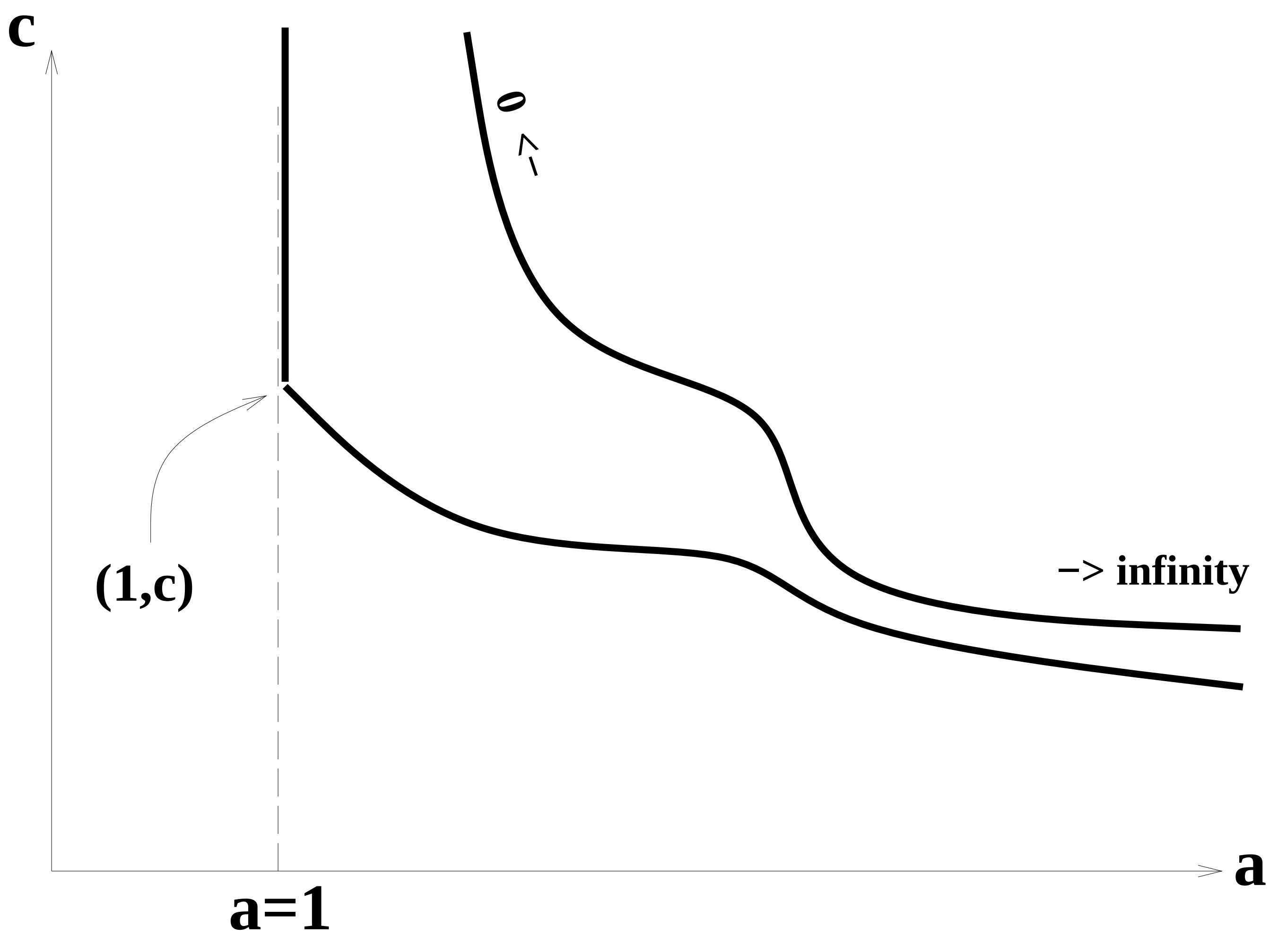
If choice of γ^t always exists, then the total loss satisfies:
$$\sum_t \lambda(\omega^t, \gamma^t) \leq -c \ln \sum_i W_i^{T+1} \leq a L_{\min} + c \ln N$$

Vovk's result: *yes!* a choice for γ_t always exists!

Vovk's algorithm is optimal

Theorem (Vovk 95)
The pair (a, c) is achievable (by some algorithm) if and only if it is achieved by Vovk's algorithm.
The separation curve is

$$\left\{ \left(a(\eta), \frac{a(\eta)}{\eta} \right) \middle| \eta \in [0, \infty] \right\}$$



Special case: $a(\eta) = 1$ for $\eta < \infty$

Vovk's condition: Choose γ so that, for all $\omega \in \Omega$:

$$\lambda(\omega, \gamma^t) - c \ln \sum_i W_i^t \leq -c \ln \left(\sum_i W_i^t e^{-\eta \lambda(\omega, \gamma_i^t)} \right)$$

Can be re-written as:

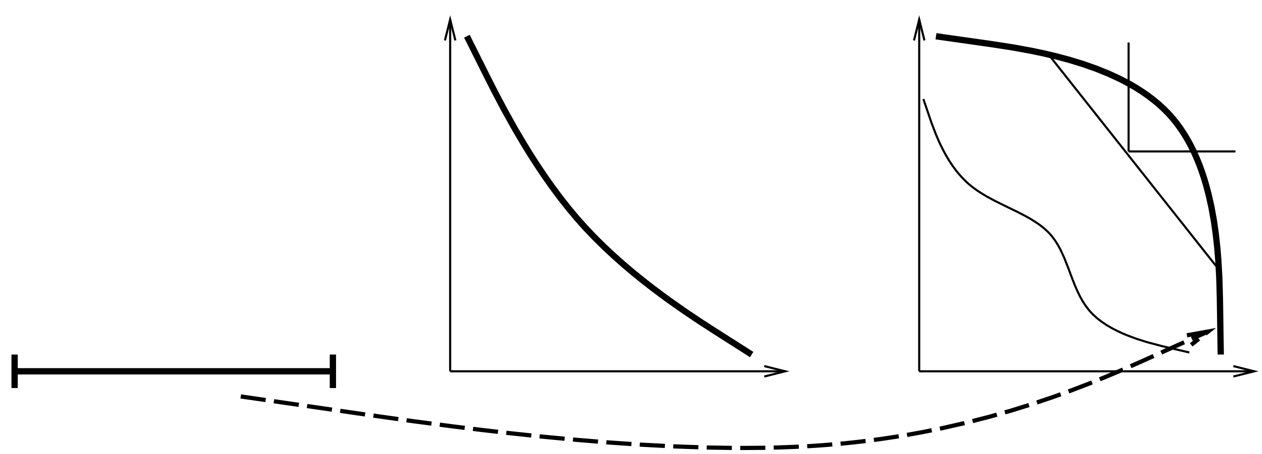
$$e^{-\frac{1}{c}\lambda(\omega,\gamma)} \geq \sum_i \left(\frac{W_i}{\sum_j W_j} \right) e^{-\eta \lambda(\omega, \gamma_i)}$$

- $a(\eta) = 1 \Leftrightarrow \frac{1}{c} = \eta$.
- **Assumption:** fix $\lambda(\omega, \gamma_i)$ for all but $i \notin \{j, k\}$ then increasing $\lambda(\omega, \gamma_j)$ decreases $\lambda(\omega, \gamma_k)$.
- We get a convexity condition on the image of Γ under the function

$$F(\gamma) = \left\langle e^{-\eta\lambda(\omega,\gamma)} \right\rangle_{\omega \in \Omega}$$

Example: Suppose $\Omega = \{0, 1\}$, $\Gamma = [0, 1]$. then

$$F(\gamma) = \left\langle e^{-\eta\lambda(0,\gamma)}, e^{-\eta\lambda(1,\gamma)} \right\rangle$$



Negating Vovk's condition implies a "bad" distribution

Vovk's condition:

$$\begin{aligned}
&\forall \ N, \\
&\forall \ \langle(\gamma_1, W_1), \ldots, (\gamma_N, W_N)\rangle, \\
&\exists \ \gamma, \ \forall\omega : \\
&\quad e^{-\frac{1}{c}\lambda(\omega,\gamma)} \geq \sum_i \left(\frac{W_i}{\Sigma_j W_j}\right) e^{-\eta\lambda(\omega,\gamma_i)}
\end{aligned}$$

Negation is

$$\begin{aligned}
&\exists \ N, \\
&\exists \ \langle(\gamma_1, W_1), \ldots, (\gamma_N, W_N)\rangle, \\
&\forall \ \gamma, \ \exists\omega : \\
&\quad e^{-\frac{1}{c}\lambda(\omega,\gamma)} < \sum_i \left(\frac{W_i}{\Sigma_j W_j}\right) e^{-\eta\lambda(\omega,\gamma_i)}
\end{aligned}$$

Defines a "BAD" distribution over Γ (prediction space).
The distribution has a finite support.

Adversarial construction for Vovk's algorithm

- **Experts**: Each expert predicts IID according to the "BAD" distribution over Γ.
- **Sequence:** Choose an outcome ω which satisfies the bad inequality *and*:
  - If $\exists \omega$ such that some expert suffers non-zero loss then use it.
  - Else choose ω that maximizes loss of prediction algorithm.

Vovk's homepage: HTTP://casbs.stanford.edu/~vovk/

Explicit analysis for special cases

Vovk's meta-algorithm is not constructive:

Given $\lambda(\omega, \gamma)$ find:

- Set of achievable pairs. $\{(a, c)\}$
  In particular, is there an achievable $(1, c)$, $c < \infty$?
- How to calculate a good prediction γ^t?
  (easy when $|\Omega| < \infty$).

Vovk's algorithm for binary outcomes, log loss

$$\ell^t = -x^t \ln p_i^t - (1-x^t)\ln(1-p_i^t)$$

- The pair $(1,1)$ is achieved by $\eta = 1$.
- Prediction = weighted average:

$$\frac{\Sigma_i W_i^t \gamma_i^t}{\Sigma_i W_i^t}$$

- The update rule is:

$$\begin{aligned} W_i^{t+1} &= W_i^t \, \exp\left(x^t \ln p_i^t + (1-x^t)\ln(1-p_i^t)\right) \\ &= W_i^t (p_i^t)^{x^t}(1-p_i^t)^{1-x^t} \end{aligned}$$

This algorithm is *identical* to the Bayes prediction algorithm with a uniform prior.

Theorem: For *any* sequence the cumulative log loss of the algorithm is larger by at most $\ln N$ from the loss of the best expert.

This result is well known in Information theory as a "Universal coding" folk-theorem.

Vovk's algorithm for Binary outcomes, square loss

$$\ell^t = (x^t - p_i^t)^2$$

- The pair $(1, 1/2)$ is achieved by $\eta = 2$.
- **Weight Update rule:**

$$W_i^{t+1} = W_i^t \exp\left(-2(x^t - p_i^t)^2\right)$$

- **Prediction rule:** any choice from the range

$$1 - \sqrt{-\frac{1}{2}\ln \sum_i V_i^t e^{-2(1-p_i^t)^2}} \leq p^t \leq \sqrt{-\frac{1}{2}\ln \sum_i V_i^t e^{-2(p_i^t)^2}}$$

where

$$V_i^t = \frac{W_i^t}{\sum_s W_i^s} .$$

This algorithm is quite different from the Bayes algorithm.

Theorem: For *any* sequence the cumulative square loss of the algorithm is larger by at most $\frac{1}{2}\ln N$ from the loss of the best expert.

Bayes algorithm is sub-optimal for square loss

Bayes is optimal only if data is generated by a model from the class.

Example:

$N = 3$

$\forall t \;\; p_1^t = 0, \; p_2^t = 1/2, \; p_3^t = 1$

Source: biased coin with $p = 0.9$.

For *any* prior: Bayes algorithm quickly converges to $p = 1/2$.

Expected total loss of Bayes:

$$\approx (1/2)^2 * T = 0.25 * T$$

Expected total loss of expert 3:

$$(0.9 * (1-1)^2 + 0.1 * (1-0)^2) * T = 0.1 * T$$

Expected total loss for Vovk's algorithm:

$$\leq 0.1 * T + \ln(3)/2.$$

A prediction algorithm for binary outcome, absolute loss

[Cesa-Bianchi et. al. 93]

$$\ell^t = |x^t - p_i^t|$$

- There is no achievable pair $(1, c)$, $c < \infty$.
- By selecting η as a function of T we can achieve

$$L_A - L_{\min} \leq \sqrt{\frac{T \ln(N+1)}{2}} + \frac{\log_2(N+1)}{2}$$

- *Lower bound:* If all predictions and outcome are chosen to be 0 or 1 with equal probability then, for *any* algorithm:

$$L_A - L_{\min} \geq (1 - o(1))\sqrt{\frac{T \ln N}{2}} \text{ when } T, N \to \infty$$

A lower bound construction for absolute loss

- **Sequence:** Random, IID, $p = \frac{1}{2}$.
- N **Experts:** Random, IID, $p = \frac{1}{2}$.

Expected loss of *any* algorithm $= \frac{T}{2}$.

Expected loss of **best** expert $= \frac{T}{2} - (1 - o(1))\sqrt{\frac{T \ln N}{2}}$.
when $T, N \to \infty$.

- $T \to \infty :\ L'_i = \frac{L_i - T/2}{\sqrt{T}}$ converges to normal.
- $N \to \infty :\ \sqrt{2 \ln N} \max(L'_1, \ldots, L'_N) + 2 \ln N$ converges to a limit distribution with finite expected value.

Planting the lower bound inside a less trivial case

Summary of part I, N experts

- For any bounded loss "Hedge" achieves

$$L_A - L_{\min} = O(\sqrt{T \ln N})$$

- For any continuous loss Vovk's algorithm achieves the best bounds of the form

$$L_A \leq a L_{\min} + c \ln N$$

- For log loss and square loss Vovk achieves

$$L_A - L_{\min} = O(\ln N)$$

- For absolute loss, any algorithm satisfies

$$L_A - L_{\min} = \Omega(\sqrt{T \ln N})$$

A continuous class of models

- Each expert corresponds to a biased coin, predicts with some fixed $\theta \in [0, 1]$.
- All values of θ allowed.
- Uncountably infinite set of experts.
- Bound of the form

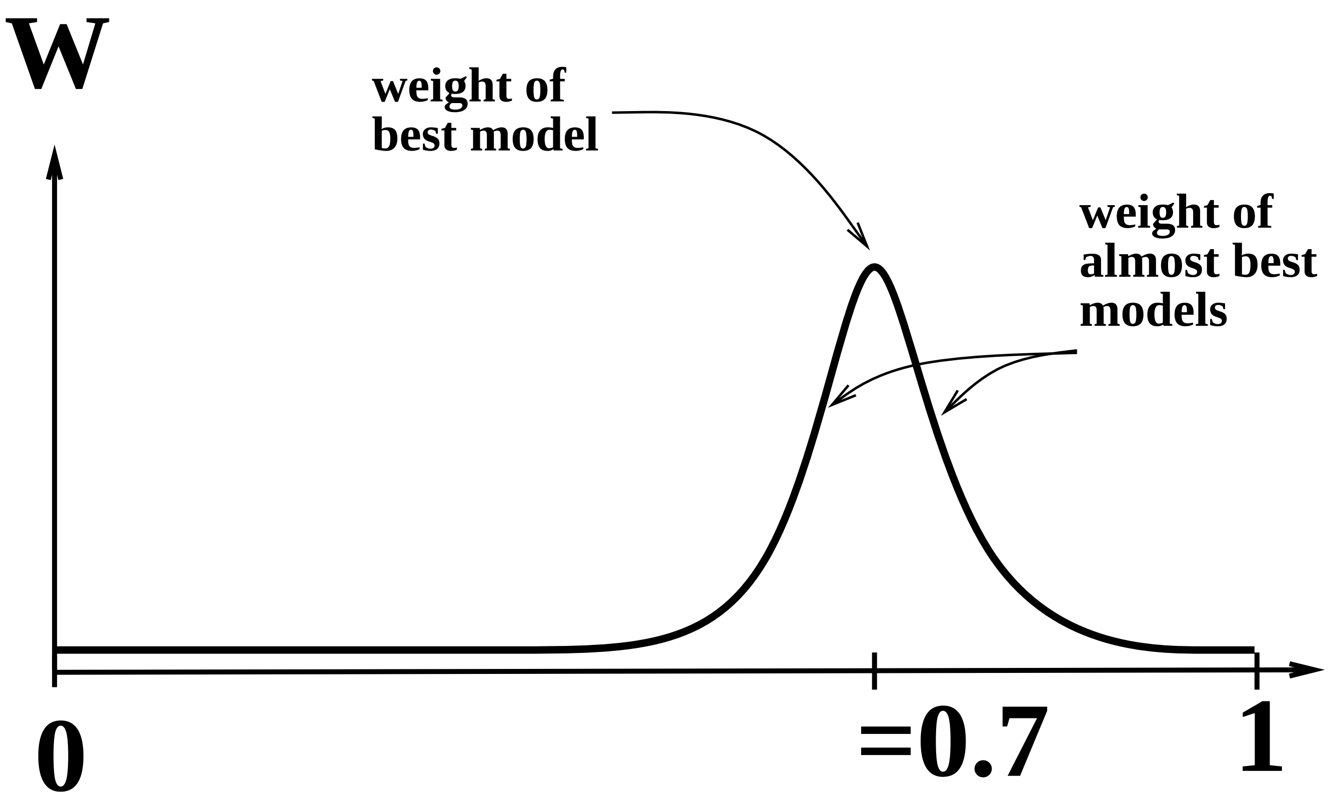
$$L_A \leq aL_{\min} + c \ln N$$

  is meaningless.
- Can we still get a meaningful bound?

- Replace the initial weight by a density measure $w(\theta) = w^1(\theta)$, $\int_0^1 w(\theta)d\theta = 1$.
- **Upper bound** on final total weight holds translates directly:

$$L_A \leq -c \ln \int_0^1 w(\theta) e^{-\eta L_\theta^{T+1}} d\theta$$

- We need a new **lower bound** on the final total weight
- **Idea:** If $w^t(\theta)$ is large then $w^t(p+\epsilon)$ is also large.



Rewriting the integral in exponential form

- For log loss and square loss best θ is empirical distribution of seq.

$$\hat{\theta} = \frac{\#\{x^t = 1;\ \ 1 \le t \le T\}}{T}$$

- The total loss scales with T:

$$L_\theta = T \cdot (\hat{\theta}\ell(\theta,1) + (1-\hat{\theta})\ell(\theta,0)) \doteq T \cdot g(\hat{\theta},\theta)$$

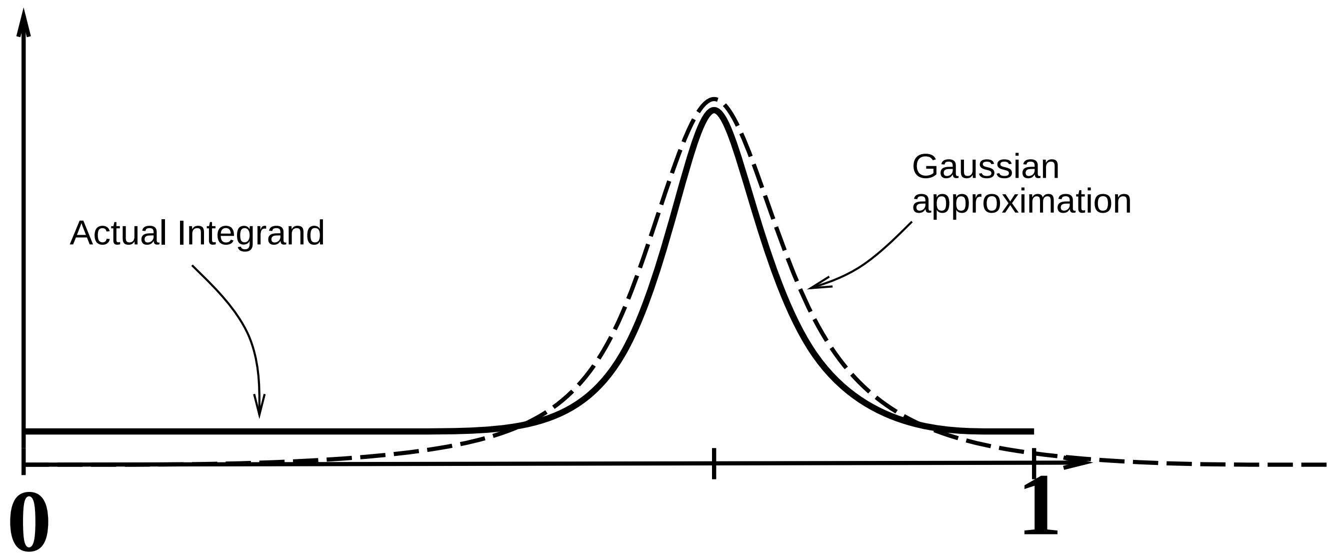
- If $a = 1$ then $\eta = 1/c$ and we can put everything in the exponent:

$$\begin{aligned} L_A - L_{\min} &\le -c\ln\int_0^1 w(\theta)e^{-\eta L_\theta}d\theta - c\ln e^{(1/c)L_{\min}} \\ &= -c\ln\int_0^1 w(\theta)e^{-\eta(L_\theta - L_{\min})}d\theta \\ &= -c\ln\int_0^1 w(\theta)e^{-\eta T(g(\hat{\theta},\theta)-g(\hat{\theta},\hat{\theta}))}d\theta \end{aligned}$$

Approximating the area under the peak

Expanding the exponent around $\theta = \hat{\theta}$:

- First and second term in the expansion of $g(\hat{\theta}, \theta) - g(\hat{\theta}, \hat{\theta})$ around $\hat{\theta}$ are zero.
- Third term gives a quadratic expression in the exponent
  $\Rightarrow$ a gaussian.



$$\int_0^1 w(\theta) e^{-\eta T (g(\hat{\theta},\theta) - g(\hat{\theta},\hat{\theta}))} d\theta$$

$$= w(\hat{\theta}) \sqrt{\frac{-2\pi c}{T \left.\frac{d^2}{d\theta^2}\right|_{\theta=\hat{\theta}} (g(\hat{\theta},\theta) - g(\hat{\theta},\hat{\theta}))}} + O(T^{-3/2})$$

Choosing the optimal prior

- Would like to choose $w(\theta)$ to maximize

$$\min_{\hat\theta} w(\hat\theta)\sqrt{\frac{-2\pi c}{T\left.\frac{d^2}{d\theta^2}\right|_{\theta=\hat\theta}\left(g(\hat\theta,\theta)-g(\hat\theta,\hat\theta)\right)}}$$

- Make bound equal for all $\hat\theta\in[0,1]$ by choosing

$$w^*(\hat\theta)=\frac{1}{Z}\sqrt{\frac{\left.\frac{d^2}{d\theta^2}\right|_{\theta=\hat\theta}\left(g(\hat\theta,\theta)-g(\hat\theta,\hat\theta)\right)}{-2\pi c}}\ ,$$

where Z is the normalization factor:

$$Z=\sqrt{\frac{1}{2\pi c}}\int_0^1\sqrt{\left.\frac{d^2}{d\theta^2}\right|_{\theta=\hat\theta}\left(g(\hat\theta,\hat\theta)-g(\hat\theta,\theta)\right)}\ d\hat\theta$$

- The bound becomes:

$$\begin{aligned}
L_A-L_{\min} &\leq -c\ln\int_0^1 w^*(\theta)e^{-\eta T(g(\hat\theta,\theta)-g(\hat\theta,\hat\theta))}d\theta \\
&= -c\ln\left(\sqrt{\frac{2\pi Z}{T}}+O(T^{-3/2})\right) \\
&= \frac{c}{2}\ln\frac{T}{2\pi}-\frac{c}{2}\ln Z+O(1/T)\ .
\end{aligned}$$

Biased coins with log loss

- The exponent in the integral is

$$g(\hat{\theta},\theta)-g(\hat{\theta},\hat{\theta}) = \hat{\theta}\ln\frac{\hat{\theta}}{\theta}+(1-\hat{\theta})\ln\frac{1-\hat{\theta}}{1-\theta} = D_{KL}(\hat{\theta}||\theta)$$

- The second derivative

$$\left.\frac{d^2}{d\theta^2}\right|_{\theta=\hat{\theta}} D_{KL}(\hat{\theta}||\theta)$$

Is called the *empirical Fisher information*

- The optimal prior:

$$w^*(\hat{\theta}) = \frac{1}{Z}\sqrt{\frac{T\left.\frac{d^2}{d\theta^2}\right|_{\theta=\hat{\theta}}(g(\hat{\theta},\theta)-g(\hat{\theta},\hat{\theta}))}{-2\pi c}}\ ,$$

is known as *Jeffrey's prior* which, for this class, is the *Dirichlet*-$(1/2, 1/2)$ *prior*.

Bounds for biased coins, log loss

- The bound, for Bayes algorithm using the Dirichlet-(1/2, 1/2) distribution, for $\hat{\theta} \in [\epsilon, 1-\epsilon]$, is:

$$L_A - L_{\min} \leq \frac{1}{2}\ln(T+1) + \frac{1}{2}\ln\frac{\pi}{2} + O(1/T)$$

- Bound asymptotically equal to the min/max bound when T known in advance.

- The bound when seq. **generated by a biased coin**, for same algorithm [Xie and Barron, 95] is:

$$\max_{\theta} E(L_A - L_{\theta}) \leq \frac{1}{2}\ln(T) + \frac{1}{2}\ln\frac{\pi}{2e} + O(1/T)$$

- The prediction is very simple:

$$p_t = \frac{\#\{x^t = 1;\ \ 1 \leq t \leq T\} + 1/2}{T+1}$$

Biased coins with square loss

- The exponent in the integral is

$$g(\hat{\theta}, \theta) - g(\hat{\theta}, \hat{\theta}) = (\theta - \hat{\theta})^2$$

- The second derivative is a constant:

$$\left.\frac{d^2}{d\theta^2}\right|_{\theta=\hat{\theta}} (\theta - \hat{\theta})^2 = 2$$

- The optimal prior is uniform over $[0, 1]$.
- The bound for $\hat{\theta} \in [\epsilon, 1-\epsilon]$, is:

$$L_A - L_{\min} \leq \frac{1}{4} \ln T - \frac{1}{4} \ln \frac{\pi}{2} + O(1/T)$$

- **Prediction:** complicated to write, but $O(\log T)$ to compute.

Summary

N models:

1. A very simple algorithm to predict almost as well as the best expert.
2. No probabilistic assumptions about the outcomes or the experts.
3. Algorithm is identical to Bayes for log-loss.
4. Algorithm is new and better than Bayes for square loss and absolute loss.

Biased coins:

1. Bayes with Jeffrey's prior is also optimal in the worst case for log loss.
2. Uniform prior is best in the worst case for square loss.

Relation to on-line competitive analysis

Positive Aspects:

- Competitive ratio $\rightarrow 1$ as $T \rightarrow \infty$.
- Good bounds on the competitive *difference*.

Negative aspects:

- Off-line algorithm restricted to choose one strategy from a (finite) class.
- Loss on each iteration must depend only on action taken on same iteration.
  Example: Servicing Page Faults. Loss depends on previous actions.
  Dependence can be weakened by making each iteration correspond to a large number of page faults.

Further work

Ordered by my familiarity with the subject

1. Uncountably infinite classes of experts (parameterized experts).
2. Observing only the loss of the selected action (multiple arm bandit problem).
3. Learning to play repeated matrix games.
4. Relation to boosting.
5. Efficient calculation of exponentially many experts.
6. Relations with universal coding and universal portfolios. (Gallager, Risannen, Feder, Merhav, Cover).
7. "Specialists" - Allowing experts to abstain from predicting.
8. Allowing the identity of the best expert to change from time to time (non-stationarity).
9. Competing against the best *linear combination* of experts.
10. Relations with calibration methods (Foster, Vohra).

More efficient versions

- Generally, $\Omega(N)$ computation time per iteration.
- Some expert classes can be done in $O(\log N)$ time per iteration.

1. [Littlestone ??]: "Winnow": Expert = a disjunction over k out of n elements (Time= $O(n)$ instead of $O(n^k)$).
2. [Warmuth, Maass ??]: Expert = an indicator function of an axis-parallel box in $[1 \ldots m]^d$.
3. [Willems Starkov ??],[Helmbold, Schapire ??]: Expert = a pruning of a fixed decision tree.

Restricted feedback

- $\ell_i^t \in [-1, 0]$= loss of action i at time t.
- Learner chooses action i_t.
- Learner observes only the loss $\ell_{i_t}^t$.
- Exploration vs. exploitation achieved by choosing i_t with probability

$$p_i^t = \mu \frac{W_i^t}{\Sigma_{j=1}^K W_j^t} + (1 - \mu)$$

- Only weight of selected action is updated

$$W_{i_t}^{t+1} = W_{i_t}^t \ \exp(-\eta \ell_{i_t}^t / p_{i_t}^t)$$

- Upper bound:

$$E[L_A - L_{\min}] = O(\sqrt{TK \ln K})$$

- Lower bound: for any algorithm

$$E[L_A - L_{\min}] = \Omega(\sqrt{TK})$$