Sequential Prediction

Andrew Nobel

October, 2023

Overview

Idea: Game of prediction that evolves in an ordered sequence of rounds, denoted by $t \geq 1$, with outcomes y_t

Agents and actions: At each round $t \geq 1$

- Panel of experts offer their predictions of next outcome y_t
- Forecaster predicts y_t using expert advice
- Environment generates outcome y_t
- Forecaster and experts incur some loss based on their predictions

Goal: Strategies enabling Forecaster to perform nearly as well as best expert

Setting of Prediction with Expert Advice

Basic components: A triple

- Outcome space $\mathcal{Y}$
- Decision space $\mathcal{V}$, a convex subset of a real vector space
- Loss function $\ell : \mathcal{V} \times \mathcal{Y} \to \mathbb{R}$ such that $\ell(\cdot, y)$ is convex for all $y \in \mathcal{Y}$

Predicting outcome $y \in \mathcal{Y}$ by element $v \in \mathcal{V}$ incurs loss $\ell(v, y)$

Idea: Prediction proceeds in a sequence of rounds. At round t goal is to predict outcome $y_t \in \mathcal{Y}$ based on expert advice and previous outcomes

Examples

1. Squared loss: $\mathcal{V} = \mathcal{Y} = \mathbb{R}$, loss $\ell(v, y) = (v - y)^2$

2. Absolute loss: $\mathcal{V} = \mathcal{Y} = \mathbb{R}$, loss $\ell(v, y) = |v - y|$

3. Relative entropy loss: $\mathcal{V} = \mathcal{Y} = [0, 1]$, loss

$$\ell(v, y) = y \log \frac{y}{v} + (1 - y) \log \frac{1 - y}{1 - v}$$

4. Log loss: $\mathcal{V} = [0, 1]$ and $\mathcal{Y} = \{0, 1\}$

$$\ell(v, y) = \mathbb{I}(y = 1) \log \frac{1}{v} + \mathbb{I}(y = 0) \log \frac{1}{1 - v}$$

Experts

Informal: An expert f_k is an unspecified entity generating, at each round t, a prediction $f_{k,t} \in \mathcal{V}$ to which a forecaster has access

Definition: An expert f_k is *static* if its predictions $\{f_{k,t} : t \geq 1\} \subseteq \mathcal{V}$ are fixed in advance and only depend on the round t

Idea: To handle general experts we establish regret bounds that hold for all possible sequences of expert advice, i.e., all static experts

Definition: An expert panel is a collection $\mathcal{F} = \{f_1, \ldots, f_N\}$ of experts f_k, which are not necessarily static

Forecasting Strategy

Definition: A forecasting strategy F leveraging a panel $\mathcal{F} = \{f_1, \ldots, f_N\}$ of N experts is a sequence of functions $F_1, F_2, \ldots$ where

$$F_t : \mathcal{Y}^{t-1} \times (\mathcal{V}^N)^t \to \mathcal{V}$$

Idea: Forecast F_t of strategy F at round t depends on

- Previous outcomes $y_1, \ldots, y_{t-1}$
- Previous and current predictions of experts $\{(f_{1,s}, \ldots, f_{N,s}) : 1 \leq s \leq t\}$

Prediction with Expert Advice

Given: Panel of experts $\mathcal{F} = \{f_1, \ldots, f_N\}$ and forecasting strategy F

At each round $t = 1, 2, \ldots$

- Each expert f_k makes a prediction $f_{k,t} \in \mathcal{V}$ of the next outcome
- Forecaster makes prediction $F_t \in \mathcal{V}$ of next outcome based on previous outcomes and expert advice
- Environment generates next outcome $y_t \in \mathcal{Y}$
- Forecaster incurs loss $\ell(F_t, y_t)$, expert f_k incurs loss $\ell(f_{k,t}, y_t)$

Regret

Given: Strategy F leveraging an expert panel $\mathcal{F} = \{f_1, \ldots, f_N\}$

Cumulative loss: For $n \geq 1$ define

$$L_n \;=\; \sum_{t=1}^{n} \ell(F_t, y_t) \quad \text{and} \quad L_{k,n} \;=\; \sum_{t=1}^{n} \ell(f_{k,t}, y_t)$$

Definition: The regret of strategy F at round n is given by

$$R_n \;=\; L_n - \min_{k \in [N]} L_{k,n}$$

Thus $R_n =$ the difference between the cumulative loss of F and that of the best expert at round n

Exponential Weighted Average Forecaster (EWAF)

Exponential Weighted Average Forecaster (η-EWAF)

Initialize: Fix $\eta > 0$. Assign weight $w_0(k) = 1$ to each expert $f_k \in \mathcal{F}$

Iterate: At each round $t \geq 1$

1. Forecaster's prediction is the average of the expert predictions $f_{k,t}$ under the normalized weight distribution

$$F_t \;=\; \frac{\sum_{k=1}^{N} w_{t-1}(k) f_{k,t}}{\sum_{k=1}^{N} w_{t-1}(k)}$$

2. When y_t revealed, the weight of each expert is reduced exponentially by its loss on that round

$$w_t(k) \;=\; w_{t-1}(k) \exp\{-\eta \ell(f_{k,t}, y_t)\} \;=\; \exp\{-\eta L_{k,t}\}$$

Regret Bound for EWAF: Bounded Convex Loss

Assume that the loss function $\ell : \mathcal{V} \times \mathcal{Y} \to \mathbb{R}$ satisfies

1. $\ell(\cdot, y)$ is convex for each $y \in \mathcal{Y}$
2. $\ell(v, y) \in [0, 1]$ for each $v \in \mathcal{V}$ and $y \in \mathcal{Y}$.

Theorem: Fix $\eta > 0$ and let F be the η-EWAF. Then for all $n \geq 1$, all panels $\mathcal{F}$ of N experts, and all outcome sequences $y_1^n \in \mathcal{Y}^n$

$$R_n \;\leq\; \frac{\log N}{\eta} + \frac{n\eta}{8}$$

Choosing $\eta = \sqrt{(8 \log N)/n}$ gives fixed horizon regret bound

$$R_n \;\leq\; \sqrt{\frac{n \log N}{2}}$$

Regret Bound for EWAS, cont.

Bound above requires knowing horizon n in advance in order to select η

We can avoid this using "doubling trick": divide $1, 2, \ldots$ into epochs

$$E_k \;=\; \left\{ \sum_{l=0}^{k-1} 2^l + 1, \ldots, \sum_{l=0}^{k-1} 2^l + 2^k \right\}$$

Within each epoch E_k

- Reset all weights $w_t(k) = 1$, and use EWAF with $\eta_k = \sqrt{8 \log N / 2^k}$

Upshot: Using the doubling trick, for all $n \geq 1$

$$R_n \;\leq\; \frac{\sqrt{2}}{\sqrt{2}-1} \sqrt{\frac{n}{2} \log N}$$

Exp-Concave Loss

Definition: A loss function $\ell : \mathcal{V} \times \mathcal{Y} \to \mathbb{R}$ is exp-concave for $\eta > 0$ if $G_\eta(v) := \exp\{-\eta \ell(v, y)\}$ is concave for all $y \in \mathcal{Y}$

Fact: If ℓ is exp-concave for some $\eta > 0$ then $\ell(\cdot, y)$ is convex for each $y \in \mathcal{Y}$

Examples

1. If $\mathcal{V} = \mathcal{Y} = [0, 1]$ then squared loss is exp-concave for $\eta = 1/2$
2. Absolute loss is not exp-concave for any η
3. Relative entropy loss is exp-concave for $\eta = 1$
4. Log loss is exp-concave for $\eta = 1$

Regret Vectors and Exponential Potential

Definition: Let F be any prediction scheme leveraging a panel of N experts $\mathcal{F} = \{f_1, \ldots, f_N\}$. The regret vector for F at round $t \geq 1$ is given by

$$U_t \;=\; (L_t - L_{1,t}, \ldots, L_t - L_{N,t})$$

Definition: The exponential potential function $\Phi_\eta : \mathbb{R}^N \to \mathbb{R}$ is given by

$$\Phi_\eta(u) \;=\; \frac{1}{\eta} \log \left(\sum_{k=1}^{N} e^{\eta u_k} \right)$$

Regret Bound for EWAF Under Exp-Concave Loss

Theorem: Assume that ℓ is exp-concave for $\eta > 0$, and let F be η-EWAF.

1. For each $n \geq 1$, every panel $\mathcal{F}$ of N experts, and each sequence of outcomes $y_1^n \in \mathcal{Y}^n$, the regret vector of F satisfies $\Phi_\eta(U_n) \leq \Phi_\eta(\mathbf{0})$

2. η-EWAF satisfies the risk bound

$$R_n \leq \frac{\log N}{\eta}$$

Note:

- Choose largest η such that ℓ is exp-concave for η
- Bound holds for some unbounded losses (relative entropy, log)

Minimax Regret

Assume that components $\mathcal{V}, \mathcal{Y}, \ell$ of prediction problem are fixed

Definition

1. Let $R_n(F, \mathcal{F}, y_1^n) =$ round n regret of a forecasting strategy F leveraging a panel of experts $\mathcal{F}$ on the outcome sequence y_1^n.

2. Minimax regret at round n for any strategy leveraging N experts is

$$V_n^N = \inf_F \sup_{\mathcal{F}:|\mathcal{F}|=N} \sup_{y_1^n \in \mathcal{Y}^n} R_n(F, \mathcal{F}, y_1^n)$$

   where the inf is over all forecasting strategies, and the first sup is over all panels of N static experts

Lower Bound for Absolute Loss

Fact: Let $Z_1, Z_2, \ldots$ be iid $\mathcal{N}(0,1)$, and let $(U_{k,t})_{k,t\geq 1}$ be an array of iid bounded random variables with mean 0 and variance 1. Then for $N \geq 1$

$$\lim_{n\to\infty} \mathbb{E}\left[\max_{1\leq k\leq N} \frac{1}{\sqrt{n}} \sum_{t=1}^{n} U_{k,t}\right] = \mathbb{E}\left[\max_{1\leq k\leq N} Z_k\right]$$

Moreover, as N tends to infinity,

$$\lim_{N\to\infty} \frac{\mathbb{E}\left[\max_{1\leq k\leq N} Z_k\right]}{\sqrt{2\log N}} = 1$$

Lower Bound for Absolute Loss

Proposition: If $\mathcal{V} = [0, 1]$, $\mathcal{Y} = \{0, 1\}$, and $\ell(v, y) = |v - y|$ then

$$\sup_{n \geq 1} \sup_{N \geq 1} \frac{V_n^N}{\sqrt{(n/2) \log N}} \;\geq\; 1$$

Randomized Prediction with Constant Experts

Setting for Randomized Prediction

Basic components

- Outcome space $\mathcal{Y}$
- Decision space $\mathcal{V} = \{1, 2, \ldots, N\}$
- Loss function $\ell : \{1, 2, \ldots, N\} \times \mathcal{Y} \to [0, 1]$
- Constant experts $\mathcal{F} = \{f_1, \ldots, f_N\}$ with $f_{k,t} = k$ for all $t \geq 1$

Example: Suppose $\mathcal{Y} = \mathcal{V} = \{1, 2\}$ and $\ell(k, y) = \mathbb{I}(k \neq y)$. For each $n \geq 1$ there is a sequence y_1^n such that

$$L_n = n \text{ and } \min(L_{1,n}, L_{2,n}) \leq n/2$$

Thus worst case regret $R_n \geq n/2$ is linear in n. Solution is to randomize.

Randomized Prediction

Given: Source of randomness $U_1, U_2, \ldots$ iid $\sim$ Unif$(0,1)$

Initialize: Let $y_0 \in \mathcal{Y}$ and $i_0 \in [N]$ be fixed

Iterate: At each time $t \geq 1$

1. Forecaster selects pmf p_t on $[N]$ based on past outcomes Y_0^{t-1}
2. Nature selects outcome Y_t based on past actions I_0^{t-1} of forecaster
3. Forecaster chooses action $I_t \in [N]$ using randomization U_t: for $k \in [N]$

$$I_t = k \text{ if } \sum_{j=1}^{k-1} p_t(j) \leq U_t < \sum_{j=1}^{k} p_t(j)$$

4. Forecaster incurs loss $\ell(I_t, Y_t)$

Randomized Prediction, cont.

Note: Under the setting above, for each $t \geq 1$

- Decision I_t is random and $I_t \sim p_t$
- $Y_1, \ldots, Y_t$ fully determined by $U_1, \ldots, U_{t-1}$
- $p_1, \ldots, p_t$ fully determined by $U_1, \ldots, U_{t-1}$
- $I_1, \ldots, I_{t-1}$ fully determined by $U_1, \ldots, U_{t-1}$

Unconditional and Conditional Regret

Definition: Regret of randomized forecaster at round n is

$$R_n \;=\; \sum_{t=1}^{n} \ell(I_t, Y_t) \;-\; \min_{k\in[N]} \sum_{t=n}^{n} \ell(k, Y_t)$$

Definition: Conditional regret of randomized forecaster at round n is

$$\overline{R}_n \;=\; \sum_{t=1}^{n} \overline{\ell}(p_t, Y_t) \;-\; \min_{k\in[N]} \sum_{t=n}^{n} \ell(k, Y_t)$$

where

$$\overline{\ell}(p_t, Y_t) \;:=\; \mathbb{E}\left[\ell(I_t, Y_t) \,|\, U_1^{t-1}\right] \;=\; \sum_{k=1}^{N} p_t(k)\,\ell(k, Y_t)$$

Randomized Prediction via EWAF

Idea: Apply η-EWAF with

- Outcome space $\mathcal{Y}$
- Decision space $\mathcal{V}$ = probability simplex in $\mathbb{R}^N$
- Loss function $\overline{\ell}(v, y) = \sum_{k=1}^{N} v_k \, \ell(k, y)$
- Panel $\mathcal{F} = \{f_1, \ldots, f_N\}$ of constant experts: $f_{k,t} = k$ for all $t \geq 1$

Constant experts represent "pure strategies", randomized forecaster employs a "mixed strategy"

Randomized Prediction via EWAF

Upshot: At round t Forecaster uses probability mass function

$$p_t(k) \;=\; \frac{\exp(-\eta \sum_{s=1}^{t-1} \ell(k, Y_s))}{\sum_{l=1}^{N} \exp(-\eta \sum_{s=1}^{t-1} \ell(l, Y_s))}$$

Choosing $\eta = \sqrt{8 \log N / n}$ gives bound on conditional regret

$$\overline{R}_n \;\leq\; \sqrt{\frac{n}{2} \log N}$$

Cor: For each $\delta \in (0, 1)$, with probability at least $1 - \delta$

$$R_n \;\leq\; \sqrt{\frac{n}{2} \log N} + \sqrt{\frac{n}{2} \log \frac{1}{\delta}}$$

Some Connections with Game Theory

Minimax Theorem

Thm: Let $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be continuous. Assume that

1. $\mathcal{X} \subseteq \mathbb{R}^k$ is convex and compact
2. $\mathcal{Y} \subseteq \mathbb{R}^l$ is convex and compact
3. $f(\cdot, y)$ is convex for each $y \in \mathcal{Y}$
4. $f(x, \cdot)$ is concave for each $x \in \mathcal{X}$

Then

$$\inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} f(x, y) \;=\; \sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} f(x, y)$$

Two-Player Zero Sum Games

Ingredients: Two players and loss matrix

- Player P1 with finite action space $[M] = \{1, \ldots, M\}$
- Player P2 with finite action space $[N] = \{1, \ldots, N\}$
- Loss function $\ell : [M] \times [N] \to [0, 1]$

One-round game

- P1 selects action $i \in [M]$ and P2 selects action $j \in [N]$
- P1 incurs loss $\ell(i, j)$ and P2 incurs gain $-\ell(i, j)$ (zero sum)

Two-Player Zero Sum Games, cont.

Competing goals of players

- P1: Choose action i to minimize his loss $\ell(i, j)$
- P2: Choose action j to minimize her payoff $-\ell(i, j)$

Examples: Games described by loss matrix

	1	2
1	.3	0
2	.6	.4

	1	2
1	.25	.9
2	.5	0

Mixed Strategies

Upshot: In general, playing the pure, conservative strategies

$$i^* = \operatorname*{argmin}_{i \in [M]} \max_{j \in [N]} \ell(i,j) \quad \text{and} \quad j^* = \operatorname*{argmax}_{j \in [N]} \min_{i \in [M]} \ell(i,j)$$

is not stable: one player may be incentivized to choose another action

Mixed strategies: Actions of players are random

- Mixed strategy for P1 is a pmf p on $[M]$
- Mixed strategy for P2 is a pmf q on $[N]$
- Mixed strategy profile is product pmf $p \otimes q$ on $[M] \times [N]$

Nash Equilibria

Definition: Given mixed strategies p on $[M]$ and q on $[N]$ let

$$\overline{\ell}(p,q) \;=\; \sum_{i=1}^{M}\sum_{j=1}^{N} p(i)\,q(j)\,\ell(i,j)$$

be the expected loss of P1 (gain of P2) under the profile $p \otimes q$

Definition: A profile $p \otimes q$ is a Nash equilibrium if for all p' and q'

$$\overline{\ell}(p,q') \;\leq\; \overline{\ell}(p,q) \;\leq\; \overline{\ell}(p',q)$$

Interpretation: If P1 plays strategy p and P2 plays strategy q, neither player has an incentive to change their strategy

Nash Equilibria, cont.

Definition: By minimax theorem applied to $f(p,q) = \overline{\ell}(p,q)$ we have

$$\min_{p'} \max_{q'} \overline{\ell}(p',q') \;=\; \max_{q'} \min_{p'} \overline{\ell}(p',q') \;:=\; V$$

where V is called the *value of the game*

Fact: Profile $p \otimes q$ is a Nash equilibrium iff $\overline{\ell}(p,q) = V$

Repeated Two-Player Zero Sum Games

At each round $t \geq 1$

- P1 chooses action $I_t \in [M]$ according to p_t
- P2 chooses action $J_t \in [N]$ according to q_t
- P1 incurs loss $\ell(I_t, J_t)$ and P2 incurs gain $-\ell(I_t, J_t)$

Exchange of Information

- Strategies p_t, q_t may depend on previous actions I_1^{t-1} and J_1^{t-1}
- Actions I_t and J_t are independent given p_t and q_t
- At the end of each round players can assess their hypothetical losses $\ell(i, J_t)$ and $\ell(I_t, j)$ had they taken other actions

Regret and Hannan Consistency

Goal for P1: Minimize regret relative to best fixed action *in retrospect*

$$R_n \;:=\; \sum_{t=1}^{n} \ell(I_t, J_t) \;-\; \min_{1 \le i \le M} \sum_{t=1}^{n} \ell(i, J_t)$$

Definition: An action strategy $I_1, I_2, \ldots$ for P1 is Hannan consistent if for all possible actions $j_1, j_2, \ldots$ of P2

$$\limsup_{n \to \infty} \left[\frac{1}{n} \sum_{t=1}^{n} \ell(I_t, j_t) \;-\; \min_{1 \le i \le M} \frac{1}{n} \sum_{t=1}^{n} \ell(i, j_t) \right] \;=\; 0 \quad \text{wp1}$$

Note: Selecting $I_t \sim p_t$ where p_t derived from EWAF with $\eta_t = \sqrt{(8 \log N)/t}$ and panel of M constant experts is Hannan consistent

Limiting Average Loss for Hannan Consistent Play

Fact: Consider a zero-sum game with players P1 and P2 and value V

1. If P1 plays a Hannan consistent strategy $I_1, I_2, \ldots$ then

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \ell(I_t, J_t) \;\leq\; V \quad \text{wp1}$$

2. If P1 and P2 play Hannan consistent strategies then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \ell(I_t, J_t) \;=\; V \quad \text{wp1}$$

Blackwell's Approachability Theorem

Setting: Two player zero sum game with *vector-valued* loss

- P1 has action space $[M]$, P2 has action space $[N]$
- Loss $\ell : [M] \times [N] \to B_m$ where $B_m = \{v \in \mathbb{R}^m : ||v|| \leq 1\}$

Question: When can P1 force his average loss to be close, asymptotically, to a given convex subset $S \subseteq B_m$?

Approachability

Definition: A set $S \subseteq B_m$ is approachable by P1 if there is a strategy $I_1, I_2, \ldots$ such that for all actions $j_1, j_2, \ldots \in [N]$ of P2

$$d\left(\frac{1}{n}\sum_{t=1}^{n} \ell(I_t, j_t), S\right) \to 0 \text{ wp1} \quad \text{where} \quad d(u, S) = \min_{v \in S} \|u - v\|$$

Note: In case $m = 1$ with $\ell \in [0, 1]$ result on Hannan consistent play shows that the interval $S = [0, s]$ is approachable if $s \geq V$

Blackwell's Approachability Theorem

Lemma: A halfspace $H = \{u : \langle a, u \rangle \leq c\}$ is approachable iff there is pmf p on $[M]$ such that

$$\max_{1 \leq j \leq N} \langle a, \overline{\ell}(p, j) \rangle \ \leq \ c$$

Theorem: A closed convex set $S \subseteq B_m$ is approachable iff every halfspace H containing S is approachable.