

Machine Learning Theory 2026

Lecture 13

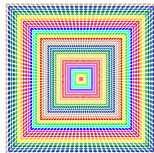
Wouter M. Koolen

Download these slides now from elo.mastermath.nl!

- ▶ Prediction with log-loss:
 - ▶ NML/Shtarkov
 - ▶ Bayes Uniform Prior/Jeffreys Prior
 - ▶ Finite Θ /Parametric Θ

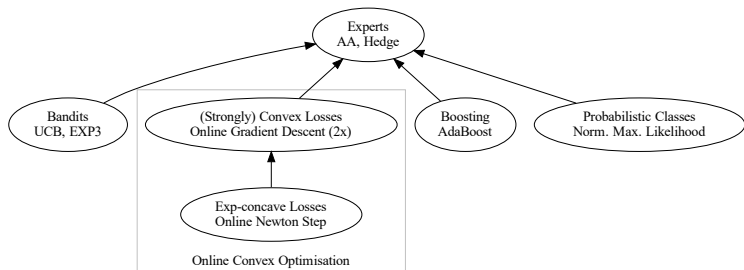
Application:

- ▶ Markov and CTW prediction



Recap

Overview of Second Half of Course



Material: course notes on MLT website.

Background Material: Chapter 9 from *Prediction, Learning and Games* by Cesa-Bianchi and Lugosi.

Outlook

Today: **adversarial online learning** with **statistical models** as our hypotheses

Main points:

- ▶ Minimax analysis tractable, elegant, insightful
- ▶ Bayesian methods can get very close
- ▶ Foundation for practical methods

Log-loss prediction

Log Loss Prediction Setup

Start with a class Θ of *simulatable* predictors for outcomes y_1, y_2, \dots

After seeing past y^{n-1} , each $\theta \in \Theta$ assigns a probability p_θ to the next outcome y_n denoted by

$$p_\theta(y_n | y^{n-1})$$

Interesting examples:

- ▶ Finite class
- ▶ Bernoulli
- ▶ Mixtures (categorical distributions)
- ▶ Markov chains
- ▶ Logistic regression

Conditional vs Joint Equivalence

A sequential one-step-ahead forecaster (aka conditional distribution)

$$p(y_t|y^{t-1})$$

induces a distribution on length- T sequences (aka joint distribution)

$$p(y^T) := \prod_{t=1}^T p(y_t|y^{t-1})$$

Conditional vs Joint Equivalence

A sequential one-step-ahead forecaster (aka conditional distribution)

$$p(y_t|y^{t-1})$$

induces a distribution on length- T sequences (aka joint distribution)

$$p(y^T) := \prod_{t=1}^T p(y_t|y^{t-1})$$

Conversely, any distribution over full T -length outcome sequences

$$p(y^T)$$

induces a one-step forecaster (by integrating out the future)

$$p(y_t|y^{t-1}) := \frac{\sum_{y_{t+1}^T} p(y^{t-1}, y_t, y_{t+1}^T)}{\sum_{y_t^T} p(y^{t-1}, y_t^T)}$$

So: two **equivalent representations** of the same object

Log Loss Prediction Notation

A predictor θ assigns to sequence y^T probability

$$p_{\theta}(y^T) = \prod_{t=1}^T p_{\theta}(y_t | y^{t-1})$$

Definition

The **maximum likelihood estimator** (MLE) for data y^T is

$$\hat{\theta}(y^T) = \arg \max_{\theta \in \Theta} p_{\theta}(y^T),$$

and the **maximum likelihood** is

$$p_{\hat{\theta}(y^T)}(y^T) = \max_{\theta \in \Theta} p_{\theta}(y^T).$$

NB: $\sum_{y^T} p_{\hat{\theta}(y^T)}(y^T) \gg 1$.

Log-loss Prediction Game

Fix a class Θ of simulatable predictors

Protocol

- ▶ For $t = 1, 2, \dots, T$
 1. The learner assigns probability $\tilde{p}_t \in \Delta_{\mathcal{Y}}$ to the next outcome.
 2. The next outcome $y_t \in \mathcal{Y}$ is revealed
 3. Learner incurs **log loss** $-\ln \tilde{p}_t(y_t)$.

NB: \tilde{p}_t typically improper (not itself in Θ)

Definition (Regret)

After T rounds, the regret is

$$\underbrace{\sum_{t=1}^T -\ln \tilde{p}_t(y_t)}_{\text{Learner's log loss}} - \underbrace{\min_{\theta \in \Theta} \sum_{t=1}^T -\ln p_{\theta}(y_t | y^{t-1})}_{\text{log loss of MLE: } -\ln p_{\hat{\theta}(y^T)}(y^T)}$$

Data compression connection

Intuition

#bits \approx log-loss

Key words:

- ▶ Shannon-Fano code : code lengths are $-\log(p)$ rounded-up
- ▶ Kraft Inequality : $2^{-\text{bit length}}$ sums to ≤ 1 for any code
- ▶ arithmetic coding: bits $\approx -\log(p^T)$ *sequentially*

What we already know: Experts

Theorem

For finite $|\Theta| < \infty$, there is an algorithm for the log loss game with regret at most $\ln|\Theta|$.

Proof.

By reduction to the mix loss game. Consider running the Agregating Algorithm from Lecture 8 on experts Θ with losses

$$\ell_t^\theta = -\ln p_\theta(y_t|y^{t-1})$$

and using w_t to form the predictions

$$\tilde{p}_t(y) = \sum_{\theta \in \Theta} w_t^\theta p_\theta(y|y^{t-1}).$$

Then log loss equals mix loss

$$-\ln \tilde{p}_t(y_t) = -\ln \sum_{\theta \in \Theta} w_t^\theta e^{-\ell_t^\theta}$$

and the $\ln|\Theta|$ regret bound follows.



What we already know: Experts

AA-based strategy takes a particularly simple form

$$\begin{aligned}\tilde{p}_t(y) &= \sum_{\theta \in \Theta} w_t^\theta p_\theta(y|y^{t-1}) \\ &= \frac{\sum_{\theta \in \Theta} e^{-\sum_{s=1}^{t-1} \ell_s^\theta} p_\theta(y|y^{t-1})}{\sum_{\theta \in \Theta} e^{-\sum_{s=1}^{t-1} \ell_s^\theta}} \\ &= \frac{\sum_{\theta \in \Theta} e^{-\sum_{s=1}^{t-1} -\ln p_\theta(y_s|y^{s-1})} p_\theta(y|y^{t-1})}{\sum_{\theta \in \Theta} e^{-\sum_{s=1}^{t-1} -\ln p_\theta(y_s|y^{s-1})}} \\ &= \frac{\sum_{\theta \in \Theta} \prod_{s=1}^{t-1} p_\theta(y_s|y^{s-1}) p_\theta(y|y^{t-1})}{\sum_{\theta \in \Theta} \prod_{s=1}^{t-1} p_\theta(y_s|y^{s-1})} \\ &= \frac{\sum_{\theta \in \Theta} p_\theta(y^{t-1}) p_\theta(y|y^{t-1})}{\sum_{\theta \in \Theta} p_\theta(y^{t-1})}\end{aligned}$$

Average of predictions $p_\theta(y|y^{t-1})$ with weights $\propto p_\theta(y^{t-1})$.

Bayes rule (uniform prior on Θ).

What we already know: Exp-concavity

Log loss is a **1-exp concave** function of the prediction $\tilde{p}_t \in \Delta_{\mathcal{Y}}$.

With $f_t(\tilde{p}_t) = -\ln \tilde{p}_t(y_t)$, we have gradient

$$\nabla f_t(\tilde{p}_t) = \nabla -\ln \tilde{p}_t(y_t) = -\frac{e_{y_t}}{\tilde{p}_t(y_t)}.$$

Potentially **unbounded** gradient (as we saw in Homework 11.2). Online Newton Step may need additional assumptions.

Questions for Today

- ▶ Is $\text{regret} \leq \ln|\Theta|$ good for this problem?
- ▶ And what if $|\Theta| = \infty$?

Minimax Regret for Log Loss

Log Loss Prediction Minimax Regret

Fix a model Θ .

Definition

The minimax regret of the T -round log-loss game on Θ is

$$\mathcal{V}_T(\Theta) := \min_{\tilde{p}_1} \max_{y_1} \min_{\tilde{p}_2} \max_{y_2} \dots \min_{\tilde{p}_T} \max_{y_T} \text{Regret}$$

Note: can be linear if Θ is too large.

Normalised Maximum Likelihood

Easier to solve the problem in whole-sequence-at-once form:

$$\begin{aligned}\mathcal{V}_T(\Theta) &= \min_{\tilde{p}_1} \max_{y_1} \min_{\tilde{p}_2} \max_{y_2} \dots \min_{\tilde{p}_T} \max_{y_T} \text{Regret} \\ &= \min_{\tilde{p}(y^T)} \max_{y^T} -\ln \tilde{p}(y^T) + \ln p_{\hat{\theta}(y^T)}(y^T)\end{aligned}$$

Normalised Maximum Likelihood

Theorem (Shtarkov)

The minimax predictor is **Normalised Maximum Likelihood**

$$p_{NML}(y^T) = \frac{\max_{\theta \in \Theta} p_{\theta}(y^T)}{\sum_{y^T} \max_{\theta \in \Theta} p_{\theta}(y^T)}$$

and the minimax regret is

$$\mathcal{V}_T(\Theta) = \ln \left(\sum_{y^T} \max_{\theta \in \Theta} p_{\theta}(y^T) \right)$$

Game-theoretic measure of capacity of Θ called **Stochastic Complexity**

Counts number of parameters $\theta \in \Theta$ that are “essentially different” at horizon T .

Rate at which you need to grow cardinality when using finite discretisation.

Proof

See Theorem 9.1 in the material.

Minimax regret

Consider again the finite Θ case. Then

$$\begin{aligned}\mathcal{V}_T(\Theta) &= \ln \left(\sum_{y^T} \max_{\theta \in \Theta} p_{\theta}(y^T) \right) \\ &\leq \ln \left(\sum_{y^T} \sum_{\theta \in \Theta} p_{\theta}(y^T) \right) \\ &= \ln |\Theta|\end{aligned}$$

Can be much smaller in practise.

Asymptotic Expansion for Minimax Regret I

Now consider the i.i.d. Bernoulli model $\Theta = [0, 1]$ where $p_\theta(1|y^{t-1}) = \theta$.

Theorem

$$\mathcal{V}_T(\Theta) = \frac{1}{2} \ln \frac{T\pi}{2} + o(1)$$

Asymptotic Expansion for Minimax Regret II

Proof.

$$\begin{aligned}\mathcal{V}_T(\Theta) &= \ln \left(\sum_{y^T} \max_{\theta \in \Theta} p_{\theta}(y^T) \right) \\ &= \ln \left(\sum_{i=0}^T \binom{T}{i} \left(\frac{i}{T}\right)^i \left(\frac{T-i}{T}\right)^{T-i} \right) \\ &\stackrel{\text{Stirling}}{\approx} \ln \left(\sum_{i=0}^T \sqrt{\frac{T}{2\pi i(T-i)}} \right) \stackrel{\text{Integral}}{\approx} \ln \left(\sqrt{\frac{T\pi}{2}} \right)\end{aligned}$$

Where the approximation is Stirling's $n! \approx \sqrt{2\pi n} \left(\frac{n}{e}\right)^n$. So that

$$\binom{T}{i} \approx \frac{\sqrt{2\pi T} \left(\frac{T}{e}\right)^T}{\sqrt{2\pi i} \left(\frac{i}{e}\right)^i \sqrt{2\pi(T-i)} \left(\frac{T-i}{e}\right)^{T-i}} = \sqrt{\frac{T}{2\pi i(T-i)}} \left(\frac{T}{i}\right)^i \left(\frac{T}{T-i}\right)^{T-i}$$

□

Asymptotic Expansion for Categorical

Consider the k -outcome categorical model $\Theta = \Delta_k$ with $p_\theta = \theta$.
Bernoulli is the case $k = 2$

Theorem

$$\mathcal{V}_T(\Theta) = \frac{k-1}{2} \ln \frac{T}{2\pi} + \ln \frac{\Gamma(1/2)^k}{\Gamma(k/2)} + o(1)$$

Proof.

See reading material



Asymptotic Expansion for i.i.d. Classes

NB: This is **just for context**

Theorem

Consider any “suitably regular” model $\Theta \subseteq \mathbb{R}^k$ of i.i.d. predictors. Then

$$\mathcal{V}_T(\Theta) = \frac{k}{2} \ln \frac{T}{2\pi} + \log \int \sqrt{\det I(\theta)} d\theta + o(1)$$

where $I(\theta)$ is the Fisher information matrix (Hessian of negative entropy)

$$I(\theta) = \mathbb{E}_{Y \sim p_\theta} [\nabla_\theta^2 \ln p_\theta(Y)].$$

Bayesian Predictors

Idea

For finite classes Θ , we saw that AA reduces to a Bayesian mixture.

Do Bayesian mixtures also control the regret for infinite Θ ?

For example, what about Bernoulli? How good is e.g. the uniform average

$$p(y^T) = \int_0^1 p_\theta(y^T) d\theta$$

Uniform Average aka Laplace Mixture

Theorem

The uniform average predictor has predictions

$$p_t(1|y^{t-1}) = \frac{n_1(y^{t-1}) + 1}{t + 1}$$

and worst-case regret equal to

$$\max_{y^T} \text{Regret} = \ln(T + 1)$$

About twice $\mathcal{V}_T(\Theta) \dots$

Jeffreys' Average

Jeffreys proposed (based on invariance considerations) the prior

$$p(\theta) = \frac{1}{\pi\sqrt{\theta(1-\theta)}}$$

Theorem

The Jeffreys predictor is equivalent to the Krichevsky-Trofimoff predictor

$$p_t(1|y^{t-1}) = \frac{n_1(y^{t-1}) + 1/2}{t}$$

and has worst-case regret equal to

$$\max_{y^T} \text{Regret} \leq \frac{1}{2} \ln(T) + \ln 2$$

Matches $\mathcal{V}_T(\Theta)$ up to lower-order constant.

General Bayesian Mixures

NB: **this is just for context**

For a general model, Jeffreys' prior is

$$p(\theta) = \frac{\sqrt{\det I(\theta)}}{\int \sqrt{\det I(\theta)} d\theta}$$

Where $I(\theta)$ is the Fisher Information matrix.

Theorem

Consider a suitably regular i.i.d. $\Theta \subseteq \mathbb{R}^k$. The worst-case regret of Bayesian model averaging with Jeffreys' prior is

$$\max_{y^T} \text{Regret} = \frac{k}{2} \ln \frac{T}{2\pi} + \log \int \sqrt{\det I(\theta)} d\theta + o(1)$$

Equal to minimax regret $\mathcal{V}(\Theta)$ up to $o(1)$.

Practice: Bayesian methods easier to interpret/compute than minimax.

Applications

Markov Models

k th order Markov model can be summarised by a table

context	prediction
00	θ_{00}
01	θ_{01}
10	θ_{10}
11	θ_{11}

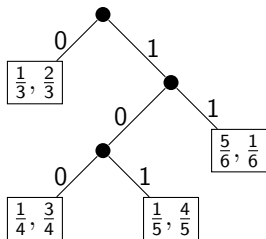
In context x , assign probability θ_x to seeing outcome 1 next.

0 1 0 1 0 0 1 0 1 0 1 0 1 0 1 ?
context

2^k parameters.

Bayesian average can be maintained efficiently. Regret is about $2^{k-1} \ln T$.

Application: Context Tree Weighting (CTW)



To predict next symbol: look up context right-to-left from root, use leaf dist.

$$\overbrace{01101}^{\text{context}} \underbrace{001}_{\text{used}} ? \Rightarrow \underbrace{\frac{1}{4}, \frac{3}{4}}_{\text{prediction}}$$

- ▶ 2^{k+1} parameters for maximum context length k .
- ▶ $O(k)$ per round implementation of Bayesian model average over all context tree predictors
- ▶ Excellent data compression performance.

Conclusion

- ▶ Prediction with log loss has elegant exact minimax solution:
normalized maximum likelihood
- ▶ **Bayesian mixtures** (version of AA) with carefully selected priors can often match the minimax regret
- ▶ Can tackle complex models with (hierarchical) Bayesian mixtures