DATA 37200: Learning, Decisions, and Limits (Winter 2026)

Lecture 9: The Online Ridge Forecaster

Instructor: Frederic Koehler

References

Cesa-Bianchi and Lugosi, Chapters 3 and 11.

Recap

- We saw how to use ϵ-Greedy to reduce contextual bandits to online prediction/forecasting/learning (with squared loss).
- For finite classes, we saw how to solve online prediction (even in the "adversarial" setting with no probabilities involved) using multiplicative updates. We showed how to derive the algorithm via a simple Bayesian model.
- For large/infinite classes, we can forecast with low regret by discretizing the function class (ϵ-net argument), using the $\log |\mathcal{F}|$ scaling of multiplicative weights.
- However, this is often not algorithmically practical.
- For a "real world" setting like linear models, can we find a faster forecasting strategy?

From Finite to Infinite Experts

Recap: Finite Experts

- We had K discrete experts.
- Prior: Uniform $1/K$.
- Algorithm: Exponential Weights (Posterior Mean).
- Regret: $O(\log K)$.

New Setting: Linear Experts

- Experts are now vectors $u \in \mathbb{R}^d$.
- At step t, we see feature vector $x_t \in \mathbb{R}^d$.
- Expert u predicts $f_u(t) = u^\top x_t$.
- **Goal:** Compete with the best fixed vector u^*:

$$\min_{u \in \mathbb{R}^d} \sum_{t=1}^{T} (y_t - u^\top x_t)^2$$

The Bayesian Linear Model

We apply the same Bayesian strategy: *Predict the Posterior Mean*.

1. The Prior (Gaussian): Instead of a uniform prior, we place a Gaussian prior on the weight vector u, centered at 0 with variance parameter $a > 0$:

$$u \sim \mathcal{N}(\mathbf{0}, aI)$$

$$p(u) \propto \exp\left(-\frac{\|u\|^2}{2a}\right)$$

2. The Likelihood (Gaussian): We model the data generation as linear with Gaussian noise (variance σ^2):

$$y_t \mid x_t, u \sim \mathcal{N}(u^\top x_t, \sigma^2)$$

$$p(y_t \mid x_t, u) \propto \exp\left(-\frac{(y_t - u^\top x_t)^2}{2\sigma^2}\right)$$

Deriving the Posterior

After observing data $D_{t-1} = \{(x_1, y_1), \ldots, (x_{t-1}, y_{t-1})\}$, the posterior $p(u \mid D_{t-1})$ is proportional to Prior $\times$ Likelihoods.

The exponent looks like:

$$-\frac{1}{2}\left(\frac{\|u\|^2}{a} + \sum_{s=1}^{t-1} \frac{(y_s - u^\top x_s)^2}{\sigma^2}\right)$$

This is a quadratic form in u, meaning the posterior is also Gaussian:

$$u \mid D_{t-1} \sim \mathcal{N}(\mu_{t-1}, \Sigma_{t-1})$$

The posterior mean is a **ridge regression** estimator (next slide).

Online Ridge: Predict the Posterior Mean

We define the correlation matrix A_{t-1} (inverse covariance):

$$A_{t-1} = \frac{\sigma^2}{a} I + \sum_{s=1}^{t-1} x_s x_s^\top$$

The mean of the posterior μ_{t-1} minimizes the Ridge objective:

$$\mu_{t-1} = \arg\min_u \left(\frac{\sigma^2}{a} \|u\|^2 + \sum_{s=1}^{t-1} (y_s - u^\top x_s)^2 \right)$$

The Algorithm (Online Ridge Forecaster)

At time t:

1. Receive x_t.
2. Predict $\hat{y}_t = \mu_{t-1}^\top x_t$.
3. Receive y_t.
4. Compute μ_t (updated Ridge solution).

Regret Bound

We analyze the regret against any fixed comparator u.

Theorem (Regret of Ridge Forecaster)

Let $\lambda = \sigma^2/a$ and suppose $\sigma^2 \geq 1$. The cumulative squared loss of the algorithm satisfies:

$$\sum_{t=1}^{T}(\hat{y}_t - y_t)^2 - \sum_{t=1}^{T}(u^\top x_t - y_t)^2 \leq \lambda\|u\|^2 + \sigma^2 \log\det(A_T) - \sigma^2 \log\det(\lambda I)$$

Simplification: This bound is best when we take $\sigma^2 = 1$ (same as in finite case). Suppose for simplicity/by rescaling that $\|x_t\| \leq 1$ always, then

$$\log\det A_T = \sum_{i=1}^{d} \log \lambda_i(A_T) \leq d \log(\lambda + T)$$

Take $\lambda = 1$; the regret against $\|u\| \leq R$ is $O(R^2 + d\log(T+1))$.

Recall: Exp-Concavity

Like last time, to analyze the regret we we use the property of **exp-concavity**.

Exp-concavity of Squared Loss

For domains $[0,1]$ and any outcome $y \in [0,1]$, the function:

$$G(x) = \exp\left(-\eta(x-y)^2\right)$$

is concave in x provided that $\eta \le \frac{1}{2}$.

We will apply this with $\eta = 1/2\sigma^2$, in which case the condition $\eta \le 1/2$ becomes $\sigma^2 \ge 1$.

Exp-concavity: posterior mean dominates posterior

Since our algorithm predicts the posterior mean

$$\hat{y}(t) = \mathbb{E}_{u \sim p(u|\mathcal{F}_{t-1})}[u^T x(t)],$$

by exp-concavity and Jensen's inequality, we know that:

$$\mathbb{E}_{u \sim p(u|\mathcal{F}_{t-1})}\left[e^{-\eta(u^T x(t) - y(t))^2}\right] \leq e^{-\eta(\hat{y}(t) - y(t))^2}$$

In English: the likelihood of the response y under the model $N(\hat{y}, 1/2\eta)$ is always higher than the likelihood under the posterior $\int N(u^T x, 1/2\eta) dp(u \mid \mathcal{F}_{t-1})$. Taking logs,

$$\log \mathbb{E}_{u \sim p(u|\mathcal{F}_{t-1})}\left[e^{-\eta(u^T x(t) - y(t))^2}\right] \leq -\eta(\hat{y}(t) - y(t))^2$$

Note: we can improve on the posterior because in reality $y \in [0, 1]$, but the posterior does not know this, it is based on a Gaussian assumption. Here the misspecification of our model is "useful".

Potential analysis

Follow the pattern from last time:

- If the data is fit well by a some linear model $u \in \mathbb{R}^d$, the log-likelihood of the data under the Bayesian model is high.
- By exp-concavity, the log-likelihood of the data under the posterior mean model is always better !
- Log-likelihood under the posterior mean model is the same as squared loss.

Now we go through these steps in detail and see how it yields the regret bound.

Proof via Potential Functions (Step 1)

We define the **Potential Function** as the negative log-marginal likelihood (normalizing constant):

$$\Phi_t = -\sigma^2 \log \left(\int_{\mathbb{R}^d} \prod_{s=1}^{t} e^{-\frac{(y_s - u^\top x_s)^2}{2\sigma^2}} \cdot e^{-\frac{\|u\|^2}{2a}} du \right)$$

(Note the scaling factor σ^2 to match the squared loss scale).

This integral can be computed exactly for Gaussians:

$$\int e^{-\frac{1}{2\sigma^2}(\sum (y_s - u^\top x_s)^2 + \lambda \|u\|^2)} du = \sqrt{\frac{(2\pi\sigma^2)^d}{\det(A_t)}} e^{-\frac{1}{2\sigma^2} \min_u L_t(u)}$$

where $L_t(u)$ is the cumulative Ridge loss.

Proof via Potential (Step 2)

Taking the log of the integral:

$$\Phi_t = \frac{\sigma^2}{2} \log \det(A_t) + \min_u \frac{1}{2} \left(\lambda \|u\|^2 + \sum_{s=1}^{t} (y_s - u^\top x_s)^2 \right) + \text{const}$$

On the other hand, consider the incremental update $\Phi_t - \Phi_{t-1}$. By Bayes rule, the difference is given by the log likelihood of the observation

$$\Phi_t - \Phi_{t-1} = -\sigma^2 \log P(y_t \mid x_t, D_{t-1}) \geq \frac{1}{2} (\hat{y}_t - y_t)^2$$

where we used $\eta = 1/2\sigma^2$ and the last inequality was the key conclusion from exp-concavity.

Proof via Potential (Conclusion)

Telescoping the last inequality, we find

$$\Phi_T - \Phi_0 = \sum_{t=1}^{T} (\Phi_t - \Phi_{t-1}) \geq \frac{1}{2} \sum_{t=1}^{T} (\hat{y}_t - y_t)^2.$$

We also computed that

$$\Phi_T = \frac{\sigma^2}{2} \log\det(A_T) + \min_u \frac{1}{2} \left(\lambda \|u\|^2 + \sum_{s=1}^{T} (y_s - u^\top x_s)^2 \right) + \text{const}$$

and similarly $\Phi_0 = \frac{\sigma^2}{2} \log\det(A_0) + \text{const}$. So indeed,

$$\frac{1}{2} \sum_{t=1}^{T} (\hat{y}_t - y_t)^2 \leq \frac{\sigma^2}{2} \log\det(A_T) - \frac{\sigma^2}{2} \log\det(\lambda I) + \min_u \frac{1}{2} \left(\lambda \|u\|^2 + \sum_{s=1}^{t} (y_s - u^\top x_s)^2 \right)$$

An algorithmic improvement

The Computational Bottleneck

- In the naive implementation, computing the posterior mean $\mu_t = A_t^{-1} \sum_{s=1}^{t} y_s x_s$ requires inverting a $d \times d$ matrix at every step.
- Naive inversion takes $O(d^3)$. Total time for T rounds: $O(Td^3)$.
- For high-dimensional features ($d \gg 1$), this is impractical.

Solution: Rank-One Updates

- Recall that $A_t = A_{t-1} + x_t x_t^\top$.
- We can update the inverse matrix $P_t = A_t^{-1}$ directly using the Sherman-Morrison formula:

$$P_t = P_{t-1} - \frac{P_{t-1} x_t x_t^\top P_{t-1}}{1 + x_t^\top P_{t-1} x_t}$$

- This reduces the cost to $O(d^2)$ per step.
- This trick is called **Recursive Least Squares (RLS)**. Invented by Gauss in early 1800s?

Final remarks

- We studied online ridge, using exp-concavity, because this works very nicely under the assumption that responses (rewards) are $[0, 1]$ valued.
- $O(d \log T)$ regret turns out to be minimax.
- With ϵ-greedy: yields $\tilde{O}(d^{1/3} T^{2/3} (\log T)^{1/3})$ regret for CB.

More advanced topics:

- There is a well-known variant of online ridge called the Vovk-Azoury-Warmuth (VAW) forecaster. It has better constant factors, and if y_t are drawn from an unbounded domain, VAW is more elegant than online ridge.
- VAW is tied to Vovk's Aggregating Algorithm and related concept of "mixability" (more general/sophisticated concept than exp-concavity). See textbook.

extra slides

Why is the rank-one update true? (Intuition)

We expect $(A + uv^T)^{-1}$ may be similar to A^{-1}.
The **Sherman-Morrison formula** gives the exact correction:

$$(A + uv^\top)^{-1} = A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{1 + v^\top A^{-1}u}$$

To motivate this, observe in the scalar case that

$$\frac{1}{a + uv} = \frac{1}{a} - \frac{uv}{a(a + uv)}$$

which is easy to check.

Verification of Sherman-Morrison (Part 1)

To prove $(A + uv^\top)^{-1} = A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{1+v^\top A^{-1}u}$, we multiply the matrix by the claimed inverse and check if we get I.

Let $\gamma = 1 + v^\top A^{-1}u$ (a scalar) and $B = A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{\gamma}$.

$$
\begin{aligned}
(A + uv^\top)B &= (A + uv^\top)\left(A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{\gamma}\right) \\
&= AA^{-1} - \frac{AA^{-1}uv^\top A^{-1}}{\gamma} + uv^\top A^{-1} - \frac{uv^\top A^{-1}uv^\top A^{-1}}{\gamma} \\
&= I - \frac{uv^\top A^{-1}}{\gamma} + uv^\top A^{-1} - \frac{u(v^\top A^{-1}u)v^\top A^{-1}}{\gamma}
\end{aligned}
$$

Key Observation: The term in the middle $(v^\top A^{-1}u)$ is exactly the scalar $(\gamma - 1)$.

Verification of Sherman-Morrison (Part 2)

Continuing from the previous slide, we substitute $v^\top A^{-1}u = \gamma - 1$:

$$(A + uv^\top)B = I + uv^\top A^{-1} - \left(\frac{uv^\top A^{-1} + u(\gamma - 1)v^\top A^{-1}}{\gamma}\right)$$

Factor out $uv^\top A^{-1}$ in the numerator:

$$\begin{aligned}
&= I + uv^\top A^{-1} - \left(\frac{u(1 + \gamma - 1)v^\top A^{-1}}{\gamma}\right) \\
&= I + uv^\top A^{-1} - \left(\frac{\gamma uv^\top A^{-1}}{\gamma}\right) \\
&= I + uv^\top A^{-1} - uv^\top A^{-1} \\
&= I \quad \blacksquare
\end{aligned}$$

Since the product is the Identity matrix, the formula for the inverse is correct.