SD = ONLINE LEARNING

I. PREDICTION WITH EXPERT ADVICE

- Framework: Sequential predictions: time matters!

  Observations are $y_1, y_2, \dots, y_T$

  At time $t-1$, we predict y_t using $\hat{p}_t$, and we measure the quality of the prediction by means of the loss $\ell(\hat{p}_t, y_t)$.
  The decision maker's prediction is based on expert advices:
  At time $t-1$, expert k $(k=1,\dots,K)$ predicts f_t^k, and their prediction is revealed to the decision maker.
  ↳ $\hat{p}_t$ is then constructed from $f_t^1, \dots, f_t^K$.

  Goal: construct a meaningful estimator of y_t using expert advice.

- Loss at time t is $\ell(\hat{p}_t, y_t)$
  Loss of expert k at time t is $\ell(f_t^k, y_t)$.

  ⇒ Cumulative loss after T rounds is $\sum_{t=1}^{T} \ell(\hat{p}_t, y_t)$.

  This loss is compared to a benchmark value: the cumulative loss of the best expert $\min_{1\le k\le K} \sum_{t=1}^{T} \ell(f_t^k, y_t)$.

  ⇒ Aim is to minimize the cumulative regret R_T:

$$R_T := \sum_{t=1}^{T} \ell(\hat{p}_t, y_t) - \min_{1\le k\le K} \left(\sum_{t=1}^{T} \ell(f_t^k, y_t) \right)$$

$\sum_{t=1}^{T} \ell(\hat{p}_t, y_t) =: L_T$; $\sum_{t=1}^{T} \ell(f_t^k, y_t) =: L_T^k$

I.1. Exponential Weighting.

- Consider a weighted average prediction:

  For $t \ge 1$, $\hat{p}_t := \sum_{k=1}^{K} \left\{ \frac{\omega_{t-1}^k}{\sum_{j=1}^{K} \omega_{t-1}^j} \right\} f_t^k$,

  with the convention that $\omega_0^1 = \dots = \omega_0^K = 1$, for non-negative weights ω_{t-1}^k.

- We use exponential weights $\omega_{t-1}^k = e^{-\beta L_{t-1}^k} = \exp\left(-\beta \sum_{s=1}^{t-1} \ell(f_s^k, y_s)\right)$.

  For some $\beta > 0$

  The weights depend only on the cumulative performance of the experts, and not on the decision maker's performance.

- We derive an upper bound for the cumulative regret R_T using exponential weights.

  Put $W_0 := K \ge W_t = \sum_{k=1}^{K} e^{-\beta L_t^k}$, $t \ge 1$

  After T rounds,

$$\ln\left(\frac{W_T}{W_0}\right) = \ln\left(\sum_{k=1}^{K} e^{-\beta L_T^k}\right) - \ln K$$

$$\ge \ln\left(\max_{1\le k\le K} e^{-\beta L_T^k}\right) - \ln K$$

$$= -\beta \underbrace{\min_{1\le k\le K} L_T^k}_{\text{cumulative loss of the best expert}} - \ln K \qquad (*)$$

↳ We derive an upper bound for $\ln\left(W_T/W_0\right)$.

- $\forall t = 1, \dots, T$,

$$\ln\left(\frac{W_t}{W_{t-1}}\right) = \ln\left(\frac{\sum_{k=1}^K e^{-\beta L_t^k}}{\sum_{k=1}^K e^{-\beta L_{t-1}^k}}\right)$$

$$= \ln\left(\frac{\sum_{k=1}^K e^{-\beta L_{t-1}^k} \, e^{-\beta \ell(f_t^k, y_t)}}{\sum_{k=1}^K e^{-\beta L_{t-1}^k}}\right)$$

$$= \ln\left(\sum_{k=1}^K \left\{\frac{e^{-\beta L_{t-1}^k}}{\sum_{j=1}^K e^{-\beta L_{t-1}^j}}\right\} e^{-\beta \ell(f_t^k, y_t)}\right)$$

$\uparrow$ = probability distribution

= Expected value of a random variable $\equiv \mathbb{E}(\exp(-\beta X))$ for X taking values in $[0,1]$, with mean $\sum_{k=1}^K \{\dots\} \ell(f_t^k, y_t)$

Hoeffding's lemma $\leq \frac{\beta^2}{8} - \beta \sum_{k=1}^K \{\dots\} \ell(f_t^k, y_t)$

Assuming a loss function ℓ convex in its first argument,

$$\geq \ell\left(\sum_{k=1}^K \frac{w_{t-1}^k}{\sum_j w_{t-1}^j} f_t^k, y_t\right)$$

$$= \ell(\hat{p}_t, y_t)$$

Thus $\ln\left(\frac{W_t}{W_{t-1}}\right) \leq \frac{\beta^2}{8} - \beta \ell(\hat{p}_t, y_t)$, and

$$\ln\left(\frac{W_T}{W_0}\right) \leq \frac{\beta^2 T}{8} - \beta \sum_{t=1}^T \ell(\hat{p}_t, y_t).$$

$\uparrow$ Together with (*) page 2 gives:

$$\frac{\beta^2 T}{8} - \beta \sum_{t=1}^T \ell(\hat{p}_t, y_t) \geq -\beta \min_{1 \leq k \leq K} L_T^k - \ln K$$

$$\Leftrightarrow \sum_{t=1}^T \ell(\hat{p}_t, y_t) - \min_{1 \leq k \leq K} L_T^k \leq \underbrace{\frac{\ln K}{\beta} + \frac{\beta T}{8}}.$$

Minimizing this term as a function of β gives $\beta = \sqrt{8 \log K / T}$, for which we obtain

$$R_T = \sum_{t=1}^T \ell(\hat{p}_t, y_t) - \min_{1 \leq k \leq K} L_T^k \leq \sqrt{\frac{T \log K}{2}}$$

Regret bound for exponential weighting / loss function convex in its first argument.

- Remarks (i) The bound is obtained for a choice of β that depends on the horizon T. Bounds that hold uniformly over time can be derived using the so-called "doubling trick", see Section 2.3 in N. Cesa-Bianchi & G. Lugosi, Prediction, Learning, and Games.

(ii) Other weighted average forecaster may be considered, such as the Polynomially Weighted Average Forecaster.

I.2. Follow the perturbed leader.

Consider again a weighted average strategy:

$$\hat{p}_t = \sum_{k=1}^K \hat{\pi}_{t-1}^k f_t^k = \hat{\pi}_{t-1}^t f_t \quad ; \quad \hat{\pi}_{t-1} = (\hat{\pi}_{t-1}^1, \dots, \hat{\pi}_{t-1}^K)^t$$

t = transpose

$$f_t := (f_t^1, \dots, f_t^K)^t,$$

where $\hat{\pi}_{t-1} \in \Delta^K := \left\{ x \in \mathbb{R}^K \mid x_i \geq 0 \ \sum_{i=1}^K x_i = 1 \right\}$.

- An alternative to exponential weighting is to select $\hat{\pi}_{t-1} \in \Delta^K$ such that

$$\hat{\pi}_{t-1} = \underset{\pi \in \Delta^K}{\operatorname{argmin}} \sum_{s=1}^{t-1} \pi^t z_s \quad ; \quad z_s := \begin{pmatrix} \ell(f_s^1, y_s) \\ \vdots \\ \ell(f_s^K, y_s) \end{pmatrix}$$

& predict $\hat{p}_t = \hat{\pi}_{t-1}^t f_t$ "FOLLOW THE LEADER"

↑ Indeed, the total cumulated loss over T rounds is

$$\sum_{t=1}^T \ell(\hat{p}_t, y_t) = \sum_{t=1}^T \ell\left(\sum_{k=1}^K \hat{\pi}_{t-1}^k f_t^k, y_t \right)$$

Assuming a loss convex in its first argument ↳

$$\leqslant \sum_{t=1}^T \sum_{k=1}^K \hat{\pi}_{t-1}^k \ell(f_t^k, y_t)$$

$$= \sum_{t=1}^T \underbrace{\hat{\pi}_{t-1}^t z_t}$$

$\pi \in \Delta^K$ is selected at each time step so as to minimize this upper bound.

However, this strategy is not guaranteed to get an average regret that goes to 0 with T (counter-examples can be constructed).
↳ We regualize the follow the leader strategy by adding a small amount of noise:

$$\hat{\pi}_{t-1} = \underset{\pi \in \Delta^K}{\operatorname{argmin}} \sum_{s=1}^{t-1} \pi^t (z_s + \xi) \text{ , where } \xi \sim \mathcal{U}([0, M]^K)$$

"FOLLOW THE PERTURBED LEADER"

Then, with $M = \sqrt{KT}$, $\mathbb{E}\{R_T\} \leq 2\sqrt{2TK}$

↑ Average regret is of order $O\left(\sqrt{\frac{K}{T}}\right)$

II- MULTI-ARMED BANDITS

II.1. The model (i.i.d. case)

- We consider a multi-armed bandit problem with K arms. At each discrete time step $t = 1, 2, 3, \ldots$, an arm is pulled ≡ an action $A_t \in \{1, \ldots, K\}$ is taken, generating a reward $r_t \in [0, 1]$.
The distribution of $r_t \mid A_t = k$ is unknown. Denote $m_k := \mathbb{E}(r_t \mid A_t = k)$ its mean (also unknown).
For a fixed time horizon T, the goal is to maximize the total reward $\sum_{t=1}^T r_t$ ← If we knew $m_1, \ldots, m_K$, this would be easy.

- i.i.d. assumption: conditionally on the arm pulled, rewards are assumed i.i.d.
Used to estimate which arm is the best

- The performance of a strategy (≡ an algorithm) is measured in terms of the regret:
Put $m_* = m_{k^*} = \mathbb{E}(r_t \mid A_t = k^*) = \max_{1 \leqslant k \leqslant K} \mathbb{E}(r_t \mid A_t = k)$.

Highest expected reward

index of the arm generating the highest expected reward

The cumulative regret up to time T is $R_T := \sum_{t=1}^T (m_* - m_{A_t})$.

(expected value)

R_T is a RV since the choice of the arm is random. Consider instead the expected regret $\mathbb{E}(R_T)$

Note that

$$R_T = \sum_{t=1}^{T} (m_* - m_{A_t}) = \sum_{k=1}^{K} \sum_{t | A_t = k} (m_* - m_k) = \sum_{k=1}^{K} n_T(k)\,(m_* - m_k)$$

$(m_* - m_k) = \Delta_k$

Introduce $n_T(k)$ = # times arm k is pulled up to time T

Δ_k = gap (unknown)
= difference in value between action k and optimal action k^*.

We want to discard arms with large gaps.

In this notation,

$$\mathbb{E}(R_T) = \sum_{k=1}^{K} \Delta_k \, \mathbb{E}(n_T(k))$$

Used to assess the quality of an algorithm (given past obs, which arm to pull?)

× Lower bound on regret.

The next theorem provides a lower bound on the expected regret, in the special case that $R_t | A_t = k \sim B(m_k)$.

Thm (Lai & Robbins (1985))

Consider a K-multi-armed bandit problem, such that the m_k are not all equal.
If $\forall \alpha > 0$, holds $\mathbb{E}(R_T) = o(T^\alpha)$,
Then

$$\liminf_{T \to +\infty} \frac{\mathbb{E}(R_T)}{\ln T} \geq \sum_{k | m_* > m_k} \frac{m_* - m_k}{KL(m_k \| m_*)},$$

where $KL(m_k \| m_*)$ denotes the KL divergence between the two binomial distributions with parameter m_k & m_*,

$$KL(m_k \| m_*) = m_k \ln\left(\frac{m_k}{m_*}\right) + (1-m_k) \ln\left(\frac{1-m_k}{1-m_*}\right)$$

× Remarks (i) Condition $\mathbb{E}(R_T) = o(T^\alpha)$ $\forall \alpha > 0$ is required for the lower bound to hold. Indeed, consider an algorithm that always pulls the first arm. If this arm turns out to be the best arm, the algorithm makes no mistake, and the regret is zero; the lower bound cannot hold in this case. However, for such an algorithm, if the first arm is not the best one, $\Delta_1 > 0$, $n_T(1) = T$, and the regret $R_T = \Delta_1 T$ is linear in T.

(ii) In words, Lai & Robbins 's result show that any good algorithm must make at least a logarithmic number of mistakes.
In addition, observe that $2\Delta_k^2 \leq KL(m_k \| m_*) \leq \frac{\Delta_k^2}{m_*(1-m_*)}$

Pinsker's inequality. Using $\ln x \leq x - 1$

This indicates that an algorithm such that

$$\mathbb{E}(R_T) = O\left(\log T \sum_{k | m_k \neq m_*} \Delta_k^{-1}\right)$$ must be close to optimal.

(iii) The bound presented in the Bernoulli setting can be extended to more general distributions, see Apostolos Burnetas & Michael Katehakis (1996).

• The average reward when pulling arm k can be naturally estimated using the sample mean:

$$\overline{X}_{k,t} := \frac{1}{n_t(k)} \sum_{s=1}^{t} r_s \, \mathbb{1}(A_s = k)$$

• First attempts :

↘ Greedy algorithm : (little exploration)

After some fixed amount of exploration, keep pulling the arm with the largest empirical mean $\overline{X}_{k,t}$.

Issue : the regret of this strategy is linear, since there is always a strictly positive probability of not selecting the best arm.

↘ ε-greedy algorithm : ($\varepsilon > 0$ small) (more exploration)

With probability $(1-\varepsilon)$, pull the arm with the highest sample mean, & with probability ε, select a random arm with probability $1/K$.

Issue : the regret is linear here as well, since each arm is pulled on average $\geq \frac{\varepsilon T}{K}$ after T rounds, yielding $\mathbb{E}(R_T) \geq \frac{\varepsilon T}{K} \sum_{k : m_* > m_k} \Delta_k$.

We explore in sections II.2 and II.3 two better alternatives: Thompson sampling & UCB.

II.2. Thompson sampling

The idea of Thompson sampling is to randomly draw each arm according to its (posterior) probability of being optimal → Bayesian framework. We expose the ideas in the context of Bernoulli bandits, and then present the sampling algorithm in the general case.

× Assumption : $(r_t \mid A_t = k) \sim B(\theta_k)$, $1 \leq k \leq K$

& put $\theta := (\theta_1, \dots, \theta_K)$ = vector of mean rewards

Put a Beta prior on each of the mean rewards :

$$p(\theta_k) = \frac{\Gamma(\alpha_k + \beta_k)}{\Gamma(\alpha_k)\Gamma(\beta_k)} \theta_k^{\alpha_k - 1} (1-\theta_k)^{\beta_k - 1}, \qquad 1 \leq k \leq K$$

$$(\sim B(\alpha_k, \beta_k))$$

The posterior distribution of θ_k, provided arm k is pulled at time $t=1$ ($A_1 = k$) & reward r_1 is observed, is :

$\pi_{k,1}(\theta_k)$ ← posterior of arm k, time $t=1$

$\propto$ likelihood × prior

$\propto \underbrace{\theta_k^{r_1}(1-\theta_k)^{1-r_1}}_{B(\theta_k)} \underbrace{\theta_k^{\alpha_k - 1}(1-\theta_k)^{\beta_k - 1}}_{\propto B(\alpha_k, \beta_k)}$

$\sim B(\alpha_k + r_1, \beta_k + 1 - r_1)$

↳ still a Beta distribution.

Thus, update the distribution according to the following rule :

$$(\alpha_k, \beta_k) \leftarrow \begin{cases} (\alpha_k, \beta_k) & \text{if } A_t \neq k \\ (\alpha_k + r_t, \beta_k + 1 - r_t) & \text{if } A_t = k \end{cases}$$

(5) Only the parameter of the selected arm is updated.

BERNOULLI THOMPSON SAMPLING

Consider a $B(\alpha_k, \beta_k)$ prior on θ_k, $1 \leq k \leq K$.

For $t = 1, 2, \dots$

(i) Draw $\theta_{k,t} \sim B(\alpha_k, \beta_k)$, $1 \leq k \leq K$

(ii) Select arm $j = \operatorname{argmax}_k \theta_{k,t}$ & observe reward r_t

(iii) Update $(\alpha_k, \beta_k) \leftarrow \begin{cases} (\alpha_k, \beta_k) & \text{if } k \neq j \\ (\alpha_k + r_t, \beta_k + 1 - r_t) & \text{if } k = j \end{cases}$

* Remarks: (i) For $\alpha_k = \beta_k = 1$, we have a uniform prior.

(ii) $B(\alpha_k, \beta_k)$ has mean $\frac{\alpha_k}{\alpha_k + \beta_k}$. As an alternative to step (i) in the Thompson sampling algorithm, we may take $\theta_{k,t} = \frac{\alpha_k}{\alpha_k + \beta_k}$ equal to the posterior mean.

(iii) Let n_k denote the total number of times arm k was pulled, and S_k the total number of successes. The posterior distribution of arm k is $B(\alpha_k + S_k, \beta_k + n_k - S_k)$, with mean

$$\frac{\alpha_k + S_k}{\alpha_k + \beta_k + n_k} = \left\{ \alpha_k \left(\frac{\alpha_k + \beta_k}{\alpha_k + \beta_k} \right) + S_k \left(\frac{n_k}{n_k} \right) \right\} \times (\alpha_k + \beta_k + n_k)^{-1}$$

$$= \frac{n_k}{\alpha_k + \beta_k + n_k} \left(\frac{S_k}{n_k} \right) + \frac{\alpha_k + \beta_k}{\alpha_k + \beta_k + n_k} \left(\frac{\alpha_k}{\alpha_k + \beta_k} \right)$$

$\frac{S_k}{n_k}$ = MLE; $\frac{\alpha_k}{\alpha_k + \beta_k}$ = prior mean; the two coefficients sum to one.

The posterior mean is a weighted combination of the MLE & the prior mean, and converges to the MLE as the number of observations increase.

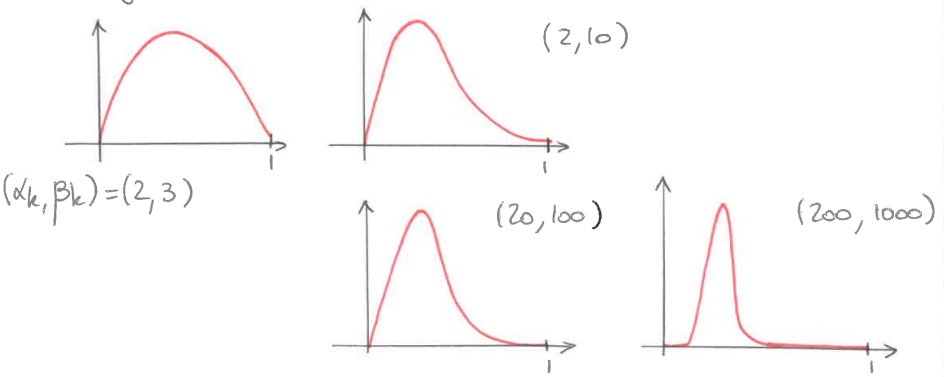
(iv) The update rule for arm k, provided arm k is pulled, is:

$$(\alpha_k, \beta_k) \to \begin{cases} (\alpha_k + 1, \beta_k) & \text{if } r_t = 1 \\ (\alpha_k, \beta_k + 1) & \text{if } r_t = 0 \end{cases}$$

Generally, as the α parameter increases, the distribution is shifted to the left, and as β increases, it is shifted to the right. In addition, the beta distribution becomes more concentrated as $(\alpha + \beta)$ grows.

• Ex: Consider a sequence (α_k, β_k) such that $\frac{\alpha_k}{\alpha_k + \beta_k} \to 0.2$

A rough sketch of the associated beta distributions:



the beta distribution gets more & more concentrated around its mode $(\alpha_k - 1)/(\alpha_k + \beta_k - 2)$

(v) Regret analysis.

S. Agrawal & N. Goyal (2012) showed that Thompson sampling achieves a logarithmic expected regret. Specifically, assuming that the first arm is optimal, and denoting $\Delta_k = \theta_1 - \theta_k$, $2 \le k \le K$, we have

$$\mathbb{E}(R_T) = O\left(\left(\sum_{k=2}^{K} \frac{1}{\Delta_k^2} \right)^2 \log T \right)$$

[REF] Theorem 2 in S. Agrawal & N. Goyal, Analysis of Thompson Sampling for the Multi-armed bandit problem. JMLR Vol 23 (2012) 39.1 - 39.26

Thompson sampling for general stochastic bandits :

Consider a prior $\pi_0(\theta_k)$ on θ_k, $1 \le k \le K$.

For $t = 1, 2, \dots$

(i) Draw $\theta_{k,t} \sim \pi_{t-1}(\theta_k)$, $1 \le k \le K$

(ii) Pull arm $A_t = j$ that maximizes the expected reward,

$$j = \underset{1 \le k \le K}{\operatorname{argmax}} \; \mathbb{E}(r_t \mid A_t = k, \theta_{k,t})$$

and observe reward r_t

(iii) Update the posterior

$$\pi_t(\theta_j) \propto \underbrace{\ell(r_t ; A_t = j, \theta_{j,t})}_{\text{likelihood}} \pi_{t-1}(\theta_j).$$

II.3. Upper Confidence Bound (UCB)

The more uncertain we are about an action value and the more important it is to explore that action.

⇒ Estimate an upper confidence for each action value/mean, and select an action based on this upper bound (instead of the sample mean $\overline{X}_{k,t}$).

- Hoeffding's inequality implies that w.p. $\ge 1 - \frac{2}{t^2}$,

$$|\overline{X}_{k,t} - m_k| < \sqrt{\frac{\ln t}{n_t(k)}}$$

Recall: $\mathbb{P}(|\overline{X} - \mu| \ge \delta) \le 2e^{-2n\delta^2}$, $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Put $U_t(k) := \overline{X}_{k,t} + \sqrt{\frac{\ln t}{n_t(k)}}$

Note that

- $U_t(k) > m_k$ w.p. $\ge 1 - \frac{2}{t^2}$ ——— (1)
- $\overline{X}_{k,t} < m_k + \frac{\Delta_k}{2}$ w.p. $\ge 1 - \frac{2}{t^2}$, ——— (2)

provided $n_t(k) \ge \frac{4 \ln t}{\Delta_k^2}$

- Consider the following algorithm.

UCB

Input = K arms, T rounds

(i) For $t = 1, \dots, K$, play arm t

(ii) For $t = K+1, \dots, T$, play arm $A_t = \underset{1 \le k \le K}{\operatorname{argmax}} \; U_{t-1}(k)$

Observations made up to time $t-1$.

We derive a bound on the expected regret for the UCB algorithm.

Lemma. Suppose that a sub-optimal arm k ($m_k < m_*$) has been played $n_t(k) \ge \frac{4 \ln t}{\Delta_k^2}$ times after t rounds.

Then $U_t(k) < U_t(k^*)$ w.p. $\ge 1 - \frac{4}{t^2}$.

Consequently,

$$\mathbb{P}\left(A_{t+1} = k \mid n_t(k) \ge \frac{4 \ln t}{\Delta_k^2}\right) \le \frac{4}{t^2}$$

proof =

$$\overline{X}_{k,t} + \sqrt{\frac{\ln t}{n_t(k)}} \leqslant \overline{X}_{k,t} + \frac{\Delta_k}{2} \quad \text{since } n_t(k) \geqslant \frac{4\ln t}{\Delta_k^2}$$

$$< \left(m_k + \frac{\Delta_k}{2}\right) + \frac{\Delta_k}{2} \quad \text{from (2) page 10}$$

$$= m_* \quad \text{by definition of } m_*/\Delta_k$$

$$< \overline{X}_{k^*,t} + \sqrt{\frac{\ln t}{n_t(k^*)}} \quad \text{from (1) page 10.}$$

Indeed, we obtain $u_t(k) < u_t(k^*)$ w.p. $\geqslant 1 - \frac{4}{t^2}$, using a union bound. ∎

Lemma Let $n_T(k)$ be the number of times the k-th arm is pulled after T rounds. Then, $\forall k$ s.t. $m_k < m_*$,

$$\mathbb{E}(n_T(k)) \leqslant \frac{4\ln T}{\Delta_k^2} + 8$$

proof:

$$\mathbb{E}(n_T(k)) = 1 + \mathbb{E}\left\{\sum_{t=K}^{T} \mathbb{1}(A_{t+1} = k)\right\}$$

$$= 1 + \mathbb{E}\left\{\sum_{t=K}^{T} \mathbb{1}\left(A_{t+1} = k, \, n_t(k) < \frac{4\ln t}{\Delta_k^2}\right)\right\}$$

$$+ \mathbb{E}\left\{\sum_{t=K}^{T} \mathbb{1}\left(A_{t+1} = k, \, n_t(k) \geqslant \frac{4\ln t}{\Delta_k^2}\right)\right\}$$

$$= \sum_{t=K}^{T} \mathbb{P}\left(A_{t+1} = k \,\middle|\, n_t(k) \geqslant \frac{4\ln t}{\Delta_k^2}\right) \mathbb{P}\left(n_t(k) \geqslant \frac{4\ln t}{\Delta_k^2}\right)$$

$$\leqslant \sum_{t=K}^{T} \frac{4}{t^2} \qquad \left(\leqslant \frac{4}{t^2} \text{ (previous lemma)}, \quad \leqslant 1\right)$$

$$\leqslant 8$$

For the other term, note that the number of times the indicator $\mathbb{1}\left(A_{t+1} = k, \, n_t(k) < \frac{4\ln t}{\Delta_k^2}\right)$ equals 1 must be less than $\frac{4\ln t}{\Delta_k^2} - 1$.

Thus $\mathbb{E}(n_T(k)) \leqslant \frac{4\ln t}{\Delta_k^2} + 8$. ∎

Theorem.
Let R_T denote the cumulative regret associated with a K-armed bandit problem, following the UCB algorithm. Then, $\forall T \geqslant N$, we have

$$\mathbb{E}(R_T) \leqslant \sum_{k \,|\, m_* > m_k} \left\{\frac{4\ln T}{\Delta_k} + 8\Delta_k\right\}$$

proof: It follows directly from the previous lemma and the expression $\mathbb{E}(R_T) = \sum_{k=1}^{K} \Delta_k \, \mathbb{E}(n_T(k))$ ∎

Bound is logarithmic in T
↳ This is optimal; recall Lai & Robbins's theorem

× **Remark**: The bound is called "instance-dependent", since it depends on Δ_k. We can obtain an "instance-independent" bound of the form $\mathbb{E}(R_T) \leqslant 8K + 5\sqrt{KT\log T}$.

↳ Indeed, let $K_1 = \left\{1 \leqslant k \leqslant K \,|\, \Delta_k < \sqrt{\frac{K}{T}\ln T}\right\}$

$$K_2 = \left\{1 \leqslant k \leqslant K \,|\, \Delta_k \geqslant \sqrt{\frac{K}{T}\ln T}\right\}.$$

Then

$$\sum_{k \in K_1} n_T(k)\Delta_k < \sqrt{\frac{K}{T}\ln T} \underbrace{\sum_{k \in K_1} n_T(k)}_{\leqslant T} \leqslant \sqrt{KT\ln T}$$

And

$$\sum_{k \in K_2} \mathbb{E}(n_T(k))\,\Delta_k \leq \sum_{k \in K_2} \left(\frac{4 \ln T}{\Delta_k^2} + 8 \right) \Delta_k$$

(lemma p. 15)

$$\leq \sum_{k \in K_2} \left(\frac{4 \ln T}{\Delta_k} + 8 \right) \qquad (\Delta_k \leq 1)$$

$$\leq 8K + \sum_{k \in K_2} 4 \ln T \sqrt{\frac{T}{K \ln T}}$$

$$\leq 8K + 4\sqrt{KT \ln T} \; .$$

II. 4. Contextual bandits

In the previous sections, the distribution of the reward is context-free : it depends solely on the arm pulled. In many interesting applications, we can make use of extra contextual information to decide on which arm to pull, in addition to previously observed rewards. For example, in personalized News article recommendation, selecting which article (≡ arm) to show to a particular user can greatly benefit from additional information such as the page visited, the user profile, the geo-location, the type of article, ... on top of the past rewards (click or no click).

We present a procedure called LinUCB for contextual bandits, proposed by Li et al (2010) A Contextual-bandit approach to Personalized News Article Recommendation. WWW 2010.

× Notation : At trial t, let $x_{t,k}$ be the context-vector summarizing information of both the user u_t and arm / article k. ($x_{t,k} \in \mathbb{R}^d$)

× LinUCB with disjoint linear models : Assume that the expected payoff of an arm k is linear in its d-dimensional feature vector $x_{t,k}$, with some unknown vector of coefficients θ_k^* :

$$\mathbb{E}(r_{t,k} \mid A_t = k, x_{t,k}) = x_{t,k}^T \theta_k^*$$

Compare this expression with the mean-reward of context-free bandits

$\mathbb{E}(r_{t,k} \mid A_t = k, \cancel{x_{t,k}}) = m_k$ (page 6)

Assume that $r_{t,k} \in \{0, 1\}$ = no click vs click (when a presented article is clicked, a reward of 1 is incurred ; otherwise the payoff is 0).

× let D_k ($m \times d$) $= \begin{pmatrix} \vdots \\ - \; x_{t,k}^T \; - \\ \vdots \end{pmatrix}$ } e.g. all m training inputs containing the m context vectors for arm / article k

y_k ($m \times 1$) $= \begin{pmatrix} \vdots \\ 0 \\ 1 \\ \vdots \end{pmatrix}$ } the corresponding click / no click feedback

The disjoint linear model above implies that

$$\mathbb{E}\, y_k = D_k \theta_k^* \; .$$

$\Rightarrow (D_k, y_k)$ = our training data for arm k.

↳ Ridge estimate of θ_k^* is $\hat{\theta}_k = (D_k^t D_k + I_d)^{-1} D_k^t y_k$

We show below that w.p. $\geq 1-\delta$,

$$\left| x_{t,k}^t \hat{\theta}_k - x_{t,k}^t \theta_k^* \right| \leq \alpha \sqrt{x_{t,k}^t (D_k^t D_k + I_d)^{-1} x_{t,k}},$$

where $\alpha = 1 + \sqrt{\log(2/\delta)/2}$, $\forall \delta > 0$. (*)

↑ Upper Confidence Bound (UCB) for the expected reward of arm k.
⇒ Use this UCB to select which arm to pull.

Pull arm $A_t = j$, where

$$j = \underset{1 \leq k \leq K}{\text{argmax}} \left\{ x_{t,k}^t \hat{\theta}_k + \alpha \sqrt{x_{t,k}^t A_k^{-1} x_{t,k}} \right\},$$

with $A_k := D_k^t D_k + I_d$.

We summarize the procedure below — **Lin UCB**

For $t = 1, 2, \dots$

- Observe features $x_{t,k}$ for all arms $1 \leq k \leq K$.
- For all $k \in \{1, \dots, K\}$
  - ↘ If k is new
    $A_k \leftarrow I_d$
    $b_k \leftarrow 0_{(d \times 1)}$
  - ↘ $\hat{\theta}_k \leftarrow A_k^{-1} b_k$
    $p_{t,k} \leftarrow x_{t,k}^t \hat{\theta}_k + \alpha \sqrt{x_{t,k}^t A_k^{-1} x_{t,k}}$

- Choose an arm $j = \underset{1 \leq k \leq K}{\text{argmax}}\ p_{t,k}$ & observe reward r_t
- $A_k \leftarrow A_k + x_{t,k} x_{t,k}^t$
  $b_k \leftarrow b_k + r_t x_{t,k}$

- Proof of (*)

$$x_{t,k}^t \hat{\theta}_k - x_{t,k}^t \theta_k^*$$

$$= x_{t,k}^t \underbrace{A_k^{-1} D_k^t y_k}_{= \hat{\theta}_k} - x_{t,k}^t \underbrace{A_k^{-1} (D_k^t D_k + I_d)}_{= I_d} \theta_k^*$$

$$= x_{t,k}^t A_k^{-1} D_k^t (y_k - D_k \theta_k^*) - x_{t,k}^t A_k^{-1} \theta_k^*$$

$$\Rightarrow \left| x_{t,k}^t \hat{\theta}_k - x_{t,k}^t \theta_k^* \right| \leq \left| x_{t,k}^t A_k^{-1} D_k^t (y_k - D_k \theta_k^*) \right| + \underbrace{\| x_{t,k}^t A_k^{-1} \theta_k^* \|}_{\leq \| x_{t,k}^t A_k^{-1} \|, \text{ assuming } \| \theta_k^* \| \leq 1}.$$

The second term is

$$\| x_{t,k}^t A_k^{-1} \| = \sqrt{x_{t,k}^t A_k^{-1} A_k^{-1} x_{t,k}}$$

$$\leq \sqrt{x_{t,k}^t A_k^{-1} (I_d + D_k^t D_k) A_k^{-1} x_{t,k}}$$

$$= \sqrt{x_{t,k}^t A_k^{-1} x_{t,k}} =: s_{t,k}$$

Note that

$$\begin{aligned} s_{t,k}^2 &= x_{t,k}^t A_k^{-1} x_{t,k} \\ &= x_{t,k}^t A_k^{-1} (I + D_k^t D_k) A_k^{-1} x_{t,k} \\ &\geq x_{t,k}^t A_k^{-1} D_k^t D_k A_k^{-1} x_{t,k} \\ &= \| D_k A_k^{-1} x_{t,k} \|. \end{aligned}$$

We turn our attention to the term $|x_{t,k}^t A_k^{-1} D_k^t (y_k - D_k \theta_k^*)|$.

$$|x_{t,k}^t A_k^{-1} D_k^t y_k - x_{t,k}^t A_k^{-1} D_k^t D_k \theta_k^*|$$

$x_{t,k}^t A_k^{-1} D_k^t y_k$ can be rewritten $\alpha_{t,k}^t y_k$, with $\alpha_{t,k} := x_{t,k}^t A_k^{-1} D_k^t \in \mathbb{R}^m$.

$x_{t,k}^t A_k^{-1} D_k^t D_k \theta_k^* = \mathbb{E}(x_{t,k}^t A_k^{-1} D_k^t y_k)$ since $\mathbb{E} y_k = D_k \theta_k^*$, see page 18.

Assuming clicks are independent given the context vector, and since each entry of $y_k \in [0,1]$, each weighted entry contributing to the sum $\alpha_{t,k}^t y_k$ is bounded by the corresponding entry in $\alpha_{t,k}$,

$$\mathbb{P}(0 \leq (\alpha_{t,k})_i (y_k)_i \leq (\alpha_{t,k})_i) = 1$$

(i-th entry; $= l_i$)

and $\sum_{i=1}^m l_i^2 = \| \alpha_{t,k} \|^2 = \| x_{t,k}^t A_k^{-1} D_k^t \|^2$

Applying Azuma inequality yields

$$\mathbb{P}(|x_{t,k}^t A_k^{-1} D_k^t (y_k - D_k \theta_k^*)| > \alpha s_{t,k}) \leq 2 e^{-\frac{2\alpha^2 s_{t,k}^2}{\| x_{t,k}^t A_k^{-1} D_k^t \|}}$$

$\mathbb{P}(\dots > \alpha s_{t,k}) \leq 2 e^{-2\alpha^2}$, since $s_{t,k}^2 \geq \| D_k A_k^{-1} x_{t,k} \|$,

which concludes the proof. ∎

II.5. Adversarial bandits.

In sections I.1 to I.4, we assumed that each arm is associated with a reward distribution.
We consider here a different setup:

for each arm k, there is a sequence of rewards $r_{k,1}, \dots, r_{k,T}$ pre-determined in advance.

↑ However, the reward received is not selected by the environment after an action is performed; the sequence of rewards is fixed in advance → "oblivious adversary".

We compare our strategy with the best strategy that always perform the same action / pick the same arm.
The total regret is

$$R_T := \max_{1 \leq k \leq K} \sum_{t=1}^T r_{k,t} - \sum_{t=1}^T r_{A_t,t}.$$

↑ The performance of our strategy will be evaluated using $\mathbb{E}(R_T)$.

where the expected value is with respect to the randomness in our strategy.

- Consider first a simplified setting, where after an action is performed, we observe the rewards of all other arms.
→ we will relax this assumption later.

This set up is similar as learning with expert advice (section I)

⇒ Consider an algorithm picking arm k (at time t) with probability $p_{k,t}$, where

$$p_{k,t} = \frac{w_{t-1}^k}{\sum_{j=1}^K w_{t-1}^j},$$

and weights are updated according to $w_t^k = w_{t-1}^k \exp(\beta r_{k,t})$, $1 \le k \le K$, after observing all rewards $r_{1,t}, \dots, r_{K,t}$. (For some $\beta > 0$)

(initialize $w_0^1 = \dots = w_0^K = 1$)

aka the **HEDGE ALGORITHM**

Arm with the largest reward at time t gets a largest positive update for the next round.

× Remark: Similar to the problem of prediction with expert advice, with $r_{k,t} = -\ell(f_t^k, y_t)$

Theorem. Assuming $r_{k,t} \in [0,1]$, taking $\beta = \sqrt{\frac{\log K}{T}} < 1$, we have $\mathbb{E}(R_T) \le 3\sqrt{T \log K}$

proof: We proceed as on pages 2/3/4.

Put. $W_0 := K$ and $W_t := \sum_{k=1}^K w_t^k$

. $r_t := \sum_{k=1}^K p_{k,t} r_{k,t}$.

. First, note that $W_t = \sum_{k=1}^K w_{t-1}^k e^{\beta r_{k,t}}$, so that

$$\begin{aligned}
\frac{W_t}{W_{t-1}} &= \sum_{k=1}^K p_{k,t} e^{\beta r_{k,t}} \\
&\le \sum_k p_{k,t}\left(1 + \beta r_{k,t} + \beta^2 r_{k,t}^2\right) \quad \text{since } \beta r_{k,t} \le 1 \\
&= 1 + \beta \sum_k p_{k,t} r_{k,t} + \beta^2 \sum_k p_{k,t} r_{k,t}^2 \\
&\le 1 + \beta \sum_k p_{k,t} r_{k,t} + \beta^2 \sum_k p_{k,t} r_{k,t} \quad r_{k,t} \le 1 \\
&= 1 + \beta r_t + \beta^2 r_t \\
&\le \exp\{(\beta + \beta^2) r_t\}.
\end{aligned}$$

. Next, observe that $\frac{W_T}{W_0} \le \exp\left\{(\beta+\beta^2)\sum_{t=1}^T r_t\right\}$,

so that

$$\log\left(\frac{W_T}{W_0}\right) \le (\beta+\beta^2)\sum_{t=1}^T r_t = \mathbb{E}\left(\sum_{t=1}^T r_{A_t,t}\right)$$

= expected reward obtained following the Hedge algorithm.

. Lower bound for W_T:

$$\begin{aligned}
W_T &= \sum_{k=1}^K w_T^k \\
&= \sum_k \exp\{\beta(r_{k,1} + \dots + r_{k,T})\} \\
&\ge \max_{1\le k\le K} \exp\{\beta(r_{k,1} + \dots + r_{k,T})\}
\end{aligned}$$

$$\Rightarrow \log W_T \ge \beta \max_{1\le k\le K}\left(\sum_{t=1}^T r_{k,t}\right)$$

- Thus, $\beta \max_k \left(\sum_{t=1}^T r_{k,t} \right) - \log K \leq (\beta + \beta^2) \sum_{t=1}^T r_t$

$$\max_k \left(\sum_{t=1}^T r_{k,t} \right) - \frac{\log K}{\beta} \leq (1+\beta) \sum_{t=1}^T r_t .$$

$$\frac{1}{1+\beta} \left\{ \max_k \left(\sum r_{k,t} \right) - \frac{\log K}{\beta} \right\} \leq \sum_{t=1}^T r_t$$

$$(1-2\beta) \left\{ \max_k \left(\sum r_{k,t} \right) - \frac{\log K}{\beta} \right\} \leq \sum_{t=1}^T r_t .$$

$1-2\beta \leq \frac{1}{1+\beta}$ for $\beta \in [0,1]$

$$\Rightarrow \boxed{\mathbb{E}(R_T) \leq 2\beta \max_k \left(\sum_{t=1}^T r_{k,t} \right) + \underbrace{\frac{\log K}{\beta}}_{\leq T}}$$

+ Selecting $\beta = \sqrt{\frac{\log K}{T}}$ yields the desired bound ∎

- We consider now the more realistic scenario where once we pull arm k, at time t, only the reward $r_{k,t}$ is observed. (partial information)

↳ At each time step, consider the estimate

$$\hat{r}_{k,t} = \begin{cases} r_{k,t} / p_{k,t} & \text{if } A_t = k \\ 0 & \text{otherwise.} \end{cases}$$

Then we see that $\mathbb{E}(\hat{r}_{k,t}) = r_{k,t}$

$\hat{r}_{k,t}$ is an unbiased estimate $\forall k, \forall t$.

↳ The algorithm EXP3, introduced by Auer, Cesa-Bianchi, Freund, Schapire (1995) uses these estimates as substitutes for the rewards used in the Hedge algorithm.

EXP3 Algorithm.

- Input: $\beta \in [0,1]$, $\gamma \in [0,1]$

$\omega_{1,0} = \dots = \omega_{1,K} = 1$.

- For $t \geq 1$, do

  - Play arm k with probability

$$p_{k,t} = (1-\gamma) \frac{\omega_{t-1}^k}{\sum_j \omega_{t-1}^j} + \frac{\gamma}{K}$$

  - Observe reward $r_{k,t}$
  - Update

$$\omega_t^k = \omega_{t-1}^k e^{\beta \hat{r}_{k,t}} \quad ; \quad \hat{r}_{k,t} = \begin{cases} \frac{r_{k,t}}{p_{k,t}} & \text{if } A_t = k \\ 0 & \text{o/w.} \end{cases}$$

Theorem: Assuming $r_{k,t} \in [0,1]$, then EXP3 achieves a regret bound

$$\mathbb{E}(R_T) \leq 3\sqrt{KT \log T} \text{ , with } \beta = \sqrt{\frac{\log K}{KT}}$$

proof = Put $W_t := \sum_{k=1}^K \omega_t^k$.

- Take $\gamma = K\beta \in [0,1]$

Since $p_{k,t} \geq \frac{\gamma}{K} = \beta$, we see that $\hat{r}_{k,t} \in [0, \frac{1}{\beta}]$.

Then

$$\frac{W_t}{W_{t-1}} = \frac{\sum_k \omega_{t-1}^k e^{\beta \hat{r}_{k,t}}}{W_{t-1}} = \sum_{k=1}^K \frac{p_{k,t} - \beta}{1 - K\beta} e^{\beta \hat{r}_{k,t}}$$

$\beta = \frac{\gamma}{K}$

Thus

$$\frac{W_t}{W_{t-1}} \leq \sum_{k=1}^{K} \left\{ \frac{p_{k,t} - \beta}{1 - K\beta} \right\} \left(1 + \beta \hat{r}_{k,t} + \beta^2 \hat{r}_{k,t} \right) \quad \text{since } \beta \hat{r}_{k,t} \leq 1$$

$$= 1 + \sum_k \left\{ \frac{p_{k,t} - \beta}{1 - K\beta} \right\} \left(\beta \hat{r}_{k,t} + \beta^2 \hat{r}_{k,t} \right)$$

$$\leq 1 + \beta \sum_k \left(\frac{p_{k,t} \hat{r}_{k,t}}{1 - K\beta} \right) + \beta^2 \sum_k \left(\frac{p_{k,t} \hat{r}_{k,t}^2}{1 - K\beta} \right).$$

$$= 1 + \frac{\beta}{1 - K\beta} r_{A_t,t} + \frac{\beta^2}{1 - K\beta} r_{A_t,t} \hat{r}_{A_t,t}$$

≤ 1

Since $p_{k,t} \hat{r}_{k,t} = \begin{cases} r_{k,t} & \text{if } A_t = k \\ 0 & \text{otherwise} \end{cases}$

$$\leq 1 + \frac{\beta}{1 - K\beta} r_{A_t,t} + \frac{\beta^2}{1 - K\beta} \hat{r}_{A_t,t}$$

$$\frac{W_t}{W_{t-1}} \leq \exp\left\{ \frac{1}{1 - K\beta} \left(\beta r_{A_t,t} + \beta^2 \hat{r}_{A_t,t} \right) \right\}$$

- Cascading this inequality, we obtain

$$\log\left(\frac{W_T}{W_0} \right) \leq \frac{1}{1 - K\beta} \left(\beta \sum_{t=1}^{T} r_{A_t,t} + \beta^2 \sum_{t=1}^{T} \hat{r}_{A_t,t} \right)$$

Cumulative reward of EXP3 strategy.

- Lower bound for W_T:

$$W_T = \sum_{k=1}^{K} \exp\left(\sum_{t=1}^{T} \beta \hat{r}_{k,t} \right) \geq \exp\left(\sum_{t=1}^{T} \beta \hat{r}_{j,t} \right) \quad \forall j \in \{1, \dots, K\}$$

Thus

$$\log\left(\frac{W_T}{W_0} \right) \geq \beta \left(\sum_{t=1}^{T} \hat{r}_{j,t} \right) - \log K \,, \quad \forall j \in \{1, \dots, K\}.$$

- Putting terms together:

$$\beta \left(\sum_{t=1}^{T} \hat{r}_{j,t} \right) - \log K \leq \frac{1}{1 - K\beta} \left(\beta \sum_{t=1}^{T} r_{A_t,t} + \beta^2 \sum_{t=1}^{T} \hat{r}_{A_t,t} \right).$$

Taking $\mathbb{E}(\dots)$

$$\beta \left(\sum_{t=1}^{T} r_{j,t} \right) - \log K \leq \frac{1}{1 - K\beta} \left(\beta \, \mathbb{E} \sum_{t=1}^{T} r_{A_t,t} + \beta^2 \sum_{k,t} r_{k,t} \right)$$

Since $\mathbb{E}\, \hat{r}_{j,t} = r_{j,t} \;\; \forall j$

Since $\mathbb{E}\, \hat{r}_{A_t,t} = \sum_k \frac{r_{k,t}}{p_{k,t}} \mathbb{P}(A_t = k) = \sum_k r_{k,t}$

True $\forall j \in \{1, \dots, K\}$. Thus

$$\beta \left(\max_{1 \leq j \leq K} \sum_{t=1}^{T} r_{j,t} \right) - \log K \leq \frac{1}{1 - K\beta} (\text{---} " \text{---})$$

Best strategy performing always the same action.

$$\Rightarrow \max_{1 \leq j \leq K} \sum_{t=1}^{T} r_{j,t} - \frac{\log K}{\beta} \leq \frac{1}{1 - K\beta} \left(\mathbb{E} \sum_{t=1}^{T} r_{A_t,t} + \beta \sum_{k,t} r_{k,t} \right)$$

Rearranging the terms yields

$$\underbrace{\max_{1 \leq j \leq K} \sum_{t=1}^{T} r_{j,t} - \mathbb{E} \sum_{t=1}^{T} r_{A_t,t}}_{= \mathbb{E}(R_T)} \leq 2K\beta \underbrace{\max_j \sum_{t=1}^{T} r_{j,t}}_{\leq T} + \frac{\log K}{\beta}$$

Plugging in the value $\beta = \sqrt{\frac{\log K}{KT}}$ gives the desired bound. ∎

III. APPLICATION : DISPLAY ADVERTISING

We present an application of Thompson sampling & LinUCB to online advertising.

[REF] O. Chapelle & L. Li (Yahoo)
An Empirical Evaluation of Thompson Sampling.

Given a user visiting a webpage, the task is to select the best advertisement for that user:
What is the probability that an ad will be clicked given some context (user, page visited, ...).

× Bayesian setting + regularized logistic regression for predicting this probability.
↑ corresponding to the Click-Through Rate (CTR)

• LR model : $\mathbb{P}(Y_i = y_i \mid X_i = x_i, \beta) \in \mathbb{R}^d$

$= \sigma_i^{y_i} (1-\sigma_i)^{1-y_i}$

where $\sigma_i = \sigma(\beta^t x_i)$

σ = sigmoïd

$y_i \in \{0, 1\}$
click (1) or no click (0)

features representing the user, page, ad, ...
$x_i \in \mathbb{R}^d$

β = model parameter to be learned.
A feature vector x_i is constructed for every (context, ad) pair & the policy decides which ad to show.

The likelihood associated with a sample $(x_1, y_1), \dots, (x_n, y_n)$ of size n is

$$\prod_{i=1}^n \sigma_i^{y_i} (1-\sigma_i)^{1-y_i}$$

↳ maximizing the likelihood is equivalent to minimizing

$$\sum_{i=1}^n \log\left(1 + e^{-y_i' \beta^t x_i}\right),$$

where $y_i' = 2y_i - 1 \in \{-1, 1\}$,
see p.42 in SL : FOUNDATIONS

• Prior on β

Assume a Gaussian prior with diagonal covariance matrix :

$$\beta_j \sim \mathcal{N}(m_j, q_j^{-1}), \quad 1 \le j \le d, \quad \beta = (\beta_1, \dots, \beta_d)^t$$

• Posterior distribution

$$\sum_{i=1}^n \log\left(1 + e^{-y_i' \beta^t x_i}\right) + \frac{1}{2} \sum_{i=1}^d q_i (\beta_i - m_i)^2$$

↑ • "minus" the posterior
• up to some additive constants

↳ The MAP estimate of β minimizes the above expression.
The posterior distribution is then approximated using a Gaussian density (aka Laplace Approximation).
⇒ We need to calculate the curvature of the posterior.

$$\frac{\partial \text{ posterior}}{\partial \beta_j} = q_j (\beta_j - m_j) + \sum_{i=1}^n \frac{-y_i' x_{ij} e^{-y_i' \beta^t x_i}}{1 + e^{-y_i' \beta^t x_i}}$$

$$\frac{\partial^2 \text{posterior}}{\partial \beta_j^2} = q_j + \sum_{i=1}^{n} x_{ij}^2 \, p_i(1-p_i) \; ; \; p_i = \frac{1}{1+e^{y_i \beta^t x_i}}$$

After calculations, using $(y_i')^2 = 1$

= variance^{-1} of the laplace approx. distribution.

The mean of the laplace approx distribution is given by the mode of the posterior: $m_j = \beta_j$, $1 \le j \le d$.

- Algorithm: Regularized LR for online advertising
- Init: $m_j = 0$, $q_j =$ some $\lambda > 0$
  Each weight $w_j \sim \mathcal{N}(m_j, q_j^{-1})$
- For $t = 1, 2, \dots$
  - ↘ Get a new batch of training data (x_i, y_i), $i = 1, \dots, n$
  - ↘ Find the minimizer β of
    $$\sum_{i=1}^{n} \log\left(1 + e^{-y_i \beta^t x_i}\right) + \frac{1}{2} \sum_{i=1}^{d} q_i (\beta_i - m_i)^2$$
  - ↘ Update the parameters:
    - $m_j = \beta_j$
    - $q_j = q_j + \sum_{i=1}^{n} x_{ij}^2 \, p_i(1-p_i)$.

↑ It remains to decide on which ads to show to a given user; i.e. on a procedure to collect the training data ⇒ explore / exploit trade off.

(a) Thompson Sampling:
Each weight β_j is sampled according to its Gaussian approximation $\mathcal{N}(m_j, q_j^{-1})$.

(b) Lin UCB
Select the ad for which $\sum_{j=1}^{d} m_j x_{ij} + \alpha \sqrt{\sum_{j=1}^{d} q_j^{-1} x_{ij}^2}$ is maximum.

↑ Compare with the expression p. 19

(c) Exploit-Only: Select the ad with the highest mean.

(d) ε-greedy: Select a random ad w.p. ε and the ad with the highest mean w.p. $1-\varepsilon$.

→ Simulation study can be found in the referenced paper.