HW5

Kevin Chang

November 26, 2025

1

Consider a medical decision making setting with a population of patients all suffering from the same health condition. The condition, left untreated, causes each individual patient i cost $c_i(0)$. A drug for the condition is available, but has known side effects quantified by cost $c_i(1)$, and unknown and probabilistic individual treatment effect quantified by $\tau_i \in \mathbb{R}$, drawn from some unknown distribution with mean τ. If patient i is left untreated, they suffer cost $c_i(0)$; on the other hand, if they are treated with the drug, they suffer cost $ci(1) - \tau_i$.

(i) Suppose only the population average treatment effect $\tau := \mathbb{E}[\tau_i]$ is known. What is the treatment decision rule that minimizes expected cost for a new patient j?

For a new patient j, the cost when untreated is $c_j(0)$ (deterministic). If treated, the cost is $c_j(1) - \tau_j$, whose expectation is

$$\mathbb{E}[c_j(1) - \tau_j] = c_j(1) - \tau.$$

To minimize expected cost, we treat whenever

$$\tau > c_j(1) - c_j(0).$$

Decision rule: Treat patient j if $\tau > c_j(1) - c_j(0)$, and do not treat otherwise.

(ii) Suppose that members of the population are parametrized by features $X \in \{-1, 1\}^d$, and in addition to the population average treatment effect τ, the conditional average treatment effects $\tau_x = \mathbb{E}[\tau_i | X_i = x]$ for all x are known. What is the decision rule that minimizes expected cost for a new person j with feature vector x_j?

For a new patient j with feature vector x_j, the cost when untreated is

$$C_{\text{untreated}} = c_j(0).$$

If treated, the cost is $c_j(1) - \tau_j$, and using the conditional average treatment effect

$$\tau_{x_j} := \mathbb{E}[\tau_i \mid X_i = x_j],$$

the expected treated cost is

$$\mathbb{E}[C_{\text{treated}} \mid X_j = x_j] = c_j(1) - \tau_{x_j}.$$

To minimize expected cost, we treat whenever

$$\tau_{x_j} > c_j(1) - c_j(0).$$

Decision rule: For a patient j with feature vector x_j, treat if $\tau_{x_j} > c_j(1) - c_j(0)$, and do not treat otherwise.

(iii) Suppose instead that we make a prediction for individual j in the form of an interval $(\hat{\tau}_j^{\min}, \hat{\tau}_j^{\max}) \subset \mathbb{R}$, with the guarantee that

$$\mathbb{P}[\tau_j \notin (\tau_j^{\min}, \tau_j^{\max})] \leq \epsilon.$$

Is it possible to design a treatment decision rule for patient j that minimizes the expected cost for j? If so, give such a rule. If not, explain why and design a rule that achieves a different notion of optimality.

We are given an interval prediction $(\hat{\tau}_j^{\min}, \hat{\tau}_j^{\max})$ with coverage guarantee

$$\mathbb{P}\big[\tau_j \notin (\hat{\tau}_j^{\min}, \hat{\tau}_j^{\max})\big] \leq \epsilon.$$

The untreated cost is $c_j(0)$, while the treated cost is $c_j(1) - \tau_j$.

The interval guarantee does *not* identify the expected value $\mathbb{E}[\tau_j]$, since many distributions of τ_j are compatible with the same interval. Hence no rule can uniquely minimize the *expected* cost for all distributions consistent with the interval.

A natural alternative is a *robust* (minimax) rule that minimizes the worst-case treated cost over all τ_j in the interval. The worst-case treated cost is

$$\sup_{\tau_j \in [\hat{\tau}_j^{\min}, \hat{\tau}_j^{\max}]} (c_j(1) - \tau_j) = c_j(1) - \hat{\tau}_j^{\min}.$$

Robust decision rule: Treat patient j if

$$c_j(1) - \hat{\tau}_j^{\min} < c_j(0) \quad \Longleftrightarrow \quad \hat{\tau}_j^{\min} > c_j(1) - c_j(0),$$

and do not treat otherwise.

(iv) In words, describe the meaning of optimality for the following scenarios. That is, state the source(s) of randomness, and the degree to which a guarantee holds for a particular person j.

(a) Scenario described in Part 1.

The only randomness comes from the unknown treatment effect τ_j, of which only the population mean τ is known. Optimality means minimizing the *expected* cost for patient j, where the expectation is taken over the population distribution of τ_i.

(b) Scenario described in Part 2.

Randomness again arises from τ_j, but we now know the conditional mean τ_{x_j} given the patient's features. Optimality means minimizing the *expected* cost conditional on $X_j = x_j$, i.e., on average over all patients with the same feature vector.

(c) Scenario described in Part 3, where the marginal distribution of τ_i over the entire population is known, and the prediction interval $(\hat{\tau}_j^{\min}, \hat{\tau}_j^{\max}))$ is constructed with this information about the distribution.

The interval $(\hat{\tau}_j^{\min}, \hat{\tau}_j^{\max})$ is constructed so that it contains τ_j with probability at least $1-\epsilon$ under the known marginal distribution of treatment effects. However, the interval does not identify the mean of τ_j, so expected-cost minimization is not well-defined. Optimality instead refers to minimizing the *worst-case* cost over all τ_j in the high-probability interval.

(d) Scenario described in Part 3, where the prediction interval $(\hat{\tau}_j^{\min}, \hat{\tau}_j^{\max})$ is constructed by post-processing a point predictor for τ_j with the standard split conformal prediction procedure.

The interval is constructed via split conformal prediction, which guarantees marginal coverage

$$\mathbb{P}(\tau_j \in (\hat{\tau}_j^{\min}, \hat{\tau}_j^{\max})) \geq 1 - \epsilon.$$

As in (c), the mean is not identified, so optimality is defined in a *minimax* sense–minimizing the worst-case cost over all τ_j in the conformal prediction interval.

2

Consider a randomized trial with n units. Let $a_i \in \{0,1\}$ denote the action taken on unit $i \in [n]$. Let $y_i(j) \in \{0,1\}$ denote the potential outcome for unit i when the action is j. Recall that the average treatment effect (ATE) is the quantity

$$ATE = \frac{1}{n}\sum_{i=1}^{n} y_i(1) - y_i(0)$$

1. Suppose n is even. Assume that half of the units are selected uniformly at random and assigned action 1, and the other half are assigned action 0. Prove that the Horvitz–Thompson estimator,

$$A\hat{T}E = \frac{2}{n}\sum_{\{i:a_i=1\}} y_i(1) - \frac{2}{n}\sum_{\{j:a_j=0\}} y_j(0),$$

is unbiased. What is its variance?

Unbiasedness. Write the estimator using indicators:

$$A\hat{T}E = \frac{2}{n}\sum_{i=1}^{n} \mathbb{I}\{a_i = 1\} y_i(1) - \frac{2}{n}\sum_{i=1}^{n} \mathbb{I}\{a_i = 0\} y_i(0).$$

Under complete randomization with $n/2$ treated and $n/2$ control units, each unit is equally likely to be treated, so

$$\mathbb{P}(a_i = 1) = \mathbb{P}(a_i = 0) = \frac{1}{2} \quad \Rightarrow \quad \mathbb{E}[\mathbb{I}\{a_i = 1\}] = \mathbb{E}[\mathbb{I}\{a_i = 0\}] = \frac{1}{2}.$$

Thus

$$\mathbb{E}[A\hat{T}E] = \frac{2}{n}\sum_{i=1}^{n} \mathbb{E}[\mathbb{I}\{a_i = 1\}]\, y_i(1) - \frac{2}{n}\sum_{i=1}^{n} \mathbb{E}[\mathbb{I}\{a_i = 0\}]\, y_i(0) = \frac{1}{n}\sum_{i=1}^{n} y_i(1) - \frac{1}{n}\sum_{i=1}^{n} y_i(0) = ATE,$$

Variance.

Rewrite the estimator as

$$A\hat{T}E = -\frac{2}{n}\sum_{i=1}^{n} y_i(0) + \frac{2}{n}\sum_{i=1}^{n} Z_i\big(y_i(1) + y_i(0)\big),$$

where $Z_i = \mathbb{I}\{a_i = 1\}$. Let

$$w_i := y_i(1) + y_i(0), \qquad \bar{w} := \frac{1}{n}\sum_{i=1}^{n} w_i,$$

and denote the finite-population variance

$$S_w^2 = \frac{1}{n-1}\sum_{i=1}^{n} (w_i - \bar{w})^2.$$

Since exactly $n/2$ units are sampled without replacement to form the treated group, the random sum

$$\sum_{i=1}^{n} Z_i w_i$$

is the sample total of a simple random sample of size $m = n/2$ without replacement from $\{w_i\}$.

A standard finite-population result gives

$$\operatorname{Var}\left(\sum_{i=1}^{n} Z_i w_i\right) = m(1 - m/n)\, S_w^2 = \frac{n}{2}\left(1 - \frac{1}{2}\right) S_w^2 = \frac{n}{4} S_w^2.$$

Therefore

$$\mathrm{Var}(A\hat{T}E) = \left(\frac{2}{n}\right)^2 \mathrm{Var}\left(\sum_{i=1}^{n} Z_i w_i\right) = \frac{4}{n^2} \cdot \frac{n}{4} S_w^2 = \frac{S_w^2}{n}.$$

Thus the variance of the Horvitz–Thompson estimator is

$$\boxed{\mathrm{Var}(A\hat{T}E) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \big((y_i(1) + y_i(0)) - (\bar{y}(1) + \bar{y}(0))\big)^2}$$

where

$$\bar{y}(1) = \frac{1}{n}\sum_{i=1}^{n} y_i(1), \qquad \bar{y}(0) = \frac{1}{n}\sum_{i=1}^{n} y_i(0).$$

2. Now consider a different estimator of the ATE. Suppose m_1 of the units are assigned to action 1 and m_0 to action zero. For $j \in \{0,1\}$, define

$$\hat{\mu}_j := \frac{1}{m_j} \sum_{i:a_i=j} y_i(j)$$

The $z-score$ is the quantity

$$z = \frac{\hat{\mu}_1 - \hat{\mu}_0}{std(y)\sqrt{\frac{1}{m_0} + \frac{1}{m_1}}}$$

Under the $z-test$, a result is statistically significant if z has magnitude larger than 1.96. Show that no matter how a is chosen, z is proportional to the correlation coefficient of a and y. That is, prove

$$z = \sqrt{n}R(a,y) = \frac{\sqrt{n}cov(a,y)}{std(a)std(y)}.$$

3

Consider a k-armed bandit problem with $[0,1]$-bounded rewards, which means $X^(t) \in [0,1]^k$ for all rounds t. Let $\mu_i = \mathbb{E}[X_i^{(1)}]$ be the expected reward from arm $i \in [k]$ and let $\mu_* = \max_{i\in[k]} \mu_i$. The ϵ-greedy algorithm takes as input an exploration schedule $\{\epsilon_t\}$, where $\epsilon_t \in [0,1]$ for all t, and works as follows. On the first k rounds, the algorithm plays each arm once. On each round $t > k$, with probability ϵ_t , the algorithm plays $i_t \sim \mathit{Unif}(1,\ldots,k)$, and with probability $1-\epsilon_t$, it plays

$$i_t = \arg\max_{i\in[k]} \frac{1}{\sum_{s<t} \mathbb{I}(i_s = i)} \sum_{s<t} X_i^{(s)} \mathbb{I}(i_s = i).$$

(i) Suppose $\epsilon_t = \epsilon \in (0,1)$ for all t. Show that

$$\lim_{t\to\infty} \frac{1}{t}\left(t\mu_* - \mathbb{E}\left[\sum_{s\le t} X_{i_s}^{(s)}\right]\right) = \frac{\epsilon}{k}\sum_{i\in[k]} (\mu_* - \mu_i).$$

What does this result imply about the expected regret of the ϵ-greedy algorithm with any constant exploration schedule?

On each round $s > k$, the ϵ-greedy algorithm explores with probability ϵ and exploits with probability $1-\epsilon$. By the law of large numbers, the empirical means converge almost surely to their true means, so on exploitation rounds the optimal arm is selected with probability tending to 1. Thus, for each arm $i \in [k]$,

$$\Pr(i_s = i) \longrightarrow \begin{cases} (1-\epsilon) + \epsilon/k, & i \text{ optimal}, \\ \epsilon/k, & i \text{ suboptimal}. \end{cases}$$

Therefore the expected reward per round converges to

$$r_\infty = (1-\epsilon)\mu_* + \epsilon \cdot \frac{1}{k}\sum_{i=1}^{k} \mu_i.$$

Since the rewards lie in $[0,1]$, the sequence $\{r_s\}$ is bounded, and a Cesàro convergence argument gives

$$\lim_{t\to\infty} \frac{1}{t}\mathbb{E}\left[\sum_{s\le t} X_{i_s}^{(s)}\right] = \lim_{t\to\infty} \frac{1}{t}\sum_{s\le t} r_s = r_\infty.$$

Hence

$$\begin{aligned}\lim_{t\to\infty} \frac{1}{t}\left(t\mu_* - \mathbb{E}\left[\sum_{s\le t} X_{i_s}^{(s)}\right]\right) &= \mu_* - r_\infty \\ &= \mu_* - \left((1-\epsilon)\mu_* + \epsilon\cdot\frac{1}{k}\sum_{i=1}^{k}\mu_i\right) \\ &= \frac{\epsilon}{k}\sum_{i=1}^{k}(\mu_* - \mu_i).\end{aligned}$$

Thus,

$$\boxed{\lim_{t\to\infty} \frac{1}{t}\left(t\mu_* - \mathbb{E}\left[\sum_{s\le t} X_{i_s}^{(s)}\right]\right) = \frac{\epsilon}{k}\sum_{i\in[k]}(\mu_* - \mu_i)}.$$

Implication for regret. The limit above is the asymptotic average regret per round. Since it is a positive constant (unless all arms have the same mean), we have

$$R(t) = \Theta(t).$$

Therefore, ϵ-greedy with any constant $\epsilon \in (0,1)$ incurs *linear regret* and is not a no-regret algorithm.

(ii) Suppose $\epsilon_t = \left(\frac{Klog(t)}{t}\right)^{1/3}$. Show that there is a universal constant $C > 0$ such that

$$t\mu_* - \mathbb{E}\left[\sum_{s\le t} X_{t_s}^{(s)}\right] \le C(K\log{(t)})^{1/3}\cdot t^{2/3}$$

for all t. What does this result imply about the expected regret of the ϵ-greedy algorithm with a decaying exploration schedule? Compare and contrast this result with the result of the previous part.

The (pseudo-)regret up to time t is

$$R(t) = t\mu_* - \mathbb{E}\Big[\sum_{s\le t} X_{i_s}^{(s)}\Big].$$

Each time a suboptimal arm is played, the instantaneous regret is at most 1 since rewards are in $[0,1]$. Thus, the total regret is at most the expected number of pulls of suboptimal arms:

$$R(t) \;\le\; \mathbb{E}\Big[\#\{\text{suboptimal arm pulls up to time } t\}\Big].$$

At round s, the algorithm explores with probability ϵ_s and chooses an arm uniformly at random; during exploration, each arm is chosen with probability ϵ_s/k. Ignoring for the moment any suboptimal choices during exploitation (which can be shown to contribute at most a comparable lower-order term), we have

$$\mathbb{E}\Big[\#\{\text{suboptimal arm pulls up to time } t\}\Big] \;\lesssim\; \sum_{s=1}^{t}\epsilon_s.$$

With the given schedule $\epsilon_s = \big(K\log(s)/s\big)^{1/3}$,

$$\sum_{s=1}^{t}\epsilon_s = K^{1/3}\sum_{s=1}^{t}\frac{(\log s)^{1/3}}{s^{1/3}}.$$

Since $f(x) = (\log x)^{1/3} x^{-1/3}$ is positive and decreasing for $x \geq e$, we can bound the sum by an integral:

$$\sum_{s=1}^{t} \frac{(\log s)^{1/3}}{s^{1/3}} \leq 1 + \int_1^t \frac{(\log x)^{1/3}}{x^{1/3}}\, dx.$$

Consider the integral

$$I(t) := \int_1^t \frac{(\log x)^{1/3}}{x^{1/3}}\, dx.$$

Use the change of variables $x = u^{3/2}$, so $dx = \frac{3}{2}u^{1/2}du$ and $x^{-1/3} = u^{-1/2}$. Then

$$I(t) = \int_1^{t^{2/3}} \frac{\big(\log u^{3/2}\big)^{1/3}}{u^{1/2}} \cdot \frac{3}{2}u^{1/2}\, du = \frac{3}{2}\int_1^{t^{2/3}} \big(\tfrac{3}{2}\log u\big)^{1/3}\, du.$$

For all $1 \leq u \leq t^{2/3} \leq t$, we have $(\log u)^{1/3} \leq (\log t)^{1/3}$, hence

$$I(t) \leq \frac{3}{2}\Big(\tfrac{3}{2}\Big)^{1/3} (\log t)^{1/3} \int_1^{t^{2/3}} du \leq C_1 (\log t)^{1/3} t^{2/3},$$

for some universal constant $C_1 > 0$. Therefore,

$$\sum_{s=1}^{t} \epsilon_s \leq K^{1/3}\big(1 + C_1(\log t)^{1/3}t^{2/3}\big) \leq C_2(K\log t)^{1/3}t^{2/3},$$

for another universal constant $C_2 > 0$ and all t large enough (and we can adjust C_2 to make the inequality hold for all t).

Since each pull of a suboptimal arm contributes at most 1 to regret, we obtain

$$R(t) = t\mu_* - \mathbb{E}\Big[\sum_{s\leq t} X_{i_s}^{(s)}\Big] \leq C(K\log t)^{1/3}t^{2/3},$$

for some universal constant $C > 0$. This gives the desired bound

$$\boxed{t\mu_* - \mathbb{E}\Big[\sum_{s\leq t} X_{i_s}^{(s)}\Big] \leq C(K\log t)^{1/3}t^{2/3} \quad \text{for all } t.}$$

Implication for regret. We have shown that

$$R(t) = O\big((K\log t)^{1/3}t^{2/3}\big).$$

In particular, this is *sublinear* in t, so the average regret satisfies

$$\frac{R(t)}{t} = O\big((K\log t)^{1/3}t^{-1/3}\big) \xrightarrow[t\to\infty]{} 0.$$

Thus, ϵ-greedy with the decaying schedule $\epsilon_t = (K\log t/t)^{1/3}$ is a *no-regret* algorithm.

Comparison with part (i). In part (i), with constant $\epsilon_t = \epsilon \in (0, 1)$, we found that the average regret per round converges to a positive constant, so $R(t)$ grows *linearly* with t:

$$R(t) = \Theta(t).$$

In contrast, with the decaying schedule $\epsilon_t = (K\log t/t)^{1/3}$, the regret grows only as

$$O\big((K\log t)^{1/3}t^{2/3}\big)$$

, which is sublinear. Therefore the decaying exploration schedule yields much better long-run performance: the average regret goes to 0, whereas under constant ϵ it converges to a positive constant.

4

Minimizing regret to the best fixed arm in hindsight is only one way to formalize what it means to perform well in a k-armed bandit problem. An alternative goal is best-arm identification: at the end of T rounds, predict the index y_T of the arm with highest expected reward. The goal of best-arm identification can be equivalently stated as minimizing instantaneous regret $\mu_* - \mu_{y_T}$. When a k-armed bandit problem is interpreted as a clinical trial involving k treatments, the regret minimization problem can be interpreted as minimizing harm done to study participants, while best-arm identification can be interpreted as identifying the best treatment for all future patients, possibly at the cost of greater harm done to study participants.

(i) Suppose you alternate between playing each of the k arms over T rounds, such that each arm is played $\lfloor T/k \rfloor$ times. Prove that for sufficiently large T, there exists constant $C > 0$ such that

$$\mu_* - \mathbb{E}[\mu_{y_T}] \leq e^{-C \cdot T}.$$

Can this algorithm achieve sublinear regret as well?

Let $\mu_1, \dots, \mu_k$ be the arm means, and let

$$\mu_* = \max_{i \in [k]} \mu_i, \quad i_* \in \arg\max_i \mu_i.$$

Assume there is a unique best arm, so that for all $i \neq i_*$ we can define the gaps

$$\Delta_i := \mu_* - \mu_i > 0, \qquad \Delta_{\min} := \min_{i \neq i_*} \Delta_i > 0.$$

Over T rounds, the algorithm plays each arm exactly

$$m := \left\lfloor \frac{T}{k} \right\rfloor$$

times. Let $\widehat{\mu}_i$ be the empirical mean of arm i after m plays. At the end, the algorithm outputs

$$y_T := \arg\max_{i \in [k]} \widehat{\mu}_i.$$

Step 1: A high-probability "good" event. Consider the event

$$E := \bigcap_{i=1}^{k} \left\{ |\widehat{\mu}_i - \mu_i| \leq \frac{\Delta_{\min}}{2} \right\}.$$

On E, we compare the empirical means:

For the optimal arm i_*,

$$\widehat{\mu}_{i_*} \geq \mu_* - \frac{\Delta_{\min}}{2}.$$

For any suboptimal arm $i \neq i_*$,

$$\widehat{\mu}_i \leq \mu_i + \frac{\Delta_{\min}}{2} = \mu_* - \Delta_i + \frac{\Delta_{\min}}{2} \leq \mu_* - \frac{\Delta_{\min}}{2},$$

since $\Delta_i \geq \Delta_{\min}$. Hence on E we must have

$$\widehat{\mu}_{i_*} > \widehat{\mu}_i \quad \forall\, i \neq i_*,$$

so $y_T = i_*$ and

$$E \subseteq \{y_T = i_*\}.$$

Thus

$$\Pr(y_T \neq i_*) \leq \Pr(E^c) \leq \sum_{i=1}^{k} \Pr\left(|\widehat{\mu}_i - \mu_i| > \frac{\Delta_{\min}}{2} \right).$$

Step 2: Apply Hoeffding and bound the error probability. Each $\widehat{\mu}_i$ is the average of m i.i.d. $[0,1]$-bounded rewards, so by Hoeffding's inequality,

$$\Pr\left(|\widehat{\mu}_i - \mu_i| > \frac{\Delta_{\min}}{2}\right) \le 2\exp\left(-2m\left(\frac{\Delta_{\min}}{2}\right)^2\right) = 2\exp\left(-\frac{m\Delta_{\min}^2}{2}\right).$$

Therefore,

$$\Pr(y_T \neq i_*) \le 2k\exp\left(-\frac{m\Delta_{\min}^2}{2}\right).$$

For sufficiently large T, we have $m = \lfloor T/k \rfloor \ge \frac{T}{2k}$ (e.g., for $T \ge 2k$), so

$$\Pr(y_T \neq i_*) \le 2k\exp\left(-\frac{T\Delta_{\min}^2}{4k}\right) = 2k\, e^{-C'T}, \quad \text{where } C' := \frac{\Delta_{\min}^2}{4k} > 0.$$

Since $2k$ is just a constant, we can absorb it into the exponential for large T: there exist T_0 and $C > 0$ such that for all $T \ge T_0$,

$$\Pr(y_T \neq i_*) \le e^{-CT}.$$

Step 3: From misidentification probability to instantaneous regret. We have

$$\mu_* - \mathbb{E}[\mu_{y_T}] = \sum_{i=1}^{k}(\mu_* - \mu_i)\Pr(y_T = i) = \sum_{i \neq i_*}\Delta_i \Pr(y_T = i).$$

Using $\Delta_i \le 1$ (since rewards lie in $[0,1]$),

$$\mu_* - \mathbb{E}[\mu_{y_T}] \le \sum_{i \neq i_*}\Pr(y_T = i) = \Pr(y_T \neq i_*) \le e^{-CT},$$

for sufficiently large T. This proves the desired exponential best-arm identification bound.

Can this algorithm achieve sublinear regret? No. Let $N_i(T)$ be the number of times arm i is pulled up to time T. For this round-robin algorithm,

$$N_i(T) = m = \left\lfloor \frac{T}{k} \right\rfloor \quad \text{for all } i.$$

The (expected) cumulative regret is

$$R_T := T\mu_* - \mathbb{E}\left[\sum_{t=1}^{T} X_{i_t}^{(t)}\right] = \sum_{i=1}^{k}(\mu_* - \mu_i)\,\mathbb{E}[N_i(T)] = \sum_{i \neq i_*}\Delta_i \left\lfloor \frac{T}{k} \right\rfloor.$$

Since each $\Delta_i > 0$ for $i \neq i_*$, we get

$$R_T = \Theta(T),$$

which is *linear* in T and hence not sublinear.

Thus, this simple uniform exploration strategy achieves exponentially small best-arm identification error but suffers linear cumulative regret.

(ii) Suppose you have a blackbox algorithm whose regret to the best fixed arm in hindsight after T rounds is bounded by a sublinear function of T. That is, the algorithm guaranteees that

$$T\mu_* - \mathbb{E}\left[\sum_{t \le T} X_{i_t}^{(t)}\right] \le \phi(T)$$

for some $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ such that $\lim_{T\to\infty} \phi(T)/T = 0$. Show how this algorithm can be used to obtain the best-arm identification guarantee

$$\mu_* - \mathbb{E}[\mu_{y_T}] \le \frac{\phi(T)}{T}.$$

What does this result imply about the compatibility of regret minimization and best-arm identification, as goals? Compare and contrast with the result of the previous part.

Construction of y_T and proof of the bound. Run the blackbox algorithm for T rounds, obtaining actions $i_1, \dots, i_T$. After round T, define y_T as follows: pick an index U uniformly at random from $\{1, \dots, T\}$, independently of the rewards, and set

$$y_T := i_U.$$

We first relate $\mathbb{E}[\mu_{y_T}]$ to the average performance of the algorithm. By definition of U,

$$\mathbb{E}[\mu_{y_T}] = \mathbb{E}[\mu_{i_U}] = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\mu_{i_t}].$$

Since $X_{i_t}^{(t)}$ has mean μ_{i_t},

$$\mathbb{E}\Big[\sum_{t=1}^{T} X_{i_t}^{(t)}\Big] = \sum_{t=1}^{T} \mathbb{E}[\mu_{i_t}].$$

Hence

$$\mathbb{E}[\mu_{y_T}] = \frac{1}{T}\,\mathbb{E}\Big[\sum_{t=1}^{T} X_{i_t}^{(t)}\Big].$$

Using the regret guarantee,

$$T\mu_* - \mathbb{E}\Big[\sum_{t=1}^{T} X_{i_t}^{(t)}\Big] \le \phi(T),$$

we divide both sides by T:

$$\mu_* - \frac{1}{T}\,\mathbb{E}\Big[\sum_{t=1}^{T} X_{i_t}^{(t)}\Big] \le \frac{\phi(T)}{T}.$$

Substituting the expression for $\mathbb{E}[\mu_{y_T}]$ gives

$$\mu_* - \mathbb{E}[\mu_{y_T}] \le \frac{\phi(T)}{T},$$

which is the desired best-arm identification guarantee.

Interpretation and comparison with part (i).

Since $\phi(T)$ is sublinear, we have $\phi(T)/T \to 0$ as $T \to \infty$, so this construction yields

$$\mu_* - \mathbb{E}[\mu_{y_T}] \to 0.$$

Thus, *any* algorithm with sublinear regret can be turned into a best-arm identification procedure whose instantaneous regret vanishes asymptotically.

In contrast, in part (i), the simple round–robin algorithm achieves *excellent* best-arm identification (the error decays exponentially in T), but its cumulative regret is $\Theta(T)$ (linear), so it *does not* achieve sublinear regret.

Therefore:

- Sublinear regret $\Rightarrow$ good best-arm identification (via the random-time trick above).
- Good best-arm identification $\not\Rightarrow$ low regret (as shown by the round–robin example).

This shows that regret minimization is a strictly stronger requirement: algorithms designed to minimize regret are compatible with best-arm identification, but algorithms optimized solely for best-arm identification may incur large cumulative regret.

5

We aim to improve fantasy football score prediction. Progress in this task has direct financial relevance in the form of gambling. This is a prediction problem that is quite challenging due to highly complex data that depends on many unmodeled features, as well as a general scarcity of valid data. For this problem, our overarching objective is to maximize prediction accuracy in determining the relative ordering of player performance (within various position groups) for a given week of NFL games. In doing so, we consider varied sources of data to determine which data and features are most successful in predicting week-to-week performance. We plan to investigate the efficacy of using embeddings of text-based sports articles as a feature. The idea behind this method is that articles may capture additional context and better inform previous statistics. If successful, a similar methodology may be used to inform trading decisions by analyzing the analysis and articles published by financial analysts. We intend to compare our prediction performance against the analysis and rankings from fantasy-focused sports media outlets, e.g., [Fantasy Pros](https://www.fantasypros.com/nfl/rankings/half-point-ppr-rb.php).