Contents

3

Bandit methods

When you say you're going to do an A/B test somebody usually suggests using bandit methods instead. And vice versa.

In the bandit framework you try to optimize as you go and ideally spend next to no time on the suboptimal choice between A and B, or other options. It can be as good as having only $O(\log(n))$ tries of any sub-optimal treatment in n trials.

We review some theory of bandit methods. The main point is to learn the goals and methods and tradeoffs with bandits. We also go more deeply into Thompson sampling proposed originally in Thompson (1933).

Puzzlers and opinions

Most texts and articles are all about facts, not opinions. Facts are better. Sometimes opinions fill in where we don't have a desired fact. I'll be putting some of those in, and I expect you will be able to tell. I have been reading the blog of Andrew Gelman `https://statmodeling.stat.columbia.edu/` for many years and have benefitted enormously. Some of the opinions he shares are things I have long believed but never saw in print. It was nice to know that I was not alone. Sometimes my opinion is different from his.

One opinion I'm planning to describe is that theorems have issues of external validity. We have to think carefully about when and how to apply them.

I'm also going to put in some 'puzzlers'. These are things that are potentially confusing or apparently contradictory about the methods and their properties. Sometimes there is a momentary puzzle while we figure things out. Other times we do not get a clean answer. One of my goals is for students to learn to find and solve their own puzzlers. When you spot and resolve a puzzler it deepens

your understanding. So, get confused and then get out of it. Spotting and resolving puzzlers is also a way to find research ideas.

Here is a quote from Paul Halmos about reading mathematics:

> Don't just read it; fight it! Ask your own questions, look for your own examples, discover your own proofs. Is the hypothesis necessary? Is the converse true? What happens in the classical special case? What about the degenerate cases? Where does the proof use the hypothesis?

It is good to poke at statistical ideas in much the same way, with a view to which problems they suit.

3.1 Exploration and exploitation

In a regular experiment we get data on n subjects estimate $\mathbb{E}(Y \mid A)$ and $\mathbb{E}(Y \mid B)$. We pick what seems to be the better of A and B from our data and retain that choice hypothetically forever. Perhaps only for some $N \gg n$ future uses before we contemplate another change.

In this setting, we use the first n subjects to ***explore*** treatment differences. Once we have the apparent best one, we ***exploit*** that knowledge for the next N subjects by using the winning treatment.

One problem with experimentation is that something like $n/2$ of the subjects will be getting the suboptimal treatment in the experiment. Maybe we can avoid much of tha cost by biasing the experiment towards the seemingly better treatment at each stage as the data come in.

The theoretically most effective way to do this is through what are called ***bandit methods***. The name comes from slot machines for gambling that are sometimes called one-armed bandits. Each time you pull that arm you win a random amount of money that has expected value less than what you paid to play. The image for a multi-armed bandit is such a machine that offers you $K \geqslant 2$ arms to pick from. Each arm has its own distribution of random payoffs. In the gambling context of pulling n arms, the goal might be to minimize your expected loss. Of course, the best move is not to play at all, so the metaphor is imperfect.

In the experimental settings we care about, the goal is to maximize your expected winnings. If we knew the best arm, we'd choose it every time. But we don't. Instead we sample from the arms to learn about payoffs on the fly while also trying to get the best payoff. We will see bandit methods that pick a suboptimal arm only $O(\log(n))$ times as the number n of subjects goes to infinity.

A very old method is called 'play the winner'. Suppose that $Y_i \in \{0, 1\}$ with 1 being the desired outcome. Then if $Y_i = 1$ we could take $W_{i+1} = W_i$, while for $Y_i = 0$ we switch to $W_{i+1} = 1 - W_i$ when the choices are $W \in \{0, 1\}$ or, for $K > 2$ choices switch to a random other arm. These methods were studied intensely in the 1960s and 1970s and the term 'play the winner' seems to be

used to describe numerous different strategies. If there is a really great strategy in the mix then play the winner can have long streaks of $Y_i = 1$. If instead the best arm has a small payoff, like $\Pr(Y_i = 1) = 0.03$, and the other arm (out of 2) has $\Pr(Y_i = 1) = 0$, then play the winner will alternate too much between the best and worst arms.

3.2 Regret

These definitions are based on Bubeck and Cesa-Bianchi (2012). Suppose that at time $i = 1, 2, 3, \ldots$ we have arms $j = 1, 2, \ldots, K$ to pick from. If at time i we pick arm j then we would get $Y_{i,j} \sim \nu_j$. Notice that the distribution ν_j here is assumed to not depend on i. We let $\mu_j = \mathbb{E}(Y_{i,j})$ be the mean of ν_j and

$$\mu_* = \max_{1 \leqslant j \leqslant K} \mu_j \equiv \mu_{j_*}.$$

So μ_* is the optimal expected payout and j_* is the optimal arm (or one that is tied for optimal).

If we knew the μ_j we would choose arm j_* every time and get expected payoff $n\mu_*$ in n tries. Instead we randomize our choice of arm, searching for the optimal one. At time i we choose a random arm $J_i \in \{1, 2, \ldots, K\}$ and get payoff Y_{i,J_i}. Because we choose just one arm, we do not get to see what would have happened for the other $K - 1$ arms. That is, we never see $Y_{i,j'}$ for any $j' \neq J_i$, so we cannot learn from those values.

There are various ways to quantify how much worse off we are than optimal play would be. The ***regret*** at time n is

$$R_n = \max_j \sum_{i=1}^{n} Y_{i,j} - \sum_{i=1}^{n} Y_{i,J_i}.$$

This is how much worse off we are compared to whatever arm would have been the best one to use continually for the first n tries. Be sure that you understand why $\Pr(R_n < 0) > 0$ with this defniition. A harsher definition is

$$\sum_{i=1}^{n} \max_j Y_{i,j} - \sum_{i=1}^{n} Y_{i,J_i}.$$

This is how much worse off we would be compared to a psychic who knew the future data. It is not a reasonable comparison so it is not the focus of our study.

The expected regret is

$$\mathbb{E}(R_n) = \mathbb{E}\left(\max_j \sum_{i=1}^{n} Y_{i,j} - \sum_{i=1}^{n} Y_{i,J_i} \right).$$

It is awkward to study because the maximizing j is inside the expectation.

Bubeck and Cesa-Bianchi (2012) define the pseudo-regret

$$\begin{aligned}\bar{R}_n &= \max_j \mathbb{E}\left(\sum_{i=1}^n Y_{i,j} - \sum_{i=1}^n Y_{i,J_i}\right)\\ &= \max_j n\mu_j - \sum_{i=1}^n \mathbb{E}(Y_{i,J_i})\\ &= n\mu_* - \sum_{i=1}^n \mathbb{E}(\mu_{J_i})\end{aligned}$$

Each time we move $\max_j$ outside of a sum or expectation, things get easier. What is random in the $\mathbb{E}(\cdot)$ of $\bar{R}_n$ is the sequence $J_1, \ldots, J_n$ of chosen arms. Other authors call $\bar{R}_n$ the expected regret.

Now let $\Delta_j = \mu_* - \mu_j \geqslant 0$ be the suboptimality of arm j and define $T_j(s) = \sum_{i=1}^s 1\{J_i = j\}$. This is the number of times that arm j was chosen in the first s tries. Then

$$\bar{R}_n = n\mu_* - \sum_{i=1}^n \mathbb{E}(\mu_{J_i}) = \sum_{j=1}^K \mathbb{E}(T_j(n))\mu_* - \sum_{j=1}^K \mathbb{E}(T_j(n))\mu_j = \sum_{j=1}^K \mathbb{E}(T_j(n))\Delta_j.$$

Our pseudo-regret comes from the expected number of each kind of suboptimal pulls time their suboptimality. To derive this notice that

$$n = \sum_{j=1}^K T_j(n) = \sum_{j=1}^K \mathbb{E}(T_j(n))$$

because exactly one arm is chosen for every i.

3.3 Upper confidence limit

Figure 3.1 depicts a hypothetical setting with three treatment arms, there denoted by A, B and C with confidence intervals for the expected value of Y in all three arms.

Based on this information, which arm should we choose? There is an argument for arm B because the point estimate at the center of its confidence interval is the highest of the three. But that does not take account of uncertainty. If we would consider two restaurants one with six ratings that were all 5 stars and another with 999 5 star ratings and one 4 star rating, we would be more confident about the second restaurant. One way to judge that and play it safe is to rank by a lower confidence limit on the expected value. By that measure, treatment C has the highest lower limit and so it seems best for the cautious user.

The answer however, spoiled by the title of this section, is to choose arm A because it has the highest upper confidence limit. Suppose that we went

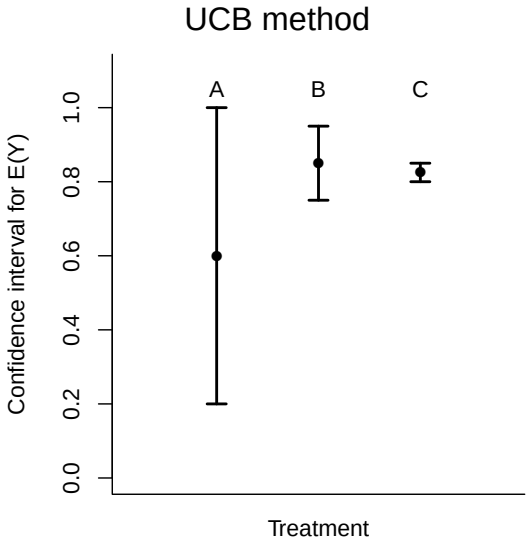

Figure 3.1: Hypothetical confidence intervals for $\mathbb{E}(Y \mid A)$, $\mathbb{E}(Y \mid B)$ and $\mathbb{E}(Y \mid C)$.

with B. Then it's confidence interval would tend to get narrower with further samples. It's center could also shift up or down as sampling goes on, tending towards the true mean which we anticipate to be somewhere inside the current confidence interval, though that won't always hold. What could happen is that the confidence interval converges on a value above the center for A but below the upper limit for A. Then if A were really as good as its upper limit, we would never sample it and find out. The same argument holds for sampling from C.

Now suppose that we sample from A. Its confidence interval will narrow and the center could move up or down. If A is really bad then sampling will move the mean down and narrow the confidence interval and it will no longer keep the top upper confidence limit. We would then stop playing it, at least for a while. If instead, A was really good, we would find that out.

Given that we want to use the upper limit of a confidence interval, what confidence level should we choose? The definitive sources on that point are Gittins (1979), Lai and Robbins (1985) and Lai (1987). We can begin with a finite horizon n, for instance the next n patients or visitors to a web page. At step i we could use the $100(1-\alpha_i)\%$ upper confidence limit.

It is easy to choose the treatment for the n'th subject. We just take the arm that we think has the highest mean. There is no $n+1$'st subject to benefit from what we learn from subject n. So we could take $\alpha_n = 1/2$. That would be the center of our confidence interval (if it is symmetric). If we are picking a fixed sequence $\alpha_1, \ldots, \alpha_n$ then it makes sense to have α_i increasing towards 0.5 because as time goes on, there is less opportuntity to take advantage of any learning. The α_i should start small, especially if n is large.

A finite horizon might not be realistic. We might not know how many

subjects will be in the study. Another approach is to define the discounted regret

$$\sum_{i=1}^{\infty}(\mu_* - \mathbb{E}(\mu_{J_i}))\theta^{i-1} \qquad 0 < \theta < 1.$$

This regret is the immediate regret plus θ times a similar future quantity

$$\mu_* - \mathbb{E}(\mu_{J_1}) + \theta\sum_{i=1}^{\infty}(\mu_* - \mathbb{E}(\mu_{J_{i+1}}))\theta^{i-1}.$$

It is much more reasonable to use some constant α in the discounted setting than in the fixed n setting. Choosing θ can be complicated. The average index is

$$\sum_{i=1}^{\infty} i\theta^{i-1} \Big/ \sum_{i=1}^{\infty}\theta^{i-1} = \frac{1}{1-\theta}$$

by considering the mean of a geometric distribution. So if we pick $\theta = 0.99$ then the weighted average index is 100. Or, if we have a time horizon like $n = H$ in mind we can set $\theta = 1 - 1/H$.

Finding the critical α_i values is complicated and depends on the underlying parametric distribution one might assume for the distributions ν_j. For us, the main idea is that betting on optimism paid off by keeping pseudo-regret $O(\log(n))$.

3.4 Thompson sampling

Thompson sampling goes back to Thompson (1933). The method was forgotten and rediscovered a few times. It is comparatively recently that most articles are online and findable over the internet, so the reinvention is understandable. The idea in Thompson sampling is to choose arm i with probability $\Pr(\mu_i \text{ is the best})$. This probability is a Bayesian one, so that it can be updated based on the observations so far. Thompson's motivation was for medical problems. The current surge in interest is from internet services.

Agrawal and Goyal (2012) showed that Thompson sampling can have a pseudo-regret of $O(\log(n))$. That puts it in the same performance league as UCB methods and Thompson is easier to deploy at least in simple settings.

We will focus on the Bernoulli case. The response values are $Y_{i,j} \in \{0, 1\}$. Let $\mu_j = \Pr(Y_{i,j} = 1)$. Then the likelihood function that we will use is

$$L(\mu_1, \ldots, \mu_K) = \prod_{i=1}^{n} p(Y_{i,J_i} \,|\, J_i) = \prod_{i=1}^{n} \mu_{J_i}^{Y_{i,J_i}}(1-\mu_{J_i})^{1-Y_{i,J_i}}. \tag{3.1}$$

Each factor $\mu_{J_i}^{Y_{i,J_i}}(1-\mu_{J_i})^{1-Y_{i,J_i}}$ is a likelihood contribution for $(\mu_1, \ldots, \mu_K)$ based on the conditional distribution of $Y_i = Y_{i,J_i}$ given J_i. There's a brief discussion about using a conditional likelihood below.

Taking this conditional likelihood (3.1) as our likelihood, we then pick a conjugate prior distribution in the beta family. Taking $\mu_j \overset{\text{ind}}{\sim} \text{Beta}(a_j, b_j)$ they have joint prior distribution proportional to

$$\prod_{j=1}^{K} \mu_j^{a_j-1}(1-\mu_j)^{b_j-1} \qquad 0 \leqslant \mu_j \leqslant 1.$$

Taking $a_j = b_j = 1$ makes $\mu_j \sim \mathbb{U}[0,1]$ independently. This is a popular choice. If K is very large and we know that the μ_j are likely to be very near zero from past experience then we could work with $a_j < b_j$. The mean of $\text{Beta}(a,b)$ is $\mu = a/(a+b)$ and the variance is $\mu(1-\mu)/(a+b+1)$ and these facts might help us settle on (a_j, b_j).

Let $S_j = \sum_{i=1}^{n} Y_{i,J_i} 1\{J_i = j\}$ and $F_j = \sum_{i=1}^{n}(1-Y_{i,J_i})1\{J_i = j\}$ be the numbers of successes and failures, respectively, observed in arm j. Then we can write our conditional likelihood as

$$\prod_{j=1}^{K} \mu_j^{S_j}(1-\mu_j)^{F_j}.$$

Notice that although (S_j, F_j) are defined by summing over all $i = 1, \ldots, n$, they do not depend on any unobserved Y values. Multiplying our conditional likelihood by the prior density we find a posterior density proportional to

$$\prod_{j=1}^{K} \mu_j^{a_j+S_j-1}(1-\mu_j)^{b_j+F_j-1}.$$

This means that the posterior distribution has $\mu_j \overset{\text{ind}}{\sim} \text{Beta}(a_j + S_j, b_j + F_j)$. This expression also shows that a_j and b_j can be viewed as numbers of prior pseudo-counts. We are operating as if we had already seen a_j successes and b_j failures from arm j before starting.

Figure 3.2 has pseudo-code for running the Thompson sampler for Bernoulli data and beta priors. In this problem it is easy to pick arm j with probability equal to the probability that μ_j is largest. We sample $\mu_1, \ldots, \mu_K$ one time each and let J be the index of the biggest one we get.

Thompson sampling is convenient for web applications where we might not be able to update S_j and F_j as fast as the data come in. Maybe the logs can only be scanned hourly or daily to get the most recent (J_i, Y_{i,J_i}) pairs. Then we just keep sampling with the fixed posterior distribution between updates. If instead we were using UCB then we might have to sample the arm with the highest upper confidence limit for a whole day between updates. That could be very suboptimal if that arm turns out to be a poor one.

Puzzler: the UCB analysis is pretty convincing that we win by betting on optimism. How does optimism enter the Thompson sampler? We get just one draw from the posterior for arm j. That draw could be better or worse than the mean. Just taking the mean would not bake in any optimism and

Initialize:
$S_j \leftarrow a_j$, $F_j \leftarrow b_j$, $j = 1, \dots, K$ # $a_j = b_j = 1$ starts $\mu_j \sim \mathbb{U}[0,1]$

Run:
for $i \geqslant 1$
  for $j = 1, \dots, K$
    $\theta_j \sim \text{Beta}(S_j, F_j)$ # make sure $\min(S_j, F_j) > 0$
  $J \leftarrow \arg\max_j \theta_j$ # call it J_i if you plan to save them
  $S_J \leftarrow S_J + X_{i,J}$
  $F_J \leftarrow F_J + 1 - X_{i,J}$

Figure 3.2: Pseudo-code for the Thompson sampler with Bernoulli responses and beta priors. As written it runs forever.

would fail to explore. We could bake in more optimism by letting each arm take $m > 1$ draws and report its best result. I have not seen this proposal analyzed (though it might well be in the literature somewhere). It would play more towards optimism but that does not mean it will work better; optimism was just one factor. Intuitively, taking $m > 1$ should favor the arms with less data, other things being equal. Without some theory, we cannot be sure that $m > 1$ doesn't actually slow down exploration. Maybe it would get us stuck in a bad arm forever (I doubt that on intuitive grounds only). If we wanted, we could take some high quantile of the beta distributions but deciding what quantile to use would involve the complexity that we avoided by moving from UCB to Thompson. For Bernoulli responses with a low success rate, the beta distributions will initially have a positive skewness. That is a sort of optimism.

Puzzler/rabbit hole: are we leaving out information about μ_j from the distribution of J_i? I think not, because the distribution of J_i is based on the past Y_i which already contribute to the conditional likelihood terms. A bit of web searching did not turn up the answer. It is clear that if you were given $J_1, \dots, J_n$ it would be possible at the least to figure out which μ_j was μ_*. But that doesn't mean they carry extra information. The random variables are observed in this order:

$$J_1 \to Y_1 \to J_2 \to Y_2 \to \cdots \to J_i \to Y_i \to \cdots \to J_n \to Y_n.$$

Each arrow points to new information about μ. The distribution of J_1 does not depend on $\mu = (\mu_1, \dots, \mu_K)$. The likelihood is

$$p(y_1 \mid J_1; \mu) p(J_2 \mid J_1, y_1; \mu) p(y_2 \mid J_2, J_1, y_1; \mu) p(J_3 \mid y_2, J_2, J_1, y_1; \mu) \cdots$$

Now in our Bernoulli Thompson sampler our algorithm for choosing J_3 was just based on a random number generator that was making our beta random variables. That convinces me that $p(J_3 \mid y_2, J_2, J_1, y_1; \mu)$ has nothing to do with μ. So the conditional likelihood is ok. At least for the Bernoulli bandit. Phew!

3.5 Theoretical findings on Thompson sampling

Agrawal and Goyal (2012) generalize Bernoulli Thompson sampling to handle bounded non-binary inputs. They get a value $Y \in \{0,1\}$ from $Y' \in [0,1]$ by randomly taking $Y = 1$ with probabilty Y'. This is pure randomness coming from their algorrithm not their data. If the response is actually $Y'' \in [a,b]$ for known $b > a$ we can start by setting $Y' = (Y'' - a)/(b-a)$ and then sampling $Y \sim \mathrm{Bern}(Y')$.

Theorem 1. *For $K = 2$ and $Y'_{i,j} \in [0,1]$ and $Y_{i,j} \sim \mathrm{Bern}(Y'_{i,j})$ the pseudo-regret is*

$$R_n = O\Big(\frac{\log(n)}{\Delta} + \frac{1}{\Delta^3}\Big),$$

as $n \to \infty$, where Δ is the suboptimality of the second best treatment.

Proof. This is Theorem 1 of Agrawal and Goyal (2012). They refer to expected regret but it appears to be pseudo-regret in the terminology of Bubeck and Cesa-Bianchi (2012). □

The pseudo-regret grows like $O(\log(n))$ for fixed Δ. If arm 1 is the suboptimal one then this means that $\mathbb{E}(T_1(n)) = O(\log(n))$. If the cumulative number of mistakes grows logarithmically then the typical gap between mistake times has to be growing exponentially. For instance $\mathbb{E}(T_1(2n) - T_1(n)) = O(\log(2n) - \log(n)) = O(1)$ (because the constant $\log(2)$ is $O(1)$). Each doubling of n brings at most a constant expected number of suboptimal arm choices.

Now let's look into the denominator Δ. The pseudo-regret is larger when Δ is smaller. If arms return 2% and 3% respectively, the bound leads us to expect much worse results than if they are 2.99% and 3%. The reason is that a suboptimal but nearly optimal arm will get chosen much more often than one that is very suboptimal. What about $\Delta = 0$? Something discontinuous happens here. The pseudo-regret bound is ∞. The actual pseudo-regret is exactly 0. It is not a contradiction: $0 < \infty$.

I promised a discussion of ***external validity of theorems***. The big-O in our theorem means that there exist constants $C < \infty$ and $N < \infty$ such that

$$\bar{R}_n \leqslant C \times \Big(\frac{\log(n)}{\Delta} + \frac{1}{\Delta^3}\Big) \quad \text{for all } n \geqslant N.$$

Sometimes it holds for $N = 1$. In a given situation we might expect $\bar{R}_n$ to grow like $\log(n)$ but be disappointed. Maybe it happens for extremely large N (in our problem). Or maybe the value of C is so large that $C\log(n)/n$ is not very small for any n that we can afford. We need more information than the $O(\cdot)$ result in the theorem to know if things are going to be good.

It can be valuable to spot-check theorems with some simulated examples. Simulated examples on their own are unsatisfying, also for external validity reasons. The ones in the literature might have been cherry-picked. Ideally the examples are known to be similar to our use case or perhaps to cover a wide range of possibilities.

A theorem can also be misleadingly pessimistic. If an error quantity $E_n = o(g(n))$ it means that $\lim_{n\to\infty} g(n)|E_n| = 0$. Anything that is $o(g(n))$ is automatically $O(g(n))$. In this case there is a theorem from Lai (1987) showing that no method could be $o(\log(n))$, so that doesn't happen here.

In this case we can think of what is perhaps the best possible case for Thompson sampling. One arm has $\mu_j = 1$ and the other has $\mu_j = 0$.

Puzzler: In class, I wondered what would happen if instead of adding Y_{i,J_i} to S_j and $1 - Y_{iJ_i}$ to F_j we added the probabilities Y'_{I,J_i} to S_j and $1 - Y'_{iJ_i}$ to F_j. That takes some noise out of the algorithm. It leads to beta distributions with non-integral parameters, but those are ok. It would complicate the analysis behind the theorem. We know from Lai and Robbins that we would not get a better convergence rate than $O(\log(n))$. Specifically, their bound is of the form

$$\left(\sum_{j=2}^{K} \frac{\Delta_j}{\mathrm{KL}(\nu_j||\nu_*)} + o(1)\right)\log(n),$$

for Kullback-Leibler divergence

$$\mathrm{KL}(P||Q) = \int_{-\infty}^{\infty} p(x)\log\Big(\frac{p(x)}{q(x)}\Big)\,\mathrm{d}x$$

in the case of continuous distributions P and Q with a natural modification for discrete distributions. Perhaps we get a better constant in the rate from using Y' instead of Y.

Agrawal and Goyal (2012) have additional results to cover the case $K > 2$.

Theorem 2. *For $K > 2$ and optimal arm $j^* = 1$*

$$\bar{R}_n = O\left(\left(\sum_{j=2}^{K}\frac{1}{\Delta_j^2}\right)^2\log(n)\right)$$

for suboptimalities Δ_j. Also

$$\bar{R}_n = O\left(\frac{\Delta_{\max}}{\Delta_{\min}^3}\sum_{j=2}^{K}\frac{1}{\Delta_j^2}\right)\log(n)\Bigg)$$

where $\Delta_{\min} = \min_{2\leqslant j\leqslant K}\Delta_j$ and $\Delta_{\max} = \max_{2\leqslant j\leqslant K}\Delta_j$.

Proof. The first result is Theorem 2 of Agrawal and Goyal (2012) and the second is their Remark 3. □

The second bound is better for large K, while the first is better for small Δ_j.

3.6 More about bandits

Suppose that A is better than B. Then from a bandit we only get $O(\log(n))$ samples from B. Therefore we do not get a good estimate of

$$\Delta = \mathbb{E}(Y\,|\,A) - \mathbb{E}(Y\,|\,B).$$

We can be confident that we are taking the best arm but we cannot get a good estimate of the amount of improvement. For some purposes we might want to know how much better the best arm is.

Maybe $W_i = (W_{i1}, \ldots, W_{i,10}) \in \{0,1\}^{10}$ because we have 10 decisions to make for subject i. We could run a bandit with $K = 2^{10}$ arms but that is awkward. An alternative is to come up with a model, such as $\Pr(Y = 1 \,|\, W) = \Phi(W^\mathsf{T}\beta)$ for unknown β. Or maybe a logistic regression would be better. We can place a prior on β and update it as data come in. Then we need a way to sample a W with probability proportional to it being the best one. Some details for this example are given in Scott (2010). These can be hard problems but the way forward via Thompson sampling appears easier than generalizing UCB. This setting has an interesting feature. Things we learn from one of the 1024 arms provide information on β and thereby update our prior on some of the other arms.

For contextual bandits, we have a feature vector X_i that tells us something about subject i before we pick a treatment. Now our model might be $\Pr(Y = 1 \mid X, W; \beta)$ for parameters β. See Agrawal and Goyal (2013) for Thompson sampling and contextual bandits.

In restless bandits, the distributions ν_j can be drifting over time. See Whittle (1988). Clearly we have to explore more often in this case because some other arm might have suddenly become much more favorable than the one we usually choose. It also means that the very distant past observations might not be relevant, and so the upper confidence limits or parameter distributions should be based on recent data with past data downweighted or omitted.

Bibliography

Agrawal, S. and Goyal, N. (2012). Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1.

Agrawal, S. and Goyal, N. (2013). Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135.

Bubeck, S. and Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*.

Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):148–164.

Lai, T. L. (1987). Adaptive treatment allocation and the multi-armed bandit problem. *The Annals of Statistics*, pages 1091–1114.

Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22.

Scott, S. L. (2010). A modern bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294.

Whittle, P. (1988). Restless bandits: Activity allocation in a changing world. *Journal of applied probability*, pages 287–298.