

A How can adaptively collected data lead to biased estimates?

“Adaptive” data is data that is collected such that the decision about what to sample next depends on the results of previous samples. For example, in a multi-arm bandit setting, data collected by strategies like ϵ -greedy and Thompson sampling are adaptive, since these strategies sample arms that gave high rewards in the past more than arms that previously gave lower rewards. Adaptively collected data have a temporal dependency and thus the samples are not independent. The dependency between samples makes constructing unbiased estimators more difficult for adaptively collected data than for i.i.d. data.

We describe one illustrative example about how the dependency between adaptively collected samples can lead to biased estimates when using estimators that assume i.i.d. data. Specifically, we will show how the sample mean of adaptive data can be biased [7].

Suppose we have two identical arms that emit rewards that are drawn i.i.d. from a standard normal distribution. We draw a total of three samples from these arms: X_1, X_2, X_3 . X_1 is always sampled from arm 1 and X_2 is always sampled from arm 2. We follow a greedy strategy and sample X_3 from the arm that emits a higher reward from the first two samples. Let M_1 and M_2 represent the sample means of arms 1 and 2 respectively.

If $X_1 > X_2$:

$$\begin{cases} M_1 = \frac{1}{2}(X_1 + X_3) \\ M_2 = X_2 \end{cases}$$

Nice!

else ($X_1 < X_2$):

$$\begin{cases} M_1 = X_1 \\ M_2 = \frac{1}{2}(X_2 + X_3) \end{cases}$$

We let $Z_1, Z_2, Z_3 \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$ and use the following order statistics notation $Z_{(1)} \sim \min(Z_1, Z_2)$ and $Z_{(2)} \sim \max(Z_1, Z_2)$.

$$M_1 \sim P(X_1 > X_2) \frac{Z_{(2)} + Z_3}{2} + P(X_1 < X_2) Z_{(1)}$$

Thus, since by symmetry $P(X_1 > X_2) = P(X_1 < X_2) = \frac{1}{2}$,

$$\begin{aligned} E[M_1] &= P(X_1 > X_2) E[M_1 | X_1 > X_2] + P(X_1 < X_2) E[M_1 | X_1 < X_2] \\ &= P(X_1 > X_2) E\left[\frac{Z_{(2)} + Z_3}{2}\right] + P(X_1 < X_2) E[Z_{(1)}] \\ &= \frac{1}{2} E\left[\frac{Z_{(2)} + Z_3}{2}\right] + \frac{1}{2} E[Z_{(1)}] \\ &= \frac{1}{4} E[Z_{(2)}] + \frac{1}{2} E[Z_{(1)}] < 0 \end{aligned}$$

since by symmetry, $E[Z_{(2)}] = -E[Z_{(1)}]$.

Thus we have shown in this simple three sample case, that the sample mean is negatively biased when using a greedy strategy. Note that by symmetry, if we use an “anti-greedy” strategy of sampling, in which we sample the arm that gives a *lower* reward more, the expected sample mean *overestimates* the true mean.

Nie et al. [7] show that for bandit algorithms that have the properties “exploit” and “independence of irrelevant options”, the sample mean will be negatively biased. The exploit property means that if arm K is chosen, then in an alternative sample history, if the sample mean of arm K were higher, then arm K would still be chosen. The IIO property means that if arm K is not chosen, which other arm is selected only depends on the histories of the other arms.