Evaluation of Early Stopping in A/B Testing

Shan Huang

June 12, 2017

This documentation presents the evaluation of early stopping algorithms.

1 Problem

Given samples **x** from treatment group, samples **y** from control group, we want to know whether there is a significant difference between the means $\delta = \mu(y) - \mu(x)$.
To save the cost of long-running experiments, we want to stop the test early if we are already certain that there is a statistically significant result.

2 Significance Decision

We can decide whether the difference is statistically significant using either:

- **Confidence interval**: If 0 is outside confidence interval of δ, it is statistically significant.
- or **credible interval**: Credible interval is the Bayesian version of confidence interval. If 0 is outside credible interval of δ, it is statistically significant.
- or **Bayes factor**: Theoretically, Bayes factors higher than 3 can be interpreted as support for the null hypothesis (significant no difference), whereas values smaller than 1/3 can be interpreted as support for the alternative hypothesis (significant difference). Values between 1/3 and 3 are inconclusive. **In our test, we will only use Bayes factor less than 1/3 as decision criteria**. See reasons below.

Given the null hypothesis H_0 representing no difference, the alternative hypothesis H_1 representing a difference of the means, the ability of each metric, i.e., types of significance it can detect, is shown in the following table.

	Paradigm	Significant H_1	Significant H_0	No Significant Result
Confidence Interval	frequentist	✓		✓
Credible Interval	Bayesian	✓		✓
Bayes factor	Bayesian	✓	✓	✓

Table 2.1: Comparison of decision ability

To be consistent with all mehods, we draw a binary conclusion that either there is a significant difference or not. In other words, the first column means that there is a significant difference, combining the second and third column means that there is no significant difference. Thus, we only use the part where Bayes factor less than 1/3 as decision criteria.
It is worth noting that a **typical conclusion of Bayes factor** would be for instance "*There is a significant difference corresponds to Cauchy prior and a threshold of Bayes factor=3*". This might be quite difficult to explain to non-tech users. On the other hand, a **conclusion based on interval** can be more intuitive such as "*You can be 95% sure that the significant difference is not due to chance*".

3 Early Stopping Criteria

We can stop the experiment by either

- **Confidence interval**: Calculate the new significance level for each day based on alpha-spending function in group sequential method. Use new significance level to compute confidence interval. Stop the test if 0 is outside confidence interval of δ.
- or **credible interval**: Stop the test if 0 is outside credible interval of δ.
- or **Bayes factor**: Stop the test if Bayes factor is smaller than 1/3.
- or **Bayes precision**: Stop the test if credible interval width is smaller than 0.08.

4 Metric

4.1 Evaluate Significance Decision

We are going to compare false positive (type I error), false negative (type II error), true positive, true negative rates for the three significance decision criteria described in Section 2. These metrics will be evaluated only on simulation data, where we know the true value of significance.

4.2 Evaluate Early Stopping Criteria

It is obvious that if we use frequentist's approach — in our case is confidence interval — for early stopping, we should also use frequentist's approach for significance decision. On the other hand, when using Bayes factor for early stopping, we find a conflicting result if we draw significance conclusion based on credible interval. Therefore we don't evaluate all

combinations of early stopping algorithms and significance decisions, find below a table of combination we evaluated on.

Significant Based On	Early Stopping Based On			
	Confidence Interval	Credible Interval	Bayes factor	Bayes precision
Confidence Interval	✓			
Credible Interval		✓		
Bayes factor			✓	✓

Table 4.1: Combinations of early stopping and significance decision for evaluation

For each combination shown in the table above, we compute the following metrics:

- **False positive rate**: Percentage of tests which are wrongly stopped. i.e., Early stop says there is a significant difference and stops the test, but the test to the end day (as if there is no early stopping) will tell you it is not significant.
- **Run time reduced (all)**: Percentage of run time reduced for all tests.
- **Run time reduced (true positive)**: Percentage of run time reduced for only the correctly stopped tests.
- **Bias**: Average difference of effect size (delta) between stopping day and end day of test (as if there is no early stopping).

5 Evaluation

The evaluation is conducted on both simulation data and real data. Code can be found here.

5.1 Simulation Data

We generate 2000 simulation tests based on Gaussian distributed KPIs, of which 1000 tests have a significant difference between control and treatment, while the other 1000 tests have no significant difference. The effect size of significant difference satisfies the minimal detectable effect size from power analysis.
We simulate the experiment data for a period of 20 days. We run an analysis and decide whether to stop on each day. The frequency of visits per entity is modelled by a Poisson distribution. It is simulated in a way that an entity will visit the experiment 3 times in average.
First, the table below shows the performance of **significance decision** algorithms.

	FPR	TNR	FNR	TPR
confidence interval	4.6%	95.4%	0.9%	99.1%
credible interval	4.6%	95.4%	1.0%	99.0%
Bayes factor	0.2%	99.8%	18.9%	81.1%

Table 5.1: Performance of significance decision on simulation data

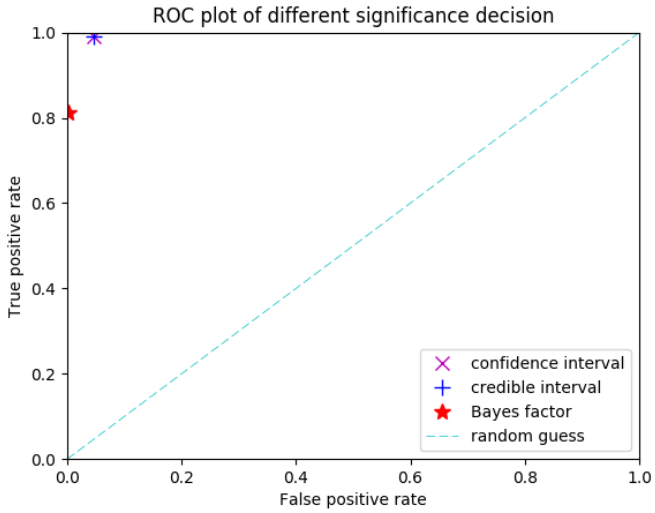

Figure 5.1: ROC plot on simulation data

We can observe in Figure 5.1 that Bayes factor works very good on false positive rate, but the two interval based approaches works better on true positive rate. In general, the two interval based approaches are a little bit better because of larger distances to the line of random guess, while the error rate of all three methods are acceptable (FPR < 5% and FNR < 80%). So there is no strong arguements that one method is better than the others in terms of significance decision on the last day.
Next, we show the results of **early stopping** algorithms in the following table . To make notations compact, we use FPR for false positive rate, RTR (all) for run time reduced (all), RTR (TP) for run time reduced (true positive) and CI for confidence interval. The definition of each metric is described previously in Section 4.2.

Significance	Early Stopping	Paradigm	FPR	RTR(all)	RTR(TP)	Bias
CI	CI	frequentist	3.3%	1.60%	16.74%	0.00
Credible Interval	Credible Interval	Bayesian	**22.6**%	17.56%	42.28%	0.00
Bayes factor	Bayes factor	Bayesian	0.7%	0.57%	42.5%	-0.02
Bayes factor	Bayes precision	Bayesian	0.2%	78.02%	78.02%	0.00

Table 5.2: Performance of early stopping in A/A test on simulation data

Significance	Early Stopping	Paradigm	FPR	RTR(all)	RTR(TP)	Bias
CI	CI	frequentist	1.7%	51.58%	51.74%	-0.01
Credible Interval	Credible Interval	Bayesian	0.7%	78.53%	78.78%	0.00
Bayes factor	Bayes factor	Bayesian	7.8%	46.44%	52.96%	0.00
Bayes factor	Bayes precision	Bayesian	**67.8**%	77.95%	77.76%	-0.02

Table 5.3: Performance of early stopping in A/B test on simulation data

Note that false positive rate is the most critical metric when evaluating early stopping. This is because it makes no sense how much run time we have saved, if we stop the test with a wrong conclusion.
We observe that there is a large false positive rate by credible interval on A/A tests and also by Bayes precision on A/B tests. It would be dangerous to roll these two methods out in production without further study. In general, the confidence interval approach works best on simulation data.

5.2 Real Data

We first fetch raw data in BigQuery from previous A/B testings. After some cleanup and data processing, we got 10 datasets with nice properties for evaluating A/B testing. You can download processed datasets in my GitHub repository. For a compact notation, we refer to each dataset using the following abbreviation.

- DTP_CH_web (**DTP**)
- Editorial_Assortment_Entries_treatment1 (**EAE1**)
- Editorial_Assortment_Entries_treatment2 (**EAE2**)
- Editorial_Catalog_entries_Msite (**ECEM**)
- lipstick_catalog_naviTracking_bunchbox_IT (**LCIT**)
- lipstick_catalog_naviTracking_bunchbox_NL (**LCNL**)
- segmented_sorting_fasion_floor_fashion (**FFF**)
- segmented_sorting_fasion_floor_modern (**FFM**)
- segmented_sorting_fasion_floor_no_floor (**FFNF**)
- segmented_sorting_fasion_floor_trend (**FFT**)

We test early stopping algorithms with the most frequently used primary KPI in the past, which is **conversion rate per session (CR)**.
A summary of dataset properties can be found below, where **traffic split** represents the percentage of samples in treatment group, and **NaN in CR** represents the percentage of missing values after computing conversion rate per session.

Dataset	#Days	#Samples	#Control	#Treatment	Traffic Split	NaN in CR
DTP	20	552995	284448	268547	48.56%	1.84%
EAE1	21	329960	164833	165127	50.04%	0.00%
EAE2	21	329910	164833	165077	50.04%	0.00%
ECEM	35	250186	126106	124080	49.60%	**7.83%**
LCIT	14	970017	485157	484860	49.98%	0.00%
LCNL	14	**1027470**	513266	514204	50.05%	0.00%
FFF	**56**	135995	68855	67140	49.37%	0.00%
FFM	**56**	68596	34778	33818	49.30%	0.00%
FFNF	**56**	134800	67583	67217	49.86%	0.00%
FFT	**56**	34265	16776	17489	51.04%	0.00%

Table 5.4: Properties of selected real-world datasets for evaluation

The **conversion rate per session** for visit i is calculated as simple as $CR_i = \frac{O_i}{S_i}$, where O_i is the sum of orders during visit i, and S_i is the number of sessions during visit i.
Given n samples of visits, it is worth noting that there are two ways to compute the overall conversion rate, we can calculate the average ratio per entity, i.e.

$$CR^{(pe)} = \frac{1}{n}\sum_{i=1}^{n} CR_i = \frac{1}{n}\sum_{i=1}^{n} \frac{O_i}{S_i} \tag{5.1}$$

This approach assumes equal weights on the contribution of each visit, which might lead to huge bias in practice. In fact, the product analysts in Zalando use the ratio of totals to reflect overall equal contributions to the conversion rate, which can be formulated as:

$$CR^{(rt)} = \frac{\sum_{i=1}^{n} O_i}{\sum_{i=1}^{n} S_i} \tag{5.2}$$

Nevertheless, this will just be one value in the end. Our statistical hypothesis requires that we have a group of samples in both control and treatment group. To be able to have samples of $CR^{(rt)}$ from $CR^{(pe)}$, we implemented a re-weighting trick (proof is trivial):

$$CR^{(rt)} = \frac{1}{n}\sum_{i=1}^{n} \alpha_i \frac{O_i}{S_i} \tag{5.3}$$

where

$$\alpha_i = n\frac{S_i}{\sum_{i=1}^{n} S_i} \tag{5.4}$$

Now that we have seen all components to evaluate real-world usecases, we can show the result of performance on real data. The result is illustrated in the Table 5.5.
We observe that the frequentist early stopping approach works on real data as on simulation data. Bayes factor also works fine in most cases.
Another interesting observation is that Bayes precision always saves a large amount of time, this is because Bayes precision implicitly stops on both side — supporting H_0 and supporting H_1 — whereas the other three methods only stops when there is a significant difference, i.e., only stops supporting H_1.

Dataset	Frequentist	Credible Interval	Bayes Factor	Bayes Precision
DTP	Correctly stopped: saved 35% time	Correctly stopped: saved 65% time	Correctly stopped: saved 20% time	**Wrongly** stopped
EAE1	No stop	**Wrongly** stopped	No stop	Correctly stopped: saved 95% time
EAE2	No stop	Correctly stopped: saved 95% time	No stop	Correctly stopped: saved 95%
ECEM	No stop	No stop	No stop	Correctly stopped: saved 94% time
LCIT	No stop	**Wrongly** stopped	No stop	Correctly stopped: saved 93% time
LCNL	No stop	No stop	No stop	Correctly stopped: saved 93% time
FFF	Correctly stopped: saved 1% time	Correctly stopped: saved 79% time	No stop	Correctly stopped: saved 91% time
FFM	No stop	No stop	No stop	Correctly stopped: saved 84% time
FFNF	No stop	**Wrongly** stopped	No stop	Correctly stopped: saved 91% time
FFT	No stop	No stop	No stop	Correctly stopped: saved 70% time

Table 5.5: Performance of early stopping on real data

6 Conclusion

Given the performance of Bayes precision and credible interval on both simulation data and real data, there is still much space to improve. We will need to investigate further on the theory and implementation of these two methods.
Bayes factors and the frequentist approach both perform good. However, we have to change the whole concept of our A/B testing to roll out the Bayesian approach in live system. Therefore, we suggest the frequentist approach — i.e., use group sequential method for early stopping and confidence interval for significance decision.