Shipping for profit in A/B-testing

A Preprint

Artem V. Vorozhtsov
London,
artem.voroztsov@gmail.com

July 29, 2022

Abstract

This paper presents another instance of the exploration vs. exploitation trade-off problem, which arises in A/B testing or changes evaluation as a problem to maximize profit over time. The specificity of our setup is that: (a) there are infinitely many independent versions (arms) to test, (b) the aim is to maximize profit without accounting for p-values, (c) new data continuously arrives, and we monitor test results continuously, employing early stopping. This problem can be reformulated as a Bayesian multi-armed bandit problem with continuous time, an infinite number of arms, and a separation between exploitation and exploration, where exploitation occurs only once in an arm's life, and an arm can be not exploited, i.e. rejected. We consider case of of gaussian arms, where unexplored arm has a priori parameters (a_0, σ_0), and after some exploration has parameters $(\hat{a}, \hat{\sigma})$.

We compare perfomance of several stopping criteria involving conditions on UCB-like indexes, for instance MOSS index, and new index $\mathbf{g}$ with condition $\mathbf{g}(\hat{a}, \hat{\sigma}^2) < g_0$ for rejecting, and $\hat{a} > \mathbf{g}(\hat{a}, \hat{\sigma}^2) - g_0$ for launching, where $\mathbf{g}(\hat{a}, \hat{\sigma}^2) = \hat{\sigma} \cdot \phi(\hat{a},\ \hat{\sigma}^2) + \hat{a} \cdot \Phi(\hat{a}, \hat{\sigma}^2) - \xi/\hat{\sigma}^2$.

Keywords multi-armed bandit · sequential A/B testing · Gittins index · Network evaluation

Contents

1 Introduction

Companies typically optimize their product offerings using randomized controlled trials (RCTs), in industry parlance known as A/B testing.

A/B testing is exposing a product versions with some changes change to some percent of users and/or assortment and running it for some time period to get statistically significant answers about the hypothesis if the change is profitable or not.

But experiments require resources, i. e. time and users that you have to put under an experiment. And here we come to the important question: what is the best strategy for early stopping in A/B testing when you have shortage of resources, numerous experimental versions, and want to maximize company's profit?

This is one more variant of exploration vs exploitation balance problem, that is not covered by existing numerous formalizations of multi-armed bandit problem [Bubeck et al.(2012)].

In A/B testing we test versions, which are basically divided into three classes: good (with positive impact on product) and bad (with negative impact), and neutral (with zero impact). Real tangible changes with zero impact are quite hypothetical, so the third class could be empty.

It is important to have one main metric measuring this impact on product. It should combine all aspects of you product. Here we call it "profit", but in practice it could be complex combination of several metrics about product quality, user behaviour, and company's profit.

This metric is often arises as Lagrangian for some optimization task and corresponds to balance between "happiness" of all parties (business itself, partners, users). And task of miximizing absolute value of this metric during the next year is often s task of a high priority for a business.

It is worth being very specific here: you task is to maximize this metric, rather than number of launched good versions with predefined thresholds on falsity of lauched and unlanched versions.

Let's consider two hypotheses "version is good" and "version is bad". There are two false rates corresponding to these hypotheses: probability of type I error, i. e. launching bad version, end probability of type II error, i. e. rejecting good version.

The well-known state-of-art approach for A/B-testing is test with precalculated fixed sample size that guarantees certain values for these two false rates. And usually, to collect fixed size data, you need fixed period of time. But, as we will show, fixed test data size is not optimal for maximizing profit. The intuition bihind this is simple: some of really bad and really good versions can be rejected and launched faster, before planned sample size is reached.

Naive approach of continuously monitoring profit's mean value and its variance, and deriving p-values from them, provides unreliable results. But there is approach for sequential analysis that provides "always valid p-values" estimation [Johari(2015), Johari(2016)].

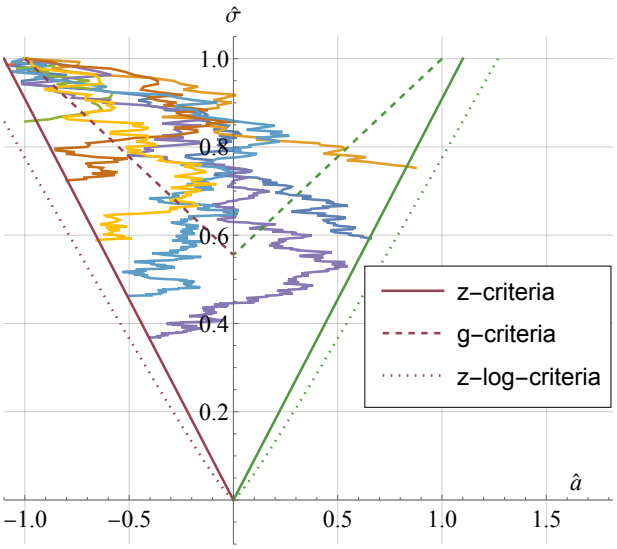
From one hand it's an interesting mathematical result and tangible step toward practionieers' demand. But from the another hand it is step in the opposite direction to what we propose here. "Always valid p-values" approach results in more strict conditions for rejecting and launching. Meanwile "shipping for profit" approach make these conditions less strict.

And again, it is important to say, that any aproach about trade-off between false rates and average experiment run time is not equvalent to the task of maximizing profit. It could be, but it is not.

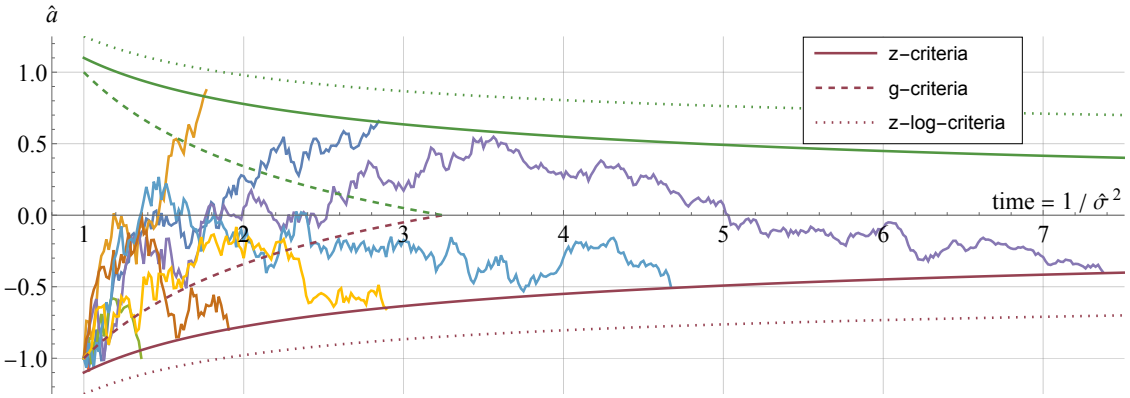
Here we describe sequential A/B testing approach with goal to maximize profit per time. It does not account for mentioned error probabilities, although they can be calculated. Proposed criteria can reject and launch versions with different false rates, and these false rates can be close to 0.5.

In the section 2 we define the basic "shipping for profit" problem $\texttt{GaussSP}(a_0, \sigma_0^2, \delta^2, \tau)$, where versions' values are i.i.d gaussian random variables. Lauching a version leeds to addition of its value to the global value. One can think about global value as log(profit). This means that each version provides multiplier to global profit, and logarithm of this multiplier is sampled from the normal distribution $\mathcal{N}(a_0, \sigma_0^2)$ with known a_0, σ_0. Values of a_0 and σ_0 are the only prior knowledge we have about a new version.

Let's use symbols $\Phi(a, \sigma)$ and $\phi(a, \sigma)$ for CDF and PDF functions of $\mathcal{N}(a, \sigma^2)$.



(a) Experiments trace on the plane $(\hat{a},\ \hat{\sigma})$.



(b) Experiments trace on the plane (time, $\hat{a}$)).

Figure 1: Continuous monitoring of an experiment gives us a trace of estimated a posteriori profit value $\hat{a}$ and its variance $\hat{\sigma}^2$. Traces are plotted in two planes: $(\hat{a},\ \hat{\sigma})$ and (time, $\hat{a}$), where time $= 1/\hat{\sigma}^2$. We have a priori values $\hat{a} = -1$, $\hat{\sigma} = 1$, thats why traces start at $\hat{a} = -1$, rather than $\hat{a} = 0$. Here we have three types of borders: (a) z-criteria, (b) g-criteria, (c) z-log-criteria. The z-criteria is typical threshold on "number of sigmas", usually denoted as z, and used by physisists as value for statistical significance. This criteria does not provides statistical significance corresponding to selected z value, if you do early stoping instead of waiting for fixed amount of time before applying criteria. The g-criteria is new proposed criteria which is very close to straight segments on the plane $(\hat{a},\ \hat{\sigma})$, connecting some point $(0,\ \sigma_f)$ with two symmetric points $(a_0,\ \sigma_0)$ and $(-a_0,\ \sigma_0)$. The z-log-criteria is modification of z-criteria that allows early stopping and guarantees statistical significance corresponding to selected z value.

The idea behind "shipping for profit" problem $\texttt{GaussSP}(a_0, \sigma_0^2, \delta^2, \tau)$ is simple. We have a posteriori value $\hat{a}$ and $\hat{\sigma}$ of an explored version. At the begining of the experiment, parameters $(\hat{a}, \hat{\sigma})$ are equal to (a_0, σ_0). During the experiment they do "random walk" to final point $(a, 0)$ not ever reaching it.

On the plane $(\hat{a}, \hat{\sigma})$ you can draw "reject" and "launch" border lines. And intersecting these borders triggers signals "stop and reject" or "stop and launch".

The commonly used p-value based borders for a gaussian metric are

$$\Phi(\hat{a}, \hat{\sigma}) < \alpha,$$
$$\Phi(\hat{a}, \hat{\sigma}) > 1 - \beta$$

Let's name them "z-criteria" and rewrite in equivalent form via z coefficients:

The z-criteria:

Parameters: z_α, z_β.

$$\begin{array}{rl} \text{reject if:} & \hat{a}/\hat{\sigma} < z_\alpha, \\ \text{launch if:} & -\hat{a}/\hat{\sigma} > z_\beta, \\ \text{continue:} & \text{otherwise} \end{array}$$

The z-criteria are not the most profitable criteria.

The following criteria perform better:

The **g**-criteria:

Parameters: ξ, a_0, σ_0

$$\begin{array}{rl} \text{reject if:} & \mathbf{g}(\hat{a}, \hat{\sigma}^2) - g_0 < 0, \\ \text{launch if:} & \mathbf{g}(\hat{a}, \hat{\sigma}^2) - g_0 < \hat{a}, \\ \text{continue:} & \text{otherwise} \end{array}$$

where

$$\mathbf{g}(a, \sigma^2) = \sigma \cdot \phi(a,\ \sigma^2) + a \cdot \Phi(a, \sigma^2) - \xi/\sigma^2,$$
$$g_0 = \mathbf{g}(a_0, \sigma_0^2).$$

We prove their optimality for continuous time in section 3.

The launch condition is equivalent to the mirrowed reject condition:

$$\mathbf{g}(-\hat{a}, \hat{\sigma}^2) < g_0,$$

So two borders for launching and rejecting are symmetric on plane $(\hat{a}, \hat{\sigma})$ with respect to the line $\hat{a} = 0$.

We can put $-g_0$ inside function $\mathbf{g}_{\text{full}}$:

$$\mathbf{g}_{\text{full}}(a, \sigma) = \mathbf{g}(a, \sigma) - g_0,$$

and then launch and reject conditions are simplified to the following:

$$\begin{array}{rl} \text{reject if:} & \mathbf{g}_{\text{full}}(\hat{a}, \hat{\sigma}^2) < 0, \\ \text{launch if:} & \mathbf{g}_{\text{full}}(\hat{a}, \hat{\sigma}^2) < \hat{a}, \\ \text{continue:} & \text{otherwise.} \end{array}$$

Borders, corresponding **g**-criteria are very close to straight lines, which we define as **q**-criteria:

The **q**-criteria:

Parameters: z, a_0, σ_0.

$$\begin{array}{rl} \text{reject if:} & \mathbf{q}(\hat{a}, \hat{\sigma}) < \mathbf{q}_0, \\ \text{launch if:} & \mathbf{q}(-\hat{a}, \hat{\sigma}) < \mathbf{q}_0 \\ \text{continue:} & \text{otherwise} \end{array}$$

where $\mathbf{q}(\hat{a}, \hat{\sigma}) = \frac{1}{2}\hat{a} + \frac{1}{2}z \cdot \hat{\sigma}$, and $\mathbf{q}_0 = \mathbf{q}(a_0, \sigma_0)$.

The **q**-criteria corresponds to two straight lines at the plane $(\hat{a}, \hat{\sigma})$ with intersect at some point $(0, \sigma_{min})$ rather than at the point $(0, 0)$ as in the case of z-criteria.

These conditions can be rewriten in the same way as **g**-criteria:

$$\begin{array}{rl} \text{reject if:} & \mathbf{q}_{\text{full}}(\hat{a}, \hat{\sigma}^2) - \mathbf{q}_0 < 0, \\ \text{launch if:} & \mathbf{q}_{\text{full}}(\hat{a}, \hat{\sigma}^2) - \mathbf{q}_0 < \hat{a}, \\ \text{continue:} & \text{otherwise.} \end{array}$$

This form is the reason why multiplier $\frac{1}{2}$ was introduced in the definition of **q**.

Numerical experiments with $a_0 = -1$, $\sigma_0 = 1$, $\delta^2 = 35$ reveals small but statisticaly significant difference between performance of **g**-creteria and **q**-criteria (with the best fitted parameters ξ, z, $\mathbf{q}_0$, $\mathbf{q}_0$).

2 The bayesian ship-for-profit problem $\texttt{GaussSP}(a_0, \sigma_0^2, \delta^2, \tau)$

Let's consider basic model for sequential A/B testing where versions' values are sampled from given distribution $\mathcal{N}(a_0, \sigma_0)$:

- There is fixed percent of users for which you can run one experiment at a time. We denote this experimental group as B, and control group as A. Each moment you can explore only one version, exposing it to users from group B.
- You have one strictly positive metrics profit that you should maximize during the year launching different versions. Value of profit is some **quantity per unit time per potential user**.
- You have an infinite pool of versions. Each version i has assigned to it value a_i, where a_i is sampled from $\mathcal{N}(a_0, \sigma_0^2)$, i.e. all values a_i are i.i.d. When launched, an version i adds a_i to the value of log(profit), so an version affects profit in multiplicative way, and logarithm of the multiplier is sampled from $\mathcal{N}(a_0, \sigma_0^2)$.
- You have discrete time moments t_n separated by τ seconds: $t_{n+1} = t_n + \tau$. After each n-th round $[t_n, t_{n+1})$ you can look at current profit in A (control) and B (experimental version i) test groups for the period of exploration of the version i, combine it in bayesian way with prior knowledge about an version, and make decision about experiment: stop and launch, stop and reject, or continue the experiment. Measurement of profit in A and B within time interval $[t_n, t_{n+1})$ gives

  $$b_{i,n} = b_{i,n,n+1} = b(i, t_n, t_{n+1}) = \log\left(\frac{\text{profit}(B, t_n, t_{n+1})}{\text{profit}(A, t_n, t_{n+1})}\right)$$

  as sample from $\mathcal{N}(a_i, \delta^2)$. Change in logarithmic relative profit over the longer time period $[t_k, t_{k+m})$, $k = n - m + 1$, of exploration of the version i is

  $$b_{i,k,k+m} = b(i, t_k, t_{k+m}) = \log\left(\frac{\text{profit}(B, t_k, t_{k+m})}{\text{profit}(A, t_k, t_{k+m})}\right).$$

  As it is avarage of m samples from $\mathcal{N}(a_i, \delta^2)$, it has distribution $\mathcal{N}(a_i, \delta^2/m)$. Combination of measurement and prior knowledge gives us the following estimated values:

  $$\hat{a}_i := \frac{\sigma_0^{-2} \cdot a_0 + m \cdot \delta^{-2} \cdot b(i, t_k, t_{k+m})}{\sigma_0^{-2} + m \cdot \delta^{-2}}$$
  $$\hat{\sigma}_i := (\sigma_0^{-2} + m \cdot \delta^{-2})^{-2}$$

  Value of δ^2 depends on the A and B groups sizes and time quant length τ:

  $$\delta^2 \propto \left(\frac{1}{\text{size}(A)} + \frac{1}{\text{size}(B)}\right) \cdot \frac{1}{\tau}$$

- Preparing an version for experiment, launching or rejecting it do not require time and do not cost anything. Cost of maintaining launched version can be incorporated into a.
- Control is automatically updated each time an version is launched.
- Result is measured as sum of a_i of launched versions. One can argue that lauched versions continue contributing to the profit each moment after launching, and it's true, but using this definition of result gives equivalent optimization problem.
- We do not consider changes in profit in group B during exploration for two reasons: experimental group B is usually small (for instance, 1% of total potential users) and thus does not effect much the total profit, and even after taking it into consideration we, again, get equivalent optimization problem.

In practice, logarithm of relative profits $b_{i,k,k+m}$ are quite close to gaussian, because usually profit is a sum of a large number of payments, so values profit(A) and profit(B) are close to $\mathcal{N}(a_{total}, \sigma_{total}^2)$ with low relative variance: $\sigma_{total}/a_{total} \ll 1$. From low relative variance we get that log(profit(A)) and log(profit(A)) are also close to gaussian, and so is the value of their difference log(profit(B)) − log(profit(B)).

The most questionable propositions in this model are the independence of versions values a_i and identity of their distributions. Its far from what we have in practice:

- Versions do correlate, often strongly, and after launching / rejecting of an version some other versions should be reconsidered, and some experiments should be restarted.
- Versions have different estimated potential in profit grouth, and different non zero costs for testing, launching and maintaining.

Still, we adhere to this simplified model, hoping to get a simple answer. Let's denote log(profit) grouth induced by launched versions during time $n \cdot \tau$ as total value V, and relative grouth as $v = V/\tau \cdot n$.

Let's reformulate model in a more strict and concise way as an instance of multi-armed bandit problem (MAB, arm = version) with infinitly many unreusable arms, and separate exploration and exploitation actions:

The bayesian ship-for-profit problem $\texttt{GaussSP}(a_0, \sigma_0^2, \delta^2, \tau)$

Input:
Infinitly many arms $i = 1, 2, \ldots$ with unknown values $a_i \sim \mathcal{N}(a_0, \sigma_0^2)$.
Number of rounds N.
Measurement variance: δ^2.

Output: v

Goal: Continuously explore arms and launch the most promising of them trying to maximize expected launched value v per round.

1 Initialize:
Choose the arm $i := 1$.
Set a posteriori arm parameters to $\hat{a}_i := a_0$ and $\hat{\sigma}_i := \sigma_0$.
Set absolute and relative values $V := 0$ and $v := 0$.
Set $t := \tau$.

2 For each round $n = 1, 2, \ldots, n_{max}$:

2.1 Chooses one of thee actions:

(a) launch the arm i and take the next arm and update total value V:

$$i := i + 1;\ \hat{a}_i := a_0;\ \hat{\sigma}_i := \sigma_0$$
$$V := V + a_i$$
$$v := V/t$$

(b) reject the arm i, and take the next arm:

$$i := i + 1;\ \hat{a}_i := a_0;\ \hat{\sigma}_i := \sigma_0$$

(c) do nothing

2.2 Continue exploring the arm i for time τ; get new measurment $b_{i,n}$ sampled from $\mathcal{N}(a_i, \delta^2)$ and update $(\hat{a}_i,\ \hat{\sigma}_i)$:

$$t := t + \tau$$
$$\hat{a}_i := \frac{\hat{\sigma}_i^{-2} \cdot \hat{a}_i + \delta^{-2} \cdot b_{i,n}}{\hat{\sigma}_i^{-2} + \delta^{-2}}$$
$$\hat{\sigma}_i := (\hat{\sigma}_i^{-2} + \delta^{-2})^{-1/2}$$

Arm exploitation could be done only once, and this is the reason why term "exploitation" is replaced by "launching".

Notice 1: Time flows only in step 2.2.

Notice 2: It can be shown that $1/\hat{\sigma}_i^2 = m \cdot 1/\delta^2$, where m is the number of exploration rounds of the version i. If we denote $w_i = 1/\hat{\sigma}_i^2$ and $w_0 = 1/\delta^2$, then update rules can be rewritten as

$$\hat{a}_i := \frac{w_i \cdot \hat{a}_i + w_0 \cdot b_{i,n}}{w_i + w_0}$$
$$w_i := w_i + w_0$$

Notice 3: The case of positive a_0 and $u_r = u_l = 0$ is degenerate, because the optimal solution is to launch "all" versions without exploration, i.e. set $\tau = 0$ or just skip 2.2. (it is allowed), and get infinite V for zero time.

Notice 4: The case $a_0 = 0$ and $u_r = u_l = 0$ has the following optimal solution: at round n we explore the version $i = n$, and launch or reject it depending on $b_{i,n} > 0$ or $b_{i,n} \le 0$.

Experiments and simple calculations for $(a_0,\ \sigma_0) = (0,\ 1)$ shows, that optimal solution has

$$v \approx \frac{C}{\tau \cdot \delta^2}, \quad \text{where } C = \frac{1}{\sqrt{2\pi}} \approx 0.4.$$

But standard deviation δ^2 of measurement $a(n, t_n, t_{n+1})$ for a time period of length τ is proportional to $1/\tau$, so, let's introduce proportionality constant C_δ by the equation

$$\delta^2 = C_\delta/\tau.$$

We will show that

$$v \to \frac{C}{\tau \cdot \delta^2} = \frac{1}{\sqrt{2\pi} \cdot C_\delta},$$

as $\tau \to 0$.

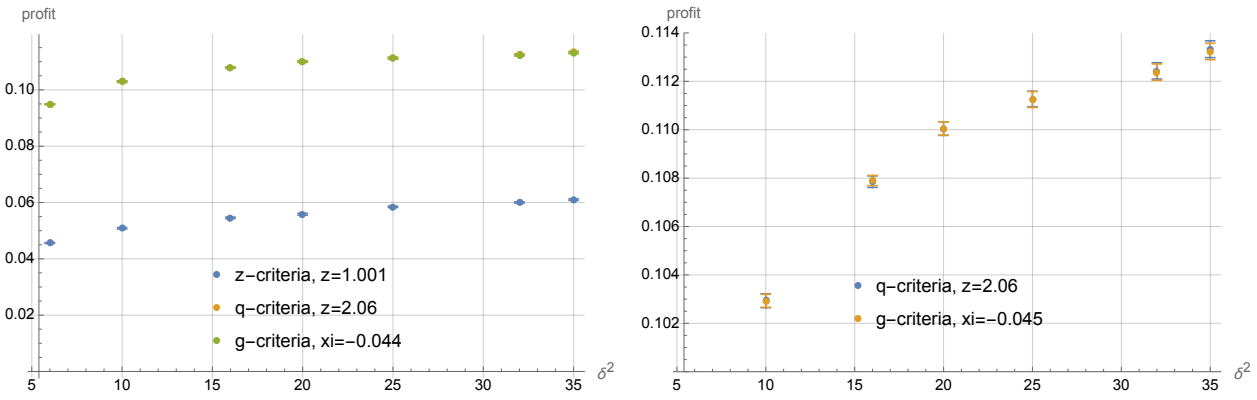
As a result, with prior $(a_0,\ \sigma_0) = (0,\ 1)$ one should test each version for one round and then launch or reject it. The time quant τ should be as small as possible, preserving the property of measurements $b_{i,n}$ being gaussian. In practise it means that τ should be chosen small, but rather big so that profit contains tens of nonzero components. Although, it is quite improbabale to have a_0 excactly equal to 0.

3 Case of bad in average versions. Experimental results

TODO!!!

$\texttt{GaussSP}(a_0, \sigma_0^2, \delta^2, \tau)$

Prior $(a_0,\ \sigma_0)$ with negative a_0 endows us with not trivial interesting case.



(a) Profit vs measurement error δ^2 for the three criteria: z-criteria, q-criteria, and g-criteria

(b) Profit vs measurement error δ^2 for the two best criteria: q-criteria and g-criteria

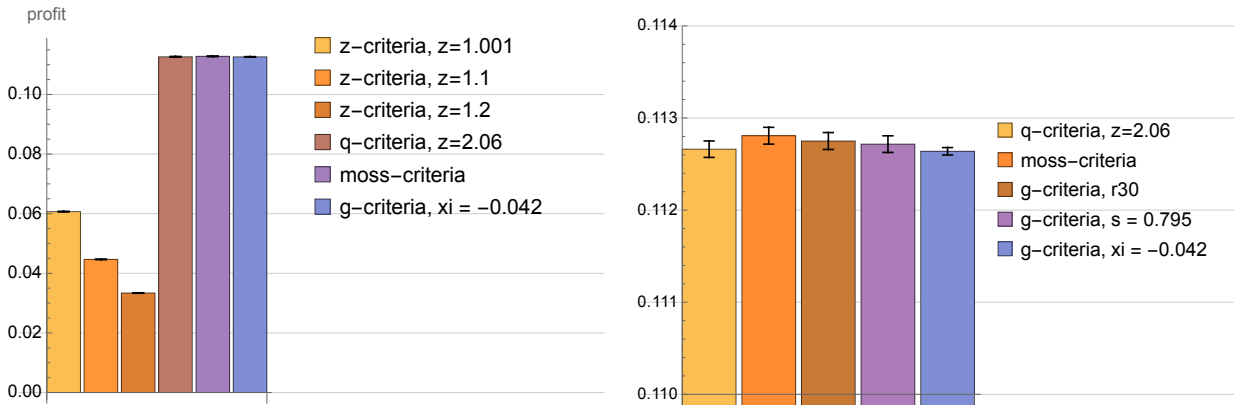
Figure 2: Profit vs. error δ^2

4 Inference of g

Theorem 1 *The optimal criteria for* $\texttt{GaussSP}(a_0, \sigma_0^2, C_\delta/\tau, \tau)$ *problem in limit* $\tau \to 0$ *are defined as*

$$\mathbf{g}(\hat{a}, \hat{\sigma}^2) > g_0, \quad \hat{a} > \mathbf{g}(\hat{a}, \hat{\sigma}^2).$$

where function $\mathbf{g}$ *is defined as* $\mathbf{g}(a, \sigma^2) = \sigma \cdot \phi(a,\ \sigma^2) + \hat{a} \cdot \Phi(a, \sigma^2) - \xi/\sigma^2$,



(a) Profit for $\delta^2 = 35$ for several criteria

(b) Profit for $\delta^2 = 35$ for the best criteria

Figure 3: Profit for error $\delta^2 = 35$

Before proving it, let's note that for $\xi = 0$ and $t = \sigma^2$, the function $\mathbf{g}$ satisfies diffusion differential equation:

$$\frac{\partial \mathbf{g}(a,t)}{\partial t} = \frac{1}{2} \cdot \frac{\partial^2 \mathbf{g}(a,t)}{\partial a^2}. \tag{1}$$

And the $\mathbf{g}$ with term ξ/σ^2 satisfies equation

$$\frac{\partial \mathbf{g}(a,t)}{\partial t} = \frac{1}{2} \cdot \frac{\partial^2 \mathbf{g}(a,t)}{\partial a^2} - \frac{\xi}{\sigma^2}. \tag{2}$$

The inference of $\mathbf{g}(a, \sigma^2)$ is based on this fact.

4.1 The idea behind g

Suppose, we have an optimal launching and rejecting criteria for the problem $\texttt{GaussSP}(a_0, \sigma_0^2, C_\delta/\tau, \tau)$. And suppose we have a version i explored for some period, with a poseriori parameters $(\hat{a},\ \hat{\sigma})$.

Let's define total expected value, that we get during next n rounds using this optimal algorithm as

$$V_{opt}(\hat{a}, \hat{\sigma}, a_0, \sigma_0, n).$$

This function satisfies the following reccurent formula:

$$\begin{array}{l} V_{opt}(\hat{a}, \hat{\sigma}, a_0, \sigma_0, n) = \max(\\ \quad \hat{a} + V_{opt}(a_0, \sigma_0, a_0, \sigma_0, n), \\ \quad V_{opt}(a_0, \sigma_0, a_0, \sigma_0, n), \\ \quad \mathbb{E}\left[V_{opt}(a', \sigma', a_0, \sigma_0, n-1)\right] \\ ), \end{array} \tag{3}$$

where (a', σ') is pair of random variables equal to a posteriori parameters that we can have after next round of exploration.

Three expressions in max function correspond to three actions that we can undertake:

(a) launch version i with parameters $(\hat{a}, \hat{\sigma})$ and take next version with parameters (a_0, σ_0); it is reasonable only if $\hat{a} > 0$

(b) reject version i and take next version with parameters (a_0, σ_0); it is reasonable only when $\hat{a} < 0$;

(c) explore version i one more round and get updated a posteriori parameters (a', σ').

In the case (c) we need to avarage over all probable measurement outcomes. The value of σ' is not random and determined by $1/\sigma'^2 = 1/\hat{\sigma}^2 + 1/\delta^2$. The value of a' is random, and we have to calculate expectation integrating over $a' \sim \mathcal{N}(\hat{a}, \sigma'^2 \cdot \hat{\sigma}^2/\delta^2)$.

Now we decompose $V_{opt}(\hat{a}, \hat{\sigma}, a_0, \sigma_0, n)$ into sum of two components:

- $V_{opt}(a_0, \sigma_0, a_0, \sigma_0, n)$ – expected value we get for n rounds using optimal strategy subject to we have no any explored unlaunched versions; for big n is grows linealy, and $V_{opt}(a_0, \sigma_0, a_0, \sigma_0, n) = v_0 \cdot n + c_0 + o(1)$. The value of v_0 depends on (a_0, σ_0).
- $\mathbf{g}(\hat{a}, \hat{\sigma}, a_0, \sigma_0, n)$ – the rest part of $V_{opt}(\hat{a}, \hat{\sigma}, a_0, \sigma_0, n)$; it can be interpreted as reward for having explored version with a posteriori parameters $(\hat{a}, \hat{\sigma})$.

Subtracting $V_{opt}(a_0, \sigma_0, a_0, \sigma_0, n)$ from both sides of reccurent equation, and omitting, for the sake of simplicity, arguments (a_0, σ_0), we get

$$\mathbf{g}(\hat{a}, \hat{\sigma}^2) = \max(\hat{a},\ 0,\ -v_0 + \mathbb{E}\left[\mathbf{g}(a', \sigma'^2)\right])$$

Expression $\mathbb{E}\left[\mathbf{g}(a', \sigma'^2)\right]$ corresponds to gaussian smoothing, also known as Weierstrass transform, of function $\mathbf{g}(a, \sigma^2)$ over argument a. It is defined by the convolution of $\mathbf{g}(a, \sigma^2)$ with the PDF function of the normal dstribution $\mathcal{N}(0, \frac{(\sigma^{-2}+\delta^{-2})^{-1} \cdot \sigma^2}{\delta^2})$ over the argument a.

Let's account for τ and do substitutions:

$$\begin{gathered} \delta \mapsto C_\delta / \sqrt{\tau}, \\ v_0 \mapsto C_v \cdot \tau. \end{gathered}$$

We get

$$\mathbf{g}(\hat{a}, \hat{\sigma}^2) = \max(\hat{a},\ 0,\ -C_v \cdot \tau + \mathbb{E}\left[\mathbf{g}(a', \sigma'^2)\right]) \tag{4}$$

and

$$\begin{gathered} a' = \hat{a} + \nu, \\ \nu \sim \mathcal{N}(0, \tau \cdot s^2), \\ \sigma'^2 = 1/(1/\hat{\sigma}^2 + \tau / C_\delta^2), \\ s^2 = \sigma'^2 \cdot \hat{\sigma}^2 / C_\delta^2. \end{gathered}$$

Let's take series upto $o(\tau^2)$:

$$\mathbf{g}(\hat{a}, \hat{\sigma}^2) = \max(\hat{a},\ 0,\ \mathbf{g}(\hat{a}, \hat{\sigma}^2) - \tau \cdot C_v + \tau \cdot \frac{\hat{\sigma}^4}{C_\delta^2} \cdot \frac{1}{2} \cdot \partial_{2,0} g(\hat{a}, \hat{\sigma}^2) - \tau \cdot \frac{\hat{\sigma}^4}{C_\delta^2} \cdot \partial_{0,1} \mathbf{g}(\hat{a}, \hat{\sigma}^2))$$

where $\partial_{2,0}\mathbf{g}$ is the second derivative on the first argument, i.e. $\frac{\partial^2 \mathbf{g}(a,t)}{\partial a^2}$, and $\partial_{0,1}\mathbf{g}$ is the first derivative on the second argument, i.e. $\frac{\partial \mathbf{g}(a,t)}{\partial t}$.

In the region, where $\mathbf{g}(a, \sigma) > 0$ and $\mathbf{g}(a, \sigma) > a$ we have

$$\frac{\partial g(a,t)}{\partial t} = 1/2 \cdot \frac{\partial^2 g(a,t)}{\partial a^2} - \frac{\xi}{t^2},$$

where $t = \hat{\sigma}^2$, $\xi = C_\delta^2 \cdot C_v$.

The border condition is $g(a, 0) = \max(0,\ a)$ follows from definition of $V_{opt}(\hat{a}, \hat{\sigma}, a_0, \sigma_0, n)$. Indeed, if we have an version with zero value variance, then we reject or launch it immediately and sample new version from $\mathcal{N}(a_0, \sigma_0)$:

$$V_{opt}(\hat{a}, 0, a_0, \sigma_0, n) = \max(0,\ \hat{a}) + V_{opt}(a_0, \sigma_0, a_0, \sigma_0, n).$$

If we assume $\xi = 0$, we get the solution

$$\mathbf{g}(a, \sigma^2) = \sigma \cdot \phi(a, \sigma^2) + a \cdot \Phi(a, \sigma^2).$$

The function

$$\mathbf{g}(a, \sigma^2) = \frac{\xi}{\sigma^2} + \sigma \cdot \phi(a, \sigma^2) + a \cdot \Phi(a, \sigma^2)$$

satisfies differential equation with non zero ξ, but does not satisfy border conditions. It is contradiction, because differencial equation works only in the region $\mathbf{g}(a, \sigma) > 0$ and $\mathbf{g}(a, \sigma) > \hat{a}$.

So, complete solution of 4 under limit $\tau \to 0$ is

$$\mathbf{g}(a, \sigma^2) = \max(0,\ a,\ \frac{\xi}{\sigma^2} + \sigma \cdot \phi(a, \sigma^2) + a \cdot \Phi(a, \sigma^2)).$$

5 Appendix

5.1 Proof of g-criteria optimality

TODO!!!

5.2 Proof of optimality of 1-round algorithm for $a_0 = 0$

Let's consider $\texttt{GaussSP}(a_0 = 0, \sigma_0^2, \delta^2 = C_\delta^2/\tau, \tau)$ and estimate v for 1-round algorithm, when we explore each version just for 1 round. Lets denote mean value of $\hat{a}$ after one round as a_1, and value of $\hat{\sigma}$ as σ_1. We launch version if $\hat{a} > 0$, and reject otherwise. This model can be described by equations:

$$\begin{array}{l} a \sim \mathcal{N}(0, \sigma_0^2), \\ m = 1/\sigma_0^2, \;\; m_\delta = 1/\delta^2 = \tau/C_\delta^2, \\ a_1 = \frac{m \cdot 0 + m_\delta \cdot a}{m + m_\delta}, \\ m_1 = m + m_\delta; \sigma_1^2 = 1/m_1, \\ \hat{a} \sim \mathcal{N}(a_1, \sigma_1^2), \\ V := V + \mathbb{E}\left[a \cdot \mathbf{1}_{\hat{a}>0}\right] \end{array} \tag{5}$$

The expectation $\mathbb{E}\left[a \cdot \mathbf{1}_{\hat{a}>0}\right]$ is the expectation of added value. The expression $\mathbf{1}_c$ is equal to 1 when the condition c is true, and 0 otherwise.

$$\begin{aligned} V_{add} &= \mathbb{E}\left[a \cdot \mathbf{1}_{\hat{a}>0} | \hat{a} \sim \mathcal{N}(a_1, \sigma_1^2)\right] = \\ &= \tfrac{1}{2} \cdot \left(1 + \operatorname{erf}\left(\tfrac{a \cdot m_\delta}{\sqrt{2} \cdot \sqrt{m + m_\delta}}\right)\right) = \\ &= \tfrac{1}{2} a + \tfrac{a^2 \cdot m_\delta}{\sqrt{2\pi} \cdot \sqrt{m}} + o(m_\delta) \end{aligned}$$

Now we can take the expecation of this value over $a \sim \mathcal{N}(0, \sigma_0^2)$:

$$\begin{aligned} V_{add} &= \mathbb{E}\left[\tfrac{1}{2} a + \tfrac{a^2 \cdot m_\delta}{\sqrt{2\pi} \cdot \sqrt{m}} + o(m_\delta) | a \sim \mathcal{N}(0, \sigma_0^2)\right] = \\ &= \tfrac{m_\delta}{\sqrt{2\pi} \cdot m^{3/2}} + o(m_\delta) \\ &= \tau \cdot \tfrac{\sigma_0^3}{\sqrt{2\pi} \cdot C_\delta^2} + o(\tau) \end{aligned}$$

If we take series up to $o(m_\delta)$ we get

$$\begin{aligned} V_{add} &= \tfrac{m_\delta}{\sqrt{2\pi} \cdot m^{3/2}} - \tfrac{m_\delta^2}{2\sqrt{2\pi} \cdot m^{5/2}} + o(m_\delta^2) = \\ &= \tau \cdot \tfrac{\sigma_0^3 \cdot \sigma_0}{\sqrt{2\pi} \cdot C_\delta^2} - \tau^2 \cdot \tfrac{\sigma_0^5 \cdot \sigma_0}{\sqrt{2\pi} \cdot C_\delta^4} + o(\tau^2) \end{aligned}$$

So, additional value per time is

$$v_{add} = V_{add}/\tau = \frac{\sigma_0^3}{\sqrt{2\pi} \cdot C_\delta^2} - \tau \cdot \frac{\sigma_0^5}{\sqrt{2\pi} \cdot C_\delta^4} + O(\tau)$$

So the less τ the better is result. This result does not prove the optimality of 1 step algorithm. We should prove that there is no 2 or k-steps approach which performes better in the limit $\tau \to 0$.

TODO!!!

6 Gaussian processes on the plane $(a,\ \sigma)$

Experiment traces on the plane (a, σ) can be considered as random walk processes with Poincare metrics of Lobachevskyi space. Randome steps are 1-dimentional (axes a) and axe σ corresponds to time: $time = \log(\sigma^2)$. Poincare metrics coinsides with Fisher metrics for the family of gaussian distributions $\mathcal{N}(a, \sigma)$. We can try to solve excactly the same problem on Euclidian space and find out in which way optimal solutions resamble each other. TODO!!!

References

[Johari(2015)] Johari, Ramesh and Pekelis, Leo and Walsh, David J. (2015) Always Valid Inference: Bringing Sequential Analysis to A/B Testing *arXiv preprint arXiv:1512.04922* .

[Johari(2016)] Johari, Ramesh, Stanford Seminar: Peeking at A/B Tests – Why It Matters and What to Do About It, `https://youtu.be/AJX4W3MwKzU`

[Bubeck et al.(2012)] Bubeck S, Cesa-Bianchi N, et al. (2012) Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning* 5(1):1–122.