Adaptive Experimentation at Scale

Ethan Che and Hongseok Namkoong
Columbia Business School

Overview

Adaptive sampling can improve statistical power of experiments.

Standard adaptive algorithms (e.g. bandits) are narrowly designed for unit-level reallocation

However, **unit-level reallocation is hard!**
- Delayed feedback
- Engineering cost

Modeling real-world experiments, we consider
- **Batch** evaluations of treatments
- **Limited** number of reallocation epochs
- **Low signal-to-noise** for broad KPIs

Main contributions
- **adaptive** policies with flexible batches
- **scalable** optimization-based algo
- **near-optimal** for the # of reallocations
- can incorporate **prior** knowledge.

Model

Find best option out of K treatment arms
- Treatment reward R_a with mean μ_a
- Few reallocation epochs (T), each with flexible batch sizes
- Choose allocation $\boldsymbol{\pi_t} \in \Delta_K$ at each epoch t

Goal: minimize **Bayes Simple Regret,** the optimality gap of final selection compared to the best arm, averaged over a prior over means

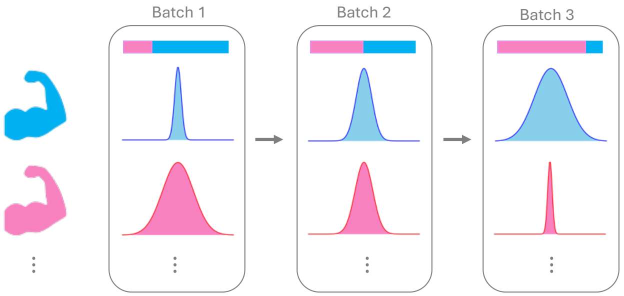
We take a prior over arm means (possibly from prior experiments), and pick the arm with **the highest posterior mean** after T epochs

Note: Prior only required on the gaps between means, not parameters of the reward distribution. **Prior only informs experimental design; we take a frequentist view to inference**

Gaussian Sequential Experiment

Challenge: How does the sampling policy affect uncertainty?

- The more one samples an arm, the more precise the measurement
- When we aggregate the samples in a batch, the measurement can be approximated by a normal with variance $\sim \pi_{t,a}^{-1}$.
- Experimenter observes a sequence of these measurements -> **Gaussian Sequential Experiment**



Theorem: This picture is a good approximation for large batches.

Assuming normal prior $N(\mu_{0,a}, \sigma_{0,a})$ over arm means, posterior beliefs in the Gaussian sequential experiment follows a Markov Decision Process (MDP) with known transitions:

Updated Posterior mean
Current Posterior mean
Sampling allocation to arm at epoch t

$$\mu_{t+1,\text{arm}} = \mu_{t,\text{arm}} + \sigma_{t,\text{arm}} \sqrt{\frac{\pi_{t,\text{arm}} \sigma_{t,\text{arm}}^2}{s_{\text{arm}}^2 + \pi_{t,\text{arm}} \sigma_{t,\text{arm}}^2}} Z_{t,\text{arm}}$$

$$\sigma_{t+1,\text{arm}}^2 = \left(\sigma_{t,\text{arm}}^{-2} + \frac{\pi_{t,\text{arm}}}{s_{\text{arm}}^2} \right)^{-1}$$

N(0,1) r.v.
Reward measurement variance
Updated Posterior variance
Current Posterior variance

The more one samples an arm, the more one's beliefs can change:
As $\pi \to 1$, variance of update increases to σ^2
As $\pi \to 0$, variance of update decreases to 0, no update in beliefs

Algorithm

- A policy $\pi = \{\pi_t(\mu_t, \sigma_t)\}$ determines the allocation based on current beliefs summarizing measurements seen so far
- Minimizing Bayes simple regret is equivalent to

Q-function at t = 0
policy affects future beliefs
posterior mean

$$Q_0^{\pi}\underbrace{(\mu_0, \sigma_0)}_{} = \mathbb{E}_0^{\pi}[\max_{\text{arm}} \mu_{T,\text{arm}}]$$

prior
arm with highest posterior mean at the end of the experiment

Algo 1: Policy Gradient
- Parameterize the policy $\pi_\theta = \{\pi_t^\theta\ (\mu_t, \sigma_t)\}$ using a neural network
- Directly optimize the objective by stochastic gradient descent on θ:

$$\theta \leftarrow \theta + \alpha \nabla_\theta Q_0^{\pi_\theta}(\mu_\theta, \sigma_\theta)$$

Algo 2: Iterated Static Optimization (Q-myopic):
- At epoch t, solve for the **best static allocation** π_t over remaining batches.

$$\pi_t(\mu_t, \sigma_t) = \arg\max_{\pi \in \Delta} E_t^{\pi}\ [\max_{arm} \mu_{T,arm}].$$

Theorem: Q-myopic obtains lower regret than any non-adaptive allocation, including Uniform, and the static allocation problem is strongly concave for $T - t$ large.

Results

$K = 10$ arms, $\text{B} = 100$ samples per batch.

Left: Bernoulli rewards, Beta prior.
- Achieves **strong improvement** over **uniform** and standard adaptive algos
- **Despite small effective batch size.**

Right: Gumbel rewards, Gamma prior.
- Each bar is a different reward measurement noise level.
- Achieves **strong improvements** over uniform **despite large measurement noise**, other policies struggle to eliminate arms