BOTS: Batch Bayesian Optimization of Extended Thompson Sampling for Severely Episode-Limited RL Settings

Karine[1], Susan A. Murphy[2], Benjamin M. Marlin[1]

1. University of Massachusetts Amherst, 2. Harvard University

Motivations and challenges

- There is an increasing interest in using Reinforcement learning (RL) to learn policies for just-in-time adaptive interventions (JITAIs).
- Physical activity adaptive intervention can be framed as **RL**: a mobile health app (**agent**) sends messages (**actions**) to a participant (**environment**) to encourage the participant to exercise more (**reward** = walking step count).

However, there are challenges:

- The number of episodes available for learning can be severely limited due to cost or time constraints.
- Thompson Sampling (TS) can be used in severely episode-limited RL setting, but TS selects actions based on distributions of immediate rewards only.

Contributions

1. We introduce **BOTS**, a novel method that offers a **practical method** for the **severely episode-limited RL settings** (e.g., health adaptive interventions).
2. BOTS is batch Bayesian Optimization of extended TS:
   - We introduce the **action bias term βa**. xTS extends TS to select actions based on TS's estimate of the expected immediate reward combined with βa.
   - We use **batch Bayesian optimization (BO)** over episodes to **learn βa** with the goal of maximizing the expected return of extended TS.
3. BOTS outperforms RL methods in the severely episode-limited RL settings.
4. Our code is available at: github.com/reml-lab/BOTS

Methods: Configurations

BO optimizes costly functions that only requires the ability to evaluate the function. BO works by iteratively optimizing an approximation (GP regression) to the true costly objective function. We explore the use of batch BO, including local and global BO.

Table 1: Summary of the BOTS methods

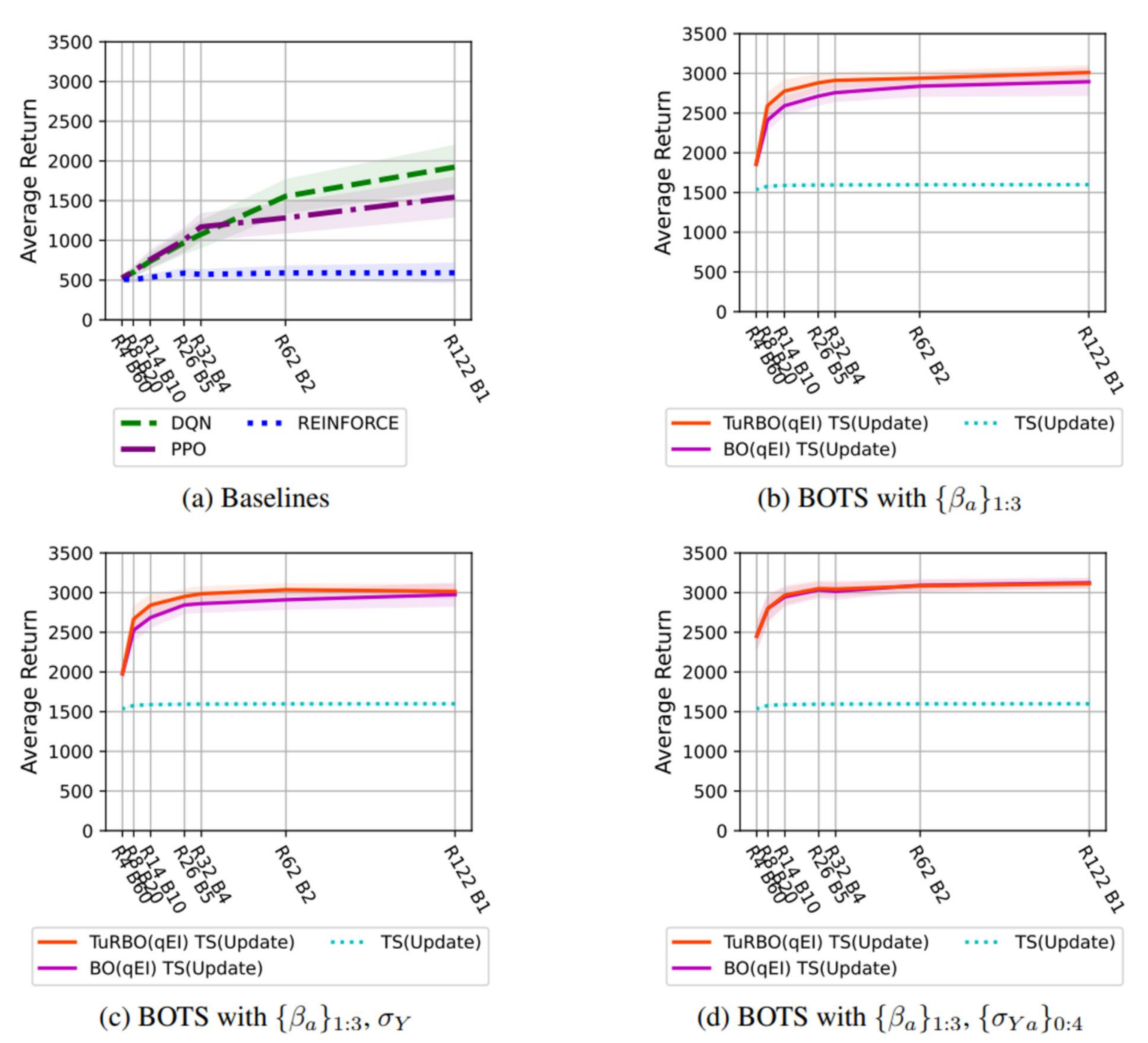
Method	Type	Description
BO(qEI)	global BO + extended TS	Batch BO with qEI acquisition function
TuRBO(qEI)	local BO + extended TS	Batch TuRBO with qEI acquisition function

Table 2: Summary of the BOTS parameter space configuration

Configuration	BO parameters	Description
$\{\beta_a\}_{1:3}$	$[\beta_1, \beta_2, \beta_3]$	action bias terms (for actions 1, 2 and 3)
$\{\beta_a\}_{1:3}, \sigma_Y$	$[\beta_1, \beta_2, \beta_3, \sigma_Y^2]$	action bias terms, shared reward variance
$\{\beta_a\}_{1:3}, \{\sigma_{Ya}\}_{0:4}$	$[\beta_1, \beta_2, \beta_3, \sigma_{Y_0}^2, \sigma_{Y_1}^2, \sigma_{Y_2}^2, \sigma_{Y_3}^2]$	action bias terms, per-action reward variance

Experiments

BOTS offers a practical method for the severely episode-limited RL settings.



(a) Baselines

(b) BOTS with $\{\beta_a\}_{1:3}$

(c) BOTS with $\{\beta_a\}_{1:3}, \sigma_Y$

(d) BOTS with $\{\beta_a\}_{1:3}, \{\sigma_{Ya}\}_{0:4}$

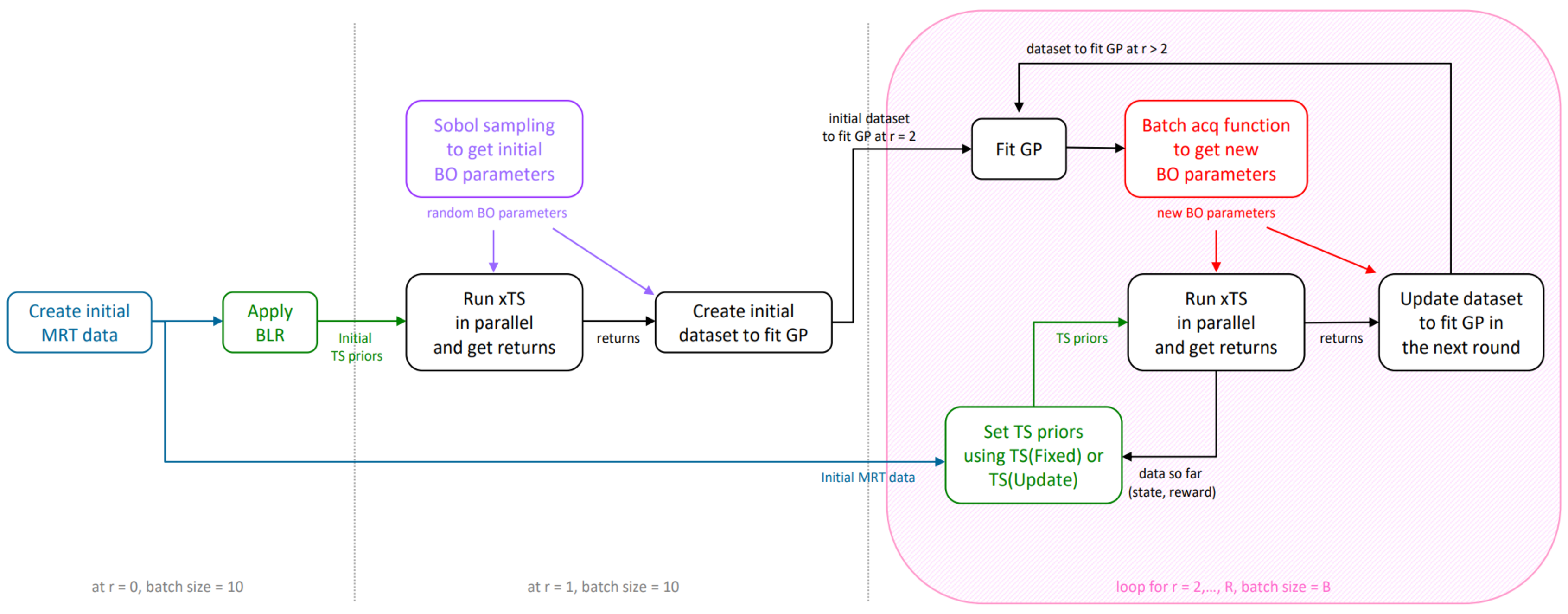
Results for severely episode-limited settings. Note: the x-axes show the tuple (number of rounds, batch size) combinations, not round index.
In all experiments, BOTS shows a better performance in a low number of rounds.

This work is supported by National Institutes of Health National Cancer Institute, Office of Behavior and Social Sciences, and National Institute of Biomedical Imaging and Bioengineering through grants U01CA229445 and 1P41EB028242.

Methods: BOTS overview in the JITAI setting

BOTS is batch Bayesian Optimization (BO) of extended TS.

- BO parameters (i.e., action bias terms βa) are updated via BO.
- TS priors can be initialized using an MRT. Then, TS priors are propagated across rounds using one of two strategies: TS(Fixed) where we fix the same prior for all rounds, or TS(Update) where we update the prior from round to round.



BOTS overview in the JITAI setting, including setting TS initial priors using an MRT.

Extended TS and BOTS Algorithm

We introduce an extended TS, which selects actions via an expected utility that includes fixed action bias terms β_a.

$$u_{ta} = r_{ta} + \boxed{\beta_a}$$
$$p(r_{ta}|a, \mathbf{s}_t) = \mathcal{N}(r_{ta}; \theta_{ta}^\top \mathbf{s}_t, \sigma_{Ya}^2)$$
$$p(\theta_{ta}|\mu_{ta}, \Sigma_{ta}) = \mathcal{N}(\theta_{ta}; \mu_{ta}, \Sigma_{ta})$$

We use batch Bayesian optimization (BO) to learn the action bias terms β_a, to maximize the expected return of the extended TS policy.

Algorithm 1 BOTS: Batch Bayesian Optimization of Extended Thompson Sampling

Inputs $\{B_i\}_{0:R}, \mu_0, \Sigma_0, \sigma_Y^2$:
for $i = 0 : R$ **do**
  Use batch acq function to select β_{ib} for $1 \le b \le B_i$
  for all $b = 1 : B_i$ **do in parallel**
    Run episode using policy $\pi_{xTS}(\beta_{ib}, \mu_0, \Sigma_0, \sigma_Y^2)$
    Obtain return R_{ib}
  end for
  Update GP using $\{(\beta_{ib}, R_{ib}) | 1 \le b \le B_i\}$
end for

Algorithm 2 Extended Thompson Sampling policy π_{xTS}

Inputs $\{\beta_a\}_{0:A}, \{\mu_{0a}\}_{0:A}, \{\Sigma_{0a}\}_{0:A}, \{\sigma_{Ya}^2\}_{0:A}$
for t=0:T **do**
  Observe state $\mathbf{s}_t$
  for a=0:A **do**
    $\hat{\theta}_{ta} \sim \mathcal{N}(\theta_{ta}; \mu_{ta}, \Sigma_{ta})$
    $\hat{r}_{ta} \leftarrow \hat{\theta}_{ta}^\top \mathbf{s}_t$
    $\hat{u}_{ta} \leftarrow \hat{r}_{ta} + \beta_a$
  end for
  $a_t \leftarrow \arg\max_a \hat{u}_{ta}$
  Take action a_t. Observe r_t.
  for a=0:A **do**
    $\Sigma_{(t+1)a} \leftarrow \sigma_{Ya}^2 \left([a_t = a]\mathbf{s}_t^\top \mathbf{s}_t + \sigma_{Ya}^2 \Sigma_{ta}^{-1}\right)^{-1}$
    $\mu_{(t+1)a} \leftarrow \Sigma_{(t+1)a} \left([a_t = a](\sigma_{Ya}^2)^{-1} r_t \mathbf{s}_t + \Sigma_{ta}^{-1} \mu_{ta}\right)$
  end for
end for

Conclusion

- ✓ We introduce BOTS, a novel method that uses extended TS to select actions via an expected utility that includes fixed action bias terms. We use batch Bayesian optimization (BO) to learn the action bias terms over multiple rounds, to maximize the expected return of the extended TS policy.
- ✓ We have also presented a practical method to set the linear reward model priors using a small micro-randomized trial (MRT).
- ✓ Our results show that BOTS outperforms RL methods in severely episode-limited RL settings. BOTS is a promising approach for deployment in real studies.

References

- Karine Karine and Benjamin Marlin (2024). "StepCountJITAI: simulation environment for RL with application to physical activity adaptive intervention". In: NeurIPS 2024 Workshop on Behavioral Machine Learning.
- Karine Karine, Susan Murphy, Benjamin Marlin (2024). "BOTS: Batch Bayesian Optimization of Extended Thompson Sampling for Severely Episode-Limited RL Settings". In: NeurIPS 2024 Workshop on Bayesian Decision-making and Uncertainty.