Diffusion Prior for Online Decision Making: A Case Study of Thompson Sampling

Yu-Guan Hsieh[1], Patrick Bloebaum[2], Shiva Kasiviswanathan[2], Branislav Kveton[2] [1]Université Grenoble Alpes [2]AWS AI Labs

Introduction

TL;DR: We show that diffusion model is able to learn a prior that can helps reduce the regret in multi-armed bandit problems

Multi-Armed Bandits

A model for online decision making:

- Learner pulls arm a_t at round t
- Learner receives rewards r_t drawn from the arm's distribution
- The goal is to maximize the cumulative rewards, i.e., $\sum_t r_t$

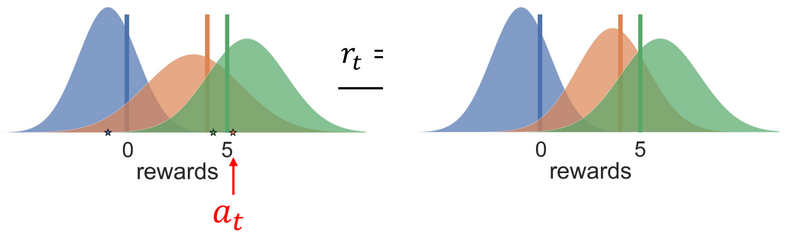
Thompson Sampling (TS)

A Bayesian approach to tackle exploration-exploitation tradeoff

- Given:
  - A prior $p(\mu)$ over mean reward vector μ
  - Interaction history $\mathcal{H}_t = (a_s, r_s)_{s\in\{1,\dots,t\}}$
- Maintain posterior distribution

$$p(w \mid \mathcal{H}_t) \propto p(\mathcal{H}_t \mid w)p(w)$$

- Sample $\tilde{\mu}_t$ from the posterior distribution
- Pull $a_t \in \operatorname{argmax}_{a\in\mathcal{A}} \tilde{\mu}_t^a$ and update posterior

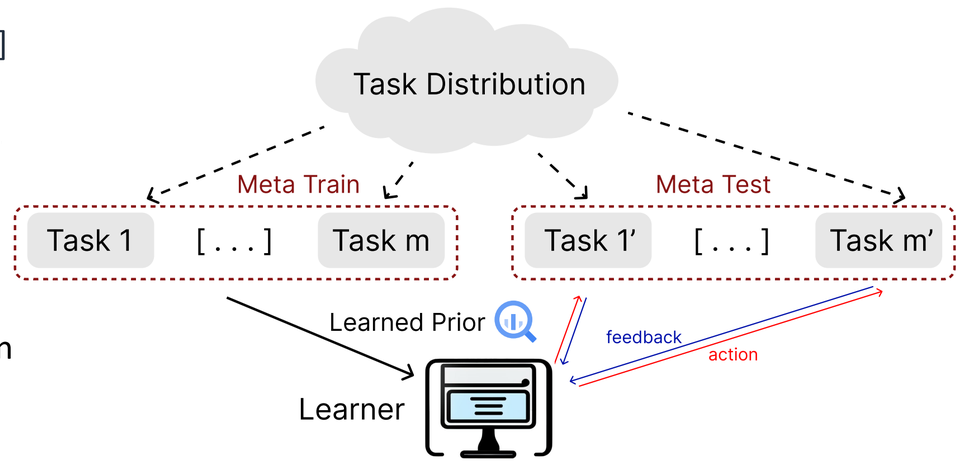


TS is sensitive to the choice of prior:
Can we learn a good prior to minimize regret?

Meta-Learning for Bandits[1]

Different bandit instances can have similar patterns

- Recommend items to different customers
- Assign price to different items using an online pricing algorithm



Goal: Meta-learn a prior from similar tasks using diffusion models

Diffusion Models [2]

- Noise is gradually added in the forward diffusion process

$$q(X_{\ell+1} \mid x_\ell) = \mathcal{N}(X_{\ell+1}; \sqrt{\alpha_{\ell+1}}x_\ell, (1-\alpha_{\ell+1})\mathrm{I})$$

- The model is trained to reverse the process

$$p_\theta(X_\ell \mid x_{\ell+1}) = q(X_\ell \mid x_{\ell+1}, X_0 = h_\theta(x_{\ell+1}, \ell+1))$$

where h_θ is a denoiser and predicts x_0

- **Variance calibration**

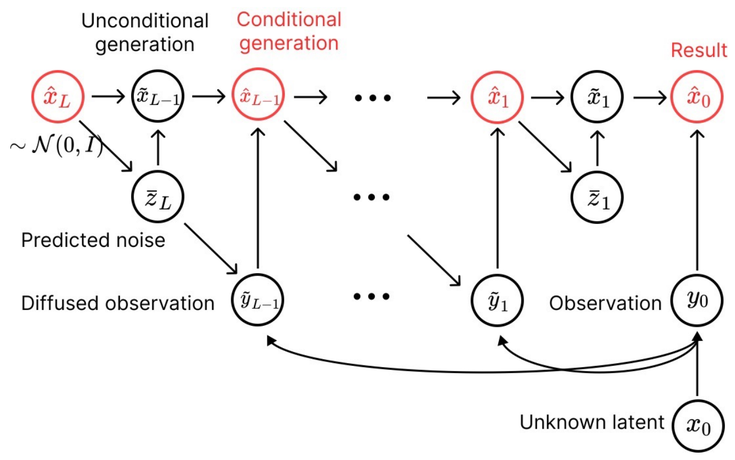
$$p_{\theta,\tau}(X_\ell \mid x_{\ell+1}) = \int q(X_\ell \mid x_{\ell+1}, x_0)p_{\theta,\tau}(x_0 \mid x_{\ell+1})dx_0$$

with $p_{\theta,\tau}(x_0 \mid x_{\ell+1}) = \mathcal{N}(h_\theta(x_{\ell+1}, \ell+1), \tau^2_{\ell+1})$

and mean squared reconstruction error

$$\tau^2_\ell = \mathbb{E}_{x_0,x_\ell}[\|x_0 - h_\theta(x_\ell, \ell)\|^2]$$

TS with Diffusion Prior



Let X_0 denote the vector of expected rewards

The goal is to sample $\tilde{\mu}_t$ from $X_0 \mid \mathcal{H}_t$

1. Sample $\hat{x}_L$ from $\mathcal{N}(0,1)$
2. Sample $\hat{x}_\ell$ from $X_\ell \mid \hat{x}_{\ell+1}, y_0$
   - Arm never pulled: $\hat{x}^a_\ell \sim p_{\theta,\tau}(X^a_\ell \mid \hat{x}_{\ell+1})$
   - Arm pulled: compute empirical mean $\hat{\mu}^a_t$ and

$$\tilde{y}^a_\ell = \sqrt{\bar{\alpha}_\ell}\hat{\mu}^a_t + \sqrt{1-\bar{\alpha}_\ell}\bar{z}^a_{\ell+1}$$

$$\tilde{q}(X^a_\ell \mid \hat{x}_{\ell+1}, \mathcal{H}_t) \propto p_{\theta,\tau}(X^a_\ell \mid \hat{x}_{\ell+1})\,\mathcal{N}(\tilde{y}^a_\ell, \sigma^a_{t,\ell})$$

Sample : $\hat{x}^a_\ell \sim \tilde{q}(X^a_\ell \mid \hat{x}_{\ell+1}, \mathcal{H}_t)$

Numerical Experiments

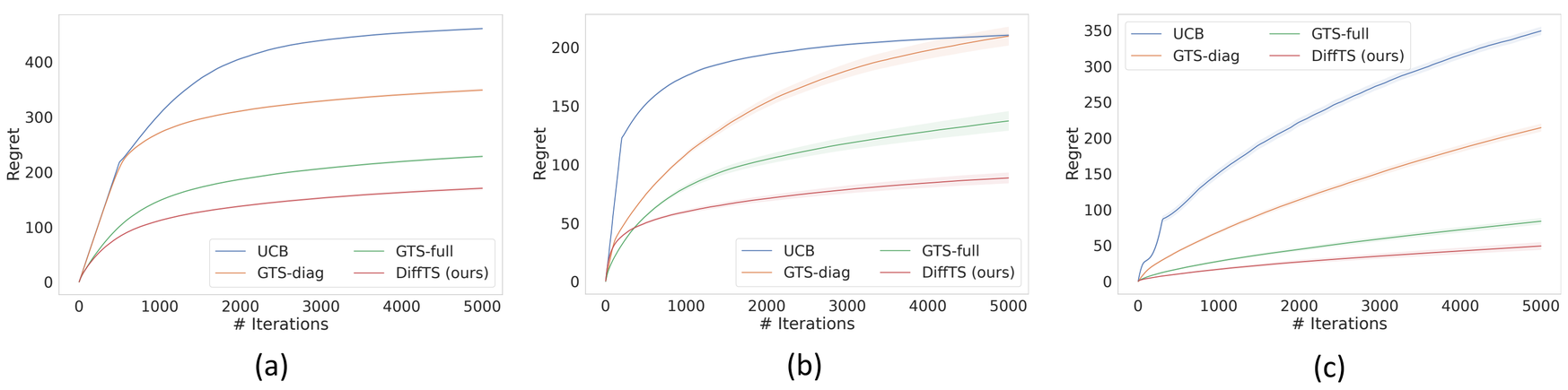
Handcrafted tasks with underlying patterns. We consider three settings:

(a) Arms are labeled and the rewards are related to these labels
(b) Arms are separated into different groups, some groups have higher means but are never optimal (hard for Gaussian)
(c) Modeling the reward of a simple bidding problem with a real-world auction dataset

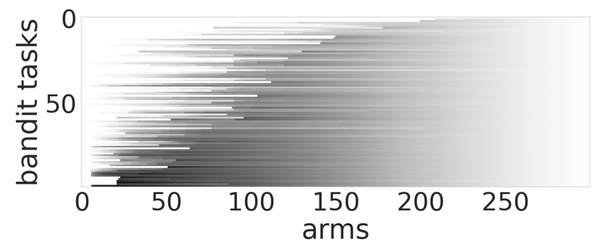
Algorithm Meta Learning for Bandits with Diffusion Models

1: `Meta Training`
2: **Input:** A set of expected means $(\mu_B)_B$ from different tasks $B \sim \mathcal{T}$
3: Train a diffusion model (a denoiser) h_θ to model the distribution of the mean rewards
4: `Calibration`
5: **Input:** A set of expected means $(\mu_B)_B$ from different tasks $B \sim \mathcal{T}$
6: Compute $(\tau^2_\ell)_{\ell\in\{1,\dots,L\}}$ for the denoiser h_θ at different noise levels
7: `Meta Test / Deployment`
8: For any new task B, run Thompson sampling with diffusion prior using the trained model

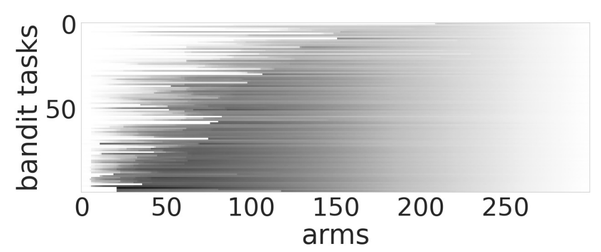
Regret = difference of cumulative rewards between an algorithm and the one that consistently chooses the best action
The algorithms are provided presumed noise variance σ'
The means and covariance of gaussian prior are estimated using training + calibration data



Real data



Generated by our trained diffusion model



Refs. [1] Kveton Branislav et al. Meta Thompson sampling. In ICML, 2021 [2] Jascha Sohl-Dickstein et al. Deep unsupervised learning using nonequilibrium thermodynamics. In ICML, 2015