Gaussian Process Bandits with Aggregated Feedback

We consider the continuum-armed bandits problem [2], where an agent adaptively chooses a sequence of N options from a continuous set (*arm space*) in order to minimise some objective given an oracle that provides noisy observations of the objective evaluated at the options. In this work we consider a novel setting that an oracle that provides noisy average evaluations of the objective over some grid. Our goal is to recommend a local area with the highest possible aggregated reward with a fixed budget of N reward evaluations. Our setting can be motivated by applications where the precise rewards are impossible or expensive to obtain, while an aggregated reward or feedback, such as the average over a subset, is available. For example, in sensor hardware designs, survey sampling methodologies and privacy-preserving data sharing motivate data analysis techniques that account for smooth or average rather than point or instantaneous measurements [5]. A concrete example is the charge-coupled device [1], which when used as an imaging device, consists of a sequence of capacitors whose electrical charge is proportional to the total number of photons incident over an area and time.

We consider one important gap in the literature, best arm(s) identification for continuum-armed bandits with average rewards under a fixed budget. For the problem of black-box optimisation of a function f under single point stochastic feedback, [3] proposed a continuum-armed bandit algorithm called Stochastic Optimistic Optimisation (StoOO) with adaptive hierarchical partitioning of arm space, under the *optimism in the face of uncertainty principle*. For bandits with aggregated feedback, [4] studied finite-armed case for the combinatorial bandits under full-bandit feedback. To the best of our knowledge, we are the first work address the continumm-armed bandits function optimisation problem under aggregated feedback.

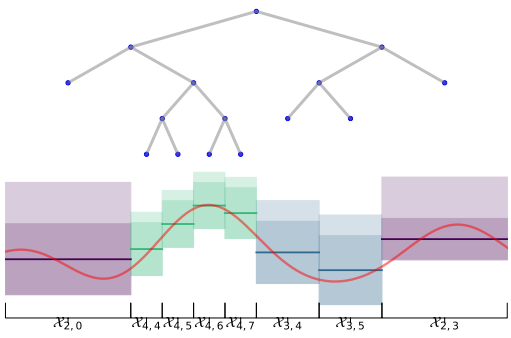

Figure 1: GPOO illustration.

Hierarchically forming arms allows adaptive discretisation over the arm space, which provides a computationally efficient approach for exploring the continuous arm space. Insprised by [3], we propose the Gaussian Process Optimistic Optimisation (GPOO) algorithm for the case where the f is sampled from an unknown Gaussian Process (GP) and the reward feedback is an average over representatives in a subset. Using a GP allows us to encode smoothness assumptions on the function f through a choice of kernel. It also allows us to exploit the closure of Gaussian vectors under affine maps to update our belief of f under aggregated feedback in a Bayesian framework. In Figure 1, GPOO adaptively constructs a tree where the value associated with each node is an estimate of the aggregated reward over a cell. Red shows the reward function to be optimised. Solid horizontal lines show estimated mean aggregated reward. Dark shaded regions shows probable objective function ranges based on Bayesian uncertainty. Light shaded regions additionally account for potential function variation due to smoothness assumptions.

Our **contributions** are (i) a new continuum-armed bandits setting under the aggregated feedback and corresponding new simple regret notion, (ii) the first fixed budget best arms identification algorithm (GPOO) for continuum-armed bandit with noisy average feedback, (iii) theoretical analysis for the proposed algorithm, and (iv) empirical illustrations of the proposed algorithm.

[1] W. S. Boyle and G. E. Smith. Charge coupled semiconductor devices. *The Bell System Technical Journal*, 49(4):587–593, 1970.

[2] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412(19):1832–1852, April 2011.

[3] Rémi Munos. From bandits to monte-carlo tree search: The optimistic principle applied to optimization and planning. 2014.

[4] Idan Rejwan and Yishay Mansour. Top-k combinatorial bandits with full-bandit feedback. In *Algorithmic Learning Theory*, pages 752–776. PMLR, 2020.

[5] Yivan Zhang, Nontawat Charoenphakdee, Zhenguo Wu, and Masashi Sugiyama. Learning from aggregate observations. *Advances in Neural Information Processing Systems*, 33, 2020.