Gaussian Process Planning with Lipschitz Continuous Reward Functions: Towards Unifying Bayesian Optimization, Active Learning, and Beyond

Chun Kai Ling chunkai@comp.nus.edu.sg
Bryan Kian Hsiang Low lowkh@comp.nus.edu.sg
Patrick Jaillet jaillet@mit.edu

NUS National University of Singapore

SMART Singapore-MIT Alliance for Research and Technology

FUTURE URBAN MOBILITY

1 Motivation

- Bayesian Optimization (BO) and some Active Learning (AL) problems often assume Gaussian Process (GP) priors.
- Identifying a common *planning* framework allows us to tackle both problems more effectively by utilizing planning techniques
- Provide basis for **theoretical guarantees** and **non-myopic** decision making
- Identify and solve **novel problems** with similar rewards and priors
- Potential applications: robot exploration in spatially correlated fields, robotic energy harvesting, hyperparameter tuning

2 Gaussian Process Planning (GPP)

Offers Flexibility in Reward fuctions:
- Weak restriction of R to be Lipschitz in Z
- Encompasses several existing formulations
- Generalizes to new interesting tasks

Examples:
- Maximum Entropy Sampling (Shewry, 1987)
- UCB selection criterion (Srinivas et al, 2010)
- Diminishing Rewards: $\log(Z)$ for $Z > 1$, 0 otherwise

$$V_t^\pi(d_t) \triangleq Q_t^\pi(\pi(d_t), \boxed{d_t})$$ Information state

$$Q_t^\pi(s_{t+1}, d_t) \triangleq \mathbb{E}[\underbrace{R(Z_{t+1}, \mathbf{s}_{t+1})}_{\text{Immediate reward}} + \underbrace{V_{t+1}^\pi(\langle \mathbf{s}_{t+1}, \mathbf{z}_t \oplus Z_{t+1}\rangle)}_{\text{Future reward}} \underbrace{| s_{t+1}, d_t]}_{\text{Dependency on past observations}}$$

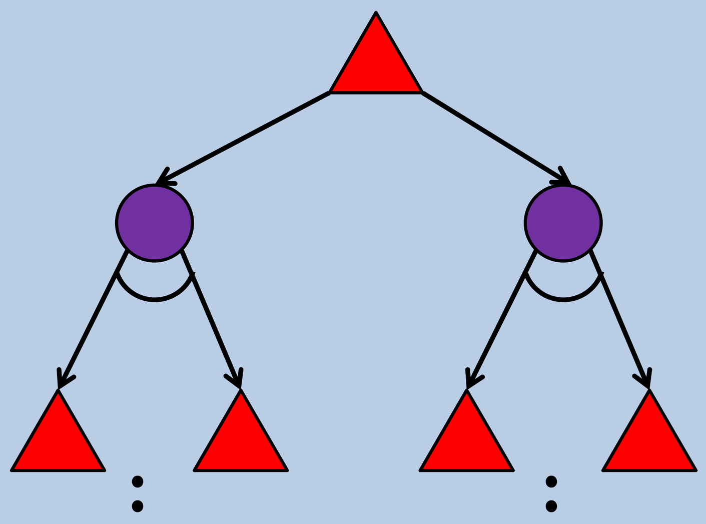
H-stage Bellman Equations
- Resolves exploration-exploitation tradeoff
- Removes the need to explicitly encourage exploration e.g., acquisition functions

3 ϵ-GPP

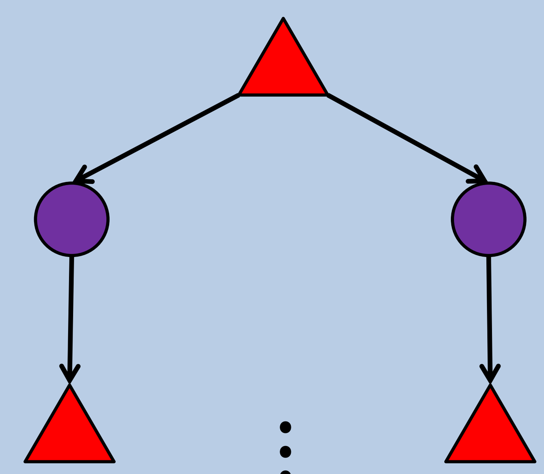
Key idea: By exploiting L1-continuous reward functions, a *finite* tree search may be used to approximate the search over all possible sample observations/actions

Guarantees policy is ϵ-optimal in specified horizon H. An anytime branch and bound extension of ϵ-GPP is proposed to suit real-time applications.

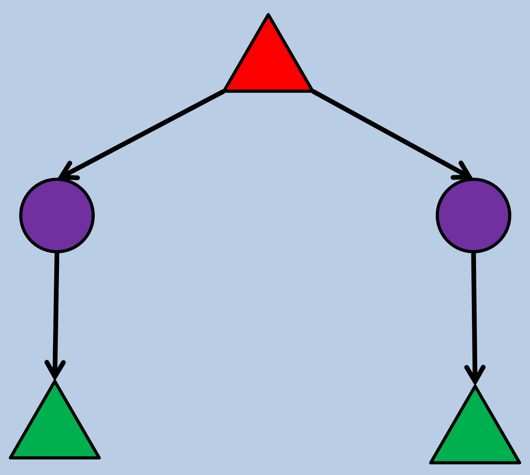
Comparison to existing methods



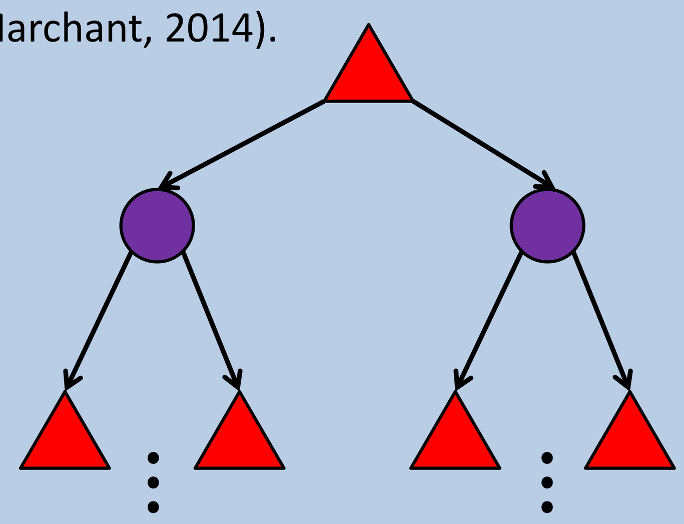
A) Theoretically optimal GPP policy
Observations(purple to red) are unaccountably infinite and is computationally unfeasible.



C) Maximum likelihood observations
The most likely observation is assumed for planning. No performance guarantees in the general case. Acquisition functions may be used to further encourage exploration (Marchant, 2014).



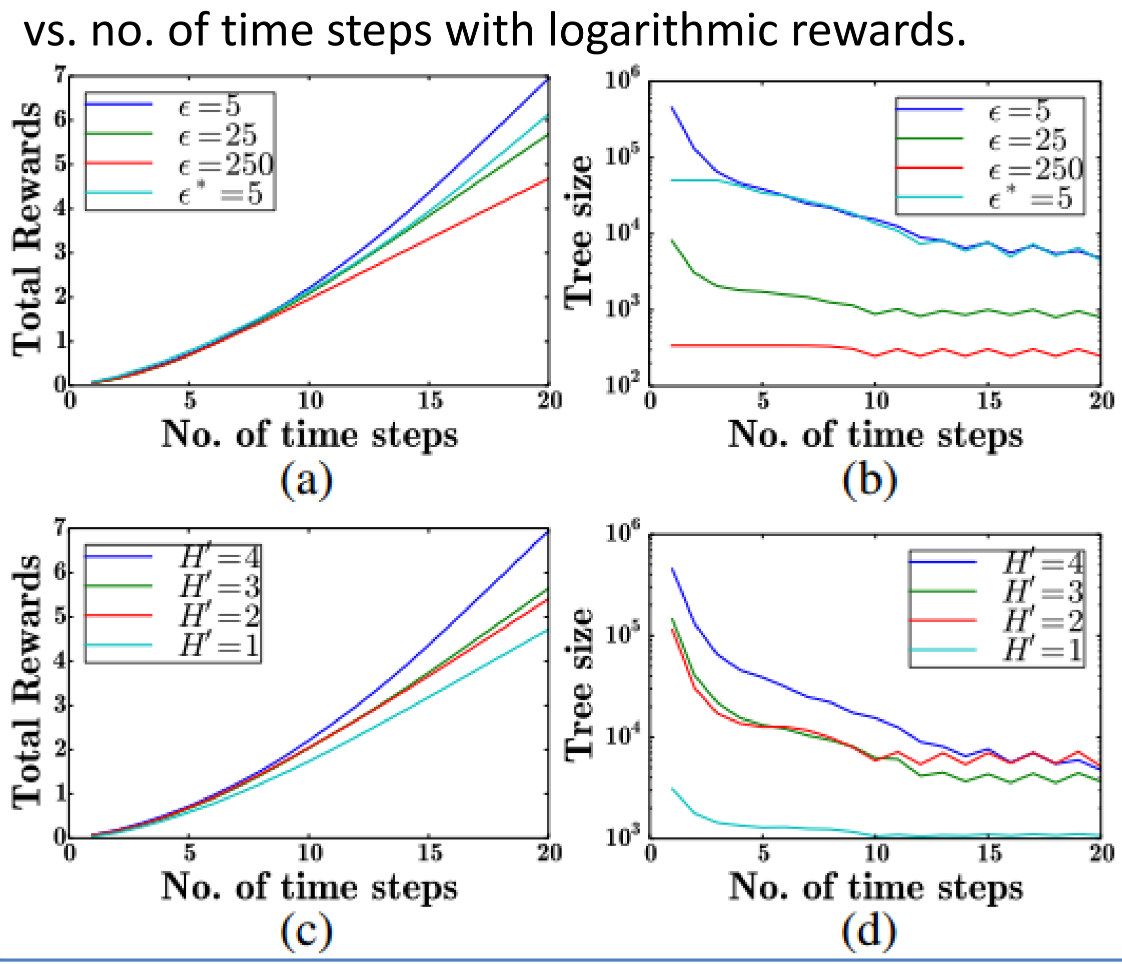
B) Myopic search with surrogate rewards
To avoid search over deeper horizons while encouraging exploration, acquisition functions may be used for planning e.g., GP-UCB.



D) ϵ-GPP
Searches a finite set of observations at each stage. This guarantees near-optimality with provable error bounds.

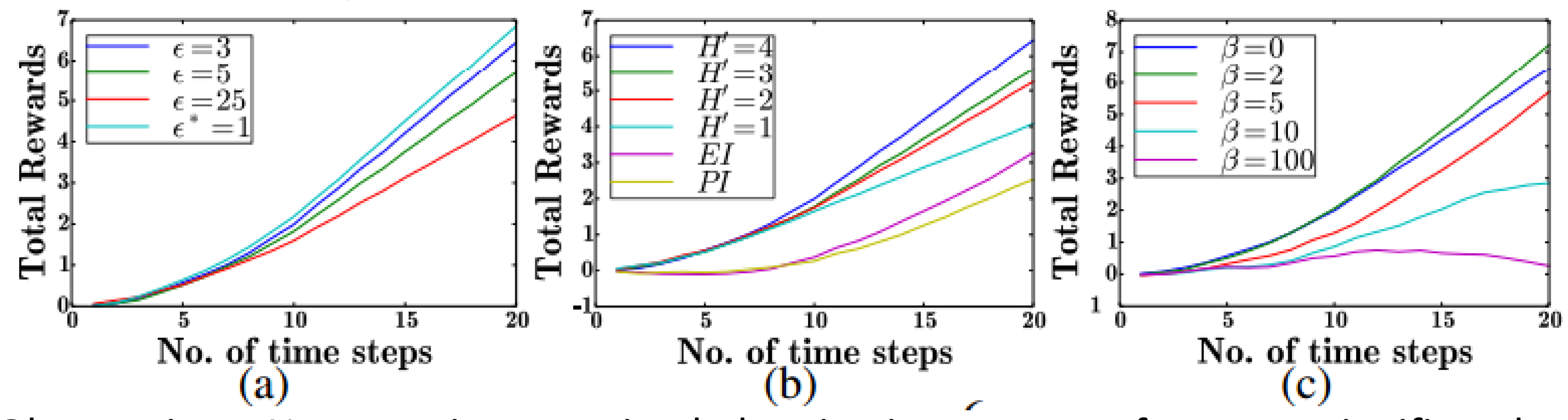
4 Empirical Results

Figure 1: Optimal planning on synthetically generated environments. Graphs of total rewards and tree size of ϵ -GPP policies with (a-b) online planning horizon H' = 4 and varying ϵ and (c-d) varying H' = 1, 2, 3, 4 (respectively, ϵ = 0.002, 0.06, 0.8, 5) vs. no. of time steps with logarithmic rewards.



The plot of ϵ^* = 5 uses our anytime variant with a maximum tree size of 50000 nodes while the plot of ϵ = 250 effectively assumes maximum likelihood observations during planning (Marchant, 2014).

Figure 2: BO on real world log potassium concentration field. Graphs of total normalized rewards of ϵ -GPP policies using UCB-based rewards with (a) H' = 4, β = 0, and varying ϵ, (b) varying H' = 1, 2, 3, 4 (respectively, ϵ = 0.002, 0.003, 0.4, 2) and β = 0, and (c) H' = 4, ϵ = 1, and varying β vs. no. of time steps. The plot of ϵ^* = 1 uses our anytime variant with a maximum tree size of 30000 nodes while the plot of ϵ = 25 effectively assumes maximum likelihood observations.



Observations: Nonmyopic, ϵ –optimal planning improves performance significantly over myopic rewards employing EI/PI. Setting a small value of beta improves performance slightly as exploration is encouraged. However, ϵ-GPP is competitive and does not require tuning of the parameter β.

Acknowledgments: This work was supported by SingaporeMIT Alliance for Research and Technology Subaward Agreement No. 52 R-252-000-550-592.