Adversarial Attacks on Gaussian Process Bandits

Eric Han,[1] Jonathan Scarlett[1,2]

[1]School of Computing, National University of Singapore (NUS)
[2]Department of Mathematics & Institute of Data Science, NUS
eric_han@nus.edu.sg, scarlett@comp.nus.edu.sg

39th International Conference on Machine Learning (Jul 17-23, 2022)

Motivation

GP bandits is the problem of optimizing a black-box function f by using derivative-free queries guided by a GP surrogate model; where f is assumed to be in the RKHS:

$$\max_x f(x).$$

Function observations are typically subject to **corruptions** in the real-world, which are not adequately captured by random noise alone:

1. Rare outliners - i.e. equipment failures,
2. Bad actors - i.e. malicious users.

Related Work

In literature, methods primarily **focused on proposing methods that defend against the proposed uncertainty model** to improve robustness for GP optimization:

- Presence of outliers,
- Random perturbations to sampled points,
- Adversarial perturbations to the final point / samples.

Minimal work studying the problem from an **attacker's perspective**.

Our Goal

Examine from an attacker's perspective, focusing on adversarial perturbations.

Setup

At time t, with random Noise $z_t \sim \mathcal{N}(0, \sigma^2)$, adversarial noise c_t and budget C:

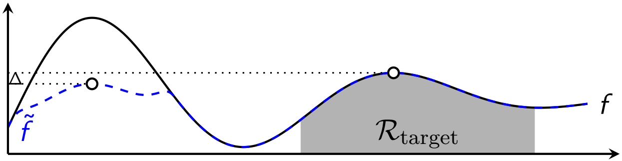
$$y_t = f(\mathbf{x}_t) + c_t + z_t, \qquad \text{where } \sum_{t=1}^{n} |c_t| \leq C.$$

With various levels of knowledge available to the adversary:

1. Targeted Attack - make the player choose actions in a particular region $\mathcal{R}_{\mathrm{target}}$.
2. Untargeted Attack - make the player's cumulative regret as high as possible.

Theoretical Study

Theory applies[1] to **any** algorithm that gets sublinear regret in non-corrupted setting.



Theorem 1 (Rough Sketch)

Adversary performs an attack shifting the original function f to $\tilde{f}$, with sufficient conditions, resulting in linear regret with high probability.

[1] Also even in certain cases where the attacker doesn't know f.

Subtraction Attack (Known f)

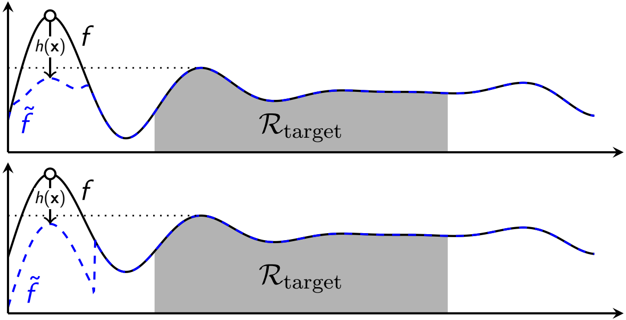
Idea is to 'swallow' the peaks of the function f.

Set $\tilde{f}(\mathbf{x}) = f(\mathbf{x}) - h(\mathbf{x})$, where h:

- Subtraction Rnd - bump fn.
- Subtraction Sq - indicator fn.

Discussion:

1. Strong theoretical guarantees[2].
2. Requiring knowledge of f.
3. Difficult to construct h.



Subtraction Rnd (top) and Sq (bottom).

[2] Only for Subtraction Rnd; depending on the properties of h.

Clipping Attack (Known f)

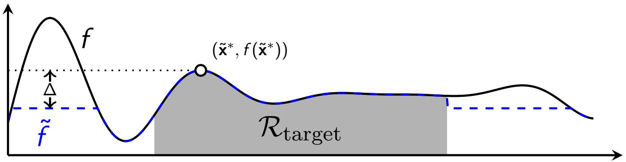
Idea is to 'cut' the rest of the function f off by Δ from the peak in $\mathcal{R}_{\mathrm{target}}$.

Clipping Attack by setting:

$$\tilde{f}(\mathbf{x}) = \begin{cases} f(\mathbf{x}) & \mathbf{x} \in \mathcal{R}_{\mathrm{target}} \\ \min\left\{f(\mathbf{x}), f(\widetilde{\mathbf{x}}^*) - \Delta\right\} & \mathbf{x} \notin \mathcal{R}_{\mathrm{target}}, \end{cases}$$

Discussion:

1. Practical, easy to implement.
2. $\tilde{f}$ not in RKHS.



Aggressive Subtraction Attack (Unknown f)

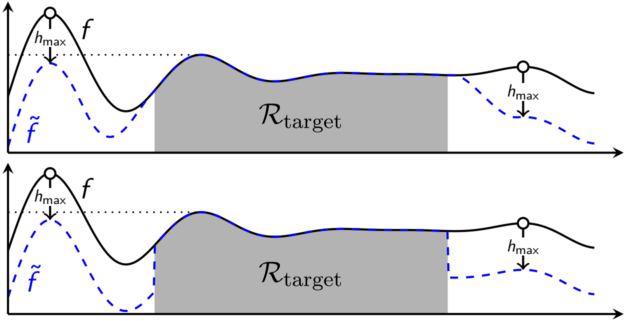
Idea is to subtract *all* points outside $\mathcal{R}_{\text{target}}$ by roughly the same value h_{max}.

Simplified Aggressive Subtraction, without "transition region":

$$\tilde{f}(\mathbf{x}) = \begin{cases} f(\mathbf{x}) & \mathbf{x} \in \mathcal{R}_{\text{target}} \\ f(\mathbf{x}) - h_{\text{max}} & \mathbf{x} \notin \mathcal{R}_{\text{target}}. \end{cases}$$

Discussion:

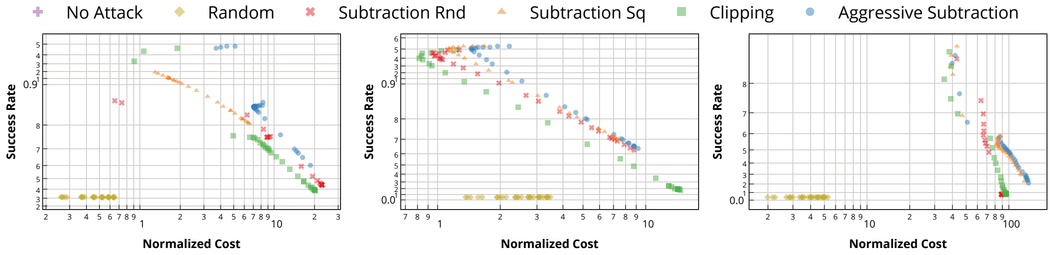
1. Strong theoretical guarantees[3].



With "transition region" (top) and without (bottom).

[3] Only for Aggressive Subtraction with "transition region".

Results for Synthetic1D, Forrester1D, Levy-Hard1D



- Clipping works consistently.
- Aggressive Subtraction works, but with higher cost.
- Subtraction Rnd and Subtraction Sq is ‘in between’.
- Subtraction Rnd tends to narrowly beat Subtraction Sq (due to smooth $h(\mathbf{x})$).

Key Contributions

1. Study conditions under which an adversarial attack can succeed.
2. Present various attacks:
   2.1 Known f: Subtraction Rnd and Subtraction Sq, Clipping Attack.
   2.2 Unknown f: Aggressive Subtraction.

   Demonstrated their effectiveness on a diverse range of objective functions.

Adversarial Attacks on Gaussian Process Bandits
Eric Han, Jonathan Scarlett
ICML 2022
arXiv: https://arxiv.org/abs/2110.08449