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

Highlights

We study robustness for GP optimization from an attacker's perspective, focusing on adversarial perturbations.

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

Demonstrated their effectiveness on a diverse range of functions.

Introduction

GP bandits is the problem of optimizing a black-box function f by using derivative-free queries guided by a GP surrogate model,

$$\max_x f(x).$$

- Function observations can be subject to **corruptions** in the applications, which are not adequately captured by random noise.
- Current literature focused on proposing methods that **defend** against the proposed uncertainty model to improve robustness for GP opt.

Setup: With random noise $z_t \sim \mathcal{N}(0, \sigma^2)$, adversarial noise c_t and attack budget C:

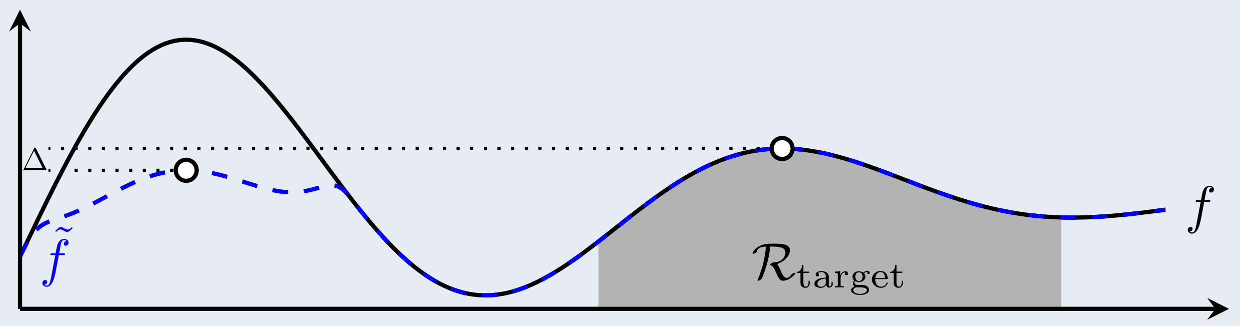
$$y_t = f(\mathbf{x}_t) + c_t + z_t, \qquad \text{where } \sum_{t=1}^{n} |c_t| \le C.$$

- Two distinct attack goals:
  1. **Targeted** - make the player choose actions in a particular $\mathcal{R}_{\text{target}}$.
  2. **Untargeted** - make the player's cumulative regret high.

Theoretical Study

Theorm 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.

- Under sufficient conditions, optimizer finds peak of $\tilde{f}$, so we can bound the number of actions that fall outside $\mathcal{R}_{\text{target}}$,
- Can then bound the budget needed for such perturbation,
- Since $\arg\max f \neq \arg\max \tilde{f}$, regret is linear.



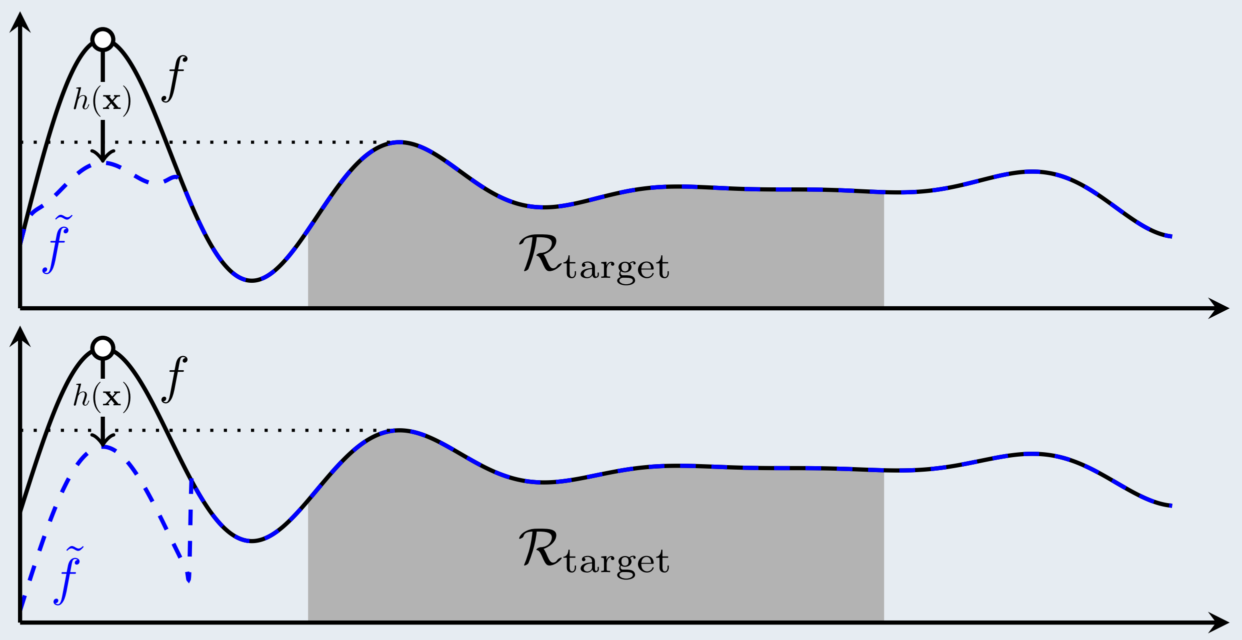
Theory applies (even in certain cases where the attacker doesn't know f) to **any** algorithm that gets sublinear regret in non-corrupted setting.

Attack Methods (Known f)

Subtraction Attack: 'swallow' the peaks of the function f.

$$\tilde{f}(\mathbf{x}) = f(\mathbf{x}) - h(\mathbf{x})$$

- Subtraction Rnd (Top) - Let function h to be a bump fn.
- Subtraction Sq (Bottom) - Let function h to be an indicator fn.

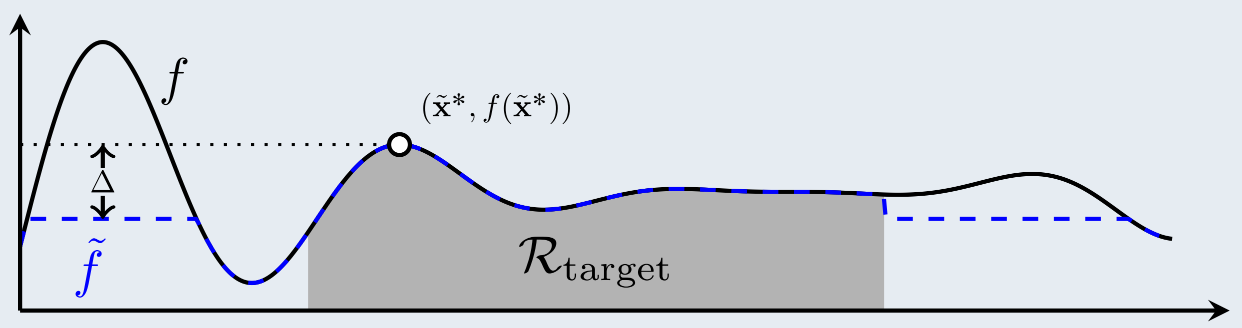


Discussion:

1. Strong theoretical guarantees for Rnd (depends on h).
2. Requiring knowledge of f.
3. Difficult to construct h in practice.

Clipping Attack: 'cut' the rest of the fn f off by Δ from the peak in $\mathcal{R}_{\text{target}}$.

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



Discussion:

1. Practical, easy to implement and performs well.
2. $\tilde{f}$ not in RKHS; our theoretical analysis does not follow.

Further Information

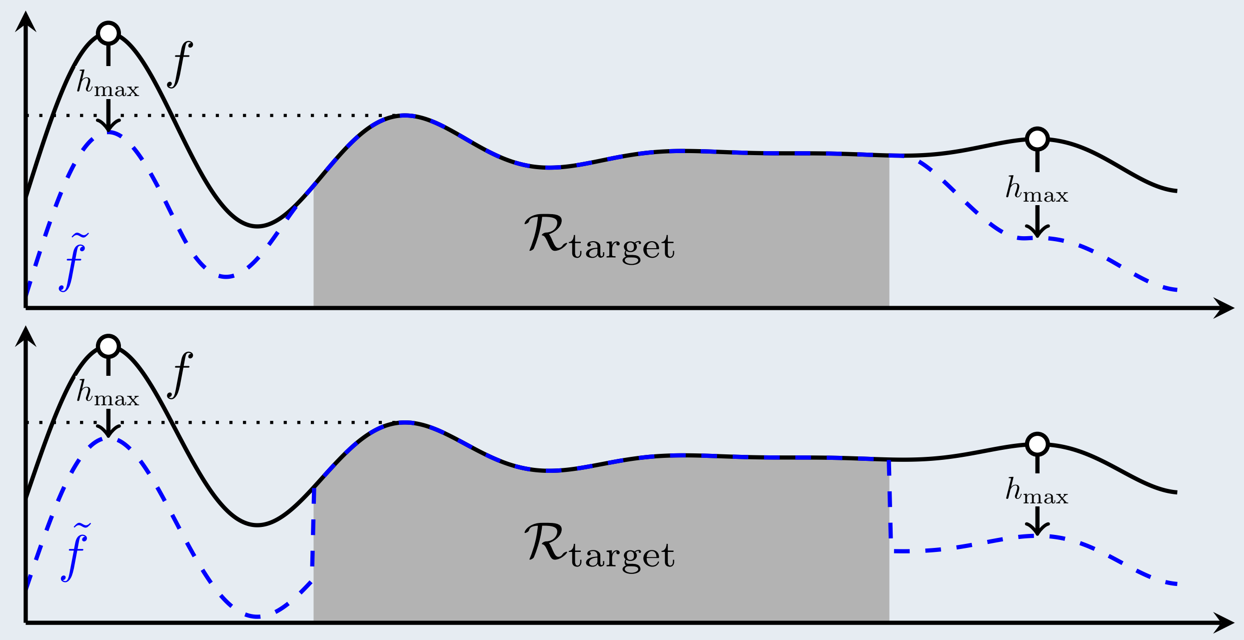
Full paper at https://arxiv.org/abs/2110.08449.

Attack Methods (Unknown f)

Aggressive Subtraction Attack: subtract *all* points outside $\mathcal{R}_{\text{target}}$ by roughly the same value $h_{\max}$.

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

- With 'transition region' (Top) - So that $\tilde{f}$ is smooth to match theory.
- Without (Bottom) - Simplified version used for experiments.



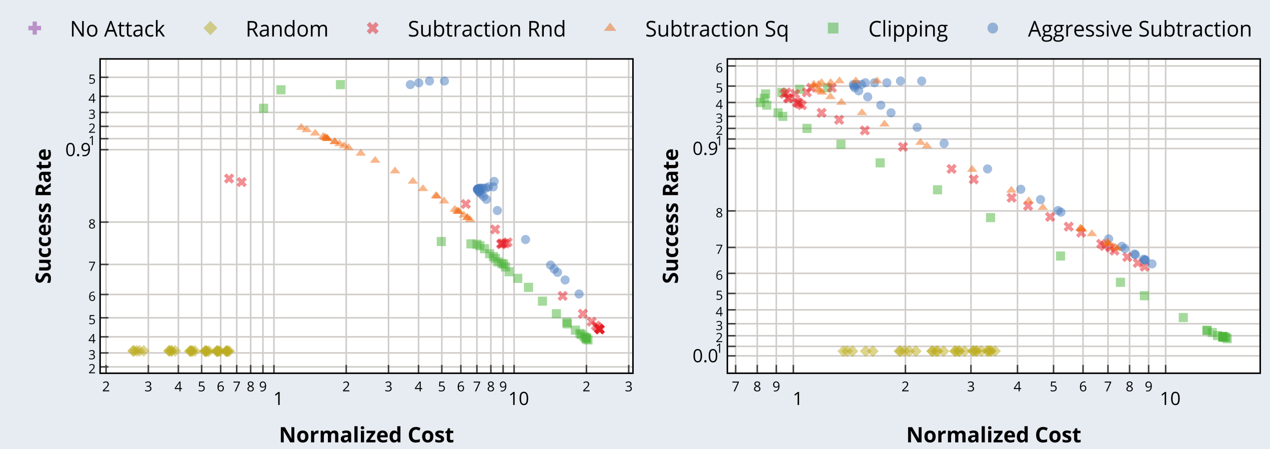
Discussion:

1. Strong theoretical guarantees with 'transition region'.
2. Just knowing that $\mathcal{R}_{\text{target}}$ has a local maximum is sufficient.

Experiments and Results

Each point on the plots correspond a particular attack hyperparameter; averaged over several runs, where the metrics are measured:

$$\text{Success-Rate}(t) = \frac{|\mathcal{R}_{\text{target}} \cap X_t|}{t}, \quad \text{Normalized-Cost}(t) = \sum_{a \in A_t} \frac{a}{f_{\max} - f_{\min}}.$$



Summary of key findings:

- Clipping works consistently.
- Aggressive Subtraction works, but with higher cost.
- Subtraction Rnd and Sq is 'in between', with Rnd narrowly beating Sq.

Additional Experiments can be found in our paper; with more synthetic experiments up to 6 dimensions and robot pushing experiments.