Robust Super-Level Set Estimation using Gaussian Processes

Junzi Zhang

Stanford University, ICME

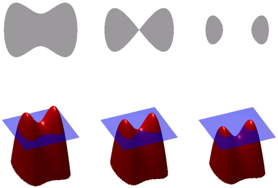
junziz@stanford.edu

Joint work with Andrea Zanette and Mykel J. Kochenderfer

November 6, 2018

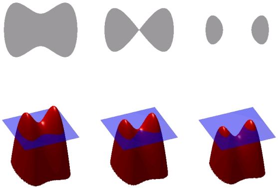
Overview

What is online super level set estimation?



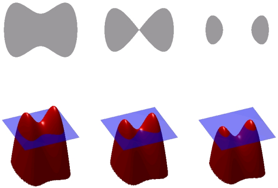
- **Super level set**: Determining the subregion where a function f exceeds a given threshold t, i.e., where $f(x) > t$.

What is online super level set estimation?



- **Super level set**: Determining the subregion where a function f exceeds a given threshold t, i.e., where $f(x) > t$.
- **Estimation**: assume that function evaluations are **costly**, and we only have access to the **noisy observations**: $y(x) = f(x) + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$.

What is online super level set estimation?



- **Super level set**: Determining the subregion where a function f exceeds a given threshold t, i.e., where $f(x) > t$.
- **Estimation**: assume that function evaluations are **costly**, and we only have access to the **noisy observations**: $y(x) = f(x) + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$.
- **Online**: No batch data (i.e., no prior dataset); instead actively collect data and adjust sampling plan based on observations.

Why online super level set estimation? (Applications)

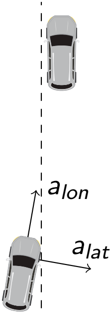
- Often not interested at finding the maximum of $f(\cdot)$, i.e., $f(x) \geq t$ is sufficient.

Why online super level set estimation? (Applications)

- Often not interested at finding the maximum of $f(\cdot)$, i.e., $f(x) \geq t$ is sufficient.
- Surpassing a particular value t indicates that the system meets the requirements, not interested in the actual value of $f(x)$

Why online super level set estimation? (Applications)

- Often not interested at finding the maximum of $f(\cdot)$, i.e., $f(x) \geq t$ is sufficient.
- Surpassing a particular value t indicates that the system meets the requirements, not interested in the actual value of $f(x)$
- Example: what is the minimum performance (e.g., accuracy, reproducibility, ...) of the sensors that will ensure reliable collision avoidance?



Mathematical formulation

- Given a function $f : \Omega \to \mathbb{R}$ and a threshold $t \in \mathbb{R}$, we consider the problem of finding the region Ω such that:

$$P\{f(\mathbf{x}) > t\} \geq 1 - \delta. \tag{1}$$

Mathematical formulation

- Given a function $f : \Omega \to \mathbb{R}$ and a threshold $t \in \mathbb{R}$, we consider the problem of finding the region Ω such that:

$$P\{f(\mathbf{x}) > t\} \geq 1 - \delta. \tag{1}$$

- No gradient information ($f(\cdot)$ is accessed through a black box).

Mathematical formulation

- Given a function $f : \Omega \rightarrow \mathbb{R}$ and a threshold $t \in \mathbb{R}$, we consider the problem of finding the region Ω such that:

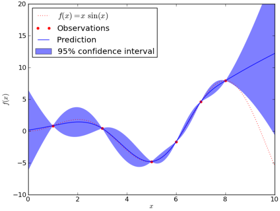
$$P\{f(\mathbf{x}) > t\} \geq 1 - \delta. \tag{1}$$

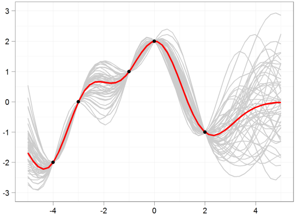
- No gradient information ($f(\cdot)$ is accessed through a black box).
- Model: can obtain noise-corrupted measurements

$$y(\mathbf{x}) = f(\mathbf{x}) + \epsilon$$

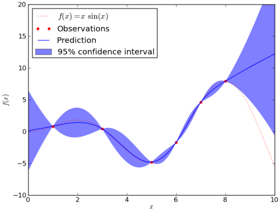
where $\epsilon \sim N(0, \sigma_\epsilon^2)$.

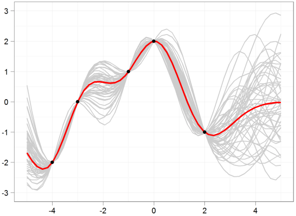
Gaussian Processes





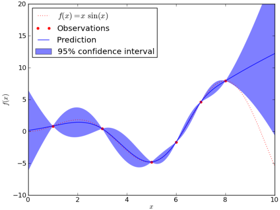
- Data efficiency: we adopt a **Bayesian** framework.

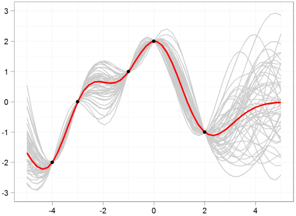




- Data efficiency: we adopt a **Bayesian** framework.
- We model $f(\mathbf{x})$ as a sample from a **Gaussian process** (GP) with prior mean $\mu_0(\mathbf{x})$ and kernel $k_0(\mathbf{x}, \mathbf{x}')$. What if the prior is wrong? more on this later.

Gaussian Processes

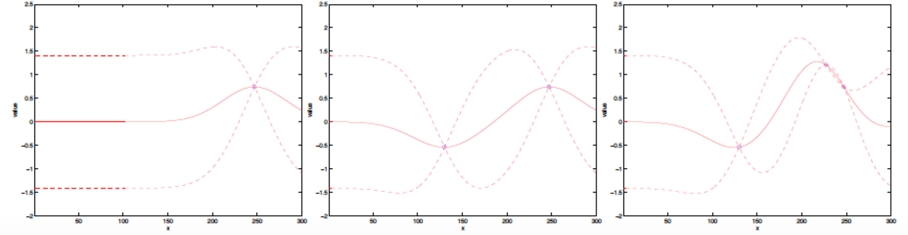




- Data efficiency: we adopt a **Bayesian** framework.
- We model $f(\mathbf{x})$ as a sample from a **Gaussian process** (GP) with prior mean $\mu_0(\mathbf{x})$ and kernel $k_0(\mathbf{x}, \mathbf{x}')$. What if the prior is wrong? more on this later.
- If we query at point $x \in \Omega$, then we obtain a noisy measurement $y = f(x) + \epsilon$, where $\epsilon \sim N(0, \sigma_\epsilon^2)$ are independent noises.

Gaussian Process Update

Illustrative example of GP update: the essence is just **linear algebra** (**Schur complement** computation).



Existing Algorithms

Variance based algorithms:

- **Straddle** (NIPS 2005, no theory): heuristically samples close to threshold t + Exploration (acquisition func: **Straddle score**)

Existing Algorithms

Variance based algorithms:

- **Straddle** (NIPS 2005, no theory): heuristically samples close to threshold t + Exploration (acquisition func: **Straddle score**)
- **LSE** (ICML 2013, Bayesian PAC):
  Identifies the level sets with high probability with some ϵ error (acquisition func: **ambiguity**, generalization of Straddle score)

Existing Algorithms

Variance based algorithms:

- **Straddle** (NIPS 2005, no theory): heuristically samples close to threshold t + Exploration (acquisition func: **Straddle score**)
- **LSE** (ICML 2013, Bayesian PAC):
  Identifies the level sets with high probability with some ϵ error (acquisition func: **ambiguity**, generalization of Straddle score)
- **TruVaR** (NIPS 2016, Bayesian PAC):
  Aims at global variance reduction, similar performance to LSE (acquisition func: truncated variance reduction)

Existing Algorithms (continued)

Volume based algorithms:

- **AAS** (AISTATS 2014, no theory):
  Aims at identifying the a large volume of (super)-level set (similar to our approach), but no convergence theory is established

Existing Algorithms (continued)

Volume based algorithms:

- **AAS** (AISTATS 2014, no theory):
  Aims at identifying the a large volume of (super)-level set (similar to our approach), but no convergence theory is established
- **APPS** (AISTATS 2015, no theory): extension of AAS.

Our Contributions

- Propose a statistically and computationally efficient algorithm to identify the super-level set which is **robust with respect to model misspecification**.

Our Contributions

- Propose a statistically and computationally efficient algorithm to identify the super-level set which is **robust with respect to model misspecification**.
- Usually better numerical performance than state-of-the-art algorithms with **provable exploration guarantees**.

RMILE Algorithm

- Measure the current volume above the threshold:

$$|I_{GP}^{(t)}| = \sum_{\mathbf{x}\in\Omega} \mathbb{1}\left\{P_{GP}\left(f(\mathbf{x}) > t\right) > 1 - \delta\right\}$$

RMILE Algorithm

- Measure the current volume above the threshold:

$$|I_{GP}^{(t)}| = \sum_{\mathbf{x} \in \Omega} \mathbb{1}\left\{P_{GP}\left(f(\mathbf{x}) > t\right) > 1 - \delta\right\}$$

- Choose the next query point $\mathbf{x}^+$ that yields the maximum (expected) improvement:

$$\arg\max_{\mathbf{x}^+} \mathbb{E}_{y^+} \left|I_{GP^+}^{(t)}\right| - \left|I_{GP}^{(t)}\right| = \arg\max_{\mathbf{x}^+} \mathbb{E}_{y^+} \left|I_{GP^+}^{(t)}\right|$$

RMILE Algorithm

- Measure the current volume above the threshold:

$$|I_{GP}^{(t)}| = \sum_{\mathbf{x}\in\Omega} \mathbb{1}\left\{P_{GP}\left(f(\mathbf{x}) > t\right) > 1-\delta\right\}$$

- Choose the next query point $\mathbf{x}^+$ that yields the maximum (expected) improvement:

$$\arg\max_{\mathbf{x}^+} \mathbb{E}_{y^+}\left|I_{GP^+}^{(t)}\right| - \left|I_{GP}^{(t)}\right| = \arg\max_{\mathbf{x}^+} \mathbb{E}_{y^+}\left|I_{GP^+}^{(t)}\right|$$

  Here the expectation is taken with respect to the random outcome y^+ resulting from sampling at $\mathbf{x}^+$ and is conditioned on the filtration up to the current time step.

RMILE Algorithm

- Measure the current volume above the threshold:

$$|I_{GP}^{(t)}| = \sum_{\mathbf{x}\in\Omega} \mathbb{1}\left\{P_{GP}\left(f(\mathbf{x}) > t\right) > 1-\delta\right\}$$

- Choose the next query point $\mathbf{x}^+$ that yields the maximum (expected) improvement:

$$\arg\max_{\mathbf{x}^+} \mathbb{E}_{y^+}\left|I_{GP^+}^{(t)}\right| - \left|I_{GP}^{(t)}\right| = \arg\max_{\mathbf{x}^+} \mathbb{E}_{y^+}\left|I_{GP^+}^{(t)}\right|$$

  Here the expectation is taken with respect to the random outcome y^+ resulting from sampling at $\mathbf{x}^+$ and is conditioned on the filtration up to the current time step.
- **Robustification**: Incorporate an exploration term $\gamma\sigma_{GP}(\mathbf{x}^+)$ in the acquisition function:

$$E_{GP}(\mathbf{x}^+) := \max\{\mathbb{E}_{y^+}\left|I_{GP^+}^{(t)}\right| - \left|I_{GP}^{(t-\epsilon_0)}\right|, \gamma\sigma_{GP}(\mathbf{x}^+)\}$$

- **Acquisition Function**

$$\arg\max_{\mathbf{x}^+} E_{GP}(\mathbf{x}^+) := \arg\max_{\mathbf{x}^+} \max\{\mathbb{E}_{y^+} \left| I_{GP^+}^{(t)} \right| - \left| I_{GP}^{t-\epsilon_0} \right|, \gamma\sigma_{GP}(\mathbf{x}^+)\} \quad (*)$$

where $\gamma > 0$, $\epsilon_0 > 0$ are two small user-defined constants.

Algorithm 1 Robust Max Improvement Level-set Estimation (RMILE)

Input: prior mean μ_0, kernel k_0, objective function
for $i = 1, 2, \ldots$ **do**
  Choose $\mathbf{x}^+$ according to $(*)$
  Query the objective function at $\mathbf{x}^+$ to obtain y^+
  Update $GP^+ \leftarrow GP$ using $(\mathbf{x}^+, y^+)$
Estimate super-level set I_{GP} as $I_{GP} := \{\mathbf{x} \in \Omega \mid P_{GP}(f(\mathbf{x}) > t) > \delta\}$.

Provable exploration guarantees

Lemma (informal)

If RMILE is run on a finite grid, and if a point x is sampled K times, then its RMILE score $E_{GP}(x) = \Omega(1/K)$. In particular, if RMILE is run without termination, then no point is sampled only finitely often.

Provable exploration guarantees

Lemma (informal)

If RMILE is run on a finite grid, and if a point x is sampled K times, then its RMILE score $E_{GP}(x) = \Omega(1/K)$. In particular, if RMILE is run without termination, then no point is sampled only finitely often.

- Maximizing the known volume above the threshold drives more pro-active discovery of the super-level set (**exploitation**)

Provable exploration guarantees

Lemma (informal)

If RMILE is run on a finite grid, and if a point x is sampled K times, then its RMILE score $E_{GP}(x) = \Omega(1/K)$. In particular, if RMILE is run without termination, then no point is sampled only finitely often.

- Maximizing the known volume above the threshold drives more pro-active discovery of the super-level set (**exploitation**)
- Asymptotically all points are sampled infinitely often with the help of the robustification variance term (unknown function is gradually revealed at the grid points) (**exploration**)

Provable exploration guarantees

Lemma (informal)

If RMILE is run on a finite grid, and if a point x is sampled K times, then its RMILE score $E_{GP}(x) = \Omega(1/K)$. In particular, if RMILE is run without termination, then no point is sampled only finitely often.

- Maximizing the known volume above the threshold drives more pro-active discovery of the super-level set (**exploitation**)
- Asymptotically all points are sampled infinitely often with the help of the robustification variance term (unknown function is gradually revealed at the grid points) (**exploration**)
- In some sense, asymptotic convergence occurs even if the initial model is **misspecified** (as is typically the case)

Provable exploration guarantees

Lemma (informal)

If RMILE is run on a finite grid, and if a point x is sampled K times, then its RMILE score $E_{GP}(x) = \Omega(1/K)$. In particular, if RMILE is run without termination, then no point is sampled only finitely often.

- Maximizing the known volume above the threshold drives more pro-active discovery of the super-level set (**exploitation**)
- Asymptotically all points are sampled infinitely often with the help of the robustification variance term (unknown function is gradually revealed at the grid points) (**exploration**)
- In some sense, asymptotic convergence occurs even if the initial model is **misspecified** (as is typically the case)
  - Purely algebraic proof based on taking limits in Schur complement update formula of Gaussian processes

Intuition for why it works

Acquisition Function: $\arg\max_{\mathbf{x}^+} \max\{\mathbb{E}_{y^+} |I_{GP^+}| - |I_{GP}|, \gamma\sigma_{GP}(\mathbf{x}^+)\}$

- initially $\mathbb{E}_{y^+} |I_{GP^+}| - |I_{GP}| \gg \gamma\sigma_{GP}(\mathbf{x}^+)$ because $\gamma > 0$ is small
  - $\mathbb{E}_{y^+} |I_{GP^+}| - |I_{GP}|$ drives **exploitation**
  - coupled with a good prior makes the method efficient

Intuition for why it works

Acquisition Function: $\arg\max_{\mathbf{x}^+} \max\{\mathbb{E}_{y^+} |I_{GP^+}| - |I_{GP}|, \gamma\sigma_{GP}(\mathbf{x}^+)\}$

- initially $\mathbb{E}_{y^+} |I_{GP^+}| - |I_{GP}| \gg \gamma\sigma_{GP}(\mathbf{x}^+)$ because $\gamma > 0$ is small
  - $\mathbb{E}_{y^+} |I_{GP^+}| - |I_{GP}|$ drives **exploitation**
  - coupled with a good prior makes the method efficient
- at some point it may happen that $\mathbb{E}_{y^+} |I_{GP^+}| - |I_{GP}| \lessapprox 0$, i.e., the algorithm is pessimistic about getting new samples. This would make the algorithm stall (i.e., not try new sampling locations)
  - $\mathbb{E}_{y^+} |I_{GP^+}| - |I_{GP}| < \gamma\sigma_{GP}(\mathbf{x}^+)$, so $\gamma\sigma_{GP}(\mathbf{x}^+)$ pushes for **exploration**.

Intuition for why it works

Acquisition Function: $\arg\max_{\mathbf{x}^+} \max\{\mathbb{E}_{y^+} |I_{GP^+}| - |I_{GP}|, \gamma\sigma_{GP}(\mathbf{x}^+)\}$

- initially $\mathbb{E}_{y^+} |I_{GP^+}| - |I_{GP}| \gg \gamma\sigma_{GP}(\mathbf{x}^+)$ because $\gamma > 0$ is small
  - $\mathbb{E}_{y^+} |I_{GP^+}| - |I_{GP}|$ drives **exploitation**
  - coupled with a good prior makes the method efficient
- at some point it may happen that $\mathbb{E}_{y^+} |I_{GP^+}| - |I_{GP}| \lessapprox 0$, i.e., the algorithm is pessimistic about getting new samples. This would make the algorithm stall (i.e., not try new sampling locations)
  - $\mathbb{E}_{y^+} |I_{GP^+}| - |I_{GP}| < \gamma\sigma_{GP}(\mathbf{x}^+)$, so $\gamma\sigma_{GP}(\mathbf{x}^+)$ pushes for **exploration**.
- this robustification modification can work with any acquisition function that satisfies mild conditions, i.e., this approach can be extended beyond the objective of this paper

Necessity for Robustification

- MILE: $\gamma = -\infty$, $\epsilon_0 = 0$;
- RMILE: $\gamma = \epsilon_0 = 1\text{e-}8$ (similar results for other positive γ and ϵ_0).

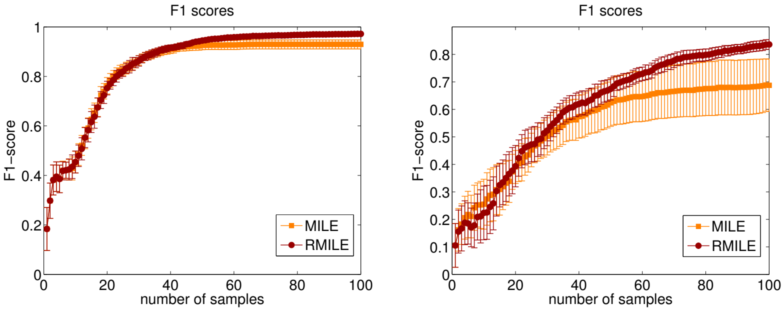


Figure: Himmelblau's function. Left: small noise. Right: large misspecified noise.

Necessity for Robustification (Continued)

A snapshot of intermediate steps:

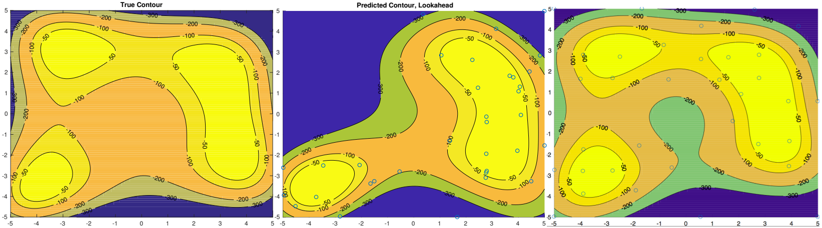


Figure: Left: Himmelbleu's function. Middle: MILE. Right: RMILE.

2D Synthetic Problems

Plotting F_1 score of RMILE vs Straddle[1] and LSE [2].

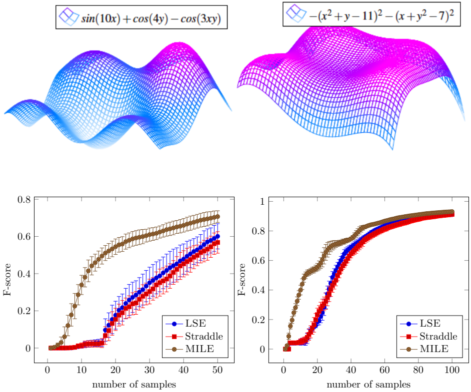


Figure: Sinusoidal function (left), and Himmelblau's function (right).

[1] Bryan et al. Active learning for identifying function threshold boundaries. NIPS 2005

[2] A. Gotovos at al. Active learning for level set estimation. IJCAI 2013

How does the algorithm sample?

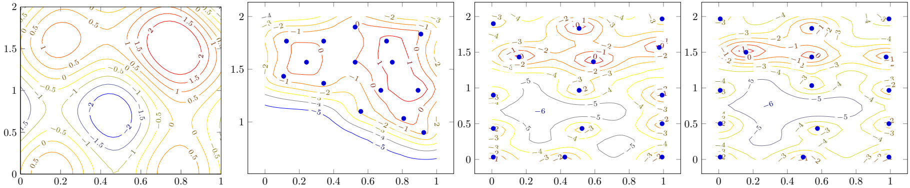
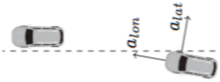


Figure: Far left: true contours for the sinusoidal function. Location of the first 15 samples along with the contours given by the GP for $\mu_{GP}(\mathbf{x}) - 1.96\sigma_{GP}(\mathbf{x})$ for RMILE (middle left), Straddle (middle right) and LSE (far right).

Simulation Problem

Consider estimating actuator performance requirements in an automotive setting. We seek to determine the necessary **precision** for **longitudinal** and **lateral** acceleration maneuvers of simulated vehicles such that the likelihood of **hard braking** events is below a threshold.



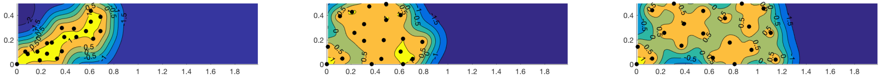


Figure: Contours for $\mu_{GP}(\mathbf{x}) - 1.96\sigma_{GP}(\mathbf{x})$. 20 points budget. RMILE is on left, LSE in the center, and Straddle om the right. The yellow region is the area identified as above the threshold by the Gaussian process.

Summary

- **Objective**: identify regions that meet requirements $f(x) > t$ with high probability.
- To enable high statistical efficiency we use **Gaussian processes**.
- **Provable exploration**: we provide some simple exploration guarantees that address the model misspecification both in theory and in practice.
- **Robustification** also improves practical performance in terms of accuracy.
- **Future directions**: extend this robustification to other acquisition functions as well as to safe exploration.

Thanks for listening!