Local policy search with Bayesian optimization

Sarah Müller *[1,4] Alexander von Rohr *[1,2,3] Sebastian Trimpe [1,2]

[1]Max Planck Institute for Intelligent Systems, Stuttgart, Germany

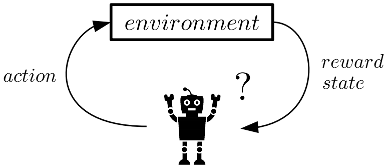
[2]Institute for Data Science in Mechanical Engineering, RWTH Aachen University, Germany

[3]IAV GmbH, Gifhorn, Germany

[4]Institute for Ophthalmic Research, University of Tübingen, Tübingen, Germany

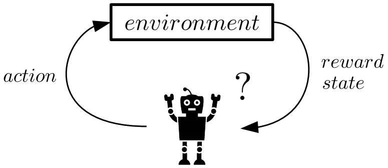
* **Equal contribution**

- Main principle in reinforcement learning: Find an optimal policy by interaction with an environment.

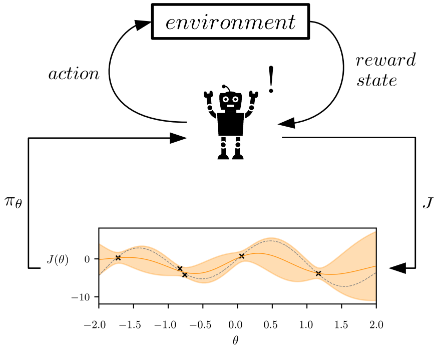


Motivation

- Main principle in reinforcement learning: Find an optimal policy by interaction with an environment.
- Local **gradient-based** policy optimization achieves state-of-the-art performance.
  - Exploration is usually done via random samples.

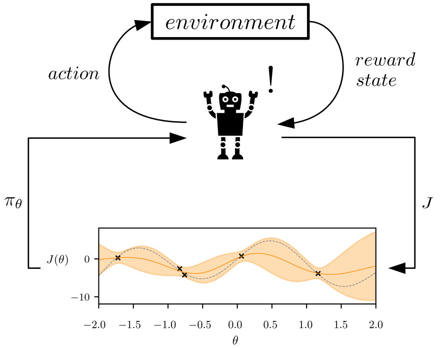


- Main principle in reinforcement learning: Find an optimal policy by interaction with an environment.
- Local **gradient-based** policy optimization achieves state-of-the-art performance.
    - Exploration is usually done via random samples.
- Global **Bayesian optimization** (BO) promises sample-efficient optimization through active exploration.
    - Global optimization in high-dimensional search spaces is a challenging problem to solve.



Motivation

- Main principle in reinforcement learning: Find an optimal policy by interaction with an environment.
- Local **gradient-based** policy optimization achieves state-of-the-art performance.
  - Exploration is usually done via random samples.
- Global **Bayesian optimization** (BO) promises sample-efficient optimization through active exploration.
  - Global optimization in high-dimensional search spaces is a challenging problem to solve.
- Our proposed algorithm (GIBO) reduces gradient uncertainty through active sampling.
  - GIBO improves sample-efficiency of gradient-based methods compared to non-active sampling baselines.



Policy search & Bayesian optimization

Policy search

- Find a *local* optimal policy in the space that maps policy parameters to their episodic reward:

$$J(\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{i=0}^{I} r_i\right].$$

- Update parameter with gradient-based optimizer:

$$\theta_{t+1} = \theta_t + \eta \cdot \nabla_\theta J\big|_{\theta=\theta_t}.$$

Policy search & Bayesian optimization

Policy search

- Find a *local* optimal policy in the space that maps policy parameters to their episodic reward:

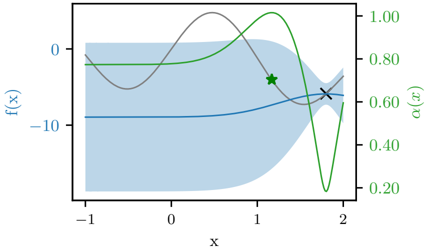
$$J(\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{i=0}^{I} r_i\right].$$

- Update parameter with gradient-based optimizer:

$$\theta_{t+1} = \theta_t + \eta \cdot \nabla_\theta J\big|_{\theta=\theta_t}.$$

Bayesian optimization

- *Global* black-box optimization method.
- Probabilistic model of the objective function $f(x)$, e.g. Gaussian process (GP).
- Acquisition function $\alpha(x)$ that determines points with the most information for the global optimum.



Policy search

- Find a *local* optimal policy in the space that maps policy parameters to their episodic reward:

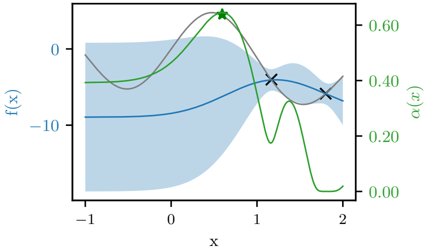
$$J(\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{i=0}^{I} r_i\right].$$

- Update parameter with gradient-based optimizer:

$$\theta_{t+1} = \theta_t + \eta \cdot \nabla_\theta J\big|_{\theta=\theta_t}.$$

Bayesian optimization

- *Global* black-box optimization method.
- Probabilistic model of the objective function $f(x)$, e.g. Gaussian process (GP).
- Acquisition function $\alpha(x)$ that determines points with the most information for the global optimum.



Policy search

- Find a *local* optimal policy in the space that maps policy parameters to their episodic reward:

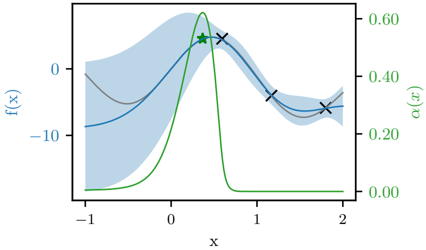
$$J(\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{i=0}^{I} r_i\right].$$

- Update parameter with gradient-based optimizer:

$$\theta_{t+1} = \theta_t + \eta \cdot \nabla_\theta J\big|_{\theta=\theta_t}.$$

Bayesian optimization

- *Global* black-box optimization method.
- Probabilistic model of the objective function $f(x)$, e.g. Gaussian process (GP).
- Acquisition function $\alpha(x)$ that determines points with the most information for the global optimum.



Policy search

- Find a *local* optimal policy in the space that maps policy parameters to their episodic reward:

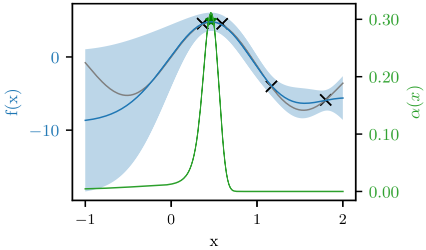
$$J(\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{i=0}^{I} r_i\right].$$

- Update parameter with gradient-based optimizer:

$$\theta_{t+1} = \theta_t + \eta \cdot \nabla_\theta J\big|_{\theta=\theta_t}.$$

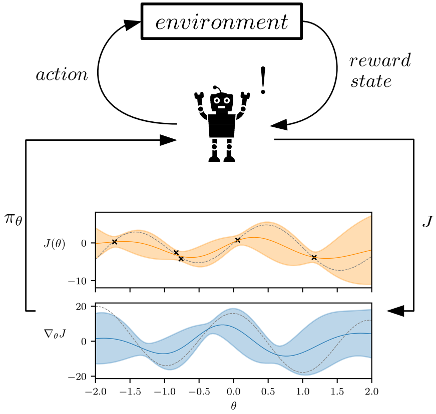
Bayesian optimization

- *Global* black-box optimization method.
- Probabilistic model of the objective function $f(x)$, e.g. Gaussian process (GP).
- Acquisition function $\alpha(x)$ that determines points with the most information for the global optimum.



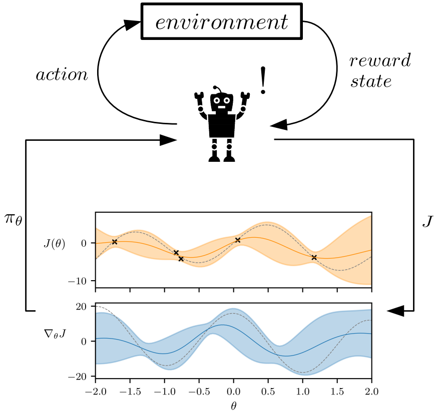
Gradient Information with BO

- Combine the strengths of both worlds: Local search can handle high-dimensional search spaces and global BO is sample-efficient with active exploration.



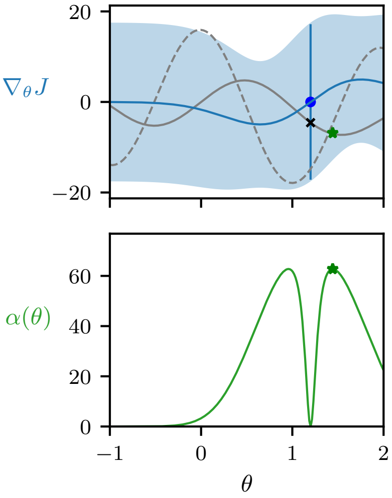
Gradient Information with BO

- Combine the strengths of both worlds: Local search can handle high-dimensional search spaces and global BO is sample-efficient with active exploration.
- Probabilistic surrogate model of objective function $J(\theta)$ and its Jacobian $\nabla_\theta J$.



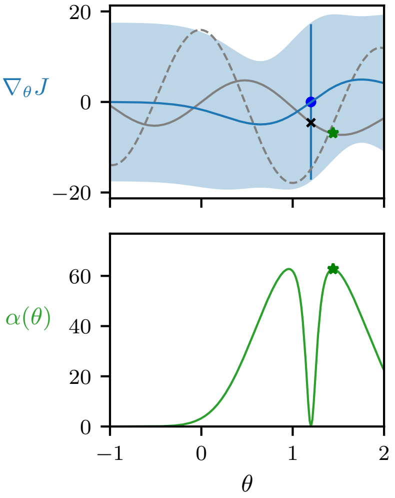
Gradient Information with BO

- Combine the strengths of both worlds: Local search can handle high-dimensional search spaces and global BO is sample-efficient with active exploration.
- Probabilistic surrogate model of objective function $J(\theta)$ and its Jacobian $\nabla_\theta J$.
- Acquisition function $\alpha(\theta)$ that determines points for an accurate gradient estimate at the current point θ_t.
  - Measures the decrease in the Jacobian's variance at θ_t when observing a new point θ.



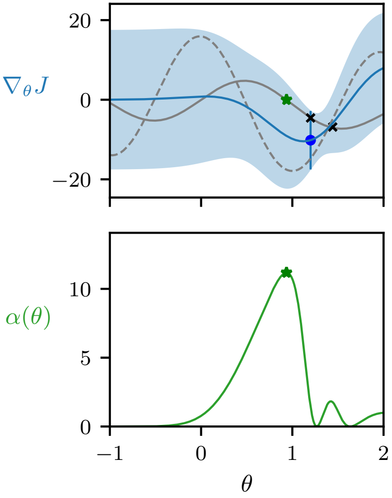
Gradient Information with BO

- Combine the strengths of both worlds: Local search can handle high-dimensional search spaces and global BO is sample-efficient with active exploration.
- Probabilistic surrogate model of objective function $J(\theta)$ and its Jacobian $\nabla_\theta J$.
- Acquisition function $\alpha(\theta)$ that determines points for an accurate gradient estimate at the current point θ_t.
    - Measures the decrease in the Jacobian's variance at θ_t when observing a new point θ.
- Iterative algorithm:
    1. Sample points with acquisition function for a gradient estimate.
    2. Update with gradient based optimizer.



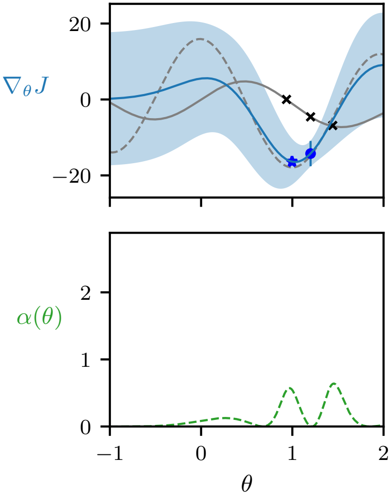
Gradient Information with BO

- Combine the strengths of both worlds: Local search can handle high-dimensional search spaces and global BO is sample-efficient with active exploration.
- Probabilistic surrogate model of objective function $J(\theta)$ and its Jacobian $\nabla_\theta J$.
- Acquisition function $\alpha(\theta)$ that determines points for an accurate gradient estimate at the current point θ_t.
  - Measures the decrease in the Jacobian's variance at θ_t when observing a new point θ.
- Iterative algorithm:
  1. Sample points with acquisition function for a gradient estimate.
  2. Update with gradient based optimizer.



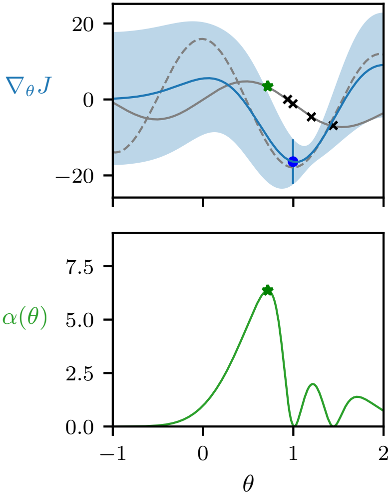
Gradient Information with BO

- Combine the strengths of both worlds: Local search can handle high-dimensional search spaces and global BO is sample-efficient with active exploration.
- Probabilistic surrogate model of objective function $J(\theta)$ and its Jacobian $\nabla_\theta J$.
- Acquisition function $\alpha(\theta)$ that determines points for an accurate gradient estimate at the current point θ_t.
  - Measures the decrease in the Jacobian's variance at θ_t when observing a new point θ.
- Iterative algorithm:
  1. Sample points with acquisition function for a gradient estimate.
  2. Update with gradient based optimizer.



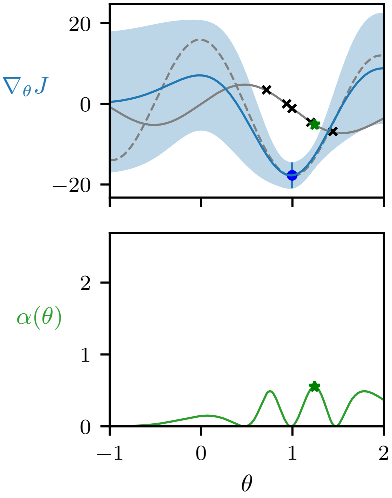
Gradient Information with BO

- Combine the strengths of both worlds: Local search can handle high-dimensional search spaces and global BO is sample-efficient with active exploration.
- Probabilistic surrogate model of objective function $J(\theta)$ and its Jacobian $\nabla_\theta J$.
- Acquisition function $\alpha(\theta)$ that determines points for an accurate gradient estimate at the current point θ_t.
  - Measures the decrease in the Jacobian's variance at θ_t when observing a new point θ.
- Iterative algorithm:
  1. Sample points with acquisition function for a gradient estimate.
  2. Update with gradient based optimizer.



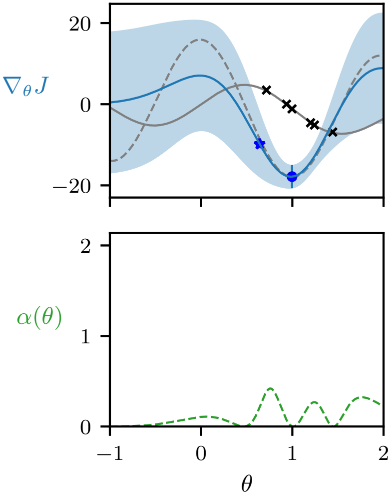
Gradient Information with BO

- Combine the strengths of both worlds: Local search can handle high-dimensional search spaces and global BO is sample-efficient with active exploration.
- Probabilistic surrogate model of objective function $J(\theta)$ and its Jacobian $\nabla_\theta J$.
- Acquisition function $\alpha(\theta)$ that determines points for an accurate gradient estimate at the current point θ_t.
  - Measures the decrease in the Jacobian's variance at θ_t when observing a new point θ.
- Iterative algorithm:
  1. Sample points with acquisition function for a gradient estimate.
  2. Update with gradient based optimizer.



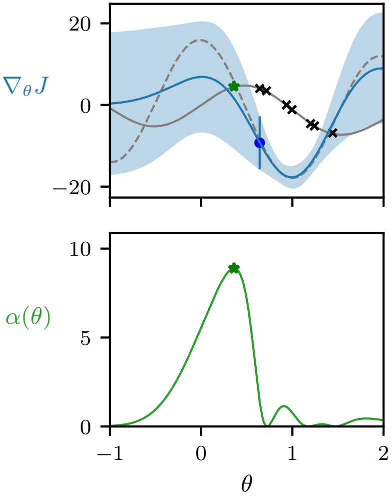
Gradient Information with BO

- Combine the strengths of both worlds: Local search can handle high-dimensional search spaces and global BO is sample-efficient with active exploration.
- Probabilistic surrogate model of objective function $J(\theta)$ and its Jacobian $\nabla_\theta J$.
- Acquisition function $\alpha(\theta)$ that determines points for an accurate gradient estimate at the current point θ_t.
  - Measures the decrease in the Jacobian's variance at θ_t when observing a new point θ.
- Iterative algorithm:
  1. Sample points with acquisition function for a gradient estimate.
  2. Update with gradient based optimizer.



Gradient Information with BO

- Combine the strengths of both worlds: Local search can handle high-dimensional search spaces and global BO is sample-efficient with active exploration.
- Probabilistic surrogate model of objective function $J(\theta)$ and its Jacobian $\nabla_\theta J$.
- Acquisition function $\alpha(\theta)$ that determines points for an accurate gradient estimate at the current point θ_t.
  - Measures the decrease in the Jacobian's variance at θ_t when observing a new point θ.
- Iterative algorithm:
  1. Sample points with acquisition function for a gradient estimate.
  2. Update with gradient based optimizer.



Acquisition function

- Measures the decrease in the Jacobian's variance at θ_t when observing a new point θ:

$$\alpha(\theta \mid \theta_t, \mathcal{D}) = \mathbb{E}\left[\mathrm{Tr}\left(\Sigma'(\theta_t \mid \mathcal{D})\right) - \mathrm{Tr}\left(\Sigma'\left(\theta_t \mid \{\mathcal{D}, (\theta, y)\}\right)\right)\right].$$

- Expected difference between the Jacobian's variance $\Sigma'(\theta_t \mid \mathcal{D})$ *before* and the Jacobian's variance $\Sigma'(\theta_t \mid \{\mathcal{D}, (\theta, y)\})$ *after* observing a new point (θ, y).
- Where $\Sigma'(\theta_t \mid \mathcal{D})$ is the variance of the Jacobian's GP model evaluated at θ_t.

Acquisition function

- Measures the decrease in the Jacobian's variance at θ_t when observing a new point θ:

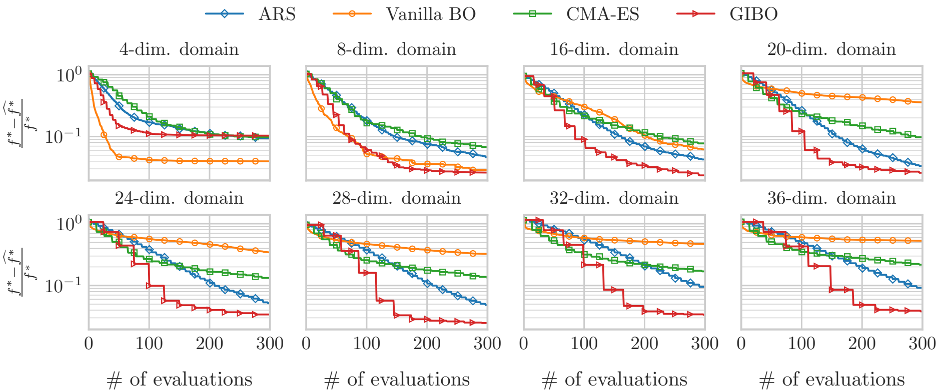
$$\alpha(\theta \,|\, \theta_t, \mathcal{D}) = \mathbb{E}\left[\mathrm{Tr}\left(\Sigma'(\theta_t \,|\, \mathcal{D})\right) - \mathrm{Tr}\left(\Sigma'\left(\theta_t \,|\, \{\mathcal{D}, (\theta, y)\}\right)\right)\right].$$

- Expected difference between the Jacobian's variance $\Sigma'(\theta_t \,|\, \mathcal{D})$ *before* and the Jacobian's variance $\Sigma'(\theta_t \,|\, \{\mathcal{D}, (\theta, y)\})$ *after* observing a new point (θ, y).
- Where $\Sigma'(\theta_t \,|\, \mathcal{D})$ is the variance of the Jacobian's GP model evaluated at θ_t.
- A property of the Gaussian distribution is that the covariance function is independent of the observed targets y:

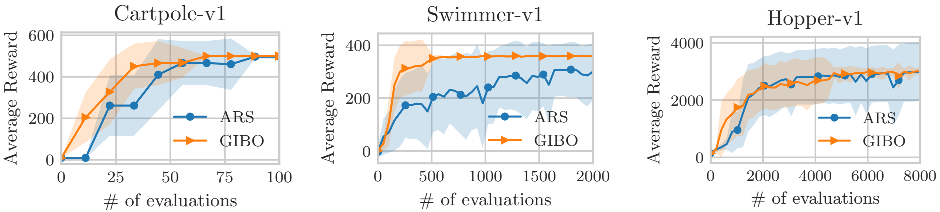
$$\arg\max_{\theta} \alpha(\theta \,|\, \theta_t, \mathcal{D}) = \arg\min_{\theta} \mathrm{Tr}\left(\Sigma'\left(\theta_t \,|\, [X, \theta]\right)\right),$$

with the virtual data set $[\theta_1, \ldots, \theta_n, \theta] =: [X, \theta]$.

Synthetic experiments



Gym and MuJoCo



Summary and contributions

- Novel policy search algorithm that combines
  - active sampling,
  - surrogate modeling,
  - local search with approximate gradient descent.

Summary and contributions

- Novel policy search algorithm that combines
  - active sampling,
  - surrogate modeling,
  - local search with approximate gradient descent.
- Contributions
  - Significantly improved sample complexity on **synthetic objective functions**.
  - Solved RL benchmarks in a **sample efficient** manner.
  - **Reduce reward variance** compared to non-active sampling baselines.