Multi-Fidelity Bayesian Optimization with Unreliable Information Sources

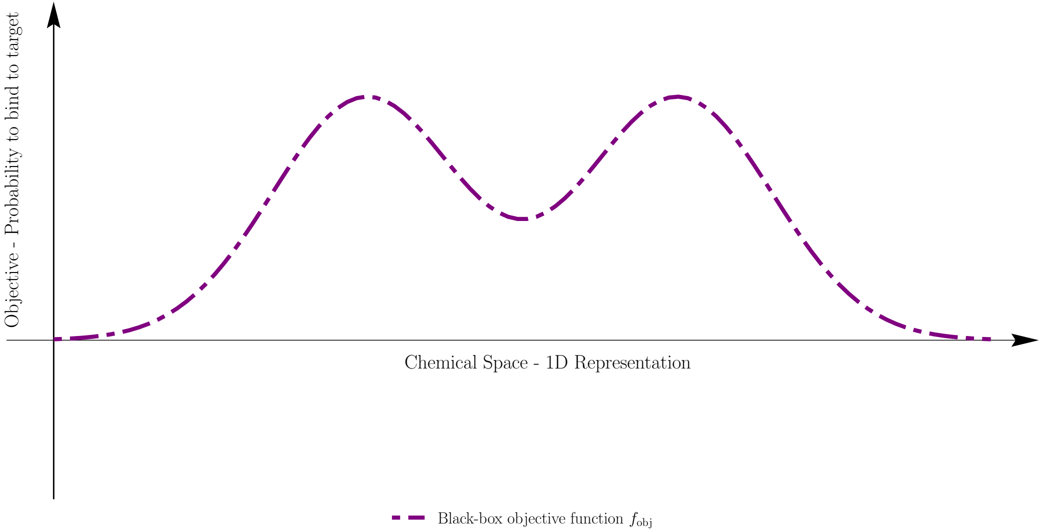
Petrus Mikkola, **Julien Martinelli**, Louis Filstroff, Samuel Kaski

Aalto University

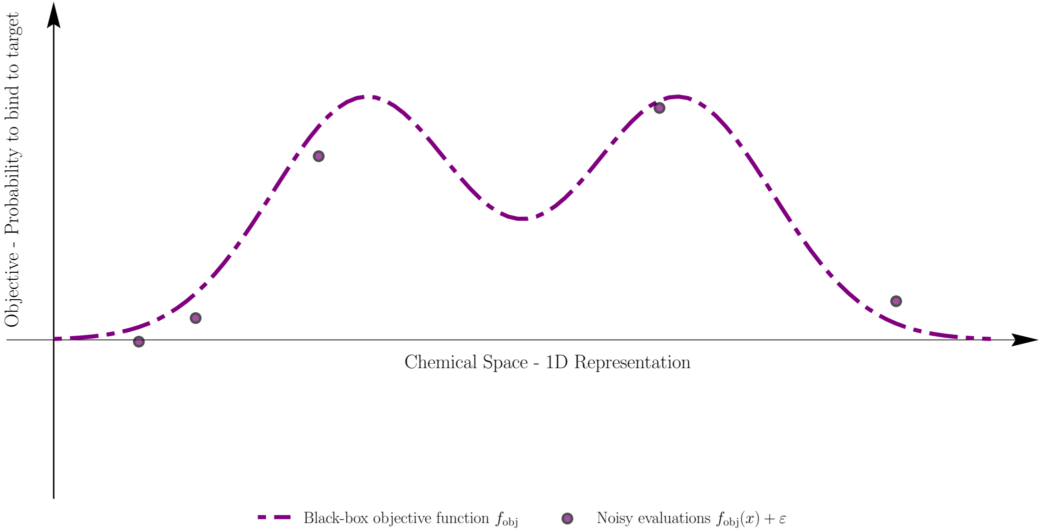
May 15^{th}, 2023

Bayesian Optimization 101

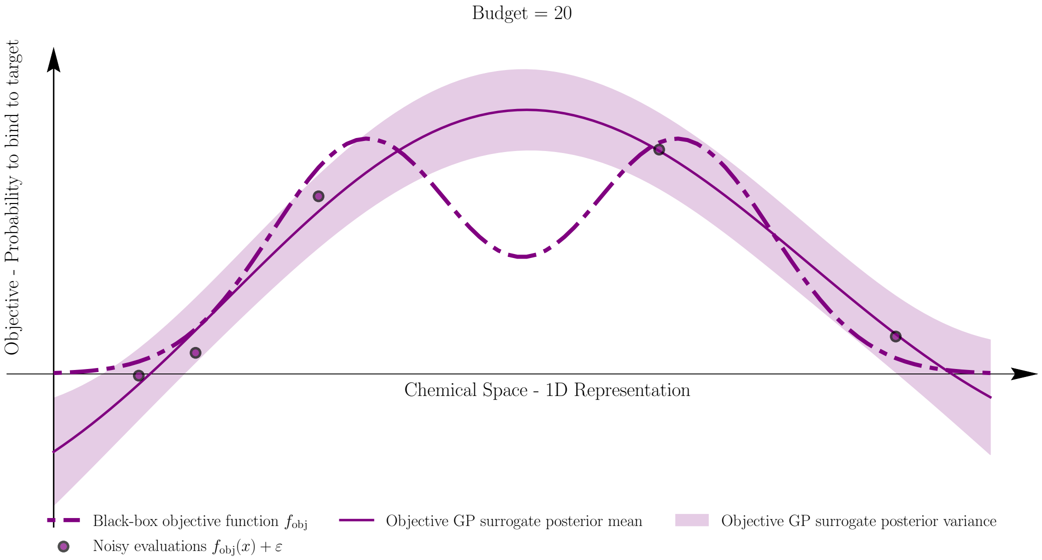
Bayesian Optimization 101



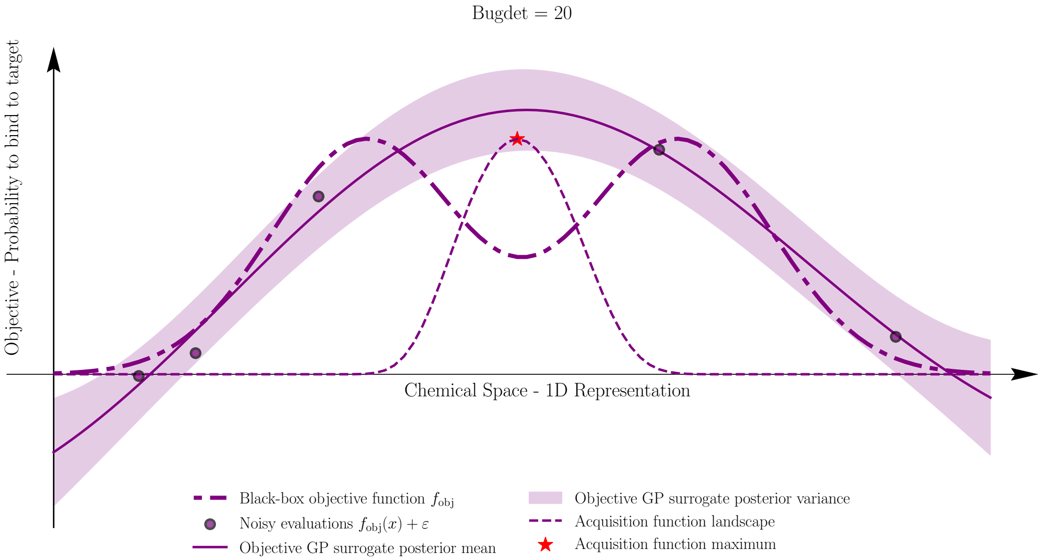
Bayesian Optimization 101



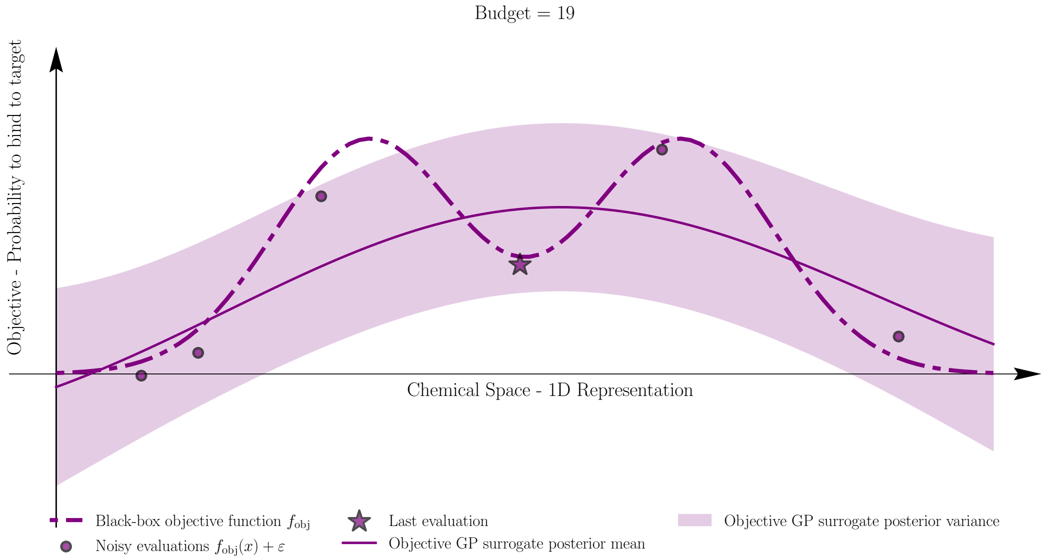
Bayesian Optimization 101



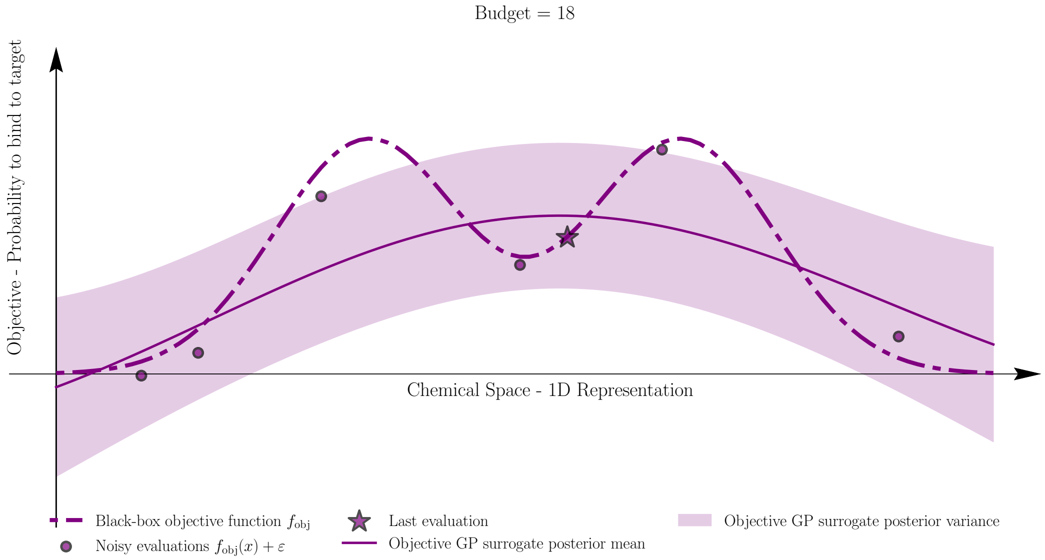
Bayesian Optimization 101



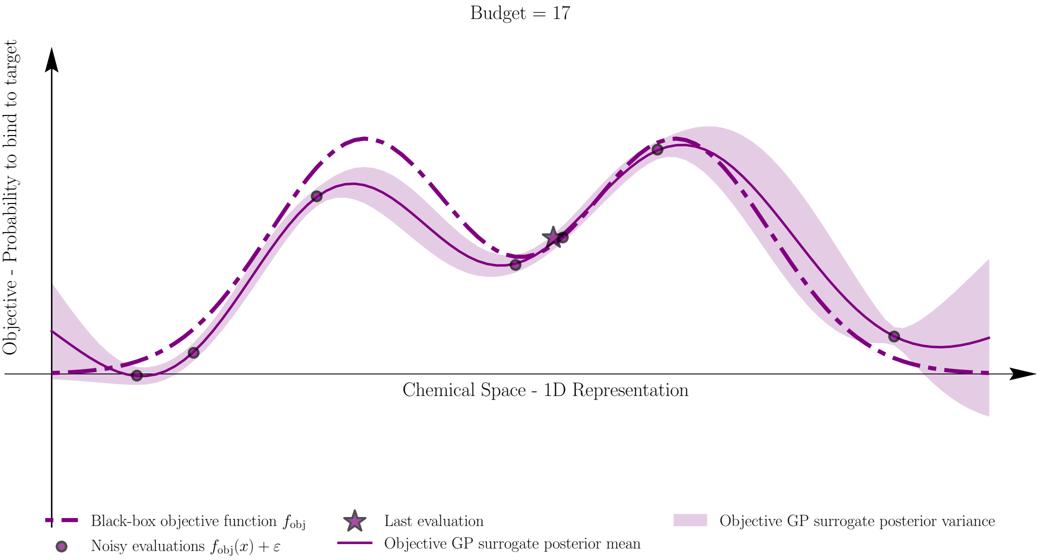
Bayesian Optimization 101



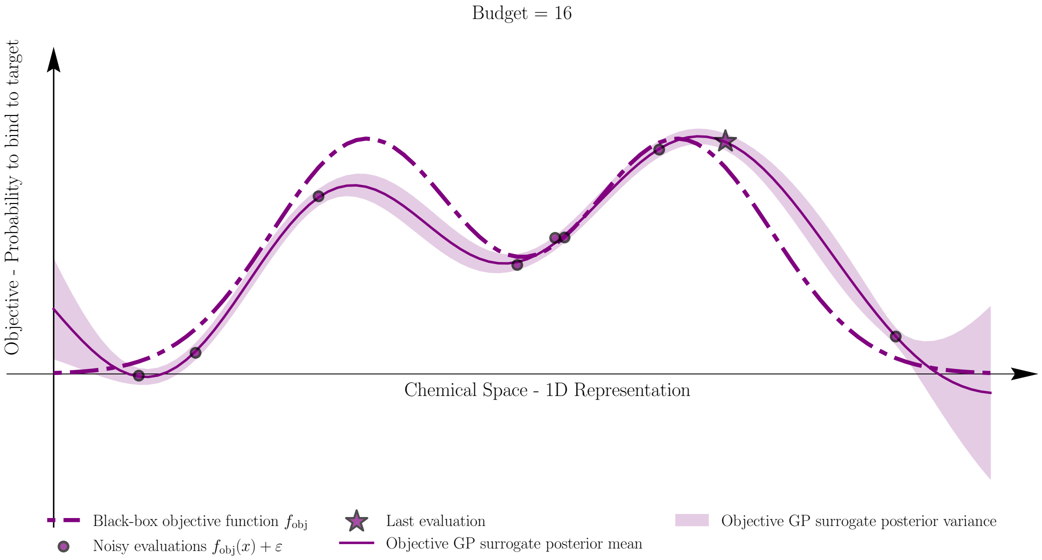
Bayesian Optimization 101



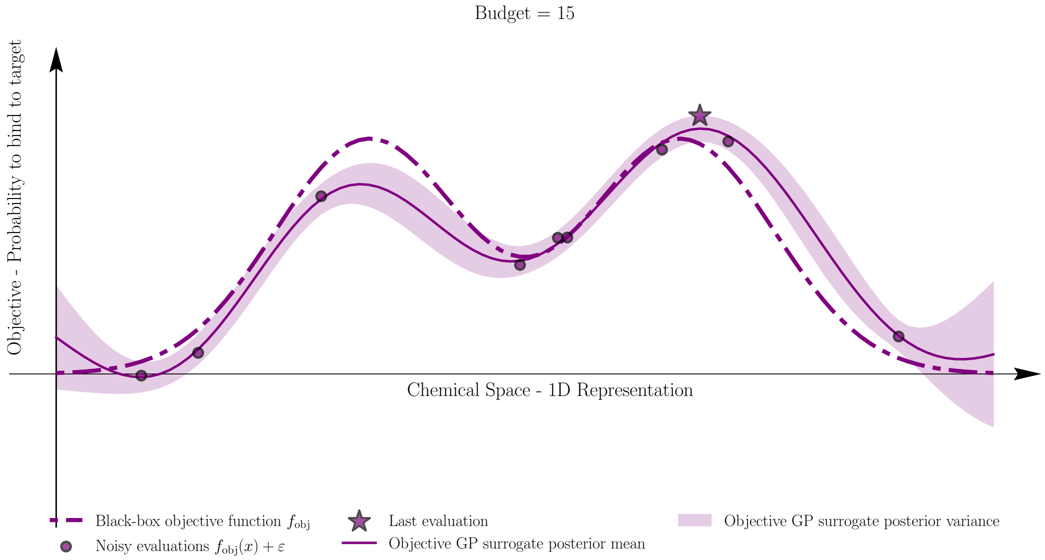
Bayesian Optimization 101



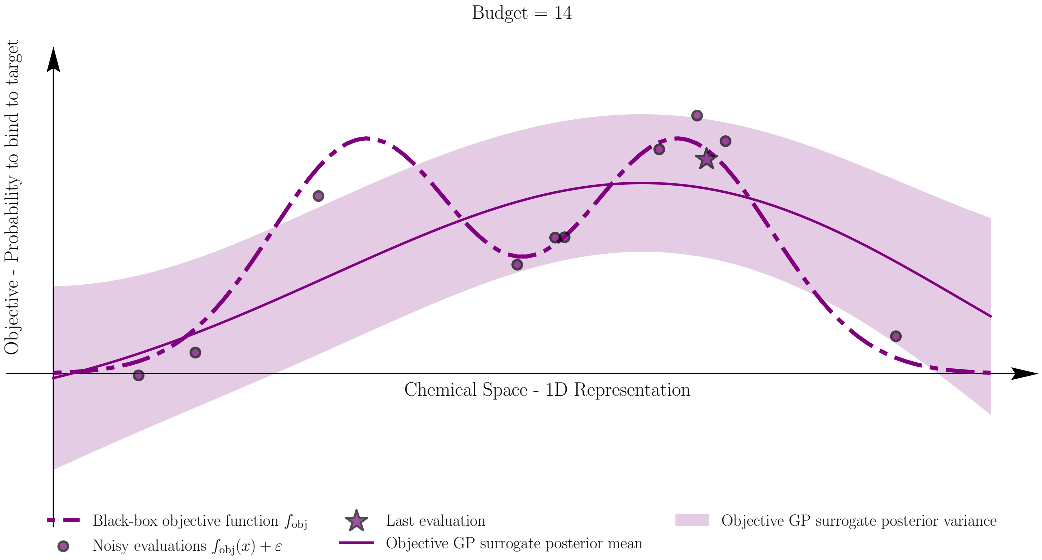
Bayesian Optimization 101



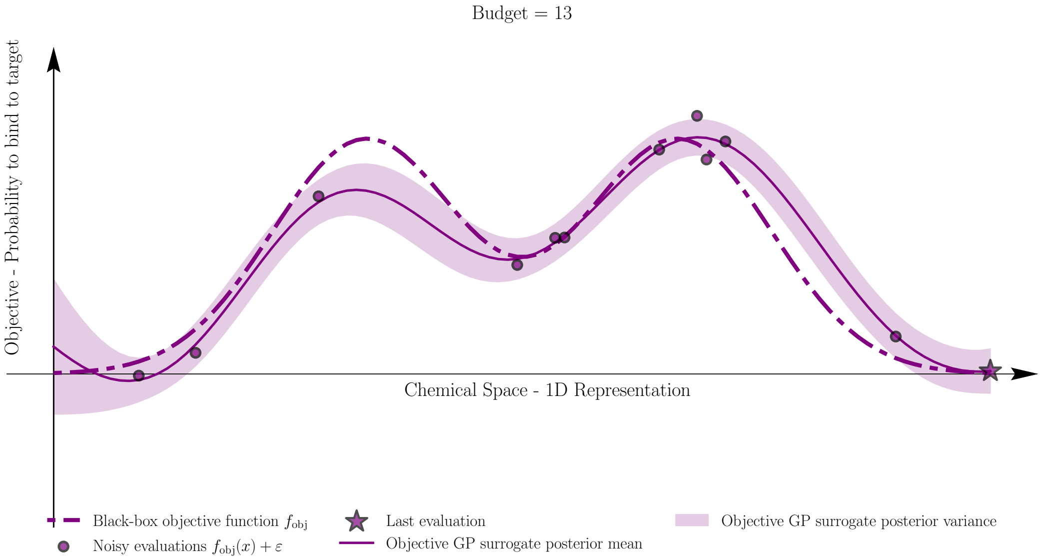
Bayesian Optimization 101



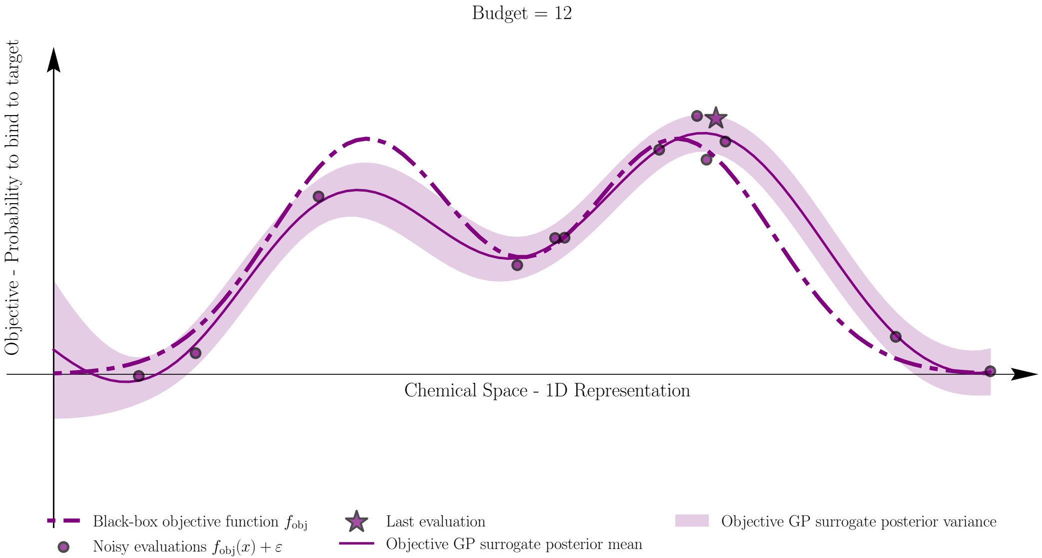
Bayesian Optimization 101



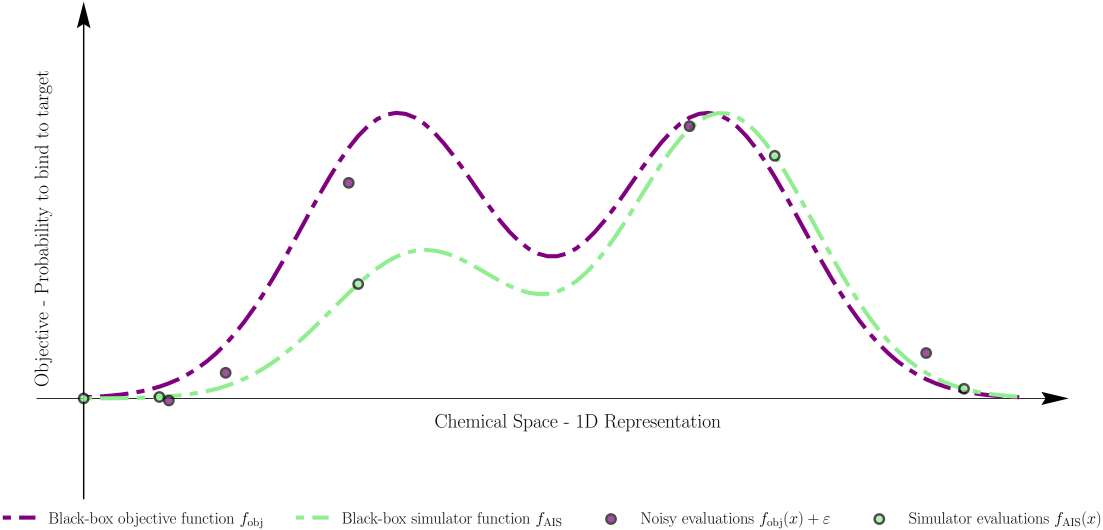
Bayesian Optimization 101



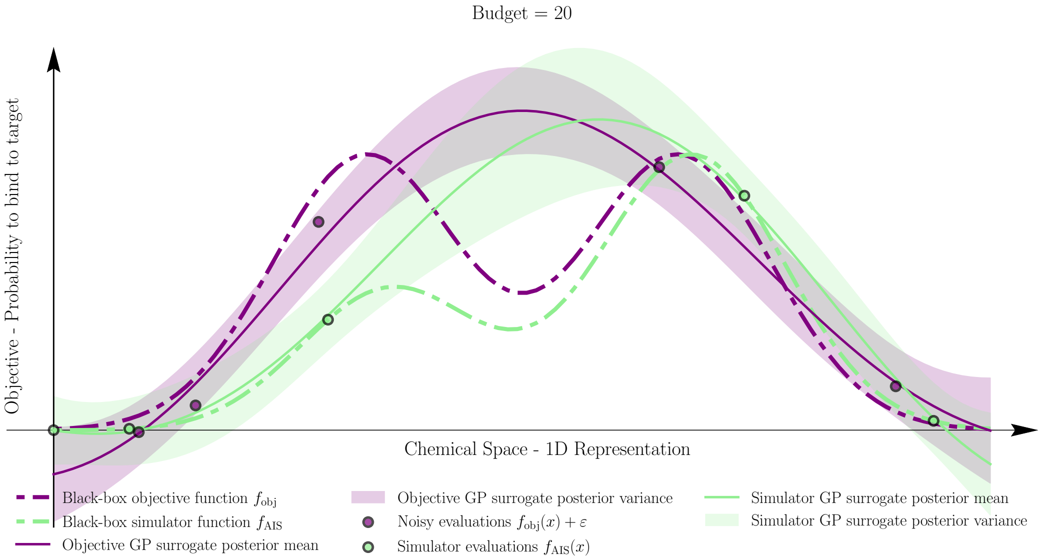
Bayesian Optimization 101



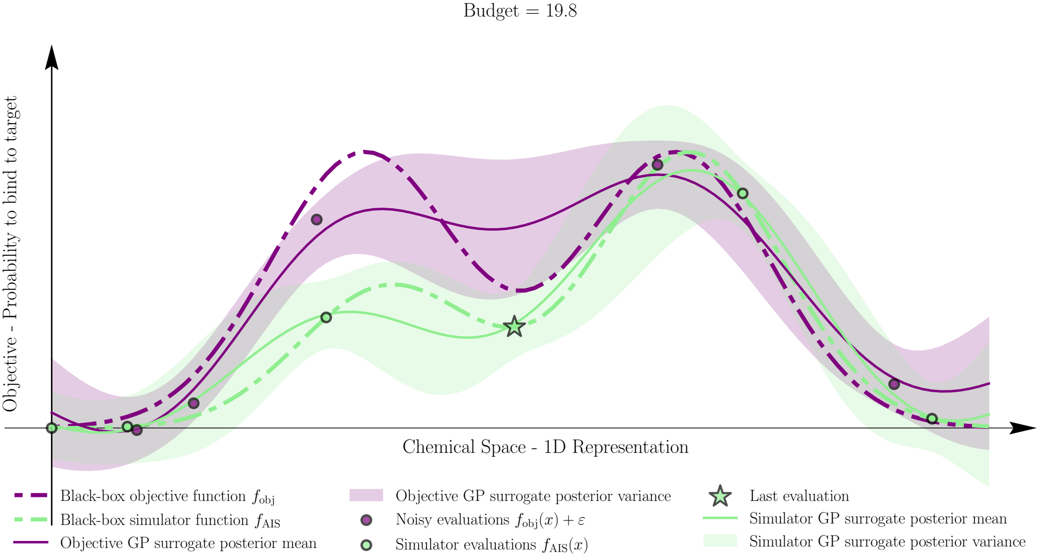
Multi Fidelity Bayesian Optimization 101



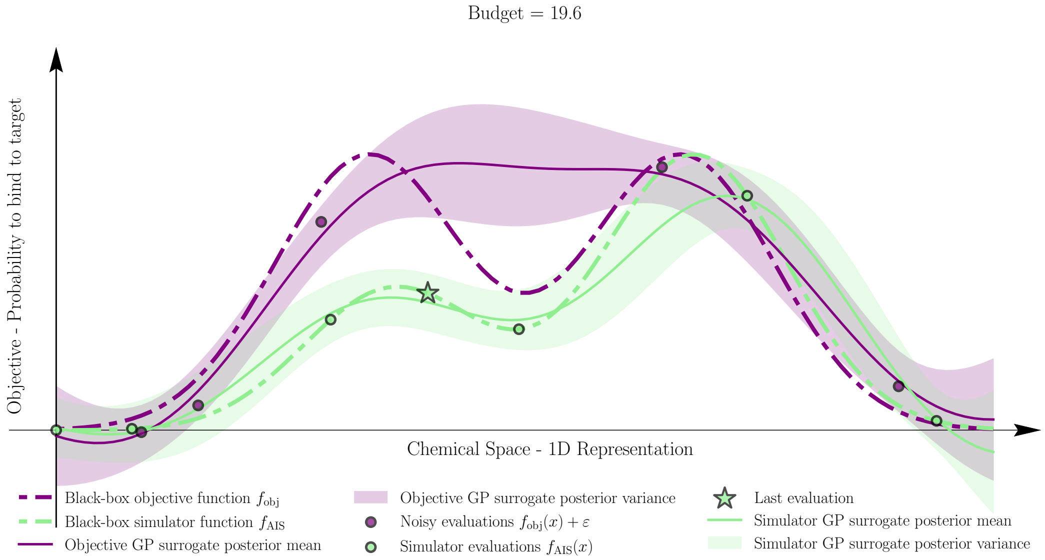
Multi Fidelity Bayesian Optimization 101



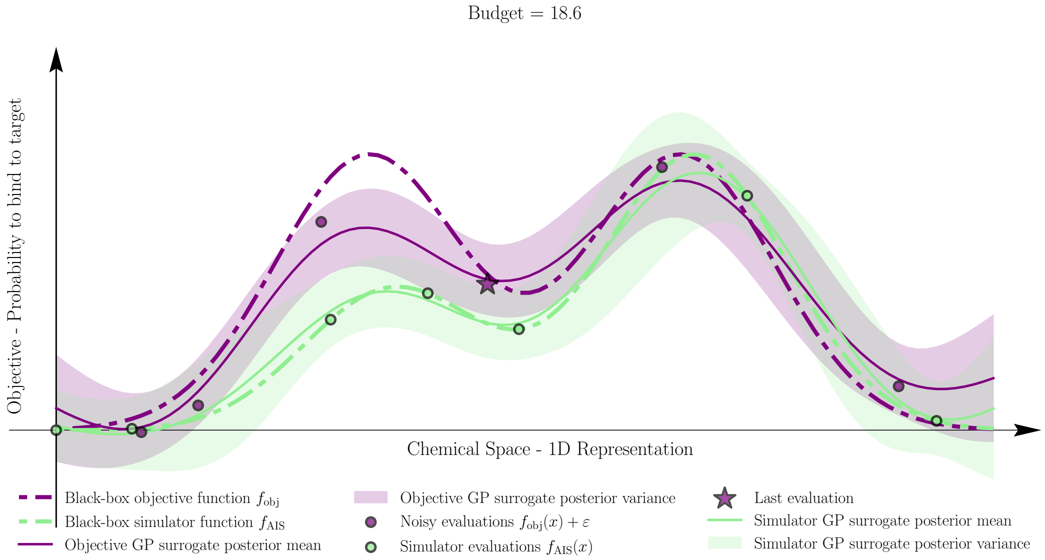
Multi Fidelity Bayesian Optimization 101



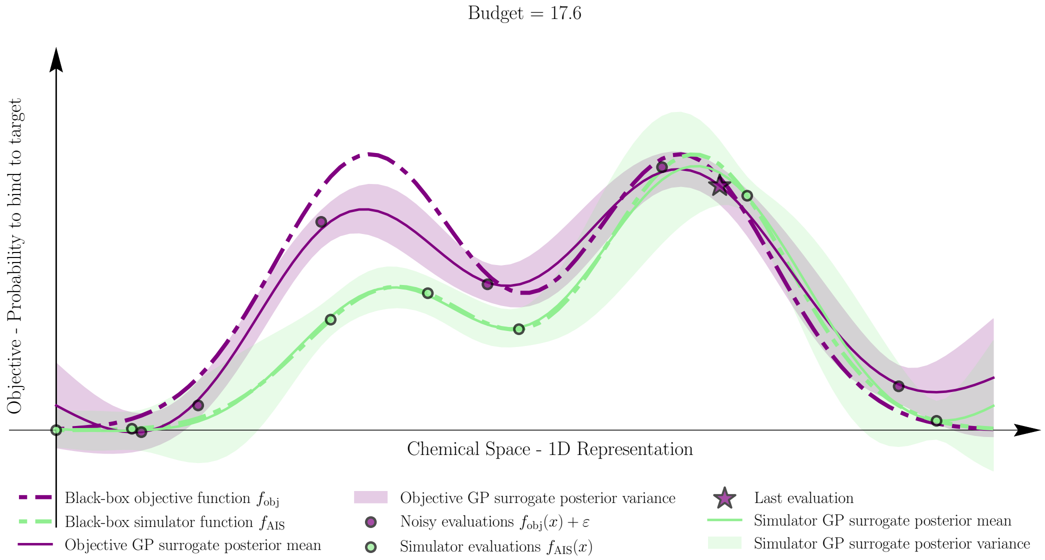
Multi Fidelity Bayesian Optimization 101



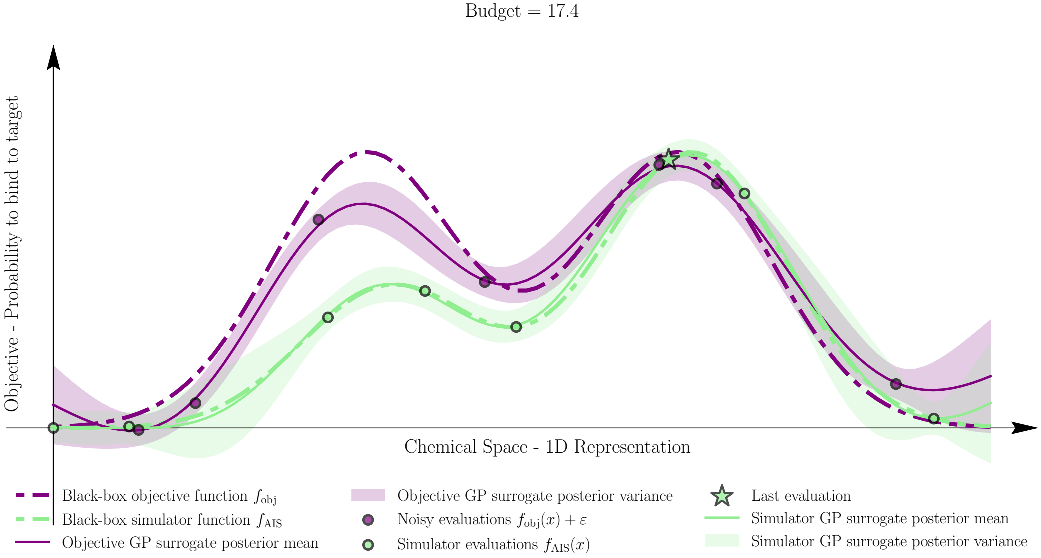
Multi Fidelity Bayesian Optimization 101



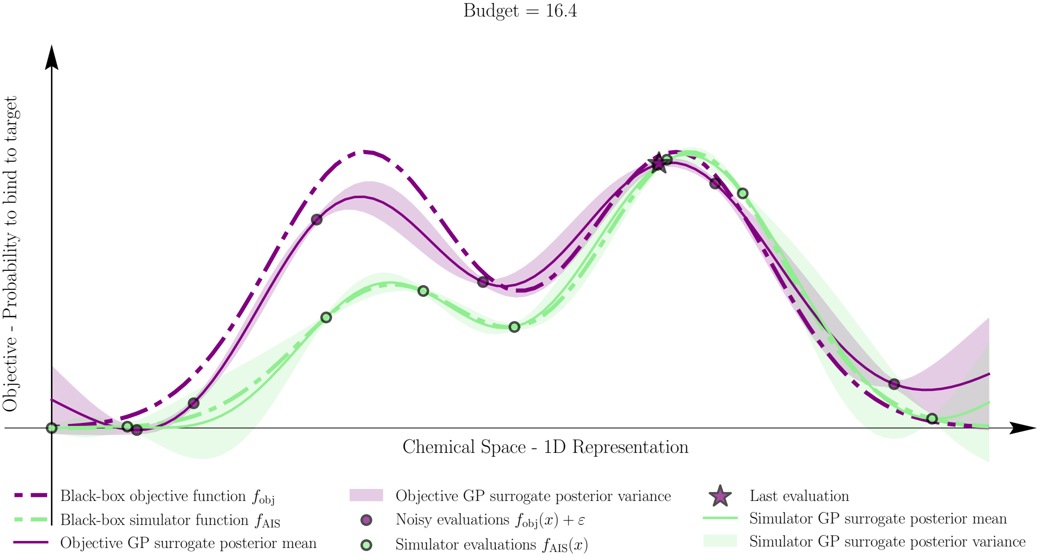
Multi Fidelity Bayesian Optimization 101



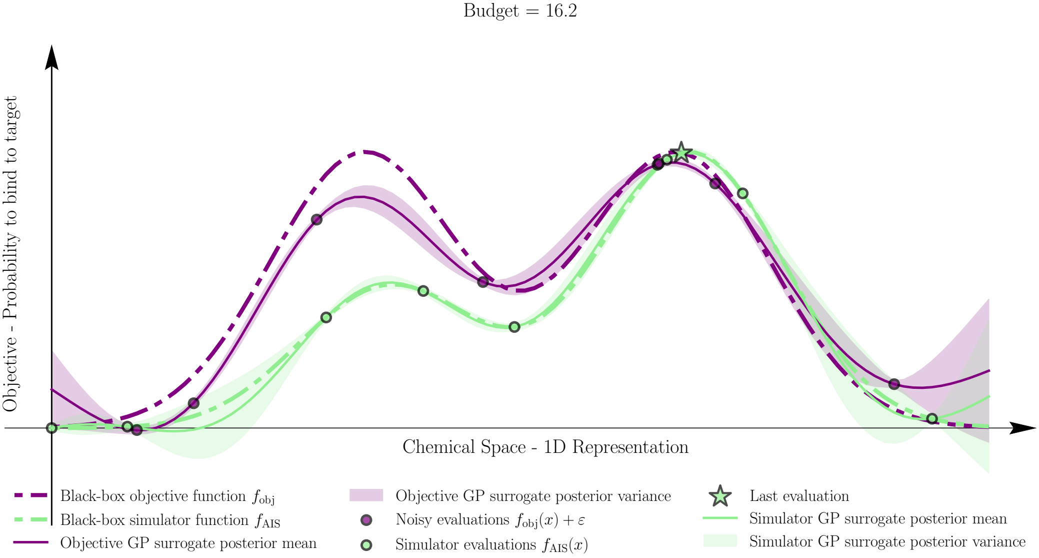
Multi Fidelity Bayesian Optimization 101



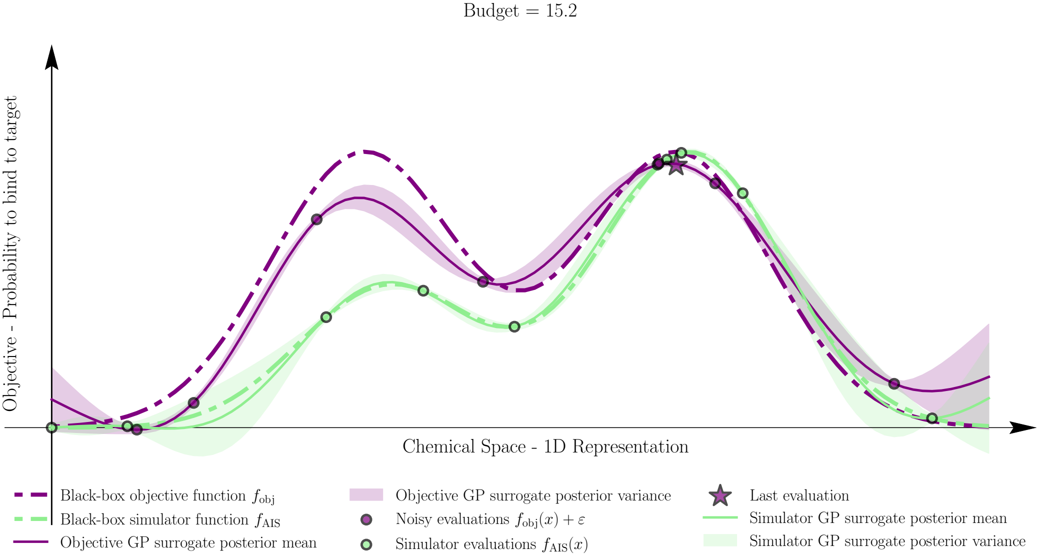
Multi Fidelity Bayesian Optimization 101



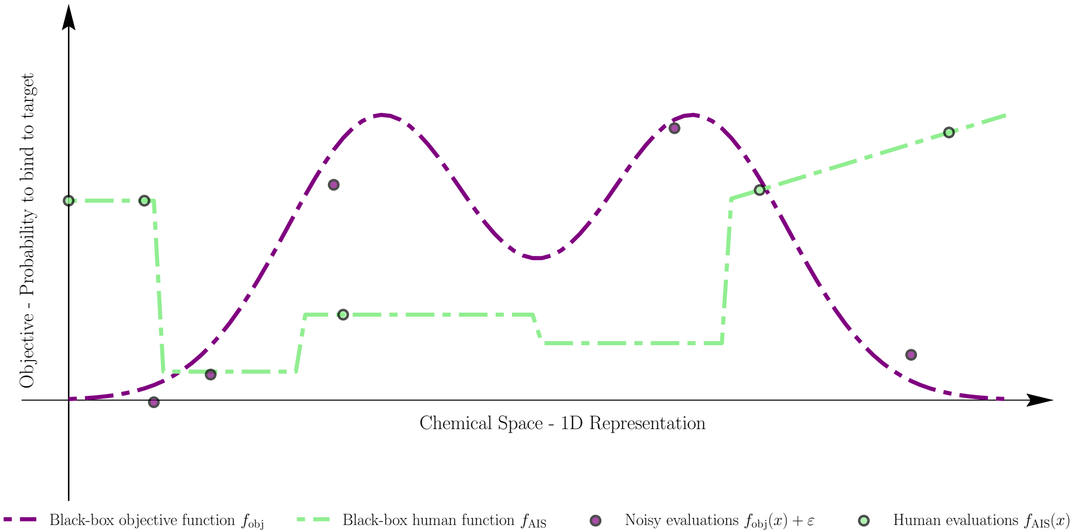
Multi Fidelity Bayesian Optimization 101



Multi Fidelity Bayesian Optimization 101

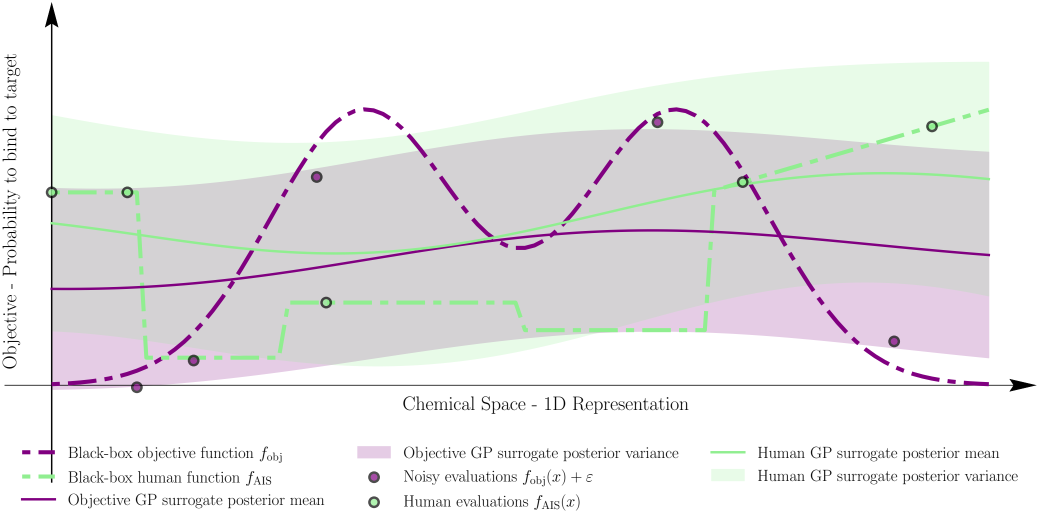


Multi Fidelity Bayesian Optimization with Unreliable Sources

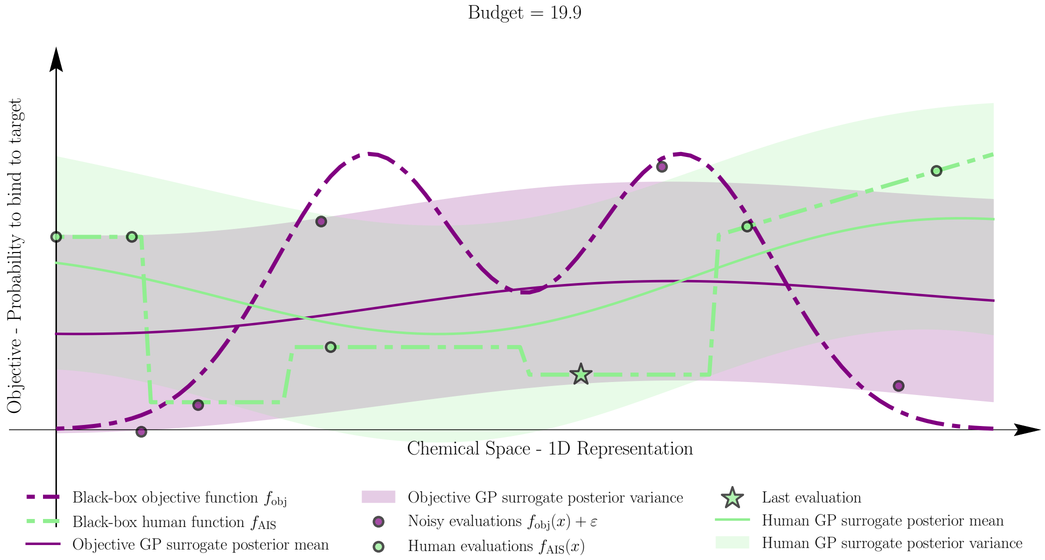


Multi Fidelity Bayesian Optimization with Unreliable Sources

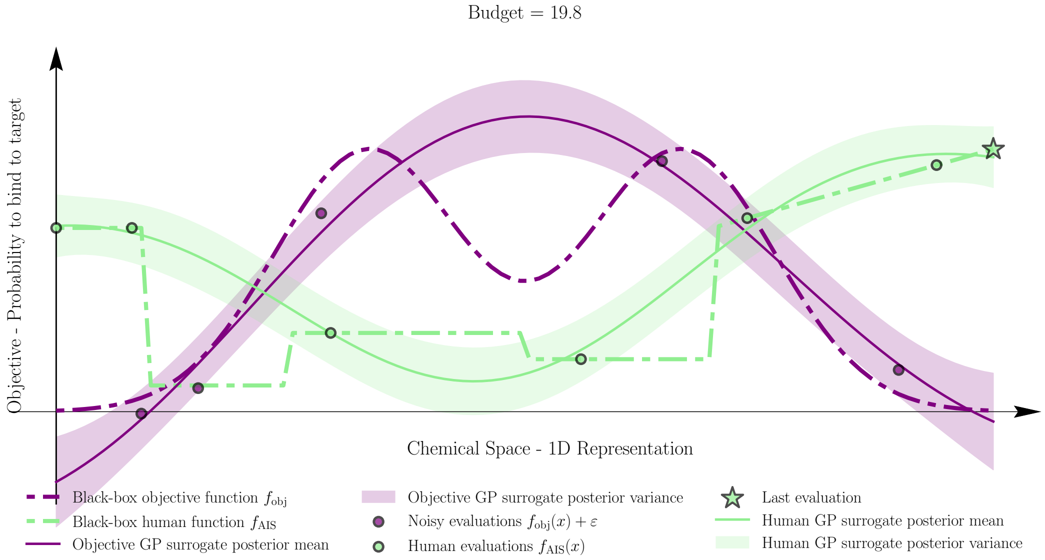
Budget = 20



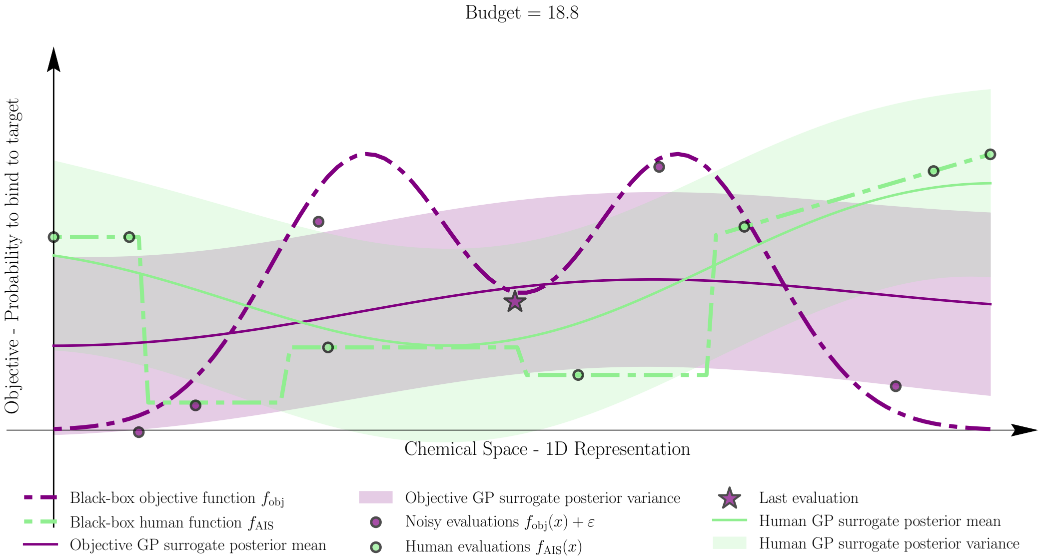
Multi Fidelity Bayesian Optimization with Unreliable Sources



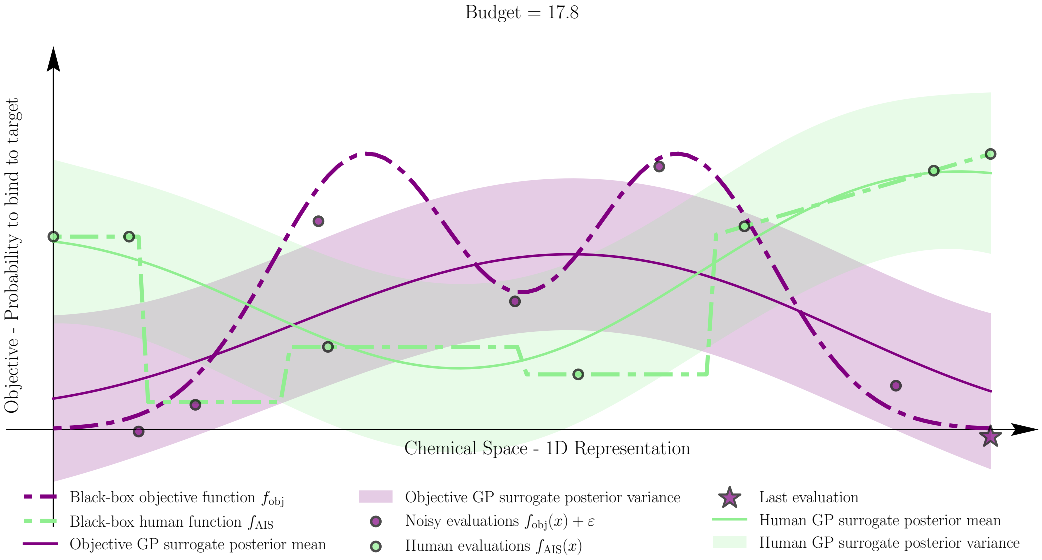
Multi Fidelity Bayesian Optimization with Unreliable Sources



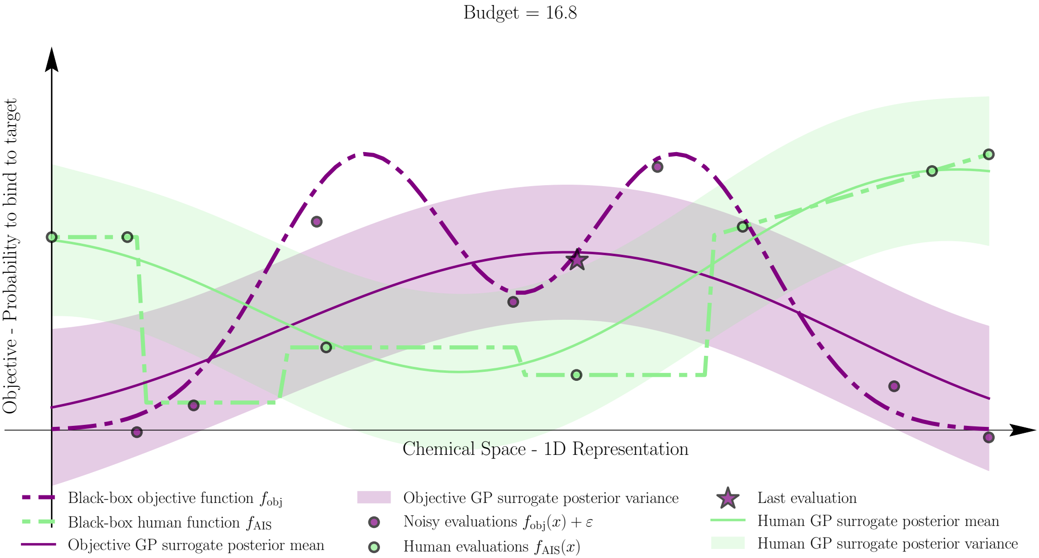
Multi Fidelity Bayesian Optimization with Unreliable Sources



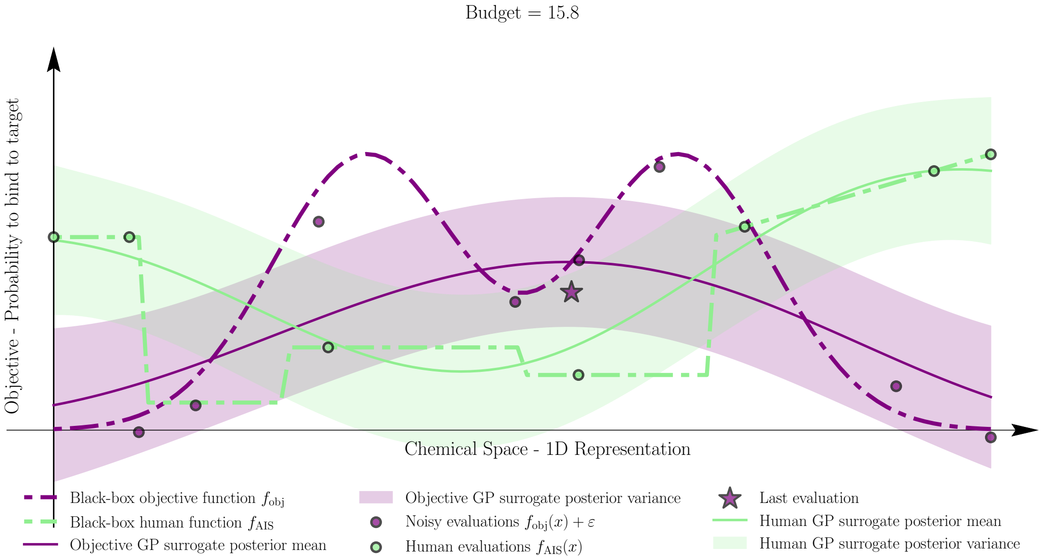
Multi Fidelity Bayesian Optimization with Unreliable Sources



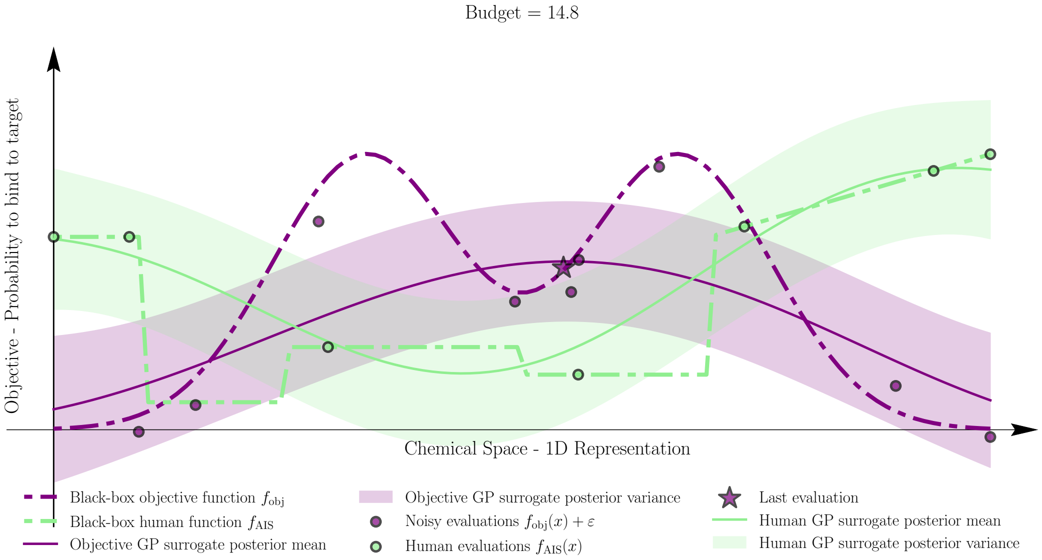
Multi Fidelity Bayesian Optimization with Unreliable Sources



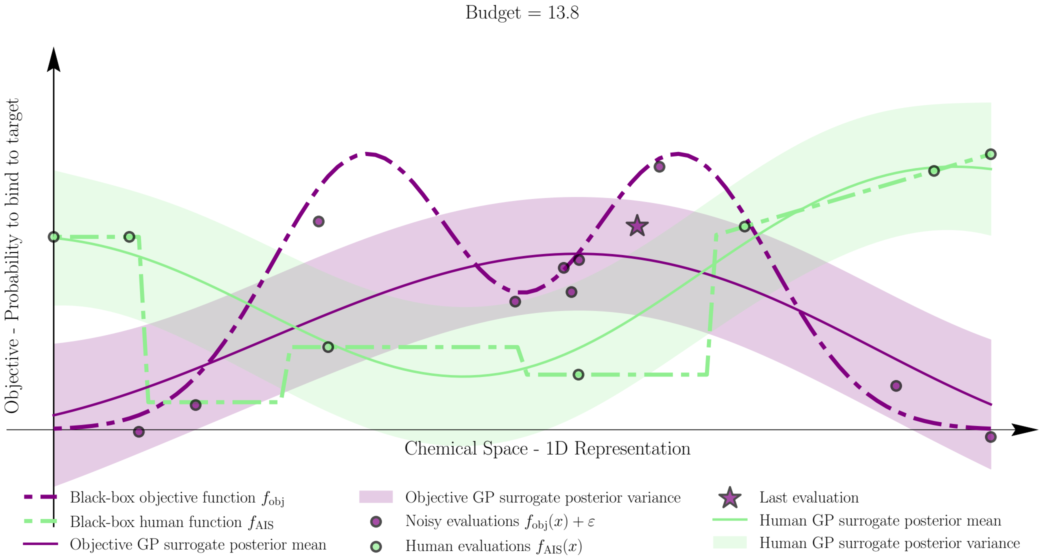
Multi Fidelity Bayesian Optimization with Unreliable Sources



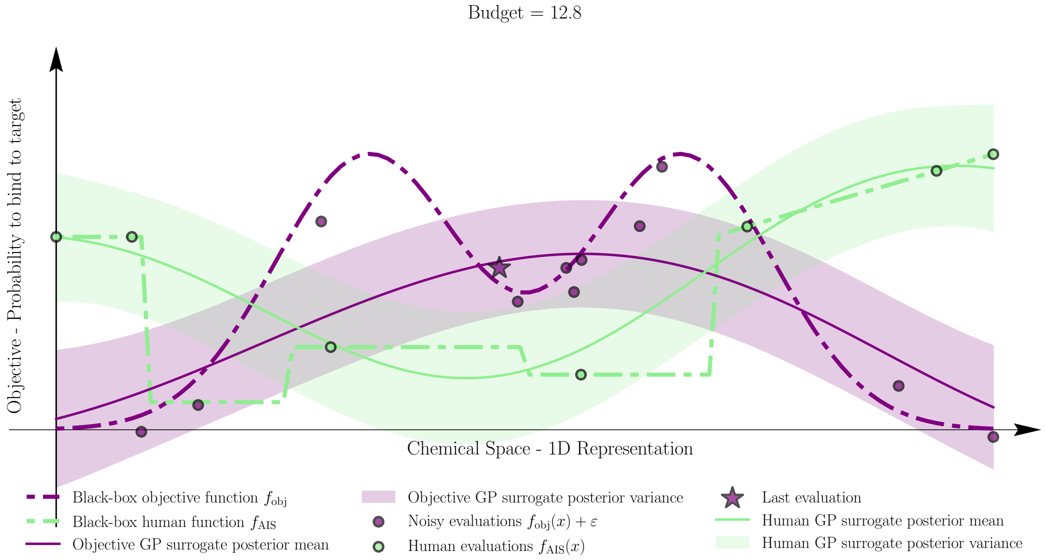
Multi Fidelity Bayesian Optimization with Unreliable Sources



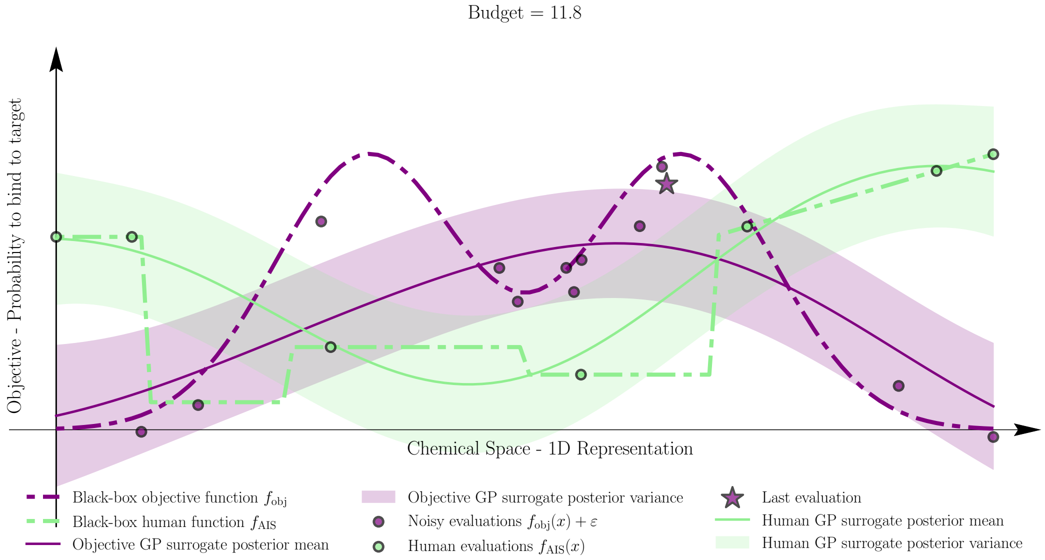
Multi Fidelity Bayesian Optimization with Unreliable Sources



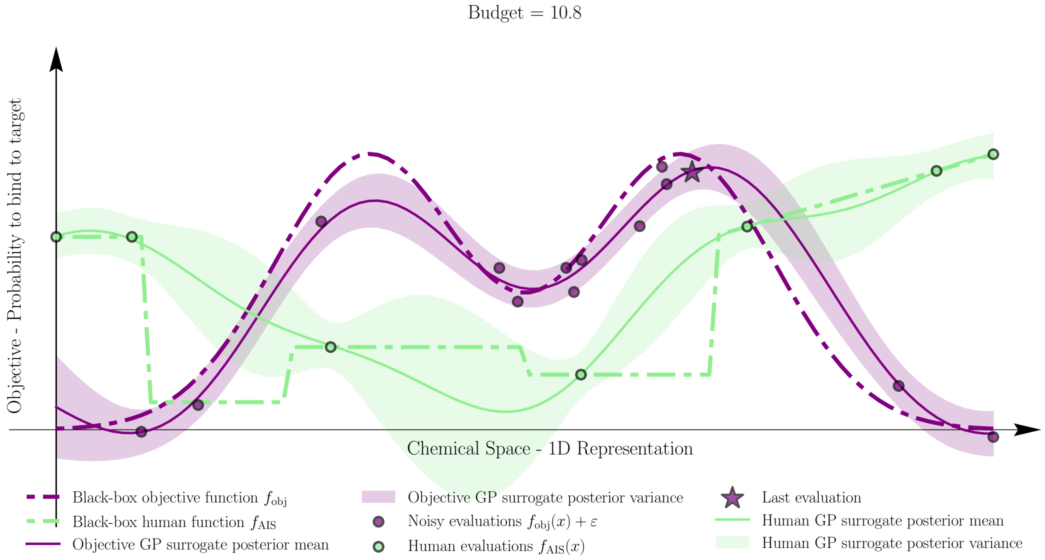
Multi Fidelity Bayesian Optimization with Unreliable Sources



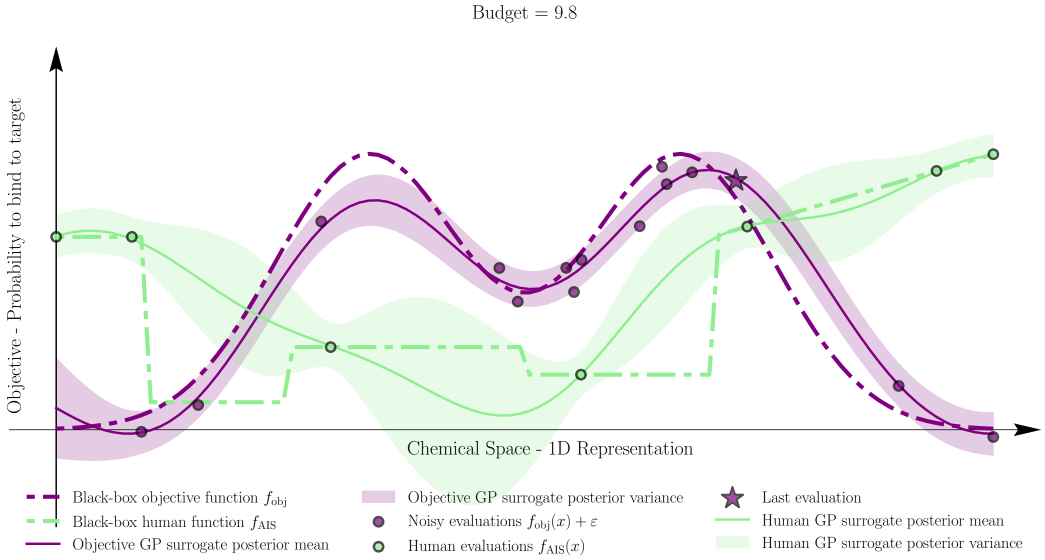
Multi Fidelity Bayesian Optimization with Unreliable Sources



Multi Fidelity Bayesian Optimization with Unreliable Sources



Multi Fidelity Bayesian Optimization with Unreliable Sources



An overlooked problem

Poloczek *et al.*[1] introduced a novel MFBO algorithm. For evaluation, they considered the Rosenbrock function and a low fidelity version over $[-5, 5]^2$

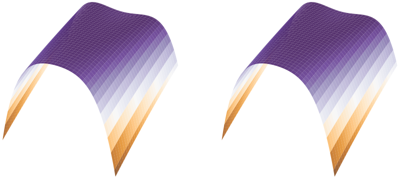
$$f^{\text{obj}}(\mathrm{x}) = -(100(x_2 - x_1^2)^2 + (x_1 - 1)^2)$$
$$f^{\text{AIS}}(\mathrm{x}) = f^{\text{obj}}(\mathrm{x}) + 2\sin(10x_1 + 5x_2)$$

[1] Multi-Information Source Optimization, NeurIPS'17

An overlooked problem

Poloczek *et al.*[1] introduced a novel MFBO algorithm. For evaluation, they considered the Rosenbrock function and a low fidelity version over $[-5, 5]^2$

$$f^{\mathrm{obj}}(\mathrm{x}) = -(100(x_2 - x_1^2)^2 + (x_1 - 1)^2)$$
$$f^{\mathrm{AIS}}(\mathrm{x}) = f^{\mathrm{obj}}(\mathrm{x}) + 2\sin(10x_1 + 5x_2)$$

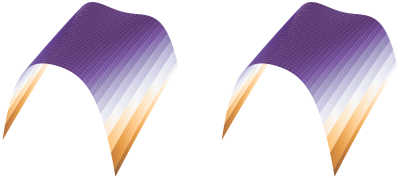


[1]Multi-Information Source Optimization, NeurIPS'17

An overlooked problem

Poloczek *et al.*[1] introduced a novel MFBO algorithm. For evaluation, they considered the Rosenbrock function and a low fidelity version over $[-5, 5]^2$

$$f^{\text{obj}}(\mathrm{x}) = -(100(x_2 - x_1^2)^2 + (x_1 - 1)^2)$$
$$f^{\text{AIS}}(\mathrm{x}) = f^{\text{obj}}(\mathrm{x}) + 2\sin(10x_1 + 5x_2)$$



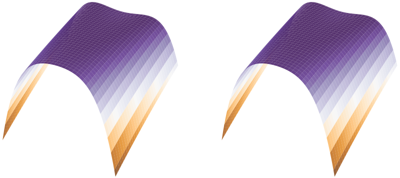
Notice any difference? No? That's normal

[1]Multi-Information Source Optimization, NeurIPS'17

An overlooked problem

Poloczek *et al.*[1] introduced a novel MFBO algorithm. For evaluation, they considered the Rosenbrock function and a low fidelity version over $[-5, 5]^2$

$$f^{\text{obj}}(\mathrm{x}) = -(100(x_2 - x_1^2)^2 + (x_1 - 1)^2)$$
$$f^{\text{AIS}}(\mathrm{x}) = f^{\text{obj}}(\mathrm{x}) + 2\sin(10x_1 + 5x_2)$$



Notice any difference? No? That's normal

Query cost for f^{obj}: 1000. For f^{AIS} ? 1 ¯_(ツ)_/¯

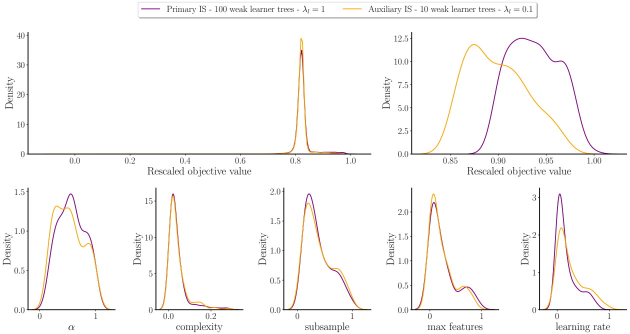
[1]Multi-Information Source Optimization, NeurIPS'17

Hyperparameter tuning with highly reliable lower fidelities

XGBoost hyperparameter optimization benchmark performed by Shibo Li *et al.*[2]

$$f^{\text{obj}}(x) = \text{rMSE using XGB with 100 weak learners trees at cost 10}$$

$$f^{\text{AIS}}(x) = \text{rMSE XGB with 10 weak learners trees at cost 1}$$



Current methods always consider the lower fidelity as (extremely) relevant

[2] Batch Multi-Fidelity Bayesian Optimization with Deep Auto-Regressive Networks, NeurIPS'21

What does it mean to be *unreliable*?

Hand-waving definition based-on inference regret. Define

$$\mathrm{x}_{\star}^{\mathsf{SF}} = \operatorname*{argmax}_{\mathrm{x} \in \mathscr{X}} \mu^{\mathsf{SF}}(\mathrm{x} | \mathscr{D}_t^{\mathsf{SF}})$$

$$\mathrm{x}_{\star}^{\mathsf{MF}} = \operatorname*{argmax}_{\mathrm{x} \in \mathscr{X}} \mu^{\mathsf{MF}}(\mathrm{x} | \mathscr{D}_t^{\mathsf{MF}})$$

An unreliable information source is s.t. $f^{\mathsf{obj}}(\mathrm{x}_{\star}^{\mathsf{SF}}) \geq f^{\mathsf{obj}}(\mathrm{x}_{\star}^{\mathsf{MF}})$ for the same budget.

What does it mean to be *unreliable*?

Hand-waving definition based-on inference regret. Define

$$x_\star^{\mathsf{SF}} = \operatorname*{argmax}_{x \in \mathscr{X}} \mu^{\mathsf{SF}}(x | \mathscr{D}_t^{\mathsf{SF}})$$

$$x_\star^{\mathsf{MF}} = \operatorname*{argmax}_{x \in \mathscr{X}} \mu^{\mathsf{MF}}(x | \mathscr{D}_t^{\mathsf{MF}})$$

An unreliable information source is s.t. $f^{\mathsf{obj}}(x_\star^{\mathsf{SF}}) \geq f^{\mathsf{obj}}(x_\star^{\mathsf{MF}})$ for the same budget.

Not model-free? Depends on kernel, acquisition function...

What does it mean to be *unreliable*?

Hand-waving definition based-on inference regret. Define

$$\mathrm{x}_\star^{\mathsf{SF}} = \operatorname*{argmax}_{\mathrm{x}\in\mathscr{X}} \mu^{\mathsf{SF}}(\mathrm{x}|\mathscr{D}_t^{\mathsf{SF}})$$

$$\mathrm{x}_\star^{\mathsf{MF}} = \operatorname*{argmax}_{\mathrm{x}\in\mathscr{X}} \mu^{\mathsf{MF}}(\mathrm{x}|\mathscr{D}_t^{\mathsf{MF}})$$

An unreliable information source is s.t. $f^{\mathsf{obj}}(\mathrm{x}_\star^{\mathsf{SF}}) \geq f^{\mathsf{obj}}(\mathrm{x}_\star^{\mathsf{MF}})$ for the same budget.

Not model-free? Depends on kernel, acquisition function...

But most importantly, depends on $\mathscr{D}_t$!

What does it mean to be *unreliable*?

Hand-waving definition based-on inference regret. Define

$$x_\star^{\mathsf{SF}} = \operatorname*{argmax}_{x \in \mathscr{X}} \mu^{\mathsf{SF}}(x|\mathscr{D}_t^{\mathsf{SF}})$$

$$x_\star^{\mathsf{MF}} = \operatorname*{argmax}_{x \in \mathscr{X}} \mu^{\mathsf{MF}}(x|\mathscr{D}_t^{\mathsf{MF}})$$

An unreliable information source is s.t. $f^{\mathsf{obj}}(x_\star^{\mathsf{SF}}) \geq f^{\mathsf{obj}}(x_\star^{\mathsf{MF}})$ for the same budget.

Not model-free? Depends on kernel, acquisition function...

But most importantly, depends on $\mathscr{D}_t$!

$$\mathbb{E}_{\mathscr{D}_t^{\mathsf{SF}} \sim q_t^{\mathsf{SF}}(\cdot)}\left[f(x_\star^{\mathsf{SF}})\right] \geq \mathbb{E}_{\mathscr{D}_t^{\mathsf{MF}} \sim q_t^{\mathsf{MF}}(\cdot)}\left[f(x_\star^{\mathsf{MF}})\right]$$

Intractable.

What does it mean to be *unreliable*?

Hand-waving definition based-on inference regret. Define

$$x_\star^{\mathsf{SF}} = \operatorname*{argmax}_{x \in \mathscr{X}} \mu^{\mathsf{SF}}(x | \mathscr{D}_t^{\mathsf{SF}})$$

$$x_\star^{\mathsf{MF}} = \operatorname*{argmax}_{x \in \mathscr{X}} \mu^{\mathsf{MF}}(x | \mathscr{D}_t^{\mathsf{MF}})$$

An unreliable information source is s.t. $f^{\mathsf{obj}}(x_\star^{\mathsf{SF}}) \geq f^{\mathsf{obj}}(x_\star^{\mathsf{MF}})$ for the same budget.

Not model-free? Depends on kernel, acquisition function...

But most importantly, depends on $\mathscr{D}_t$!

$$\mathbb{E}_{\mathscr{D}_t^{\mathsf{SF}} \sim q_t^{\mathsf{SF}}(\cdot)} \left[f(x_\star^{\mathsf{SF}}) \right] \geq \mathbb{E}_{\mathscr{D}_t^{\mathsf{MF}} \sim q_t^{\mathsf{MF}}(\cdot)} \left[f(x_\star^{\mathsf{MF}}) \right]$$

Intractable.
$\longrightarrow$ Not straightforward to define

What does it mean to be *unreliable*?

Hand-waving definition based-on inference regret. Define

$$x_\star^{\text{SF}} = \operatorname*{argmax}_{x \in \mathscr{X}} \mu^{\text{SF}}(x|\mathscr{D}_t^{\text{SF}})$$

$$x_\star^{\text{MF}} = \operatorname*{argmax}_{x \in \mathscr{X}} \mu^{\text{MF}}(x|\mathscr{D}_t^{\text{MF}})$$

An unreliable information source is s.t. $f^{\text{obj}}(x_\star^{\text{SF}}) \geq f^{\text{obj}}(x_\star^{\text{MF}})$ for the same budget.

Not model-free? Depends on kernel, acquisition function...

But most importantly, depends on $\mathscr{D}_t$!

$$\mathbb{E}_{\mathscr{D}_t^{\text{SF}} \sim q_t^{\text{SF}}(\cdot)}\left[f(x_\star^{\text{SF}})\right] \geq \mathbb{E}_{\mathscr{D}_t^{\text{MF}} \sim q_t^{\text{MF}}(\cdot)}\left[f(x_\star^{\text{MF}})\right]$$

Intractable.
$\longrightarrow$ Not straightforward to define

Our approach: start from an "unreliable belief" and develop a defensive strategy

A 6D case: the Hartmann problem

For $A, P \in \mathscr{M}_{4,6}(\mathbb{R})$ two matrices, $\mathrm{x} \in [0,1]^6$ $\ell \in [0,1]$, we define

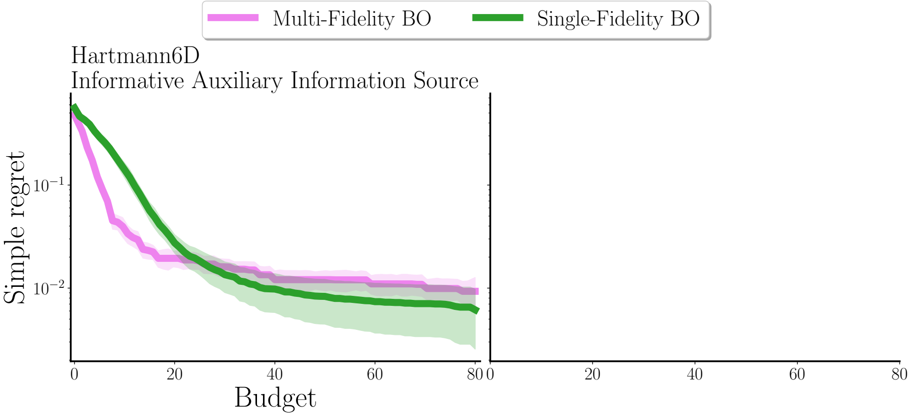
$$f^{(\ell)}(\mathrm{x}) = -\sum_{i=1}^{4} \alpha_i \exp\left(-\sum_{j=1}^{6} A_{ij}(x_j - P_{ij})\right)$$

$$\alpha = (1.0 - 0.1(1-\ell), 1.2, 3.0, 3.2)^T$$

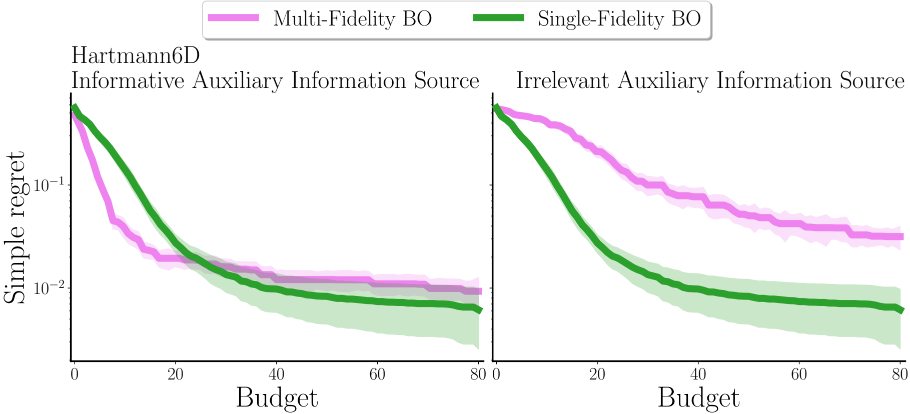
The objective is $f^{(1)}$, with query cost $\lambda_{\text{obj}} = 1$. BO is performed in 3 scenarios, using:

- Only $f^{(1)}$ (Single-Fidelity BO)
- $f^{(1)}$ and an **informative** AIS: $f^{(0.2)}$, $\lambda_{\text{AIS}} = 0.2$ (Multi-Fidelity BO)
- $f^{(1)}$ and an **irrelevant** AIS: $f^{\text{irr}}(\mathrm{x}) = \sum_{i=1}^{5} \left(100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2\right)$, $\lambda_{\text{AIS}} = 0.2$

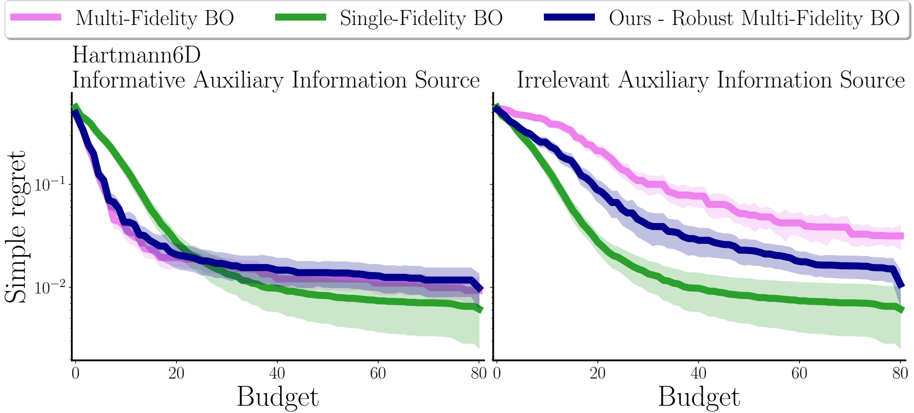
Multi-Fidelity BO is not robust to unreliable Information Sources



Multi-Fidelity BO is not robust to unreliable Information Sources



Multi-Fidelity BO is not robust to unreliable Information Sources



Not far from a well-known problem in transfer learning: **negative transfer**

- Main aim of our contribution: **robustness** to irrelevant AIS...
- ...While still accelerating convergence for relevant AIS

Introducing robust MFBO (rMFBO)

1st idea: perform a test on multi-fidelity proposals to ensure relevant information is added

$$(\mathrm{x}_t^{\mathsf{MF}}, \ell_t) = \underset{\mathrm{x}\in\mathscr{X}, \ell\in\{\text{obj, AIS}\}}{\operatorname{argmax}} \alpha(\mathrm{x}, \ell | \mu_{\mathsf{MF}}, \sigma_{\mathsf{MF}}, \mathscr{D}^{\mathsf{MF}})/\lambda_{l_t}$$

$$s(\mathrm{x}^{\mathsf{MF}}, \ell_t) \geq c_2$$

- Prevent misleading information to flow into the joint GP model
- Guarantee budget is not wasted

For s, we consider information gain: $s(\mathrm{x}, \ell) = \frac{\mathrm{I}(f^{(\ell)}(\mathrm{x}), f_*^{\mathrm{obj}} | \mathscr{D}^{\mathsf{MF}})}{\lambda_\ell}$

This step can be seen as a more demanding acquisition strategy:

1. Compute the acquisition function maximizer
2. Ensure that the found maximum is large enough

Introducing robust MFBO (rMFBO)

2nd idea: maintain a single-fidelity track alongside the multi-fidelity one. Revert to it if needed.

- Keep track of a single-output GP...
- ...And of what would have looked like the acquisition trajectory without AIS

At iteration t, choose between two queries:

$$(\mathrm{x}_t^{\mathrm{MF}}, \ell_t) = \underset{\mathrm{x}\in\mathscr{X}, \ell\in\{\mathrm{obj}, \mathrm{AIS}\}}{\operatorname{argmax}} \quad \alpha(\mathrm{x}, \ell | \mu_{\mathrm{MF}}, \sigma_{\mathrm{MF}}, \mathscr{D}^{\mathrm{MF}})/\lambda_{l_t}$$

$$(\mathrm{x}_t^{\mathrm{pSF}}, \mathrm{obj}) = \underset{\mathrm{x}\in\mathscr{X}}{\operatorname{argmax}}\, \alpha(\mathrm{x} | \mu_{\mathrm{SF}}, \sigma_{\mathrm{SF}}, \mathscr{D}^{\mathrm{pSF}})$$

If $\sigma_{\mathrm{MF}}(\mathrm{x}_t^{\mathrm{pSF}}, \mathrm{obj}) \leq c_1$ $\wedge\, s(\mathrm{x}^{\mathrm{MF}}, \ell_t) \geq c_2$, choose $(\mathrm{x}_t^{\mathrm{MF}}, \ell_t)$

Because $\sigma_{\mathrm{MF}}(\mathrm{x}_t^{\mathrm{pSF}}, \mathrm{obj}) \leq c_1 \implies$ joint model reliable at $\mathrm{x}_t^{\mathrm{pSF}}$.

Therefore $\mathscr{D}^{\mathrm{pSF}} \leftarrow \left(\mathrm{x}_t^{\mathrm{pSF}}, \mu_{\mathrm{MF}}^{\mathrm{obj}}\left(\mathrm{x}_t^{\mathrm{pSF}}\right)\right)$: creating a *pseudo* single fidelity track

Summary

$$\sigma_{\mathrm{MF}}(\mathrm{x}_t^{\mathrm{pSF}}, \mathrm{obj}) \le c_1 \;\wedge\; s(\mathrm{x}^{\mathrm{MF}}, \ell_t) \ge c_2 \implies \begin{cases} \text{pick } (\mathrm{x}_t^{\mathrm{MF}}, \ell_t) \\ \mathscr{D}^{\mathrm{MF}} \leftarrow (\mathrm{x}_t^{\mathrm{MF}}, f^{(\ell_t)}(\mathrm{x}_t^{\mathrm{MF}})) \\ \mathscr{D}^{\mathrm{pSF}} \leftarrow (\mathrm{x}_t^{\mathrm{pSF}}, \mu_{\mathrm{MF}}^{\mathrm{obj}}(\mathrm{x}_t^{\mathrm{pSF}})) \end{cases}$$

Summary

$$\sigma_{\mathrm{MF}}(\mathrm{x}_t^{\mathrm{pSF}}, \mathrm{obj}) \leq c_1 \;\wedge\; s(\mathrm{x}^{\mathrm{MF}}, \ell_t) \geq c_2 \implies \begin{cases} \text{pick } (\mathrm{x}_t^{\mathrm{MF}}, \ell_t) \\ \mathscr{D}^{\mathrm{MF}} \leftarrow (\mathrm{x}_t^{\mathrm{MF}}, f^{(\ell_t)}(\mathrm{x}_t^{\mathrm{MF}})) \\ \mathscr{D}^{\mathrm{pSF}} \leftarrow (\mathrm{x}_t^{\mathrm{pSF}}, \mu_{\mathrm{MF}}^{\mathrm{obj}}(\mathrm{x}_t^{\mathrm{pSF}})) \end{cases}$$

If not satisfied:

1. Pick $(\mathrm{x}_t^{\mathrm{pSF}}, \mathrm{obj})$
2. $\mathscr{D}^{\mathrm{MF}} \leftarrow \left(\mathrm{x}_t^{\mathrm{pSF}}, f^{\mathrm{obj}}\left(\mathrm{x}_t^{\mathrm{pSF}}\right)\right)$
3. $\mathscr{D}^{\mathrm{pSF}} \leftarrow \left(\mathrm{x}_t^{\mathrm{pSF}}, f^{\mathrm{obj}}\left(\mathrm{x}_t^{\mathrm{pSF}}\right)\right)$

Summary

$$\sigma_{\mathrm{MF}}(\mathrm{x}_t^{\mathrm{pSF}}, \mathrm{obj}) \leq c_1 \wedge s(\mathrm{x}^{\mathrm{MF}}, \ell_t) \geq c_2 \implies \begin{cases} \text{pick } (\mathrm{x}_t^{\mathrm{MF}}, \ell_t) \\ \mathscr{D}^{\mathrm{MF}} \leftarrow (\mathrm{x}_t^{\mathrm{MF}}, f^{(\ell_t)}(\mathrm{x}_t^{\mathrm{MF}})) \\ \mathscr{D}^{\mathrm{pSF}} \leftarrow (\mathrm{x}_t^{\mathrm{pSF}}, \mu_{\mathrm{MF}}^{\mathrm{obj}}(\mathrm{x}_t^{\mathrm{pSF}})) \end{cases}$$

Ensures pseudo-queries added to SFBO are trustworthy: **unreliable case**

If not satisfied:

1. Pick $(\mathrm{x}_t^{\mathrm{pSF}}, \mathrm{obj})$
2. $\mathscr{D}^{\mathrm{MF}} \leftarrow \left(\mathrm{x}_t^{\mathrm{pSF}}, f^{\mathrm{obj}}\left(\mathrm{x}_t^{\mathrm{pSF}}\right)\right)$
3. $\mathscr{D}^{\mathrm{pSF}} \leftarrow \left(\mathrm{x}_t^{\mathrm{pSF}}, f^{\mathrm{obj}}\left(\mathrm{x}_t^{\mathrm{pSF}}\right)\right)$

Summary

$$\sigma_{\mathrm{MF}}(x_t^{\mathrm{pSF}}, \mathrm{obj}) \leq c_1 \;\wedge\; s(x^{\mathrm{MF}}, \ell_t) \geq c_2 \implies \begin{cases} \text{pick } (x_t^{\mathrm{MF}}, \ell_t) \\ \mathscr{D}^{\mathrm{MF}} \leftarrow (x_t^{\mathrm{MF}}, f^{(\ell_t)}(x_t^{\mathrm{MF}})) \\ \mathscr{D}^{\mathrm{pSF}} \leftarrow (x_t^{\mathrm{pSF}}, \mu_{\mathrm{MF}}^{\mathrm{obj}}(x_t^{\mathrm{pSF}})) \end{cases}$$

Ensures pseudo-queries added to SFBO are trustworthy: **unreliable case**

But does not say anything about the relevance of x_t^{MF} as potential maximizer!

If not satisfied:

1. Pick $(x_t^{\mathrm{pSF}}, \mathrm{obj})$
2. $\mathscr{D}^{\mathrm{MF}} \leftarrow \left(x_t^{\mathrm{pSF}}, f^{\mathrm{obj}}\left(x_t^{\mathrm{pSF}}\right)\right)$
3. $\mathscr{D}^{\mathrm{pSF}} \leftarrow \left(x_t^{\mathrm{pSF}}, f^{\mathrm{obj}}\left(x_t^{\mathrm{pSF}}\right)\right)$

Summary

This does:
reliable case

$$\sigma_{\mathrm{MF}}(\mathrm{x}_t^{\mathrm{pSF}}, \mathrm{obj}) \leq c_1 \wedge s(\mathrm{x}^{\mathrm{MF}}, \ell_t) \geq c_2 \implies \begin{cases} \text{pick } (\mathrm{x}_t^{\mathrm{MF}}, \ell_t) \\ \mathscr{D}^{\mathrm{MF}} \leftarrow (\mathrm{x}_t^{\mathrm{MF}}, f^{(\ell_t)}(\mathrm{x}_t^{\mathrm{MF}})) \\ \mathscr{D}^{\mathrm{pSF}} \leftarrow (\mathrm{x}_t^{\mathrm{pSF}}, \mu_{\mathrm{MF}}^{\mathrm{obj}}(\mathrm{x}_t^{\mathrm{pSF}})) \end{cases}$$

Ensures pseudo-queries added to SFBO are trustworthy: **unreliable case**

But does not say anything about the relevance of $\mathrm{x}_t^{\mathrm{MF}}$ as potential maximizer!

If not satisfied:

1. Pick $(\mathrm{x}_t^{\mathrm{pSF}}, \mathrm{obj})$
2. $\mathscr{D}^{\mathrm{MF}} \leftarrow \left(\mathrm{x}_t^{\mathrm{pSF}}, f^{\mathrm{obj}}\left(\mathrm{x}_t^{\mathrm{pSF}}\right)\right)$
3. $\mathscr{D}^{\mathrm{pSF}} \leftarrow \left(\mathrm{x}_t^{\mathrm{pSF}}, f^{\mathrm{obj}}\left(\mathrm{x}_t^{\mathrm{pSF}}\right)\right)$

Summary

This does:
reliable case

$$\sigma_{\mathrm{MF}}(x_t^{\mathrm{pSF}}, \mathrm{obj}) \leq c_1 \;\wedge\; s(x^{\mathrm{MF}}, \ell_t) \geq c_2 \implies \begin{cases} \text{pick } (x_t^{\mathrm{MF}}, \ell_t) \\ \mathscr{D}^{\mathrm{MF}} \leftarrow (x_t^{\mathrm{MF}}, f^{(\ell_t)}(x_t^{\mathrm{MF}})) \\ \mathscr{D}^{\mathrm{pSF}} \leftarrow (x_t^{\mathrm{pSF}}, \mu_{\mathrm{MF}}^{\mathrm{obj}}(x_t^{\mathrm{pSF}})) \end{cases}$$

Ensures pseudo-queries added to SFBO are trustworthy: **unreliable case**

But does not say anything about the relevance of x_t^{MF} as potential maximizer!

If not satisfied:

1. Pick $(x_t^{\mathrm{pSF}}, \mathrm{obj})$
2. $\mathscr{D}^{\mathrm{MF}} \leftarrow \left(x_t^{\mathrm{pSF}}, f^{\mathrm{obj}}\left(x_t^{\mathrm{pSF}}\right)\right)$
3. $\mathscr{D}^{\mathrm{pSF}} \leftarrow \left(x_t^{\mathrm{pSF}}, f^{\mathrm{obj}}\left(x_t^{\mathrm{pSF}}\right)\right)$

$\mathscr{D}^{\mathrm{pSF}}$ and $\mathscr{D}^{\mathrm{SF}}$ only differ at the points where we inputed $\mu_{\mathrm{MF}}^{\mathrm{obj}}(x_t^{\mathrm{pSF}})$!

rMFBO regret can be tied to that of SFBO

Assumptions:

- f^{obj} is drawn from a GP with zero-mean and covariance function $\kappa(\mathrm{x}, \mathrm{x}')$
- κ is known and twice differentiable
- $\mathbb{P}\left(\sup_{\mathrm{x}\in\mathscr{X}} \left|\frac{\partial f^{\text{obj}}}{\partial x_j}\right| > L\right) \leq ae^{-(L/b_j)^2} \quad \forall\ j \in \{1, \dots, d\}$, for $a, b_j > 0$

rMFBO regret can be tied to that of SFBO

Assumptions:

- f^{obj} is drawn from a GP with zero-mean and covariance function $\kappa(\mathrm{x}, \mathrm{x}')$
- κ is known and twice differentiable
- $\mathbb{P}\left(\sup_{\mathrm{x}\in\mathscr{X}} \left|\frac{\partial f^{\text{obj}}}{\partial x_j}\right| > L\right) \leq ae^{-(L/b_j)^2} \quad \forall\ j \in \{1, \ldots, d\}$, for $a, b_j > 0$
- The mapping $(\mathrm{x}, \mathscr{D}) \mapsto \alpha(\mathrm{x}|p(f|\mathscr{D}))$ is twice differentiable

rMFBO regret can be tied to that of SFBO

Assumptions:

- f^{obj} is drawn from a GP with zero-mean and covariance function $\kappa(\mathrm{x}, \mathrm{x}')$
- κ is known and twice differentiable
- $\mathbb{P}\left(\sup_{\mathrm{x} \in \mathscr{X}} \left| \frac{\partial f^{\text{obj}}}{\partial x_j} \right| > L\right) \leq a e^{-(L/b_j)^2} \quad \forall\, j \in \{1, \ldots, d\}$, for $a, b_j > 0$
- The mapping $(\mathrm{x}, \mathscr{D}) \mapsto \alpha(\mathrm{x} | p(f | \mathscr{D}))$ is twice differentiable

Theorem:

for any AIS, the difference in regrets achieved by SFBO and rMFBO can be bounded.

rMFBO regret can be tied to that of SFBO

Assumptions:

- f^{obj} is drawn from a GP with zero-mean and covariance function $\kappa(\mathrm{x}, \mathrm{x}')$
- κ is known and twice differentiable
- $\mathbb{P}\left(\sup_{\mathrm{x}\in\mathscr{X}} \left|\frac{\partial f^{\text{obj}}}{\partial x_j}\right| > L\right) \leq a e^{-(L/b_j)^2} \quad \forall\ j \in \{1, \dots, d\}$, for $a, b_j > 0$
- The mapping $(\mathrm{x}, \mathscr{D}) \mapsto \alpha(\mathrm{x}|p(f|\mathscr{D}))$ is twice differentiable

Theorem:
for any AIS, the difference in regrets achieved by SFBO and rMFBO can be bounded.

$R(\Lambda, \mathrm{x}_T^{\text{rMF}}) \leq R(\Lambda, \mathrm{x}_T^{\text{SF}}) + \varepsilon \max\left\{T\hat{M}_T d^{T+1}, 2\right\}$ with probability $\geq q\left(1 - da \exp\left(-\frac{1}{b^2}\right)\right)$

$c_1(\varepsilon, q) = \frac{\varepsilon}{\sqrt{-2\log(1-q)}}$. Theorem does not depend on c_2.

In practice, bound really useful the first few rounds...

Key ideas for proof

- Bound $\|\mathrm{x}_{t+1}^{\mathrm{pSF}} - \mathrm{x}_{t+1}^{\mathrm{SF}}\|_\infty$ (induction over t). Then use $\mathbb{P}\left(\sup_{\mathrm{x}\in\mathscr{X}}\left|\frac{\partial f^{\mathrm{obj}}}{\partial x_j}\right| > L\right) \leq ae^{-(L/b_j)^2}$

Key ideas for proof

- Bound $\|\mathrm{x}_{t+1}^{\mathsf{pSF}} - \mathrm{x}_{t+1}^{\mathsf{SF}}\|_\infty$ (induction over t). Then use $\mathbb{P}\left(\sup_{\mathrm{x}\in\mathscr{X}} \left|\frac{\partial f^{\mathsf{obj}}}{\partial x_j}\right| > L\right) \leq ae^{-(L/b_j)^2}$
- Consider $\mathscr{D}_t$ as a $t(d+1)$-dimensional vector $\mathscr{D}_t = (x_1^{(1)}, \dots, x_t^{(d)}, y_1, \dots, y_t)$
  View $\mathrm{x}_{t+1} = \operatorname{argmax}_{\mathrm{x}\in\mathscr{X}} \alpha(\mathrm{x}|\mathscr{D}_t)$ as an implicit function $\mathscr{D}_t \mapsto \mathrm{x}_{t+1}(\mathscr{D}_t)$

$$M_t = \max_{\mathscr{D}\in\mathbb{D}_t} \left\|\frac{\partial \mathrm{x}_{t+1}}{\partial \mathscr{D}}\right\|_{\mathsf{op}}$$

  For $\mathbb{D}_t := \{\mathscr{D} \mid \mathscr{D} = (1-u)\mathscr{D}_t^{\mathsf{pSF}} + u\mathscr{D}_t^{\mathsf{SF}}, u \in [0,1]\}$.
  M_t is the sensitivity of the next query to change in the dataset. Allows to bound

$$\|\mathrm{x}_{t+1}^{\mathsf{pSF}} - \mathrm{x}_{t+1}^{\mathsf{SF}}\|_\infty = \|\mathrm{x}_{t+1}(\mathscr{D}_t^{\mathsf{SF}}) - \mathrm{x}_{t+1}(\mathscr{D}_t^{\mathsf{pSF}})\|_\infty \leq \|\mathscr{D}_t^{\mathsf{SF}} - \mathscr{D}_t^{\mathsf{pSF}}\|_\infty M_t$$

Key ideas for proof

- Bound $\|\mathrm{x}_{t+1}^{\mathrm{pSF}} - \mathrm{x}_{t+1}^{\mathrm{SF}}\|_\infty$ (induction over t). Then use $\mathbb{P}\left(\sup_{\mathrm{x}\in\mathscr{X}} \left|\frac{\partial f^{\mathrm{obj}}}{\partial x_j}\right| > L\right) \leq a e^{-(L/b_j)^2}$

- Consider $\mathscr{D}_t$ as a $t(d+1)$-dimensional vector $\mathscr{D}_t = (x_1^{(1)}, \dots, x_t^{(d)}, y_1, \dots, y_t)$
  View $\mathrm{x}_{t+1} = \operatorname{argmax}_{\mathrm{x}\in\mathscr{X}} \alpha(\mathrm{x}|\mathscr{D}_t)$ as an implicit function $\mathscr{D}_t \mapsto \mathrm{x}_{t+1}(\mathscr{D}_t)$

$$M_t = \max_{\mathscr{D}\in\mathbb{D}_t} \|\frac{\partial \mathrm{x}_{t+1}}{\partial \mathscr{D}}\|_{\mathrm{op}}$$

For $\mathbb{D}_t := \{\mathscr{D} | \mathscr{D} = (1-u)\mathscr{D}_t^{\mathrm{pSF}} + u\mathscr{D}_t^{\mathrm{SF}}, u \in [0,1]\}$.
M_t is the sensitivity of the next query to change in the dataset. Allows to bound

$$\|\mathrm{x}_{t+1}^{\mathrm{pSF}} - \mathrm{x}_{t+1}^{\mathrm{SF}}\|_\infty = \|\mathrm{x}_{t+1}(\mathscr{D}_t^{\mathrm{SF}}) - \mathrm{x}_{t+1}(\mathscr{D}_t^{\mathrm{pSF}})\|_\infty \leq \|\mathscr{D}_t^{\mathrm{SF}} - \mathscr{D}_t^{\mathrm{pSF}}\|_\infty M_t$$

f^{obj} is a draw from a GP $\implies \mathbb{P}\left(\frac{f^{\mathrm{obj}}(\mathrm{x}) - \mu(\mathrm{x})}{\sigma(\mathrm{x})} > C\right) \leq \frac{1}{2}\exp\left(-\frac{C^2}{2}\right)$ exercise :)

Results on 2D case

Acquisition function: max-value entropy search

$$\alpha(\mathrm{x}, \ell) = \frac{I(f_*; f^{(\ell)} | \mathscr{D}_t)}{\lambda_\ell}$$

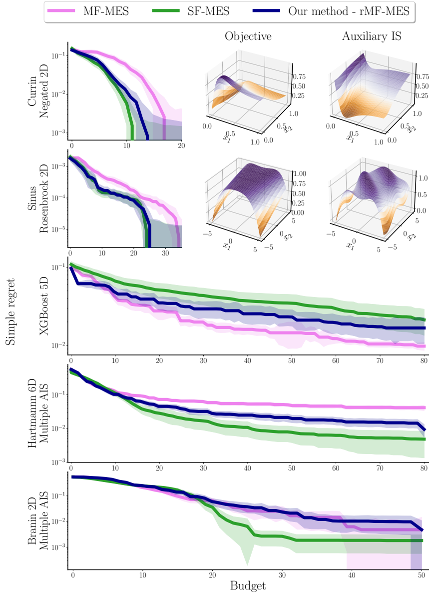
Kernel:

$$k((\mathrm{x}, \ell), (\mathrm{x}', \ell')) = k_{\mathrm{input}}(\mathrm{x}, \mathrm{x}') k_{\mathrm{IS}}(\ell, \ell')$$

$c_1 = c_2 = 0.1$ throughout all experiments

Obj cost = 1. AIS cost: 0.1 (rows 1-3), 0.2 (rows 2-4-5)

More acquisition functions, kernels and ablation studies in the paper!



Open questions

- Instead of keeping track of two separate GPs, can we come up with a joint model that does the same job?
- Definition of an *unreliable* information source...
- Find a principled way to benchmark MFBO algorithms with IS of any relevance

Open questions

- Instead of keeping track of two separate GPs, can we come up with a joint model that does the same job?
- Definition of an *unreliable* information source...
- Find a principled way to benchmark MFBO algorithms with IS of any relevance
- Concrete case with human experts
  (FCAI: Real AI for **Real People** in the Real World)

Open questions

- Instead of keeping track of two separate GPs, can we come up with a joint model that does the same job?
- Definition of an *unreliable* information source...
- Find a principled way to benchmark MFBO algorithms with IS of any relevance
- Concrete case with human experts
  (FCAI: Real AI for **Real People** in the Real World)

Questions?

Some stuff ¯_(ツ)_/¯

$$s(\mathrm{x}, \ell) = \frac{\mathrm{I}(f^{(\ell)}(\mathrm{x}), f_*^{\mathsf{obj}} \mid \mathscr{D}^{\mathsf{MF}})}{\lambda_\ell} = \mathbb{H}(f^{(l)}(\mathrm{x})|\mathscr{D}_t) - \mathbb{E}_{f_*^{\mathsf{obj}}|\mathscr{D}_t}\left[\mathbb{H}(f^{(l)}(\mathrm{x})|f_*^{\mathsf{obj}}\mathscr{D}_t)\right]$$

We have that

$$\mathrm{I}(\{\mathrm{x}, y\}; y_* \mid \mathscr{D}_t) \approx \frac{\gamma_{y_*}(\mathrm{x})\psi(\gamma_{y_*}(\mathrm{x}))}{2\Psi(\gamma_{y_*}(\mathrm{x}))} - \log(\Psi(\gamma_{y_*}(\mathrm{x})))$$

ψ is the normal p.d.f. and Ψ normal c.d.f. ; $\gamma_{y_*}(\mathrm{x}) = \frac{y_* - \mu_t(\mathrm{x})}{\sigma_t(\mathrm{x})}$. I is unbounded above but rarely in practice greater than $-\log(1/2)$, for $\gamma_{y_*}(\mathrm{x}) = 0$.

Roughly speaking, $c_2 = 0.1 \implies$ AIS query should give at least about 15% of the max info gain.

We set $c_2 = -u\log(1/2)$, where u is the percent of the maximum information gain required for a cost-adjusted AIS query.

We found $u = 15\%$ works as a good default value.

Some stuff (continued) ¯_(ツ)_/¯

$$k_{\mathrm{MISO}}((\mathrm{x},\ell),(\mathrm{x}',\ell')) = \begin{cases} k_{\mathrm{input}}(\mathrm{x},\mathrm{x}') + k_{\ell}(\mathrm{x},\mathrm{x}') & \ell = \ell' \neq 1 \\ k_{\mathrm{input}}(\mathrm{x},\mathrm{x}') & \text{otherwise} \end{cases}$$

$$k_{\mathrm{LT}}((\mathrm{x},\ell),(\mathrm{x}',\ell')) = \begin{cases} k_{\mathrm{input}}(\mathrm{x},\mathrm{x}') + (1-\ell)(1-\ell')k_{\mathrm{IS}}(\mathrm{x},\mathrm{x}') & \ell \neq 1,\ \ell' \neq 1 \\ k_{\mathrm{input}}(\mathrm{x},\mathrm{x}') & \text{otherwise} \end{cases}$$

$$k_{\mathrm{DS}}((\mathrm{x},\ell),(\mathrm{x}',\ell')) = \begin{cases} c k_{\mathrm{input}}(\mathrm{x},\mathrm{x}') + (1-\ell)(1-\ell')k_{\mathrm{input}}(\mathrm{x},\mathrm{x}') & \ell \neq 1,\ \ell' \neq 1 \\ c k_{\mathrm{input}}(\mathrm{x},\mathrm{x}') & \text{otherwise} \end{cases}$$