# This Too Shall Pass: Removing Stale Observations in Dynamic Bayesian Optimization


Anthony Bardou[1], Patrick Thiran[1], and Giovanni Ranieri[1]

[1]IC, EPFL, Lausanne, Switzerland


GITHUB VIDEO PAPER

## Dynamic Bayesian Optimization

Dynamic Bayesian Optimization (DBO) is used to optimize a **dynamic black-box** function $f : \mathcal{S} \times \mathcal{T} \to \mathbb{R}$, where $\mathcal{S} \subseteq \mathbb{R}^d$ and $\mathcal{T} \subseteq \mathbb{R}$, which is **expensive to evaluate** and **noisy**. In this setting, $f$ is discovered by successive queries that feed a **Gaussian Process (GP)** controlled by a covariance function $k : (\mathcal{S} \times \mathcal{T})^2 \to \mathbb{R}^+$ with parameters $\boldsymbol{\theta} = (\lambda, \boldsymbol{\theta}_S, \boldsymbol{\theta}_T)$ such that

$$k((\boldsymbol{x}, t), (\boldsymbol{x}', t')) = \overbrace{\lambda}^{\text{variance}} \overbrace{k_S(||\boldsymbol{x} - \boldsymbol{x}'||_2; \boldsymbol{\theta}_S)}^{\text{spatial correlation}} \overbrace{k_T(|t - t'|; \boldsymbol{\theta}_T)}^{\text{temporal correlation}} .$$

The posterior GP conditioned on the dataset of observations $\mathcal{D} = \{((\boldsymbol{x}_1, t_1), y_1), \cdots, ((\boldsymbol{x}_n, t_n), y_n)\}$, denoted $\mathcal{GP}_\mathcal{D}\left(\mu_\mathcal{D}, \sigma^2_\mathcal{D}\right)$, is used to find the next query $(\boldsymbol{x}_{n+1}, t_{n+1})$. An **exploration-exploitation dilemma** is solved by maximizing an **acquisition function**, e.g. GP-UCB

$$\boldsymbol{x}_{n+1} = \arg\max_{\boldsymbol{x} \in \mathcal{S}} \overbrace{\mu_\mathcal{D}(\boldsymbol{x}, t)}^{\mathcal{GP}_\mathcal{D} \text{ mean}} + \beta_t^{1/2} \overbrace{\sigma_\mathcal{D}(\boldsymbol{x}, t)}^{\mathcal{GP}_\mathcal{D} \text{ std. dev.}} .$$

## Inherent Challenges

A DBO task is harder than its static counterpart for three reasons:

**No Time Travel** At time $t_0$, only $f(\cdot, t_0)$ is observable. $f(\cdot, t)$ with $t < t_0$ is no longer observable, $f(\cdot, t')$ with $t' > t_0$ is not observable yet.

**Stale Observations** As time $t$ goes by, a given observation becomes less and less relevant to keep track of $\arg\max_{\boldsymbol{x} \in \mathcal{S}} f(\boldsymbol{x}, t)$.

**Sampling Frequency** It is crucial that the DBO algorithm keeps sampling $f$ at a high frequency to properly track $\arg\max_{\boldsymbol{x} \in \mathcal{S}} f(\boldsymbol{x}, t)$.

Because the GP inference is in $\mathcal{O}(n^3)$, where $n = |\mathcal{D}|$ is the dataset size, a DBO algorithm **must pinpoint and remove irrelevant observations** to avoid becoming prohibitive to use in the long run.

## The Big Question

Can we pinpoint irrelevant observations in the dataset $\mathcal{D}$ and remove them in an online fashion?

## The Wasserstein Distance as a Measure of Relevancy

**Remark** An observation is **irrelevant** if removing it from $\mathcal{D}$ **does not significantly impact the future predictions** of the GP posterior.

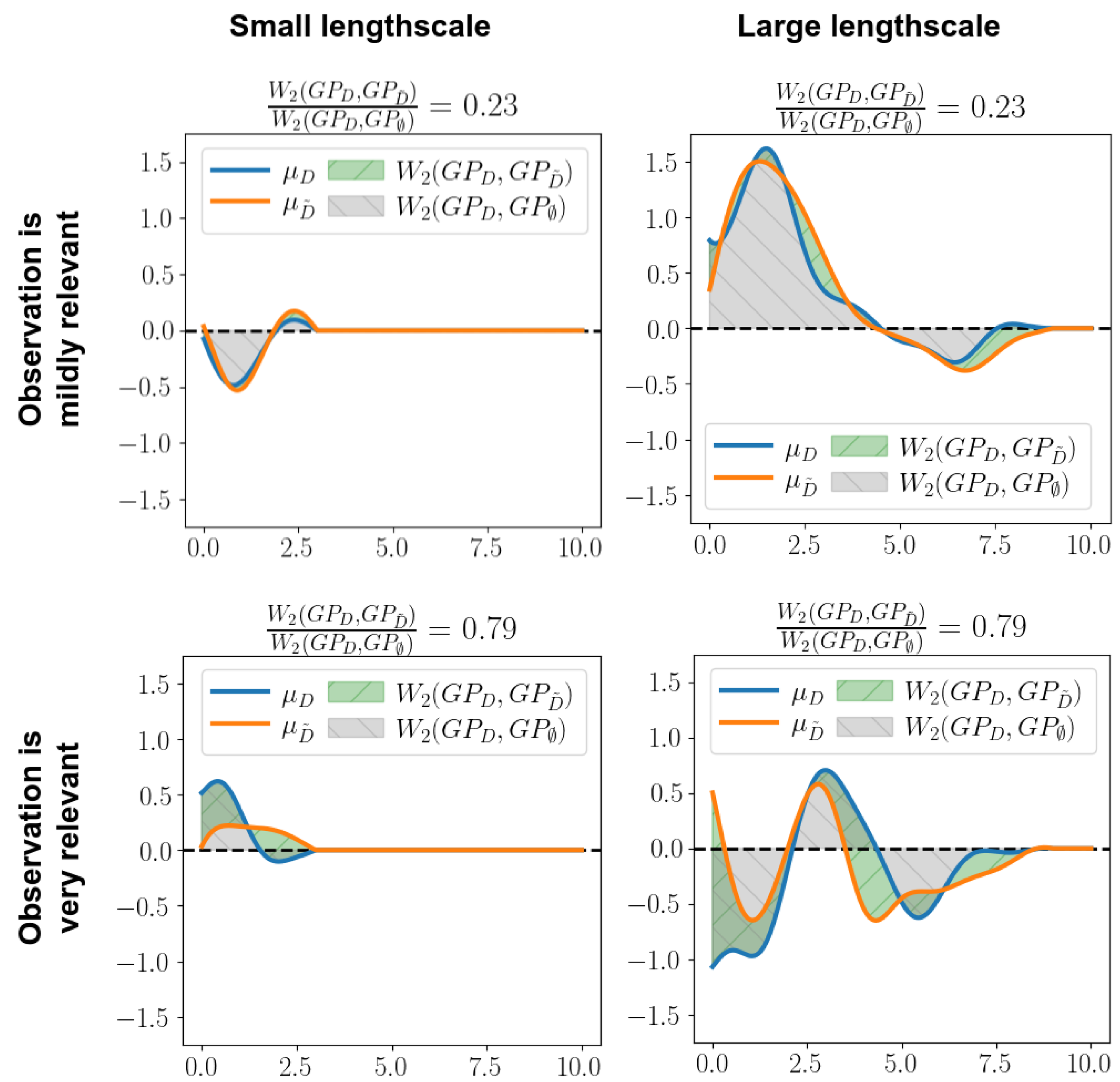
Given an observation $((\boldsymbol{x}_i, t_i), y_i) \in \mathcal{D}$ at the current running time $t_0$, consider the alternative dataset $\tilde{\mathcal{D}} = \mathcal{D} \setminus \{((\boldsymbol{x}_i, t_i), y_i)\}$ , as well as the two posteriors $\mathcal{GP}_\mathcal{D}\left(\mu_\mathcal{D}, \sigma^2_\mathcal{D}\right)$ and $\mathcal{GP}_{\tilde{\mathcal{D}}}\left(\mu_{\tilde{\mathcal{D}}}, \sigma^2_{\tilde{\mathcal{D}}}\right)$. We measure the similarity between the posteriors with **the 2-Wasserstein distance**

$$W_2^2(\mathcal{GP}_\mathcal{D}, \mathcal{GP}_{\tilde{\mathcal{D}}}) = \oint_\mathcal{S} \int_{t_0}^{\infty} ((\mu_\mathcal{D}(\boldsymbol{x}, t) - \mu_{\tilde{\mathcal{D}}}(\boldsymbol{x}, t))^2 + (\sigma_\mathcal{D}(\boldsymbol{x}, t) - \sigma_{\tilde{\mathcal{D}}}(\boldsymbol{x}, t))^2) d\boldsymbol{x} dt$$

**Definition** The relevancy $R(\boldsymbol{o}_i)$ of an observation $\boldsymbol{o}_i = ((\boldsymbol{x}_i, t_i), y_i) \in \mathcal{D}$ is given by the normalized Wasserstein distance

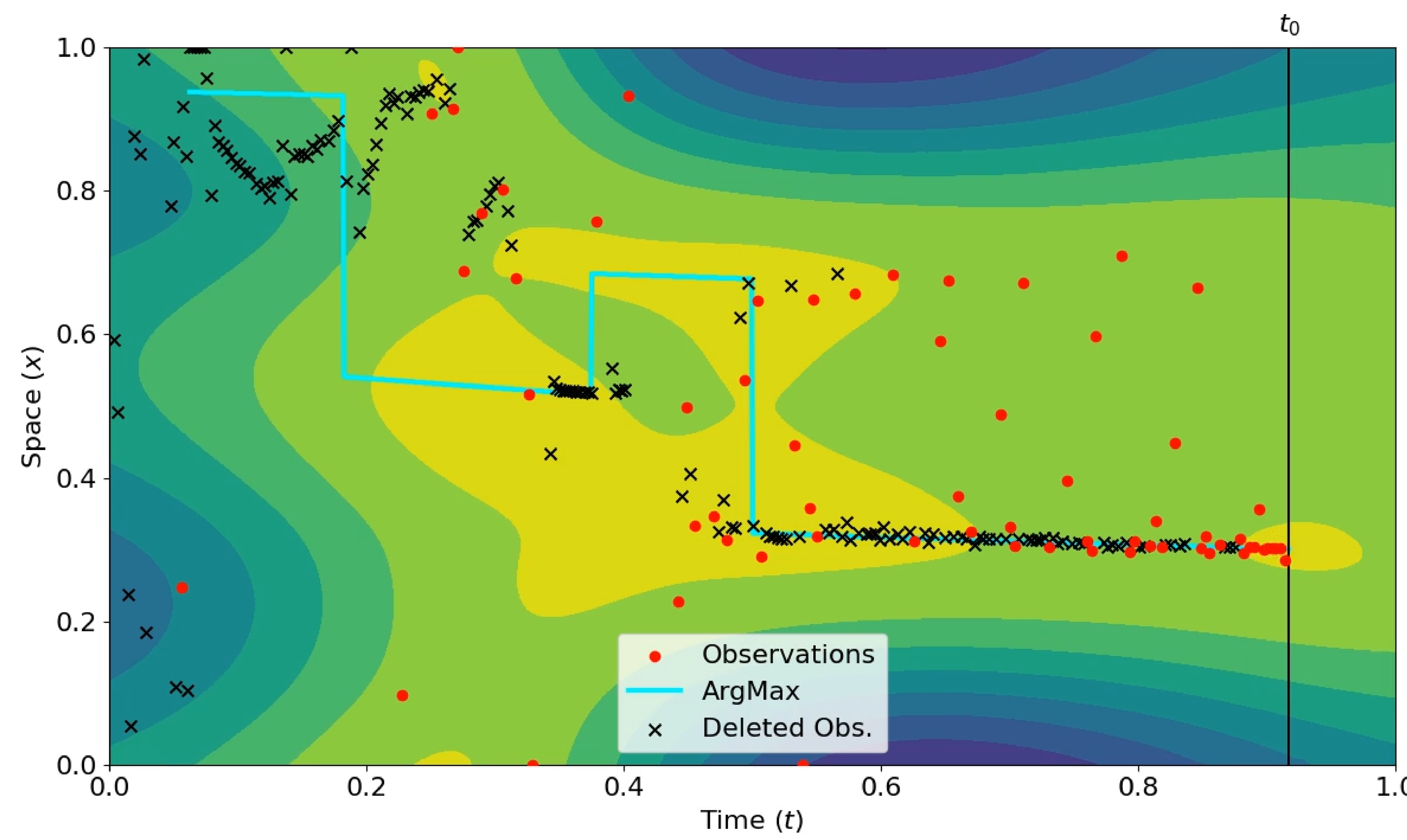
$$R(\boldsymbol{o}_i) = \frac{W_2(\mathcal{GP}_\mathcal{D}, \mathcal{GP}_{\tilde{\mathcal{D}}})}{W_2(\mathcal{GP}_\mathcal{D}, \mathcal{GP}_\emptyset)}$$

where $\mathcal{GP}_\emptyset$ is the prior $\mathcal{GP}_\emptyset(0, \lambda)$ and $W_2(\mathcal{GP}_\mathcal{D}, \mathcal{GP}_\emptyset)$ is a normalization constant that cancels the influence of $\boldsymbol{\theta}$ on $W_2(\mathcal{GP}_\mathcal{D}, \mathcal{GP}_{\tilde{\mathcal{D}}})$.

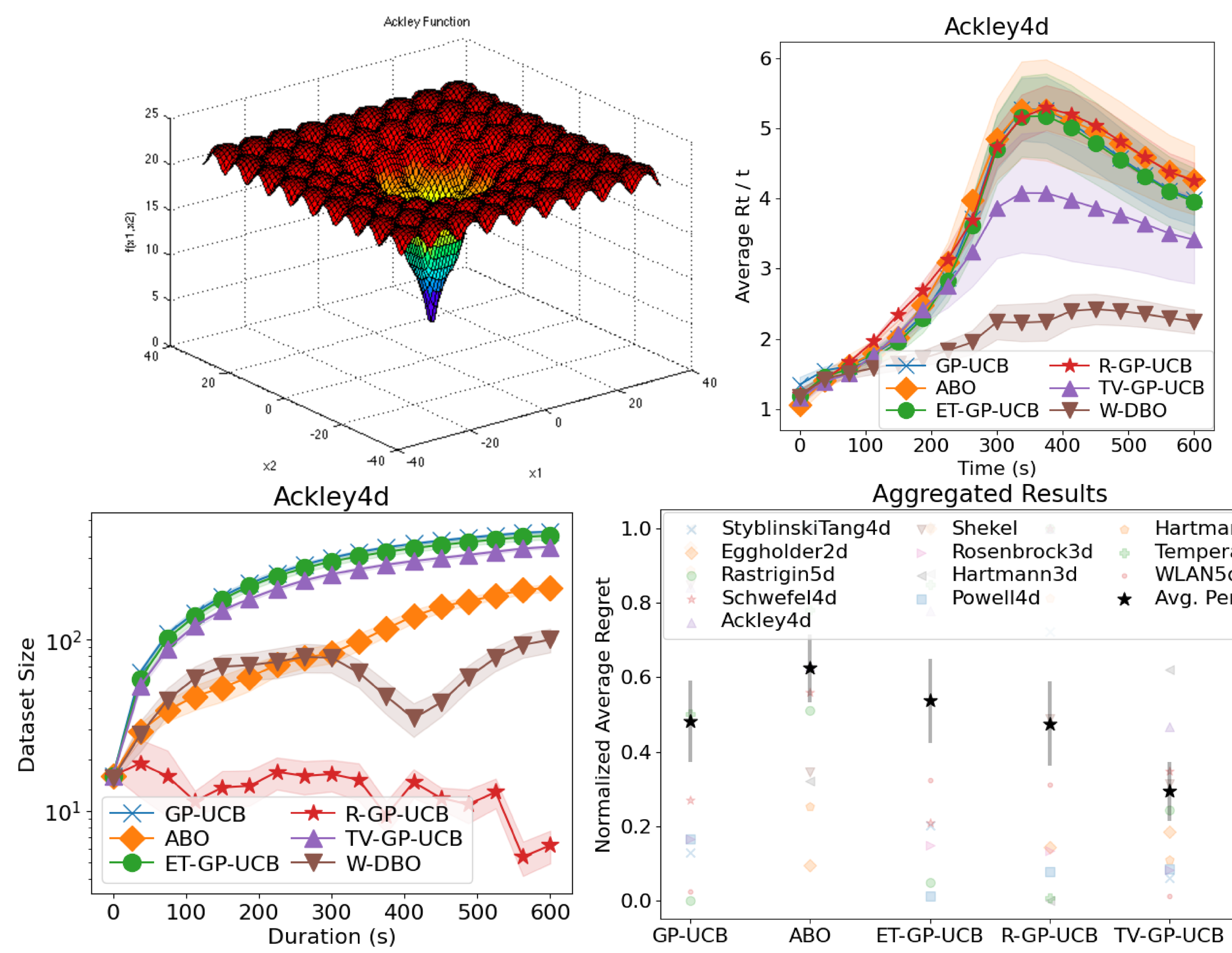



## The W-DBO Algorithm

At the end of each iteration of **any** BO algorithm, iteratively remove the observation $\boldsymbol{o}^* = \arg\min_{\boldsymbol{o}_i \in \mathcal{D}} R(\boldsymbol{o}_i)$ until a deletion budget is consumed.




## Numerical Results




## Summary

- A DBO algorithm that does not remove irrelevant observations **performs poorly on tasks with large time horizons**, because it gets prohibitive to use.
- **Observation relevancy** can be measured on the fly with an **intuitive criterion** involving the Wasserstein distance.
- **Removing irrelevant observations yields a significant performance boost**, both on the average regret and on the sampling frequency of the DBO algorithm.


**Contact**: {anthony.bardou, patrick.thiran, giovanni.ranieri}@epfl.ch