Forgetting to Improve: Principled Data Removal in Active Learning

Manuel Wendl [*1,2] **Erik Englesson** [1] **Andreas Krause** [2] **Carl Henrik Ek** [1,3]
[1] University of Cambridge [2] ETH Zürich [3] Karolinska Institutet

Abstract

The uncertainty of a statistical model is most commonly factorized into an aleatoric and an epistemic part. This factorization changes how predictions are interpreted in downstream decision tasks. Importantly, except for the idealistic scenario with no model mismatch, the quantification is a characteristic of the model and not the data generating process. In this paper, we propose Forgetting to Improve a method that reduces this discrepancy by incorporating the task into the modeling framework. Our key insight is to acknowledge that in scenarios of model mismatch, data can have a detrimental effect on the modeling for a specific task. Based on this insight, we propose an influence function for Gaussian process models that allows for principled removal of detrimental data samples. We showcase the flexibility of this approach by demonstrating significant improvements across a range of tasks, including Bayesian optimization, model-based reinforcement learning, and transductive learning.

1 Introduction

Gaussian process (GP) models [59] are a popular choice for probabilistic modeling. Given their strong priors they are especially popular in low data regimes and their interpretable uncertainties makes them ideally suited for sequential decision tasks such as Bayesian optimization [13, 21] and Reinforcement Learning [16, 35]. For the modeling assumption of GPs, the inference procedure leads to a natural factorization of the uncertainty into an *aleatoric* and an *epistemic* part. As we often specify the model before seeing the data, there are scenarios where the model does not correspond to the data generation process, which we refer to as *model mismatch*. Box [11] famously stated "all models are wrong but some are useful" which means that these assumptions very rarely match the data generation process. In *active learning* scenarios such as Bayesian optimization, modeling and decision making are often seen as two separate tasks. We first model the data, then make decisions on where to model next. In scenarios of model mismatch this can have severely limiting effects, as we might introduce data, which has detrimental effects on the model that is relevant to the task.

More formally, the interaction of the *aleatoric* and *epistemic* components of the predictive uncertainty is not intrinsic to the data alone, but is induced by the choice of model and the task [64]. First, in most cases, the choice of model is made a priori, and may be misaligned for the considered task, while we retain full control over the subset of data used during training. Secondly, many machine learning tasks only require accurate predictions in a *target region* of the input domain, whereas available observations may originate from a broader region with a potentially different distribution. This distribution shift may lead to a misattribution between the aleatoric and epistemic uncertainty within the target region. As a result, incorporating additional observations can have adverse effects: **(i)** the aleatoric noise level in the target region may increase significantly when certain observations are included, and **(ii)** may cause some samples to *counteract* variance reduction precisely in the target region. Both effects arise because observations that improve the global model fit may be uninformative or even detrimental to predictions in the target region. This is particularly interesting, since in many machine learning tasks, such as *transductive learning* [47] and *experimental design*

*Correspondence to: mwendl@ethz.ch.

Accepted at ICML 2026 Workshop on Decision-Making from Offline Datasets to Online Adaptation: Black-Box Optimization to Reinforcement Learning.

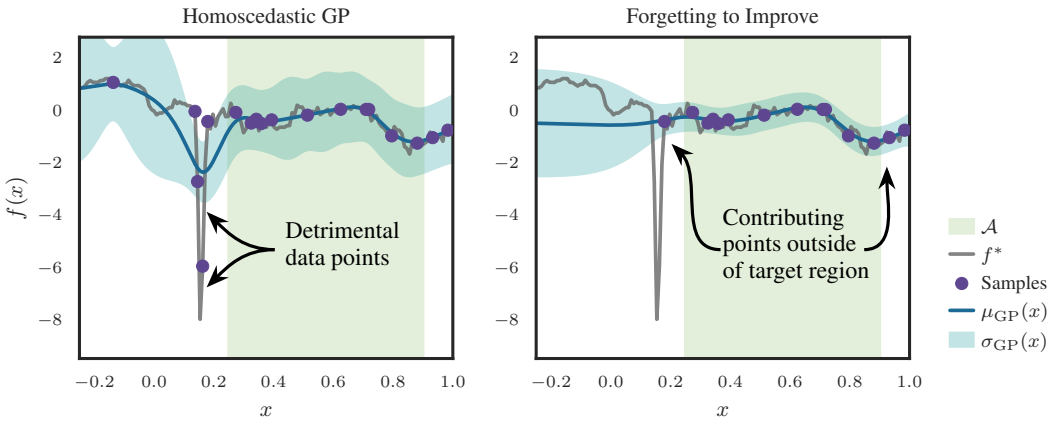


Figure 1: Influence of detrimental data structures based on the Dow Jones data set [4]. (Left) Homoscedastic GP absorbs detrimental structures in the noise, resulting in uncertain predictions in $\mathcal{A}$. (Right) Forgetting to Improve (F2I) removes detrimental data to improve predictions in $\mathcal{A}$.

[12], the target region in which predictions are made is known beforehand. This knowledge is used in many sequential learning algorithms for BO and RL to select new samples according to information-theoretic acquisition functions, under the assumption that more observed samples strictly improve performance. However, in scenarios of model mismatch, we obtain the described non-trivial model-data dependency of the aleatoric and epistemic uncertainty, which results in potentially detrimental influences of early observed samples for the entire learning process.

To address these limitations, we propose Forgetting to Improve (F2I), a principled data removal algorithm that *"forgets"* data points with detrimental influence on the predictions within the target region in transductive learning tasks. As we cannot rely on having access to labels to measure generalization in the target region, we focus instead on reducing predictive uncertainty. This is achieved by determining the influence of individual samples on the predictive uncertainty, consisting of the model-data dependent and location-specific epistemic and global aleatoric uncertainties. Samples with detrimental influence are removed to improve predictions and confidence in transductive learning, as visualized for a real-world stock market data set in Figure 1. Especially sequential learning tasks, such as BO and RL, can benefit significantly from this removal paradigm when detrimental samples have been evaluated in the beginning of the optimization process.

Our contributions include theoretical grounding and practical evidence of Forgetting to Improve:

- We propose a principled measure to quantify the contribution of individual samples to the epistemic and aleatoric components of the predictive uncertainty.
- We introduce the greedy data removal algorithm F2I to "forget" detrimental data points in transductive learning tasks. In addition, we derive a near-optimality guarantee for F2I and provide a rigorous analysis by establishing approximation bounds on the predictive uncertainty with respect to the optimal data subset.
- We extend F2I to Forgetful BO and Forgetful RL, formulating BO and RL as sequential transductive learning tasks with dynamical target regions.
- We conduct an extensive empirical evaluation of F2I and Forgetful BO on several different transductive and sequential learning tasks.

2 Related Literature

Prior works on probabilistic models and their use in uncertainty-aware learning algorithms fall broadly into two categories: **(i)** principled methods that focus on choosing the best next unseen sample to improve predictions and **(ii)** algorithms aiming to improve predictions given a fixed set of samples.

Targeted Data Acquisition A canonical line of work focuses on *"where to sample next"*. While active learning aims to optimally select informative sampling locations, transductive active learning has additional knowledge about the desired target region, where predictions will be made. Early work [47] introduced information-theoretic criteria for selecting samples that are most informative for a fixed target region. This perspective has recently been revisited empirically for different settings [38, 63, 70],

and information-theoretically extended in [62, 73]. Recently, optimality guarantees were derived in [30] for these decision rules in transductive learning and BO. Closely related ideas appear in optimal sensor placement and Bayesian experimental design, for which acquisition functions have been designed to maximize information gain or reduce posterior uncertainty [29, 41, 62]. However, these approaches require strict assumptions on well-behaved objectives and focus exclusively on selecting new samples and leveraging all collected data. Meaning that once a sample is acquired, it is never revisited, even if it later proves detrimental under model mismatch for the specific task. In contrast, our work addresses the complementary and largely unexplored question *"which samples should be removed"*.

Robustness via Correction and Reweighting A large body of work addresses robustness in probabilistic modeling with data preprocessing, alternative likelihoods, or noise modeling. Data curation and preprocessing methods [14] aim to remove corrupted samples before training. These methods are typically model-agnostic, assume access to the full dataset beforehand, may introduce unintended biases, and have limited applicability in sequential tasks. In contrast, robust GP formulations modify the likelihood or noise model to mitigate the influence of detrimental data structures. This includes heavy-tailed likelihoods, such as Student-t [34], Laplace [58], and Huber losses [1]. These have been proven to be effective against global noise misspecification. However, noise can also be locally dependent, which is addressed in a different line of work by introducing noise models to determine input-dependent noise distributions using heteroscedastic [25, 36, 44, 71] or twinned GPs [51]. Although these approaches allow local noise variations, the underlying noise distribution has to be smooth with respect to the input. Additionally, these methods do not take into account task-specific prediction regions and allow modeling structures of interest with the noise model. More closely related methods remove or down-weight data points. Trimmed likelihood approaches select subsets of data that maximize the marginal likelihood [5, 7], however require a prespecified fraction of detrimental samples [7], or converge to a locally stationary point without optimality guarantees [5]. Therefore, these methods may unnecessarily remove points for well-behaved objectives or in the target region of our predictions. Other strategies rely on residual-based heuristic pruning [46] or structured noise models that assume clustered outliers [2]. Latent variable and input warping approaches aim to absorb the model mismatch by reshaping the input space [10, 65]. The latent GP model separates inputs with detrimental structure by introducing an additional dimension and performing computationally expensive posterior sampling with MCMC over the separation, while input warping is only capable of transforming the entire input space globally, ignoring local structures. Related work also considers modifications in the output space, by assuming biased observations of the form $y(x) = f(x) + \delta(x) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ denotes random noise and $\delta(x)$ captures the sample specific bias [4, 55]. While both methods introduce learnable data-point-specific biases, [55] optimizes the negative marginal log likelihood in a single step, which generally does not lead to sparsity in the biases, in contrast to [4]. Consequently, [10, 4] are most closely related to our work, as they induce effects similar to the removal of detrimental samples by distancing them in the input space or adding a large bias in the output space. Unlike latent-space or bias-based methods, which implicitly reduce the influence of detrimental samples through additional modeling assumptions, our approach explicitly selects a task-optimal subset of data with theoretical guarantees. Furthermore, our per-sample influence is related to classical influence-functions and data-valuation approaches [37, 23], that quantify single sample attributions on a global estimator or loss. Crucially, our measure enables principled task-specific removals rather than global reweighting.

3 Problem Setting

We consider in this work a standard Gaussian process regression setting. Let $\mathcal{X}$ be the input space and $f : \mathcal{X} \to \mathbb{R}$ be an unknown function. We model f as a GP $f \sim \mathrm{GP}(0, k(\cdot, \cdot))$, with an a priori chosen kernel k. We are given a dataset $\mathcal{S} \coloneqq \{x_i, y_i\}_{i=1}^n \subseteq \mathcal{X} \times \mathbb{R}$, with observations $y_i = f(x_i) + \epsilon_i$, and $\epsilon_i \sim \mathcal{N}(0, \sigma_w^2)$. In practice the noise variance σ_w^2 depends on the measurement noise and model-data mismatch. The strength of this dependence may be controlled by placing a prior $p(\sigma_w^2)$ on the noise variance, and solving for the *type II maximum a posteriori* (MAP II) estimate [50]

$$\sigma_w^2(\mathcal{S}) = \arg\max_{\sigma_w^2} \mathcal{L}(\sigma_w^2, \mathcal{S}) \coloneqq \log \prod_{i=1}^n p(y_i | x_i, \sigma_w^2) + \log p(\sigma_w^2). \tag{1}$$

The predictive distribution of a GP trained on the subset $\mathcal{S}' \subseteq \mathcal{S}$ is Gaussian at any x with variance

$$\sigma_{\mathcal{S}'}^2(x) = k(x, x) - k_{x,\mathcal{S}'}(K + \sigma_w^2(\mathcal{S}')I)^{-1} k_{\mathcal{S}',x} + \sigma_w^2(\mathcal{S}')I, \tag{2}$$

where $K = k(\mathcal{S}', \mathcal{S}')$ is the Gram matrix over $\mathcal{S}'$, $k_{x,\mathcal{S}'} = [k(x, x_i)]_{x_i \in \mathcal{S}'}$, and we explicitly denote the dependence of the noise estimate on the subset by $\sigma_w^2(\mathcal{S}')$.

Our goal is to find a subset $\mathcal{S}' \subseteq \mathcal{S}$ that improves the GP predictive in a target region $\mathcal{A} \subseteq \mathcal{S}$ without having access to ground truth labels in the entire region. Without labels, it is hard to guide the subset selection to improve the predictive mean. Instead, the *predictive uncertainty* becomes a natural objective, as it characterizes the information content of the posterior distribution in $\mathcal{A}$. Therefore, we aim to *find the subset $\mathcal{S}' \subseteq \mathcal{S}$ that minimizes the predictive uncertainty in $\mathcal{A}$.*

$$\min_{\mathcal{S}' \subseteq \mathcal{S}} \mathcal{J}(\mathcal{S}'), \qquad \mathcal{J}(\mathcal{S}') = \sum_{x \in \mathcal{A}} \sigma_{\mathcal{S}'}^2(x). \tag{3}$$

This closely relates to established principles in *Bayesian Optimization* and *Reinforcement Learning*, where uncertainty guides exploration and balances exploitation (see Sections 5 and 6).

4 Forgetting to Improve – Principled Data Removal

We propose Forgetting to Improve (F2I), a principled data-removal algorithm to find the subset $\mathcal{S}'$ in Equation (3) by greedily removing data points based on our predictive uncertainty influence measure. The influence measure naturally decomposes into the sum of per-example aleatoric and epistemic terms. In this section, we derive these terms, present the full algorithm, and discuss its optimality.

Tracing Epistemic Uncertainty Back to Data A natural strategy for data selection, deeply rooted in related works of active learning, is to select data with the *maximum gain in information* for the given task. For transductive learning, this reduces to choosing the sample that *minimizes the epistemic uncertainty* in the target region $\mathcal{A}$ as much as possible. The subset selection in Equation (3) is combinatorial and computationally infeasible for large data sets. Leave-one-out (LOO) methods instead greedily remove the sample with the highest influence, which for example i is defined by $\sigma^2_{\mathcal{S}\setminus\{x_i,y_i\}} - \sigma^2_{\mathcal{S}}$. However, this requires a model refit for each individual removal candidate from the current subset. To solve this, we first generalize the discrete LOO case via a continuous removal direction h where $k(x, \mathcal{S}, h) = k_{x,\mathcal{S}} - h \odot k_{x,\mathcal{S}}$ in Proposition 5. Thus, the predictive uncertainty $\sigma^2_{\mathcal{S}}(h)$ depends on h, where h being the zero vector or the standard basis e_i corresponds to $\sigma^2_{\mathcal{S}}$ and $\sigma^2_{\mathcal{S}\setminus\{x_i,y_i\}}$, respectively. Now, to get an efficient approximation to the LOO influence, we do a first-order Taylor expansion of $\sigma^2_{\mathcal{S}}(h)$ around the full dataset $\mathcal{S}$ ($h = 0$), and evaluate at $h = e_i$: $\sigma^2_{\mathcal{S}\setminus\{x_i,y_i\}} - \sigma^2_{\mathcal{S}} \approx e_i \cdot \nabla_h \sigma^2_{\mathcal{S}}|_{h=0}$. Thus, we only require one gradient computation rather than several LOO evaluations per removal.

Proposition 1. *Suppose we have a single test point $\mathcal{A} = \{x\}$ and the corresponding GP prediction with mean $\mu_{\mathcal{S}}(x)$ and variance $\sigma^2_{\mathcal{S}}(x)$, for a given set of samples $\mathcal{S} = \{x_i, y_i\}_{i=1}^n$. The epistemic influence $u_i(x)$ of an individual sample $x_i \in \mathcal{S}$ on the predictive uncertainty is defined as the partial derivative of $\sigma^2_{\mathcal{S}}(x)$ w.r.t. the removal in h_i for sample $x_i \in \mathcal{S}$ under fixed aleatoric noise $\sigma^2_w(\mathcal{S})$*

$$u_i(x) = 2k_{x,x_i}(A^{-1}k_{\mathcal{S},x})_i - 2(A^{-1}k_{\mathcal{S},x})_i(k_{x_i,\mathcal{S}}A^{-1}k_{\mathcal{S},x}),$$
$$\textit{where } k_{x,\mathcal{S}} = k(x, \mathcal{S}), \textit{ and } A = K + \sigma^2_w(\mathcal{S})I, \textit{ with } K = k(\mathcal{S}, \mathcal{S}).$$

Furthermore, in Proposition 2 we can interpret our derived influence measure u_i as the directional derivative of the uncertainty difference for sample removal x_i into the kernel direction of test points x, unifying both dependance of task and model in the dependence on x and $k(\cdot)$.

Proposition 2. *Suppose the uncertainty difference at $\mathcal{A} = \{x\}$, when removing sample i is given by*

$$\Delta_i(x) = \sigma^2_{\mathcal{S}\setminus\{x_i,y_i\}}(x) - \sigma^2_{\mathcal{S}}(x) = \frac{(A^{-1}k_{\mathcal{S},x})_i^2}{A^{-1}_{i,i}}.$$

Then, our uncertainty measure u in Proposition 1 captures the half directional derivative along the kernel direction $k_{x,\mathcal{S}}$, and k_{x,x_i} for an individual component of u:

$$k_{x,x_i}\frac{\partial \Delta_i}{\partial k_{x,x_i}} = 2k_{x,x_i}(A^{-1}k_{\mathcal{S},x})_i - 2(A^{-1}k_{\mathcal{S},x})_i(k_{x_i,\mathcal{S}}A^{-1}k_{\mathcal{S},x}) = u_i(x).$$

We refer to Section B for the formal proofs of Propositions 1 and 2. Please note that the absolute uncertainty reduction $\Delta_i(x)$ is strictly positive due to the quadratic term in the nominator and the positive elements of A, whereas u_i can be negative. This indicates that an infinitesimal *increase* in the coupling k_{x,x_i} *decreases* the variance reduction and is therefore *locally redundant* or *counteracting*.

How Data drives Aleatoric Uncertainty under Model Mismatch The derived epistemic influence measure u_i captures uncertainty differences, explainable by the kernel k. The detrimental structures in the data that cannot be captured by the kernel are modeled using independent white noise $\sigma_w^2(\mathcal{S})$ in a homoscedastic GP setting. This creates a strong dependence of the *training data* on the *model-data mismatch* absorbed by the noise $\sigma_w^2(\mathcal{S})$. We place a Gamma prior $p(\sigma_w^2|a,b)$ on the noise variance and derive in Proposition 3 the derivative of the optimal noise MAP II estimate in the removal direction h_i, that influences both the reproducing kernel Hilbert space (RKHS) and the output space (Proposition 5).

Proposition 3. *Let $A(\sigma_w(\mathcal{S})) = K(\mathcal{S},\mathcal{S}) + \sigma_w^2(\mathcal{S})I$ and suppose that $\sigma_w^2(\mathcal{S})$ is the optimal noise variance of the log MAP II (Equation (1)) $\frac{\partial}{\partial \sigma_w^2}\mathcal{L}(\sigma_w^2,\mathcal{S},a,b) = \mathcal{D}(\sigma_w^2,\mathcal{S},a,b) = 0$ for the samples in $\mathcal{S}$. Then we determine the derivative of the optimal noise floor $\sigma_w^2(\mathcal{S})$ into the removal direction h_i using implicit differentiation $\frac{\partial \sigma_w^2}{\partial h_i} = -\frac{\partial}{\partial h_i}\mathcal{D}(\sigma_w^2,\mathcal{S},a,b)/\frac{\partial}{\partial \sigma_w^2}\mathcal{D}(\sigma_w^2,\mathcal{S},a,b)$, which results in*

$$\frac{\partial \sigma_w^2(\mathcal{S})}{\partial h_i} = \frac{-(A^{-1}y)_i(k_{x_i,\mathcal{S}}(A^{-2}y)) - (A^{-2}y)_i(k_{x_i,\mathcal{S}}(A^{-1}y)) + k_{x_i,\mathcal{S}}(A^{-2})_{\cdot,i}}{-y^\top A^{-3}y + \frac{1}{2}\operatorname{tr}(A^{-2}) - \frac{a-1}{(\sigma_w^2)^2}}.$$

Given this sensitivity, we now derive the aleatoric influence measure $v_i(x)$, given by the partial derivative of the aleatoric contribution of the predictive uncertainty with respect to h_i. This measures how individual removals influence the global noise floor $\sigma_w^2(\mathcal{S})$ and change the predictive uncertainty.

Proposition 4. *Suppose we have a single test point $\mathcal{A} = \{x\}$ and the corresponding GP prediction with mean $\mu_{\mathcal{S}}(x)$ and variance $\sigma_{\mathcal{S}}^2(x)$, for a given set of samples $\mathcal{S} = \{x_i, y_i\}_{i=1}^n$. The aleatoric influence $v_i(x)$ of an individual sample $x_i \in \mathcal{S}$ on the predictive uncertainty is defined as the partial derivative of all aleatoric noise terms $\sigma_w^2(\mathcal{S})$ in $\sigma_{\mathcal{S}}^2(x)$ w.r.t. the removal in h_i for sample $x_i \in \mathcal{S}$.*

$$v_i(x) = (1 + k_{x,\mathcal{S}}A^{-2}k_{\mathcal{S},x})\frac{\partial \sigma_w^2(\mathcal{S})}{\partial h_i},$$

where the partial derivative of the noise parameter is taken from Proposition 3.

We provide the mathematical derivations of Propositions 3 and 4 in Section C.

Forgetting to Improve Propositions 1 and 4 establish how removing samples from $\mathcal{S}$ influences the epistemic and aleatoric components of the predictive uncertainty in a single-point target region $\mathcal{A} = \{x\}$. Building on this, our greedy algorithm Forgetting to Improve (Algorithm 1), iteratively discards samples whose removal reduces the predictive uncertainty in a multi-point target region $\mathcal{A}$, and terminates once no further uncertainty reduction is possible. To do this, we extend Propositions 1 and 4 in two ways. First, we combine $u(x)$ and $v(x)$ to approximate the total change in predictive uncertainty when removing a sample. Here, $u(x)$ captures the reduction due to information gain in the kernel, while $v(x)$ reflects changes in the estimated aleatoric noise. Second, we compute the influence on multi-point target regions by summing the individual influences of the test points $x \in \mathcal{A}$. Finally, as the aleatoric influence relies on the local type II MAP noise variance estimate around the current data subset, we recompute the estimate after each removal. This greedy algorithm is based

Algorithm 1 Forgetting to Improve (F2I)

Require: Target region $\mathcal{A}$, all samples $\mathcal{S}$, and the GP model with kernel k.
  Initialize $h = 0$, and $\mathcal{S}^{(0)} \leftarrow \mathcal{S}$.
  Compute influence measures $u^{(0)}(\mathcal{A})$, and $v^{(0)}(\mathcal{A})$ ▷ Propositions 1 and 4
  while $\min(u^{(h)}(\mathcal{A}) + v^{(h)}(\mathcal{A})) \leq 0$ **do** ▷ Samples that reduce uncertainty
    Determine worst sample: $i = \arg\min_j(u_j^{(h)}(\mathcal{A}) + v_j^{(h)}(\mathcal{A}))$
    Update set of samples $\mathcal{S}^{(h+1)} \leftarrow \mathcal{S}^{(h)} \setminus x_i$ and GP noise MAP II
    Increase $h \leftarrow h+1$ and compute influences: $u^{(h)}(\mathcal{A}), v^{(h)}(\mathcal{A})$ ▷ Propositions 1 and 4
  end while
  return $\mathcal{S}' = \mathcal{S}^{(h+1)}$

on the derivatives of the aleatoric and epistemic uncertainty of Propositions 1 and 4 and can hence be interpreted as a "*reverse*" orthogonal matching pursuit algorithm [69] for the predictive uncertainty in $\mathcal{A}$. We therefore consider the induced set-function $\bar{\mathcal{J}}(T) = \mathcal{J}(\mathcal{S}) - \mathcal{J}(\mathcal{S} \setminus T)$, which measures the decrease in predictive uncertainty (gain) from removing T. While $\mathcal{J}$ itself is not submodular, $\bar{\mathcal{J}}$ can be shown to exhibit approximate diminishing returns under suitable curvature conditions. Building

on approximate submodularity theory [15, 18, 40], we derive a curvature condition for the near optimality of the removals of Algorithm 1. The result relies on a curvature condition that ensures approximate submodularity of the objective. In Lemma 7 in Section L, we provide a sufficient condition on the curvature of $-\mathcal{J}$ (based on Lemma 4), which depends on the Gamma noise prior parameter a, under which these assumptions hold for all potential removals R.

Theorem 1. *Let the set-function $\bar{\mathcal{J}}$ be defined as $\bar{\mathcal{J}}(T) = \mathcal{J}(\mathcal{S}) - \mathcal{J}(\mathcal{S} \setminus T)$ for a compact $\mathcal{A}$ and bounded, twice continuously differentiable kernel k. If the inner MAP II for σ_w^2 has a unique solution and $-\mathcal{J}$ is m_R restricted strong concave and M_R restricted smooth for all potential removals $R \supseteq T$ (Definition 1), satisfying Lemma 7, then approximate submodularity of $\bar{\mathcal{J}}$ guarantees*

$$\bar{\mathcal{J}}(\mathcal{S}') \geq (1 - e^{-\gamma}) \max_{\substack{\mathcal{S}^* \subseteq \mathcal{S} \\ |\mathcal{S}^*| = |S'|}} \bar{\mathcal{J}}(\mathcal{S}^*), \quad \textit{where } \gamma = \tfrac{m_R}{M_R}.$$

Proof Sketch A detailed outline of the proof is given in Section L. First, we define the corresponding set-function $\bar{\mathcal{J}}$ for the objective $\mathcal{J}$ in Lemma 1. Further, we show that $-u_i(\mathcal{A}) - v_i(\mathcal{A})$ corresponds to the gradient of $-\mathcal{J}$ in Lemma 2. Given that $-\mathcal{J}$ is restricted strongly concave and restricted smooth, we use Lemmas 8 and 9 to establish approximate submodularity of $\bar{\mathcal{J}}$. Combining this with the sample selection rule in Algorithm 1, we obtain the near-optimality guarantee of Theorem 1.

5 Forgetful Bayesian Optimization

We extend F2I to Bayesian optimization by viewing BO as a sequence of transductive learning problems, each defined by a dynamically shrinking region of interest. The key idea is to remove samples that distort posterior uncertainty exactly where the acquisition function relies on it. This allows us to improve optimization performance without modifying the acquisition function itself. Therefore, we consider an unknown objective function $f(x) : \mathcal{X} \to \mathbb{R}$, with the goal of solving the global optimization problem $x_{max} = \arg\max_{x \in \mathcal{X}} f(x)$ under a limited evaluation budget.

Model Mismatch in Bayesian Optimization The model-mismatch effects studied in Section 4 are particularly problematic in Bayesian optimization. Since BO relies on uncertainty-aware acquisition functions, such as the GP Upper Confidence Bound (GP-UCB) [66], logarithmic Expected Improvement (logEI) [3], Expected Improvement (EI) and Probability of Improvement [33, 49], misleading posterior uncertainty can directly harm exploration and exploitation decisions. Prior work addresses this issue through alternative likelihoods [48], latent or warped input spaces [10, 65], tree-based structure [32], or preference-based feedback [26]. In contrast to these approaches, which modify the model or acquisition function, we address the model mismatch by explicitly revising the set of observations used to form the posterior uncertainty, in each iteration.

Target Regions in Bayesian Optimization In order to address this problem, we propose to view BO as sequential transductive learning [30]. At each iteration, only a subset of the search space can plausibly contain the maximum. Improving predictive uncertainty outside this subset is therefore unnecessary and can even be harmful under model mismatch. We formalize this intuition by defining a region of interest and applying F2I for this region. The GP-UCB algorithm [66] provides a natural way to define such regions of interest. At each iteration t, future samples are selected by maximizing the upper confidence bound within regions where improvement is possible; specifically, where the upper confidence bound exceeds the current maximum lower confidence bound. Improving the model predictions on $\mathcal{A}_t$ with F2I results in better uncertainty bounds, which is traded for a potentially increased epistemic uncertainty outside of $\mathcal{A}_t$ due to sample removals. Instead of keeping $\mathcal{A}_t$ fixed throughout the removal process, we leverage the improved confidence bounds for the samples $\mathcal{S}_t^{(k)}$ after the k-th removal to update $\mathcal{A}_t^{(k)}$ accordingly. However, removing samples may increase uncertainty outside $\mathcal{A}_t^{(k)}$, potentially creating artificial improvement regions. To prevent this effect, we enforce that the region of interest can only shrink over forgetting iterations and propose the following intersecting update for the target region

$$\mathcal{A}_t^{(k)} \leftarrow \{x_i \in \mathcal{X} : \underbrace{\mu_{\mathcal{S}_t^{(k)}}(x_i) + \beta_t \sigma_{\mathcal{S}_t^{(k)}}(x_i)}_{\mathrm{UCB}(x_i)} > \max_{x_m \in \mathcal{X}} \underbrace{\mu_{\mathcal{S}_t^{(k)}}(x_m) - \beta_t \sigma_{\mathcal{S}^{(k)}}(x_m)}_{\mathrm{LCB}(x_m)}\} \cap \mathcal{A}_t^{(k-1)}. \quad (4)$$

Accordingly chosen constants β_t ensure that the model is well calibrated [66] and therefore $\mathcal{A}_t$ contains all points that could outperform the current best estimate. We combine the data removal algorithm F2I (Algorithm 1) together with the recursive target region update in a standard BO pipeline. The full algorithm Forgetful BO is detailed in Algorithm 2 in Section D.

6 Forgetful Reinforcement Learning

Building on F2I, we investigate by introducing Forgetful RL how the data removal paradigm mitigates model-data mismatch in model-based reinforcement learning. We consider a discounted Markov Decision Process (MDP) $(\mathcal{X}, \mathcal{A}, p, r, \gamma, \rho_0)$ with state space $\mathcal{X}$, action space $\mathcal{A}$, transition probabilities $p(x'|x,a)$, rewards $r(x,a)$, discounting $\gamma \in (0,1)$, and initial state distribution ρ_0. In MBRL [16, 31], a world model is learned to simulate trajectories, enabling in each episode $n = 1, \dots, N$ the optimization of the policy π_n to maximize the sum of rewards $R(\pi_n) = \mathbb{E}_{\pi_n}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t, a_t)\right]$ in this learned environment, using the corresponding state-action value function [68]

$$Q^{\pi_n}(x,a) = \mathbb{E}_{\pi_n}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t, a_t) \mid x_0 = x, a_0 = a\right], \quad V^{\pi_n}(x) = \mathbb{E}_{a \sim \pi_n(x)}[Q^{\pi_n}(x,a)]. \tag{5}$$

Since we only have limited data at episode n, the convergence and quality of the resulting policy critically depend on both the accuracy and calibrated uncertainty of this probabilistic model. Since the planning trajectories are sampled from the predictive distribution of the world model, low uncertainty in the trajectory estimates is desired. Motivated by this, we transfer the idea of a dynamically determined target region, which is defined as the set of state-action pairs (x,a) with potential policy improvement and from visitation state distributions of equal or better policies. In each iteration, we apply the data removal paradigm of F2I for the target region of improvement and reduce the predictive uncertainty of the model for the state-action pairs in the target region

$$\mathcal{A}_n = \{(x,a) \in \mathcal{X} \times \mathcal{A} : \bar{Q}^{\pi_n}(x,a) \geq \min_{x_m \sim \rho_0} \underline{V}^{\pi_n}(x_m) \wedge \bar{Q}^{\pi_n}(x,a) \geq \underline{V}^{\pi_n}(x)\}, \tag{6}$$

where $\bar{Q}^{\pi_n}$ is an upper confidence bound on all possible state-action value functions, as well as $\underline{V}^{\pi_n}$ being a lower confidence bound under the currently available data. We describe our GP-based Forgetful RL implementation, based on the MBPO implementation of [57] in detail in Section E.

7 Experiments

We evaluate the empirical performance of F2I, Forgetful BO and Forgetful RL, using our implementation, based on GPyTorch [20], BoTorch [8], and MBRL-Lib [57]. Our results focus on **(i)** transductive regression tasks, improving predictions in a priori known target regions; **(ii)** BO, by removing detrimental data structures in the sequential learning tasks; and **(iii)** MBRL under global kernel misspecification of the GP world model.

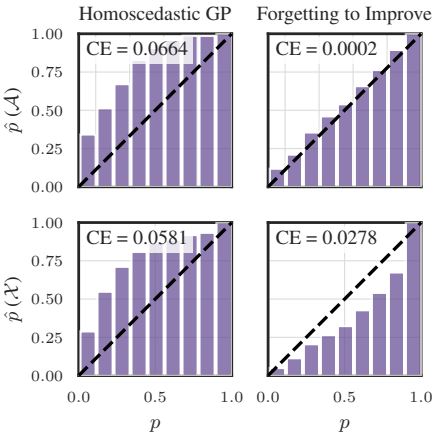


Figure 2: Empirical confidence $\hat{p}$ compared to nominal p. Calibration curves for (left) homoscedastic GP and improvement for (right) F2I on (top) the target region $\mathcal{A}$ and (bottom) the entire domain $\mathcal{X}$ for the example in Figure 1.

Transductive Learning For many real-world applications, the desired prediction location is known a priori, which we refer to as target region $\mathcal{A}$, while the training data $\mathcal{S}$ is distributed across the entire domain $\mathcal{X}$. In the following *transductive learning experiments*, our aim is to improve the predictions in $\mathcal{A}$. Therefore, we evaluate the prediction performance in $\mathcal{A}$ based on the mean squared error (MSE), mean absolute error (MAE), test negative log-likelihood (NLL) [17], and *sharpness*, which is defined as the expected predictive variance in $\mathcal{A}$. Additionally, we evaluate calibration, based on previous work that extends calibration measures from classification to regression [43, 45, 22], to quantify the coverage of predicted confidence intervals (CE) [43], as well as the local calibration comparing the observed residuals (ECE) [45]. Based on these metrics, we compare F2I with a homoscedastic and a heteroscedastic GP [36], domain *warping GP* [65] and *relevance pursuit* [4] on four transductive learning tasks. These include two synthetic test functions, described in Section G, and two real-world tasks using the Boston Housing dataset [28] and the intra-day data from the Dow Jones Industrial Average index [4]. We report in Figure 3 the mean relative performance across 100 different randomly sampled training points using the hyperparameters listed in Section G. The relative performance is computed with respect to the best and worst performing method. The numerical results, including standard deviations, are provided in Section G. As we minimize predictive

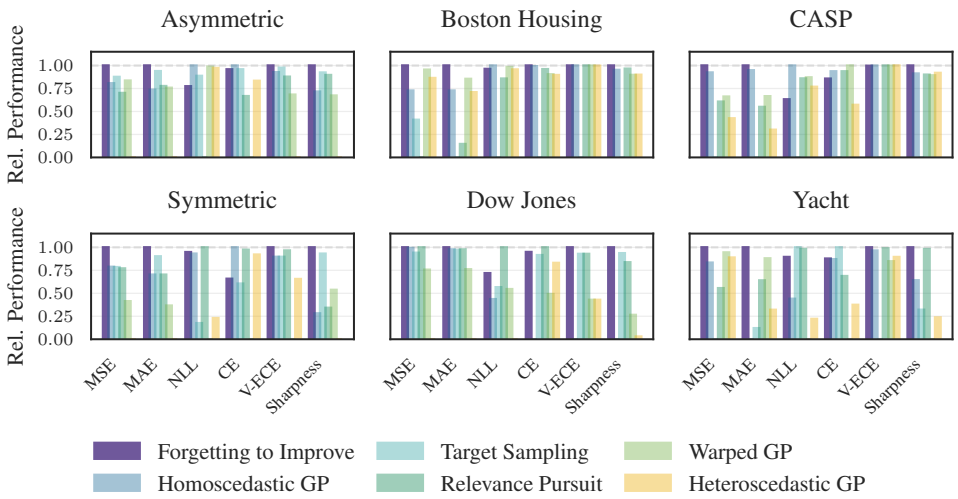


Figure 3: Benchmark comparison (mean over 100 training sample draws). F2I consistently achieves the lowest sharpness, while also improving calibration and accuracy compared to all baselines.

uncertainty with our objective in Equation (3), F2I outperforms all baselines consistently in terms of sharpness in $\mathcal{A}$. We additionally observe that the calibration measures (CE, ECE) benefit significantly from the tighter confidence bounds. Figure 2 shows this improvement for the task in Figure 1. The top row shows the calibration improvement for unseen test samples in the target region $\mathcal{A}$, while the right column shows the tradeoff: underconfident predictions on the full domain $\mathcal{X}$. Remarkably, we also observe a consistent improvement in the accuracy metrics, which we do not explicitly optimize.

Bayesian Optimization We next investigate how these task-oriented prediction improvements translate to the sequential learning process of Forgetful BO. The BO performance is measured using *simple regret*, the difference between the best evaluation up to the current iteration t and the optimal achievable function value $r(t) = f(x^*) - \max_{1 \leq t' \leq t} f(x_t)$. Accumulating simple regret over all T iterations yields the cumulative regret, which is provided in Section I. We compare Forgetful BO to a homoscedastic GP model, the *warped* GP model [65] and the *relevance pursuit* model [4], based on the BoTorch implementations [8], as well as additional baselines in Section H, using the Matérn $5/2$ kernel and a LogNormal$(0, 1)$ for the prior lengthscale. In Figure 4, we use the GP-UCB acquisition function [66], while we also report results for the log expected improvement acquisition function [3] in Section H. We provide statistical results for 10 runs, with random initializations and a limited evaluation budget of $T = 50$. We report in Figure 4 the mean simple regret and standard deviation for 8 test-functions [67], using the default domains, provided together with the dimensionality in Section H. By forgetting detrimental data during the optimization task, Forgetful BO achieves a consistently low simple regret, while the performance of competing baselines strongly varies between test-functions. To demonstrate the advantage of removals, we visualize this behavior for a two dimensional test function in Figure 6 of Section D. In Sections H and J, we additionally demonstrate superior performance of Forgetful BO for noisy tasks with sparse outliers, and the computational efficiency of Forgetful BO.

Reinforcement Learning We demonstrate the full potential of data removal under model-data mismatch, by learning the overall *non-linear Cartpole Balance* task using a *linear* GP for model-based reinforcement learning. From classical control theory, it is well known that the task relevant part of the dynamics around the upright pole position can be described with a linearized model [6]. Since RL learns from trial and error and may fail in early episodes, the system may not remain in this approximately linear domain. We show in Figure 5 (top) that Forgetful RL converges using a linear kernel (left) by constantly removing detrimental nonlinear experiences (middle) for predictions in the shrinking target region $\mathcal{A}_n$ (right). In contrast MBPO [31], which retains all samples, does not solve the task. Also for the Matérn kernel, we observe learning improvements in Section K, as Forgetful RL aims to only model the domain parts that are relevant for solving the task. We visualize this behavior in Figure 5 (bottom) for the experiment in Figure 22 of Section K, where the retained data points lie within the dynamics part relevant to reaching the upward pole position. Forgetful RL automatically removes samples from the attractive regime of the undesired stable poles, compared with the pendulum dynamics plot in the right-hand subplot.

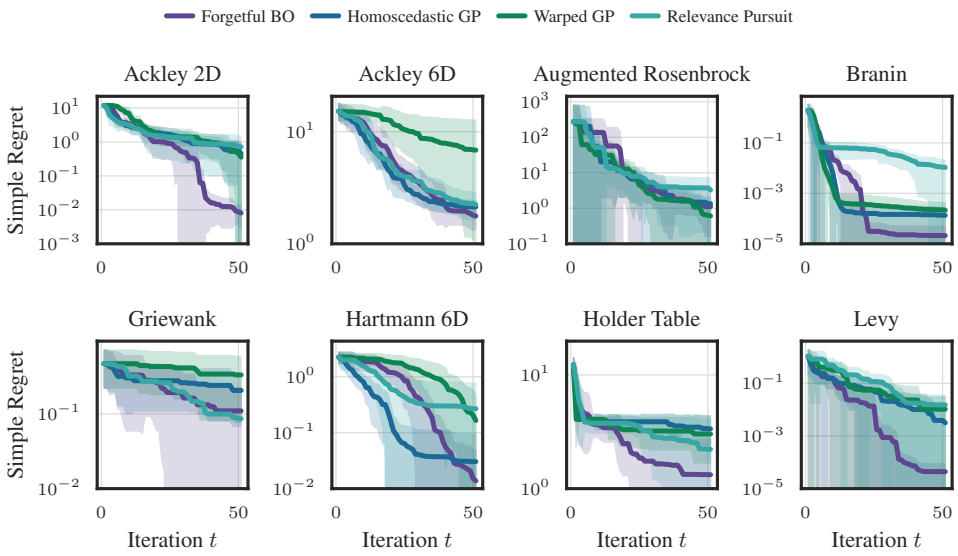


Figure 4: Simple regret across different test functions. Forgetful BO consistently achieves low simple regret, while competing baselines exhibit strong variability across tasks.

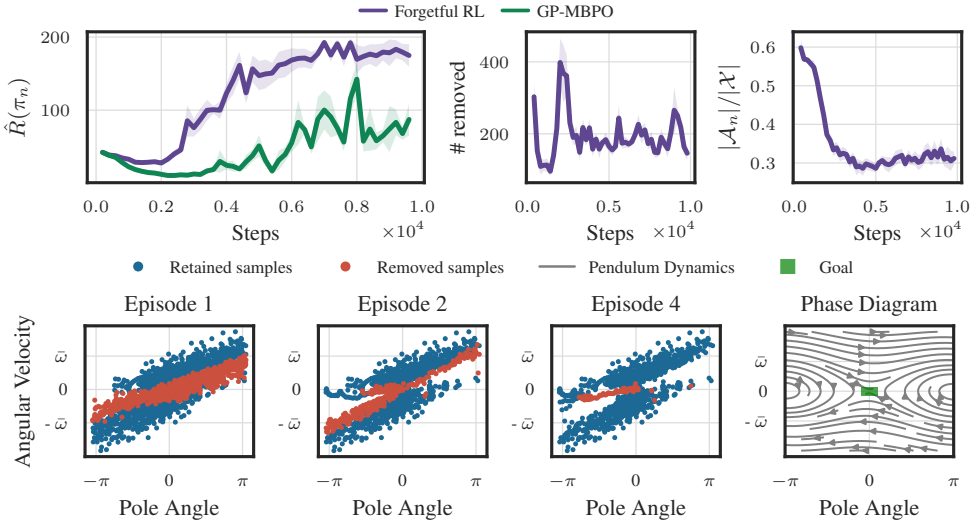


Figure 5: (Top) Forgetful RL succeeds to learn *Cartpole Balance* with a linear GP model, while standard methods suffer from retained nonlinear observations. (Bottom) Visualization of removed and retained samples for Forgetful RL using a Matérn 5/2 kernel. The method keeps task-relevant data near the upright pole and removes samples from attracting dynamics to the undesired stable regimes.

8 Conclusion

This work revisits the fundamental assumption on the irreducibility of aleatoric uncertainty in GP regression, from which the well known paradigm of incorporating all available data to improve the predictive uncertainty originates. For task-specific predictions, we demonstrate that this assumption breaks under model mismatch. To address this, we introduce Forgetting to Improve, a principled data-removal algorithm that explicitly minimizes predictive uncertainty in the *task-specific target region* by identifying and removing samples that are detrimental to these areas under model mismatch. We further extend this paradigm to sequential decision making, reformulating BO and RL as iterative transductive learning problems. Our improvements on regression, BO and RL tasks highlight *data removal* as a complementary and largely underexplored direction for improving uncertainty-aware learning.

Acknowledgement

We would like to thank Amanda Prorok, Jasmine Bayrooti, Massimiliano Tamborski and Yarden As for insightful discussions during this project. We thank the anonymous reviewers for their valuable comments and suggestions. This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) founded by the Knot and Alice Wallenberg Foundation.

References

[1] Pooja Algikar and Lamine Mili. Robust gaussian process regression with huber likelihood. *arXiv preprint arXiv:2301.07858*, 2023.

[2] Matias Altamirano, François-Xavier Briol, and Jeremias Knoblauch. Robust and conjugate gaussian process regression. *arXiv preprint arXiv:2311.00463*, 2023.

[3] Sebastian Ament, Samuel Daulton, David Eriksson, Maximilian Balandat, and Eytan Bakshy. Unexpected improvements to expected improvement for bayesian optimization. *Advances in Neural Information Processing Systems*, 36, 2023.

[4] Sebastian Ament, Elizabeth Santorella, David Eriksson, Ben Letham, Maximilian Balandat, and Eytan Bakshy. Robust gaussian processes via relevance pursuit. *Advances in Neural Information Processing Systems*, 37, 2024.

[5] Daniel Andrade and Akiko Takeda. Robust gaussian process regression with the trimmed marginal likelihood. In *Uncertainty in Artificial Intelligence*, 2023.

[6] Karl Johan Åström and Richard Murray. *Feedback systems: an introduction for scientists and engineers*. Princeton university press, 2021.

[7] Pranjal Awasthi, Abhimanyu Das, Weihao Kong, and Rajat Sen. Trimmed maximum likelihood estimation for robust generalized linear model. *Advances in Neural Information Processing Systems*, 35, 2022.

[8] Maximilian Balandat, Brian Karrer, Daniel R. Jiang, Samuel Daulton, Benjamin Letham, Andrew Gordon Wilson, and Eytan Bakshy. BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization. In *Advances in Neural Information Processing Systems 33*, 2020.

[9] Bernd Bischl, Giuseppe Casalicchio, Matthias Feurer, Frank Hutter, Michel Lang, Rafael G Mantovani, Jan N Van Rijn, and Joaquin Vanschoren. Openml benchmarking suites and the openml100. *stat*, 1050, 2017.

[10] Erik Bodin, Markus Kaiser, Ieva Kazlauskaite, Zhenwen Dai, Neill Campbell, and Carl Henrik Ek. Modulating surrogates for bayesian optimization. In *International Conference on Machine Learning*, 2020.

[11] George E. P. Box. Science and statistics. *Journal of the American Statistical Association*, 71, 1976.

[12] Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical science*, 1995.

[13] Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, 2017.

[14] Xu Chu, Ihab F Ilyas, Sanjay Krishnan, and Jiannan Wang. Data cleaning: Overview and emerging challenges. In *Proceedings of the 2016 international conference on management of data*, 2016.

[15] Abhimanyu Das and David Kempe. Approximate submodularity and its applications: Subset selection, sparse approximation and dictionary selection. *Journal of Machine Learning Research*, 19, 2018.

[16] Marc Deisenroth and Carl E Rasmussen. Pilco: A model-based and data-efficient approach to policy search. In *Proceedings of the 28th International Conference on machine learning (ICML-11)*, 2011.

[17] Sameer K Deshpande, Soumya Ghosh, Tin D Nguyen, and Tamara Broderick. Are you using test log-likelihood correctly? *arXiv preprint arXiv:2212.00219*, 2022.

[18] Ethan R Elenberg, Rajiv Khanna, Alexandros G Dimakis, and Sahand Negahban. Restricted strong convexity implies weak submodularity. *The Annals of Statistics*, 46, 2018.

[19] Jerome H Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33, 2010.

[20] Jacob Gardner, Geoff Pleiss, Kilian Q Weinberger, David Bindel, and Andrew G Wilson. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. *Advances in neural information processing systems*, 31, 2018.

[21] Roman Garnett. *Bayesian optimization*. Cambridge University Press, 2023.

[22] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 2023.

[23] Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, 2019.

[24] David Ginsbourger and Cédric Schärer. Fast calculation of gaussian process multiple-fold cross-validation residuals and their covariances. *Journal of Computational and Graphical Statistics*, 2025.

[25] Paul Goldberg, Christopher Williams, and Christopher Bishop. Regression with input-dependent noise: A gaussian process treatment. *Advances in neural information processing systems*, 10, 1997.

[26] Javier González, Zhenwen Dai, Andreas Damianou, and Neil D Lawrence. Preferential bayesian optimization. In *International Conference on Machine Learning*, 2017.

[27] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, 2018.

[28] David Harrison Jr and Daniel L Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, 5, 1978.

[29] Xun Huan, Jayanth Jagalur, and Youssef Marzouk. Optimal experimental design: Formulations and computations. *Acta Numerica*, 33, 2024.

[30] Jonas Hübotter, Bhavya Sukhija, Lenart Treven, Yarden As, and Andreas Krause. Transductive active learning: Theory and applications. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[31] Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. *Advances in neural information processing systems*, 32, 2019.

[32] Rodolphe Jenatton, Cedric Archambeau, Javier González, and Matthias Seeger. Bayesian optimization with tree-structured dependencies. In *International Conference on Machine Learning*, 2017.

[33] Donald R Jones, Matthias Schonlau, and William J Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13, 1998.

[34] Pasi Jylänki, Jarno Vanhatalo, and Aki Vehtari. Robust gaussian process regression with a student-t likelihood. *Journal of Machine Learning Research*, 12, 2011.

[35] Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information theoretic regret bounds for online nonlinear control. *Advances in Neural Information Processing Systems*, 33, 2020.

[36] Kristian Kersting, Christian Plagemann, Patrick Pfaff, and Wolfram Burgard. Most likely heteroscedastic gaussian process regression. In *Proceedings of the 24th international conference on Machine learning*, 2007.

[37] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, 2017.

[38] Suraj Kothawade, Nathan Beck, Krishnateja Killamsetty, and Rishabh Iyer. Similar: Submodular information measures based active learning in realistic scenarios. *Advances in Neural Information Processing Systems*, 34, 2021.

[39] Steven George Krantz and Harold R Parks. *The implicit function theorem: history, theory, and applications*. Springer Science & Business Media, 2002.

[40] Andreas Krause and Daniel Golovin. Submodular function maximization. *Tractability*, 3, 2014.

[41] Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9, 2008.

[42] Daniel Kühn, Philipp Probst, Janek Thomas, and Bernd Bischl. Automatic exploration of machine learning experiments on openml. *arXiv preprint arXiv:1806.10961*, 2018.

[43] Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*, 2018.

[44] Miguel Lázaro-Gredilla and Michalis K Titsias. Variational heteroscedastic gaussian process regression. In *International conference on machine learning*, 2011.

[45] Dan Levi, Liran Gispan, Niv Giladi, and Ethan Fetaya. Evaluating and calibrating uncertainty prediction in regression tasks. *Sensors*, 2022.

[46] ZZ Li, L Li, and Z Shao. Robust gaussian process regression based on iterative trimming, astron. comput., 36, 100483. *arXiv preprint arXiv:2011.11057*, 2021.

[47] David JC MacKay. Information-based objective functions for active data selection. *Neural computation*, 4, 1992.

[48] Ruben Martinez-Cantin, Michael McCourt, and Kevin Tee. Robust bayesian optimization with student-t likelihood. *arXiv preprint arXiv:1707.05729*, 2017.

[49] Jonas Močkus. On bayesian methods for seeking the extremum. In *IFIP Technical Conference on Optimization Techniques*, 1974.

[50] Kevin P Murphy. *Probabilistic machine learning: Advanced topics*. MIT press, 2023.

[51] Andrew Naish-Guzman and Sean Holden. Robust regression with twinned gaussian processes. *Advances in neural information processing systems*, 20, 2007.

[52] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[53] Diane Valérie Ouellette. Schur complements and statistics. *Linear Algebra and its Applications*, 1981.

[54] Brendan O'Donoghue, Ian Osband, Remi Munos, and Volodymyr Mnih. The uncertainty bellman equation and exploration. In *International conference on machine learning*, 2018.

[55] Chiwoo Park, David J Borth, Nicholas S Wilson, Chad N Hunter, and Fritz J Friedersdorf. Robust gaussian process regression with a bias model. *Pattern Recognition*, 124, 2022.

[56] Valerio Perrone, Rodolphe Jenatton, Matthias W Seeger, and Cédric Archambeau. Scalable hyperparameter transfer learning. *Advances in neural information processing systems*, 31, 2018.

[57] Luis Pineda, Brandon Amos, Amy Zhang, Nathan O. Lambert, and Roberto Calandra. Mbrl-lib: A modular library for model-based reinforcement learning. *Arxiv*, 2021.

[58] Rishik Ranjan, Biao Huang, and Alireza Fatehi. Robust gaussian process modeling using em algorithm. *Journal of Process Control*, 42, 2016.

[59] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2005.

[60] Jonas Rothfuss, Dominique Heyn, Andreas Krause, et al. Meta-learning reliable priors in the function space. *Advances in Neural Information Processing Systems*, 34, 2021.

[61] Felix Schur, Parnian Kassraie, Jonas Rothfuss, and Andreas Krause. Lifelong bandit optimization: no prior and no regret. In *Uncertainty in Artificial Intelligence*, 2023.

[62] Sambu Seo, Marko Wallat, Thore Graepel, and Klaus Obermayer. Gaussian process regression: Active data selection and test point rejection. In *Proceedings of the IEEE-INNS-ENNS international joint conference on neural networks. Neural computing: New challenges and perspectives for the New Millennium*, volume 3, 2000.

[63] Freddie Bickford Smith, Andreas Kirsch, Sebastian Farquhar, Yarin Gal, Adam Foster, and Tom Rainforth. Prediction-oriented bayesian active learning. In *International conference on artificial intelligence and statistics*, 2023.

[64] Freddie Bickford Smith, Jannik Kossen, Eleanor Trollope, Mark Van Der Wilk, Adam Foster, and Tom Rainforth. Rethinking aleatoric and epistemic uncertainty. In *International Conference on Machine Learning*, 2025.

[65] Jasper Snoek, Kevin Swersky, Rich Zemel, and Ryan Adams. Input warping for bayesian optimization of non-stationary functions. In *International conference on machine learning*, 2014.

[66] Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

[67] S. Surjanovic and D. Bingham. Virtual library of simulation experiments: Test functions and datasets. Retrieved January 11, 2026, from http://www.sfu.ca/~ssurjano.

[68] Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

[69] Joel A Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information theory*, 50, 2004.

[70] Chaoqi Wang, Shengyang Sun, and Roger Grosse. Beyond marginal uncertainty: How accurately can bayesian regression models estimate posterior predictive correlations? In *International Conference on Artificial Intelligence and Statistics*, 2021.

[71] Chunyi Wang and Radford M Neal. Gaussian process regression with heteroscedastic or non-gaussian residuals. *arXiv preprint arXiv:1212.6246*, 2012.

[72] Manuel Wendl, Yarden As, Manish Prajapat, Anton Pollak, Stelian Coros, and Andreas Krause. Safe exploration via policy priors. *The Fourteenth International Conference on Learning Representations*, 2026.

[73] Kai Yu, Jinbo Bi, and Volker Tresp. Active learning via transductive experimental design. In *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

A Proof for Removal Direction

Proposition 5. *Let k be a positive-definite kernel with RKHS $\mathcal{H}_k$ and feature map $\phi(x) = k(x,\cdot)$. For a dataset $\mathcal{S} = \{(x_j, y_j)\}_{j=1}^n$, define the kernel vector and targets by $k_{\mathcal{S}}(x) \coloneqq [k(x,x_1),\dots,k(x,x_n)]^\top, \quad y \coloneqq [y_1,\dots,y_n]^\top$. Fix an index i, and define the continuous removal direction $h_i \in [0,1]$ by scaling the i-th basis function $\phi(x_i)$ by $1-h_i$. Equivalently,*

$$k_{\mathcal{S}}(x;h_i) \coloneqq k_{\mathcal{S}}(x) - h_i\, k(x,x_i)e_i, \qquad y(h_i) \coloneqq y - h_i\, y_i e_i,$$

where e_i is the i-th standard basis vector. Then, for fixed aleatoric noise σ_w^2, setting $h_i = 1$ recovers the leave-one-out predictor:

$$\mu_{\mathcal{S}}(x;1) = \mu_{\mathcal{S}\setminus\{x_i,y_i\}}(x), \qquad \sigma^2_{\mathcal{S}}(x;1) = \sigma^2_{\mathcal{S}\setminus\{x_i,y_i\}}(x).$$

Proof. Write the regularized kernel matrix as

$$A(h_i) \coloneqq K(h_i) + \sigma_w^2 I, \tag{7}$$

and the corresponding predictor in kernel form as

$$\mu_{\mathcal{S}}(x;h_i) = k_{\mathcal{S}}(x;h_i)^\top A(h_i)^{-1} y(h_i), \tag{8}$$

$$\sigma^2_{\mathcal{S}}(x;h_i) = k(x,x) - k_{\mathcal{S}}(x;h_i)^\top A(h_i)^{-1} k_{\mathcal{S}}(x;h_i) + \sigma_w^2. \tag{9}$$

At $h_i = 1$, the i-th basis function is removed, so the i-th row and column of $K(1)$ vanish. After permuting the i-th coordinate to the last position, $A(1)$ has the block form

$$A(1) = \begin{bmatrix} K_{\mathcal{S}\setminus\{x_i,y_i\}} + \sigma_w^2 I & 0 \\ 0^\top & \sigma_w^2 \end{bmatrix}, \qquad A(1)^{-1} = \begin{bmatrix} (K_{\mathcal{S}\setminus\{x_i,y_i\}} + \sigma_w^2 I)^{-1} & 0 \\ 0^\top & \sigma_w^{-2} \end{bmatrix}, \tag{10}$$

$$k_{\mathcal{S}}(x;1) = \begin{bmatrix} k_{\mathcal{S}\setminus\{x_i,y_i\}}(x) \\ 0 \end{bmatrix}, \qquad y(1) = \begin{bmatrix} y_{\mathcal{S}\setminus\{x_i,y_i\}} \\ 0 \end{bmatrix}. \tag{11}$$

Substituting these expressions into the formulas above gives

$$\mu_{\mathcal{S}}(x;1) = k_{\mathcal{S}\setminus\{x_i,y_i\}}(x)^\top (K_{\mathcal{S}\setminus\{x_i,y_i\}} + \sigma_w^2 I)^{-1} y_{\mathcal{S}\setminus\{x_i,y_i\}} = \mu_{\mathcal{S}\setminus\{x_i,y_i\}}(x), \tag{12}$$

$$\sigma^2_{\mathcal{S}}(x;1) = k(x,x) - k_{\mathcal{S}\setminus\{x_i,y_i\}}(x)^\top (K_{\mathcal{S}\setminus\{x_i,y_i\}} + \sigma_w^2 I)^{-1} k_{\mathcal{S}\setminus\{x_i,y_i\}}(x) + \sigma_w^2 = \sigma^2_{\mathcal{S}\setminus\{x_i,y_i\}}(x). \tag{13}$$

To complete the proof for the output space, we examine the weight vector $v(h_i) = A(h_i)^{-1} y(h_i)$. At the limit $h_i = 1$, the target vector is $y(1) = [y_1,\dots,0,\dots,y_n]^\top$. Using the previously derived block-inverse structure of $A(1)^{-1}$:

$$v(1) = \begin{bmatrix} (K_{\mathcal{S}\setminus\{x_i,y_i\}} + \sigma_w^2 I)^{-1} & 0 \\ 0^\top & \frac{1}{\sigma_w^2} \end{bmatrix} \begin{bmatrix} y_{\mathcal{S}\setminus\{x_i,y_i\}} \\ 0 \end{bmatrix} = \begin{bmatrix} (K_{\mathcal{S}\setminus\{x_i,y_i\}} + \sigma_w^2 I)^{-1} y_{\mathcal{S}\setminus\{x_i,y_i\}} \\ 0 \end{bmatrix} \tag{14}$$

Let $v_{\mathcal{S}\setminus\{x_i,y_i\}}$ be the weights obtained by training a GP on the subset $\mathcal{S}\setminus\{x_i,y_i\}$. It shows that:

$$v(1) = \begin{cases} (v_{\mathcal{S}\setminus\{x_i,y_i\}})_j & j \neq i \\ 0 & j = i \end{cases} \tag{15}$$

The predictive mean at x under the removal direction is $\mu_{\mathcal{S}}(x,1) = k(x,\mathcal{S},1)^\top v(1)$. Substituting the zeroed i-th components:

$$\mu_{\mathcal{S}}(x,1) = \sum_{j\neq i} k(x,x_j) v_j(1) + k(x,x_i)\cdot 0 = \sum_{j\neq i} k(x,x_j)(v_{\mathcal{S}\setminus\{x_i,y_i\}})_j \tag{16}$$

This is identically the LOO predictive mean $\mu_{\mathcal{S}\setminus\{x_i,y_i\}}(x)$. Since both the mean and the variance recover the LOO statistics at $h_i = 1$ under fixed σ_w^2, the proposition holds. □

B Proofs for Epistemic Influence Forgetting

Proposition 1. *Suppose we have a single test point $\mathcal{A} = \{x\}$ and the corresponding GP prediction with mean $\mu_{\mathcal{S}}(x)$ and variance $\sigma^2_{\mathcal{S}}(x)$, for a given set of samples $\mathcal{S} = \{x_i, y_i\}_{i=1}^n$. The epistemic influence $u_i(x)$ of an individual sample $x_i \in \mathcal{S}$ on the predictive uncertainty is defined as the partial derivative of $\sigma^2_{\mathcal{S}}(x)$ w.r.t. the removal in h_i for sample $x_i \in \mathcal{S}$ under fixed aleatoric noise $\sigma^2_w(\mathcal{S})$*

$$u_i(x) = 2k_{x,x_i}(A^{-1}k_{\mathcal{S},x})_i - 2(A^{-1}k_{\mathcal{S},x})_i(k_{x_i,\mathcal{S}}A^{-1}k_{\mathcal{S},x}),$$
$$\textit{where } k_{x,\mathcal{S}} = k(x,\mathcal{S}), \textit{ and } A = K + \sigma^2_w(\mathcal{S})I, \textit{ with } K = k(\mathcal{S},\mathcal{S}).$$

Proof. Let the predictive uncertainty of the GP for a set of samples $\mathcal{S}$ be defined as in Equation (2), and we consider the aleatoric noise estimate $\sigma^2_w(\mathcal{S})$ to be fixed.

$$\sigma^2_{\mathcal{S}}(x) = k(x,x) - k_{x,\mathcal{S}}(K + \sigma^2_w(\mathcal{S})I)^{-1}k_{\mathcal{S},x} + \sigma^2_w(\mathcal{S})I \tag{17}$$

Furthermore, we consider the RKHS and output removal directions in h_i introduced in Proposition 5:

$$\sigma^2_{\mathcal{S}}(x) = k(x,x) - k(x,\mathcal{S},h_i)(K(h_i) + \sigma^2_w(\mathcal{S})I)^{-1}k(\mathcal{S},x,h_i) + \sigma^2_w(\mathcal{S})I, \tag{18}$$

where we define the Gram matrix under the removal is defined using the as

$$K(h_i) = k(\mathcal{S},\mathcal{S},h_i) = [(1-h_j)K_{j,k}(1-h_k)]_{j,k} = \begin{cases} (1-h_i)k(x_j,x_k) & \text{if } j = i \vee k = i \\ (1-h_i)^2k(x_j,x_k) & \text{if } i = j = k \\ k(x_j,x_k) & \text{otherwise.} \end{cases} \tag{19}$$

Hence, after deriving the continuous surrogate, we differentiate w.r.t. h_i. Thereby, we observe that all noise terms $\sigma^2_w(\mathcal{S})$ drop out, as we assume the aleatoric component to be fixed:

$$\begin{aligned}
\frac{\partial\sigma^2_{\mathcal{S}}(x)}{\partial h_i} =& \frac{\partial}{\partial h_i}k(x,x) - \frac{\partial}{\partial h_i}k(x,\mathcal{S},h_i)(K(h_i) + \sigma^2_w(\mathcal{S})I)^{-1}k(\mathcal{S},x,h_i) + \frac{\partial}{\partial h_i}\sigma^2_w(\mathcal{S})I && (20)\\
=& -\frac{\partial}{\partial h_i}k(x,\mathcal{S},h_i)(K(h_i) + \sigma^2_w(\mathcal{S})I)^{-1}k(\mathcal{S},x,h_i) && (21)\\
=& -\frac{\partial k(x,\mathcal{S},h_i)}{\partial h_i}(K(h_i) + \sigma^2_w(\mathcal{S})I)^{-1}k(\mathcal{S},x,h_i) && (22)\\
& - k(x,\mathcal{S},h_i)\frac{\partial(K(h_i) + \sigma^2_w(\mathcal{S})I)^{-1}}{\partial h_i}k(\mathcal{S},x,h_i)\\
& - k(x,\mathcal{S},h_i)(K(h_i) + \sigma^2_w(\mathcal{S})I)^{-1}\frac{\partial k(\mathcal{S},x,h_i)}{\partial h_i}\\
=& -2\frac{\partial k(x,\mathcal{S},h_i)}{\partial h_i}(K(h_i) + \sigma^2_w(\mathcal{S})I)^{-1}k(\mathcal{S},x,h_i) && (23)\\
& + k(x,\mathcal{S},h_i)A(h_i)^{-1}\frac{\partial K(h_i)}{\partial h_i}A(h_i)^{-1}k(\mathcal{S},x,h_i)\\
=& 2k(x,x_i)e_i^\top A(h_i)^{-1}k(\mathcal{S},x,h_i) && (24)\\
& + k(x,\mathcal{S},h_i)A(h_i)^{-1}(e_ik(x_i,\mathcal{S}) + k(\mathcal{S},x_i)e_i^\top)A(h_i)^{-1}k(\mathcal{S},x,h_i).
\end{aligned}$$

We can further simplify this expression by differentiating the Gram matrix $K(h_i)$ of Equation (19) and introducing the short notation $\alpha = A(h_i)^{-1}k(\mathcal{S},x,h_i)$. We hence obtain

$$\begin{aligned}
\frac{\partial\sigma^2_{\mathcal{S}}(x)}{\partial h_i} =& 2k(x,x_i)\alpha_i + \alpha^\top\left[-e_ik(x_i,\mathcal{S},h_i) - k(\mathcal{S},x_i,h_i)e_i^\top\right]\alpha && (25)\\
=& 2k(x,x_i)\alpha_i - \left[(\alpha^\top e_i)(k(x_i,\mathcal{S},h)\alpha) + (\alpha^\top k(\mathcal{S},x_i,h))(e_i^\top\alpha)\right] && (26)\\
=& 2k(x,x_i)\alpha_i - 2\alpha_i(k(x_i,\mathcal{S},h)\alpha) && (27)
\end{aligned}$$

Since we evaluate the partial derivative at $h_i = 0$, the term simplifies to

$$\begin{aligned}
\left.\frac{\partial\sigma^2_{\mathcal{S}}(x)}{\partial h_i}\right|_{h_i=0} =& 2k(x,x_i)\alpha_i - 2\alpha_i(k(x_i,\mathcal{S})\alpha) && (28)\\
=& 2k(x,x_i)(A^{-1}k(\mathcal{S},x))_i - 2(A^{-1}k(\mathcal{S},x))_i(k(x_i,\mathcal{S})A^{-1}k(\mathcal{S},x)), && (29)
\end{aligned}$$

which completes the proof, when writing the kernel vectors in index notation. □

Proposition 2. *Suppose the uncertainty difference at $\mathcal{A} = \{x\}$, when removing sample i is given by*

$$\Delta_i(x) = \sigma^2_{\mathcal{S}\setminus\{x_i,y_i\}}(x) - \sigma^2_{\mathcal{S}}(x) = \frac{(A^{-1}k_{\mathcal{S},x})_i^2}{A^{-1}_{i,i}}.$$

Then, our uncertainty measure u in Proposition 1 captures the half directional derivative along the kernel direction $k_{x,\mathcal{S}}$, and k_{x,x_i} for an individual component of u:

$$k_{x,x_i}\frac{\partial \Delta_i}{\partial k_{x,x_i}} = 2k_{x,x_i}(A^{-1}k_{\mathcal{S},x})_i - 2(A^{-1}k_{\mathcal{S},x})_i(k_{x_i,\mathcal{S}}A^{-1}k_{\mathcal{S},x}) = u_i(x).$$

Proof. Let $\mathcal{S} = \{x_1, \dots, x_n\}$ be the training set, fix an index i and a single test point x_t. Recall

$$A = K(\mathcal{S},\mathcal{S}) + \sigma_w^2(\mathcal{S})I, \qquad k := k_{\mathcal{S},x_t} = k(\mathcal{S}, x_t) \in \mathbb{R}^n, \tag{30}$$

and partition A and k conformably isolating index i:

$$A = \begin{bmatrix} A_{i,i} & A_{i,-i} \\ A_{-i,i} & A_{-i,-i} \end{bmatrix}, \qquad k = \begin{bmatrix} k_i \\ k_{-i} \end{bmatrix}. \tag{31}$$

Define the Schur complement [53]

$$S_{i,i} := A_{i,i} - A_{i,-i}A^{-1}_{-i,-i}A_{-i,i}. \tag{32}$$

Epistemic uncertainty difference for leave one-out: Using the block inverse identities in [24, Prop. 2], the ii-block of A^{-1} equals $A^{-1}_{i,i} = S^{-1}_{i,i}$ and the off-blocks satisfy

$$A^{-1}_{i,-i} = -S^{-1}_{i,i}A_{i,-i}A^{-1}_{-i,-i}, \qquad A^{-1}_{-i,i} = -A^{-1}_{-i,-i}A_{-i,i}S^{-1}_{i,i}, \tag{33}$$

$$A^{-1}_{-i,-i} = A^{-1}_{-i,-i} + A^{-1}_{-i,-i}A_{-i,i}S^{-1}_{i,i}A_{i,-i}A^{-1}_{-i,-i} \tag{34}$$

The GP predictive variance with the full set $\mathcal{S}$ is

$$\sigma^2_{\mathcal{S}}(x_t) = k(x_t, x_t) - k^\top A^{-1}k, \tag{35}$$

and the variance when training on $\mathcal{S} \setminus \{x_i, y_i\}$ (i.e. removing index i) is

$$\sigma^2_{\mathcal{S}\setminus\{x_i,y_i\}}(x_t) = k(x_t, x_t) - k_{-i}^\top A^{-1}_{-i,-i}k_{-i}. \tag{36}$$

Therefore the variance drop due to sample i is

$$\Delta_i(x_t) = \sigma^2_{\mathcal{S}\setminus\{x_i,y_i\}}(x_t) - \sigma^2_{\mathcal{S}}(x_t) = k^\top A^{-1}k - k_{-i}^\top A^{-1}_{-i,-i}k_{-i}. \tag{37}$$

Expanding $k^\top A^{-1}k$ using the block-inverse factorization yields the Schur-form:

$$\begin{aligned}
k^\top A^{-1}k &= \begin{bmatrix} k_i & k_{-i}^\top \end{bmatrix} \begin{bmatrix} S^{-1}_{i,i} & A^{-1}_{i,-i} \\ A^{-1}_{-i,i} & A^{-1}_{-i,-i} \end{bmatrix} \begin{bmatrix} k_i \\ k_{-i} \end{bmatrix} && (38)\\
&= S^{-1}_{i,i}k_i^2 + 2k_iA^{-1}_{i,-i}k_{-i} + k_{-i}^\top P^{-1}_{-i,-i}k_{-i} && (39)\\
&= S^{-1}_{i,i}k_i^2 + 2k_i(-S^{-1}_{i,i}A_{i,-i}A^{-1}_{-i,-i})k_{-i} && (40)\\
&\quad + k_{-i}^\top(A^{-1}_{-i,-i} + A^{-1}_{-i,-i}A_{-i,i}S^{-1}_{i,i}A_{i,-i}A^{-1}_{-i,-i})k_{-i} \\
&= S^{-1}_{i,i}k_i^2 - 2k_iS^{-1}_{i,i}A_{i,-i}A^{-1}_{-i,-i}k_{-i} && (41)\\
&\quad + k_{-i}^\top A^{-1}_{-i,-i}k_{-i} + k_{-i}^\top A^{-1}_{-i,-i}A_{-i,i}S^{-1}_{i,i}A_{i,-i}A^{-1}_{-i,-i}k_{-i} \\
&= k_{-i}^\top A^{-1}_{-i,-i}k_{-i} + S^{-1}_{i,i}(k_i^2 - 2k_iA_{i,-i}A^{-1}_{-i,-i}k_{-i} && (42)\\
&\quad + (A_{i,-i}A^{-1}_{-i,-i}k_{-i})^\top(A_{i,-i}A^{-1}_{-i,-i}k_{-i})) \\
&= k_{-i}^\top A^{-1}_{-i,-i}k_{-i} + S^{-1}_{i,i}(k_i - A_{i,-i}A^{-1}_{-i,-i}k_{-i})^2. && (43)
\end{aligned}$$

Further, the i-th component of $A^{-1}k$ is given by

$$(A^{-1}k)_i = A^{-1}_{ii}k_i + A^{-1}_{i,-i}\,k_{-i} = S^{-1}_{i,i}\big(k_i - A_{i,-i}A^{-1}_{-i,-i}k_{-i}\big). \tag{44}$$

Finally, we express the variance drop using Equation (43):

$$\Delta_i(x_t) = S^{-1}_{i,i}\big(k_i - A_{i,-i}A^{-1}_{-i,-i}k_{-i}\big)^2 \overset{\text{(i)}}{=} S^{-1}_{i,i}\big(S_{i,i}(A^{-1}k)_i\big)^2 \overset{\text{(ii)}}{=} \frac{(A^{-1}k_{\mathcal{S},x_t})_i^2}{A^{-1}_{i,i}}. \tag{45}$$

We obtain step (i) by substituting Equation (44) and (ii) from $S^{-1}_{i,i} = A^{-1}_{i,i}$.

Directional-derivative: Let us abbreviate $\alpha = A^{-1}k$ and $v = A^{-1}_{ii} = \mathbf{e}_i^\top A^{-1}\mathbf{e}_i$. We compute the exact directional derivative of $\Delta_i = \alpha_i^2/v$ with respect to a change in the i-th kernel direction. By following the quotient rule:

$$\partial_{k_i}\Delta_i = \frac{2\alpha_i}{v}\partial_{k_i}\alpha_i - \frac{\alpha_i^2}{v^2}\partial_{k_i}v. \tag{46}$$

Using the identity $\partial_{k_i}(A^{-1}) = -A^{-1}(\partial_{k_i}A)A^{-1}$, the differentials for $\alpha_i = \mathbf{e}_i^\top A^{-1}k$ and v are:

$$\partial_{k_i}\alpha_i = \mathbf{e}_i^\top(-A^{-1}(\partial_{k_i}A)A^{-1}k + A^{-1}\partial_{k_i}k) = -\mathbf{e}_i^\top A^{-1}(\partial_{k_i}A)\alpha + v\partial_{k_i}k, \tag{47}$$

$$\partial_{k_i}v = \mathbf{e}_i^\top(-A^{-1}(\partial_{k_i}A)A^{-1})\mathbf{e}_i = -v\mathbf{e}_i^\top(\partial_{k_i}A)A^{-1}_{:,i} - (A^{-1}_{i,:}\partial_{k_i}A)\mathbf{e}_i v. \tag{48}$$

For a directional change along the i-th component, the variations are $k_i\partial_{k_i}k = \mathbf{e}_i k_i$ and $k_i\partial_{k_i}A = \mathbf{e}_i K_{i,\mathcal{S}} + K_{\mathcal{S},i}\mathbf{e}_i^\top$. Substituting these into the differentials:

$$\partial_{k_i}\alpha_i = -(v(K_{i,\mathcal{S}}\alpha) + (A^{-1}_{i,\mathcal{S}}K_{\mathcal{S},i})\alpha_i) + vk_i, \tag{49}$$

$$\partial_{k_i}v = -(v(K_{i,\mathcal{S}}A^{-1}_{:,i}) + (A^{-1}_{i,\mathcal{S}}K_{\mathcal{S},i})v) = -2v(K_{i,\mathcal{S}}A^{-1}_{:,i}). \tag{50}$$

Now, substitute $\partial_{k_i}\alpha_i$ and $\partial_{k_i}v$ back into the total differential $d\Delta_i$:

$$\partial_{k_i}\Delta_i = \frac{2\alpha_i}{v}\left(-v(K_{i,\mathcal{S}}\alpha) - \alpha_i(K_{i,\mathcal{S}}A^{-1}_{:,i}) + vk_i\right) - \frac{\alpha_i^2}{v^2}\left(-2v(K_{i,\mathcal{S}}A^{-1}_{:,i})\right) \tag{51}$$

$$= -2\alpha_i(K_{i,\mathcal{S}}\alpha) - \frac{2\alpha_i^2}{v}(K_{i,\mathcal{S}}A^{-1}_{:,i}) + 2\alpha_i k_i + \frac{2\alpha_i^2}{v}(K_{i,\mathcal{S}}A^{-1}_{:,i}) \tag{52}$$

$$= 2\alpha_i k_i - 2\alpha_i(K_{i,\mathcal{S}}\alpha). \tag{53}$$

Hence, we recover the epistemic influence $u_i(x)$:

$$u_i(x) = 2k_{x,x_i}(A^{-1}k)_i - 2(A^{-1}k)_i(K_{i,\mathcal{S}}A^{-1}k). \tag{54}$$

□

C Proofs for Aleatoric Influence

Proposition 3. *Let $A(\sigma_w(\mathcal{S})) = K(\mathcal{S},\mathcal{S}) + \sigma_w^2(\mathcal{S})I$ and suppose that $\sigma_w^2(\mathcal{S})$ is the optimal noise variance of the log MAP II (Equation (1)) $\frac{\partial}{\partial\sigma_w^2}\mathcal{L}(\sigma_w^2,\mathcal{S},a,b) = \mathcal{D}(\sigma_w^2,\mathcal{S},a,b) = 0$ for the samples in $\mathcal{S}$. Then we determine the derivative of the optimal noise floor $\sigma_w^2(\mathcal{S})$ into the removal direction h_i using implicit differentiation $\frac{\partial\sigma_w^2}{\partial h_i} = -\frac{\partial}{\partial h_i}\mathcal{D}(\sigma_w^2,\mathcal{S},a,b)/\frac{\partial}{\partial\sigma_w^2}\mathcal{D}(\sigma_w^2,\mathcal{S},a,b)$, which results in*

$$\frac{\partial\sigma_w^2(\mathcal{S})}{\partial h_i} = \frac{-(A^{-1}y)_i(k_{x_i,\mathcal{S}}(A^{-2}y)) - (A^{-2}y)_i(k_{x_i,\mathcal{S}}(A^{-1}y)) + k_{x_i,\mathcal{S}}(A^{-2})_{\cdot,i}}{-y^\top A^{-3}y + \frac{1}{2}\operatorname{tr}(A^{-2}) - \frac{a-1}{(\sigma_w^2)^2}}.$$

Proof. We abbreviate $A = A(\sigma_w^2(\mathcal{S}))$ and $\sigma_w^2 = \sigma_w^2(\mathcal{S})$ for shorter notation. By definition of the maximal log marginal likelihood the pair $(\sigma_w^2(\mathcal{S}), y)$ satisfies

$$\mathcal{D}(\sigma_w^2,\mathcal{S},a,b) = \frac{\partial(\log p(y|\mathcal{S},\sigma_w^2) + \log p(\sigma_w^2(\mathcal{S})))}{\partial\sigma_w^2} = 0 \tag{55}$$

$$= \tfrac{1}{2}y^\top A^{-1}\frac{\partial A^{-1}}{\sigma_w^2}A^{-1}y - \frac{1}{2}\operatorname{tr}\left(A^{-1}\frac{\partial A}{\partial\sigma_w^2}\right) + \frac{\partial p(\sigma_w^2)}{\partial\sigma_w^2} = 0 \tag{56}$$

$$= \frac{1}{2}(y^\top A^{-2}y - \operatorname{tr}(A^{-1})) + \frac{a-1}{\sigma_w^2} - b = 0. \tag{57}$$

We apply the implicit function theorem [39, Theorem 1.3.1]: differentiate $\mathcal{D}(\sigma_w^2,\mathcal{S},a,b) = 0$ with respect to y_i while treating $\sigma_w^2 = \sigma_w^2(y)$ as an implicit function of y. We obtain

$$\frac{\partial\mathcal{D}(\sigma_w^2,\mathcal{S},a,b)}{\partial\sigma_w^2}\frac{\partial\sigma_w^2}{\partial h_i} + \frac{\partial\mathcal{D}(\sigma_w^2,\mathcal{S},a,b)}{\partial h_i} = 0 \qquad \Longrightarrow \qquad \frac{\partial\sigma_w^2}{\partial h_i} = -\frac{\frac{\partial\mathcal{D}(\sigma_w^2,\mathcal{S},a,b)}{\partial h_i}}{\frac{\partial\mathcal{D}(\sigma_w^2,\mathcal{S},a,b)}{\partial\sigma_w^2}}. \tag{58}$$

Since $\mathcal{D}(\sigma_w^2, \mathcal{S}, a, b) = y^\top A(h_i)^{-2} y - \operatorname{tr}(A(h_i)^{-1}) + \frac{a-1}{\sigma_w^2} - b$, we get for the partial derivative of $\mathcal{D}(\sigma_w^2, \mathcal{S}, a, b)$ with respect to h_i, that the enumerator is given by

$$\begin{aligned}
\frac{\partial \mathcal{D}(\sigma_w^2, \mathcal{S}, a, b)}{\partial h_i} =& \frac{\partial}{\partial h_i}((y^\top A(h_i)^{-2} y) - \operatorname{tr}(A(h_i)^{-1})) && (59)\\
=& y^\top \frac{\partial A(h_i)^{-2}}{\partial h_i} y - \frac{\partial}{\partial h_i} \operatorname{tr}(A(h_i)^{-1}) && (60)\\
=& 2((A^{-1}y)_i (k_{x_i,\mathcal{S}}(A^{-2}y)) + (A^{-2}y)_i (k_{x_i,\mathcal{S}}(A^{-1}y))) - 2k_{x_i,\mathcal{S}}(A^{-2})_{\cdot,i}. && (61)
\end{aligned}$$

Hence, we can use the results of to compute the derivative of $\mathcal{D}(\sigma_w^2, \mathcal{S}, a, b)$ with respect to σ_w^2 as:

$$\begin{aligned}
\frac{\partial \mathcal{D}(\sigma_w^2, \mathcal{S}, a, b)}{\partial \sigma_w^2} &= \frac{\partial}{\partial \sigma_w^2} \left((y^\top A(\sigma_w^2)^{-2} y) - \operatorname{tr}(A(\sigma_w^2)^{-1}) + \frac{2(a-1)}{\sigma_w^2} \right) && (62)\\
&= y^\top \big(-2A^{-3} \big) y - \operatorname{tr}\big(-A^{-2}\big) + \frac{-2(a-1)}{(\sigma_w^2)^2} && (63)\\
&= -2y^\top A^{-3} y + \operatorname{tr}(A^{-2}) - \frac{2(a-1)}{(\sigma_w^2)^2} && (64)
\end{aligned}$$

Finally, we combine the derived terms in the implicit differentiation, evaluated at $h_i = 0$

$$\begin{aligned}
\frac{\partial \sigma_w^2}{\partial h_i} &= -\frac{2((A^{-1}y)_i (k_{x_i,\mathcal{S}}(A^{-2}y)) + (A^{-2}y)_i (k_{x_i,\mathcal{S}}(A^{-1}y))) - 2k_{x_i,\mathcal{S}}(A^{-2})_{\cdot,i}}{-2y^\top A^{-3} y + \operatorname{tr}(A^{-2}) - \frac{2(a-1)}{(\sigma_w^2)^2}} && (65)\\
&= \frac{-(A^{-1}y)_i (k_{x_i,\mathcal{S}}(A^{-2}y)) - (A^{-2}y)_i (k_{x_i,\mathcal{S}}(A^{-1}y)) + k_{x_i,\mathcal{S}}(A^{-2})_{\cdot,i}}{-y^\top A^{-3} y + \frac{1}{2}\operatorname{tr}(A^{-2}) - \frac{a-1}{(\sigma_w^2)^2}}. && (66)
\end{aligned}$$

This completes the proof for the derivative of the aleatroric MAP w.r.t. the removal direction h_i. □

Proposition 4. *Suppose we have a single test point $\mathcal{A} = \{x\}$ and the corresponding GP prediction with mean $\mu_{\mathcal{S}}(x)$ and variance $\sigma_{\mathcal{S}}^2(x)$, for a given set of samples $\mathcal{S} = \{x_i, y_i\}_{i=1}^n$. The aleatoric influence $v_i(x)$ of an individual sample $x_i \in \mathcal{S}$ on the predictive uncertainty is defined as the partial derivative of all aleatoric noise terms $\sigma_w^2(\mathcal{S})$ in $\sigma_{\mathcal{S}}^2(x)$ w.r.t. the removal in h_i for sample $x_i \in \mathcal{S}$.*

$$v_i(x) = (1 + k_{x,\mathcal{S}} A^{-2} k_{\mathcal{S},x}) \frac{\partial \sigma_w^2(\mathcal{S})}{\partial h_i},$$

where the partial derivative of the noise parameter is taken from Proposition 3.

Proof. Let the predictive uncertainty of the GP for a set of samples $\mathcal{S}$ be defined as in Equation (2), then we compute the derivative with respect to all aleatoric noise terms of $\sigma_{\mathcal{S}}^2$, assuming that the kernel contributions and Gram matrix are fixed

$$\sigma_{\mathcal{S}}^2(x) = k(x,x) - k_{x,\mathcal{S}}(K + \sigma_w^2(\mathcal{S})I)^{-1} k_{\mathcal{S},x} + \sigma_w^2(\mathcal{S})I. \tag{67}$$

We further consider the RKHS and output space direction h_i and differentiate all aleatoric terms, assuming that the kernel and Gram matrix are fixed.

$$\begin{aligned}
\frac{\partial \sigma_{\mathcal{S}}^2(x)}{\partial h_i} &= \frac{\partial}{\partial h_i} k_{x,\mathcal{S}}(K + \sigma_w^2(\mathcal{S})I)^{-1} k_{\mathcal{S},x} + \frac{\partial}{\partial h_i} \sigma_w^2(\mathcal{S}) && (68)\\
&= k_{x,\mathcal{S}} A^{-1} \frac{\partial \sigma_w^2(\mathcal{S}) I}{\partial h_i} A^{-1} k_{\mathcal{S},x} + \frac{\partial \sigma_w^2(\mathcal{S})}{\partial h_i} && (69)\\
&= (1 + k_{x,\mathcal{S}} A^{-2} k_{\mathcal{S},x}) \frac{\partial \sigma_w^2(\mathcal{S})}{\partial h_i}, && (70)
\end{aligned}$$

which completes the proof for the aleatoric influence measure $v_i(x)$. □

Algorithm 2 Forgetful BO

Require: Objective f, domain $\mathcal{X}$, acquisition function α, initial dataset size n_0, budget T
  Initialize dataset $\mathcal{S}_0 = \{(x_i, y_i)\}_{i=1}^{n_0}$ with $y_i = f(x_i)$
  Fit initial GP model $(\mu_{\mathcal{S}_0}(\cdot), \sigma_{\mathcal{S}_0}(\cdot))$ on $\mathcal{S}_0$
  Set $x^* = \arg\max_{x_i \in \mathcal{S}_0} y_i$, $y^* = \max_{x_i \in \mathcal{S}_0} y_i$
  for $t = 1, \ldots, T$ **do**
    Optimize acquisition function: $x_t = \arg\max_{x \in \mathcal{X}} \alpha(x|(\mu_{\mathcal{S}'_{t-1}}(\cdot), \sigma_{\mathcal{S}'_{t-1}}(\cdot)))$
    Evaluate objective: $y_t = f(x_t)$ and update $\mathcal{S}_t = \mathcal{S}_{t-1} \cup \{(x_t, y_t)\}$
    $(x^*, y^*) \leftarrow \max((x_t, y_t), (x^*, y^*)$
    Initialize $h = 0$, $\mathcal{S}_t^{(0)} = \mathcal{S}_t$ and fit noise MAP II
    Determine $\mathcal{A}_t^{(0)}$ and compute $u^{(0)}(\mathcal{A}_t^{(0)})$, $v^{(0)}(\mathcal{A}_t^{(0)})$ ▷ Equation (4) and propositions 1 and 4
    while $\min(u^{(h)}(\mathcal{A}_t^{(h)}) + v^{(h)}(\mathcal{A}_t^{(h)})) \leq 0$ **do**
      Determine worst sample: $i = \arg\min_j (u_j^{(h)}(\mathcal{A}_t^{(h)}) + v_j^{(h)}(\mathcal{A}_t^{(h)}))$
      Update $\mathcal{S}_t^{(h+1)} = \mathcal{S}_t^{(h)} \setminus x_i$ and refit GP noise MAP II
      Update $h \leftarrow h + 1$, $\mathcal{A}_t^{(h)}$, $u^{(h)}(\mathcal{A}_t^{(h)})$, $v^{(h)}(\mathcal{A}_t^{(h)})$ ▷ Equation (4) and propositions 1 and 4
    end while
    Set $\mathcal{S}'_t = \mathcal{S}_t^{(h)}$
  end for
  return x^*, y^*

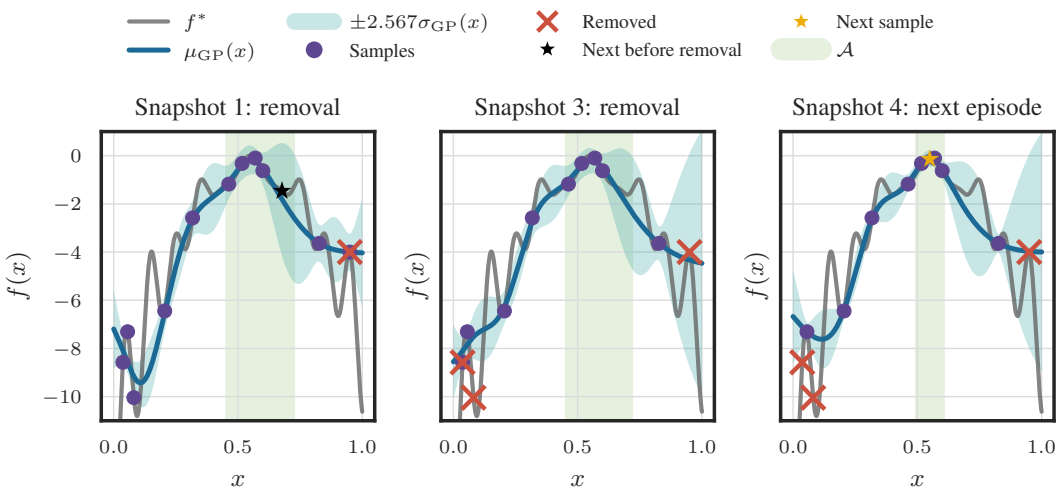


Figure 6: Forgetful BO: Visualization of sequential removal mechanism integrated in BO pipeline.

D Forgetful Bayesian Optimization

Starting from an initial sample set $\mathcal{S}_0$ of size n_0, a Gaussian process model is fit to $\mathcal{S}_0$ and iteratively updated as new evaluations are acquired by maximizing an acquisition function α. Before each function evaluation, except for the first, the algorithm identifies a region of interest $\mathcal{A}$ using Equation (4) and assesses the contribution of each data point in $\mathcal{S}$ to the predictive uncertainty within this region $\mathcal{A}$. Samples with negative influence are removed iteratively as in Algorithm 1, and the GP noise MLE estimate is refit on the reduced dataset $\mathcal{S}'$ while recursively updating the region of interest according to Equation (4). This forgetting step continues until all remaining samples provide a positive contribution to uncertainty reduction in the target region $\mathcal{A}$. Using this improved model, we next maximize the acquisition function α and determine the next probing location x_t, which is evaluated and added to the overall data set $\mathcal{S}$. From this data-set, we start the next iteration by fitting the GP with the updated $\mathcal{S}$ and repeat this process until reaching the evaluation budget T.

In Figure 6, we visualize the application of our iterative removal scheme F2I in our Forgetful BO pipeline. In *Snapshot 1*, we visualize the default predictive distribution based on all samples and indicate with the black star the next chosen query point according to the GP-UCB acquisition function. The first red cross in *Snapshot 1* indicates that Forgetful BO is removing the right-most sample. As

we can see from *Snapshot 3*, this removal causes the uncertainty bands to inflate at the location of the removal, which are outside of the previous target area $\mathcal{A}$, for which we have optimized the samples. Now it is important to consider the intersection rule of the target region update in Equation (4), to not include previously excluded regions in the target region. Further, we see in *Snapshot 3* that the detrimental data structure, composed of the three data points in the bottom left corner, is removed, resulting in the improved posterior predictive distribution visualized in *Snapshot 4*. The recursive target region in Equation (4) has substantially decreased the size of the target region further. Within the target region, we now apply the GP-UCB acquisition function to query the next point at the location indicated with the yellow star, which coincides with the global optimum of our test function. This demonstrates visually how our removal scheme in Forgetful BO improves the predictive distribution of the model, and thereby results in better acquisition.

E Forgetful Reinforcement Learning

Algorithm 3 Forgetful RL

Require: Initial real experience $\mathcal{S}_0 = \{(x_i, a_i, x_{i+1}, r_i)\}_{i=1}^{n_0}$, GP prior, SAC buffer $\mathcal{D}_{\text{SAC}} \leftarrow \{\}$
  Fit initial GP model $(\mu_{\mathcal{S}_0}(\cdot), \sigma_{\mathcal{S}_0}(\cdot))$ on $\mathcal{S}_0$ (type II MAP)
  Initialize policy π_0, value $Q^{\pi_0}(x, a)$
  for $n = 1, \ldots, N$ **do**
    Fit GP on current dataset $\mathcal{S}_{n-1}$
    Compute target region $\mathcal{A}_n$ and $u(\mathcal{A}_n), v(\mathcal{A}_n)$ ▷ Equation (4) and propositions 1 and 4
    if there exist samples $j \in \mathcal{A}_n$ with $u_j + v_j \leq 0$ **then**
      $R \leftarrow \{j \in \mathcal{S}_{n-1} \mid u_j + v_j \leq 0\}$ ▷ indices of negatively influential samples
      Remove samples: $\mathcal{S}_{n-1} \leftarrow \mathcal{S}_{n-1} \setminus \{(x_j, a_j, x_{j+1}, r_j)\}_{j \in R}$
      Refit GP on the pruned dataset $\mathcal{S}_{n-1}$ to obtain $(\mu_{\mathcal{S}_{n-1}}(\cdot), \sigma_{\mathcal{S}_{n-1}}(\cdot))$
    end if

    for $m = 1, \ldots, M$ **do** ▷ Generate M synthetic rollouts of length H
      Sample start state $\tilde{x}_0 \sim \rho_0$
      for $t = 0, \ldots, H-1$ **do**
        Sample action $\tilde{a}_t \sim \pi_{n-1}(\cdot \mid \tilde{x}_t)$
        Predict via GP: $\tilde{x}_{t+1} \sim p_{\text{GP}}(\cdot \mid \tilde{x}_t, \tilde{a}_t), \quad \tilde{r}_t \leftarrow r_{\text{GP}}(\tilde{x}_t, \tilde{a}_t)$
        Store synthetic transition $(\tilde{x}_t, \tilde{a}_t, \tilde{x}_{t+1}, \tilde{r}_t)$ in $\mathcal{D}_{\text{SAC}}$
      end for
    end for

    Learn $\pi_n = \arg\max_{\pi \in \Pi} \mathbb{E}_{\pi, p_{\text{GP}}} \left[\sum_{h=0}^{H-1} r(s_t, a_t) \right]$ with SAC on $\mathcal{D}_{\text{SAC}}$.

    for $t = 0, \ldots, H-1$ **do**
      Execute action $a_t \sim \pi_n(s_t)$ in the environment
      Append collected transition to the real buffer: $\mathcal{S}_n \leftarrow \mathcal{S}_{n-1} \cup \{s_t, a_t, s_{t+1}, r_t\}$
    end for
  end for

Our MBRL implementation builds upon Model-Based Policy Optimization (MBPO) [31] and employs GPs to model both the dynamics and reward function. In each episode n, the model is trained on the set of real-word transitions stored in the replay buffer $\mathcal{S}_n$. The learned model is then used to simulate trajectories for improving the current policy π_n. Policy optimization is performed using a Soft Actor-Critic (SAC) [27] agent with replay buffer $\mathcal{D}_{SAC}$ and corresponding state-action value function $Q^{\pi_n}(x, a)$. We in practice approximate the upper and lower confidence bound values in Equation (6) using the respective highest and lowest ensemble member of value functions. For a more principled approach to obtain these optimistic and pessimistic Q-functions, one can implement an uncertainty Bellman equation [54] or an approximation by for example using an optimistic and pessimistic uncertainty bonus, based on the model uncertainty [72].

To enhance model reliability, we introduce an additional removal step prior to generating the synthetic rollouts. Specifically, we determine the target region for potential improvements in Equation (6). For this target region, we next determine the influence measures u and v for all samples in $\mathcal{S}_n$ and remove all samples with negative total influence. After pruning these samples, we refit the GP by recomputing

		Homoscedastic GP	Target Sampling	Relevance Pursuit	Warped GP	Heteroscedastic GP	F2I
Dow Jones	MSE	0.049 ± 0.023	0.057 ± 0.048	0.048 ± 0.029	0.085 ± 0.105	0.200 ± 0.129	$\mathbf{0.048} \pm 0.026$
	MAE	0.164 ± 0.033	0.165 ± 0.048	0.164 ± 0.037	0.202 ± 0.095	0.336 ± 0.127	$\mathbf{0.160} \pm 0.035$
	NLL	0.267 ± 0.460	0.178 ± 0.807	$\mathbf{-0.120} \pm 0.214$	0.192 ± 0.269	0.566 ± 0.340	0.073 ± 0.448
	CE	0.065 ± 0.059	0.011 ± 0.011	$\mathbf{0.006} \pm 0.005$	0.036 ± 0.022	0.016 ± 0.016	0.009 ± 0.008
	ECE	0.035 ± 0.028	0.009 ± 0.004	0.009 ± 0.003	0.023 ± 0.010	0.023 ± 0.010	$\mathbf{0.007} \pm 0.003$
	Sharpness	0.511 ± 0.301	0.189 ± 0.063	0.223 ± 0.040	0.420 ± 0.183	0.501 ± 0.184	$\mathbf{0.167} \pm 0.039$
Symmetric	MSE	0.284 ± 0.130	0.285 ± 0.208	0.289 ± 0.131	0.390 ± 0.161	0.507 ± 0.092	$\mathbf{0.224} \pm 0.164$
	MAE	0.397 ± 0.136	0.336 ± 0.176	0.397 ± 0.136	0.500 ± 0.149	0.612 ± 0.053	$\mathbf{0.306} \pm 0.157$
	NLL	0.501 ± 0.450	1.272 ± 8.276	$\mathbf{0.431} \pm 0.440$	1.450 ± 0.793	1.216 ± 0.486	0.484 ± 1.422
	CE	$\mathbf{0.018} \pm 0.030$	0.033 ± 0.038	0.019 ± 0.030	0.056 ± 0.039	0.021 ± 0.018	0.031 ± 0.034
	ECE	0.018 ± 0.006	0.018 ± 0.012	0.016 ± 0.006	0.044 ± 0.064	0.025 ± 0.006	$\mathbf{0.015} \pm 0.009$
	Sharpness	0.523 ± 0.117	0.287 ± 0.148	0.501 ± 0.102	0.430 ± 0.119	0.625 ± 0.128	$\mathbf{0.262} \pm 0.095$
Asymmetric	MSE	0.253 ± 0.193	0.234 ± 0.249	0.282 ± 0.196	0.245 ± 0.215	0.474 ± 0.210	$\mathbf{0.200} \pm 0.199$
	MAE	0.335 ± 0.174	0.269 ± 0.175	0.322 ± 0.155	0.328 ± 0.178	0.575 ± 0.153	$\mathbf{0.249} \pm 0.137$
	NLL	$\mathbf{0.536} \pm 1.577$	2.543 ± 9.314	18.306 ± 51.006	0.771 ± 0.859	1.018 ± 0.492	4.551 ± 11.286
	CE	$\mathbf{0.026} \pm 0.028$	0.027 ± 0.027	0.034 ± 0.033	0.050 ± 0.039	0.030 ± 0.025	0.027 ± 0.025
	ECE	0.018 ± 0.009	0.016 ± 0.012	0.020 ± 0.010	0.028 ± 0.008	0.056 ± 0.277	$\mathbf{0.015} \pm 0.010$
	Sharpness	0.394 ± 0.167	0.226 ± 0.164	0.248 ± 0.188	0.429 ± 0.065	0.974 ± 2.792	$\mathbf{0.164} \pm 0.095$
Boston House	MSE	0.058 ± 0.045	0.072 ± 0.031	0.090 ± 0.179	0.048 ± 0.019	0.052 ± 0.025	$\mathbf{0.046} \pm 0.017$
	MAE	0.177 ± 0.062	0.217 ± 0.045	0.209 ± 0.112	0.170 ± 0.037	0.178 ± 0.045	$\mathbf{0.162} \pm 0.034$
	NLL	$\mathbf{-0.035} \pm 0.441$	4.004 ± 5.679	0.540 ± 1.379	0.034 ± 0.093	0.140 ± 0.396	0.115 ± 0.538
	CE	0.019 ± 0.016	0.143 ± 0.089	0.023 ± 0.028	0.030 ± 0.017	0.031 ± 0.024	$\mathbf{0.018} \pm 0.017$
	ECE	0.004 ± 0.002	1.725 ± 2.295	0.004 ± 0.003	0.005 ± 0.002	0.007 ± 0.003	$\mathbf{0.003} \pm 0.001$
	Sharpness	0.238 ± 0.058	1.862 ± 2.354	0.215 ± 0.100	0.331 ± 0.022	0.328 ± 0.093	$\mathbf{0.156} \pm 0.035$
Yacht	MSE	0.003 ± 0.002	0.018 ± 0.089	0.008 ± 0.059	0.001 ± 0.000	0.002 ± 0.005	$\mathbf{0.000} \pm 0.000$
	MAE	0.041 ± 0.016	0.038 ± 0.130	0.025 ± 0.082	0.019 ± 0.007	0.033 ± 0.027	$\mathbf{0.016} \pm 0.005$
	NLL	-0.985 ± 0.249	$\mathbf{-1.993} \pm 0.971$	-1.960 ± 0.632	-0.193 ± 0.008	-0.591 ± 0.601	-1.804 ± 0.201
	CE	0.120 ± 0.042	$\mathbf{0.101} \pm 0.049$	0.147 ± 0.055	0.248 ± 0.010	0.193 ± 0.052	0.119 ± 0.034
	ECE	0.011 ± 0.003	0.176 ± 0.847	0.007 ± 0.021	0.031 ± 0.001	0.023 ± 0.013	$\mathbf{0.005} \pm 0.001$
	Sharpness	0.160 ± 0.037	0.244 ± 0.971	0.071 ± 0.100	0.328 ± 0.002	0.266 ± 0.142	$\mathbf{0.066} \pm 0.014$

Table 1: Benchmark comparison showing mean and standard deviation across 100 different training sample draws. Best (lowest mean) values per row are highlighted in bold. Sharpness measures the average prediction uncertainty (lower is better, indicating sharper/more confident predictions).

its MAP estimate for the reduced dataset. This procedure aims to reduce predictive uncertainty and mitigate the effect of detrimental data points before model-based rollouts are performed.

Finally, this model is used to roll out the current policy π_n, generating synthetic transitions that are stored in the buffer $\mathcal{D}_{SAC}$. The current policy is then further optimized using both real and synthetic experiences through the learned state-action value function $Q^{\pi_n}(x, a)$.

F Limitations and Future Work

Although the proposed removal procedure is substantially more efficient than exhaustive search or extensive leave-one-out evaluations, it still incurs computational overhead compared to a standard homoscedastic GP because each greedy step requires updating the model and evaluating the removal influence under the current posterior. Compared to other robust Gaussian process Bayesian Optimization baselines, we have comparable or lower computational costs Figure 21.

The accompanying near-optimality guarantee is based on a restricted curvature condition that serves as a theoretical certificate rather than an algorithmic constraint; in practice, these conditions may be conservative, and the method can remain effective even when they are not verified explicitly. Finally, the current formulation is tailored to Gaussian process models, and it would be interesting to investigate analogous removal influences for parametric models.

G Experimental Setup - Transductive Learning

We provide in Table 1, the absolute values with standard deviations for the results in Figure 3.

We evaluate F2I in the transductive learning setting, in addition to the real world data sets, on two synthetic functions with a detrimental wavelength structure. Therefore, we consider a varying lengthscale across the domain, where our target region-aware algorithm F2I significantly improves in prediction accuracy and calibration. We therefore consider the test-functions *Symmetric Lengthscale (LS)* $f_{symmetric}(x)$ and *Asymmetric Lengthscale (LS)* $f_{asymmetric}(x)$, which are formally defined

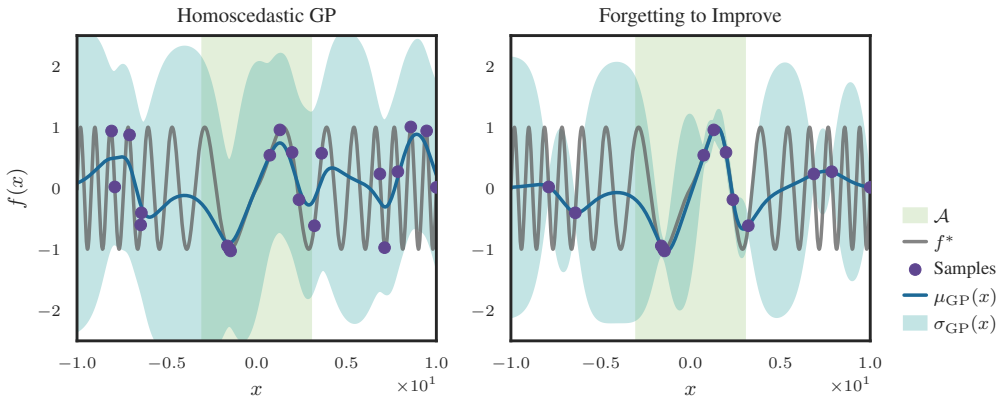


Figure 7: Influence of detrimental data structures based on *Symmetric LS* test functions. (Left) A standard homoscedastic GP absorbs the detrimental lengthscale structures in the noise. (Right) We visualize the result of applying F2I.

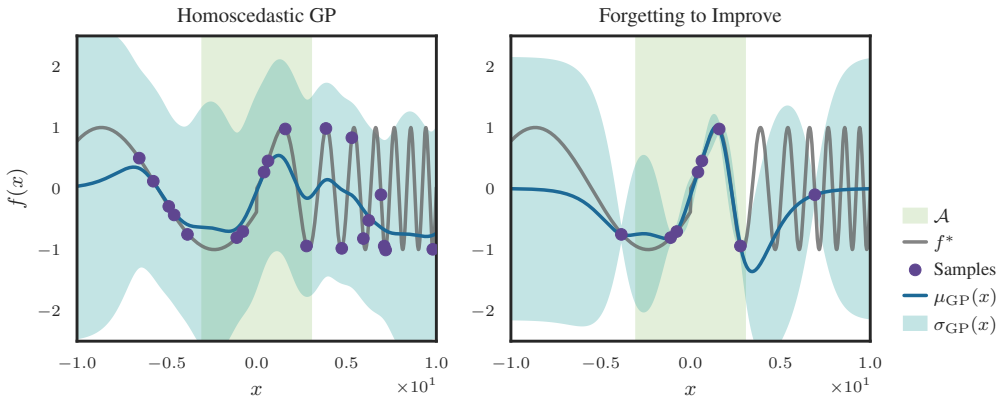


Figure 8: Influence of detrimental data structures based on *Asymmetric LS* test functions. (Left) A standard homoscedastic GP absorbs the detrimental lengthscale structures in the noise. (Right) We visualize the result of applying F2I.

by

$$f_{symmetric}(x) = \sin(0.5 + 0.39 \cdot |x|), \tag{71}$$

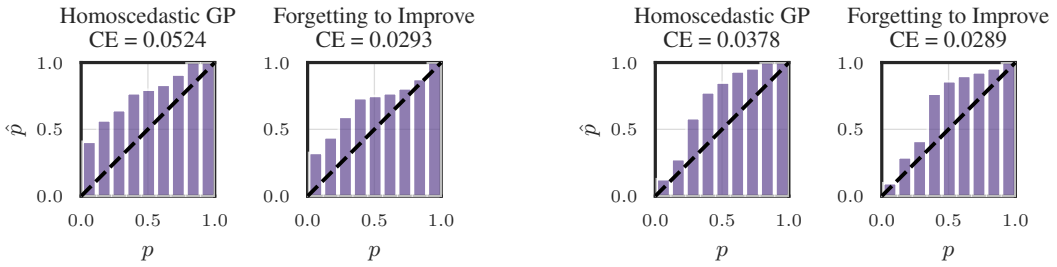
$$f_{assymetric}(x) = \begin{cases} \sin(0.5 + 0.39x) & \text{if } x \geq 0 \\ \sin(0.5 + 0.39\frac{1}{x}) & \text{otherwise.} \end{cases} \tag{72}$$

We provide visualizations comparing the homoscedastic GP and F2I for both synthetic test functions in Figures 7 and 8. Further, we also provide the calibration metrics for the two examples in Figure 9. The two designed test functions do not have specific outliers, but are rather detrimental in their overall design for the chosen kernel with a single lengthscale. We observe, that *relevance pursuit* is outperformed by F2I, which is also acknowledged in the limitations section of [4] by not considering the location of removed samples in *relevance pursuite*.

In addition, for all benchmarks in the transductive learning setting, we report the domain and task specifications in Table 2.

Additionally, we show for the Dow Jones data set, how F2I handles different extreme target region specifications that correspond to having $\mathcal{A} = \mathcal{X}$ and a very small concentrated target region $\mathcal{A} = [0.1, 0.23]$ around the spike in the stock price. We demonstrate in Figures 10 and 11 that F2I is capable of handling both cases, due to its task formulation using the target region $\mathcal{A}$.

We evaluated all methods for 100 random seeds for an RBF kernel with a log prior $log(0, 1)$ on the lengthscale and a Gamma prior with parameters $a = 2$ and $b = 0.15$ for the aleatoric noise variance. All transductive learning experiments are run on an Intel Core Ultra 9 185H CPU with a NVIDIA RTX 3000 Ada GPU.



(a) Symmetric lengthscale increase

(b) Asymmetric lengthscale increase

Figure 9: Calibration comparison of (left) homoscedastic GP and (right) F2I for (a) symmetric and (b) asymmetric lengthscale increases test functions.

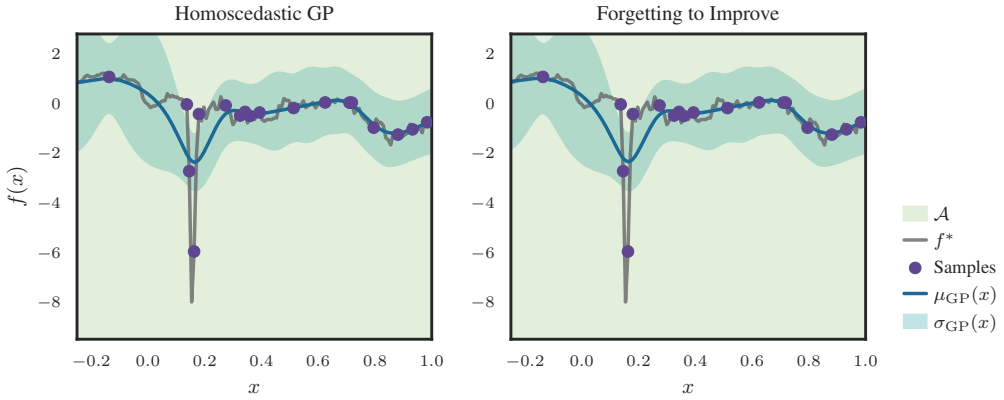


Figure 10: Influence of detrimental data structures based on *Dow Jones* data set with target region $\mathcal{A} = \mathcal{X}$. F2I retains all data points to have the best global fit, since the target region $\mathcal{A}$ spans $\mathcal{X}$.

H Experimental Setup - Bayesian Optimization

For our BO benchmarks, we provide the standard domain limits, the dimensionality and the global optimum in Table 3. During the optimization process, we rescale all objective functions into the hypercube $[0, 1]^n$, with dimensionality n. Further, we start all experiments from 5 initial random observations and proceed with a sampling budget of 50 for the remaining optimization. We perform all BO experiments on an Intel Core Ultra 9 185H CPU with a NVIDIA RTX 3000 Ada GPU.

I Bayesian Optimization - Additional Experiments

As a challenging real-world benchmark for Bayesian optimization, we consider hyperparameter tuning for the GLMNET training algorithm [19] across multiple datasets. Following [61, 56, 60], we

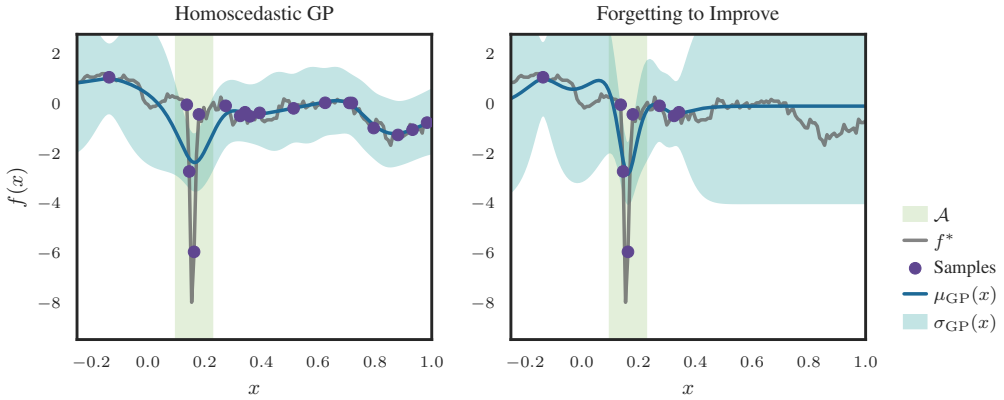


Figure 11: Influence of detrimental data structures based on *Dow Jones* data set with aconcentarted target region $\mathcal{A} = [0.1, 0.23]$. F2I retains data points within the spike.

Table 2: Hyperparameters and settings for the transductive learning benchmarks.

Setting	Dow Jones	Boston Housing	Symmetric LS	Asymmetric LS
$\mathcal{X}$ limits	$[-1.5,\ 1.5]$	$[-15,\ 15]^{13}$	$[-10,\ 10]$	$[-10,\ 10]$
$\mathcal{A}$ limits	$[0.75,\ 1.25]$	$[-0.7,\ 0.4]^{13}$	$[-3,\ 3]$	$[-3,\ 3]$
$\lvert\mathcal{X}\rvert$	405	506	1000	1000
$\lvert\mathcal{S}\rvert$	50	40	20	20

Table 3: Bayesian Optimization Test Functions: Domain Limits and Optimal Values

Function	Dim.	Domain	Optimum	Optimizer
Ackley	2	$[-10, 30]^2$	0.0	$(0, 0)$
Ackley	6	$[-10, 30]^6$	0.0	$(0, \ldots, 0)$
A. Rosenbrock	4	$[-5, 10]^2 \times [1, 0]^2$	0.0	$(1, \ldots, 1)$
Branin	2	$[-5, 0] \times [10, 15]$	-0.3979	$(-3.14, 12.27)$
Griewank	2	$[-50, 20]^2$	0.0	$(0, 0)$
Holder Table	2	$[-10, 10]^2$	19.21	$(8.06, 9.67)$
Hartmann	6	$[0, 1]^6$	-3.322	$(0.2, 0.15, 0.48, 0.28, 0.31, 0.66)$
Levy	2	$[-10, 10]^2$	0.0	$(1, 1)$

replace the expensive evaluation step with a table lookup constructed from a large set of precomputed evaluations [42]. Using the first seven classification datasets from OpenML [9], we study in Figure 12 whether Forgetful BO can improve transfer learning of hyperparameters across tasks. We evaluate the first seven tasks sequentially, using 25 samples per task. To highlight the effect of forgetting potentially detrimental information from previous datasets, and thus improving transfer across tasks, we compare our method against all baselines in two settings: one that retains all observations from previous tasks, and one that uses only data from the current task. Most importantly, all baselines benefit from retaining data from previous tasks, since this allows them to transfer knowledge about well-performing configurations across datasets. At the same time, we observe that Forgetful BO outperforms all baselines by selectively discarding information that is harmful for subsequent tasks. Both the cumulative regret in the left panel and the simple regret in the right panel confirm the value of forgetting in this setting.

For Figure 4, we add in Figure 13 the two additional baselines for robust Bayesian Optimization, Student-t likelihood [48] and trimmed marginal log likelihood [5], are compared to Forgetful BO. We observe that both trimmed MLL and Student-t likelihood are outperformed by Forgetful BO. We extend our evaluation in Figure 14 to the outlier setting of [4], where the test functions are perturbed

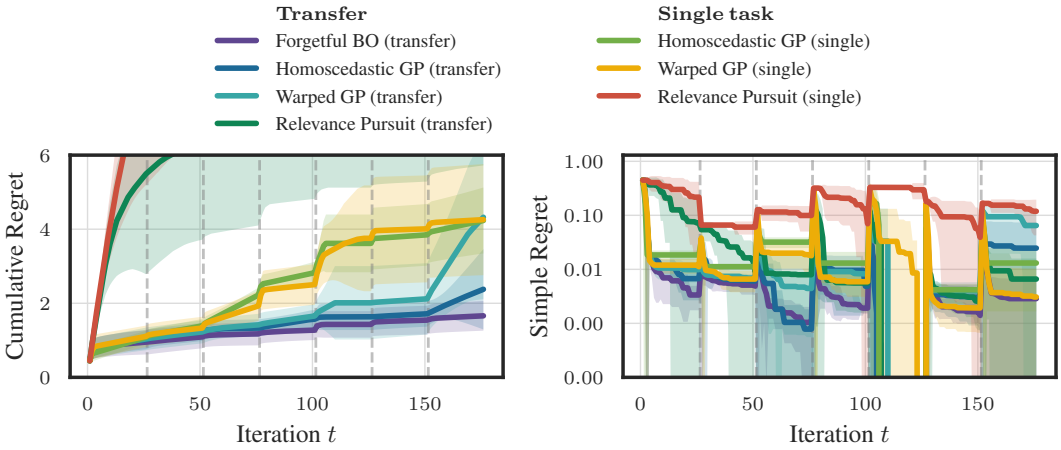


Figure 12: Transfer learning AutoML benchmark for seven sequential GLMNET hyper-parameter training tasks. Comparison of (left) cumulative regret and (right) simple regret

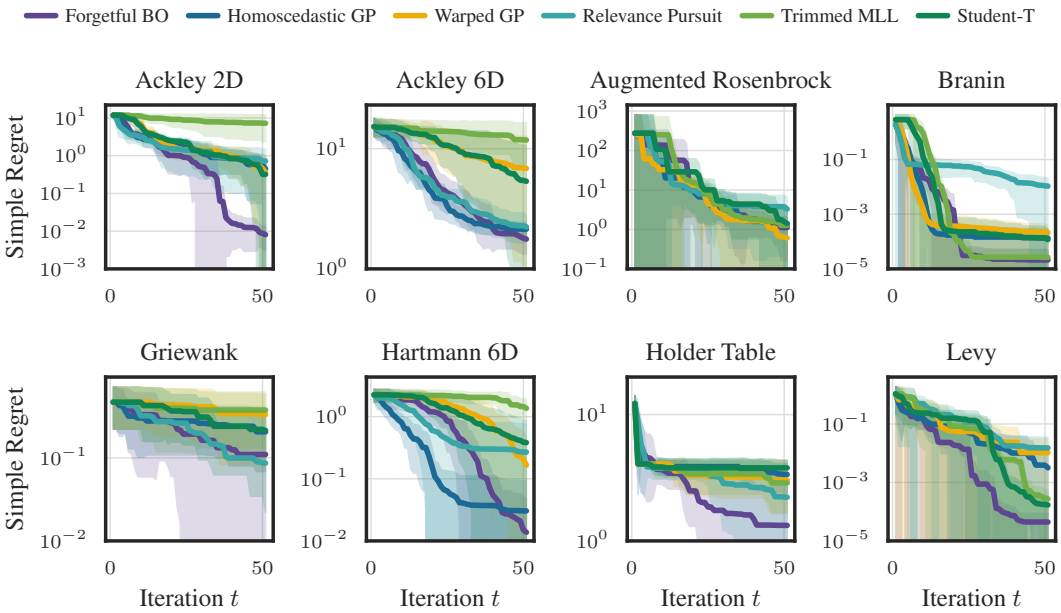


Figure 13: Simple regret of Forgetful BO and competing baselines for test functions.

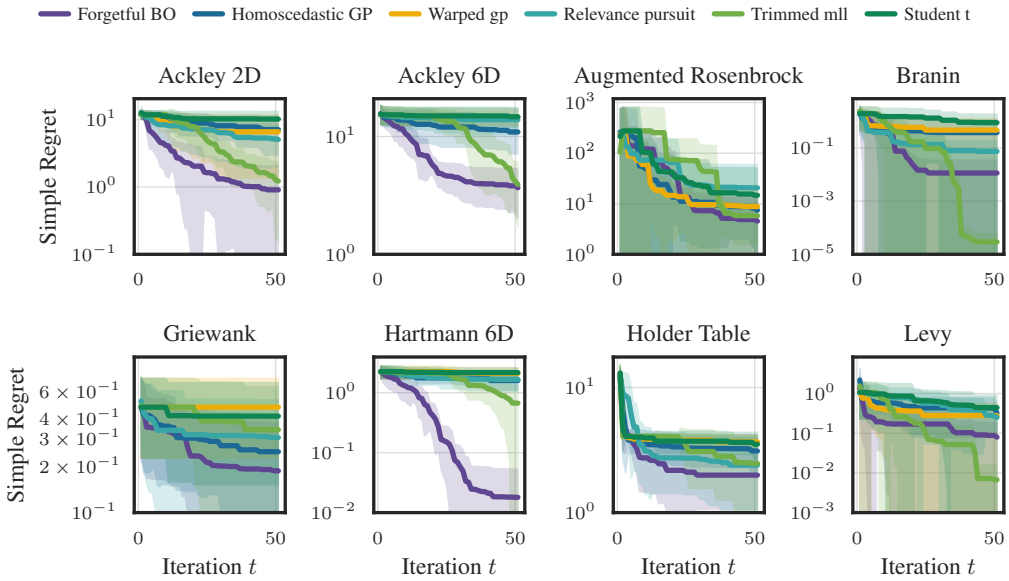


Figure 14: Simple regret of Forgetful BO and competing baselines for *perturbed* test functions.

using constant outliers with probability $p = 0.1$. This experiment demonstrates that Forgetfull BO also improves optimization in the perturbed setting and outperforms all competing baselines. Also note that *relevance search* performs better relative to baselines in the perturbed setting, compared to the unperturbed, as mentioned in [4], the algorithm might remove important samples related to the optimum by mistaking them for outliers, due to the dependence on the missing target region. Additionally, we provide the cumulative regret results for the BO experiments from Figures 4 and 14. Figure 15 demonstrates that Forgetful BO is not only reaching good simple regret, but also has a fast diminishing regret, so that the accumulated value over all iterations $t = 1 \dots T$ is lower or on par compared to all competing baselines. We also report the cumulative regret for the perturbed setting of Figure 14 in Figure 16. In addition, Forgetful BO can seamlessly integrate different acquisition functions than *GP-UCB* [66]. TO demonstrate this, we evaluate Forgetful BO on all benchmarks using the *log expected improvement* (LEI) acquisition function [3]. Therefore, we also use LEI for all competitive baselines and visualize the achieved regret in Figure 17. We note that, except for

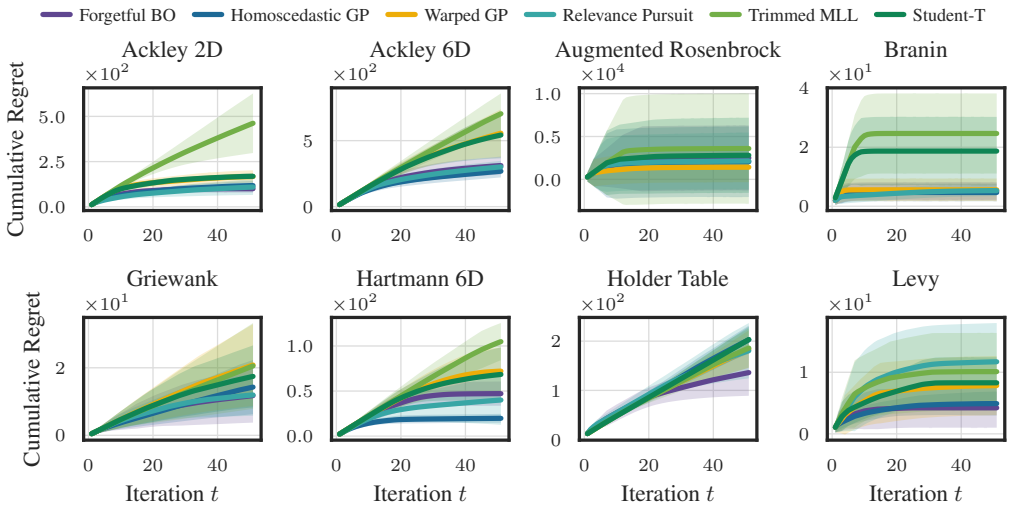


Figure 15: Cumulative regret of Forgetful BO and competing baselines for different test functions.

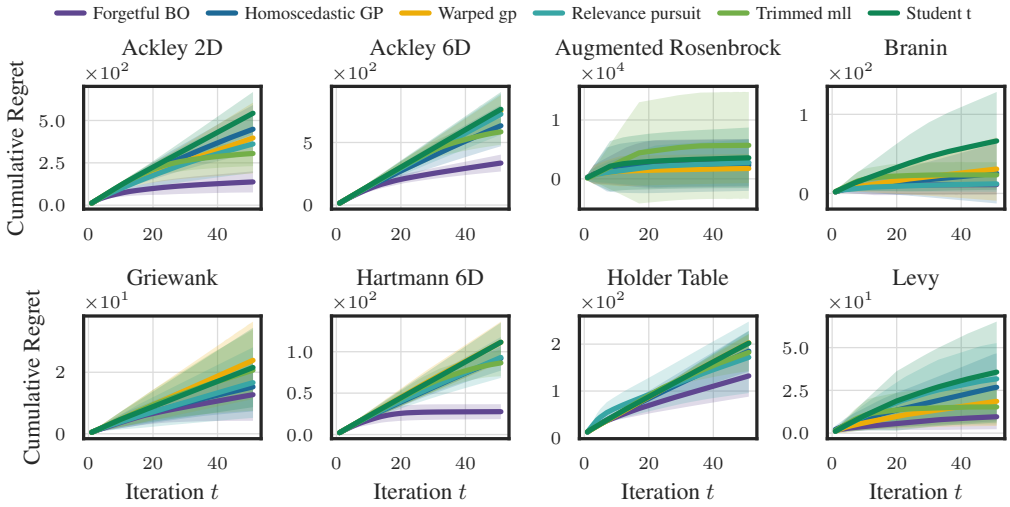


Figure 16: Cumulative regret of Forgetful BO and competing baselines for different test functions with constant corruptions.

the *Ackley 6D* test-function, Forgetful BO performs better or on par with competing baselines. The cumulative regret for the LEI experiments are also provided in Figure 18. To complete this evaluation, we provide the performance of Forgetful BO on the noisy benchmarks using the LEI acquisition function in Figures 19 and 20. We observe, that also in this setting Forgetful BO demonstrates superior or equal performance across all benchmarks and compared to all baselines.

J Bayesian Optimization Computational Complexity

We demonstrate in Figure 21, that Forgetful BO, based on the removal algorithm F2I is computationally efficient. We report the mean and standard deviation of the acquisition time per sample. We visualize the mean across all random seeds and the entire optimization process of Figure 4 in Figure 21. Note that Forgetful BO outperforms *relevance pursuite* by nearly one order of magnitude in the log-scale, while having small computational overhead compared to the *homoscedastic GP* and the *Warped GP*. The large variance of the Forgetful BO runtime is caused due to the potentially higher number of removals in later stages of the optimization process.

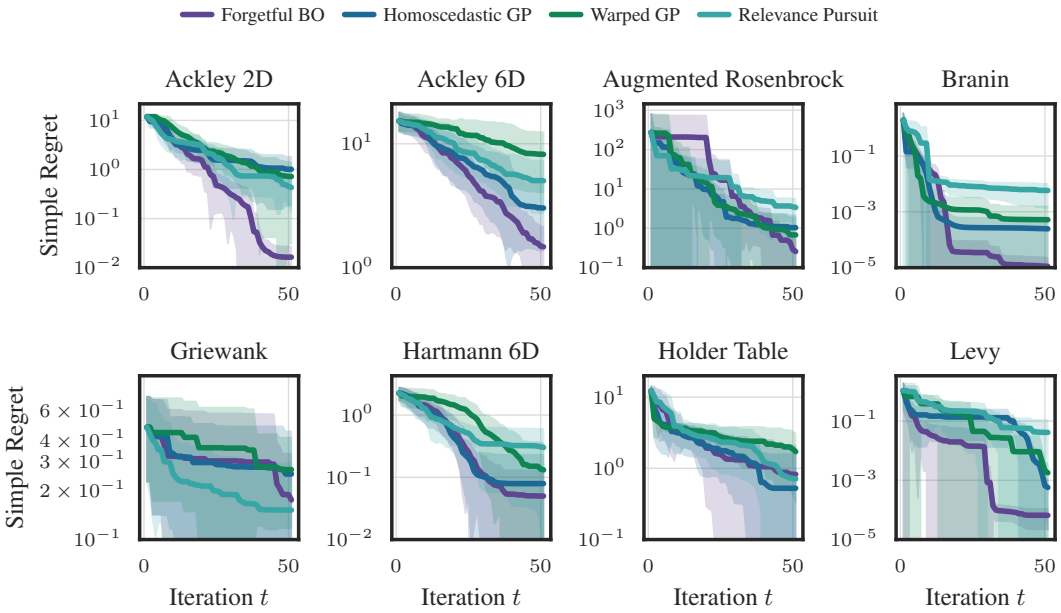


Figure 17: Simple regret of Forgetful BO and competing baselines for different test functions using log expected improvement acquisition function.

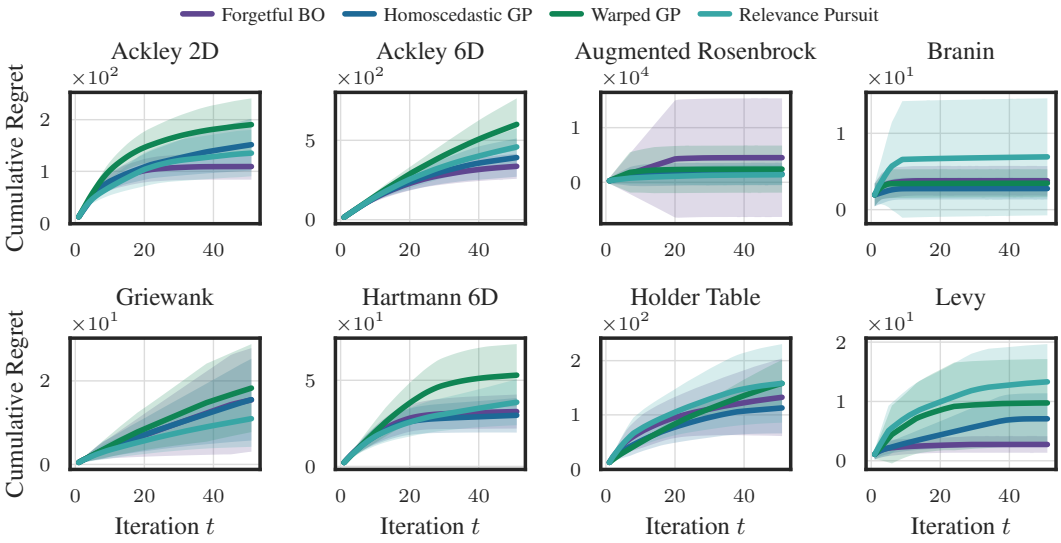


Figure 18: Cumulative regret of Forgetful BO and competing baselines for different test functions using log expected improvement acquisition function.

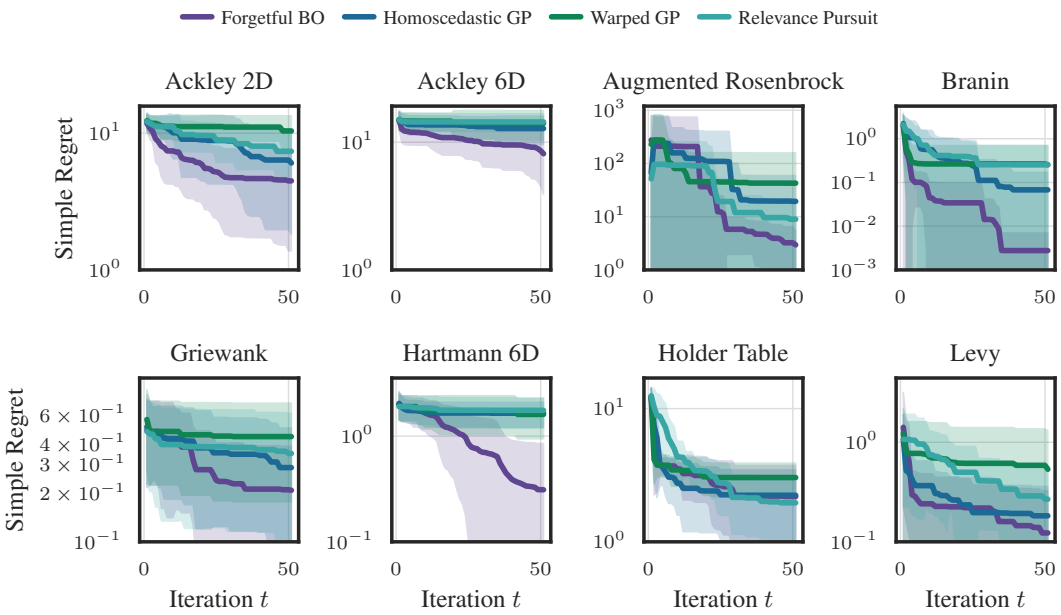


Figure 19: Simple regret of Forgetful BO and competing baselines for perturbed test functions using log expected improvement acquisition function.

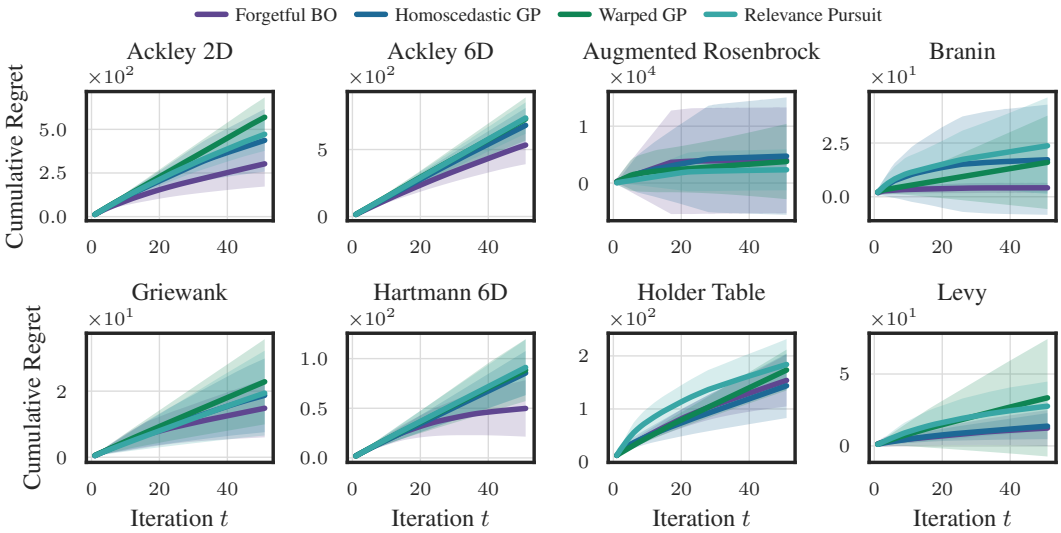


Figure 20: Cumulative regret of Forgetful BO and competing baselines for perturbed test functions using log expected improvement acquisition function.

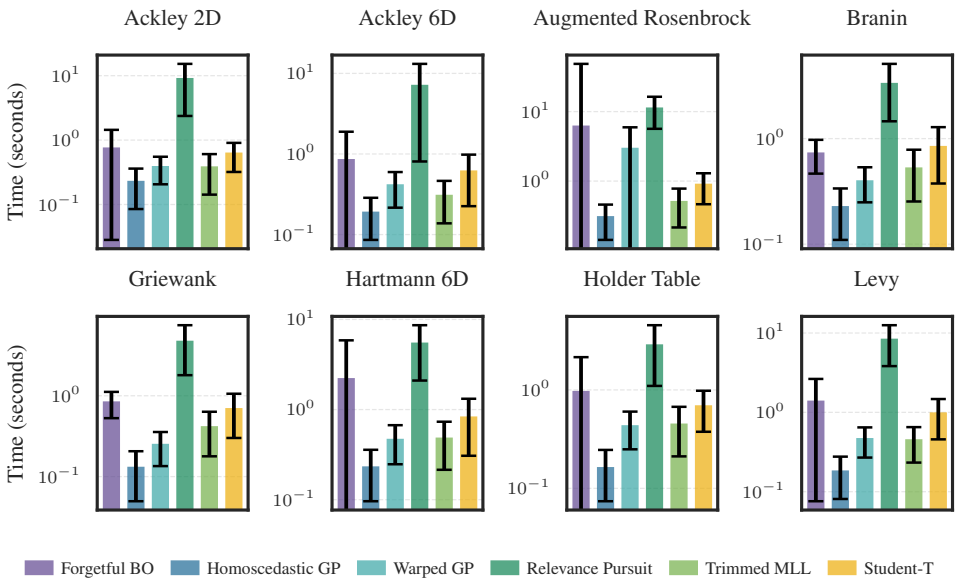


Figure 21: Acquisition time comparisons for the different BO algorithms. Comparing mean and standard deviation across all evaluated non-corrupted benchmarks and baselines.

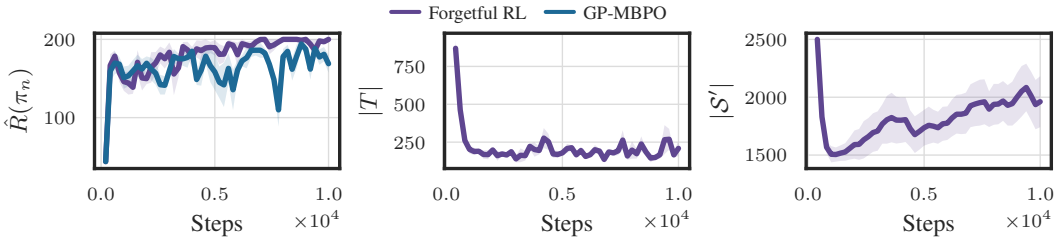


Figure 22: Cartpole Balance comparing Forgetful RL with vanilla GP-based MBPO. Performance throughout learning (left), number of removed samples (middle), and current buffer size (right).

K Reinforcement Learning - Additional Experiments

We also conducted additional experiments for Forgetful RL using a Matérn $5/2$ kernel and a single NVIDIA RTX 4090 GPU per run, with 5 CPU cores. Comparing vanilla MBPO with Forgetful RL on *Cartpole Balance*, we observe in Figure 22 a much better convergence to the globally optimal performance. We observe, that Forgetful RL removes especially in the beginning of the learning process, while continuous to remove detrimental experiences throughout the learning process. Compared to the vanilla MBPO implementation with 2500 experiences in the buffer throughout, the replay buffer is not entirely used, only containing up to 2000 experiences at convergence. Further, we also consider tasks, which start in a downward pendulum position and require to swing up. This task nature requires a significantly larger target region, where the dynamics should be modeled well, and which can definitely no longer be approximated with a linear model. We first consider the *inverted pendulum* task in Figure 23. We observe, that relatively little data removal occurs due to a much larger target region, but these lead to better convergence compared to the vanilla MBPO. We observe in Figure 23, that we initially remove most points from the replay buffer, that we obtained from the initial random samples prefilling the buffer. We ablate in Figure 24 the removal behavior by not filling the buffer with random experiences before the initial episode. We observe that we converge slower and require to remove more detrimental experiences throughput learning, which have been gathered during suboptimal episodes. Further, we extend the swingup task in Figure 25 to *Cartpole Swingup*. Analogous to the other *swingup*, tasks we also remove fewer samples than in the *Cartpole Balance* task, because large parts of the state space are visited of the optimal policy, and require detailed modeling. We additionally increase the task difficulty of *Inverted Pendulum* by introducing a sparse reward function. In Figure 26, we observe that the task is significantly harder with a sparse reward and requires longer to converge. Forgetful RL also improves in this problem setting the learning behavior and reaches the optimum.

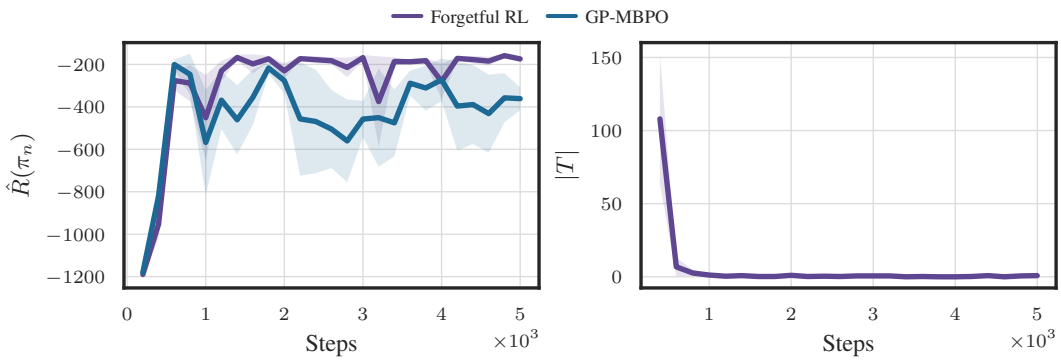


Figure 23: Comparing the performance and data removal of Forgetful RL and vanilla GP-based MBPO for *Inverted Pendulum*.

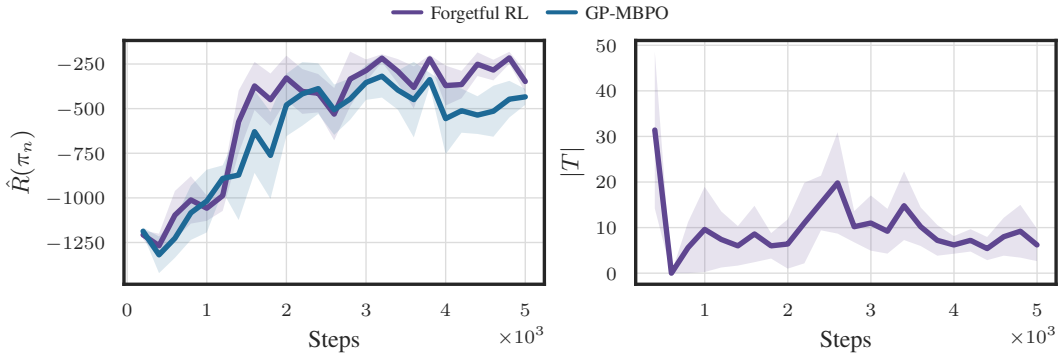


Figure 24: Comparing the performance and data removal of Forgetful RL and vanilla GP-based MBPO for *Inverted Pendulum*, beginning with an empty buffer.

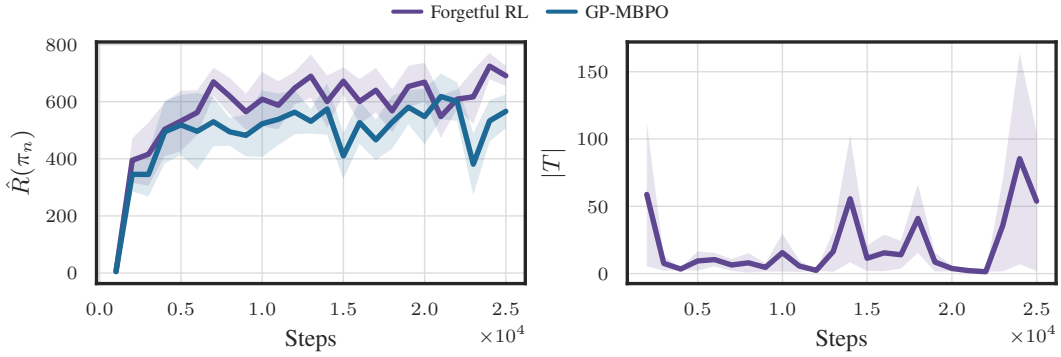


Figure 25: Comparing the performance and data removal of Forgetful RL and vanilla GP-based MBPO for *Cartpole Swingup*.

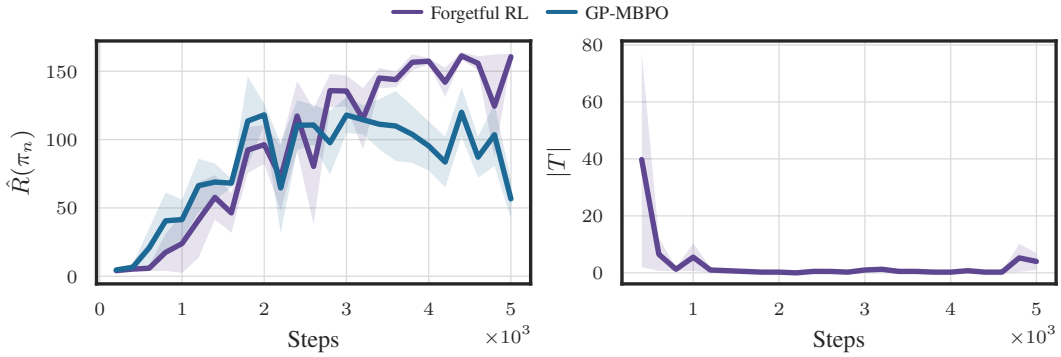


Figure 26: Comparing the performance and data removal of Forgetful RL and vanilla GP-based MBPO for *Cartpole Swingup* on a sparse reward task.

L Near-Optimality Proofs using Approximate Submodularity

In order to establish the near-optimality proof under a curvature condition on the magnitude of the prior parameter a of the gamma prior $p(\sigma_w^2(\mathcal{S})) = \frac{b^a}{\Gamma(a)}(\sigma_w^2(\mathcal{S}))^{a-1}e^{-b\sigma_w^2(\mathcal{S})}$, we proceed as follows

1. We define a set-function $\bar{\mathcal{J}}$ for the considered removal process. We thereby convert our problem formulation into a greedy selection process, by adding points to the set fo removed points. Additionally, we convert the objective into a maximization problem for the negated objective $-\mathcal{J}$, and normalize the set function $\bar{\mathcal{J}}$ to zero by adding the objective evaluated for the entire set of samples $\mathcal{J}(\mathcal{S})$. These modifications do not affect the marginal contributions of individual elements as shown in Lemma 1, and therefore preserve the approximate submodularity ratio, used in the statement of near-optimality. Furthermore, this set-function formulation exactly matches the problem setting considered in Elenberg et. al. [18].
2. Secondly, we show in Lemma 2 that F2I selects the data points to be greedily removed along the largest gradient. Therefore, we demonstrate that $-u_i(\mathcal{A}) - v_i(\mathcal{A})$ is the gradient $\frac{\partial}{\partial h_i} - \mathcal{J}(\mathcal{A}_i)$ of the negated objective $-\mathcal{J}$. This therefore recovers the maximization objective and sample selection of the *Orthogonal Matching Pursuit* (OMP) algorithm [18, Algorithm 2].
3. The near optimality guarantee of OMP [18, Algorithm 2] applies for any function that satisfies *M-restricted smoothness* (RSM) and *m-restricted strong concavity* (RSC), which we define in Definition 1.
4. In order to show that the negated objective $-\mathcal{J}$ satisfies the RSC condition, we have to analyze its curvature. Therefore, we first compute in Lemma 3 the curvature of the aleatoric noise parameter σ_w^2. Based on Lemma 3, we derive in Lemma 4 the entire Hessian of $-\mathcal{J}$, by deriving the epistemic and aleatoric contribution to the curvature.
5. Using the derived Hessian of $-\mathcal{J}$, we can now derive the minimal and maximal curvatures on the removal set in order to obtain m_R and M_R of the RSM and RSC definition in Definition 1.
6. To satisfy the RSC condition, we require the curvature of the objective to be negative on all removal directions, such that $m_R > 0$. In Lemma 6, we derive a condition on the prior parameter a to guarantee concavity of the inner optimization problem, providing a unique optimal noise level $\sigma_w^2(\mathcal{S}')$ for each retained data set. For the outer optimization process of the set-function $\bar{\mathcal{J}}$, we derive an RSC condition in Lemma 7. Thereby, the concavity of the aleatoric Hessian contribution has to outweigh the potentially convex curvature of the epistemic Hessian along all considered removal directions. The curvature condition depends on the prior parameter a of the Gamma prior.
7. Given the set-function satisfies these conditions, we obtain with the intermediate derivations in Lemma 8 and Lemma 9 similar to [18], the near optimality guarantee in Theorem 1.

In the following, we list all relevant lemmas and definitions required for Theorem 1:

Lemma 1. *Consider the set of given samples $\mathcal{S}$ and the objective $\min_{\mathcal{S}'\subseteq\mathcal{S}} \mathcal{J}(\mathcal{S}')$ in Equation (3). Let the removed set of samples be $T = \mathcal{S} \setminus \mathcal{S}'$, then we defined the set function $\bar{\mathcal{J}}(T)$ as*

$$\max_{T\subseteq\mathcal{S}} \bar{\mathcal{J}}(T) \coloneqq \mathcal{J}(\mathcal{S}) - \mathcal{J}(\mathcal{S}') = \mathcal{J}(\mathcal{S}) - \mathcal{J}(\mathcal{S}\setminus T).$$

This set function has the same form as [18, Equation 4], with the submodularity ratio $\gamma_{L,U}$, which is defined for any $L \subseteq T$ and $U \subseteq T \setminus L$ as

$$\gamma_{L,U}(\bar{\mathcal{J}}) = \frac{\sum_{j\in U}(\bar{\mathcal{J}}(L\cup\{j\}) - \bar{\mathcal{J}}(L))}{\bar{\mathcal{J}}(L\cup U) - \bar{\mathcal{J}}(L)}.$$

Proof. Consider a retained set $\mathcal{S}'$ and its complement set $T = \mathcal{S} \setminus \mathcal{S}'$. Additionally let $L \subseteq T$ and $U \subseteq T \setminus L$, and $j \in U$,

$$\bar{\mathcal{J}}(L\cup\{j\}) - \bar{\mathcal{J}}(L) = \mathcal{J}(\mathcal{S}) - \mathcal{J}(\mathcal{S}\setminus(L\cup\{j\})) - \mathcal{J}(\mathcal{S}) + \mathcal{J}(\mathcal{S}\setminus L) \tag{73}$$

$$= \mathcal{J}(\mathcal{S}\setminus L) - \mathcal{J}(\mathcal{S}\setminus(L\cup\{j\})) \tag{74}$$

Thus, the forward marginal on $\bar{\mathcal{J}}$ equals the removal marginal on $-\mathcal{J}$. The submodularity ratio defined in [18, Def. 2] depends only on the marginal combinations. Hence, the submodularity ratio for the set-function considering the complement removal sets $L \subseteq T$ and $U \subseteq T \setminus L$ is given by

$$\gamma_{L,U}(\bar{\mathcal{J}}) = \frac{\sum_{j \in U}(\bar{\mathcal{J}}(L \cup \{j\}) - \bar{\mathcal{J}}(L))}{\bar{\mathcal{J}}(L \cup U) - \bar{\mathcal{J}}(L)}. \tag{75}$$

□

Lemma 2. *Let $\mathcal{S} \subseteq \mathcal{X}$ and an individual sample $x_i \in \mathcal{S}$, then the negated objective is given by*

$$-\mathcal{J}(\mathcal{S}) = -\sum_{x \in \mathcal{A}} \sigma^2_{\mathcal{S}}(x).$$

Let the RKHS removal direction associated with a sample i be

$$k(x, \mathcal{S}, h_i) = k_{x,\mathcal{S}} - h_i k(x, x_i) e_i, \qquad h_i \in [0, 1] \subset \mathbb{R},$$

i.e. an infinitesimal decrease of the kernel coupling to x_i. Then the directional derivative of $-\mathcal{J}$ at $\mathcal{S}$ in this RKHS removal direction satisfies

$$\frac{\partial}{\partial h_i} - \mathcal{J}(\mathcal{A}_i)\Big|_{h_i=0} = -u_i(\mathcal{A}) - v_i(\mathcal{A}),$$

where $u_i(\mathcal{A})$ and $v_i(\mathcal{A})$ are the epistemic and aleatoric influences defined in Propositions 1 and 4.

Proof. We compute the gradient of $-\mathcal{J}(\mathcal{S})$ under an infinitesimal reduction of the RKHS basis component $k(\cdot, x_i)$. We decompose the predictive variance formulation as follows

$$-\mathcal{J}(\mathcal{S}) = \sum_{x \in \mathcal{A}} \left(-k(x, x) + k_{x,\mathcal{S}} A^{-1} k_{\mathcal{S},x}\right) - |\mathcal{A}| \sigma^2_w(\mathcal{S}), \tag{76}$$

where $|\mathcal{A}|$ is the Lebesgue measure of the target are $\mathcal{A}$. We next differentiate every term that is dependent on the removal parameter h_i. Since $-\mathcal{J}$ is continuous in $x \in \mathcal{A}$ and we have defined our surrogate function to be continuous in h_i, we interchange the integral over $\mathcal{A}$ and the derivative

$$\frac{\partial - \mathcal{J}(\mathcal{S})}{\partial h_i} = \sum_{x \in \mathcal{A}} \left(-\frac{\partial}{\partial h_i} k(x, x) + \frac{\partial}{\partial h_i} k(x, \mathcal{S}, h_i) A(h_i)^{-1} k(\mathcal{S}, x, h_i)\right) - |\mathcal{A}| \frac{\partial}{\partial h_i} \sigma^2_w(\mathcal{S}) \tag{77}$$

$$= \sum_{x \in \mathcal{A}} \Bigg(2 \frac{\partial k(x, \mathcal{S}, h_i)}{\partial h_i} A^{-1} k_{\mathcal{S},x} - k(x, \mathcal{S}, h_i) A(h_i)^{-1} \frac{\partial A(h_i)}{\partial h_i} A(h_i)^{-1} \tag{78}$$

$$\times k(\mathcal{S}, x, h_i)\Bigg) - |\mathcal{A}| \frac{\partial}{\partial h_i} \sigma^2_w(\mathcal{S}')$$

$$= \sum_{x \in \mathcal{A}} \Bigg(-2k(x, x_i)(A^{-1} k(\mathcal{S}, x))_i - k(x, \mathcal{S}, h_i) A(h_i)^{-1} \frac{\partial K(h_i)}{\partial h_i} A(h_i)^{-1} k(\mathcal{S}, x, h_i) \tag{79}$$

$$- k(x, \mathcal{S}, h_i) A(h_i)^{-1} \frac{\partial \sigma^2_w(\mathcal{S})}{\partial h_i} A(h_i)^{-1} k(\mathcal{S}, x, h_i)\Bigg) - |\mathcal{A}| \frac{\partial}{\partial h_i} \sigma^2_w(\mathcal{S})$$

$$= \sum_{x \in \mathcal{A}} -2k(x, x_i)(A^{-1} k(\mathcal{S}, x))_i - k(x, \mathcal{S}, h_i) A(h_i)^{-1} (-e_i k(x_i, \mathcal{S}) - k(\mathcal{S}, x_i) e_i^\top) \tag{80}$$

$$\times A(h_i)^{-1} k(\mathcal{S}, x, h_i) - \frac{\partial}{\partial h_i} \sigma^2_w(\mathcal{S})(1 + k(x, \mathcal{S}, h_i) A(h_i)^{-2} k(\mathcal{S}, x, h_i))$$

$$= \sum_{x \in \mathcal{A}} \underbrace{-2k(x, x_i)(A^{-1} k(\mathcal{S}, x))_i + 2(A^{-1} k(\mathcal{S}, x))_i (k(x_i, \mathcal{S}) A^{-1} k(\mathcal{S}, x))}_{-u_i(x)} \tag{81}$$

$$\underbrace{-(1 + k(x, \mathcal{S}, h_i) A(h_i)^{-2} k(\mathcal{S}, x, h_i)) \frac{\partial}{\partial h_i} \sigma^2_w(\mathcal{S})}_{-v_i(x)}$$

This completes the proof that $-u_i(\mathcal{A}) - v_i(\mathcal{A})$ is the directional derivative of the negated objective $-\mathcal{J}$ in the RKHS removal direction induced by the sample x_i, when evaluating at $h_i = 0$. □

Definition 1 (Restricted Strong Concavity, Restricted Smoothness)**.** *A function $-\mathcal{J}(\cdot)$ is said to be restricted strong concave (RSC) with $m_R \in \mathbb{R}_+$ and restricted smooth (RSM) with $M_R \in \mathbb{R}_+$ on the set R, if*

$$-\frac{m_R}{2}\|h\|_2^2 \geq \mathcal{J}(\mathcal{S} \setminus L) - \mathcal{J}(\mathcal{S} \setminus (L \cup U)) - \langle \nabla - \mathcal{J}(\mathcal{S} \setminus L), h \rangle \geq -\frac{M_R}{2}\|h\|_2^2,$$

where $L \cup U \subseteq R$, and h is the indicator vector of the removed points in $L \cup U$ compared to L.

Lemma 3. *Let $\mathcal{S}$ be a fixed sample set and let $\sigma_w^2(\mathcal{S})$ denote the type II MAP estimate, that satisfies $\mathcal{D}(\sigma_w^2, \mathcal{S}, a, b) = 0$ from Proposition 3. Then the Hessian of $\sigma_w^2(\mathcal{S})$ w.r.t. the removal directions in h is given by*

$$\begin{aligned}\nabla_h^2 \sigma_w^2 =& -\frac{1}{T}(\nabla_h^2 \mathcal{D}(\sigma_w^2; y) + \nabla_{\sigma_w^2,h}^2 \mathcal{D}(\sigma_w^2; y)(\nabla_h \sigma_w^2)^\top \\ &+ \nabla_h \sigma_w^2 (\nabla_{\sigma_w^2,h}^2 \mathcal{D}(\sigma_w^2; y))^\top + \nabla_{\sigma_w^2} T((\nabla_h \sigma_w^2)(\nabla_h \sigma_w^2)^\top),\end{aligned}$$

where we define the following sub-parts of the derivatives as

$$\begin{aligned}\nabla_h \sigma_w^2 =& \frac{1}{T}(-\alpha^{(1)} \odot s^{(2)} - \alpha^{(2)} \odot s^{(1)} + \operatorname{diag}(M^{(2)})), \\ \nabla_{\sigma_w^2,h}^2 \mathcal{D}(\sigma_w^2; y) =& -(2\alpha^{(2)} \odot s^{(2)} + 2\alpha^{(1)} \odot s^{(3)} + 2\alpha^{(3)} \odot s^{(1)} - 2\operatorname{diag}(M^{(3)})), \\ \nabla_h^2 \mathcal{D}(\sigma_w^2; y) =& A^{-1} \odot (s^{(2)} s^{(1)\top} + s^{(1)} s^{(2)\top}) + A^{-2} \odot s^{(1)} s^{(1)\top} \\ &+ M^{(1)} \odot \big(s^{(2)} \alpha^{(1)\top} + \alpha^{(1)} s^{(2)\top} + s^{(1)} \alpha^{(2)\top} + \alpha^{(2)} s^{(1)\top} \\ &\qquad - \sigma_w^2 (\alpha^{(1)} \alpha^{(2)\top} + \alpha^{(2)} \alpha^{(1)\top}) + \alpha^{(1)} \alpha^{(1)\top}\big) \\ &+ M^{(2)} \odot \big(s^{(1)} \alpha^{(1)\top} + \alpha^{(1)} s^{(1)\top} - \sigma_w^2\, \alpha^{(1)} \alpha^{(1)\top}\big) \\ &+ K \odot (\alpha^{(1)} \alpha^{(2)\top} + \alpha^{(2)} \alpha^{(1)\top}) \\ &- 2M^{(1)} \odot M^{(2)} - K \odot A^{-2} + \sigma_w^2 (M^{(1)} \odot A^{-2} + M^{(2)} \odot A^{-1}) - M^{(1)} \odot A^{-1}, \\ \nabla_{\sigma_w^2} T =& 3\alpha^{(2)\top} \alpha^{(2)} - \operatorname{tr}(A^{-3}) + \frac{2(a-1)}{(\sigma_w^2)^3},\end{aligned}$$

using the short hand notations

$$\begin{aligned}&\alpha^{(p)} = A^{-p} y, \quad M^{(p)} = K A^{-p}, \quad s^{(p)} = K \alpha^{(p)}, \\ &T = -\alpha^{(1)\top} \alpha^{(2)} + \frac{1}{2}\operatorname{tr}(A^{-2}) - \frac{a-1}{(\sigma_w^2)^2},\end{aligned}$$

Proof. We obtained the first derivative from the implicit function theorem

$$\frac{\partial \mathcal{D}(\sigma_w^2; y)}{\partial \sigma_w^2} \frac{\partial \sigma_w^2}{\partial h_i} + \frac{\partial \mathcal{D}(\sigma_w^2; y)}{\partial h_i} = 0. \tag{82}$$

By differentiating the identity in Equation (82) w.r.t. h_j, we obtain

$$\frac{\partial}{\partial h_j}\left(\frac{\partial \mathcal{L}(\sigma_w^2; y)}{\partial \sigma_w^2} \frac{\partial \sigma_w^2}{\partial h_i} + \frac{\partial \mathcal{L}(\sigma_w^2; y)}{\partial h_i}\right) = \left(\frac{\partial^2 \mathcal{L}(\sigma_w^2; y)}{\partial \sigma_w^2 \partial h_j} + \frac{\partial^2 \mathcal{L}(\sigma_w^2; y)}{\partial^2 \sigma_w^2} \frac{\partial \sigma_w^2}{\partial h_j}\right) \frac{\partial \sigma_w^2}{\partial h_i} \tag{83}$$

$$+ \frac{\partial^2 \mathcal{L}(\sigma_w^2; y)}{\partial h_i \partial h_j} + \frac{\partial^2 \mathcal{L}(\sigma_w^2; y)}{\partial \sigma_w^2 \partial h_i} \frac{\partial \sigma_w^2}{\partial h_j} + \frac{\partial \mathcal{L}(\sigma_w^2; y)}{\partial \sigma_w^2} \frac{\partial^2 \sigma_w^2}{\partial h_i \partial h_j} = 0. \tag{84}$$

Reordering the terms yields the desired second derivative from the implicit function theorem

$$\frac{\partial^2 \sigma_w^2}{\partial h_i \partial h_j} = -\frac{1}{\frac{\partial \mathcal{D}(\sigma_w^2; y)}{\partial \sigma_w^2}}\left(\frac{\partial^2 \mathcal{D}(\sigma_w^2; y)}{\partial \sigma_w^2 \partial h_j} \frac{\partial \sigma_w^2}{\partial h_i} + \frac{\partial^2 \mathcal{D}(\sigma_w^2; y)}{\partial^2 \sigma_w^2} \frac{\partial \sigma_w^2}{\partial h_j} \frac{\partial \sigma_w^2}{\partial h_i}\right. \tag{85}$$

$$\left. + \frac{\partial^2 \mathcal{D}(\sigma_w^2; y)}{\partial h_i \partial h_j} + \frac{\partial^2 \mathcal{D}(\sigma_w^2; y)}{\partial \sigma_w^2 \partial h_i} \frac{\partial \sigma_w^2}{\partial h_j}\right). \tag{86}$$

We now differentiate each term to obtain the full Hessian. For convenience, we introduce the abbreviation $\alpha^{(p)} = A^{-p}y$. The derivatives $\frac{\partial \sigma_w^2}{\partial h_i}$ and $\frac{\partial \sigma_w^2}{\partial h_j}$ are given by the implicit function theorem in Proposition 3 as

$$\frac{\partial \sigma_w^2}{\partial h_i} = -\frac{\frac{\partial \mathcal{D}(\sigma_w^2;y)}{\partial h_i}}{\frac{\partial \mathcal{D}(\sigma_w^2;y)}{\partial \sigma_w^2}} = \frac{-\alpha_i^{(1)}(k_{x_i,\mathcal{S}}\alpha^{(2)}) - \alpha_i^{(2)}(k_{x_i,\mathcal{S}}\alpha^{(1)}) + k_{x_i,\mathcal{S}}(A^{-2})_{\cdot,i}}{-\alpha^{(1)\top}\alpha^{(2)} + \frac{1}{2}\operatorname{tr}(A^{-2}) - \frac{a-1}{(\sigma_w^2)^2}}, \tag{87}$$

$$\frac{\partial \sigma_w^2}{\partial h_j} = -\frac{\frac{\partial \mathcal{D}(\sigma_w^2;y)}{\partial h_j}}{\frac{\partial \mathcal{D}(\sigma_w^2;y)}{\partial \sigma_w^2}} = \frac{-\alpha_j^{(1)}(k_{x_j,\mathcal{S}}\alpha^{(2)}) - \alpha_j^{(2)}(k_{x_j,\mathcal{S}}\alpha^{(1)}) + k_{x_j,\mathcal{S}}(A^{-2})_{\cdot,j}}{-\alpha^{(1)\top}\alpha^{(2)} + \frac{1}{2}\operatorname{tr}(A^{-2}) - \frac{a-1}{(\sigma_w^2)^2}}, \tag{88}$$

The denominator is given by

$$\frac{\partial \mathcal{D}(\sigma_w^2;y)}{\partial \sigma_w^2} = -y^\top A^{-3}y + \frac{1}{2}\operatorname{tr}(A^{-2}) - \frac{a-1}{(\sigma_w^2)^2} \tag{89}$$

$$= -\alpha^{(1)\top}\alpha^{(2)} + \frac{1}{2}\operatorname{tr}(A^{-2}) - \frac{a-1}{(\sigma_w^2)^2}. \tag{90}$$

The second derivative of this term is given by

$$\frac{\partial^2 \mathcal{D}(\sigma_w^2;y)}{\partial^2 \sigma_w^2} = 3y^\top A^{-4}y - \operatorname{tr}(A^{-3}) + \frac{2(a-1)}{(\sigma_w^2)^3} \tag{91}$$

$$= 3\alpha^{(2)\top}\alpha^{(2)} - \operatorname{tr}(A^{-3}) + \frac{2(a-1)}{(\sigma_w^2)^3}. \tag{92}$$

Further, the cross derivatives $\frac{\partial^2 \mathcal{D}(\sigma_w^2;y)}{\partial \sigma_w^2 \partial h_j}$ and $\frac{\partial^2 \mathcal{D}(\sigma_w^2;y)}{\partial \sigma_w^2 \partial h_i}$ follow by differentiating $\frac{\partial \mathcal{D}}{\partial h_i} = \alpha_i^{(1)} s_i^{(2)} + \alpha_i^{(2)} s_i^{(1)} - \operatorname{diag}(M^{(2)})_i$ w.r.t. σ_w^2, using $\frac{\partial \alpha^{(p)}}{\partial \sigma_w^2} = -p\,\alpha^{(p+1)}$ and $\frac{\partial s^{(p)}}{\partial \sigma_w^2} = -p\,s^{(p+1)}$:

$$\frac{\partial^2 \mathcal{D}(\sigma_w^2;y)}{\partial \sigma_w^2 \partial h_i} = -2\alpha_i^{(2)}(k_{x_i,\mathcal{S}}\alpha^{(2)}) - 2\alpha_i^{(1)}(k_{x_i,\mathcal{S}}\alpha^{(3)}) - 2\alpha_i^{(3)}(k_{x_i,\mathcal{S}}\alpha^{(1)}) + 2k_{x_i,\mathcal{S}}(A^{-3})_{\cdot,i}, \tag{93}$$

$$\frac{\partial^2 \mathcal{D}(\sigma_w^2;y)}{\partial \sigma_w^2 \partial h_j} = -2\alpha_j^{(2)}(k_{x_j,\mathcal{S}}\alpha^{(2)}) - 2\alpha_j^{(1)}(k_{x_j,\mathcal{S}}\alpha^{(3)}) - 2\alpha_j^{(3)}(k_{x_j,\mathcal{S}}\alpha^{(1)}) + 2k_{x_j,\mathcal{S}}(A^{-3})_{\cdot,j}. \tag{94}$$

Finally, the term $\frac{\partial^2 \mathcal{D}(\sigma_w^2;y)}{\partial h_i \partial h_j}$ is given by

$$\frac{\partial^2 \mathcal{D}(\sigma_w^2;y)}{\partial h_i \partial h_j} = (A^{-1})_{ij}(k_{x_j,\mathcal{S}}\alpha^{(1)})(k_{x_i,\mathcal{S}}\alpha^{(2)}) + \alpha_j^{(1)}(A^{-1}k_{\mathcal{S},x_j})_i(k_{x_i,\mathcal{S}}\alpha^{(2)}) \tag{95}$$

$$+ \alpha_i^{(1)}(k_{x_i,\mathcal{S}}(A^{-1}_{\cdot,j}k_{x_j,\mathcal{S}}\alpha^{(2)} + A^{-1}k_{\mathcal{S},x_j}\alpha_j^{(2)})). \tag{96}$$

Next, we express all terms in matrix vector notation to obtain the full Hessian matrix. Therefore, we used the following additional definitions

$$M^{(p)} = KA^{-p}, \quad s^{(p)} = K\alpha^{(p)}, \quad T = \frac{\partial \mathcal{D}(\sigma_w^2;y)}{\partial \sigma_w^2}. \tag{97}$$

First, the gradient $\nabla_h \sigma_w^2$ follows from Equation (87) is given by

$$\nabla_h \sigma_w^2 = \frac{1}{T}(-\alpha^{(1)} \odot s^{(2)} - \alpha^{(2)} \odot s^{(1)} + \operatorname{diag}(M^{(2)})) \tag{98}$$

Further, the mixed terms $\nabla^2_{\sigma_w^2,h}\mathcal{D}(\sigma_w^2;y)$ from Equation (93) simplifies to

$$\nabla^2_{\sigma_w^2,h}\mathcal{D}(\sigma_w^2;y) = -(2\alpha^{(2)} \odot s^{(2)} + 2\alpha^{(1)} \odot s^{(3)} + 2\alpha^{(3)} \odot s^{(1)} - 2\operatorname{diag}(M^{(3)})). \tag{99}$$

And the term $\frac{\partial^2 \mathcal{D}(\sigma_w^2;y)}{\partial h_i \partial h_j}$ of Equation (95) is given by

$$\begin{aligned}
\nabla_h^2 \mathcal{D}(\sigma_w^2; y) =& A^{-1} \odot (s^{(2)} s^{(1)\top} + s^{(1)} s^{(2)\top}) + A^{-2} \odot s^{(1)} s^{(1)\top} && (100)\\
&+ M^{(1)} \odot \big(s^{(2)} \alpha^{(1)\top} + \alpha^{(1)} s^{(2)\top} + s^{(1)} \alpha^{(2)\top} + \alpha^{(2)} s^{(1)\top} && (101)\\
&\qquad - \sigma_w^2 (\alpha^{(1)} \alpha^{(2)\top} + \alpha^{(2)} \alpha^{(1)\top}) + \alpha^{(1)} \alpha^{(1)\top}\big) && (102)\\
&+ M^{(2)} \odot \big(s^{(1)} \alpha^{(1)\top} + \alpha^{(1)} s^{(1)\top} - \sigma_w^2\, \alpha^{(1)} \alpha^{(1)\top}\big) && (103)\\
&+ K \odot (\alpha^{(1)} \alpha^{(2)\top} + \alpha^{(2)} \alpha^{(1)\top}) && (104)\\
&- 2M^{(1)} \odot M^{(2)} - K \odot A^{-2} + \sigma_w^2 (M^{(1)} \odot A^{-2} + M^{(2)} \odot A^{-1}) - M^{(1)} \odot A^{-1}. && (105)
\end{aligned}$$

Finally, the last term as the gradient of the denominator $\nabla_{\sigma_w^2} T$ as

$$\nabla_{\sigma_w^2} T = 3\alpha^{(2)\top} \alpha^{(2)} - \mathrm{tr}(A^{-3}) + \frac{2(a-1)}{(\sigma_w^2)^3}. \tag{106}$$

So that we can write the final Hessian as

$$\begin{aligned}
\nabla_h^2 \sigma_w^2 =& -\frac{1}{T} (\nabla_h^2 \mathcal{D}(\sigma_w^2; y) + \nabla_{\sigma_w^2, h}^2 \mathcal{D}(\sigma_w^2; y) (\nabla_h \sigma_w^2)^\top && (107)\\
&+ \nabla_h \sigma_w^2 (\nabla_{\sigma_w^2, h}^2 \mathcal{D}(\sigma_w^2; y))^\top + \nabla_{\sigma_w^2} T ((\nabla_h \sigma_w^2)(\nabla_h \sigma_w^2)^\top).
\end{aligned}$$

□

Lemma 4. *Let $\mathcal{S} \subseteq \mathcal{X}$ and sample $x_i \in \mathcal{S}$. Consider the negated objective $-\mathcal{J}(\mathcal{S})$ and let the RKHS removal path of sample x_i be defined as $k(x, \mathcal{S}, h_i) = k_{x,\mathcal{S}} - h_i\, k(x, x_i) e_i, h_i \in [0, 1]$ from (Proposition 5). Then the Hessian of $-\mathcal{J}$ evaluated at $h = 0$ is given by*

$$\nabla_h^2 - \mathcal{J}(\mathcal{S})\big|_{h=0} = H^{\mathrm{epi}} + H^{\mathrm{ale}},$$

where the epistemic and aleatoric Hessian contributions are given by

$$\begin{aligned}
H^{\mathrm{epi}}(x) =& 2(\mathrm{diag}(c_x)(M^{(2)} + M^{(2)\top})\,\mathrm{diag}(c_x) - 2c_x c_x^\top \odot M^{(1)} - M^{(1)} \odot c_x c_x^\top)\\
H^{\mathrm{ale}}(x) =& \mathrm{mix}(\nabla_h \sigma_w^2)^\top + (\nabla_h \sigma_w^2)\mathrm{mix}^\top + 2(c_x \cdot c_x')(\nabla_h \sigma_w^2)(\nabla_h \sigma_w^2)^\top - (\|c_x\|^2 + 1)\nabla_h^2 \sigma_w^2,
\end{aligned}$$

with $\nabla_h \sigma_w^2$ and $\nabla_h^2 \sigma_w^2$ defined as in Lemma 3, the mixed partial vector

$$\mathrm{mix} = -(2\, c_x' \odot (s_x - k(\mathcal{S}, x)) + 2\, s_x' \odot c_x),$$

and the following abbreviations

$$c_x = A^{-1} k(\mathcal{S}, x), \quad c_x' = A^{-2} k(\mathcal{S}, x), \quad s_x = K c_x, \quad s_x' = K c_x',$$
$$M^{(1)} = K A^{-1}, \quad M^{(2)} = A^{-1} K A^{-1}.$$

Proof. We begin by revisiting the derivative of the negated objective function in Lemma 2

$$\begin{aligned}
\frac{\partial - \mathcal{J}(\mathcal{S})}{\partial h_i} =& -2(k_{x,\mathcal{S}'} A^{-1})_i (A^{-1} k_{\mathcal{S},x}) + 2(k_{x,\mathcal{S}} A^{-1})_i (k_{x_i,\mathcal{S}} A^{-1} k_{\mathcal{S},x}) && (108)\\
&- (k_{x,\mathcal{S}} A^{-2} k_{\mathcal{S},x} + 1) \frac{\partial \sigma_w^2}{\partial h_i}\\
=& -2k(x, x_i) e_i c_x + 2(c_x)_i^\top (k_{x_i,\mathcal{S}} c_x) - (\|c_x\|^2 + 1) \frac{\partial \sigma_w^2}{\partial h_i}. && (109)
\end{aligned}$$

We first derive the epistemic part, corresponding to the measure of epistemic influence $u_i(x)$:

$$u_i(x) = -2k(x, x_i) e_i c_x + 2(c_x)_i^\top (k_{x_i,\mathcal{S}} c_x) = 2(c_x)_i (k_{x_i,\mathcal{S}} c_x - k(x, x_i)). \tag{110}$$

Differentiating this expression w.r.t. h_j results in

$$\frac{\partial u_i(x)}{\partial h_j} = 2 \frac{\partial (c_x)_i}{\partial h_j} (k_{x_i,\mathcal{S}} c_x - k(x, x_i)) + 2(c_x)_i \left(k_{x_i,\mathcal{S}} \frac{\partial c_x}{\partial h_j} + \frac{\partial k_{x_i,\mathcal{S}}}{\partial h_j} c_x \right). \tag{111}$$

Where we differentiate the individual terms as

$$\frac{\partial c_x}{\partial h_j} = \frac{\partial A^{-1} k_{\mathcal{S},x}}{\partial h_j} = -A^{-1}k(x,x_j)e_j - A^{-1}(-e_j k_{x_j,\mathcal{S}} - k_{\mathcal{S},x_j} e_j^\top)c_x \tag{112}$$

$$= (A^{-1})_{\cdot,j}(k_{x_j,\mathcal{S}}c_x - k(x,x_j)) + (c_x)_j(A^{-1}k_{\mathcal{S},x_j}), \tag{113}$$

$$\frac{\partial k_{x_i,\mathcal{S}}}{\partial h_j} = -k(x_i,x_j)e_j. \tag{114}$$

For the further derivation, let us denote

$$\alpha_i \coloneqq k_{x_i,\mathcal{S}}c_x - k(x,x_i), \tag{115}$$

such that we obtain the following derivatives

$$\frac{\partial c_x}{\partial h_j} = (A^{-1})_{\cdot j}\alpha_j + (c_x)_j A^{-1}k_{\mathcal{S},x_j}, \qquad \frac{\partial \alpha_i}{\partial h_j} = -k(x_i,x_j)(c_x)_j + k_{x_i,\mathcal{S}}\frac{\partial c_x}{\partial h_j}. \tag{116}$$

Hence, if we reformulate the derivative of u_i w.r.t. α and substitute the derived terms, we obtain

$$\frac{\partial u_i(x)}{\partial h_j} = 2\frac{\partial (c_x)_i}{\partial h_j}\alpha_i + 2(c_x)_i\frac{\partial \alpha_i}{\partial h_j} = 2\Big((A^{-1})_{ij}\alpha_j\alpha_i \tag{117}$$

$$+ (c_x)_j(A^{-1}k_{\mathcal{S},x_j})_i\alpha_i - (c_x)_i k(x_i,x_j)(c_x)_j + (c_x)_i k_{x_i,\mathcal{S}}\frac{\partial c_x}{\partial h_j}\Big).$$

Using $(A^{-1})_{ij}\alpha_j = (M^{(2)})_{ij}(c_x)_j$, and $(A^{-1}k_{\mathcal{S},x_j})_i = (M^{(1)})_{ij}$, the second derivative is given by

$$\frac{\partial^2 - \mathcal{J}(\mathcal{S})}{\partial h_i\,\partial h_j} = 2\Big((c_x)_i(M^{(2)}_{ij} + M^{(2)}_{ji})(c_x)_j - 2(c_x)_i(c_x)_j M^{(1)}_{ij} - M^{(1)}_{ij}(c_x)_i(c_x)_j\Big). \tag{118}$$

Therefore, in matrix form,

$$H^{\text{epi}} = 2\Big(\operatorname{diag}(c_x)(M^{(2)} + M^{(2)\top})\operatorname{diag}(c_x) - 2\,c_x c_x^\top \odot M^{(1)} - M^{(1)} \odot c_x c_x^\top\Big). \tag{119}$$

For the aleatoric term, we write the negated objective as $-\mathcal{J} = F(h,\sigma_w^2(h)) - \sigma_w^2(h)$, where $F(h,s) = k(h)^\top(K(h)+sI)^{-1}k(h)$. Applying the total derivative decomposition to the second-order variation gives

$$\frac{\partial^2(-\mathcal{J})}{\partial h_i\,\partial h_j} = \frac{\partial^2 F}{\partial h_i \partial h_j}\bigg|_s + \frac{\partial^2 F}{\partial s\,\partial h_j}g_i + \frac{\partial^2 F}{\partial s\,\partial h_i}g_j + \frac{\partial^2 F}{\partial s^2}g_i g_j - \left(\|c_x\|^2 + 1\right)H_\sigma[i,j], \tag{120}$$

where $g = \nabla_h \sigma_w^2$ and $H_\sigma = \nabla_h^2 \sigma_w^2$. The first term yields H^{epi}. For the mixed partial we differentiate $\frac{\partial F}{\partial h_j}\big|_s = u_j(x)$ w.r.t. s, using $\frac{\partial c_x}{\partial s} = -c_x'$ and $\frac{\partial s_{x,j}}{\partial s} = -s_{x,j}'$:

$$\frac{\partial^2 F}{\partial s\,\partial h_j} = \frac{\partial u_j}{\partial s} = -2c_{x,j}'(s_{x,j} - k(\mathcal{S},x)_j) - 2c_{x,j}s_{x,j}' = \text{mix}_j. \tag{121}$$

The pure s-second derivative of F is

$$\frac{\partial^2 F}{\partial s^2} = 2\,k(\mathcal{S},x)^\top A^{-3}k(\mathcal{S},x) = 2\,c_x \cdot c_x'. \tag{122}$$

Collecting these contributions yields the aleatoric Hessian term

$$H^{\text{ale}} = \text{mix}(\nabla_h\sigma_w^2)^\top + (\nabla_h\sigma_w^2)\text{mix}^\top + 2(c_x\cdot c_x')(\nabla_h\sigma_w^2)(\nabla_h\sigma_w^2)^\top - (\|c_x\|^2+1)\nabla_h^2\sigma_w^2. \tag{123}$$

□

Lemma 5. *Let $H(\mathcal{S}^{(k)}) = H^{\text{epi}}(\mathcal{S}^{(k)}) + H^{\text{ale}}(\mathcal{S}^{(k)})$ be the Hessian of the negated objective $-\mathcal{J}$, as defined in Lemma 4 for the retained dataset $\mathcal{S}^{(k)}$ in episode k. Then for all possible removal subsets $T^{(k)} \subseteq T \coloneqq \mathcal{S} \setminus S' \subseteq R$, we denote by $H_R(\mathcal{S}^{(k)})$ the principal sub-matrix of the $H(\mathcal{S}^{(k)})$ on R. Then, the curvature bounds across all possible removal directions are obtained by*

$$m_R \coloneqq \inf_{\mathcal{S}^{(k)}} \lambda_{\min}(-H_R(\mathcal{S}^{(k)})), \qquad M_R \coloneqq \sup_{\mathcal{S}^{(k)}} \lambda_{\max}(-H_R(\mathcal{S}^{(k)})),$$

such that for all $\mathcal{S}^{(k)}$ and $T^{(k)} \subseteq T \subseteq R$, the projected Hessians $H_R(\mathcal{S}^{(k)})$ satisfy

$$-M_R I \le H_R(\mathcal{S}^{(k)}) \le -m_R I.$$

Proof. By definition of m_R and M_R, for every retained subset $\mathcal{S}^{(k)} \subseteq \mathcal{S}'$ the principal submatrix $H_R(\mathcal{S}^{(k)})$ satisfies

$$-M_R I \leq H_R(\mathcal{S}^{(k)}) \leq -m_R I, \tag{124}$$

which is precisely the final criterion in Lemma 5.

It remains to show that this criterion implies restricted strong concavity and smoothness (RSC/RSM). Let $L \subset L \cup U \subseteq R$ and let h denote the indicator vector of the removed coordinates in U relative to L. Since h is supported on $U \subseteq R$, the quadratic form satisfies $h^\top H(\mathcal{S}^{(k)})h = h^\top H_R(\mathcal{S}^{(k)})h$ for any retained dataset. The uniform spectral bound Equation (124) therefore says that $-\bar{\mathcal{J}}$, restricted to removal directions in R, is simultaneously m_R-strongly convex and M_R-smooth. By the standard first-order characterization of strong convexity and smoothness [52, Lemma 1.2.3], this is equivalent to

$$-\frac{M_R}{2}\|h\|_2^2 \;\leq\; \bar{\mathcal{J}}(L \cup U) - \bar{\mathcal{J}}(L) - \langle \nabla \bar{\mathcal{J}}(L),\, h\rangle \;\leq\; -\frac{m_R}{2}\|h\|_2^2, \tag{125}$$

which is the RSC/RSM condition for the outer objective. □

Lemma 6. *Let K be the Gram matrix and $\lambda_i(K)$ its eigenvectors, as well as $\|y\|_2^2$ the output norm. Given the Gamma prior distribution $p(\sigma_w^2(\mathcal{S})) = \frac{b^a}{\Gamma(a)}(\sigma_w^2(\mathcal{S}))^{a-1}e^{-b\sigma_w^2(\mathcal{S})}$ has parameter a, that satisfies for an admissible range of noise parameters $\sigma_w^2 \in [\xi_{\min}, \xi_{\max}]$ that*

$$a \geq 1 + \xi_{\max}^2 \left[\frac{\|y\|_2^2}{(\lambda_{\min}(K) + \xi_{\min})^3} - \frac{1}{2}\sum_{i=1}^{|\mathcal{S}|}(\lambda_i(K) + \xi_{\max})^{-2} \right]. \tag{126}$$

we obtain that the noise optimization is concave and has a unique optimal value $\sigma_w^{2*} \in [\xi_{\min}, \xi_{\max}]$.

Proof. We consider σ_w^2 as the optimal noise level, satisfying $\frac{\partial \mathcal{L}(\sigma_w^2, \mathcal{S}, a, b)}{\partial \sigma_w^2} = 0$. Then the second derivative of the optimality condition is given by

$$\frac{\partial^2 \mathcal{L}(\sigma_w^2, \mathcal{S}, a, b)}{\partial (\sigma_w^2)^2} = -y^\top A^{-3} y + \frac{1}{2}\mathrm{tr}(A^{-2}) - \frac{a-1}{(\sigma_w^2)^2}. \tag{127}$$

Strict concavity of the inner optimization problem holds across all retained data subsets $\mathcal{S}^{(k)}$ if

$$-y^\top A^{-3} y + \frac{1}{2}\mathrm{tr}(A^{-2}) - \frac{a-1}{(\sigma_w^2)^2} \leq 0 \iff a \geq 1 + (\sigma_w^2)^2 \left(y^\top A^{-3} y - \frac{1}{2}\mathrm{tr}(A^{-2}) \right), \tag{128}$$

where $A = K(\mathcal{S}, \mathcal{S}) + \sigma_w^2 I$ and we denote the eigenvalues of the Gram matrix K by λ_i. We bound

$$\mathrm{tr}(A^{-2}) = \sum_{i=1}^{|\mathcal{S}|}(\lambda_i + \sigma_w^2)^{-2} \geq \sum_{i=1}^{|\mathcal{S}|}(\lambda_i + \xi_{\max})^{-2}, \tag{129}$$

and obtain for the worst case noise parameter $(\lambda_i + \sigma_w^2) \geq (\lambda_i + \xi_{\min})$ that

$$y^\top A^{-3} y = \sum_{i=1}^{|\mathcal{S}|} \frac{(z_i^\top y)^2}{(\lambda_i + \sigma_w^2)^3} \leq \frac{\|y\|_2^2}{(\lambda_{\min}(K) + \xi_{\min})^3}, \tag{130}$$

where z_i are the eigenvectors of K. Therefore, with $(\sigma_w^2)^2 \leq \xi_{\max}^2$ both terms combined lead to

$$(\sigma_w^2)^2 \left(y^\top A^{-3} y - \frac{1}{2}\mathrm{tr}(A^{-2}) \right) \leq \xi_{\max}^2 \left[\frac{\|y\|_2^2}{(\lambda_{\min}(K) + \xi_{\min})^3} - \frac{1}{2}\sum_{i=1}^{|\mathcal{S}|}(\lambda_i + \xi_{\max})^{-2} \right]. \tag{131}$$

Thus, a sufficient condition for the concavity of the inner optimization problem is

$$a \geq 1 + \xi_{\max}^2 \left[\frac{\|y\|_2^2}{(\lambda_{\min}(K) + \xi_{\min})^3} - \frac{1}{2}\sum_{i=1}^{|\mathcal{S}|}(\lambda_i + \xi_{\max})^{-2} \right]. \tag{132}$$

□

Lemma 7. *Let us consider the gradient denominator term*

$$A = K + \sigma_w^2 I, \qquad D(\sigma_w^2) = -y^\top A^{-3} y + \frac{1}{2}\operatorname{tr}(A^{-2}) - \frac{a-1}{(\sigma_w^2)^2}.$$

Assume $\sigma_w^2 \in [\xi_{\min}, \xi_{\max}]$, and define as in Lemma 6

$$\delta_{\text{in}}(a) := \frac{a-1}{\xi_{\max}^2} - \frac{1}{2}\sum_{i=1}^{n}(\lambda_i + \xi_{\min})^{-2}, \qquad \underline{\lambda} := \lambda_{\min}(K) + \xi_{\min}, \qquad \Lambda := \lambda_{\max}(K).$$

Then $D(\sigma_w^2) \le -\delta_{\text{in}}(a) < 0$ provides inner stability (Lemma 6). For the outer optimization, let's define the following terms

$$\underline{d}(a) := \frac{2(a-1)}{\xi_{\max}^3} - \sum_{i=1}^{n}(\lambda_i + \xi_{\min})^{-3},$$

$$\beta_{\text{ale}}(a) \le \left(\frac{1+\bar{c}_k^2}{\delta_{\text{in}}(a)}\,\underline{d}(a) - 2\bar{c}'_{cc}\right)\|g_k\|_2^2 - \frac{2\bar{w}_k}{\delta_{\text{in}}(a)}\,\bar{g}_k - \frac{1+\bar{c}_k^2}{\delta_{\text{in}}(a)}\,\rho_R,$$

$$\eta_{\text{epi}} = \left(\sum_{i=1}^{n}\left|\frac{4\lambda_i}{(\lambda_i+\sigma_w^2)^2} - \frac{6\lambda_i}{\lambda_i+\sigma_w^2}\right|\right)\left(\sum_{j=1}^{n}\frac{|(U^\top k_{\mathcal{S},x})_j|}{\lambda_j+\sigma_w^2}\right)^2,$$

using the following bounds with $\kappa := \sup_{x\in\mathcal{A}}\|k_{\mathcal{S},x}\|_\infty$

$$\|g(x)\|_2 \le \frac{\bar{g}_k}{\delta_{\text{in}}(a)}, \quad \bar{g}_k := \frac{\Lambda}{\underline{\lambda}^2} + \frac{2\Lambda\|y\|_2^2}{\underline{\lambda}^3}, \quad \bar{c}_k := \frac{\kappa\sqrt{n}}{\underline{\lambda}}, \quad \bar{c}'_k := \frac{\kappa\sqrt{n}}{\underline{\lambda}^2}, \quad \bar{c}'_{cc} := \bar{c}_k\bar{c}'_k,$$

$$\bar{m}_k := 4\xi_{\max}\bar{c}_k\bar{c}'_k + 2\bar{c}_k^2, \quad \bar{p}_k := \frac{6\Lambda\|y\|^2}{\underline{\lambda}^4} + \frac{2\Lambda}{\underline{\lambda}^3}, \quad \bar{w}_k := \bar{m}_k + \frac{(1+\bar{c}_k^2)\,\bar{p}_k}{\delta_{\text{in}}(a)},$$

$$\rho_R := \frac{3\Lambda\|y\|^2}{\underline{\lambda}^3} + \frac{2\Lambda^2}{\underline{\lambda}^3} + \frac{2\Lambda}{\underline{\lambda}^2}.$$

Hence, restricted strong concavity is given for the removal directions if the prior parameter a satisfies

$$a > 1 : \delta_{\text{in}}(a) > 0,\ \underline{d}(a) > 0,\ \beta_{\text{ale}}(a) \ge \eta_{\text{epi}}.$$

Proof. From Lemma 4 and Lemma 6, the Hessians and the inner-stability bound are

$$\begin{aligned} H^{\text{epi}}(x) = 2\Bigg(&2\operatorname{diag}\big(A^{-1}k_{\mathcal{S},x}\big)A^{-1}KA^{-1}\operatorname{diag}\big(A^{-1}k_{\mathcal{S},x}\big) - 2(A^{-1}k_{\mathcal{S},x})(A^{-1}k_{\mathcal{S},x})^\top \\ &\odot (KA^{-1}) - (KA^{-1}) \odot (A^{-1}k_{\mathcal{S},x})(A^{-1}k_{\mathcal{S},x})^\top\Bigg), \end{aligned} \tag{133}$$

$$H^{\text{ale}}(x) = \operatorname{mix}(x)\,g^\top + g\,\operatorname{mix}(x)^\top + 2(c_x\cdot c'_x)\,gg^\top - (\|c_x\|^2+1)\,\nabla_h^2\sigma_w^2, \tag{134}$$

$$D(\sigma_w^2) = -y^\top A^{-3}y + \tfrac{1}{2}\operatorname{tr}(A^{-2}) - \tfrac{a-1}{(\sigma_w^2)^2} \le -\delta_{\text{in}}(a) < 0, \tag{135}$$

where $g = \nabla_h\sigma_w^2$. The identity $K = A - \sigma_w^2 I$ gives $Kc_x = k_{\mathcal{S},x} - \sigma_w^2 c_x$ (so $s_x - k_{\mathcal{S},x} = -\sigma_w^2 c_x$) and $Kc'_x = c_x - \sigma_w^2 c'_x$ (so $s'_x = c_x - \sigma_w^2 c'_x$). Substituting into the definition of mix from Lemma 4,

$$\operatorname{mix}_i = -(2c'_{x,i}(-\sigma_w^2 c_{x,i}) + 2(c_{x,i} - \sigma_w^2 c'_{x,i})c_{x,i}) = 4\sigma_w^2\,c'_{x,i}\,c_{x,i} - 2c_{x,i}^2, \tag{136}$$

i.e. $\operatorname{mix} = 4\sigma_w^2\,(c_x\odot c'_x) - 2\,(c_x\odot c_x)$. Expanding $-(\|c_x\|^2+1)\nabla_h^2\sigma_w^2$ via Lemma 3 and defining $w(x) := \operatorname{mix}(x) + \frac{\|c_x\|^2+1}{D}\nabla_{\sigma,h}^2\mathcal{D}$, we arrive at the three-term decomposition

$$H^{\text{ale}}(x) = T_1(x) + T_2(x) + T_3(x), \tag{137}$$

where

$$T_1(x) = w(x)\,g^\top + g\,w(x)^\top, \tag{138}$$

$$T_2(x) = \frac{\|c_x\|^2+1}{D}\,\nabla_h^2\mathcal{D}, \tag{139}$$

$$T_3(x) = \left[\frac{\|c_x\|^2+1}{D}\,\nabla_\sigma D + 2(c_x\cdot c'_x)\right]g\,g^\top. \tag{140}$$

Bound on T_1: The quadratic form satisfies $u^\top T_1 u = 2(u^\top w)(u^\top g)$. By Cauchy–Schwarz $|u^\top w| \le \|w\|$. We bound $\|w\| \le \|\text{mix}\| + \frac{1+\bar{c}_k^2}{|D|}\|\nabla^2_{\sigma,h}\mathcal{D}\|$. From Equation (136), using $\|c_x\|, \|c'_x\| \le \bar{c}_k, \bar{c}'_k$ and $\sigma_w^2 \le \xi_{\max}$,

$$\|\text{mix}\| \le 4\xi_{\max}\|c_x \odot c'_x\| + 2\|c_x\|^2 \le 4\xi_{\max}\bar{c}_k\bar{c}'_k + 2\bar{c}_k^2 =: \bar{m}_k. \tag{141}$$

For $\|\nabla^2_{\sigma,h}\mathcal{D}\|$, using $\|\alpha^{(p)}\| \le \|y\|/\underline{\lambda}^p$ and $\|s^{(p)}\| \le \Lambda\|y\|/\underline{\lambda}^p$, each component of the mixed derivative is bounded elementwise, giving $\|\nabla^2_{\sigma,h}\mathcal{D}\| \le \bar{p}_k$. With $|D| \ge \delta_{\text{in}}(a)$, we get $\|w\| \le \bar{w}_k$ and therefore

$$-u^\top T_1(x)u \ge -\frac{2\bar{w}_k}{\delta_{\text{in}}(a)}\,|u^\top g(x)|. \tag{142}$$

Bound on T_2: The matrix $\nabla^2_h\mathcal{D}$ from Lemma 3 is bounded termwise using $\|A^{-m}\| \le \underline{\lambda}^{-m}$, $\|K\| = \Lambda$, $\|A^{-p}y\| \le \|y\|/\underline{\lambda}^p$, $\|KA^{-p}y\| \le \Lambda\|y\|/\underline{\lambda}^p$, and $\|M^{(p)}\| = \|KA^{-p}\| \le \Lambda/\underline{\lambda}^p$. The dominant contributions are

$$\|M^{(1)} \odot \alpha^{(1)}\alpha^{(1)\top}\| \le \frac{\Lambda\|y\|^2}{\underline{\lambda}^3}, \quad \|K \odot (\alpha^{(1)}\alpha^{(2)\top} + \alpha^{(2)}\alpha^{(1)\top})\| \le \frac{2\Lambda\|y\|^2}{\underline{\lambda}^3}, \tag{143}$$

$$\|M^{(1)} \odot M^{(2)}\| \le \frac{\Lambda^2}{\underline{\lambda}^3}, \quad \|K \odot A^{-2}\| + \|M^{(1)} \odot A^{-1}\| \le \frac{2\Lambda}{\underline{\lambda}^2}. \tag{144}$$

Collecting all terms (including the remaining outer-product and cross terms, each bounded by $O(\Lambda^2\|y\|^2/\underline{\lambda}^4)$), we obtain

$$\|\nabla^2_h\mathcal{D}\|_{\text{op}} \le \rho_R := \frac{3\Lambda\|y\|^2}{\underline{\lambda}^3} + \frac{2\Lambda^2}{\underline{\lambda}^3} + \frac{2\Lambda}{\underline{\lambda}^2}, \tag{145}$$

and therefore

$$-u^\top T_2(x)u \ge -\frac{1+\bar{c}_k^2}{\delta_{\text{in}}(a)}\,\rho_R. \tag{146}$$

Bound on T_3: Since $D < 0$ and $\nabla_\sigma D = 3\|\alpha^{(2)}\|^2 - \text{tr}(A^{-3}) + \frac{2(a-1)}{(\sigma_w^2)^3} \ge \underline{d}(a)$, the term $\frac{\|c_x\|^2+1}{D}\nabla_\sigma D \le -\frac{1+\bar{c}_k^2}{\delta_{\text{in}}(a)}\underline{d}(a)$ is negative. The additional term $2(c_x \cdot c'_x) \ge 0$ partially offsets this, with $c_x \cdot c'_x = k_{\mathcal{S},x}^\top A^{-3}k_{\mathcal{S},x} \le \bar{c}'_{cc}$. Hence,

$$-u^\top T_3(x)u \ge \left(\frac{1+\bar{c}_k^2}{\delta_{\text{in}}(a)}\,\underline{d}(a) - 2\bar{c}'_{cc}\right)(u^\top g(x))^2. \tag{147}$$

Summing the three contributions of T_1, T_2, T_3 and using $|u^\top g| \le \|g\| \le \bar{g}_k/\delta_{\text{in}}(a)$ yields

$$-u^\top H^{\text{ale}}(x)u \ge \left(\frac{1+\bar{c}_k^2}{\delta_{\text{in}}(a)}\,\underline{d}(a) - 2\bar{c}'_{cc}\right)(u^\top g(x))^2 - \frac{2\bar{w}_k}{\delta_{\text{in}}(a)}\,|u^\top g(x)| - \frac{1+\bar{c}_k^2}{\delta_{\text{in}}(a)}\,\rho_R. \tag{148}$$

From the spectral representation derived earlier, we have

$$u^\top H^{\text{epi}}(x)u \le \eta_{\text{epi}} = \left(\sum_{i=1}^n \left|\frac{4\lambda_i}{(\lambda_i+\sigma_w^2)^2} - \frac{6\lambda_i}{\lambda_i+\sigma_w^2}\right|\right)\left(\sum_{j=1}^n \frac{|(U^\top k_{\mathcal{S},x})_j|}{\lambda_j+\sigma_w^2}\right)^2. \tag{149}$$

For concavity along all u, it suffices that

$$\eta_{\text{epi}} \le \beta_{\text{ale}}(a) := \left(\frac{1+\bar{c}_k^2}{\delta_{\text{in}}(a)}\,\underline{d}(a) - 2\bar{c}'_{cc}\right)\|g_k\|_2^2 - \frac{2\bar{w}_k}{\delta_{\text{in}}(a)}\,\bar{g}_k - \frac{1+\bar{c}_k^2}{\delta_{\text{in}}(a)}\,\rho_R. \tag{150}$$

Using the bounds derived above yields the stated expression for $\beta_{\text{ale}}(a)$. If

$$\delta_{\text{in}}(a) > 0, \quad \underline{d}(a) > 0, \quad \beta_{\text{ale}}(a) \ge \eta_{\text{epi}}, \tag{151}$$

then for all $x \in \mathcal{A}$ and all removal directions u,

$$u^\top(H^{\text{epi}}(x) + H^{\text{ale}}(x))u \le 0, \tag{152}$$

which proves restricted strong concavity of the outer objective.

□

Lemma 8. *Suppose that the negated objective $-\mathcal{J}(\cdot)$ is restricted strongly concave on T, as defined in Definition 1, then the difference in the objective for two considered subsets of T is bounded by*

$$\mathcal{J}(\mathcal{S} \setminus L) - \mathcal{J}(\mathcal{S} \setminus (L \cup U)) \leq \frac{1}{2m_R} \|\nabla - \mathcal{J}(\mathcal{S} \setminus L)_U\|_2^2$$

where the sets satisfy $L \subset L \cup U \subseteq T$.

Proof. Let us begin with the restricted strong concavity criterion

$$-\frac{m_R}{2}\|h\|_2^2 \geq \mathcal{J}(\mathcal{S} \setminus L) - \mathcal{J}(\mathcal{S} \setminus (L \cup U)) - \langle \nabla - \mathcal{J}(\mathcal{S} \setminus L), h \rangle. \tag{153}$$

By bringing the inner product of the gradient to the other side, we obtain

$$\langle \nabla - \mathcal{J}(\mathcal{S} \setminus L), h \rangle - \frac{m_R}{2}\|h\|_2^2 \geq \mathcal{J}(\mathcal{S} \setminus L) - \mathcal{J}(\mathcal{S} \setminus (L \cup U)), \tag{154}$$

which we can further upper bound by the best possible subset

$$\max_h \langle \nabla - \mathcal{J}(\mathcal{S} \setminus L), h \rangle - \frac{m_R}{2}\|h\|_2^2 \geq \mathcal{J}(\mathcal{S} \setminus L) - \mathcal{J}(\mathcal{S} \setminus (L \cup U)), \tag{155}$$

so that using $h = L + \frac{1}{m_R}\nabla - \mathcal{J}(\mathcal{S} \setminus L)_U$, where $\nabla - \mathcal{J}(\mathcal{S} \setminus L)_U$ is the gradient in all directions in U.

$$\mathcal{J}(\mathcal{S} \setminus L) - \mathcal{J}(\mathcal{S} \setminus (L \cup U)) \leq \frac{1}{2m_R} \|\nabla - \mathcal{J}(\mathcal{S} \setminus L)_U\|_2^2 \tag{156}$$

□

Lemma 9. *Suppose that the set-function $\bar{\mathcal{J}}(\cdot)$ for the negated restricted strongly concave objective $-\mathcal{J}$ on R, as defined in Definition 1, then the increment in the objective for the removal along the maximal gradient according to* F2I *for points in R, is related to removing the optimal set T^* by*

$$\bar{\mathcal{J}}(\mathcal{S}^{(k+1)}) - \bar{\mathcal{J}}(\mathcal{S}^{(k)}) \geq \frac{m_R}{(|R| - |\mathcal{S} \setminus \mathcal{S}^{(k)}|)M_R}(\bar{\mathcal{J}}(T^*) - \bar{\mathcal{J}}(\mathcal{S} \setminus \mathcal{S}^{(k)})).$$

Proof. We again begin this proof with the strong restricted concavity on R from Definition 1.

$$\mathcal{J}(\mathcal{S}^{(k)}) - \mathcal{J}(\mathcal{S}^{(k+1)}) - \langle \nabla - \mathcal{J}(\mathcal{S}^{(k)}), h \rangle \geq -\frac{M_R}{2}\|h\|_2^2. \tag{157}$$

The removal direction h is chosen according to the largest gradient in F2I (Algorithm 1) $h = \alpha e_j$, where it's the j-th sample that is selected for removal.

$$\mathcal{J}(\mathcal{S}^{(k)}) - \mathcal{J}(\mathcal{S}^{(k+1)}) \geq \langle \nabla - \mathcal{J}(\mathcal{S}^{(k)}), \alpha e_j \rangle - \frac{M_R}{2}\alpha^2 \tag{158}$$

$$= \alpha \|\nabla - \mathcal{J}(\mathcal{S}^{(k)})\|_\infty - \frac{M_R}{2}\alpha^2. \tag{159}$$

We replace the inner product with the infinity norm, due to removing the next sample with the largest gradient. Further, we obtain for $\alpha = \frac{\|\nabla - \mathcal{J}(\mathcal{S}^{(k)})\|}{M_R}$ the tight bound

$$\mathcal{J}(\mathcal{S}^{(k)}) - \mathcal{J}(\mathcal{S}^{(k+1)}) \geq \frac{1}{2M_R}\|\nabla - \mathcal{J}(\mathcal{S}^{(k)})\|_\infty^2. \tag{160}$$

Next, we can lower bound this further using the optimal removal set $T^* = \mathcal{S} \setminus \mathcal{S}^* \subseteq R$. Let us therefore define $S^R = T^* \setminus (\mathcal{S} \setminus \mathcal{S}^{(k)})$, which is the set of samples that are in the optimal removal set but have not yet been removed. We can bound the infinity norm using the mean ℓ_2 norm among all of the candidates in this remaining set:

$$\mathcal{J}(\mathcal{S}^{(k)}) - \mathcal{J}(\mathcal{S}^{(k+1)}) \geq \frac{1}{2|S^R|M_R} \sum_{j \in S^R} \langle \nabla - \mathcal{J}(\mathcal{S}^{(k)}), e_j \rangle^2 \tag{161}$$

$$= \frac{1}{2|S^R|M_R} \|\nabla - \mathcal{J}(\mathcal{S}^{(k)})_{S^R}\|_2^2. \tag{162}$$

Substituting Lemma 8 into Equation (161), yields:

$$\mathcal{J}(\mathcal{S}^{(k)}) - \mathcal{J}(\mathcal{S}^{(k+1)}) \geq \frac{m_R}{|S^R| M_R}(\mathcal{J}(\mathcal{S}^{(k)}) - \mathcal{J}(S^R \cup (\mathcal{S}^{(k)}))) \tag{163}$$

$$\geq \frac{m_R}{(|R| - |\mathcal{S} \setminus \mathcal{S}^{(k)}|) M_R}(\mathcal{J}(\mathcal{S}^{(k)}) - \mathcal{J}(\mathcal{S} \setminus T^*)). \tag{164}$$

From the definition of the set-function in Lemma 1, we obtain that

$$\bar{\mathcal{J}}(\mathcal{S}^{(k+1)}) - \bar{\mathcal{J}}(\mathcal{S}^{(k)}) \geq \frac{m_R}{(|R| - |\mathcal{S} \setminus \mathcal{S}^{(k)}|) M_R}(\bar{\mathcal{J}}(T^*) - \bar{\mathcal{J}}(\mathcal{S} \setminus \mathcal{S}^{(k)})). \tag{165}$$

□

Theorem 1. *Let the set-function $\bar{\mathcal{J}}$ be defined as $\bar{\mathcal{J}}(T) = \mathcal{J}(\mathcal{S}) - \mathcal{J}(\mathcal{S} \setminus T)$ for a compact $\mathcal{A}$ and bounded, twice continuously differentiable kernel k. If the inner MAP II for σ_w^2 has a unique solution and $-\mathcal{J}$ is m_R restricted strong concave and M_R restricted smooth for all potential removals $R \supseteq T$ (Definition 1), satisfying Lemma 7, then approximate submodularity of $\bar{\mathcal{J}}$ guarantees*

$$\bar{\mathcal{J}}(\mathcal{S}') \geq (1 - e^{-\gamma}) \max_{\substack{\mathcal{S}^* \subseteq \mathcal{S} \\ |\mathcal{S}^*| = |S'|}} \bar{\mathcal{J}}(\mathcal{S}^*), \quad \textit{where } \gamma = \tfrac{m_R}{M_R}.$$

Proof. In Lemma 2, we have shown that the point selection in F2I, according to $-u(x) - v(x)$ has been along the maximal gradient of the complement forward process $\bar{\mathcal{J}}$ according to Lemma 1. Therefore, given that our objective satisfies the restricted strong convexity and smoothness in Definition 1, with the Hessian in Lemma 4 we bound the k-th incremental greedy removal along the maximum gradient $-u(x) - v(x)$, with the incremental gain from adding the optimal set $(\bar{\mathcal{J}}(T^*) - \bar{\mathcal{J}}(\mathcal{S} \setminus \mathcal{S}^{(k)}))$ by Lemma 9:

$$\bar{\mathcal{J}}(\mathcal{S} \setminus \mathcal{S}^{(k+1)}) - \bar{\mathcal{J}}(\mathcal{S} \setminus \mathcal{S}^{(k)}) \geq \frac{m_R}{(|R| - |\mathcal{S} \setminus \mathcal{S}^{(k)}|) M_R}(\bar{\mathcal{J}}(T^*) - \bar{\mathcal{J}}(\mathcal{S} \setminus \mathcal{S}^{(k)})). \tag{166}$$

Further, we observe analogous to the proof of [18, Theorem 3], that

$$\bar{\mathcal{J}}(\mathcal{S} \setminus \mathcal{S}^{(k+1)}) - \bar{\mathcal{J}}(\mathcal{S} \setminus \mathcal{S}^{(k)}) = (\bar{\mathcal{J}}(T^*) - \bar{\mathcal{J}}(\mathcal{S} \setminus \mathcal{S}^{(k)})) - (\bar{\mathcal{J}}(T^*) - \bar{\mathcal{J}}(\mathcal{S} \setminus \mathcal{S}^{(k+1)})), \tag{167}$$

so that we can rewrite using Equation (166) from Lemma 9

$$(\bar{\mathcal{J}}(T^*) - \bar{\mathcal{J}}(\mathcal{S} \setminus \mathcal{S}^{(k+1)})) = (\bar{\mathcal{J}}(T^*) - \bar{\mathcal{J}}(\mathcal{S} \setminus \mathcal{S}^{(k)})) - (\bar{\mathcal{J}}(\mathcal{S} \setminus \mathcal{S}^{(k+1)}) - \bar{\mathcal{J}}(\mathcal{S} \setminus \mathcal{S}^{(k)})) \tag{168}$$

$$(\bar{\mathcal{J}}(T^*) - \bar{\mathcal{J}}(\mathcal{S} \setminus \mathcal{S}^{(k+1)})) \overset{166}{\leq} \left(1 - \frac{m_R}{(|R| - |\mathcal{S} \setminus \mathcal{S}^{(k)}|) M_R}\right)(\bar{\mathcal{J}}(T^*) - \bar{\mathcal{J}}(\mathcal{S} \setminus \mathcal{S}^{(k)})). \tag{169}$$

Hence, from this follows that we can bound the k-th optimality gap with

$$(\bar{\mathcal{J}}(T^*) - \bar{\mathcal{J}}(\mathcal{S} \setminus \mathcal{S}^{(k)})) \leq \left(1 - \frac{m_R}{(|R| - |\mathcal{S} \setminus \mathcal{S}^{(k)}|) M_R}\right)^k (\bar{\mathcal{J}}(T^*) - \bar{\mathcal{J}}(\mathcal{S} \setminus \mathcal{S}^{(0)})). \tag{170}$$

Hence, we obtain according to the exponential inequality that

$$\bar{\mathcal{J}}(\mathcal{S} \setminus \mathcal{S}^{(k)}) \geq (\bar{\mathcal{J}}(T^*) - \bar{\mathcal{J}}(\mathcal{S} \setminus \mathcal{S}^{(k)}))(1 - (1 - C)^k) \tag{171}$$

$$\geq \left(1 - e^{-\frac{m_R}{M_R}}\right) \max_{\substack{T \subseteq \mathcal{S} \\ |T| = |\mathcal{S} \setminus \mathcal{S}^{(k)}|}} \bar{\mathcal{J}}(T). \tag{172}$$

□

M Experimental Validation of Concavity

In this section, we demonstrate that our theoretical derivations match the empirically observed behavior of the removal process. Therefore, we investigate the curvature of the negated objective $-\mathcal{J}(T)$, restricted to the removal directions T. We have derived in Lemma 2 that $-u - v$ is the gradient direction for this optimization problem. Hence, the curvature of the *maximization* problem has to be *concave* in the removal direction to achieve an improvement in the uncertainty. All possible directions in $\mathcal{S}$, may not be concave, however we are only interested into the points that we remove

$T \subseteq R$. In each removal step k one more sample of T is removed from $\mathcal{S}$, resulting in $\mathcal{S}^{(k)}$ until the algorithm terminates and returns $\mathcal{S}' = \mathcal{S} \setminus T$. To verify the curvature condition of Lemma 5 empirically, we evaluate the principal submatrix $H_R(\mathcal{S}^{(k)})$ of

$$H^{(k)} = H^{\mathrm{epi},(k)} + H^{\mathrm{ale},(k)} \tag{173}$$

restricted to the indices $R_k = T \cap \mathcal{S}^{(k)}$ of all removal candidates still present at step k. We then track the *restricted curvature* at each step via the scalar curvature $H^{(k)}_{R_k}[i_k, i_k]$ of the point currently being removed, and additionally report the spectral bounds

$$\lambda_{\min}\left(H_{R_k}(\mathcal{S}^{(k)})\right) \leq \lambda_{\max}\left(H_{R_k}(\mathcal{S}^{(k)})\right) \tag{174}$$

which directly correspond to $-M_R$ and $-m_R$ in Lemma 5. We demonstrate these quantities in Figure 27 for the provided Forgetting to Improve examples in Figs. 1, 7 and 8. We observe that the curvature bounds depend on the prior parameter a, as described in Lemma 7. Thereby, larger prior parameter a lead to more stable curvature behavior throughout the removal process. While for the Dow Jones as well as the asymmetric task the restricted curvatures are clearly convex for $a = 1$, which corresponds to having no prior, increasing the prior parameter a improves concavity by stabilizing the aleatoric curvature.

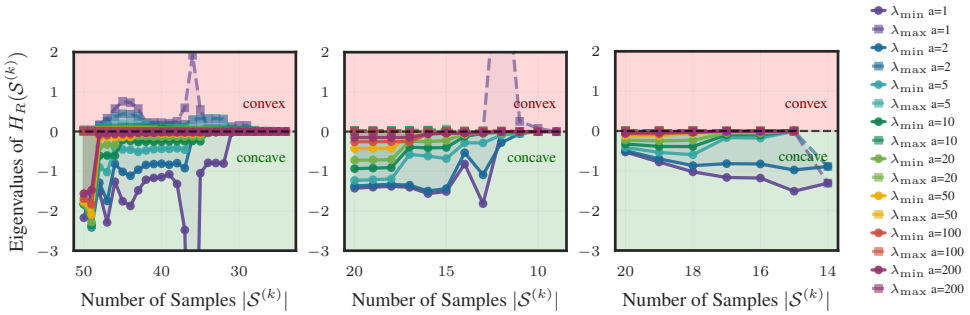


Figure 27: Visualization of the restricted curvature bounds (Equation (174)) for Forgetting to Improve on the Dow Jones data set in Figure 1, the asymmetric test-function in Figure 8, and the symmetric test-function in Figure 7.