THE UNIVERSITY of EDINBURGH

Afonso Eduardo

November 2022

1 Active Learning Strategies for Optimization and Inference

Active learning plays a key role in optimization under uncertainty. In particular, the Expected Improvement (EI) [1] and the Upper/Lower Confidence Bound (UCB/LCB) [2, 3] enjoy wide popularity among machine learning practitioners due to their computational simplicity. However, these two strategies are also susceptible to local minima.

Despite being governed by a parameter that promotes exploration, careful tuning of UCB/LCB is difficult. For instance, the theoretical rules derived by Srinivas et al. [3] and Chowdhury et al. [2], based on regret bounds, postulate certain conditions that rarely occur in practice; e.g., that global optima of UCB/LCB can be found, the objective function belongs to a known Reproducing Kernel Hilbert Space (RKHS), or a bound on the RKHS norm is known. More recently, Berkenkamp et al. [4] and Bogunovic et al. [5] propose modifications to handle unknown (hyper)parameters or model misspecification, but not without their own limitations, e.g. stationarity or known error (ϵ-misspecification). Fundamentally, if the value of the parameter in UCB/LCB increases rapidly, the acquisition step can become very exploratory, also compromising sample efficiency.

As for EI, it favors the point that offers the greatest expected improvement upon a threshold, typically the value associated with the incumbent solution.[1] This greedy strategy has empirically been shown to be more efficient than UCB/LCB when properties of the objective function are unknown [e.g. 6, 2, 7]. As with UCB/LCB, some modifications have also been proposed [see e.g. 8, 9]. Importantly, I have proposed the Collapsed Expected Improvement (CEI, Appendix A) that chooses more informative points by design. In CEI, I target one mode at a time and collapse them to approximate the remaining, provided that the actively sampled point is not sufficiently informative.[2] The same idea can be applied to other acquisition functions. In fact, the initial aim is to investigate in more detail these mode-collapsed strategies, comparing them with alternatives, and possibly to develop a more theoretical analysis.

Furthermore, the usefulness of active learning strategies extends to (likelihood-free) inference, whose goal is to identify regions of non-negligible density and to correctly estimate regions of interest, typically those of highest density; that is, once optima are identified, the objective is then to efficiently explore modes. Previous work has used random perturbations [e.g. stochastic LCB 11] to explore these modes, not taking into account the amount of information gained. The initial aim is again to investigate mode-collapsed strategies, comparing them with posterior sampling, e.g. with MCMC [12], and possibly other more sample-efficient but costlier strategies for LFI. For instance, strategies based on the variance in the posterior approximation [maxvar/expintvar 13] or entropy-based strategies, such as Bayesian Active Learning by Disagreement [BALD 14, MaxMI 15] where active samples are the points for which the parameters (of the conditional density model) under the posterior disagree the most.

[1] Technically, it is a parameter that can be controlled, allowing e.g. relaxation, which I explore in CEM-EI (Appendix B).

[2] It can be seen as an extension of the Top-Two Expected Improvement (TTEI) proposed by Qin et al. [10], generalized to handle continuous spaces and to select the top "arm" (mode) that meets a certain condition based on information gain. By contrast, in TTEI, "the idea is to identify in each period the two most promising arms based on current observations, and randomize to choose which to sample. A tuning parameter [...] controls the probability assigned to the top arm."

2 Search-Aware Bayesian Optimization

2.1 Motivation

Bayesian Optimization depends on the choice of surrogate models and acquisition functions. While surrogates can heavily influence the acquisition step [e.g., 16], the decision itself requires finding the optimum of an acquisition function over the domain or, alternatively, a trust region [e.g., 17]. The resulting optimization problem is typically nonconvex, being a computational challenge not only in high dimensions, but possibly also in lower-dimensional domains. For instance, despite a proven record of good empirical performance [e.g., 6, 2, 7], the Expected Improvement (`EI`) landscape is often multimodal and characterized by flat regions, severely hampering the effectiveness of gradient-based optimizers.

As shown in [Figure 11 1], the nonconvex issues that affect `EI`[3] have been known since its inception, leading Jones et al. [1] to propose a branch-and-bound algorithm, followed by a Lipschitz global optimizer [`DIRECT` 18]. Despite the popularity of the latter, the curse of dimensionality limits its applicability. An alternative is then to combine local optimization with global heuristics. In this setting, a standard practice is to adopt gradient-based optimization with multiple restarts, where initial points are chosen from a large pool of candidates [e.g., 19, 20, 16]. Importantly, the cost of the acquisition step typically dominates, without providing guarantees that acquisition points are global optima. As a side effect, theoretical regret bounds and convergence rates have also limited applicability [see, e.g., Appendix A 16].[4]

2.2 Research Questions

Q1: Is it possible to design acquisition strategies that are more amenable to high-dimensional optimization without compromising sample efficiency?

For this purpose, it may be useful to note that optimization can be seen as the estimation of a posterior distribution over optima $p(\boldsymbol{x}^\star|\mathcal{D}_n)$, where $\mathcal{D}_n$ is the evidence set containing previous observations $\{(\boldsymbol{x}_i, y_i)\}_{i\leq n}$. In fact, the problem can be cast as an expectation over a search model $q^\star = \arg\max_q \mathbb{E}_{\boldsymbol{x}\sim q}[\mathbb{E}_{y\sim f(\boldsymbol{x})}[u(y)]]$, where f is the target to be approximated by $p(y|\boldsymbol{x})$ from a surrogate and u is a utility. The integrated utility then becomes an acquisition function $\alpha(\boldsymbol{x})$. However, in BO, the search model is implicit.[5]

Q2: Can the acquisition cost be amortized if an explicit search model is maintained during optimization?

Perhaps the search model can be used to sample more densely from the highest density regions without having to solve a high-dimensional optimization problem at each acquisition step.

2.3 BO with Adaptive Sampling (CEM-PI)

Given a set of desired properties $S_\tau = \{y : y \leq \tau\}$, the original optimization problem, in which we can sample $y \sim f(\boldsymbol{x})$, may be seen as a rare event estimation problem that is cast as an expectation over a search model,

$$\max_{\boldsymbol{x}} \mathbb{P}(Y \in S_\tau \mid \boldsymbol{x}) \geq \max_{\boldsymbol{\theta}} \mathbb{E}_{q_{\boldsymbol{\theta}}}[\mathbb{P}(Y \in S_\tau \mid \boldsymbol{x})]. \tag{2.1}$$

Then, a surrogate provides $p(y|\boldsymbol{x})$ that approximates $\mathbb{P}(Y \in S_\tau \mid \boldsymbol{x}) \approx \int p(y|\boldsymbol{x}) \mathbb{1}_{y\leq\tau} dy = \mathrm{CDF}_Y(\boldsymbol{x}, \tau)$.[6]

In [`CbAS` 21], the goal is to estimate $p(\boldsymbol{x} \mid S_\tau, \boldsymbol{\theta}_0) \overset{\propto}{\sim} \mathrm{CDF}_Y(\boldsymbol{x}, \tau) q(\boldsymbol{x}; \boldsymbol{\theta}_0)$.[7] By minimizing the (forward) KL divergence $\mathrm{KL}(p(\boldsymbol{x} \mid S_\tau, \boldsymbol{\theta}_0) \parallel q(\boldsymbol{x}; \boldsymbol{\theta}))$, the objective becomes $\boldsymbol{\theta}^\star = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{q(\boldsymbol{x};\, \boldsymbol{\theta}_0)} [\mathrm{CDF}_Y(\boldsymbol{x}, \tau) \log q(\boldsymbol{x}; \boldsymbol{\theta})]$. However, since $\mathrm{CDF}_Y(\boldsymbol{x}, \tau)$ can be vanishingly small for $\boldsymbol{x} \sim q_{\boldsymbol{\theta}_0}$ and small τ, the method[8] relies on adaptive importance sampling by iterating according to

$$\boldsymbol{\theta}_{t+1} = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{q(\boldsymbol{x};\, \boldsymbol{\theta}_t)} \left[\frac{q(\boldsymbol{x}; \boldsymbol{\theta}_0)}{q(\boldsymbol{x}; \boldsymbol{\theta}_t)} \mathrm{CDF}_Y(\boldsymbol{x}, \tau_t) \log q(\boldsymbol{x}; \boldsymbol{\theta}) \right], \tag{2.2}$$

[3] Other acquisition functions can similarly lead to nonconvex optimization problems [e.g., `UCB`/`LCB` 3].

[4] Another assumption that may not hold in practice is related to concentration inequalities [see, e.g., Figure 1 16].

[5] Technically, it is given by a point-mass distribution (Dirac delta).

[6] This is an acquisition function $\alpha(\boldsymbol{x}; \tau) = \mathrm{CDF}_Y(\boldsymbol{x}, \tau)$. In fact, it is the same as the Probability of Improvement (PI) with utility $u(y) = \mathbb{1}_{y\leq\tau}$. The threshold is given by τ, not necessarily the best observed value $y_{\min}$.

[7] Technically, `CbAS` performs maximization with $1 - \mathrm{CDF}_Y(\boldsymbol{x}, \tau)$.

[8] It is equivalent to the cross-entropy method [`CEM` 22] for optimization with performance function $\mathrm{CDF}_Y(\boldsymbol{x}, \tau)$.

where the condition becomes increasingly stringent $\tau_{t-1} \geq \tau_t \geq y_{\min}$. This is essentially an iterative weighted maximum likelihood method that updates parameters by sampling from $q(\boldsymbol{x}; \boldsymbol{\theta}_t)$ [Appendix S6.1. 21]. As a result, closed-form updates are available if $q_{\boldsymbol{\theta}}$ belongs to the exponential family, thereby avoiding gradient computation.

Importantly, `CbAS` has been proposed for the fixed data setting. For BO purposes, the current initial implementation runs `CbAS` at each new acquisition step, starting from a multivariate Gaussian $q_{\boldsymbol{\theta}_0}$ with mean given by the incumbent solution $\boldsymbol{x}_{\text{best}}$. Several issues were found during initial development. For instance, sampling from the search model to acquire data may lead to no improvement, even in relatively low-dimensional problems (<10). For this reason, the current strategy acquires the mean of the final search model $q_{\boldsymbol{\theta}}$ at each acquisition step, unless it is already in the training set. Meanwhile, small importance weights $q(\boldsymbol{x}; \boldsymbol{\theta}_0)/q(\boldsymbol{x}; \boldsymbol{\theta}_t)$ may cause early termination of `CbAS`, often resulting in poor performance, as shown in Figure 2.1 (all `AS` without suffix `-IW`). Variants without these search weights (i.e., with `-IW`), or equivalently with $q_{\boldsymbol{\theta}_0} = q_{\boldsymbol{\theta}_t}$, correspond to the update given by `DbAS` [Equation 12 21]. The nondecreasing thresholds τ may also need to be reset at each new acquisition step.

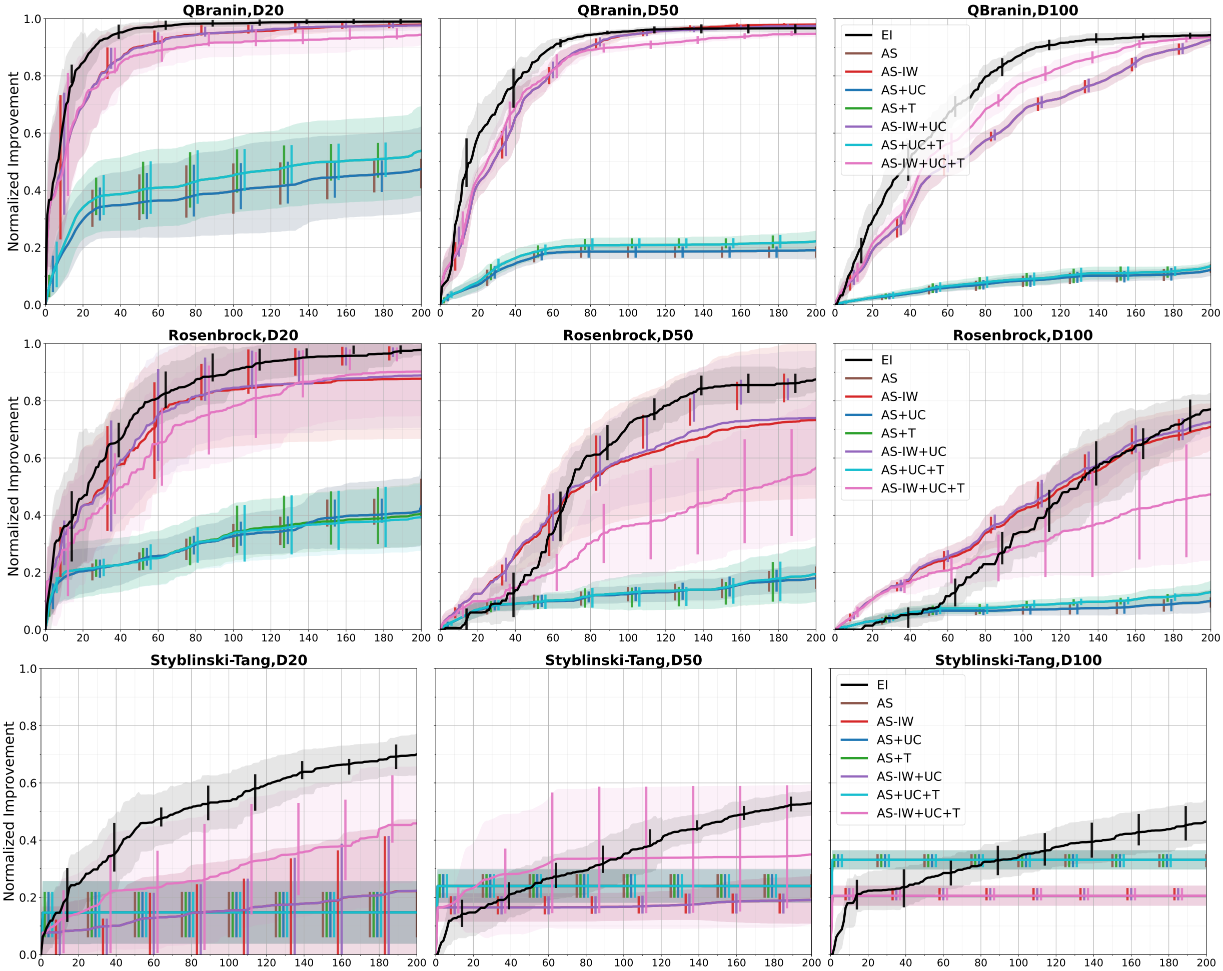

Figure 2.1. Performance of acquisition strategies, `EI` vs `AS` (and variants), on objectives ranging from 20 to 100 dimensions [Appendix B.3 16]. The total budget is 200 acquisitions, excluding initial observations. Solid curves and shaded regions represent the normalized improvement mean and one standard deviation over 10 trials, each with different initial conditions/observations. Solid vertical lines show the interquartile range. The type of surrogate model is the same in all experiments [`I+XA` 16]. Search models in `AS` are multivariate Gaussians. Variants `-IW` significantly outperform others, whereas unbiased estimates `+UC` only lead to marginal gains and nondecreasing thresholds `+T` can negatively affect performance. Overall, `AS` requires further research and development, as revealed by its lower improvement mean and higher variance. **Abbreviations:** Expected Improvement (`EI`); Adaptive Sampling/CEM-PI (`AS`); without search Importance Weights (`-IW`); Unbiased estimate of Covariance matrix (`+UC`); nondecreasing Threshold during optimization (`+T`).

2.4 Other Strategies

Fundamentally, Adaptive Sampling (`AS`) computes an approximation q to the implicit posterior over optima $p(\boldsymbol{x}^\star|\mathcal{D}_n) := p(\boldsymbol{x}|S_\tau, \boldsymbol{\theta}_0)$.[9] If the initial search model is uniform, then this posterior is proportional to a nonnegative acquisition function (see definition of $p(\boldsymbol{x}|S_\tau, \boldsymbol{\theta}_0)$ and footnote 6). Naturally, the next acquisition in standard BO is $\boldsymbol{x}_{n+1} = \arg\max p(\boldsymbol{x}^\star|\mathcal{D}_n) = \arg\max \alpha(\boldsymbol{x}|m_n, v_n)$ for a surrogate with posterior predictive mean m_n and variance v_n.[10] More broadly, it is possible to apply a monotonically increasing function g that preserves the optima, but changes the shape of the posterior $p(\boldsymbol{x}^\star|\mathcal{D}_n) \propto g(\alpha(\boldsymbol{x}^\star|m_n, v_n))$. This effectively extends the strategy to any real-valued α by using a nonnegative g, as long as the normalizing constant Z is finite. Interestingly, $p(\boldsymbol{x}^\star|\mathcal{D}_n) \propto \exp(\alpha(\boldsymbol{x}^\star|m_n, v_n))p_0(\boldsymbol{x}^\star)$ leads to an energy-based model,[11] where the partition function $Z(\boldsymbol{\theta})$ depends on the parameters of the surrogate model, and training data $\mathcal{D}_n$ if the latter is nonparametric. It seems that any algorithm that can sample from energy-based models is applicable.

Importantly, the aim is to sample from the highest density regions of $p(\boldsymbol{x}^\star|\mathcal{D}_n)$. For this purpose, one more strategy is based on the sequential application of the Laplace approximation to, e.g, $p(\boldsymbol{x}^\star|\mathcal{D}_n) \propto \texttt{EI}(\boldsymbol{x}^\star|m_n, v_n)$, leading to a Gaussian mixture search model.[12] In fact, this model can be obtained as a byproduct of an acquisition function I developed in February 2022, referred to as the collapsed `EI` (Appendix A). The problem is that we are required to optimize over multiple turns, but the cost can perhaps be amortized over subsequent iterations, e.g., once estimated, it can later be adapted with `AS` and Multiple Importance Sampling [`MIS` 25].

[9] Technically, a posterior over promising points since we are not interested in sampling data that have already been observed.

[10] It follows the same notation as that in [Section 2.4 16].

[11] This formulation is also related to belief distributions [23] (and LFI [e.g., 24]), where priors are updated to posterior beliefs through a negative loss/acquisition/performance function. More generally, there may be a tempering parameter [Section 3 23].

[12] Most importantly, it is a general strategy to estimate multimodal distributions without the problems generally associated with 1-turn VI (mass-covering or mode-seeking). Here, I target one mode at a time and collapse them to approximate the remaining. The resulting mixture approximation can lead to multiple importance sampling. Being a general strategy, its significance extends beyond BO and LFI.

References

[1] D. R. Jones et al. “Efficient Global Optimization of Expensive Black-Box Functions”. In: *Journal of Global optimization* 13.4 (1998), pages 455–492.

[2] S. R. Chowdhury and A. Gopalan. “On Kernelized Multi-armed Bandits”. In: *International Conference on Machine Learning*. PMLR. 2017, pages 844–853.

[3] N. Srinivas et al. “Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design”. In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*. 2010, pages 1015–1022.

[4] F. Berkenkamp et al. “No-regret Bayesian Optimization with Unknown Hyperparameters”. In: *JMLR* 20.50 (2019), pages 1–24.

[5] I. Bogunovic and A. Krause. “Misspecified Gaussian process bandit optimization”. In: *Advances in Neural Information Processing Systems* 34 (2021), pages 3004–3015.

[6] J. Snoek et al. “Practical Bayesian Optimization of Machine Learning Algorithms”. In: *Advances in Neural Information Processing Systems*. 2012, pages 2951–2959.

[7] E. Merrill et al. “An Empirical Study of Bayesian Optimization: Acquisition Versus Partition”. In: *JMLR* 22 (2021), pages 4–1.

[8] C. Hvarfner et al. “πBO: Augmenting Acquisition Functions with User Beliefs for Bayesian Optimization”. In: *International Conference on Learning Representations*. 2021.

[9] S. Gupta et al. “Regret Bounds for Expected Improvement Algorithms in Gaussian Process Bandit Optimization”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pages 8715–8737.

[10] C. Qin et al. “Improving the Expected Improvement Algorithm”. In: *Advances in Neural Information Processing Systems* 30 (2017).

[11] M. U. Gutmann, J. Corander, et al. “Bayesian Optimization for Likelihood-Free Inference of Simulator-Based Statistical Models”. In: *JMLR* (2016).

[12] R. Oliveira et al. “No-regret Approximate Inference via Bayesian Optimisation”. In: *Uncertainty in Artificial Intelligence*. PMLR. 2021, pages 2082–2092.

[13] M. Järvenpää et al. “Efficient acquisition rules for model-based approximate Bayesian computation”. In: *Bayesian Analysis* 14.2 (2019), pages 595–622.

[14] N. Houlsby et al. “Bayesian active learning for classification and preference learning”. In: *arXiv preprint arXiv:1112.5745* (2011).

[15] J.-M. Lueckmann et al. “Likelihood-free inference with emulator networks”. In: *Symposium on Advances in Approximate Bayesian Inference*. PMLR. 2019, pages 32–53.

[16] A. Eduardo and M. U. Gutmann. “Bayesian Optimization with Informative Covariance”. In: *arXiv preprint arXiv:2208.02704* (2022).

[17] D. Eriksson et al. “Scalable Global Optimization via Local Bayesian Optimization”. In: *Advances in Neural Information Processing Systems* 32 (2019).

[18] D. R. Jones and J. R. Martins. “The DIRECT algorithm: 25 years Later”. In: *Journal of Global Optimization* 79.3 (2021), pages 521–566.

[19] C. Oh et al. “BOCK: Bayesian optimization with cylindrical kernels”. In: *ICML*. PMLR. 2018, pages 3868–3877.

[20] M. Balandat et al. “BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization”. In: *Advances in Neural Information Processing Systems* 33 (2020), pages 21524–21538.

[21] D. Brookes et al. “Conditioning by adaptive sampling for robust design”. In: *ICML*. PMLR. 2019, pages 773–782.

[22] R. Y. Rubinstein and D. P. Kroese. *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*. Volume 133. Springer, 2004.

[23] P. G. Bissiri et al. “A General Framework for Updating Belief Distributions”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78.5 (2016), pages 1103–1130.

[24] O. Thomas et al. “Misspecification-Robust Likelihood Free Inference in High Dimensions”. In: *arXiv preprint arXiv:2002.09377* (2020).

[25] V. Elvira and L. Martino. “Advances in Importance Sampling”. In: *arXiv preprint arXiv:2102.05407* (2021).

[26] Z. Wang and N. de Freitas. “Theoretical Analysis of Bayesian Optimisation with Unknown Gaussian Process Hyper-Parameters”. In: *arXiv preprint arXiv:1406.7758* (2014).

A Collapsed Expected Improvement

Expected Improvement (`EI`) and Lower Confidence Bound (`LCB`) are susceptible to local minima. There is no intrinsic mechanism that prevents the rules from choosing points with already small predictive variances. In `LCB`, this is typically addressed by increasing β_n and decreasing lengthscales over time [3, 4], ensuring that the entire domain is eventually explored. Increasing β_n is functionally almost the same as increasing the prior variance. So, a similar strategy can carry over to `EI`. However, if the prior variance increases carelessly, or if the lengthscales decrease rapidly, BO becomes very exploratory before it has the chance to exploit any information given by the posterior predictive mean. So, unconditional updates should be avoided.[13]

Wang et al. [26] provide a conditional rule for such update that seems reasonable: Lengthscales only decrease if the predictive variances of successive future acquisitions are less than a certain threshold, e.g., noise variance. The problem with this approach is that the amount of decrease that leads to mode collapse of `EI` is unknown, so it may take multiple updates and evaluations before this main acquisition mode vanishes. Here, we modify the acquisition function directly by removing this acquisition mode, and the procedure is repeated until the predictive variance of the acquisition is larger than the threshold. To this end, we use the Laplace method, estimating a log quadratic model centered at the location that maximizes the current acquisition function.

Collapsed EI (`CEI`) may be computationally more expensive than `EI`. Still, larger posterior predictive variances translate into more informative acquisitions, making better use of the limited evaluation budget. As a byproduct, note that repeated application of the Laplace method with mode collapse retrieves a Gaussian mixture search model $q(\boldsymbol{x}) = \sum_i w_i \mathcal{N}(\boldsymbol{x}; \boldsymbol{x}^{(i)}, \mathbf{H}^{-1\,(i)})$, with $w_i \propto \mathrm{EI}(\boldsymbol{x}^{(i)})$, that approximates the implicit posterior over promising points $p(\boldsymbol{x}^\star | \mathcal{D}_n) \propto \mathrm{EI}(\boldsymbol{x}^\star)$. This information about promising points can also later be fed to informative covariance functions [16] via the shaping function ϕ that induces the nonstationary effects.

Algorithm 1 Collapsed EI (`CEI`)

```
Input: posterior predictive variance $v_n$, threshold $\epsilon$
Output: acquisition $\boldsymbol{x}^{(i)}$
$\boldsymbol{x}^{(0)} = \arg\max \mathrm{EI}(\boldsymbol{x})$
$i = 0$
$\mathrm{CEI}^{(i)} = \mathrm{EI}$
while $v_n(\boldsymbol{x}^{(i)}) \leq \epsilon$ do
    $\mathbf{H}^{(i)} = \nabla^2_{\boldsymbol{x}} - \log \mathrm{CEI}^{(i)}(\boldsymbol{x})|_{\boldsymbol{x}=\boldsymbol{x}^{(i)}}$
    $\mathrm{CEI}^{(i+1)}(\boldsymbol{x}) = \mathrm{CEI}^{(i)}(\boldsymbol{x}) - \mathrm{CEI}^{(i)}(\boldsymbol{x}^{(i)}) \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^{(i)})^\intercal \mathbf{H}^{(i)} (\boldsymbol{x} - \boldsymbol{x}^{(i)})\right)$
    $\boldsymbol{x}^{(i+1)} = \arg\max \mathrm{CEI}^{(i+1)}(\boldsymbol{x})$
    $i = i + 1$
end while
```

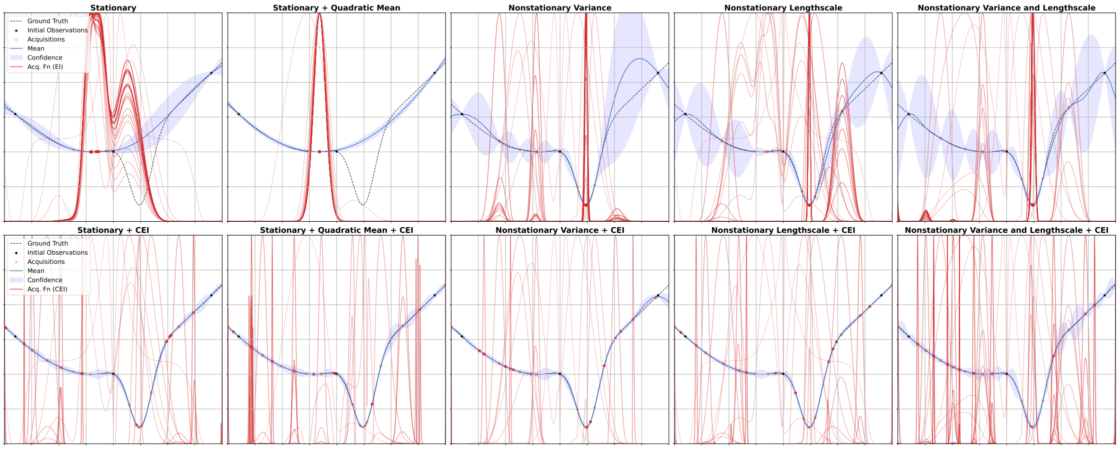

Figure A.1. A comparison of `EI` and `CEI`. In this example, `CEI` can solve the problem that is due to overconfident stationary models. Overall, it leads to more informative acquisitions as it avoids locations with small predictive variances.

[13] Interestingly, informative covariance functions [16] can too be of help in this case: Information from the posterior predictive mean can be directly incorporated in the proposed covariance functions through the shaping function ϕ.

B BO with Adaptive Sampling (CEM-EI)

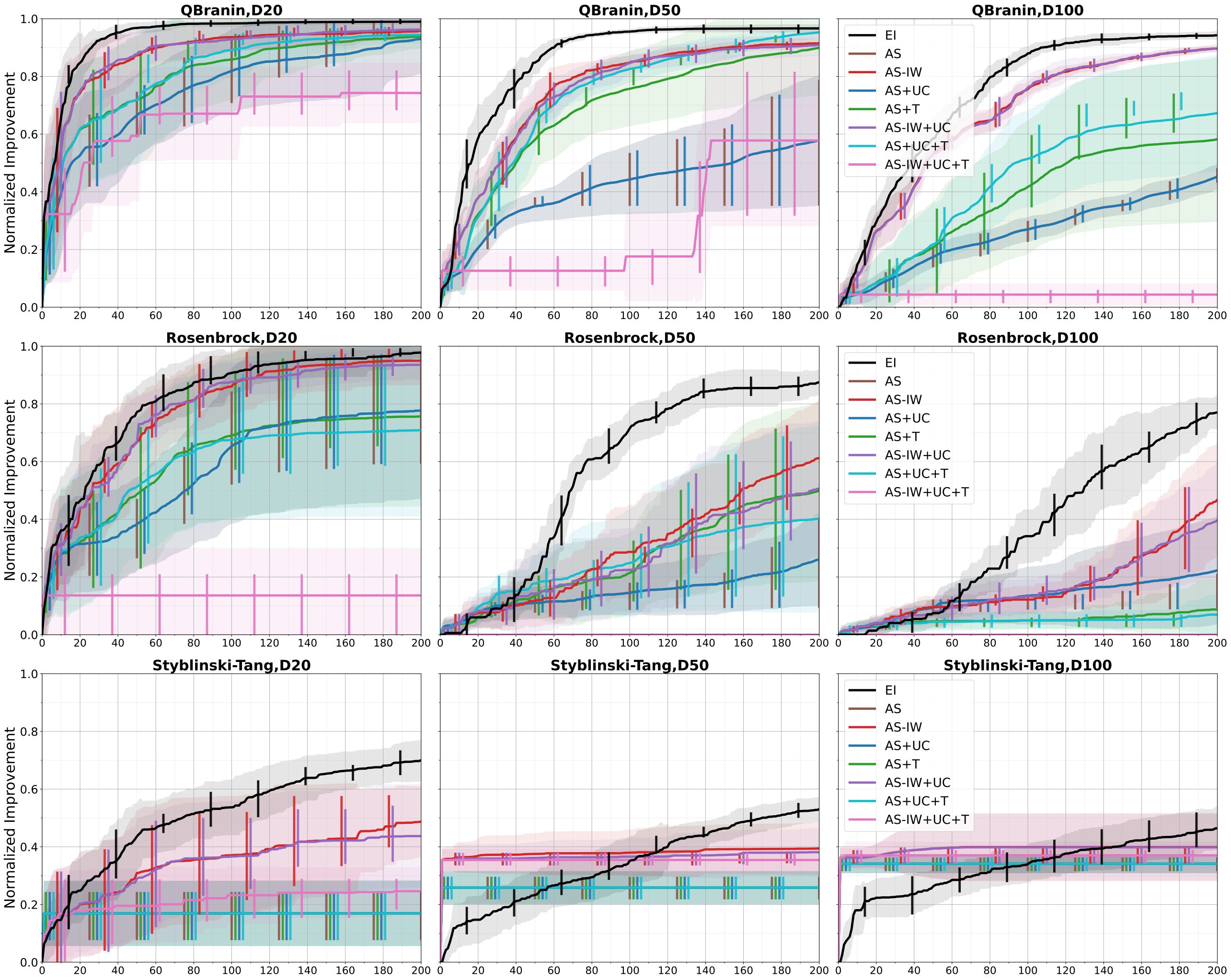

Figure B.1. Performance of acquisition strategies, EI vs CEM-EI (and variants), referred here again as AS, on objectives ranging from 20 to 100 dimensions [Appendix B.3 16]. The total budget is 200 acquisitions, excluding initial observations. Solid curves and shaded regions represent the normalized improvement mean and one standard deviation, computed over 10 trials, each with different initial conditions/observations. Solid vertical lines show the interquartile range. The type of surrogate model is the same in all experiments [I+XA 16]. Search models in AS are multivariate Gaussians. Variants -IW can significantly outperform others. Overall, AS requires more research and development, as revealed by its lower improvement mean and in some cases higher variance. **Abbreviations:** Expected Improvement (EI); Adaptive Sampling/CEM-EI (AS); without search Importance Weights (-IW); Unbiased estimate of Covariance matrix (+UC); nondecreasing Threshold during optimization (+T).