On Multi-information source Constraint Active Search

Gustavo Malkomes[*]
Intel SigOpt

Santiago Miret
Intel Labs

Bolong Cheng
Intel SigOpt

Abstract

Constraint active search is a promising sample-efficient multiobjective experimental design formulation that aims to aid scientists and engineers in searching for new materials. In this proposal, we extend this formulation to situations where one can obtain observations from multiple sources each with a given cost, such as when both computer simulations and a laboratory experiments can be used to calculate (or estimate) properties of a materials. We present a novel cost-efficient policy that balances the cost of obtaining observations with the benefit of evaluating a more expensive-to-compute source. Initial results on a synthetic problem show that our proposed methodology is more selective when searching for expensive source.

1 Introduction

Multi-information source experimental design: We consider the problem of intelligent experimentation on an expensive-to-evaluate black-box function where we can access cheaper approximations objectives, similar to the optimization setting considered on [Poloczek et al., 2017]. For instance, this scenario arises when one seeks to design new materials with the assistance of computer simulations. Performing laboratory experiments is expensive, whereas outcomes of computer simulations are relatively cheap-to-obtain, when labor and time are considered.

In recent years, machine learning techniques have been applied to similar scientific problems in material science, drug discovery materials design, drug discovery, and chemical engineering Forrester et al. [2008], Negoescu et al. [2011], Molesky et al. [2018].In particular, adaptive methods gained popularity due to their sample efficiency Song et al. [2018], Attia et al. [2020], Haghanifar et al. [2020], Duris et al. [2020]. In particular, we focus on constraint active search (CAS), a recently proposed and promising formulation for experimental design [Malkomes et al., 2021]. In this formulation, the goal is to search for *diverse* and *high-performing* parameter configurations. As opposed to searching for the *single best parameter configuration* as in standard optimization problems, CAS seeks to find parameter configurations that satisfy a given predefined performance criterion. Searching for multiple and diverse candidate configurations increases the chances of finding new materials.

Multi-fidelity experimentation in materials design: In this paper, we introduce an automated way to perform constrained active search across multiple fidelities of data acquisition methods. This can be particularly relevant to AI-guided materials design when having multiple data sources including different simulation techniques and experimental results. One concrete example may be that the lowest fidelity technique leverages semiempirical techniques, such as PM6, with mid-tier fidelities spanning several types of density functional theory-based simulations and the highest fidelity being a costly experimental measurement. Our method maps best to simulation based settings where cost be easily defined as the compute cost of a simulation but can be extended to include real-world experimental evaluation assuming that cost metrics can be defined.

[*] Correspondence to: <gustavo.malkomes@intel.com>

36th Conference on Neural Information Processing Systems (NeurIPS 2022) AI4Mat Workshop.

2 Multi-information source constraint active search

Suppose that we have access to an expensive-to-evaluate objective function $f : \mathcal{X} \rightarrow \mathbb{R}$ that represents the quality of given materials describe by features $\mathbf{x} \in \mathcal{X}$. We wish to search for designs that satisfy a certain performance criteria, for example, we seek for $f(\mathbf{x}) \geq \tau$, $\tau \in \mathbb{R}$[2]. Now, consider that we have access to other auxiliary functions $\{f_l\}$, $l \in \mathcal{I}$ from other sources such that each $f_l \colon \mathcal{X} \rightarrow \mathbb{R}$ may reveal information about f at a lower cost $c_l \colon \mathcal{X} \rightarrow \mathbb{R}$. Let us index our desired so-called *full-fidelity* objective function f with a special index $* \in \mathcal{I}$ such that $f = f_*$. As a practical example, consider that f_l is a computer simulation and f_* is the outcome of an laboratory experiment.

In multi-information source constraint active search, we seek to sequentially select an input location $\mathbf{x} \in \mathcal{X}$ and an index $l \in \mathcal{I}$ to obtain an observation $f_l(\mathbf{x})$ with cost $\textsc{cost}(l)$. Our goal is to find the *satisfactory region*[3] of the full-fidelity function, $\mathcal{S}^\star = \{\mathbf{x}^\star \mid f^\star(\mathbf{x}) \geq \tau\}$, with a restricted total budget B of observations to: $\sum_{t:(i,l)} \textsc{cost}_{t(l)}(\mathbf{x}_t) \leq B$.

2.1 Cost-aware expected coverage increase

To efficiently search for the satisfactory region of each source $\mathcal{S}_l$, we extend the Expected Coverage Increase (ECI) policy to our setting [Malkomes et al., 2021], in which we can query multiple information source with different costs. First, we review the concept of coverage.

Coverage neighborhood [Malkomes et al., 2021]. The *coverage neighborhood* of any $\mathbf{x}$ is defined as $\mathbb{N}_r(\mathbf{x}) = \{\mathbf{x}' : d(\mathbf{x}, \mathbf{x}') < r\}$, for an *a priori* fixed $r \in \mathbb{R}^+$ and an appropriate distance function $d : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}^+$. The *coverage neighborhood of a set of points* $\mathbf{X}$ is defined as

$$\mathbb{N}_r(\mathbf{X}) = \bigcup_{\mathbf{x} \in \mathbf{X}} \mathbb{N}_r(\mathbf{x}).$$

For Euclidean distance, $\mathbb{N}_r(\mathbf{x})$ is simply a ball of radius r. Notice the relationship between the radius r and the diversity of a set of points $\mathbf{X}$: points $\mathbf{x} \in \mathbf{X}$ that have distance larger than r, i.e. points that are dissimilar according to our distance function d, will lead to large $\mathbb{N}_r(\mathbf{X})$; whereas points within r (similar points) will have smaller impact in the coverage neighborhood.

Coverage recall. An important metric for constraint active search is how much we recover from a given satisfactory region $\mathcal{S}$. Given a known set $\mathcal{S}$ and samples $\mathbf{X}$, the *coverage recall* is $|\mathbb{N}_r(\mathbf{X}) \cap \mathcal{S}| \, / |\mathcal{S}|$. This is a key metric of progress for control experiments.

In practice, we do not know the satisfactory region; therefore, we will use probabilistic models to capture our prior beliefs about f_l and thus $\mathcal{S}_l$. Assume that we have probabilistic models $\mathcal{M}_l$ for each objective function f_l. When appropriate, we can use observations from other sources to inform a given model, for example, our full-fidelity model $\mathcal{M}_*$.

Let Z_l be a variable that indicates if an observation $\mathbf{x}_l$ will satisfy the thresholds for a given source l, $Z(\mathbf{x}_l) = \mathbf{1}\big[f_l(\mathbf{x}_l) \geq \tau\big]$. Our probabilistic models can compute the probability of any point $\mathbf{x}_l$ belonging to the satisfactory region $\mathcal{S}_l$ given the observed data $\mathcal{D}$. We denote this as $p(Z(\mathbf{x}_l) = 1 \mid \mathcal{D}, \mathcal{M}_l)$. For a given source l, the *expected coverage increase* acquisition function is

$$\textsc{eci}(\mathbf{x} \mid \mathcal{D}, l) = \mathbb{E}_{Z_l}\Bigg[\sum_{\mathbf{x}' \in \mathbb{N}(\mathbf{x}) \setminus \mathbb{N}_r(\mathbf{X})} Z_l(\mathbf{x}')\Bigg] = \sum_{\mathbf{x}' \in \mathbb{N}(\mathbf{x}) \setminus \mathbb{N}_r(\mathbf{X})} p(Z_l(\mathbf{x}') = 1 \mid \mathcal{D}, \mathcal{M}_l).$$

To extend ECI to multi-information source we propose the *cost-aware expected coverage increase*, which selects the next point by finding the input location $\mathbf{x}_l$ that maximizes ECI over cost

$$\textsc{ca-eci}(\mathbf{x}, l \mid \mathcal{D}) = \frac{\textsc{eci}(\mathbf{x} \mid \mathcal{D},\, l)}{\textsc{cost}(l)}.$$

To further improve cost-efficiency, we only evaluate higher cost sources after obtaining a given number of satisfactory points in the lower cost objective. For example, in our demonstration below,

[2] We could generalize this to multiple criteria as long as they have the same information cost, but we will only consider single-metric for simplicity in this proposal.

[3] Satisfactory region is the set of input locations that yield good results according to the black-box performance criteria. Notice that this definition is different than feasible region.

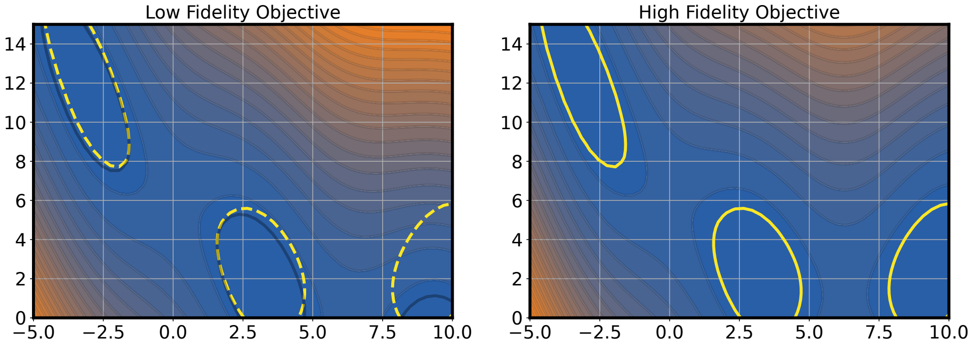

Figure 1: Augmented Branin function with fidelity parameter 0.1 (low fidelity) and 1 (high fidelity). The region in yellow depicts the satisfactory region, where the function values are below 10, of the full-fidelity function on both figures; our goal is to find this region. The functions are correlated in the regions with low function values for two of the three regions (yellow circles).

we perform ECI on the low-fidelity function until we find five satisfactory points. Next, we evaluate these satisfactory points in the higher-cost source, and only then we use CA-ECI to automatically select $\mathbf{x}_l$ (input and source).

3 Demonstration of multi-information source CAS

We now demonstrate our proposed methodology and compare it with other approaches. We consider an augmented version of the Branin function that is commonly used for testing multi-fidelity methods Wu et al. [2020].

Figure 1 shows the two objectives we considered here. We will assume the cost of obtaining a full-fidelity observation is one unit and that of a low-fidelity observation is 0.1. In other words, one sample of the full-fidelity function is equivalent to 10 samples of the low-fidelity function in terms of cost. In yellow, we label the satisfactory region of the *full-fidelity function*; these are regions with the high performing observations. We are interested in finding diverse configurations inside this region. Notice the two functions are highly correlated, except for the rightmost region of the input space.

Figure 2 shows the sample pattern of four methodologies. We ran each method 100 times with a total cost budget of 15. We compare to random search, expected improvement (EI), and the original ECI; these three methods only have access to the full-fidelity function.

For EI, we observe that it focuses on finding the center of the satisfactory regions, i.e., the likely local optima of the function. For ECI, we can see that the sample pattern is denser around the satisfactory region; instead of concentrating in the middle as Expected Improvement, it spreads points inside and around the satisfactory region.

Cost-Aware ECI, on the other hand, had the same total budget of 15, but it can sample points from the low fidelity objective with 1/10 the cost; as a result, this method can be very selective about sampling the more costly full-fidelity function. On average, Cost-Aware ECI select only uses 80% of the budget on sampling the full-fidelity function. However, we see that the full-fidelity observations are more densely spread within the satisfactory regions. Note that the third yellow region (on the right side) was not sampled by Cost-Aware ECI due to the mismatch between the low and high fidelity functions.

In Figure 3, we show the best full-fidelity function value found and the coverage recall for the methodologies described above. With the restricted budget of 15, both ECI methods seem to perform well for finding the best function value (dashed grey line), but this is not the goal of these methods. We expected optimization methods to find the optimal value, while ECI methods do not have this goal. Nevertheless, we can see that all methods find high-performing observations, i.e., values below the threshold 10, fairly quickly. When considering the volume covered inside the satisfactory region, we can see that both ECI methods perform well in this metric, with CA-ECI being the most efficient.

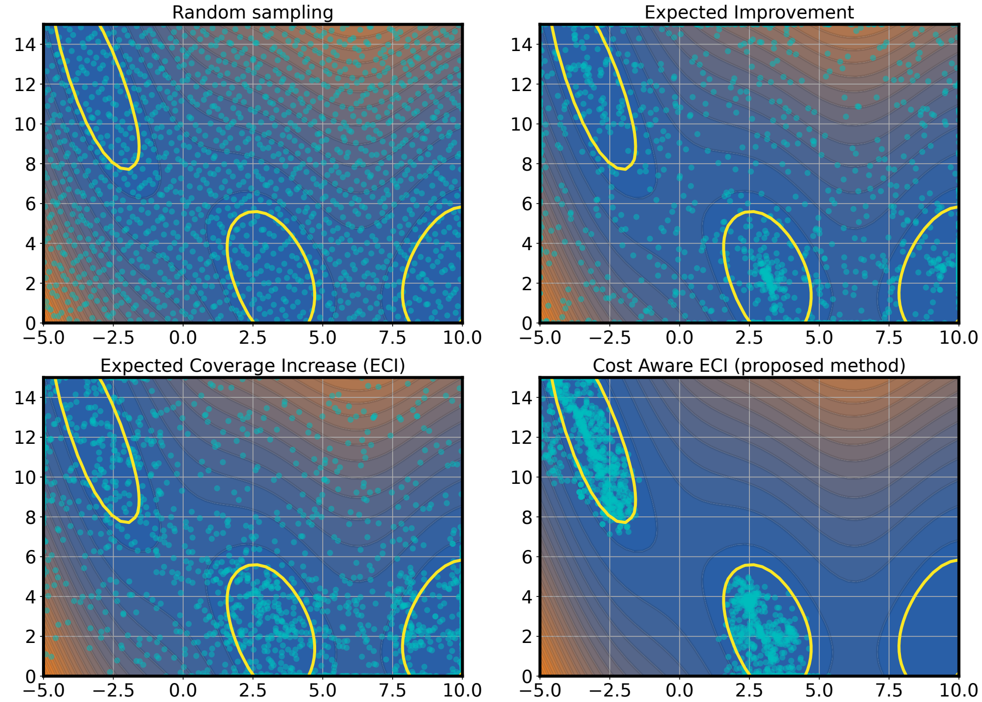

Figure 2: Sample pattern of the full-fidelity observations for different methodologies across 100 experiments, each with a total cost of 15.

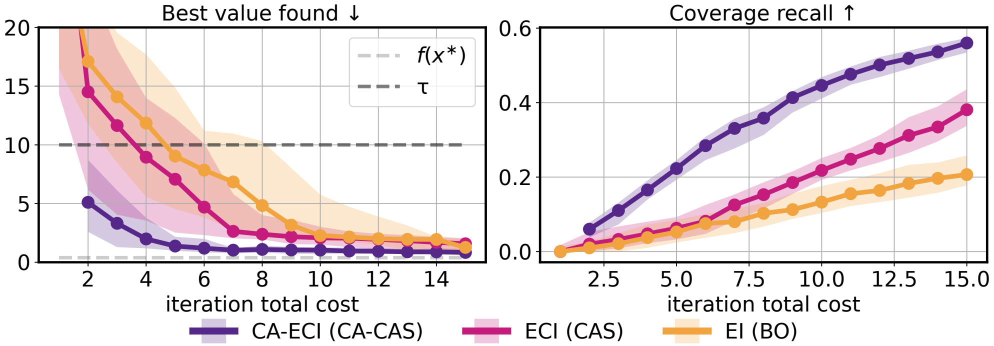

Figure 3: Quantitative results for the Augmented Branin over 100 experiments. We show the median of the best value found and the coverage recall computed per iteration, the total cost is 15, and the results are computed for the full-fidelity function. The shaded region represents the interquartile range of each metric.

4 Discussion and future work

In this work-in-progress we introduce multi-information source constraint active search for experimental design, which is an experimentation formulation that seeks to automatically find diverse and high-performing configurations using different sources. We extend ECI to this setting by proposing the Cost-Aware ECI. Crucially, our method is very selective about querying full-fidelity source.

In future work, we would like to investigate the effectiveness of our approach in real-world problems, contrasting it with other multi-fidelity methods. There are a few topics that we would like to investigate further. First, we want to generalize this methodology to more than two information sources; this is occurs of accessing a simulator with different mesh sizes or Monte-Carlo sample sizes. In particular, we want to study the relationship between efficacy of our proposed algorithm (the ability to identify high performing regions in the full-fidelity model) and the number of information sources (and perhaps more importantly, the relative costs of each).

Next, we want to quantify the impact of the correlation between the different information sources on the final performance of the algorithms. We believe that this can provide us with intuition on how to set up the multiple costs/fidelities setting. This is still an open question in the community.

Lastly, we want to consider multiple objectives along with multiple information sources.

References

Peter M. Attia, Aditya Grover, Norman Jin, Kristen A. Severson, Todor M. Markov, Yang-Hung Liao, Michael H. Chen, Bryan Cheong, Nicholas Perkins, Zi Yang, Patrick K. Herring, Muratahan Aykol, Stephen J. Harris, Richard D. Braatz, Stefano Ermon, and William C. Chueh. Closed-loop optimization of fast-charging protocols for batteries with machine learning. *Nature*, 578(7795): 397–402, Feb 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-1994-5.

J. Duris, D. Kennedy, A. Hanuka, J. Shtalenkova, A. Edelen, P. Baxevanis, A. Egger, T. Cope, M. McIntire, S. Ermon, and D. Ratner. Bayesian optimization of a free-electron laser. *Phys. Rev. Lett.*, 124:124801, Mar 2020. doi: 10.1103/PhysRevLett.124.124801.

Alexander Forrester, Andras Sobester, and Andy Keane. *Engineering design via surrogate modelling: a practical guide*. John Wiley & Sons, 2008.

Sajad Haghanifar, Michael McCourt, Bolong Cheng, Jeffrey Wuenschell, Paul Ohodnicki, and Paul W. Leu. Discovering high-performance broadband and broad angle antireflection surfaces by machine learning. *Optica*, 7(7):784–789, Jul 2020. doi: 10.1364/OPTICA.387938.

Gustavo Malkomes, Bolong Cheng, Eric Hans Lee, and Mike Mccourt. Beyond the pareto efficient frontier: Constraint active search for multiobjective experimental design. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7423–7434. PMLR, 2021. URL `http://proceedings.mlr.press/v139/malkomes21a.html`.

Sean Molesky, Zin Lin, Alexander Y Piggott, Weiliang Jin, Jelena Vucković, and Alejandro W Rodriguez. Inverse design in nanophotonics. *Nature Photonics*, 12(11):659–670, 2018.

Diana M Negoescu, Peter I Frazier, and Warren B Powell. The knowledge-gradient algorithm for sequencing experiments in drug discovery. *INFORMS Journal on Computing*, 23(3):346–363, 2011.

Matthias Poloczek, Jialei Wang, and Peter Frazier. Multi-information source optimization. *Advances in neural information processing systems*, 30, 2017.

Jialin Song, Yury S. Tokpanov, Yuxin Chen, Dagny Fleischman, Kate T. Fountaine, Harry A. Atwater, and Yisong Yue. Optimizing photonic nanostructures via multi-fidelity Gaussian processes. NeurIPS Workshop on Machine Learning for Molecules and Materials, December 2018.

Jian Wu, Saul Toscano-Palmerin, Peter I Frazier, and Andrew Gordon Wilson. Practical multi-fidelity bayesian optimization for hyperparameter tuning. In *Uncertainty in Artificial Intelligence*, pages 788–798. PMLR, 2020.