BOtied: Multi-objective Bayesian optimization using multivariate ranks and quantiles

Ji Won Park

Joint work with **Natasa Tagasovska**
Principal Machine Learning Scientist
Prescient Design, Genentech/Roche

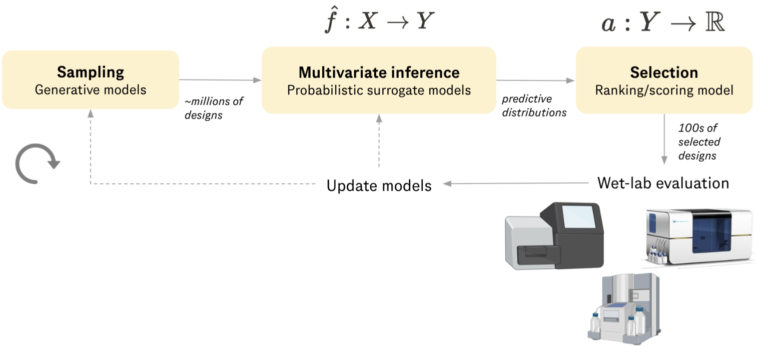
1 March 2024
SIAM Conference on Uncertainty Quantification

Prescient Design
A Genentech Accelerator

Genentech
A Member of the Roche Group

About my research

High-dimensional inference and adaptive decision making



Interests: MCMC sampling, Bayesian modeling, causal representation learning, extreme values, uncertainty quantification and calibration, ...

Content

Based on:

Park, J. W.*, Tagasovska, N.*, Maser, M., Ra, S., & Cho, K. BOtied: Multi-objective Bayesian optimization with tied multivariate ranks, an application to active drug discovery. ICML 2024. arXiv: 2306.00344

Motivation and Background

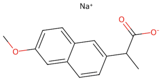
- Drug design: jointly optimizing multiple (tailed) molecular properties
- A quick primer on multi-objective Bayesian optimization (MOBO)

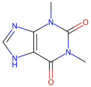
Method

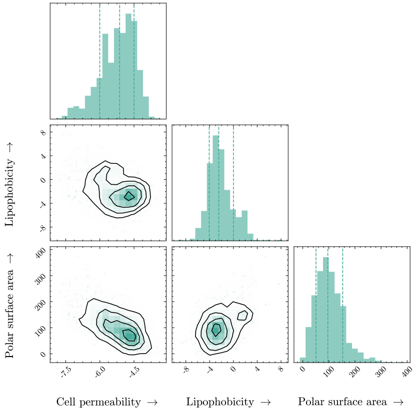
Empirical results

Molecular design: tale of correlated tails

- Goal: jointly optimize molecule for multiple competing properties
- Molecular properties tend to have long tails[1] and tail correlations[2]







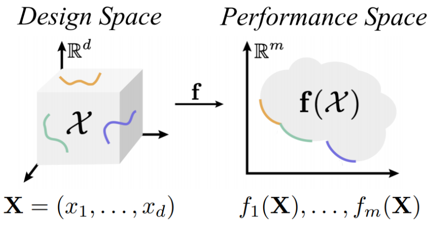
[1] Jain et al., "Biophysical properties of the clinical-stage antibody landscape" (2017).
[2] Wang et al., "ADME properties evaluation in drug discovery: prediction of Caco-2 cell permeability using a combination of NSGA-II and boosting" (2016).

Multi-objective Bayesian optimization (MOBO)

Problem

Optimizing a vector-valued objective $f : \mathbb{R}^d \to \mathbb{R}^M$ with $f(x) = (f_1(x), \ldots, f_M(x))$ over a bounded set $\mathcal{X} \subset \mathbb{R}^d$.

When f is an expensive black-box function (e.g., wet lab protocol), Bayesian optimization offers a sample-efficient method.[3]



[4]

[3] Jones, Schonlau, and Welch, "Efficient global optimization of expensive black-box functions" (1998).

[4] Konakovic Lukovic, Tian, and Matusik, "Diversity-guided multi-objective bayesian optimization with batch evaluations" (2020).

The MOBO algorithm

Key components:

- **Surrogate** $\hat{f} : \mathbb{R}^d \to \mathbb{R}^M$ tractably approximating f, with $p(\hat{f}|\mathcal{D})$
- **Acquisition function** $a^{\hat{f}} : \mathcal{X} \to \mathbb{R}$ capturing the "usefulness" of each design, used to determine which design to evaluate next
  - exploration (of highly uncertain designs)
  - exploitation (of designs believed to be optimal)

The MOBO algorithm

Key components:

- **Surrogate** $\hat{f} : \mathbb{R}^d \to \mathbb{R}^M$ tractably approximating f, with $p(\hat{f}|\mathcal{D})$
- **Acquisition function** $a^{\hat{f}} : \mathcal{X} \to \mathbb{R}$ capturing the "usefulness" of each design, used to determine which design to evaluate next
    - exploration (of highly uncertain designs)
    - exploitation (of designs believed to be optimal)

MOBO proceeds in repeating cycles of

1. Fitting the surrogate on $\mathcal{D} = \{(x^{(i)}, f(x^{(i)})\}_{i=1}^{N}$, to obtain $p(\hat{f}|\mathcal{D})$
2. Optimizing to obtain $x^* = \mathrm{argmax}_{x\in\mathcal{X}}\ a^{\hat{f}}(x)$
3. Appending the resulting measurement: $\mathcal{D} \leftarrow \mathcal{D} \cup \{(x^*, f(x^*))\}$

Dominance operators: notation

How to compare vectors in Euclidean spaces when $M > 2$?
Assume minimization. For $y = (y_1, \ldots, y_M), z = (z_1, \ldots, z_M) \in \mathbb{R}^M$,

- z **weakly** dominates y
  $z \preccurlyeq y \iff z_i \le y_i \ \forall i \in [M]$
- z **strictly** dominates y
  $z \lneq y \iff z_i \le y_i \ \forall i \in [M]$ and $\exists k \in [M] : z_k < y_k$
  $\iff z \preccurlyeq y$ and $z \neq y$

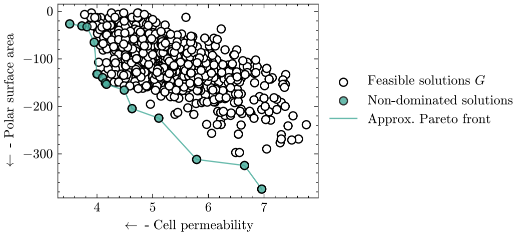
Pareto front

Pareto front $\mathcal{P}$ of the set G is the subset containing points which are not strictly dominated:

$$y \in \mathcal{P} \iff \forall z \in G, \neg(z \precneqq y),$$

and equiv. the subset of G that are weakly dominated only by themselves:[5]

$$y \in \mathcal{P} \iff \{z \in G, z \preccurlyeq y\} = \{y\}.$$



MOBO aims to obtain a finite approximation $\hat{\mathcal{P}}$ to the true Pareto front $\mathcal{P}$.

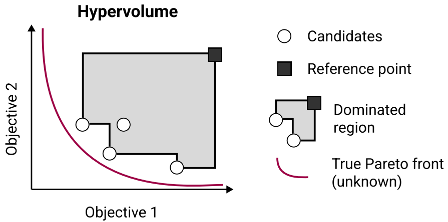
[5] Warburton, "Quasiconcave vector maximization: connectedness of the sets of Pareto-optimal and weak Pareto-optimal alternatives" (1983).

Quality indicators

Quality indicator $I : 2^{\mathcal{Y}} \to \mathbb{R}$
evaluates the quality of approximation set $\hat{\mathcal{P}}$.

Hypervolume indicator

Example: **hypervolume** (HV)[6] of polytope bounded from below by $\hat{\mathcal{P}}$ and from above by a reference point



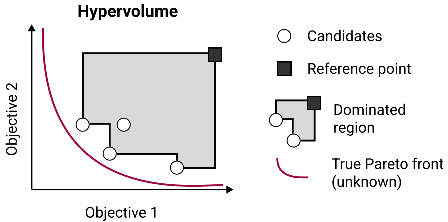
[6] Emmerich, Deutz, and Klinkenberg, "Hypervolume-based expected improvement: Monotonicity properties and exact computation" (2011).

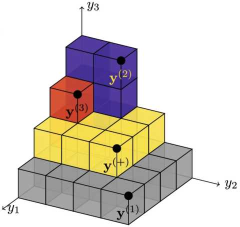
[7] Yang et al., "A multi-point mechanism of expected hypervolume improvement for parallel multi-objective bayesian global optimization" (2019).

Hypervolume indicator: limitations

Example: **hypervolume** (HV)[6] of polytope bounded from below by $\hat{\mathcal{P}}$ and from above by a reference point

- HV $\sim \mathcal{O}(n^{\lfloor \frac{M}{2} \rfloor}) \rightarrow$ impractical for $M>4$ despite box decomposition[7]



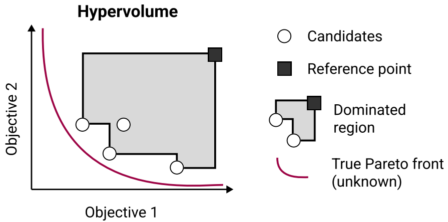


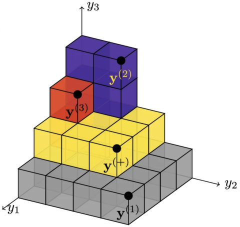
[6] Emmerich, Deutz, and Klinkenberg, "Hypervolume-based expected improvement: Monotonicity properties and exact computation" (2011).
[7] Yang et al., "A multi-point mechanism of expected hypervolume improvement for parallel multi-objective bayesian global optimization" (2019).

Hypervolume indicator: limitations

Example: **hypervolume** (HV)[6] of polytope bounded from below by $\hat{\mathcal{P}}$ and from above by a reference point

- HV $\sim \mathcal{O}(n^{\lfloor \frac{M}{2} \rfloor}) \rightarrow$ impractical for $M > 4$ despite box decomposition[7]
- Sensitive to rescaling of the objectives, with different natural units





[6] Emmerich, Deutz, and Klinkenberg, "Hypervolume-based expected improvement: Monotonicity properties and exact computation" (2011).

[7] Yang et al., "A multi-point mechanism of expected hypervolume improvement for parallel multi-objective bayesian global optimization" (2019).

Content

Motivation and Background

- Drug design: jointly optimizing multiple (tailed) molecular properties
- A quick primer on multi-objective Bayesian optimization (MOBO)
  - Acquisition function $a^{\hat{f}} : \mathcal{X} \to \mathbb{R}$
  - Quality indicator $I : 2^{\mathcal{Y}} \to \mathbb{R}$

Method

- Connection between the CDF ranks and the Pareto front
- BOtied: MOBO based on the CDF

Empirical results

Probabilistic perspective

Let $Y = f(X) \in \mathbb{R}^M$, where X is a random vector with values in $\mathcal{X}$.

[8] Binois, Rullière, and Roustant, "On the estimation of Pareto fronts from the point of view of copula theory" (2015).

Probabilistic perspective

Let $Y = f(X) \in \mathbb{R}^M$, where X is a random vector with values in $\mathcal{X}$.

$$y \in \mathcal{P} \implies \mathbb{P}[\ Y \in \underbrace{\{z \in G, z \preccurlyeq y\}}_{\text{set weakly dom. } y}\ ] \stackrel{\text{def}}{=} \mathbb{P}[\ Y \in \{y\}\ ] = 0$$

[8] Binois, Rullière, and Roustant, "On the estimation of Pareto fronts from the point of view of copula theory" (2015).

Probabilistic perspective

Let $Y = f(X) \in \mathbb{R}^M$, where X is a random vector with values in $\mathcal{X}$.

$$y \in \mathcal{P} \implies \mathbb{P}[\ Y \in \underbrace{\{z \in G, z \preccurlyeq y\}}_{\text{set weakly dom. } y}\ ] \stackrel{\text{def}}{=} \mathbb{P}[\ Y \in \{y\}\ ] = 0$$

$$\implies \mathbb{P}[Y \preccurlyeq y] := F_Y(y) = 0,$$

with $F_Y(y)$ the cumulative distribution function (CDF) of Y.

[8] Binois, Rullière, and Roustant, "On the estimation of Pareto fronts from the point of view of copula theory" (2015).

Connection between the CDF and the Pareto front

Let $Y = f(X) \in \mathbb{R}^M$, where X is a random vector with values in $\mathcal{X}$.

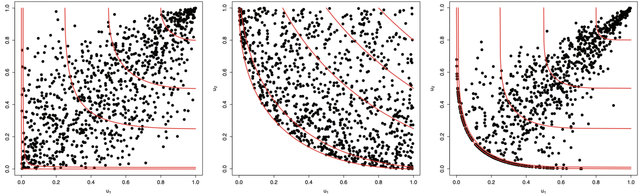
$$y \in \mathcal{P} \implies \mathbb{P}[\ Y \in \underbrace{\{z \in G, z \preccurlyeq y\}}_{\text{set weakly dom. } y}\ ] \stackrel{\text{def}}{=} \mathbb{P}[\ Y \in \{y\}\ ] = 0$$

$$\implies \mathbb{P}[Y \preccurlyeq y] := F_Y(y) = 0,$$

with $F_Y(y)$ the cumulative distribution function (CDF) of Y.

Consider the α level line of F_Y, $\partial\mathcal{L}_\alpha^F = \{y' \in G, F_Y(y') = \alpha\}$.

The Pareto front belongs to the zero ($\alpha = 0$) level line of F_Y![8]



[8]Binois, Rullière, and Roustant, "On the estimation of Pareto fronts from the point of view of copula theory" (2015).

CDF vs. PDF

$$F_{Y_1,\ldots,Y_M}(y) = \mathbb{P}[Y_1 \leq y_1, \ldots, Y_M \leq y_m] = \int_{(-\infty,\ldots,-\infty)}^{(y_1,\ldots,y_M)} f_Y(s)ds.$$

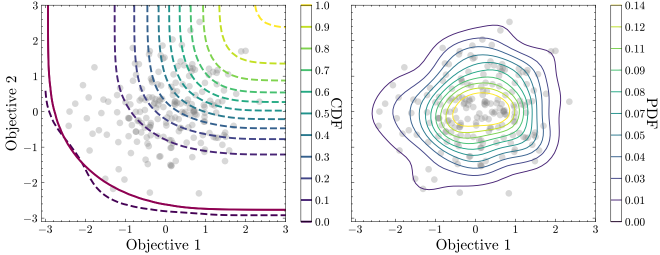


Figure 1: Level lines of the CDF (left) and the PDF (right) from kernel density estimation based on 200 observations (gray dots). The zero level line of the CDF closely traces the true Pareto front (solid red curve).

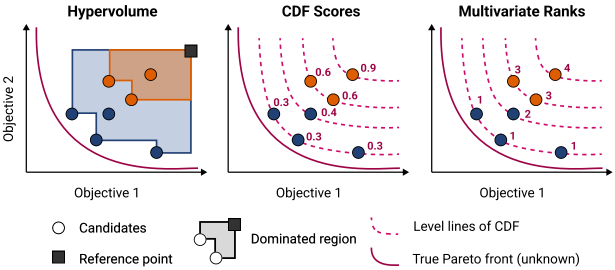
Enter the CDF indicator[9]

$$\text{We propose } I_{\mathrm{CDF}}(A) \coloneqq \min_{y\in A}\left(F_Y(y)\right).$$

Weak Pareto compliance (Theorem 4.1)

For any arbitrary approximation sets $A, B \in 2^{\mathcal{Y}}$,

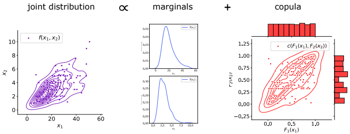
$$A \precsim B \implies I_{CDF}(A) \leq I_{CDF}(B).$$

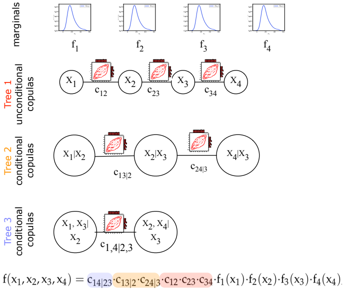


[9] Park et al., "BOtied: Multi-objective Bayesian optimization with tied multivariate ranks" (2023).

Efficient fitting of CDF with vine copulas

We can pairwise decompose an M-dim copula density into a product of $M(M-1)/2$ bivariate conditional densities ("pair copulas") organized in a sequence of trees ("vine")[10] $\sim \mathcal{O}(nML)$, where $L \in \{1, \ldots, M\}$ is depth



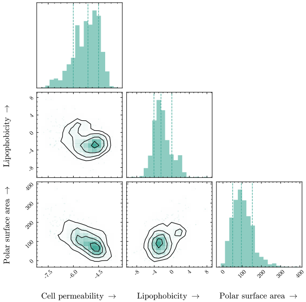


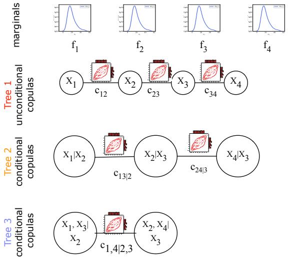
[10] Joe, *Multivariate Models and Dependence Concepts* (1997).

Model-based Pareto front

Domain knowledge or information from unpaired observations of Y (without X associations) can be encoded in the choices of

- marginal distributions
- pair copula models
- vine structure

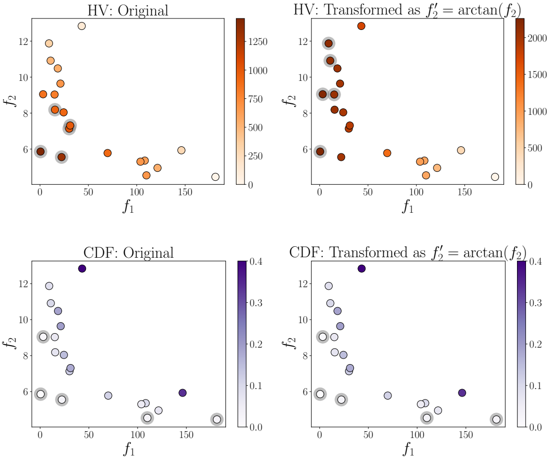




$$f(x_1, x_2, x_3, x_4) = c_{14|23} \cdot c_{13|2} \cdot c_{24|3} \cdot c_{12} \cdot c_{23} \cdot c_{34} \cdot f_1(x_1) \cdot f_2(x_2) \cdot f_3(x_3) \cdot f_4(x_4).$$

Desirable invariance properties

CDF is invariant to arbitrary monotonic transformations of objectives, while HV is very sensitive. Important due to common unit conversions (e.g., linear μm $\rightarrow$ nm, loglike KD $\rightarrow$ pKD to remove tails)!



MOBO acquisition function

The acquisition function $a^{\hat{f}} : \mathcal{X} \to \mathbb{R}$ quantifies the expected utility of each design based on predictions by the surrogate $\hat{f}$.

MOBO acquisition function

Recall each cycle of the MOBO algorithm:

1. Fitting the surrogate to obtain $p(\hat{f}|\mathcal{D})$
2. Optimizing to obtain $x^* = \mathrm{argmax}_{x \in \mathcal{X}}\ a^{\hat{f}}(x)$
   - Gradient-based (exact or estimated)
   - Gradient-free[11] ✓
3. Appending the resulting measurement: $\mathcal{D} \leftarrow \mathcal{D} \cup \{(x^*, f(x^*))\}$

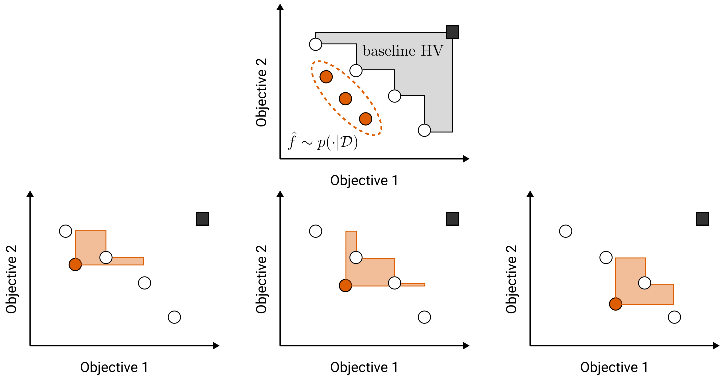
[11] Hansen, "The CMA evolution strategy: a comparing review" (2006).

Acquisition function: general form

$$a^{\hat{f}}(x) = \mathbb{E}_{\hat{f} \sim p(\cdot|\mathcal{D})}[\ \underbrace{u^{\hat{f}}(x)}_{\text{utility}}\ ] = \int u^{\hat{f}}(x)\ p(\hat{f}|\mathcal{D}) d\hat{f}$$

Expected hypervolume improvement (EHVI)[12]

$$a_{EHVI}(x) = \int \underbrace{\left[I_{\mathrm{HV}}\left(\hat{P} \cup \{\hat{f}(x)\}\right) - I_{\mathrm{HV}}\left(\hat{P}\right) \right]}_{\substack{\text{HV improvement:} \\ \text{how much } x \text{ is predicted to improve on current PF}}} p(\hat{f}|\mathcal{D}) d\hat{f}$$



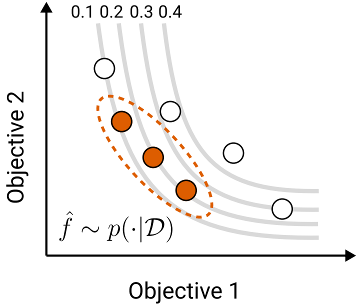
[12] Emmerich, Deutz, and Klinkenberg, "Hypervolume-based expected improvement: Monotonicity properties and exact computation" (2011).

Acquisition function based on the CDF indicator

We propose BOtied, the expected CDF:

$$a_{CDF}(x) = \int \left[1 - \hat{F}_Y\left(\hat{f}(x)\right)\right] p(\hat{f}|\mathcal{D})d\hat{f},$$

where $\hat{F}_Y$ is the CDF fit on $\{y : (x, y) \in \mathcal{D}\} \cup \{\hat{f}(\mathcal{X})\}$.*



*In practice, we draw samples $x' \sim \mathcal{X}$.

Content

Background

Method

- Connection between the CDF ranks and the Pareto front
- BOtied: MOBO based on the CDF

Empirical results

Empirical results

BOtied outperforms EHVI on standard synthetic benchmark problems for MOBO, even in terms of HV.

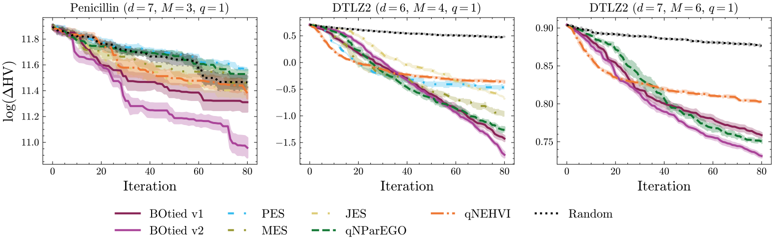


Figure 2: Metric vs. iterations for two synthetic problems.
Metric: $\log(\Delta \mathrm{HV}) := \log\left(HV(\mathcal{P}) - HV(\hat{\mathcal{P}})\right)$ (lower is better)

Empirical results

BOtied effectively acquires samples along the Pareto front for the Branin-Currin[13] (d=2, M=2) test function.

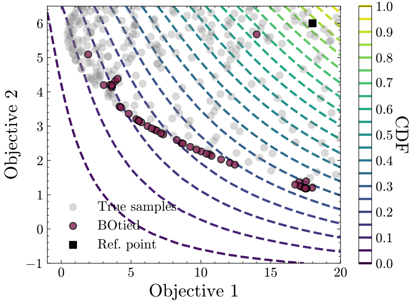
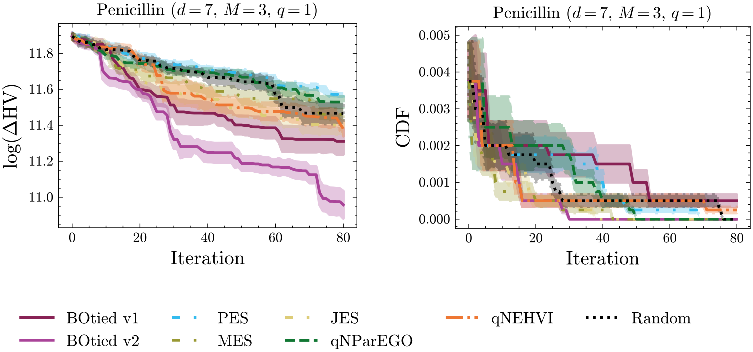


Figure 3: **1K samples from Branin-Currin** overlaid with **BOtied-acquired $\hat{\mathcal{P}}$** and level lines of CDF fit on 86 datapoints at final iteration (not shown).

[13]Belakaria, Deshwal, and Doppa, "Max-value entropy search for multi-objective Bayesian optimization" (2019).

Empirical results

As metrics, the CDF and HV indicators are consistent.



Empirical results

BOtied outperforms EHVI on a real-world dataset of measured hemical properties carrying long tails.

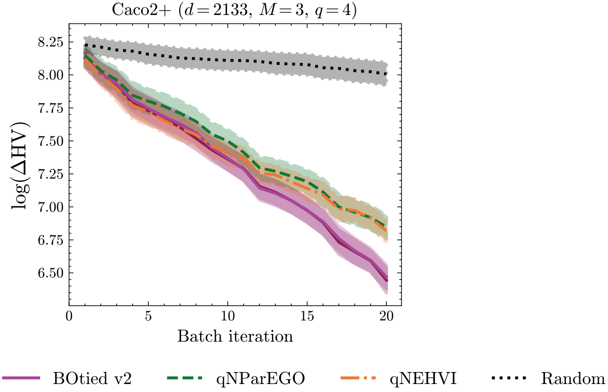
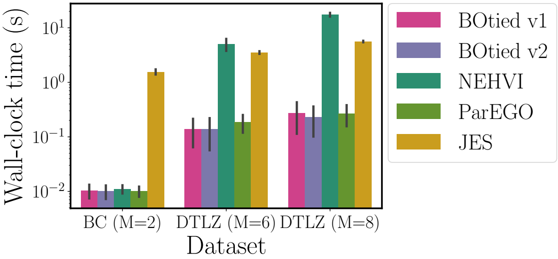


Figure 4: Metric vs. iterations for the modified Caco2 dataset

Computational efficiency

- Vine copula implementation makes BOtied very fast relative to EHVI and joint entropy search (JES), both involving M-dim integrals
- BOtied has competitive wall-clock time with ParEGO, which randomly scalarizes the objectives (effectively $M = 1$)

Per function evaluation:



Summary and outlook

We propose BOtied, a multi-objective acquisition function that leverages CDF-based multivariate ranking.

Summary and outlook

We propose BOtied, a multi-objective acquisition function that leverages CDF-based multivariate ranking.

- efficiently implemented using vine copulas for $M > 3$ objectives

Summary and outlook

We propose BOtied, a multi-objective acquisition function that leverages CDF-based multivariate ranking.

- efficiently implemented using vine copulas for $M > 3$ objectives
- invariant to monotonic transformations of objectives

Summary and outlook

We propose BOtied, a multi-objective acquisition function that leverages CDF-based multivariate ranking.

- efficiently implemented using vine copulas for $M > 3$ objectives
- invariant to monotonic transformations of objectives
- enables integration of domain knowledge in model-based construction of Pareto front

Summary and outlook

We propose BOtied, a multi-objective acquisition function that leverages CDF-based multivariate ranking.

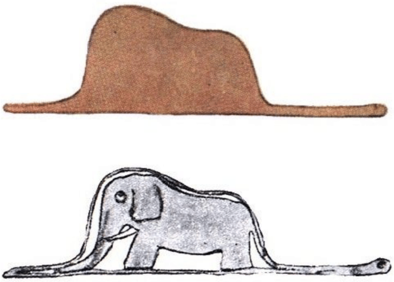
- efficiently implemented using vine copulas for $M > 3$ objectives
- invariant to monotonic transformations of objectives
- enables integration of domain knowledge in model-based construction of Pareto front

Framework is general → hierarchical Bayesian inference, constrained, mixed/discrete...
Follow-up work in progress: *differentiable* BOtied for efficient gradient-based optimization over high-dimensional design space $\mathcal{X}$ (guided sampling)

$$\max_{x \in \mathcal{X}} a(x)$$

Thank you!

park.ji_won@gene.com



References

Jain, Tushar et al. **"Biophysical properties of the clinical-stage antibody landscape".** In: *Proceedings of the National Academy of Sciences* 114.5 (2017), pp. 944–949.

Wang, Ning-Ning et al. **"ADME properties evaluation in drug discovery: prediction of Caco-2 cell permeability using a combination of NSGA-II and boosting".** In: *Journal of Chemical Information and Modeling* 56.4 (2016), pp. 763–773.

Jones, Donald R, Matthias Schonlau, and William J Welch. **"Efficient global optimization of expensive black-box functions".** In: *Journal of Global optimization* 13 (1998), pp. 455–492.

Konakovic Lukovic, Mina, Yunsheng Tian, and Wojciech Matusik. **"Diversity-guided multi-objective bayesian optimization with batch evaluations".** In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 17708–17720.

Warburton, Arthur R. **"Quasiconcave vector maximization: connectedness of the sets of Pareto-optimal and weak Pareto-optimal alternatives".** In: *Journal of optimization theory and applications* 40 (1983), pp. 537–557.

Emmerich, Michael TM, André H Deutz, and Jan Willem Klinkenberg. **"Hypervolume-based expected improvement: Monotonicity properties and exact computation".** In: *2011 IEEE Congress of Evolutionary Computation (CEC)*. IEEE. 2011, pp. 2147–2154.

Yang, Kaifeng et al. **"A multi-point mechanism of expected hypervolume improvement for parallel multi-objective bayesian global optimization".** In: *Proceedings of the Genetic and Evolutionary Computation Conference*. 2019, pp. 656–663.

Binois, Mickaël, Didier Rullière, and Olivier Roustant. **"On the estimation of Pareto fronts from the point of view of copula theory".** In: *Information Sciences* 324 (2015), pp. 270–285.

Park, Ji Won et al. **"BOtied: Multi-objective Bayesian optimization with tied multivariate ranks".** In: *arXiv preprint arXiv:2306.00344* (2023).

Joe, Harry. ***Multivariate Models and Dependence Concepts*.** Chapman & Hall/CRC, 1997.

Hansen, Nikolaus. **"The CMA evolution strategy: a comparing review".** In: *Towards a new evolutionary computation: Advances in the estimation of distribution algorithms* (2006), pp. 75–102.

Belakaria, Syrine, Aryan Deshwal, and Janardhan Rao Doppa. **"Max-value entropy search for multi-objective Bayesian optimization".** In: *Advances in neural information processing systems* 32 (2019).

Huard, David, Guillaume Evin, and Anne-Catherine Favre. **"Bayesian copula selection".** In: *Computational Statistics & Data Analysis* 51.2 (2006), pp. 809–822.