Multi-criteria Optimization

Bayesian Optimization

Bernd Bischl   Frank Hutter   Lars Kotthoff
Marius Lindauer   Joaquin Vanschoren

Recap: Bayesian Optimization I

Advantages of BO

- Sample efficient
- Can handle noise
- Native incorporation of priors
- Does not require gradients
- Theoretical guarantees

We will now extend BO to multiple cost functions.

Recap: Bayesian Optimization II

Bayesian optimization loop

Require: Search space $\mathbf{\Lambda}$, cost function c, acquisition function u, predictive model $\hat{c}$, maximal number of function evaluations T

Result : Best configuration $\hat{\boldsymbol{\lambda}}$ (according to $\mathcal{D}$ or $\hat{c}$)

```
Initialize data D^(0) with initial observations
for t = 1 to T do
    Fit predictive model ĉ^(t) on D^(t-1)
    Select next query point: λ^(t) ∈ argmax_{λ∈Λ} u(λ; D^(t-1), ĉ^(t))
    Query c(λ^(t))
    Update data: D^(t) ← D^(t-1) ∪ {⟨λ^(t), c(λ^(t))⟩}
```

Initialize data $\mathcal{D}^{(0)}$ with initial observations

for $t = 1$ **to** T **do**

- Fit predictive model $\hat{c}^{(t)}$ on $\mathcal{D}^{(t-1)}$
- Select next query point: $\boldsymbol{\lambda}^{(t)} \in \arg\max_{\boldsymbol{\lambda} \in \mathbf{\Lambda}} u(\boldsymbol{\lambda}; \mathcal{D}^{(t-1)}, \hat{c}^{(t)})$
- Query $c(\boldsymbol{\lambda}^{(t)})$
- Update data: $\mathcal{D}^{(t)} \leftarrow \mathcal{D}^{(t-1)} \cup \{\langle \boldsymbol{\lambda}^{(t)}, c(\boldsymbol{\lambda}^{(t)}) \rangle\}$

Multi-Criteria Bayesian Optimization

Goal: Extend Bayesian optimization to multiple cost functions

$$\min_{\boldsymbol{\lambda} \in \boldsymbol{\Lambda}} c(\boldsymbol{\lambda}) \Leftrightarrow \min_{\boldsymbol{\lambda} \in \boldsymbol{\Lambda}} \left(c_1(\boldsymbol{\lambda}), c_2(\boldsymbol{\lambda}), ..., c_m(\boldsymbol{\lambda})\right).$$

There are two basic approaches:

1. Simplify the problem by scalarizing the cost functions, or
2. define acquisition functions for multiple cost functions.

Scalarization

Idea: Aggregate all cost functions

$$\min_{\boldsymbol{\lambda} \in \boldsymbol{\Lambda}} \sum_{i=1}^{m} w_i c_i(\boldsymbol{\lambda}) \qquad \text{with} \quad w_i \geq 0$$

- **Obvious problem:** How to choose $w_1, \ldots, w_m$?
  - Expert knowledge?
  - Systematic variation?
  - Random variation?
- If expert knowledge is not available a-priori, we need to ensure that different trade-offs between cost functions are explored.
- Simplifies multi-criteria optimization problem to single-objective
  $\longrightarrow$ Bayesian optimization can be used without adaption of the general algorithm.

Scalarization: ParEGO [Knowles et al. 2006]

Scalarize the cost functions using the augmented Tchebycheff norm / achievement function

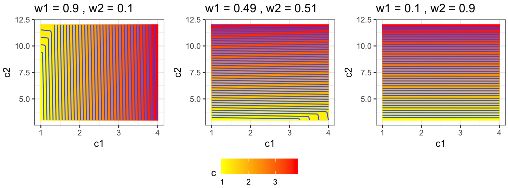
$$c = \max_{i=1,\ldots,m} (w_i c_i(\boldsymbol{\lambda})) + \rho \sum_{i=1}^{m} w_i c_i(\boldsymbol{\lambda}),$$

- The weights $w \in W$ are drawn from

$$W = \left\{ w = (w_1, \ldots, w_m) | \sum_{i=1}^{m} w_i = 1, w_i = \frac{l}{s} \wedge, l \in 0, \ldots, s \right\},$$

  with $|W| = \binom{s+m-1}{k-1} 1$.
- New weights are drawn in every BO iteration.
- ρ is a small parameter suggested to be set to 0.05.
- s selects the number of different weights to draw from.

Why the Tchebycheff norm?



$$c = \max_{i=1,\ldots,m} \left(w_i c_i(\boldsymbol{\lambda}) \right) + \rho \sum_{i=1}^{m} w_i c_i(\boldsymbol{\lambda}),$$

- The norm consists of two components:
  - $\max_{i=1,\ldots,m} \left(w_i c_i(\boldsymbol{\lambda}) \right)$ takes only the maximum weighted cost into account.
  - $\sum_{i=1}^{m} w_i c_i(\boldsymbol{\lambda})$ is the weighted sum of all cost functions.
- ρ describes the trade-off between these components.
- By the randomized weights in each iteration and the usually small value of $\rho = 0.05$, this allows exploration of extreme points of single cost functions.
- One can prove: **Every solution of the scalarized problem is pareto-optimal!**

ParEGO Algorithm

ParEGO loop

Require: Search space $\boldsymbol{\Lambda}$, cost function c, acquisition function u, predictive model $\hat{c}$, maximal number of function evaluations T, ρ, l, s

Result : Best configuration $\hat{\boldsymbol{\lambda}}$ (according to $\mathcal{D}$ or $\hat{c}$)

```
Initialize data 𝒟^(0) with initial observations
for t = 1 to T do
    Sample w from { w = (w_1, ..., w_m) | Σ_{i=1}^m w_i = 1, w_i = l/s ∧, l ∈ 0, ..., s };
    Compute scalarization c^(t) = max_{i=1,...,m} (w_i c_i(λ)) + ρ Σ_{i=1}^m w_i c_i(λ);
    Fit predictive model ĉ^(t) on 𝒟^(t−1)
    Select next query point: λ^(t) ∈ arg max_{λ∈Λ} u(λ; 𝒟^(t−1), ĉ^(t))
    Query c(λ^(t))
    Update data: 𝒟^(t) ← 𝒟^(t−1) ∪ {⟨λ^(t), c(λ^(t))⟩}
```

Initialize data $\mathcal{D}^{(0)}$ with initial observations

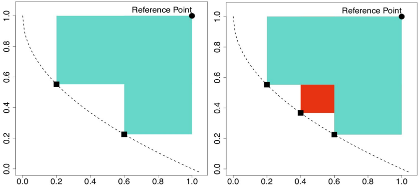
for $t = 1$ **to** T **do**

- Sample w from $\left\{w = (w_1, \ldots, w_m) | \sum_{i=1}^{m} w_i = 1, w_i = \frac{l}{s} \wedge, l \in 0, \ldots, s\right\}$;
- Compute scalarization $c^{(t)} = \max_{i=1,\ldots,m} (w_i c_i(\boldsymbol{\lambda})) + \rho \sum_{i=1}^{m} w_i c_i(\boldsymbol{\lambda})$;
- Fit predictive model $\hat{c}^{(t)}$ on $\mathcal{D}^{(t-1)}$
- Select next query point: $\boldsymbol{\lambda}^{(t)} \in \arg\max_{\boldsymbol{\lambda} \in \boldsymbol{\Lambda}} u(\boldsymbol{\lambda}; \mathcal{D}^{(t-1)}, \hat{c}^{(t)})$
- Query $c(\boldsymbol{\lambda}^{(t)})$
- Update data: $\mathcal{D}^{(t)} \leftarrow \mathcal{D}^{(t-1)} \cup \{\langle \boldsymbol{\lambda}^{(t)}, c(\boldsymbol{\lambda}^{(t)}) \rangle\}$

Hypervolume based Acquisition Functions

Idea: Define acquisition function that directly models contribution to dominated HV.

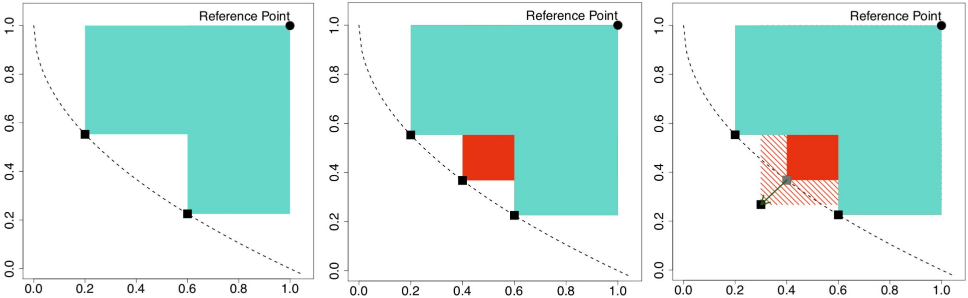
$$\max(0, S(\mathcal{P} \cup \boldsymbol{\lambda}, R) - S(\mathcal{P}, R))$$



- Fit m single-objective surrogate models $\hat{c}_1, \ldots, \hat{c}_m$
- Acquisition function takes all surrogate models into account.
- Single-criteria optimization of acquisition function.

S-Metric Selection-based EGO I

Using the Lower Confidence bound $u_{\text{LCB},1}(\boldsymbol{\lambda}), \ldots, u_{\text{LCB},m}(\boldsymbol{\lambda})$, an optimistic estimate of hypervolume contribution can be calculated.



S-Metric Selection-based EGO II

Problem: Based on the way the hypervolume contribution is measured large plateaus of zero improvement are present.

- These make optimization much harder.
- An adaptive penalty is added to regions in which the lower confidence bound is dominated.

This method is referred to as SMS-EGO [Ponweiser et al. 2008].

Further Hypervolume based Acquisition Functions

Expected Hypervolume Improvement (EHI) [Yang et al. 2019]

$$u_{EI,\mathcal{H}}(\boldsymbol{\lambda}) = \int_{-\infty}^{\infty} p(c \mid \boldsymbol{\lambda}) \times \mathcal{H}(\boldsymbol{\lambda}) \;\; dc,$$

with $\mathcal{H}(\boldsymbol{\lambda}) = S(\mathcal{P} \cup \boldsymbol{\lambda}, R) - S(\mathcal{P}, R))$.

- Direct extension of u_{EI} to the hypervolume.
- $p(c \mid \boldsymbol{\lambda})$ is the joint density of the surrogate model predictions at $\boldsymbol{\lambda}$.
- As the surrogates are GPs and modeled independently of each other, this is just an integral over m univariate normal distributions.
- Efficient computations for $m \leq 3$ exist, beyond that expensive simulation-based computation is required.

Further hypervolume based acquisition functions:

- **Stepwise Uncertainty Reduction** (SUR) based on the probability of improvement.
- **Expected Maximin Improvement** (EMI) based on the ϵ-indicator.

Hypervolume based BO Algorithm

Hypervolume based Bayesian optimization loop

Require: Search space $\mathbf{\Lambda}$, cost function c, acquisition function u, predictive model $\hat{c}$, maximal number of function evaluations T

Result : Best configuration $\hat{\boldsymbol{\lambda}}$ (according to $\mathcal{D}$ or $\hat{c}$)

1 Initialize data $\mathcal{D}^{(0)}$ with initial observations

2 **for** $t = 1$ **to** T **do**

3 Fit predictive models $\hat{c}_1^{(t)}, \dots, \hat{c}_m^{(t)}$ on $\mathcal{D}^{(t-1)}$

4 Select next query point: $\boldsymbol{\lambda}^{(t)} \in \arg\max_{\boldsymbol{\lambda} \in \mathbf{\Lambda}} u(\boldsymbol{\lambda}; \mathcal{D}^{(t-1)}, \hat{c}_1^{(t)}, \dots, \hat{c}_m^{(t)}))$

5 Query $c(\boldsymbol{\lambda}^{(t)})$

6 Update data: $\mathcal{D}^{(t)} \leftarrow \mathcal{D}^{(t-1)} \cup \{\langle \boldsymbol{\lambda}^{(t)}, c(\boldsymbol{\lambda}^{(t)}) \rangle\}$