Multi-criteria Optimization

Introduction

Bernd Bischl   Frank Hutter   Lars Kotthoff
Marius Lindauer   Joaquin Vanschoren

Introductory example I

Often we want to solve optimization problems concerning several goals.

General applications:

- Medicine: maximum effect, but minimum side effect of a drug.
- Finances: maximum return, but minimum risk of an equity portfolio.
- Production planning: maximum revenue, but minimum costs.
- Booking a hotel: maximum rating, but minimum costs.

In machine learning:

- Sparse models: maximum predictive performance, but minimal number of features.
- Fast models: maximum predictive performance, but short prediction time.
- ...

Introductory example II

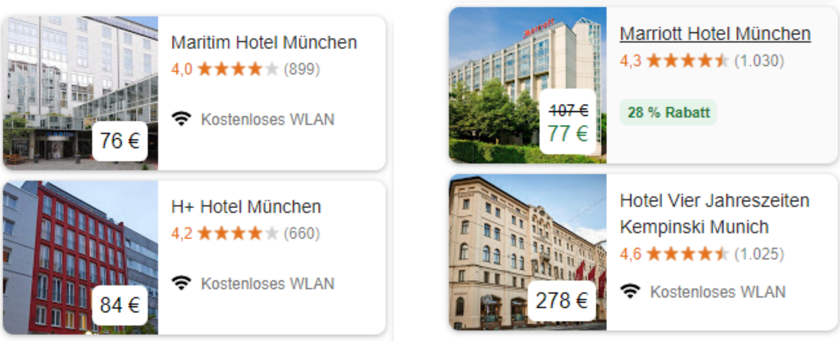
Example:
Choose the best hotel to stay at by maximizing ratings subject to a maximum price per night.

Problems:

- The result depends on how we select the maximum price and usually returns different solutions for different maximum price values.
- We could also choose a minimum rating and optimize the price per night.
- The more objectives we optimize, the more difficult such a definition becomes.

Goal:
Find a more general approach to solve multi-criteria problems.

Introductory example III



When booking a hotel: find the hotel with

- minimum price per night (**costs**) and
- maximum user rating (**performance**).

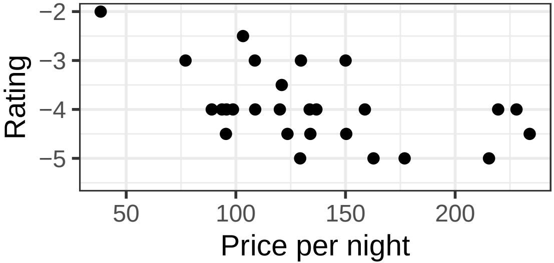
Since our standard is to minimize objectives, we minimize negative ratings.

Introductory example IV

The objectives often conflict with each other:

- Lower price $\rightarrow$ usually lower hotel rating.
- Better rating $\rightarrow$ usually higher price.

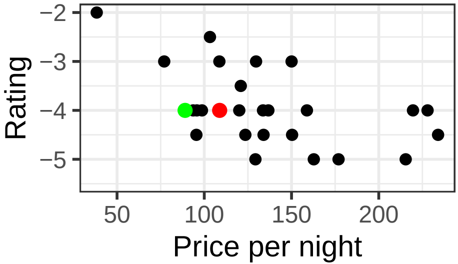
Example: (negative) average rating by hotel guests (1 - 5) vs. average price per night (excerpt).

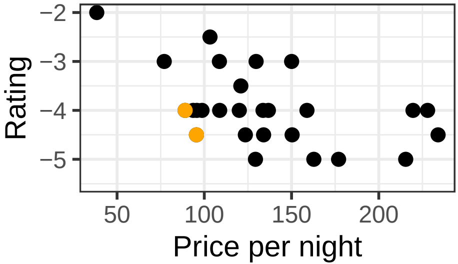


Introductory example V

Often, objectives are not directly comparable as they are measured on different scales:

- Left: A hotel with rating 4 for 89 Euro ($c^{(1)} = (89, -4.0)$) would be preferred to a hotel for 108 Euro with the same rating ($c^{(2)} = (108, -4.0)$).
- Right: How to decide if $c^{(1)} = (89, -4.0)$ or $c^{(1)} = (95, -4.5)$ is preferred?
- How much is one *rating point* worth?





Definition: multi-criteria optimization problem

A **multi-criteria optimization problem** is defined by

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

with $\boldsymbol{\Lambda} \subset \mathbb{R}^n$ and multi-criteria objective function $c : \boldsymbol{\Lambda} \to \mathbb{R}^m$, $m \geq 2$.

- **Goal:** minimize multiple target functions simultaneously.
- $(c_1(\boldsymbol{\lambda}), ..., c_m(\boldsymbol{\lambda}))^\top$ maps each candidate $\boldsymbol{\lambda}$ into the objecive space $\mathbb{R}^m$.
- Often no clear best solution, as objective are usually conflicting and we cannot totally order in $\mathbb{R}^m$.
- W.l.o.g. we always minimize.
- Alternative names: multi-criteria optimization, multi-objective optimization, Pareto optimization.

Pareto sets and Pareto optimality

Definition:
Given a multi-criteria optimization problem

$$\min_{\boldsymbol{\lambda} \in \boldsymbol{\Lambda}} \left(c_1(\boldsymbol{\lambda}), ..., c_m(\boldsymbol{\lambda}) \right), \quad c_i : \boldsymbol{\Lambda} \to \mathbb{R}.$$

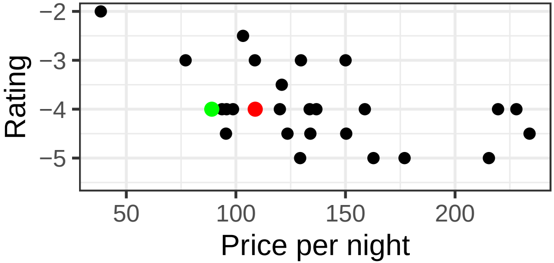
- A candidate $\boldsymbol{\lambda}^{(1)}$ **(Pareto-) dominates** $\boldsymbol{\lambda}^{(2)}$, if $c(\boldsymbol{\lambda}^{(1)}) \prec c(\boldsymbol{\lambda}^{(2)})$, i.e.
  1. $c_i(\boldsymbol{\lambda}^{(1)}) \leq c_i(\boldsymbol{\lambda}^{(2)})$ for all $i \in \{1, 2, ..., m\}$ and
  2. $c_j(\boldsymbol{\lambda}^{(1)}) < c_j(\boldsymbol{\lambda}^{(2)})$ for at least one $j \in \{1, 2, ..., m\}$
- A candidate $\boldsymbol{\lambda}^*$ that is not dominated by any other candidate is called **Pareto optimal**.
- The set of all Pareto optimal candidates is called **Pareto set**
  $\mathcal{P} := \{\boldsymbol{\lambda} \in \boldsymbol{\Lambda} | \nexists\ \tilde{\boldsymbol{\lambda}}$ with $c(\tilde{\boldsymbol{\lambda}}) \prec c(\boldsymbol{\lambda})\}$
- $\mathcal{F} = c(\mathcal{P}) = \{c(\boldsymbol{\lambda}) | \boldsymbol{\lambda} \in \mathcal{P}\}$ is called **Pareto front**.

How to define optimality? I

Let $c = (\text{price}, -\text{rating})$. For some cases it is *clear* which point is the better one:

- The candidate $c^{(1)} = (89, -4.0)$ dominates $c^{(2)} = (108, -4.0)$: $c^{(1)}$ is not worse in any dimension and is better in one dimension. Therefore, $c^{(2)}$ gets **dominated** by $c^{(1)}$

$$c^{(2)} \prec c^{(1)}.$$

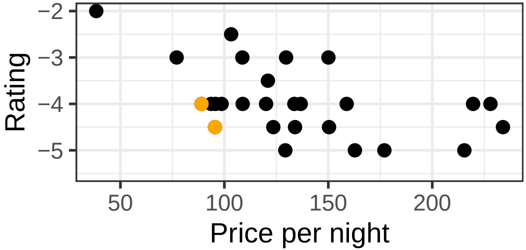


How to define optimality? II

For the points $c^{(1)} = (89, -4.0)$ and $c^{(2)} = (95, -4.5)$ we cannot say which one is better.

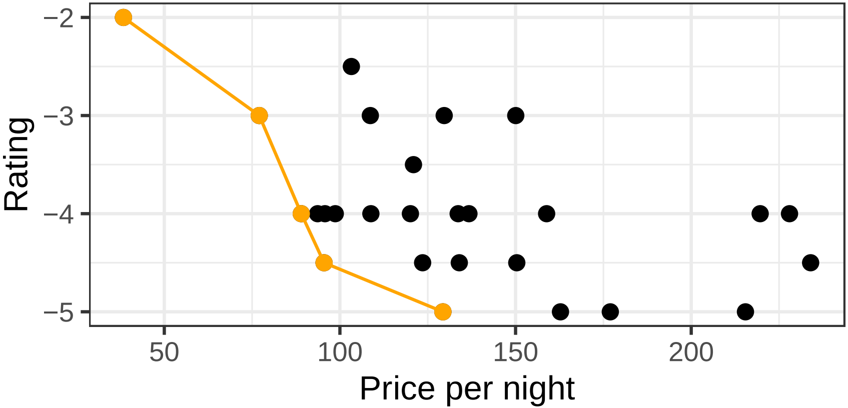
- We define the points as **equivalent** and write

$$c^{(1)} \not\prec c^{(2)} \text{ and } c^{(2)} \not\prec c^{(1)}.$$



How to define optimality? III

- The set of all equivalent points that are not dominated by another point is called the **Pareto front**.

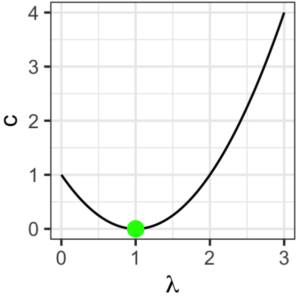


Example: One objective function

We consider the minimization problem

$$\min_{\boldsymbol{\lambda}} c(\boldsymbol{\lambda}) = (\boldsymbol{\lambda} - 1)^2, \qquad 0 \leq \boldsymbol{\lambda} \leq 3.$$

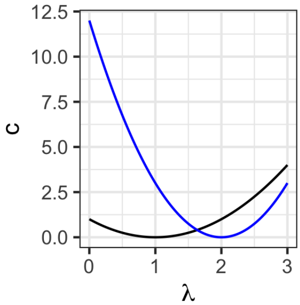
The optimum is at $\boldsymbol{\lambda}^* = 1$.



Example: Two target functions I

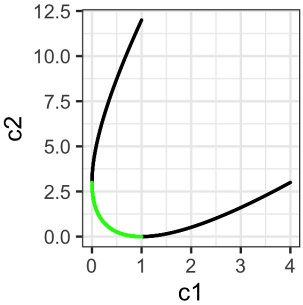
We extend the above problem to two objective functions $c_1(\boldsymbol{\lambda}) = (\boldsymbol{\lambda} - 1)^2$ and $c_2(\boldsymbol{\lambda}) = 3(\boldsymbol{\lambda} - 2)^2$, thus

$$\min_{\boldsymbol{\lambda}} c(\boldsymbol{\lambda}) = \left(c_1(\boldsymbol{\lambda}), c_2(\boldsymbol{\lambda})\right), \qquad 0 \leq \boldsymbol{\lambda} \leq 3.$$



Example: Two target functions II

We consider the functions in the objective function space $c(\mathbf{\Lambda})$ by drawing the objective function values $(c_1(\boldsymbol{\lambda}), c_2(\boldsymbol{\lambda}))$ for all $0 \leq \boldsymbol{\lambda} \leq 3$.



The Pareto front is shown in green. The Pareto front cannot be *left* without getting worse in at least one objective function.

A-priori vs. A-posteriori

- The Pareto set is a set of equally optimal solutions.
- In many applications one is often interested in a **single** optimal solution.
- Without further information no unambiguous optimal solution can be determined.
  $\rightarrow$ The decision must be based on other criteria.

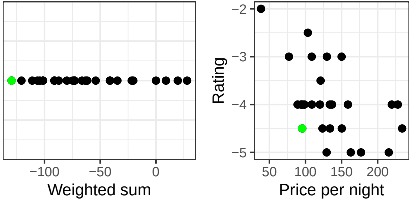
There are two possible approaches:

- **A-priori approach**: User preferences are considered **before** the optimization process
- **A-posteriori approach**: User preferences are considered **after** the optimization process

A-priori procedure I

Example: Weighted total
Prior knowledge: One rating point is worth 50 Euro to a customer.
$\rightarrow$ We optimize the weighted sum:

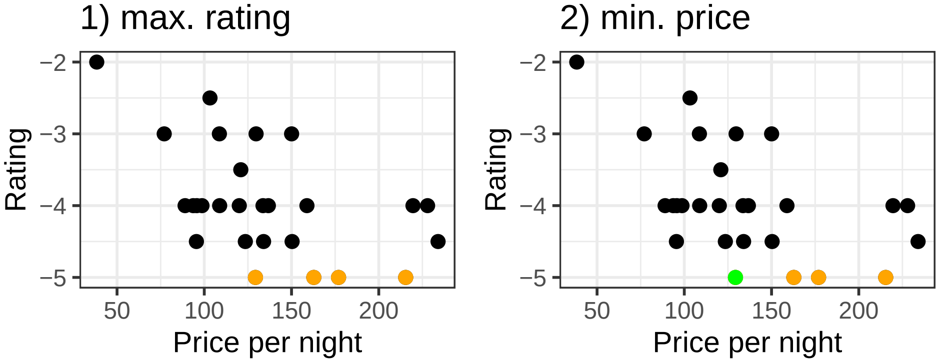
$$\min_{\text{Hotel}} (\text{Price / Night}) - 50 \cdot \text{Rating}$$



Alternative a weighted sum: $\min_{\boldsymbol{\lambda} \in \boldsymbol{\Lambda}} \sum_{i=1}^{m} w_i c_i(\boldsymbol{\lambda})$ with $w_i \geq 0$

A-priori procedure II

Example: Lexicographic method
Prior knowledge: Customer prioritizes rating over price.
$\rightarrow$ Optimize target functions one after the other.



A-priori procedure III

A-priori approach: Lexicographic method

$$
\begin{aligned}
c_1^* &= \min_{\boldsymbol{\lambda} \in \boldsymbol{\Lambda}} c_1(\boldsymbol{\lambda}) \\
c_2^* &= \min_{\boldsymbol{\lambda} \in \{\boldsymbol{\lambda} \mid c_1(\boldsymbol{\lambda}) = c_1^*\}} c_2(\boldsymbol{\lambda}) \\
c_3^* &= \min_{\boldsymbol{\lambda} \in \{\boldsymbol{\lambda} \mid c_1(\boldsymbol{\lambda}) = c_1^* \wedge c_2(\boldsymbol{\lambda}) = c_2^*\}} c_3(\boldsymbol{\lambda}) \\
&\vdots
\end{aligned}
$$

But: Different sequences provide different solutions.

A-priori procedure IV

Summary a-priori approach:

- Implicit assumption: Single-objective optimization is *easy*.
- Only one solution is obtained, which depends on a-priori weights, order, etc.
- Several solutions can be obtained if weights, order, etc. are systematically varied.
- Usually not all non-dominated candidates can be found by these methods.

A-posteriori procedure I

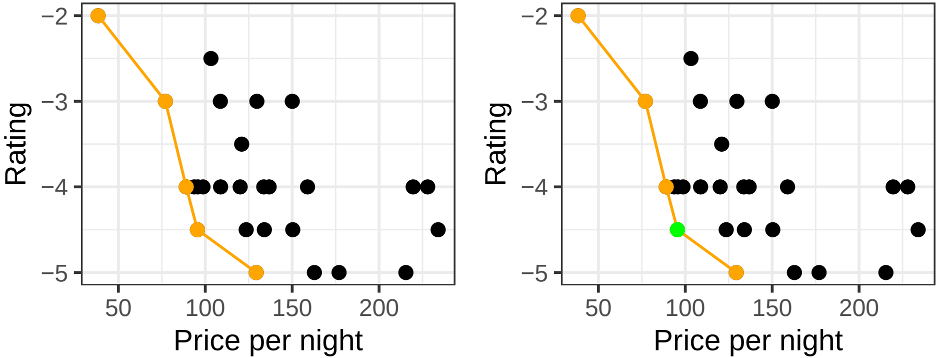
A-posteriori methods try to

- find the set of **all** optimal candidates (the Pareto set),
- select (if necessary) an optimal candidate based on prior knowledge or individual preferences.
- Implicit assumption: Specifying your hidden preferences / making a selection from a pool of candidates is easier, if you see the non-dominated solutions.

A-posteriori methods are therefore the more generic approach to solving a multi-criteria optimization problem.

A-posteriori procedure II

Example: A user is displayed all Pareto optimal hotels (left) and chooses an optimal candidate (right) based on his hidden preferences or additional criteria (e.g. location of the hotel).

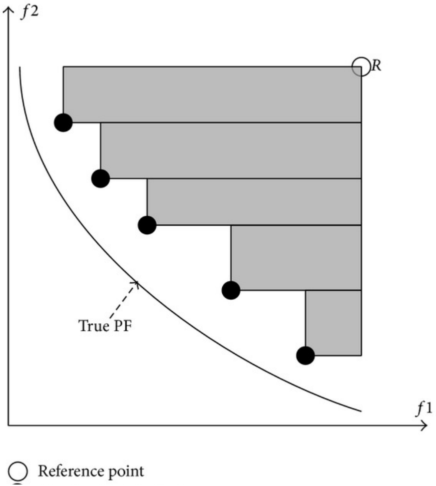


Evaluation of solutions I

A common metric for evaluating the performance of a set of candidates $\mathcal{P} \subset \mathbf{\Lambda}$ is the **dominated hypervolume**

$$S(\mathcal{P}, R) = \Lambda\left(\bigcup_{\tilde{\boldsymbol{\lambda}} \in \mathcal{P}} \left\{\boldsymbol{\lambda} | \tilde{\boldsymbol{\lambda}} \prec \boldsymbol{\lambda} \prec R\right\}\right),$$

where Λ is the Lebuesge measure.



Evaluation of solutions II

- HV is calculated w.r.t the reference point R, which often reflects in each component the natural maximum of the respective objective – if possible
- The dominated hypervolume is also often called **S-Metric**.
- Computation of HV scales exponentially in the number of objective functions $\mathcal{O}(n^{m-1})$.
- Fast approximations exist for small values of m and especially for machine learning applications we rarely optimize $m > 3$ objectives.