A divide and conquer approach to determine the Pareto frontier for optimization of protein engineering experiments

L. He, A.M. Friedman, and C. Bailey-Kellogg

Supplementary Material

Divide-and-conquer for two objectives

Algorithm 1 provides pseudocode for the algorithm.

Algorithm 1 PEPFR divide-and-conquer algorithm for two objectives with fixed $\boldsymbol{\alpha}$

```
Divide-And-Conquer($B = (l_1, u_1) \times (l_2, u_2)$):
$F \leftarrow \emptyset$        // Pareto frontier
// Conquer
$\lambda \leftarrow g(\boldsymbol{\alpha}, B)$
if $\lambda$ exists then
    $F \leftarrow F \cup \{\lambda\}$
    // Divide
    $\mathbf{p} \leftarrow \mathbf{f}(\lambda)$
    $B_{01} \leftarrow (l_1, p_1) \times (p_2, u_2)$
    $B_{10} \leftarrow (p_1, u_1) \times (l_2, p_2)$
    for all $B'$ in $\{B_{01}, B_{10}\}$ do
        $F' \leftarrow$ Divide-And-Conquer($B'$)
        $F \leftarrow F \cup F'$
    end for
end if
return $F$
```

The conquer step depends on the fact that the result of g is undominated within B. We now prove that.

Claim 1. *The λ in B minimizing hyperplane function $\boldsymbol{\alpha} \cdot \mathbf{f}(\lambda)$, for any $\boldsymbol{\alpha} \in (\mathbb{R}^+)^2$, is undominated within B.*

Proof. Let us assume for contradiction that λ' in B dominates λ. According to the definition of "dominates", $\forall i \in \{1, 2\}$, $f_i(\lambda') \leq f_i(\lambda)$ and $\exists j \in \{1, 2\}$, $f_j(\lambda') < f_j(\lambda)$. Therefore $\sum_{i=1}^{2} \alpha_i f_i(\lambda') < \sum_{i=1}^{2} \alpha_i f_i(\lambda)$, so $\boldsymbol{\alpha} \cdot \mathbf{f}(\lambda') < \boldsymbol{\alpha} \cdot \mathbf{f}(\lambda)$, contradicting the optimality of λ. □

Divide-and-conquer for multiple design objectives

We extend the hierarchical divide-and-conquer approach for two objectives (in $\mathbb{R}^2$) to handle multiple objectives (in $\mathbb{R}^n$). The same basic structure holds.

Initialize. This works exactly as in the 2-objective case: independently minimize and maximize each objective, and relax the determined minimum and maximum values by ϵ, to yield an open box $B = (\min(f_1) - \epsilon, \max(f_1) + \epsilon) \times (\min(f_2) - \epsilon, \max(f_2) + \epsilon) \times \ldots \times (\min(f_n) - \epsilon, \max(f_n) + \epsilon)$.

Conquer. Claim 1 can be straightforwardly generalized to three or more objectives. Our instantiations of box-constrained optimization via integer programming and dynamic programming are also immediately extensible to optimize linear combinations of multiple objectives. The one tricky part is guaranteeing the overall Pareto optimality of a discovered design that is undominated within a box; we cover that below, based on a lattice structure of the divided boxes.

Divide. Let λ be the undominated design uncovered by the conquer step in the box $B = (l_1, u_1) \times (l_2, u_2) \times \ldots \times (l_n, u_n)$. Let $\mathbf{p} = (p_1, p_2, \ldots, p_n) = \mathbf{f}(\lambda)$. We use $\mathbf{p}$ to divide B into 2^n portions, denoted $B_{c_1c_2\ldots c_n}$, $c_i \in \{0, 1\}$, where $c_i = 0$ denotes the range (l_i, p_i) and $c_i = 1$ denotes the range (p_i, u_i). Since $\mathbf{p}$ is undominated in B, no design is in $B_{00\ldots0}$ and $\mathbf{p}$ dominates everything in $B_{11\ldots1}$, the same as in two dimensions. However, points in some of the remaining $2^n - 2$ portions may dominate some of those in some others. For example, in a three-dimensional objective space, a Pareto optimal point $\mathbf{p} = (p_1, p_2, p_3)$ divides a box $B = (l_1, u_1) \times (l_2, u_2) \times (l_3, u_3)$ into 2^3 boxes, $B_{000}, B_{001}, \ldots, B_{111}$. Let λ' be a design in B_{001} at point $\mathbf{f}(\lambda') = \mathbf{a} = (a_1, a_2, a_3)$. Now consider a sub-box of B_{011}, defined as $\bar{B} = (a_1, p_1) \times (p_2, u_2) \times (a_3, u_3)$. We can see that λ' dominates every design in $\bar{B}$, being better in all three dimensions (in the second dimension, note that $a_2 < p_2$ by the definition of B_{001}).

More generally, suppose $B_{c_1c_2\ldots c_n}$ and $B_{c'_1c'_2\ldots c'_n}$ are two distinct boxes yielded by dividing box $B = (l_1, u_1) \times (l_2, u_2) \times \ldots \times (l_n, u_n)$ according to Pareto optimal point $\mathbf{p} = (p_1, p_2, \ldots, p_n)$. We have the following claims.

Claim 2. *If $\forall i : c_i \le c'_i$, then point $\mathbf{a} = (a_1, a_2, \ldots, a_n)$ in $B_{c_1c_2,\ldots,c_n}$ dominates sub-box $\bar{B} = (\bar{l}_1, \bar{u}_1) \times (\bar{l}_2, \bar{u}_2) \times \ldots \times (\bar{l}_n, \bar{u}_n)$ in $B_{c'_1c'_2,\ldots,c'_n}$, where*

$$\forall i,\ (\bar{l}_i, \bar{u}_i) = \begin{cases} (a_i, p_i), & \textit{if } c_i = c'_i = 0 \\ (p_i, u_i), & \textit{if } c_i = 0 < c'_i = 1 \\ (a_i, u_i), & \textit{if } c_i = c'_i = 1 \end{cases}$$

Proof. Let $\mathbf{a}' = (a'_1, a'_2, \ldots, a'_n) \in \bar{B}$. Since we assume $\forall i : c_i \le c'_i$, there are three cases:

1. $c_i = 0 < c'_i = 1$. Then $p_i < a'_i < u_i$, $l_i < a_i < p_i$, so $a_i < a'_i$.
2. $c_i = c'_i = 0$. Then $a_i < a'_i < p_i$.
3. $c_i = c'_i = 1$. Then $a_i < a'_i < u_i$.

Since $\forall i : a_i < a'_i$, point $\mathbf{a}$ dominates point $\mathbf{a}'$. Since $\mathbf{a}'$ is an arbitrary point in $\bar{B}$, $\mathbf{a}$ dominates all of $\bar{B}$. □

Claim 3. *If $\exists i : c_i > c'_i$, then no point in $B_{c_1c_2\ldots c_n}$ dominates any point in $B_{c'_1c'_2\ldots c'_n}$.*

Proof. Consider any point $\mathbf{a} = (a_1, a_2, \ldots, a_n)$ in $B_{c_1c_2\ldots c_n}$ and any point $\mathbf{a}' = (a'_1, a'_2, \ldots, a'_n)$ in $B_{c'_1c'_2\ldots c'_n}$. In dimension i, $B_{c_1c_2\ldots c_n}$ has range (p_i, u_i) while $B_{c'_1c'_2\ldots c'_n}$ has range (l_i, p_i). Then since $l_i < a'_i < p_i < a_i < u_i$, we see that $\mathbf{a}$ cannot dominate $\mathbf{a}'$. Consequentially no point in $B_{c_1c_2\ldots c_n}$ dominates any point in $B_{c'_1c'_2\ldots c'_n}$. □

We say that $B_{c_1c_2\ldots c_n}$ and $B_{c'_1c'_2\ldots c'_n}$ are *disjoint* if no point in $B_{c_1c_2\ldots c_n}$ dominates any point of $B_{c'_1c'_2\ldots c'_n}$ and *vice-versa*.

Based on these relationships, we define an order in which to search the divided boxes, in order to eliminate dominated sub-boxes, and ensure that the designs identified in the conquer step are guaranteed to be Pareto optimal. We form a lattice structure (Fig. 1), such that points in ancestor boxes dominate sub-boxes in descendant boxes, but points in descendant boxes do not dominate any points in ancestor boxes, and other pairs of boxes are disjoint. We then recurse into the boxes in breadth-first order according to the lattice structure, applying Claim 2 to eliminate sub-boxes of descendants as we go along. When sub-boxes are eliminated from a box, the remaining sub-boxes are likewise explored in breadth-first order of their lattice structure. Thus by the time we reach a box (or sub-box), all its ancestors have been handled, and all its dominated points already eliminated, so any discovered undominated design is indeed Pareto optimal.

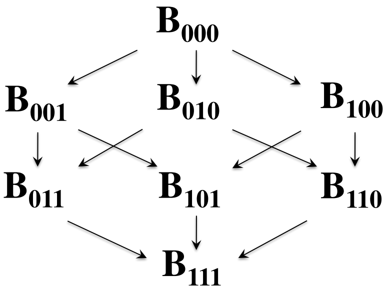

Figure 1: Lattice structure for boxes divided from a single box by a Pareto optimal point $\mathbf{p}$ in $\mathbb{R}^3$. No design exists in B_{000} and designs in B_{111} are all dominated by $\mathbf{p}$. Designs in ancestor boxes can dominate subboxes of descendant boxes (e.g., a point in B_{001} dominates a part of B_{011}), while designs in descendants do not dominate any part of ancestors (e.g., no point in B_{011} dominate any point of B_{001}). Boxes with no ancestor/descendant relationship are disjoint (e.g., B_{001} and B_{110}).