# On the theoretical convergence properties of the SIMP method

José Mario Martínez*

March 4, 2004

**Abstract**

The SIMP (Solid Isotropic Material with Penalization) method is used in Topology Optimization to solve problems where the variables are 0 or 1. The theoretical convergence properties are not exhaustively studied. In this paper a convergence theorem with weaker assumptions than earlier conditions is given. Extensions and counterexamples are also provided.

## 1 Introduction

Many times, in Topology Optimization, one has to decide whether there must be material or not at every point of a two- or three-dimensional region. Usually, the region is divided in 2D-pixels (or 3D-voxels) and to each of them a variable is associated which takes the value 1 if the pixel (voxel) must be filled and 0 otherwise. The decision depends on the solution of an optimization problem. See ([3], Ch. 1). Therefore, the optimization problem is of the form

$$\text{Minimize } f(x) \text{ s.t. } x \in D, \ \ x \in \{0,1\}^n. \tag{1}$$

In (1) $D \subset I\!R^n$ represents the constraints of the problem and $f$ is the objective function. Throughout this paper $\{0,1\}$ is the set whose elements are 0 and 1 and $[0,1]$ is the interval $\{t \in I\!R \mid 0 \leq t \leq 1\}$.

*Institute of Mathematics, University of Campinas, CP 6065, 13081-970 Campinas SP, Brazil. This work was supported by PRONEX, FAPESP (grant 2001-04597-4), CNPq and FAEP-UNICAMP.

Problems of this type are usually solved by means of the SIMP (Solid Isotropic Material with Penalization) method. See [1, 7]. The SIMP method consists in replacing $f(x)$ by $f(x^p)$ (where $x^p = (x_1^p, \ldots, x_n^p)$ and $p > 0$) in (1) and replacing the constraint $x \in \{0,1\}^n$ by $x \in [0,1]^n$. This approach seems to be surprisingly efficient([3], p. 5), although it obviously fails for very simple "toy problems". Suppose that the constraints $x \in D$ of (1) are not modified, so that we obtain the problem:

$$\text{Minimize } f(x^p) \text{ s.t. } x \in D, \quad x \in [0,1]^n. \tag{2}$$

Define, for $n = 1$,

$$f(x) = (x - \frac{1}{4})^2, \quad D = I\!R^1.$$

The unique solution of (1) is $x = 0$. However, the unique solution of (2) tends monotonically to 1 when $p$ tends to infinity. The same behavior occurs if the constraints $x \in D$ are replaced by $x^p \in D$.

Until 2001, no theoretical justification of the SIMP method had been given [2, 4]. Only in 2001, Rietz [4] proved that, under some sufficient conditions, the SIMP method with the formulation (2) is guaranteed to be successful. Rietz conditions are:

(a) The problem (1) has only one constraint, which has the form

$$\sum_{i=1}^{n} x_i = V;$$

(b) The problem (1) has a unique solution.

(c) The function $f$ is differentiable and there exist constants $c_1, c_2 < 0$ such that

$$c_1 \leq \frac{\partial f}{\partial x_i}(x) \leq c_2$$

for all $x \in [0,1]^n$.

Under these assumptions, Rietz proved that, for some $p > 0$, the solution of (2) coincides with the solution of (1). He also observed that his proof applies when, instead of $t^p$, one uses different "penalty" functions with similar properties.

The fact that the SIMP method is used successfully in more general situations than the ones studied by Rietz motivated the present research. In Section 2 we relax Rietz's assumptions (b) and (c). As a result, we prove that, if $p$ is large enough, any solution of (2) approaches to a solution of (1) as much as desired. If, after solving (2), the solution is "rounded" to 0

or 1, a solution of (1) is obtained for a finite value of $p$. This result is also valid for one-dimensional functions different from $t^p$. In Section 3 we remove difficulties related to the lack of fulfillment of the assumptions of Section 2. Conclusions are stated in Section 4.

## 2 A convergence result

Consider the optimization problem

$$\text{Minimize } f(x) \text{ s.t. } \sum_{i=1}^{n} x_i = V, \ \ x \in \{0,1\}^n. \tag{3}$$

We assume that $V \in \{1, \ldots, n\}$. So, the set of points $x \in \{0,1\}^n$ such that $\sum_{i=1}^{n} x_i = V$ is nonempty. Therefore, (3) has a solution, although not necessarily unique.

For all $x, y \in I\!R^n$ we say that $x \leq y$ if $x_i \leq y_i$ for all $i = 1, \ldots, n$. We say that $x < y$ if $x \leq y$ and $x \neq y$. So, if $x < y$ each component of $x$ is less than or equal to the correspondent component of $y$ and, for at least one index $i$, one has $x_i < y_i$.

**Assumption 1.** *The function $f : [0,1]^n \to I\!R$ is continuous and, for all $x, y \in [0,1]^n$,*

$$x < y \Rightarrow f(x) > f(y).$$

It is easy to show that, when $f$ is differentiable and its derivatives satisfy Rietz's assumption

$$c_1 \leq \frac{\partial f}{\partial x_i}(x) \leq c_2 < 0 \ \forall \ x \in [0,1]^n, \tag{4}$$

Assumption 1 also holds. However, Assumption 1 is obviously weaker than (4).

Now, for all $p \in \{1, 2, \ldots\}$ we define the SIMP problem associated to (3). We need an auxiliary family of continuous functions

$$\varphi_p : [0,1] \to [0,1], \ p = 1, 2, \ldots$$

for which the following assumptions hold:

**Assumption 2.** *For all $p = 1, 2, \ldots$,*

$$\varphi_p(0) = 0, \ \varphi_p(1) = 1.$$

**Assumption 3.** *For all* $t \in [0,1]$, $p = 1, 2, \ldots$,

$$\varphi_p(t) \leq t.$$

**Assumption 4.** *For all* $s, t \in [0,1]$, $p = 1, 2, \ldots$,

$$s < t \Rightarrow \varphi_p(s) < \varphi_p(t).$$

**Assumption 5.** *For all* $t \in [0,1)$,

$$\lim_{p \to \infty} \varphi_p(t) = 0.$$

Clearly, the assumptions above are satisfied by the family of functions $\varphi_p(t) = t^p$, which is the one traditionally used in SIMP.

For all $x \in [0,1]^n$ we define

$$\varphi_p(x) = (\varphi_p(x_1), \ldots, \varphi_p(x_n))$$

and

$$g_p(x) = f(\varphi_p(x)).$$

The SIMP problem associated to (3) is:

$$\text{Minimize } g_p(x) \text{ s.t. } \sum_{i=1}^{n} x_i = V, \quad x \in [0,1]^n. \tag{5}$$

Since $f$ is continuous and $\varphi_p$ is continuous for all $p$, the function $g_p$ is continuous for all $p$. Therefore, (5) is the minimization of a continuous function in a nonempty compact set. Then (5) has at least one global solution for all $p$.

In the following theorem we prove that, for $p$ large enough, solutions of (5) are approximations of the solutions of (3).

**Theorem 1.** *Suppose that Assumptions 1–5 hold. For all* $p = 1, 2, \ldots$ *let* $x(p)$ *be a solution of (5). Then, the sequence* $\{x(p), p = 1, 2, \ldots\}$ *admits at*

*least a limit point and every limit point of this sequence is a solution of (3).*

*Proof.* For all $p = 1, 2, \ldots$ the point $x(p)$ belongs to the compact set defined by

$$\sum_{i=1}^{n} x_i = V, \quad x \in [0,1]^n,$$

therefore the sequence admits at least a limit point that belongs to this set.

Let $\bar{x}$ be a limit point of $\{x(p), p = 1, 2, \ldots\}$. Then, there exists an infinite set

$$K \subset \{1, 2, \ldots\}$$

such that

$$\lim_{p \in K} x(p) = \bar{x}.$$

We will prove first that $\bar{x}$ is a feasible point of (3). Since $\sum_{i=1}^{n} x_i(p) = V$ for all $p$, taking limits for $p \in K$ it turns out that

$$\sum_{i=1}^{n} \bar{x}_i = V.$$

Therefore, for proving feasibility, we only need to prove that $\bar{x} \in \{0,1\}^n$.

Suppose, by contradiction, that $\bar{x} \notin \{0,1\}^n$. Then, there exist $r$ components of $\bar{x}$, $r \geq 1$, that are strictly between 0 and 1. Therefore, defining

$$I_0 = \{i \in \{1, \ldots, n\} \mid \bar{x}_i = 0\},$$

$$I_1 = \{i \in \{1, \ldots, n\} \mid \bar{x}_i = 1\},$$

$$I_+ = \{i \in \{1, \ldots, n\} \mid 0 < \bar{x}_i < 1\},$$

we have that $I_+$ is nonempty (it has $r$ elements) and

$$I_0 \cup I_1 \cup I_+ = \{1, \ldots, n\}.$$

Since $\sum_{i=1}^{n} \bar{x}_i = V$ and $\sum_{i \in I_0} \bar{x}_i = 0$ we have that

$$\sum_{i \in I_1 \cup I_+} \bar{x}_i = V.$$

Since $I_+$ is nonempty, it turns out that

$$\sum_{i \in I_1} \bar{x}_i < V.$$

But $\sum_{i\in I_1} \bar{x}_i$ is integer, so

$$W \equiv \sum_{i\in I_0\cup I_1} \bar{x}_i = \sum_{i\in I_1} \bar{x}_i \leq V - 1.$$

Therefore,

$$r \geq \sum_{i\in I_+} \bar{x}_i = V - W \geq 1.$$

Since $V-W$ is an integer and $1 \leq V-W \leq r$ we can choose $V-W$ elements of $I_+$ and store these elements in a set $I_{+1} \subset I_+$. Defining $I_{+0}$ the set of the remaining elements of $I_+$ we have that

$$I_+ = I_{+1} \cup I_{+0},$$

where $I_{+1}$ is nonempty and has $V - W$ elements.

This fact allows us to define a vector $w \in \{0,1\}^n$ such that

$$w_i = 0 \ \ \forall \ \ i \in I_0,$$

$$w_i = 1 \ \ \forall \ \ i \in I_1,$$

$$w_i = 0 \ \ \forall \ \ i \in I_{+0},$$

$$w_i = 1 \ \ \forall \ \ i \in I_{+1}.$$

Clearly,

$$\sum_{i=1}^{n} w_i = V,$$

therefore $w$ is a feasible point of (3).

Let us consider now the sequence $\{\varphi_p(x(p))\}_{p\in K}$. We denote $[\varphi_p(x(p))]_i$ the $i-$th coordinate of $\varphi_p(x(p))$. So,

$$[\varphi_p(x(p))]_i = \varphi_p(x_i(p)), \ i = 1, \ldots, n.$$

For all $i \in I_0$,

$$\bar{x}_i = \lim_{p\in K} x_i(p) = 0.$$

But, by Assumption 3, $\varphi_p(x_i(p)) \leq x_i(p)$, so

$$\lim_{p\in K} \varphi_p(x_i(p)) = 0 \ \ \forall\, i \in I_0. \tag{6}$$

For all $i \in I_+$,

$$\bar{x}_i = \lim_{p \in K} x_i(p) < 1.$$

So, there exists $c < 1$ and $k_0 \in K$ such that

$$x_i(p) \leq c \ \ \forall\, p \in K, p \geq k_0.$$

So, by Assumption 4,

$$\varphi_p(x_i(p)) \leq \varphi_p(c) \ \ \forall\, p \in K, p \geq k_0.$$

Then, by Assumption 5,

$$\lim_{p \in K} \varphi_p(x_i(p)) = 0 \ \ \forall\, i \in I_+. \tag{7}$$

Finally, if $i \in I_1$, since $\varphi_p(x_i(p)) \in [0,1]$ and $[0,1]$ is compact, there exists an infinite set $K_1 \subset K$ such that the sequence $\{\varphi_p(x_i(p))\}_{p \in K_1}$ is convergent. Taking successive subsequences, we may assume that $K_1$ is the same subset of $K$ for all $i \in I_1$.

Then,

$$\lim_{p \in K_1} \varphi_p(x_i(p)) = \xi_i \leq 1 \ \forall\, i \in I_1. \tag{8}$$

Let us define $v \in [0,1]^n$ by:

$$v_i = 0 \ \text{ if } i \in I_0 \cup I_+,$$

$$v_i = \xi_i \ \text{ if } i \in I_1.$$

By (6, 7, 8) we have that

$$\lim_{p \in K_1} \varphi_p(x(p)) = v.$$

Therefore, by the continuity of $f$,

$$\lim_{p \in K_1} g_p(x(p)) = \lim_{p \in K_1} f(\varphi_p(x(p))) = f(v). \tag{9}$$

But, by the definition of $w$ and $v$ we see that $v < w$. Therefore, by Assumption 1,

$$f(w) < f(v).$$

So, by (9) and the continuity of $f$, for $p \in K_1$ large enough we have that

$$f(w) < g_p(x(p)).$$

But, since $w \in \{0,1\}^n$, by Assumption 2 we have that

$$f(\varphi_p(w)) = f(w),$$

so:

$$g_p(w) = f(\varphi_p(w)) = f(w) \; \forall \; p = 1, 2, \ldots.$$

Then, $w$ is a feasible point of (5) such that $g_p(w) < g_p(x(p))$ for all $p \in K_1$ large enough. This contradicts the fact that $x(p)$ is a solution of (5) for all $p$.

So far, we proved that every limit point $\bar{x}$ of $\{x(p)\}$ is a feasible point of (3). That is, it satisfies $\sum_{i=1}^{n} \bar{x}_i = V$ and $\bar{x} \in \{0,1\}^n$. Assume now that $\bar{x}$ is a limit point of $\{x(p)\}$ and that $x^*$ is a solution of (3). Let $K$ be an infinite set of $\{1, 2, \ldots\}$ such that

$$\lim_{p \in K} x(p) = \bar{x}.$$

Since $x(p)$ is a solution of (5) and $x^*$ is a feasible point of (5), we have, for all $p = 1, 2, \ldots$,

$$g_p(x(p)) \leq g_p(x^*).$$

But, since $x^* \in \{0,1\}^n$,

$$x^* = \varphi_p(x^*) \;\text{ and }\; f(x^*) = f(\varphi_p(x^*)) = g_p(x^*),$$

so:

$$g_p(x(p)) \leq f(x^*) \;\; \forall \; p = 1, 2, \ldots. \tag{10}$$

But, by Assumption 3, $\varphi_p(x(p)) \leq x(p)$. Then, by Assumption 1,

$$g_p(x(p)) = f(\varphi_p(x(p))) \geq f(x(p)) \; \forall \; p = 1, 2, \ldots.$$

Therefore, by (10),

$$f(x(p)) \leq f(x^*) \;\; \forall \; p = 1, 2, \ldots.$$

Taking limits for $p \in K$ in the last inequality, by the continuity of $f$, we get:

$$f(\bar{x}) \leq f(x^*).$$

Since $\bar{x}$ is a feasible point of (3), it turns out that it is also a global minimizer of this problem. So, the theorem is proved. □

The next theorem shows that, under Assumptions 1–5, the result of Theorem 1 is is true when the constraint is $\sum_{i=1}^{n} x_i \leq V$. Consider the problem

$$\text{Minimize } f(x) \text{ s.t. } \sum_{i=1}^{n} x_i \leq V, \;\; x \in \{0,1\}^n. \tag{11}$$

If $V$ is not integer, the problem is equivalent to the one that results of changing $V$ by the largest number in $\{1,\ldots,n\}$ smaller than $V$. If $V = 0$ the problem has the trivial solution $x = 0$. Therefore, we assume $V \in \{1,\ldots,n\}$.

Assume that $\varphi_p$ satisfies Assumptions 2–5 and define the associated SIMP problem as:

$$\text{Minimize } g_p(x) \text{ s.t. } \sum_{i=1}^{n} x_i \leq V, \;\; x \in [0,1]^n. \tag{12}$$

**Theorem 2.** *Suppose that Assumptions 1–5 hold. For all $p = 1, 2, \ldots$ let $x(p)$ be a solution of (12). Then, the sequence $\{x(p), p = 1, 2, \ldots\}$ admits at least a limit point and every limit point of this sequence is a solution of (11).*

*Proof.* The idea is to reduce (11) to (3) and (12) to (5). First we are going to prove that, for all $p = 1, 2, \ldots$,

$$\sum_{i=1}^{n} x_i(p) = V. \tag{13}$$

Assume that (13) is not true. Then, for some $p$,

$$\sum_{i=1}^{n} x_i(p) < V \leq n. \tag{14}$$

So, there exists $j \in \{1,\ldots,n\}$ such that $x_j(p) < 1$. By (14) there exists $\delta > 0$ such that $x_j(p) + \delta \leq 1$ and

$$\sum_{i=1}^{n} x_i(p) + \delta \leq V.$$

Define $y_i = x_i(p)$ if $i \neq j$ and $y_j = x_j(p) + \delta$ . Then, by Assumption 4,

$$\varphi_p(y_j) > \varphi_p(x_j(p)),$$

so,

$$\varphi_p(y) > \varphi_p(x(p)).$$

Therefore, by Assumption 1,

$$g_p(y) = f(\varphi_p(y))) < f(\varphi(x(p))) = g_p(x(p)).$$

This contradicts the fact that $x(p)$ is minimizer of (12).

So, we proved that $\sum_{i=1}^{n} x_i(p) = V$ for all $p = 1, 2, \ldots$. Since $x(p)$ is a minimizer of (12) it is also a minimizer of (5). Therefore, by Theorem 1, every limit point of $\{x(p)\}$ is a solution of (3). Therefore, it only remains to prove that every solution of (3) is a solution of (11). Assume that this is not true. Then, there exists a solution $x$ of (11) such that

$$\sum_{i=1}^{n} x_i < V \leq n.$$

Take $j$ such that $x_j < 1$ and $\delta > 0$ such that $x_j + \delta \leq 1$ and

$$\sum_{i=1}^{n} x_i + \delta \leq V.$$

Define $y = x_i$ if $i \neq j$ and $y_j = x_j + \delta$. Then, $y$ is a feasible point of (11) and $y > x$. By Assumption 1, $f(y) < f(x)$. This contradicts the fact that $x$ was a minimizer of (11). So, the proof is complete. □

Now we show that an exact solution of (3) (or (11)) can be obtained for a finite value of $p$. In the rest of this section, when we say "the original problem" we mean (3) or (11) and when we say "the SIMP problem" we mean, respectively, (5) or (12).

For each $p = 1, 2, \ldots$, if $x(p)$ is a solution of the SIMP problem, we define $\tilde{x}(p)$ as the "rounded" SIMP solution, in the following sense: For all $i = 1, \ldots, n$, if $x_i(p) < 1/2$, then $\tilde{x}_i(p) = 0$ and if $x_i(p) \geq 1/2$, then $\tilde{x}_i(p) = 1$.

**Theorem 3.** *Suppose that Assumptions 1–5 hold. Then, there exists $p_0 \in \{1, 2, \ldots\}$ such that, for all $p \geq p_0$, $\tilde{x}(p)$ is a solution of the original problem.*

*Proof.* Assume that the thesis is not true. Then, for all $p_0 \in \{1, 2, \ldots\}$ there exists $p \geq p_0$ such that $\tilde{x}(p)$ is not a solution of the original problem. In particular, taking $p_0 = 1$, we see that there exists $p_1 \in \{1, 2, \ldots\}$ such that

$\tilde{x}(p_1)$ is not a solution of the original problem. Now, taking $p_0 = p_1 + 1$ we obtain that there exists $p_2 > p_1$ such that $\tilde{x}(p_2)$ is not a solution of the original problem. Repeating this procedure, we obtain $p_3 > p_2$ such that $\tilde{x}(p_3)$ is not a solution of the original problem, and so on. In this way, we construct an infinite set $K = \{p_1, p_2, p_3, \ldots\}$ such that, for all $p \in K$, $\tilde{x}(p)$ is not a solution of the original problem. Take a convergent subsequence of $\{x(p_1), x(p_2), x(p_3), \ldots\} \equiv \{x(p)\}_{p \in K}$. Namely, let $K_1$ be an infinite subset of $K$ such that

$$\lim_{k \in K_1} x(p) = \bar{x}. \tag{15}$$

By Theorems 1 and 2, $\bar{x}$ is a solution of the original problem. But, by (15), since $\bar{x} \in \{0,1\}^n$, we have that

$$\tilde{x}(p) = \bar{x}$$

for $p \in K_1$ large enough. This contradicts the construction of the set $K$. So, the thesis is proved. □

# 3 Removing drawbacks

The observations of the present section will justify the application of the SIMP method to most optimization problem where some of its variables are $\{0,1\}$.

## 3.1 A counterexample

Assume that, in our problem (3), Assumptions 2–5 hold but Assumption 1 does not. In this case, Theorem 1 does not apply and the SIMP method can fail. For example take $n = 3$, $V = 2$, $\varphi_p(t) = t^p$ and

$$f(x_1, x_2, x_3) = x_1^2 + x_2^2 + (x_3 - 1)^2 + 10(x_1 - x_2)^2$$

The feasible points of

$$\text{Minimize } f(x_1, x_2, x_3) \text{ s.t. } \sum_{i=1}^{n} x_i = 2,\ x \in \{0,1\}^3 \tag{16}$$

are three: $(1,0,1), (0,1,1)$ and $(1,1,0)$. Now,

$$f(1,0,1) = 11, f(0,1,1) = 11, f(1,1,0) = 3.$$

So, $(1,1,0)$ is the unique solution of (3). We associate to (16) the classical SIMP problem

$$\text{Minimize } f(\varphi_p(x)) \text{ s.t. } \sum_{i=1}^{n} x_i = V, x \in [0,1]^n. \tag{17}$$

As always, $f(\varphi_p(1,1,0)) = 3$ for all $p$. However,

$$f(\varphi(\frac{1}{2},\frac{1}{2},1)) = \frac{1}{2^{p-1}}.$$

Since $(1/2,1/2,1)$ is a feasible point of (17), this implies that

$$f(\varphi_p(x(p))) \leq \frac{1}{2^{p-1}},$$

where $x(p)$ denotes a solution of (17). This implies that, for $p$ large enough,

$$(x_3(p)^p - 1)^2 < \frac{1}{4}.$$

So, for $p$ large enough,

$$|x_3(p)^p - 1| < \frac{1}{2}.$$

Therefore,

$$-\frac{1}{2} < x_3(p)^p - 1 < \frac{1}{2}.$$

So,

$$x_3(p)^p > \frac{1}{2}$$

and this implies that $x_3(p) > 1/2$ for $p$ large enough. Therefore, no subsequence of $\{x(p)\}$ can converge to $(1,1,0)$. Moreover, since $x_3(p) > 1/2$, we have that $\tilde{x}_3(p) = 1$ and the rounded approximation $\tilde{x}(p)$ is different from $(1,1,0)$ for all $p$ large enough.

### 3.2 Forcing the monotonicity assumption

The example above shows that Assumption 1 is essential for the success of the SIMP technique and cannot be removed. However, a trivial modification of the objective function that does not change the problem forces the assumption to be satisfied.

We say that the function $f : [0,1]^n \to I\!R$ has bounded positive variation (BPV) if there exists $M > 0$ such that, whenever $x, y \in [0,1]^n$ are such that $x_i = y_i$ except for one coordinate $j$ such that $x_j < y_j$, we have:

$$f(y) < f(x) + M(y_j - x_j).$$

It is easy to see that the BPV property implies that the function is continuous. Moreover, if $f$ has continuous derivatives in $[0,1]^n$, the Mean Value Theorem guarantees that the BPV property is satisfied, taking $M$ as a positive bound of the derivatives.

If $f$ satisfies the BPV property, a modification that does not modify problem (3) can be introduced. On one hand, observe that the problem

$$\text{Minimize } f(x) \text{ s.t. } \sum_{i=1}^{n} x_i = V, x \in \{0,1\}^n$$

is equivalent to

$$\text{Minimize } f(x) - M\sum_{i=1}^{n} x_i \text{ s.t. } \sum_{i=1}^{n} x_i = V, x \in \{0,1\}^n.$$

On the other hand, if $f$ satisfies the BPV property, defining $f(x) = f(x) - M\sum_{i=1}^{n} x_i$ for all $x \in [0,1]^n$, and taking $x < y$ it is easy to verify that $f(y) < f(x)$. Therefore, Assumption 1 is satisfied in the new problem.

### 3.3 Case with inequality volume constraint

If the original problem is (11) and $f$ does not satisfy Assumption 1, the situation is slightly more involved. In this case, assuming that the BPV property is satisfied, subtraction of $M\sum_{i=1}^{n} x_i$, although forces the fulfillment of Assumption 1, does not produce an equivalent problem. Therefore, we must first transform a problem of the form (11) into a problem of the form (3). Observe first that (11) is equivalent to

$$\text{Minimize } f(x) \text{ s.t. } \sum_{i=1}^{n} x_i + y_i = V, x \in \{0,1\}^n, y \in \{0,1\}^n,$$

and, as before, this is equivalent to

$$\text{Minimize } f(x) - M\sum_{i=1}^{n} x_i + y_i \text{ s.t. } \sum_{i=1}^{n} x_i + y_i = V, x \in \{0,1\}^n, y \in \{0,1\}^n. \tag{18}$$

With this transformation, we may safely apply the SIMP technique to (18) for obtaining a solution of (11).

### 3.4 Dealing with additional constraints

The presence of additional constraints deserves careful consideration. Suppose that the original problem is:

$$\text{Minimize } f(x) \text{ s.t. } \sum_{i=1}^{n} x_i = V,\ x \in \{0,1\}^n, x \in \Omega, \tag{19}$$

and that the SIMP problem is defined as:

$$\text{Minimize } g_p(x) \text{ s.t. } \sum_{i=1}^{n} x_i = V,\ x \in [0,1]^n, x \in \Omega, \tag{20}$$

where Assumptions 1–5 hold. We emphasize that, in (20), we maintained the constraint $x \in \Omega$ without modification and we will see that this is intrinsically wrong. Let $n = 3$, $V = 2$, $f(x) = -x_3 - 0.1(x_1 + x_2)$ and

$$\Omega = \{(1,1,0), (\frac{1}{2}, \frac{1}{2}, 1)\}.$$

The point $(1,1,0)$ is the only feasible point of (19), so it is the solution of this problem.

Taking $\varphi(t) = t^p$ we have that

$$g_p(1,1,0) = f(1,1,0) = -0.2 \ \forall\, p = 1,2,\ldots.$$

But $(1/2, 1/2, 1)$ is a feasible point of (20) and

$$g_p(\frac{1}{2}, \frac{1}{2}, 1) = -1 - 0.1 \times 2^{1-p} < g_p(1,1,0) \ \ \forall\, p = 1,2,\ldots.$$

Since the feasible set of (20) is $\{(1,1,0), (1/2,1/2,1)\}$ it turns out that $x(p) = (1/2, 1/2, 1)$ for all $p$. Therefore, $(1/2, 1/2, 1)$ is the only possible limit point of $\{x(p)\}$.

In what follows we explain how this difficulty can be circumvented. Assume that our original problem is

$$\text{Minimize } f(x) \ \text{ s.t. } \ \sum_{i=1}^{n} x_i = V,\ C(x) \leq 0,\ x \in \{0,1\}^n, \tag{21}$$

where $C : I\!R^n \to I\!R^q$ is continuous. Assume that the feasible set of (21) is nonempty. Moreover, we assume $V \in \{1, \ldots, n\}$. Defining $C_+(x) = \sum_{i=1}^q \max\{0, C_i(x)\}$, the constraint $C(x) \leq 0$ can be expressed as $C_+(x) = 0$.

Since the feasible set of (21) is finite, we can define

$$\mu = \min\{C_+(x) \mid \sum_{i=1}^n x_i = V, x \in \{0,1\}^n,\ C_+(x) > 0\}.$$

Let $x^*$ be a solution of (21). If $\rho > 0$ is large enough, we have that $f(x) + \rho\mu > f(x^*)$ for all $x \in \{0,1\}^n$. Therefore, for $\rho$ large enough, (21) is equivalent to

$$\text{Minimize } f(x) + \rho C_+(x) \ \text{ s.t. } \ \sum_{i=1}^n x_i = V,\ x \in \{0,1\}^n.$$

Then, we can apply the already obtained results to see that, for $M$ large enough and $p$ large enough, any rounded solution of the problem below coincides with a solution of (21):

$$\text{Minimize } f(\varphi_p(x)) + \rho C_+(\varphi_p(x)) - M\sum_{i=1}^n \varphi_p(x_i) \ \text{ s.t. } \ \sum_{i=1}^n x_i = V,\ x \in [0,1]^n. \tag{22}$$

The observation above remains if one considers $C_+(x)^2$ instead of $C_+(x)$. Therefore, (22) provides a way of dealing with arbitrary constraints.

Observe that (22) remembers the application of a penalty method to the nonlinear programming problem:

$$\text{Minimize } f(\varphi_p(x)) - M\sum_{i=1}^n \varphi_p(x_i) \ \text{ s.t. } \ \sum_{i=1}^n x_i = V,\ C(\varphi_p(x)) \leq 0,\ x \in [0,1]^n. \tag{23}$$

Comparing (23) with (21) we see that the main warning that must be done when one applies the SIMP technique to problems with arbitrary constraints is that these constraints must also be transformed from $C(x) \leq 0$ to $C(\varphi_p(x)) \leq 0$. Observe that in (23) the BPV must be applied to the new objective function.

Finally, it must be mentioned that, when the original problem does not have volume constraints ($\sum_{i=1}^n x_i = (\leq) V$), the problem does not change if one artificially imposes the constraint $\sum_{i=1}^n x_i \leq n$, so that the SIMP technique can be applied.

### 3.5 Problems with continuous variables

Here we consider problems of the form

$$\text{Minimize } F(x,y) \tag{24}$$

$$\text{s.t. } x \in \{0,1\}^n,\ \sum_{i=1}^{n} x_i = V,\ C(x) \le 0,\ \ D(x,y) \le 0. \tag{25}$$

where $y \in I\!R^m$, $C : I\!R^n \to I\!R^q$, $D : I\!R^{n+m} \to I\!R^\nu$. We assume that, for all $x \in [0,1]^n$, a solution $y(x)$ of the problem

$$\text{Minimize } F(x,y) \ \text{ s. t. } \ D(x,y) \le 0$$

exists and that $f(x) \equiv F(x, y(x))$ defines a continuous function. Therefore, (24–25) is equivalent to

$$\text{Minimize } f(x) \ \text{ s.t. } \ x \in \{0,1\}^n,\ \sum_{i=1}^{n} x_i = V,\ C(x) \le 0. \tag{26}$$

In the previous section we saw that solving (26) is essentially equivalent to solve (23) for suitable values of $p$ and $M$. But (23) is equivalent to

$$\text{Minimize } F(\varphi_p(x)), y) - M \sum_{i=1}^{n} \varphi_p(x_i)$$

$$\text{s.t. } \sum_{i=1}^{n} x_i = V,\ C(\varphi_p(x)) \le 0,\ x \in [0,1]^n,\ \ D(\varphi_p(x), y) \le 0.$$

The case with the constraint $\sum_{i=1}^{n} x_i \le V$ is analogous. Therefore, we see again that the key point is to consider the additional constraints with $\varphi_p(x)$ replacing $x$.

## 4 Final remarks

The SIMP method has been traditionally used in Topology Optimization but it can be applied to any problem in which some of the variables are $\{0,1\}$. In this paper we saw that the main warnings related to this application are:

(a) The problem must be in the form

$$\text{Minimize } f(x) \ \ s.t. \ \sum_{i=1}^{n} x_i = V,\ x \in \Omega \tag{27}$$

or

$$\text{Minimize } f(x) \;\; s.t. \;\; \sum_{i=1}^{n} x_i \leq V, \; x \in \Omega. \tag{28}$$

(b) Any problem of the form

$$\text{Minimize } f(x) \;\; s.t. \;\; x \in \Omega$$

can be expressed in the form (28) adding the artificial constraint $\sum_{i=1}^{n} x_i \leq n$.

(c) Assumption 1 (monotonicity) must hold, both in (27) and (28). It is easy to force this assumption in (27) (subtracting $M \sum_{i=1}^{n} x_i$ in the objective function), but not in (28). However, adding auxiliary variables, a problem of the form (28) can be reduced to (27).

(d) In the SIMP problem, the constraint $x \in \Omega$ must be replaced by $g_p(x) \in \Omega$. The volume constraint must remain without modification.

With these safeguards, the main difficultly regarding the effectiveness of the SIMP approach lies in the resolution of the SIMP subproblems. In fact, we need global solutions, which might be very difficult to obtain, since these subproblems are nonconvex. See [5] for an analysis of the limitations of continuation methods for obtaining global solutions in SIMP subproblems. Neither the non-uniqueness of the solution nor the presence of non-volume constraints are serious drawbacks for the application of SIMP. In [6] a material interpolation model is proposed by means of which the probability of obtaining global solutions of $0-1$-related subproblems is increased.

The improvement of SIMP methods can be linked to the choice of alternative univariate functions $\varphi_p(t)$. Since differentiability of the objective function of SIMP is not necessary for proving convergence, it should be interesting to try piecewise linear one-dimensional functions like the one given by

$$\varphi_p(t) = \frac{t}{p} \quad \text{if} \quad t \leq \frac{p}{1+p},$$

$$\varphi_p(t) = pt - p + 1 \quad \text{if} \quad t \geq \frac{p}{1+p}.$$

Perhaps a smoothed version of this function could also be interesting.

**Acknowledgements.** The author is indebted to Prof. Eduardo Fancello and two anonymous referees for their careful reading and useful remarks on the first draft of this paper.

## References


[1] M. P. Bendsøe. Optimal shape design as a material distribution problem. *Structural Optimization* 1, 193-202 (1989).

[2] M. P. Bendsøe. *Optimization of Structural Topology: shape and material*. Springer-Verlag, Berlin-Heidelberg-New York, 1995.

[3] M. P. Bendsøe and O. Sigmund. *Topology Optimization: Theory, Methods and Applications*. Springer-Verlag, Berlin-Heidelberg-New York, 2003.

[4] A. Rietz. Sufficiency of a finite exponent in SIMP (power law) methods. *Structural and Multidisciplinary Optimization* 21, 159-163 (2001).

[5] M. Stolpe and K. Svanberg. On the trajectories of penalization methods for topology optimization. *Structural and Multidisciplinary Optimization* 21, 128-139 (2001).

[6] M. Stolpe and K. Svanberg. An alternative interpolation scheme for minimum compliance topology optimization. *Structural and Multidisciplinary Optimization* 22, 116-124 (2001).

[7] M. Zhou and G. I. N. Rozvany. The COC algorithm, part I. Topological, geometrical and generalized shape optimization. *Comp. Meth. Appl. Mech. Engrg.* 89, 309-336 (1991).