Topological Optimization Using the SIMP Method

Mikal William Nelson

B.S. Mathematics, University of Minnesota, 2013

Submitted to the graduate degree program in Department of Mathematics and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Master of Arts.

Paul Cazeaux, Chairperson

Committee members

Mat Johnson

Dionyssios Mantzavinos

Yannan Shen

Date defended: July 26, 2021

The Thesis Committee for Mikal William Nelson certifies
that this is the approved version of the following thesis :

Topological Optimization Using the SIMP Method

Paul Cazeaux, Chairperson

Date approved: July, 2021

Abstract

Insert Abstract Here

Acknowledgements

Dedicated to my mother, my first math teacher.

Thank you to my wife, father, brothers, and all the teachers and professors who have helped me grow along the way.

Contents

List of Figures

Introduction

Chapter 1

Background

1.1 PDE Discretization

Multidimensional topological optimization problems often involve the use of partial differential equations (PDEs) which model the physical properties of the materials involved. Most of these PDEs cannot be uniquely solved analytically, so we turn to numerical methods in order to approximate their solutions. The first step in many of these methods is to discretize our domain; that is, we want to choose some scheme to divide our continuous domain into a finite number of pieces over which we will apply a particular method to approximate solutions to the PDE.

In the implementation of the SIMP method that is used throughout this paper, the Finite Volume Method is used to discretize and approximate solutions to the heat equation for our heat generating medium. We will introduce the Heat Equation and then proceed to give an overview of the Finite Volume Method.

1.1.1 The Heat Equation

Consider a stationary object of that has heat flowing between its interior regions. The temperature at any point in the interior of the object will depend on the spatial position chosen as well as the time we measure the temperature at that point. Therefore, the temperature (T) at any point in such an object is a function of both space ($\vec{x}$) and time (t) coordinates: $T(\vec{x}, t)$. Physical principles demand that such a temperature function must satisfy

the equation

$$\frac{\partial T}{\partial t} = \nabla \cdot (k(\vec{x})\nabla T)\,, \tag{1.1}$$

where ∇ is the gradient operator and the function k represents the thermal diffusivity at a point in our object.

Equation (1.1) is commonly referred to as the Heat or Diffusion Equation. If we were to have a constant thermal diffusivity throughout our object on a simple domain (such as a square or circle), it would be possible to analytically find a solution to this partial differential equation. However, as in the problems of interest throughout this paper, when k is not constant we must turn to numerical methods to find approximate solutions for the function T.

1.1.2 The Finite Volume Method

For the numerical approximations of PDEs in this paper the Finite Volume Method (FVM) was implemented, which will be described in this section.

As with any other numerical method to solve PDEs, we must first discretize our domain by creating a mesh. One major advantage of the finite volume method is that it allows for a great amount of freedom in mesh choice. When using FVM the domain can be discretized into a mesh of arbitrary polygons, but we chose uniform squares or rectangles in our work to simplify the resulting calculations.

Given a mesh of polygons on a domain Ω with sample points at $\{x_i\} \subset \Omega$, we create a set of *control volumes* around each x_i. The resulting set of control volumes discretize the partial differential equation. The finite volume method has us integrate our PDE over each control volume and then use the Divergence Theorem (Theorem 1) to convert volume integrals into surface integrals involving the fluxes across the boundaries of the control volumes. We then approximate those fluxes across the boundaries to calculate approximate solutions to the PDE of interest, such as (1.1).

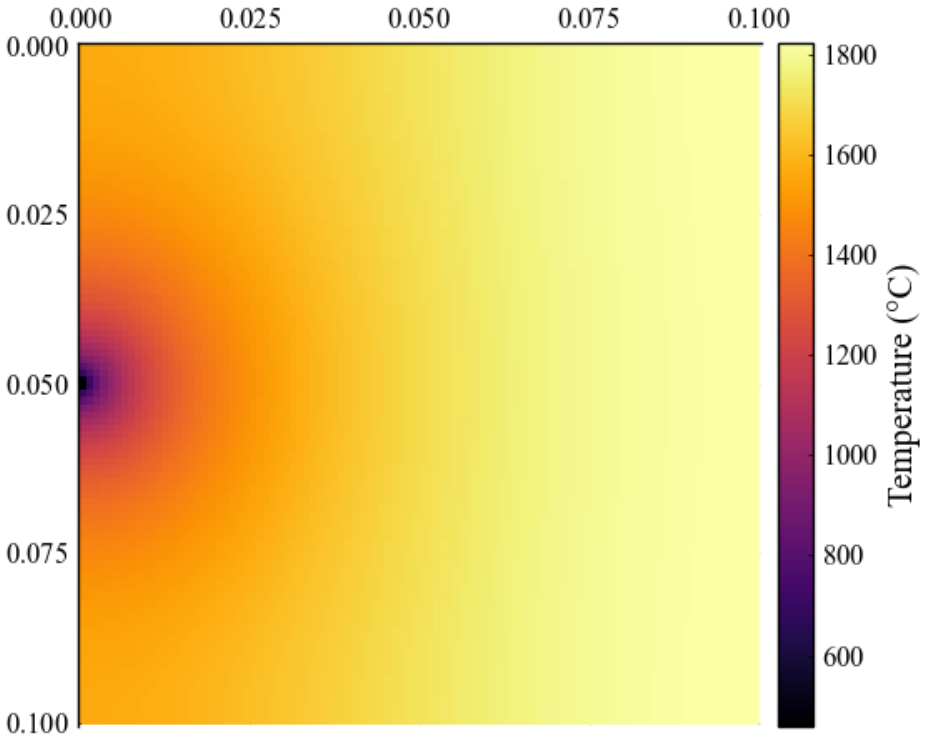

Figure 1.1: Heatmap for a 0.1 m $\times$ 0.1 m object with uniform heat generation and a heat sink at the center of its west boundary. This map was produced via the Finite Volume Method using 100 $\times$ 100 uniform control volumes.

Theorem 1 (The Divergence Theorem)**.** *Suppose that $\mathcal{V}$ is a compact subset of $\mathbb{R}^n$ that has a piecewise smooth boundary $\mathcal{S}$ (i.e. $\partial\mathcal{V} = \mathcal{S}$). If $\mathbf{F}$ is a continously differentiable vector field defined on a neighborhood of $\mathcal{V}$, then*

$$\iiint_{\mathcal{V}} (\nabla \cdot \mathbf{F}) \, \mathrm{d}\mathcal{V} = \oiint_{\mathcal{S}} (\mathbf{F} \cdot \hat{\mathbf{n}}) \, \mathrm{d}\mathcal{S}. \tag{1.2}$$

The divergence theorem is the key component in the finite volume method because it allows us to look at fluxes across the boundaries of each control volume, rather than the control volume itself.

Let us look at the finite volume method applied to the heat equation in two dimensions. Suppose we have discretized our space by dividing it up into a mesh of control volumes $\{V_i\}$. We integrate the heat equation PDE over each control volume, using the divergence theorem to convert the volume integral into a surface integral:

$$\int_{V_i} \frac{\partial T}{\partial t} \, \mathrm{d}\vec{x} = \int_{V_i} \nabla \cdot (k(\vec{x}) \nabla T) \, \mathrm{d}\vec{x} \underset{(1.2)}{=} \int_{\partial V_i} k(\vec{x}) \nabla T \cdot \mathrm{d}s,$$

where s represents the lines that form the boundary of the control volume. Then, applying an approximation scheme to this result, we obtain a sparse and structured linear system.

One other major advantage of the finite volume method is that boundary conditions can easily be taken into account on general domains. For example, adding a heat sink by applying a Dirichlet boundary condition ($T_s = 0\,^\circ\mathrm{C}$) can be thought of as zeroing out our algebraic equations by introducing a ghost cell that, when interpolated with the boundary cell, causes the temperature across the boundary to be zero.

1.2 The Optimization Problem

In much of mathematics, our goal is to seek some sort of solution. In cases where there are multiple solutions, it is desirable to determine the "best" solution judged against some set of criteria. Mathematical optimization is the study of solving such problems. In its simplest case, mathematical optimization is the practice of minimizing or maximizing a given function over a certain set and possibly subject to some constraints.

1.2.1 Optimization Problem Definition

Definition 1. *An optimization problem (in standard form) has the form*

$$\begin{array}{ll} \textit{minimize} & f_0(x) \\ \textit{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_i(x) = 0, \quad i = 1, \ldots, p \end{array} \tag{1.3}$$

where

- $x = (x_1, \ldots, x_n)$ *are the optimization variables,*
- $f_0 : \mathbb{R}^n \to \mathbb{R}$ *is the objective function,*
- $f_i : \mathbb{R}^n \to \mathbb{R}$ *are the inequality constraint functions, and*
- $h_i : \mathbb{R}^n \to \mathbb{R}$ *are the equality constraint functions.*

If there are no constraints ($m = p = 0$), then the problem is called *unconstrained*. (Boyd & Vandenberghe, 2004, p. 127)

We call a vector $x^\star$ *globally optimal* if it has the smallest objective value among all vectors that satisfy the constraints. That is, for any z with $f_1(z) \leq 0, \ldots, f_m(z) \leq 0$, then $f_0(z) \geq f_0(x^\star)$. A point x that is in the domains of each function f_i and h_i is called *feasible*

if it satisfies all the constraints. Finally, the *optimal value* $p^\star$ of the problem is defined as

$$p^\star = \{f_0(x) \mid f_i(x) \leq 0, i = 1, \ldots, m, h_i(x) = 0, i = 1, \ldots, p\}\,.$$

Therefore, $p^\star = f_0(x^\star)$, the objective function value at a feasible and globally optimal vector $x^\star$.

Notice that the optimization problem in standard form is a minimization problem. We can easily change it into a maximization problem by minimizing the objective function $-f_0$ subject to the same constraints.

The optimization problem is *linear* or called a *linear program* if the objective and constraint functions are all linear. An optimization problem involving a quadratic objective function and linear constraints is *quadratic* or a *quadratic program*. If the optimization problem is not linear or quadratic, it is referred to as a *nonlinear program*.

There are exists efficient methods for solving linear programming and many quadratic programming problems.

1.2.2 Convex Optimization

A set C is *convex* if the line segment between any two points in C lies in C. That is if for any $x_1, x_2 \in C$ and any θ with $0 \leq \theta \leq 1$, we have $\theta x_1 + (1-\theta)x_2 \in C$.

A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if the domain of f is a convex set and if for all x, y in the domain of f, and θ with $0 \leq \theta \leq 1$, we have

$$f\left(\theta x + (1-\theta)\, y\right) \leq \theta f(x) + (1-\theta) f(y). \tag{1.4}$$

A *convex optimization problem*, therefore, is an optimization problem of the form

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & a_i^T = b_i, \qquad i = 1, \ldots, p \end{array} \tag{1.5}$$

where $f_0, \ldots, f_m$ are convex functions.

Notice that there are three requirements that differentiate a convex optimization problem from a general optimization problem:

- the objective function must be convex
- the inequality constraint functions must be convex
- the equality constraint functions must be affine

In a convex optimization problem we minimize a convex objective function over a convex set and any locally optimal point is also globally optimal. This is a *very* useful fact!

Why might local optimality implying global optimality be useful? Consider a situation where we have a convex objective function. If we are able to find any minimum value for the objective function, then we know this value is not just a local minimum, but indeed a global minimum. If we do not have a convex function, we cannot be as assured that we have found the global optimal value. For example, consider a function such as the one shown in Figure 1.2. An optimization strategy may find the local minimum near $x = 1$, but since the function is not convex everywhere on its domain, we cannot conclude that this value is the truly optimal value. In fact, we see that the global minimum, and hence the actual optimal value, is between $x = -1$ and $x = -0.5$. On the other hand, if we have a function that is everywhere convex, as soon as we find a local minima, we can be assured that it is also the global minima!

So we can see that convexity is a very powerful and useful property in terms of optimization problems. As a result, any time we can take advantage of convexity or approximate

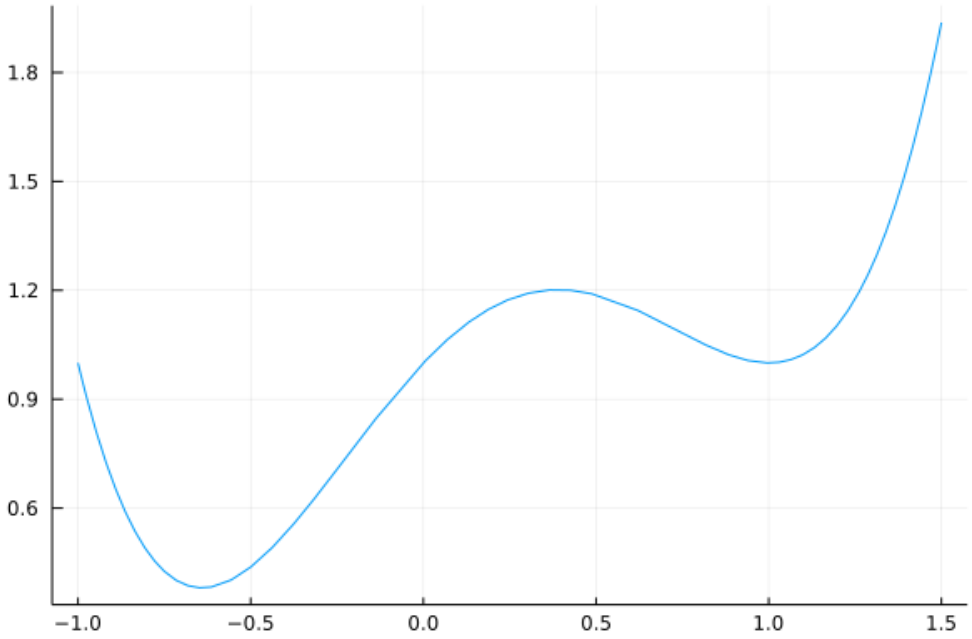

Figure 1.2: A graph of the non-convex function $f(x) = x^4 - x^3 - x^2 + x + 1$. Notice it has two local minima and that the right local minima is not equal to the global minimum.

functions using convex functions, we often do so.

1.3 Optimization Methods

In this section I will present a few optimization algorithms.

1.3.1 Line Search Methods

The *line search* is a strategy that selects the step size (commonly represented by t) that determines where along the line $\{x + t\Delta x \mid t \in \mathbb{R}_+\}$ the next iterate in the descent method will be. (Δx represents the *descent direction*, .) Line search strategies can either be *exact* or *inexact*.

Exact Line Search

An *exact line search* chooses the value t along the ray $\{x + t\Delta x \mid t \in \mathbb{R}_+\}$ that exactly minimizes the function of interest f:

$$t = \arg\min_{s \geq 0} f(x + s\Delta x)$$

An exact line search is almost never practical. In very special cases, such as some quadratic optimization problems, where computing the cost of the minimization problem is low compared to actually calculating the search direction, one might employ an exact line search.

Backtracking Line Search

Most often in practice we use *inexact line searches*. In an inexact line search, we choose t such that f is *approximately* minimized or reduced "enough" along $\{x + t\Delta x \mid t \in \mathbb{R}_+\}$.

One inexact line search strategy is the *Backtracking Line Search*.

Algorithm 1 Backtracking Line Search (Boyd & Vandenberghe, 2004)

given a descent direction Δx for f at $x \in \mathbf{dom} f, \alpha \in (0, 0.5), \beta \in (0, 1)$.

$t := 1$.

while $f(x + t\Delta x) > f(x) + \alpha t \nabla f(x)^T \Delta x$ **do**

  $t := \beta t$

end while

"Backtracking" in the name is refers to the fact that the method starts with a unit step size ($t = 1$) and then reduces the step size ("backtracks") by the factor β until we meet the stopping criterion $f(x + t\Delta x) \leq f(x) + \alpha t \nabla f(x)^T \Delta x$.

Figure 9.1 (Boyd & Vandenberghe, 2004, p. 465) demonstrates the Backtracking Line Search visually for a parabola.

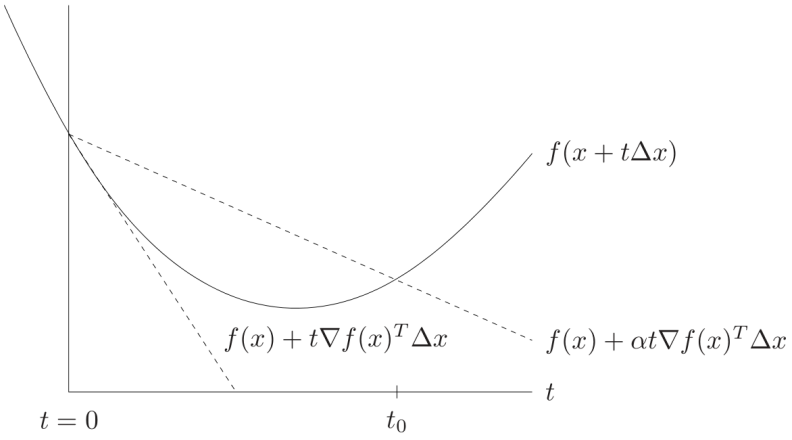

Figure 9.1 *Backtracking line search.* The curve shows f, restricted to the line over which we search. The lower dashed line shows the linear extrapolation of f, and the upper dashed line has a slope a factor of α smaller. The backtracking condition is that f lies below the upper dashed line, *i.e.*, $0 \leq t \leq t_0$.

Notice that the backtracking search will find a step size t such that $0 \leq t \leq t_0$ and that for such t, $f(x + t\Delta x)$ is smaller relative to $f(x)$. However, the step size we choose may not exactly be the minimum of the function, but we have funneled it down to be closer to the minimum of f.

1.3.2 Gradient Descent

The gradient descent method chooses the search direction to be the negative gradient. That is, in this method we set $\Delta x = -\nabla f(x)$, where f is the function we seek to optimize. Since the gradient of a function gives the direction of greatest increase, naturally the negative gradient will give the direction of the most rapid decline.

Algorithm 2 Gradient Descent Method (Boyd & Vandenberghe, 2004)

given a starting point $x \in \mathbf{dom} f$.
repeat

1. $\Delta x := -\nabla f(x)$.
2. *Line search.* Choose step size t via exact or backtracking line search.
3. *Update.* $x := x + t\Delta x$.

until stopping criterion is satisfied.

Notice in Algorithm 2 that essentially a line search is used to determine a step size and then we update our iterate in the direction of the steepest descent. This is repeated until we meet some sort of stopping criterion, typically something of the form $\|\nabla f(x)\|_2 \leq \eta$, where η is small and positive. Another common stopping criterion is to stop the algorithm when no significant progress is made between iterates by stopping the algorithm when $\|f_{k+1} - f_k\| < \eta$.

An implementation of the gradient descent algorithm in the Julia language can be found in the appendix.

1.3.3 Nonlinear Conjugate Gradient

The Nonlinear Conjugate Gradient method works similarly to the gradient descent algorithm, but adds the additional requirement that in each iteration gradient is orthogonal to each previous search direction.

Suppose we have a function $f(x)$ of N variables. Let x_0 be an initial guess value for the minimum. The opposite of the gradient will give the direction of steepest descent. Therefore, we start off by setting $\Delta x_0 = -\nabla f(x_0)$.

We set an adjustable step length α and perform a line search in the direction $d_0 = \Delta x_0$ until a minimum of f is reached:

$$\alpha_0 := \arg\min_{\alpha} f(x_0 + \alpha \Delta x_0),$$
$$x_1 = x_0 + \alpha_0 \Delta x_0.$$

Suppose we were to simply iterate this process and for each step i we do the following:

1. Set $\Delta x_i = -\nabla f(x_i)$
2. Calculate $\alpha_i = \arg\min_{\alpha} f(x_i + \alpha \Delta x_i)$
3. Compute $x_{i+1} = x_i + \alpha_i \Delta x_i$

However, there is an issue with this proposed iterative scheme: We have moved α_i in direction Δx_i to find the minimum value in that direction, but by moving α_{i+1} in direction Δx_{i+1} we may have accidentally *undone* the progress made in the previous iteration so that we no longer have a minimum value in direction Δx_i. We can fix this problem by making sure that successive direction vectors have no influence in the directions of previous iterations. That is, we require our directions in each iteration to be conjugate (with respect to the matrix of coefficient for our system) to one another. Therefore, rather than taking Δx_{i+1} to be $-\nabla f(x_i)$, we compute a direction conjugate to all previous directions by some pre-chosen methodology. This suggests the following iterative scheme:

After the first iteration, the following steps constitute one iteration along a conjugate direction:

1. Calculate the new steepest descent direction: $\Delta x_i = -\nabla f(x_i)$,
2. Compute β_i using some formulation. Two options are below:
   - Fletcher–Reeves: $\beta_i^{FR} = \frac{\Delta x_i^T \Delta x_i}{\Delta x_{i-1}^T \Delta x_{i-1}}$
   - Polak–Ribière: $\beta_i^{PR} = \frac{\Delta x_i^T (\Delta x_i - \Delta x_{i-1})}{\Delta x_{i-1}^T \Delta x_{i-1}}$
3. Update the conjugate direction: $d_i = \Delta x_i + \beta_i d_{i-1}$,
4. Line search: Optimize $\alpha_i = \arg\min_{\alpha} f(x_i + \alpha d_i)$,
5. Update iterate value: $x_{i+1} = x_i + \alpha_i d_i$.

Algorithm 3 uses the Newton-Raphson method to find the values of α_i.

To get some intuition of how this algorithm operates, let's look at it applied to a quadratic function.

Example: Rosenbrock Function

Algorithm 3 Nonlinear Conjugate Gradient Using Newton-Raphson (Shewchuk, 1994)

Given a function f, a starting value x, a maximum number of CG iterations $i_{\max}$, a CG error tolerance $\epsilon < 1$, a maximum number of Newton-Raphson iterations $j_{\max}$, and a Newton-Raphson error tolerance $\varepsilon < 1$:

$i \Leftarrow 0$
$k \Leftarrow 0$
$r \Leftarrow -f'(x)$
$d \Leftarrow r$
$\delta_{\text{new}} \Leftarrow r^T r$
$\delta_0 \Leftarrow \delta_{\text{new}}$
while $i < i_{\max}$ and $\delta_{\text{new}} > \epsilon^2 \delta_0$ **do**
    $j \Leftarrow 0$
    $\delta_d \Leftarrow d^T d$
    while true **do**
        $\alpha \Leftarrow -\frac{\left[f'(x)\right]^T d}{d^T f''(x) d}$
        $x \Leftarrow x + \alpha d$
        $j \Leftarrow j + 1$
        $j < j_{\max}$ and $\alpha^2 \delta_d > \varepsilon^2$ OR **Break**
    end while
    $r \Leftarrow -f'(x)$
    $\delta_{\text{old}} \Leftarrow \delta_{\text{new}}$
    $\delta_{\text{new}} \Leftarrow r^T r$
    $\beta \Leftarrow \frac{\delta_{\text{new}}}{\delta_{\text{old}}}$
    $d \Leftarrow r + \beta d$
    $k \Leftarrow k + 1$
    if $k = n$ or $r^T d \leq 0$ **then**
        $d \Leftarrow r$
        $k \Leftarrow 0$
    end if
    $i \Leftarrow i + 1$
end while

1.4 The Method of Moving Asymptotes (MMA)

The Method of Moving Asymptotes (MMA) is a method of non-linear programming, originally developed for structural optimization. The method uses an iterative process which creates a convex subproblem that is solved in each iteration. Each of these subproblems is an approximation of the original problem with parameters that change the curvature of the approximation. These parameters act as asymptotes for the subproblem and moving the asymptotes between iterations stablizes the convergence of the entire process.

1.4.1 General Method Description

Consider an optimization problem of the following general form

$$\begin{array}{lllll} P\text{:} & \text{minimize} & f_0(\vec{x}) & & (\vec{x} \in \mathbb{R}^n) \\ & \text{subject to} & f_i(\vec{x}) \leq \hat{f}_i, & \text{for } i = 1, \ldots, m & \\ & & \underline{x}_j \leq x_j \leq \overline{x}_j, & \text{for } j = 1, \ldots, n & \end{array} \tag{1.6}$$

where

- $\vec{x} = (x_1, \ldots, x_n)^T$ is the vector of variables
- $f_0(\vec{x})$ is the objective function
- $f_i(\vec{x}) \leq \hat{f}_i$ are behavior constraints
- $\underline{x}_j$ and $\overline{x}_j$ are given lower and upper bounds on the variables

The general approach for solving such optimization problems is to split it up and solve a sequence of subproblems using the following iteration:

Step 0: Choose a starting point $\vec{x}^{(0)}$, and let the iteration index $k = 0$.

Step 1: Given an iteration point $\vec{x}^{(k)}$, calculate $f_i(\vec{x}^{(k)})$ and the gradients $\nabla f_i(\vec{x}^{(k)})$ for $i = 0, 1, \ldots, m$.

Step 2: Generate a subproblem $P^{(k)}$ by replacing, in (6), the (usually implicit) functions f_i by approximating explicit functions $f_i^{(k)}$, based on calculations from Step 1.

Step 3: Solve $P^{(k)}$ and let the optimal solution of this subproblem be the next iteration point $\vec{x}^{(k+1)}$. Let $k = k + 1$ and go to Step 1.

In MMA, each $f_i^{(k)}$ is obtained by a linearization of f_i in variables of the type

$$\frac{1}{x_j - L_j} \quad \text{or} \quad \frac{1}{U_j - x_j}$$

dependent on the signs of the derivatives of f_i at $\vec{x}^{(k)}$. The values of L_j and U_j are normally changed between iterations and are referred to as moving asymptotes.

Defining The Functions $f_i^{(k)}$

Given the iteration point $\vec{x}^{(k)}$ at an iteration k, values of the parameters $L_j^{(k)}$ and $U_j^{(k)}$ are chosen, for $j = 1, \ldots, n$, such that $L_j^{(k)} < x_j^{(k)} < U_j^{(k)}$.

For each $i = 0, 1, \ldots, m$, $f_i^{(k)}$ is defined by

$$f_i^{(k)}(\vec{x}) = r_i^{(k)} + \sum_{j=1}^{n} \left(\frac{p_{ij}^{(k)}}{U_j^{(k)} - x_j} + \frac{q_{ij}^{(k)}}{x_j - L_j^{(k)}} \right)$$

where

$$p_{ij}^{(k)} = \begin{cases} \left(U_j^{(k)} - x_j^{(k)} \right)^2, & \text{if } \frac{\partial f_i}{\partial x_j} > 0 \\ 0, & \text{if } \frac{\partial f_i}{\partial x_j} \leq 0 \end{cases}$$

$$q_{ij}^{(k)} = \begin{cases} 0, & \text{if } \frac{\partial f_i}{x_j} \geq 0 \\ -\left(x_j^{(k)} - L_j^{(k)} \right)^2 \frac{\partial f_i}{\partial x_j}, & \text{if } \frac{\partial f_i}{\partial x_j} < 0 \end{cases}$$

$$r_i^{(k)} = f_i(\vec{x}^{(k)}) - \sum_{j=1}^{n} \left(\frac{p_{ij}^{(k)}}{U_j^{(k)} - x_j^{(k)}} + \frac{q_{ij}^{(k)}}{x_j^{(k)} - L_j^{(k)}} \right)$$

and where all $\frac{\partial f_i}{\partial x_j}$ are evaluated at $\vec{x} = \vec{x}^{(k)}$.

Notice that $f_i^{(k)}$ is a first-order approximation of f_i at $\vec{x}^{(k)}$. Additionally, by construction, $f_i^{(k)}$ is a convex function.

Looking at the second derivatives, the closer $L_j^{(k)}$ and $U_j^{(k)}$ are chosen to $x_j^{(k)}$, the larger the second derivatives become and hence the more curvature is given to the approximating function $f_i^{(k)}$. This means that the closer $L_j^{(k)}$ and $U_j^{(k)}$ are chosen to $x_j^{(k)}$, the more conservative becomes the approximation of the original problem. If $L^{(k)}$ and $U^{(k)}$ are chosen 'far away' from $\vec{x}^{(k)}$, then the approximation $f_i^{(k)}$ becomes close to linear.

We always choose the values of $L_j^{(k)}$ and $U_j^{(k)}$ to be finite. As a result each $f_i^{(k)}$ becomes strictly convex except when $\frac{\partial f_i}{\partial x_j} = 0$ at $\vec{x} = x^{(k)}$.

Now, with the approximating functions $f_i^{(k)}$ as defined earlier, we have the following subproblem $P^{(k)}$:

$$
\begin{array}{lll}
P^{(k)}: & \text{minimize} & \displaystyle\sum_{j=1}^{n} \left(\frac{p_{oj}^{(k)}}{U_j^{(k)} - x_j} + \frac{q_{oj}^{(k)}}{x_j - L_j^{(k)}} \right) + r_o^{(k)} \\
& \text{subject to} & \displaystyle\sum_{j=1}^{n} \left(\frac{p_{ij}^{(k)}}{U_j^{(k)} - x_j} + \frac{q_{ij}^{(k)}}{x_j - L_j^{(k)}} \right) + r_i^{(k)} \leq \hat{f}_i, \quad \text{for } i = 1, \ldots, m \\
& \text{and} & \max\{\underline{x}_j, \alpha_j^{(k)}\} \leq x_j \leq \min\{\overline{x}_j, \beta_j^{(k)}\}, \quad \text{for } j = 1, \ldots, n
\end{array}
\tag{1.7}
$$

(The parameters $\alpha_j^{(k)}$ and $\beta_j^{(k)}$ are called move limits.)

$\alpha_j^{(k)}$ and $\beta_j^{(k)}$ should at least be chosen such that $L_j^{(k)} < \alpha_j^{(k)} < x_j^{(k)} < \beta_j^{(k)} < U_j^{(k)}$.

General Rule for how to choose $L_j^{(k)}$ and $U_j^{(k)}$:

(a) If the process tends to oscillate, then it needs to be stabilized and this can be accomplished by moving the asymptotes closer to the current iteration point.

(b) If, instead, the process is monotone and slow, it needs to be "relaxed". This can be accomplished by moving the asymptotes away from the current iteration point.

1.4.2 The Dual Problem

$P^{(k)}$ is a convex, separable problem, so we can create a dual problem using a Lagrangian function. The Lagrangian function corresponding to $P^{(k)}$ is given by

$$\ell(x, y) = f_0^{(k)}(\vec{x}) + \sum_{i=1}^{m} y_i f_i^{(k)}(\vec{x})$$

Letting $\vec{y}$ be the vector of Lagrange multipliers or "dual variables" and doing some derivations, we get the dual objective function W defined (for $\vec{y} \geq 0$), as below:

$$\begin{aligned} W(\vec{y}) &= \min_x \{\ell(\vec{x}, \vec{y}); \alpha_j \leq x_j \leq \beta_j \text{ for all } j\} \\ &= r_0 - \vec{y}^T \vec{b} + \sum_{j=1}^{n} W_j(\vec{y}) \end{aligned}$$

where $W_j(\vec{y}) = \min_{x_j} \{l_j(x_j, \vec{y}); \alpha_j \leq x_j \leq \beta_j\}$

This formulation is beneficial since it "eliminates" $\vec{x}$.

The dual problem corresponding to $P^{(k)}$ is given as follows:

$$\begin{aligned} D\text{:} \quad & \text{maximize} \quad W(\vec{y}) \\ & \text{subject to} \quad \vec{y} \geq 0 \end{aligned} \tag{1.8}$$

D is a "nice" problem which may be solved by an arbitrary gradient method.

Once the dual problem has been solved the optimal solution of the primal subproblem $P^{(k)}$ is directly obtained by just pluggin in the optimal dual solution $\vec{y}$ in to the following expression:

$$x_j(\vec{y}) = \frac{\left(p_{0j} + \vec{y}^T \vec{p}_j\right)^{1/2} L_j + \left(q_{0j} + \vec{y}^T \vec{q}_j\right)^{1/2} U_j}{\left(p_{0j} + \vec{y}^T \vec{p}_j\right)^{1/2} + \left(q_{0j} + \vec{y}^T \vec{q}_j\right)^{1/2}}.$$

Chapter 2

SIMP Optimization

2.1 Solid Isotropic Material with Penalization (SIMP)

Volume-to-Point (VP) Heat Conduction Problem

Consider a finite-size volume in which heat is being generated at *every* point, and which is cooled through a small patch (the heat sink) located on its boundary. Suppose that we have a finite amount of high-conductivity (k_+) material available. Our goal is to determine the optimal distribution of the k_+ material through the given volume such that *the highest temperature is minimized*.

Solid Isotropic Material with Penalization (SIMP) is a method based on topology optimization that can be used to solve the VP Heat Conduction Problem. SIMP is what is called a soft kill method, meaning that in each step of the method we increase or decrease high-conductivity material by a small quantity. This allows us to apply methods designed for continuous optimization problems to this discrete problem.

2.1.1 Preliminary Parameters

Assumptions

In order to develop the method, we need to make a couple of assumptions.

First of all, the energy differential equation driving the heat-flux inside the finite-volume requires:

1. All calculations are run under steady-state conditions. That is, all heat produced in the volume is evacuated through the heat sink.

2. Low-conductivity materials (k_0) and high-conductivity materials (k_+) are treated as homogeneous and isotropic on their respective conductivities.

Throughout the article (Marck et al., 2012), the authors also set the following conditions:

- Thermal conductivities are constant:

$$k_0 = 1\frac{\mathrm{W}}{\mathrm{m^2\,K}} \qquad \text{and} \qquad k_+ = 100\frac{\mathrm{W}}{\mathrm{m^2\,K}}$$

- All structures have a square aspect ratio with $L = H = 0.1\mathrm{m}$
- The heat-sink is located on the middle of the west side of the structure
- The heat-sink has Dirichlet boundary conditions: $T_S = 0°\mathrm{C}$
- All other boundaries are adiabatic: $\nabla T = 0$

Notation

We have the following sets to describe the VP-problem:

- $\vec{x} \in \Omega$ = two-dimensional spatial field

  We set $\Omega = \Omega_0 \cup \Omega_+$ where

  - Ω_0 = portion of Ω that has conductivity k_0. This is the portion of the space where we have not placed any high-conductivity material.
  - Ω_+ = portion of Ω with conductivity k_+. This is the portion of the space with the high-conductivity material applied.

- $\vec{k} \in \mathbb{K}$ = distribution of high-conductivity material

- $T \in \overline{\mathbb{T}}$ = temperature field that satisfies the energy equation $\nabla \cdot \left(\vec{k} \nabla T_{\vec{k}}\right) + q = 0$ where q is the local heat-generating rate.

2.1.2 The Optimization Problem

Using the above established notation, we develop the following optimization problem:

$$\begin{aligned} &\text{minimize} && f(T) && \text{for } \; T \in \mathbb{T}, \vec{k} \\ &\text{subject to} && \nabla \cdot \left(\vec{k} \nabla T_{\vec{k}}\right) + q = 0 && \\ & && \vec{x} \in \Omega, \quad \vec{k} \in \mathbb{K}_{\text{ad}} && \end{aligned} \tag{2.1}$$

Penalization Process

The problem of whether to place high conductivity material in a particular location or not is discrete in nature. This is unfortunate as continuous optimization problems are generally easier to solve. In particular, we cannot apply some of the optimization methods described earlier, such as gradient descent, to a discrete optimization problem as they require the optimization variables to be continuous.

The SIMP method has a clever way of getting around this particular issue of discrete variables: create a continuous function that allows for a "mix" of the two conductive materials. This function turns our discrete variables into a continuous one, allowing us to apply the nice methods used in continuous optimization problems. However, in reality, we cannot actually mix the two conductive materials and therefore need a solution that produces a 1—0 structure. That is, our final result needs to either have conductivity k_0 or k_+ at each point, not some fraction of each. Therefore, in each iteration of the SIMP process, we *penalize* the mixing of the material. Keeping this in mind we introduce a design parameter $\eta \in [0, 1]$ that controls the amount of mixing of the two materials:

$$k(\eta) = k_0 + (k_+ - k_0)\, \eta^p \qquad \text{with} \qquad 0 \leq \eta \leq 1 \qquad \text{and} \qquad p \geq 1. \tag{2.2}$$

An added bonus of this formulation of $k(\eta)$ is that it is of the form (1.4), and hence a convex function! Notice that when $\eta = 0$, $k(0) = k_0$ and when $\eta = 1$, $k(1) = k_+$. The value p in (2.2) is the *penalization parameter*. p aids in the convergence process; without p the SIMP method converges to a structure that is not 1—0: a composite structure where finite-volumes are made up of different proportions of k_0 and k_+ materials.

To converge to a 1—0 structure, we gradually increase p beginning from $p = 1$. Increasing p from 1 puts the objective function in (2.1) at a disadvantage if $\eta \neq 1$. Once p gets much larger than 1 the second term in $k(\eta)$ of (2.2) becomes much smaller than k_0 and hence $k(\eta) \approx k_0$ for values of $\eta \neq 1$. As a result, when trying to optimize $f(T)$, value of η in $(0, 1)$ are penalized which leads to design parameters taking on values of 0 or 1, creating a 1—0 structure.

"King Me": Avoiding the Checkerboard Problem

If one were to optimize conductive material placement to increase heat transfer on a standard rectangular grid, a simplistic solution would be to create a grid of alternating material types in each adjacent rectangle, like a checkerboard. This way, the heat is always flowing to adjacent cells. While this may indeed increase heat transfer, it doesn't necessarily decrease the average temperature in our object (or transfer the heat towards the heat-sink). Hence, avoiding this non-physical checkerboard solution is of concern.

In order to solve the heat equation (1.1), we employ the Finite Volume Method (described earlier). This involves splitting up our space into a finite number of control volumes. The checkerboard pattern emerges when the solution of our optimization process converges to a 1—0 structure which has some meshes that successively belong to the Ω_0 and Ω_+ sets (i.e.: adjacent grid volumes have alternating thermal conductivites). As a result, the heat transfer within the structure between k_+ and k_0 regions is maximized, artificially increasing the impact of adding k_+ material on the temperature T, which in turn minimizes the objective function in (2.1) (Versteeg, 2007). Typically, this pattern occurs locally but then spreads

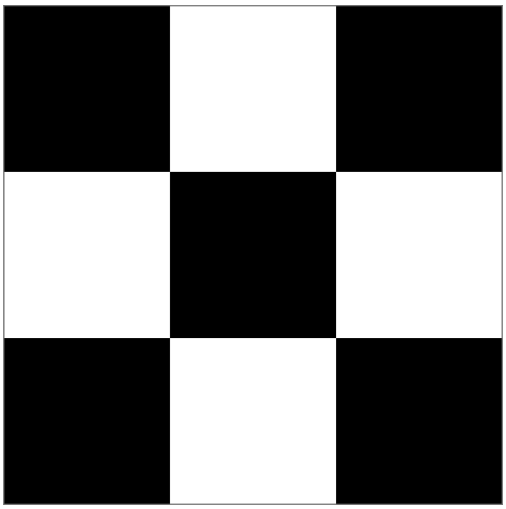

Figure 2.1: A checkerboard pattern result on a 3-by-3 control volume grid. The black spaces represent areas where $\eta = 1$ and the white spaces represent areas where $\eta = 0$ in (2.2). This results in adjacent regions of alternating thermal conductivites k_+ and k_0, artificially maximizing heat transfer between control volumes.

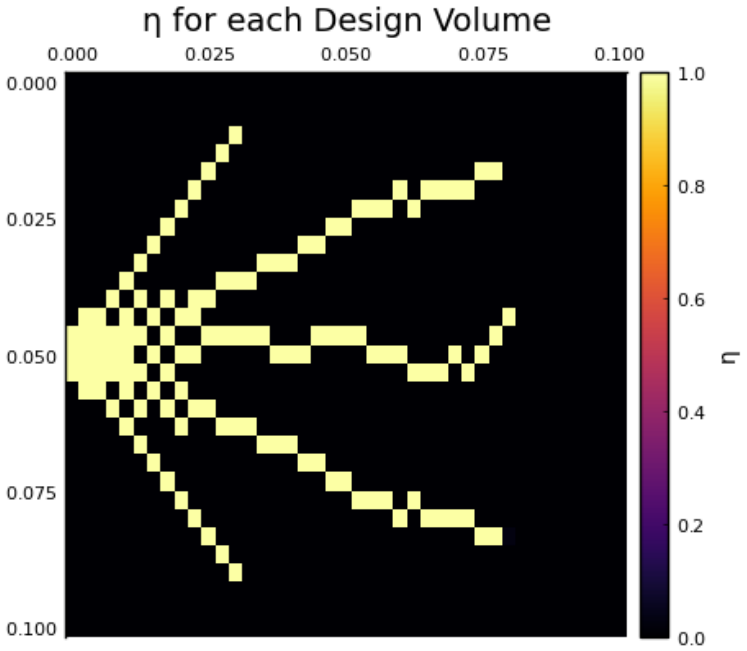

Figure 2.2: Output of SIMP algorithm for 30×40 temperature control volumes. Notice the local checkerboarding around $(0.05, 0.015)$.

throughout the entire structure through successive iterations of the optimization process. However, in the real world, these checkerboard placements of our conductive materials do not actually have the effect of lowering the average temperature in structures. In fact, the checkerboard example in Figure 2.1 doesn't even direct heat towards the heat-sink on the left wall of the structure.

In order to avoid obtaining checkerboard solutions from our optimization process, we employ two separate staggered grids for our temperature and design variables. We need to employ some extra equations in order to translate between design and temperature variables, but this strategy adequately solves the issue of convergence to checkerboard solutions.

One of the grids contains the information related to the temperature scalars and the other stores information related to the design parameters, η.

2.1.3 Finite Volume Method Discretization

Any discretization method could be used to numerically solve the heat equation (1.1). In our implementation of the SIMP algorithm, the Finite Volume Method (described earlier in §1.1.2) was used. FVM is used to discretize the formation of the heat equation in (2.1):

$$\nabla \cdot (k \nabla T) + q = 0. \tag{2.3}$$

We create a rectangular grid of $N_T = m \times n$ temperature control volumes of size $\Delta x \times \Delta y$ (solid squares in Figure 2.3). Each element is indexed by $1 \leq i \leq m$ and $1 \leq j \leq n$, where i refers to the row and j the column of the control volume. The volume indexed $(i, j) = (1, 1)$ is located in the upper-left and $(i, j) = (m, n)$ in the bottom-right corner of the object. The temperature over the area of the temperature control volume (i, j) is considered to be $T^{i,j}$.

Around the upper left corner of each temperature control volume we create a corresponding design element (dashed squares in Figure 2.3). The area within each (i, j) design element has conductivity $k^{i,j} = k(\eta^{i,j})$ as evaluated by (2.2).

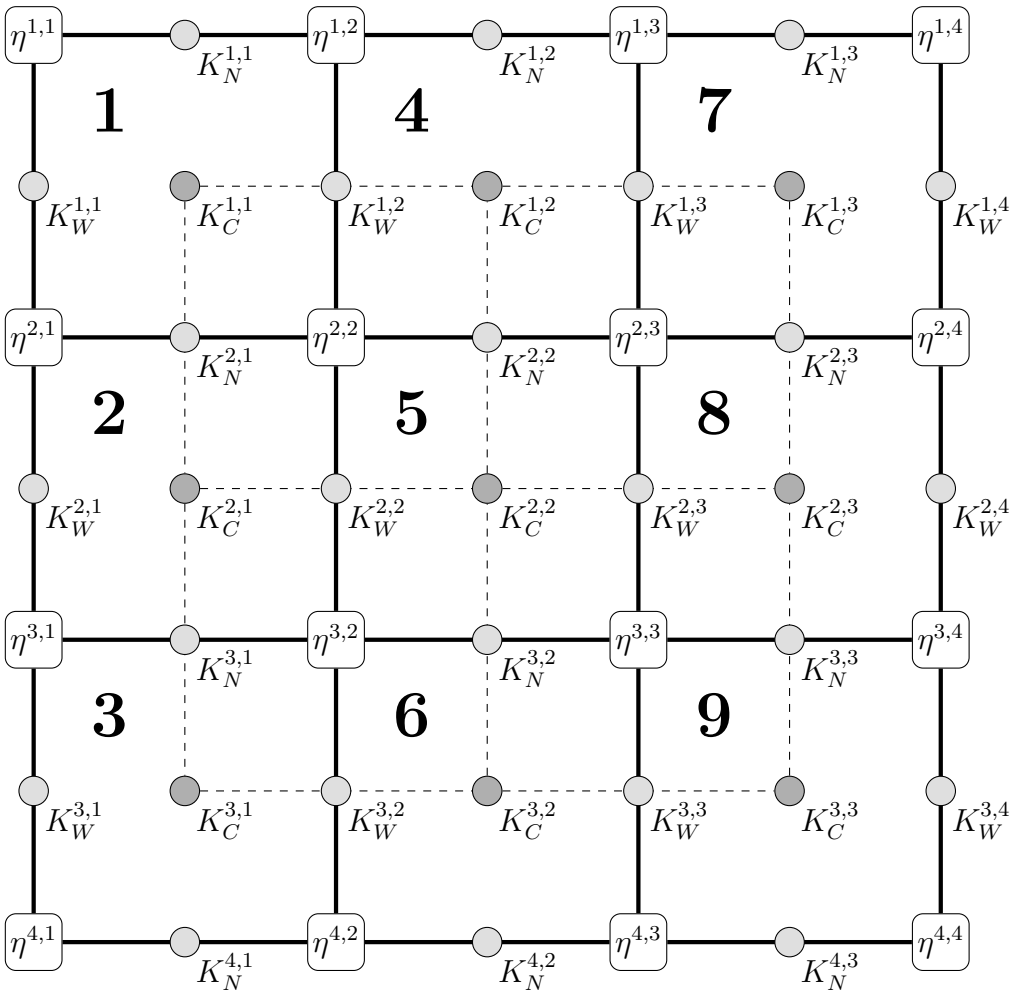

Figure 2.3: Overlayed Temperature (—) and Design (– – –) grids with 4×4 Design Element ($\eta^{i,j}$) and 3×3 Temperature Control Volume ($K_C^{i,j}$) Nodes. $K_N^{i,j}$ and $K_W^{i,j}$ indicate nodes at the North and West boundaries, respectively, of each Temperature Control Volume. Each Temperature control volume is numbered beginning in the upper left and continuing column-by-column, left-to-right.

In order to employ the finite volume method, it is necessary to be able to calculate the temperature fluxes along the boundaries of each control volume. To do this, we need to have a value for the conductivity along the faces of each control volume. Notice that the faces of the temperature control volumes lie within two adjacent design regions, which implies that there are two different conductivites along that face. To create a consistent conductivity along the control volume wall, we average (using either an arithmetic or harmonic mean) the conductivity of the two adjacent design nodes. Hence the conductivity along the West face of control volume (i, j), denoted by $k_W^{i,j}$, is given by

$$\text{Arithmetic Mean: } k_W^{i,j} = \frac{k^{i,j} + k^{i+1,j}}{2} \qquad \text{or} \qquad \text{Harmonic Mean: } k_W^{i,j} = 2\left(\frac{1}{k^{i,j}} + \frac{1}{k^{i+1,j}}\right)^{-1}. \tag{2.4}$$

Similarly, the conductivity along the North face of control volume (i, j), denoted by $k_N^{i,j}$, is given by

$$\text{Arithmetic Mean: } k_N^{i,j} = \frac{k^{i,j} + k^{i,j+1}}{2} \qquad \text{or} \qquad \text{Harmonic Mean: } k_N^{i,j} = 2\left(\frac{1}{k^{i,j}} + \frac{1}{k^{i,j+1}}\right)^{-1}. \tag{2.5}$$

For temperature control volume (i, j), the finite volume method discretizes (2.3) into the following linear equation

$$K_C^{i,j} T^{i,j} = K_W^{i,j} T^{i,j-1} + K_W^{i,j+1} T^{i,j+1} + K_N^{i,j} T^{i-1,j} + K_N^{i+1,j} T^{i+1,j} + Q^{i,j}, \tag{2.6}$$

where the $K^{i,j}$ terms represent the diffusive flux coefficients, $T^{i,j}$ the temperature of control volume (i, j), and $Q^{i,j}$ the heat generation of volume (i, j).

The value of the flux at the center node of the control volume is equal to the total flux through the volume faces, so we have an additional equation to pair with (2.6):

$$K_C^{i,j} = K_W^{i,j} + K_W^{i,j+1} + K_N^{i,j} + K_N^{i+1,j} \tag{2.7}$$

The K_W and K_N coefficients are dependent on the thermal conductivity and cross-sectional area of their corresponding faces:

$$K_W^{i,j} = \frac{k_W^{i,j}\Delta y}{\Delta x} \qquad \text{and} \qquad K_N^{i,j} = \frac{k_N^{i,j}\Delta x}{\Delta y} \tag{2.8}$$

Putting together (2.4), (2.5), (2.6), (2.7), and (2.8) for all N_T control volumes gives us a system of equations that discretize (2.3). Collecting the coefficients K into a matrix, representing the T and Q values as vectors, and doing a little reorganizing, we can represent the system of equations as a matrix equation:

$$\mathbf{KT} = \mathbf{Q}. \tag{2.9}$$

Let us take a moment to analyze the structure of the matrix $\mathbf{K}$, as it might not be immediately evident to the reader what the elements of this matrix represent. (It took the author some time to interpret the meaning of this matrix.) $\mathbf{K}$ is a sparse, symmetric, and pentadiagonal $mn \times mn$ matrix.

The entries in the matrix $\mathbf{K}$ indicate the coefficient of diffusive flux between numbered temperature control volumes. Notice in Figure 2.3 how the temperature control volumes are numbered down the columns. We can convert between volume index (i, j) and control volume number, #, using a simple function

$$\#(i, j) = i + m(j - 1) \tag{2.10}$$

Entry $\mathbf{K}[\alpha, \beta]$ is the flux coefficient between volumes number α and β. Since the flux coefficient between volumes α and β is the same as that between β and α, $\mathbf{K}[\alpha, \beta] = \mathbf{K}[\beta, \alpha]$, producing the symmetry of matrix $\mathbf{K}$. Since a particular control volume only interfaces with adjacent cells (and itself), each row/column will have (up to) five non-zero entries (all other entries are zero since there is no flux between cells that are not in contact with one another),

which produces the pentadiagonality and sparsity of $\mathbf{K}$. The i^{th} elements of the size mn vectors $\mathbf{T}$ and $\mathbf{Q}$ are the values of the temperature and heat generation of control volume number i.

Equation (2.9) is solved for $\mathbf{T}$ using any appropriate method, such as QR factorization. In our implementation we used the standard "\" operator in Julia: $\mathbf{T} = \mathbf{K} \backslash \mathbf{Q}$.

2.1.4 Discretized Optimization Problem

Now that we have been able to discretize (2.3), we can update the optimization problem (2.1):

$$\begin{aligned} \text{minimize} \quad & f(\mathbf{T}) \\ \text{subject to} \quad & \mathbf{KT} = \mathbf{Q} \\ & \frac{1}{N_T} \sum_{i=1}^{N_T} \boldsymbol{\eta}_i \leq \overline{\phi} \\ & \mathbf{k} = k_0 \mathbf{1} + (k_+ - k_0) \boldsymbol{\eta}^p \\ & 0 \leq \boldsymbol{\eta} \leq 1 \text{ and } p \geq 1 \end{aligned} \tag{2.11}$$

Here $\boldsymbol{\eta}$ represents the vector of design parameter values for each control volume. Additionally, we introduce the variable $\overline{\phi}$ which represents the maximum porosity, the maximal fraction of high-conductivity material allowed within the domain.

For our implementation of the SIMP method for this problem, the average temperature was chosen as the objective function:

$$f_{av}(\mathbf{T}) = \frac{1}{N_T} \sum_{i=1}^{N_T} \mathbf{T}_i. \tag{2.12}$$

Additionally, we opted to have no heat generation for each control volume:

$$\mathbf{Q} = \mathbf{0}. \tag{2.13}$$

We use the Method of Moving Asymptotes (MMA) to update the design parameters $\boldsymbol{\eta}$ throughout the optimization process. To create a local and convex approximation of the problem (see §1.4), MMA requires both function and constraint evaluations, as well as evaluations of the respective gradients. Hence, we need to calculate the gradients of the average temperature function and porosity constraint. The gradients of the functions with respect to the design parameters indicate the *sensitivity* of those functions to changes in $\boldsymbol{\eta}$, and hence analysis of these derivatives is called *sensistivity analysis*.

Sensitivity Analysis

We seek to find the partial derivatives of f_{av} and $\overline{\phi}$ with respect to an arbitrary design element η_ℓ so that we can form the gradients. That is, we are looking to find expressions for

$$\frac{\partial f_{av}}{\partial \eta_\ell} \quad \text{and} \quad \frac{\partial \overline{\phi}}{\partial \eta_\ell}.$$

The adjoint method is employed to make the calculation of the partial derivative of f_{av} easier to find. Similar to the method of Lagrange multipliers (Johnson, 2021), by assuming (2.9) is true, we add a clever form of $\mathbf{0}$ to the objective function f_{av}:

$$f_{av}(\mathbf{T}) = \frac{1}{N_T}\sum_{i=1}^{N_T}\mathbf{T}_i + \lambda \cdot \underbrace{(\mathbf{KT}-\mathbf{Q})}_{=\mathbf{0}}. \tag{2.14}$$

References

Boyd, S. P. & Vandenberghe, L. (2004). *Convex Optimization*. Cambridge Univ. Pr.

Johnson, S. G. (2021). Notes on adjoint methods for 18.335. WebPage.

Marck, G., Nemer, M., Harion, J.-L., Russeil, S., & Bougeard, D. (2012). Topology optimization using the SIMP method for multiobjective conductive problems. *Numerical Heat Transfer, Part B: Fundamentals*, 61(6), 439–470.

Shewchuk, J. R. (1994). An introduction to the conjugate gradient method without the agonizing pain. *Quake Project*.

Versteeg, H. K. (2007). *An Introduction to Computational Fluid Dynamics : The Finite Volume Method*. Harlow, England New York: Pearson Education Ltd.

Appendix A

Julia Codes

A.1 Backtracking Line Search

Here is an implementation of the Backtracking Line Search in Julia with default values for the parameters being $\alpha = 0.25$ and $\beta = 0.5$.

```
function ln_srch(d_dir,x,f,fx,dfx;alpha=0.25,beta=0.5)
t = 1
x1 = x+t*d_dir
y1 = f(x1)
y2 = fx+alpha*t*(dfx)'*d_dir
while y1 > y2
t = beta*t
x1 = x+t*d_dir
y1 = f(x1)
y2 = fx+alpha*t*(dfx)'*d_dir
end
return t
end
```

A.2 Gradient Descent

```
using LinearAlgebra


#Function to Optimize
```

```
f(x)=(x[2])^3-x[2]+(x[1])^2-3x[1]

#Gradient of Function
df(x)=[2x[1]-3,3x[2]^2-1]

#Initial Point
x=[0,0]

#Gradient Descent Algorithm
function grad_d(f,df,x)
d_dir = -df(x)
t = ln_srch(d_dir,x,f,f(x),df(x))
x = x + t*d_dir
return x
end

#Compute Minimum for Defined Tolerance
while norm(df(x))>0.00001
global x = grad_d(f,df,x)
end

display(x)
```