Space Mapping: Models, Sensitivities, and Trust-Regions Methods

Luís N. Vicente *

May 24, 2002

Abstract

The goal of this paper is to organize some of the mathematical and algorithmic aspects of the space-mapping technique for continuous optimization with expensive function evaluations. First, we consider the mapping from the fine space to the coarse space when the models are vector-valued functions and when the space-mapping (nonlinear) least-squares residual is nonzero. We show how the sensitivities of the space mapping can be used to deal with space-mapping surrogates of the fine model. We derive a framework where it is possible to design globally convergent trust-region methods to minimize such fine-model surrogates.

We consider also a different perspective of space mapping and apply it, for sake of simplicity, to the situation where the models are scalar functions. The space mapping is defined in a way where it is reasonable to assume that it is point-to-point. We prove that the surrogate model built by composition of the space mapping and the coarse model is a regular function. We also discuss trust-region methods in this context.

Keywords. space mapping, surrogate-based optimization, trust-region methods, global convergence, sensitivities

AMS subject classifications. 49M37, 90C06, 90C30, 90C31

1 Introduction

New techniques have been recently developed to deal with optimization problems that involve expensive function evaluations that may require long cpu calculations. Space mapping assumes the existence of two models for the same physical phenomenon: a fine model, accurate and expensive, and a coarse model, significantly cheaper and considerably less accurate. The idea behind space mapping is to construct a mapping between the fine-model space of parameters or variables and the coarse-model space that allows to defer the minimization process to the coarse model, where most function evaluations should take place. Space-mapping techniques are typically iterative as the mapping is unknown *a priori* and it is calculated for a sequence of points in the fine space.

The space-mapping technique was introduced first by Bandler et al. [5] in 1994. It has been modified and enhanced by classical optimization methods for nonlinear optimization. Bandler et al [6] proposed the use of Broyden's method to construct linear approximations for the space mapping and Bakr et al. [2] applied the trust-region technique to globalize the minimization process. These

*Departamento de Matemática, Universidade de Coimbra, 3001-454 Coimbra, Portugal (lvicente@mat.uc.pt). Support for this work was provided by Centro de Matemática da Universidade de Coimbra, by FCT under grant POCTI/35059/MAT/2000, and by the European Union under grant IST-2000-26063.

and other approaches are reviewed in the paper by Bakr et al. [3] and in the masters thesis of Søndergaard [15]. Leary, Bhaskar, and Keane [11] introduced space-mapping techniques for the treatment of models that appear as constraints. New space-mapping applications are reported in the papers collected in the volume edited by Nielsen [13] (see also [9]).

We address first in this paper the mapping from the fine space to the coarse space when the models are vector-valued functions, as analyzed in the work by Bakr et al. [4]. We show that the sensitivities of the space mapping P, defined in (1), can be calculated provided first-order derivatives of the fine model f and first-order and second-order derivatives of the coarse model c are given and some invertibility is assumed related with the size of the space-mapping (nonlinear) least-squares residual. The sensitivities of the space mapping P define the linearization P_ℓ of this mapping. Thus, we can use $c \circ P_\ell$ to locally minimize the surrogate $c \circ P$ that the space mapping P provides for the fine model f. An alternative surrogate introduced by Bakr et al. [4] is $w(c \circ P) + (1-w) f_\ell^{app}$, where $w \in [0,1]$ and f_ℓ^{app} is an approximation to the linearized model of the fine model f. In a similar way, we can work with $w(c \circ P_\ell) + (1-w) f_\ell^{app}$ to minimize $w(c \circ P) + (1-w) f_\ell^{app}$. The idea behind this linear combination is to introduce more accurate local information of the fine model. We show how to develop trust-region methods that are globally convergent to stationary points of these surrogates.

We address then a different situation where the fine and the coarse models are scalar functions, denoted by g and $\hat{g}$, respectively. The shape of the surrogate $\hat{g} \circ P$, defined by the composition of the space mapping P and the coarse model $\hat{g}$, is investigated. Given a point x in the fine space, the space-mapping image $P(x)$ is defined in this context by minimizing, in the coarse space, the distance to x subject to the matching of the coarse model to the fine-model value $g(x)$, see (21) and (22). It is possible to observe that such definition of space mapping yields a point-to-point map in several instances where space mapping based only on the matching of the models is point-to-set. When P is point-to-point, it is proved that the surrogate $\hat{g} \circ P$ is a regular function, i.e., that it has always first-order directional derivatives. The surrogate $\hat{g} \circ P$ coincides with the fine model except possibly near minimizers of the coarse model where it may become flat. The transition can create kinks, the source of non-differentiability. We also discuss trust-region methods to minimize this type of surrogate models.

We have structured this paper in two main sections, corresponding to the two space-mapping approaches mentioned above. At the end of each section, we draw some conclusions and discuss possible extensions. Norms and inner products used in this paper are the ℓ_2 ones.

2 Space mapping using vector-valued models

2.1 The space-mapping definition

Let us consider a physical phenomenon where the variables defining it belong to a subset of $\mathbb{R}^n$ and the function values that define it belong to $\mathbb{R}^m$. We follow the approach in Bakr et al. [4] and define space mapping by considering a fine model of this phenomenon denoted by f with $f : S^{(f)} \to \mathbb{R}^m$, and a coarse model represented by c with $c : S^{(c)} \to \mathbb{R}^m$, where $S^{(f)}, S^{(c)} \subset \mathbb{R}^n$. The fine model f is expensive to evaluate but the coarse model c is relatively cheap. It is assumed that $m > n$ (the practical situation in mind is when $m \gg n$). The case $m \leq n$ requires a different, more general approach, gives rise to nondifferentiable surrogates, and will be discussed in section 2.5.

The performance of both models is measured by a merit function $H : \mathbb{R}^m \to \mathbb{R}$. In several engineering applications H is not differentiable as it may result from the use of the ℓ_∞ norm. We will assume in this paper that H is quadratic: for instance the squared ℓ_2 norm in $\mathbb{R}^m$, or other

quadratic variants based on the squared ℓ_2 distance (as it is the case in several data fitting and parameter estimation problems).

The goal is to minimize

$$H_f \stackrel{\text{def}}{=} H \circ f$$

by considering the surrogate

$$H_c \stackrel{\text{def}}{=} H \circ c$$

and the composition of H_c with a mapping relating the models f and c.

The space mapping $P : S^{(f)} \to S^{(c)}$ is based on the solution of a nonlinear least-squares minimization problem, in the following way:

$$P(x) \stackrel{\text{def}}{=} \text{argmin}_{\hat{x} \in S^{(c)}} \frac{1}{2} \|c(\hat{x}) - f(x)\|^2 . \tag{1}$$

(We will assume in this paper that the minimal argument is always unique and therefore we can consider the notation where argmin returns a point and not a singleton.)

Bakr et al. [4] consider also a linear approximation $p(x)$ for $P(x)$ constructed by Broyden's method, looking then at the surrogate $H \circ c \circ p$ that take values in the fine space $S^{(f)}$. The surrogate they work with is actually given by

$$H\left((w)c(p(x)) + (1-w)\ell(x)\right) \tag{2}$$

where $w \in [0,1]$ is a weighted parameter and ℓ is a linear approximation for $f : S^{(f)} \to \mathbb{R}^m$. Their linear approximation ℓ is computed using once again Broyden's method and the term $(1-w)\ell$ provides to the surrogate more local accurate information about the fine model.

2.2 Space-mapping sensitivities and adjoints

Assuming that P is well defined as a point-to-point map and assuming appropriate smoothness for f and c, the space-mapping image $P(x)$ is given by the first-order necessary conditions for (1):

$$J_c(P(x))^\top (c(P(x)) - f(x)) \;=\; 0\,, \tag{3}$$

where J_c denotes the Jacobian of c. We will assume that $S^{(f)}$ and $S^{(c)}$ are open domains and that (3) is true for all $x \in S^{(f)}$.

2.2.1 Sensitivities of the space mapping

To compute the sensitivities of P, $J_P : S^{(f)} \to \mathbb{R}^{n \times n}$, we now differentiate (3) with respect to x, yielding

$$\sum_{i=1}^{m} [c_i(P(x)) - f_i(x)] \nabla^2 c_i(P(x)) J_P(x) + J_c(P(x))^\top \left(J_P(x)^\top J_c(P(x))^\top - J_f(x)^\top \right)^\top \;=\; 0\,, \tag{4}$$

where J_f and J_c denote the Jacobians of f and c, respectively. Thus, $J_P(x)$ can be computed from

$$G(x)\, J_P(x) \;=\; J_c(P(x))^\top J_f(x)\,,$$

where

$$G(x) \stackrel{\text{def}}{=} \sum_{i=1}^{m} [c_i(P(x)) - f_i(x)] \nabla^2 c_i(P(x)) + J_c(P(x))^\top J_c(P(x))\,.$$

Since $G(x)$ is symmetric, we also have

$$J_P(x)^\top G(x) \;=\; J_f(x)^\top J_c(P(x))\,. \tag{5}$$

Had we assumed that $f(x) = c(P(x))$ for all $x \in S^{(f)}$, which is ideally the underlying motivation, we would have obtained $J_f(x) = J_c(P(x))J_P(x)$, consistently with (4,5).

The calculation (5) of the sensitivities $J_P(x)$ requires the solution of n systems of linear equations with the matrix $G(x)$. It also requires the evaluation of first-order derivatives of the fine model and the evaluation of first-order and second-order derivatives of the coarse model. We will see later that it is not $J_P(x)$ but rather its action on appropriate vectors that needs to be computed.

We will assume that $G(x)^{-1}$ exists for all x in $S^{(f)}$. The case where $J_P(x)$ and $J_f(x)$ are approximated, say by $J_P^{app}(x)$ and $J_f^{app}(x)$, respectively, will be discussed later.

2.2.2 Gradient of the surrogate $H_{c(P)}$

The space mapping provides a surrogate model $H_{c(P)} \stackrel{\text{def}}{=} H \circ c \circ P$ for the fine-model function H_f. The next iteration involves solving

$$\min_{x \in S^{(f)}} \; H_{c(P)}(x) = H(c(P(x)))\,.$$

The sensitivities of P provide the gradient for the surrogate $H_{c(P)}$:

$$\nabla H_{c(P)}(x) \;=\; J_P(x)^\top J_c(P(x))^\top \nabla H(c(P(x)))\,,$$

where $\nabla H(c(P(x)))$ is the gradient of H at $c(P(x))$. One can see that the gradient $\nabla H_{c(P)}(x)$ can be also computed by an adjoint-type calculation (note that $G(x)$ is symmetric):

$$\nabla H_{c(P)}(x) \;=\; J_f(x)^\top J_c(P(x)) G(x)^{-1} J_c(P(x))^\top \nabla H(c(P(x)))\,,$$

requiring the solution of a single system of linear equations with $G(x)$.

2.3 Trust-region methods for minimizing the surrogate $H_{c(P)}$

2.3.1 A quadratic model for the surrogate $H_{c(P)}$

Given the sensitivities $J_P(x)$, one can consider a local linear model $P_\ell(x + \cdot)$ for $P(x)$ near x:

$$P_\ell(x + s) \;\stackrel{\text{def}}{=}\; P(x) + J_P(x)s\,. \tag{6}$$

The minimization of the surrogate $H_{c(P)}$ can be carried out by a trust-region approach. To compute a step s from x, we introduce a trust-region subproblem of the type

$$\min_{\|s\| \le \Delta} H_{c_\ell(P_\ell)}(x + s) \stackrel{\text{def}}{=} (H \circ c_\ell \circ P_\ell)\,(x + s)\,, \tag{7}$$

where $\Delta > 0$ is the trust radius. Here $c_\ell(P(x) + \cdot)$ denotes a local linear model of the coarse model near $P(x)$ with exact first-order information, i.e., a linear model of the form

$$c_\ell(P(x) + \hat{s}) \;\stackrel{\text{def}}{=}\; c(P(x)) + J_c(P(x))\hat{s}\,. \tag{8}$$

At this point it is important to remark that we are using c_ℓ instead of c. Since the coarse model c is cheap to evaluate, it is reasonable to expect that c could be used directly instead of

being approximated, as happens in [4]. The algorithmic approaches that we develop next could be carried out in that way, with $H_{c_\ell(P_\ell)}$ replaced by $H_{c(P_\ell)}$. However, we remark that for global convergence purposes, the surrogate $H_{c(P_\ell)}$ would be required to yield a condition of the type (11) and it is not clear that that would hold for every coarse model c. We will return to this point later.

Since $H_{c_\ell(P_\ell)} = H \circ c_\ell \circ P_\ell$ has been defined by the composition of two linear models (6,8) holding exact first-order information with the quadratic H, we obtain that $H_{c_\ell(P_\ell)}$ is itself a quadratic model, of the form

$$H_{c_\ell(P_\ell)}(x+s) \;=\; a(x) + \langle b(x), s\rangle + \frac{1}{2}\langle s, B(x)s\rangle\,,$$

where

$$\begin{aligned} a(x) &= H_{c(P)}(x)\,, \\ b(x) &= \nabla H_{c(P)}(x) \;=\; J_P(x)^\top J_c(P(x))^\top \nabla H(c(P(x)))\,, \\ B(x) &= J_P(x)^\top J_c(P(x))^\top\, \nabla^2 H\, J_c(P(x)) J_P(x)\,, \end{aligned}$$

$\nabla H(c(P(x)))$ is the gradient of H at $c(P(x))$, and $\nabla^2 H$ is the Hessian of the quadratic H.

2.3.2 Cauchy decrease

The step s can be required to satisfy a fraction of Cauchy decrease:

$$H_{c_\ell(P_\ell)}(x) - H_{c_\ell(P_\ell)}(x+s) \;\geq\; \kappa \left(H_{c_\ell(P_\ell)}(x) - H_{c_\ell(P_\ell)}(x+s^C) \right)\,, \tag{9}$$

where $\kappa \in (0,1]$. Here s^C is the Cauchy step defined by $s^C = -\alpha^C \nabla H_{c(P)}(x)$, with α^C given by the solution of the one-dimensional problem:

$$\alpha^C \;=\; \operatorname{argmin}_{\alpha>0,\, \|-\alpha \nabla H_{c(P)}(x)\| \leq \Delta} H_{c_\ell(P_\ell)}(x - \alpha \nabla H_{c(P)}(x))\,.$$

There are several algorithms that produce steps satisfying the fraction of Cauchy decrease condition (9); see [8].

Since $H_{c_\ell(P_\ell)}$ is quadratic, a result due to Powell [14, theorem 4] (see also [8, section 6.3], [12, lemma 4.8]) implies

$$H_{c_\ell(P_\ell)}(x) - H_{c_\ell(P_\ell)}(x+s^C) \;\geq\; \frac{1}{2}\|\nabla H_{c(P)}(x)\| \min\left\{\Delta, \frac{\|\nabla H_{c(P)}(x)\|}{\|B(x)\|}\right\}\,. \tag{10}$$

where $B(x)$ is the Hessian of the quadratic model $H_{c_\ell(P_\ell)}$ of the surrogate $H_{c(P)}$ as defined before. The Hessian $B(x)$, or a symmetric approximation thereof, will be assumed uniformly bounded across all iterations of the trust-region methods. The lower bound (10) for the decrease obtained in $H_{c_\ell(P_\ell)}$ by s^C, together with the fraction of Cauchy decrease condition (9) imply

$$H_{c_\ell(P_\ell)}(x) - H_{c_\ell(P_\ell)}(x+s) \;\geq\; \frac{\kappa}{2}\|\nabla H_{c(P)}(x)\| \min\left\{\Delta, \frac{\|\nabla H_{c(P)}(x)\|}{\|B(x)\|}\right\}\,. \tag{11}$$

This estimate is key to prove global convergence of trust-region methods to stationary points of $H_{c(P)}$.

One can replace c_ℓ by c and still retain global convergence provided (11), or alternatively (9) and (10), is valid for c instead of c_ℓ. More elaborated model managing techniques (see Alexandrov et al. [1]) could be applied to enforce (11) with c_ℓ replaced by c.

2.3.3 Minimization of the surrogate $H_{c(P)}$ using surrogate function values

If the goal is only to minimize the surrogate $H_{c(P)}$, then we have all the ingredients we need to identify a class of trust-region methods that are able to converge to stationary points of $H_{c(P)}$.

In fact, all it takes is to require the step s to satisfy the fraction of Cauchy decrease condition (9) and to accept the step s and possibly increase Δ if

$$\frac{\text{ared}(x,s)}{\text{pred}(x,s)} \stackrel{\text{def}}{=} \frac{H_{c(P)}(x) - H_{c(P)}(x+s)}{H_{c_\ell(P_\ell)}(x) - H_{c_\ell(P_\ell)}(x+s)} \geq \eta_1 \tag{12}$$

and, otherwise, to reject the step s (reducing Δ to $\gamma_1\Delta$ and recomputing a new step s yielding (9)). The constants γ_1 and η_1 must belong to $(0,1)$ and be fixed across all iterations. We describe next this family of trust-region methods.

Algorithm 2.1 *Trust-region methods for the minimization of $H_{c(P)}$ using surrogate function values*
Let $x_0 \in \mathbb{R}^n$, $\Delta_0 > 0$, and $\gamma_1, \eta_1 \in (0,1)$ be given.

For $k = 0, 1, 2, \ldots$

- *Compute a step s_k from the trust-region subproblem (7) that satisfies (9), for $x = x_k$.*
- *Let*
$$\rho_k = \frac{ared(x_k, s_k)}{pred(x_k, s_k)} = \frac{H_{c(P)}(x_k) - H_{c(P)}(x_k + s_k)}{H_{c_\ell(P_\ell)}(x_k) - H_{c_\ell(P_\ell)}(x_k + s_k)}.$$
- *If $\rho_k \geq \eta_1$ then $x_{k+1} = x_k + s_k$ and Δ_{k+1} is chosen so that $\Delta_{k+1} \geq \Delta_k$.*
- *If $\rho_k < \eta_1$ then $x_{k+1} = x_k$ and $\Delta_{k+1} = \gamma_1 \Delta_k$.*

end

The rules to update the trust radius Δ_k are in practice more sophisticated. What we have just described enables the method to achieve global convergence and it is verified by most implementations.

Any such trust-region method generates a sequence of iterates $\{x_k\}$ that verifies an asymptotic result of the following form [8, section 6.4]:

Theorem 2.1 *Let $H_{c(P)}$ be a continuously differentiable function with uniformly continuous gradient in $S^{(f)}$. Consider a sequence $\{x_k\}$ generated by a trust-region method of the form of algorithm 2.1. Let also $H_{c(P)}$ be bounded below on*

$$L(x_0) = \{x \in S^{(f)} : H_{c(P)}(x) \leq H_{c(P)}(x_0)\}.$$

Finally, let $\{B(x_k)\}$ be a bounded sequence. Then

$$\lim_{k \to +\infty} \|\nabla H_{c(P)}(x_k)\| = 0. \tag{13}$$

The assumptions of theorem 2.1 are posed in terms of the surrogate $H_{c(P)}$. Those assumptions are satisfied provided:

Conditions 2.1

- *$S^{(f)}$ and $S^{(c)}$ are open domains;*

- *the Jacobian of f is uniformly continuous in $S^{(f)}$;*
- *the Hessians of c_i, $i = 1, \ldots, m$, are uniformly continuous in $S^{(c)}$;*
- *$P : S^{(f)} \to S^{(c)}$ is a well defined point-to-point map, $H_{c(P)} = H \circ c \circ P$ is bounded below on $S^{(f)}$ (which is trivially satisfied when H is the squared ℓ_2 norm), (3) is true for all $x \in S^{(f)}$, and $G(\cdot)^{-1}$ exists in $S^{(f)}$.*

The strongest assumptions concern the well-definiteness and the smoothness of the mapping P. Assuming that $P : S^{(f)} \to S^{(c)}$ is a well defined point-to-point map and that $G(\cdot)^{-1}$ exists in $S^{(f)}$ is reasonable when $m \gg n$. It is difficult to establish scenarios under which these assumptions are verified, as the situation is highly problem dependent. It is reasonable to say that the chances of satisfying such assumptions increase as m becomes bigger and bigger than n, as the contribution of the semi-positive definite term $J_c^\top J_c$ in G becomes more and more relevant.

2.3.4 Minimization of the surrogate $H_{c(P)}$ using fine-model function values

Since we need to evaluate the fine model f to compute $H_{c(P)}(x)$ and $H_{c(P)}(x+s)$, we could think of replacing the actual reduction

$$H_{c(P)}(x) - H_{c(P)}(x+s)$$

given in (12), by

$$H_f(x) - H_f(x+s)\,.$$

Since we are inexactly approximating $H_{c(P)}$ by H_f in this algorithmic context, we need to impose the following conditions:

$$\begin{aligned} |H_f(x) - H_{c(P)}(x)| &\leq \eta_0 \operatorname{pred}(x,s), \\ |H_f(x+s) - H_{c(P)}(x+s)| &\leq \eta_0 \operatorname{pred}(x,s), \end{aligned} \tag{14}$$

where $0 < \eta_0 < \frac{1}{2}\eta_1$.

It is proved in [8, section 10.6] that the limit result (13) is retained if conditions (14) are satisfied. However, the satisfaction of these conditions might be problematic and expensive. They can be expensive because they may force the recomputation of H_f or $H_{c(P)}$ at x or $x+s$ more accurately (see [8, section 10.6]). But, more importantly, they can be problematic because there is no guarantee that the surrogate $H_{c(P)}$ agrees with the fine model H_f.

2.4 Trust-region methods for minimizing a surrogate based on $c \circ P$ and on the fine model f

Based on the work by Bakr et al. [4], in particular in what has been developed for their surrogate (2), we consider now the surrogate

$$H_w \stackrel{\text{def}}{=} H\{(w)c \circ P + (1-w)f\}\,,$$

with $w \in [0,1]$, and the corresponding quadratic model

$$q_w \stackrel{\text{def}}{=} H\{(w)c_\ell \circ P_\ell + (1-w)f_\ell\}\,,$$

where $f_\ell(x+\cdot)$ is a local linear model of the fine model f. (The brackets are used to easy notation. $H\{f\}$ represents $H \circ f$.)

We will assume now that both $J_P(x)$ and $J_f(x)$ are computed inexactly, and consider a local linear model of the fine model f with inexact first-order information, of the form

$$f_\ell^{app}(x+s) \;=\; f(x) + J_f^{app}(x)s\,,$$

using an approximation $J_f^{app}(x)$ for the Jacobian $J_f(x)$ of the fine model, and a local linear model of the space mapping with inexact first-order information, of the form

$$P_\ell^{app}(x+s) \;=\; P(x) + J_P^{app}(x)s\,,$$

using an approximation $J_P^{app}(x)$ for the sensitivities $J_P(x)$.

Thus, we get

$$q_w^{app}(x+s) \;=\; a_w(x) + \langle b_w(x), s\rangle + \frac{1}{2}\langle s, B_w(x)s\rangle\,,$$

where

$$\begin{aligned} b_w(x) &= \nabla_s H\left\{(w)c_\ell \circ P_\ell^{app} + (1-w)f_\ell^{app}\right\}(x+s) \\ &= \left((w)J_P^{app}(x)^\top J_c(P(x))^\top + (1-w)J_f^{app}(x)^\top\right)\nabla H((w)c(P(x)) + (1-w)f(x))\,, \end{aligned}$$

and $\nabla H((w)c(P(x)) + (1-w)f(x))$ is the gradient of H at $(w)c(P(x)) + (1-w)f(x)$. The Hessian $B_w(x)$, or a symmetric approximation thereof, is assumed to be uniformly bounded across all iterations of the trust-region methods.

We consider now a trust-region subproblem of the type

$$\min_{\|s\|\le\Delta}\; q_w^{app}(x+s)\,, \tag{15}$$

with $\Delta > 0$, and require the step s to satisfy the following fraction of Cauchy decrease condition:

$$q_w^{app}(x) - q_w^{app}(x+s) \;\ge\; \kappa_w\left(q_w^{app}(x) - q_w^{app}(x+s_w^C)\right)\,, \tag{16}$$

where $\kappa_w \in (0,1]$. Here s_w^C is the Cauchy step defined by $s_w^C = -\alpha_w^C b_w(x)$, with α_w^C given by the solution of the one-dimensional problem:

$$\alpha_w^C \;=\; \mathrm{argmin}_{\alpha>0,\,\|-\alpha b_w(x)\|\le\Delta}\; q_w^{app}(x - \alpha b_w(x))\,.$$

The step s is accepted and Δ is possibly increased if

$$\frac{\mathrm{ared}(x,s;w)}{\mathrm{pred}(x,s;w)} \;\stackrel{\mathrm{def}}{=}\; \frac{H_w(x) - H_w(x+s)}{q_w^{app}(x) - q_w^{app}(x+s)} \;\ge\; \eta_1\,.$$

Otherwise, the step s is rejected and Δ is reduced to $\gamma_1\Delta$. The constants γ_1 and η_1 must belong to $(0,1)$ and be fixed across all iterations.

The gradient of H_w at x is given by

$$\nabla H_w(x) \;=\; \left((w)J_P(x)^\top J_c(P(x))^\top + (1-w)J_f(x)^\top\right)\nabla H((w)c(P(x)) + (1-w)f(x))\,.$$

Since the term $b_w(x)$ used in $q_w^{app}(x+s)$ is not exactly the gradient $\nabla H_w(x)$, we need to impose the following condition [7],[8, section 8.4] on this first-order approximation:

$$\frac{\|\nabla H_w(x) - b_w(x)\|}{\|b_w(x)\|} \;\le\; \frac{\kappa_w(1-\eta_1)}{2}\,. \tag{17}$$

One can see that the term $b_w(x)$ is different from $\nabla H_w(x)$ because of the approximations $J_f^{app}(x)$ for $J_f(x)$ and $J_P^{app}(x)$ for $J_P(x)$. The approximation $J_f^{app}(x)$ for $J_f(x)$ is used explicitly in the formula for $b_w(x)$ and in the computation of $J_P^{app}(x)$. The sensitivities $J_P(x)$ can be inexact just because of the inexactness of $J_f^{app}(x)$. But even when exact first-order derivatives are available for the fine model f, the sensitivities computation can be inexact (e.g., it might result of the application of iterative linear solvers or of Broyden's method). Thus, condition (17) is controlling both the quality of the approximation of the first-order derivatives of the fine model f and the quality of the approximation of the sensitivities of P. We describe next this family of trust-region methods, this time for the surrogate H_w.

Algorithm 2.2 *Trust-region methods for the minimization of H_w*
Let $x_0 \in \mathbb{R}^n$, $\Delta_0 > 0$, and $\gamma_1, \eta_1 \in (0,1)$ be given.

For $k = 0, 1, 2, \ldots$

- *Compute $b_w(x_k)$ such that*

$$\frac{\|\nabla H_w(x_k) - b_w(x_k)\|}{\|b_w(x_k)\|} \;\leq\; \frac{\kappa_w(1-\eta_1)}{2}\,.$$

- *Compute a step s_k from the trust-region subproblem (15) that satisfies (16), for $x = x_k$.*
- *Let*

$$\rho_k \;=\; \frac{ared(x_k, s_k)}{pred(x_k, s_k)} \;=\; \frac{H_w(x_k) - H_w(x_k + s_k)}{q_w^{app}(x_k) - q_w^{app}(x_k + s_k)}.$$

- *If $\rho_k \geq \eta_1$ then $x_{k+1} = x_k + s_k$ and Δ_{k+1} is chosen so that $\Delta_{k+1} \geq \Delta_k$.*
- *If $\rho_k < \eta_1$ then $x_{k+1} = x_k$ and $\Delta_{k+1} = \gamma_1 \Delta_k$.*

end

The comment about the trust radius Δ_k made after algorithm 2.1 also applies here.

The global convergence result [7],[8, section 8.4] for any trust-region method in this family is summarized in the next theorem.

Theorem 2.2 *Let H_w be a continuously differentiable function with uniformly continuous gradient in $S^{(f)}$. Consider a sequence $\{x_k\}$ generated by a trust-region method of the form of algorithm 2.2. Let also H_w be bounded below on*

$$L_w(x_0) \;=\; \{x \in S^{(f)} :\; H_w(x) \leq H_w(x_0)\}\,.$$

Finally, let $\{B_w(x_k)\}$ be a bounded sequence. Then

$$\lim_{k \to +\infty} \|\nabla H_w(x_k)\| \;=\; 0\,.$$

The assumptions of theorem 2.2 are posed in terms of the surrogate H_w. Those assumptions are satisfied provided conditions 2.1 hold.

By tuning the parameter w iteratively, replacing w by w_k in algorithm 2.2, and by forcing w_k to converge to zero, we get an asymptotic result for the fine model:

Corollary 2.1 *Under the assumptions of the previous theorem, if $\lim_{k \to +\infty} w_k = 0$ then*

$$\lim_{k \to +\infty} \|\nabla H_f(x_k)\| \;=\; 0\,.$$

2.5 Discussion and extensions

Results describing global convergence to points satisfying second-order necessary conditions could also be proved for modified versions of algorithms 2.1 and 2.2, but such modifications are less realistic from a practical point of view since they would require, among other things, one more order of differentiability for f and c. Several other algorithmic enhancements could be considered. One could, for instance, use line-search techniques instead of the trust-region approach, developing algorithms also globally convergent. Quasi-Newton methods, such as the BFGS or SR1, or their limited memory versions, could be applied to improve the numerical behavior related with local convergence, without requiring more differentiability.

When $m \leq n$, the definition of the space mapping given by (1) gives easily rise to a point-to-set map, as it is expected that the system $c(\hat{x}) = f(x)$, for fixed x, has nonunique solutions in $S^{(c)}$. In this case, one could instead define $P(x)$ by looking at the problem

$$\begin{array}{rl} \min_{\hat{x} \in S^{(c)}} & \frac{1}{2}\|\hat{x} - x\|^2 \\ \text{s.t.} & c(\hat{x}) = f(x)\,. \end{array} \tag{18}$$

If $S(x) \stackrel{\text{def}}{=} \{\hat{x} \in S^{(c)} : \hat{c}(\hat{x}) = f(x)\} \neq \emptyset$, one could define $P(x)$ as the (uniquely assumed) solution of (18). Otherwise, $P(x)$ would be the (uniquely assumed) least-squares solution of the constraints in (18), already defined in (1). Such definition does not lead to a smooth mapping, even when the models f and c are smooth. In the next section, we will discuss space mapping using scalar-valued models, where we will consider a space mapping given by a problem of the form (18) with only one constraint. It will become clear from the context of the next section what type of nondifferentiability arises when space mapping is based on (18).

3 Space mapping using scalar-valued models

Let us consider a coarse model $\hat{g} : \hat{X} \subset \mathbb{R}^n \to \mathbb{R}$ of a fine model $g : X \subset \mathbb{R}^n \to \mathbb{R}$. A parallel to the previous notation can be drawn by considering $\hat{g} = H_c = H \circ c$, $S^{(c)} = \hat{X}$, $g = H_f = H \circ f$, and $S^{(f)} = X$. The goal is to minimize the fine model $g(x)$ in X.

3.1 The space-mapping definition

Let us assume that X and $\hat{X}$ are open sets of $\mathbb{R}^n$. If

$$S(x) \stackrel{\text{def}}{=} \{\hat{x} \in \hat{X} : \hat{g}(\hat{x}) = g(x)\} \neq \emptyset \tag{19}$$

then, assuming that the problem

$$\begin{array}{rl} \min_{\hat{x} \in \hat{X}} & \frac{1}{2}\|\hat{x} - x\|^2 \\ \text{s.t.} & \hat{g}(\hat{x}) = g(x) \end{array} \tag{20}$$

has an unique solution, we define $P(x)$ as

$$P(x) \stackrel{\text{def}}{=} \operatorname{argmin}_{\hat{x} \in \hat{X}} \frac{1}{2}\|\hat{x} - x\|^2 \ \text{ s.t. } \ \hat{g}(\hat{x}) = g(x)\,. \tag{21}$$

If the set $S(x)$ given in (19) is empty then $P(x)$ is given by the solution, assumed unique, of the unconstrained problem that consists of the minimization of the least-squares norm of the constraint $\hat{g}(\hat{x}) = g(x)$:

$$P(x) \stackrel{\text{def}}{=} \operatorname{argmin}_{\hat{x} \in \hat{X}} \frac{1}{2}\left(\hat{g}(\hat{x}) - g(x)\right)^2\,. \tag{22}$$

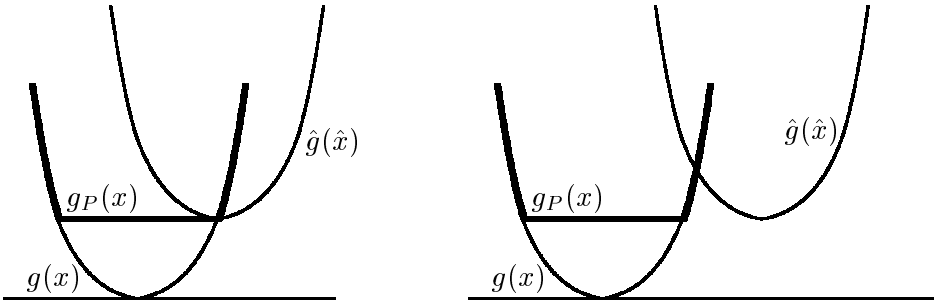

Figure 1: The surrogate $g_P = \hat{g} \circ P$ is the same in both examples.

3.2 The surrogate g_P

The space mapping provides a surrogate model $g_P \stackrel{\text{def}}{=} \hat{g} \circ P$ for the fine model. The next step involves solving

$$\min_{x \in X} g_P(x) = \hat{g}(P(x)) \,.$$

We analyze now the differentiability properties of the surrogate g_P. We show first that g_P is a regular function, i.e., that it has directional (or Gâteaux) derivatives along any direction and at any point in X. The proof of the next theorem is itself an introduction to the shape of the surrogate g_P.

For a better understanding of the proof let us first introduce a simple example. Let $X = \hat{X} = \mathbb{R}$, $g(x) = x^2$, and $\hat{g}(\hat{x}) = (\hat{x} - 1)^2 + 1$. In this example, $P(x) = 1$ and $g_P(x) = \hat{g}(P(x)) = 1$ for $x \in [-1, 1]$. Outside $[-1, 1]$, the fine model g and the surrogate $g_P = \hat{g} \circ P$ coincide. This example is depicted in figure 1 (left).

Theorem 3.1 *Let g and $\hat{g}$ be continuously differentiable functions in X and $\hat{X}$, respectively. Let us assume also that P is a well defined point-to-point map from X to $\hat{X}$.*

Then g_P is regular.

Proof: We will show that g_P has directional derivatives for every x in X. The proof is divided in three parts: in the first part we deal with the case where g_P coincides with g; in the second we will look at the case where g_P is flat; the last part analyzes the kinks.

Part 1.

In a neighborhood N_1 of X where for all $x \in N_1$ one has $S(x) \neq \emptyset$ and $\nabla \hat{g}(P(x)) \neq 0$, $P(x)$ must satisfy the first-order necessary conditions for (20):

$$P(x) - x + \lambda(x) \nabla \hat{g}(P(x)) \;=\; 0 \,, \tag{23}$$

$$\hat{g}(P(x)) \;=\; g(x) \,, \tag{24}$$

where $\lambda(x)$ is the multiplier corresponding to the constraint $\hat{g}(\hat{x}) = g(x)$. The fact that $\nabla \hat{g}(P(x)) \neq 0$ acts like the constraint qualification needed for the necessary conditions. Thus, in N_1, g_P coincides with g, and g_P is differentiable with a gradient given by

$$\nabla g_P(y) \;=\; \nabla g(y) \,.$$

Part 2.
If for a given x we have that $S(x) = \emptyset$, then $P(x)$ must verify the first-order necessary conditions for (22):

$$[\hat{g}(P(x)) - g(x)]\nabla\hat{g}(P(x)) \;=\; 0\,.$$

Since $\hat{g}(P(x)) \neq g(x)$ we obtain

$$\nabla\hat{g}(P(x)) \;=\; 0\,,$$

i.e., $P(x)$ is a stationary point for the coarse model $\hat{g}$. Moreover, we can easily prove by contradiction that $P(x)$ is either a local minimizer of $\hat{g}$ (when $\hat{g}(P(x)) > g(x)$) or a local maximizer of $\hat{g}$ (when $\hat{g}(P(x)) < g(x)$). A continuity argument shows that there exists a neighborhood N_2 of x where $S(y) = \emptyset$, $P(y) = P(x)$, and $\nabla\hat{g}(P(y)) = 0$ for all $y \in N_2$. Thus, P and g_P are constant in N_2. As a consequence, g_P is differentiable in N_2, and its gradient is given by

$$\nabla g_P(y) \;=\; 0\,.$$

Part 3.
We are left with situations characterized by the existence of points $z \in X$ where one has

$$\nabla\hat{g}(P(z)) \;=\; 0\,, \tag{25}$$

$$\hat{g}(P(z)) \;=\; g(z)\,. \tag{26}$$

In this situation one cannot appeal to (23)-(24) due to the apparent absence of a constraint qualification. Two cases can occur here and we analyze them separately.

The first case is when $S(\cdot)$ is still nonempty in a neighborhood of z. In this case we fall in the N_1-neighborhood situation described above, where g and g_P coincide, with the particularity that $\nabla g(z) = 0$, i.e., z is a stationary point for the fine model g.

The second case is when there is no neighborhood of z where the set $S(\cdot)$ is nonempty. One can also show here by contradiction that $P(z)$ is either a local minimizer or a local maximizer of $\hat{g}$. Furthermore, for any direction d either

$$g_P'(z;d) \;=\; \langle\nabla g(z), d\rangle$$

(when $\langle\nabla g(z), d\rangle \geq 0$ and $P(z)$ is a local minimizer of $\hat{g}$ or when $\langle\nabla g(z), d\rangle < 0$ and $P(z)$ is a local maximizer of $\hat{g}$), or

$$g_P'(z;d) \;=\; 0$$

(when $\langle\nabla g(z), d\rangle < 0$ and $P(z)$ is a local minimizer of $\hat{g}$ or when $\langle\nabla g(z), d\rangle \geq 0$ and $P(z)$ is a local maximizer of $\hat{g}$). We conclude that the directional derivative $g_P'(z;d)$ exists for all directions d. We remark that when $P(z)$ is a local minimizer of $\hat{g}$, we have

$$0 \;\in\; \partial g_P(z) \stackrel{\text{def}}{=} \{r \in \mathbb{R}^n : \langle r, d\rangle \leq g_P'(z;d) \text{ for all } d \in \mathbb{R}^n\}\,,$$

i.e., z is a stationary point for the surrogate function g_P. It can be proved here that z is a local minimizer of g_P, although not unique, since g_P is flat along directions d for which $\langle\nabla g(z), d\rangle < 0$. ∘

In the example where $X = \hat{X} = \mathbb{R}$, $g(x) = x^2$, and $\hat{g}(\hat{x}) = (\hat{x} - 1)^2 + 1$, there are two kinks, -1 and 1. We have that $P(-1) = 1$ and the gradient of $\hat{g}$ at $P(-1)$ is zero: there is no Lagrange

multiplier $\lambda(-1)$ that solves (23). At the other kink, we observe that $P(1) = 1$, but the gradient of $\hat{g}$ at $P(1)$ is also zero. Despite the lack of the linear independence constraint qualification, any real multiplier $\lambda(1)$ solves condition (23).

The proof provides significant insight about the surrogate g_P. There is however one point that has not been analyzed explicitely in the proof and that is relevant for the numerical minimization of g_P. Consider a sequence of points in a N_1-neighborhood that is converging to a kink point $z \in \mathrm{cl}(N_1)$, where z satisfies (25)–(26). We have that

$$\lim_{k \to +\infty} \|\nabla \hat{g}(P(y_k))\| \;=\; 0\,.$$

In such a situation, two cases can happen. The first case is when

$$\lim_{k \to +\infty} \|P(y_k) - y_k\| \;=\; 0\,,$$

and in this case the behavior of $\lambda(y_k)$ is not relevant, provided

$$\lim_{k \to +\infty} \lambda(y_k) \nabla \hat{g}(P(y_k)) \;=\; 0\,.$$

(In the example analyzed in this section this case corresponds to $z = 1$.) The second case corresponds to

$$\lim_{k \to +\infty} \|P(y_k) - y_k\| \;\neq\; 0\,, \tag{27}$$

where we must have

$$\lim_{k \to +\infty} \lambda(y_k) \;=\; +\infty\,.$$

(In the example analyzed in this section this case corresponds to $z = -1$.)

Thus, the sizes of the multiplier $\lambda(y_k)$ and of the distance $\|P(y_k) - y_k\|$ are an indication of the convergence to a kink point z, where $0 \in \partial g_P(z)$ and z is a local minimizer of g_P or where $0 \in -\partial g_P(z)$ and z is a local maximizer of g_P.

In the example that we have been considering, if we change the coarse model to $\hat{g}(x) = (\hat{x}-2)^2+1$ then we can see that $P(x) = 2$ for $x \in [-1, 1]$ but $g_P = \hat{g} \circ P$ does not change. The kinks -1 and 1 are now both of the second case (27). There is now a point $x = 5/4$ in a N_1-neighborhood for which $P(5/4) = 5/4$, $\nabla \hat{g}(P(5/4)) \neq 0$, and $\lambda(5/4) = 0$. This example is depicted in figure 1 (right). We illustrate also, in figure 2, a situation where the fine model has no minimizer but where the surrogate g_P can be successfully minimized.

Another relevant aspect is that condition (23) provides a local linear model for P around x:

$$P_\ell^{app}(y) \;=\; y - \lambda(x) \nabla \hat{g}(P(x))\,,$$

that might be useful to build a new (local) surrogate $\hat{g} \circ P_\ell^{app}$.

3.3 Discussion and extensions

The results of section 3 can be generalized in various ways. The approach is not restricted to $\mathbb{R}^n$ and could be easily developed in infinite dimensional spaces (Banach reflexive or Hilbert), by requiring Fréchet differentiability of the models g and $\hat{g}$ and by assuming the same type of well-definiteness for the space mapping. The norm used in (21) should be smooth to allow differentiability. The approach described here for $\mathbb{R}^n$ also works with ellipsoidal norms of the form $\|x\| = \|Q^{1/2}x\|$, where Q is a symmetric positive definite matrix.

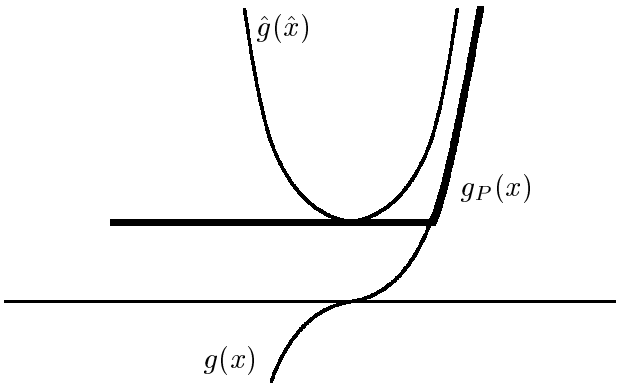

Figure 2: The surrogate $g_P = \hat{g} \circ P$ is bounded below despite the fact that the fine model g is unbounded.

Assuming that $P : X \to \hat{X}$ is point-to-point is certainly strong, problem dependent, and only guaranteed under special convexity assumptions. But such assumption allowed us to study the main properties of $g_P = \hat{g} \circ P$ (see the last paragraph in this section) whose flavor is also present when $P : X \to \hat{X}$ is point-to-set.

The regularity of g_P allows the application of the approach and global convergence results of Dennis, Li, and Tapia [10]. This paper considers a trust-region step that is an optimal solution of the trust-region subproblem. Conn, Gould, and Toint [8, chapter 11] generalized their approach for the case where the trust-region steps satisfy only a fraction of Cauchy decrease condition.

The analysis of section 3 for scalar-valued models has shown that the surrogate $g_P = \hat{g} \circ P$ may be flat when the image of P is close to a minimizer of $\hat{g}$. Thus, space-mapping techniques solely based on the minimization of g_P should be applied with caution and abandoned when flatness is encountered. The same comment applies to space-mapping techniques for vector-valued models when $m \leq n$, as discussed in section 2.5.

Acknowledgments

The author would like to thank John E. Dennis (Rice University, Houston, USA) and Kaj Madsen and Jacob Søndergaard (Technical University of Denmark) for their comments and suggestions on an earlier draft of this paper.

References

[1] N. M. Alexandrov, J. E. Dennis, R. M. Lewis, and V. Torczon, *A trust region framework for managing the use of approximation models in optimization*, Structural Optimization, 15 (1998), pp. 16–23.

[2] M. H. Bakr, J. W. Bandler, R. M. Biernacki, S. H. Chen, and K. Madsen, *A trust region agressive space mapping algorithm for EM optimization*, IEEE Trans. Microwave Theory Tech., 46 (1998), pp. 2412–2425.

[3] M. H. Bakr, J. W. Bandler, K. Madsen, and J. Søndergaard, *Review of the space mapping approach to engineering optimization and modeling*, Optimization and Engineering, 1 (2000), pp. 241–276.

[4] ——, *An introduction to the space mapping technique*, (2002). To appear in Optimization and Engineering.

[5] J. W. Bandler, R. M. Biernacki, S. H. Chen, P. A. Grobelny, and R. H. Hemmers, *Space mapping technique for electromagnetic optimization*, IEEE Trans. Microwave Theory Tech., 42 (1994), pp. 2536–2544.

[6] J. W. Bandler, R. M. Biernacki, S. H. Chen, R. H. Hemmers, and K. Madsen, *Electromagnetic optimization exploiting agressive space mapping*, IEEE Trans. Microwave Theory Tech., 43 (1995), pp. 2874–2882.

[7] R. G. Carter, *On the global convergence of trust region algorithms using inexact gradient information*, SIAM J. Numer. Anal., 28 (1991), pp. 251–265.

[8] A. R. Conn, N. I. M. Gould, and P. L. Toint, *Trust-Region Methods*, MPS-SIAM Series on Optimization, SIAM, Philadelphia, 2000.

[9] J. E. Dennis, *Surrogate Modelling and Space Mapping for Engineering Optimization. A summary of the Danish Technical University November 2000 Workshop*, Tech. Rep. TR00–35, Department of Computational and Applied Mathematics, Rice University, 2000.

[10] J. E. Dennis, S.-B. B. Li, and R. A. Tapia, *A unified approach to global convergence of trust region methods for nonsmooth optimization*, Math. Programming, 68 (1995), pp. 319–346.

[11] S. Leary, A. Bhaskar, and A. Keane, *A constraint mapping approach to the structural optimization of an expensive model using surrogates*, in Surrogate Modelling and Space Mapping for Engineering Optimization, H. B. Nielsen, ed., DK-2800, Lyngby – Denmark, 2000, Department of Mathematical Modelling, Technical University of Denmark.

[12] J. J. Moré, *Recent developments in algorithms and software for trust regions methods*, in Mathematical programming. The state of art, A. Bachem, M. Grotschel, and B. Korte, eds., Springer Verlag, New York, 1983, pp. 258–287.

[13] H. B. Nielsen, ed., *Surrogate Modelling and Space Mapping for Engineering Optimization*, DK-2800, Lyngby – Denmark, 2000, Department of Mathematical Modelling, Technical University of Denmark.

[14] M. J. D. Powell, *Convergence properties of a class of minimization algorithms*, in Nonlinear Programming 2, O. L. Mangasarian, R. R. Meyer, and S. M. Robinson, eds., Academic Press, New York, 1975, pp. 1–27.

[15] J. Søndergaard, *Non-Linear Optimization Using Space Mapping*, Master's thesis, Department of Mathematical Modelling, Technical University of Denmark, 1999.