A note on an interaction between penalization and discretization

Tomáš Roubíček

Institute of Information Theory and Automation
Czechoslovak Academy of Sciences
Pod vodárenskou věží 4, 182 08 Praha 8,
Czechoslovakia

The aim of this note is to investigate phenomena appearing when a state-constrained optimal control (or optimal shape design, etc.) problem governed by some differential equation is handled numerically. Then we are forced to approximate the problem on finite dimension spaces by some discretization method like finite diferences or finite elements, and simultaneously to cope with the state space constraints by some dual method – here we confine ourselves to the simplest one, namely to the penalty function method, but the augmented Lagrangean method will behave essentially by the same manner. By author's knowledge, an interaction between discretization and penalization has not been studied yet, except some investigations in soviet literature collected in the book by F.P.Vasilev [3] which does not deal directly with the dual treatment of the state constraints, however. Though the matter is not too complicated, it is perhaps worth mentioning briefly here because, by author's experience, all possible events are not sometimes realized well by those who use discretization with penalization simultaneuously.

As most of the phenomena appear already on an abstract level, we may begin with the following abstract optimization problem

$$(P) \qquad \begin{cases} \text{minimize } f(u) \text{ on } u \in U \\ \text{subject to } g(u) \in C \end{cases}$$

where $f : U \to \bar{R}$ is a cost function, $g : U \to Y$ a state operator, U a set of admissible controls, Y a space of states, and $C \subset Y$ a set of admissible states. From now on, we shall suppose controlability of (P), that is $g(U) \cap C \neq \emptyset$. After penalization (with a parameter $\varepsilon > 0$) and discretization (with a parameter $h > 0$) we get a family of unconstrained optimization problems, each of which can be written in an abstract form:

$$(P_\varepsilon^h) \qquad \text{minimize } f_\varepsilon^h(u) = f^h + \varepsilon^{-1} p(g^h(u)) \text{ on } u \in U^h,$$

where $f^h : U^h \to \bar{R}$, $g^h : U^h \to Y$ are an approximate cost function and state operator, respectively, $U^h \subset U$ is an internal approximation of the set of admissible controls, and $p : Y \to \bar{R}$ is an appropriate penalty function; for simplicity we suppose that p is so easy to be evaluated that it need not be approximated by some p^h, which is often case, indeed.

To simplify the problem as much as possible, we will assume the following, quite strong assumptions:

(1) U is compact, its topology being denoted by τ,

(2) Y is a metric space, ρ its metric, C its closed subset,

(3) f, g are continuous, $f > -\infty$,

(4) p is continuous, $p(C) = 0$, $p(Y \backslash C) > 0$,

(5) U^h is closed in U, f^h, g^h are continuous in the (relativized) topology τ,

(6) $U^{h_1} \subset U^{h_2}$ for $h_1 \geq h_2 > 0$, $\bigcup_{h>0} U^h$ is dense in U, and

(7) $f^h \to f$, $g^h \to g$ uniformly in the sense:

$$\forall \varepsilon > 0 \ \exists h_0 > 0 \ \forall 0 < h \leq h_0 \ \forall u \in U^h : \ |f^h(u) - f(u)| \leq \varepsilon, \ \rho(g^h(u), g(u)) \leq \varepsilon.$$

Note that the assumptions (1)–(5) obviously guarantee existence of a minimizer both of (P_ε^h) and of (P), which is, however, not too much important because all phenomena studied below appear also in more general setting of the problem where compactness (1) need not be used, cf. [2].

Though the assumptions (1)–(7) may seem quite powerful on a first look, they cannot ensure the convergence of the minima of (P_ε^h) to the minimum of (P) (and *a fortiori* the convergence of minimizers, either) if only $\varepsilon, h \searrow 0$, as shown by the following example.

Example 1. Consider a very simple situation: $U = [-1, 1]$, $Y = R$, $f(u) = g(u) = u$, $C = \{+1, -1\}$, $U^h = [-1 + h, 1]$, $f^h \equiv f$, $g^h \equiv g$ on U^h, and $p = 1 - |u|$. All the assumptions (1)–(7) are fulfilled trivially, and clearly $\min(P) = -1$, and $\operatorname{Argmin}(P) = \{-1\}$. On the other hand, it is easy to compute that, for $\varepsilon < h/2$, $\min(P_\varepsilon^h) = 1$ and $\operatorname{Armin}(P_\varepsilon^h) = \{1\}$, which shows that neither the minimum, nor the minimizer of (P_ε^h) converge respectively to the minimum or the only minimizer of (P) when $\varepsilon, h \searrow 0$ and $\varepsilon < h/2$, that means when ε tends to zero too quickly in comparison with h.

What the assumptions (1)–(7) can guarantee is only the existence of a stability criterion "$h \leq \eta(\varepsilon)$" under which the convergence is ensured:

Theorem 1. *Under the assumptions (1)–(7) there exists $\eta : R^+ \to R^+$ such that*

(8) $$\lim_{\varepsilon, h \searrow 0, h \leq \eta(\varepsilon)} \min(P_\varepsilon^h) = \min(P) \ , \quad \textit{and}$$

(9) $$\limsup_{\varepsilon, h \searrow 0, \ h \leq \eta(\varepsilon)} \operatorname{Argmin}(P_\varepsilon^h) \subset \operatorname{Argmin}(P) \ ,$$

where "limsup" has the usual meaning, i.e. it contains all τ-cluster points of all chosen subnets.

The proof is, in fact, contained as a part of the proof of Theorem 4.3 in [2] and will be thus omitted here (however, Theorem 4.3 there itself is stated in terms of so-called minimizing filters instead of the sets of minimizers, not supposing any compactness).

It should be emphasize that Theorem 1 has a little practical usage because it does not say anything about the stability criterion "$h \le \eta(\varepsilon)$" except its mere existence. The following Theorems 2 and 3 provide us with more information, the former one dealing even with the extreme situation when no stability criterion is needed:

Theorem 2. *If (1)-(7) are fulfilled and moreover*

(10) $\quad C = \mathrm{cl}_Y \mathrm{int}_Y C$ *and* $g(U) \cap \mathrm{int}_Y C \neq \emptyset$ *and* ,

(11) $\quad \forall$ *uniform neighbourhood* B *of* $g^{-1}(\mathrm{int}_Y C)$ $\exists \delta > 0 :\ g^{-1}(C_\delta) \subset B$

where "cl_Y" *and* "int_Y" *stand respectively for the closure and the interior in* Y *and* C_δ *for* δ*-neighbourhood of* C *in the metric* ρ . *Then (8) and (9) hold with* $\eta \equiv 1$, *that means the convergence is unconditional.*

Again, the proof is essentially contained in [2] as a part of the proof of Theorem 4.4 and will be omitted here.

Remark 1. The hypothesis (10) is particularly satisfied if Y is a linear metric space, C is convex with nonempty interior and $g(U) \cap \mathrm{int}_Y C \neq \emptyset$; then we come to the standard Slater constraint qualification. As for (11), it is particularly satisfied if g^{-1} is uniformly continuous, possibly in the Haussdorff sense provided g^{-1} is multivalued.

Unfortunately, (11) is typically not fulfilled in optimization problems for systems governed by differential equations where usually Y is a normed linear space with a norm strictly coarser than the corresponding energetic norm; e.g. $Y = L^2(.)$ while the energetic space is some Sobolev space $H^k(.)$ with $k > 0$. In such case we have to perform the analysis more in detail, introducing also the auxiliary penalized problem without any discretization:

$$(P_\varepsilon) \qquad \text{minimize } f_\varepsilon(u) = f + \varepsilon^{-1} p(g(u)) \text{ on } u \in U.$$

Theorem 3. *Let (1)-(7) be fulfilled and the following discretization error is known:*

$$(12) \qquad \forall h \le h_0 : |\min(P_\varepsilon^h) - \min(P_\varepsilon)| \ \le E(\varepsilon, h_0).$$

Then every $\eta : R^+ \to R^+$ *such that* $\lim_{\varepsilon \searrow 0} E(\varepsilon, \eta(\varepsilon)) = 0$ *will guarantee (8) and (9).*

The proof of (8) follows from the fact that $\min(P_\varepsilon)$ converges for $\varepsilon \searrow 0$ to $\min(P)$ and from the obvious estimate:

$$|\min(P_\varepsilon^h) - \min(P)| \leq E(\varepsilon, \eta(\varepsilon)) + |\min(P_\varepsilon) - \min(P)|$$

provided $h \leq \eta(\varepsilon)$. As soon as (8) is proved, (9) is ensured simply by standard compactness arguments.

Example 2. We outline a rather model situation dealing with an optimal distributed-control problem for a nonlinear elliptic equation to illustrate how Theorem 3 can be applied. Let Ω be a bounded, polyhedral domain in R^n, $\partial\Omega$ its boundary, $U = \{u \in L^\infty(\Omega); -1 \leq u(x) \leq 1 \text{ for a.a. } x \in \Omega\}$, τ is the topology induced on U from $H^1(\Omega)^*$ (which obviously guarantees (1), "$*$" stands for the topological dual space), $Y = L^2(\Omega)$, and $g(u) = y \in H^1(\Omega)$ is the weak solution of the nonlinear boundary value problem:

$$\nabla(a(|\nabla y|)\nabla y) = u \text{ on } \Omega, \tag{13}$$

$$a(|\nabla y|)\frac{\partial y}{\partial \nu} + y = 0 \text{ on } \partial\Omega \tag{14}$$

with some nonlinearity $a(.)$ such that the function $\xi \mapsto a(\xi)\xi$ is uniformly increasing with a linear growth, ν is the outward unit normal to $\partial\Omega$. In other words, $g(u) = y$ should fulfil the integral identity:

$$\int_\Omega a(|\nabla y|)\nabla y\ \nabla v\ dx + \int_{\partial\Omega} yv\ dS = \int_\Omega uv\ dx \qquad \forall v \in H^1(\Omega).$$

Furthermore, let

$$f(u) = \int_{\partial\Omega} g(u)\ dS, \tag{15}$$

C be a closed subset of $L^2(\Omega)$, and $p(y) = \inf_{\tilde{y}\in C} \|y - \tilde{y}\|^2_{L^2(\Omega)}$. In view of the cost function (15) together with the boundary conditions (14), we can see that, speaking in terms of a heat-transfer interpretation, we are to choose heat sources distributed around Ω in order to minimize the heat flux through the boundary $\partial\Omega$ representing a lost of energy outside the domain Ω, subject to some constraints imposed on the heat sources and on the temperature distribution. Hence our model problem has a quite reasonable practical interpretation.

We discretize the problem (13)–(14) by a standard manner, using the finite element method (any numerical integration is not needed here). Let $\{\Upsilon_h\}_{h>0}$ be a regular family of triangulations of Ω, $U^h = \{u \in U; u \text{ is piecewise constant on } \Upsilon_h\}$, $V^h = \{y \in H^1(\Omega); y \text{ is piecewise linear on } \Upsilon_h\}$, $f^h \equiv f$ on U^h, and $g^h(u) \in V^h$ is the unique solution of the integral identity:

$$\int_\Omega a(|\nabla g^h(u)|)\nabla g^h(u)\ \nabla v\ dx + \int_{\partial\Omega} g^h(u)v\ dS = \int_\Omega uv\ dx \qquad \forall v \in V^h.$$

To derive the estimate of the type (12) we employ the following facts:

i) f, g, and p are Lipschitz continuous on their respective sets of admissible arguments.

ii) The rate-of-error estimates which are uniform with respect to the control are known:

$$\|g(u) - g^h(u)\|_{L^2(\Omega)} \le c\ h^\alpha \qquad \forall u \in U^h, \text{and}$$

$$|f(u) - f^h(u)| \le c\ h^\beta \qquad \forall u \in U^h.$$

If the regularity $g(u) \in H^2(\Omega)$ is valid, by [3] it is well known that $\alpha = 1$, and in the linear case (i.e. $a \equiv const. > 0$) even $\alpha = 2$. As for β, its expected value is $\frac{1}{2}$ (or $\frac{3}{2}$ in the linear case), but we shall see (cf. Remark 2) that its concrete value has no influence on mere convergence (8) and (9).

iii) The uniform approximation error estimate is known:

$$\inf_{u^h \in U^h} \|u - u^h\|_{H^1(\Omega)^*} \le c\ h^\gamma \qquad \forall u \in U. \tag{16}$$

Let us outline the proof of (16). For $u \in L^\infty(\Omega)$ denote by $u^h \in U^h$ the function defined by $\int_\Delta u^h\ dx\ = \int_\Delta u\ dx$ for every simplex $\Delta \in \Upsilon_h$. It is easy to verify that $\|v - v^h\|_{L^2(\Omega)} \le const.h\|v\|_{H^1(\Omega)}$ for every $v \in H^1(\Omega)$. Realizing that $\langle u - u^h, v^h\rangle = 0$ because evidently $\int_\Delta (u - u^h)\ dx\ = 0$ and v^h is constant on Δ for every $\Delta \in \Upsilon_h$, we obtain the estimate $|\langle u - u^h, v\rangle| = |\langle u - u^h, v - v^h\rangle| \le\ const.(\|u\|_{L^2(\Omega)} + \|u^h\|_{L^2(\Omega)})\ h\ \|v\|_{H^1(\Omega)}$. Taking into account that $u, u^h \in U$ and the definition of the standard dual norm, we can see that $\|u - u^h\|_{H^1(\Omega)^*} \le 2\ const.\sqrt{\text{meas}\Omega}\ h$, and put $\gamma = 1$ in (16).

Now we will employ the facts i)–iii) to derive the estimate (12). Taking some $u \in \text{Argmin}(P_\varepsilon)$, by (16) we can find some $u^h \in U^h$ with $\|u - u^h\|_{H^1(\Omega)^*} \le (c + 1)h^\gamma$. By i) we can then see that $f_\varepsilon(u^h) \le \min(P_\varepsilon) + (c+1)h^\gamma(L + \frac{L^2}{\varepsilon})$, where L stands for the common Lipschitz constant of f, g, and p. By ii) we come to

$$\min(P_\varepsilon^h) \le f_\varepsilon^h(u^h) \le \min(P_\varepsilon) + (c+1)(L + \frac{L^2}{\varepsilon})h^\gamma + c\ h^\beta + \frac{cL}{\varepsilon}h^\alpha. \tag{17}$$

Conversely, let us take some $u \in \text{Argmin}(P_\varepsilon^h)$. By ii) we get immediately

$$\min(P_\varepsilon) \le f_\varepsilon(u) \le \min(P_\varepsilon^h) + c\ h^\beta + \frac{cL}{\varepsilon}h^\alpha. \tag{18}$$

Joining (17) and (18), we come to the error estimate (12) with

$$E(\varepsilon, h) = Const.(h^\gamma + h^\beta + \frac{1}{\varepsilon}(h^\gamma + h^\alpha)).$$

Then by Theorem 3, for the stability criterion function η we can take arbitrary function

$$\eta(\varepsilon) = \varepsilon^q \qquad \text{with } q > \max(\frac{1}{\alpha}, \frac{1}{\gamma}).$$

Remark 2. Note that β has no influence to a freedom of the choice of η, which is due to the fact that we investigated only mere convergence of the problem (P_ϵ^h) to (P), not any rate of convergence. Note also that the optimal case is $\alpha = \gamma$, particularly the case $\alpha = 2$ has here the same efficiency as $\alpha = 1$.

Remark 3. It is known that without the compactness hypothesis (1), the penalized problem (P_ϵ) does not generally approximate the original problem (P), but some extended problem (roughly speaking, a "relaxed control" problem). In such case, our considerations are also well fitted to approach relaxed controls by solving numerically the problems (P_ϵ^h); cf. [2] for a general treatment of this idea.

References

[1] P.Ciarlet,*The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[2] T.Roubíček, Constrained optimization: a general tolerance approach. *Apl. Mat.* (to appear).

[3] F.P.Vasiljev, *Methods of Solving Extremal Problems* (in Russian), Nauka, Moscow, 1981.