OPEN GEOMECHANICS

Frictional Plasticity in a Convex Analytical Setting

Guy T. Houlsby *

* *Department of Engineering Science, University of Oxford,*

Published
12th September 2019
https://doi.org/10.5802/ogeo.2

Edited by
Claudio Tamagnini
Dipartimento di Ingegneria
Civile e Ambientale
University of Perugia
Italy

Reviewed by
Denis Caillerie
Laboratoire 3SR
Université Grenoble Alpes
France

Francesco Solombrino
Dipartimento di Matematica e
Applicazioni "Renato Caccioppoli"
University of Naples "Federico II"
Italy

Correspondence
Guy T. Houlsby
Department of Engineering Science,
University of Oxford,
Parks Road, Oxford, OX1 3PJ
United Kingdom
guy.houlsby@eng.ox.ac.uk

Abstract.
A very simple frictional plasticity model for a granular material is presented, including the effects of dilation. The novelty lies in the fact that this is described within the hyperplasticity framework, expressed using the terminology of convex analysis. This allows a consistent mathematical treatment of the dilation constraint. The Fenchel Dual is used to link the force and flow potentials. The resulting model accommodates non-associated flow within a rigorous mathematical framework that ensures compliance with the Laws of Thermodynamics.

Keywords. Convex analysis, friction, dilation, plasticity

1. Introduction

The language of engineering calculations is of course mathematics, and it is the Author's experience that when the appropriate branch of mathematics is applied to a particular class of engineering problems, it brings greater clarity to those problems and can lead to new insights. Over the last 50 years or more, plasticity theory has proven to be an extremely useful mathematical framework within which to describe the constitutive behaviour of soils. However, plasticity theory can itself be expressed using a number of different mathematical techniques. In some quarters the terminology of "Convex Analysis" is used to express plasticity theory, although this has yet to become widespread, and the techniques of convex analysis are unfamiliar to most practising geotechnical engineers. It is the Author's view, however, that the terminology of convex analysis is the mathematical language that is most suitable for expressing plasticity theory, and specifically the frictional plasticity theories that need to be employed to describe soils. The purpose of this paper is therefore to express a very simple frictional plasticity theory using the language of convex analysis. The model itself is not novel; it is the language that is used to express it that is new. Convex analytical terminology provides a concise and rigorous language to describe a frictional plasticity model. These advantages mean that this approach serves as a useful starting point for development of more sophisticated models, or for establishing (for instance) theorems about the behaviour of frictional materials.

Convex analytical terminology that may be unfamiliar to most readers is introduced as it is used, at a level of detail that should be sufficient for those familiar with differential calculus as applied in geotechnical engineering. For more rigorous proofs of the definitions and results, see a standard mathematical textbook such as Rockafellar [1970].

It is widely recognised that coarse-grained granular materials should be treated as "frictional" materials. The very simplest model of a sand would be a material that is elastic (or even rigid) within a certain range of stresses, and then flows plastically when the stresses satisfy a simple frictional criterion, usually specified by an angle of friction. For all but the loosest of sands, plastic flow is accompanied by dilation.

While countless plasticity theories exist that describe the above behaviour in broad terms, a fundamental problem lies at the heart of most conventional plasticity theories that embrace friction. For most non–frictional materials (such as ductile metals) Drucker's quasi-thermodynamic "stability" criterion is adopted [Drucker, 1951, 1959], and this leads directly to "associated flow": that is to say the yield surface and plastic potential are identical, *i.e.* the plastic strain increment vector is normal to the yield surface when plotted in terms of appropriate work-conjugate variables. Frictional soils manifestly disobey this condition, as the angle of dilation is invariably much less than the angle of friction. So associated flow, and with it Drucker's criterion, must be rejected for frictional materials. Unfortunately, this leads to a number of undesirable consequences, for instance the loss of proof of the uniqueness of incremental response. Furthermore, it is not in general possible to show that frictional plasticity theories are consistent with the Laws of Thermodynamics.

There is therefore a strong motivation to seek a way of formulating frictional plasticity theory in a more rigorous way, at the very least demonstrating that it is consistent with thermodynamics. Hyperplasticity [Collins and Houlsby, 1997, Houlsby, 1981, Houlsby and Puzrin, 2000, 2006], which implements the orthogonality principle of Ziegler Ziegler [1977], offers such a possibility. The use of Ziegler's principle ensures that the models are consistent with thermodynamics and (unlike the case of Drucker's criterion) it proves possible to formulate realistic frictional models that satisfy Ziegler's principle. In brief Ziegler's orthogonality principle is a stronger statement than the Second Law of Thermodynamics, so that materials which obey Ziegler's principle automatically satisfy the Second Law.

Use of Ziegler's orthogonality leads to the important result that the entire constitutive response of a material can be specified through knowledge of just two scalar functions – one that represents the stored energy, and one that specifies how energy is dissipated. The constitutive behaviour is then obtained by differentiation of these functions. There is considerable freedom, however, in the choice of the functions, as a series of Legendre Transforms can be used to move between different functions, interchanging the role of independent and dependent variables.

The Legendre Transform is defined as follows. We start from a function $X = X(x_i)$ which acts as a potential, such that $y_i = \partial X/\partial x_i = X'(x_i)$, which establishes a correspondence between values of x_i and the conjugate values of y_i. The purpose of the transform is to interchange the roles of x_i and y_i. To do so it is necessary to establish the form of a function $Y = Y(y_i)$, which is achieved by first defining

$$Y = \langle y_i, x_i \rangle - X$$

where the notation $\langle y_i, x_i \rangle$ indicates the inner product between the variables; for vectors in n–dimensional space this would simply be

$$\langle y_i, x_i \rangle = \sum_{i=1}^{n} y_i x_i.$$

We then establish Y as a function of y_i by the following procedure. For sufficiently smooth functions, the relationship $y_i = X'(x_i)$ will be invertible, *i.e.* it can be written $x_i = (X')^{-1}(y_i)$ such that $(X')^{-1}\big(X'(x_i)\big) = x_i$. We therefore write

$$Y = \langle y_i, x_i \rangle - X = \langle y_i, (X')^{-1}(y_i) \rangle - X\big((X')^{-1}(y_i)\big) = Y(y_i).$$

It then follows that the conjugacy between points x_i and y_i can equivalently be written as $x_i = \partial Y/\partial y_i$. Indeed (once the functional form of $Y = Y(y_i)$ is known) the three statements relating conjugate points

$$y_i = \frac{\partial X}{\partial x_i} \qquad x_i = \frac{\partial Y}{\partial y_i} \qquad X + Y = \langle y_i, x_i \rangle$$

are mathematically equivalent [Moreau, 1974].

Thus, for example, a hyperelastic material (with no dissipation) can be specified either by knowledge of the "strain energy", a function of the strains $E = E(\epsilon_{ij})$ such that $\sigma_{ij} =$

$\partial E/\partial \epsilon_{ij}$, or the "complementary energy", a function of the stresses $C = C(\sigma_{ij})$ such that $\epsilon_{ij} = \partial C/\partial \sigma_{ij}$. The two are related by a Legendre Transform $C + E = \sigma_{ij}\epsilon_{ij}$, and each can be derived from the other. In the context of thermodynamics, the strain energy can be identified with the Helmholtz free energy $f(\epsilon_{ij}, \theta)$ at a constant temperature θ_0, $E(\epsilon_{ij}) = f(\epsilon_{ij}, \theta_0)$, and the complementary energy can be similarly identified with the (negative) Gibbs free energy at constant temperature, $C(\sigma_{ij}) = -g(\sigma_{ij}, \theta_0)$.

In hyperplasticity (as opposed to hyperelasticity) theory, two scalar potentials are employed, an energy function and the dissipation function. A Legendre Transform of the dissipation function leads to definition of the yield function. However, for rate–independent materials, the dissipation is a homogeneous first order function of the plastic strain rates, and as a result, derivation of its Legendre Transform involves a degenerate special case for which *ad hoc* procedures are necessary (see Appendix C of Houlsby and Puzrin [2006]). The special case is better treated within the terminology of convex analysis, in which the Legendre Transform is generalised to the Fenchel Dual.

A further complication arises in the case of frictional plasticity, in that the dilation is most straightforwardly imposed as a "constraint" on the plastic strain rates [Houlsby, 1992]. When implemented in a hyperplasticity formulation in its simplest form, constraints are implemented by the method of Lagrangian multipliers, resulting in an augmented dissipation function that can then be differentiated to give the (generalised) stresses. Any constraint on the plastic strain rates must of course be consistent with the requirement of dissipation of energy. In principle the yield surface can be obtained by applying a Legendre Transform to the augmented dissipation function, but again this can be treated more consistently using convex analysis.

Convex analysis [Rockafellar, 1970] is a technique that is very well developed in the area of optimisation, as well as in several other fields. A small number of researchers have adopted convex analytical terminology for plasticity theory (see *e.g.* Han and Reddy [1999], Maugin [1992]), but this approach has not yet become routine. However, convex analysis terminology is ideally suited for plasticity theory, and specifically for hyperplasticity. It is the natural language to use to express plasticity theories in a consistent and rigorous way, and a brief introduction in this context is given in Appendix D of Houlsby and Puzrin [2006]. It will be seen below that where constraints and Lagrangian multipliers would be introduced in a more conventional approach, convex analysis makes use of "Indicator Functions" (instead of constraints) and their "Normal Cones", which can be expressed through so–called Karush–Kuhn–Tucker conditions. There is of course a close equivalence (indeed isomorphism) between the two approaches, but the Author believes that the convex analytical terminology is the more versatile, and in particular allows the employment of the Fenchel Dual, which is a generalisation of the Legendre Transform.

The purpose of this paper is therefore to set out the simplest of frictional–dilatant models in the terminology of convex analysis, showing how this leads to a simple and self–consistent mathematical framework, within which a thermodynamically rigorous soil model can be expressed. The simplicity of the model is emphasised: this is merely the starting point for more realistic (and necessarily more complex) models.

2. Analysis

We start by setting out hyperplasticity theory without recourse to convex analysis. We choose just one version of hyperplasticity in which we focus on the Gibbs free energy g and the "force potential" z. The force potential is identical to the dissipation d for the special case of rate–independence which we consider here, but we use this terminology for consistency with the more general rate–dependent case. It is related to the dissipation d through the relationship

$$d = \frac{\partial z}{\partial \dot{\alpha}} \dot{\alpha}$$

(see Houlsby and Puzrin [2006], Ch. 11). Other formulations are possible using different potentials, which are Legendre Transforms of these. The Gibbs free energy $g = g(\sigma, \alpha)$ is a function of the stresses σ and some kinematic "internal variables" α, which in the following will be seen to play exactly the same role as plastic strains. The force potential $z = z(\sigma, \alpha, \dot{\alpha})$ is a function of the stresses, internal variables and the rates of the internal variables. Since the force potential for a rate–independent material is identical to the dissipation, which must be non–negative, we require the functional form of z to satisfy $z(\sigma, \alpha, \dot{\alpha}) \geqslant 0$. For the more general rate–dependent case we require $d = (\partial z/\partial \dot{\alpha})\dot{\alpha} \geqslant 0$.

It can then be shown from the First and Second Laws of Thermodynamics (see for instance Houlsby and Puzrin [2006]) that:

$$\epsilon = -\frac{\partial g}{\partial \sigma} \tag{1}$$

$$0 = \left(\frac{\partial g}{\partial \alpha} + \frac{\partial z}{\partial \dot{\alpha}}\right)\dot{\alpha} \tag{2}$$

One way of expressing Ziegler's orthogonality principle is to replace (2) by the stronger statement:

$$0 = \frac{\partial g}{\partial \alpha} + \frac{\partial z}{\partial \dot{\alpha}} \tag{3}$$

Any material which obeys (3) therefore also obeys (2) and thus satisfies the Laws of Thermodynamics. Ziegler's orthogonality principle can be expressed in a variety of different ways, but it can be understood as the concept that whilst the Second Law requires the dissipation to be non–negative, Ziegler's principle states that, subject to any relevant constraints, dissipation is maximal.

In practice it is convenient to rewrite (3) by defining a pair of new variables, χ and $\bar{\chi}$, both called the "generalised stress", and defined by:

$$\bar{\chi} = -\frac{\partial g}{\partial \alpha} \tag{4}$$

$$\chi = \frac{\partial z}{\partial \dot{\alpha}} \tag{5}$$

and impose the side condition $\chi = \bar{\chi}$, which is in effect Ziegler's orthogonality condition. Although the two new variables are always equal, it is necessary to treat them separately for some formal mathematical purposes.

The above approach is satisfactory when g and z are differentiable (C^1 continuous). However, when they are not, and in the special case when z is homogeneous first order in the rates, it is convenient to use the terminology of convex analysis and introduce the subdifferential (essentially a generalisation of the derivative for the case of convex functions that are not C^1 continuous).

At this stage it is necessary to introduce some of the essential language of convex analysis. A real–valued function $y = f(x)$, that may take values on the extended line $[-\infty, +\infty]$, is *convex* if

$$f\big((1-\lambda)x_1 + \lambda x_2\big) \leqslant (1-\lambda) f(x_1) + \lambda f(x_2), \quad \forall\, 0 \leqslant \lambda \leqslant 1$$

The notation $\forall\, 0 \leqslant \lambda \leqslant 1$ means "for all λ between 0 and 1". This concept is illustrated for a function of one variable in Fig. 1, in which the inequality states that point Q lies above (or at least not below) point P. This corresponds to the conventional notion of convexity.

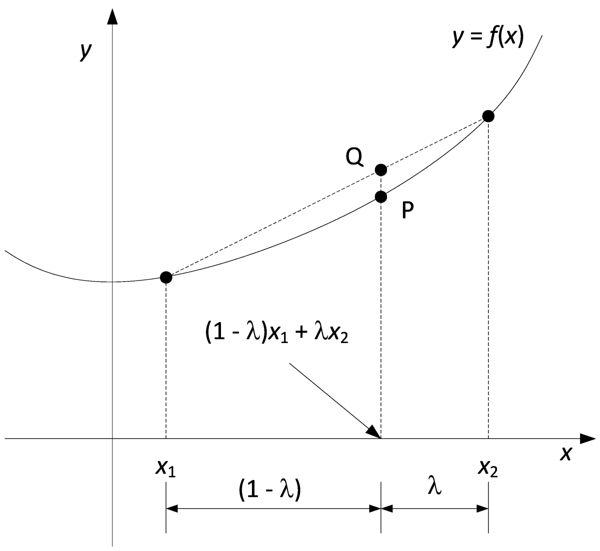

Figure 1. A convex function of one variable.

The derivative $f'(x)$ is familiar as the slope of a tangent to a curve $y = f(x)$. For convex functions that are continuous but non–smooth (*i.e.* C^0 but not C^1 continuous) the concept can be extended to the concept of a *subdifferential*. First of all, for a function of one variable, a *subgradient* is defined as the slope x^* of any line that touches a curve at x_1 but does not pass above it, *i.e.* such that

$$x^*(x_2 - x_1) \leqslant f(x_2) - f(x_1), \quad \forall x_2$$

see Fig. 2.

The subdifferential $\partial f(x)$ is then defined as the set of all subgradients at a given point on the curve. For the function of one variable illustrated in Fig. 3 the definition is

$$\partial f(x_1) = \Big\{x^* \;\Big|\; x^*(x_2 - x_1) \leqslant f(x_2) - f(x_1),\ \forall x_2\Big\},$$

where the notation $\{a \,|\, b\}$ means "the set of all values of a that satisfy condition b". For the more general case this is

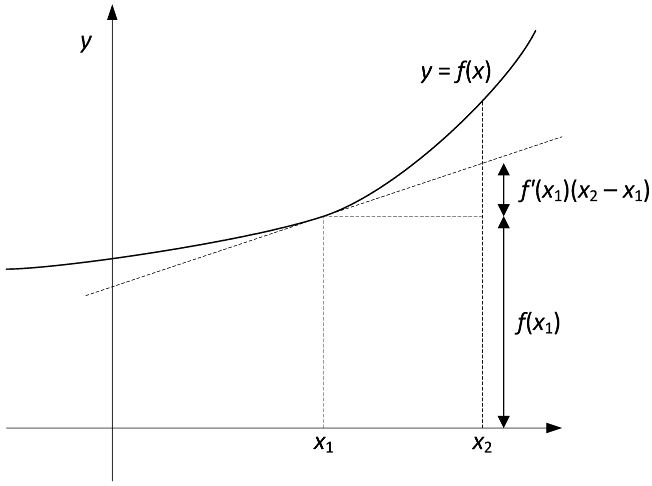

Figure 2. Subgradient of a function of one variable.

written

$$\partial f(x_1) = \Big\{x^* \in V' \;\Big|\; \langle x^*, (x_2 - x_1)\rangle \leqslant f(x_2) - f(x_1),\ \forall x_2\Big\},$$

where x is a member of the vector space V and x^* a member of the dual space V' under the inner product $\langle x^*, x\rangle$. Noting that the subdifferential is a set that consists of all possible values of the subgradient at a particular point; at any point x_1 where the function is smooth (C^1 continuous) the set is a "singleton" that contains just one value which is identical to the conventional derivative $\partial f(x_1) = \{f'(x_1)\}$.

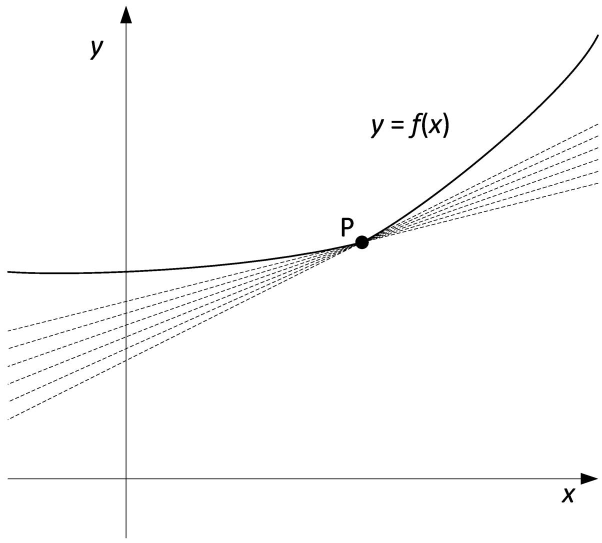

Figure 3. Subdifferential of a function of one variable.

Where $f(x, y)$ is a function of more than one variable we use the notation $\partial_x f$ to denote the subdifferential of f with respect to x.

In convex analytical notation equations (1), (4) and (5) now become:

$$\epsilon \in \partial_\sigma(-g) \tag{6}$$

$$\bar{\chi} \in -\partial_\alpha g \tag{7}$$

$$\chi \in \partial_{\dot{\alpha}} z \tag{8}$$

together with $\chi = \bar{\chi}$ as before. The notation $x \in S$ means "x is a member of the set S". The positions of the minus signs in (6) and (7) are important: in general $-g$ is convex in σ and g is convex in α.

2.1. Frictional plasticity

We now express a very simple frictional plasticity theory using the above approach. We shall use the Cambridge triaxial effective stress parameters (p', q), but to emphasise the connection with the notation already introduced we shall call these (σ_p, σ_q). The corresponding strains are (ϵ_p, ϵ_q), the internal variables (plastic strains) are (α_p, α_q) and the generalised stresses (χ_p, χ_q), $(\bar{\chi}_p, \bar{\chi}_q)$. Note that we use the compressive positive convention usual in geomechanics. The stress σ_p is equal to the mean effective stress (*i.e.* $\sigma_{ii}/3$ in tensorial subscript notation) and σ_q is the deviatoric stress, for the special case of triaxial stress states considered here. More generally it is equal to

$$\sigma_q = \sqrt{\frac{3}{2} s_{ij} s_{ij}} \qquad s_{ij} = \sigma_{ij} - \frac{1}{3}\sigma_{kk}\delta_{ij}$$

We specify the Gibbs free energy:

$$g = -\frac{\sigma_p^2}{2K} - \frac{\sigma_q^2}{6G} - \sigma_p \alpha_p - \sigma_q \alpha_q \tag{9}$$

We could alternatively have specified the Helmholtz free energy, the Legendre Transform of the Gibbs free energy, as

$$f = \frac{K}{2}\left(\epsilon_p - \alpha_p\right)^2 + \frac{3G}{2}\left(\epsilon_q - \alpha_q\right)^2$$

but we do not pursue that alternative here.

As g is in fact C^1 continuous, the subdifferential at any point would be a singleton, and equations (6) and (7) reduce to (1) and (4) so we can write:

$$\epsilon_p = -\frac{\partial g}{\partial \sigma_p} = \frac{\sigma_p}{K} + \alpha_p \tag{10}$$

$$\epsilon_q = -\frac{\partial g}{\partial \sigma_q} = \frac{\sigma_q}{3G} + \alpha_q \tag{11}$$

$$\bar{\chi}_p = -\frac{\partial g}{\partial \alpha_p} = \sigma_p \tag{12}$$

$$\bar{\chi}_q = -\frac{\partial g}{\partial \alpha_q} = \sigma_q \tag{13}$$

We can therefore deduce that for this particular model:

a) the internal variables (α_p, α_q) play the role of plastic strains, as they are simply additive terms to the elastic strains $(\sigma_p/K, \sigma_q/3G)$, and
b) the generalised stresses (χ_p, χ_q), $(\bar{\chi}_p, \bar{\chi}_q)$ are equal to the true stresses (σ_p, σ_q). This will be true for many models, hence the terminology "generalised stress", although for some formal mathematical purposes σ, χ and $\bar{\chi}$ need to be treated separately. In some models that employ a "back–stress" the generalised stress differs from the true stress, but such models are not pursued here. Kinematic hardening models that employ a back–stress involve additional energy terms in the Gibbs free energy (sometimes termed "frozen energy").

We now turn to the force potential, which we must treat using convex analysis as it is not C^1 continuous. For the frictional model we write the force potential as:

$$z = M\sigma_p \left|\dot{\alpha}_p\right| + I_{[-\infty;0]}\left(\dot{\alpha}_p + N\left|\dot{\alpha}_q\right|\right) \tag{14}$$

The term $M\sigma_p\left|\dot{\alpha}_p\right|$ simply represents a frictional dissipation – a dissipation proportional to the normal stress and the plastic shear strain rate. The notation $I_C(x)$ describes the Indicator Function[(1)] (also sometimes called the Characteristic Function) of the convex set C. The indicator function is defined by

$$I_C(x) = \begin{cases} 0 & x \in C; \\ +\infty & x \notin C. \end{cases}$$

It is therefore a rather curious function which takes the value zero if x is a member of the set C and is infinite otherwise. The use of this indicator function effectively imposes the constraint that its argument $\dot{\alpha}_p + N\left|\dot{\alpha}_q\right|$ is a member of the set with range $[-\infty;0]$, in other words it is non–positive. Effectively this is the dilation constraint, so that plastic shearing is always accompanied by volumetric expansion, in proportion to the shear strain as specified by the factor N, as is commonly observed for dense granular materials. Note, however, the important feature that a unilateral constraint $\dot{\alpha}_p + N\left|\dot{\alpha}_q\right| \leqslant 0$ is imposed rather than an equality constraint $\dot{\alpha}_p + N\left|\dot{\alpha}_q\right| = 0$. It later follows that this allows a consistent treatment of the behaviour at zero stress. It will be seen that the combination of the two terms in Equation (14) results in a response in which the apparent frictional strength is made up from two terms, a frictional dissipation expressed through the factor M and an additional component due to dilation and defined by the factor N. Such a concept has been familiar since the work of Taylor [1948].

The subdifferential of the indicator function $I_C(x)$ of a convex set C is a set called the Normal Cone $\mathcal{N}_C(x)$, and it can be shown from the general definition of the subdifferential that the normal cone is defined by

$$\mathcal{N}_C(x) = \partial I_C(x) = \left\{x^* \;\middle|\; \langle x^*, (y-x)\rangle \leqslant 0,\ \forall y \in C\right\}.$$

Furthermore, for the special case where $C = [-\infty;0]$, and therefore the indicator function is just a function of one variable, any member Λ of the Normal Cone at x, $\Lambda \in \mathcal{N}_{[-\infty;0]}(x)$, satisfies the conditions:

$$\begin{cases} \Lambda x = 0 \\ x \leqslant 0 \\ \Lambda \geqslant 0 \end{cases} \tag{15}$$

which are sometimes called the Karush–Kuhn–Tucker (KKT) conditions. Although these properly belong to a slightly different context, we use this terminology here. Exploiting Theorem 3.7 of Romano [1995][(2)], which is analogous to the

(1) Confusingly, in other mathematical contexts both the terms Indicator Function and Characteristic Function have been defined in other ways.

(2) The theorem is briefly stated as follows. Let $m : \mathbb{R} \cup \{+\infty\} \mapsto \mathbb{R} \cup \{+\infty\}$ be a monotone convex function with $m(+\infty) = +\infty$ and $k : X \mapsto \mathbb{R} \cup \{+\infty\}$ be a proper convex functional continuous at $x \in X$. Then, if x is not a minimum point of k and m is subdifferentiable at $k(x)$, setting $f = m \circ k$ results in $\partial f(x) = \partial m[k(x)]\partial k(x)$.

chain rule of conventional differential calculus, it follows that:

$$\begin{aligned}(\chi_p, \chi_q) &\in \partial z(\dot\alpha_p, \dot\alpha_q) \\ &= \Big\{ \big(\Lambda, M\sigma_p \Sigma_1 + \Lambda N \Sigma_2\big) \,\Big|\, \\ &\qquad \Lambda \in \mathscr{N}_{[-\infty;0]}\left(\dot\alpha_p + N\left|\dot\alpha_q\right|\right), \\ &\qquad \Sigma_1, \Sigma_2 \in S(\dot\alpha_q) \Big\}\end{aligned} \tag{16}$$

where we define a signum function:

$$S(x) = \begin{cases} \{-1\} & x < 0 \\ [-1;+1] & x = 0 \\ \{+1\} & x > 0. \end{cases}$$

It is necessary to use this (set valued) function rather than the conventionally defined signum function

$$\mathrm{sgn}(x) = \begin{cases} -1 & x < 0 \\ 0 & x = 0 \\ +1 & x > 0 \end{cases}$$

in order to describe the subdifferential correctly at $x = 0$. Note specifically that a member of $S(x)$ takes an indeterminate value at $x = 0$. It follows that although $\Sigma_1 \in S(\dot\alpha_q)$ and $\Sigma_2 \in S(\dot\alpha_q)$ there is no *a priori* reason to assume $\Sigma_1 = \Sigma_2$. However, note that for all values of $\dot\alpha_q$ and any positive multipliers A and B it is always possible to write

$$A\Sigma_1 + B\Sigma_2 \in (A+B)S(\dot\alpha_q),$$

This means we can simplify equation (16) to:

$$\begin{aligned}(\chi_p, \chi_q) \in \Big\{ \big(\Lambda, \left(M\sigma_p + \Lambda N\right)\Sigma\big) \,\Big|\, \\ \Lambda \in \mathscr{N}_{[-\infty;0]}\left(\dot\alpha_p + N\left|\dot\alpha_q\right|\right),\ \Sigma \in S(\dot\alpha_q) \Big\}\end{aligned} \tag{17}$$

Furthermore, Λ in equation (17) satisfies the KKT conditions:

$$\begin{cases} \Lambda\left(\dot\alpha_p + N\left|\dot\alpha_q\right|\right) = 0 \\ \dot\alpha_p + N\left|\dot\alpha_q\right| \leqslant 0 \\ \Lambda \geqslant 0 \end{cases} \tag{18}$$

2.2. Fenchel dual

Very importantly, it is possible to interchange the roles of independent and dependent variables by use of a Fenchel Dual, which is the convex analytical generalization of the Legendre Transform. The Fenchel Dual is defined as follows. We start from a convex function $f(x)$ which acts as a potential, such that $x^* \in \partial f(x)$, which establishes a correspondence between values of x and the conjugate values of x^*. The dual interchanges the roles of x and x^*. To do so it is necessary to establish the form of the dual function $f^*(x^*)$, which is defined as

$$f^*(x^*) = \sup_x \left(\langle x^*, x\rangle - f(x)\right)$$

where $\langle x^*, x\rangle$ denotes the inner product. It then follows that the conjugacy between points x and x^* can equivalently be written as $x \in \partial f^*(x^*)$. Indeed (once the functional form of $f^*(x^*)$ is known) the three statements relating conjugate points:

$$x^* \in \partial f(x) \qquad x \in \partial f^*(x^*) \qquad f(x) + f^*(x^*) = \langle x^*, x\rangle$$

are mathematically equivalent [Moreau, 1974]. The dual of the dual is the original function: $f^{**}(x) = f(x)$. It can readily be proven that for smooth (C^1 continuous) functions the Fenchel Dual reduces to the Legendre Transform.

Although the constitutive behaviour can be derived entirely from the force potential z , it is more convenient in some applications to use the flow potential w, which is closely related to the conventional yield surface. The two potentials are not independent, but related as Fenchel Duals. Applying the above definition we therefore seek the dual to z , defined by:

$$\begin{aligned} w(\chi_p, \chi_q) &= \sup_{(\dot\alpha_p, \dot\alpha_q)} \left\{\chi_p\dot\alpha_p + \chi_q\dot\alpha_q - z(\dot\alpha_p, \dot\alpha_q)\right\} \\ &= \sup_{(\dot\alpha_p, \dot\alpha_q)} \Big\{\chi_p\dot\alpha_p + \chi_q\dot\alpha_q \\ &\qquad - \left[M\sigma_p\left|\dot\alpha_q\right| + I_{[-\infty;0]}\left(\dot\alpha_p + N\left|\dot\alpha_q\right|\right)\right]\Big\} \end{aligned} \tag{19}$$

It can be shown (see Appendix) that:

$$w(\chi_p, \chi_q) = I_{[-\infty;0]}\left(\left|\chi_q\right| - N\chi_p - M\sigma_p\right) + I_{[-\infty;0]}\left(-\chi_p\right) \tag{20}$$

It follows, from the fact that w is the Fenchel Dual of z, that $\dot\alpha \in \partial w$ and

$$\begin{aligned}(\dot\alpha_p, \dot\alpha_q) &\in \partial w(\chi_p, \chi_q) \\ &= \left(-N, S(\chi_q)\right)\mathscr{N}_{[-\infty;0]}\left(\left|\chi_q\right| - N\chi_p - M\sigma_p\right) \\ &\quad + (-1, 0)\,\mathscr{N}_{[-\infty;0]}\left(-\chi_p\right) \\ &= \left\{\left(-\lambda_1 N - \lambda_2, \lambda_1 S(\chi_q)\right)\right\}\end{aligned} \tag{21}$$

where λ_1, λ_2 satisfy the KKT conditions:

$$\begin{cases} \lambda_1\left(\left|\chi_q\right| - N\chi_p - M\sigma_p\right) = 0 \\ \left|\chi_q\right| - N\chi_p - M\sigma_p \leqslant 0 \\ \lambda_1 \geqslant 0 \end{cases} \tag{22}$$

$$\begin{cases} \lambda_2\left(-\chi_p\right) = 0 \\ -\chi_p \leqslant 0 \\ \lambda_2 \geqslant 0 \end{cases} \tag{23}$$

It can immediately be seen that if $\chi_p > 0$ (in which case $\lambda_2 = 0$) the plastic strains satisfy the dilation constraint, and that w expresses both the yield function (in generalized stress space) and the plastic potential of conventional plasticity theory in a compact form. Note specifically that it is the partition between the terms $M\sigma_p$ and $N\chi_p$ in the first indicator function of w that defines the frictional and dilational components of strength. Although in this particular model σ_p and χ_p are always numerically equal, σ_p simply serves as a parameter in equation (20).

The relationships between the force potential z and the flow potential w are illustrated in Fig. 4. On the left the force potential z is illustrated in Fig. 4(a). The domain of z is shaded, and within this domain $z = M\sigma_p\left|\dot\alpha_q\right|$. Outside the shaded region $z = +\infty$, indicating unattainable values of $(\dot\alpha_p, \dot\alpha_q)$. On the right the flow potential w is illustrated in Fig. 4(b). The domain of w is shaded, and within this domain $w = 0$. Outside the shaded region $w = +\infty$, indicating unattainable values of (χ_p, χ_q).

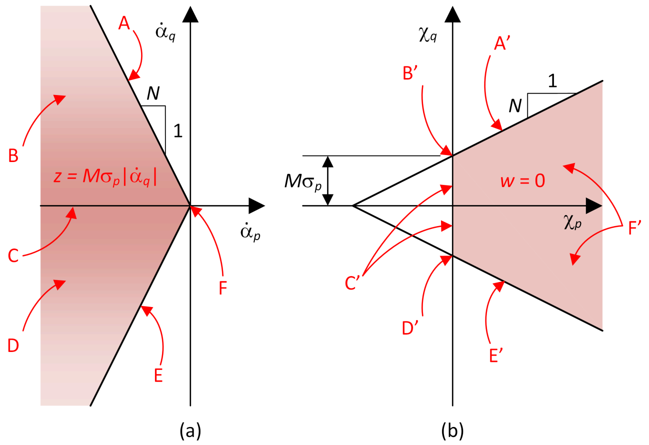

Figure 4. Relationships between the force potential z and the flow potential w: a) domain of z in $(\dot{\alpha}_p, \dot{\alpha}_q)$ plastic strain rate space; b) domain of w in (χ_p, χ_q) generalised stress space.

Any points on the lines A and A' marked in Figs. 4(a) and 4(b) are conjugate points (*i.e.*, they represent corresponding pairs of values $(\dot{\alpha}_p, \dot{\alpha}_q)$ and (χ_p, χ_q)). Importantly, this mapping is many–to–many: each point on A is conjugate to any point on A' and *vice versa*. A similar observation is made with respect to the lines E and E'. These lines represent regular shearing, accompanied by dilation, in the positive (A and A') and negative (E and E') directions. The entire shaded region B in Fig. 4(a) is conjugate to the corner B' of the shaded region in Fig. 4(b): this state only occurs when $\chi_p = 0$, and for the model considered there are side conditions (imposed by the functional form of g) that require $\sigma_p = \chi_p$, so this only occurs when in fact all the stresses are zero. At this point the plastic strain rates become indeterminate, although constrained in direction. A similar discussion applies to the region D conjugate to point D'. The line C ($\dot{\alpha}_p < 0, \dot{\alpha}_q = 0$) is conjugate to the line segment C', but note again that because of the side condition $\sigma_p = \chi_p$ this segment collapses to a single point. Finally the apex F of the shaded region at the origin in Fig. 4(a) is conjugate to the entire shaded region F' in Fig. 4(b): this represents elastic behaviour for which the yield condition is not satisfied and all plastic strain rates are zero.

Finally we can evaluate the dissipation $d = \langle \chi, \dot{\alpha} \rangle = z + w$, and we obtain $d = M\sigma_p |\dot{\alpha}_q|$ as the indicator functions evaluate to zero within the effective domains of the variables.

2.3. Interpretation in stress space

We now turn to an interpretation of the above results in true stress space, which is of course more familiar to most readers. The form of equation (20) requires that

$$|\chi_q| - N\chi_p - M\sigma_p \leqslant 0 \qquad \text{and} \qquad \chi_p \geqslant 0.$$

Noting (12) and (13) and orthogonality, the functional form of g allows us to identify that $\chi_p = \sigma_p$ and $\chi_q = \sigma_q$, so that these conditions can be rewritten as

$$|\sigma_q| - N\sigma_p - M\sigma_p \leqslant 0 \qquad \text{and} \qquad \sigma_p \geqslant 0.$$

However, note that any set of stresses that obeys the first condition automatically obeys the second, so that the domain of accessible stresses is given by:

$$f(\sigma) = |\sigma_q| - N\sigma_p - M\sigma_p \leqslant 0 \tag{24}$$

where $f(\sigma)$ is the yield function in (true) stress space.

Whilst the strength is defined by the ratio

$$|\sigma_q| / \sigma_p = M + N$$

the dilation rate is defined by the ratio $-\dot{\alpha}_p / |\dot{\alpha}_q| = N$, and clearly the model exhibits "non–associated" flow in true stress space, with (in loose terms) the apparent coefficient of friction $M + N$ being the sum of a constant volume coefficient of friction M and a coefficient of dilation N. The yield surface and plastic strain increment vectors are illustrated in Fig. 5, with yield in the positive and negative directions on lines A', E' (the lettering corresponds to the equivalent on Fig. 4(b)) and the dark blue arrows indicating the direction of the plastic strain vectors. Note that at the origin B'C'D' the second of the KKT conditions in (18) ensures that the plastic strain increment vectors fall within the fan of arrows illustrated. The pale blue shaded region F' represents elastic states within the yield surface.

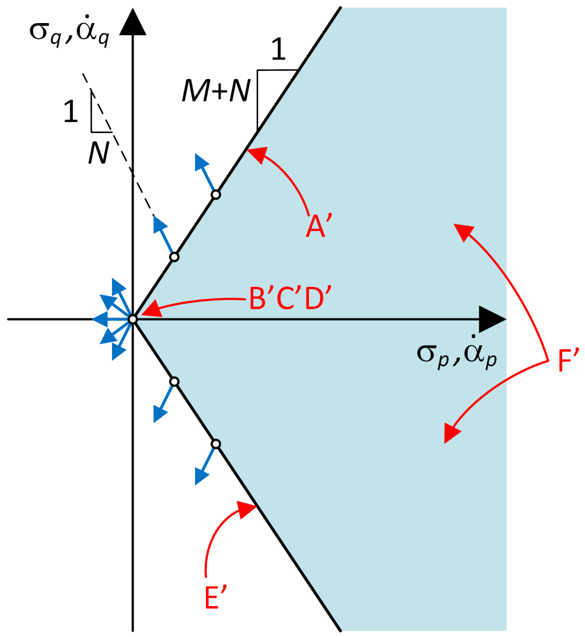

Figure 5. Yield function and plastic strain increment vectors in true stress space.

Alternatively the equation

$$y = |\chi_q| - N\chi_p - M\sigma_p = 0$$

can be understood as defining a family of yield functions in generalized stress $(\chi_p, \chi_q$ space, parameterized by the true stress σ_p. These surfaces are those shown on Fig. 4(b) as A' and E'. The "normality" relationships $\dot{\alpha} \in \lambda \partial y$ are true in this generalized stress space, but not in true stress space.

3. Discussion

3.1. Comparison with critical state models

The dissipation expression $d = M\sigma_p |\dot{\alpha}_q|$ is of course the same as the plastic work rate expression $\dot{W}_p = M\sigma_p |\dot{\alpha}_q|$ adopted in the original Cam Clay model [Schofield and Wroth, 1968], where $\dot{W}_p = \sigma_p\dot{\alpha}_p + \sigma_q\dot{\alpha}_q$ is the plastic work rate. It was implicit in the original Cam Clay model that the plastic work $\dot{W}_p$ was the same as the dissipation rate d, although it had been noted by Palmer [1967] that this was not necessarily the case, a point discussed in more detail by Collins and Houlsby [1997]: the two are only the same if the true stress and generalized stress are equal. Importantly, the Cam Clay model was developed on the basis that the "flow rule" implied by the plastic work rate expression could be converted to a plastic potential, and then Drucker's normality invoked to deduce the shape of the yield surface. It is the Author's view that the combination of frictional dissipation with Drucker's normality is inconsistent with a rigorous thermodynamic formulation of plasticity theory.

The above observation immediately raises the issue that, if such an approach is inconsistent theoretically, why does the original Cam Clay model represent the yielding of clays with some success, and the Modified Cam Clay model [Roscoe and Burland, 1968], which is based on an alternative frictional plastic work rate $\dot{W}_p = \sigma_p\sqrt{\dot{\alpha}_p^2 + M^2\dot{\alpha}_q^2}$, arguably represent soft clay behaviour even better?

Houlsby [1981] showed that Modified Cam Clay can be derived within the hyperplasticity approach by assuming that the dissipation is proportional to the preconsolidation pressure rather than the mean effective stress: $d = (p_c/2)\sqrt{\dot{\alpha}_p^2 + M^2\dot{\alpha}_q^2}$. This result points to a fundamental difference between the behaviour of clays, with dissipation proportional to preconsolidation pressure ("cohesive materials", or perhaps better "pseudo–frictional materials"), and coarse granular soils, with dissipation proportional to pressure ("frictional materials"). Collins and Kelly [2002] explore, using a slightly different terminology, transitions between these two extremes.

3.2. Incremental stress–strain response

The model described in this paper is perhaps the very simplest idealization of the behaviour of a dense sand or other frictional material. It is defined completely by the two expressions:

$$g = -\frac{\sigma_p^2}{2K} - \frac{\sigma_q^2}{6G} - \sigma_p\alpha_p - \sigma_q\alpha_q \tag{25}$$

$$w(\chi_p, \chi_q) = I_{[-\infty;0]}\left(|\chi_q| - N\chi_p - M\sigma_p\right) + I_{[-\infty;0]}\left(-\chi_p\right) \tag{26}$$

The derivation of the incremental response is as follows. First we note equations (9) and (10) in incremental form:

$$\dot{\epsilon}_p = \frac{\dot{\sigma}_p}{K} + \dot{\alpha}_p \qquad \text{and} \qquad \dot{\epsilon}_q = \frac{\dot{\sigma}_q}{3G} + \dot{\alpha}_q.$$

We then note that equation (21), once we make use of $\chi_p = \sigma_p$, gives

$$\left(\dot{\alpha}_p, \dot{\alpha}_q\right) = \left(-\lambda N, \lambda S(\chi_q)\right),$$

with the KKT conditions noted in equation (22). These require that if $|\chi_q| - N\chi_p - M\sigma_p < 0$ then $\lambda = 0$ and trivially we obtain

$$\dot{\epsilon}_p = \frac{\dot{\sigma}_p}{K} \qquad \text{and} \qquad \dot{\epsilon}_q = \frac{\dot{\sigma}_q}{3G}$$

(elastic behaviour). If on the other hand $|\chi_q| - N\chi_p - M\sigma_p = 0$ then $\lambda \geqslant 0$ and

$$\dot{\epsilon}_p = \frac{\dot{\sigma}_p}{K} - \lambda N \qquad \text{and} \qquad \dot{\epsilon}_q = \frac{\dot{\sigma}_q}{3G} + \lambda S(\chi_q).$$

The plastic multiplier λ of course remains undetermined in this perfectly plastic model. However, during plastic deformation the (generalized) stress point must remain on the yield surface, so we deduce from $|\chi_q| - N\chi_p - M\sigma_p = 0$ the incremental continuity condition

$$|\dot{\chi}_q| - N\dot{\chi}_p - M\dot{\sigma}_p = 0.$$

In view of equations (11) and (12) which give $\sigma = \chi$ and our fundamental assumption (Ziegler's orthogonality condition) $\chi = \bar{\chi}$, it follows that the continuity condition requires that

$$|\dot{\sigma}_q| - (M+N)\dot{\sigma}_p = 0.$$

This completes the incremental stress-strain relationship.

In the above very simple model the distinction between generalized stresses and true stresses may seem superfluous, but in more complex models, *e.g.*, those that involve kinematic hardening, the two quantities are not the same (they differ by the "back stress") and the distinction is essential.

Equations (25) and (26) serve to provide an extremely compact mathematical description of a model that embodies linear elasticity, and yield defined by a purely frictional criterion accompanied by dilation. Crucially, it involves an angle of dilation (related to the factor N) lower than the angle of friction (related to the factor M), *i.e.* non–associated flow conditions, within a theoretical framework that guarantees compliance with the Laws of Thermodynamics. Such a model is not new, and it bears only superficial comparison with the behaviour of real sands, but it can serve as the basis for more complex models. For instance, a more realistic model would need to include non–linear elasticity (see for instance Houlsby et al. [2005] for a suitable approach employing hyperelasticity), as well as a dependence of the dilation rate on the density (as addressed for instance by Houlsby [1992]).

The purpose of this paper has, however, not been to describe a sophisticated model, but to introduce the terminology of Convex Analysis for the description of frictional materials, with a view to using that language for more advanced constitutive modelling.

4. Conclusion

Frictional plasticity, including dilation, has been defined within the mathematical framework of Convex Analysis, allowing a rigorous mathematical treatment that defines elastic and plastic behaviour, including dilation and non–associated flow. The Laws of Thermodynamics are embodied in the hyperplasticity approach that is used. Fenchel Duals are used to interchange the roles of dependent and independent variables. Whilst the convex analytical terminology used here may be unfamiliar to many readers, it is suggested that it has advantages as the natural mathematical language for expressing plasticity theories.

Conflicts of Interest

There are no potential conflicts of interest.

Acknowledgements

The author is grateful to Prof. Alexander Puzrin and an anonymous reviewer (*Editor's note: Prof. Denis Caillerie*) for their helpful comments on this paper.

Review History

This paper was sent to two Reviewers: Prof. Denis Caillerie, Université Grenoble Alpes and Prof. Francesco Solombrino, University of Naples "Federico II". The two reviewers remained anonymous during the entire revision process. After the reviewing process was completed, both reviewers decided to disclose their identity. There was a total of three rounds before both reviewers and the editor decided to accept the paper. Please see the Review History online for the complete reviews: ReviewHistory.pdf.

References

Collins, I. F. and Houlsby, G. T. (1997). Application of thermomechanical principles to the modelling of geotechnical materials. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 453:1975–2001.

Collins, I. F. and Kelly, P. A. (2002). A thermomechanical analysis of a family of soil models. *Géotechnique*, 52(7):507–518.

Drucker, D. C. (1951). A more fundamental approach to plastic stress-strain relations. In *Proc. of 1st US National Congress of Applied Mechanics, 1951*, pages 487–491.

Drucker, D. C. (1959). A definition of a stable inelastic material. *Journal of Applied Mechanics*, pages 101–106.

Han, W. and Reddy, B. D. (1999). *Plasticity: mathematical theory and numerical analysis.* Springer.

Houlsby, G. T. (1981). *Study of plasticity theories and their applicability to soils.* PhD thesis, University of Cambridge.

Houlsby, G. T. (1992). Interpretation of dilation as a kinematic constraint. In *Modern approaches to plasticity*, pages 19–38. Elsevier: New York.

Houlsby, G. T., Amorosi, A., and Rojas, E. (2005). Elastic moduli of soils dependent on pressure: a hyperelastic formulation. *Géotechnique*, 55(5):383–392.

Houlsby, G. T. and Puzrin, A. M. (2000). A thermomechanical framework for constitutive models for rate–independent dissipative materials. *International journal of Plasticity*, 16(9):1017–1047.

Houlsby, G. T. and Puzrin, A. M. (2006). *Principles of hyperplasticity: an approach to plasticity theory based on thermodynamic principles.* Springer Verlag, London.

Maugin, G. (1992). *The thermodynamics of plasticity and fracture.* Cambridge University Press.

Moreau, J. J. (1974). On unilateral constraints, friction and plasticity. In *New variational techniques in mathematical physics*, pages 173–322. CIME.

Palmer, A. C. (1967). Stress-strain relations for clays: an energy theory. *Géotechnique*, 17(4):348–358.

Rockafellar, R. T. (1970). *Convex Analysis.* Princeton University Press, NJ.

Romano, G. (1995). New results in subdifferential calculus with applications to convex optimization. *Applied Mathematics and Optimization*, 32(3):213–234.

Roscoe, K. H. and Burland, J. B. (1968). On the generalized behaviour of 'wet' clay. In Heyman, J. and Leckie, F. A., editors, *Engineering Plasticity*, pages 535–610. Cambridge University Press.

Schofield, A. N. and Wroth, C. P. (1968). *Critical state soil mechanics.* McGraw–Hill London.

Taylor, D. W. (1948). *Fundamentals of Soil Mechanics.* Wiley, New York.

Ziegler, H. (1977). *An Introduction to Thermomechanics.* North Holland, Amsterdam. (2nd edition 1983).

Appendix: Derivation of Equation (20)

We seek the Fenchel Dual of the force potential given in Equation (14):

$$z(\dot\alpha_p,\dot\alpha_q) = M\sigma_p\left|\dot\alpha_q\right| + I_{[-\infty;0]}\left(\dot\alpha_p + N\left|\dot\alpha_q\right|\right)$$

The dual is defined as:

$$z_1^*(\chi_p,\chi_q) = \sup_{(\dot\alpha_p,\dot\alpha_q)}\Big\{\chi_p\dot\alpha_p + \chi_q\dot\alpha_q - M\sigma_p\left|\dot\alpha_q\right| - I_{[-\infty;0]}\left(\dot\alpha_p + N\left|\dot\alpha_q\right|\right)\Big\}$$

Which by virtue of the indicator function becomes:

$$z_1^*(\chi_p,\chi_q) = \sup_{\substack{(\dot\alpha_p,\dot\alpha_q)\\ \dot\alpha_p+N|\dot\alpha_q|\leqslant 0}}\Big\{\chi_p\dot\alpha_p + \chi_q\dot\alpha_q - M\sigma_p\left|\dot\alpha_q\right|\Big\}$$

We now consider the value of the supremum. We first observe that if $\chi_p < 0$ then the values $\dot\alpha_p = -\infty$, $\dot\alpha_q = 0$ result in the supremum equal to $+\infty$. Considering the case when $\chi_p \geqslant 0$, the first term in the supremum is maximized when $\dot\alpha^p$ takes its maximum allowable value, which is $-N\left|\dot\alpha_q\right|$. We can also observe that the second term is maximized (without affecting the value of the third term) when χ_q has the same sign as $\dot\alpha_q$ rather than the opposite, so this term could be written as $\left|\chi_q\right|\left|\dot\alpha_q\right|$. The supremum can therefore be rewritten in the form

$$\sup_{\dot\alpha_q}\Big\{\left(\left|\chi_q\right| - N\chi_p - M\sigma_p\right)\left|\dot\alpha_q\right|\Big\}.$$

Clearly this takes the value zero if $\left|\chi_q\right| - N\chi_p - M\sigma_p \leqslant 0$ and $+\infty$ if $\left|\chi_q\right| - N\chi_p - M\sigma_p > 0$. We can therefore identify $z_1^*(\chi_p,\chi_q)$ as the indicator function of the set

$$X(\chi_p,\chi_q) = \Big\{(\chi_p,\chi_q)\;\Big|\;\chi_p \geqslant 0,\ \left(\left|\chi_q\right| - N\chi_p - M\sigma_p\right) \leqslant 0\Big\},$$

and we can therefore write:

$$z^*(\chi_p,\chi_q) = I_{[-\infty;0]}\left(\left|\chi_q\right| - N\chi_p - M\sigma_p\right) + I_{[-\infty;0]}\left(-\chi_p\right).$$

Notation

C	Complementary energy
d	Dissipation
f	(1) Helmholtz free energy; (2) yield function in true stress space
g	Gibbs free energy
E	Strain energy
G	Shear modulus
$I_C(x)$	Indicator Function of a set C
K	Bulk modulus
M	Friction constant
N	Dilation constant
$\mathcal{N}_C(x)$	Normal Cone of a set C
$S(x)$	Generalised Signum Function
p, q	Subscripts associating variable with Cambridge triaxial parameters
p_c	Preconsolidation pressure
w	Flow potential
$\dot{W}_p$	Plastic work rate
y	Yield function in generalised stress space
z	Force potential
α	Internal variable (in this paper equal to plastic strain)
ϵ	Strain
λ, Λ	Factors in KKT conditions
σ	Stress
$\chi, \bar{\chi}$	Generalised stress
$\partial f(x)$	Subdifferential
$\langle x^*, x\rangle$	Inner product

Manuscript received 13th July 2018, accepted 16th August 2019.