Convergence of Augmented Lagrangian Methods in Extensions Beyond Nonlinear Programming

R. Tyrrell Rockafellar[1]

Abstract

The augmented Lagrangian method (ALM) is extended to a broader-than-ever setting of generalized nonlinear programming in convex and nonconvex optimization that is capable of handling many common manifestations of nonsmoothness. With the help of a recently developed sufficient condition for local optimality, it is shown to be derivable from the proximal point algorithm through a kind of local duality corresponding to an optimal solution and accompanying multiplier vector that furnish a local saddle point of the augmented Lagrangian. This approach leads to surprising insights into stepsize choices and new results on linear convergence that draw on recent advances in convergence properties of the proximal point algorithm. Local linear convergence is shown to be assured for a class of model functions that covers more territory than before.

Keywords: *generalized augmented Lagrangians, generalized nonlinear programming, multiplier methods, ALM, localized proximal point algorithm, linear convergence criteria, sufficient conditions for local optimality, variational convexity, strong variational sufficiency, second-order variational analysis, local duality, conic programing, second-order cone programming, semidefinite programming, convex and nonconvex optimization, nonsmooth optimization*

Version of 26 February 2022

[1]University of Washington, Department of Mathematics, Box 354350, Seattle, WA 98195-4350;
E-mail: *rtr@uw.edu*, URL: sites.math.washington.edu/∼rtr/mypage.html

1 Introduction

In the "method of multipliers" proposed in 1969 independently by Hestenes [12] and Powell [17] (and slightly later also by Haarhoff and Buys [11]) for minimizing a function $f_0(x)$ of $x \in \mathbb{R}^n$ subject to equality constraints $f_i(x) = 0$ for $i = 1, \ldots, m$, the ordinary Lagrangian function

$$L(x, y) = f_0(x) + y_1 f_1(x) + \cdots + y_m f_m(x) \tag{1.1}$$

was augmented by terms $\frac{r}{2} f_i(x)^2$. In step k the augmented Lagrangian was minimized in x for given multipliers y_i^k to get x^{k+1}, and then y_i^k was updated to $y_i^{k+1} = y_i^k + r f_i(x^{k+1})$. An extension of the procedure to inequality constraints $f_i(x) \leq 0$ was devised in 1970 by Rockafellar [19] and taken up by Buys in his 1972 thesis in Leiden [5]. That version, in the case of convex functions f_i, was shown in 1973 [20] to correspond to solving an associated dual problem by a method that later was recognized as an application of the proximal point algorithm [25]; see [21] and [26]. But even in nonconvex nonlinear programming, a kind of local duality based on sufficient conditions for local optimality emerged as the key to understanding convergence of the augmented Lagrangian method, as revealed by Bertsekas in his landmark 1982 book [3].

Connections with duality and the proximal point algorithm will be front and center here as we explore convergence properties of augmented Lagrangian methods in an extended setting. We focus on solving *generalized* nonlinear programming problems of the form

$$(P) \qquad \begin{array}{l} \text{minimize } \ \varphi(x, u) = f_0(x) + g(F(x) + u) \ \text{ subject to } \ u = 0, \\ \text{with } g \text{ closed proper convex and } F(x) = (f_1(x), \ldots, f_m(x)), \end{array}$$

where the vector u has the role of a canonical perturbation variable. The perturbation structure leads to associating with (P) the *generalized* Lagrangian function

$$l(x, y) := \inf_u \{\, \varphi(x, u) - y \cdot u \} = L(x, y) - g^*(y), \tag{1.2}$$

where g^* is the convex function conjugate to g, as well as the *generalized augmented* Lagrangian function

$$l_r(x, y) := \inf_u \Big\{\, \varphi(x, u) - y \cdot u + \frac{r}{2}|u|^2 \Big\} \ \text{ for } r > 0 \text{ and } \ |u| = ||u||_2, \tag{1.3}$$

cf. [32, Sections 11I+11K]. In terms of the auxiliary convex functions

$$\begin{array}{l} g^r(u) := \min_{u'} \Big\{\, g(u') + \frac{r}{2}|u' - u|^2 \Big\} \ \text{ with conjugate } \ g^{r\,*}(y) = g^*(y) + \frac{1}{2r}|y|^2, \\ g_r(u) \ := g(u) + \frac{r}{2}|u|^2 \ \text{ with conjugate } \ g_r^*(y) \ = \min_{y'} \Big\{\, g^*(y') + \frac{1}{2r}|y' - y|^2 \Big\}, \end{array} \tag{1.4}$$

the augmented Lagrangian has the alternative expressions

$$\begin{array}{l} l_r(x, y) = f_0(x) + g^r(F(x) + \frac{1}{r}y) - \frac{1}{2r}|y|^2, \ \text{ or} \\ l_r(x, y) = L(x, y) + \frac{r}{2}|F(x)|^2 - g_r^*(y + rF(x)). \end{array} \tag{1.5}$$

Problem (P) becomes *conic* programming when g is the indicator δ_K of a closed convex cone K. Then g^* is the indicator of the polar cone $Y = K^*$, and in terms of the distance functions d_K and d_Y associated with those cones, the expressions for $l_r(x, y)$ have $g^r = \frac{r}{2} d_K^2$ and $g_r^* = \frac{1}{2r} d_Y^2$. Classical nonlinear programming is recovered by taking K to be the standard constraint cone there. Second-order cone programming has K being the Lorenz cone (the epigraph of the Euclidean norm), while semidefinite programming has K being the cone of positive definite symmetric matrices. But the

composite term in (P) can encompass much more than a constraint $F(x) \in K$. For illustration, it's possible to have in terms of a closed convex set C that

$$\begin{array}{c} g(F(x)) = \delta_C(f_1(x), \ldots, f_a(x)) + \max\Big\{f_{a+1}(x), \ldots, f_b(x)\Big\} + \|(f_{b+1}(x), \ldots, f_m(x))\|_1, \\ \text{in which case} \quad g^*(y) = \sigma_C(y_1, \ldots, y_a) + \delta_\Sigma(y_{a+1}, \ldots, y_b) + \delta_{\boldsymbol{B}_\infty}(y_{b+1}, \ldots, y_m) \end{array} \tag{1.6}$$

for the support function σ_C of C, the unit simplex Σ, and the unit ball for $\|\cdot\|_\infty$.

Although earlier problem formats on these lines, as in [27], have typically included also an abstract constraint $x \in X$ for a closed convex set X, that has been omitted from (P) in alignment with our work in [30], where such a constraint was judged to be unduly distracting and vision-obscuring for the second-order variational analysis that was needed. (An alternative for enforcing $x \in X$ is replacing $F(x)$ by $\overline{F}(x) = (F(x), x)$ and $g(u)$ by $\overline{g}(u, u') = g(u) + \delta_X(u')$.)

The *convex case* of (P), in which $\varphi(x, u)$ is a convex function of (x, u), corresponds to the Lagrangian $l(x, y)$ being convex in $x \in I\!R^n$ for each $y \in I\!R^m$. That case will be important to us, both directly and as a template for developing augmented Lagrangian methodology. Of course $l(x, y)$ and $l_r(x, y)$ are always concave in y for every x, because of the convexity of g, which also holds for g^r and g_r^*. Outside of the convex case of (P), we exploit situations where, as it turns out from [30] in ways not well appreciated in the past, the augmented Lagrangian $l_r(x, y)$ will have a useful amount of *local* convexity in x, but not for every y. In that picture, properties of derivatives of $l_r(x, y)$ in x and y that come from the functions g^r and g_r^* will be crucial. Those functions are themseves $\mathcal{C}^{1+}$, i.e., differentiable with gradient mappings that are locally Lipschitz continuous. In fact, those mappings ∇g^r and ∇g_r^* are globally Lipschitz continuous with Lipschitz constants r and r^{-1}, respectively, in consequence of the inf-convolution expressions for g^r and g_r^* in (1.4). Therefore, $l_r(x, y)$ is $\mathcal{C}^1$ with respect to (x, y) when the functions $f_0, f_1, \ldots, f_m$ are $\mathcal{C}^1$, which is our baseline assumption, and $\mathcal{C}^{1+}$ when they are actually $\mathcal{C}^2$. We'll refer to that as *the* $\mathcal{C}^2$ *case* of problem (P).

Algorithm. The extended form of the augmented Lagrangian method (ALM) that we employ for solving problem (P) follows the pattern that sequences of primal vectors $x^1, x^2, \ldots$ and dual vectors $y^0, y^1, y^2, \ldots$ are generated by

$$x^{k+1} \approx \bar{x}^{k+1} = \operatorname*{argmin}_{x \in \mathcal{X}} l_{r_k}(x, y^k), \qquad y^{k+1} = y^k + r'_k \nabla_y l_{r_k}(x^{k+1}, y^k) \tag{1.7}$$

with respect to nondecreasing sequences of parameter values $r_k > 0$ and $r'_k > 0$. Here "$\approx$" refers to allowing the minimization to be inexact — as quantified by some stopping criterion to be specified later along with the set $\mathcal{X}$ that localizes the minimization and supports its attainment at a unique point. The updating rule for y^k in (1.7) reduces to the traditional one when (P) specializes to classical nonlinear programming and $r'_k = r_k$. However, our analysis will uncover reasons why taking $r'_k < r_k$ may lead to improved rates of convergence, not only in the original ALM setting, but also for ALM forays into cases of (P) beyond traditional nonlinear programming.

One of our main goals in this paper is demonstrating how, through the local convexity detected in [30], the augmented Lagrangian method can be derived by applying the proximal point algorithm to a dual problem of concave maximization, even for nonconvex (P). This contrasts with the entire nonconvex ALM research literature after Bertsekas [3], which instead has looked for direct estimates of the iterations, as in the innovative contribution of Fernandes and Solodov [8] that allowed weakening of some of the customary assumptions in the classical NLP setting. Our different approach brings the r'_k stepsize question to light and leads, we hope, to more transparency about the fundamental underpinnings of the method. It relies on a refined view of second-order sufficiency, which imposes a small restriction in some situations compared to other research results, but makes it possible to

establish convergence *for all problems in our* (P) *format* without imposing a constraint qualification-type condition or requiring a unique multiplier element in optimality.

Another of our goals, even for the convex case, is gaining insights into the deep underlying circumstances that support linear rates of convergence and identifying forms of (P) in which those circumstances can be counted on to be present. We are able in this, through our different approach, to draw on new results about the behavior of the proximal point algorithm itself [31], including how it can be articulated in a merely local manner and, when the solution set is more than a singleton, must converge directionally. Second-order variational analysis [32] helps us to translate such properties in the small to properties of the local dual objective function around the dual solution set. The challenge faced then is how to pass to verifiable properties of (P) itself. Our main accomplishment in that direction is tying this to the model function g. We show in particular that linear convergence is guaranteed when g is "fully amenable" [32, 10F]. That category of modeling covers most of the forms of (P) that researchers have worked on until now, and much more. However, it leaves out semidefinite programming extensions of the augmented Lagrangian method as in [33].

Much has already been written about nonconvex ALM in certain extensions beyond nonlinear programming that fit into our format as particular cases. A full review would turn into a lengthy digression, but two recent contributions on the forefront may, with their many references, serve to indicate where research currently stands. In [9] by Hang, Mordukhovich and Sarabi, the topic is second-order cone programming in the line of [2, 4, 13], the case here of $g - \delta_K$ for the Lorenz cone, while in [10] by Hang and Sarabi, g is instead be any piecewise linear-quadratic convex function in the sense of [32, Sec. 10E]). A milder second-order sufficient condition than ours is imposed in [10],[2] but both papers restrict in other ways. Uniqueness of the multiplier vector is called for in [9],[3] whereas the minimization is required to be exact ($x^k = \bar{x}^k$) in [10]. Both arrive at Q-linear convergence of the primal-dual sequence of pairs (x^k, y^k) in a pattern pioneered in [8]. That implies R-linear convergence of the x^k and y^k sequences individually. Our proximal point approach proceeds from Q-linear convergence of y^k to R-linear convergence of $\bar{x}^k$ as a tag-along, and with a tighter stopping criterion (still short of the exact minimization in [10]) gets R-linear convergence of x^k. That automatically yields also R-linear, although not Q-linear, convergence of the pairs (x^k, y^k). This distinction prevails also in relating our results here to previous work in classical nonlinear programming itself, such as in [8]. From the angle of such linear convergnce details, our results are complementary, therefore, to those in the existing literature, but they cover more optimization territory and bring unnoticed fundamentals of the algorithm to view.

There is no need to go into which numerical procedure might be invoked for the approximate minimization, but properties of the augmented Lagrangian as a function of x would certainly be relevant to that. With $l_r(x, y)$ generally not being $\mathcal{C}^2$ even in the $\mathcal{C}^2$ case of (P), "second-order" methods would require adaptation. The $\mathcal{C}^{1+}$ property of $l_r(x, y)$ in the $\mathcal{C}^2$ case does, at least, imply twice differentiabiity almost everywhere in an extended sense [32, Sec. 13A]. Some of the potentially useful implications of that for computations have been brought out in [30] together with the development of a sufficient condition for local optimality in (P) at the second-order level.

That sufficient condition will be our bridge for crossing from from global ALM behavior in the convex case of (P) to closely parallel, but only local, behavior in the nonconvex case. Innovations in dualization are a crucial part of this and need some explanation before we can go further with comparing our contribution to past work.

[2]The definite difference between the two conditions is confirmed by the example that answers [30, Question 2].

[3]Although the statement of their convergence result [9, Theorem 5.3] explicitly assumes this, the authors say that their proof mostly goes through without it. For more, see the discussion in Section 5 after our Example 5.3.

The central role of duality. Equivalent first-order optimality conditions for (P) with respect to a primal-dual pair $(\bar{x}, \bar{y})$ can be expressed in several ways, as elaborated in [30], namely

$$(0, \bar{y}) \in \partial\varphi(\bar{x}, 0), \quad \text{or} \quad (0, \bar{y}) \in \partial\varphi_r(\bar{x}, 0) \quad \text{where} \quad \varphi_r(x, u) = \varphi(x, u) + \frac{r}{2}|u|^2, \tag{1.8a}$$

or in Lagrangian terms by

$$0 = \nabla_x l(\bar{x}, \bar{y}) \ \text{ and } \ 0 \in \partial_y[-l](\bar{x}, \bar{y}), \tag{1.8b}$$

which comes down to

$$\nabla_x L(\bar{x}, \bar{y}) = 0 \ \text{ with } \ \bar{y} \in \partial g(F(\bar{x})), \tag{1.8c}$$

or through the augmented Lagrangian (1.3) as

$$\nabla_x l_r(\bar{x}, \bar{y}) = 0 \ \text{ and } \ \nabla_y l_r(\bar{x}, \bar{y}) = 0, \tag{1.8d}$$

Under a constraint qualification, these equivalent conditions are necessary for having a local minimum in (P) at $\bar{x}$ [32, 11.43], but here we'll only be concerned with their role in sufficiency.

In the convex case of (P), the first-order conditions guarantee that $\bar{x}$ gives not just a local minimum but a global minimum. As seen from the version in (1.8b) and the convexity of $l(x, y)$ in x, along with the ever-present concavity in y, they correspoond to $(\bar{x}, \bar{y})$ being a *global saddle point* of l on $\mathbb{R}^n \times \mathbb{R}^m$. The multiplier vectors $\bar{y}$ that enter are then the solutions to a dual problem,

$$(D) \qquad \text{maximize } \ h(y) \ \text{ over } \ y \in \mathbb{R}^m, \ \text{ where } \ h(y) = \inf_{x \in \mathbb{R}^m} l(x, y).$$

It was shown in [21, 26], that the augmented Lagrangian iterations (1.7) with $r'_k = r_k$ correspond in classical convex programming to iterations of the proximal point algorithm in maximizing the concave function h in problem (D). A linear rate of convergence of y^k to $\bar{y}$ was derived under the assumption that $\bar{y}$ is the unique solution to (D) and h has a quadratic growth property there. Uniqueness of $\bar{x}$ as a solution to (P) then entails x^k converging to $\bar{x}$.

In this paper, those long-standing results will be extended to the general convex case of (P) with significant improvements. Criteria for linear convergence of x^k to $\bar{x}$ will be obtained, even in circumstances of nonuniqueness of $\bar{y}$ as a solution to (D).

Variational sufficiency. In passing beyond the convex case of (P), our efforts will likewise revolve around identifying the augmented Lagrangian iterations (1.7) with proximal point iterations, but in a *localized* framework of saddle points and duality. This framework is based on the sufficient condition for local optimality in nonconvex optimization that we introduced in [29] and studied in depth in [30] for problem (P). That condition utilizes the notion of the *variational convexity* of a function with respect to a pair of elements in the graph of its subgradient mapping. It builds on the first-order subgradient condition (1.8a) by stipulating for the augmented objective function φ_r that

$$\exists\, \bar{r} \text{ such that } \varphi_{\bar{r}} \text{ is variationally convex at } (\bar{x}, 0) \text{ for } (0, \bar{y}). \tag{1.9}$$

This is the *variational* sufficient condition (at level $\bar{r}$) for local optimaity in (P). Its enhancement with $\varphi_{\bar{r}}$ variationally *strongly* convex is the *strong* variational sufficient condition. When one of these holds for $\bar{r}$, it also holds for every $r > \bar{r}$, so the properties are of "threshold" type.

Variational convexity was introduced in [28] in an echo of an unnamed but mostly stronger property that was crucial earlier in results on tilt stability in [16]. It captures the situation in which *the values and subgradients of a function are locally indistinguishable from those of a convex function*, the localization being in a "primal-dual" sense. In describing what that means exactly for us here, there

are simplications because the functions φ and φ_r are amenable, prox-regular and subdifferentially continuous, as ascertained in [30], and all their subgradients are *regular* in the sense of variational analysis [32]. The variational convexity in (1.9) comes down then to having open convex neighborhoods $\mathcal{W}$ of $(\bar{x},0)$ and $\mathcal{Z}$ of $(0,\bar{y})$ such that

$$\begin{array}{l}\exists\,\bar{r} \text{ and a proper lsc convex function } \psi \le \varphi_{\bar{r}} \text{ on } \mathcal{W} \text{ such that}\\ \qquad [\mathcal{W}\times\mathcal{Z}]\cap \operatorname{gph}\partial\psi \;=\; [\mathcal{W}\times\mathcal{Z}]\cap \operatorname{gph}\partial\varphi_{\bar{r}}\\ \text{and, for } (x,u;v,y) \text{ belonging to this common set, } \psi(x,u)=\varphi_{\bar{r}}(x,u).\end{array} \tag{1.10}$$

Variational *strong* convexity has ψ *strongly* convex on $\mathcal{W}$. With $s>0$ as the modulus of that strong convexity, variational strong convexity of $\varphi_{\bar{r}}$ at $(\bar{x},0)$ for $(0,\bar{y})$ can be identified with the "parametric quadratic growth condition" that

$$\begin{array}{c}\varphi_{\bar{r}}(x',u')\ge \varphi_{\bar{r}}(x,u)+(v,y)\cdot[(x',u')-(x,u)]+\frac{s}{2}|(x',u')-(x,u)|^2\\ \text{for } (x',u')\in\mathcal{W} \text{ when } (v,y)\in\mathcal{Z}\cap\partial\varphi_r(x,u).\end{array} \tag{1.11}$$

This can be contrasted with the more commonly studied "quadratic growth condition" on local optimality that it includes as the case where $x=\bar{x}$ and $u=u'=0$.

Although the strong varational sufficient condition for local optimality might seem more theoretical than practical, it is in fact *equivalent* to standard strong sufficiency in classical nonlinear programming,[4] the SOSC in second-order cone programming,[5] and still other instances of (P) identified in [30]. In general, through the tilt stability property in (1.11), it stands as an enrichment of other known second-order sufficient conditions in the sense of offering an extra degree of support for algorithm development.

The key point for our purposes here is that variational sufficiency precisely captures the possibility of locally reducing a nonconvex problem (P) to a problem fully in the realm of convex analysis. This is seen in the following results from [30], which will guide us to our portrayal of the augmented Lagrangian method (1.7) as a dual application of the proximal point algorithm even in the *nonconvex* case of (P).

Theorem 1.1 [30, Theorem 1] (Lagrangian characterization of variational sufficiency)**.** *With respect to $\bar{x}$ and $\bar{y}$ satisfying the first-order optimality condition in (P), the variational sufficient condition for local optimality holds at level $\bar{r}$ if and only if there is a closed convex neighborhood $\mathcal{X}\times\mathcal{Y}$ of $(\bar{x},\bar{y})$ such that $l_{\bar{r}}(x,y)$ is convex in $x\in\mathcal{X}$ if $y\in\mathcal{Y}$ and concave in $y\in\mathcal{Y}$ if $x\in\mathcal{X}$. Then $l_r(x,y)$ for every $r\ge\bar{r}$ enjoys those properties and has $(\bar{x},\bar{y})$ as a saddle point relative to $\mathcal{X}\times\mathcal{Y}$.*

Theorem 1.2 [30, Theorem 2] (Lagrangian characterization of strong variational sufficiency)**.** *The strong version of the variational sufficient condition for local optimality corresponds to strengthening the characterization of variational sufficiency in Theorem 1.1 to include augmented tilt stability, namely, the existence of a neighborhood $\mathcal{V}$ of 0 such that the mapping*

$$(v,y)\mapsto \operatorname*{argmin}_{x\in\mathcal{X}}\{\,l_{\bar{r}}(x,y)-v\cdot x\} \;\text{ for }\; (v,y)\in\mathcal{V}\times\mathcal{Y} \tag{1.12}$$

is single-valued and Lipschitz continuous. It corresponds equally to having the functions $l_{\bar{r}}(\cdot,y)$ on $\mathcal{X}$ for $y\in\mathcal{Y}$ be strongly convex, all with the same modulus of strong convexity. The modulus $s>0$ for

[4] As shown in Example 1 of [30]

[5] When $F(\bar{x})$ isn't at the apex of the constraint cone, i.e., $F(\bar{x})\neq 0$, as shown in Example 3 of [30]. For more on this condition see [9, Proposition 2.1] and the discussion after it.

that strong convexity[6] *then yields, as* s^{-1}*, a modulus for the Lipschitz continuity in the augmented tilt stability.*

Theorems 1.1 and 1.2 only depend on the functions f_i in (P) being $\mathcal{C}^1$, but when those functions are $\mathcal{C}^2$, strong variational sufficiency can be identified with conditions involving generalized second derivatives, cf. [30, Sections 3 and 4].

The properties of $l_{\bar r}$ on $\mathcal{X}\times\mathcal{Y}$ in Theorem 1.2 carry over to l_r on $\mathcal{X}\times\mathcal{Y}$ with the same modulus for all $r>\bar r$, because[7] $\lambda_r(x,y)=\max_{y'}\{\,l_{\bar r}(x,y')-\frac{1}{2(r-\bar r)}|y'-y|^2\}$; the pointwise supremum of a family of functions that are strongly convex on $\mathcal{X}$ with modulus s inherits that same strong convexity. The augmented tilt stability and guaranteed strong convexity have obvious significance in that way for the approximate minimization step in the ALM iteration (1.7). If the localization set $\mathcal{X}$ in (1.7) agrees with the one in (1.12) (or a smaller neighborhood of $\bar x$ within it), and $r_k\geq\bar r$, two candidates x and x' for approximate minimizers, as judged by the size of the gradients $v=\nabla_x l_{r_k}(x,y^k)$ and $v'=\nabla_x l_{r_k}(x',y^k)$, will have $|x'-x|\leq s^{-1}|v'-v|$. This is valuable for controlling errors in the approximation. The strong convexity in the mimization may furthermore help in supporting the minimization procedure that may be applied in the subproblems..

Other generalizations of the classical strong second-order sufficient condition (SOSSC), in situations where they truly turn out to be milder than strong variational sufficiency, are unable to provide those benefits. Even if convergence of a form of ALM is supported, the minimization steps will, through the associated lack of local strong convexity and augmented tilt stability, be more delicate and less under control, whether or not that is apparent from how convergence results are written up.

Local duality. The convex-concave-type saddle point property in Theorem 1.1 corresponds in the duality format of convex analysis [18], [23], [32, Chap. 11], to saying that

$$\begin{array}{l}\bar x \text{ minimizes over } x\in\mathcal{X} \text{ the convex function } \sup_{y\in\mathcal{Y}} l_{\bar r}(x,y),\\ \bar y \text{ maximizes over } y\in\mathcal{Y} \text{ the concave function } \inf_{x\in\mathcal{X}} l_{\bar r}(x,y),\\ \text{and the optimal values in these paired problems are equal.}\end{array}\tag{1.13}$$

Solving (P) by determining a pair $(\bar x,\bar y)$ that satisfies the variational sufficient condition for local optimality at a level $\bar r$ thus relates to solving primal and dual problems as in (1.13). When the proximal point algorithm is invoked for the dual problem, the iterations turn out reduce to those of the augmented Lagrangian method (1.7), and that will enable us to derive ALM convergence properties from those of the proximal point algorithm. Through recent advances in [31], we will obtain linear convergence of x^k to $\bar x$ without having to assume there is only one $\bar y$ partnered with $\bar x$. Moreover, despite the indirectness of the setting in (1.13), we will be able in this manner to tie such convergence to local properties in problem (P) itself.

Stepsize implications. In the literature on versions of the augmented Lagrangian method that have been implemented in nonconvex optimization, the convergence analysis starts with x^k and y^k "almost" satisfying some local optimality condition, and with the minimization step having $r_k\geq\bar r$ for a threshold value $\bar r>0$ drawn from that condition. The practical question of how to know in the course of computations whether these circumstances are being met, is left open, and that will be the case here as well. However, an interesting difference will emerge. Up to now, the stepsize r'_k for

[6] It is the same as the modulus of strong convexity invoked in the assumption of strong variational convexity, as seen in [30] in the theorem's proof, although not brought out in the theorem's statement.

[7] This is dual to the formula $\varphi_r(x,\cdot)=\varphi_{\bar r}(x,\cdot)+\frac{r-\bar r}{2}|\cdot|^2$ through the conjugacy of $\varphi_r(x,\cdot)$ and $\varphi_{\bar r}(x,\cdot)$ with the functions $-l_r(x,\cdot)$ and $-l_{\bar r}(x,\cdot)$ in the definition (1.3), along with the fact that addition of convex functions dualizes to infimal convolution [32, 11.23(a)].

updating from y^k to y^{k+1} in (1.7) has been taken to be r_k, but our approach via the proximal point algorithm with its stepsizes c_k, making r_k come out at $\bar{r} + c_k$, will suggest taking $r'_k = c_k = r_k - \bar{r}$ instead. Might that really help convergence?

It's difficult to be sure in general, because the arguments that researchers use to tease out convergence rates have often been exceedingly complicated, producing results perhaps more qualitative than quantitative. Nonetheless, support can be found in the analysis of original method of multipliers by Bertsekas [3], when examined from the perspective of the 1992 innovation of Eckstein and Bertsekas [7] that allows the stepsize c_k in the proximal point algorithm to be relaxed to any $c'_k \in (0, 2c_k)$. For us, that would translate in (1.7) to updating with any $r'_k = c'_k$ instead of $r'_k = c_k$ and covers the choice $r'_k = r_k = \bar{r} + c_k$ as long as

$$r_k \in (0, 2c_k) = (0, 2[r_k - \bar{r}]), \;\; \text{i.e., } \; r_k > 2\bar{r}. \tag{1.14}$$

In fact, Bertsekas in [3, Proposition 2.7] does require $r_k > 2\bar{r}$ with respect to the threshold value we would deem as $\bar{r}$ in that classical setting. Relaxed proximal point stepsizes do, therefore, appear to be operating behind ALM iterations in that case. But Pennanen [15] determined in 2002 that the best rate in the proximal point algorithm would be obtained only with the relaxation ultimately infinitesimal, i.e., with $c'_k/c_k \to 1$. Whether the same phenomenon is operating in ALM extensions like those in [9] and [10] would take more effort to pin down. The convergence analysis there goes through chains of delicate estimates of Lipschitz constants, quadratic growth constants, and the like. It's hard to trace the quantitative effects, but there are hints of what Bertsekas saw.

The implication for ALM seems anyway to be that updating in (1.7) with $r'_k = r_k - \bar{r}$ instead of the traditional $r'_k = r_k$ could offer an improvement. That will be confirmed through our analysis.

This issue of stepsize is consequential also for the standards adopted for the approximate minimization in (1.7). In terms of error parameters ε_k, three levels of increasing tightness will basically come into play here for the acceptability of x^{k+1}:

$$\Big(2r'_k\Big[l_{r_k}(x^{k+1}, y^k) - \inf\nolimits_{\mathcal{X}} l_{r_k}(\cdot, y^k)\Big]\Big)^{1/2} \leq \begin{cases} \text{(a)} & \varepsilon_k \\ \text{(b)} & \varepsilon_k \min\{1, |r'_k \nabla_y l_{r_k}(x^{k+1}, y^k)|\} \\ \text{(c)} & \varepsilon_k \min\{1, |r'_k \nabla_y l_{r_k}(x^{k+1}, y^k)|^2\}. \end{cases} \tag{1.15}$$

In the past, levels (a) and (b) have been utilized (in their reduction to special cases) in taking $r'_k = r_k$, but now, for the reasons explained, we permit $r'_k \neq r_k$. The (c) level is new for ALM, being brought in from [31]. It will support linear convergence in partnership with strong variational sufficiency, where strong convexity of the augmented Lagrangian expressians $l_{r_k}(x, y^k)$ in x will be available in consequence of Theorem 1.2. That strong convexity with modulus s provides the estimate[8] $l_{r_k}(x^{k+1}, y^k) - \inf_{\mathcal{X}} l_{r_k}(\cdot, y^k) \leq \frac{1}{2s}|\nabla_x l_{r_k}(x^{k+1}, y^k)|^2$, which allows (1.15) to be replaced by

$$\sqrt{r'_k}\,\Big|\nabla_x l_{r_k}(x^{k+1}, y^k)\Big| \leq \begin{cases} \text{(a)} & \varepsilon'_k \\ \text{(b)} & \varepsilon'_k \min\{1, |r'_k \nabla_y l_{r_k}(x^{k+1}, y^k)|\} \\ \text{(c)} & \varepsilon'_k \min\{1, |r'_k \nabla_y l_{r_k}(x^{k+1}, y^k)|^2\} \end{cases} \tag{1.16}$$

by taking $\varepsilon'_k = \varepsilon_k \sqrt{s}$. Without any conditions having been imposed so far on ε_k, it may seem that utilizing (1.16) instead of (1.15) in the case of strong variational sufficiency involves no more than passing to a different sequence of error parameters ε'_k, with explicit knowledge of s not being required. There is truth to that, but the error parameters ε_k in (1.15) enter subtly in the description that will

[8] The strong convexity inequality $l_{r_k}(x, y^k) \geq l_{r_k}(x^{k+1}, y^k) + \nabla_x l_{r_k}(x^{k+1}, y^k) \cdot (x - x^{k+1}) + \frac{s}{2}|x - x^{k+1}|^2$ leads to this by minimizing on both sides with respect to $x \in \mathcal{X}$.

be given about just how close to $(\bar{x},\bar{y})$ the algorithm needs to be initiated in order to succeed. That will need closer attention in due course.

Outline. Section 2 will establish the deep connection with the proximal point algorithm under the variational sufficient condition for local optimality. Section 3 will proceed, through strong variational sufficiency, with translating linear convergence properties of the proximal point algorithm to the augmented Lagrangian method under assumptions about the local dual problem that's implicitly present in the background. Those assumptions undergo further translation then into conditions that can be checked in terms of the problem structure explicitly available. That is the subject of Section 4.

Readers wishing to avoid all that heavy technology at first pass can skip ahead to Section 5 to see in summary, on a more immediately understandable level, what ultimately is achieved about the augmented Lagrangian method.

2 Proximal point derivation of the augmented Lagrangian method

The augmented Lagrangian method aims at determining a point $\bar{x}$ that is locally (perhaps globally) optimal in problem (P), but instead of trying to generate a sequence $\{x^k\}$ that converges to such $\bar{x}$ with objective values $\varphi(x^k,0)$ decending to $\varphi(\bar{x},0)$, it generates a sequence $\{(x^k,y^k)\}$ targeted to

$$S = \text{set of all } (\bar{x},\bar{y}) \text{ satisfying the equivalent conditions (1.8).} \tag{2.1}$$

Although the values $\varphi(x^k,0)$ might then be ∞, signaling that x^k isn't even a feasible solution to (P), it is desired at least to have *asymptotic feasibility* in the sense that

$$\exists\, u^k \to 0 \;\text{ with }\; \varphi(x^k,u^k) < \infty. \tag{2.2}$$

The goal also, of course, is that a pair $(\bar{x},\bar{y}) \in S$ obtained as the limit, or maybe a cluster point, of the sequence $\{(x^k,y^k)\}$ will have $\bar{x}$ locally optimal, and that's where a second-order condition beyond (2.1) must come in. For us, that condition will be variational sufficiency now, but later, especially in Sections 3 and 4 when working on linear convergence, it will be strong variational sufficiency.

To take advantage of the local duality coming from the variational sufficient condition through Theorem 1.1, we adopt for this section the framework of a pair of closed convex sets $\mathcal{X} \subset \mathbb{R}^n$ and $\mathcal{Y} \subset \mathbb{R}^m$ having nonempty interior combined with a value $\bar{r} > 0$ such that the augmented Lagrangian $l_{\bar{r}}(x,y)$ is convex in $x \in \mathcal{X}$ for $y \in \mathcal{Y}$ (as well as concave in $y \in \mathcal{Y}$ for $x \in \mathcal{X}$). We suppose that

$$S \cap [\operatorname{int}\mathcal{X} \times \operatorname{int}\mathcal{Y}] \neq \emptyset, \tag{2.3}$$

but we don't for the moment fix on a particular pair $(\bar{x},\bar{y})$ in the intersection. The nonemptiness of S itself can be assured by feasibility and growth or compactness conditions in (P) plus a constraint qualification, but that's not the issue here. We are taking that nonemptiness for granted and posing in (2.3) a localization of it to work from.

In the mode of convex analysis in [23] and [32, Sec. 11H], we associate with this localization to $\mathcal{X} \times \mathcal{Y}$ of the augmented Lagrangian $l_{\bar{r}}$ the primal and dual problems of optimization in (1.13), but with perturbation variables. We define

$$\begin{aligned}
&\widehat{\varphi}(x,u) = \sup_{y\in\mathcal{Y}}\{\, l_{\bar{r}}(x,y) + y\cdot u\,\} \;\text{ for }\; x \in \mathcal{X}, \qquad \widehat{\varphi}(x,u) = \infty \;\text{ for }\; x \notin \mathcal{X},\\
&\widehat{\psi}(v,y) = \inf_{x\in\mathcal{X}}\{\, l_{\bar{r}}(x,y) - v\cdot x\,\} \;\text{ for }\; y \in \mathcal{Y}, \qquad \widehat{\psi}(v,y) = -\infty \;\text{ for }\; y \notin \mathcal{Y},
\end{aligned} \tag{2.4}$$

noting that then $\widehat{\varphi}$ and $-\widehat{\psi}$ are lsc proper convex functions conjugate to each other on $\mathbb{R}^n \times \mathbb{R}^m$:

$$-\widehat{\psi}(v,y) = \widehat{\varphi}^*(v,y) = \sup_{x,u}\{\, v\cdot x + y\cdot u - \widehat{\varphi}(x,u)\}. \tag{2.5}$$

The associated local primal problem is

$$(\widehat{P}) \qquad \text{minimize } \widehat{f}(x) \text{ over } x \in \mathcal{X}, \text{ where } \widehat{f}(x) = \widehat{\varphi}(x,0) = \sup_{y\in\mathcal{Y}} l_{\bar{r}}(x,y) \text{ for } x \in \mathcal{X},$$

while the local dual problem is

$$(\widehat{D}) \qquad \text{maximize } \widehat{h}(y) \text{ over } y \in \mathcal{Y}, \text{ where } \widehat{h}(y) = \widehat{\psi}(0,y) = \inf_{x\in\mathcal{X}} l_{\bar{r}}(x,y) \text{ for } y \in \mathcal{Y}.$$

We are headed toward applying the proximal point algorithm to $(\widehat{D})$ and showing how that can result in to solving (P). Understanding the connection between $(\widehat{P})$, $(\widehat{D})$, and (P) will be vital.

Theorem 2.1 (foundation for primal-dual developments)**.** *Under (2.3), the problems $(\widehat{P})$ and $(\widehat{D})$ have optimal solutions with $\min(\widehat{P}) = \max(\widehat{D})$, and*

$$\bar{x} \text{ solves } (\widehat{P}) \quad \Longleftrightarrow \quad \bar{x} \text{ minimizes in } (P) \text{ relative to } \mathcal{X}. \tag{2.6}$$

Moreover the following conditions on a pair

$$(\bar{x},\bar{y}) \in \operatorname{int}\mathcal{X} \times \operatorname{int}\mathcal{Y} \tag{2.7}$$

are equivalent and guarantee that $\bar{x}$ is locally optimal relative to $\mathcal{X}$ in (P) with the objective value $\varphi(\bar{x},0)$ agreeing with the common optimal values in $(\widehat{P})$ and $(\widehat{D})$ as well as with $\widehat{l}(\bar{x},\bar{y})$ and $l_{\bar{r}}(\bar{x},\bar{y})$:

(a) $(\bar{x},\bar{y}) \in S$,
(b) *$\bar{x}$ minimizes in $(\widehat{P})$ and $\bar{y}$ maximizes in $(\widehat{D})$,*
(c) *$(\bar{x},\bar{y})$ is a saddle point of $\widehat{l}$ on $\mathbb{R}^n \times \mathbb{R}^m$,*
(d) *$(\bar{x},\bar{y})$ is a saddle point of $l_{\bar{r}}$ on $\mathcal{X}\times\mathcal{Y}$,*
(e) *$(\bar{x},\bar{y})$ is a saddle point of l_r on $\mathcal{X}\times\mathcal{Y}$ for every $r \geq \bar{r}$.*

Proof. Except for (2.6), this just summarizes, from the perspective of the set $\mathcal{X}\times\mathcal{Y}$ rather than one particular pair $(\bar{x},\bar{y}) \in S$, the facts about variational sufficiency in Theorem 1.1 and their implications in (1.13). In particular, the optimal solutions to $(\widehat{P})$ belonging to $\operatorname{int}\mathcal{X}$ are the points yielding a minimum in (P) over $\operatorname{int}\mathcal{X}$, which then by convexity is a minimum over all of $\mathcal{X}$. Thus, in denoting by C and D the closed convex sets of $\bar{x}$ vectors on the left and right sides of (2.6), we have $C \cap \operatorname{int}\mathcal{X} = D \cap \operatorname{int}\mathcal{X}$. The common intersection is nonempty in consequence of assumption (2.3). Since C and D are subsets of $\mathcal{X}$, the nonemptiness implies $C = \operatorname{cl}[C \cap \operatorname{int}\mathcal{X}]$ and $D = \operatorname{cl}[D \cap \operatorname{int}\mathcal{X}]$, so $C = D$ and (2.6) is correct. □

A valuable but somewhat curious consequence of Theorem 2.1 is that the intersection of S with the interior of $\mathcal{X}\times\mathcal{Y}$ is the product of a convex set of vectors $\bar{x}$ and a convex set of vectors $\bar{y}$. In nonconvex optimization, there would be no reason in general to expect product structure in S.

Theorem 2.1 reveals a potential dual approach to solving (P). If $\bar{y}$ is a solution to $(\widehat{D})$ that lies in the interior of $\mathcal{Y}$, then for any $r \geq \bar{r}$, the solutions to (P) relative to $\mathcal{X}$ that lie in the interior of $\mathcal{X}$ are among the minimizers $\bar{x}$ of $l_r(x,\bar{y})$ over $x \in \mathcal{X}$. Specifically, they are the minimizers $\bar{x}$ for which $l_r(\bar{x},\bar{y}) = \varphi(\bar{x},0)$, since that pair of conditions means that $(\bar{x},\bar{y})$ is a saddle point on $\mathcal{X}\times\mathcal{Y}$. In practice, we can only know $\bar{y}$ approximately as the limit of a sequence $\{y^k\}$ generated by some

method for solving $(\widehat{D})$. But we can combine each y^k with an x^{k+1} that approximately minimizes $l_{r_k}(x, y^k)$ over $x \in \mathcal{X}$ to generate in tandem a sequence $\{x^k\}$ that might have an optimal solution $\bar{x}$ to (P) relative to $\mathcal{X}$ as a limit or cluster point. The proximal point algorithm as a means of solving $(\widehat{D})$ will bring this scheme to fruition.

The proximal point algorithm is designed to find a zero of a maximal monotone mapping T. It iterates with resolvent mappings $P_k = (I + c_k T)^{-1}$ which are single-valued and nonexpansive. In applying it to solve $(\widehat{D})$, we take $T = \partial[-\widehat{h}]$, which is maximal monotone because $\widehat{h}$ is an upper semicontinuous concave function [32, 12.17]. Then the zeros of T, as the points $\bar{y}$ where $0 \in T(\bar{y})$, are the optimal solutions to $(\widehat{D})$. The zero set is

$$Z = \operatorname{argmax}_y \widehat{h}(y), \tag{2.8}$$

and the iterations select

$$y^{k+1} \approx P_k(y^k) \;\text{ with }\; P_k(y^k) = \operatorname{argmax}_y \Big\{ \widehat{h}^k(y) := \widehat{h}(y) - \frac{1}{2c_k}|y - y^k|^2 \Big\}. \tag{2.9}$$

The proximal parameters c_k will be taken here to satisfy

$$1 \le c_k \to c_\infty \le \infty, \tag{2.10}$$

while the approximation will be controlled at three possible levels by stopping criteria of the form

$$|y^{k+1} - P_k(y^k)| \le \begin{cases} \text{(a)} & \varepsilon_k \\ \text{(b)} & \varepsilon_k \min\{1, |y^{k+1} - y^k|\} \\ \text{(c)} & \varepsilon_k \min\{1, |y^{k+1} - y^k|^2\} \end{cases} \tag{2.11}$$

in which the error parameters ε_k satisfy

$$\varepsilon_k \in (0, 1) \;\text{ with }\; \textstyle\sum_{k=0}^{\infty} \varepsilon_k = \sigma < \infty. \tag{2.12}$$

In this section, we'll only be concerned with the basic level (a). Levels (b) and (c) will be important later in supporting a linear rate of convergence.

It's known from the proximal point theory in [25] that the size of $|y^{k+1} - P_k(y^k)|$ can be estimated from above by $c_k \operatorname{dist}(0, \partial[-\widehat{h}^k](y^{k+1}))$ for the strongly concave function $\widehat{h}^k$ in (2.9). That expression could therefore be substituted on the left of (2.11). But on our way to the augmented Lagrangian method and its stopping criteria in (1.15), a different upper estimate for $|y^{k+1} - P_k(y^k)|$ will eventually be needed instead.

It will be essential for our purposes to have $P_k(y^k)$ and y^{k+1} keep to the interior of $\mathcal{Y}$, even though the procedure, as described, is known from [25] to exhibit global convergence from any starting point y^0 in $\mathbb{R}^m$. The refined localization that will serve our needs is available from [31], as we record next.

Theorem 2.2 [25, 31] (basic proximal point convergence)**.** *Let the initial point y^0 and the value σ in (2.12) satisfy the following closeness condition relative to the closed convex set $Z = \operatorname{argmax} \widehat{h}$:*

$$\exists\, \rho > \operatorname{dist}(y^0, Z) + \sigma \;\text{ such that }\; \mathcal{Y} \supset \Big\{ y \;\Big|\; |y - y^0| < 3\rho \Big\}. \tag{2.13}$$

Then the sequence $\{y^k\}$ generated by the proximal point iterations (2.9) under (2.10), (2.11a) and (2.12) will belong to $\operatorname{int} \mathcal{Y}$ and converge to a particular point $\bar{y} \in Z$ in the ball $\Big\{ y \;\Big|\; |y - \bar{y}^0| < \rho \Big\} \subset \operatorname{int} \mathcal{Y}$,

where $\bar{y}^0$ is the point of Z closest to y^0. In the course of this, neither y^k nor $P_k(y^k)$ will ever leave that ball, and the dual objective values $\widehat{h}(y^k)$ will converge to the optimal value $\widehat{h}(\bar{y})$ in $(\widehat{D})$.

Proof. This is a simplification of [31, Theorem 2.1] based on the mapping $\partial[-\widehat{h}]$ being maximal monotone globally. (The result in its full form addresses the possibility of a subgradient mapping that is maximal monotone only locally.) The fact that y^k and $P_k(y^k)$ never leave the ball in question didn't appear in the original statement in [31] but is shown in displays in the theorem's proof. □

The formula for $P_k(y^k)$ in (2.9) isn't the only way to think of how the proximal point algorithm proceeds. Another interpretation, which will soon have a role in our developments, revolves around the functions generated by the maximization in (2.9), namely

$$\widehat{h}_{c_k}(y^k) = \max_y\Big\{\, \widehat{h}^k(y) := \widehat{h}(y) - \frac{1}{2c_k}|y-y^k|^2\Big\}. \tag{2.14}$$

Through their definition and the fact that $P_k(y^k)$ furnishes the maximum, these functions are concave and differentiable on $\mathbb{R}^m$ with

$$\nabla\widehat{h}_{c_k}(y^k) = c_k^{-1}[P_k(y^k)-y^k], \;\text{ so that }\; P_k(y^k) = y^k + c_k\nabla\widehat{h}_{c_k}(y^k). \tag{2.15}$$

Another consequence of the sup-convolution formula defining $\widehat{h}_{c_k}$ is that

$$\widehat{h}_{c_k}(y) \geq \widehat{h}_{c_k}(y^k) + \nabla\widehat{h}_{c_k}(y^k)\cdot[y-y^k] - \frac{1}{2c_k}|y-y^k|^2 \;\text{ for all }\; y\in\mathbb{R}^m, \tag{2.16}$$

from which $P_k(y^k)$ receives a further description out of (2.15),

$$P_k(y^k) = \operatorname{argmax}_y\Big\{\, \widehat{h}_{c_k}(y^k) + \nabla\widehat{h}_{c_k}(y^k)\cdot[y-y^k] - \frac{1}{2c_k}|y-y^k|^2\Big\}. \tag{2.17}$$

The guarantee in Theorem 2.2 that the sequence $\{y^k\}$ stays in $\operatorname{int}\mathcal{Y}$ will be crucial in what comes next. It enables us to concentrate on the behavior of the dual objective $\widehat{h}$ just on $\operatorname{int}\mathcal{Y}$. There is the question then of whether, in the formula $\widehat{h}(y) = \inf_{x\in\mathcal{X}} l_{\bar{r}}(x,y)$, the minimum is attained, and not just at a boundary point of $\mathcal{X}$. From the convexity of $l_{\bar{r}}(\cdot,y)$ we know that

$$\operatorname*{argmin}_{x\in\mathcal{X}} l_{\bar{r}}(x,y) = \{\, x \mid -\nabla_x l_{\bar{r}}(x,y) \in N_{\mathcal{X}}(x)\}, \tag{2.18}$$

where $N_{\mathcal{X}}(x)$ is the normal cone to the closed convex set $\mathcal{X}$ at x. We will benefit from having the sets (2.18) be nonempty and bounded, at least when $y\in\operatorname{int}\mathcal{Y}$. This certainly is the case if $\mathcal{X}$ is bounded, for instance, which could be imposed more or less harmlessly for our purposes. But it could also stem from growth properies of $l_{\bar{r}}(x,y)$ in x that might be tied to the original Lagrangian $l(x,y)$ n (1.2).

Theorem 2.3 (fundamental ALM characterization)**.** *Suppose the argmin sets (2.18) are nonempty and bounded when $y\in\operatorname{int}\mathcal{Y}$. Let the augmented Lagrangian method (1.7) with stopping criterion (1.15a), error parameters ε_k as in (2.12), $r_k\in(\bar{r},\infty)$ with $r_k\to r_\infty\in(\bar{r},\infty]$ and stepsizes $r'_k = r_k-\bar{r}$, be initiated with y^0 satisfying the prescription in Theorem 2.2 (so executability of the steps is assured). Then, by the estimate*

$$|y^{k+1}-P_k(y^k)|^2 \leq 2c_k\Big[l_{r_k}(x^{k+1},y^k) - \inf_{\mathcal{X}} l_{r_k}(\cdot,y^k)\Big] \;\text{ for }\; c_k = r_k-\bar{r}, \tag{2.19}$$

the resulting sequence $\{y^k\}$ can be interpreted as being generated by the proximal point algorithm (2.9) with $c_k = r'_k$ under the stopping criterion (2.11a) for the same error parameters ε_k. It will thus, as in Theorem 2.2, converge within $\operatorname{int}\mathcal{Y}$ to a particular solution $\bar{y}$ to $(\widehat{D})$ that lies in $\operatorname{int}\mathcal{Y}$.

On the other hand, the sequence $\{x^k\}$ in $\mathcal{X}$ will be bounded and asymptotically feasible in (P). Each of its cluster points will be a solution $\bar{x}$ to $(\widehat{P})$ furnishing also a minimum in (P) relative to $\mathcal{X}$ and thus be locally optimal in $(\widehat{P})$ if it belongs to $\operatorname{int}\mathcal{X}$.

Executing the augmented Lagrangian method with stopping criterion (1.15b) or (1.15c) instead of (1.15a) correponds in this to executing the proximal point algorithm with (2.11b) or (2.11c).

Proof. The functions $\widehat{h}_{c_k}$ and their properties in (2.15)–(2.17) in connection with the proximal point algorithm will lead the way in this. In parallel with our scheme of problems $(\widehat{P})$ and $(\widehat{D})$ associated with $\widehat{l}$, we can associate with the convex-concave function

$$\widehat{l}^k(x,y) := l_{\bar{r}}(x,y) - \frac{1}{2c_k}|y-y^k|^2 \;\text{ for }\; x\in\mathcal{X} \;\text{ and }\; y\in\mathcal{Y} \tag{2.20}$$

the primal problem

$$(\widehat{P}^k) \qquad \text{minimize over } x\in\mathcal{X} \text{ the function} \quad \sup_{y\in\mathcal{Y}} \widehat{l}^k(x,y) =: f^k(x)$$

and the dual problem

$$(\widehat{D}^k) \qquad \text{maximize over } y\in\mathcal{Y} \text{ the function} \quad \inf_{x\in\mathcal{X}} \widehat{l}^k(x,y) = \widehat{h}(y) - \frac{1}{2c_k}|y-y^k|^2,$$

which corresponds to the proximal point maximization in (2.9). Our assumption that the sets (2.18) are nonempty and bounded when $y\in\operatorname{int}\mathcal{Y}$ makes the convex functions $l_{\bar{r}}(\cdot,y)$ be level-bounded [32, 3.23], and that passes over to the functions $\widehat{l}^k(\cdot,y)$, causing the convex objective function f^k in $(\widehat{P}^k)$ to be level bounded as well.[9] The concave objective function in $(\widehat{D}^k)$ is likewise level-bounded (from below instead of from above), due to its quadratic term. Because of this, optimal solutions to both $(\widehat{P}^k)$ and $(\widehat{D}^k)$ exist, characterized by forming saddlepoints in (2.20), and the optimal values in these problems agree [32, 11.40]. We already know, of course, that $(\widehat{D}^k)$ has $P_k(y^k)$ as its unique optimal solution, and the optimal value in $(\widehat{D}^k)$ has been defined in (2.14) to be $\widehat{h}_{c_k}(y^k)$. The new information now is that

$$\begin{aligned} \exists\,\hat{x}^k\in\mathcal{X} \;\text{ such that }\; \widehat{h}_{c_k}(y^k) &= f^k(\hat{x}^k) = l^k(\hat{x}^k, P_k(y^k)) \\ &= \max_{y\in\mathcal{Y}} \widehat{l}^k(\hat{x}^k,y) = \max_{y\in\mathcal{Y}}\Big\{ l_{\bar{r}}(\hat{x}^k,y) - \frac{1}{2c_k}|y-y^k|^2\Big\}, \end{aligned} \tag{2.21}$$

where by concavity

$$\max_{y\in\mathcal{Y}}\Big\{ l_{\bar{r}}(\hat{x}^k,y) - \frac{1}{2c_k}|y-y^k|^2\Big\} = \max_{y\in\boldsymbol{R}^m}\Big\{ l_{\bar{r}}(\hat{x}^k,y) - \frac{1}{2c_k}|y-y^k|^2\Big\} \;\text{ if }\; P_k(y^k)\in\operatorname{int}\mathcal{Y}. \tag{2.22}$$

A simple relationship between the augmented Lagrangians $l_{\bar{r}}$ and l_{r_k} for $r_k=\bar{r}+c_k$ can next be brought in. It comes from the convex functions $-l_{\bar{r}}(x,\cdot)$ and $-l_{r_k}(x,\cdot)$ being conjugate to $\varphi_{\bar{r}}(x,\cdot)$ and $\varphi_{r_k}(x,\cdot)$. Because $\varphi_{\bar{r}+c_k}(x,u) = \varphi_{\bar{r}}(x,u) + \frac{c_k}{2}|u|^2$ and the addition of convex functions dualizes to inf-convolution, $-l_{\bar{r}+c_k}(x,\cdot)$ is obtained from $-l_{\bar{r}}(x,\cdot)$ by inf-convolution with the function conjugate to $\frac{c_k}{2}|u|^2$, which is $\frac{1}{2c_k}|y|^2$. Therefore

$$\max_{y\in\boldsymbol{R}^m}\Big\{ l_{\bar{r}}(x,y) - \frac{1}{2c_k}|y-y^k|^2\Big\} = l_{r_k}(x,y^k). \tag{2.23}$$

[9] In fact, to reach this conclusion it would suffice to assume the nonemptiness and boundedness of just one of the sets in (2.18) for $y\in\operatorname{int}\mathcal{Y}$.

The combination of (2.21), (2.22) and (2.23) yields

$$\exists\, \hat{x}^k \in \mathcal{X} \ \text{ such that } \ \widehat{h}_{c_k}(y^k) = l_{r_k}(\hat{x}^k, y^k) \ \text{ if } \ P_k(y^k) \in \operatorname{int} \mathcal{Y}. \tag{2.24}$$

In comparison, the definition (2.14) of $\widehat{h}_{c_k}$, in which $\widehat{h}(y) \le l_{\bar{r}}(x,y)$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, implies through (2.23) that

$$\widehat{h}_{c_k}(y) \le l_{r_k}(x,y) \ \text{ for all } \ x \in \mathcal{X} \ \text{ and } \ y \in I\!R^m. \tag{2.25}$$

In particular, we learn from this and (2.24) that

$$\widehat{h}_{c_k}(y^k) = \min_{x \in \mathcal{X}} l_{r_k}(x, y^k) \ \text{ if } \ P_k(y^k) \in \operatorname{int} \mathcal{Y}. \tag{2.26}$$

We can turn now to the vectors x^{k+1} and y^{k+1} in the augmented Lagrangian method (1.7). The concavity of $l_{r_k}(x^{k+1}, y)$ in y gives us

$$l_{r_k}(x^{k+1}, y) \le l_{r_k}(x^{k+1}, y^k) + \nabla_y l_{r_k}(x^{k+1}, y^k) \cdot (y - y^k) \ \text{ for all } \ y \in I\!R^m.$$

This can be partnered through (2.25) with the inequality in (2.16) to obtain

$$\widehat{h}_{c_k}(y^k) + \nabla \widehat{h}_{c_k}(y^k) \cdot [y - y^k] - \frac{1}{2c_k}|y - y^k|^2 \le l_{r_k}(x^{k+1}, y^k) + \nabla_y l_{r_k}(x^{k+1}, y^k) \cdot (y - y^k),$$

from which it follows that

$$l_{r_k}(x^{k+1}, y^k) - \widehat{h}_{c_k}(y^k) \ge [\nabla_y l_{r_k}(x^{k+1}, y^k) - \nabla \widehat{h}_{c_k}(y^k)] \cdot (y - y^k) - \frac{1}{2c_k}|y - y^k|^2 \ \text{ for all } \ y \in I\!R^m,$$

where, by (2.16) and the rule for obtaining y^{k+1} in (1.7),

$$\nabla_y l_{r_k}(x^{k+1}, y^k) - \nabla \widehat{h}_{c_k}(y^k) = c_k^{-1}[y^{k+1} - y^k] - c_k^{-1}[P_k(y^k) - y^k)] = c_k^{-1}[y^{k+1} - P_k(y^k)].$$

Therefore, in terms of $z = y - y^k$,

$$c_k\Big[l_{r_k}(x^{k+1}, y^k) - \widehat{h}_{c_k}(y^k)\Big] \ge \max_{z \in \boldsymbol{R}^m}\Big\{\, [y^{k+1} - P_k(y^k)] \cdot z - \tfrac{1}{2}|z|^2 \Big\} = \tfrac{1}{2}|y^{k+1} - P_k(y^k)|^2.$$

Under (2.26), this becomes the estimate claimed in (2.19).

Recalling now from Theorem 2.2 that the proximal point algorithm, when initiated as prescribed, always has $P_k(y^k) \in \operatorname{int} \mathcal{Y}$, we reach comfirmation of the statements about the sequence $\{y^k\}$ generated by the augmented Lagrangian method with stopping criterion (1.15a).

What happens in the meantime to the sequence $\{x^k\}$? The objective f^k in $(\widehat{P}^k)$ has been determined to be $\le l_{r_k}(\cdot, y^k)$, but we have also verified that

$$\min_{x \in \mathcal{X}} f^k(x) = \min_{x \in \mathcal{X}} l_{r_k}(x, y^k) = \widehat{h}_{c_k}(y^k). \tag{2.27}$$

Since x^{k+1} is chosen under the stopping criterion (1.15a) to have $l_{r_k}(x^{k+1}, y^k) - \widehat{h}_{c_k}(y^k) \le \varepsilon_k^2/2c_k$, it follows that

$$f^k(x^{k+1}) \le l_{\bar{r}}(x^{k+1}, y^k) \le \alpha_k := \widehat{h}_{c_k}(y^k) + \frac{\varepsilon_k^2}{2c_k}. \tag{2.28}$$

Here, from the definition of $\widehat{h}_{c_k}(y^k)$ in (2.14), we know $\widehat{h}(y^k) \le \widehat{h}_{c_k}(y^k) \le \max \widehat{h}$, but $\widehat{h}(y^k) \to \max \widehat{h}$ according to Theorem 2.2, hence

$$\alpha_k \to \bar{\alpha} = \max(\widehat{D}) = \min(\widehat{P}) = \min_{x \in \mathcal{X}} l_{\bar{r}}(x, \bar{y}). \tag{2.29}$$

On another front, the definition of f^k entails the lower bound $f^k(x) \geq l_{\bar{r}}(x, y^k)$, which in the limit as $y^k \to \bar{y}$ implies

$$\{\, x \in \mathcal{X} \,|\, f^k(x) \leq \alpha\} \subset \{\, x \in \mathcal{X} \,|\, l_{\bar{r}}(x, \bar{y}) \leq \alpha\} \;\text{ for all }\; \alpha \in I\!R,$$

where sets on the right are bounded under the argmin assumption in the theorem [32, 3.23]. From (2.28) we therefore have

$$x^{k+1} \in \{\, x \in \mathcal{X} \,|\, l_{\bar{r}}(x, \bar{y}) \leq \alpha\} \;\text{ for any }\quad \alpha \geq \alpha_k.$$

and can confirm through (2.29) that the sequence $\{x^k\}$ is bounded with all its cluster points belong to $\operatorname{argmin}_{\mathcal{X}} l_{\bar{r}}(\cdot, \bar{y})$.

It will be demonstrated now that the vectors x^k can be paired with vectors $u^k \to 0$ in $I\!R^m$ for which $\varphi(x^k, u^k)$ converges to the minimum in $(\widehat{P})$. This will justify the claim of asymptotic feasibility in the sense of (2.2) and also show that every cluster point $\bar{x}$ of $\{x^k\}$ is an optimal soultion to $(\widehat{P})$, inasmuch as $\widehat{\varphi}(\bar{x}, 0) \leq \liminf_k \widehat{\varphi}(x^k, u^k)$ by the lower semicontinuity of $\widehat{\varphi}$. Then, with a look back at (2.6), the proof of the theorem will be complete.

From the ALM update in (1.7) we have $\nabla_y l_{r_k}(x^{k+1}, y^k) = c_k^{-1}[y^{k+1} - y^k]$, where $y^{k+1} - y^k \to 0$ as $c_k \to c_\infty \leq \infty$. Taking $u^{k+1} = \nabla_y l_{r_k}(x^{k+1}, y^k)$, we get a sequence $\{u^k\}$ partnered with $\{x^k\}$ such that not only $u^k \to 0$ but in fact

$$c_k u^{k+1} \to 0. \tag{2.30}$$

Because the convex function $-l_{r_k}(x^{k+1}, \cdot)$ is conjugate to $\varphi_{r_k}(x^{k+1}, \cdot)$, with u^{k+1} being in particular a subgradient of the former at y^k, we also have

$$u^{k+1} \cdot y^k + l_{r_k}(x^{k+1}, y^k) = \varphi_{r_k}(x^{k+1}, u^{k+1}) = \varphi(x^{k+1}, u^{k+1}) + \frac{\bar{r} + c_k}{2} |u^{k+1}|^2,$$

where the initial and final terms tend to 0 by (2.30). We have determined via (2.28) and (2.29) that $l_{r_k}(x^{k+1}, y^k)$ tends to the optimal value in $(\widehat{P})$, and we see now that $\varphi(x^{k+1}, u^{k+1})$ does the same, as claimed.

The assertion at the end of the theorem about the stronger stopping criteria (1.15b) and (1.15c) is justified obviously by (2.19). □

Corollary 2.3.1 (convergence to an isolated local minimium)**.** *Suppose $\bar{x}$ satisfies the variational sufficient condition for local optimality in (P), but is isolated from any other such point (i.e., is unique with respect to a neighborhood). This holds in particular if a dual vector $\bar{y}$ paired with $\bar{x}$ in the Lagrangian characterization of variational sufficiency in Theorem 1.1 has $\bar{x}$ as the only minimizer of $l_{\bar{r}}(\cdot, \bar{y})$ over $\mathcal{X}$, as for instance under strong variational sufficiency. Then, in the associated framework of local duality passed from Theorem 1.1 to Theorem 2.1 and utilized in Theorem 2.3, the sequence $\{x^k\}$ generated by the augmented Lagrangian method must converge to that local solution $\bar{x}$.*

Proof. This refers to the situation in which $\bar{x}$ is unique initial component of the pairs in the set in assumption (2.3). The claim about how that can be a consequence of a minimization property of $l_{\bar{r}}(\cdot, \bar{y})$ is based on the characterization in Theorem 2.1(d). When the sequence $\{x^k\}$ in Theorem 2.3 has only one possible cluster point $\bar{x}$, it must converge to that point. □

Corollary 2.3.2 (application to the convex case)**.** *In the convex case of problem (P), the sets $\mathcal{X}$ and $\mathcal{Y}$ can be taken to be all of $I\!R^n$ and $I\!R^m$. The augmented Lagrangian method, as implemented in Theorem 2.3 under the assumptions there, can then have $\bar{r}$ arbitrarily near to 0, and it can start from*

any y^0, inasmuch as ρ can be arbitrarily large in (2.13). Moreover the locally optimal solutions $\bar{x}$ it generates will be globally optimal solutions to (P).

In asking that $\bar{r}$ be positive, even if arbitrarily close to 0, the statement in Corollary 2.3.2 falls a bit short of what might be wished in the convex case. Why can't $\bar{r}$ just be 0, so that the problems $(\widehat{P})$ and $(\widehat{D})$ reduce to (P) and its dual (D) that maximizes $h(y) = \inf_y l(x, y)$? The update from y^k to y^{k+1} could then be simplified from $r'_k = r_k - \bar{r}$ to $r'_k = r_k$. This is merely artificial trouble coming from the choices adopted for the exposition in this paper. The Lagrangian l in (1.2) can be identified as the case of the augmented Lagrangian l_r in (1.3) in which $r = 0$, but as l_0 it wouldn't have the differentiability properties in y that we have utilized so conveniently. The special treatment needed to take care of that didn't seem worth the effort here. What could be better is a separate, if parallel, development of the augmented Lagrangian method for optimization problems like (P) having convex $\varphi(x, u)$, but not necessarily limited to the generalized nonlinear programming form chosen here.

But in fact, confirmation of convergence when taking $r'_k = r_k$ instead of $r'_k = r^k - \bar{r}$ can be provided even in the nonconvex case of (P) by appealing to the relaxed version of the proximal point algorithm developed by Eckstein and Bertsekas [7]. That version differs from (2.9) under (2.11a) in having

$$y^{k+1} = (1 - \theta_k)y^k + \theta_k \widehat{y}^{k+1} \;\text{ with }\; |\widehat{y}^{k+1} - P_k(y^k)| \leq \varepsilon_k. \tag{2.31}$$

The original version corresponds to $\theta_k \equiv 1$. As translated to the maximization in Theorem 2.2, the result in [7, Theorem 3] gives global convergence of the relaxed iterations (2.31) to some point $\bar{y} \in Z = \operatorname{argmax} \widehat{h}$, as long as

$$\theta_k \in (0, 2), \qquad \limsup_{k\to\infty} \theta_k < 2, \qquad \liminf_{k\to\infty} \theta_k > 0. \tag{2.32}$$

Although Eckstein and Bertsekas didn't address the specifics of a localization to a set $\mathcal{Y}$, that topic was taken up by Pennanen. His result in [15, Proposition 6] tells us in the case of $\theta_k \geq 1$ that if

$$\exists\, \delta > 0 \;\text{ such that }\; \operatorname{dist}(y, Z) \leq \delta \implies y \in \operatorname{int} \mathcal{Y}, \tag{2.33}$$

and the procedure is initiated with y^0 close enough to Z, the sequence $\{y^k\}$ will stay inside $\mathcal{Y}$ while converging to $\bar{y}$.[10]

How would the replacement of the original proximal point algorithm by the relaxed version play out in the ALM derivation argument in the proof of Theorem 2.3? Nothing changes except that the updating expression $y^k + c_k \nabla_y l_{r_k}(x^{k+1}, y^k)$ in (1.7) now designates the approximation $\widehat{y}^{k+1}$ in (2.31) rather than y^{k+1}, which is given instead then by

$$y^{k+1} = (1 - \theta_k)y^k + \theta_k[y^k + c_k \nabla_y l_{r_k}(x^{k+1}, y^k)] = y_k + r'_k \nabla_y l_{r_k}(x^{k+1}, y^k) \;\text{ for }\; r'_k = \theta_k c_k. \tag{2.34}$$

Theorem 2.4 (relaxed stepsizes in the ALM derivation)**.** *Suppose the argmin sets (2.18) are nonempty and bounded when $y \in \operatorname{int} \mathcal{Y}$, and that (2.33) holds. Let the augmented Lagrangian algorithm (1.7) with stopping criterion (1.15a), error parameters ε_k as in (2.13), parameters $r_k \in (\bar{r}, \infty)$ with $r_k \to r_\infty \in (\bar{r}, \infty]$ and stepsizes r'_k such that*

$$1 \leq \frac{r'_k}{r_k - \bar{r}} < 2, \qquad \limsup_{k\to\infty} \frac{r'_k}{r_k - \bar{r}} < 2, \tag{2.35}$$

[10] Pennanen has an assumption in [15, Proposition 6] that is needed to get a rate of linear convergence, but that assumption is irrelevant to his proof of localization of the generated sequence.

be initiated with y^0 close enough to Z. Then the convergence properties in Theorem 2.3 will prevail along with their specialization in Corollary 2.3.1.

In particular, this applies with the stepsize r'_k taken to be r_k itself, provided that $r_k > 2\bar{r}$ and, in the limit, also $r_\infty > 2\bar{r}$. In the application to the convex case of (P) in Corollary 2.3.2, $r'_k = r_k$ suffices always, and the initial y^0 can be arbitrarily far from Z.

Proof. The discussion leading up to the statement of the theorem covers everything up to some details about r'_k. The ALM derivation in the proof of Theorem 2.3 has $r_k = \bar{r} + c_k$, so the relaxed stepsize $r'_k = \theta_k c_k$ deduced in (2.34) coresponds to $\theta_k = r'_k/(r_k - \bar{r})$. That turns the conditions on θ_k in (2.32) and Pennanen's restriction to $\theta_k \geq 1$ into the conditions in (2.35). In specializing r'_k to r_k, (2.32) reduces to requiring $r_k > 2\bar{r}$ and $r_\infty > 2\bar{r}$. In Corollary 2.3.2, $\bar{r}$ can be assigned any positive value, no matter how small, so taking r'_k to be r_k works under the stipulation that $\liminf_k r_k > 0$. The nearness condition on y^0, aimed at making sure the procedure keeps within $\mathcal{Y}$, is superfluous becuase $\mathcal{Y}$ is all of $\mathbb{R}^m$. □

Note that the condition (2.33) assumed in Theorem 2.4 trivializes when there is only one $\bar{y}$ in $Z \cap \mathcal{Y}$. That uniqueness of the multiplier vector has been a typical assumption in much of the literature on the augmented Lagrangian method.

Theorem 2.4 provides an extension of Theorem 2.3 that is attractive especially in reconciling the stepsize rule with the traditional one, but how useful is it really? In the linear convergence result of Pennanen in [15, Proposition 6] for the proximal point algorithm, getting the optimal rate when $c_k \to c_\infty < \infty$ requires $\theta_k \to 1$. But $c_k \to c_\infty < \infty$ corresponds in Theorem 2.4 to $r_k \to r_\infty < \infty$, while $r'_k = r_k$ corresponds to $\theta_k = r_k/(r_k - \bar{r})$, and $\theta_k \to r_\infty/(r_\infty - \bar{r}) < 1$. In other words, the traditional ALM stepsize is somehow *inherently incompatible* with achieving the best convergence rate unless stepsizes are forced toward ∞.

The connection with the proximal point algorithm in Theorem 2.3 involved implementing the augmented Lagrangian method with the stopping criteria in (1.15), but the alternative stopping criteria in (1.16) could well be more convenient. Can they safely be used instead? For that, the duality framework leading to Theorem 2.1 must incorporate the strong convexity in x furnished by Theorem 2.1 under strong variational sufficiency.

Theorem 2.5 (convergence with alternative stopping criteria)**.** *Let strong variational sufficiency hold, the duality framework behind Theorem 2.1 being that of Theorem 1.2 with its strong convexity modulus s, instead of just Theorem 1.1. The stopping criteria (1.15) in Theorem 2.3 and its corollaries, as well as Theorem 2.4, can be replaced then respectively by the stopping criteria (1.16) with*

$$\varepsilon'_k > 0, \qquad \sum_{k=1}^{\infty} \varepsilon'_k =: \sigma' < \infty, \tag{2.36}$$

as long as $\varepsilon'_k \leq 1/\sqrt{s}$ and the initialization condition (2.13) in Theorem 2.2 is fulfilled with $\sigma = \sigma'/\sqrt{s}$. These extra conditions on ε'_k are not needed in the convex case covered in Corollary 2.3.2.

Proof. As explained when introducing (1.16) at the end of Section 1 as a potential substitute for (1.15), the key is an estimate based on the augmented Lagrangians being strongly convex in $x \in \mathcal{X}$ with modulus s. The error parameters are related in this by $\varepsilon'_k = \varepsilon_k \sqrt{s}$. Given (2.36), the question then is whether, in taking $\varepsilon_k = \varepsilon'_k/\sqrt{s}$, the conditions on ε_k in (2.12) and (2.13) will hold. That is answered by the indicated limitations on ε'_k. □

3 Linear convergence set-up under strong variational sufficiency

From here on in this paper, the variational sufficiency we have been working with will be replaced by *strong* variational sufficiency at a level $\bar{r} > 0$. We retain the localized primal-dual framework of the preceding section, but now have a single $\bar{x}$ in our view without designating a particular accompanying $\bar{y}$. In other words, we continue with $\mathcal{X}$, $\mathcal{Y}$, $\widehat{l}(x,y)$, and the problems $(\widehat{P})$ and $(\widehat{D})$ under assumption (2.3) on the first-order set S in (2.1), with

$$\emptyset \neq S \cap [\operatorname{int}\mathcal{X} \times \operatorname{int}\mathcal{Y}] = \{\bar{x}\} \times [Z \cap \operatorname{int}\mathcal{Y}], \;\; \text{where} \;\; Z = \operatorname{argmax}_{\mathcal{Y}} \widehat{h}. \tag{3.1}$$

Because solutions to the dual problem are secondary to the goal of solving (P), it's best to avoid, as far as possible, supposing that Z is a singleton $\{\bar{y}\}$.

Strong variational sufficiency puts at our disposal the extra tools in Theorem 1.2. For $r \geq \bar{r}$, the functions $l_r(\cdot, y)$ on $\mathcal{X}$ are strongly convex with the same modulus s, and the mappings

$$A_r : (v,y) \to \operatorname*{argmin}_{x\in\mathcal{X}}\{\, l_r(x,y) - v\cdot x\} \;\; \text{on} \;\; \mathcal{V}\times\mathcal{Y} \;\; \text{for a neighborhood } \mathcal{V} \text{ of } 0 \tag{3.2}$$

are all Lipschitz continuous with the same modulus s^{-1}. This has a big effect on the dual objective function $\widehat{h} = \widehat{\psi}(0,y)$ in $(\widehat{D})$ as derived from the minimization formula for $\widehat{\psi}$ in (2.4). Because the function $(v,y) \mapsto l_{\bar{r}}(x,y) - v\cdot x$ is differentiable with gradient $(-x, \nabla_y l_{\bar{r}}(x,y))$, that formula with the minimum attained uniquely by $x = A_{\bar{r}}(v,y)$ makes the concave function $\widehat{\psi}$ be differentiable at $(v,y) \in \mathcal{V}\times\mathcal{Y}$ with its gradient being the evaluation of $(-x, \nabla_y l_{\bar{r}}(x,y))$ at $x = A_{\bar{r}}(v,y)$ and thus depending Lipschitz continuously on $(v,y) \in \mathcal{V}\times\mathcal{Y}$. Hence

$$\widehat{h} \text{ is a } \mathcal{C}^{1+} \text{ concave function on } \operatorname{int}\mathcal{Y} \text{ with } \nabla\widehat{h}(y) = \nabla_y l_{\bar{r}}(x,y) \;\; \text{for} \;\; x = A_{\bar{r}}(0,y). \tag{3.3}$$

In the maximization of $\widehat{h}$ over $\mathcal{Y}$ by the proximal point algorithm that turns into the augmented Lagrangian method, such extra properties can be beneficial in getting linear or superlinear convergence of different kinds, Q-linear and R-linear. Recall that a sequence of values $\alpha_k > 0$ converges *Q-linearly* to 0 at a rate ρ if $\limsup_k[\alpha_{k+1}/\alpha_k] \leq \rho < \infty$, this being Q-superlinear convergence if $\rho = 0$. A sequence of values $a_k \geq 0$ converges *R-linearly* to 0 at a rate ρ if $\alpha_k \leq \beta_k$ for values $\beta_k > 0$ that Q-linearly converge to 0 at that rate. Linear or superlinear convergence of a sequence of vectors w^k to a vector w refers to such convergence of the norms $|w^k - \bar{w}|$.

The main concern in the augmented Lagrangian method is ordinarily with the convergence characteristics of the primal sequence $\{x^k\}$ that it produces. Convergence characteristics of the dual sequence $\{y^k\}$ are important mostly for their impact on $\{x^k\}$. But now, in light of the minimization step in each iteration having a unique exact solution, due to strong convexity, our attention is drawn also to the sequence of vectors $\bar{x}^k$ defined by

$$\bar{x}^{k+1} := \operatorname*{argmin}_{x\in\mathcal{X}} l_{r_k}(x, y^k), \;\; \text{so that} \;\; l_{r_k}(\bar{x}^{k+1}, y^k) = \min_{x\in\mathcal{X}} l_{r_k}(x, y^k). \tag{3.4}$$

Interestingly, there are circumstances in which $\bar{x}^k$ can be guaranteed to converge R-linearly to $\bar{x}$ without a parallel guarantee of R-linear convergence for x^k.

Although the $\bar{x}^k$ sequence is only "implicit," in contrast to the x^k sequence, its rate of convergence can anyway have genuine practical significance. The algorithm can be implemented with approximate minimization while generating the vectors x^k, but when terminated at some point, the minimization (3.4) in the final iteration can be carried out with more precision to get $\bar{x}^{k+1}$ in effect as the end product. In other words, *the augmented Lagrangian method can rightly be interpreted as terminating in iteration k, not just with x^{k+1}, but effectively also $\bar{x}^{k+1}$* from just a bit of final push. This observation affects how the attainment of linear convergence is assessed.

Theorem 3.1 (ALM primal convergence from ALM dual convergence)**.**

(a) *Under strong variational sufficiency, the convergence $y^k \to \bar{y} \in Z$ in the augmented Lagrangian method (1.7), as implemented in Theorem 2.3 (with the stopping criterion (1.15a) potentially replaced by (1.16a) on the basis of Theorem 2.5), induces both $x^k \to \bar{x}$ and $\bar{x}^k \to \bar{x}$. It entails that $\nabla_x l_{r_k}(\bar{x}^{k+1}, y^k) = 0$ for k beyond some $\bar{k}$, along with*

$$\frac{s}{2}|x^{k+1} - \bar{x}^{k+1}| \le l_{r_k}(x^{k+1}, y^k) - l_{r_k}(\bar{x}^{k+1}, y^k) \le \frac{1}{2s}|\nabla_x l_{r_k}(x^{k+1}, y^k)|^2. \tag{3.5}$$

(b) *If $\operatorname{dist}(y^k, Z) \to 0$ Q-linearly at a rate ρ as $y^k \to \bar{y}$, then $\bar{x}^k \to \bar{x}$ R-linearly at that rate.*

(c) *If $y^k \to \bar{y}$ Q-linearly at a rate ρ, then $x^k \to \bar{x}$ R-linearly at that rate — as long as the stopping criterion in approximate minimization is supplemented by the proviso that*

$$|\nabla_x l_{r_k}(x^{k+1}, y^k)| \le c|y^{k+1} - y^k| \quad \text{for some fixed } c. \tag{3.6}$$

Proof. In (a) we know from Corollary 2.3.1 that $x^k \to \bar{x}$. However, that doesn't immediately tell us that $\bar{x}^{k+1} \to \bar{x}$, or even that eventually $\bar{x}^{k+1} \in \operatorname{int} \mathcal{X}$. Because $\bar{x}^{k+1}$ minimizes $l_{r_k}(\cdot, y^k)$ over $\mathcal{X}$, we know that $\nabla_x l_{r_k}(\bar{x}^{k+1}, y^k)\cdot(\bar{x} - \bar{x}^{k+1}) \ge 0$. On the other hand, the strong convexity of $l_{r_k}(\cdot, y^k)$ yields

$$l_{r_k}(\bar{x}, y^k) \ge l_{r_k}(\bar{x}^{k+1}, y^k) + \nabla_x l_{r_k}(\bar{x}^{k+1}, y^k)\cdot(\bar{x} - \bar{x}^{k+1}) + \frac{s}{2}|\bar{x} - \bar{x}^{k+1}|^2,$$

where also $l_{r_k}(\bar{x}, y^k) \ge l_{r_k}(\bar{x}, \bar{y}) = l_{\bar{r}}(\bar{x}, \bar{y}$. Therefore,

$$l_{\bar{r}}(\bar{x}, \bar{y}) \ge l_{r_k}(\bar{x}^{k+1}, y^k) + \frac{s}{2}|\bar{x} - \bar{x}^{k+1}|^2. \tag{3.7}$$

In the proof of Theorem 2.3, the values $\widehat{h}_{c_k}(y^k)$ in (2.26), identifiable now as $l_{r_k}(\bar{x}^{k+1}, y^k)$, were shown through (2.27) to converge to the value in (2.28), which is in turn identifiable now as $l_{\bar{r}}(\bar{x}, \bar{y})$. Thus $l_{r_k}(\bar{x}^{k+1}, y^k) \to l_{\bar{r}}(\bar{x}, \bar{y})$ in (3.7), and this confirms that $\bar{x}^{k+1} \to \bar{x}$.

Eventually then, $\bar{x}^{k+1}$ must belong to $\operatorname{int} \mathcal{X}$ and have $\nabla_x l_{r_k}(\bar{x}^{k+1}, y^k) = 0$ as the condition for it to give the minimum. The left side of (3.5) follows from that and the strong convexity. For the right side of (3.5), we again make use of strong convexity to see that

$$l_{r_k}(x, y^k) - l_{r_k}(x^{k+1}, y^k) \ge \nabla_x l_{r_k}(x^{k+1}, y^k)\cdot(x - x^{k+1}) + \frac{s}{2}|x - x^{k+1}|^2 \ \text{ for all } \ x \in \mathcal{X}.$$

By taking the minimum over $x \in \mathcal{X}$ on the left and the minimum over $x \in I\!R^n$ on the right, we get

$$l_{r_k}(\bar{x}, y^k) - l_{r_k}(x^{k+1}, y^k) \ge \min_{\xi \in \boldsymbol{R}^n} \Big\{ \nabla_x l_{r_k}(x^{k+1}, y^k)\cdot\xi + \frac{s}{2}|\xi|^2 \Big\} = \frac{1}{2s}|\nabla_x l_{r_k}(x^{k+1}, y^k)|^2,$$

as claimed.

For part (b), we use the fact that, when y^k is close enough to $\bar{y}$, its projection $\bar{y}^k$ on Z gives $\operatorname{dist}(y^k, Z) = |y^k - \bar{y}^k|$. Because $\bar{x}^{k+1}$ minimizes $l_{r_k}(\cdot, y^k)$ on $\mathcal{X}$ whereas $\bar{x}$ minimizes $l_{r_k}(\cdot, \bar{y}^k)$ on $\mathcal{X}$, the augmented tilt stability property in Theorem 1.2 gives us $|\bar{x}^{k+1} - \bar{x}| \le s^{-1}|y^k - \bar{y}^k| = s^{-1}\operatorname{dist}(y^k, Z)$. Thus, if $\operatorname{dist}(y^k, Z) \to 0$ Q-linearly at a rate ρ, then $|\bar{x}^{k+1} - \bar{x}| \to 0$ R-linearly at the rate ρ.

Augmented tilt stability acts similarly for part (c). We have

$$x^{k+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \{ l_{r_k}(x, y^k) - v^k\cdot x \} \ \text{ for } \ v^k = \nabla_x l_{r_k}(x^{k+1}, y^k),$$
$$\text{whereas} \quad \bar{x} = \operatorname*{argmin}_{x \in \mathcal{X}} \{ l_{r_k}(x, \bar{y}) - \bar{v}\cdot x \} \ \text{ for } \ \bar{v} = 0,$$

and therefore $|x^{k+1} - \bar{x}| \leq s^{-1}|(v^k, y^k) - (\bar{v}, \bar{y})|$. Bringing in (3.6) we get

$$s^2|x^{k+1} - \bar{x}|^2 \leq |v^k|^2 + |y^k - \bar{y}|^2 \leq c^2|y^{k+1} - y^k|^2 + |y^k - \bar{y}|^2,$$

where $|y^{k+1} - y^k| \leq |y^{k+1} - \bar{y}| + |y^k - \bar{y}| \leq (\theta_k + 1)|y^k - \bar{y}|$ for $\theta_k := |y^{k+1} - \bar{y}|/|y^k - \bar{y}|$, hence

$$s^2|x^{k+1} - \bar{x}|^2 \leq (c^2(1 + \theta_k)^2 + 1)|y^k - \bar{y}|^2. \tag{3.8}$$

The assumption that y^k converges to $\bar{y}$ at the Q-linear rate ρ means $\limsup_k \theta_k \leq \rho$. Thus, (3.8) implies the existence of a bound $|x^{k+1} - \bar{x}| \leq b|y^k - \bar{y}|$ for some b. This confirms that $x^{k+1} \to \bar{x}$ at the R-linear rate ρ. □

The dual ALM sequence $\{y^k\}$ comes from applying the proximal point algorithm in the maximization of $\widehat{h}$. We can therefore make use of convergence properies of that algorithm (as translated from the customary format of minimizing a convex function to that of maximizing a concave function). The original linear convergence result for the proximal point algorithm in minimization, in [25] with the stopping criterion (2.11b) invoked instead of (2.11a), depended on having a unique maximizer $\bar{y}$. It deduced a rate of Q-linear convergence out of assuming a quadratic growth condition on $\widehat{h}$ at $\bar{y}$. This is unsatisfying for our purposes, which include enabling nonuniqueness of solutions to $(\widehat{D})$. Luque [14] developed an alternative to the result in [25] that gets around uniquenss by assuming $\widehat{h}(y) \leq \widehat{h}(\bar{y}) - b\,\mathrm{dist}^2(y, Z)$ for some $b > 0$ when $\mathrm{dist}^2(y, Z) < \delta$. But that growth condition can be problematical when Z is unbounded, and it only guarantees Q-linear convergence of $\mathrm{dist}(y^k, Z)$ to 0, not that of $|y^k - \bar{y}|$ to 0. This could still be of value to us through part (b) of Theorem 3.1, but there has been a recent advance. We showed in [31] that Luque's growth condition doesn't need to apply to all of Z. In the context of the proximal point algorithm in Theorem 2.2, it only needs to hold for y in a neighborhood of $\bar{y}$. We showed moreover that Q-linear convergence of y^k to $\bar{y}$ can be obtained by employing the tighter stopping criterion (2.11c).

That result will be recalled precisely in Theorem 3.2 below, along with a criterion for the underlying growth condition in terms of generalized second derivatives. The derivatives in question are usually articulated in terms of epigraphs of difference quotient functions, but here we are dealing with a concave function $\widehat{h}$, for which hypographs would be appropriate instead. Rather than entering that parallel universe, we can pass to the convex function $-\widehat{h}$. But because the concept will be utilized later for convex functions other than just $-\widehat{h}$, we'll pose the definition neutrally in terms of a closed proper convex function k on $\mathbb{R}^m$. The *second-order difference quotient* functions $\Delta_\tau^2 k(y\,|\,u)$ associated with having $u \in \partial k(y)$ take the form

$$\Delta_\tau k(y\,|\,u)(\eta) = \Big[k(y + \tau\eta) - k(y) - \tau\eta\cdot u\Big]\Big/\tfrac{1}{2}\tau^2 \;\text{ for }\; \tau > 0, \tag{3.9}$$

and the corresponding *second subderivative* function is given by

$$d^2 k(y\,|\,u)(\eta) = \liminf_{\substack{\eta' \to \eta \\ \tau \searrow 0}} \Delta_\tau k(y\,|\,u)(\eta'). \tag{3.10}$$

Much more about these concepts will enter the discussion in the lead-up to Theorem 4.3 in the next section. For the moment we are focused on $k = -\widehat{h}$, $y = \bar{y}$ and $u = 0 = \nabla\widehat{h}(\bar{y})$.

Theorem 3.2 [31] (linear convergence of the proximal point algorithm in maximization)**.**

(a) *In the circumstances of Theorem 2.2 with stopping criterion (1.15a) strengthened to (1.15b) (or (1.16b) on the basis of Theorem 2.5) to get $y^k \to \bar{y} \in Z = \mathrm{argmax}_Y \widehat{h}$, suppose*

$$\exists\, b > 0,\ \lambda > 0,\ \textit{ such that }\ \widehat{h}(y) \leq [\max_Y \widehat{h}] - b\,\mathrm{dist}^2(y, Z)\ \textit{ when }\ |y - \bar{y}| < \lambda. \tag{3.11}$$

Then $\mathrm{dist}(y^k, Z) \to 0$ *at the Q-linear rate* $\rho = 1/\sqrt{1+b^2c_\infty^2}$*, which is 0 when* $c_\infty = \infty$

(b) *If the still tighter stopping criterion (1.15c) is used (or (1.16c) on the basis of Theorem 2.5), then* $y^k \to \bar{y}$ *at that Q-linear rate* ρ*. Moreover the growth condition (3.11) can be replaced by the weaker condition, in terms of second subderivatives and the normal cone* $N_Z(\bar{y})$*, that*

$$0 < b \leq \min\Big\{ \tfrac{1}{2} d^2[-\widehat{h}](\bar{y}\,|\,0)(\eta) \;\Big|\; \eta \in N_Z(\bar{y}),\ |\eta| = 1 \Big\}, \tag{3.12}$$

this minimum being positive as long as

$$d^2[-\widehat{h}](\bar{y}\,|\,0)(\eta) > 0 \ \text{ for all nonzero } \ \eta \in N_Z(\bar{y}). \tag{3.13}$$

That holds in particular if only the vectors η *in the tangent cone* $T_Z(\bar{y})$ *have* $d^2[-\widehat{h}](\bar{y}\,|\,0)(\eta) = 0$*.*

(c) *If the growth condition (3.11) in (a) holds at a boundary point* $\bar{y}_0$ *of* $Z \cap \operatorname{int} \mathcal{Y}$*, then for any boundary* $\bar{y}'$ *with* $|\bar{y}' - \bar{y}| < \lambda$*, it holds at* $\bar{y}$ *with* λ *replaced by* $\lambda' = \lambda - |\bar{y}' - \bar{y}|$*.*

Proof. Part (a) is taken from [31, Theorem 3.2], whereas part (b) rests on [31, Theorem 3.3]. The b here is $1/a$ there. Part (c) is an elementary observation about (3.12) that deserves to be recorded. □

The observation in (c) of Theorem 3.2 is valuable in the light of Theorem 2.2 and the uncertainty about where the sequence generated by the proximal point algorithm will end up when the solution isn't unique. If the procedure is initiated at a point y^0 in the circumstances specified, it will end up at a point $\bar{y}$ within distance ρ from the projection $\bar{y}^0$ of y^0 on Z. Quadratic growth of $\widehat{h}$ out of Z around $\bar{y}^0$ can therefore be inherited by $\bar{y}$, if ρ is small enough, and then linear convergence to $\bar{y}$ is achieved.

In the convex case of (P), where the unaugmented Lagrangian l is concave-convex on $\mathbb{R}^n \times \mathbb{R}^m$ with global saddle point at $(\bar{x}, \bar{y})$ and the algorithm specializes as in Corollary 2.3.2, there is a possible boost toward verifying the conditions in Theorem 3.2. They can be tested on an underlying concave function h which might be determined explicitly in some situations. This h is the objective function of the dual problem (D) associated with the convex case of (P) in Section 1.

Thorem 3.3 (simplification in the convex case)**.** *For the convex case of* (P)*, the concave function* $\widehat{h}$ *can essentially be replaced in the conditions in (3.11), (3.12), (3.13), by the concave function* $h(y) = \inf_y l(x,y)$*. Specifically, if (3.11) holds for* h *with a value* b_0*, then it holds for* $\widehat{h}$ *for* $b = b_0/(1+\bar{r}b_0)$*, and likewise in (3.12), and this is true for* $\bar{r}$ *arbitrarily close to 0.*

Proof. In this case, where $\mathcal{X} \times \mathcal{Y}$ can be taken to be all of $\mathbb{R}^n \times \mathbb{R}^m$ as explained at the end of Section 2, we have

$$\begin{aligned} \widehat{h}(y) &= \min_x l_{\bar{r}}(x,y) = \min_x \max_{y'} \Big\{ l(x,y') - \tfrac{1}{2\bar{r}}|y'-y|^2 \Big\} \\ &= \max_{y'} \inf_x \Big\{ l(x,y') - \tfrac{1}{2\bar{r}}|y'-y|^2 \Big\} = \max_{y'} \{ h(y') - \tfrac{1}{2\bar{r}}|y'-y|^2 \}, \end{aligned} \tag{3.14}$$

where min max = max inf because the bracketed expression is now convex in x as well as strongly concave in y'.[11] In particular, this relationship between $\widehat{h}$ and h implies that they have the same max value μ and the same argmax set Z containing $\bar{y}$. To simplify, we can harmlessly suppose in what follows that $\mu = 0$ and $\bar{y} = 0$.

In terms of the self-conjugate function $j(y) = \frac{1}{2}|y|^2$, (3.14) says $-\widehat{h} = (-h) \mathbin{\#} \bar{r}^{-1} j$, where $\#$ denotes infimal convolution. The mapping from y to the unique y' giving the max at the end of (3.14) is the prox mapping $(I + \bar{r}\partial[-h])^{-1}$, which is single-valued and nonexpansive with the elements of Z, in particular our $\bar{y} = 0$, as its fixed points [32, 12.12+12.17]. Hence for any $\lambda > 0$,

$$y \in \lambda \mathbb{B} \ \implies \ \exists\, y' \in \lambda \mathbb{B} \ \text{ with } \ \widehat{h}(y) = h(y') - \frac{1}{2\bar{r}}|y'-y|^2, \tag{3.15}$$

[11] This is a special case of the minimax rule in [18, Theorem 37.3(b)].

where $\mathbb{B}$ denotes the closed unit ball. Suppose (3.11) holds for h and the value b_0. This can be expressed under our simplification as $-h \geq \delta_Z \,\#\, 2b_0 j$ on $\lambda\mathbb{B}$. Then, for any $y \in \lambda\mathbb{B}$, we have from (3.15) the existence of $y' \in \lambda\mathbb{B}$ such that

$$-\widehat{h}(y) \geq [\delta_Z \,\#\, 2b_0 j](y') + \bar{r}^{-1} j(y' - y) \geq (\,[\delta_Z \,\#\, 2b_0 j] \,\#\, \bar{r}^{-1} j)(y). \tag{3.16}$$

But the operation $\#$, being dual to addition, is commutative and associative, with $\alpha^{-1} j \,\#\, \beta^{-1} j = (\alpha + \beta)^{-1} j$ (as seen through conjugacy). Therefore the right side of (3.16) is $\delta_Z \,\#\, ((2b_0^{-1} + \bar{r})^{-1} j(y)$, which is $b\,\mathrm{dist}^2(y, Z)$ for the indicated choice of b.

The claim about (3.12) and (3.13) is based on identifying the function $\widehat{k} := \frac{1}{2} d^2[-\widehat{h}](\bar{y}\,|\,0)$ with the function $k \,\#\, \bar{r}^{-1} j$ for $k := \frac{1}{2} d^2[-h](\bar{y}\,|\,0)$, since that leads in parallel to b_0 for h in (3.12) being replaced by the indicated b for $\widehat{h}$. In our simplified setting, $\Delta_\tau^2[-h](\bar{y}\,|\,0) = k_\tau$ in the notation $k_\tau(\eta) = \tau^{-2}[-h](\tau\eta)$, so that

$$k(\eta) = \liminf_{\substack{\eta' \to \eta \\ \tau \searrow 0}} k_\tau(\eta'), \quad \text{or equivalently,} \quad \mathrm{epi}\, k = \limsup_{\tau \searrow 0} \mathrm{epi}\, k_\tau.$$

By the cluster description of outer limits in [32, 4.18], this corresponds to

$$\mathrm{epi}\, k = \bigcup \Big\{ \,\mathrm{epi}\, k_0 \,\Big|\, \exists \tau_k \searrow 0 \;\text{ with }\; \mathrm{epi}\, k_{\tau_k} \to \mathrm{epi}\, k_0 \Big\}, \tag{3.17}$$

where $\mathrm{epi}\, k_{\tau_k} \to \mathrm{epi}\, k_0$ means that k_{τ_k} epi-converges to k and implies that k_0 is convex [32, 7B+4.15]. In the same way, in terms of $\widehat{k}_\tau(\eta) = \tau^{-2}[-\widehat{h}](\tau\eta)$, we have

$$\mathrm{epi}\, \widehat{k} = \bigcup \Big\{ \,\mathrm{epi}\, \widehat{k}_0 \,\Big|\, \exists \tau_k \searrow 0 \;\text{ with }\; \mathrm{epi}\, \widehat{k}_{\tau_k} \to \mathrm{epi}\, \widehat{k}_0 \Big\}. \tag{3.18}$$

Moreover, direct calculation reveals that

$$\widehat{k}_\tau = k_\tau \,\#\, [\bar{r}^{-1} j], \;\text{ or dually, }\; \widehat{k}_\tau^* = k_\tau^* + \bar{r} j. \tag{3.19}$$

Confirming that $\widehat{k} = k \,\#\, \bar{r}^{-1} j$ amounts to confirming geometrically that $\mathrm{epi}\, \widehat{k} = \mathrm{epi}\, k + \mathrm{epi}[\bar{r}^{-1} f]$, where the latter, by (3.16), is the union of all sets $\mathrm{epi}\, k_0 + \mathrm{epi}[\bar{r}^{-1} j]$ such that a sequence of functions k_{τ_k} with $\tau_k \searrow 0$ epi-converges to k_0. In view of (3.18), we can do that by demonstrating that the functions $\widehat{k}_0$ occuring there are the functions of the form $k_0 \,\#\, [\bar{r}^{-1} j]$ for k_0 obtainable as one of the epi-limits in (3.17). For this we can rely on (3.19) and the fact that epi-convergence is preserved in passing to conjugates [32, 11.34]. The conclusion follows this way because epi-convergence of $k_{\tau_k}^* + \bar{r} j$ to some $\widehat{k}_0$ is equivalent to epi-convergence of $k_{\tau_k}^*$ to some k_0 [32, 7.8(a)]. □

4 Model support for confirming linear convergence

The obvious challenge in confirming linear convergence of the augmented Lagrangian method by applying Theorems 3.2 to Theorem 3.1 is the obscurity of the dual objective function $\widehat{h}$. Aside from circumstances in the convex case of (P) in which Theorem 3.3 might yield a convenient alternative function h, which will be illustrated at the end of Section 5, we only have at our disposal conditions like (3.11), (3.12) and (3.13). How can such conditions be understood as induced from verifiable properties of problem (P) itself? Conditions on the model function g can have a powerful role in answering this question.

As a first step toward demonstrating that, we look at a more general version of the quadratic growth property in (3.11) with the aim of eventually invoking it for g^* at $\bar{y}$ with respect to $F(\bar{x}) \in \partial g(\bar{y})$. To maintain versatility, because the facts to be reported will have applications beyond just that one situation, we pass to a general notational formulation in terms of a conjugate pair of functions f and f^* on $\mathbb{R}^m$, where in particular f^* might stand for $-\widehat{h}$. Suppose $\bar{u} \in \partial f^*(\bar{y})$, or equivalently $\bar{y} \in \partial f(\bar{u})$, so that

$$f^*(y) \geq f^*(\bar{y}) + \bar{u}\cdot(y - \bar{y}), \;\; \text{with equality} \iff y \in \partial f(\bar{u}), \tag{4.1}$$

and consider the property

$$\exists\, b > 0,\, \lambda > 0 \;\text{ such that }\; |y - \bar{y}| \leq \lambda \implies f^*(y) \,\geq\, f^*(\bar{y}) + \bar{u}\cdot(y - \bar{y}) + b\,\mathrm{dist}^2(y, \partial f(\bar{u})). \tag{4.2}$$

Aragon and Geoffroy have shown in [1] that this local growth condition on f is equivalent to a property of ∂f called *calmness* at $\bar{u}$ with respect to the subgradient $\bar{y} \in \partial f(\bar{u})$ [32, Sec. 9I]. With $\mathbb{B}$ denoting the closed unit ball (in any Euclidean space at hand), so that adding $\varepsilon\mathbb{B}$ to a point or cloed set creates a "closed ε-neighborhood around it, the property in question can be expressed as

$$\exists\, a > 0, \delta > 0,\, \lambda > 0, \;\text{ such that }\; u \in [\bar{u} + \delta\mathbb{B}] \implies \partial f(u) \cap [\bar{y} + \lambda\mathbb{B}] \subset \partial f(\bar{u}) + a|u - \bar{u}|\mathbb{B}. \tag{4.3}$$

The sufficient criterion for the latter that we develop next will provide a valuable handle on verifying the quadratic growth in (4.2).

Theorem 4.1 (a sufficient condition for subdifferential calmness)**.** *For a closed proper convex f having $\bar{y} \in \partial f(\bar{u})$, the following circumstances, in particular, guarantee that ∂f is calm at $\bar{u}$ for $\bar{y}$ in the sense of (4.3). On some neighborhood of $\bar{u}$, there is a composite representation $f(u) = \gamma(\Gamma(u))$ for a $\mathcal{C}^2$ mapping Γ and a closed proper convex function γ under the constraint qualification*

$$\zeta \in N_{\mathrm{cl\,dom}\,\gamma}(\Gamma(\bar{u})), \quad \nabla\Gamma(\bar{u})^*\zeta = 0 \implies \zeta = 0 \tag{4.4}$$

and the assumption that $\partial\gamma$ is calm at $\bar{w} = \Gamma(\bar{u})$ in the sense that

$$\exists\, \kappa \geq 0 \;\; \textit{that} \;\; \partial\gamma(w) \subset \partial\gamma(\bar{w}) + \kappa|w - \bar{w}|\mathbb{B} \;\; \textit{for } w \textit{ near } \bar{w}. \tag{4.5}$$

Proof. The constraint qualification (4.4) activates the chain rule for subgradients in [32, 10.6], here in the "regular" case, since γ is convex. That gives us, for u near $\bar{u}$, the formula

$$\partial f(u) = \nabla\Gamma(u)^*\partial\gamma(\Gamma(u)) = \Big\{\, y \,\Big|\, \exists\, z \in \partial\gamma(\Gamma(u)) \;\text{ with }\; \nabla\Gamma(u)^*z = y \Big\}. \tag{4.6}$$

Let $Z(u, y)$ denote the set of z corresponding to a pair (u, y) in this formula. We assert that

$$\exists\, \rho \;\text{ such that }\; Z(u, y) \subset \rho\mathbb{B} \;\text{ when } (u, y) \text{ is near to } (\bar{u}, \bar{y}) \text{ in } \mathrm{gph}\,\partial f. \tag{4.7}$$

Otherwise there would exist $(u^k, y^k) \to (\bar{u}, \bar{y})$ in $\mathrm{gph}\,\partial f$ with $z^k \in Z(u^k, y^k)$ having $|z^k| \to \infty$. Then for $\varepsilon_k = 1/|z^k| \to 0$ and $\zeta^k = \varepsilon_k z^k$ with $|\zeta^k| = 1$ we would have $\zeta^k \in \partial(\varepsilon_k\gamma)(\Gamma(u^k))$ and $\nabla\Gamma(u^k)^*\zeta^k = \varepsilon_k y^k \to 0$. Passing to a subsequence if necessary, we can suppose that ζ^k converges to some $\bar{\zeta}$ with $|\bar{\zeta}| = 1$ and $\nabla\Gamma(\bar{u})^*\bar{\zeta} = 0$. The functions $\varepsilon_k\gamma$ epi-converge to the indicator of $\mathrm{cl\,dom}\,\gamma$ as $\varepsilon_k \to 0$ [32, 7.3], and their subdifferential mappings $\partial(\varepsilon\gamma)$ therefore converge graphically by Attouch's Theorem [32, 12.35] to the subdifferential of that indicator, which is the normal cone mapping $N_{\mathrm{cl\,dom}\,\gamma}$. Then from $(\Gamma(u^k), \zeta^k) \in \mathrm{gph}\,\partial(\varepsilon_k\gamma)$ with $(\Gamma(u^k), \zeta^k) \to (\Gamma(\bar{u}), \bar{\zeta})$ we have $(\Gamma(\bar{u}), \bar{\zeta}^k) \in \mathrm{gph}\,N_{\mathrm{cl\,dom}\,\gamma}$. But this constitutes for $z = \bar{\zeta}$ a violation of the constraint qualification (4.4).

Therefore (4.7) is correct, so that for δ and λ chosen small enough, we have

$$|u-\bar{u}| \le \delta \implies \partial f(u) \cap [\bar{y} + \lambda I\!B] = \nabla\Gamma(u)^*[\rho I\!B \cap \partial\gamma(\Gamma(u))] \tag{4.8}$$

and can estimate further that

$$\begin{aligned}\nabla\Gamma(u)^*\Big[\rho I\!B \cap \partial\gamma(\Gamma(u))\Big] &\subset \nabla\Gamma(\bar{u})^*\Big[\rho I\!B \cap \partial\gamma(\Gamma(u))\Big] + \Big[\nabla\Gamma(u)^* - \nabla\Gamma(\bar{u})^*\Big]\Big[\rho I\!B \cap \partial\gamma(\Gamma(u))\Big] \\ &\subset \nabla\Gamma(\bar{u})^*\partial\gamma(\Gamma(u)) \;+\; [\nabla\Gamma(u)^* - \nabla\Gamma(\bar{u})^*][\rho I\!B],\end{aligned} \tag{4.9}$$

Because Γ is $\mathcal{C}^2$, there exist $\alpha > 0$ and $\beta > 0$ such that

$$|\Gamma(u) - \Gamma(\bar{u})| \le \alpha|u-\bar{u}| \;\text{ and }\; ||\nabla\Gamma(u) - \nabla\Gamma(\bar{u})|| \le \beta|u-\bar{u}| \;\text{ when }\; |u-\bar{u}| \le \delta.$$

Then at the end of (4.9) we have $[\nabla\Gamma(u)^* - \nabla\Gamma(\bar{u})][\rho I\!B] \subset \beta\rho|u-\bar{u}|I\!B$, and on the other hand, by the calmness assumed in (4.5), $\partial\gamma(\Gamma(u)) \subset \partial\gamma(\Gamma(\bar{u})) + \alpha|u-\bar{u}|I\!B$. That lets us to propagate the combination of (4.8) and (4.9) into

$$\begin{aligned}|u-\bar{u}| \le \delta \implies \partial f(u) \cap [\bar{y} + \lambda I\!B] &\subset \nabla\Gamma(\bar{u})^*\Big[\partial\gamma(\Gamma(\bar{u})) + \alpha\kappa|u-\bar{u}|I\!B\Big] \\ &\subset \nabla\Gamma(\bar{u})^*\partial\gamma(\Gamma(\bar{u})) + ||\nabla\Gamma(\bar{u})||\alpha\kappa|u-\bar{u}|I\!B + \beta\rho|u-\bar{u}|I\!B \\ &= \partial f(\bar{u}) + a|u-\bar{u}|I\!B \;\text{ for }\; a = \beta\rho + \alpha\kappa||\nabla\Gamma(\bar{u})||.\end{aligned}$$

This confirms the property in (4.3) that was our goal. □

The calmness property on $\partial\gamma$ in (4.5) is stronger than the one on ∂f in (4.3) in being localized only in the domain and not around a pair in the graph. It is known to hold in particular for set-valued mappings that are *piecewise polyhedral* in having a graph that is the union of finitely many polyhedral convex sets [32, 9.57]. For mappings that are the subdifferentials of convex functions, that property corresponds precisely to the function being *piecewise linear-quadratic* in the sense that its domain is the union of finitely many polyhedral convex sets, on each of which it is given by a polynomial of degree no more than 2 [32, Sec. 10E].

A function f that has a local representation at $\bar{u}$ as described in Theorem 4.1 in which γ is piecewise linear-quadratic is, by definition, *fully amenable* at $\bar{u}$ [32, 10F]. Theorem 4.1 therefore covers fully amenable convex functions and brings to view a rich class of examples of functions having the equivalent properties in (4.2) and (4.3).

Corollary 4.1.1 (dual quadratic growth from full amenability)**.** *If $\bar{y} \in \partial f(\bar{u})$ for a convex function f on $I\!R^m$ that is fully amenable at $\bar{u}$, then the mapping ∂f is calm at $\bar{u}$ for $\bar{y}$ as in (4.3), and the conjugate function f^* has the local quadratic growth property in (4.2).*

With these results about quadratic growth and subdifferential calmness in hand, we can proceed to establish a means of verifying the growth assumption in Theorem 3.2(a) in terms of a property of the model function g.

Theorem 4.2 (criterion for the growth condition)**.** *Express Z as $G \cap M$ for $G = \partial g(F(\bar{x}))$ and $M = \{\, y \,|\, 0 = \nabla_x L(\bar{x}, y) = \nabla f_0(\bar{x}) + \nabla F(\bar{x})^* y\}$, noting that $G = \{\, y \,|\, g^*(y) = g^*(\bar{y}) + F(\bar{x})\cdot(y-\bar{y})\}$. Suppose that G is polyhedral, and*

$$\begin{aligned}&\exists\, b_0 > 0,\ \lambda_0 > 0 \text{ such that, when } |y-\bar{y}| < \lambda_0, \\ &\qquad g^*(y) \;\ge\; g^*(\bar{y}) + \nabla F(\bar{x})^*(y-\bar{y}) + b_0\,\mathrm{dist}^2(y, G).\end{aligned} \tag{4.10}$$

Assume $\nabla F(\bar z) \neq 0$ to avoid the degenerate case where the affine set M is all of $\mathbb{R}^m$, and with respect to $M^\perp$, the subpace orthogonal to M, let

$$\beta(\nabla F(\bar x)) = \min\Big\{\, |\nabla F(\bar x)^*\eta| \,\Big|\, \eta \in M^\perp,\ |\eta| = 1\Big\}. \tag{4.11}$$

Then $\beta(\nabla F(\bar x)) > 0$ and there exists $\kappa_{G,M} > 0$ such that condition (3.11) in Theorem 3.2(a) holds for

$$b = \frac{\kappa_{G,M}}{a_G + a_M} \ \text{ with } \ a_G = b_0^{-1} + 2\bar r \ \text{ and } \ a_M = \frac{2||\nabla^2_{xx} L(\bar x, \bar y) + \bar r I||}{\beta(\nabla F(\bar x))^2}. \tag{4.12}$$

Proof. Let $\mu = \max_{\mathcal{Y}} \widehat{h}$, hence $\mu = l(\bar x, \bar y) = f_0(\bar x) + g(F(\bar x))$ as well, through Theorem 2.1. We want to identify b such that $\mu - \hat h(y)$ is bounded below by $b\,\mathrm{dist}^2(y, Z)$ on a neighborhood of $\bar y$.

The functions $d_G^2(y) := \mathrm{dist}^2(y, G)$ and $d_M^2(y) := \mathrm{dist}^2(y, M)$ will have a crucial role to play. Our assumption that G is polyhedral makes d_G^2 be piecewise linear-quadratic [32, 11.28] as well as $\mathcal{C}^{1+}$. On the other hand, d_M^2 is quadratic, since M is affine. Then $d_G^2 + d_M^2$ is piecewise linear-quadratic and $\mathcal{C}^{1+}$, moreover with

$$\min(d_G^2 + d_M^2) = 0, \quad \mathrm{argmin}(d_G^2 + d_M^2) = Z, \quad 0 \in \partial(D_g + D_m)(\bar y).$$

The conjugate function $f = (d_G^2 + d_M^2)^*$, having $\bar y \in \partial f(0)$, is then piecewise linear-quadratic, too, because that property is dual to itself [32, 11.14]. Piecewise linear-quadratic functions are fully amenable everywhere, so Corollary 4.1.1 is applicable and yields for $f^* = d_G^2 + d_M^2$ the quadratic growth property that

$$\exists\, \kappa_{G,M} > 0 \ \text{ such that } \ d_G(y) + d_M(y) \geq \kappa_{G,M}\,\mathrm{dist}^2(y, Z) \ \text{ for } y \text{ near } \bar y. \tag{4.13}$$

If we can establish local bounds of the form

$$\mu - \widehat{h}(y) \geq a_G^{-1} d_G^2(y), \qquad \mu - \widehat{h}(y) \geq a_M^{-1} d_M^2(y), \qquad \text{for some } \ a_G > 0,\ a_M > 0, \tag{4.14}$$

for a_G and a_M as described in (4.12), we will have $(a_G + a_M)[\mu - \widehat{h}] \geq d_G^2 + d_M^2$ around $\bar y$ and be able to conclude through (4.13) that the value b designated in (4.12) is valid for (3.11).

A shift in perspective will facilitate the development of the estimates in (4.14). Recalling that $g(F(\bar x)) + g^*(\bar y) = F(\bar x)\cdot\bar y$, because $\bar y \in \partial g(F(\bar x)$, consider the conjugate pair

$$\begin{array}{l} \bar g(u) = g(F(\bar x) + u) - g(F(\bar x)), \qquad \bar g^*(y) = g^*(y) - g^*(\bar y) - F(\bar x)\cdot(y - \bar y), \\ \text{with} \quad \bar g(0) = 0 = \min \bar g^* = \bar g^*(\bar y), \quad \mathrm{argmin}\, \bar g = G, \quad \bar g^*(y) \geq b_0\,\mathrm{dist}^2(y, G), \end{array} \tag{4.15}$$

the last by our assumption (4.10). Switching from functions of x to functions of $\xi = x - \bar x$, define

$$\begin{array}{l} \bar f_0(\xi) = f_0(\bar x + \xi) - f_0(\bar x) \ \text{ with } \ \bar f_0(0) = 0,\ \nabla \bar f_0(0) = \nabla f_0(\bar x), \\ \bar F(\xi) = F(\bar x + \xi) - F(\bar x) \ \text{ with } \ \bar F(0) = 0,\ \nabla \bar F(0) = \nabla F(\bar x). \end{array} \tag{4.16}$$

Then $g(F(x) + u) = g\Big(F(\bar x) + \bar F(\xi) + u\Big) = \bar g(\bar F(\xi) + u) + g(F(\bar x))$, so that

$$\begin{array}{l} \varphi(\bar x + \xi, u) = \mu + \bar f_0(\xi) + \bar g(\bar F(\xi) + u), \\ l(\bar x + \xi, y) = \mu + \bar f_0(\xi) + y\cdot\bar F(\xi) - \bar g(y), \\ l_{\bar r}(\bar x + \xi, y) = \mu + \bar f_0(\xi) + y\cdot\bar F(\xi) + \frac{\bar r}{2}|\bar F(\xi)|^2 - \bar g(y + \bar r \bar F(\xi)). \end{array} \tag{4.17}$$

According to the definition of $\hat{h}$, this gives us

$$\mu - \widehat{h}(y) = \max_{\bar{x}+\xi\in\mathcal{X}}\Big\{ \bar{g}_{\bar{r}}^*(y + \bar{r}\bar{F}(\xi)) - \Big(\bar{f}_0(\xi) + y\cdot\bar{F}(\xi) + \tfrac{\bar{r}}{2}|\bar{F}(\xi)|^2\Big)\Big\}. \tag{4.18}$$

In particular, in taking $\xi = 0$, we see from (4.18) that $\mu - \bar{h}(y) \geq \bar{g}_{\bar{r}}^*(y)$. Here $\bar{g}_{\bar{r}}^*$, like g_r^*, is given by infimal convolution of $\bar{g}^*$ with $\bar{r}^{-1}j$ for the function $j = \frac{1}{2}|\cdot|^2$, so by the final part of (4.15), $\bar{g}_{\bar{r}}^*$ is bounded above by the infimal convolution of $\bar{r}^{-1}j$ with $b_0\,\mathrm{dist}^2(\cdot, Z)$, which is itself generated by infimal convolution of δ_G with $2b_0 j$. Infimal convolution of $\alpha^{-1}j$ with $\beta^{-1}j$ produces $(\alpha+\beta)^{-1}j$, since in conjugacy with $j^* = j$ this corresponds to $\alpha j + \beta j = (\alpha+\beta)j$. In the present case where $\alpha = \bar{r}$ and $\beta = (2b_0)^{-1}$, we are looking at the infimal convolution of δ_G with $\alpha^{-1}j$ and $\beta^{-1}j$, hence that of δ_G with $(\alpha+\beta)^{-1}j$, which results in $\frac{1}{2}(\alpha+\beta)^{-1}\,\mathrm{dist}^2(\cdot, Z)$. This function, therefore, is $\leq \bar{g}_{\bar{r}}^*$, hence $\leq \mu - \widehat{h}$. Thus, the first inequality in (4.14) holds for $a_G = 2(\bar{r} + (2b_0)^{-1}) = b_0^{-1} + 2\bar{r}$, which is the value of a_G indicated in (4.12).

Working now towards the second inequality in (4.13), we take advantage of the fact that $\min \bar{g}^* = 0$ to derive from (4.18) the estimate

$$\mu - \widehat{h}(y) \geq -\min_{\bar{x}+\xi\in\mathcal{X}} k(\xi, y) \;\text{ for }\; k(\xi, y) = \bar{f}_0(\xi) + y\bar{F}(\xi) + \tfrac{\bar{r}}{2}|\bar{F}(\xi)|^2. \tag{4.19}$$

The function k is strongly convex in ξ when $\bar{x} + \xi \in \mathcal{X}$, inasmuch as $l_{\bar{r}}(\bar{x}+\xi, y)$ has this property when $y \in \mathcal{Y}$ (from our set-up with strong variational sufficiency). It has

$$k(0, y) = 0 \;\text{ and }\; \nabla_\xi k(0, y) = \nabla f_0(\bar{x}) + \nabla F(\bar{x})^* y = \nabla F(\bar{x})^*(y - \bar{y}),$$

since $\nabla f_0(\bar{x}) + \nabla F(\bar{x})^*\bar{y} = \nabla_x L(\bar{x}, \bar{y}) = 0$, as well as

$$\nabla^2_{\xi\xi} k(0, y) = \nabla^2_{xx} L(\bar{x}, \bar{y}) + \bar{r}I + \textstyle\sum_{i=1}^m (y_i - \bar{y}_i)\nabla^2 f_i(\bar{x}),$$

where the matrix $\nabla^2_{xx} L(\bar{x}, \bar{y}) + \bar{r}I$ is positive-definite (by virtue of the strong concavity of $l_{\bar{r}}$ in its primal argument). But the norm of the matrix term at the end can be brought below any ε by restricting y to a small-enough neighborhood of $\bar{y}$. Hence through quadratic expansion,

$$\begin{array}{l} \forall r > \bar{r},\ \exists\, \rho, \delta > 0 \;\text{ such that, for the matrix }\; H_r = \nabla^2_{xx} L(\bar{x}, \bar{y}) + rI, \\ k(\xi, y) \;\leq\; \xi\cdot\nabla F(\bar{x})^*(y - \bar{y}) + \frac{1}{2}\xi\cdot H_r\xi \;\text{ for } |\xi| \leq \rho \text{ when } |y - \bar{y}| \leq \delta. \end{array} \tag{4.20}$$

To gain understanding of the vector $\nabla F(\bar{x})^*(y - \bar{y})$ in (4.20), we appeal to the subspace $M^\perp$ orthogonal to the affine set M, which is the range of the linear transformation $\xi \mapsto \nabla F(\bar{x})\xi$ and is complementary to the subspace M_0 parallel to M, that being in turn the kernel of the linear transformation $\eta \mapsto \nabla F(\bar{x})^*\eta$. The projection mappings P_M and $P_{M^\perp}$ furnish the decomposition

$$y = P_M(y) + P_{M^\perp}(y) \;\text{ with }\; |P_{M^\perp}(y)| = d_M(y) \tag{4.21}$$

and yield $\nabla F(\bar{x})^*(y - \bar{y}) = \nabla F(\bar{x})^* P_{M^\perp}(y)$, since both $P_M(y)$ and $\bar{y}$ belong to M, so their difference belongs to M_0. Applying this insight to (4.20) and looking back to (4.19), we see that

$$\begin{array}{rcl} |y - \bar{y}| \leq \delta \;\;\Longrightarrow\;\; \mu - \widehat{h}(y) &\geq& -\min\limits_{|\xi|\leq\rho}\Big\{ \xi\cdot\nabla F(\bar{x})^* P_{M^\perp}(y) + \frac{1}{2}\xi\cdot H_r\xi \Big\} \\ &\geq& -\min\limits_{|\xi|\leq\rho}\Big\{ \xi\cdot\nabla F(\bar{x})^* P_{M^\perp}(y) + \frac{1}{2}||H_r||\,|\xi|^2 \Big\}. \end{array} \tag{4.22}$$

By letting η stand for $\nabla F(\bar{x})^* P_{M^\perp}(y)$ in the last minimization for simplicity, it can be seen that the minimizing ξ equals $-||H_r||^{-1}\eta$ when $||H_r||^{-1}|\eta| < \rho$, in which case the minimum is global over $\mathbb{R}^m$ by convexity. Then

$$\begin{aligned}-\min_{|\xi|\le\rho}\Big\{\,\xi\cdot\nabla F(\bar{x})^* P_{M^\perp}(y) + \tfrac{1}{2}||H_r||\,|\xi|^2\Big\} &= \max_{|\xi|\in R^m}\Big\{\,\xi\cdot[-\nabla F(\bar{x})^* P_{M^\perp}(y)] - \tfrac{1}{2}||H_r||\,|\xi|^2\Big\}\\ &= \frac{1}{2||H_r||}|\nabla F(\bar{x})^* P_{M^\perp}(y)|^2 \ge \frac{1}{2||H_r||}\beta(\nabla F(\bar{x}))^2|P_{M^\perp}(y)|^2\end{aligned}$$

by (4.11), as continued from (4.22). But $|P_{M^\perp}(y)^2| = \mathrm{dist}(y, M)$. Thus, the second estimate in (4.14) holds for the value of a_M indicated in (3.15). This completes the proof of the theorem. □

The result in Theorem 4.2 is formulated to bring out the influence of the threshold value $\bar{r}$ in the variational sufficiency. Both a_G and a_M increase with $\bar{r}$, and that lowers b and the accompanying rate of linear convergence.

In the degenerate case excluded from Theorem 4.2, where $\nabla F(\bar{x}) = 0$ and $M = \mathbb{R}^m$, there is no distinction between Z and G. Then linear convergence comes out anyway under (4.10) with $b = a_G^{-1}$, and no polyhedral assumption on G. That's the situation also when $G \subset M$, even if $\nabla F(\bar{x}) \ne 0$.

An example illustrating the general reason for assuming G to be polyhedral in Theorem 4.2 is produced in $\mathbb{R}^2$ by taking G to be the closed disk of radius 1 centered at $(1, 0)$ and M to be the vertical axis. Then Z is just $\{(0, 0)\}$, so that $\mathrm{dist}^2(y, Z) = y_1^2 + y_2^2$, but $\mathrm{dist}^2(y, G) + \mathrm{dist}^2(y, M) = y_1^2$ when $(\frac{1}{2}y_1, y_2) \in G$. That precludes the existence of the positive $\kappa_{G,M}$ in (4.13) that enters in the growth conditionn (4.12). This trouble could be avoided by replacing the assumption of polyhedality by the assumption that M meets the relative interior of G. But that wouldn't offer much practical assistance in the absence of a tool of verification.

Second-order difference quotients and second subderivatives as in (3.9) and (3.10) have already been put to important use in formulating Theorem 3.2 and proving Theorem 3.3, but will help even more in the next stage of development. The vehicle for explaining additional background most conveniently will now be the model function g, although other convex functions like g^*, g^r, g_r^*, and the primal/dual functions $\widehat{\varphi}$ and $-\widehat{\psi}$ will benefit later as well. Shifting from the k notation in (3.9) and (3.10), consider u and y with $y \in \partial g(u)$ for $\tau > 0$ the second-order difference quotient function

$$\Delta_\tau^2 g(u\,|\,y)(\omega) = \Big[g(u + \tau\omega) - g(u) - \tau\omega\cdot y\Big]/\tfrac{1}{2}\tau^2 \tag{4.23}$$

and the corresponding second subderivative function

$$d^2 g(u\,|\,y)(\omega) = \liminf_{\substack{\omega'\to\omega\\ \tau\searrow 0}} \Delta_\tau^2 g(u\,|\,y)(\omega'), \tag{4.24}$$

the formula for which means that the epigraph of $d^2 g(u\,|\,y)$ is the outer limit of the epigraphs of the functions $\Delta_\tau^2 g(u\,|\,y)$ as $\tau\searrow 0$ in the sense of set convergence [32, Chap. 4]. If that epigraph is also the inner limit, so that $\Delta_\tau^2 g(u\,|\,y)$ *epi-converges* to $d^2 g(u\,|\,u)$ as $\tau\searrow 0$, then g is said to be *twice epi-differentiable at u for y*. That's especially of interest in our situation, because epi-convergence preserves convexity [32, 7.17] and is itself preserved in conjugacy [32, 11.34]. The conjugate relationship between the convex functions $\frac{1}{2}\Delta_\tau^2 g(u\,|\,y)$ and $\frac{1}{2}\Delta_\tau^2 g^*(y\,|\,u)$ leads to the striking fact that

$$\begin{aligned}&g \text{ is twice epi-diff. at } u \text{ for } y \quad\Longleftrightarrow\quad g^* \text{ is twice epi-diff. at } y \text{ for } u,\\ &\text{and then } \tfrac{1}{2}d^2 g(u\,|\,y) \text{ and } \tfrac{1}{2}d^2 g^*(y\,|\,u) \text{ are conjugate convex functions.}\end{aligned} \tag{4.25}$$

Another remarkable thing about second-order epi-differentiabiliy is the prescription it affords for differentiating subgradient mappings. The mapping ∂g is said to be proto-differentiable at u for $y \in \partial g(u)$

if the set-valued difference quotient mapping $\Delta_\tau[\partial g](u\,|\,y) : \omega \mapsto \tau^{-1}[\partial g(u+\tau\omega) - y]$ converges graphically at $\tau \to 0$; the limit mapping, the graphical derivative, is denoted then by $D[\partial g](u\,|\,y)$ [32, 10G].. In fact $\Delta_\tau[\partial g](u\,|\,y)$ is the subgradient mapping for the function $\frac{1}{2}\Delta_\tau^2 g(u\,|\,y)$, so it's immediate from Attouch's Theorem [32, 12.35+13.40] that

$$\begin{array}{c} g \text{ is twice epi-diff. at } u \text{ for } y \iff \partial g \text{ is proto-diff. at } u \text{ for } y, \\ \text{and then } \partial[\tfrac{1}{2}d^2 g(u\,|\,y)] = D[\partial g](u\,|\,y). \end{array} \tag{4.26}$$

A special aspect of the duality in (3.28) that we'll need to draw on here is the role of nonnegativity. Convexity of g and g^* with $y \in \partial g(u)$, equivalently $u \in \partial g^*(y)$, makes the conjugate convex functions $\frac{1}{2}d^2 g(u\,|\,y)$ and $\frac{1}{2}d^2 g^*(y|u)$ be nonnegative. Because those functions are positively homogeneous of degree two, the convex sets $\operatorname{dom} d^2 g(u\,|\,y) = \{\,\omega \,|\, d^2 g(u\,|\,y)(\omega) < \infty\}$ and

$$\ker d^2 g^*(y\,|\,u) := \{\,\eta \,|\, d^2 g^*(y\,|\,u)(\eta) = 0\} = \operatorname{argmin} d^2 g^*(y\,|\,u)$$

are cones. Moreover from a rule of convex analysis (specializing [32, 11.5] to nonnegagive functions that are positively homegeneous of degree 2), we have the polarity relations

$$[\operatorname{dom} d^2 g(u\,|\,y)]^* = \ker d^2 g^*(y\,|\,u), \qquad [\ker d^2 g^*(y\,|\,u)]^* = \operatorname{cl}[\operatorname{dom} d^2 g(u\,|\,y)]. \tag{4.27}$$

It's also known from [32, 13.5] that $\operatorname{dom} d^2 g(u\,|\,y)$ lies in the normal cone $N_{\partial g(u)}(y)$, which through convexity is polar to the tangent cone $T_{\partial g(u)}(y)$. That leads through (4.27) to the conclusion in the context of (4.25) that

$$\begin{array}{l} \operatorname{cl}[\operatorname{dom} d^2 g(u\,|\,y)] \subset N_{\partial g(u)}(y) \;\text{ and }\; \ker d^2 g^*(y\,|\,u) \supset T_{\partial g(u)}(y) \;\text{ always,} \\ \text{with } \operatorname{cl}[\operatorname{dom} d^2 g(u\,|\,y)] = N_{\partial g(u)}(y) \iff \ker d^2 g^*(y\,|\,u) = T_{\partial g(u)}(y). \end{array} \tag{4.28}$$

Theorem 4.3 (criterion for the second-order condition)**.** *Under the assumption that $\partial g(F(\bar{x}))$ is polyhedral, condition (3.13) in Theorem 3.2(b) will be satisfied if g^* is twice epi-differentiable at $\bar{y}$ for $F(\bar{x}) \in \partial g^*(\bar{y})$ and*

$$d^2 g^*(\bar{y}|F(\bar{x}))(\eta) = 0 \implies \eta \in T_{\partial g(F(\bar{x}))}(\bar{y}), \tag{4.29}$$

which is equivalent to having g twice epi-differentiable at $F(\bar{x})$ for $\bar{y} \in \partial g(F(\bar{x}))$ and the closure of $\operatorname{dom} d^2 g(F(\bar{x})\,|\,\bar{y})]$ being all of the normal cone $N_{\partial g(F(\bar{x}))}(\bar{y})$. Furthermore, this sufficient condition is necessary for (3.13) to hold without drawing on further information about $\nabla F(\bar{x})$ in the relationship between Z and $\partial g(F(\bar{x}))$.

Proof. This will depend on connecting $d^2[-\widehat{h}](\bar{y}\,|\,0)$ with $d^2 g^*(\bar{y}\,|\,F(\bar{x}))$, which will be possible from the perspective of $\widehat{h}$ in $(\widehat{D})$ coming from a restriction operation on $\widehat{\psi}$. It's important from that angle to understand the subgradients and second subderivatives of the convex function $-\widehat{\psi}$, which are partnered through the conjugacy (2.5) with those of $\widehat{\varphi}$.

The argument based on the Lipschitz continuity of the argmin mappings in (3.2), which led to the conclusion about $-\widehat{h}$ in (3.3), leads equally through the formula for $\widehat{\psi}$ in (2.4) to the conclusion that

$$\widehat{\psi} \text{ is } \mathcal{C}^{1+} \text{ concave on } \operatorname{int}[\mathcal{V} \times \mathcal{Y}] \text{ with } \nabla\widehat{\psi}(y) = (x, \nabla_y l_{\bar{r}}(x,y)) \text{ for } x = A_{\bar{r}}(0,y). \tag{4.30}$$

This simplifies consideration of second subderivatives of $-\widehat{\psi}$ from the perspective of their tie to proto-derivatives of the subgradient mapping $\partial[-\widehat{\psi}]$ in the pattern of (4.26). With that subgradient mapping reduced to $-\nabla\widehat{\psi}$ and being Lipschitz continuous, the formula for $d^2[-\widehat{\psi}](v,y\,|\,x,u)$, in which $(x,u) = -\nabla\widehat{\psi}(v,y)$, reduces to

$$d^2[-\widehat{\psi}](v,y\,|\,x,u)(\theta,\eta) = \liminf_{\tau\to 0} \Delta_\tau^2[-\widehat{\psi}](v,y\,|\,x,u)(\theta,\eta).$$

Twice epi-diffferentiability corresponds to $\liminf = \lim$, because pointwise convergence of functions of (θ,η) as $\tau \to 0$ automatically results in uniform convergence on bounded sets of (θ,η) vectors by the Lipschitz continuity. It follows then from having $\widehat{h}(y) = \widehat{\psi}(0,y)$ that

$$\begin{array}{c} \text{twice epi-differentiability of } -\widehat{\psi} \text{ at } (0,\bar{y}) \text{ for } (\bar{x},0) = -\nabla\widehat{\psi}(0,\bar{y}) \\ \Longrightarrow \quad \text{twice epi-differentiability of } -\widehat{h} \text{ at } \bar{y} \text{ for } 0 = -\nabla\widehat{h}(\bar{y}), \\ \text{moreover with } \ d^2[-\widehat{h}](\bar{y}\,|\,0)(\eta) = d^2[-\widehat{\psi}](0,\bar{y}\,|\,\bar{x},0)(0,\eta). \end{array} \tag{4.31}$$

On the other hand, because $-\widehat{\psi}$ and $\widehat{\varphi}$ are conjugate convex functions (2.5), we have in the pattern of (4.25) that

$$\begin{array}{c} [\text{twice epi-differentiability of } -\widehat{\psi} \text{ at } (v,y) \text{ for } (x,u) = -\nabla\widehat{\psi}(v,y)] \\ \Longleftrightarrow \quad [\text{twice epi-differentiability of } \widehat{\varphi} \text{ at } (x,u) \text{ for } (v,y) \in \partial\widehat{\varphi}(x,u)] \\ \text{and then } \ \tfrac{1}{2}d^2[-\widehat{\psi}](v,y\,|\,x,u) \ \text{ is conjugate to } \ \tfrac{1}{2}d^2\widehat{\varphi}(x,u\,|\,v,y). \end{array} \tag{4.32}$$

In applying this to (4.31) we can appeal to the duality between the operations of restriction and inf-projection of convex functions [32, 11.23(c)] to see that

$$\begin{array}{l} \text{twice epi-differentiability of } \widehat{\varphi} \text{ at } (\bar{x},0) \text{ for } (0,\bar{y}) \in \partial\widehat{\varphi}(\bar{x},0) \\ \quad \Longrightarrow \quad \text{twice epi-differentiability of } -\widehat{h} \text{ at } \bar{y} \text{ for } 0 = -\nabla\widehat{h}(\bar{y}), \text{ and then} \\ \tfrac{1}{2}d^2[-\widehat{h}](\bar{y}\,|\,0) \ \text{ is conjugate to the function } \ \omega \mapsto \min_\xi[\tfrac{1}{2}\widehat{\varphi}(\bar{x},0\,|\,0,\bar{y})](\xi,\omega). \end{array} \tag{4.33}$$

In carrying this further, we look at the close relationship between $\widehat{\varphi}$ and $\varphi_{\bar{r}}$. The definition of $l_{\bar{r}}$ in (1.3) as $\inf_u\{\,\varphi_{\bar{r}}(x,u) - y{\cdot}u\}$ says that $-l_{\bar{r}}(x,\cdot)$ is the convex function on $\mathbb{R}^m$ conjugate to $\varphi_{\bar{r}}(x,\cdot)$, whereas the definition of $\widehat{\varphi}$ in (2.4) says that $\widehat{\varphi}(x,\cdot)$ is the convex function conjugate to $-l_{\bar{r}}(x,\cdot)$, when $x \in \mathcal{X}$. For such x, therefore, $\widehat{\varphi}(x,\cdot)$ is the biconjugate of $\varphi_{\bar{r}}(x,\cdot)$, which is its convex hull. In our setting of strong variational sufficiency, however, $\varphi_{\bar{r}}$ is variationally strongly convex at $(\bar{x},0)$ for $(0,\bar{y}) \in \partial\varphi_{\bar{r}}(\bar{x},0)$; the graph of $\partial\varphi_{\bar{r}}$ in some localization around $(\bar{x},0;0,\bar{y})$, and the associated values of $\varphi_{\bar{r}}$ there are indistinguishable from those of some convex function. The convex hull connection with $\widehat{\varphi}$ requires that convex function to agree locally with $\widehat{\varphi}$. In that localization, then,

$$\begin{array}{c} [\text{twice epi-differentiability of } \widehat{\varphi} \text{ at } (x,u) \text{ for } (v,y) \in \partial\widehat{\varphi}(x,u)] \\ \Longleftrightarrow \quad [\text{twice epi-differentiability of } \varphi_{\bar{r}} \text{ at } (x,u) \text{ for } (v,y) \in \partial\varphi_{\bar{r}}(x,u)] \\ \text{and then } \ d^2\widehat{\varphi}(x,u\,|\,v,y) = d^2\varphi_{\bar{r}}(x,u\,|\,v,y). \end{array} \tag{4.34}$$

In particular, $d^2\widehat{\varphi}(\bar{x},0\,|\,0,\bar{y})$ can be replaced in (4.33) by $d^2\varphi_{\bar{r}}(\bar{x},0\,|\,0,\bar{y})$.

To arrive at a formula for $d^2\varphi_{\bar{r}}(\bar{x},0\,|\,0,\bar{y})$, we calculate directly with the second-order difference quotients:

$$\begin{aligned} \Delta^2_\tau\varphi_{\bar{r}}(\bar{x},0\,|\,0,\bar{y})(\xi,\omega) &= \Big[\varphi(\bar{x}+\tau\xi,0+\tau\omega) - \varphi_{\bar{r}}(\bar{x},0) - \tau(\xi,\omega){\cdot}(0,\bar{y})\Big]/\tfrac{1}{2}\tau^2 \\ &= \Big[f_0(\bar{x}+\tau\xi) + g(F(\bar{x}+\tau\xi)+\tau\omega) + \tfrac{\bar{r}}{2}|\tau\omega|^2 - f_0(\bar{x}) - g(F(\bar{x})) - \tau\omega{\cdot}\bar{y}\Big]/\tfrac{1}{2}\tau^2. \end{aligned} \tag{4.35}$$

In terms of $\Delta_\tau F(\bar{x})(\xi) = \tau^{-1}[F(\bar{x}+\tau\xi) - F(\bar{x})]$, which converges to $\nabla F(\bar{x})\xi$ as $\tau \searrow 0$, we can express $g(F(\bar{x}+\tau\xi)+\tau\omega)$ as $g(F(\bar{x}) + \tau[\Delta_\tau F(\bar{x})(\xi)+\omega]\,)$ and aim to incorporate the second-order difference quotient

$$\Delta^2_\tau g(F(\bar{x})\,|\,\bar{y})(\omega') = \Big[g(F(\bar{x})+\tau\omega') - g(F(\bar{x})) - \tau\omega'{\cdot}\bar{y}\Big]/\tfrac{1}{2}\tau^2$$

into a reformulation of (4.35) by way of

$$\begin{aligned}\Big[g(F(\bar{x})+\tau[\Delta_\tau F(\bar{x})(\xi)+\omega])-g(F(\bar{x}))-\tau\omega\cdot\bar{y}\Big]/\tfrac{1}{2}\tau^2\\ =\Delta_\tau^2 g(F(\bar{x})\,|\,\bar{y})(\Delta_\tau F(\bar{x})(\xi)+\omega)+\Big[\Delta_\tau F(\bar{x})(\xi)/\tfrac{1}{2}\tau\Big].\end{aligned}\tag{4.36}$$

Assistance will come from recognizing, through $0=\nabla L(\bar{x},\bar{y})=\nabla f_0(\bar{x})+\nabla F(\bar{x})^*\bar{y}$, that

$$\Big[L(\bar{x}+\tau\xi,\bar{y})-L(\bar{x},\bar{y})-\tau\xi\cdot\nabla_x L(\bar{x},\bar{y})\Big]/\tfrac{1}{2}\tau^2=\Big[f_0(\bar{x}+\tau\xi)-f_0(\bar{x})+\tau\Delta_\tau F(\bar{x})(\xi)\Big]/\tfrac{1}{2}\tau^2,$$

and consequently

$$\frac{f_0(\bar{x}+\tau\xi)-f_0(\bar{x})}{\frac{1}{2}\tau^2}+\frac{\Delta_\tau F(\bar{x})(\xi)}{\frac{1}{2}\tau}=\xi\cdot\nabla^2_{xx}L(\bar{x},\bar{y})\xi+\frac{o(\tau^2)}{\tau^2}.$$

We are able from this and (4.36) to continue (4.35) as the expression

$$\xi\cdot\nabla^2_{xx}L(\bar{x},\bar{y})\xi+\bar{r}|\omega|^2+\Delta_\tau^2 g(F(\bar{x})\,|\,\bar{y})(\Delta_\tau F(\bar{x})(\xi)+\omega)+\frac{o(\tau^2)}{\tau^2}$$

and get from the definition of second subderivatives that

$$\begin{aligned}d^2\varphi_{\bar{r}}(\bar{x},0\,|\,0,\bar{y})(\xi,\omega)&=\liminf_{\substack{(\xi',\omega')\to(\xi,\omega)\\ \tau\searrow 0}}\Delta_\tau^2\varphi_{\bar{r}}(\bar{x},0\,|\,0,\bar{y})(\xi',\omega')\\ &=\xi\cdot\nabla^2_{xx}L(\bar{x},\bar{y})\xi+\bar{r}|\omega|^2+\liminf_{\substack{(\xi',\omega')\to(\xi,\omega)\\ \tau\searrow 0}}\Delta_\tau^2 g(F(\bar{x})\,|\,\bar{y})(\Delta_\tau F(\bar{x})(\xi')+\omega').\end{aligned}$$

This makes clear that

$$\begin{aligned}&\text{[twice epi-differentiability of } \varphi_{\bar{r}} \text{ at } (x,u) \text{ for } (v,y)\in\partial\varphi_{\bar{r}}(x,u)]\\ &\qquad\Longleftrightarrow\quad \text{[twice epi-differentiability of } g \text{ at } F(\bar{x}) \text{ for } \bar{y}],\ \text{ and then}\\ &d^2\varphi_{\bar{r}}(\bar{x},0\,|\,0,\bar{y})(\xi,\omega)=\xi\cdot\nabla^2_{xx}L(\bar{x},\bar{y})\xi+\bar{r}|\omega|^2+d^2g(F(\bar{x})\,|\,\bar{y})(\nabla F(\bar{x})\xi+\omega).\end{aligned}\tag{4.37}$$

In consolidating with (4.33), we have learned that

$$\begin{aligned}&\text{[twice epi-differentiability of } g \text{ at } F(\bar{x}) \text{ for } \bar{y}]\\ &\qquad\Longrightarrow\quad \text{[twice epi-differentiability of } -\widehat{h} \text{ at } \bar{y} \text{ for } 0=-\nabla\widehat{h}(\bar{y})], \text{ and then}\\ &\Big(\tfrac{1}{2}d^2[-\widehat{h}](\bar{y}\,|\,0)\Big)^*(\omega)=\min_\xi\Big\{\,\xi\cdot\nabla^2_{xx}L(\bar{x},\bar{y})\xi+\bar{r}|\omega|^2+d^2g(F(\bar{x})\,|\,\bar{y})(\nabla F(\bar{x})\xi+\omega)\,\Big\}.\end{aligned}\tag{4.38}$$

In particular then, through the equivalence of the assumed twice epi-differentiability of g at $F(\bar{x})$ for $\bar{y}$ with that of g^* at $\bar{y}$ for $F(\bar{x})$, we see that

$$\operatorname{dom}(d^2[-\widehat{h}](\bar{y}\,|\,0))^*=\{\,\omega\,|\,\exists\,\xi\ \text{ with }\ \nabla F(\bar{x})\xi+\omega\in\operatorname{dom}d^2g(F(\bar{x})\,|\,\bar{y})\},$$

By taking polars to get kernels as indicated in (4.27), we obtain

$$\ker g^*(\bar{y}\,|\,F(\bar{x}))\cap\ker\nabla F(\bar{x})^*=\ker d^2[-\widehat{h}](\bar{y}\,|\,0)=\{\,\eta\,|\,d^2[-\widehat{h}](\bar{y}\,|\,0)(\eta)=0\},\tag{4.39}$$

where $\ker\nabla F(\bar{x})^*:=\{\,\eta\,|\,\nabla F(\bar{x})^*\eta=0\}$.

We are close to our goal of establishing under the assumed twice epi-differentiability that (4.29) furnishes a criterion for condition (3.13) of Theorem 3.2(b) to hold. Tbat condition can be stated as $\ker[-\widehat{h}]\cap N_Z(\bar{y})=\{0\}$. In the G, M, notation of Theorem 4.2 expressing Z as $G\cap M$, our assumption

that G is polyhedral (whereas M is affine) makes $N_Z(\bar{y}) = N_G(\bar{y}) + M^\perp$. Also, $\ker \nabla F(\bar{x})^* = M^{\perp\perp}$. This translates (3.13) via (4.39) into the condition that

$$\ker d^2 g^*(\bar{y}\,|\,F(\bar{x})) \cap M^{\perp\perp} \cap [N_G(\bar{y}) + M^\perp] = \{0\}, \tag{4.40}$$

while (4.29) takes the form that

$$\ker g^*(\bar{y}\,|\,F(\bar{x})) \subset T_G(\bar{y}). \tag{4.41}$$

Our claim is that (4.41) implies (4.40). Indeed, if we had (4.41) but not (4.40), there would exist nonzero $a \in T_G(\bar{y}) \cap M^{\perp\perp}$ and $u \in M^\perp$ such that $a - u \in N_G(\bar{y})$. But then $a{\cdot}(a-u) \leq 0$ from the usual relationship between tangent vectors and normal vectors, and yet $a{\cdot}u = 0$ because these vectors belong to complementary subspaces, so $a{\cdot}a = 0$ in contradiction to $a \neq 0$.

The final task is illustrating that, if (4.38) doesn't hold, i.e., if $\ker g^*(\bar{y}\,|\,F(\bar{x}))\backslash T_G(\bar{y})$ contains a vector $a \neq 0$, then there exists F such that (4.40) fails. Take F to be an affine mapping $x \mapsto \bar{y} + A(x - \bar{x})$ for which the range of $x \mapsto Ax$ is the hyperplane $a^\perp$. Then $\nabla F(\bar{x}) = A$, $M^\perp = a^\perp$ and $M^{\perp\perp}$ is the one-dimensional subspace generated by a. In this case, $a \in \ker d^2 g^*(\bar{y}\,|\,F(\bar{x}) \cap M^{\perp\perp}$ by choice. However, since $a \notin T_G(\bar{x})$, there must exist b in the polar cone $N_G(\bar{x})$ such that $a{\cdot}b > 0$. Choose λ to make $(a - \lambda b){\cdot}a = 0$, i.e., $\lambda = |a|^2/a{\cdot}b$. Then $a - \lambda b \in M^\perp$ whilc $\lambda b \in N_G(\bar{y})$, hence $a \in N_G(\bar{x}) + M^\perp$. Thus, a violates (4.40), and the claim of necessity is confirmed. □

5 Linear convergence examples and conclusions

The hard work in the two preceding sections has brought insights that enable us now to reach helpful conclusions about the behavior of the augmented Lagrangian method for solving (P) in various situations..

In this section our focus will continue to be on vector pairs $(\bar{x}, \bar{y})$ that satisfy, along with the first-order optimality conditions $\bar{y} \in g(F(\bar{x}))$ and $\nabla_x L(\bar{x}, \bar{y}) = 0$, the strong variational sufficient condition for local optimality. We'll refer to these as *strongly optimal pairs* for short. Theorem 3.1 established that the pairs (x^k, y^k) generated by the algorithm (under the given details of implementation with inexact minimization steps) are sure to converge to some strongly optimal pair $(\bar{x}, \bar{y})$, if initiated in adequate proximiity. The question of when that convergence might be linear, in one way or another, was addressed at a basic level in Theorems 3.1 and 3.2, but the answer there hinged on properties of a localized dual objective function, rather than on aspects of (P) that might be checked directly. Criteria developed in Theorems 4.2 and 4.3, however, opened a path to verification through properties just of the model function g or its conjugate g^*. Our plan now is to follow that path toward results which can more readily be appreciated for their applicability. Afterward we'll take up the alternative possibilities for the convex case of (P) offered by Theorem 3.3.

The picture of linear convergence in Theorem 3.1 embraced not only x^k and y^k, but also the vectors $\bar{x}^k$ that tacitly follow along in (2.14) and converge to $\bar{x}$ as well. When the decision is made to terminate the x^k sequence, the vector at the same stage in the $\bar{x}^k$ sequence can be effectively be pulled up by "exact minimization" in that final iteration, without any attention to such exactness having to be paid in earlier iterations. The superior linear convergence properties of the $\bar{x}^k$ sequence can therefore be of real practical interest.

To juggle all of this for the task at hand, we pose a catch-all definition which covers both of the stopping criteria (1.15b) and (1.15c) used in Theorem 2.3.

Definition (full model support for linear convergence)**.** *The model function g in (P) will be said to provide full support for linear and superlinear convergence at a strongly optimal pair $(\bar{x}, \bar{y})$ when the*

sequences $x^k \to \bar{x}$, $y^k \to \bar{y}$ and $\bar{x}^k \to \bar{x}$ generated by the augmented Lagrangian method in accordance with Theorem 2.3 are guaranteed to have the following convergence properties, without need of any accompaning assumptions on $\nabla F(\bar{x})$:

(a) *under (1.15b),* $\operatorname{dist}(y^k, Z)$ *converges Q-linearly to 0 and* $|\bar{x}^k - \bar{x}|$ *converges R-linearly to 0.*

(b) *under (1.15c),* $|y^k - \bar{y}|$ *converges Q-linearly to 0 and* $|x^k - \bar{x}|$ *converges R-linearly to 0.*

Here the alternative stopping criteria (1.16b) and (1.16c) can replace (1.15b) and (1.15c) under the stipulations in Theorem 2.5.

Superlinear convergence in each case corresponds to the ultimate rate of linear convergence in Theorem 3.2 being 0, which is guaranteed when the limit c_∞ of the proximal parameters c_k is ∞.

We have already brought up, ahead of Corollary 4.1.1, the concept of a function being fully amenable in the sense of being representable as the composite of a piecewise linear-quadratic function with a $\mathcal{C}^2$ mapping under a basic constraint qualification; see [32, 10F] for background. Full amenability has many favorable consequences in second-order variational analysis and is enjoyed by large classes of functions that commonly appear in optimization, cf. [32, 10.24]. Now we consider that property for the model function g itself.

Example 5.1 (linear convergence with fully amenable modeling)**.** *If g is fully amenable at $F(\bar{x})$, then g provides full support for linear and superlinear convergence.*

Detail. In this case Corollary 4.1.1 is applicable with $f = g$ and $\bar{u} = F(\bar{x})$, and the growth condition (4.2) it yields for f^* becomes the growth condition (4.10) assummed in Theorem 4.2 for g^*. Full amenability furthermore makes ∂g be polyhedral-valued through the chain rule in [32, 10.6] (as articulated for f in (4.6)). Thus, Theorem 4.2 is applicable and guarantees the growth condition (3.11) in Theorem 3.2(a) — as long as $\nabla F(\bar{x}) \neq 0$. But, as observed after the proof of Theorem 4.2, that growth condition holds trivially in the degenerate case when $\nabla F(\bar{x}) = 0$.

The second-order condition (3.13) in Theorem 3.2(b) is always satisfied under full amenability, too. To confirm that, we can appeal to the chain rule for second subderivatives in [32, 13.14]. According to that result, full amenability of g at $F(\bar{x})$ implies twice epi-differentiability with the function $d^2g(F(\bar{x})\,|\,\bar{y})$ having $N_{\partial g(F(\bar{x}))}(\bar{y})$ as its domain. That property is equivalent to (3.13), as pointed out in Theorem 3.2(b). □

Example 5.1 greatly adds to the range of problems (P) accompanied by results on linear convergence of the augmented Lagrangian method. Even for previously studied cases like classical nonlinear programming, it offers, under the definition of "full model support," new conclusions and perspectives. This will be seen from various specializations of full amenability.

Example 5.2 (linear convergence with piecewise linear-quadratic modeling)**.** *If g is piecewise linear-quadratic, then g provides full support for linear and superlinear convergence. In particular that covers models where $g = \delta_K$ for a polyhedral convex set, as for instance in classical nonlinear programming.*

Detail. This just specializes Example 5.1 to the case of the full amenability representation $g = \gamma(\Gamma(u))$ in which Γ is the identity mapping. □

Linear convergence results in classical nonlinear programming, which Example 5.2 enhances in several ways, have already been discussed in Section 1. This is only one of many versions of (P) in which g is piecewise linear-quadratic and the strong variational sufficient condition for local optimality is equivalent to, or a bit sharper than, the previously promoted second-order sufficient condition, as explained in [30]. The g behind the modeling possibility in (1.6), for instance, is piecewise linear-quadratic when the C component is polyhedral. In fact it's then piecewise linear (a polyhedral convex function in the sense of [18]).

A recent paper of Hang and Sarabi [10] focuses precisely on linear convergence of the augmented Lagrangian method for the class of generalized nonlinear programming problems in Example 5.2, and the results there can be compared to the ones obtained here. Instead of strong variational sufficiency with its tie to the strong convexity in Theorem 1.2, that paper utilizes a "uniform quadratic growth condition" on the augmented Lagrangian $l_r(x,y)$ locally in x and obtains Q-linear convergence of (x^k,y^k) to $(\bar{x},\bar{y})$ for some $\bar{y}$ in Z, which doesn't need to be a singleton. The relationship between that kind of convergence and the kinds in our definition of "full support" has already been discussed in Section 1. The exact relationship in [10] between the ALM parameter r_k and our threshold $\bar{r}$ is not easy to determine. But $r_k - \bar{r}$ is required to be high enough to cover estimates which, we think, might well imply $r^k > 2\bar{r}$ and thus signal a connection with relaxed proximal point step-sizes, as discussed around (1.14) and addressed in Theorem 2.4.

A key point in the comparison with our results, however, is that in [10] exact execution of the minimization in every iteration of the augmented Lagrangian method is demanded ($x^k = \bar{x}^k$), instead of stopping criteria as in (2.11). But a feature of the piecewise linear-quadratic model in [10] that's absent in our model is the incorporation of an abstract constraint $x \in X$ for a polyhedral convex set X, but their culminating results require X to be affine. In the framework we've limited ourselves to, an abstract constraint can be handled only by building it into the model function g with a term δ_X. That triggers an extra perturbation component u', shifting $\delta_X(x)$ to $\delta_X(x+u')$, and an extra multiplier component y' for it in the ALM computations. (Future elaborations could get around this.)

Example 5.1 covers much more than the piecewise linear-quadratic modeling in Example 5.2. For instance, it allows g to be any convex function that can be characterized locally around $\bar{u}$ as

$$\begin{array}{l} g = g_0 + \delta_C \text{ with } C \text{ described by finitely many } \mathcal{C}^2 \text{ constraints under a basic constraint} \\ \text{qualification and } g_0 \text{ a } \mathcal{C}^2 \text{ function or the pointwise max of finitely many } \mathcal{C}^2 \text{ functions.} \end{array} \tag{5.1}$$

Furthermore, full amenability is preserved under various operations like addition and some kinds of composition. Particularly of interest in our modeling framework is the fact that when $u \in \mathbb{R}^m$ is comprised of components $u^j \in \mathbb{R}^{m_j}$, $j = 1,\ldots,s$,

$$\begin{array}{l} \text{if } g(u^1,\ldots,u^s) = g^1(u^1) + \cdots + g^s(u^s) \text{ and each } g^j \text{ is fully} \\ \text{amenable at } \bar{u}^j, \text{ then } g \text{ is fully amenable at } \bar{u} = (\bar{u}^1,\ldots,\bar{u}^s). \end{array} \tag{5.2}$$

A very special case of (5.1) lies behind *second-order cone programming*, as we point out next.

Example 5.3 (linear convergence in second-order cone programming)**.** *If $g = \delta_K$ for the second-order cone K, then g provides full support for linear and superlinear convergence, as long as $F(\bar{x}) \neq 0$.*

Detail. The cone K in question can be descrbed as consisting of the points $u = (u_1,\ldots,u_m)$ satisfying $|(u_2,\ldots,u_m)| - u_1 \leq 0$. This fits Example 5.1 as a case of full amenability within (4.2) described by a single $\mathcal{C}^2$ constraint, provided that we are not at the origin. □

The results for second-order cone programming in the recent article of Hang, Mordukhovich and Sarabi [9] differ from ours in Example 5.3 in several ways. There, as in [10], a sort of Q-linear convergence of (x^k,y^k) is obtained,[12] hence R-linear convergence of x^k and y^k individually, but not our Q-linear convergence of $\bar{x}^k$ and y^k. In contrast to [10] and here, [9, Theorem 5.3] requires $\bar{y}$ be the *unique* element of Z; the authors communicate, however, that the proof actually only utilizes that assumption in handling the very case we've left out of Example 5.3 as unresolved, where $F(\bar{x}) = 0$.

[12] Specifically, it is shown that, by taking r_k large enough, it can be arranged with respect to the norm $||(x,y)|| = |x|+|y|$ that $||(x^{k+1},y^{k+1}) - (\bar{x},\bar{y})|| \leq \frac{1}{2}||(x^k,y^k) - (\bar{x},\bar{y})||$.

Another assumption in [9, Theorem 5.3], having no counterpart here, is the calmness of a certain "multiplier mapping." Interestingly, the stepsize r_k in [9] must exceed a positive-definiteness constant of the augmented Lagrangian at $\bar{x}$ that resembles our threshold $\bar{r}$, and it must also be above other quantities that emerge in technical estimates. Again, that could relate to relaxed proximal point stepsizes as in Theorem 2.4.

A different category of optimization problems in the frame of Example 5.1 emphasizes norms.

Example 5.4 (linear convergence with norm modeling)**.** *If $g = ||\cdot||_p$ for $p \in [1, \infty]$ then g provides full support for linear and superlinear convergence, as long as $F(\bar{x}) \neq 0$. Likewise, g provides such support if it is the indicator of a closed ball with respect to one of these norms.*

Detail. When $p = 1$ or $p = \infty$, the norm in question is piecewise linear and falls into a special case of Example 5.2. Otherwise it is a $\mathcal{C}^2$ function away from the origin and thus is fully amenable there, fitting with Example 5.1 through (5.1). The case where g is the indicator of a closed ball is covered by Example 5.2 when $p = 1$ or $p = \infty$ and otherwise corresponds to full amenability from a single $\mathcal{C}^2$ constraint in specialization of (5.1). □

Norms, second-order cones, and many other objective or constraint elements can, of course, come together in setting up a fully amenable model function g in the pattern of (5.2).

Applications to the convex case. When the function $\varphi(x, u)$ in (P) is convex in (x, u), which corresponds to the Lagrangian $l(x, y)$ being convex in x for each y, the augmented Lagrangian method simplifies in the manner described in Corollary 2.3.2. The complication of whether a minimum might only be local goes away. Variational sufficiency is automatic, and strong variational sufficiency can be easier to establish on the basis of the tools furnished in [30]. The choice of the model function g may be enough, in itself, to guarantee linear convergence, since the examples presented so far in this section cover the convex case of (P) as well. But there is also now another route to establishing linear convergence, through Theorem 3.3. Our final example shows the way.

Example 5.5 (linear convergence in convex-affine modeling)**.** *Suppose in (P) that f_0 is convex and F is affine, $F(x) = b - Ax$, so the goal in (P) is to*

$$\text{determine a minimizer } \bar{x} \text{ of the convex function } \; f_0(x) + g(b - Ax). \tag{5.3}$$

Suppose strong variational sufficiency holds, which reduces in this setting to confirming a condition on how $\nabla^2 f_0(\bar{x})$ relates to certain generalized quadratic forms assocated with g at $F(\bar{x})$ for $\bar{y}$ [30, Theorem 5]. Then the criteria for linear convergence in Theorem 3.3 are applicable, moreover with the concave function h having the expression

$$h(y) = b{\cdot}y - f_0^*(A^*y) - g^*(y). \tag{5.4}$$

Detail. In this version of the convex case of (P) with $\varphi(x, u) = f_0(x) + g(b - Ax)$ and $l(x, y) = f_0(x) = y{\cdot}[b - Ax] - g^*(y)$, direct calculation of $h(y) = \inf_x \{\, l(x, y) - y{\cdot}u\}$ yields the expression for h indicated in (5.4). □

The generalized quadratic forms associated with g that come into the characterization of strong variational sufficiency in this example can, for chosen specializations of g, be understood very well, but that is a subject we can't get into here; see [30, Sec. 4]. Likewise, the accessibility of the conjugate functions f_0^* and g^* can depend on the form of these functions, as must the checking of the conditions in Theorem 3.3 on the function h in (5.4). Nonetheless, the results in this paper contribute new perspectives to this popular area of convex optimization.

References

[1] Aragon, F.J., Geoffroy, M.F., "Characterization of metric regularity of subdifferential mappings," *J. Convex Analysis* **15** (2008), 365–380.

[2] Alizateh,f., Goldfarb, D., "Second-order cone programming," *Math. Programming* **95** (2003), 3-51.

[3] Bertsekas, D. P., *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, 1982.

[4] Bonnans, J.F., and Ramirez, H.C., "Perturbation analysis of second-order cone programming problems," *Math. Programming* **104** (2005), 205–227.

[5] Buys, J. D., "Dual algorithms for constrained optimization," Thesis, Leiden, 1972.

[6] Dontchev, A.D., Rockafellar, R.T., *Implicit Functions and Solution Mappings*, Springer Verlag, second edition: 2014.

[7] Eckstein, J., Bertsekas, D. P., "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Math. Programming* **55** (1992), 293–318.

[8] Fernandez, D., and Solodov, M.V., "Local convergence of exact and inexact augmented Lagrangian methods under the second-order sufficient optimality condition," *SIAM J. Optimization* **22** (2012), 384–407.

[9] Hang, N.V.T., Mordukhovich, B.S., Sarabi, M.E., "Augmented Lagrangian method for second-order cone programs under second-order sufficiency," J. Global Optimization **82** (2022), 51–81.

[10] Hang, N.V.T., Sarabi, M.E., "Local convergence analysis of augmented Lagrangian methods for piecewise linear-quadratic composite optimization problems," Math. of Operations Research, to appear; arXiv: 2010.11379.

[11] Haarhoff, P. C., Buys, J. D., "A new method for the optimization of a nonlinear function subject to nonlinear constraints," *Computer J.* **13** (1970), 178–184.

[12] Hestenes, M., "Multiplier and gradient methods," *J. Optimization Theory Appl.* **4** (1969), 303–320.

[13] Liu, Y.J., and Zhang, L., "Convergence analysis of the augmented Lagrangian method for second-order cone optimization problems," *Nonlinear Analysis* **67** (2007),1359–1373.

[14] Luque, F.J., "Asymptotic convergence analysis of the proximal point algorithm," *SIAM J. Control Opt.* **22** (1984), 277–293.

[15] Pennanen, T., "Local convergence of the proximal point algorithm and multiplier methods without monotonicity." *Mathematics of Operations Research* **27** (2002), 170–191.

[16] Poliquin, R. A., and Rockafellar, R. T., "Tilt stability of a local minimum," *SIAM J. Optimization* **8** (1998), 287–299.

[17] Powell, M. J. D., "A method for nonlinear optimization in minimization problems," in *Optimization* (R. Fletcher, ed.), Academic Press, 1969, 283–298.

[18] Rockafellar, R. T., *Convex Analysis*, Princeton University Press, 1970.

[19] Rockafellar, R. T., "New applications of duality in convex programming," Proc. 4th Conference on Probability, brasov, Romania, 1971. (This is the written version of a talk given at several conferences, including the 7th International Symposium on Mathematical Programming in the Hague, 1970.

[20] Rockafellar, R. T., "A dual approach to solving nonlinear programming problems by unconstrained optimization," *Math. Programming* **5** (1973), 354–373.

[21] Rockafellar, R. T., "The multiplier method of Hestenes and Powell applied to convex programming," *J. Optimization Theory* **12** (1973), 555–562.

[22] Rockafellar, R. T., "Augmented Lagrange muliplier functions and duality in nonconvex programming." *SIAM J. Control* **12** (1974), 268–285.

[23] Rockafellar, R. T., *Conjugate Duality and Optimization*, No. 16 in Conference Board of MathṠciences Series, SIAM Publications, 1974.

[24] Rockafellar, R. T., "Solving a nonlinear programming problem by way of a dual problem." *Symposia Mathematica* **19** (1976), 135–160.

[25] Rockafellar, R. T., "Monotone operators and the proximal point algorithm." *SIAM J. Control Opt.* **14** (1976), 877–898.

[26] Rockafellar, R. T., "Augmented Lagrangians and applications of the proximal point algorithm in convex programming." *Math. of Operations Research* **1** (1976), 97–116.

[27] Rockafellar, R. T., "Lagrange multipliers and optimality," *SIAM Review* **35** (1993), 183–238.

[28] Rockafellar, R. T., "Variational convexity and local monotonicity of subgradient mappings," *Vietnam Journal of Mathematics*, *Vietnam J. Math.* **47** (2019), 547–561.

[29] Rockafellar, R. T., "Progressive decoupling of linkages in optimization and variational inequalities with elicitable convexity or monotonicity," *Set-Valued and Variational Analysis* **27** (2019), 863–893.

[30] Rockafellar, R. T., "Augmented Lagrangians and hidden convexity in sufficient conditions for local optimality," *Math. Programming*, published online January 2022; doi.org/10.1007/s10107-022-01768-w.

[31] Rockafellar, R. T., "Advances in convergence and scope of the proximal point algorithm," *J. Nonlinear and Convex Analysis* **22** (2021), 2347–2375.

[32] Rockafellar, R. T., and Wets, R. J-B, *Variational Analysis*, No. 317 in the series *Grundlehren der Mathematischen Wissenschaften*, Springer-Verlag, 1997.

[33] Sun, D., Sun J., and Zhang, L., "The rate of convergence of the augmented Lagrangian method for nonlinear semidefinite programming," *Math. Programming* **114** (2008), 349–381.