Advances in Progressive Decoupling Methodology For Problem Decomposition in Optimization

R. Tyrrell Rockafellar[1]

Abstract

The progressive decoupling algorithm solves problems of optimization by iteratively suppressing complications coming from linear linkage relations. Here its formulation is significantly extended and a sharper theory of convergence is provided. Minimization steps no longer have to be executed exactly, and proximal terms in broadened form, potentially for numerical preconditioning, can vary from one interation to the next. Linear convergence, for which no results were previously available, is tied precisely to a condition of metric subregularity at the solution, with its rate given by a formula based on the modulus of that property. It is established that, for problems with piecewise linear-quadratic objectives, linear convergence is guaranteed, and that for convex general objectives the lack of it is, in a sense, rare. The methodology is shown to be extendable also to problem decomposition with nonconvex objectives, but with shortcomings so far in practical implementability.

Keywords: *convex optimization, nonconvex optimization, linkage problems, progressive decoupling, problem decomposition, linear convergence, metric subregularity, variational sufficiency, variable-metric prox-terms.*

Version of 21 July 2025

[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

There's a long story to ideas for solving a difficult problem of optimization by solving an associated sequence of simpler problems. Different approaches typically focus on exploiting different kinds of problem structure. Our focus here is on *linkage* problems of the form

$$\text{(P)} \qquad \text{minimize } \varphi(z) \text{ subject to } z \in S \subset H,$$

where H is a Euclidean space (finite-dimensional Hilbert space), S is a subspace of H, and φ is an extended-real-valued function on H that is closed (lower semicontinuous) and proper in not taking on $-\infty$ and having the set

$$\operatorname{dom}\varphi = \{\, z \in H \mid \varphi(z) < \infty\}$$

be nonempty. The constraint $z \in \operatorname{dom}\varphi$ is implicit in (P), but kept in the background. The thinking is that minimization subject only to the implicit constraint would relatively easy, but there are linear relationships beyond that which are vital to maintain, yet seriously get in the way. Those relationships are posed by specifying S.[2] The hope is to find procedures for solving (P) in which, iteratively, some substitute for φ is minimized in z without minding whether or not z belongs to S.

To assist in such a procedure, the complementary subspace $S^\perp$ orthogonal to S can be brought in together with the mappings

$$P = \text{projection onto } S, \qquad P^\perp = I - P = \text{projection onto } S^\perp.$$

Subgradients of φ can be utilized and have a critical role from the start in furnishing the first-order optimality condition

$$\bar z \in S,\ \bar w \in S^\perp,\ \bar w \in \partial\varphi(\bar z), \ \text{ or } \ (\bar z, \bar w) \in (S \times S^\perp) \cap \operatorname{gph}\partial\varphi. \tag{1.1}$$

In the general framework of variational analysis [16], the existence of $\bar w$ satisfing (1.1) in tandem with $\bar z$ is necessary, under a constraint condition, for $\bar z$ to be a local minimizer in (P); then (1.1) corresponds to 0 being a subgradient of the sum of φ and the indicator of S.

When φ is convex, which is assumed from now on in this section and the next two, but will be relaxed later, in Section 4, the subgradients in (1.1) are those of convex analysis. Duality provides powerful support in pairing (P) with the dual linkage problem

$$\text{(D)} \qquad \text{maximize } -\varphi^*(w) \text{ subject to } w \in S^\perp \subset H$$

for the convex function φ^* conjugate to φ.[3] Because

$$w \in \partial\varphi(z) \iff z \in \partial\varphi^*(w), \tag{1.2}$$

the first-order optimality condition (1.1) for (P) is simultaneously the first-order optimality condition for (D). This joint condition is always sufficient for $\bar z$ to solve (P) and $\bar w$ to solve (D) and it guarantees that the minimum value of the objective in (P) equals the maximum value of the objective in (D). It is moreover necessary under constraint qualifications such as the nonemptiness of $S \cap \operatorname{ri}\operatorname{dom}\varphi$ and $S^\perp \cap \operatorname{ri}\operatorname{dom}\varphi^*$. Thus, there is no harm in taking for granted in what follows the nonemptiness of

$$\begin{aligned} &Z = \{\,\text{all solutions } \bar z \text{ to (P)}\}, \quad W = \{\,\text{all solutions } \bar w \text{ to (D)}\}, \ \text{ with} \\ &Z \times W = \{\,\text{all } (\bar z, \bar w) \text{ that satisfy the subgradient condition (1.1)}\}. \end{aligned} \tag{1.3}$$

[2] This takes the linear relations to be homogeneous, but that can always be arranged by a shift in variables.

[3] The duality fits as a case of [16, 11.41]. The symmetry in this scheme, through having $S^{\perp\perp} = S$ and $\varphi^{**} = \varphi$, is a big incentive for posing the linkages in (P) as given by a subspace rather than a general affine set.

In this setting with subgradients of convex analysis, it's easy to see from (1.1) that a solution algorithm for (P) of the kind envisioned ought to be possible. Having $\bar{w} \in \partial\varphi(\bar{z})$ corresponds to having $0 \in \partial\bar{\varphi}(\bar{z})$ for $\bar{\varphi} = \varphi(z) - \langle \bar{w}, x\rangle$, and that's equivalent to $\bar{z}$ minimizing $\bar{\varphi}$ over H. Thus, a minimizer of φ on S is, in principle at least, obtainable by minimizing $\bar{\varphi}$ with no attention paid to S. Progressive decoupling, as a way of implementing this iteratively with tilts terms that approximate the ideal $\bar{w}$ tilt, was first proposed in [8].[4]

Progressive Decoupling Algorithm PDA$_0$ (original version for convex optimization [8]). *From any initial $z^0 \in S$ and $w^0 \in S^\perp$, generate sequences of vectors $z^k \in S$ and $w^k \in S^\perp$ by*

$$\begin{aligned}&(\mathrm{a}_0)\ \ \hat{z}^{k+1} = \mathrm{argmin}\ \varphi^k \ \ \textit{for}\ \ \varphi^k(z) = \varphi(z) - \langle w^k, z\rangle + \tfrac{r}{2}\|z - z^k\|^2,\\ &(\mathrm{b}_0)\ \ z^{k+1} = P\hat{z}^{k+1}, \quad w^{k+1} = w^k - rP^\perp \hat{z}^{k+1}.\end{aligned}$$

Note that $\hat{z}^{k+1}$ is uniquely determined in the global minimization in step (a$_0$) because the proximal term makes φ^k be strongly convex. The theory for PDA$_0$ in [8] established that (z^k, w^k) must converge to some particular pair $(\bar{z}, \bar{w}) \in Z \times W$ even if there are others. It was unable, however, to establish conditions for the convergence to be linear, or even to confirm convergence when r may replaced by an adjustable sequence of proximal parameter values r_k. Furthermore, the exactness of the minimization in step (a$_0$) couldn't be relaxed.

The reason for these shortcomings in [8] was that the engine for this procedure is the proximal point algorithm, PPA, but the application is inherently delicate. The proximal point theory avaliable from [5] at the time of [8] was inadequate for coping with some of the key challengers. For instance, replacing r by r_k amounted, in the delicate underlying circumstances, to changing from one Euclidean space to another in each iteration. It was hard to see how to make sense of that.

Subsequent work on extending and refining the proximal point algorithm itself has opened a path for moving forward. A variable-metric version was devised in [11] that could be implemented with stopping criteria for inexactness that mimicked those originally in [5]. It tightly tied Q-linear convergence to a property of metric subregularity at the limit pair $(\bar{z}, \bar{w})$ which holds "almost always" with respect to canonical perturbation parameters in the problem. The progressive decoupling algorithm of [8] was adapted to these proximal point advances in [13, Section 4], but in only in monotone mapping mode. Here the improvements are worked out in optimization mode to get an algorithm with many improvements over PDA$_0$. It accepts inexact minimization in the iteration subproblems, allows far more general proximal terms than just $\frac{r}{2}\|z - z^k\|^2$, and is supported by a much sharper theory of convergence.

Progressive Decoupling Algorithm PDA (improved version for convex optimization). *Sequences of vectors $z^k \in S$ and $w^k \in S^\perp$ are generated from any $z^0 \in S$ and $w^0 \in S^\perp$ by steps of the form*

$$\begin{aligned}&(\mathrm{a})\ \ \hat{z}^{k+1} \approx \mathrm{argmin}\ \varphi^k \ \ \textit{for}\ \ \varphi^k(z) = \varphi(z) - \langle w^k, z\rangle + \tfrac{1}{2}\|z - z^k\|^2_{J_k},\\ &(\mathrm{b})\ \ z^{k+1} = P\hat{z}^{k+1}, \quad w^{k+1} = w^k - J_k(P^\perp \hat{z}^{k+1}).\end{aligned}$$

Here "≈" refers to inexact minimization guided by stopping criteria, the details of which will be provided in Section 2. The generalized proximal terms in (a) use norms

$$\|z\|_{J_k} = \sqrt{\langle z, J_k z\rangle} \ \text{ for positive-definite self-adjoint linear } \ J_k : H \to H, \tag{1.4}$$

and the same mappings J_k then enter the dual update rule in (b). These mappings will be required to carry the complementary spaces S and $S^\perp$ into themselves and, as pinned down in Section 2, satisfy

[4] A version for specially structured convex φ preceded it slightly in [6].

bounds on the extent to which J_{k+1} can differ from J_k. Those bounds will in particular induce J_k to approach a positive-definite mapping J_∞ as $k \to \infty$. Taking $J_k = rI$ would reduce the proximal terms in PDA to the one in PDA$_0$. But the generality not only accommodates varying r_k, it can encompass preconditioning and, eventually perhaps even quasi-Newton-type adaptations.

The better theory to go with PDA establishes basic convergence of (z^k, w^k) to some solution pair $(\bar{z}, \bar{w}) \in Z \times W$ when the approximate minimization in each iteration is guided by a stopping criterion based on the J_k-distance of 0 from the subgradient set $\partial\varphi^k(\hat{z}^{k+1})$. Linear convergence is shown to prevail under a stricter condition on that distance, as long as a certain set-valued mapping enjoys metric subregularity at the limit pair, which is seen to hold "usually." The rate of that Q-linear convergence is given by a formula that depends only on the modulus of metric subregularity there as expressed using J_∞. Section 2 will lay that all out.

Section 3 will offer perspectives on how realistic it is to expect such linear convergence. For piecewise-linear-quadratic φ, it is seen to be guaranteed. Without any restriction on φ, it is anyway seen to be generic in the sense of embedding (P) in the family of linkage problems parameterizes by small translation or tilt perturbations of φ. Also answered in Section 3 is the question of whether applying progressive decoupling to (D) instead of to (P) might be a way of reaching solutions to the two problems by a different route. It is demonstrated that, with exact minimization, the two approaches in fact produce exactly the same z^k and w^k sequences.

Section 4 will take up extensions to the case where φ is no longer convex, the subgradients are those of variational analysis [16], and the approximate minimization in step (a) is merely local. It was proved in [8] that, with exact minimization, PDA$_0$ successfully in that way reaches a pair in the solution set $Z \times W$ when initiated with (z^0, w^0) close enough to $Z \times W$, as long as care is taken not to stray away from that locality, and the proximal parameter is high enough to "elicit" a primal-dual-localized property called *variational* convexity. Fundamentally, that was through Pennanen's observation in [3] that the proximal point algorithm only requires local information in its implementation. Variational convexity and the associated *variational sufficient condition* for a local minimum in [8] have since then undergone much scrutiny and elaboration; see [7], [9], [14], [15]. The variable-metric extension of the proximal point algorithm in [11] was confirmed to operate with just local information as well. However, such localization was not taken up in the progressive decompling extension in [13]. Only globally monotone mappings were addressed there, not the local monotonicity that, for subgradient mappings has been determined to correspond to variational convexity.

Section 4 will nonetheless provide some counterparts to the results for PDA in Section 2 that persist locally when the convexity of φ is relaxed. It must be emphasized, though, that those results remain, at this stage of technology, more theoretical than practical. A major obstacle to implementing the progressive decoupling in a realistic numerical application is not understanding how to locate a pair (z^0, w^0) that's near enough to a solution pair. Furthermore, the proximal terms need to be strong enough to elicit the required variational convexity, but there's little guidance, so far, on how to estimate that from accessible knowledge about the ingredients of problem (P).

Before getting into the details of all this, something general needs to be explained about the linkage format chosen for (P) and the extent to which progressive decoupling is "new and different." Other approaches to problem decomposition in convex optimization have long been popular. Many take the format instead to be that of minimizing $f(z) + g(Az)$ for convex f and g and linear A.[5] The format in (P) can be regarded as the special case of that in which $A = I$ and $g = \delta_S$, the indicator of the subspace S. Those other approaches can, in that way, be applied to solve (P). Applying ADMM, for instance, produces in fact the procedure in PDA$_0$. But ADMM is likewise basically a special

[5] That format comes from this author's 1967 paper [4].

application of the proximal point algorithm, PPA, while the PPA can itself be portrayed in turn as a special application of progressive decoupling. From that angle, all these procedures and their formats are "equivalent," yet that doesn't imply equality in theory and practice. The powerful advantage of progressive decoupling is its very direct connection in [8] and [13] to the PPA through Spingarn's device in [17] of taking a partial inverse of a monotone mapping as the vehicle for computations but then translating the implementation back to the original mapping. That makes it possible to propagate almost any advance in PPA theory into an advance in decoupling theory. The same can't be said for ADMM and other decomposition procedures for which the derivation from PPA is longer and more complicated. That's behind the more tenuous results about linear convergence for those procedures, and it may also be why they are typically articulated only with exact minimization.

The linkage format of (P) moreover has a powerful advantage over the extended Fenchel duality format $f(z) + g(Az)$ when developing special application-driven schemes of problem decomposition, at least in this writer's experience. It's easier — and liberating — in efforts toward getting the right decomposition to have only to think about linear relationships which might be decoupled, than it is to somehow identify f, g, and A to achieve that end.

2 Formulation details and results for convex optimization

Under the assumption that φ is convex in (P), which remains in force throughout this section, the function φ^k in step (a) of the improved progressive decoupling algorithm, PDA, is strongly convex along with being closed and proper. The condition $0 \in \partial\varphi^k(z)$ is necessary and sufficient then for φ^k to have its global minimum attained at z. However, instead of insisting on $\widehat{z}^{k+1}$ being the unique global minimizer, we want to allow the minimization to be approximate. A convenient approach that is to require the subgradient set $\varphi^k(\widehat{z}^{k+1})$, which is closed and convex, to contain a vector "near enough" to 0, or in other words, for the distance from 0 to the closed convex set $\partial\varphi(\widehat{z}^{k+1})$ to be small. In our variable-metric setting, however, calibrating that distance by the norm $\|\cdot\|$ of the space H isn't the ideal approach. The iteration-dependent norms $\|\cdot\|_{J_k}$ in (1.4) can do better service. More must be clarified about them before proceeding with minization criteria.

A fundamental requirement imposed on the mappings $J_k : H \to H$ in (1.4), beyond computability without a need to check for membership in S, is that they must preserve the subspaces S and $S^\perp$:

$$J_k(S) \subset S, \quad J_k(S^\perp) \subset S^\perp, \quad \text{so that } \; P \circ J_k = J_k \circ P \; \text{ and } \; P^\perp \circ J_k = J_k \circ P^\perp. \tag{2.1}$$

Then the subspace complementary to S with respect to the J_k-inner product $\langle w, z\rangle_{J_k} = \langle w, J_k z\rangle$ is still its complement $S^\perp$ for the inner product of H. Since every $u \in H$ can be represented as $Pu + P^\perp u$, (2.1) means that

$$\begin{gathered} J_k(u) = B_k(Pu)) + C_k(P^\perp u) \;\text{ and }\; J_k^{-1}(u) = B_k^{-1}(Pu) + C_k^{-1}(P^\perp u) \;\text{ for} \\ \text{positive-definite self-adjoint mappings } \; B_k : S \to S \;\text{ and }\; C_k : S^\perp \to S^\perp, \end{gathered} \tag{2.2}$$

where the self-adjointness is from the perspective of S and $S^\perp$ being Euclidean spaces in themselves with respect to the restrictions of the norm on H. This might seem unwieldy and give the impression that in numerical implementation of the minimization in PDA step (a), direct attention might have to be paid, via P and $P^\perp$, to the subspaces S and $S^\perp$ after all. But it can be perfectly easy to bypass that, as illustrated by the following, widely applicable scheme.

Example 2.1 (adapting to a simple scheme of complementarity). *Suppose that $H = H_+ \times H_0 \times H_-$, a product of three other Euclidean spaces, so that each $u \in H$ has the form $u = (u_+, u_0, u_-)$. Let S_0*

be a subspace of H_0 having, with respect to H_0, orthogonal complement $S_0^\perp$, and let P_0 and $P_0^\perp$ be the projections from H_0 onto S_0 and $S_0^\perp$. Take, as complementary subspaces in H,

$$\begin{array}{l} S = \{\, u = (u_+, u_0, 0) \,|\, u_0 \in S_0\}, \quad S^\perp = \{\, u = (0, u_0, u_-) \,|\, u_0 \in S_0^\perp\}, \;\; \textit{so that} \\ Pu = (u_+, P_0 u_0, 0), \quad P^\perp u = (0, P_0^\perp u_0, u_-), \;\; \textit{for any } u = (u_+, u_0, u_-) \in H. \end{array} \tag{2.3}$$

Next let $D_{k,+} : H_+ \to H_+$ and $D_{k,-} : H_- \to H_-$ be linear mappings that, with respect to the inner products on those spaces, are self-adjoint and positive-definite with associated norms $\|\cdot\|_{D_{k,+}}$ and $\|\cdot\|_{D_{k,-}}$. Let $r_k > 0$ and construct J_k by taking

$$\begin{array}{l} J_k(u_+, u_0, u_-) = (D_{k,+}(u_+), r_k u_0, D_{k,-}(u_-)), \;\; \textit{so} \\ J_k^{-1}(u_+, u_0, u_-) = (D_{k,+}^{-1}(u_+), r_k^{-1} u_0, D_{k,-}^{-1}(u_-)), \end{array} \tag{2.4}$$

noting that J_k conforms to (2.1) and corresponds in (2.2) to taking

$$\begin{array}{l} B_k(u) = D_{k,+}(u_+), r_k u_0, 0) \;\; \textit{for} \;\; u = (u_+, u_0, 0) \in S, \\ C_k(u) = (0, r_k u_0, D_{k,-}(u_-) \;\; \textit{for} \;\; u = (0, u_0, u_-) \in S^\perp. \end{array} \tag{2.5}$$

Then $\|u\|_{J_k}^2 = \|u_+\|_{D_{k,+}}^2 + r_k\|u_0\|^2 + \|u_-\|_{D_{k,-}}^2$ and is trouble-free because, in evaluating it, there is no need to execute the projections P and P^+.

Along with the subspace compatibility requirment on the mappings J_k, it's essential that the shifts in metric from one iteration to the next not be too wild. For that, the B_k and C_k mappings are required here to satisfy the following mutual norm bounds for $z \in S$ and $w \in S^\perp$:

$$\begin{array}{l} \beta_k^{-1}\|z\|_{B_{k-1}} \le \|z\|_{B_k} \le \beta_k\|z||_{B_{k-1}} \;\text{ and }\; \beta_0^{-1}\|z\| \le \|z\|_{B_0} \le \beta_0\|z\|, \\ \gamma_k^{-1}\|w\|_{C_{k-1}^{-1}} \le \|w\|_{C_k^{-1}} \le \gamma_k\|w\|_{C_{k-1}^{-1}} \;\text{ and }\; \gamma_0^{-1}|w| \le \|w\|_{C_0^{-1}} \le \gamma_0\|w\|, \\ \text{where } \beta_k \ge 1,\ \gamma_k \ge 1, \text{ with } \Pi_{k=0}^\infty \beta_k =: \beta < \infty,\ \Pi_{k=0}^\infty \gamma_k =: \gamma < \infty. \end{array} \tag{2.6}$$

The bounds (2.6) have a pattern that originated in [11] and imply, as demonstrated there, that

$$\begin{array}{l} J_k \text{ converges as } k \to \infty \text{ to a positive-definite self-adjoint mapping } J_\infty, \\ \text{while } B_k \to B_\infty \text{ and } C_k \to C_\infty \text{ in (2.2) for such mappings } B_\infty \text{ and } C_\infty. \end{array} \tag{2.7}$$

The variations in geometry must, in that way, settle down eventually. But only the *existence* in (2.6) is supposed. The β_k and γ_k values won't be needed in computations, only in proofs.

The mappings J_k could be specified at the outset, but they might also be generated while the algorithm is executed. Information in iteration k could suggest in some quasi-Newton manner how J_k might advantageously be updated to J_{k+1}. The conditions in (2.2) would, from that angle, exercize some long-run control over the updating.

Example 2.2 (conforming bounds in an elementary situation). *Suppose that $J_k = r_k I$ for a sequence of values $r_k > 0$. Then the bounds in (2.6), implying in particular that $r_k \to r_\infty \in (0, \infty)$, are available if and only if*

$$\textstyle\prod_{k=0}^\infty \alpha_k < \infty \;\; \textit{for} \;\; \alpha_k = \max\{r_k/r_{k-1},\, r_{k-1}/r_k\}. \tag{2.8}$$

Detail. The demand on β_k in (2.6) in this case is that $\beta_k^{-1} r_{k-1} \le r_k \le \beta_k r_{k-1}$, or in other words, $\beta_k \ge r_k/r_{k-1}$ and $\beta_k \ge r_{k-1}/r_k$. That's the same as $\beta_k \ge \alpha_k$ for the $\alpha_k \ge 1$ in (2.8). The demand on γ_k is that $\gamma_k^{-1} r_{k-1}^{-1} \le r_k^{-1} \le \gamma_k r_{k-1}^{-1}$, which is the same as $\gamma_k^{-1} r_{k-1} \le r_k \le \gamma_k r_{k-1}$, so γ_k like β_k can be taken to equal α_k. Then the finite product conditions in (2.6) are the one on α_k in (2.6). □

With these clarifications, we can proceed to our basic convergence result for the improved version of the progressive decompling algorithm presented in Section 1. More background will then have to be explained, however, before our result on linear convergence can even be formulated.

Theorem 2.3 (global convergence of PDA in the convex case)**.** *Under the assumptions on J_k in (2.1) and (2.6) and the underlying assumption that the primal-dual optimality condition in (1.1) can be satisfied, and for tolerances*

$$\varepsilon_k > 0 \ \text{ with } \ \textstyle\sum_{k=0}^{\infty} \varepsilon_k < \infty, \tag{2.9}$$

let the sequences of vectors $z^k \in S$ and $w^k \in S^\perp$ be generated by PDA from any initial $z^0 \in S$ and $w^0 \in S^\perp$ with the stopping criterion for the minimization in step (a) being

$$\begin{array}{l} \operatorname{dist}_{J_k^{-1}}(0, \partial\varphi^k(\widehat{z}^{k+1})) \leq \varepsilon_k, \ \text{ or equivalently through (2.2),} \\ \exists w \in \partial\varphi^k(\widehat{z}^{k+1})) \ \text{ such that } \ \|Pw\|^2_{B_k^{-1}} + \|P^\perp w\|^2_{C_k^{-1}} \leq \varepsilon_k^2, \end{array} \tag{2.10}$$

noting that, in terms of (2.2), the dual update in the PDA step (b) has the alternative expressions

$$w^{k+1} = w^k - C_k(P^\perp \widehat{z}^{k+1}) = w^k - C_k(\widehat{z}^{k+1} - z^{k+1}). \tag{2.11}$$

Then (z^k, w^k) is sure to converge to some solution pair $(\bar{z}, \bar{w}) \in Z \times W$ in (1.1),

Proof. This will be deduced as a specialization of Theorem 4.1 of [5]. That theorem pertains to progressive decoupling in the framework of determining a pair $(\bar{z}, \bar{w})$ such that

$$\bar{z} \in S, \quad \bar{w} \in S^\perp, \quad \bar{w} \in M(\bar{z}), \tag{2.12}$$

for a maximal monotone mapping $M : H \rightrightarrows H$. For convex φ, the subgradient mapping $\partial\varphi$ is maximal monotone, as is well known, and M can therefore be taken as $\partial\varphi$, with (2.12) then being the relations (1.1) that simultaneously solve the optimization problems (P) and (D). Also, the mappings $M^k(z) = M(z) + J_k(z) - w^k$ are the subgradient mappings $\partial\varphi^k$. The earlier stopping criterion in the M^k form reduces then to the one in (2.10), except that the ε_k here is twice the size of the one in the earlier result. That scaling makes no difference in (2.9), of course, so the conclusion about convergence is thereby justified. □

Once more in (2.10) a threat of having to evaluate projections onto S and $S^\perp$ in order to execute the approximation seems to appear. Example 2.1 again indicates a way of bypassing that.

After the basic convergence in Theorem 2.3 the question of linear convergence comes up. What extra circumstances and numerical precautions, if any, are needed to achieve it? But also, what exactly might be meant by linear convergence and how might the rate of it be calibrated? A worthy goal might be Q-linear convergence of (z^k, w^k) to $(\bar{z}, \bar{w})$, which would ordinarily be interpreted in the usual norm on H as saying that[6]

$$\exists\, b \ \text{ such that } \ \|(z^{k+1}, w^{k+1}) - (\bar{z}, \bar{w})\| \leq b\|(z^k, w^k) - (\bar{z}, \bar{w})\| \ \text{ for large enough } k. \tag{2.13}$$

The *rate* of that Q-linear convergence is the inf of all possible b. This version of convergence will generally be out of reach, however, because of the possible multiplicity of solutions to (P) and (D). Instead, the focus will be on getting the distance of (z^k, w^k) from the solution set $Z \times W$ in (1.3) to converge Q-linearly to 0. When both Z and W are singletons, this reduces obviously to the Q-linear convergence in (2.13), but not otherwise, despite our knowing that (z^k, w^k) tends to a particular pair in $Z \times W$.[7] That limit pair $(\bar{z}, \bar{w})$ will nonethless have a crucial role in whether a key property of metric subregularity holds there or not.

[6] Here $\|(z,w)\|^2 = \|z\|^2 + \|w\|^2 = \|z+w\|^2$ because the z-vectors are in S and the w-vectors are in $S^\perp$.

[7] Imagine $(\bar{z}, \bar{w})$ being a surface point of $Z \times W$ that is approached by (z^k, w^k) in a very slow spiral which very quickly nears the surface.

In general, a mapping $T : H \rightrightarrows H$ is said to be *metrically subregular* at $\bar{u}$ for a vector $\bar{v} \in T(\bar{u})$ if there exists $a \in (0,\infty)$ such that, for (u,v) in some small enough neighborhood $\mathcal{U} \times \mathcal{V}$ of $(\bar{u},\bar{v})$,

$$\operatorname{dist}(u, T^{-1}(\bar{v})) \leq a \operatorname{dist}(\bar{v}, T(u) \cap \mathcal{V}) \;\text{ when }\; u \in \mathcal{U}, \tag{2.14}$$

where the distance on the right of the estimate is considered to be ∞ when $T(u) \cap \mathcal{V} = \emptyset$. The *modulus* of that metric subregularity is defined to be the liminf of the available sizes of a as the neighborhood shrinks down to $(\bar{u},\bar{v})$. The neighborhood V turns out to be superfluous in this, however, according [1, 3H.4]. The same condition is captured when $T(u) \cap \mathcal{V}$ in (2.14) is replaced by $T(u)$, with only a neighborhood $\mathcal{U}$ of $\bar{u}$ being specified for nearness.

The mapping T for which metric subregularity will be invoked here is the "partial inverse" of the mapping $\partial\varphi$ obtained as follows, in taking advantage of the fact that every element of H can be represented uniquely as a sum of something in S and something in $S^\perp$:

$$\begin{array}{l} z' + w' \in T(z+w) \;\longleftrightarrow\; z' + w \in \partial\varphi(z+w') \;\text{ for } z, z' \in S,\; w, w' \in S^\perp, \\ \text{or equivalently: } \operatorname{gph} T = R(\operatorname{gph}\varphi),\; \operatorname{gph}\varphi = R(\operatorname{gph} T), \;\text{ for the reflection} \\ R \text{ that takes } (u,v) = (Pu + P^\perp u, Pv + P^\perp v) \text{ to } (Pu + P^\perp v, Pv + P^\perp u). \end{array} \tag{2.15}$$

The modulus for the subregularity will enter a formula for the rate of convergence. But we are operating in a variable-metric environment. It turns out best from that angle, for the sake of gaining the tightest rate, to adapt the metric subregularity and its modulus to a different norm than the given one on H. That goes back to the theory of linear convergence for the variable-metric version of the proximal point algorithm in [11]. The alternative norm isn't $\|\cdot\|_{J_\infty}$, as might be anticipated from (2.7), but rather

$$\begin{array}{l} J^{\#}_\infty(u) = B_\infty(Pu) + C^{-1}_\infty(P^\perp u), \;\text{ with associated distances tied to} \\ \|u\|^2_{J^{\#}_\infty} = \|Pu\|^2_{B_\infty} + \|P^\perp u\|^2_{C^{-1}_\infty}, \qquad \|u\|^2_{J^{\#-1}_\infty} = \|Pu\|^2_{B^{-1}_\infty} + \|P^\perp u\|^2_{C_\infty}. \end{array} \tag{2.16}$$

This is a good time to recall that, because of finite-dimensionality, any norm on H can be estimated in a uniform manner from any other norm on H. In terms of "mineig" referring to lowest eigenvalue and "maxeig" to highest, distances with respect to the norms given by J_k and J_k^{-1} can be estemated in terms of distances with respect to the given norm $\|\cdot\|$ by

$$\begin{array}{c} \sqrt{\operatorname{mineig}(J_k)}\cdot \operatorname{dist} \leq \operatorname{dist}_{J_k} \leq \sqrt{\operatorname{maxeig}(J_k)}\cdot \operatorname{dist}, \\ \dfrac{1}{\sqrt{\operatorname{mineig}(J_k)}}\cdot \operatorname{dist} \leq \operatorname{dist}_{J_k^{-1}} \leq \dfrac{1}{\sqrt{\operatorname{maxeig}(J_k)}}\cdot \operatorname{dist}, \end{array} \tag{2.17}$$

and similarly for $J^{\#}_\infty$ and $J^{\#-1}_\infty$. Shifting to a different norm doesn't affect whether metric subregularity is present or not, but it alters the associated modulus, and a lower value is better..

With all this in mind, we turn now to the extended definition in [11], according to which T is *metrically subregular* at $\bar{u}$ with respect to $\bar{v} \in T(\bar{u})$ *as expressed in the* $J^{\#}_\infty$ *norm* when

$$\exists\, a \in (0,\infty) \;\text{ such that }\; \operatorname{dist}_{J^{\#}_\infty}(u, T^{-1}(\bar{v})) \leq a \operatorname{dist}_{J^{\#-1}_\infty}(\bar{v}, T(u)) \;\text{ for } u \text{ near } \bar{u}. \tag{2.18}$$

The corresponding *modulus* of metric subregularity is then

$$\operatorname{subreg}_{J^{\#}_\infty} T(\bar{u}\,|\,\bar{v}) := \text{liminf of } a \text{ values in (2.18) as the neighborhood shrinks.} \tag{2.19}$$

The case where $\bar{v} = 0$ will be paramount, for reasons explained next.

Proposition 2.4 (properties of the partial inverse). *For the mapping T defined in (2.15),*

$$T^{-1}(0) = Z + W \text{ for the solution sets } Z \text{ and } W \text{ in (1.3).} \tag{2.20}$$

The inequality in (2.18) for $\bar{v} = 0$ translates then, in terms of $z, z' \in S$ and $w, w' \in S^{\perp}$, to

$$\begin{aligned} &\operatorname{dist}^2_{B_\infty}(z, Z) + \operatorname{dist}^2_{C_\infty^{-1}}(w, W) \\ &\qquad \leq a^2 \min\Big\{ \|z'\|^2_{B_\infty^{-1}} + \|w'\|^2_{C_\infty} \,\Big|\, z' + w \in \partial\varphi(z + w') \Big\}. \end{aligned} \tag{2.21}$$

Proof. Because of orthogonality, having $z' + w' = 0$ in (2.15) requires both $z' = 0$ and $w' = 0$. But this corresponds to having $w \in \partial\varphi(z)$ with $z \in S$ and $w \in S^{\perp}$, which in (1.3) means $z \in Z$, $w \in W$. That takes care of (2.20). Then (2.21) follows from (2.15) and the definition of $J^{\#}_{\infty}$ in (2.16). □

In combination with $\bar{v} = 0$ in (2.18) and (2.19), we'll take as $\bar{u}$ the sum $\bar{z} + \bar{w}$ associated with the PDA limit pair $(\bar{z}, \bar{w})$ in Theorem 2.3. The metric subregularity corresponds then to having the estimate (2.21) available when z is near to $\bar{z}$ and w is near to $\bar{z}$. That estimate will therefore be available eventually for the z^k and w^k generated by the algorithm and be able to come into play in convergence analysis.

Theorem 2.5 (linear convergence of PDA in the convex case and its rate). *In the execution of PDA in Theorem 2.3 under the assumptions there, strengthen the stopping criterion in (2.10) to*

$$\operatorname{dist}_{J_k^{-1}}(0, \partial\varphi^k(\widehat{z}^{k+1})) \leq \varepsilon_k \min\{1, \|\widehat{z}^{k+1} - z^k\|_{J_k}\}, \tag{2.22}$$

where alternatively, from the PDA updating rule (b),

$$\|\widehat{z}^{k+1} - z^k\|_{J_k} = \sqrt{\|z^{k+1} - z^k\|^2_{B_k} + \|w^{k+1} - w^k\|^2_{C_k^{-1}}}. \tag{2.23}$$

Suppose at the limit $(\bar{z}, \bar{w})$ approached by (z^k, w^k) that the mapping T in (2.15) is metrically subregular at $\bar{z} + \bar{w}$ for 0. Then the distances

$$\begin{aligned} d^k &= \sqrt{\operatorname{dist}^2_{J_k}(z^k, Z) + \operatorname{dist}^2_{J_k^{-1}}(w^k, W)} \\ &= \sqrt{\operatorname{dist}^2_{B_k}(z^k, Z) + \operatorname{dist}^2_{C_k^{-1}}(w^k, W)} \end{aligned} \tag{2.24}$$

will converge Q-linearly to 0 at the rate

$$b_{\#} = \frac{a_{\#}}{\sqrt{1 + a_{\#}^2}} \text{ for } a_{\#} = \operatorname{subreg}_{J^{\#}_{\infty}} T\Big(\bar{z} + \bar{w} \,\Big|\, 0\Big). \tag{2.25}$$

Proof. Like Theorem 2.3, this specializes a part of [13, Theorem 4.1],[8] but with serious challenges in making the connection and getting the most out of it. That's reflected in the developments above, where almost everything needing argument or explanation has already been addressed. All that remains is verifying that the stopping criterion agrees with the one utilized in [13, Theorem 4.1], which in place of $\|\widehat{z}^{k+1} - (z^k + w^k)\|_{J_k}$ in (2.22) had the square root of $\|z^{k+1} - z^k\|^2_{J_k} + \|w^{k+1} - w^k\|^2_{J_k^{-1}}$.

[8] In (4.22) of [13], which stands behind the rate formula now in (2.25), the subregularity modulus was intended to be the one in (4.18) of [13] for a mapping that specializes here to $J^{\#}_{\infty}$, but a different mapping D_∞ was accidentally indicated through a cloning glitch.

That square root can be identified with the right side of (2.23) on the basis of the structure introduced in (2.2), inamuch as $z^{k+1} - z^k \in S$ and $w^{k+1} - w^k \in S^\perp$. The same structure likewise yields (2.25). At issue therefore is just the equality claimed in (2.23).

In the algorithm's updating rule (b) as elaborated in (2.11), we have $P\widehat{z}^{k+1} = z^{k+1}$ in S and $P^\perp \widehat{z}^{k+1} = C_k^{-1}(w^{k+1} - w^k)$ in $S^\perp$. On the other hand, $P\widehat{z}^{k+1} + P^\perp \widehat{z}^{k+1} = \widehat{z}^{k+1}$. Hence, through (2.2),

$$\begin{aligned} \|\widehat{z}^{k+1} - z^k\|^2_{J_k} &= \|z^{k+1} + C_k^{-1}(w^{k+1} - w^k) - z^k\|^2_{J_k} \\ &= \|z^{k+1} - z^k\|^2_{B_k} + \|C_k^{-1}(w^{k+1} - w_k)\|^2_{C_k}. \end{aligned}$$

But $\|C_k^{-1}(w^{k+1} - w^k)\|^2_{C_k} = \langle C_k^{-1}(w^{k+1} - w^k), C_k(C_k^{-1}(w^{k+1} - w^k))\rangle = \|w^{k+1} - w^k\|^2_{C_k^{-1}}$. In that way, the validity of (2.23) is seen as confirmed. □

The characteristic features of linear convergence of the progressive decoupling algorithm depicted in Theorem 2.5 are *definitive*, because they are based squarely on those of the proximal point algorithm in [11] from which the procedure is derived, where the rate is show to be generically tight. It's interesting that the rate ultimately depends only on the pair $(\bar{z}, \bar{w})$ to which the sequence tends and the T subregularity modulus there, as calibrated by the limit mappings B_∞ and C_∞. Even the failure of linear convergence can be viewed as embodied in this formula: the modulus $a_\#$ can be regarded as equaling ∞ when subregularity is absent, with b in (2.25) being interpreted then as 1.

3 Linear convergence prospects and dualization

How might it be determined in some given instance of the linkage problem (P) whether linear convergence is assured? In principle that would have to be gleaned from a close examination of whether (2.21) is satified by some a when z and w are close to $\bar{z}$ and $\bar{w}$. This could be difficult and would likely have to come down to a universal property of the mapping $\partial\varphi$. One simplification at least is that the mappings B_∞ and C_∞, while affecting the modulus, don't matter in assertaining the *presence* of the subregularity. Tests on satisfying (2.21) can proceed therefore with the norms and distances in (2.21) being just those of H.

Some help can then come from another direction, through connections between metric subregularity and calmness. The mapping T^{-1} is *calm* at 0 for the vector $\bar{z} + \bar{w} \in T^{-1}(0)$ if there are neighborhoods $\mathcal{U}$ of $\bar{z} + \bar{w}$ and $\mathcal{V}$ of 0 for which

$$\exists\, a \in (0, \infty) \;\text{ such that }\; T^{-1}(v) \cap \mathcal{U} \subset T^{-1}(0) + a\|v\|I\!B \;\text{ when }\; v \in \mathcal{V}, \tag{3.1}$$

where $I\!B$ denotes the unit ball in H. The modulus of that calmness is defined to be the liminf of the available a as $\mathcal{U}$ and $\mathcal{V}$ shink to $\bar{z} + \bar{w}$ and 0. This calmness property of T^{-1} is known from [1, 2H.3] to be equivalent to the metric subregularity of T at $\bar{z} + \bar{w}$ for 0, moreover with the calmness modulus equaling the subregularity modulus as calibrated by the norm $\|\cdot\|$ on H. The calibration of calmness could be adapted to $\|\cdot\|_{J^\#_\infty}$, but leaving that aside, it would anyway be possible in principle to get a guarantee of linear convergence by verifying that the calmness in (3.1) is sure to be available when applying PDA to a given φ.

Here's an illustration that covers the case of φ being *piecewise linear-quadratic* in the sense of $\operatorname{dom}\varphi$ being representable as a union of polyhedral convex sets on each of which φ reduces, in a linear coordinate system for H, to a polynomial of no more than degree 2. The theory of such functions is available in [16, Section 10E]. The key feature for us here is that a convex function φ is piecewise linear-quadratic if and only if its subgradient mapping $\partial\varphi$ is piecewise polyhedral in the sense of its graph being the union of a finite collection of polyhedral convex sets.

Theorem 3.1 (guaranteed linear convergence in the piecewise linear-quadratic case)**.** *When φ is piecewise linear-quadratic, there exists $\bar{a} \in (0,\infty)$ such that partial inverse mapping T has*

$$\operatorname{subreg}_{J^{\#}_{\infty}} T(\bar{u} \,|\, \bar{v}) \leq \bar{a} \;\; \textit{for every pair} \;\; (\bar{u},\bar{v}) \in \operatorname{gph} T. \tag{3.2}$$

Then, regardless of anything about the particular limit pair $(\bar{z},\bar{w})$, the distances d^k in Theorem 2.5 will surely converge Q-linearly to 0 at the rate $\bar{b} := \bar{a}/\sqrt{1+\bar{a}^2}$ or better.

Proof. In this case the mapping $\partial\varphi$ is piecewise polyhedral, as noted above. Then T and T^{-1} are piecewise polyhedral too, because that graphical property is maintained when passing to inverses and partial inverses, which simply apply to graphs the isometric linear transformation R seenr in (2.15). Piecewise polyhedral mappings are everywhere "outer Lipschitz continuous" [1, 3D1], that being for T^{-1} the property stronger than the calmness in (3.1) in which the left side is just $T^{-1}(v)$. Moreover, modulus for it is universal because of only finitely many pieces making up the graph of T^{-1}. Here we also appeal to the fact that, because H is finite-dimensional, relations like those in (2.17) are available for estimating a modulus calibated any given norm in terms of a different norm. □

Another sort of guarantee of linear convergence in progressive decoupling is obtained by showing that its failure is, from some perspective, "rare."

Theorem 3.2 (a generic guarantee of linear convergence)**.** *Consider the perturbation-parameterized family of linkage problems $(\mathrm{P}_{\zeta,\omega})$ in which φ is replaced by*

$$\varphi_{\zeta,\omega}(z) = \varphi(z+\zeta) - \langle \omega, z\rangle \;\; \textit{for} \;\; (\zeta,\omega) \in S \times S^{\perp} \;\; \textit{near} \; (0,0), \tag{3.3}$$

where $(\mathrm{P}_{0,0}) = (\mathrm{P})$ and the optimality conditions for $(\mathrm{P}_{\zeta,\omega})$ paralleling (1.1) are

$$\bar{z}_{\zeta,\omega} \in S, \qquad \bar{w}_{\zeta,\omega} \in S^{\perp}, \qquad \bar{w}_{\zeta,\omega} \in \partial\varphi_{\zeta,\omega}(\bar{z}_{\zeta,\omega}). \tag{3.4}$$

Suppose the solutions sets Z and W to (P) and (D) in (1.3) are bounded, which holds if and only the projection of $\operatorname{dom}\varphi$ on $S^{\perp}$ has the origin in its interior relative to $S^{\perp}$ and, for some α, the set $\{\, z \in S \,|\, \varphi(z) \leq \alpha\}$ is nonempty and bounded.

There will then be a neighborhood $\mathcal{O}$ of $(0,0)$ in $S \times S^{\perp}$ such that, for almost all[9] $(\zeta,\omega) \in \mathcal{O}$, the conditions (3.4) determine a unique pair $(\bar{z}_{\zeta,\omega},\bar{w}_{\zeta,\omega})$ and the mapping $T_{\zeta,\omega}$ associated with $\partial\varphi_{\zeta,\omega}$ as in (2.15) will be metrically subregular at $\bar{z}_{\zeta,\omega} + \bar{w}_{\zeta,\omega}$ for 0. The sequence generated by the algorithm in solving $(\mathrm{P}_{\zeta,\omega})$ and its dual $(\mathrm{D}_{\zeta,\omega})$ will itslef then converge Q-linearly to $(\bar{z}_{\zeta,\omega},\bar{w}_{\zeta,\omega})$.

Proof. This translates [13, Theorem 5.2] to our optimization context, but the criterion given for the boundedness of Z and W is particular to that and requires its separate justification. It helps to recall that an orthonormal basis for a coordinate system can be chosen which identifies the elements of H with vector pairs in $I\!R^n \times I\!R^m$, moreover in such a way that S consists of the pairs where the second component is 0, while $S^{\perp}$ consists of those where the first component is 0. In writing $z \in H$ accordingly as (x,u), the linkage problem (P) becomes the problem of minimizing the closed convex function $x \mapsto \varphi(x,0)$. It's well known in convex analysis that if any of the level sets of such a function is bounded, then they all are, and the minimum is attained. On the other hand, the set of solutions to the dual problem is nonempty and bounded if and only if the projection of $\operatorname{dom}\varphi$ in the second argument is a neighborhood of $u=0$. These two conditions are the coordinatized forms of the ones claimed in the theorem for the boundedness of the primal and dual solution sets. □

[9] In the sense of Lebeguese measure, with the exception thus constituting a so-called negligible set.

The genericity in the theorem comes ultimately from a property of maximal monotone mappings that was brought to light in [12].

To complete the findings in convex optimization, it's interesting to return to duality and with the observation that, since progressive decoupling as applied to the linkage problem (P) solves both the dual linkage problem (D), it could also be applied to (D) and likewise then solve both (P) and (D). To what degree might there be an advantage in that approach?

Dual Progressive Decoupling Algorithm PDA* (for convex optimization)**.** *Sequences of vectors $z^k \in S$ and $w^k \in S^\perp$ are generated from any $z^0 \in S$ and $w^0 \in S^\perp$ by steps of the form*

$$\begin{aligned}&(\text{a}*)\;\; \widehat{w}^{k+1} \approx \operatorname{argmin} \varphi^{*k} \;\; \textit{for} \;\; \varphi^{*k}(w) = \varphi^*(w) - \langle z^k, w\rangle + \tfrac{1}{2}\|w - w^k\|^2_{J_k^{-1}},\\ &(\text{b}*)\;\; w^{k+1} = P^\perp \widehat{w}^{k+1}, \quad z^{k+1} = z^k - J_k^{-1}(P\widehat{w}^{k+1}).\end{aligned}$$

Theorem 3.3 (PDA versus PDA* in convex optimization)**.** *When PDA and PDA* start from the same z^0 and w^0 and are executed with exact minimization, they produce the same sequences of vectors z^k and w^k converging in the same manner to the same solutions $\bar{z}$ and $\bar{w}$.*

Proof. In exact minimization, step (a*) of PDA* has $0 \in \partial\varphi^*(\widehat{w}^{k_1}) - z^k + J_k^{-1}[\widehat{w}^{k+1} - w^k]$. This means $z^k - J_k^{-1}[\widehat{w}^{k+1} - w^k] \in \partial\varphi^*(\widehat{w}^{k+1})$, which, since the subgradient mappings φ and φ^* are the inverses of each other, can be written as

$$\widehat{w}^{k+1} \in \partial\varphi(\widehat{z}^{k+1}) \;\text{ for }\; \widehat{z}^{k+1} := z^k - J_k^{-1}(\widehat{w}^{k+1} - w^k). \tag{3.5}$$

Then $0 \in \partial\varphi(\widehat{z}^{k+1}) - w^k + J_k(\widehat{z}^{k+1} - z^k) = \partial\varphi(\widehat{z}^{k+1})$, which identifies $\widehat{z}^{k+1}$ as the argmin in PDA step (a). On the other hand, in applying P and $P^\perp$ to formula for $\widehat{z}^{k+1}$ in (3.5), we see from the fact in (2.1) that these projection commute with J_k, likewise then with J_k^{-1}, that

$$P\widehat{z}^{k+1} = z^k - J_k^{-1}(P\widehat{w}^{k+1}), \qquad P^\perp\widehat{z}^{k+1} = -J_k^{-1}(P^\perp\widehat{w}^{k+1} - w^k), \tag{3.6}$$

inasmuch as $z^k \in S$ and $w^k \in S^\perp$. The two right sides are, respectively, z^{k+1} and $-J_k^{-1}(w^{k+1} - w^k)$ according to PDA* step (b*). The two equations in (3.6) come out then as the update rule in PDA step (b). The steps in PDA* are, in this mode of execution, the same then as the steps in PDA. □

Of course, this pretty fact doesn't rule out differences between the primal and dual procedures when minimization is only approximate, or that either might have an advantage over the other in particular circumstances. Sometimes minimizing might be easier in (a*) than in (a).

4 Extending to nonconvex optimization

The challenge next is understanding how far the results about progressive decoupling might extend when the function φ in the linkage problem (P) isn't convex on H, just closed and proper. Locating a global minimizer can no longer then be the algorithmic goal.

The projection mappings P and $P^\perp$ are still available, and so too is the first-order condition (1.1) with subgradients of convex analysis replaced by the general subgradients of variational analysis [16, 8B], but in association now with local optimality, not global optimality. For $\bar{z}$ to be a local minimizer in (P), it is necessary, under a constraint qualification,[10] to have some $\bar{w}$ satisfy (1.1) with $\bar{z}$, but

[10] The basic one is that φ has no nonzero "horizon" subgradient in $S^\perp$ at $\bar{z}$, see [16, 8.15]. This reduces for convex φ to having the projection of $\operatorname{dom}\varphi$ on $S^\perp$ be a neighborhood of the origin relative to $S^\perp$. It will be assumed to hold in what follows.

for sufficiency something "second-order" must be added. In variational analyis there's more than one approach to that, however, and it's essential to choose one that best supports the needs at hand.

Progressive decoupling revolves not only around the degree to which the limit $\bar{z}$ of a sequence of vectors z^k might be regarded as solving the minimization problem (P), but also the solution qualities of each z^k in the minimization subproblem producing it. Perhaps optimality wishes could be relaxed to just accepting $\bar{z}$ and z^k as first-order "stationary points," but it's hard to see how that could be made to work successfully, with so little in structure to utilize. The alternative is that $\bar{z}$ and z^k should be local minimizers of a strong-enough sort to ensure stability in computations and provide a solid foundation for convergence analysis. A framework for that was devised in [8] by taking a novel approach to sufficiency for local optimality. Instead of relying on generalized second derivatives, as in [16, Chapter 13] and subsequent developments recently reviewed and extended by Mordukhovich in [2], it took advantage of the property of *variational convexity* in [7], [14], that characterizes when a function's subgradient mapping has a graphical localization that is maximal monotone.

Such a localization is exactly what permits the proximal point algorithm to be applied to that mapping locally. Since progressive decoupling is based on that algorithm and we are indeed applying to a subgradient mapping, variational convexity will naturally be the key again here. Variational convexity indicates, in a *primal-dual*-local sense, a function's subgradients and the its values associated with them behave just as if the function is convex. Here is a precise statement.

Definition 4.1 (variational convexity). *A closed proper function ψ on H is variationally convex at $\bar{z}$ for $\bar{w} \in \partial\psi(\bar{z})$ if the following holds with respect to some closed proper function $\hat{\psi}$ that is also convex. There is a neighborhood $\mathcal{Z} \times \mathcal{W}$ of $(\bar{z}, \bar{w})$ and $\alpha > \psi(\bar{z})$ such that, for $\mathcal{Z}_\alpha = \{\, z \in \mathcal{Z} \,|\, \psi(z) < \alpha\}$,*

$$\begin{array}{l} \widehat{\psi} \leq \psi \text{ on } \mathcal{Z} \text{ with } [\mathcal{Z} \times \mathcal{W}] \cap \operatorname{gph} \partial\widehat{\psi} \;=\; [\mathcal{Z}_\alpha \times \mathcal{W}] \cap \operatorname{gph} \partial\psi \text{ and} \\ \widehat{\psi}(z) = \psi(z) \text{ at all } (z, w) \text{ in the joint intersection, such as } (\bar{z}, \bar{w}). \end{array} \tag{4.1}$$

Variational strong convexity has $\hat{\psi}$ strongly convex.

Besides the original developments about this mysterious-at-first concept in [7] and [8, Section 3], with examples, much more has become available in [14] and [15]. Variational strong convexity is characterized for instance by a strong quadratic growth condition on function values holding in a locally uniform sense, and it makes the subgradient mapping be strongly max monotone in a graphical localization.

Definition 4.2 (variational sufficiency in the linkage problem). *The strong variational sufficient condition for local optimality of $\bar{z}$ in (P) holds if for some $\bar{w}$ satisfying with $\bar{z}$ the first-order condition (1.1) there exists $e \geq 0$ (elicitation parameter) such that the function*

$$\varphi_e(z) = \varphi(z) + \frac{e}{2} \operatorname{dist}^2(z, S) \tag{4.2}$$

is variationally strongly convex at $\bar{z}$ for $\bar{w}$. Here $\bar{w} \in \partial\varphi_e(\bar{z})$ as well as $\bar{w} \in \partial\varphi(\bar{z})$, because

$$\partial\varphi_e(z) = \partial\varphi(z) + eP^\perp z, \text{ so that } \partial\varphi_e(z) = \partial\varphi(z) \text{ for } z \in S. \tag{4.3}$$

This guarantees, through the quadratic growth associated with variational strong convexity, that $\bar{z}$ is a strong local minimizer in (P) which is locally isolated; there is no other point $\hat{z}$ near $\bar{z}$ for which even just the first-order condition (1.1) can be satisfied:

$$\begin{array}{l} [(S \times S^\perp) \cap \operatorname{gph} \partial\varphi] \cap [\mathcal{Z} \times \mathcal{W}] \;=\; \{\bar{z}\} \times [\partial\varphi(\bar{z}) \cap \mathcal{W} \cap S^\perp] \\ \text{with } \partial\varphi(\bar{z}) \cap \mathcal{W} \text{ convex and such that every } \hat{w} \in \partial\varphi(\bar{z}) \cap \mathcal{W} \\ \text{partners with } \bar{z} \text{ like } \bar{w} \text{ in giving strong variational sufficiency.} \end{array} \tag{4.4}$$

That's because φ can be replaced in this by φ_e and then in turn by the strongly convex function associated with φ_e in the strong variational convexity and variational sufficiency.

It was observed in [8] that, in the special case where φ is a $\mathcal{C}^2$ function on $H = I\!R^n$ and $\bar{w} = \nabla\varphi(\bar{z})$, this sufficient condition for local optimality comes down to the classical stipulation that $\nabla\varphi(\bar{z}) \perp S$ and the hessian $\nabla^2\varphi(\bar{z})$ is positive-definite relative to S. Explorations since [8] have demonstrated that in many other situations common to optimization a similar identification with a well established strong second-order sufficient condition for local optimality can be made. That includes standard nonlinear programming, generalized nonlinear programming, and much more; see [9], [18]. In [15], new kinds of second derivatives have recently been introduced which enable identification of variational convexity and perhaps estimation of the elicitation parameter in Definition 4.2. For example, in an echo of the case of smooth φ, if φ is "prox-regular" (a property that's commonly assured [16, 13F]), a bundle of "generalized quadratic forms" is associated with φ at $\bar{z}$ for $\bar{w} \in \varphi(\bar{z})$ that can serve as a substitute for the hessian matrix. Strong variational sufficiency corresponds then by [15, Theorem 5.1] to those forms being positive-definite relative to S.

Our adjusted goal for progressive decoupling in the face of nonconvexity is to get it to work locally when initiated with (z^0, w^0) near enough to some $(\bar{z}, \bar{w})$ satisfying the strong variational sufficient condition for local optimality. Experience in [10] with applying augmented Lagrangian methods to problems in nonconvex optimization suggests that the proximal terms in progressive decoupling might have to be coordinated with the elicitation level for the strong variational convexity, marked by the big-enough parameter e. A sort of trust region approach to computation needs to be adopted as well. It involves a small-enough *locality* parameter ρ as the radius of a ball around z^k in the J_k-norm in iteration k. Besides ρ, there will later be a small-enough *initialization* parameter ρ_0 to specify how close (z^0, w^0) ought to be to a locally optimal $(\bar{z}, \bar{w})$ for the procedure to be operate without straying off target. The size of the tolerances ε_k in the stopping conditions must be small enough as well. The details of the relationships will come in Theorem 4.3.

Progressive Decoupling Algorithm PDA′ (localized version for nonconvex optimization). *Let $\rho > 0$. Let $e \geq 0$ and let J_k^e be the mapping obtained from J_k by replacing its C_k part in (2.2) by $C_k + eI$. From $(z^0, w^0) \in S \times S^\perp$ generate sequences of $z^k \in S$ and $w^k \in S^\perp$ by*

$$\begin{aligned}&(\mathrm{a}')\ \ \widehat{z}^{k+1} \approx \operatorname{argmin}\Big\{\, \varphi(z) - \langle w^k, z\rangle + \tfrac{1}{2}\|z - z^k\|^2_{J_k^e} \,\Big|\, \|z - z^k\|_{J_k} \leq \rho \,\Big\},\\ &(\mathrm{b}')\ \ z^{k+1} = P\widehat{z}^{k+1}, \quad w^{k+1} = w^k - J_k(P^\perp \widehat{z}^{k+1}),\end{aligned}$$

with the subgradients in stopping criterion for the approximation in (a′) being those of the function φ^k given by adding the indicator of the ρ-ball trust region to the specified objective function.

Our results about progressive decoupling in the convex case in Section 2 were specialized from [13], where the variable-metric form of the proximal point algorithm in [11] was applied to the partial inverse T of $\partial\varphi$ in (2.15). Although the nonconvex case wasn't addressed in [13], its developments remain pertinent because the variable-metric PPA is able to operate with maximal monotonicity just in a graphical localization. But here, hidden under the surface, we'll in be applying a localization of PPA to a different mapping than T, namely the partial inverse T_e associated with the subgradient mapping for the function φ_e in the strong variational sufficient condition with respect to some $(\bar{z}, \bar{w})$.

From a practical standpoint, there will nevertheless be important shortcomings which will be discussed after the proof of the next theorem.

Theorem 4.3 (convergence properties of PDA′ in nonconvex optimization). *Let $e \geq 0$ and $\rho > 0$. Require the tolerances ε_k in the stopping criteria to have*

$$\textstyle\sum_{k=0}^{\infty} \varepsilon_k < \rho - \rho_0 \ \text{ for a choice of } \rho_0 \in (0, \rho). \tag{4.5}$$

Suppose in the initiation of the procedure that the strong variational sufficient condition for local optimality in problem (P) holds at elicitation level e for some $\bar{z}$ and $\bar{w}$ such that

$$\sqrt{\|z^0 - \bar{z}\|^2 + \|w^0 - \bar{w}\|^2} < \rho_0. \tag{4.6}$$

Then, as long as ρ and ρ_0 are small enough, the approximate minimizers $\widehat{z}^{k+1}$ in step (a′) will always belong to the interior of the ρ-ball giving trust region, with the subgradient set $\partial\varphi^k(\widehat{z}^{k+1})$ in the stopping criteria thus being unaffected by the trust region indicator term in the definition of φ^k in PDA′. The sequences of vectors $z^k \in S$ and $w^k \in S^\perp$ will converge to $\bar{z}$ and some $\tilde{w}$ which, like $\bar{w}$, satisfies with $\bar{z}$ the strong variational sufficient condition for local optimality in (P). Under the tighter stopping criterion in Theorem 2.5, the convergence will be Q-linear as described there, but with $\tilde{w}$ replacing $\bar{w}$ in the required metric subregularity and accompanying convergence rate formula.

Proof. For the function φ_e in Definition 4.2 of the strong variational sufficiency at $\bar{z}$ for $\bar{w}$ we have

$$\varphi_e(z) - w^k + J_k^e(z - z^k) \;=\; \varphi(z) - w^k + J_k(z - z^k), \tag{4.7}$$

so that the steps (a′) and (b′) in PDA′ for φ with the fixed J metric actually apply steps (a) and (b) of PDA to φ_e with the trust region. If φ_e were a strongly convex function on H, not just variationally strongly convex at $\bar{z}$ for $\bar{w}$, the results in Theorems 2.3 and 2.5 would immediately give the convergence properies claimed for PDA′ without any trust region having to be brought in. The challenge facing us therefore is not about convergence, but rather about establishing that, under the added locality restrictions, PDA′ behaves just as it would if applied with φ_e being strongly convex on H.

What the variational strong convexity gives us is an open convex neighborhood of $(\bar{z}, \bar{w})$, which can just as well be taken to be of the form

$$\mathcal{Z} = \mathcal{Z}_S + \mathcal{Z}_{S^\perp},\;\; \mathcal{W} = \mathcal{W}_S + \mathcal{W}_{S^\perp},\;\; \text{for}\;\; \mathcal{Z}_S, \mathcal{W}_S \subset S,\;\; \mathcal{Z}_{S^\perp}, \mathcal{W}_{S^\perp} \subset S^\perp, \tag{4.8}$$

with respect to which the mapping $\partial\varphi_e$ is strongly max monotone. As understood from the theory in [13, Section 4] behind the convex case in Section 2, applying PDA to φ_e corresponds to applying PPA, the proximal point algorithm, to the partial inverse T_e of $\partial\varphi_e$ and the $J_k^\#$-metrics

$$J_k^\#(u) = B_k(Pu) + C_k^{-1}(P^\perp u). \tag{4.9}$$

The PDA iterates on φ_e identify with PPA iterates on T_e through

$$u^k = z^k + w^k, \qquad z^k = Pu^k,\; w^k = P^\perp u^k, \tag{4.10}$$

Here, by the definition of partial inverse in terms of the reflection mapping R in (2.15), we have $\operatorname{gph} T_e = R(\operatorname{gph}\partial\varphi_e)$ and $\operatorname{gph}\partial\varphi_e = R(\operatorname{gph} T_e)$, along with

$$\bar{u} \in T_e^{-1}(0) \;\text{ for }\; \bar{u} = \bar{z} + \bar{w}, \tag{4.11}$$

since in general

$$R(z, w) = (z + w, 0) \iff (z, w) \in S \times S^\perp. \tag{4.12}$$

The neighborhood $\mathcal{Z} \times \mathcal{W}$ in (4.8) passes through R to become the open convex set

$$\mathcal{U} \times \mathcal{V} = (\mathcal{Z}_S + \mathcal{W}_{S^\perp}) \times (\mathcal{W}_S + \mathcal{Z}_{S^\perp}),\;\; \text{a neighborhood of}\;\; (\bar{u}, 0), \tag{4.13}$$

the key to computations around (4.11). Monotonicity is preserved by R, so T_e is maximal monotone relative to that neighborhood $\mathcal{U} \times \mathcal{V}$.

The issue is therefore just whether PDA, under the restrictions in the theorem, amounts in translation to executing T_e in a manner such that nothing beyond $\mathcal{U} \times \mathcal{V}$ takes part in its convergence to some $\tilde{u} \in T^{-1}(0)$. That $\tilde{u}$ would then give, through (4.12) and the projection scheme in (3.11), an alternative pair $(\tilde{z}, \tilde{w})$ that, like $(\bar{z}, \bar{w})$ satisfies the strong variational sufficient condition for local optimality in (P), necessarily therefore with $\tilde{z} = \bar{z}$, as already noted.

Guidance will come from the localized PPA theory in [11, Section 4] after we translate the PDA restrictions into the corresponding PPA restrictions via (4.10). Observe first that the initialization inequality in (4.6) means $\|u^0 - \bar{u}\| < \rho_0$. In connection with the trust region, observe next from step (b$'$) of PDA$'$ that

$$\widehat{z}^{k+1} - z^k = P\widehat{z}^{k+1} - z^k + P^{\perp}\widehat{z}^{k+1} = (z^{k+1} - z^k) - C_k^{-1}(w^{k+1} - w^k),$$

where the z-part belongs to S and the w-part belongs to $S^{\perp}$. That makes $\|\widehat{z}^{k+1} - z^k\|_{J_k}^2$ come out as

$$\begin{aligned}\|z^{k+1} - z^k\|_B^2 + \|C_k^{-1}(w^{k+1} - w^k\|_{C_k}^2 &= \|z^{k+1} - z^k\|_B^2 + \|w^{k+1} - w^k\|_{C_k^{-1}}^2 \\ &= \|(z^{k+1} + w^{k+1}) - (z^k + w^k)\|_{J_k^{\#-1}}\end{aligned}$$

because $\|C_k^{-1}w\|_{C_k}^2 = \langle C_k^{-1}w, C_k C_k^{-1}w\rangle = \|w\|_{C_k^{-1}}^2$. Thus,

$$\|\widehat{z}^{k+1} - z^k\|_{J_k} = \|u^{k+1} - u^k\|_{J_k^{\#}}.$$

We are therefore executing the PPA on T_e with trust region $\|u^{k+1} - u^k\|_{J_k^{\#}}$ starting from $\|u^0 - \bar{u}\| \leq \rho_0$.

According to [11, Theorem 3.1], the PPA interations in question will proceed in these circumstances with only the localization of T_e to the neighborhood $\mathcal{U} \times \mathcal{V}$ being involved, as long as ρ and ρ_0 are small enough that

$$\{\, v \mid \|v\| \leq 2\rho \} \subset \mathcal{V} \;\text{ and, under (4.6), }\; \{\, u \mid \|u - u^0\| \leq 2\rho \} \subset \mathcal{U}. \tag{4.14}$$

This can always be arranged with a corresponding adjustment of the tolerances ε_k in (4.5), and that completes the proof. □

It's apparent from Theorem 4.3 that much more will need to be understood before progressive decoupling in nonconvex optimization can considered useful and practical. The contribution here is only a potential door opener.

A serious shortcoming, of course, is the mystery of how to select the parameter values. How high should e be for good prospects of (P) having a locally optimal solution $\bar{z}$ satisfying the strong variational sufficent condition at that level of elicitation? How could the proximity of $\bar{z}$ be assessed in order to choose ρ, ρ_0, and ε_k that satisfy the conditions in Theorem 4.3 and the further condition that came out in the proof as (4.14), where furthermore the unknown neighborhood $\mathcal{U} \times \mathcal{V}$ in (4.13) is involved?

The sizes would be hard to estimate from problem data, but maybe they could be "discouvered" during computations through an extension of the algorithm in which they, too, could vary from one iteration to the next? That's nothing new in numerical optimization, of course. There's long and positive experience in coping with such discovery in getting quasi-Newton methods to work, for instance. They, too, are based on second-order properties that have influence in close enough proximity to a locally optimal solution. But that points to another lack, and it's a major one. Quasi-Newton methods can be combined with simpler first-order schemes that have a much wider reach in at least being able to find a sort of stationary point quasi-solution.

What might there be that's comparable here in relation to progressive decoupling for a linkage problem? Finding a quasi-solution can be identified with solving the subgradient condition (1.1) for $\bar{z}$ and $\bar{w}$ when the subgradients aren't those of convex analysis which immediately signal local optimality. Could that be accomplished by a first-order-type scheme that respected the linkage structure by having iterations where the subspace constraint is relaxed? Without that relaxation, the rationale for the computations would be missing.

Another shortcoming in PDA$'$, equally serious although not so obvious, is that the proximal term in (a$'$) can't be evaluated when $e > 0$ without utilizing the projection mapping $P^{\perp}$. That undermines the motivation for progressive decoupling: minimization steps that are able to ignore the linkage subspace structure. Maybe there's a way to get around this while still taking advantage of the intriguing reduction to the convex case offered by variational sufficiency?

One idea is to switch the role of the elicitation parameter e from step (a$'$) to step (b$'$) by leaving J_k in the minimization as in PDA but switching from J_k in the dual update formula to J_k^e, but now with $J_k^e = C_k - eI$ instead of $J_k^e = C_k + eI$. This corresponds to applying the PPA to T_e with those J_k^e norms. There's no trouble then with the projection $P^{\perp}$ coming in, because it's already implicit at that stage from the projection P being used to get z^{k+1}. In fact, that was the approach sucessfully taken to progressive decoupling originally in [8]. But, unfortunately, it emerges now that it worked there only because the metric was fixed and the minimization was exact. The trouble unleashed by dropping either of those features is that the variable-metric taming conditions in (2.6) have to hold for $C_k - eI$ instead of C_k, and the distances in the trust region and stopping criteria have to change that way as well. The altered conditions are tighter, however, and with e having a value that is truly just a tentative guess, there's not much sense in supposing they hold for a one particular level of elicitation, unless it's very high. That would cause problems from other angles.

Still, future efforts might lead to discoveries that could bring reliable success to progressive decoupling even in nonconvex optimization.

References

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

[2] Mordukhovich, B.M., *Second-Order Variational Analysis in Optimization, Variational Stability and Control*, Springer 2024.

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

[4] Rockafellar, R.T., "Duality and stability in extremum problems involving convex functions," *Pacific J. Math.* **21** (1967), 167–187.

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

[6] Rockafellar, R.T., "Progressive decoupling of linkages in monotone variational inequalities and convex optimization," *Proceedings of the 10th International Conference on Nonlinear Analysis and Convex Analysis* (Chitose, Japan, 2017), M. Hojo, M. Hoshino, W. Takahashi (eds.), Yokohama Publishers 2019, 271–291.

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

[8] 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.

[9] Rockafellar, R.T., "Augmented Lagrangians and hidden convexity in sufficient conditions for local optimality," *Math. Programming* **198** (2023), 159–194.

[10] Rockafellar, R.T., "Convergence of augmented Lagrangian methods in extensions beyond nonlinear programming," *Mathematical Programming* **199** (2023), 375–420.

[11] Rockafellar, R.T., "Generic linear convergence through metric subregularity in a variable-metric extension of the proximal point algorithm," *Computational Optimization and Applications* **86** (2023), 1327–1346.

[12] Rockafellar, R.T., "Metric regularity properties of monotone mappings," *Serdica Mathematical Journal* **49** (2023), 1–8.

[13] Rockafellar, R.T., "Generalizations of the proximal method of multipliers in convex optimization," *Computational Optimization and Applications* **87** (2024), 219–247.

[14] Rockafellar, R.T., "Variational convexity and prox-regularity," *J. Convex Analysis*, forthcoming.

[15] Rockafellar, R.T., "Derivative tests for prox-regularity and the modulus of convexity," *J. Convex Analysis*, forthcoming.

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

[17] Spingarn, J., "Partial inverse of a monotone operator," *Applied Mathematics and Optimization* **10** (1983), 247–265.

[18] Wang, W.E., Ding, C., Zhang, Y.J., Zhao, X.Y, "Strong variational sufficiency for nonlinear semidefinite programming and its implications," *SIAM J. Optimization* **33** (2023), 2988–3011.