PROBLEM DECOMPOSITION IN BLOCK-SEPARABLE CONVEX OPTIMIZATION: IDEAS OLD AND NEW

R. Tyrrell Rockafellar[1]

Abstract

Problem decomposition in convex optimization can take several forms, but one of the most important is seen in the case of block-separable constraints and objectives. Such structure is able to support solution methodology in which, in each iteration, individual agents solve separate subproblems that incorporate input from a "coordinating entity." Here the history of that approach to computation will be reviewed with its accomplishments and shortcomings, and a new primal-dual decomposition algorithm will be presented. In that procedure augmented Lagrangians for subproblems, instead of for the problem as a whole, have a key role.

Keywords: *convex optimization, block separability, decomposition into subproblems, augmented decomposition algorithm, proximal point algorithm, augmented Lagrangians*

Version of July 5, 2017

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

1 Introduction

Many problems in optimization have a structure suggestive of decisions being made by a number of "agents" who could almost act independently, except for having to coordinate over some shared resources. Such problems could be solved as a whole in various ways but it's tempting to contemplate ways in which the individual agents could be left to do most of the work on their own subject to iterative coordination on the part of a central entitity. The concept has roots in economics. A "market" could set prices for the resources on the basis of which each agent could proceed to optimize without paying attention to what the others are doing. The market could adjust those prices iteratively until supply equals demand, at which point the overall optimization problem would be solved. This attractive prospect underlies familiar free-market philosophy, but what are the mathematical assumptions and algorithmic rules under which it can be realized? Here that will be discussed in the context of convexity, which appears essential, and known developments will be augmented by new results.

The Dantzig-Wolfe method in linear programming [3] in 1961 was one of the first approaches to utilizing separability to decompose an optimization problem into smaller subproblems which could be solved in parallel. It accomplishes this effectively by solving the linear programming dual as the "master problem" that coordinates solutions and iterative modifications of the subproblems. Extensions of that approach have been made to nonlinear convex programming in which the Lagrangian dual is solved by a cutting plane method, but that leads to ever-bigger dual subproblems which could get out of hand without some heuristic "pruning."

So-called multiplier methods based on the augmented Lagrangian function and utilizing the dual problem associated with it [12] might in principle avoid such difficulties and lead to better convergence properties. However, they disrupt the separability on which decomposition relies. One way of getting around that is offered by the ADMM algorithm, which tackles the nonseparated augmented Lagrangian one variable component at a time; see [2] for a recent survey. This has worked well in some applications and is now popular in computer science, but there is little theory to support its convergence when separability leads to more than just two block-components, although some progress was recently made in [6].

Here a different method will be presented which produces augmented Lagrangians that are themselves separable, as in the scheme of Spingarn [17] but with additional flexibility. It employs proximal terms in the primal variables along with "allocation variables" in a mode traditionally associated with primal, instead of dual, approaches to decomposition. Convergence to particular solutions to both the primal and dual problems is assured even when solutions might not be unique.

This method, which we call the augmented decomposition algorithm, will be derived from the proximal point algorithm for maximal monotone mappings [11] by way of application to a special primal-dual saddle function associated with the given optimization problem. This derivation is parallel to that of the progressive hedging algorithm in stochastic programming [14] and its new extension to stochastic variational inequalities in [13].

2 Block-separable problem format and the role of duality

A common platform for utilizing separability in convex optimization is the following problem with respect to $x = (x_1, \dots, x_q)$ in which the x_j's are "subvectors" belonging to spaces $I\!R^{n_j}$:

$$(P) \qquad \begin{array}{l} \text{minimize } \; f_0(x) = f_{01}(x_1) + \cdots + f_{0q}(x_q) \; \text{ subject to } \; x_j \in X_j \; \text{ and} \\ \qquad\qquad\; f_i(x) = f_{i1}(x_1) + \cdots + f_{iq}(x_q) \begin{cases} \leq 0 \; \text{ for } i = 1, \dots, s, \\ = 0 \; \text{ for } i = s+1, \dots, m, \end{cases} \end{array}$$

where each X_j is a nonempty closed convex set in $I\!R^{n_j}$ and the functions f_{ij} are convex for $i = 1, \ldots, s$ but affine for $i = s+1, \ldots, m$. The corresponding Lagrangian function in terms of x in the convex set

$$X = X_1 \times \cdots \times X_q \subset I\!R^{n_1} \times \cdots \times I\!R^{n_q} = I\!R^n$$

and multiplier vectors $y = (y_1, \ldots, y_m)$ in the cone

$$Y = I\!R^s_+ \times I\!R^{m-s}$$

is given by

$$L(x, y) = \sum\nolimits_{j=1}^{q} f_{0j}(x_j) + \sum\nolimits_{i=1}^{m} y_i \sum\nolimits_{j=1}^{q} f_{ij}(x_j). \tag{2.1}$$

It is convex with respect to x and affine in y. Under a constraint qualification of one type or another, if needed at all, a necessary condition for global optimality of $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_q)$ is the existence of some $\bar{y}$ such that $(\bar{x}, \bar{y})$ is a saddle point of L over $X \times Y$:

$$\bar{x} \in \operatorname*{argmin}_{x \in X} L(x, \bar{y}), \qquad \bar{y} \in \operatorname*{argmax}_{y \in Y} L(\bar{x}, y). \tag{2.2}$$

Even without a constraint qualification, the saddle point condition is sufficient for global optimality. That is all standard theory, so for our purposes we can simply assume from now on that at least one saddle point pair $(\bar{x}, \bar{y})$ exists.

The important consequence of the separability in (2.1) is that the Lagrangian can be expressed as

$$L(x, y) = \sum\nolimits_{j=1}^{q} L_j(x_j, y), \quad \text{where } L_j(x_j, y) = f_{0j}(x_j) + \sum\nolimits_{i=1}^{m} y_i f_{ij}(x_j), \tag{2.3}$$

and the minimization side of the saddle point condition (2.3) breaks down then to a separate minimization for each x_j:

$$\bar{x}_j \in \operatorname*{argmin}_{x_j \in X_j} L_j(x_j, \bar{y}) \text{ for } j = 1, \ldots, q. \tag{2.4}$$

This strongly suggests that the task of solving the (P) to get $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_q)$ might be decomposed into solving the subproblems in (2.4), in which convex function $L_j(\cdot, \bar{y})$ are minimized over the convex sets X_j without imposition of other constraints. A catch, of course, is that this apparently requires knowing $\bar{y}$. Nonetheless, the idea is valuable and does turn out, in an iterative formulation, to support interesting numerical approaches to solving problem (P), as will be explained.

Decomposition into the subproblems (2.4) has an attractive interpretation in economics. Imagine that separate "agents" $j = 1, \ldots, q$ are in charge of deciding on the vectors $x_j \in X_j$, but they can't do that without coordinating because of some shared resources. Suppose that the choice of x_j incurs a cost $f_{0j}(x_j)$, in dollars, say, but also requires an input of $f_{ij}(x_j)$ units of resource i for $i = 1, \ldots m$. A negative input would be an output; agent j may already have supply of each resource, but it might not be enough, or on the other hand, it might not be fully needed. Interpret y_i as the "market price" per unit of resource i in dollars. Then $L_j(x_j, y)$ in (2.3) is the total cost incurred by agent j in choosing x_j, with respect to buying inputs and selling outputs of the resources i in a market in which the prices are y_i. The resource constraints in problem (P) insist that the aggregate demand of the agents is matched by the aggregate supply; having 0 on the right side of each constraint corresponds to the total supplies of the resources being contained in the holdings of the agents.

The existence of a pair $(\bar{x}, \bar{y})$ satisfying the saddle point condition assures in this setting that prices $\bar{y}_i$ do exist such that if the agents act independently in their own self interest the results will be able to come together without an imbalance of supply and demand, but it doesn't say that an arbitrary selection of subproblem solutions $\bar{x}_j$ in (2.4) will necessarily accomplish this. The selection must be such that the second part of the saddle point condition is satisfied as well. That part stands for complementary slackness: the aggregated demand for resource i can't be less than 0 unless the price $\bar{y}_i$ for that resource equals 0. Observe, however, that if circumstances are such that the solutions $\bar{x}_j$ in (2.4) are unique, as would hold for instance if the functions f_{0j} are *strictly* convex, then this second part of the saddle point condition must be satisfied by $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_q)$ automatically! The market prices are able in that case to coordinate the actions of the different agents, who therefore have no need to interact directly with each other.

This is the famous principle in economics of *decentralization of decision-making through markets.* But economics isn't the only area where such a notion has big implications. Many optimization problems in engineering and computer science these days have the same kind of mathematical structure. In solving those problems, different "processors" can be the agents in charge of the subproblems and acting on them in parallel.

To make progress on such computational methodology, it's essential to understand more about the multiplier vectors $\bar{y}$ that are able to play such a magical role. What can be said about the set of pairs $(\bar{x}, \bar{y})$ serving as saddle points in (2.2), in our working context where at least one such pair exists? For this the key is the problem *dual* to (P), namely

$$(D) \qquad \text{maximize } \; g(y) \; \text{ over } \; y \in Y, \;\; \text{where } \; g(y) = \inf_{x \in X} L(x, y).$$

Note that, as the infimum of a collection of affine functions $L(x, \cdot)$ on the cone $Y = I\!R^s_+ \times I\!R^{m-s}$, the function g is concave, but it might have $-\infty$ as a value. The real feasible set in (D) is thus $\{\, y \in Y \,|\, g(y) > \infty \}$. Anyway, in terms of this dual problem we have the elementary but important fact that

$$\{\,\text{set of saddle points } (\bar{x}, \bar{y})\} = \{\,\text{set of solutions } \bar{x} \text{ to } (P)\} \times \{\,\text{set of solutions } \bar{y} \text{ to } (D)\}, \qquad (2.5)$$

where the solutions sets to (P) and (D) are convex, closed and nonempty.

Any attempt at determining a multiplier vector $\bar{y}$ that supports solving (P) through the decomposition in (2.4) must therefore be tantamount to a way of solving (D). The hope is that an iterative approach to solving (D), for which many ideas could be investigated, can be elaborated into a method where subproblems akin to those in (2.4), perhaps with modifications, are solved at each step, and the resulting solutions to those subproblems yield in the limit a solution to (P).

3 Known approaches to decomposition via the dual problem

The Dantzig-Wolfe method [3], which was the first major decomposition algorithm in optimization as mentioned already in Section 1, was couched in the technology of the simplex method for linear programmming with the sets X_j in (P) being nonnegative orthants and $s = 0$ (i.e., no inequalities, only equations). Effectively it did act by solving the dual problem alongside the primal problem, since that is a general characteristic of the simplex method. But interpretations in economics were recognized and the procedure was articulated in terms of (D) being the "master problem" in which a master agent (the market) would interact repeatedly with the agents j in striving to get the prices right for decomposition.

Eventually all of that was seen more clearly without the imposition of the simplex apparatus. The particular way of solving the dual problem that came to be popular in articulating this was a version of the *cutting plane method.* That method required some adapting because in original concept it revolved around minimizing a convex function over a convex set for which there wasn't necessarily a convenient constraint representation, but only a way of determining at each boundary point a supporting half-space. Every closed convex set is the intersection of the closed half-spaces that support it, so by generating an increasing collection of such half-spaces and taking their intersection an "outer approximation" could be generated and improved.

In applying that to (D), the crucial convex set is the hypograph of the concave function g, and the hypographs of the affine functions $L(x, \cdot)$ can supply the supporting half-spaces to it. In explaining how that works, it will help to make some simplifying assumptions that avoid some inessential technical issues. Let's suppose that the sets X_j, and therefore X as well, are bounded, and that there is a bounded set $B \subset Y$ (closed and convex) which is known to include the solution set to (D). Then the constraint $y \in B$ can harmlessly be added to (D) during the process of solving it.

The boundedness of X ensures that in the formula for g in (D) the infimum is attained, and in particular that the concave function g is finite (hence continuous). The critical observation is that if x^* attains that minimum in the case of y^*, then

$$g(y) \leq L(x^*, y) \text{ for all } y, \text{ with } g(y^*) = L(x^*, y^*). \tag{3.1}$$

The geometric meaning of this "subgradient-type" inequality is that the hypograph of the affine function $L(x^*, \cdot)$ is a supporting half-space to the hypograph of g at the point $(y^*, g(y^*))$ on its boundary. This is the main tool in the following proceedure, where the iterations are indexed by $\nu = 1, 2, \ldots$.

Cutting-plane-type decomposition algorithm (raw form)**.** *In iteration ν, having a vector x^ν belonging to a current finite subset X^ν of X and a vector $y^\nu \in B \subset Y$, maximize the concave function*

$$g^\nu(y) = \min_{x_j \in X^\nu} L(x, y) \geq g(y) \tag{3.2}$$

over $y \in B$ to get $y^{\nu+1}$ and the corresponding maximum value $g^\nu(y^{\nu+1})$, Then let

$$x_j^{\nu+1} \in \operatorname*{argmin}_{x_j \in X_j} L_j(x_j, y^{\nu+1}) \text{ for } j = 1, \ldots, q \tag{3.3}$$

and add $x^{\nu+1} = (x_1^{\nu+1}, \ldots, x_q^{\nu+1})$ to X^ν to get $X^{\nu+1}$.

What does this algorithm accomplish, in its stated form? The decreasing sequence of max values $g^\nu(y^{\nu+1})$ is known to converge to the common optimal value in the primal and dual problems. Moreover every cluster point of the bounded sequence of vectors y^ν must be a solution $\bar{y}$ to (D). Also, the values $L(x^\nu, y^\nu)$ provide lower bounds to the common optimal value. But does the bounded sequence of vectors x^ν likewise have its cluster points in the set of solutions to (P)? That's not such a simple matter. If however the objective f_0 in (P) is *strictly* convex, that does have to be the case, and then in fact x^ν must converge to the unique solution $\bar{x}$ to (P).

A refinement can take advantage of the special "polyhedral" structure of the function g^ν maximized in (3.2). There will be a subcollection $\bar{X}^\nu$ of vectors $x \in X^\nu$ such that $g^\nu(y^{\nu+1}) = L(x, y^{\nu+1})$ along with coefficients $\lambda_x^\nu \geq 0$, adding to 1, such that

$$\textstyle\sum_{x \in \bar{X}^\nu} \lambda_x^\nu L(x, \cdot) \equiv 0.$$

Utilizing those coefficients to define

$$\bar{x}^\nu = \textstyle\sum_{x \in \bar{X}^\nu} \lambda_x^\nu x$$

one does get, even without strict convexity, a sequence having its cluster points in the solution set of (P). Furthermore the vectors $\bar{x}^\nu$ in this sequence are *feasible* solutions to (P), which could not be said for the sequence of vectors x^ν.

Being able to solve (P) and (D) only in terms of cluster points, in general, is an obvious drawback of the cutting-plane approach. Another drawback is that the process can be slow. Still worse, the sets X^ν could in principle grow indefinitely in size. After a while, this could make the subproblems too large to solve, because most ways of solving them would convert the affine functions $L(x, \cdot)$ for $x \in X^\nu$ into linear constraints in the epigraphical space. A potential remedy for that is to "prune" inessential elements from X^ν to keep it small enough. Then, though, the procedure may devolve into heuristics without solid guarantees.

An alternative approach to solving (D) is to invoke the proximal point algorithm, which ends up replacing L by an *augmented* Lagrangian. The proximal point algorithm [11] in this setting amounts to generating a sequence of vectors y^ν by

$$y^{\nu+1} = \operatorname*{argmax}_{y \in Y} \Big\{ g(y) - \frac{1}{2r}||y - y^\nu||^2 \Big\}, \tag{3.4}$$

as laid out in [12]. Here $r > 0$ is a parameter, and the "proximal term" it controls furnishes strong concavity that ensures not only that the maximum is attained at a unique point but also that the sequence of vectors y^ν converges to some particular solution $\bar{y}$ to (D), even though there might be other solutions. It also makes the maximum value of the espression in question, namely

$$\max_{y \in Y} \Big\{ \inf_{x \in X} \Big\{ L(x, y) - \frac{1}{2r}||y - y^\nu||^2 \Big\} \Big\},$$

be amenable to a minimax interchange that identifies it with

$$\inf_{x \in X} \Big\{ \max_{y \in Y} \Big\{ L(x, y) - \frac{1}{2r}||y - y^\nu||^2 \Big\} \Big\}.$$

This leads to introducing, as the *augmented* Lagrangian for (P), the function

$$\begin{gathered} L_r(x, y) = \max_{y' \in Y} \Big\{ L(x, y') - \frac{1}{2r}||y' - y||^2 \Big\} = f_0(x) + \textstyle\sum_{i=1}^m \psi_i(f_i(x), y_i, r) \\ \text{where} \quad \psi_i(f_i(x), y_i, r) = \begin{cases} -\frac{1}{2r}|y_i|^2 & \text{for } i \le s \text{ with } y_i + r f_i(x) \le 0, \\ y_i f_i(x) + \frac{r}{2}|f_i(x)|^2 & \text{in all other cases,} \end{cases} \end{gathered} \tag{3.5}$$

and, through further analysis (see [12] and its references), to the recognition that the steps in (3.4) can be executed in a manner which also generates a sequence of vectors x^ν. This comes out as follows.

Method of multipliers in convex programming. *In iteration ν, having current vectors $x^\nu \in X$ and $y^\nu \in Y$ determine*

$$x^{\nu+1} \in \operatorname*{argmin}_{x \in X} L_r(x, y^\nu) \tag{3.6}$$

with respect to the chosen parameter value $r > 0$. Then update from y^ν to $y^{\nu+1}$ by

$$y_i^{\nu+1} = \theta_i(y_i^\nu + r f_i(x^{\nu+1})), \quad \textit{where} \ \theta_i(t) = \begin{cases} \max\{0, t\} & \textit{for } i = 1, \dots, s, \\ t & \textit{for } i = s+1, \dots, m. \end{cases} \tag{3.7}$$

Although the sequence of multiplier vectors y^ν corresponds to the execution of (3.4) and converges to some solution to (D) as already indicated, there are again complications that affect the sequence of vectors x^ν. If $f_0(x)$ is strongly convex in $x \in X$, then the solutions in (3.6) are uniquely determined and do converge to the unique solution to (P), but in general there could even be questions about the attainment of minimum in (3.6). Such issues suggest the modification described next, from [12]

Proximal method of multipliers in convex programming. *In iteration ν, having current vectors $x^\nu \in X$ and $y^\nu \in Y$ determine*

$$x^{\nu+1} = \operatorname*{argmin}_{x \in X} \Big\{ L_r(x, y^\nu) + \frac{1}{2r}||x - x^\nu||^2 \Big\} \tag{3.8}$$

with respect to the chosen parameter value $r > 0$. Then update from y^ν to $y^{\nu+1}$, as previously, by

$$y_i^{\nu+1} = \theta_i(y_i^\nu + rf_i(x^{\nu+1})), \quad \textit{where} \;\; \theta_i(t) = \begin{cases} \max\{0, t\} & \textit{for } i = 1, \ldots, s, \\ \quad t & \textit{for } i = s+1, \ldots, m. \end{cases} \tag{3.9}$$

In this case the unique solution in (3.8) surely exists because of strong convexity induced by the proximal term in x, and there is no trouble at all with convergence. The sequence of pairs (x^ν, y^ν) is certain to converge to some saddle point $(\bar{x}, \bar{y})$ yielding solutions to both (P) and (D), even if there are multiple solutions.

However, this approach to employing the dual problem to solve the given problem (P) does *not* afford a method of decomposition, at least not directly. That's because the separability of L into the "sub-Lagrangians" in (2.3) doesn't carry over to the augmented Lagrangian L_r. Specifically, the formula for $L_r(x, y)$ in (3.5) yields

$$L_r(x, y) = f_0(x) + \sum_{i=1}^{s} \Big[y_i \max\{f_i(x), -y_i/r\} + \frac{r}{2}\max^2\{f_i(x), -y_i/r\} \Big] + \sum_{i=s+1}^{m} \Big[y_i f_i(x) + \frac{r}{2} f_i(x)^2 \Big], \tag{3.10}$$

and the separable structure of the constraint functions in (P) is unable to help in the minimization in (3.6) or (3.7), even though the proximal term in (3.8) does, along with $f_0(x)$ break down into a sum of expressions for each j.

Confronted with this difficulty but nonetheless wanting to take advantage of the valuable features of the augmented Lagrangian, researchers have turned to algorithms in which, instead of minimizing in (3.6) with respect to $x = (x_1, \ldots, x_q)$ over $X = X_1 \times \cdots \times X_q$ all at once, they minimize with respect to one x_j at a time, moreover with the update in the multiplier vector being invoked immediately, i.e,. without waiting for minimization with respect to all the other components. Such approaches are often said to be versions of the *alternating direction method of multipliers*, ADMM, but this is something of a misnomer. ADMM was originally established only for the case of two block components x_1 and x_2, not q components as here. The original convergence results, based again on the proximal point algorithm, made that essential. Nevertheless the method has become popular, despite shortcomings in theoretical underpinnings, because of its surprising successes and simplicity. Comprehensive discussion of this along with various recent advances can be found in [2], [4] and [6].

4 A different approach to separability with augmented Lagrangians

By adopting a broader strategy, we will be able to produce, in exstnsion of an approach of Spingarn in [?], a multiplier method in terms of augmented Lagrangians for separate subproblems $j = 1, \ldots, q$ in place of the separability-destroying augmented Lagrangian for problem (P) that has been treated so far. The key is to reformulate (P) equivalently in terms of additional vector variables $w_j = (w_{1j}, \ldots, w_{mj}) \in I\!R^m$ besides the x_j's and then to draw on the general duality of theory in convex optimization to get a different Lagrangian. The reformulated problem statement is

$$(\bar{P})\qquad \begin{array}{l}\text{minimize } f_{01}(x_1)+\cdots+f_{0q}(x_q) \text{ subject to } x_j\in X_j \text{ and}\\ \qquad\qquad f_{ij}(x_j)-w_{ij}\begin{cases}\leq 0 & \text{for } i=1,\dots,s,\\ =0 & \text{for } i=s+1,\dots,m,\end{cases}\\ \text{for } j=1,\dots,q, \text{ and furthermore } w_1+\cdots+w_q=0.\end{array}$$

It's easy to confirm that any solution $(\bar{w},\bar{x})=(\bar{w}_1,\dots,\bar{w}_q,\bar{x}_1,\dots,\bar{x}_q)$ to $(\bar{P})$ yields $\bar{x}$ as a solution to (P), and conversely, that any solution $\bar{x}$ to (P) can be augmented by some $\bar{w}$ to obtain a solution $(\bar{w},\bar{x})$ to $(\bar{P})$. It's in this sense that the two problems are equivalent.

The introduction of the additional variables can be seen as a *primal* approach to decomposition in (P), in contrast to the methods we have been looking at so far, which operate through the dual problem (D). The focus in a primal approach is on the subproblems

$$(\bar{P}_j(w_j))\qquad \begin{array}{l}\text{minimize } f_{0j}(x_j) \text{ over } x_j\in X_j \text{ subject to}\\ \qquad\qquad f_{ij}(x_j)-w_{ij}\begin{cases}\leq 0 & \text{for } i=1,\dots,s,\\ =0 & \text{for } i=s+1,\dots,m.\end{cases}\end{array}$$

In the economic interpretation, w_{ij} correponds to a reallocation of some amount of resource i to agent j from the total amount present in the holdings of all the agents, which through "conservation of resources" requires these amounts to add to 0. The fact that if $(\bar{w},\bar{x})$ solves $(\bar{P})$, then $\bar{x}_j$ solves $(\bar{P}_j(\bar{w}_j))$ for each j, leads to the idea of finding some scheme that in iteration ν has a w^ν, calculates corresponding vectors x_j^ν by solving the subproblems $(\bar{P}(w_j^\nu))$, and then updates to $w^{\nu+1}$.

How might that work, at least in principle? It could be based on the "projection" of $(\bar{P})$ onto a problem in w, namely

$$(\bar{P}_0)\qquad \text{minimize } \varphi(w)=\sum\nolimits_{j=1}^q \varphi_j(w_j) \text{ with } \sum\nolimits_{j=1}^q w_j=0, \text{ where } \varphi_j(w_j)=\min \text{ in } (\bar{P}_j(\bar{w}_j)).$$

The solutions to $(\bar{P}_0)$ are the $\bar{w}$ components of the solutions to $(\bar{P})$. Although $(\bar{P}_0)$ would generally be a *nonsmooth* problem of convex optimization, hope in solving it could come from generating local information about φ at a point w^ν through solving the subproblems $(\bar{P}_j(w_j^\nu))$.

The new algorithm that will be presented at the end of this section resembles such a scheme but adds proximal terms to the subproblems with respect to current iterates x_j^ν. Furthermore it utilizes the augmented Lagrangians for these subproblems. It will actually be derived from a dual approach to $(\bar{P})$, but with a different dual problem than might be expected.

It would be natural to get a Lagrangian function for $(\bar{P})$ by attaching multipliers to the various constraints indexed by i and j as well as a multiplier vector for $w_1+\cdots+w_q=0$. However, instead of bringing in a multiplier vector for $w_1+\cdots+w_q=0$, we'll handle that condition by introducing

$$W=\{\,w=(w_1,\dots,w_q)\,|\,w_1+\cdots+w_q=0\}\subset(I\!R^m)^q. \tag{4.1}$$

as a subspace of $(I\!R^m)^q$ and imposing the constraint $w\in W$ "geometrically" alongside of the constraints $x_j\in X_j$.

This departure from traditional patterns will be accompanied by a different approach to duality. Crucial to that will be subspace $W^\perp$ that's the orthogonal complement of W, given by

$$W^\perp=\{\,w=(w_1,\dots,w_q)\,|\,w_1=\cdots=w_q\}\subset(I\!R^m)^q. \tag{4.2}$$

Every $w \in (I\!R^m)^q$ can of course be expressed uniquely as a sum of an element of W and an element of $W^\perp$, namely its orthogonal projections on those subspaces. The projection onto $W^\perp$ is given by

$$P_{W^\perp}(w) = (w_*, \ldots, w_*) \;\text{ for }\; w_* = \frac{1}{q}\sum\nolimits_{j=1}^q w_j, \tag{4.3}$$

and the projection onto W is accordingly given by

$$P_W(w) = w - P_{W^\perp}(w) = (w_1 - w_*, \ldots, w_q - w_*). \tag{4.4}$$

In the general duality theory of convex optimization as developed in [7] and [8], there are different duals for a given primal problem according to how perturbations are introduced. The canonical perturbations that yield the standard dual for problem (P) are the additions of increments u_i to the quantities being constrained by the inequalities and equations in (P). We'll follow that here by adding increments u_{ij} to the corresponding constraints in $(\bar{P})$, but besides that we'll add to $w \in W$ a perturbation $v \in W^\perp$. We'll then have an optimization problem in the variables $(w, x) \in W \times I\!R^n$ that responds to perturbations $(u, v) \in (I\!R^m)^q \times W^\perp$, where

$$u = (u_1, \ldots, u_q) \;\text{ with }\; u_j = (u_{1j}, \ldots, u_{mj}) \;\text{ and }\; v = (v_1, \ldots, v_q) \;\text{ with }\; v_j = (v_{1j}, \ldots, v_{mj}),$$

but actually

$$v_1 = \cdots = v_q = v_*, \;\text{ so that }\; v_{i1} = \cdots = v_{iq} = v_{*i}.$$

(It may seem pointless to have $v_1, \ldots, v_q$ when just $v_* = (v_{*1}, \ldots, v_{*m})$ would suffice, but the redundant notation will help to keep things straight in derivations with respect to the subspaces W and $W^\perp$, and anyway will drop out in the end.)

This optimization framework is encoded by a function $f : W \times I\!R^n \times (I\!R^m)^q \times W^\perp \to (-\infty, \infty]$ with

$$\begin{aligned} &f(w,x,u,v) = f_{01}(x_1) + \cdots + f_{0q}(x_q) \;\text{ if for each } j \text{ one has }\; x_j \in X_j \;\text{ and}\\ &\qquad\qquad f_{ij}(x_j) - w_{ij} - v_{ij} + u_{ij} \begin{cases} \le 0 \;\text{ for } i = 1, \ldots, s,\\ = 0 \;\text{ for } i = s+1, \ldots, m, \end{cases}\\ &\quad \text{but }\; f(w,x,u,v) = \infty \;\text{ for all other }\; (x,w,u,v) \in I\!R^n \times W \times (I\!R^m)^q \times W^\perp. \end{aligned} \tag{4.5}$$

The optimization problem corresponding to a given $(u, v) \in (I\!R^m)^q \times W^\perp$ is to minimize $f(w, x, u, v)$ over $(w, x) \in W \times I\!R^n$ (in which the constraints in (4.5) are enforced by an infinite penalty for their violation). Problem $(\bar{P})$ itself corresponds to $(u, v) = (0, 0)$.

The Lagrangian function associated with this framework of perturbations is an extended-real-valued function $\bar{L}$ on $W \times I\!R^n \times (I\!R^m)^q \times W^\perp$ involving a multiplier pair $(\eta, \zeta) \in (I\!R^m)^q \times W^\perp$, where, much as above,

$$\eta = (\eta_1, \ldots, \eta_q) \;\text{ with }\; \eta_j = (\eta_{1j}, \ldots, \eta_{mj}) \;\text{ and }\; \zeta = (\zeta_1, \ldots, \zeta_q) \;\text{ with }\; \zeta_j = (\zeta_{1j}, \ldots, \zeta_{mj}),$$

but actually

$$\zeta_1 = \cdots = \zeta_q = \zeta_*, \;\text{ so that }\; \zeta_{i1} = \cdots = \zeta_{iq} = \zeta_{*i}.$$

(We are employing η and ζ here in order to have y and z available for a better use which will come to light later.) It is defined by

$$\bar{L}(w, x, \eta, \zeta) = \inf_{u,v} \{\, f(w, x, u, v) - \eta{\cdot}u + \zeta{\cdot}v \,\} \tag{4.6}$$

and is convex in (w, x) and concave in (η, ζ). This works out in terms of the convex cone

$$Y^q = Y \times \cdots \times Y \subset (I\!R^m)^q, \tag{4.7}$$

and the subspace[2]

$$S = \{ (\eta, \zeta) \,|\, P_{W^\perp}(\eta) = \zeta \} \subset (I\!R^m)^q \times W^\perp, \tag{4.8}$$

and the functions

$$L_j(x_j, \eta_j) = f_{0j}(x_j) + \sum\nolimits_{i=1}^m \eta_{ij} f_{ij}(x_j) \tag{4.9}$$

to mean that

$$\bar{L}(w, x, \eta, \zeta) = \begin{cases} \sum_{j=1}^q [L_j(x_j, \eta_j) - \eta_j \cdot w_j] & \text{if } x \in X,\ (\eta, \zeta) \in S \cap [Y^q \times W^\perp], \\ -\infty & \text{if } x \in X,\ (\eta, \zeta) \notin S \cap [Y^q \times W^\perp], \\ +\infty & \text{if } x \notin X. \end{cases} \tag{4.10}$$

The corresponding dual problem is

$$(\bar{D}) \qquad \text{maximize } \bar{g}(\eta, \zeta) = \inf_{(w,x) \in X \times W} \bar{L}(w, x, \eta, \zeta) \text{ over all } (\eta, \zeta) \in (I\!R^m)^q \times W^\perp,$$

where the objective function $\bar{g}$ is concave (and upper semicontinuous) and the implicit feasible set is

$$\{ (\eta, \zeta) \,|\, g(\eta, \zeta) > -\infty \} \subset S \cap [Y^q \times W^\perp]. \tag{4.11}$$

Just as before with (P) and (D), the product of the set of solutions $(\bar{w}, \bar{x})$ to $(\bar{P})$ and the set of solutions $(\bar{\eta}, \bar{\zeta})$ to $(\bar{D})$ is the set of saddle points $(\bar{w}, \bar{x}, \bar{\eta}, \bar{\zeta})$ of $\bar{L}(w, x, \eta, \zeta)$ with respect to minimizing in (w, x) and maximizing in (η, ζ).

This saddle point charactization has a consequence for the multiplier vectors, which is important to record. If $(\bar{w}, \bar{x}, \bar{\eta}, \bar{\zeta})$ is a saddle point of the Lagrangian in (4.10), then in particular the infimum of $w \cdot \bar{\eta}$ over $w \in W$ can't be $-\infty$ and therefore has to be 0. This tells us that $\bar{\eta} \in W^\perp$, so that $\bar{\eta}_1 = \cdot = \bar{\eta}_q$. Thus, although the expanded scheme makes way for possibly different multiplier vectors in the L_j's, that flexibility, helpful for developing an algorithm, doesn't lead to anything different than the single multiplier vector in (P).

More details on expressing $\bar{g}(\eta, \zeta)$ could be worked out in particular circumstances, but in the long run they don't really matter. What matters here is that the proximal point algorithm can be applied for solving $(\bar{D})$ just as for (D) to produce an associated "method of multipliers," which moreover can then also be turned into a "proximal method of multipliers" — because the theory for all of that works very generally. According to [12], proximal methods of multipliers are obtained by applying a saddle point procedure to a Lagrangian to which both primal and dual proximal terms have been added. Here we will focus on the following case, where for now we use $\rho > 0$ as the parameter instead of r for reasons like those behind η and ζ but also introduce a second parameter $c > 0$.

Proximal saddle point algorithm in the decomposition framework. *Generate a sequence of elements* $(w^\nu, x^\nu) \in W \times X$ *and* $(\eta^\nu, \zeta^\nu) \in S \cap [Y^q \times W^\perp]$ *by letting*

$$\bar{L}^\nu(w, x, \eta, \zeta) = \bar{L}(w, x, \eta, \zeta) + \frac{\rho}{2}||w - w^\nu||^2 + \frac{1}{2c}||x - x^\nu||^2 - \frac{1}{2\rho}||\eta - \eta^\nu||^2 - \frac{1}{2\rho}||\zeta - \zeta^\nu||^2 \tag{4.12}$$

[2] Ordinarily there would be a "$-$" in front of the ζ term in (4.6) but the sign has harmlessly been switched to avoid having $-\zeta$ in place of ζ in describing this subspace.

and calculating

$$\begin{array}{l}(w^{\nu+1}, x^{\nu+1}, \eta^{\nu+1}, \zeta^{\nu+1}) = \textit{unique saddle point of } \bar{L}^{\nu}(w,x,\eta,\zeta) \textit{ with respect to} \\ \quad \textit{minimizing over } (w,x) \in W \times X \textit{ and maximizing over } (\eta,\zeta) \in S \cap [Y^q \times W^{\perp}].\end{array} \tag{4.13}$$

The unique saddle point exists because the proximal terms make $\bar{L}^{\nu}$ be strongly convex in the primal variables and strongly concave in the dual variables. On the basis of [12], *the sequence of elements* $(w^{\nu}, x^{\nu}, \eta^{\nu}, \zeta^{\nu})$ *generated in this manner from any initial* $(w^1, x^1) \in W \times X$ *and* $(\eta^1, \zeta^1) \in S \cap [Y \times W^{\perp}]$ *is sure to converge to some saddle point* $(\bar{w}, \bar{x}, \bar{\eta}, \bar{\zeta})$ *of the Lagrangian* $\bar{L}$. Then $(\bar{w}, \bar{x})$ solves $(\bar{P})$ while $(\bar{\eta}, \bar{\zeta})$ solves $(\bar{D})$, as noted.

It might be wondered if the application of convergence results in [12] to the procedure just described is legitimate, because those results correspond to having $1/2\rho$ instead of $\rho/2$ for the primal proximal terms as well as the dual proximal terms. But this version simply corresponds to a rescaling of variables. That rescaling is critical for our purposes, as will soon be evident.

Validating the convergence of the saddle point iterations in our framework is, of course, only a first step on the way towards this furnishing a decomposition method for solving $(\bar{P})$. It must be shown that the calculation of the saddle point in (4.13) can be carried out in a manner that breaks down in subproblems indexed by $j = 1, \ldots, q$ which can be solved in parallel.

Quite remarkably, the proximal saddle point algorithm comes out, in the end, as something that seems entirely different and utilizes the augmented Lagrangians for the subproblems $(\bar{P}_j(w_j))$ at a level $r > 0$ different from ρ and with multipliers y_{ij} different from η_{ij}, namely

$$\begin{array}{l}\bar{L}_{j,r}(w_j, x_j, y_j) = f_{0j}(x_j) + \sum_{i=1}^{m} \psi_i(f_{ij}(x_j) - w_{ij}, y_{ij}, r), \ \text{ where} \\ \psi_i(f_{ij}(x_j) - w_{ij}, y_{ij}, r) = \begin{cases} -\frac{1}{2r}|y_{ij}|^2 \ \text{ for } i \leq s \text{ with } y_{ij} + r[f_{ij}(x_j) - w_{ij}] \leq 0, \\ y_{ij}[f_{ij}(x_j) - w_{ij}] + \frac{r}{2}|f_{ij}(x_j) - w_{ij}|^2 \ \text{ otherwise.} \end{cases}\end{array} \tag{4.14}$$

Augmented decomposition algorithm. *In iteration* ν, *having* $w^{\nu} \in W$, $x^{\nu} \in X$ *and* $y^{\nu} \in Y^q$, *determine the components of* $x^{\nu+1} \in X$ *by*

$$x_j^{\nu+1} = \operatorname*{argmin}_{x_j \in X_j} \Big\{ \bar{L}_{j,r}(w_j^{\nu}, x_j, y_j^{\nu}) + \frac{1}{2c}||x_j - x_j^{\nu}||^2 \Big\}. \tag{4.15}$$

Then, in terms of

$$\begin{cases} \eta_{ij} = \theta_i(y_{ij}^{\nu} + r[f_{ij}(x_j^{\nu+1}) - w_{ij}^{\nu}]\,), \\ \zeta_i = \frac{1}{q}\sum_{j=1}^{q} \eta_{ij}, \end{cases} \quad \textit{where } \ \theta_i(t) = \begin{cases} \max\{0,t\} & \textit{for } i = 1, \ldots, s, \\ \quad t & \textit{for } i = s+1, \ldots, m, \end{cases} \tag{4.16}$$

update to $w^{\nu+1} \in W$ *and* $y^{\nu+1} \in Y^q$ *by*

$$w_{ij}^{\nu+1} = w_{ij}^{\nu} + \frac{1}{2r}[\eta_{ij} - \zeta_i], \qquad y_{ij}^{\nu+1} = \frac{1}{2}[\eta_{ij} + \zeta_i]. \tag{4.17}$$

Convergence theorem. *For any* $r > 0$ *and* $c > 0$, *the sequence* $\{(w^{\nu}, x^{\nu}, y^{\nu})\}_{\nu=1}^{\infty}$ *generated in* $W \times X \times Y^q$ *by the augmented decomposition algorithm from any starting point converges to some* $(\bar{w}, \bar{x}, \bar{y})$ *such that*

(a) $(\bar{w}, \bar{x})$ *solves* $(\bar{P})$, *hence* $\bar{x}$ *solves* (P),

(b) $\bar{y}_1 = \cdots = \bar{y}_q \in I\!R^m$, *and this common multiplier vector solves* (D).

In the next section we move on to analyzing the calculation of the saddle point in (4.13) and uncovering how the proximal saddle point algorithm can be executed conveniently. By arriving finally at the recognition that the steps can be identified with the ones in the augmented decomposition algorithm, we will have managed then also to prove the convergence theorem.

5 Justification of the augmented decomposition algorithm

The requirement that $(\eta,\zeta)\in S$ on the maximization side of the saddle point problem in (4.13) can be handled by subtracting the indicator term $\delta_S(\eta,\zeta)$ from $\bar{L}^\nu(x,w,\eta,\zeta)$ and using the fact that

$$-\delta_S(\eta,\zeta)=\inf_{v\in W^\perp} v{\cdot}(\zeta-\eta)\ \text{ for }\ (\eta,\zeta)\in Y^q\times W^\perp.$$

From this angle, and in bringing in the expression for $\bar{L}$ in terms of the "sub-Lagrangians" in (4.9), it is apparent that we are really dealing with a saddle point of

$$\begin{array}{l}\sum_{j=1}^q \bar{L}_j(x_j,\eta_j)-w{\cdot}\eta+v{\cdot}(\zeta-\eta)\\ +\frac{1}{2c}||x-x^\nu||^2+\frac{\rho}{2}||w-w^\nu||^2-\frac{1}{2\rho}||\eta-\eta^\nu||^2-\frac{1}{2\rho}||\zeta-\zeta^\nu||^2\end{array}\tag{5.1}$$

with respect to minimizing in $x\in X$, $w\in W$ and $v\in W^\perp$ while maximizing in $\eta\in\bar{Y}$ and $\zeta\in W^\perp$.

In this situation the various maximizations and minimizations can be carried out in any order, due to the strong convexity and strong concavity induced by the proximal terms (even though there isn't one for v). Our tactic now will be to fix $x\in X$ and $\eta\in Y^q$ temporarily and try to understand how the saddle point with respect to the other elements in (5.1) will depend on (x,η). We are concerned then just with saddle points in (w,v) and ζ of

$$-w{\cdot}\eta+v{\cdot}(\zeta-\eta)+\frac{\rho}{2}||w-w^\nu||^2-\frac{1}{2\rho}||\zeta-\zeta^\nu||^2.\tag{5.2}$$

The maximization in ζ is easy to carry out first, because it separately involves only $v{\cdot}\zeta$ and the proximal term in ζ. Since both v and ζ are in $W^\perp$, it just amounts to calculating the conjugate of that proximal term:

$$\max_{\zeta\in W^\perp}\left\{v{\cdot}\zeta-\frac{1}{2\rho}||\zeta-\zeta^\nu||^2\right\}=v{\cdot}\zeta^\nu+\frac{\rho}{2}||v||^2\ \text{ with }\ \zeta=\zeta^\nu+\rho v\ \text{ as the argmax.}\tag{5.3}$$

The elimination of ζ in this manner replaces the expression in (5.2) by

$$-w{\cdot}\eta-v{\cdot}\eta+v{\cdot}z^\nu+\frac{\rho}{2}||w-w^\nu||^2+\frac{\rho}{2}||v||^2,\tag{5.4}$$

which is to be minimized in $w\in W$ and $v\in W^\perp$. Because $w\perp z^\nu$ we have $v{\cdot}z^\nu=(w+v){\cdot}z^\nu$, and because $(w-w^\nu)\perp v$, we have $||w-w^\nu||^2+||v||^2=||(w+v)-w^\nu||^2$. Therefore (5.4) can be consolidated to

$$-(w+v){\cdot}(\eta-\zeta^\nu)+\frac{\rho}{2}||(w+v)-w^\nu||^2.\tag{5.5}$$

An important observation now is that $w+v$ stands here for a general element

$$\omega\in(I\!R^m)^q\ \text{ with }\ w=P_W(\omega)\ \text{ and }\ v=P_{W^\perp}(\omega).\tag{5.6}$$

Thus in (5.5) we are left with $-\omega{\cdot}(\eta-\zeta^\nu)+\frac{\rho}{2}||\omega-w^\nu||^2$, which is to be minimized with respect to $\omega\in(I\!R^m)^q$, where

$$\begin{array}{c}\min_\omega\left\{-\,\omega{\cdot}(\eta-\zeta^\nu)+\frac{\rho}{2}||\omega-w^\nu||^2\right\}=-(\eta-\zeta^\nu){\cdot}w^\nu-\frac{1}{2\rho}||\eta-\zeta^\nu||^\nu\\ \text{with argmax }\ \omega=w^\nu+\frac{1}{\rho}(\eta-\zeta^\nu).\end{array}\tag{5.7}$$

This brings us back to the full saddle point problem in (4.13) with the understanding that, in first passing to the minimax with respect to w and z for a given x and η, the residual expression will be

$$\textstyle\sum_{j=1}^{q} L_j(x_j,\eta_j) - (\eta-\zeta^\nu){\cdot}w^\nu - \frac{1}{2\rho}||\eta-\zeta^\nu||^2 + \frac{\rho}{2}||x-x^\nu||^2 - \frac{1}{2\rho}||\eta-\eta^\nu||^2, \tag{5.8}$$

where the "sub-Lagrangians" of (4.9) in the formula for the Lagrangian in (4.10) appear. The key now is the fact that this expression is the sum of separate expressions involving only the variable associated with each j:

$$\textstyle\sum_{j=1}^{q} \Big[L_j(x_j,\eta_j) - (\eta_j-\zeta_j^\nu){\cdot}w_j^\nu - \frac{1}{2\rho}||\eta_j-\zeta_j^\nu||^2 + \frac{\rho}{2}||x_j-x_j^\nu||^2 - \frac{1}{2\rho}||\eta_j-\eta_j^\nu||^2\Big]. \tag{5.9}$$

Finding a saddle point of this with respect to $x\in X$ and $\eta\in Y^q$ *decomposes into finding for each j a saddle point with respect to $x_j\in X_j$ and $\eta_j\in Y$.*

Pursuing this in more detail, we can omit the terms $\zeta^\nu{\cdot}w_j^\nu$ in (5.9) as being constants in the minimax, hence irrelevant. By defining

$$y_j^\nu = \frac{1}{2}[\eta_j^\nu + \zeta_j^\nu], \qquad \delta_j^\nu = \frac{1}{2}[\eta_j^\nu - \zeta_j^\nu], \tag{5.10}$$

we can invoke the identity that

$$\frac{1}{2}||\eta_j-\zeta_j^\nu||^2 + \frac{1}{2\rho}||\eta_j-\eta_j^\nu||^2 = ||\eta-y^\nu||^2 + ||\delta^\nu||^2, \tag{5.11}$$

where $||\delta^\nu||^2$ likewise behaves for the moment as a constant. Then by introducing

$$L_j^\nu(x_j,\eta_j) = f_{0j}(x_j) + \textstyle\sum_{i=1}^{m}\eta_{ij}f_{ij}^\nu(x_j) \;\text{ for }\; f_{ij}^\nu(x_j) = f_{ij}(x_j) - w_{ij}^\nu, \tag{5.12}$$

we can focus our consideration for $j=1,\dots,q$ on determining a saddle point of

$$L_j^\nu(x_j,\eta_j) - \frac{1}{2r}||\eta_j-y_j^\nu||^2 + \frac{1}{2c}||x_j-x_j^\nu||^2, \;\text{ where }\; r=\rho/2. \tag{5.13}$$

To get a saddle point we can maximize first in $\eta_j\in Y$ to get the max and argmax as functions of x_j, and then minimize that max value over $x_j\in X_j$.

In fact this returns us to territory we are already familiar with from the earlier discussion of augmented Lagrangians, since

$$\max_{\eta_j\in Y}\Big\{\, L_j^\nu(x_j,\eta_j) - \frac{1}{2r}||\eta_j-y_j^\nu||^2\Big\} = L_{j,r}(w_j^\nu,x_j,y_j^\nu), \tag{5.14}$$

where $L_{j,r}(w_j,x_j,y_j)$ is the r-augmented Lagrangian for the optimization subproblem $(\bar{P}(w_j))$ already introduced in (4.14). It confirms that $x^{\nu+1}$ is obtained by the minimization step in (4.15). Moreover the vector η_j furnishing the maximum in (5.14) with respect x_j^ν is given coordinatewise by

$$\eta_{ij}^{\nu+1} = \theta_i(y_{ij}^\nu + r[f_{ij}(x_j^\nu) - w_{ij}^\nu]), \;\text{ where }\; \theta_i(t) = \begin{cases} \max\{0,t\} & \text{for } i=1,\dots,s, \\ \quad t & \text{for } i=s+1,\dots,m. \end{cases} \tag{5.15}$$

From having determined $\eta^{\nu+1}$ this way we get $\zeta^{\nu+1}$ as its projection on $W^\perp$, namely

$$\zeta_{ij}^{\nu+1} = \zeta_{i*}^{\nu+1} \;\text{ for all } j, \text{ where }\; \zeta_{i*}^{\nu+1} = \frac{1}{q}\textstyle\sum_{j=1}^{q}\eta_{ij}^{\nu+1}, \tag{5.16}$$

and therefore, via (5.10),

$$y^{\nu+1} = \frac{1}{2}[\eta^{\nu+1} + \zeta^{\nu+1}]. \tag{5.17}$$

On the other hand, from (5.7) we get $\omega^{\nu+1} = \rho^{-1}(\eta^{\nu+1} - \zeta^{\nu})$, hence, since $z^{\nu} \in W^{\perp}$ and $\rho = 2r$,

$$w^{\nu+1} = P_W(\omega^{\nu+1}) = \frac{1}{2r}P_W(\eta^{\nu+1} - \zeta^{\nu}) = \frac{1}{2r}P_W(\eta^{\nu+1}) = \frac{1}{2r}[\eta^{\nu+1} - \zeta^{\nu+1}]. \tag{5.18}$$

This validates the updates to $w^{\nu+1}$ and $y^{\nu+1}$ in (4.17), where the notation has been simplified from $\eta_{ij}^{\nu+1}$ and $\zeta_{ij}^{\nu+1}$ because there is no need any more to carry those vectors along explicitly in the statement of the procedure.

We already know from the discussion following the introduction of the Lagrangian in (4.10), that the limit of $\bar{\eta}$ must lie in $W^{\perp}$, making $\bar{\eta}_1 = \cdots = \bar{\eta}_q$, with this the same as the vector $\bar{\zeta}_1 = \cdots = \bar{\zeta}_1$. Obviously that then entails $\bar{y}_1 = \cdots = \bar{y}_q$.

References

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

[2] Bertsekas, D. P., "Incremental aggregated proximal and augmented Lagrangian algorithms." Lab. for Information and Decision Systems Report LIDS-P-3176, Massachusetts Institute of Technology, 2015, arXiv:1608.01393.

[3] Dantzig, G., Wolfe, P., "The decomposition algorithm for linear programming," *Econometrica* **29** (1961).

[4] Eckstein, J., "A simplified form of block-iterative operator splitting and an asynchronous algorithm resembling the multiblock ADMM," *J. Optimization Theory and Applications* **173** (2017), 155–182.

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

[6] Hong, M.G., Luo, Z.Q., and Razaviyayn, M., "Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems." *SIAM J. Optimization* **26** (2016), 33–364.

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

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

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

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

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

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

[13] Rockafellar, R. T., and Sun, Jie “Solving monotone stochastic variational inequalities and complementarity problems by progressive hedging,” *Set-Valued and Variational Analysis* **26** (2017).

[14] Rockafellar, R. T., and Wets, R. J-B, “Scenarios and policy aggregation in optimization under uncertainty.” *Mathematics of Operations Research* **16** (1991), 119–147.

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

[16] Rockafellar, R. T., and Wets, R. J-B, “Stochastic variational inequalities: single-stage to multistage.” *Mathematical Programming B* (2016).

[17] Spingarn, J. E., “Applications of the method of partial inverses to convex programming: decomposition,” *Mathematical Programming* **32** (1983), 199–223.