Decomposition Techniques for Bilinear Saddle Point Problems and Variational Inequalities with Affine Monotone Operators on Domains Given by Linear Minimization Oracles

Bruce Cox* Anatoli Juditsky† Arkadi Nemirovski‡

June 15, 2015

Abstract

The majority of First Order methods for large-scale convex-concave saddle point problems and variational inequalities with monotone operators are *proximal* algorithms which at every iteration need to minimize over problem's domain X the sum of a linear form and a strongly convex function. To make such an algorithm practical, X should be *proximal-friendly* – admit a strongly convex function with easy to minimize linear perturbations. As a byproduct, X admits a computationally cheap Linear Minimization Oracle (LMO) capable to minimize over X linear forms. There are, however, important situations where a cheap LMO indeed is available, but X is not proximal-friendly, which motivates search for algorithms based solely on LMO's. For smooth convex minimization, there exists a classical LMO-based algorithm – Conditional Gradient. In contrast, known to us LMO-based techniques [2, 14] for other problems with convex structure (nonsmooth convex minimization, convex-concave saddle point problems, even as simple as bilinear ones, and variational inequalities with monotone operators, even as simple as affine) are quite recent and utilize common approach based on Fenchel-type representations of the associated objectives/vector fields. The goal of this paper is to develop an alternative (and seemingly much simpler) *LMO-based decomposition techniques* for *bilinear* saddle point problems and for variational inequalities with *affine* monotone operators.

1 Introduction

This paper is a follow-up to our paper [14] and, same as its predecessor, is motivated by the desire to develop first order algorithms for solving convex-concave saddle point problem (or variational inequality with monotone operator) on a convex domain X represented by *Linear Minimization Oracle* (LMO) capable to minimize over X, at a reasonably low cost, any linear function. "LMO-representability" of a convex domain X is an essentially weaker assumption than "proximal friendliness" of X (possibility to minimize over X, at a reasonably low cost, any linear perturbation of a properly selected strongly convex function) underlying the vast majority of known first order algorithms. There are important applications giving rise to LMO-represented domains which are *not* proximal friendly, most notably

*US Air Force

†LJK, Université J. Fourier, B.P. 53, 38041 Grenoble Cedex 9, France, Anatoli.Juditsky@imag.fr

‡Georgia Institute of Technology, Atlanta, Georgia 30332, USA, nemirovs@isye.gatech.edu

Research of the third author was supported by the NSF grants CMMI-1232623, CCF-1415498, CMMI-1262063.

- nuclear norm balls arising in low rank matrix recovery and in Semidefinite optimization; here LMO reduces to approximating the leading pair of singular vectors of a matrix, while all known proximal algorithms require much costly computationally full singular value decomposition,
- total variation balls arising in image reconstruction; here LMO reduces to solving a specific flow problem [11], while a proximal algorithm needs to solve a much more computationally demanding linearly constrained convex quadratic program,
- some combinatorial polytopes

The needs of there applications inspire the current burst of activity in developing LMO-based optimization techniques. In its major part, this activity was focused on Smooth (or Lasso-type smooth regularized) Convex Minimization over LMO-represented domains, where the classical Conditional Gradient algorithm of Frank & Wolfe [5] and its modifications are applicable (see, e.g., [3, 4, 6, 7, 10, 11, 12, 13, 16] and references therein). LMO-based techniques for large-scale Nonsmooth Convex Minimization (NCM), convex-concave Saddle Point problems (SP), even bilinear ones, and Variational Inequalities (VI) with monotone operators, even affine ones, where no classical optimization methods work, have been developed only recently; to the best of our knowledge, related literature reduces to [2] (NCM) and [14] (SP, VI), where a specific approach, based on Fenchel-type representations of convex functions and monotone operators, was developed. The goal of this paper is to develop an alternative to [14] *decomposition-based* approach to solving convex-concave SP's and monotone VI's on LMO-represented domains. In the sequel, we focus on *bilinear* SP's and on VI's with *affine* operators – the cases which, on one hand, are of primary importance in numerous applications, and, on the other hand, are the cases where our decomposition approach is easy to implement and where this approach seems to be more flexible and much simpler than the machinery of Fenchel-type representations developed in [14].

The rest of this paper is organized as follows. In section 2 we present our decomposition-based approach for bilinear SP problems, while section 3 deals with decomposition of VI's with affine monotone operators; in both cases, our emphasis is on utilizing the approach to handle problems on LMO-represented domains. We demonstrate also (section 3.6) that in the context of VI's with monotone operators on LMO-represented domains, our current approach covers the one developed in [14].
Proofs missing in the main body of the paper are relegated to Appendix.

2 Decomposition of Convex-Concave Saddle Point Problems

2.1 Situation

Let $X_i \subset E_i$, $Y_i \subset F_i$, $i = 1, 2$, be convex compact sets in Euclidean spaces, let

$$X \subset X_1 \times X_2,\, Y \subset Y_1 \times Y_2, E = E_1 \times E_2,\, F = F_1 \times F_2, Z = X \times Y, G = E \times F,$$

with convex compact X and Y such that the projections of X onto E_i are the sets X_i, and projections of Y onto F_i are the sets Y_i, $i = 1, 2$. For $x_1 \in X_1$, we set $X_2[x_1] = \{x_2 : [x_1; x_2] \in X\} \subset X_2$, and for $y_1 \in Y_1$ we set $Y_2[y_1] = \{y_2 : [y_1; y_2] \in Y\} \subset Y_2$. Similarly,

$$X_1[x_2] = \{x_1 : [x_1; x_2] \in X\},\, x_2 \in X_2, \text{ and } Y_1[y_2] = \{y_2 : [y_1; y_2] \in Y\},\, y_2 \in Y_2.$$

Let also

$$\Phi(x = [x_1; x_2]; y = [y_1; y_2]) : X \times Y \to \mathbf{R} \tag{1}$$

be Lipschitz continuous convex in $x \in X$ and concave in $y \in Y$ functions.

We call the outlined situation a *direct product* one, when $X = X_1 \times X_2$ and $Y = Y_1 \times Y_2$.

2.2 Induced convex-concave function

We associate with Φ *primal and dual induced* functions:

$$\begin{array}{rcl}\phi(x_1,y_1) & := & \min\limits_{x_2\in X_2[x_1]}\max\limits_{y_2\in Y_2[y_1]}\Phi(x_1,x_2;y_1,y_2)=\max\limits_{y_2\in Y_2[y_1]}\min\limits_{x_2\in X_2[x_1]}\Phi(x_1,x_2;y_1,y_2):X_1\times Y_1\to\mathbf{R},\\ \psi(x_2,y_2) & := & \min\limits_{x_1\in X_1[x_2]}\max\limits_{y_1\in Y_1[y_2]}\Phi(x_1,x_2;y_1,y_2)=\max\limits_{y_1\in Y_1[y_2]}\min\limits_{x_1\in X_1[x_2]}\Phi(x_1,x_2;y_1,y_2):X_2\times Y_2\to\mathbf{R}.\end{array}$$

(the equalities are due to the convexity-concavity and continuity of Φ and convexity and compactness of $X_i[\cdot]$ and $Y_i[\cdot]$).

Recall that a Lipschitz continuous convex-concave function $\theta(u,v):U\times V\to\mathbf{R}$ with convex compact U,V gives rise to the primal and dual problems

$$\begin{array}{rcl}\mathrm{Opt}(P[\theta,U,V]) & = & \min\limits_{u\in U}\left[\overline{\theta}(u):=\max\limits_{v\in V}\theta(u,v)\right]\\ \mathrm{Opt}(D[\theta,U,V]) & = & \max\limits_{v\in V}\left[\underline{\theta}(v):=\min\limits_{u\in U}\theta(u,v)\right]\end{array}$$

with equal optimal values:

$$\mathrm{SadVal}(\theta,U,V):=\mathrm{Opt}(P[\theta,U,V)]=\mathrm{Opt}(D[\theta,U,V]),$$

same as gives rise to *saddle point residual*

$$\epsilon_{\mathrm{sad}}([u,v]|\theta,U,V)=\overline{\theta}(u)-\underline{\theta}(v)=\left[\overline{\theta}(u)-\mathrm{Opt}(P[\theta,U,V])\right]+\left[\mathrm{Opt}(D[\theta,U,V])-\underline{\theta}(v)\right].$$

Lemma 1. *ϕ and ψ are convex-concave on their domains, are lower (upper) semicontinuous in their "convex" ("concave") arguments, and are Lipschitz continuous in the direct product case. Besides this, it holds*

$$\mathrm{SadVal}(\phi,X_1,Y_1)=\mathrm{SadVal}(\Phi,X,Y)=\mathrm{SadVal}(\psi,X_2,Y_2),\tag{2}$$

and whenever $\bar{x}=[\bar{x}_1;\bar{x}_2]\in X$ and $\bar{y}=[\bar{y}_1;\bar{y}_2]\in Y$, one has

$$\epsilon_{\mathrm{sad}}([\bar{x}_1;\bar{y}_1]|\phi,X_1,Y_1)\le\epsilon_{\mathrm{sad}}([\bar{x};\bar{y}]|\Phi,X,Y),\ \epsilon_{\mathrm{sad}}([\bar{x}_2;\bar{y}_2]|\psi,X_2,Y_2)\le\epsilon_{\mathrm{sad}}([\bar{x};\bar{y}]|\Phi,X,Y).\tag{3}$$

The strategy for solving SP problems we intend to develop is as follows:

1. We represent the SP problem of interest as the *dual SP problem*

$$\min_{x_2\in X_2}\max_{y_2\in Y_2}\psi(x_2,y_2)\tag{D}$$

   induced by *master SP problem*

$$\min_{[x_1;x_2]\in X}\max_{[y_1;y_2]\in Y}\Phi(x_1,x_2;y_1,y_2)\tag{M}$$

   The master SP problem is built in such a way that the associated *primal SP problem*

$$\min_{x_1\in X_1}\max_{y_1\in Y_1}\phi(x_1,y_1)\tag{P}$$

   admits First Order oracle and can be solved by a traditional First Order method (e.g., a proximal one).

2. We solve (P) to a desired accuracy by First Order algorithm producing *accuracy certificates* [15] and use these certificates to recover approximate solution of required accuracy to the problem of interest.

We shall see that the outlined strategy (originating from [1][1]) can be easily implemented when the problem of interest is a bilinear SP on the direct product of two LMO-represented domains.

2.3 Regular sub- and supergradients

Implementing the outlined strategy requires some "agreement" between the first order information of the master and the induced SP's, and this is the issue we address now.

Given $\bar{x}_1 \in X_1, \bar{y}_1 \in Y_1$, let $\bar{x}_2 \in X_2[\bar{x}_1]$, $\bar{y}_2 \in Y_2[\bar{y}_1]$ form a saddle point of the function $\Phi(\bar{x}_1, x_2; \bar{y}_1, y_2)$ (min in $x_2 \in X_2[\bar{x}_1]$, max in $y_2 \in Y_2[\bar{y}_1]$); in this situation, we say that $(\bar{x} = [\bar{x}_1; \bar{x}_2], \bar{y} = [\bar{y}_1; \bar{y}_2])$ belongs to the *saddle point frontier* of Φ, let this frontier be denoted by $\mathcal{S}$. Let now $\bar{z} = (\bar{x} = [\bar{x}_1; \bar{x}_2], \bar{y} = [\bar{y}_1; \bar{y}_2]) \in \mathcal{S}$, so that the function $\Phi(\bar{x}_1, x_2; \bar{y}_1, \bar{y}_2)$ attains its minimum over $x_2 \in X_2[\bar{x}_1]$ at $\bar{x}_2$, and the function $\Phi(\bar{x}_1, \bar{x}_2; \bar{y}_1, y_2)$ attains its maximum over $y_2 \in Y_2[\bar{y}_1]$ at $\bar{y}_2$. Consider a subgradient G of $\Phi(\cdot; \bar{y}_1, \bar{y}_2)$ taken at $\bar{x}$ along X: $G \in \partial_x \Phi(\bar{x}; \bar{y})$. We say that G is a *regular* subgradient of Φ at $\bar{z}$, if for some $g \in E_1$ it holds

$$\forall x = [x_1; x_2] \in X : \langle G, x - \bar{x} \rangle \geq \langle g, x_1 - \bar{x}_1 \rangle;$$

every g satisfying this relation is called *compatible* with G. Similarly, we say that a supergradient H of $\Phi(\bar{x}; \cdot)$, taken at $\bar{y}$ along Y, is a regular supergradient of Φ at $\bar{z}$, if for some $h \in F_1$ it holds

$$\forall y = [y_1; y_2] \in Y : \langle H, y - \bar{y} \rangle \leq \langle h, y_1 - \bar{y}_1 \rangle,$$

and every h satisfying this relation will be called *compatible* with H.

Remark 1. *Let the direct product case $X = X_1 \times X_2$, $Y = Y_1 \times Y_2$ take place. If $\Phi(x; \bar{y})$ is differentiable in x at $x = \bar{x}$, the partial gradient $\nabla_x \Phi(\bar{x}; \bar{y})$ is a regular subgradient of Φ at $(\bar{x}, \bar{y})$, and $\partial_{x_1} \Phi(\bar{x}; \bar{y})$ is compatible with this subgradient:*

$$\begin{array}{l} \forall x = [x_1; x_2] \in X_1 \times X_2 : \\ \langle \nabla_x \Phi(\bar{x}; \bar{y}), x - \bar{x} \rangle = \langle \nabla_{x_1} \Phi(\bar{x}; \bar{y}), x_1 - \bar{x}_1 \rangle + \underbrace{\langle \nabla_{x_2} \Phi(\bar{x}; \bar{y}), x_2 - \bar{x}_2 \rangle}_{\geq 0} \geq \langle \nabla_{x_1} \Phi(\bar{x}; \bar{y}), x_1 - \bar{x}_1 \rangle. \end{array}$$

Similarly, if $\Phi(\bar{x}; y)$ is differentiable in y at $y = \bar{y}$, then the partial gradient $\nabla_y \Phi(\bar{x}; \bar{y})$ is a regular supergradient of Φ at $(\bar{x}, \bar{y})$, and $\nabla_{y_1} \Phi(\bar{x}; \bar{y})$ is compatible with this supergradient.

Lemma 2. *In the situation of section 2.1, let $\bar{z} = (\bar{x} = [\bar{x}_1; \bar{x}_2], \bar{y} = [\bar{y}_1; \bar{y}_2]) \in \mathcal{S}$, let G be a regular subgradient of Φ at $\bar{z}$ and let g be compatible with G. Let also H be a regular supergradient of Φ at $\bar{z}$, and h be compatible with H. Then g is a subgradient in x_1, taken at $(\bar{x}_1, \bar{y}_1)$ along X_1, of the induced function ϕ, and h is a supergradient in y_1, taken at $(\bar{x}_1, \bar{y}_1)$ along Y_1, of the induced function ϕ:*

$$\begin{array}{ll} (a) & \phi(x_1, \bar{y}_1) \geq \phi(\bar{x}_1; \bar{y}_1) + \langle g, x_1 - \bar{x}_1 \rangle, \\ (b) & \phi(\bar{x}_1, y_1) \leq \phi(\bar{x}_1; \bar{y}_1) + \langle h, y_1 - \bar{y}_1 \rangle. \end{array}$$

for all $x_1 \in X_1$, $y_1 \in Y_1$.

[1] in hindsight, a special case of this strategy was used in [8, 9].

Regular sub- and supergradient fields of induced functions. In the sequel, we say that $\phi'_{x_1}(x_1,y_1)$, $\phi'_{y_1}(x_1,y_1)$ are *regular* sub- and supergradient fields of ϕ, if for every $(x_1,y_1)\in X_1\times Y_1$ and properly selected $\bar{x}_2$, $\bar{y}_2$ such that the point $\bar{z}=(\bar{x}=[x_1;\bar{x}_2],\bar{y}=[y_1;\bar{y}_2])$ is on the SP frontier of Φ, $\phi'_{x_1}(x_1,y_1)$, $\phi'_{y_1}(x_1,y_1)$ are the sub- and supergradients of ϕ induced, via Lemma 2, by regular sub- and supergradients of Φ at $\bar{z}$. Invoking Remark 1, we arrive at the following observation:

Remark 2. *If Φ is differentiable in x and in y and we are in the direct product case $X=X_1\times X_2$, $Y=Y_1\times Y_2$, then regular sub- and supergradients of ϕ can be built as follows: given $(x_1,y_1)\in X_1\times Y_1$, we find $\bar{x}_2$, $\bar{y}_2$ such that the point $\bar{z}=(\bar{x}=[x_1;\bar{x}_2],\bar{y}=[y_1;\bar{y}_2])$ is on the SP frontier of Φ, and set*

$$\phi'_{x_1}(x_1,y_1)=\nabla_{x_1}\Phi(x_1,\bar{x}_2;y_1,\bar{y}_2),\ \phi'_{y_1}(x_1,y_1)=\nabla_{y_1}\Phi(x_1,\bar{x}_2;y_1,\bar{y}_2). \tag{4}$$

2.3.1 Existence of regular sub- and supergradients

The notion of regular subgradient deals with Φ as a function of $[x_1;x_2]\in X$ only, the y-argument being fixed, so that the existence/description questions related to regular subgradient deal in fact with a Lipschitz continuous convex function on X. And of course the questions about existence/description of regular supergradients reduce straightforwardly to those on regular subgradients. Thus, as far as existence and description of regular sub- and supergradients is concerned, it suffices to consider the situation where

- $\Psi(x_1,x_2)$ is a Lipschitz continuous convex function on X,
- $\bar{x}_1\in X_1$, and $\bar{x}_2\in X_2[\bar{x}_1]$ is a minimizer of $\Psi(\bar{x}_1,x_2)$ over $x_2\in X_2[\bar{x}_1]$.

What we need to understand, is when a subgradient G of Ψ taken at $\bar{x}=[\bar{x}_1;\bar{x}_2]$ along X and some g satisfy the property

$$\langle G,[x_1;x_2]-\bar{x}\rangle\geq\langle g,x_1-\bar{x}_1\rangle\,\forall x=[x_1;x_2]\in X \tag{5}$$

and what can be said about the corresponding g's. The answer is as follows:

Lemma 3. *With Ψ, $\bar{x}_1$, $\bar{x}_2$ as above, $G\in\partial\Psi(\bar{x})$ satisfies (5) if and only if*

(i) G is a "certifying" subgradient of Ψ at $\bar{x}$, meaning that $\langle G,[0;x_2-\bar{x}_2]\rangle\geq 0\ \forall x_2\in X_2[\bar{x}_1]$ (the latter relation indeed certifies that $\bar{x}_2$ is a minimizer of $\Psi(\bar{x}_1,x_2)$ over $x_2\in X_2[\bar{x}_1]$);

(ii) g is a subgradient, taken at $\bar{x}_1$ along X_1, of the convex function

$$\chi_G(x_1)=\min_{x_2\in X_2[x_1]}\langle G,[x_1;x_2]\rangle$$

It is easily seen that with Ψ, $\bar{x}=[\bar{x}_1;\bar{x}_2]$ as in Lemma 3) (i.e., Ψ is convex and Lipschitz continuous on X, $\bar{x}_1\in X_1$, and $\bar{x}_2\in X_2[\bar{x}_1]$ minimizes $\Psi(\bar{x}_1,x_2)$ over $x_2\in X_2[\bar{x}_1]$) a certifying subgradient G always exists; when Ψ is differentiable at $\bar{x}$, one can take $G=\nabla_x\Psi(\bar{x})$. The function $\chi_G(\cdot)$, however, not necessary admits a subgradient at $\bar{x}_1$; when it does admit it, every $g\in\partial\chi_G(\bar{x}_1)$ satisfies (5). In particular,

1. [Direct Product case] When $X=X_1\times X_2$, representing a certifying subgradient G of Ψ, taken at $[\bar{x}_1;\bar{x}_2\in\mathrm{Argmin}_{x_2\in X_2}\Psi(\bar{x}_1,x_2)]$, as $[g;h]$, we have

   $$\langle h,x_2-\bar{x}_2\rangle\geq 0\ \forall x_2\in X_2,$$

   whence $\chi_G(x_1)=\langle g,x_1\rangle+\langle h,\bar{x}_2\rangle$, and thus g is a subgradient of χ_G at $\bar{x}_1$. In particular, in the direct product case and when Ψ is differentiable at $\bar{x}$, (5) is met by $G=\nabla\Psi(\bar{x})$, $g=\nabla_{x_1}\Psi(\bar{x})$;

2. [Polyhedral case] When X is a polyhedral set, for every certifying subgradient G of Ψ the function χ_G is polyhedrally representable with domain X_1 and as such has a subgradient at every point from X_1;
3. [Interior case] When $\bar{x}_1$ is a point from the relative interior of X_1, χ_G definitely has a subgradient at $\bar{x}_1$.

2.4 Main Lemma

2.4.1 Preliminaries: execution protocols, accuracy certificates, residuals

We start with outlining some simple concepts originating from [15]. Let Y be a convex compact set in Euclidean space L, and $F(y) : Y \to L$ be a vector field on E. A *t-step execution protocol* associated with F, Y is a collection $\mathcal{I}_t = \{y_i \in Y, F(y_i) : 1 \le i \le t\}$. A *$t$-step accuracy certificate* is a t-dimensional probabilistic vector λ. Augmenting a t-step accuracy protocol by t-step accuracy certificate gives rise to two entities:

$$\begin{array}{rl} \text{approximate solution:} & y^t = y^t(\mathcal{I}_t, \lambda) := \sum_{i=1}^t \lambda_i y_i \in Y; \\ \text{residual:} & \mathrm{Res}(\mathcal{I}_t, \lambda_t | Y) = \max\limits_{y \in Y} \sum_{i=1}^t \lambda_i \langle F(y_i), y_i - y \rangle. \end{array} \tag{6}$$

When $Y = U \times V$ and F is vector field induced by convex-concave function $\theta(u, v) : U \times V \to \mathbf{R}$, that is, $F(u, v) = [F_u(u, v); F_v(u, v)] : U \times V \to E \times F$ with $F_u(u, v) \in \partial_\theta(u, v)$, $F_v(u, v) \in \partial_v[-\theta(u, v)]$ (such a field always is monotone), an execution protocol associated with (F, Y) will be called also *protocol associated with θ, U, V*.

The importance of these notions in our context stems from the following simple observation [15]:

Lemma 4. *Let U, V be nonempty convex compact domains in Euclidean spaces E, F, $\theta(u, v) : U \times V \to \mathbf{R}$ be a convex-concave function, and F be induced monotone vector field: $F(u, v) = [F_u(u, v); F_v(u, v)] : U \times V \to E \times F$ with $F_u(u, v) \in \partial_u \theta(u, v)$, $F_v(u, v) \in \partial_v[-\theta(u, v)]$. For a t-step execution protocol $\mathcal{I}_t = \{y_i = [u_i; v_i] \in Y := U \times V, F_i = [F_u(u_i, v_i); F_v(u_i, v_i)], 1 \le i \le t\}$ associated with θ, U, V, and t-step accuracy certificate λ, it holds*

$$\epsilon_{\mathrm{sad}}(y^t(\mathcal{I}_t, \lambda) | \theta, U, V) \le \mathrm{Res}(\mathcal{I}_t, \lambda | U \times V). \tag{7}$$

Indeed, for $[u; v] \in U \times V$ we have

$$\begin{array}{l} \mathrm{Res}(\mathcal{I}_t, \lambda | U \times V) \ge \sum_{i=1}^t \lambda_i \langle F_i, y_i - [u; v] \rangle = \sum_{i=1}^t \lambda_i [\underbrace{\langle F_u(u_i, v_i), u_i - u \rangle}_{\ge \theta(u_i, v_i) - \theta(u, v_i)} - \underbrace{\langle F_v(u_i, v_i), v_i - v \rangle}_{\le \theta(u_i, v_i) - \theta(u_i, v)}] \\ \qquad\qquad \text{[inequalities are due to the origin of } F \text{ and convexity-concavity of } \theta] \\ \ge \sum_{i=1}^t \lambda_i [\theta(u_i, v) - \theta(u, v_i)] \ge \theta(u^t, v) - \theta(u, v^t), \end{array}$$

where the concluding $\ge$ is due to convexity-concavity of θ. The resulting inequality holds true for all $[u; v] \in U \times V$, and (7) follows.

2.4.2 Main Lemma

Proposition 1. *In the situation and notation of sections 2.1 – 2.3, let ϕ be the primal convex-concave function induced by Φ, and let*

$$\mathcal{I}_t = \{[x_{1,i}; y_{1,i}] \in X_1 \times Y_1, [\alpha_i := \phi'_{x_1}(x_{1,i}, y_{1,i}); \beta_i := -\phi'_{y_1}(x_{1,i}, y_{1,i})] : 1 \le i \le t\}$$

be an execution protocol associated with ϕ, X_1, Y_1, where ϕ'_{x_1}, ϕ'_{y_1} are regular sub- and supergradient fields associated with Φ, ϕ. Due to the origin of ϕ, ϕ'_{x_1}, ϕ'_{y_1}, there exist $x_{2,i} \in X_2[x_{1,i}]$, $G_i \in E$, $y_{2,i} \in Y_2[y_{1,i}]$, and $H_i \in F$ such that

$$\begin{array}{lrcl} (a) & G_i & \in & \partial_x \Phi(x_i := [x_{1,i}; x_{2,i}], y_i := [y_{1,i}; y_{2,i}]), \\ (b) & H_i & \in & \partial_y \left[-\Phi(x_i := [x_{1,i}; x_{2,i}], y_i := [y_{1,i}; y_{2,i}])\right], \\ (c) & \langle G_i, x - [x_{1,i}; x_{2,i}]\rangle & \geq & \langle \phi'_{x_1}(x_{1,i}, y_{1,i}), x_1 - x_{1,i}\rangle \forall x = [x_1; x_2] \in X, \\ (d) & \langle H_i, y - [y_{1,i}; y_{2,i}]\rangle & \geq & \langle -\phi'_{y_1}(x_{1,i}, y_{1,i}), y_1 - y_{1,i}\rangle \forall y = [y_1; y_2] \in Y, \end{array} \tag{8}$$

implying that

$$\mathcal{J}_t = \{z_i = [x_i = [x_{1,i}; x_{2,i}]; y_i = [y_{1,i}; y_{2,i}]], F_i = [G_i; H_i] : 1 \leq i \leq t\}$$

is an execution protocol associated with Φ, X, Y. For every accuracy certificate λ it holds

$$\mathrm{Res}(\mathcal{J}_t, \lambda | X \times Y) \leq \mathrm{Res}(\mathcal{I}_t, \lambda | X_1 \times Y_1). \tag{9}$$

As a result, given an accuracy certificate λ and setting

$$[x^t; y^t] = [[x_1^t; x_2^t]; [y_1^t; y_2^t]] = \sum_{i=1}^t \lambda_i \left[[x_{1,i}; x_{2,i}]; [y_{1,i}; y_{2,i}]\right],$$

we ensure that

$$\epsilon_{\mathrm{sad}}([x^t; y^t] | \Phi, X, Y) \leq \mathrm{Res}(\mathcal{I}_t, \lambda | X_1 \times Y_1), \tag{10}$$

whence also, by Lemma 1,

$$\begin{array}{rcl} \epsilon_{\mathrm{sad}}([x_1^t; y_1^t] | \phi, X_1, Y_1) & \leq & \mathrm{Res}(\mathcal{I}_t, \lambda | X_1 \times Y_1), \\ \epsilon_{\mathrm{sad}}([x_2^t; y_2^t] | \psi, X_2, Y_2) & \leq & \mathrm{Res}(\mathcal{I}_t, \lambda | X_1 \times Y_1), \end{array} \tag{11}$$

where ψ is the dual function induced by Φ.

Proof. Let $z := [[u_1; u_2]; [v_1; v_2]] \in X \times Y$. Then

$$\begin{array}{l} \sum_{i=1}^t \lambda_i \langle F_i, z_i - z\rangle = \sum_{i=1}^t \lambda_i \Big[\underbrace{\langle G_i, [x_{1,i}; x_{2,i}] - [u_1; u_2]\rangle}_{\leq \langle \phi'_{x_1}(x_{1,i}, y_{1,i}), x_{1,i} - u_1\rangle \text{ by (8.c)}} + \underbrace{\langle H_i, [y_{1,i}; y_{2,i}] - [v_1; v_2]\rangle}_{\leq \langle -\phi'_{y_1}(x_{1,i}, y_{1,i}), y_{1,i} - v_1\rangle \text{ by (8.d)}} \Big] \\ \leq \sum_{i=1}^t \lambda_i \big[\langle \alpha_i, x_{1,i} - u_1\rangle + \langle \beta_i, y_{1,i} - v_1\rangle\big] \leq \mathrm{Res}(\mathcal{I}_t, \lambda | X_1 \times Y_1), \end{array}$$

and (9) follows. □

2.5 Application: Solving bilinear Saddle Point problems on domains represented by Linear Minimization Oracles

2.5.1 Situation

Let W be a nonempty convex compact set in $\mathbf{R}^N$, Z be a nonempty convex compact set in $\mathbf{R}^M$, and let $\psi : W \times Z \to \mathbf{R}$ be bilinear convex-concave function:

$$\psi(w, z) = \langle w, p\rangle + \langle z, q\rangle + \langle z, Sw\rangle. \tag{12}$$

Our goal is to solve the convex-concave SP problem

$$\min_{w \in W} \max_{z \in Z} \psi(w, z) \tag{13}$$

given by ψ, W, Z.

2.5.2 Simple observation

Let $U \subset \mathbf{R}^n$, $V \subset \mathbf{R}^m$ be convex compact sets, and let $D \in \mathbf{R}^{m\times N}$, $A \in \mathbf{R}^{n\times M}$, $R \in \mathbf{R}^{m\times n}$. Consider bilinear (and thus convex-concave) function

$$\Phi(u,w;v,z) = \langle w, p + D^T v\rangle + \langle z, q + A^T u\rangle - \langle v, Ru\rangle : [U\times W]\times[V\times Z] \to \mathbf{R} \tag{14}$$

(the "convex" argument is (u,w), the "concave" one is (v,z)) and a pair of functions

$$\begin{array}{l}\bar{u}(w,z) : W\times Z \to U,\\ \bar{v}(w,z) : W\times Z \to V\end{array}$$

and let us make the following assumption:

(!) We have

$$\begin{array}{rcl}\forall (w,z)\in W\times Z: & Dw & = & R\bar{u}(w,z)\\ \forall (w,z)\in W\times Z: & Az & = & R^T\bar{v}(w,z)\end{array} \tag{15}$$

Assuming (!), for all $(w,z)\in W\times Z$, denoting $\bar{u}=\bar{u}(w,z)$, $\bar{v}=\bar{v}(w,z)$, we have

$$\begin{array}{llcl}(a) & \langle w, D^T\bar{v}\rangle & = & \langle Dw,\bar{v}\rangle = \langle R\bar{u},\bar{v}\rangle\\ (b) & \langle z, A^T\bar{u}\rangle & = & \langle Az,\bar{u}\rangle = \langle \bar{u}, R^T\bar{v}\rangle = \langle R\bar{u},\bar{v}\rangle\\ \Rightarrow & & & \\ (c) & \nabla_u\Phi(\bar{u},w;\bar{v},z) & = & Az - R^T\bar{v} = 0\\ (d) & \nabla_v\Phi(\bar{u},w;\bar{v},z) & = & Dw - R\bar{u} = 0\\ \Rightarrow & & & \\ (e) & \bar{\psi}(w,z) & := & \min_{u\in U}\max_{v\in V}\Phi(u,w;v,z) = \Phi(\bar{u}(w,z),w;\bar{v}(w,z),z)\\ & & = & \langle w,p\rangle + \langle z,q\rangle + \langle Dw,\bar{v}(w,z)\rangle \text{ [by } (a),\ (b)]\end{array} \tag{16}$$

We have proved

Lemma 5. *In the case of* (!)*, assuming that*

$$\langle Dw,\bar{v}(w,z)\rangle = \langle z, Sw\rangle \ \forall w\in W, z\in Z, \tag{17}$$

ψ is the dual convex-concave function induced by Φ and the domains $U\times W$, $V\times Z$.

Note that there are easy ways to ensure (!) and (17).

Example 1. Here $m=M$, $n=N$, and $D=A^T=R=S$. Assuming $U\supset W$, $V\supset Z$ and setting $\bar{u}(w,z)=w$, $\bar{v}(w,z)=z$, we ensure (!) and (17).

Example 2. Let $S=A^TD$ with $A\in\mathbf{R}^{K\times M}$, $D\in\mathbf{R}^{K\times N}$. Setting $m=n=K$, $R=I_K$, $\bar{u}(w,z)=Dw$, $\bar{v}(w,z)=Az$ and assuming that $U\supset DW$, $V\supset AZ$, we again ensure (!) and (17).

2.5.3 Implications

Assume that (!) and (17) take place. Denoting $u = x_1$, $v = y_1$, $w = x_2$, $z = y_2$ and setting $X_1 = U$, $X_2 = W$, $Y_1 = V$, $Y_2 = Z$, $X = X_1 \times X_2 = U \times W$, $Y = Y_1 \times Y_2 = V \times Z$, we find ourselves in the direct product case of the situation of section 2.1, and Lemma 5 says that the bilinear SP problem of interest (12), (13) is the dual SP problem associated with the bilinear master SP problem

$$\min_{[u;w]\in U\times W}\max_{[v;z]\in V\times Z}\left[\Phi(u,w;v,z) = \langle w, p + D^T v\rangle + \langle z, q + A^T u\rangle - \langle Ru, v\rangle\right] \tag{18}$$

Since Φ is linear in $[u; v]$, the primal SP problem associated with (18) is

$$\min_{u\equiv x_1\in U=X_1}\max_{v\equiv y_1\in V=Y_1}\left[\phi(u,v) = \min_{w\in W}\langle w, p + D^T v\rangle + \max_{z\in Z}\langle v, q + A^T u\rangle - \langle Ru, v\rangle\right].$$

Assuming that W, Z allow for cheap Linear Minimization Oracles and defining $w_*(\cdot)$, $z_*(\cdot)$ according to

$$w_*(\xi) \in \operatorname*{Argmin}_{w\in W}\langle w, \xi\rangle,\ \ z_*(\eta) \in \operatorname*{Argmin}_{z\in Z}\langle z, \eta\rangle,$$

we have

$$\begin{array}{rcl}\phi(u,v) & = & \langle w(p + D^T v), p + D^T v\rangle + \langle z(-q - A^T u), q + A^T u\rangle - \langle Ru, v\rangle,\\ \phi'_u(u,v) & := & Az_*(-q - A^T u) - R^T v \in \partial_w\phi(u,v),\\ \phi'_v(u,v) & := & Dw_*(p + D^T v) - Ru \in -\partial_v[-\phi(u,v)],\end{array} \tag{19}$$

that is, first order information on the primal SP problem

$$\min_{u\in U}\max_{v\in V}\phi(u,v), \tag{20}$$

is available. Note that since we are in the direct product case, ϕ'_u and ϕ'_v are regular sub- and supergradient fields associated with Φ, ϕ.

Now let $\mathcal{I}_t = \{[u_i; v_i] \in U \times V, [\gamma_i := \phi'_u(u_i, v_i); \delta_i := -\phi'_v(u_i, v_i)] : 1 \le i \le t\}$ be an execution protocol generated by a First Order algorithm as applied to the primal SP problem (20), and let

$$\begin{array}{l} w_i = w_*(p + D^T v_i), z_i = z_*(-q - A^T u_i),\\ \alpha_i = \nabla_w\Phi(u_i, w_i; v_i, z_i) = p + D^T v_i,\\ \beta_i = -\nabla_z\Phi(u_i, w_i; v_i, z_i) = -q - Au_i,\end{array}$$

so that

$$\mathcal{J}_t = \left\{[[u_i; w_i]; [v_i; z_i]], [\ \underbrace{[\alpha_i; \gamma_i]}_{\nabla_{[u;w]}\Phi(u_i, w_i; v_i, z_i)}\ ;\ \underbrace{[\beta_i; \delta_i]}_{-\nabla_{[v;z]}\Phi(u_i, w_i; v_i, z_i)}\ ] : 1 \le i \le t\right\}$$

is an execution protocol associated with the SP problem (18). By Proposition 1, for any accuracy certificate λ it holds

$$\mathrm{Res}(\mathcal{J}_t, \lambda|U \times W \times V \times Z) \le \mathrm{Res}(\mathcal{I}_t, \lambda|U \times V) \tag{21}$$

whence, setting

$$[[u^t; w^t]; [v^t; z^t]] = \sum_{i=1}^t \lambda_i[[u_i; w_i]; [v_i; z_i]] \tag{22}$$

and invoking Lemma 4 with Φ in the role of θ,

$$\epsilon_{\rm sad}([[u^t;w^t];[v^t;z^t]]|\Phi,\underbrace{X_1\times X_2}_{U\times W},\underbrace{Y_1\times Y_2}_{V\times Z})\le \mathrm{Res}(\mathcal{I}_t,\lambda|U\times V) \tag{23}$$

whence, by Lemma 1,

$$\epsilon_{\rm sad}([w^t;z^t]|\psi,W,Z)\le \mathrm{Res}(\mathcal{I}_t,\lambda|U\times V). \tag{24}$$

The bottom line is that

> *With* (!), (17) *in force, applying to the primal SP problem* (20) *First Order algorithm* $\mathcal{B}$ *with accuracy certificates, we get, as a byproduct, feasible solutions to the SP problem of interest* (13) *of the* $\epsilon_{\rm sad}$*-inaccuracy* $\le \mathrm{Res}(\mathcal{I}_t,\lambda|U\times V)$.

Note also that when the constructions from Examples 1,2 are used, there is a significant freedom in selecting the domain $U\times V$ of the primal problem (U, V should be convex compact sets "large enough" to ensure the inclusions mentioned in Examples), so that there is no difficulty to enforce U, V to be proximal friendly. As a result, we can take as $\mathcal{B}$ a proximal First Order method, for example, Non-Euclidean Restricted Memory Level algorithm with certificates (cf. [2]) or Mirror Descent (cf. [14]). The efficiency estimates of these algorithms as given in [2, 14] imply that the resulting procedure for solving the SP of interest (12), (13) admits non-asymptotic $O(1/\sqrt{t})$ rate of convergence, with explicitly computable factors hidden in $O(\cdot)$. The resulting complexity bound is completely similar to the one achievable with the machinery of Fenchel-type representations [2, 14].

We are about co consider a special case where the $O(1/\sqrt{t})$ complexity admits a significant improvement.

2.6 Matrix Game case

Let $S\in\mathbf{R}^{M\times N}$ be represented as

$$S=A^TD$$

with $A\in\mathbf{R}^{K\times M}$ and $D\in\mathbf{R}^{K\times N}$. Let also $W=\Delta_N=\{w\in\mathbf{R}^N_+:\sum_i w_i=1\}$, $Z=\Delta_M$. Our goal is to solve matrix game

$$\min_{w\in W}\max_{z\in Z}\left[\psi(w,z)=\langle z,Sw\rangle=\langle Az,Dw\rangle\right]. \tag{25}$$

Let U, V be convex compact sets such that

$$V\supset AZ,\ U\supset DW, \tag{26}$$

and let us set

$$\begin{array}{rcl}\Phi(u,w;v,z) &=& \langle u,Az\rangle+\langle v,Dw\rangle-\langle u,v\rangle\\ \bar u:=\bar u(w,z) &=& Dw\\ \bar v:=\bar v(w,z) &=& Az\end{array}$$

implying that

$$\begin{array}{rcl}\nabla_u\Phi(\bar u,w;\bar v,z) &=& Az-\bar v=0,\\ \nabla_v\Phi(\bar u,w;\bar v,z) &=& Dw-\bar u=0,\\ \Phi(\bar u,w;\bar v,z) &=& \langle\bar u,Az\rangle+\langle\bar v,Dw\rangle-\langle\bar u,\bar v\rangle=\langle Dw,Az\rangle+\langle Az,Dw\rangle-\langle Dw,Az\rangle\\ &=& \langle z,A^TDw\rangle=\psi(w,z).\end{array}$$

It is immediately seen that the function ψ from (25) is nothing but the dual convex-concave function associated with Φ (cf. Example 2), while the primal function is

$$\phi(u,v) = \mathrm{Max}(A^T u) + \mathrm{Min}(D^T v) - \langle u, v\rangle; \tag{27}$$

here $\mathrm{Min}(p)$ and $\mathrm{Max}(p)$ stand for the smallest and the largest entries in vector p. Applying the strategy outlined in section 2.2, we can solve the problem of interest (25) applying to the primal SP problem

$$\min_{u\in U}\max_{v\in V}\left[\phi(u,v) = \mathrm{Min}(D^T v) + \mathrm{Max}(A^T u) - \langle u, v\rangle\right] \tag{28}$$

an algorithm with accuracy certificates and using the machinery outlined in previous sections to convert the resulting execution protocols and certificates into approximate solutions to the problem of interest (25).

We intend to consider a special case when the outlined approach allows to reduce a huge, but well organized, matrix game (25) to a small SP problem (28) – so small that it can be solved to high accuracy by something like Ellipsoid method. This is the case when the matrices A, D in (25) are *well organized*.

2.6.1 The case of well organized matrices A, D

Given an $K\times L$ matrix B, we call B *well organized* if, given $x\in\mathbf{R}^K$, it is easy to identify the columns $\overline{B}[x]$, $\underline{B}[x]$ of B making the maximal, resp. the minimal, inner product with x.

When matrices A, D in (25) are well organized, the first order information for the cost function ϕ in the primal SP problem (28) is easy to get. Besides, all we need from the convex compact sets U, V participating in (28) is to be large enough to ensure that $U\supset DW$ and $V\supset AZ$, which allows to make U and V simple, e.g., Euclidean balls. Finally, when the design dimension $2K$ of (28) is small, we have at our disposal a multitude of linearly converging, with the converging ratio depending solely on K, methods for solving (28), including the Ellipsoid algorithm with certificates presented in [15]. We are about to demonstrate that the outlined situation indeed takes place in some meaningful applications.

2.6.2 Example: Knapsack generated matrices

[2] Assume that we are given *knapsack data*, namely,

- positive integer *horizon* m,
- nonnegative integer *bounds* $\bar{p}_s$, $1\le s\le m$,
- positive integer *costs* h_s, $1\le s\le m$, and positive integer *budget* H, and
- *output functions* $f_s(\cdot):\{0,1,...,\bar{p}_s\}\to\mathbf{R}^{r_s}$, $1\le s\le m$.

Given the outlined data, consider the set $\mathcal{P}$ of all integer vectors $p=[p_1;...;p_m]\in\mathbf{R}^m$ satisfying the following restrictions:

$$\begin{array}{ll} 0\le p_s\le \overline{p}_s,\ 1\le s\le m & \text{[range restriction]}\\ \sum_{s=1}^m h_s p_s\le H & \text{[budget restriction]}\end{array}$$

[2] The construction to follow can be easily extended from "knapsack generated" matrices to more general "Dynamic Programming generated" ones, see section A.4 in Appendix.

and the matrix B of the size $K \times \mathrm{Card}(\mathcal{P})$, $K = \sum_{s=1}^m r_s$, defined as follows: the columns of B are indexed by vectors $p = [p_1; ...; p_s] \in \mathcal{P}$, and the column indexed by p is the vector

$$B_p = [f_1(p_1); ...; f_m(p_m)].$$

Note that assuming m, $\overline{p}_s$, r_s moderate, matrix B is well organized – given $x \in \mathbf{R}^K$, it is easy to find $\overline{B}[x]$ and $\underline{B}[x]$ by Dynamic Programming.

Indeed, to identify $\overline{B}[x]$, $x = [x_1; ...; x_m] \in \mathbf{R}^{r_1} \times ... \times \mathbf{R}^{r_m}$ (identification of $\underline{B}[x]$ is completely similar), it suffices to run for $s = m, m-1, ...1$ the backward Bellman recurrence

$$\left.\begin{array}{rcl} U_s(h) & = & \max\limits_{r \in \mathbf{Z}} \{U_{s+1}(h - h_s r) + \langle f_s(r), x_s \rangle : 0 \le r \le \overline{p}_s, 0 \le h - h_s r\} \\ A_s(h) & \in & \mathrm{Argmax}\limits_{r \in \mathbf{Z}} \{U_{s+1}(h - h_s r) + \langle f_s(r), x_s \rangle : 0 \le r \le \overline{p}_s, 0 \le h - h_s r\} \end{array}\right\}, h = 0, 1, ..., H,$$

with $U_{m+1}(\cdot) \equiv 0$, and then to recover one by one the entries p_s in the index $p \in \mathcal{P}$ of $\overline{B}[x]$ from the forward Bellman recurrence

$$\begin{array}{l} H_1 = H, p_1 = A_1(H_1); \\ H_{s+1} = H_s - h_s p_s, p_{s+1} = A_{s+1}(H_{s+1}), 1 \le s < m. \end{array}$$

2.6.3 Illustration: Attacker vs. Defender.

The "covering story" we intend to consider is as follows. Attacker and Defender are preparing for a conflict to take place on m battlefields. A pure strategy of Attacker is a vector $q = [q_1; ...; q_m]$, where positive integer q_s, $1 \le s \le m$, is the number of attacking units to be created and deployed at battlefield s; the only restrictions on q, aside of integrality, are the bounds $q_s \le \overline{q}_s$, $1 \le s \le m$, and the budget constraint $\sum_{s=1}^m h_{sA} q_s \le H_A$ with positive integer h_{sA} and H_A. Similarly, a pure strategy of Defender is a vector $p = [p_1; ...; p_m]$, where positive integer p_s is the number of defending units to be created and deployed at battlefield s, and the only restrictions on p, aside of integrality, are the bounds $p_s \le \overline{p}_s$, $1 \le s \le m$, and the budget constraint $\sum_{s=1}^m h_{sD} q_s \le H_D$ with positive integer h_{sD} and H_D. The total loss of Defender (the total gain of Attacker), the pure strategies of the players being p and q, is

$$G_{q,p} = \sum_{s=1}^m [\Omega^s]_{q_s, p_s},$$

with given $(\overline{q}_s + 1) \times (\overline{p}_s + 1)$ matrices Ω^s. Denoting by $\mathcal{Q}$ and $\mathcal{P}$ the sets of pure strategies of Attacker, resp., Defender, and setting

$$\begin{array}{rcl} \Omega^s & = & \sum_{i=1}^{r_s} f^{is} [g^{is}]^T,\ f^{is} = [f_0^{is}; ...; f_{\overline{q}_s}^{is}],\ g^{is} = [g_0^{is}; ...; g_{\overline{p}_s}^{is}],\ r_s = \mathrm{Rank}(\Omega^s),\ K = \sum_{s=1}^m r_s, \\ A_q & = & [f_{q_1}^{1,1}; f_{q_1}^{2,1}; ...; f_{q_1}^{r_1,1}; f_{q_2}^{1,2}; f_{q_2}^{2,2}; ...; f_{q_2}^{r_2,2}; ...; f_{q_m}^{1,m}; f_{q_m}^{2,m}; ...; f_{q_m}^{r_m,m}] \in \mathbf{R}^K, q \in \mathcal{Q}, \\ D_p & = & [g_{p_1}^{1,1}; g_{p_1}^{2,1}; ...; g_{p_1}^{r_1,1}; g_{p_2}^{1,2}; g_{p_2}^{2,2}; ...; g_{p_2}^{r_2,2}; ...; g_{p_m}^{1,m}; g_{p_m}^{2,m}; ...; g_{p_m}^{r_m,m}] \in \mathbf{R}^K, p \in \mathcal{P}, \end{array}$$

we end up with $K \times M$, $M = \mathrm{Card}(\mathcal{Q})$, knapsack-generated matrix A with columns A_q, $q \in \mathcal{Q}$, and $K \times N$, $N = \mathrm{Card}(\mathcal{P})$, knapsack-generated matrix D with columns D_p, $p \in \mathcal{P}$, such that

$$G = A^T D.$$

As a result, solving the Attacker vs. Defender game in mixed strategies reduces to solving SP problem (25) with knapsack-generated (and thus well organized) matrices A, D and thus can be reduced to convex-concave SP (28) of dimension K. Note that in the situation in question the design dimension $2K$ of (28) will, typically, be rather small (few tens or at most few hundreds), while the design dimensions M, N of the matrix game of interest (25) can be huge.

Numerical illustration. With the data (quite reasonable in terms of the "Attacker vs. Defender" game)

$$m = 8,\ h_{sA} = h_{sD} = 1, 1 \le s \le m,\ H_A = H_D = 64 = \overline{p}_s = \overline{q}_s,\ 1 \le s \le m$$

and rank 1 matrices Ω_s, $1 \le s \le m$, the design dimensions of the problem of interest (25) are as huge as

$$\dim w = \dim z = 11,969,016,345$$

while the sizes of (28) are just

$$\dim u = \dim v = 8,$$

and thus (28) can be easily solved to high accuracy by the Ellipsoid method. In the numerical experiment we are about to report[3], the outlined approach allowed to solve (25) within $\epsilon_{\rm sad}$-inaccuracy as small as 5.4e-5 (by factor over 7.8e4) in just 1281 steps of the Ellipsoid algorithm (537.8 sec on a medium performance laptop) – not that bad given the huge – over 10^{10} – sizes of the matrix game of interest (25)!

3 From Bilinear Saddle Point problems to Variational Inequalities with Affine Monotone Operators

In what follows, we extend the decomposition approach (developed so far for convex-concave SP problems) to Variational Inequalities (VI's) with monotone operators, with the primary goal to handle VI's with affine monotone operators on LMO-represented domains.

3.1 Preliminaries

Recall that the (Minty's) variational inequality $\mathrm{VI}(F, Y)$ associated with a convex compact subset Y of Euclidean space E and a vector field $F : Y \to E$ is

$$\text{find } y \in Y : \langle F(y'), y' - y\rangle \ge 0\ \forall y' \in Y \qquad \mathrm{VI}(F, Y)$$

an y satisfying the latter condition is called a *weak* solution to the VI. A *strong* solution to $\mathrm{VI}(F, Y)$ is a point $y_* \in Y$ such that

$$\langle F(y_*), y - y_*\rangle \ge 0\ \forall y \in Y.$$

assuming F monotone on Y, every strong solution to $\mathrm{VI}(F, Y)$ is a weak one, and when F is continuous, every weak solution to $\mathrm{VI}(F, Y)$ is a strong solution as well. When F is monotone on Y and, as stated above, Y is convex and compact, weak solutions to $\mathrm{VI}(F, Y)$ do exist.

A natural measure of inaccuracy for an approximate solution $y \in Y$ to $\mathrm{VI}(F, Y)$ is the *dual gap function*

$$\epsilon_{\rm VI}(y|F, Y) = \sup_{y' \in Y} \langle F(y'), y - y'\rangle;$$

weak solutions to the VI are exactly the points of Y where this (clearly nonnegative everywhere on Y) function is zero.

In the sequel we utilize the following simple fact originating from [15]:

[3] for implementation details, see section A.5

Lemma 6. *Let F be monotone on Y, let $\mathcal{I}_t = \{y_i \in Y, F(y_i) : 1 \le i \le t\}$ be a t-step execution protocol associated with (F, Y), λ be a t-step accuracy certificate, and $y^t = \sum_{i=1}^t \lambda_i y_i$ be the associated approximate solution. Then*

$$\epsilon_{\rm VI}(y^t|F,Y) \le {\rm Res}(\mathcal{I}_t, \lambda|Y).$$

Indeed, we have

$$\begin{array}{l}{\rm Res}(\mathcal{I}_t,\lambda|Y) = \sup_{y'\in Y}\left[\sum_{i=1}^t \lambda_i\langle F(y_i), y_i - y'\rangle\right]\\ \ge \sup_{y'\in Y}\left[\sum_{i=1}^t \lambda_i\langle F(y'), y_i - y'\rangle\right] \ \text{[since } F \text{ is monotone]}\\ = \sup_{y'\in Y}\langle F(y'), y^t - y'\rangle = \epsilon_{\rm VI}(y^T|F,Y).\end{array}$$

3.2 Situation

In the sequel, we deal with the situation as follows. Given are

- Euclidean spaces E_ξ, E_η,
- nonempty convex compact set $\Theta \subset E_\xi \times E_\eta$ with the projections Ξ onto E_ξ, and H onto E_η. Given $\xi \in \Xi$, $\eta \in H$, we set

  $$H_\xi = \{\eta : [\xi;\eta] \in \Theta\},\ \Xi_\eta = \{\xi \in \Xi : [\xi;\eta] \in \Theta\},$$

  and denote a point from $E_\xi \times E_\eta$ as $\theta = [\xi;\eta]$ with $\xi \in E_\xi$, $\eta \in E_\eta$;
- a monotone vector field

  $$\Phi(\xi,\eta) = [\Phi_\xi(\xi,\eta); \Phi_\eta(\xi,\eta)] : \Theta \to E_\xi \times E_\eta.$$

We assume in the sequel that for once for ever properly selected functions $\overline{\eta}(\xi) : \Xi \to H$ and $\overline{\xi}(\eta) : H \to \Xi$, the point $\overline{\eta}(\xi)$, $\xi \in \Xi$, is a strong solution to the VI ${\rm VI}(\Phi_\eta(\xi,\cdot), H_\xi)$, and the point $\overline{\xi}(\eta)$, $\eta \in H$, is a strong solution to the VI ${\rm VI}(\Phi_\eta(\cdot,\eta), \Xi_\eta)$, that is,

$$\overline{\eta}(\xi) \in H_\xi \ \&\ \langle \Phi_\eta(\xi, \overline{\eta}(\xi)), \eta - \overline{\eta}(\xi)\rangle \ge 0\ \forall \eta \in H_\xi \tag{29}$$

and

$$\overline{\xi}(\eta) \in \Xi_\eta \ \&\ \langle \Phi_\xi(\overline{\xi}(\eta), \eta), \xi - \overline{\xi}(\eta)\rangle \ge 0\ \forall \xi \in \Xi_\eta. \tag{30}$$

Note that the assumptions on the existence of $\overline{\eta}(\cdot)$, $\overline{\xi}(\cdot)$ satisfying (29), (30) are automatically satisfied when the monotone vector field Φ is continuous.

3.3 Induced vector fields

Let us call Φ (more precisely, the pair $(\Phi, \overline{\eta}(\cdot))$) *$\eta$-regular*, if for every $\xi \in \Xi$, there exists $\Psi = \Psi(\xi) \in E_\xi$ such that

$$\langle \Psi(\xi), \xi' - \xi\rangle \le \langle \Phi(\xi, \overline{\eta}(\xi)), [\xi';\eta'] - [\xi; \overline{\eta}(\xi)]\rangle\ \forall [\xi';\eta'] \in \Theta. \tag{31}$$

Similarly, let us call $(\Phi, \overline{\xi}(\cdot))$ *ξ-regular*, if for every $\eta \in H$ there exists $\Gamma = \Gamma(\eta) \in E_\eta$ such that

$$\langle \Gamma(\eta), \eta' - \eta\rangle \le \langle \Phi(\overline{\xi}(\eta), \eta), [\xi';\eta'] - [\overline{\xi}(\eta);\eta]\rangle\ \forall [\xi';\eta'] \in \Theta. \tag{32}$$

When $(\Phi, \overline{\eta})$ is η-regular, we refer to the above $\Psi(\cdot)$ as to a *primal* vector field induced by Φ [4], and when $(\Phi, \overline{\xi})$ is ξ-regular, we refer to the above $\Gamma(\cdot)$ as to a *dual* vector field induced by Φ.

[4] "a primal" instead of "the primal" reflects the fact that Ψ is not uniquely defined by Φ – it is defined by Φ *and* $\overline{\eta}$ *and* by how the values of Ψ are selected when (31) does not specify these values uniquely.

Example: Direct product case. This is the case where $\Theta = \Xi \times H$. In this situation, setting $\Psi(\xi) = \Phi_\xi(\xi, \overline{\eta}(\xi))$, we have for $[\xi'; \eta'] \in \Theta$ and $\xi \in \Xi$:

$$\langle \Phi(\xi, \overline{\eta}(\xi)), [\xi'; \eta'] - [\xi; \overline{\eta}(\xi)] \rangle = \underbrace{\langle \Phi_\xi(\xi, \overline{\eta}(\xi)), \xi' - \xi \rangle}_{= \langle \Psi(\xi), \xi' - \xi \rangle} + \underbrace{\langle \Phi_\eta(\xi, \overline{\eta}(\xi)), \eta' - \overline{\eta}(\xi) \rangle}_{\geq 0 \ \forall \eta' \in H_\xi = H} \geq \langle \Psi(\xi), \xi' - \xi \rangle,$$

that is, $(\Phi, \overline{\eta}(\cdot))$ is η-regular, with $\Psi(\xi) = \Phi_\xi(\xi, \overline{\eta}(\xi))$. Setting $\Gamma(\eta) = \Phi_\eta(\overline{\xi}(\eta), \eta)$, we get by similar argument

$$\langle \Phi(\overline{\xi}(\eta), \eta), [\xi'; \eta'] - [\overline{\xi}(\eta); \eta] \rangle \geq \langle \Gamma(\eta), \eta' - \eta \rangle, \ [\xi'; \eta'] \in \Theta, \eta \in H,$$

that is, $(\Phi, \overline{\xi}(\cdot))$ is ξ-regular, with $\Gamma(\eta) = \Phi_\eta(\overline{\xi}(\eta), \eta)$.

3.4 Main observation

Proposition 2. *In the situation of section 3.2, let $(\Phi, \overline{\eta}(\cdot))$ be η-regular. Then*

(i) *Primal vector field $\Psi(\xi)$ induced by $(\Phi, \overline{\eta}(\cdot))$ is monotone on Ξ. Moreover, whenever $\mathcal{I}_t = \{\xi_i \in \Xi, \Psi(\xi_i) : 1 \leq i \leq t\}$ and $\mathcal{J}_t = \{\theta_i := [\xi_i; \overline{\eta}(\xi_i)], \Phi(\theta_i) : 1 \leq i \leq t\}$ and λ is a t-step accuracy certificate, it holds*

$$\epsilon_{\rm VI}(\sum_{i=1}^t \lambda_i \theta_i | \Phi, \Theta) \leq {\rm Res}(\mathcal{J}_t, \lambda | \Theta) \leq {\rm Res}(\mathcal{I}_t, \lambda | \Xi). \tag{33}$$

(ii) *Let $(\Phi, \overline{\xi})$ be ξ-regular, and let Γ be the induced dual vector field. Whenever $\widehat{\theta} = [\widehat{\xi}; \widehat{\eta}] \in \Theta$, we have*

$$\epsilon_{\rm VI}(\widehat{\eta} | \Gamma, H) \leq \epsilon_{\rm VI}(\widehat{\theta} | \Phi, \Theta). \tag{34}$$

3.5 Implications

In the situation of section 3.2, assume that for properly selected $\overline{\eta}(\cdot)$, $\overline{\xi}(\cdot)$, $(\Phi, \overline{\eta}(\cdot))$ is η-regular, and $(\Phi, \overline{\xi}(\cdot))$ is ξ-regular, induced primal and dual vector fields being Ψ and Γ. In order to solve the *dual VI* ${\rm VI}(\Gamma, H)$, we can apply to the *primal VI* ${\rm VI}(\Psi, \Xi)$ an algorithm with accuracy certificates; by Proposition 2.i, resulting t-step execution protocol $\mathcal{I}_t = \{\xi_i, \Psi(\xi_i) : 1 \leq i \leq t\}$ and accuracy certificate λ generate an execution protocol $\mathcal{J}_t = \{\theta_i := [\xi_i; \overline{\eta}(\xi_i)], \Phi(\theta_i) : 1 \leq i \leq t\}$ such that

$${\rm Res}(\mathcal{J}_t, \lambda | \Theta) \leq {\rm Res}(\mathcal{I}_t, \lambda | \Xi),$$

whence, by Lemma 6, for the approximate solution

$$\theta^t = [\xi^t, \eta^t] := \sum_{i=1}^t \lambda_i \theta_i = \sum_{i=1}^t \lambda_i [\xi_i; \overline{\eta}(\xi_i)]$$

it holds

$$\epsilon_{\rm VI}(\theta^t | \Phi, \Theta) \leq {\rm Res}(\mathcal{I}_t, \lambda | \Xi).$$

Invoking Proposition 2.ii, we conclude that η^t is a feasible solution to the dual VI ${\rm VI}(\Gamma, H)$, and

$$\epsilon_{\rm VI}(\eta^t | \Gamma, H) \leq {\rm Res}(\mathcal{I}_t, \lambda | \Xi). \tag{35}$$

We are about to present two examples well suited for the just outlined approach.

3.5.1 Solving affine monotone VI on LMO-represented domain

Let H be a convex compact set in Euclidean space E_η, and let H be equipped with an LMO. Assume that we want to solve the VI $\mathrm{VI}(F, H)$, where

$$F(\eta) = S\eta + s$$

is an affine monotone operator (so that $S + S^T \succeq 0$). Let us set $E_\xi = E_\eta$, select Ξ as a proximal-friendly convex compact set containing H, and set $\Theta = \Xi \times H$,

$$\Phi(\xi, \eta) = \underbrace{\left[\begin{array}{c|c} S^T & -S^T \\ \hline S & \end{array}\right]}_{\mathcal{S}} \left[\begin{array}{c} \xi \\ \eta \end{array}\right] + \left[\begin{array}{c} 0 \\ s \end{array}\right].$$

We have

$$\mathcal{S} + \mathcal{S}^T = \left[\begin{array}{c|c} S + S^T & \\ \hline & \end{array}\right] \succeq 0,$$

so that Φ is an affine monotone operator with

$$\Phi_\xi(\xi, \eta) \;=\; S^T\xi - S^T\eta, \quad \Phi_\eta(\xi, \eta) \;=\; S\xi + s.$$

Setting $\overline{\xi}(\eta) = \eta$, we ensure that $\overline{\xi}(\eta) \in \Xi$ when $\eta \in H$ and $\Phi_\xi(\overline{\xi}(\eta), \eta) = 0$, implying (30). Since we are in the direct product case, we can set $\Gamma(\eta) = \Phi_\eta(\overline{\xi}(\eta), \eta) = S\eta + s = F(\eta)$; thus, $\mathrm{VI}(\Gamma, H)$ is our initial VI of interest. On the other hand, setting

$$\overline{\eta}(\xi) \in \operatorname*{Argmin}_{\eta \in H} \langle S\xi + s, \eta \rangle,$$

we ensure (29). Since we are in the direct product case, we can set

$$\Psi(\xi) = \Phi_\xi(\xi, \overline{\eta}(\xi)) = S^T[\xi - \overline{\eta}(\xi)];$$

note that the values of Ψ can be straightforwardly computed via calls to the LMO representing H. We can now solve $\mathrm{VI}(\Psi, \Xi)$ by a proximal algorithm $\mathcal{B}$ with accuracy certificates and recover, as explained above, approximate solution to the VI of interest $\mathrm{VI}(F, H)$. With the Non-Euclidean Restricted Memory Level method with certificates [2] or Mirror Descent with certificates (see, e.g., [14]), the approach results in non-asymptotical $O(1/\sqrt{t})$-converging algorithm for solving the VI of interest, with explicitly computable factors hidden in $O(\cdot)$. This complexity bound, completely similar to the one obtained in [14], seems to be the best known under the circumstances.

3.5.2 Solving skew-symmetric VI on LMO-represented domain

Let H be an LMO-represented convex compact domain in E_η, and assume that we want to solve $\mathrm{VI}(F, H)$, where

$$F(\eta) = 2Q^T P\eta + f : E_\eta \to E_\eta$$

with $K \times \dim E_\eta$ matrices P, Q such that and the matrix $Q^T P$ is skew-symmetric:

$$Q^T P + P^T Q = 0.$$

Let $E_\xi = \mathbf{R}^K \times \mathbf{R}^K$, let Ξ_1, Ξ_2 be two convex compact sets in $\mathbf{R}^K$ such that

$$PH \subset \Xi_1, QH \subset \Xi_2. \tag{36}$$

Let us set $\Xi = \Xi_1 \times \Xi_2$, and let

$$\Phi(\xi = [\xi_1; \xi_2], \eta) = \left[\begin{array}{c|c|c} & I_K & P \\ \hline -I_K & & Q \\ \hline -P^T & -Q^T & \end{array}\right] \left[\begin{array}{c} \xi_1 \\ \xi_2 \\ \eta \end{array}\right] + \left[\begin{array}{c} 0 \\ 0 \\ f \end{array}\right].$$

Not that Φ is monotone and affine. Setting

$$\overline{\xi}(\eta) = [Q\eta; -P\eta]$$

and invoking (36), we ensure (30); since we are in the direct product case, we can take, as the dual induced vector field,

$$\Gamma(\eta) = \Phi_\eta(\overline{\xi}(\eta), \eta) = -P^T(Q\eta) - Q^T(-P\eta) + f = [Q^TP - P^TQ]\eta + f = 2Q^TP\eta + f = F(\eta);$$

so that the dual VI $\mathrm{VI}(\Gamma, H)$ is our VI of interest.

On the other hand, setting

$$\overline{\eta}(\xi = [\xi_1; \xi_2]) \in \operatorname*{Argmin}_{\eta \in H} \langle f - P^T\xi_1 - Q^T\xi_2, \eta \rangle,$$

we ensure (29). Since we are in the direct product case, we can define primal vector field as

$$\Psi(\xi_1, \xi_2) = \Phi_\xi([\xi_1, \xi_2], \overline{\eta}([\xi_1; \xi_2])) = \left[\begin{array}{c} \xi_2 + P\overline{\eta}(\xi_1, \xi_2) \\ -\xi_1 + Q\overline{\eta}(\xi_1, \xi_2) \end{array}\right].$$

Note that LMO for H allows to compute the values of Ψ, and that Ξ can be selected to be proximal-friendly. We can now solve $\mathrm{VI}(\Psi, \Xi)$ by a proximal algorithm $\mathcal{B}$ with accuracy certificates and recover, as explained above, approximate solution to the VI of interest $\mathrm{VI}(F, H)$. When the design dimension $\dim \Xi$ of the primal VI is small, other choices of $\mathcal{B}$, like the Ellipsoid algorithm, are possible, and in this case we can end up with linearly converging, with the converging ratio depending solely on $\dim \Xi$, algorithm for solving the VI of interest. We are about to give a related example, which can be considered as multi-player version of the "Attacker vs. Defender" game.

Example: Nash Equilibrium with pairwise interactions. Consider the situation as follows: there are

- $L \geq 2$ players, ℓ-th of them selecting a mixed strategy w_ℓ from probabilistic simplex Δ_{N_ℓ} of dimension N_ℓ,
- *encoding matrices* D_ℓ of sizes $m_\ell \times N_\ell$, and *loss matrices* $M^{\ell\ell'}$ of sizes $m_\ell \times m_{\ell'}$ such that

$$M^{\ell\ell} = 0, M^{\ell\ell'} = -[M^{\ell'\ell}]^T,\ 1 \leq \ell, \ell' \leq L.$$

- The loss of ℓ-th player depends on mixed strategies of the players according to

$$\mathcal{L}_\ell(\eta := [w_1; ...; w_L]) = \sum_{\ell'=1}^{L} w_\ell^T E^{\ell\ell'} w_{\ell'},\ E^{\ell\ell'} = D_\ell^T M^{\ell\ell'} D_{\ell'} + \langle g_\ell, \eta \rangle.$$

In other words, every pair of distinct players ℓ, ℓ' are playing matrix game with matrix $M^{\ell\ell'}$, and the loss of player ℓ is the sum, over the pairwise games he is playing, of his losses in these games, the "coupling constraints" being expressed by the requirement that every player uses the same mixed strategy in all pairwise games he is playing.

We have described *convex* Nash Equilibrium problem, meaning that for every ℓ, $\mathcal{L}_\ell(w_1, ..., w_L)$ is convex (in fact, linear) in w_ℓ, is jointly concave (in fact, linear) in $w^\ell := (w_1, ..., w_{\ell-1}, w_{\ell+1}, ..., w_L)$, and $\sum_{\ell=1}^L L_\ell(\eta)$ is the linear function $\langle g, \eta\rangle$, $g = \sum_\ell g_\ell$, and thus is convex. It is known (see, e.g., [15]) that Nash Equilibria in convex Nash problem are exactly the weak solutions to the VI given by monotone operator

$$F(\eta := [w_1; ...; w_L]) = [\nabla_{w_1}\mathcal{L}_1(\eta); ...; \nabla_{w_L}\mathcal{L}_L(\eta)]$$

on the domain

$$H = \Delta_{N_1} \times ... \times \Delta_{N_L}.$$

Let us set

$$Q = \frac{1}{2}\begin{bmatrix} D_1 & & & \\ & D_2 & & \\ & & \ddots & \\ & & & D_L \end{bmatrix}, \; P = \begin{bmatrix} M^{1,1}D_1 & M^{1,2}D_2 & ... & M^{1,L}D_L \\ M^{2,1}D_1 & M^{2,2}D_2 & ... & M^{2,L}D_L \\ \vdots & \vdots & \ddots & ... \\ M^{L,1}D_1 & M^{L,2}D_2 & ... & M^{L,L}D_L \end{bmatrix}.$$

Then

$$Q^TP = \frac{1}{2}\begin{bmatrix} D_1^TM^{1,1}D_1 & D_1^TM^{1,2}D_2 & ... & D_1^TM^{1,L}D_L \\ D_2^TM^{2,1}D_1 & D_2^TM^{2,2}D_2 & ... & D_2^TM^{1,L}D_L \\ \vdots & \vdots & \ddots & ... \\ D_L^TM^{L,1}D_1 & D_L^TM^{L,2}D_2 & ... & D_L^TM^{L,L}D_L \end{bmatrix}$$

so that Q^TP is skew-symmetric due to $M^{\ell\ell'} = -[M^{\ell'\ell}]^T$. Besides this, we clearly have

$$F(\eta := [w_1; ...; w_L]) = 2Q^TP\eta + f, \; f = [\nabla_{w_1}\langle g_1, \eta\rangle; ...; \nabla_{w_L}\langle g_L, \eta\rangle].$$

Observe that if $D_1, ..., D_L$ are well organized, so are Q and P.

Indeed, for Q this is evident: to find the column of Q which makes the largest inner product with $x = [x_1; ...; x_L]$, $\dim x_\ell = m_\ell$, it suffices to find, for every ℓ, the column of D_ℓ which makes the maximal inner product with x_ℓ, and then to select the maximal of the resulting L inner products and the corresponding to this maximum column of Q. To maximize the inner product of the same x with columns of P, note that

$$x^TP = \big[\underbrace{[\textstyle\sum_{\ell=1}^L x_\ell^TM^{\ell,1}]}_{y_1^T}D_1, ..., \underbrace{[\textstyle\sum_{\ell=1}^L x_\ell^TM^{\ell,L}]}_{y_\ell^T}D_L\big],$$

so that to maximize the inner product of x and the columns of P means to find, for every ℓ, the column of D_ℓ which makes the maximal inner product with y_ℓ, and then to select the maximal of the resulting L inner products and the corresponding to this maximum column of P.

We see that if D_ℓ are well organized, we can use the approach from section 3.5.2 to approximate the solution to the VI generated by F on H. Note that in the case in question the dual gap function $\epsilon_{\rm VI}(\eta|F,H)$ admits a transparent interpretation in terms of the Nash Equilibrium problem we are solving: for $\eta=[w_1;...;w_L]\in H$, we have

$$\epsilon_{\rm VI}(\eta|F,H)\geq\epsilon_{\rm Nash}(\eta):=\sum_{\ell=1}^{L}\left[\mathcal{L}_\ell(\eta)-\min_{w'_\ell\in\Delta_{N_\ell}}\mathcal{L}_\ell(w_1,...,w_{\ell-1},w'_\ell,w_{\ell+1},...,w_L)\right],\tag{37}$$

and the right hand side here is the sum, over the players, of the (nonnegative) incentives for a player ℓ to deviate from his strategy w_ℓ to another mixed strategy when all other players stick to their strategies as given by η. Thus, small $\epsilon_{\rm VI}([w_1;...;w_L]|\cdot,\cdot)$ means small incentives for the players to deviate from mixed strategies w_ℓ.

Verification of (37) is immediate: denoting $f_\ell=\nabla_{w_\ell}\langle g_\ell,w\rangle$, by definition of $\epsilon_{\rm VI}$ we have for every $\eta'=[w'_1;...;w'_L]\in H$:

$$\begin{array}{l}\epsilon_{\rm VI}(\eta|F,H)\geq\langle F(\eta'),\eta-\eta'\rangle=\sum_\ell\langle\nabla_{w_\ell}\mathcal{L}_\ell(\eta'),\eta_\ell-\eta'_\ell\rangle\\
=\sum_\ell\langle f_\ell,\eta_\ell-\eta'_\ell\rangle+\sum_{\ell,\ell'}\langle D_\ell^TM^{\ell\ell'}D_{\ell'}w'_{\ell'},w_\ell-w'_\ell\rangle\\
=\sum_\ell\langle f_\ell,\eta_\ell-\eta'_\ell\rangle+\sum_{\ell,\ell'}\langle D_\ell^TM^{\ell\ell'}D_{\ell'}w_{\ell'},w_\ell-w'_\ell\rangle\\
\qquad\qquad[\text{since }\sum_{\ell,\ell'}\langle D_\ell^TM^{\ell\ell'}D_{\ell'}z_{\ell'},z_\ell\rangle=0\text{ due to }M^{\ell\ell'}=-[M^{\ell'\ell}]^T]\\
=\sum_\ell\langle\nabla_{w_\ell}\mathcal{L}_\ell(\eta),w_\ell-w'_\ell\rangle=\sum_\ell[\mathcal{L}_\ell(\eta)-\mathcal{L}_\ell(w_1,...,w_{\ell-1},w'_\ell,w_{\ell+1},...,w_L)]\\
\qquad\qquad[\text{since }\mathcal{L}_\ell\text{ is affine in }w_\ell]\end{array}$$

and (37) follows.

3.6 Relation to [14]

Here we demonstrate that the decomposition approach to solving VI's with monotone operators on LMO-represented domains cover the approach, based on Fenchel-type representations, developed in [14]. Specifically, let H be a compact convex set in Euclidean space E_η, $G(\cdot)$ be a monotone vector field on H, and $\eta\mapsto A\eta+a$ be an affine mapping from E_η to Euclidean space E_ξ. Given a convex compact set $\Xi\subset E_\xi$, let us set

$$\Theta=\Xi\times H,\ \Phi(\xi,\eta)=[\Phi_\xi(\xi,\eta):=A\eta+a;\Phi_\eta(\xi,\eta):=G(\eta)-A^*\xi]:\Theta\to E_\xi\times E_\eta,\tag{38}$$

so that Φ clearly is a monotone vector field on Θ. Assume that $\overline{\eta}(\xi):\Xi\to H$ is a somehow selected strong solution to ${\rm VI}(\Phi_\eta(\xi,\cdot),H)$:

$$\forall\xi\in\Xi:\overline{\eta}(\xi)\in H\ \&\ \underbrace{\langle G(\overline{\eta}(\xi))-A^*\xi,\eta-\overline{\eta}(\xi)\rangle}_{=\langle\Phi_\eta(\xi,\overline{\eta}(\xi)),\eta-\overline{\eta}(\xi)\rangle}\geq0\,\forall\eta\in H;\tag{39}$$

(cf. (29)); note that required $\overline{\eta}(\xi)$ definitely exists, provided that $G(\cdot)$ is continuous and monotone. Let us also define $\overline{\xi}(\eta)$ as a selection of the point-to-set mapping $\eta\mapsto{\rm Argmin}_{\xi\in\Xi}\langle A\eta+a,\xi\rangle$, so that

$$\forall\eta\in H:\overline{\xi}(\eta)\in\Xi\ \&\ \underbrace{\langle A\eta+a,\xi-\overline{\xi}(\eta)\rangle}_{=\langle\Phi_\xi(\overline{\xi}(\eta),\eta),\xi-\overline{\xi}(\eta)\rangle}\geq0,\forall\xi\in\Xi\tag{40}$$

(cf. (30)).

Observe that with the just defined Ξ, H, Θ, Φ, $\overline{\eta}(\cdot)$, $\overline{\xi}(\cdot)$ we are in the direct product case of the situation described in section 3.2. Since we are in the direct product case, $(\Phi, \overline{\eta}(\cdot))$ is η-regular, and we can take, as the induced primal vector field associated with $(\Phi, \overline{\eta}(\cdot))$, the vector field

$$\Psi(\xi) = A\overline{\eta}(\xi) + a = \Phi_\xi(\xi, \overline{\eta}(\xi)) : \Xi \to E_\xi, \tag{41}$$

and as the induced dual vector field, the field

$$\Gamma(\eta) = G(\eta) - A^*\overline{\xi}(\eta) = \Phi_\eta(\overline{\xi}(\eta), \eta) : H \to E_\xi, \tag{42}$$

Note that in terms of [14], relations (41) and (39), modulo notation, form what in the reference is called a *Fenchel-type representation* (F.-t.r.) *of a vector field $\Psi : \Xi \to E_\xi$, the data of the representation being E_η, A, a, $\overline{\eta}(\cdot)$, $G(\cdot)$, H*; on a closer inspection, every F.-t.r. of a given monotone vector field $\Psi : \Xi \to E_\xi$ can be obtained in this fashion from some setup of the form (38).

Assume now that Ξ is LMO-representable, and we have at our disposal G-oracle which, given on input $\eta \in H$, returns $G(\eta)$. This oracle combines with LMO for Ξ to induce a procedure which, given on input $\eta \in H$, returns $\Gamma(\eta)$. As a result, we can apply the decomposition machinery presented in sections 3.2 – 3.5 to reduce solving $\mathrm{VI}(\Psi, \Xi)$ to processing (Γ, H) by an algorithm with accuracy certificates. It can be easily seen by inspection that this reduction recovers constructions and results presented in . The bottom line is that the developed in section 3 decomposition-based approach to solving VI's with monotone operators on LMO-represented domains covers the developed in [14, sections 1 – 4] approach based on Fenchel-type representations of monotone vector fields[5].

References

[1] Cox, B., "Applications of accuracy certificates for problems with convex structure" (2011) – Ph.D. Thesis, Georgia Institute of Technology
`https://smartech.gatech.edu/jspui/bitstream/1853/39489/1/cox_bruce_a_201105_phd.pdf`

[2] Cox, B., Juditsky, A., Nemirovski, A. (2013), "Dual subgradient algorithms for large-scale nonsmooth learning problems" – *Mathematical Programming Series B* (2013), Online First, DOI 10.1007/s10107-013-0725-1. E-print: arXiv:1302.2349.

[3] Demyanov, V., Rubinov, A. *Approximate Methods in Optimization Problems* – Elsevier, Amsterdam 1970.

[4] Dunn, J. C., Harshbarger, S. "Conditional gradient algorithms with open loop step size rules" – *Journal of Mathematical Analysis and Applications* **62:2** (1978), 432–444.

[5] Frank, M., Wolfe, P. "An algorithm for quadratic programming" – *Naval Res. Logist. Q.* **3:1-2** (1956), 95–110.

[6] Freund, R., Grigas, P. "New Analysis and Results for the Conditional Gradient Method" – *Mathematical Programming* (2014), http://dx.doi.org/10.1007/s10107-014-0841-6

[7] Garber, D., Hazan, E. "Faster Rates for the Frank-Wolfe Method over Strongly-Convex Sets" – to appear 32nd International Conference on Machine Learning (ICML 2015), arXiv:1406.1305 (2014).

[5] "covers" instead of "is equivalent" stems from the fact that the scope of decomposition is not restricted to the setups of the form of (38)).

[8] Gol'shtein, E.G. "Direct-Dual Block Method of Lnear Programming" – *Avtomat. i Telemekh.* 1996 No. 11, 3–9 (in Russian; English translation: *Automation and Remote Control* **57:11** (1996), 1531–1536.

[9] Gol'shtein, E.G., Sokolov, N.A. "A decomposition algorithm for solving multicommodity production-and-transportation problem" – *Ekonomika i Matematicheskie Metody*, **33:1** (1997), 112-128.

[10] Harchaoui, Z., Douze, M., Paulin, M., Dudik, M., Malick, J. "Large-scale image classification with trace-norm regularization" – In CVPR, 2012.

[11] Harchaoui, Z., Juditsky, A., Nemirovski, A. "Conditional Gradient Algorithms for Norm-Regularized Smooth Convex Optimization" – *Mathematical Programming*. DOI 10.1007/s10107-014-0778-9 Online first, April 18, 2014.

[12] Jaggi, M. "Revisiting Frank-Wolfe: Projection-free sparse convex optimization" – In ICML, 2013.

[13] Jaggi, M., Sulovsky, M. "A simple algorithm for nuclear norm regularized problems" – In ICML, 2010.

[14] Juditsky, A., Nemirovski, A. "Solving Variational Inequalities with Monotone Operators on Domains Given by Linear Minimization Oracles" – *Mathematical Programming*, Online First, March 22, 2015
`http://www.springerlink.com/openurl.asp?genre=article&id=doi:10.1007/s10107-015-0876-3`

[15] Nemirovski, A., Onn, S., Rothblum, U., "Accuracy certificates for computational problems with convex structure" – *Mathematics of Operations Research*, **35** (2010), 52-78.

[16] Pshenichnyi, B.N., Danilin, Y.M. *Numerical Methods in Extremal Problems*. Mir Publishers, Moscow 1978.

A Appendix

A.1 Proof of Lemma 1

It suffices to prove the ϕ-related statements. Lipschitz continuity of ϕ in the direct product case is evident. Further, the function $\theta(x_1, x_2; y_1) = \max\limits_{y_2 \in Y_2[y_1]} \Phi(x_1, x_2; y_1, y_2)$ is convex and Lipschitz continuous in $x = [x_1; x_2] \in X$ for every $y_1 \in Y_1$, whence $\phi(x_1, y_1) = \min\limits_{x_2 \in X_2[x_1]} \theta(x_1, x_2; y_1)$ is convex and lower semicontinuous in $x_1 \in X_1$ (note that X is compact). On the other hand, $\phi(x_1, y_1) = \max\limits_{y_2 \in Y_2[y_1]} \min\limits_{x_2 \in X_2[x_1]} \Phi(x_1, x_2; y_1, y_2) = \max\limits_{y_2 \in Y_2[y_1]} \left[\chi(x_1; y_1, y_2) := \min\limits_{x_2 \in X_2[x_1]} \Phi(x_1, x_2; y_1, y_2)\right]$, so that $\chi(x_1; y_1, y_2)$ is concave and Lipschitz continuous in $y = [y_1; y_2] \in Y$ for every $x_1 \in X_1$, whence

$$\phi(x_1, y_1) = \max_{y_2 \in Y_2[y_1]} \chi(x_1; y_1, y_2)$$

is concave and upper semicontinuous in $y_1 \in Y_1$ (note that Y is compact).

Next, we have

$$\begin{aligned}
\mathrm{SadVal}(\phi, X_1, X_2) &= \inf_{x_1\in X_1}\left[\sup_{y_1\in Y_1}\left[\sup_{y_2:[y_1;y_2]\in Y}\inf_{x_2:[x_1;x_2]\in X}\Phi(x_1,x_2;y_1,y_2)\right]\right]\\
&= \inf_{x_1\in X_1}\left[\sup_{[y_1;y_2]\in Y}\inf_{x_2:[x_1;x_2]\in X}\Phi(x_1,x_2;y_1,y_2)\right]\\
&= \inf_{x_1\in X_1}\left[\inf_{x_2:[x_1;x_2]\in X}\sup_{[y_1;y_2]\in Y}\Phi(x_1,x_2;y_1,y_2)\right] \text{ [by Sion-Kakutani Theorem]}\\
&= \inf_{[x_1;x_2]\in X}\sup_{[y_1;y_2]\in Y}\Phi(x_1,x_2;y_1,y_2) = \mathrm{SadVal}(\Phi, X, Y),
\end{aligned}$$

as required in (2). Finally, let $\bar{x} = [\bar{x}_1;\bar{x}_2] \in X$ and $\bar{y} = [\bar{y}_1;\bar{y}_2] \in Y$. We have

$$\begin{aligned}
\overline{\phi}(\bar{x}_1) - \mathrm{SadVal}(\phi, X_1, Y_1) &= \overline{\phi}(\bar{x}_1) - \mathrm{SadVal}(\Phi, X, Y) \text{ [by (2)]}\\
&= \sup_{y_1\in Y_1}\phi(\bar{x}_1, y_1) - \mathrm{SadVal}(\Phi, X, Y)\\
&= \sup_{y_1\in Y_1}\sup_{y_2:[y_1;y_2]\in Y}\inf_{x_2:[\bar{x}_1;x_2]\in X}\Phi(\bar{x}_1,x_2;y_1,y_2) - \mathrm{SadVal}(\Phi, X, Y)\\
&= \sup_{[y_1;y_2]\in Y}\inf_{x_2:[\bar{x}_1;x_2]\in X}\Phi(\bar{x}_1,x_2;y_1,y_2) - \mathrm{SadVal}(\Phi, X, Y)\\
&= \inf_{x_2:[\bar{x}_1;x_2]\in X}\sup_{y=[y_1;y_2]\in Y}\Phi(\bar{x}_1,x_2;y) - \mathrm{SadVal}(\Phi, X, Y)\\
&\le \sup_{y=[y_1;y_2]\in Y}\Phi(\bar{x}_1,\bar{x}_2;y) - \mathrm{SadVal}(\Phi, X, Y)\\
&= \overline{\Phi}(\bar{x}) - \mathrm{SadVal}(\Phi, X, Y)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{SadVal}(\phi, X_1, Y_1) - \underline{\phi}(\bar{y}_1) &= \mathrm{SadVal}(\Phi, X, Y) - \underline{\phi}(\bar{y}_1) \text{ [by (2)]}\\
&= \mathrm{SadVal}(\Phi, X, Y) - \inf_{x_1\in X_1}\phi(x_1, \bar{y}_1)\\
&= \mathrm{SadVal}(\Phi, X, Y) - \inf_{x_1\in X_1}\left[\inf_{x_2:[x_1;x_2]\in X}\sup_{y_2:[\bar{y}_1;y_2]\in Y}\Phi(x_1,x_2;\bar{y}_1,y_2)\right]\\
&= \mathrm{SadVal}(\Phi, X, Y) - \inf_{x=[x_1;x_2]\in X}\sup_{y_2:[\bar{y}_1;y_2]\in Y}\Phi(x;\bar{y}_1,y_2)\\
&\le \mathrm{SadVal}(\Phi, X, Y) - \inf_{x=[x_1;x_2]\in X}\Phi(x;\bar{y}_1,\bar{y}_2)\\
&= \mathrm{SadVal}(\Phi, X, Y) - \underline{\Phi}(\bar{y}).
\end{aligned}$$

We conclude that

$$\begin{aligned}
&\epsilon_{\mathrm{sad}}([\bar{x}_1;\bar{y}_1]|\phi, X_1, Y_1) = \left[\overline{\phi}(\bar{x}_1) - \mathrm{SadVal}(\phi, X_1, Y_1)\right] + \left[\mathrm{SadVal}(\phi, X_1, Y_1) - \underline{\phi}(\bar{y}_1)\right]\\
&\le \left[\overline{\Phi}(\bar{x}) - \mathrm{SadVal}(\Phi, X, Y)\right] + \left[\mathrm{SadVal}(\Phi, X, Y) - \underline{\Phi}(\bar{y})\right] = \epsilon_{\mathrm{sad}}([\bar{x};\bar{y}]|\Phi, X, Y),
\end{aligned}$$

as claimed in (3). □

A.2 Proof of Lemma 2

For $x_1 \in X_1$ we have

$$\begin{aligned}
&\phi(x_1;\bar{y}_1) = \min_{x_2:[x_1;x_2]\in X}\max_{y_2:[\bar{y}_1;y_2]\in Y}\Phi(x_1,x_2;\bar{y}_1,y_2) \ge \min_{x_2:[x_1;x_2]\in X}\Phi(x_1,x_2;\bar{y}_1,\bar{y}_2)\\
&\ge \min_{x_2:[x_1;x_2]\in X}\big[\underbrace{\Phi(\bar{x};\bar{y})}_{\phi(\bar{x}_1;\bar{y}_1))} + \langle G, [x_1;x_2] - [\bar{x}_1;\bar{x}_2]\big]\rangle \text{ [since } \Phi(x;\bar{y}) \text{ is convex and } G \in \partial_x\Phi(\bar{x};\bar{y})]\\
&\ge \phi(\bar{x}_1;\bar{y}_1) + \langle g, x_1 - \bar{x}_1\rangle \text{ [by definition of } g, G],
\end{aligned}$$

as claimed in (a). "Symmetric" reasoning justifies (b). □

A.3 Proof of Lemma 3

Assume that (5) holds true. Then G clearly is certifying, implying that

$$\chi_G(\bar{x}_1) = \langle G, [\bar{x}_1; \bar{x}_2] \rangle,$$

and therefore (5) reads

$$\langle G, [x_1; x_2] \rangle \geq \chi_G(\bar{x}_1) + \langle g, x_1 - \bar{x}_1 \rangle \ \forall x = [x_1; x_2] \in X,$$

where, taking minimum in the left hand side over $x_2 \in X_2[x_1]$,

$$\chi_G(x_1) \geq \chi_G(\bar{x}_1) + \langle g, x_1 - \bar{x}_1 \rangle \ \forall x_1 \in X_1,$$

as claimed in (ii).

Now assume that (i) and (ii) hold true. By (i), $\chi_G(\bar{x}_1) = \langle G, [\bar{x}_1; \bar{x}_2] \rangle$, and by (ii) combined with the definition of χ_G,

$$\forall x = [x_1; x_2] \in X : \langle G, [x_1; x_2] \rangle \geq \chi_G(x_1) \geq \chi_G(\bar{x}_1) + \langle g, x_1 - \bar{x}_1 \rangle = \langle G, \bar{x} \rangle + \langle g, x_1 - \bar{x}_1 \rangle,$$

implying (5). □

A.4 Dynamic Programming generated well-organized matrices

Consider the situation as follows. There exists an evolving in time system $\mathcal{S}$, with state ξ_s at time $s = 1, 2, ..., m$ belonging to a given finite nonempty set Ξ_s. Further, every pair (ξ, s) with $s \in \{1, ..., m\}$, $\xi \in \Xi_s$ is associated with nonempty finite *set of actions* A^s_ξ, and we set

$$\mathcal{S}_s = \{(\xi, a) : \xi \in \Xi_s, a \in A^s_\xi\}.$$

Further, for every s, $1 \leq s < m$, a *transition mapping* $\pi_{s+1}(\xi, a) : \mathcal{S}_s \to \Xi_{s+1}$ is given. Finally, we are given vector-valued functions ("outputs") $\chi_s : \mathcal{S}_s \to \mathbf{R}^{r_s}$.

A *trajectory* of $\mathcal{S}$ is a sequence $\{(\xi_s, a_s) : 1 \leq s \leq m\}$ such that $(\xi_s, a_s) \in \mathcal{S}_s$ for $1 \leq s \leq m$ and

$$\xi_{s+1} = \pi_s(\xi_s, a_s),\ 1 \leq s < m.$$

The *output* of a trajectory $\tau = \{(\xi_s, a_s) : 1 \leq s \leq m\}$ is the block-vector $\chi[\tau] = [\chi_1(\xi_1, a_1); ...; \chi_m(\xi_m, a_m)]$. We can associate with $\mathcal{S}$ the matrix $D = D[\mathcal{S}]$ with $K = r_1 + ... + r_m$ rows and with columns indexed by the trajectories of $\mathcal{S}$; specifically, the column indexed by a trajectory τ is $\chi[\tau]$.

For example, knapsack generated matrix D is of the form $D[\mathcal{S}]$ with system $\mathcal{S}$ as follows:

- Ξ_s, $s = 1, ..., m$, is the set of nonnegative integers which are $\leq H$;
- A^s_ξ is the set of nonnegative integers p_s such that $p_s \leq \bar{p}_s$ and $\xi - h_s p_s \geq 0$;
- the transition mappings are $\pi_{s+1}(\xi, a) = \xi - a h_s$;
- the outputs are $\chi_s(\xi, a) = f_s(a)$, $1 \leq s \leq m$.

Observe that matrix $D = D[\mathcal{S}]$ is well organized, provided the cardinalities of Ξ_s and A^s_ξ are reasonable. Indeed, given $x = [x_1; ...; x_m] \in \mathbf{R}^n = \mathbf{R}^{r_1} \times ... \times \mathbf{R}^{r_m}$, we can identify $\overline{D}[x]$ by Dynamic Programming, running first the backward Bellman recurrence

$$\left.\begin{array}{rcl} U_s(\xi) & = & \max_{a \in A^s_\xi} \left\{x_s^T f_s(\xi, a) + U_{s+1}(\xi - h_s a)\right\} \\ A_s(\xi) & = & \mathrm{Argmax}\ a \in A^s_\xi \left\{x_s^T f_s(\xi, a) + U_{s+1}(\xi - h_s a)\right\} \end{array} , \xi \in \Xi_s \right\}, s = m, m-1, ..., 1$$

(where $U_{m+1}(\cdot) \equiv 0$), and then identify the (trajectory indexing the) column of D corresponding to $\overline{D}[x]$ by running the forward Bellman recurrence

$$\begin{array}{rcl} \xi_1 & \in & \mathrm{Argmax}_{\xi \in \Xi_1} U_1(\xi) \Rightarrow a_1 \in A_1(\xi_1) \Rightarrow ... \\ \Rightarrow \xi_{s+1} & = & \pi_s(\xi_s, a_s) \Rightarrow a_{s+1} \in A_{s+1}(\xi_{s+1}) \Rightarrow ... \end{array} , s = 1, 2, ..., m-1.$$

A.5 Attacker vs. Defender via Ellipsoid algorithm

In our implementation,

1. Relation (26) is ensured by specifying U, V as centered at the origin Euclidean balls of radius R, where R is an upper bound on the Euclidean norms of the columns in D and in A (such a bound can be easily obtained from the knapsack data specifying the matrices D, A).
2. We processed the monotone vector field associated with the primal SP problem (28), that is, the field
$$F(u, v) = [F_u(u, v) = \overline{A}[u] - v; F_v(u, v) = u - \underline{D}[v]]$$
by Ellipsoid algorithm with accuracy certificates from [15]. For $\tau = 1, 2, ...$, the algorithm generates *search points* $[u_\tau; v_\tau] \in \mathbf{R}^K \times \mathbf{R}^K$, with $[u_1; v_1] = 0$, along with execution protocols $\mathcal{I}^\tau = \{[u_i; v_i], F(u_i, v_i) : i \in I_\tau\}$, where $I_\tau = \{i \leq \tau : [u_i; v_i] \in U \times V\}$, augmented by accuracy certificates $\lambda^\tau = \{\lambda^\tau_i \geq 0 : i \in I_\tau\}$ such that $\sum_{i \in I_\tau} \lambda^\tau_i = 1$. From the results of [15] it follows that for every $\epsilon > 0$ it holds
$$\tau \geq N(\epsilon) := O(1)K^2 \ln\left(2\frac{R + \epsilon}{\epsilon}\right) \Rightarrow \mathrm{Res}(\mathcal{I}^\tau, \lambda^\tau | U \times V) \leq \epsilon. \tag{43}$$
3. When computing $F(u_i, v_i)$ (this computation takes place only at *productive* steps – those with $[u_i; v_i] \in U \times V$), we get, as a byproduct, the columns $A^i = \overline{A}[u_i]$ and $D^i = \underline{D}[v_i]$ of matrices A, D, along with the indexes q^i, p^i of these columns (recall that these indexes, according to the construction of A, D, are collections of m nonnegative integers). In our implementation, we stored these columns, same as their indexes and the corresponding search points $[u_i; v_i]$. As is immediately seen, in the case in question the approximate solution $[w^\tau; z^\tau]$ to the SP problem of interest (25) induced by execution protocol $\mathcal{I}^\tau$ and accuracy certificate λ^τ is comprised of two sparse vectors
$$w^\tau = \sum_{i \in I_\tau} \lambda^\tau_i \delta^D_{p^i},\ z^\tau = \sum_{i \in I_\tau} \lambda^\tau_i \delta^A_{q^i} \tag{44}$$
where δ^D_p is the "p-th basic orth" in the simplex Δ_N of probabilistic vectors indexed by pure strategies of Defender, and similarly for δ^A_q. Thus, we have no difficulties with representing our approximate solutions[6], in spite of their huge ambient dimension.

[6] Note that applying Caratheodory theorem, we could further "compress" the representations of approximate solutions – make these solutions convex combinations of at most $K + 1$ of $\delta^D_{p^i}$'s and $\delta^A_{p^i}$'s.

According to our general theory and (43), the number of steps needed to get an ϵ-solution $[w; z]$ to the problem of interest (i.e., a feasible solution with $\epsilon_{\rm sad}([w; z]|\psi, W, Z) \le \epsilon$) does not exceed $N(\epsilon)$, with computational effort per step dominated by the necessity to identify $\overline{A}[u_i]$, $\underline{D}[v_i]$ by Dynamic Programming.

In fact, we used the outlined scheme with two straightforward modifications.

- First, instead of building the accuracy certificates λ^τ according to the rules from [15], we used the best, given execution protocols $\mathcal{I}^\tau$, accuracy certificates by solving the convex program

$$\min_\lambda \left\{ \mathrm{Res}(\mathcal{I}^\tau, \lambda) := \max_{y \in U \times V} \sum_{i \in I_\tau} \lambda_i \langle F(u_i, v_i), [u_i; v_i] - y \rangle : \lambda_i \ge 0, \sum_{i \in I_\tau} \lambda_i = 1 \right\}$$

  In our implementation, this problem was solved from time to time, specifically, once per $4K^2$ steps; with our simple U, V, this problem is well within the scope of `cvx`.

- Second, given current approximate solution (44) to the problem of interest, we can compute its saddle point inaccuracy exactly instead of upper-bounding it by $\mathrm{Res}(\mathcal{I}^\tau, \lambda^\tau | U \times V)$. Indeed, it is immediately seen that

$$\epsilon_{\rm sad}([w^\tau; z^\tau]|\psi, W, Z) = \mathrm{Max}(A^T[\sum_{i \in I_\tau} \lambda_i^\tau D^i]) - \mathrm{Min}(D^T[\sum_{i \in I_\tau} \lambda_i^\tau A^i]).$$

  In our implementation, we performed this computation each time when a new accuracy certificate was computed, and terminated the solution process when the saddle point inaccuracy became less than a given threshold (1.e-4).

A.6 Proof of Proposition 2

(i): Let $\xi_1, \xi_2 \in \Xi$, and let $\eta_1 = \overline{\eta}(\xi_1)$, $\eta_2 = \overline{\eta}(\xi_2)$. By (31) we have

$$\begin{array}{rcl} \langle \Psi(\xi_2), \xi_2 - \xi_1 \rangle & \ge & \langle \Phi(\xi_2, \eta_2), [\xi_2 - \xi_1; \eta_2 - \eta_1] \rangle \\ \langle \Psi(\xi_1), \xi_1 - \xi_2 \rangle & \ge & \langle \Phi(\xi_1, \eta_1), [\xi_1 - \xi_2; \eta_1 - \eta_2] \rangle \end{array}$$

Summing inequalities up, we get

$$\langle \Psi(\xi_2) - \Psi(\xi_1), \xi_2 - \xi_1 \rangle \ge \langle \Phi(\xi_2, \eta_2) - \Phi(\xi_1, \eta_1), [\xi_2 - \xi_1; \eta_2 - \eta_1] \rangle \ge 0,$$

so that Ψ is monotone.

Further, the first inequality in (33) is due to Lemma 6. To prove the second inequality in (33), let $\mathcal{I}_t = \{\xi_i \in \Xi, \Psi(\xi_i) : 1 \le i \le t\}$, $\mathcal{J}_t = \{\theta_i := [\xi_i; \overline{\eta}(\xi_i)], \Phi(\theta_i) : 1 \le i \le t\}$, and let λ be t-step accuracy certificate. We have

$$\begin{array}{l} \theta = [\xi; \eta] \in \Theta \Rightarrow \\ \sum_{i=1}^t \lambda_i \langle \Phi(\theta_i), \theta_i - \theta \rangle \le \sum_{i=1}^t \lambda_i \langle \Psi(\xi_i), \xi_i - \xi \rangle \ [\text{see (31)}] \\ \le \mathrm{Res}(\mathcal{I}_t, \lambda | \Xi) \\ \Rightarrow \mathrm{Res}(\mathcal{J}_t, \lambda | \Theta) = \sup_{\theta = [\xi; \eta] \in \Theta} \sum_{i=1}^t \lambda_i \langle \Phi(\theta_i), \theta_i - \theta \rangle \le \mathrm{Res}(\mathcal{I}_t, \lambda | \Xi). \end{array}$$

(i) is proved.

(ii): Let $\eta \in H$. Invoking (32), we have

$$\langle \Gamma(\eta), \widehat{\eta} - \eta \rangle \le \langle \Phi(\overline{\xi}(\eta), \eta), [\widehat{\xi}; \widehat{\eta}] - [\overline{\xi}(\eta); \eta] \rangle \le \epsilon_{\rm VI}(\widehat{\theta} | \Phi, \Theta),$$

and (34) follows. □