arXiv: math.PR/0809.0815v1

Solving variational inequalities with Stochastic Mirror-Prox algorithm

Anatoli Juditsky[??], **Arkadi Nemirovski**[*,??] **and Claire Tauvel**[??]

LJK, Université J. Fourier
B.P. 53
38041 Grenoble Cedex 9, France
e-mail: juditsky@imag.fr

claire.tauvel@imag.fr

ISyE
Georgia Institute of Technology
Atlanta, GA 30332, USA
e-mail: nemirovs@isye.gatech.edu

Abstract: In this paper we consider iterative methods for *stochastic variational inequalities* (s.v.i.) with monotone operators. Our basic assumption is that the operator possesses both smooth and nonsmooth components. Further, only noisy observations of the problem data are available. We develop a novel Stochastic Mirror-Prox (SMP) algorithm for solving s.v.i. and show that with the convenient stepsize strategy it attains the optimal rates of convergence with respect to the problem parameters. We apply the SMP algorithm to Stochastic composite minimization and describe particular applications to Stochastic Semidefinite Feasibility problem and deterministic Eigenvalue minimization.

AMS 2000 subject classifications: Primary 90C15, 65K10; secondary 90C47.
Keywords and phrases: variational inequalities with monotone operators, stochastic convex-concave saddle-point problem, large scale stochastic approximation, reduced complexity algorithms for convex optimization.

1. Introduction

Variational inequalities with monotone operators form a convenient framework for unified treatment (including algorithmic design) of problems with "convex structure," like convex minimization, convex-concave saddle point problems and convex Nash equilibrium problems. In this paper we utilize this framework to develop first order algorithms for *stochastic* versions of the outlined problems, where the precise first order information is replaced with its unbiased stochastic estimates. This situation arises naturally in convex Stochastic Programming, where the precise first order information is unavailable (see examples in section 4). In some situations, e.g. those considered in [4, Section 3.3] and in Section 4.4, where passing from available, but relatively computationally expensive precise

*Research of this author was partly supported by the Office of Naval Research grant # N000140811104 and the NSF grant DMS-0914785.

first order information to its cheap stochastic estimates allows to accelerate the solution process, with the gain from randomization growing progressively with problem's sizes.

Our "unifying framework" is as follows. Let Z be a convex compact set in Euclidean space $\mathcal{E}$ with inner product $\langle\cdot,\cdot\rangle$, $\|\cdot\|$ be a norm on E (not necessarily the one associated with the inner product), and $F: Z \to \mathcal{E}$ be a monotone mapping:

$$\forall (z, z' \in Z) : \langle F(z) - F(z'), z - z' \rangle \geq 0 \tag{1}$$

We are interested to approximate a solution to the variational inequality (v.i.)

$$\text{find } z_* \in Z : \langle F(z), z_* - z \rangle \leq 0 \quad \forall z \in Z \tag{2}$$

associated with Z, F. Note that since F is monotone on Z, the condition in (2) is implied by $\langle F(z_*), z - z_* \rangle \geq 0$ for all $z \in Z$, which is the standard definition of a (strong) solution to the v.i. associated with Z, F. The inverse – a solution to v.i. as defined by (2) (a "weak" solution) is a strong solution as well – also is true, provided, e.g., that F is continuous. An advantage of the concept of weak solution is that such a solution always exists under our assumptions (F is well defined and monotone on a convex compact set Z).

We quantify the inaccuracy of a candidate solution $z \in Z$ by the error

$$\mathrm{Err}_{\mathrm{vi}}(z) := \max_{u \in Z} \langle F(u), z - u \rangle; \tag{3}$$

note that this error is always ≥ 0 and equals zero iff z is a solution to (2).

In what follows we impose on F, aside of the monotonicity, the requirement

$$\forall (z, z' \in Z) : \|F(z) - F(z')\|_* \leq L\|z - z'\| + M \tag{4}$$

with some known constants $L \geq 0, M \geq 0$. From now on,

$$\|\xi\|_* = \max_{z:\|z\| \leq 1} \langle \xi, z \rangle \tag{5}$$

is the norm conjugate to $\|\cdot\|$.

We are interested in the case where (2) is solved by an iterative algorithm based on a *stochastic oracle representation* of the operator $F(\cdot)$. Specifically, when solving the problem, the algorithm acquires information on F via subsequent calls to a black box ("stochastic oracle", SO). At the ith call, $i = 0, 1, ...$, the oracle gets as input a search point $z_i \in Z$ (this point is generated by the algorithm on the basis of the information accumulated so far) and returns the vector $\Xi(z_i, \zeta_i)$, where $\{\zeta_i \in \mathbf{R}^N\}_{i=1}^{\infty}$ is a sequence of i.i.d. (and independent of the queries of the algorithm) random variables. We suppose that the Borel function $\Xi(z, \zeta)$ is such that

$$\forall z \in Z : \ \mathbf{E}\left\{\Xi(z, \zeta_1)\right\} = F(z), \ \ \mathbf{E}\left\{\|\Xi(z, \zeta_i) - F(z)\|_*^2\right\} \leq \sigma^2. \tag{6}$$

We call a monotone v.i. (1), augmented by a stochastic oracle (SO), a *stochastic* monotone v.i. (s.v.i.).

To motivate our goal, let us start with known results [7] on the limits of performance of iterative algorithms for solving large-scale stochastic monotone v.i.'s. To "normalize" the situation, assume that Z is the unit Euclidean ball in $\mathcal{E} = \mathbf{R}^n$ and that n is large. In this case, the accuracy after t steps of any algorithm for solving v.i.'s cannot be better than $O(1)\left[\frac{L}{t} + \frac{M+\sigma}{\sqrt{t}}\right]$. In other words, for a properly chosen positive absolute constant C, for every number of steps t, all large enough values of n and any algorithm $\mathcal{B}$ for solving s.v.i.'s on the unit ball of $\mathbf{R}^n$, one can point out a monotone s.v.i. satisfying (4), (6) and such that the expected error of the approximate solution $\tilde{z}_t$ generated by $\mathcal{B}$ after t steps , applied to such s.v.i., is at least $c\left[\frac{L}{t} + \frac{M+\sigma}{\sqrt{t}}\right]$ for some $c > 0$. To the best of our knowledge, no existing algorithm allows to achieve, uniformly in the dimension, this convergence rate. In fact, the "best approximations" available are given by Robust Stochastic Approximation (see [4] and references therein) with the guaranteed rate of convergence $O(1)\frac{L+M+\sigma}{\sqrt{t}}$ and extra-gradient-type algorithms for solving deterministic monotone v.i.'s with Lipschitz continuous operators (see [8, 11–13]), which attain the accuracy $O(1)\frac{L}{t}$ in the case of $M = \sigma = 0$ or $O(1)\frac{M}{\sqrt{t}}$ when $L = \sigma = 0$.

The goal of this paper is to demonstrate that a specific *Mirror-Prox* algorithm [8] for solving monotone v.i.'s with Lipschitz continuous operators can be extended onto monotone s.v.i.'s to yield, uniformly in the dimension, the optimal rate of convergence $O(1)\left[\frac{L}{t} + \frac{M+\sigma}{\sqrt{t}}\right]$. We present the corresponding extension and investigate it in detail: we show how the algorithm can be "tuned" to the geometry of the s.v.i. in question and derive bounds for the probability of large deviations of the resulting error. We also present a number of applications where the specific structure of the rate of convergence indeed "makes a difference."

The main body of the paper is organized as follows: in Section 2, we describe several special cases of monotone v.i.'s we are especially interested in (convex Nash equilibria, convex-concave saddle point problems, convex minimization). We single out these special cases since here one can define a useful "functional" counterpart $\mathrm{Err}_{\mathrm{N}}(\cdot)$ of the just defined error $\mathrm{Err}_{\mathrm{vi}}(\cdot)$; both $\mathrm{Err}_{\mathrm{N}}$ and $\mathrm{Err}_{\mathrm{vi}}$ will participate in our subsequent efficiency estimates. Our main development – the *Stochastic Mirror Prox* (SMP) algorithm – is presented in Section 3. where we also provide some general results about its performance. Then in Section 4 we present SMP for Stochastic composite minimization and discuss its applications to Stochastic Semidefinite Feasibility problem and Eigenvalue minimization. All technical proofs are collected in the appendix.

Notations. In the sequel, lowercase Latin letters denote vectors (and sometimes matrices). Script capital letters, like $\mathcal{E}$, $\mathcal{Y}$, denote Euclidean spaces; the inner product in such a space, say, $\mathcal{E}$, is denoted by $\langle\cdot,\cdot\rangle_{\mathcal{E}}$ (or merely $\langle\cdot,\cdot\rangle$, when the corresponding space is clear from the context). Linear mappings from one Euclidean space to another, say, from $\mathcal{E}$ to $\mathcal{F}$, are denoted by boldface capitals like $\mathbf{A}$ (there are also some reserved boldface capitals, like $\mathbf{E}$ for expectation, $\mathbf{R}^k$ for the k-dimensional coordinate space, and $\mathbf{S}^k$ for the space of $k \times k$ symmetric matrices). $\mathbf{A}^*$ stands for the conjugate to mapping $\mathbf{A}$: if $\mathbf{A} : \mathcal{E} \to \mathcal{F}$, then $\mathbf{A}^* : \mathcal{F} \to \mathcal{E}$ is given by the identity $\langle f, \mathbf{A}e\rangle_{\mathcal{F}} = \langle \mathbf{A}^* f, e\rangle_{\mathcal{E}}$ for $f \in \mathcal{F}, e \in \mathcal{E}$. When both the origin and the destination space of a linear map, like $\mathbf{A}$, are the standard coordinate spaces, the map is identified with its matrix A, and $\mathbf{A}^*$ is identified with A^T. For a norm $\|\cdot\|$ on $\mathcal{E}$, $\|\cdot\|_*$ stands for the conjugate norm, see (5).

For Euclidean spaces $\mathcal{E}_1, ..., \mathcal{E}_m$, $\mathcal{E} = \mathcal{E}_1 \times ... \times \mathcal{E}_m$ denotes their Euclidean direct product, so that a vector from $\mathcal{E}$ is a collection $u = [u_1; ...; u_m]$ ("MATLAB notation") of vectors $u_\ell \in \mathcal{E}_\ell$, and $\langle u, v\rangle_{\mathcal{E}} = \sum_\ell \langle u_\ell, v_\ell\rangle_{\mathcal{E}_\ell}$. Sometimes we allow ourselves to write $(u_1, ..., u_m)$ instead of $[u_1; ...; u_m]$.

2. Preliminaries and Problem of interest

2.1. Nash v.i.'s and functional error

In the sequel, we shall be especially interested in a special case of v.i. (2) – in a *Nash* v.i. coming from a *convex Nash Equilibrium* problem, and in the associated *functional* error measure. The Nash Equilibrium problem can be described as follows: there are m players, the ith of them choosing a point z_i from a given set Z_i. The loss of the ith player is a given function $\phi_i(z)$ of the collection $z = (z_1, ..., z_m) \in Z = Z_1 \times ... \times Z_m$ of players' choices. With slight abuse of notation, we use for $\phi_i(z)$ also the notation $\phi_i(z_i, z^i)$, where z^i is the collection of choices of all but the ith players. Players are interested to minimize their losses, and Nash equilibrium $\widehat{z}$ is a point from Z such that for every i the function $\phi_i(z_i, \widehat{z}^i)$ attains its minimum in $z_i \in Z_i$ at $z_i = \widehat{z}_i$ (so that in the state $\widehat{z}$ no player has an incentive to change his choice, provided that the other players stick to their choices).

We call a Nash equilibrium problem *convex*, if for every i, Z_i is a compact convex set, $\phi_i(z_i, z^i)$ is a Lipschitz continuous function convex in z_i and concave in z^i, and the function $\Phi(z) = \sum_{i=1}^m \phi_i(z)$ is convex. It is well known (see, e.g., [10]) that setting

$$F(z) = \left[F^1(z); \ldots; F^m(z)\right],\ F^i(z) \in \partial_{z_i}\phi_i(z_i, z^i),\ i = 1, ..., m$$

where $\partial_{z_i}\phi_i(z_i, z^i)$ is the subdifferential of the convex function $\phi_i(\cdot, z^i)$ at a point z_i, we get a monotone operator such that the solutions to the corresponding v.i. (2) are exactly the Nash equilibria. Note that since ϕ_i are Lipschitz continuous, the associated operator F can be chosen to be bounded. For this v.i. one can consider, along with the v.i.-accuracy measure $\mathrm{Err}_{\mathrm{vi}}(z)$, the *functional* error measure

$$\mathrm{Err}_{\mathrm{N}}(z) = \sum_{i=1}^m \left[\phi_i(z) - \min_{w_i \in Z_i} \phi_i(w_i, z^i)\right]$$

This accuracy measure admits a transparent justification: this is the sum, over the players, of the incentives for a player to change his choice given that other players stick to their choices.

2.1.1. Special case: saddle points

An important by its own right particular case of Nash Equilibrium problem is a *zero sum game*, where $m = 2$ and $\Phi(z) \equiv 0$ (i.e., $\phi_2(z) \equiv -\phi_1(z)$). The convex case of this problem corresponds to the situation when $\phi(z_1, z_2) \equiv \phi_1(z_1, z_2)$ is a Lipschitz continuous function which is convex in $z_1 \in Z_1$ and concave in $z_2 \in Z_2$, the Nash equilibria are exactly the saddle points (min in z_1, max in z_2) of ϕ on $Z_1 \times Z_2$, and the functional error becomes

$$\mathrm{Err}_{\mathrm{N}}(z_1, z_2) = \max_{(u_1,u_2)\in Z} \left[\phi(z_1, u_1) - \phi(u_2, z_2)\right].$$

Recall that the convex-concave saddle point problem $\min_{z_1\in Z_1} \max_{z_2\in Z_2} \phi(z_1, z_2)$ gives rise to the "primal-dual" pair of convex optimization problems

$$(P): \min_{z_1\in Z_1} \overline{\phi}(z_1), \qquad (D): \max_{z_2\in Z_2} \underline{\phi}(z_2),$$

where

$$\overline{\phi}(z_1) = \max_{z_2\in Z_2} \phi(z_1, z_2), \quad \underline{\phi}(z_2) = \min_{z_1\in Z_1} \phi(z_1, z_2).$$

The optimal values $\mathrm{Opt}(P)$ and $\mathrm{Opt}(D)$ in these problems are equal, the set of saddle points of ϕ (i.e., the set of Nash equilibria of the underlying convex Nash problem) is exactly the direct product of the optimal sets of (P) and (D), and $\mathrm{Err}_{\mathrm{N}}(z_1, z_2)$ is nothing but the sum of non-optimalities of z_1, z_2 considered as approximate solutions to respective optimization problems:

$$\mathrm{Err}_{\mathrm{N}}(z_1, z_2) = \left[\overline{\phi}(z_1) - \mathrm{Opt}(P)\right] + \left[\mathrm{Opt}(D) - \underline{\phi}(z_2)\right].$$

In the sequel, we refer to the v.i. (2) coming from a convex Nash Equilibrium problem as *Nash* v.i., and to the just outlined particular case as the *Saddle Point* v.i. It is easy to verify that in the Saddle Point case the functional error $\mathrm{Err}_{\mathrm{N}}(z)$ is $\leq \mathrm{Err}_{\mathrm{vi}}(z)$; this is not necessary so for a general Nash v.i.

2.2. Composite Optimization problem and its saddle point reformulation

While the algorithm we intend to develop is applicable to a general-type stochastic v.i. with monotone operator, the applications to be considered in this paper deal with (saddle point reformulation of) *convex composite optimization problem* (cf. [7]).

As the simplest motivating example, one can keep in mind the minimax problem

$$\min_{x\in X} \max_{1\leq i\leq m} \phi_\ell(x), \tag{7}$$

where $X \subset \mathbf{R}^n$ is a convex compact set and $\phi_\ell(x)$ are Lipschitz continuous convex functions on X. This problem can be rewritten as the saddle point problem

$$\min_{x\in X} \max_{y\in Y} \phi(x, y) := \sum_{ell=1}^{m} y_\ell \phi_\ell(x), \tag{8}$$

where $Y = \{y \in \mathbf{R}^m_+ : \sum_{\ell=1}^m y_\ell = 1\}$ is the standard simplex. The advantages of the saddle point reformulation are twofold. First, when all ϕ_ℓ are smooth, so is ϕ, in contrast to the objective in (7) which typically is nonsmooth; this makes the saddle point reformulation better suited for processing by first order algorithms. Starting with the breakthrough paper of Nesterov [11], this phenomenon, in its general form, is utilized in the fastest known so far first order algorithms for "well-structured" nonsmooth convex programs. Second, in the stochastic case, stochastic oracles providing unbiased estimates of the first order information on ϕ_i oracles, while not induce a similar oracle for the objective of (7), do induce such an oracle for the v.i. associated with (8) and thus make the problem amenable to first order algorithms.

2.2.1. Composite minimization problem.

In this paper, we focus on a substantial extension of the minimax problem (7), namely, on a *Composite minimization problem*

$$\min_{x\in X} \phi(x) := \Phi(\phi_1(x), ..., \phi_m(x)), \tag{9}$$

where the inner functions $\phi_\ell(\cdot)$ are vector-valued, and the outer function Φ is real-valued. We are about to impose structural restrictions which allow to reformulate the problem as a "good" convex-concave saddle point problem, specifically, as follows:

A. $X \subset \mathcal{X}$ is a convex compact;

B. $\phi_\ell(x) : X \to \mathcal{E}_\ell$, $1 \le \ell \le m$, are Lipschitz continuous mappings taking values in Euclidean spaces $\mathcal{E}_\ell$ equipped with closed convex cones K_ℓ. We assume ϕ_ℓ to be K_ℓ-convex, meaning that for any $x, x' \in X$, $\lambda \in [0,1]$,

$$\phi_\ell(\lambda x + (1-\lambda)x') \le_{K_\ell} \lambda\phi_\ell(x) + (1-\lambda)\phi_\ell(x'),$$

where the notation $a \le_K b \Leftrightarrow b \ge_K a$ means that $b - a \in K$.

C. $\Phi(\cdot)$ is a convex function on $\mathcal{E} = \mathcal{E}_1 \times ... \times \mathcal{E}_m$ given by the Fenchel-type representation

$$\Phi(u_1, ..., u_m) = \max_{y\in Y} \left\{ \sum\nolimits_{\ell=1}^m \langle u_\ell, \mathbf{A}_\ell y + b_\ell\rangle_{\mathcal{E}_\ell} - \Phi_*(y) \right\}, \tag{10}$$

for $u_\ell \in \mathcal{E}_\ell$, $1 \le \ell \le m$. Here
— $Y \subset \mathcal{Y}$ is a convex compact set,
— the affine mappings $y \mapsto \mathbf{A}_\ell y + b_\ell : \mathcal{Y} \to \mathcal{E}_\ell$ are such that $\mathbf{A}_\ell y + b_\ell \in K_\ell^*$ for all $y \in Y$ and all ℓ, K_ℓ^* being the cone dual to K_ℓ,
— $\Phi_*(y)$ is a given Lipschitz continuous convex function on Y.

Under these assumptions, the optimization problem (9) is nothing but the primal problem associated with the saddle point problem

$$\min_{x\in X} \max_{y\in Y} \left[\phi(x,y) = \sum\nolimits_{\ell=1}^m \langle \phi_\ell(x), \mathbf{A}_\ell y + b_\ell\rangle_{\mathcal{E}_\ell} - \Phi_*(y) \right] \tag{11}$$

and the cost function in the latter problem is Lipschitz continuous and convex-concave due to the convexity of Φ_*, K_ℓ-convexity of $\phi_\ell(\cdot)$ and the condition $\mathbf{A}_\ell y + b_\ell \in K_\ell^*$ whenever $y \in Y$. The associated Nash v.i. is given by the domain $Z = X \times Y$ and the monotone mapping

$$F(z) \equiv F(x,y) = \left[\sum\nolimits_{\ell=1}^m [\phi_\ell'(x)]^*[\mathbf{A}_\ell y + b_\ell];\ -\sum\nolimits_{\ell=1}^m \mathbf{A}_\ell^* \phi_\ell(x) + \Phi_*'(y)\right]. \quad (12)$$

Same as in the case of minimax problem (7), the advantage of the saddle point reformulation (11) of (9) is that, independently of whether Φ is smooth, ϕ is smooth whenever all ϕ_ℓ are so. Another advantage, instrumental in the stochastic case, is that F is linear in $\phi_\ell(\cdot)$, so that stochastic oracles providing unbiased estimates of the first order information on ϕ_ℓ induce straightforwardly an unbiased SO for F.

2.2.2. *Example: Matrix Minimax problem*

For $1 \le \ell \le m$, let $\mathcal{E}_\ell = \mathbf{S}^{p_\ell}$ be the space of symmetric $p_\ell \times p_\ell$ matrices equipped with the Frobenius inner product $\langle A, B\rangle_F = \mathrm{Tr}(AB)$, and let K_ℓ be the cone $\mathbf{S}_+^{p_\ell}$ of symmetric positive semidefinite $p_\ell \times p_\ell$ matrices. Now let $X \subset \mathcal{X}$ be a convex compact set, and $\phi_\ell : X \to \mathcal{E}_\ell$ be $\mathbf{S}_+^{p_\ell}$-convex Lipschitz continuous mappings. These data induce the *Matrix Minimax problem*

$$\min_{x\in X} \max_{1\le j\le k} \lambda_{\max}\left(\sum\nolimits_{\ell=1}^m P_{j\ell}^T \phi_\ell(x) P_{j\ell}\right), \quad (P)$$

where $P_{j\ell}$ are given $p_\ell \times q_j$ matrices, and $\lambda_{\max}(A)$ is the maximal eigenvalue of a symmetric matrix A. Observing that for a symmetric $q\times q$ matrix A one has

$$\lambda_{\max}(A) = \max_{S\in\mathcal{S}_q} \mathrm{Tr}(AS)$$

where $\mathcal{S}_q = \{S \in \mathbf{S}_+^q : \mathrm{Tr}(S) = 1\}$. When denoting by Y the set of all symmetric positive semidefinite block-diagonal matrices $y = \mathrm{Diag}\{y_1, ..., y_k\}$ with unit trace and diagonal blocks y_j of sizes $q_j \times q_j$, we can represent (P) in the form of (9), (10) with

$$\begin{array}{rl}\Phi(u) := & \max\limits_{1\le j\le k} \lambda_{\max}\left(\sum_{\ell=1}^m P_{j\ell} u_\ell P_{j\ell}^T\right) = \max\limits_{y\in Y} \sum_{j=1}^k \mathrm{Tr}\left(\sum_{\ell=1}^m P_{j\ell}^T u_\ell P_{j\ell} y_j\right)\\ = & \max\limits_{y\in Y} \sum_{\ell=1}^m \mathrm{Tr}\left(u_\ell\left[\sum_{j=1}^k P_{j\ell}^T y_j P_{j\ell}\right]\right) = \max\limits_{y\in Y} \sum_{\ell=1}^m \langle u_\ell, \mathbf{A}_\ell y\rangle_F,\\ \mathbf{A}_\ell y = & \sum_{j=1}^k P_{j\ell} y_j P_{j\ell}^T\end{array}$$

Observe that in the simplest case of $k = m$, $p_j = q_j$, $1 \le j \le m$ and $P_{j\ell}$ equal to I_p for $j = \ell$ and to 0 otherwise, the problem becomes

$$\min_{x\in X}\left[\max_{1\le\ell\le m} \lambda_{\max}(\phi_\ell(x))\right]. \quad (13)$$

If, in addition, $p_j = q_j = 1$ for all j, we arrive at the convex minimax problem (7).

Illustration: Semidefinite Feasibility problem. With X and ϕ_ℓ as above, consider the Semidefinite Feasibility problem

$$\text{find } x \in X : \psi_\ell(x) \preceq 0,\ 1 \le \ell \le m. \tag{S}$$

Choosing somehow scaling factors $\beta_\ell > 0$ and setting $\phi_\ell(x) = \beta_\ell \psi_\ell(x)$, we can pose $(\mathcal{S})$ as the Matrix Minimax problem $\min\limits_{x\in X} \max\limits_{1\le\ell\le m} \lambda_{\max}(\phi_\ell(x))$; $(\mathcal{S})$ is solvable if and only if the optimal value in the Matrix Minimax problem is ≤ 0.

3. Stochastic Mirror-Prox algorithm

We are about to present the stochastic version of the deterministic Mirror-Prox algorithm proposed in [8]. The method is aimed at solving v.i. (2) associated with a convex compact set $Z \subset \mathcal{E}$ and a *bounded* monotone operator $F : Z \to \mathcal{E}$. In contrast to the original version of the method, below we allow for errors when computing the values of F – we assume that given a point $z \in Z$, we can compute an approximation (perhaps random) $\widehat{F}(z) \in \mathcal{E}$ of $F(z)$.

3.1. Algorithm's setup

The setup for SMP (Stochastic Mirror Prox algorithm) is given by

1. a norm $\|\cdot\|$ on $\mathcal{E}$; $\|\cdot\|_*$ stands for the conjugate norm, see (5);
2. a *distance-generating function* (d.-g.f.) for Z, that is, a continuous convex function $\omega(\cdot) : Z \to \mathbf{R}$ such that
   (a) with Z^o being the set of all points $z \in Z$ such that the subdifferential $\partial\omega(z)$ of $\omega(\cdot)$ at z is nonempty, $\partial\omega(\cdot)$ admits a continuous selection on Z^o: there exists a continuous on Z^o vector-valued function $\omega'(z)$ such that $\omega'(z) \in \partial\omega(z)$ for all $z \in Z^o$;
   (b) $\omega(\cdot)$ is strongly convex, modulus 1, w.r.t. the norm $\|\cdot\|$:

$$\forall (z, z' \in Z^o) : \langle \omega'(z) - \omega'(z'), z - z' \rangle \ge \|z - z'\|^2. \tag{14}$$

In order for the SMP associated with the outlined setup to be practical, $\omega(\cdot)$ and Z should "fit" each other, meaning that one can easily solve problems of the form

$$\min_{z\in Z} \left[\omega(z) + \langle e, z\rangle\right], \quad e \in \mathcal{E}. \tag{15}$$

The prox-function associated with a setup for SMP is defined as

$$V(z,u) = \omega(u) - \omega(z) - \langle \omega'(z), u - z\rangle : Z^o \times Z \to \mathbf{R}^+.$$

We set

$$\begin{array}{lrcl} (a) & \Theta(z) & = & \max_{u\in Z} V(z,u) \quad [z \in Z^o]; \\ (b) & z_c & = & \operatorname{argmin}_Z \omega(z); \\ (c) & \Omega & = & \sqrt{2\Theta(z_c)}. \end{array} \tag{16}$$

Note that z_c is well defined (since Z is a convex compact set and $\omega(\cdot)$ is continuous and strongly convex on Z) and belongs to Z^o (since $0 \in \partial\omega(z_c)$). Note also that due to the strong convexity of ω and the origin of z_c we have

$$\forall (u \in Z) : \frac{1}{2}\|u - z_c\|^2 \leq \Theta(z_c) \leq \max_{z \in Z} \omega(z) - \omega(z_c); \tag{17}$$

in particular we see that

$$Z \subset \{z : \|z - z_c\| \leq \Omega\}. \tag{18}$$

Prox-mapping. Given a setup for SMP and a point $z \in Z^o$, we define the associated *prox-mapping* as

$$Pz(\xi) = \operatorname*{argmin}_{u \in Z} \{\omega(u) + \langle \xi - \omega'(z), u \rangle\} \equiv \operatorname*{argmin}_{u \in Z} \{V(z, u) + \langle \xi, u \rangle\} : \mathcal{E} \to Z^o.$$

Since $\mathcal{S}$ is compact and $\omega(\cdot)$ is continuous and strongly convex on ZThis mapping is clearly well defined.

3.2. Basic SMP setups

We illustrate the just-defined notions with three basic examples.

Example 1: Euclidean setup. Here $\mathcal{E}$ is $\mathbf{R}^N$ with the standard inner product, $\|\cdot\|_2$ is the standard Euclidean norm on $\mathbf{R}^N$ (so that $\|\cdot\|_* = \|\cdot\|$) and $\omega(z) = \frac{1}{2}z^T z$ (i.e., $Z^o = Z$). Assume for the sake of simplicity that $0 \in Z$. Then $z_c = 0$ and $\Omega = \max_{z \in Z} \|z\|_2^2$. The prox-function and the prox-mapping are given by $V(z, u) = \frac{1}{2}\|z - u\|_2^2$, $P_z(\xi) = \operatorname{argmin}_{u \in Z} \|(z - \xi) - u\|_2$.

Example 2: Simplex setup. Here $\mathcal{E}$ is $\mathbf{R}^N$, $N > 1$, with the standard inner product, $\|z\| = \|z\|_1 := \sum_{j=1}^N |z_j|$ (so that $\|\xi\|_* = \max_j |\xi_j|$), Z is a closed convex subset of the standard simplex

$$\mathcal{D}_N = \{z \in \mathbf{R}^N : z \geq 0, \sum\nolimits_{j=1}^N z_j = 1\}$$

containing its barycenter, and $\omega(z) = \sum_{j=1}^N z_j \ln z_j$ is the entropy. Then

$$Z^o = \{z \in Z : z > 0\} \text{ and } \omega'(z) = [1 + \ln z_1; ...; 1 + \ln z_N], \quad z \in Z^o.$$

It is easily seen (see, e.g., [4]) that here

$$z_c = [1/N; ...; 1/N], \quad \Omega \leq \sqrt{2\ln(N)}$$

(the latter inequality becomes equality when Z contains a vertex of $\mathcal{D}_N$). The prox-function is

$$V(z, u) = \sum\nolimits_{j=1}^N u_j \ln(u_j/z_j),$$

and the prox-mapping is easy to compute when $Z = \mathcal{D}_N$:

$$(P_z(\xi))_j = \left(\sum\nolimits_{i=1}^N z_i \exp\{-\xi_i\}\right)^{-1} z_j \exp\{-\xi_j\}.$$

Example 3: Spectahedron setup. This is the "matrix analogy" of the Simplex setup. Specifically, now $\mathcal{E}$ is the space of $N \times N$ block-diagonal symmetric matrices, $N > 1$, of a given block-diagonal structure equipped with the Frobenius inner product $\langle a, b\rangle_F = \mathrm{Tr}(ab)$ and the trace norm $|a|_1 = \sum_{i=1}^N |\lambda_i(a)|$, where $\lambda_1(a) \geq ... \geq \lambda_N(a)$ are the eigenvalues of a symmetric $N \times N$ matrix a; the conjugate norm $|a|_\infty$ is the usual spectral norm (the largest singular value) of a. Z is assumed to be a closed convex subset of the *spectahedron* $\mathcal{S} = \{z \in \mathcal{E} : z \succeq 0,\ \mathrm{Tr}(z) = 1\}$ containing the matrix $N^{-1}I_N$. The d.-g.f. is twice the matrix entropy

$$\omega(z) = 2\sum_{j=1}^N \lambda_j(z) \ln \lambda_j(z),$$

so that $Z^o = \{z \in Z : z \succ 0\}$ and $\omega'(z) = 2\ln(z) + 2I_N$. This setup, similarly to the Simplex one, results in $z_c = N^{-1}I_N$ and $\Omega \leq 2\sqrt{\ln N}$ [2]. When $Z = \mathcal{S}$, it is relatively easy to compute the prox-mapping (see [2, 8]); this task reduces to the singular value decomposition of a matrix from $\mathcal{E}$. It should be added that the matrices from $\mathcal{S}$ are exactly the matrices of the form

$$a = \mathcal{H}(b) \equiv (\mathrm{Tr}(\exp\{b\}))^{-1} \exp\{b\}$$

with $b \in \mathcal{E}$. Note also that when $Z = \mathcal{S}$, the prox-mapping becomes "linear in matrix logarithm": if $z = \mathcal{H}(a)$, then $P_z(\xi) = \mathcal{H}(a - \xi/2)$.

3.3. Algorithm: the construction

The t-step SMP algorithm is applied to the v.i. (2), works as follows:

Algorithm 1. *1. Initialization: Choose $r_0 \in Z^o$ and stepsizes $\gamma_\tau > 0$, $1 \leq \tau \leq t$.*

2. Step τ, $\tau = 1, 2, ..., t$: Given $r_{\tau-1} \in Z^o$, set

$$\left\{ \begin{array}{rcl} w_\tau &=& P_{r_{\tau-1}}(\gamma_\tau \widehat{F}(r_{\tau-1})), \\ r_\tau &=& P_{r_{\tau-1}}(\gamma_\tau \widehat{F}(w_\tau)) \end{array} \right. \tag{19}$$

When $\tau < t$, loop to step $t + 1$.

3. At step t, output

$$\widehat{z}_t = \left[\sum_{\tau=1}^t \gamma_\tau\right]^{-1} \sum_{\tau=1}^t \gamma_\tau w_\tau. \tag{20}$$

Here $\widehat{F}$ is the approximation of $F(\cdot)$ available to the algorithm, so that $\widehat{F}(z) \in \mathcal{E}$ is the output of the "black box" – the oracle – representing F, the input to the oracle being $z \in Z$. In what follows we assume that F is a bounded monotone operator represented by a *Stochastic Oracle*.

Stochastic Oracle (SO). At the ith call to the SO, the input being $z \in Z$, the oracle returns the vector $\widehat{F} = \Xi(z, \zeta_i)$, where $\{\zeta_i \in \mathbf{R}^N\}_{i=1}^{\infty}$ is a sequence of i.i.d. random variables, and $\Xi(z, \zeta) : Z \times \mathbf{R}^N \to \mathcal{E}$ is a Borel function satisfying the following

Assumption I: *With some $\mu \in [0, \infty)$, for all $z \in Z$ we have*

$$\begin{array}{ll} (a) & \|\mathbf{E}\{\Xi(z,\zeta_i) - F(z)\}\|_* \le \mu \\ (b) & \mathbf{E}\left\{\|\Xi(z,\zeta_i) - F(z)\|_*^2\right\} \le \sigma^2. \end{array} \tag{21}$$

which is slightly milder than (6). The associated version of Algorithm 1 will be referred to as *Stochastic Mirror Prox* (SMP) algorithm.

In some cases, we augment Assumption I by the following

Assumption II: *For all $z \in Z$ and all i we have*

$$\mathbf{E}\left\{\exp\{\|\Xi(z,\zeta_i) - F(z)\|_*^2/\sigma^2\}\right\} \le \exp\{1\}. \tag{22}$$

Note that Assumption II implies (21.b), since

$$\exp\{\mathbf{E}\left\{\|\Xi(z,\zeta_i) - F(z)\|_*^2/\sigma^2\right\}\} \le \mathbf{E}\left\{\exp\{\|\Xi(z,\zeta_i) - F(z)\|_*^2/\sigma^2\}\right\}$$

by the Jensen inequality.

3.4. Algorithm: Main result

From now on, assume that the starting point r_0 in Algorithm 1 is the minimizer z_c of $\omega(\cdot)$ on Z. Further, to avoid unnecessarily complicated formulas (and with no harm to the efficiency estimates) we stick to the constant stepsize policy $\gamma_\tau \equiv \gamma$, $1 \le \tau \le t$, where t is a fixed in advance number of iterations of the algorithm. Our main result is as follows:

Theorem 1. *Let v.i. (2) with monotone operator F satisfying (4) be solved by t-step Algorithm 1 using a SO, and let the stepsizes $\gamma_\tau \equiv \gamma$, $1 \le \tau \le t$, satisfy $0 < \gamma \le \frac{1}{\sqrt{3}L}$. Then*

(i) *Under Assumption I, one has*

$$\mathbf{E}\left\{\mathrm{Err}_{\mathrm{vi}}(\widehat{z}_t)\right\} \le K_0(t) \equiv \left[\frac{\Omega^2}{t\gamma} + \frac{7\gamma}{2}[M^2 + 2\sigma^2]\right] + 2\mu\Omega, \tag{23}$$

where M is the constant from (4) and Ω is given by (16).

(ii) *Under Assumptions I, II, one has, in addition to (23), for any $\Lambda > 0$,*

$$\mathrm{Prob}\left\{\mathrm{Err}_{\mathrm{vi}}(\widehat{z}_t) > K_0(t) + \Lambda K_1(t)\right\} \le \exp\{-\Lambda^2/3\} + \exp\{-\Lambda t\}, \tag{24}$$

where

$$K_1(t) = \frac{7\sigma^2\gamma}{2} + \frac{2\sigma\Omega}{\sqrt{t}}.$$

In the case of a Nash v.i., $\mathrm{Err}_{\mathrm{vi}}(\cdot)$ *in (23), (24) can be replaced with* $\mathrm{Err}_{\mathrm{N}}(\cdot)$.

When optimizing the bound (23) in γ, we get the following

Corollary 1. *In the situation of Theorem 1, let the stepsizes $\gamma_\tau \equiv \gamma$ be chosen according to*

$$\gamma = \min\left[\frac{1}{\sqrt{3}L}, \Omega\sqrt{\frac{2}{7t(M^2+2\sigma^2)}}\right]. \tag{25}$$

Then under Assumption I one has

$$\mathbf{E}\left\{\mathrm{Err}_{\mathrm{vi}}(\widehat{z}_t)\right\} \leq K_0^*(t) \equiv \max\left[\tfrac{7}{4}\tfrac{\Omega^2 L}{t},\ 7\Omega\sqrt{\tfrac{M^2+2\sigma^2}{3t}}\right] + 2\mu\Omega, \tag{26}$$

(see (16)). Under Assumptions I, II, one has, in addition to (26), for any $\Lambda > 0$,

$$\mathrm{Prob}\left\{\mathrm{Err}_{\mathrm{vi}}(\widehat{z}_t) > K_0^*(t) + \Lambda K_1^*(t)\right\} \leq \exp\{-\Lambda^2/3\} + \exp\{-\Lambda t\} \tag{27}$$

with

$$K_1^*(t) = \frac{7}{2}\frac{\Omega\sigma}{\sqrt{t}}. \tag{28}$$

In the case of a Nash v.i., $\mathrm{Err}_{\mathrm{vi}}(\cdot)$ in (26), (27) can be replaced with $\mathrm{Err}_{\mathrm{N}}(\cdot)$.

Remark 1. *Observe that the upper bound (26) for the error of Algorithm 1 with stepsize strategy (25), in agreement with the lower bound of [7], depends in the same way on the "size" σ of the perturbation $\Xi(z,\zeta_i) - F(z)$ and on the bound M for the non-Lipschitz component of F. From now on to simplify the presentation, with slight abuse of notation, we denote M the maximum of these quantities. Clearly, the latter implies that the bounds (21.b) and (22), and thus the bounds (26) – (28) of Corollary 1 hold with M substituted for σ.*

3.5. Comparison with Robust Mirror SA Algorithm

Consider the case of a Nash s.v.i. with operator F satisfying (4) with $L = 0$, and let the SO be unbiased (i.e., $\mu = 0$). In this case, the bound (26) reads

$$\mathbf{E}\left\{\mathrm{Err}_{\mathrm{N}}(\widehat{z}_t)\right\} \leq \frac{7\Omega M}{\sqrt{t}}, \tag{29}$$

where

$$M^2 = \max\left[\sup_{z,z'\in Z}\|F(z)-F(z')\|_*^2,\ \sup_{z\in Z}\mathbf{E}\left\{\|\Xi(z,\zeta_i)-F(z)\|_*^2\right\}\right]$$

The bound (29) looks very much like the efficiency estimate

$$\mathbf{E}\left\{\mathrm{Err}_{\mathrm{N}}(\tilde{z}_t)\right\} \leq O(1)\frac{\Omega\overline{M}}{\sqrt{t}} \tag{30}$$

(from now on, all $O(1)$'s are appropriate absolute positive constants) for the approximate solution $\tilde{z}_t$ of the t-step *Robust Mirror SA* (RMSA) algorithm [4][1]. In the latter estimate, Ω is exactly the same as in (29), and $\overline{M}$ is given by

$$\overline{M}^2 = \max\left[\sup_z \|F(z)\|_*^2;\ \sup_{z\in Z} \mathbf{E}\left\{\|\Xi(z,\zeta_i) - F(z)\|_*^2\right\}\right].$$

Note that we always have $M \leq 2\overline{M}$, and typically M and $\overline{M}$ are of the same order of magnitude; it may happen, however (think of the case when F is "almost constant"), that $M \ll \overline{M}$. Thus, the bound (29) never is worse, and sometimes can be much better than the SA bound (30). It should be added that as far as implementation is concerned, the SMP algorithm is not more complicated than the RMSA (cf. the description of Algorithm 1 with the description

$$\begin{aligned} r_t &= P_{r_{t-1}}(\widehat{F}(r_{t-1})),\\ \widehat{z}_t &= \left[\sum\nolimits_{\tau=1}^t \gamma_\tau\right]^{-1} \sum\nolimits_{\tau=1}^t \gamma_\tau r_\tau, \end{aligned}$$

of the RMSA).

The just outlined advantage of SMP as compared to the usual Stochastic Approximation is not that important, since "typically" M and $\overline{M}$ are of the same order. We believe that the most interesting feature of the SMP algorithm is its ability to take advantage of a specific structure of a stochastic optimization problem, namely, insensitivity to the presence in the objective of large, but smooth and well-observable components.

We are about to consider several less straightforward applications of the outlined insensitivity of the SMP algorithm to smooth well-observed components in the objective.

4. Application to Stochastic Composite minimization

Our present goal is to apply the SMP algorithm to the Composite minimization problem (9) in the case when the associated monotone operator (12) is given by a Stochastic Oracle.

Throughout this section, the structural assumptions **A** – **C** *from Section 2.2 are in force.*

4.1. Assumptions

We start with augmenting the description of the problem of interest (9), see Section 2.2, with additional assumptions specifying the SO and the SMP setup for the v.i. reformulation

$$\text{find } z_* \in Z := X \times Y : \langle F(z), z - z_*\rangle \geq 0\ \forall z \in Z \tag{31}$$

of (9); here F is the monotone operator (12). Specifically, we assume that

[1] In this reference, only the Minimization and the Saddle Point problems are considered. However, the results of [4] can be easily extended to s.v.i.'s.

D. The embedding space $\mathcal{X}$ of X is equipped with a norm $\|\cdot\|_x$, and X itself – with a d.-g.f. $\omega_x(x)$, the associated parameter (16.c) being some Ω_x;

E. The spaces $\mathcal{E}_\ell$, $1 \leq \ell \leq m$, where the functions ϕ_ℓ take their values, are equipped with norms (not necessarily the Euclidean ones) $\|\cdot\|_{(\ell)}$ with conjugates $\|\cdot\|_{(\ell,*)}$ such that

$$\forall v, v' \in X : \left\{ \begin{array}{ll} (a) & \|[\phi'_\ell(v) - \phi'_\ell(v')]h\|_{(\ell)} \leq [L_x\|v - v'\|_x + M_x]\|h\|_x \\ (b) & \|[\phi'_\ell(v)]h\|_{(\ell)} \leq \Omega_x L_x \|h\|_x \end{array} \right. \tag{32}$$

for certain selections $\phi'_\ell(v) \in \partial^{K_\ell}\phi_\ell(v)$, $v \in X$[2)] and certain nonnegative constants L_x, M_x.

F. Functions $\phi_\ell(\cdot)$ are represented by an unbiased SO. At the ith call to the oracle, $x \in X$ being the input, the oracle returns vectors $f_\ell(x, \zeta_i) \in \mathcal{E}_\ell$ and linear mappings $\mathbf{G}_\ell(x, \zeta_i)$ from $\mathcal{X}$ to $\mathcal{E}_\ell$, $1 \leq \ell \leq m$ ($\{\zeta_i\}$ are i.i.d. random "oracle noises") such that for any $x \in X$ and $i = 1, 2, \ldots,$

$$\begin{array}{ll} (a) & \mathbf{E}\left\{f_\ell(x,\zeta_i)\right\} = \phi_\ell(x),\ 1 \leq \ell \leq m \\ (b) & \mathbf{E}\left\{\max\limits_{1\leq \ell\leq m} \|f_\ell(x,\zeta_i) - \phi_\ell(x)\|_{(\ell)}^2\right\} \leq M_x^2\Omega_x^2; \\ (c) & \mathbf{E}\left\{\mathbf{G}_\ell(x,\zeta_i)\right\} = \phi'_\ell(x),\ 1 \leq \ell \leq m, \\ (d) & \mathbf{E}\left\{\max\limits_{\substack{h\in\mathcal{X} \\ \|h\|_x\leq 1}} \|[\mathbf{G}_\ell(x,\zeta_i) - \phi'_\ell(x)]h\|_{(\ell)}^2\right\} \leq M_x^2,\ 1 \leq \ell \leq m. \end{array} \tag{33}$$

G. The data participating in the Fenchel-type representation (10) of Φ are such that

(a) the embedding space $\mathcal{Y}$ of Y is equipped with a norm $\|\cdot\|_y$, and Y itself - — with a d.-g.f. $\omega_y(y)$, the associated parameter (16.c) being some Ω_y;

(b) we have $\|y\|_y \leq 2\Omega_y$ for all $y \in Y$;[3]

(c) The convex function $\Phi_*(y)$ is given by the precise deterministic first order oracle, and

$$\|\Phi'_*(y) - \Phi'_*(y')\|_{y,*} \leq L_y\|y - y'\|_y + M_y \tag{34}$$

for certain selection $\Phi'_*(y) \in \partial\Phi_*(y)$, $y \in Y$, and some nonnegative L_y, M_y.

---

2) For a K-convex function $\phi : X \to \mathcal{E}$ ($X \subset \mathcal{X}$ is convex, $K \subset \mathcal{E}$ is a closed convex cone) and $x \in X$, the K-*subdifferential* $\partial^K\phi(x)$ is comprised of all linear mappings $h \mapsto \mathbf{P}h : \mathcal{X} \to \mathcal{E}$ such that $\phi(u) \geq_K \phi(x) + \mathbf{P}(u - x)$ for all $u \in X$. When ϕ is Lipschitz continuous on X, $\partial^K\phi(x) \neq \emptyset$ for all $x \in X$; if ϕ is differentiable at $x \in \operatorname{int} X$ (as it is the case almost everywhere on $\operatorname{int} X$), one has $\frac{\partial\phi(x)}{\partial x} \in \partial^K\phi(x)$.

3 This requirement can be ensured by shifting Y to include the origin and the associated shift in $\omega_y(\cdot)$, see (18).

Stochastic Oracle for (31). The assumptions $\mathbf{E}-\mathbf{G}$ induce an unbiased SO for the operator F in (31), specifically, the oracle

$$\Xi(x,y,\zeta_i)=\left[\sum_{\ell=1}^m \mathbf{G}_\ell^*(x,\zeta_i)[\mathbf{A}_\ell y+b_\ell];\ -\sum_{\ell=1}^m \mathbf{A}_\ell^* f_\ell(x,\zeta_i)+\Phi_*'(y)\right], \tag{35}$$

and this is the oracle we will use when solving (9) by the SMP algorithm.

4.2. Setup for the SMP as applied to (31), (12)

In retrospect, the setup for SMP we are about to present is kind of the best – resulting in the best possible efficiency estimate (26) – we can build from the entities participating in the description of the problem (9) as given by the assumptions $\mathbf{A}-\mathbf{G}$. Specifically, we equip the space $\mathcal{E}=\mathcal{X}\times\mathcal{Y}$ with the norm

$$\|(x,y)\|\equiv\sqrt{\|x\|_x^2/\Omega_x^2+\|y\|_y^2/\Omega_y^2}\quad\left[\Rightarrow \|(\xi,\eta)\|_*=\sqrt{\Omega_x^2\|\xi\|_{x,*}^2+\Omega_y^2\|\eta\|_{y,*}^2}\right]$$

and equip $Z=X\times Y$ with the d.-g.f.

$$\omega(x,y)=\frac{1}{\Omega_x^2}\omega_x(x)+\frac{1}{\Omega_y^2}\omega_y(y)$$

(it is immediately seen that $\omega(\cdot)$ indeed is a d.-g.f. w.r.t. X, $\|\cdot\|$). The SMP-related properties of our setup are summarized in the following

Lemma 1. *Let*

$$\mathcal{A}=\max_{y\in\mathcal{Y}:\|y\|_y\le 1}\sum\nolimits_{\ell=1}^m\|\mathbf{A}_\ell y\|_{(\ell,*)},\ \mathcal{B}=\sum\nolimits_{\ell=1}^m\|b_\ell\|_{(\ell,*)}. \tag{36}$$

(i) *The parameter Ω associated with $\omega(\cdot)$, $\|\cdot\|$, Z by (16), is $\le\sqrt{2}$.*
(ii) *One has*

$$\forall(z,z'\in Z):\|F(z)-F(z')\|_*\le L\|z-z'\|+M, \tag{37}$$

where

$$\begin{aligned} L &= 4\mathcal{A}\Omega_x^2\Omega_yL_x+\Omega_x\mathcal{B}+\Omega_y^2L_y,\\ M &= [3\mathcal{A}\Omega_y+\mathcal{B}]\,\Omega_xM_x+\Omega_yM_y.\end{aligned}$$

Besides this,

$$\forall(z\in Z,i):\mathbf{E}\{\Xi(z,\zeta_i)\}=F(z);\quad \mathbf{E}\left\{\|\Xi(z,\zeta_i)-F(z)\|_*^2\right\}\le M^2. \tag{38}$$

Finally, when relations (33.b,d) *are strengthened to*

$$\begin{aligned}&\mathbf{E}\left\{\exp\left\{\max_{1\le\ell\le m}\|f_\ell(x,\zeta_i)-\phi_\ell(x)\|_{(\ell)}^2/(\Omega_xM)^2\right\}\right\}\le\exp\{1\},\\ &\mathbf{E}\left\{\exp\left\{\max_{\substack{h\in\mathcal{X},\\ \|h\|_x\le1}}\|[\mathbf{G}_\ell(x)-\phi_\ell'(x)]h\|_{(\ell)}^2/M^2\right\}\right\}\le\exp\{1\},\ 1\le\ell\le m,\end{aligned} \tag{39}$$

then

$$\mathbf{E}\left\{\exp\{\|\Xi(z,\zeta_i)-F(z)\|_*^2/M^2\}\right\}\leq\exp\{1\}. \tag{40}$$

Combining Lemma 1 with Corollary 1, we get explicit efficiency estimates for the SMP algorithm as applied to the Stochastic composite minimization problem (9); these are nothing than estimates (26) with σ, Ω replaced with M, $\sqrt{2}$, respectively.

4.3. Application to Matrix Minimax problem

Consider Matrix Minimax problem from Section 2.2.2, that is, the problem

$$\min_{x\in X}\max_{1\leq j\leq k}\lambda_{\max}\left(\sum\nolimits_{\ell=1}^m P_{j\ell}^T\phi_\ell(x)P_{j\ell}\right), \tag{41}$$

where $\phi_\ell(\cdot):X\to\mathbf{S}^{p_\ell}$ are $\succeq$-convex Lipschitz continuous mappings and $P_{j\ell}\in\mathbf{R}^{p_\ell\times q_j}$. As it was shown in Section 2.2.2, (41) admits saddle point representation

$$\begin{array}{c}\min\limits_{x\in X}\max\limits_{y=\mathrm{Diag}\{y_1,\ldots,y_k\}\in Y}\sum_{\ell=1}^m\langle\phi_\ell(x),\mathbf{A}_\ell y\rangle_F,\\ \mathbf{A}_\ell y=\sum_{j=1}^k P_{j\ell}y_jP_{j\ell}^T,\end{array} \tag{42}$$

where Y is the spectahedron in the space $\mathcal{Y}$ of block-diagonal symmetric matrices $y=\mathrm{Diag}\{y_1,...,y_k\}$ with k diagonal blocks of sizes $q_1,...,q_k$. Note that we are in the situation described by assumptions $\mathbf{A}-\mathbf{C}$ from Section 2.2, with $\Phi_*(\cdot)\equiv 0$, and $K_\ell:=\mathbf{S}^{p_\ell}_+\subset\mathcal{E}_\ell:=\mathbf{S}^{p_\ell}$. Using the spectahedron setup for Y and setting $M_y=L_y=0$, $\Omega_y=2\sqrt{\ln(q_1+...+q_k)}$, we meet all assumptions in $\mathbf{G}$. Let us specify the norms $\|\cdot\|_{(\ell)}$ on the spaces $\mathcal{E}_\ell=\mathbf{S}^{p_\ell}$ as the standard matrix norms $|\cdot|_\infty$ (maximal singular value), and let assumptions $\mathbf{D}-\mathbf{F}$ from Section 4.1 take place. Note that in the case in question (36) reads

$$\mathcal{A}=\max_{1\leq j\leq k}\max_{\xi\in\mathbf{R}^{q_j}:\|\xi\|_2=1}\sum_{\ell=1}^m\|P_{j\ell}\xi\|_2^2=\max_{1\leq j\leq k}|\sum_{\ell=1}^m P_{j\ell}^TP_{j\ell}|_\infty,\quad\mathcal{B}=0.$$

(look what are the extreme points of Y), so that the quantities L, M as given by Lemma 1 become

$$\begin{array}{rcl}L&=&O(1)\mathcal{A}\sqrt{\ln(q_1+...+q_k)}\Omega_x^2L_x,\\ M&=&O(1)\mathcal{A}\sqrt{\ln(q_1+...+q_k)}\Omega_xM_x.\end{array} \tag{43}$$

Note that in the case of problem (13) one has $\mathcal{A}=1$.

4.3.1. Application to Stochastic Semidefinite Feasibility problem

Now consider the stochastic version of the Semidefinite Feasibility problem (S):

$$\text{find } x\in X:\psi_\ell(x)\preceq 0,\ 1\leq\ell\leq m. \tag{44}$$

Assuming form now on that the latter problem is feasible, we can rewrite it as the Matrix Minimax problem

$$\mathrm{Opt} = \min_{x\in X}\max_{1\le\ell\le m}\lambda_{\max}(\phi_\ell(x)), \quad \phi_\ell(x) = \beta_\ell\psi_\ell(x), \tag{45}$$

where $\beta_\ell > 0$ are "scale factors" we are free to choose. We are about to show how to use this freedom in order to improve the SMP efficiency estimates.

Assuming, same as in the case of a general-type Matrix Minimax problem, that **D** takes place, let us modify assumptions **E**, **F** as follows:

E′: the $\succeq$-convex Lipschitz functions $\psi_\ell : X \to \mathbf{S}^{p_\ell}$ are such that

$$\begin{array}{l}\max\limits_{h\in\mathcal{X},\,\|h\|_x\le1}|[\psi'_\ell(x)-\psi'_\ell(x')]h|_\infty \le L_\ell\|x-x'\|_x + M_\ell,\\ \max\limits_{h\in\mathcal{X},\,\|h\|_x\le1}|\psi'_\ell(x)h|_\infty \le \Omega_x L_\ell\end{array} \tag{46}$$

for certain selections $\psi'_\ell(x) \in \partial^{K_\ell}\psi_\ell(x)$, $x \in X$, with some known nonnegative constants L_ℓ, M_ℓ.

F′: $\psi_\ell(\cdot)$ are represented by an SO which at the ith call, the input being $x \in X$, returns the matrices $\widehat{f}_\ell(x,\zeta_i) \in \mathbf{S}^{p_\ell}$ and the linear maps $\widehat{\mathbf{G}}_\ell(x,\zeta_i)$ from $\mathcal{X}$ to $\mathbf{S}^{p_\ell}$ ($\{\zeta_i\}$ are i.i.d. random "oracle noises") such that for any $x \in X$ it holds

$$\begin{array}{ll}(a) & \mathbf{E}\left\{\widehat{f}_\ell(x,\zeta_i)\right\} = \psi_\ell(x),\ \ \mathbf{E}\left\{\widehat{\mathbf{G}}_\ell(x,\zeta_i)\right\} = \psi'_\ell(x),\ 1\le\ell\le m\\ (b) & \mathbf{E}\left\{\max\limits_{1\le\ell\le m}|\widehat{f}_\ell(x,\zeta_i)-\psi_\ell(x)|_\infty^2/(\Omega_x M_\ell)^2\right\}\le 1\\ (c) & \mathbf{E}\left\{\max\limits_{\substack{h\in\mathcal{X},\\ \|h\|_x\le1}}|[\widehat{\mathbf{G}}_\ell(x,\zeta_i)-\psi'_\ell(x)]h|_\infty^2/M_\ell^2\right\}\le 1,\ 1\le\ell\le m.\end{array} \tag{47}$$

Given a number t of steps of the SMP algorithm, let us act as follows.

(I): We compute the m quantities $\mu_\ell = \frac{\Omega_x L_\ell}{\sqrt{t}} + M_\ell$, $\ell = 1,...,m$, and set

$$\mu = \max_{1\le\ell\le m}\mu_\ell,\ \ \beta_\ell = \frac{\mu}{\mu_\ell}, \phi_\ell(\cdot) = \beta_\ell\psi_\ell(\cdot),\ \ L_x = \Omega_x^{-1}\mu\sqrt{t},\ \ M_x = \mu. \tag{48}$$

Note that by construction $\beta_\ell \ge 1$ and $L_x/L_\ell \ge \beta_\ell$, $M_x/M_\ell \ge \beta_\ell$ for all ℓ, so that the functions ϕ_ℓ satisfy (32) with the just defined L_x, M_x. Further, the SO for $\psi_\ell(\cdot)$'s can be converted into an SO for $\phi_\ell(\cdot)$'s by setting

$$f_\ell(x,\zeta) = \beta_\ell\widehat{f}_\ell(x,\zeta),\ \ \mathbf{G}_\ell(x,\zeta) = \beta_\ell\widehat{\mathbf{G}}_\ell(x,\zeta).$$

By (47) and due to $L_x/L_\ell \ge \beta_\ell$, $M_x/M_\ell \ge \beta_\ell$, this oracle satisfies (33).

(II) We then build the Stochastic Matrix Minimax problem

$$\mathrm{Opt} = \min_{x\in X}\max_{1\le\ell\le m}\lambda_{\max}(\phi_\ell(x)), \tag{49}$$

associated with the just defined $\phi_1,...,\phi_m$ and solve this Stochastic composite problem by t-step SMP algorithm. Combining Lemma 1, Corollary 1 and taking into account the origin of the quantities L_x, M_x, and the fact that $\mathcal{A} = 1$, $\mathcal{B} = 0$, we arrive at the following result:

Proposition 1. *With the outlined construction, the t-step SMP algorithm with the setup presented in Section 4.2 (where one uses $\mathcal{A}=1,\mathcal{B}=L_y=M_y=0$ and the just defined L_x, M_x) and constant stepsizes $\gamma_\tau\equiv\gamma$ defined by (25), yields an approximate solution $\widehat{z}_t=(\widehat{x}_t,\widehat{y}_t)$ such that*

$$\begin{array}{rcl}\mathbf{E}\left\{\max\limits_{1\le\ell\le m}\max[\beta_\ell\lambda_{\max}(\psi_\ell(\widehat{x}_t),0]\right\} & \le & \mathbf{E}\left\{\max\limits_{1\le\ell\le m}\beta_\ell\lambda_{\max}(\psi_\ell(\widehat{x}_t))-\mathrm{Opt}\right\}\\ & \le & K_0(t)\equiv 80\frac{\Omega_x\mu\sqrt{\ln(\sum_{\ell=1}^m p_\ell)}}{\sqrt{t}},\end{array}\tag{50}$$

(cf. (26) and take into account that we are in the case of $\Omega=\sqrt{2}$, while the optimal value in (49) is nonpositive, since (44) is feasible).

Furthermore, if assumptions (47.b,c) are strengthened to

$$\begin{array}{rcl}\mathbf{E}\left\{\max\limits_{1\le\ell\le m}\exp\{|\widehat{f}_\ell(x,\zeta_i)-\psi_\ell(x)|_\infty^2/(\Omega_xM_\ell)^2\}\right\} & \le & \exp\{1\},\\ \mathbf{E}\left\{\exp\{\max\limits_{h\in\mathcal{X},\,\|h\|_x\le1}|[\widehat{\mathbf{G}}_\ell(x,\zeta_i)-\psi'_\ell(x)]h|_\infty^2/M_\ell^2\}\right\} & \le & \exp\{1\},\ \ 1\le\ell\le m,\end{array}$$

then, in addition to (50), we have for any $\Lambda>0$:

$$\begin{array}{l}\mathrm{Prob}\left\{\max\limits_{1\le\ell\le m}\max[\beta_\ell\lambda_{\max}(\psi_\ell(\widehat{x}_t)),0]>K_0(t)+\Lambda K_1(t)\right\}\\ \qquad\qquad\le\exp\{-\Lambda^2/3\}+\exp\{-\Lambda t\},\end{array}$$

where

$$K_1(t)=\frac{15\Omega_x\mu\sqrt{\ln(\sum_{\ell=1}^m p_\ell)}}{\sqrt{t}}.$$

Discussion Imagine that instead of solving the system of matrix inequalities (44), we were interested to solve just a single matrix inequality $\psi_\ell(x)\preceq 0$, $x\in X$. When solving this inequality by the SMP algorithm as explained above, the efficiency estimate would be

$$\begin{array}{rcl}\mathbf{E}\left\{\max[\lambda_{\max}(\psi_\ell(\widehat{x}_t^\ell)),0]\right\} & \le & O(1)\sqrt{\ln(p_\ell+1)}\Omega_x\left[\frac{\Omega_xL_\ell}{t}+\frac{M_\ell}{\sqrt{t}}\right]\\ & = & O(1)\sqrt{\ln(p_\ell+1)}\beta_\ell^{-1}\frac{\Omega_x\mu}{\sqrt{t}},\end{array}$$

(recall that the matrix inequality in question is feasible), where $\widehat{x}_t^\ell$ is the resulting approximate solution. Looking at (50), we see that the expected accuracy of the SMP as applied, in the aforementioned manner, to (44) is only by a logarithmic in $\sum_\ell p_\ell$ factor worse:

$$\begin{array}{rcl}\mathbf{E}\left\{\max[\lambda_{\max}(\psi_\ell(\widehat{x}_t),0]\right\} & \le & O(1)\sqrt{\ln(\sum_{\ell=1}^m p_\ell)}\beta_\ell^{-1}\frac{\Omega_x\mu}{\sqrt{t}}\\ & = & O(1)\sqrt{\ln(\sum_{\ell=1}^m p_\ell)}\frac{\Omega_x\mu_\ell}{\sqrt{t}}.\end{array}\tag{51}$$

Thus, as far as the quality of the SPM-generated solution is concerned, passing from solving a single matrix inequality to solving a system of m inequalities is "nearly costless". As an illustration, consider the case where some of ψ_ℓ are "easy" – smooth and easy-to-observe ($M_\ell = 0$), while the remaining ψ_ℓ are "difficult", i.e., might be non-smooth and/or difficult-to-observe ($\Omega_x L_\ell/\sqrt{t} \leq M_\ell$). In this case, (51) reads

$$\mathbf{E}\left\{\psi_\ell(\widehat{x}_t)\right\} \leq O(1)\sqrt{\ln(\sum\nolimits_{\ell=1}^m p_\ell)} \cdot \left\{ \begin{array}{ll} \frac{\Omega_x^2 L_\ell}{t}, & \psi_\ell \text{ is easy,} \\ \frac{\Omega_x M_\ell}{\sqrt{t}}, & \psi_\ell \text{ is difficult.} \end{array}\right.$$

In other words, the violations of the easy and the difficult constraints in (44) converge to 0 as $t \to \infty$ with the rates $O(1/t)$ and $O(1/\sqrt{t})$, respectively. It should be added that when X is the unit Euclidean ball in $\mathcal{X} = \mathbf{R}^n$ and X, $\mathcal{X}$ are equipped with the Euclidean setup, the rates of convergence $O(1/t)$ and $O(1/\sqrt{t})$ are the best rates one can achieve without imposing bounds on n and/or imposing additional restrictions on ψ_ℓ's.

4.4. Eigenvalue optimization via SMP

The problem we are interested in now is

$$\begin{array}{rcl} \mathrm{Opt} & = & \min\limits_{x\in X} f(x) := \lambda_{\max}(A_0 + x_1A_1 + ... + x_nA_n), \\ X & = & \{x \in \mathbf{R}^n : x \geq 0, \sum\nolimits_{i=1}^n x_i = 1\}, \end{array} \tag{52}$$

where $A_0, A_1, ..., A_n$, $n > 1$, belong to the space $\mathbf{S}$ of symmetric matrices with block-diagonal structure $(p_1, ..., p_m)$ (i.e., a matrix $A \in \mathbf{S}$ is block-diagonal with $p_\ell \times p_\ell$ diagonal blocks A^ℓ, $1 \leq \ell \leq m$). We set

$$p^{(\kappa)} = \sum\nolimits_{\ell=1}^m p_\ell^\kappa, \;\; \kappa = 1, 2, ...; \;\; p^{\max} = \max_\ell p_\ell; \;\; A_\infty = \max_{1\leq j\leq n} |A_j|_\infty.$$

Setting

$$\phi_\ell : X \mapsto \mathcal{E}_\ell = \mathbf{S}^{p_\ell}, \;\; \phi_\ell(x) = A_0^\ell + \sum\nolimits_{j=1}^n x_j A_j^\ell, \;\; 1 \leq \ell \leq m,$$

we represent (52) as a particular case of the Matrix Minimax problem (13), with all functions $\phi_\ell(x)$ being affine and X being the standard simplex in $\mathcal{X} = \mathbf{R}^n$.

Now, since A_j are known in advance, there is nothing stochastic in our problem, and it can be solved either by interior point methods, or by "computationally cheap" gradient-type methods which are preferable when the problem is large-scale and medium accuracy solutions are sought. For instance, one can apply the t-step (deterministic) Mirror Prox algorithm DMP from [8] to the saddle point reformulation (11) of our specific Matrix Minimax problem, i.e., to the saddle point problem

$$\begin{array}{c} \min\limits_{x\in X}\max\limits_{y\in Y} \langle y, A_0 + \sum\nolimits_{j=1}^n x_jA_j\rangle_F, \\ Y = \left\{y = \mathrm{Diag}\{y_1, ..., y_m\} : y_\ell \in \mathbf{S}_+^{p_\ell}, \, 1 \leq \ell \leq m, \, \mathrm{Tr}(y) = 1\right\}. \end{array} \tag{53}$$

The accuracy of the approximate solution $\tilde{x}_t$ of the DMP algorithm is [8, Example 2]

$$f(\tilde{x}_t) - \mathrm{Opt} \leq O(1)\frac{\sqrt{\ln(n)\ln(p^{(1)})}A_\infty}{t}. \tag{54}$$

This efficiency estimate is the best known so far among those attainable with "computationally cheap" deterministic methods. On the other hand, the complexity of one step of the algorithm is dominated, up to an absolute constant factor, by the necessity, given $x \in X$ and $y \in Y$,

1. to compute $A_0 + \sum_{j=1}^n x_j A_j$ and $[\mathrm{Tr}(YA_1); ...; \mathrm{Tr}(YA_n)]$;
2. to compute the eigenvalue decomposition of an $y \in \mathbf{S}$.

When using the standard Linear Algebra, the computational effort per step is

$$\mathcal{C}_{\det} = O(1)[np^{(2)} + p^{(3)}] \tag{55}$$

arithmetic operations.

We are about to demonstrate that one can equip the deterministic problem in question by an "artificial" SO in such a way that the associated SMP algorithm, under certain circumstances, exhibits better performance than deterministic algorithms. Let us consider the following construction of the SO for F (different from the SO (35)!). Observe that the monotone operator associated with the saddle point problem (53) is

$$F(x,y) = \Big[\underbrace{[\mathrm{Tr}(yA_1); ...; \mathrm{Tr}(yA_n)]}_{F^x(x,y)};\ \underbrace{-A_0 - \sum\nolimits_{j=1}^n x_j A_j}_{F^y(x,y)}\Big]. \tag{56}$$

Given $x \in X$, $y = \mathrm{Diag}\{y_1, ..., y_m\} \in Y$, we build a random estimate $\Xi = [\Xi^x; \Xi^y]$ of $F(x,y) = [F^x(x,y); F^y(x,y)]$ as follows:

1. we generate a realization $\jmath$ of a random variable taking values $1, ..., n$ with probabilities $x_1, ..., x_n$ (recall that $x \in X$, the standard simplex, so that x indeed can be seen as a probability distribution), and set
$$\Xi^y = A_0 + A_\jmath; \tag{57}$$
2. we compute the quantities $\nu_\ell = \mathrm{Tr}(y_\ell)$, $1 \leq \ell \leq m$. Since $y \in Y$, we have $\nu_\ell \geq 0$ and $\sum_{\ell=1}^m \nu_\ell = 1$. We further generate a realization $\imath$ of random variable taking values $1, ..., m$ with probabilities $\nu_1, ..., \nu_m$, and set
$$\Xi^x = [\mathrm{Tr}(A_1^\imath \bar{y}_\imath); ...; \mathrm{Tr}(A_n^\imath \bar{y}_\imath)],\ \bar{y}_\imath = (\mathrm{Tr}(y_\imath))^{-1} y_\imath. \tag{58}$$

The just defined random estimate Ξ of $F(x,y)$ can be expressed as a deterministic function $\Xi(x,y,\eta)$ of (x,y) and random variable η uniformly distributed on $[0,1]$. Assuming all matrices A_j directly available (so that it takes $O(1)$ arithmetic operations to extract a particular entry of A_j given j and indexes of the entry) and given x, y and η, the value $\Xi(x,y,\xi)$ can be computed with the arithmetic cost $O(1)(n(p^{\max})^2+p^{(2)})$ (indeed, $O(1)(n+p^{(1)})$ operations are needed to convert η into $\imath$ and $\jmath$, $O(1)p^{(2)}$ operations are used to write down the y-component $-A_0 - A_\jmath$ of Ξ, and $O(1)n(p^{\max})^2$ operations are needed to compute Ξ^x). Now consider the SO's Ξ_k (k is a positive integer) obtained by averaging the outputs of k calls to our basic oracle Ξ. Specifically, at the ith call to the oracle Ξ_k, $z=(x,y)\in Z = X\times Y$ being the input, the oracle returns the vector

$$\Xi_k(z,\zeta_i) = \frac{1}{k}\sum\nolimits_{s=1}^{k}\Xi(z,\eta_{is}),$$

where $\zeta_i = [\eta_{i1};...;\eta_{ik}]$ and $\{\eta_{is}\}_{1\leq i,\,1\leq s\leq k}$ are independent random variables uniformly distributed on $[0,1]$. Note that the arithmetic cost of a single call to Ξ_k is

$$\mathcal{C}_k = O(1)k(n(p^{\max})^2+p^{(2)}).$$

The Nash v.i. associated with (53) and the stochastic oracle Ξ_k (k is the first parameter of our construction) specify a Nash s.v.i. on the domain $Z=X\times Y$. We equip the standard simplex X and its embedding space $\mathcal{X}=\mathbf{R}^n$ with the Simplex setup, and the spectahedron Y and its embedding space $\mathbf{S}$ with the Spectahedron setup (see Section 3.2). Let us next combine the x- and the y-setups, exactly as explained in the beginning of Section 4.2, into an SMP setup for the domain $Z=X\times Y$ – a d.-g.f. $\omega(\cdot)$ and a norm $\|\cdot\|$ on the embedding space $\mathbf{R}^n\times(\mathbf{S}^{p_1}\times...\times\mathbf{S}^{p_\ell})$ of Z. The SMP-related properties of the resulting setup are summarized in the following statement.

Lemma 2. *Let $n\geq 3$, $p^{(1)}\geq 3$. Then*

(i) *The parameter of the just defined d.-g.f. ω w.r.t. the just defined norm $\|\cdot\|$ is $\Omega=\sqrt{2}$.*

(ii) *For any $z,z'\in Z$ one has*

$$\|F(z)-F(z')\|_*\leq L\|z-z'\|,\quad L=\sqrt{2}\left[\ln(n)+\ln(p^{(1)})\right]A_\infty. \tag{59}$$

Besides this, for any $(z\in Z,\ i=1,2,...,$

$$\begin{array}{ll}(a) & \mathbf{E}\left\{\Xi_k(z,\zeta_i)\right\}=F(z);\\ (b) & \mathbf{E}\left\{\exp\{\|\Xi(z,\zeta_i)-F(z)\|_*^2/M^2\}\right\}\leq\exp\{1\},\\ & M=27[\ln(n)+\ln(p^{(1)})]A_\infty/\sqrt{k}.\end{array} \tag{60}$$

Combining Lemma 2 and Corollary 1, we arrive at the following

Proposition 2. *With properly chosen positive absolute constants $O(1)$, the t-step SMP algorithm with constant stepsizes*

$$\gamma_\tau = O(1)\frac{\min[1,\sqrt{k/t}]}{\ln(np^{(1)})A_\infty},\ 1\leq\tau\leq t \qquad [A_\infty=\max_{1\leq j\leq n}|A_j|_\infty]$$

as applied to the saddle point reformulation of problem (52), the stochastic oracle being Ξ_k, produces a random feasible approximate solution $\widehat{x}_t$ to the problem with the error

$$\epsilon(\widehat{x}_t)=\lambda_{\max}\left(A_0+\sum\nolimits_{j=1}^{n}[\widehat{x}_t]_jA_j\right)-\mathrm{Opt}$$

satisfying

$$\mathbf{E}\left\{\epsilon(\widehat{x}_t)\right\} \le O(1)\ln(np^{(1)})A_\infty \left[\frac{1}{t} + \frac{1}{\sqrt{kt}}\right], \tag{61}$$

and for any $\Lambda > 0$*:*

$$\mathrm{Prob}\left\{\epsilon(\widehat{x}_t) > O(1)\ln(np^{(1)})A_\infty \left[\frac{1}{t} + \frac{1+\Lambda}{\sqrt{kt}}\right]\right\} \le \exp\{-\Lambda^2/3\} + \exp\{-\Lambda t\}.$$

Further, assuming that all matrices A_j *are directly available, the overall computational effort to compute* $\widehat{x}_t$ *is*

$$\mathcal{C} = O(1)t\left[k(n(p^{\max})^2 + p^{(2)}) + p^{(3)}\right] \tag{62}$$

arithmetic operations.

To justify the bound (62) it suffices to note that $O(1)k(n(p^{\max})^2 + p^{(2)})$ operations per step is the price of two calls to the stochastic oracle Ξ_k and $O(1)(n+p^{(3)})$ operations per step is the price of computing two prox mappings.

Discussion Let us find out whether randomization can help when solving a large-scale problem (52), that is, whether, given quality of the resulting approximate solution, the computational effort to build such a solution with the Stochastic Mirror Prox algorithm SMP can be essentially less than the one for the deterministic Mirror Prox algorithm DMP. To simplify our considerations, assume from now on that $p_\ell = p$, $1 \le \ell \le m$, and that $\ln(n) = O(1)\ln(mp)$. Assume also that we are interested in a (perhaps, random) solution $\widehat{x}_t$ which with probability $\ge 1-\delta$ satisfies $\epsilon(\widehat{x}_t) \le \epsilon$. We fix a tolerance $\delta \ll 1$ and the *relative accuracy* $\nu = \frac{\epsilon}{\ln(mnp)A_\infty} \le 1$ and look what happens when (some of) the sizes m, n, p of the problem become large.

Observe first of all that the overall computational effort to solve (52) within relative accuracy ν with the DMP algorithm is

$$\mathcal{C}^{\mathrm{DMP}}(\nu) = O(1)m(n+p)p^2\nu^{-1}$$

operations (see (54), (55)). As for the SMP algorithm, let us choose k which balances the per step computational effort $O(1)k(n(p^{\max})^2 + p^{(2)}) = O(1)k(n+m)p^2$ to produce the answers of the stochastic oracle and the per step cost of prox mappings $O(1)(n+mp^3)$, that is, let us set $k = \mathrm{Ceil}\left(\frac{mp}{m+n}\right)$ [4]. With this choice of k, Proposition 2 says that to get a solution of the required quality, it suffices to carry out

$$t = O(1)\left[\nu^{-1} + \ln(1/\delta)k^{-1}\nu^{-2}\right] \tag{63}$$

steps of the method, provided that this number of steps is $\ge \sqrt{\ln(2/\delta)}$. The latter assumption is automatically satisfied when the absolute constant factor

[4]The rationale behind balancing is clear: with the just defined k, the arithmetic cost of an iteration still is of the same order as when $k=1$, while the left hand side in the efficiency estimate (61) becomes better than for $k=1$.

in (63) is ≥ 1 and $\nu\sqrt{\ln(2/\delta)} \leq 1$, which we assume from now on. Combining (63) and the upper bounds on the arithmetic cost of an SMP step stated in Proposition 2, we conclude that the overall computational effort to produce a solution of the required quality with the SMP algorithm is

$$\mathcal{C}^{\rm SMP}(\nu,\delta) = O(1)k(n+m)p^2\left[\nu^{-1} + \ln(1/\delta)k^{-1}\nu^{-2}\right]$$

operations, so that

$$R := \frac{\mathcal{C}^{\rm DMP}(\nu)}{\mathcal{C}^{\rm SMP}(\nu,\delta)} = O(1)\frac{m(n+p)}{(m+n)(k+\ln(1/\delta)\nu^{-1})} \qquad \left[k = \mathrm{Ceil}\left(\tfrac{mp}{m+n}\right)\right]$$

We see that *when ν, δ are fixed, $m \geq n/p$ and n/p is large, then R is large as well*, that is, the randomized algorithm significantly outperforms its deterministic counterpart.

Another interesting observation is as follows. In order to produce, with probability $\geq 1-\delta$, an approximate solution to (52) with relative accuracy ν, the just defined SMP algorithm requires t steps, with t given by (63), and at every one of these steps it "visits" $O(1)k(m+n)p^2$ randomly chosen entries in the data matrices $A_0, A_1, ..., A_n$. The overall number of data entries visited by the algorithm is therefore $N^{\rm SMP} = O(1)tk(m+n)p^2 = O(1)\left[k\nu^{-1} + \ln(1/\delta)\nu^{-2}\right](m+n)p^2$. At the same time, the total number of data entries is $N^{\rm tot} = m(n+1)p^2$. Therefore

$$\vartheta := \frac{N^{\rm SMP}}{N^{\rm tot}} = O(1)\left[k\nu^{-1} + \ln(1/\delta)\nu^{-2}\right]\left[\frac{1}{m} + \frac{1}{n}\right] \qquad \left[k = \mathrm{Ceil}\left(\tfrac{mp}{m+n}\right)\right]$$

We see that *when δ, ν are fixed, $m \geq n/p$ and n/p is large, ϑ is small, i.e., the approximate solution of the required quality is built when inspecting a tiny fraction of the data.* This *sublinear time behavior* [14] was already observed in [4] for the Robust Mirror Descent Stochastic Approximation as applied to a matrix game (the latter problem is the particular case of (52) with $p_1 = ... = p_m = 1$ and $A_0 = 0$). Note also that an "ad hoc" sublinear time algorithm for a matrix game, in retrospect close to the one from [4], was discovered in [3] as early as in 1995.

References

[1] Azuma, K. Weighted sums of certain dependent random variables. *Tökuku Math. J.*, **19** (1967), 357-367.

[2] Ben-Tal, A., Nemirovski, A. "Non-Euclidean restricted memory level method for large-scale convex optimization" – *Math. Progr.* **102** (2005), 407–456.

[3] Grigoriadis, M.D., Khachiyan, L.G. A sublinear-time randomized approximation algorithm for matrix games. *Operations Research Letters* **18** (1995), 53–58.

[4] Juditsky, A. Lan, G., Nemirovski, A., Shapiro, A., Stochastic Approximation Approach to Stochastic Programming, SIAM Journal on Opt. **19** (2009) 1574-1609.

[5] Juditsky, A., Nemirovski, A. (2008), Large Deviations of Vector-valued Martingales in 2-Smooth Normed Spaces
E-print: http://www.optimization-online.org/DB_HTML/2008/04/1947.html

[6] Korpelevich, G. "Extrapolation gradient methods and relation to Modified Lagrangeans" *Ekonomika i Matematicheskie Metody*, **19** (1983), 694-703 (in Russian; English translation in *Matekon*).

[7] Nemirovski, A., Yudin, D., *Problem complexity and method efficiency in Optimization* J. Wiley & Sons (1983).

[8] A. Nemirovski, "Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems" – *SIAM J. Optim.* **15** (2004), 229-251.

[9] Lu, Z., Nemirovski A., Monteiro, R. "Large-Scale Semidefinite Programming via Saddle Point Mirror-Prox Algorithm", *Math. Progr.*, **109** (2007), 211-237.

[10] Nemirovski, A., Onn, S., Rothblum, U. (2007), "Accuracy certificates for computational problems with convex structure" – submitted to *Mathematics of Operations Research*. E-print: http://www.optimization-online.org/DB_HTML/2007/04/1634.html

[11] Nesterov, Yu. "Smooth minimization of non-smooth functions", *Math. Progr.*, **103** (2005), 127-152.

[12] Nesterov, Yu. "Excessive gap technique in nonsmooth convex minimization", *SIAM J. Optim.*, **16** (2005), 235-249.

[13] Nesterov, Yu. "Dual extrapolation and its applications to solving variational inequalities and related problems", *Math. Progr.* **109** (2007) 319-344.

[14] Rubinfeld, R., Sublinear time algorithms. – Marta Sanz-Solé, Javier Soria, Juan Luis Varona, Joan Verdera, Eds. *International Congress of Mathematicians, Madrid 2006*, Vol. III, 1o95–11110. European Mathematical Society Publishing House, 2006.

5. Appendix

5.1. Preliminaries

We need the following technical result about the algorithm (19), (20):

Theorem 2. *Consider t-step algorithm 1 as applied to a v.i. (2) with a monotone operator F satisfying (4). For $\tau = 1, 2, ...$, let us set*

$$\Delta_\tau = F(w_\tau) - \widehat{F}(w_\tau);$$

for z belonging to the trajectory $\{r_0, w_1, r_1, ..., w_t, r_t\}$ of the algorithm, let

$$\epsilon_z = \|\widehat{F}(z) - F(z)\|_*,$$

and let $\{y_\tau \in Z^o\}_{\tau=0}^t$ *be the sequence given by the recurrence*

$$y_\tau = P_{y_{\tau-1}}(\gamma_\tau \Delta_\tau), \;\; y_0 = r_0. \tag{64}$$

Assume that

$$\gamma_\tau \leq \frac{1}{\sqrt{3}L}, \tag{65}$$

Then

$$\mathrm{Err}_{\mathrm{vi}}(\widehat{z}_t) \leq \left(\sum\nolimits_{\tau=1}^t \gamma_\tau\right)^{-1} \Gamma(t), \tag{66}$$

where $\mathrm{Err}_{\mathrm{vi}}(\widehat{z}_t)$ *is defined in* (3),

$$\begin{array}{rcl}\Gamma(t) & = & 2\Theta(r_0) + \sum\nolimits_{\tau=1}^t \frac{3\gamma_\tau^2}{2}\left[M^2 + (\epsilon_{r_{\tau-1}} + \epsilon_{w_\tau})^2 + \frac{\epsilon_{w_\tau}^2}{3}\right] \\ & & + \sum\nolimits_{\tau=1}^t \langle \gamma_\tau \Delta_\tau, w_\tau - y_{\tau-1}\rangle \end{array} \tag{67}$$

and $\Theta(\cdot)$ *is defined by* (16).

Finally, when (2) *is a Nash v.i., one can replace* $\mathrm{Err}_{\mathrm{vi}}(\widehat{z}_t)$ *in* (66) *with* $\mathrm{Err}_{\mathrm{N}}(\widehat{z}_t)$.

Proof of Theorem 2 $\mathbf{1^0}$**.** We start with the following simple observation: if r_e is a solution to (15), then $\partial_Z \omega(r_e)$ contains $-e$ and thus is nonempty, so that $r_e \in Z^o$. Moreover, one has

$$\langle \omega'(r_e) + e, u - r_e\rangle \geq 0 \;\forall u \in Z. \tag{68}$$

Indeed, by continuity argument, it suffices to verify the inequality in the case when $u \in \mathrm{rint}(Z) \subset Z^o$. For such an u, the convex function

$$f(t) = \omega(r_e + t(u - r_e)) + \langle r_e + t(u - r_e), e\rangle, \;\; t \in [0,1]$$

is continuous on $[0,1]$ and has a continuous on $[0,1]$ field of subgradients

$$g(t) = \langle \omega'(r_e + t(u - r_e)) + e, u - r_e\rangle.$$

It follows that f is continuously differentiable on $[0,1]$ with the derivative $g(t)$. Since the function attains its minimum on $[0,1]$ at $t = 0$, we have $g(0) \geq 0$, which is exactly (68).

$\mathbf{2^0}$**.** At least the first statement of the following Lemma is well-known:

Lemma 3. *For every* $z \in Z^o$, *the mapping* $\xi \mapsto P_z(\xi)$ *is a single-valued mapping of* $\mathcal{E}$ *onto* Z^o, *and this mapping is Lipschitz continuous, specifically,*

$$\|P_z(\zeta) - P_z(\eta)\| \leq \|\zeta - \eta\|_* \quad \forall \zeta, \eta \in \mathcal{E}. \tag{69}$$

Besides this, for all $u \in Z$,

$$\begin{array}{lll}(a) & V(P_z(\zeta), u) & \leq V(z,u) + \langle \zeta, u - P_z(\zeta)\rangle - V_z(z, P_z(\zeta)) \\ (b) & & \leq V(z,u) + \langle \zeta, u - z\rangle + \frac{\|\zeta\|_*^2}{2}.\end{array} \tag{70}$$

Proof: Let $v \in P_z(\zeta)$, $w \in P_z(\eta)$. As $V'_u(z,u) = \omega'(u) - \omega'(z)$, invoking (68), we have $v, w \in Z^o$ and

$$\langle \omega'(v) - \omega'(z) + \zeta, v - u \rangle \le 0 \quad \forall u \in Z. \tag{71}$$

$$\langle \omega'(w) - \omega'(z) + \eta, w - u \rangle \le 0 \quad \forall u \in Z. \tag{72}$$

Setting $u = w$ in (71) and $u = v$ in (72), we get

$$\langle \omega'(v) - \omega'(z) + \zeta, v - w \rangle \le 0, \ \langle \omega'(w) - \omega'(z) + \eta, v - w \rangle \ge 0,$$

whence $\langle \omega'(w) - \omega'(v) + [\eta - \zeta], v - w \rangle \ge 0$, or

$$\|\eta - \zeta\|_* \|v - w\| \ge \langle \eta - \zeta, v - w \rangle \ge \langle \omega'(v) - \omega'(w), v - w \rangle \ge \|v - w\|^2,$$

and (69) follows. This relation, as a byproduct, implies that $P_z(\cdot)$ is single-valued.

To prove (70), let $v = P_z(\zeta)$. We have

$$\begin{array}{l} V(v,u) - V(z,u) \\ = [\omega(u) - \langle \omega'(v), u - v \rangle - \omega(v)] - [\omega(u) - \langle \omega'(z), u - z \rangle - \omega(z)] \\ = \langle \omega'(v) - \omega'(z) + \zeta, v - u \rangle + \langle \zeta, u - v \rangle - [\omega(v) - \langle \omega'(z), v - z \rangle - \omega(z)] \\ \le \langle \zeta, u - v \rangle - V(z,v) \text{ (due to (71))}, \end{array}$$

as required in (70.a). The bound (70.b) is obtained from (70.a) using the Young inequality:

$$\langle \zeta, z - v \rangle \le \frac{\|\zeta\|_*^2}{2} + \frac{1}{2}\|z - v\|^2.$$

Indeed, observe that by definition, $V(z,\cdot)$ is strongly convex modulus 1 w.r.t. $\|\cdot\|$, and $V(z,v) \ge \frac{1}{2}\|z - v\|^2$, so that

$$\langle \zeta, u - v \rangle - V(z,v) = \langle \zeta, u - z \rangle + \langle \zeta, z - v \rangle - V(z,v) \le \langle \zeta, u - z \rangle + \frac{\|\zeta\|_*^2}{2}. \quad \blacksquare$$

3^0. We have the following simple corollary of Lemma 3:

Corollary 2. *Let $\xi_1, \xi_2, ...$ be a sequence of elements of $\mathcal{E}$. Define the sequence $\{y_\tau\}_{\tau=0}^\infty$ in Z^o as follows:*

$$y_\tau = P_{y_{\tau-1}}(\xi_\tau), \quad y_0 \in Z^o.$$

Then y_τ is a measurable function of y_0 and $\xi_1, ..., \xi_\tau$ such that

$$(\forall u \in Z): \quad \langle -\sum\nolimits_{\tau=1}^t \xi_\tau, u \rangle \le V(y_0, u) + \sum\nolimits_{\tau=1}^t \zeta_\tau, \tag{73}$$

with $|\zeta_\tau| \le r\|\xi_\tau\|_*$ (here $r = \max_{u \in Z} \|u\|$). Further,

$$\sum\nolimits_{\tau=1}^t \zeta_\tau \le -\sum\nolimits_{\tau=1}^t \langle \xi_\tau, y_{\tau-1} \rangle + \frac{1}{2}\sum\nolimits_{\tau=1}^t \|\xi_\tau\|_*^2. \tag{74}$$

Proof: Using the bound (70.a) with $\zeta = \xi_\tau$ and $z = y_{\tau-1}$, so that $y_\tau = P_{y_{\tau-1}}(\xi_\tau)$, we obtain for any $u \in Z$:

$$V(y_\tau, u) - V(y_{\tau-1}, u) - \langle \xi_\tau, u\rangle \leq -\langle \xi_\tau, y_\tau\rangle - V(y_{\tau-1}, y_\tau) \equiv \zeta_\tau; \tag{75}$$

summing up these inequalities over τ we get (73). Further, by definition of $P_z(\xi)$ we have

$$\zeta_\tau = \max_{v\in Z}[-\langle \xi_\tau, v\rangle - V(y_{\tau-1}, v)],$$

so that $\zeta_\tau \leq r\|\xi_\tau\|_*$ due to $V \geq 0$, and

$$-r\|\xi_\tau\|_* \leq -\langle \xi_\tau, y_{\tau-1}\rangle = [-\langle \xi_\tau, y_{\tau-1}\rangle - V(y_{\tau-1}, y_{\tau-1})] \leq \zeta_\tau$$

due to $V(y_{\tau-1}, y_{\tau-1}) = 0$. Thus, $|\zeta_\tau| \leq r\|\xi_\tau\|_*$, as claimed. Further, by (70.b), where one should set $\zeta = \xi_\tau$, $z = y_{\tau-1}$, $u = y_\tau$, we have

$$\zeta_\tau \leq -\langle \xi_\tau, y_{\tau-1}\rangle + \frac{\|\xi_\tau\|_*^2}{2}.$$

Summing up these inequalities over τ, we get (74). ■

4^0. We also need the following result.

Lemma 4. *Let $z \in Z^o$, let ζ, η be two points from $\mathcal{E}$, and let*

$$w = P_z(\zeta), \qquad r_+ = P_z(\eta)$$

Then for all $u \in Z$ one has

$$\begin{array}{ll} (a) & \|w - r_+\| \leq \|\zeta - \eta\|_* \\ (b) & V(r_+, u) - V(z, u) \leq \langle \eta, u - w\rangle + [\langle \eta, w - r_+\rangle - V(z, r_+)] \\ & \leq \langle \eta,\, u - w\rangle + \frac{1}{2}\|\zeta - \eta\|_*^2 - \frac{1}{2}\|w - z\|^2. \end{array} \tag{76}$$

Proof: (a): this is nothing but (69).
(b): Using (70.a) in Lemma 3 we can write for $u = r_+$:

$$V(w, r_+) \leq V(z, r_+) + \langle \zeta, r_+ - w\rangle - V(z, w).$$

This results in

$$V(z, r_+) \geq V(w, r_+) + V(z, w) + \langle \zeta, w - r_+\rangle. \tag{77}$$

Now using (70.a) with η substituted for ζ we get

$$\begin{array}{l} V(r_+, u) \leq V(z, u) + \langle \eta, u - r_+\rangle - V(z, r_+) \\ = V(z, u) + \langle \eta, u - w\rangle + \langle \eta, w - r_+\rangle - V(z, r_+) \\ \leq V(z, u) + \langle \eta, u - w\rangle + \langle \eta - \zeta, w - r_+\rangle - V(z, w) - V(w, r_+) \text{ [by (77)]} \\ \leq V(z, u) + \langle \eta, u - w\rangle + \langle \eta - \zeta, w - r_+\rangle - \frac{1}{2}[\|w - z\|^2 + \|w - r_+\|^2], \end{array}$$

where the concluding inequalities are due to the strong convexity of $\omega(\cdot)$. To conclude the bound (b) of (76) it suffices to note that by the Young inequality,

$$\langle \eta - \zeta, w - r_+ \rangle \leq \frac{\|\eta - \zeta\|_*^2}{2} + \frac{1}{2}\|w - r_+\|^2. \qquad \blacksquare$$

5^0. We are able now to prove Theorem 2. By (4) we have

$$\begin{aligned} \|\widehat{F}(w_\tau) - \widehat{F}(r_{\tau-1})\|_*^2 &\leq (L\|r_{\tau-1} - w_\tau\| + M + \epsilon_{r_{\tau-1}} + \epsilon_{w_\tau})^2 \\ &\leq 3L^2\|w_\tau - r_{\tau-1}\|^2 + 3M^2 + 3(\epsilon_{r_{\tau-1}} + \epsilon_{w_\tau})^2. \end{aligned} \tag{78}$$

Applying Lemma 4 with $z = r_{\tau-1}$, $\zeta = \gamma_\tau \widehat{F}(r_{\tau-1})$, $\eta = \gamma_\tau \widehat{F}(w_\tau)$ (so that $w = w_\tau$ and $r_+ = r_\tau$), we have for any $u \in Z$

$$\begin{array}{l} \langle \gamma_\tau \widehat{F}(w_\tau), w_\tau - u \rangle + V(r_\tau, u) - V(r_{\tau-1}, u) \\ \leq \frac{\gamma_\tau^2}{2}\|\widehat{F}(w_\tau) - \widehat{F}(r_{\tau-1})\|^2 - \frac{1}{2}\|w_\tau - r_{\tau-1}\|^2 \\ \leq \frac{3\gamma_\tau^2}{2}\left[L^2\|w_\tau - r_{\tau-1}\|^2 + M^2 + (\epsilon_{r_{\tau-1}} + \epsilon_{w_\tau})^2\right] - \frac{1}{2}\|w_\tau - r_{\tau-1}\|^2 \text{ [by (78)]} \\ \leq \frac{3\gamma_\tau^2}{2}\left[M^2 + (\epsilon_{r_{\tau-1}} + \epsilon_{w_\tau})^2\right] \text{ [by (65)]} \end{array}$$

When summing up from $\tau = 1$ to $\tau = t$ we obtain

$$\begin{array}{l} \sum_{\tau=1}^t \langle \gamma_\tau \widehat{F}(w_\tau), w_\tau - u \rangle \\ \leq V(r_0, u) - V(r_t, u) + \sum_{\tau=1}^t \frac{3\gamma_\tau^2}{2}\left[M^2 + (\epsilon_{r_{\tau-1}} + \epsilon_{w_\tau})^2\right] \\ \leq \Theta(r_0) + \sum_{\tau=1}^t \frac{3\gamma_\tau^2}{2}\left[M^2 + (\epsilon_{r_{\tau-1}} + \epsilon_{w_\tau})^2\right]. \end{array}$$

Hence, for all $u \in Z$,

$$\begin{array}{l} \sum_{\tau=1}^t \langle \gamma_\tau F(w_\tau), w_\tau - u \rangle \\ \leq \Theta(r_0) + \sum_{\tau=1}^t \frac{3\gamma_\tau^2}{2}\left[M^2 + (\epsilon_{r_{\tau-1}} + \epsilon_{w_\tau})^2\right] + \sum_{\tau=1}^t \langle \gamma_\tau \Delta_\tau, w_\tau - u \rangle \\ = \Theta(r_0) + \sum_{\tau=1}^t \frac{3\gamma_\tau^2}{2}\left[M^2 + (\epsilon_{r_{\tau-1}} + \epsilon_{w_\tau})^2\right] + \sum_{\tau=1}^t \langle \gamma_\tau \Delta_\tau, w_\tau - y_{\tau-1} \rangle \\ \quad + \sum_{\tau=1}^t \langle \gamma_\tau \Delta_\tau, y_{\tau-1} - u \rangle \end{array} \tag{79}$$

where y_τ are given by (64). Since the sequences $\{y_\tau\}$, $\{\xi_\tau = \gamma_\tau \Delta_\tau\}$ satisfy the premise of Corollary 2, we have

$$\begin{aligned} (\forall u \in Z): \quad \sum_{\tau=1}^t \langle \gamma_\tau \Delta_\tau, y_{\tau-1} - u \rangle &\leq V(r_0, u) + \sum_{\tau=1}^t \frac{\gamma_\tau^2}{2}\|\Delta_\tau\|_*^2 \\ &\leq \Theta(r_0) + \sum_{\tau=1}^t \frac{\gamma_\tau^2}{2}\epsilon_{w_\tau}^2, \end{aligned}$$

and thus (79) implies that for any $u \in Z$

$$\sum\nolimits_{\tau=1}^t \langle \gamma_\tau F(w_\tau), w_\tau - u \rangle \leq \Gamma(t) \tag{80}$$

with $\Gamma(t)$ defined in (67). To complete the proof of (66) in the general case, note that since F is monotone, (80) implies that for all $u \in Z$,

$$\sum\nolimits_{\tau=1}^t \gamma_\tau \langle F(u), w_\tau - u \rangle \leq \Gamma(t),$$

whence

$$\forall (u \in Z) : \langle F(u), \widehat{z}_t - u\rangle \leq \left[\sum\nolimits_{\tau=1}^{t} \gamma_\tau\right]^{-1} \Gamma(t).$$

When taking the supremum over $u \in Z$, we arrive at (66).

In the case of a Nash v.i., setting $w_\tau = (w_{\tau,1}, ..., w_{\tau,m})$ and $u = (u_1, ..., u_m)$ and recalling the origin of F, due to the convexity of $\phi_i(z_i, z^i)$ in z_i, for all $u \in Z$ we get from (80):

$$\sum_{\tau=1}^{t} \gamma_\tau \sum\nolimits_{i=1}^{m} [\phi_i(w_\tau) - \phi_i(u_i, (w_\tau)^i)] \leq \sum_{\tau=1}^{t} \gamma_\tau \sum_{i=1}^{m} \langle F^i(w_\tau), (w_\tau)_i - u_i\rangle \leq \Gamma(t).$$

Setting $\phi(z) = \sum_{i=1}^m \phi_i(z)$, we get

$$\sum\nolimits_{\tau=1}^{t} \gamma_\tau \left[\phi(w_\tau) - \sum\nolimits_{i=1}^{m} \phi_i(u_i, (w_\tau)^i)\right] \leq \Gamma(t).$$

Recalling that $\phi(\cdot)$ is convex and $\phi_i(u_i, \cdot)$ are concave, $i = 1, ..., m$, the latter inequality implies that

$$\left[\sum\nolimits_{\tau=1}^{t} \gamma_\tau\right] \left[\phi(\widehat{z}_t) - \sum\nolimits_{i=1}^{m} \phi_i(u_i, (\widehat{z}_t)^i)\right] \leq \Gamma(t),$$

or, which is the same,

$$\sum\nolimits_{i=1}^{m} \left[\phi_i(\widehat{z}_t) - \sum\nolimits_{i=1}^{m} \phi_i(u_i, (\widehat{z}_t)^i)\right] \leq \left[\sum\nolimits_{\tau=1}^{t} \gamma_\tau\right]^{-1} \Gamma(t).$$

This relation holds true for all $u = (u_1, ..., u_m) \in Z$; taking maximum of both sides in u, we get

$$\mathrm{Err_N}(\widehat{z}_t) \leq \left[\sum\nolimits_{\tau=1}^{t} \gamma_\tau\right]^{-1} \Gamma(t). \qquad \blacksquare$$

5.2. Proof of Theorem 1

In what follows, we use the notation from Theorem 2. By this theorem, in the case of constant stepsizes $\gamma_\tau \equiv \gamma$ we have

$$\mathrm{Err_{vi}}(\widehat{z}_t) \leq [t\gamma]^{-1} \Gamma(t), \tag{81}$$

where

$$\begin{aligned} \Gamma(t) &= \Omega^2 + \frac{3\gamma^2}{2} \sum_{\tau=1}^{t} \left[M^2 + (\epsilon_{r_{\tau-1}} + \epsilon_{w_\tau})^2 + \frac{\epsilon_{w_\tau}^2}{3}\right] + \gamma \sum_{\tau=1}^{t} \langle \Delta_\tau, w_\tau - y_{\tau-1}\rangle \\ &\leq \Omega^2 + \frac{7\gamma^2}{2} \sum_{\tau=1}^{t} \left[M^2 + \epsilon_{r_{\tau-1}}^2 + \epsilon_{w_\tau}^2\right] + \gamma \sum_{\tau=1}^{t} \langle \Delta_\tau, w_\tau - y_{\tau-1}\rangle. \end{aligned} \tag{82}$$

For a Nash v.i., $\mathrm{Err_{vi}}$ in this relation can be replaced with $\mathrm{Err_N}$.

Let us suppose that the random vectors ζ_i are defined on the probability space $(\Omega, \mathcal{F}, \mathbf{P})$. We define two nested families of σ-fields $\mathcal{F}_i = \sigma(r_0, \zeta_1, \zeta_2, \dots, \zeta_{2i-1})$ and $\mathcal{G}_i = \sigma(r_0, \zeta_1, \zeta_2, \dots, \zeta_{2i})$, $i = 1, 2, ...$, so that $\mathcal{F}_1 \subset ... \mathcal{G}_{i-1} \subset \mathcal{F}_i \subset \mathcal{G}_i \subset$ .... Then by description of the algorithm $r_{\tau-1}$ is $\mathcal{G}_{\tau-1}$-measurable and w_τ is $\mathcal{F}_\tau$-measurable. Therefore $\epsilon_{r_{\tau-1}}$ is $\mathcal{F}_\tau$-measurable, and ϵ_{w_τ} and Δ_τ are $\mathcal{G}_\tau$-measurable. We conclude that under Assumption I we have

$$\mathbf{E}\left\{\epsilon^2_{r_{\tau-1}} | \mathcal{G}_{\tau-1}\right\} \leq \sigma^2,\ \mathbf{E}\left\{\epsilon^2_{w_\tau} | \mathcal{F}_\tau\right\} \leq \sigma^2,\ \|\mathbf{E}\left\{\Delta_\tau | \mathcal{F}_\tau\right\}\|_* \leq \mu, \tag{83}$$

and under Assumption II, in addition,

$$\begin{array}{l} \mathbf{E}\left\{\exp\{\epsilon^2_{r_{\tau-1}}\sigma^{-2}\}\big|\mathcal{G}_{\tau-1}\right\} \leq \exp\{1\}, \\ \mathbf{E}\left\{\exp\{\epsilon^2_{w_\tau}\sigma^{-2}\}\big|\mathcal{F}_\tau\right\} \leq \exp\{1\}. \end{array} \tag{84}$$

Now, let

$$\Gamma_0(t) = \frac{7\gamma^2}{2} \sum\nolimits_{\tau=1}^t \left[M^2 + \epsilon^2_{r_{\tau-1}} + \epsilon^2_{w_\tau}\right].$$

We conclude by (83) that

$$\mathbf{E}\left\{\Gamma_0(t)\right\} \leq \frac{7\gamma^2 t}{2}[M^2 + 2\sigma^2]. \tag{85}$$

Further, $y_{\tau-1}$ clearly is $\mathcal{F}_{\tau-1}$-measurable, whence $w_\tau - y_{\tau-1}$ is $\mathcal{F}_\tau$-measurable. Therefore

$$\begin{array}{rcl} \mathbf{E}\left\{\langle \Delta_\tau, w_\tau - y_{\tau-1}\rangle \big| \mathcal{F}_\tau\right\} & = & \langle \mathbf{E}\left\{\Delta_\tau | \mathcal{F}_\tau\right\}, w_\tau - y_{\tau-1}\rangle \\ & \leq & \mu\|w_\tau - y_{\tau-1}\| \leq 2\mu\Omega, \end{array} \tag{86}$$

where the concluding inequality follows from the fact that Z is contained in the $\|\cdot\|$-ball of radius Ω centered at z_c, see (18). From (86) it follows that

$$\mathbf{E}\left\{\gamma \sum\nolimits_{\tau=1}^t \langle \Delta_\tau, w_\tau - y_{\tau-1}\rangle\right\} \leq 2\mu\gamma t\Omega.$$

Combining the latter relation, (81), (82) and (85), we arrive at (23). (i) is proved.

To prove (ii), observe, first, that setting

$$J_t = \sum\nolimits_{\tau=1}^t \left[\sigma^{-2}\epsilon^2_{r_{\tau-1}} + \sigma^{-2}\epsilon^2_{w_\tau}\right],$$

we get

$$\Gamma_0(t) = \frac{7\gamma^2 M^2 t}{2} + \frac{7\gamma^2\sigma^2}{2} J_t. \tag{87}$$

At the same time, setting $\mathcal{H}_j = \sigma(r_0, \xi_1, ..., \xi_j\}$, we can write

$$J_t = \sum\nolimits_{j=1}^{2t} \xi_j,$$

where $\xi_j \geq 0$ is $\mathcal{H}_j$-measurable, and

$$\mathbf{E}\left\{\exp\{\xi_j\}|\mathcal{H}_{j-1}\right\} \leq \exp\{1\},$$

see (84). It follows that

$$\begin{array}{l}\mathbf{E}\left\{\exp\{\sum_{j=1}^{k+1}\xi_j\}\right\} = \mathbf{E}\left\{\mathbf{E}\left\{\exp\{\sum_{j=1}^{k}\xi_j\}\exp\{\xi_{k+1}\}\right\}|\mathcal{H}_k\right\} \\ = \mathbf{E}\left\{\exp\{\sum_{j=1}^{k}\xi_j\}\mathbf{E}\left\{\exp\{\xi_{k+1}\}|\mathcal{H}_k\right\}\right\} \leq \exp\{1\}\mathbf{E}\left\{\exp\{\sum_{j=1}^{k}\xi_j\}\right\}.\end{array} \tag{88}$$

Whence $\mathbf{E}[\exp\{J\}] \leq \exp\{2t\}$, and applying the Tchebychev inequality, we get

$$\forall \Lambda > 0: \ \mathrm{Prob}\left\{J > 2t + \Lambda t\right\} \leq \exp\{-\Lambda t\}.$$

Along with (87) it implies that

$$\forall \Lambda \geq 0: \mathrm{Prob}\left\{\Gamma_0(t) > \frac{7\gamma^2 t}{2}[M^2 + 2\sigma^2] + \Lambda\frac{7\gamma^2\sigma^2 t}{2}\right\} \leq \exp\{-\Lambda t\}. \tag{89}$$

Let now $\xi_\tau = \langle \Delta_\tau, w_\tau - y_{\tau-1}\rangle$. Recall that $w_\tau - y_{\tau-1}$ is $\mathcal{F}_\tau$-measurable. Besides this, we have seen that $\|w_\tau - y_{\tau-1}\| \leq D \equiv 2\Omega$. Taking into account (84) and (86), we get

$$\begin{array}{ll}(a) & \mathbf{E}\left\{\xi_\tau|\mathcal{F}_\tau\right\} \leq \rho \equiv \mu D, \\ (b) & \mathbf{E}\left\{\exp\{\xi_\tau^2 R^{-2}\}|\mathcal{F}_\tau\right\} \leq \exp\{1\}, \ \text{ with } R = \sigma D.\end{array} \tag{90}$$

Observe that $\exp\{x\} \leq x + \exp\{9x^2/16\}$ for all x. Thus (90.b) implies for $0 \leq s \leq \frac{4}{3R}$

$$\begin{array}{rcl}\mathbf{E}\left\{\exp\{s\xi_\tau\}|\mathcal{F}_\tau\right\} & \leq & \mathbf{E}\{s\xi_\tau|\mathcal{F}_\tau\} + \mathbf{E}\left\{\exp\left\{\frac{9s^2\xi_\tau^2}{16}\right\}|\mathcal{F}_\tau\right\} \\ & \leq & s\rho + \exp\left\{\frac{9s^2R^2}{16}\right\} \leq \exp\left\{s\rho + \frac{9s^2R^2}{16}\right\}.\end{array} \tag{91}$$

Further, we have $s\xi_\tau \leq \frac{3}{8}s^2R^2 + \frac{2}{3}\xi_\tau^2 R^{-2}$, hence for all $s \geq 0$,

$$\mathbf{E}\left\{\exp\{s\xi_\tau\}|\mathcal{F}_\tau\right\} \leq \exp\{3s^2R^2/8\}\mathbf{E}\left\{\exp\left\{\frac{2\xi_\tau^2}{3R^2}\right\}|\mathcal{F}_\tau\right\} \leq \exp\left\{\frac{3s^2R^2}{8} + \frac{2}{3}\right\}.$$

When $s \geq \frac{4}{3R}$, the latter quantity is $\leq \exp\{3s^2R^2/4\}$, which combines with (91) to imply that for $s \geq 0$,

$$\mathbf{E}\left\{\exp\{s\xi_\tau\}|\mathcal{F}_\tau\right\} \leq \exp\{s\rho + 3s^2R^2/4\}. \tag{92}$$

Acting as in (88), we derive from (92) that

$$s \geq 0 \Rightarrow \mathbf{E}\left\{\exp\{s\sum_{\tau=1}^{t}\xi_\tau\}\right\} \leq \exp\{st\rho + 3s^2tR^2/4\},$$

and by the Tchebychev inequality, for all $\Lambda > 0$,

$$\mathrm{Prob}\left\{\sum_{\tau=1}^{t}\xi_\tau > t\rho + \Lambda R\sqrt{t}\right\} \le \inf_{s\ge 0}\exp\{3s^2tR^2/4 - s\Lambda R\sqrt{t}\} = \exp\{-\Lambda^2/3\}.$$

Finally, we arrive at

$$\mathrm{Prob}\left\{\gamma\sum_{\tau=1}^{t}\langle\Delta_\tau, w_\tau - y_{\tau-1}\rangle > 2\gamma\left[\mu t + \Lambda\sigma\sqrt{t}\right]\Omega\right\} \le \exp\{-\Lambda^2/3\}. \quad (93)$$

for all $\Lambda > 0$. Combining (81), (82), (89) and (93), we get (24). ∎

5.3. Proof of Lemma 1

Proof of (i) We clearly have $Z^o = X^o \times Y^o$, and $\omega(\cdot)$ is indeed continuously differentiable on this set. Let $z = (x, y)$ and $z' = (x', y')$, $z, z' \in Z$. Then

$$\begin{aligned}
&\langle\omega'(z) - \omega'(z'), z - z'\rangle \\
&= \tfrac{1}{\Omega_x^2}\langle\omega_x'(x) - \omega_x'(x'), x - x'\rangle + \tfrac{1}{\Omega_y^2}\langle\omega_y'(y) - \omega_y'(y'), y - y'\rangle \\
&\ge \tfrac{1}{\Omega_x^2}\|x - x'\|_x^2 + \tfrac{1}{\Omega_y^2}\|y - y'\|_y^2 \ge \|[x' - x; y' - y]\|^2.
\end{aligned}$$

Thus, $\omega(\cdot)$ is strongly convex on Z, modulus 1, w.r.t. the norm $\|\cdot\|$. Further, the minimizer of $\omega(\cdot)$ on Z clearly is $z_c = (x_c, y_c)$, and it is immediately seen that $\max_{z\in Z} V(z_c, z) = 1$, whence $\Omega = \sqrt{2}$.

Proof of (ii) **1^0.** Let $z = (x, y)$ and $z' = (x', y')$ with $z, z' \in Z$. Note that by assumption **G** in Section 4.1 we have

$$\|y'\|_y \le 2\Omega_y. \quad (94)$$

Further, we have from (12) $F(z') - F(z) = [\Delta_x; \Delta_y]$, where

$$\begin{aligned}
\Delta_x &= \sum_{\ell=1}^{m}[\phi_\ell'(x') - \phi_\ell'(x)]^*[\mathbf{A}_\ell y' + b_\ell] + \sum_{\ell=1}^{m}[\phi_\ell'(x)]^*\mathbf{A}_\ell[y' - y], \\
\Delta_y &= -\sum_{\ell=1}^{m}\mathbf{A}_\ell^*[\phi_\ell(x) - \phi_\ell(x')] + \Phi_*'(y') - \Phi_*'(y).
\end{aligned}$$

We have

$$\begin{aligned}
&\|\Delta_x\|_{x,*} \\
&= \max_{h\in\mathcal{X}\ \|h\|_x\le 1}\langle h, \sum_{\ell=1}^{m}[[\phi_\ell'(x') - \phi_\ell'(x)]^*[\mathbf{A}_\ell y' + b_\ell] + [\phi_\ell'(x)]^*\mathbf{A}_\ell[y' - y]]\rangle_{\mathcal{X}} \\
&\le \sum_{\ell=1}^{m}\left[\max_{\substack{h\in\mathcal{X}\\ \|h\|_x\le 1}}\langle h, [\phi_\ell'(x') - \phi_\ell'(x)]^*[\mathbf{A}_\ell y' + b_\ell]\rangle_{\mathcal{X}} + \max_{\substack{h\in\mathcal{X},\\ \|h\|_x\le 1}}\langle h, [\phi_\ell'(x)]^*\mathbf{A}_\ell[y' - y]\rangle_{\mathcal{X}}\right] \\
&= \sum_{\ell=1}^{m}\left[\max_{\substack{h\in\mathcal{X}\\ \|h\|_x\le 1}}\langle[\phi_\ell'(x') - \phi_\ell'(x)]h, \mathbf{A}_\ell y' + b_\ell\rangle_{\mathcal{X}} + \max_{h\in\mathcal{X}, \|h\|_x\le 1}\langle[\phi_\ell'(x)]h, \mathbf{A}_\ell[y' - y]\rangle_{\mathcal{X}}\right] \\
&\le \sum_{\ell=1}^{m}\Big[\max_{h\in\mathcal{X}\ \|h\|_x\le 1}\|[\phi_\ell'(x') - \phi_\ell'(x)]h\|_{(\ell)}\|\mathbf{A}_\ell y' + b_\ell\|_{(\ell,*)} \\
&\qquad\qquad + \max_{h\in\mathcal{X}\ \|h\|_x\le 1}\|\phi_\ell'(x)h\|_{(\ell)}\|\mathbf{A}_\ell[y' - y]\|_{(\ell,*)}\Big].
\end{aligned}$$

Then by (32),

$$\begin{array}{l}\|\Delta_x\|_{x,*}\\ \leq \sum\limits_{\ell=1}^m \left[[L_x\|x-x'\|_x+M_x][\|\mathbf{A}_\ell y'\|_{(\ell,*)}+\|b_\ell\|_{(\ell,*)}]+\Omega_x L_x\|\mathbf{A}_\ell[y-y']\|_{(\ell,*)}\right]\\ =[L_x\|x-x'\|_x+M_x]\sum\limits_{\ell=1}^m[\|\mathbf{A}_\ell y'\|_{(\ell,*)}+\|b_\ell\|_{(\ell,*)}]+\Omega_x L_x\sum\limits_{\ell=1}^m\|\mathbf{A}_\ell[y-y']\|_{(\ell,*)}\\ \leq [L_x\|x-x'\|_x+M_x][\mathcal{A}\|y'\|_y+\mathcal{B}]+\Omega_x L_x\mathcal{A}\|y-y'\|_y,\end{array}$$

by definition of $\mathcal{A}$ and $\mathcal{B}$. Next, due to (94) we get by definition of $\|\cdot\|$

$$\begin{array}{rcl}\|\Delta_x\|_{x,*} &\leq& [L_x\|x-x'\|_x+M_x][2\mathcal{A}\Omega_y+\mathcal{B}]+\Omega_x L_x\mathcal{A}\|y-y'\|_y\\ &\leq& [\Omega_x L_x\|z-z'\|+M_x][2\mathcal{A}\Omega_y+\mathcal{B}]+\Omega_x L_x\mathcal{A}\Omega_y\|z-z'\|,\end{array}$$

which implies

$$\|\Delta_x\|_{x,*}\leq[3\mathcal{A}\Omega_x L_x\Omega_y+\mathcal{B}]\,\|z-z'\|+[2\mathcal{A}\Omega_y+\mathcal{B}]M_x. \tag{95}$$

Further,

$$\begin{array}{rcl}\|\Delta_y\|_{y,*} &=& \max\limits_{\eta\in\mathcal{Y},\|\eta\|_y\leq 1}\langle\eta,-\sum_{\ell=1}^m\mathbf{A}_\ell^*[\phi_\ell(x)-\phi_\ell(x')]+\Phi_*'(y')-\Phi_*'(y)\rangle_{\mathcal{Y}}\\ &\leq& \max\limits_{\eta\in\mathcal{Y},\|\eta\|_y\leq 1}\sum_{\ell=1}^m\langle\eta,\mathbf{A}_\ell^*[\phi_\ell(x)-\phi_\ell(x')]\rangle_{\mathcal{Y}}+\|\Phi_*'(y')-\Phi_*'(y)\|_{y,*}\\ &=& \max\limits_{\eta\in\mathcal{Y},\|\eta\|_y\leq 1}\sum_{\ell=1}^m\langle\mathbf{A}_\ell\eta,\phi_\ell(x)-\phi_\ell(x')\rangle_{\mathcal{E}_\ell}+\|\Phi_*'(y')-\Phi_*'(y)\|_{y,*}\\ &\leq& \max\limits_{\eta\in\mathcal{Y},\|\eta\|_y\leq 1}\sum_{\ell=1}^m\|\mathbf{A}_\ell\eta\|_{(\ell,*)}\|\phi_\ell(x)-\phi_\ell(x')\|_{(\ell)}+\|\Phi_*'(y')-\Phi_*'(y)\|_{y,*}\\ &\leq& \max\limits_{\eta\in\mathcal{Y},\|\eta\|_y\leq 1}\sum_{\ell=1}^m\|\mathbf{A}_\ell\eta\|_{(\ell,*)}\Omega_x L_x\|x-x'\|_x+[L_y\|y-y'\|_y+M_y],\end{array}$$

by (32.b) and (34). Therefore

$$\|\Delta_y\|_{y,*}\leq\mathcal{A}\Omega_x L_x\|x-x'\|_x+L_y\|y-y'\|_y+M_y,$$

whence

$$\|\Delta_y\|_{y,*}\leq[\mathcal{A}\Omega_x^2 L_x+\Omega_y L_y]\|z-z'\|+M_y. \tag{96}$$

From (95), (96) it follows that

$$\begin{array}{l}\|F(z)-F(z')\|_*\leq\Omega_x\|\Delta_x\|_{x,*}+\Omega_y\|\Delta_y\|_{y,*}\\ \leq\left[4\mathcal{A}L_x\Omega_x^2\Omega_y+\Omega_x\mathcal{B}+\Omega_y^2 L_y\right]\|z-z'\|+[2\mathcal{A}\Omega_x\Omega_y+\Omega_x\mathcal{B}]\,M_x+\Omega_y M_y.\end{array}$$

We have justified (37)

2^0. Let us verify (38). The first relation in (38) is readily given by (33.a,c). Let us fix $z=(x,y)\in Z$ and i, and let

$$\begin{array}{rcl}\Delta &=& F(z)-\Xi(z,\zeta_i)\\ &=& \Big[\underbrace{\sum_{\ell=1}^m[\phi_\ell'(x)-\mathbf{G}_\ell(x,\zeta_i)]^*\overbrace{[\mathbf{A}_\ell y+b_\ell]}^{\psi_\ell}}_{\Delta_x};\underbrace{-\sum_{\ell=1}^m\mathbf{A}_\ell^*[\phi_\ell(x)-f_\ell(x,\zeta_i)]}_{\Delta_y}\Big].\end{array} \tag{97}$$

As we have seen,

$$\sum_{\ell=1}^{m}\|\psi_\ell\|_{(\ell,*)} \leq 2\mathcal{A}\Omega_y + \mathcal{B}. \tag{98}$$

Besides this, for $u_\ell \in \mathcal{E}_\ell$ we have

$$\begin{array}{l}
\|\sum_{\ell=1}^{m}\mathbf{A}_\ell^* u_\ell\|_{y,*} = \max\limits_{\eta\in\mathcal{Y},\,\|\eta\|_y\leq 1}\langle\sum_{\ell=1}^{m}\mathbf{A}_\ell^* u_\ell,\eta\rangle_{\mathcal{Y}} = \max\limits_{\eta\in\mathcal{Y},\,\|\eta\|_y\leq 1}\langle\sum_{\ell=1}^{m}u_\ell,\mathbf{A}_\ell\eta\rangle_{\mathcal{Y}}\\
\leq \max\limits_{\eta\in\mathcal{Y},\,\|\eta\|_y\leq 1}\left[\sum_{1\leq\ell\leq m}\|u_\ell\|_{(\ell)}\|\mathbf{A}_\ell\eta\|_{(\ell,*)}\right]\\
\leq \max\limits_{\eta\in\mathcal{Y},\,\|\eta\|_y\leq 1}\left[\max_{1\leq\ell\leq m}\|u_\ell\|_{(\ell)}\right]\sum_{1\leq\ell\leq m}\|\mathbf{A}_\ell\eta\|_{(\ell,*)} = \mathcal{A}\max_{1\leq\ell\leq m}\|u_\ell\|_{(\ell)}.
\end{array} \tag{99}$$

Hence, setting $u_\ell = \phi_\ell(x) - f_\ell(x,\zeta_i)$ we obtain

$$\|\Delta_y\|_{y,*} = \|\sum_{\ell=1}^{m}\mathbf{A}_\ell^*[\phi_\ell(x) - f_\ell(x,\zeta_i)]\|_{y,*} \leq \mathcal{A}\underbrace{\max_{1\leq\ell\leq m}\|\phi_\ell(x) - f_\ell(x,\zeta_i)\|_{(\ell)}}_{\xi=\xi(\zeta_i)}. \tag{100}$$

Further,

$$\begin{array}{rcl}
\|\Delta_x\|_{x,*} &=& \max\limits_{h\in\mathcal{X},\,\|h\|_x\leq 1}\langle h,\sum_{\ell=1}^{m}[\phi_\ell'(x) - \mathbf{G}_\ell(x,\zeta_i)]^*\psi_\ell\rangle_{\mathcal{X}}\\
&=& \max\limits_{h\in\mathcal{X},\,\|h\|_x\leq 1}\sum_{\ell=1}^{m}\langle[\phi_\ell'(x) - \mathbf{G}_\ell(x,\zeta_i)]h,\psi_\ell\rangle_{\mathcal{X}}\\
&\leq& \max\limits_{h\in\mathcal{X},\,\|h\|_x\leq 1}\sum_{\ell=1}^{m}\|[\phi_\ell'(x) - \mathbf{G}_\ell(x,\zeta_i)]h\|_{(\ell)}\|\psi_\ell\|_{(\ell,*)}\\
&\leq& \sum_{\ell=1}^{m}\underbrace{\max\limits_{h\in\mathcal{X},\,\|h\|_x\leq 1}\|[\phi_\ell'(x) - \mathbf{G}_\ell(x,\zeta_i)]h\|_{(\ell)}}_{\xi_\ell=\xi_\ell(\zeta_i)}\underbrace{\|\psi_\ell\|_{(\ell,*)}}_{\rho_\ell}
\end{array}$$

Invoking (98), we conclude that

$$\|\Delta_x\|_{x,*} \leq \sum_{\ell=1}^{m}\rho_\ell\xi_\ell, \tag{101}$$

where all $\rho_\ell \geq 0$, $\sum_\ell \rho_\ell \leq 2\mathcal{A}\Omega_y + \mathcal{B}$ and

$$\xi_\ell = \xi_\ell(\zeta_i) = \max_{h\in\mathcal{X},\,\|h\|_x\leq 1}\|[\phi_\ell'(x) - \mathbf{G}_\ell(x,\zeta_i)]h\|_{(\ell)}$$

Denoting by $p^2(\eta)$ the second moment of a scalar random variable η, observe that $p(\cdot)$ is a norm on the space of square summable random variables representable as deterministic functions of ζ_i, and that

$$p(\xi) \leq \Omega_x M_x,\; p(\xi_\ell) \leq M_x$$

by (33.b,d). Now by (100), (101),

$$\begin{array}{rcl}\left[\mathbf{E}\left\{\|\Delta\|_*^2\right\}\right]^{\frac{1}{2}} &=& \left[\mathbf{E}\left\{\Omega_x^2\|\Delta_x\|_{x,*}^2+\Omega_y^2\|\Delta_y\|_{y,*}^2\right\}\right]^{\frac{1}{2}}\\ &\leq& p\left(\Omega_x\|\Delta_x\|_{x,*}+\Omega_y\|\Delta_y\|_{y,*}\right)\leq p\left(\Omega_x\sum_{\ell=1}^m\rho_\ell\xi_\ell+\Omega_y\mathcal{A}\xi\right)\\ &\leq& \Omega_x\sum_\ell\rho_\ell\max_\ell p(\xi_\ell)+\Omega_y\mathcal{A}p(\xi)\\ &\leq& \Omega_x[2\mathcal{A}\Omega_y+\mathcal{B}]M_x+\Omega_y\mathcal{A}\Omega_xM_x,\end{array}$$

and the latter quantity is $\leq M$, see (37). We have established the second relation in (38).

3^0. It remains to prove that in the case of (39), relation (40) takes place. To this end, one can repeat word by word the reasoning from item 2^0 with the function $p_e(\eta)=\inf\left\{t>0:\mathbf{E}\left\{\exp\{\eta^2/t^2\}\right\}\leq\exp\{1\}\right\}$ in the role of $p(\eta)$. Note that similarly to $p(\cdot)$, $p_e(\cdot)$ is a norm on the space of random variables η which are deterministic functions of ζ_i and are such that $p_e(\eta)<\infty$. ∎

5.4. Proof of Lemma 2

Item (i) can be verified exactly as in the case of Lemma 1; the facts expressed in (i) depend solely on the construction from Section 4.2 preceding the latter Lemma, and are independent of what are the setups for $X,\mathcal{X}$ and $Y,\mathcal{Y}$.

Let us verify item (ii). Note that we are in the situation

$$\begin{array}{l}\|(x,\,y)\|=\sqrt{\|x\|_1^2/(2\ln(n))+|y|_1^2/(4\ln(p^{(1)}))},\\ \|(\xi,\,\eta)\|_*=\sqrt{2\ln(n)\|\xi\|_\infty^2+4\ln(p^{(1)})|\eta|_\infty^2}.\end{array}\tag{102}$$

For $z=(x,y)$, $z'=(x',y')\in Z$ we have

$$F(z)-F(z')=\Big[\underbrace{[\mathrm{Tr}((y-y')A_1);\ldots;\mathrm{Tr}((y-y')A_n)]}_{\Delta_x};\underbrace{-\sum_{j=1}^n(x_j-x_j')A_j}_{\Delta_y}\Big].$$

whence

$$\begin{array}{rcl}\|\Delta_x\|_\infty &\leq& |y-y'|_1\max\limits_{1\leq j\leq n}|A_j|_\infty\leq 2\sqrt{\ln(p^{(1)})}A_\infty\|z-z'\|,\\ |\Delta_y|_\infty &\leq& \|x-x'\|_\infty\max\limits_{1\leq j\leq n}|A_j|_\infty\leq\sqrt{2\ln(n)}A_\infty\|z-z'\|,\end{array}$$

and

$$\|(\Delta_x,\,\Delta_y)\|_*\leq 2\sqrt{2\ln(n)\ln(p^{(1)})}A_\infty\|z-z'\|,$$

as required in (59). Further, relation (60.a) is clear from the construction of Ξ_k. To prove (60.b), observe that when $(x,y)\in Z$, we have $\|\Xi^x(x,y,\eta)\|_\infty\leq\leq A_\infty$, $|\Xi^y(x,y,\eta)|_\infty\leq A_\infty$ (see (57), (58)), whence

$$\|\Xi^x(x,y,\eta)-F^x(x,y)\|_\infty\leq 2A_\infty,\ |\Xi^y(x,y,\eta)-F^y(x,y)|_\infty\leq 2A_\infty\tag{103}$$

due to $F(x,y) = \mathbf{E}_\eta\left\{\Xi(x,y,\eta)\right\}$, Applying [5, Theorem 2.1(iii), Example 3.2, Lemma 1], we derive from (103) that for every $(x,y)\in Z$ and every $i=1,2,...$ it holds

$$\begin{array}{l}\mathbf{E}\left\{\exp\{\|\Xi_k^x(x,y,\zeta_i)-F^x(x,y)\|_\infty^2/N_{k,x}^2\}\right\}\leq\exp\{1\},\\ N_{k,x}=2A_\infty\left(2\exp\{1/2\}\sqrt{\ln(n)}+3\right)k^{-1/2}\end{array}$$

and

$$\begin{array}{l}\mathbf{E}\left\{\exp\{\|\Xi_k^y(x,y,\zeta_i)-F^y(x,y)\|_\infty^2/N_{k,y}^2\}\right\}\leq\exp\{1\},\\ N_{k,y}=2A_\infty\left(2\exp\{1/2\}\sqrt{\ln(p^{(1)})}+3\right)k^{-1/2}.\end{array}$$

Combining the latter bounds with (102) we get (60.b). ∎