Solving variational inequalities with Stochastic Mirror-Prox algorithm

Anatoli B. Juditsky, Arkadii S. Nemirovskii, Claire Tauvel

SOLVING VARIATIONAL INEQUALITIES WITH STOCHASTIC MIRROR-PROX ALGORITHM

ANATOLI JUDITSKY*, ARKADI NEMIROVSKI†, AND CLAIRE TAUVEL‡

September 4, 2008

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 Feasability problem and Eigenvalue minimization.

Key words. Nash variational inequalities, stochastic convex-concave saddle-point problem, large scale stochastic approximation, reduced complexity algorithms for convex optimization

AMS subject classifications. 90C15, 65K10, 90C47

1. Introduction.

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:

$$\forall(z, z' \in Z) : \langle F(z) - F(z'), z - z'\rangle \geqslant 0) \tag{1.1}$$

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

$$\text{find } z_* \in Z : \langle F(z), z_* - z\rangle \leqslant 0 \quad \forall z \in Z \tag{1.2}$$

associated with Z, F. Note that since F is monotone on Z, the condition in (1.2) is implied by $\langle F(z_*), z - z_*\rangle \geqslant 0$ 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 (1.2) (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

$$\mathrm{Err}_{\mathrm{vi}}(z) := \max_{u\in Z}\langle F(u), z - u\rangle; \tag{1.3}$$

note that this error is always $\geqslant 0$ and equals zero iff z is a solution to (1.2).

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

$$\forall(z, z' \in Z) : \|F(z) - F(z')\|_* \leqslant L\|z - z'\| + M \tag{1.4}$$

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

$$\|\xi\|_* = \max_{z:\|z\|\leqslant 1} \langle \xi, z\rangle \tag{1.5}$$

*LJK, Université J. Fourier, B.P. 53, 38041 Grenoble Cedex 9, France, anatoli.juditsky@imag.fr

†Georgia Institute of Technology, Atlanta, Georgia 30332, USA, nemirovs@isye.gatech.edu,
research of this author was partly supported by the NSF award DMI-0619977.
‡LJK, Université J. Fourier, B.P. 53, 38041 Grenoble Cedex 9, France, claire.tauvel@imag.fr

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

We are interested in the case where (1.2) 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 i-th 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

$$\forall z \in Z: \ \mathbf{E}\{\Xi(z, \zeta_1)\} = F(z), \ \ \mathbf{E}\left\{\|\Xi(z, \zeta_i) - F(z)\|_*^2\right\} \leqslant N^2. \tag{1.6}$$

We call a monotone v.i. (1.1), augmented by a stochastic oracle (SO), a *stochastic* monotone v.i. (s.v.i.).

To motivate our goal, let us start with known results [5] on the limits of performance of iterative algorithms for solving large-scale stochastic 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 rate of convergence of a whatever algorithm for solving v.i.'s cannot be better than $O(1)\left[\frac{L}{t} + \frac{M+N}{\sqrt{t}}\right]$. 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 (1.4), (1.6) 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 $c\left[\frac{L}{t} + \frac{M+N}{\sqrt{t}}\right]$ for some $c > 0$. To the best of our knowledge, no one of existing algorithms allows to achieve, uniformly in the dimension, this convergence rate. In fact, the "best approximations" available are given by Robust Stochastic Approximation (see [3] and references therein) with the guaranteed rate of convergence $O(1)\frac{L+M+N}{\sqrt{t}}$ and extra-gradient-type algorithms for solving deterministic monotone v.i.'s with Lipschitz continuous operators (see [6, 9, 10, 11]), which attains the accuracy $O(1)\frac{L}{t}$ in the case of $M = N = 0$ or $O(1)\frac{M}{\sqrt{t}}$ when $L = N = 0$.

The goal of this paper is to demonstrate that a specific *Mirror-Prox* algorithm [6] 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 $O(1)\left[\frac{L}{t} + \frac{M+N}{\sqrt{t}}\right]$. We present the corresponding extension and investigate it in details: we show how the algorithm can be "tuned" to the geometry of the s.v.i. in question, derive bounds for the probability of large deviations of the resulting error, etc. 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. Some general results obout the performance of the SMP are presented in Section 3.2. Then in Section 4 we present SMP for Stochastic composite minimization and discuss its applications to Stochastic Semidefinite Feasability 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 (1.5).

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.

2.1. Nash v.i.'s and functional error. In the sequel, we shall be especially interested in a special case of v.i. (1.2) – 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, i-th of them choosing a point z_i from a given set Z_i. The loss of i-th 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 player's 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 i-th 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., [8]) that setting

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

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. (1.2) 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.

Special cases: saddle points and minimization. An important by its own right particular case of Nash Equilibrium problem is an *antagonistic 2-person 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

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

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:

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

Finally, the "trivial" case $m = 1$ of the convex Nash Equilibrium is the problem of minimizing a Lipschitz continuous convex function $\phi(z) = \phi_1(z_1)$ over the convex compact set $Z = Z_1$, In this case, the functional error becomes the usual residual in terms of the objective:

$$\mathrm{Err}_{\mathrm{N}}(z) = \phi(z) - \min_{Z} \phi.$$

In the sequel, we refer to the v.i. (1.2) coming from a convex Nash Equilibrium problem as *Nash* v.i., and to the two just outlined particular cases of the Nash v.i. as the *Saddle Point* and the *Minimization* v.i., respectively. It is easy to verify that in the Saddle Point/Minimization case the functional error $\mathrm{Err}_{\mathrm{N}}(z)$ is $\geqslant \mathrm{Err}_{\mathrm{vi}}(z)$; this is not necessary so for a general Nash v.i.

2.2. Prox-mapping. We once for ever fix a norm $\|\cdot\|$ on $\mathcal{E}$; $\|\cdot\|_*$ stands for the conjugate norm, see (1.5). A *distance-generating function* for Z is, by definition, a continuous convex function $\omega(\cdot) : Z \to \mathbf{R}$ such that

1. if Z^o be the set of all points $z \in Z$ such that the subdifferential $\partial\omega(z)$ of $\omega(\cdot)$ at z is nonempty, then the subdifferential of ω 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 $z \in Z^o$;
2. for certain $\alpha > 0$, $\omega(\cdot)$ is strongly convex, modulus α, w.r.t. the norm $\|\cdot\|$:

$$\forall(z, z' \in Z^o) : \langle\omega'(z) - \omega'(z'), z - z'\rangle \geqslant \alpha\|z - z'\|^2. \tag{2.1}$$

In the sequel, we fix a distance-generating function $\omega(\cdot)$ for Z and assume that $\omega(\cdot)$ and Z "fit" each other, meaning that one can easily solve problems of the form

$$\min_{z\in Z} \left[\omega(z) + \langle e, z\rangle\right], \quad e \in \mathcal{E}. \tag{2.2}$$

The *prox-function* associated with the distance-generating function ω 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

$$(2.3)\quad \begin{array}{llllllll} (a) & \Theta(z) & = & \max_{u\in Z} V(z,u) & [z \in Z^o]; & (b) & z_{\rm c} & = & \operatorname{argmin}_Z \omega(z); \\ (c) & \Theta & = & \Theta(z_{\rm c}); & & (d) & \Omega & = & \sqrt{2\Theta/\alpha}. \end{array}$$

Note that $z_{\rm c}$ 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 $0 \in \partial\omega(z_{\rm c})$). Note also that due to the strong convexity of ω and the origin of $z_{\rm c}$ we have

$$(2.4)\qquad \forall (u \in Z) : \frac{\alpha}{2}\|u - z_{\rm c}\|^2 \leqslant \Theta \leqslant \max_{z\in Z} \omega(z) - \omega(z_{\rm c});$$

in particular we see that

$$(2.5)\qquad Z \subset \{z : \|z - z_{\rm c}\| \leqslant \Omega\}.$$

Prox-mapping. Given $z \in Z^o$, we associate with this point and $\omega(\cdot)$ the *prox-mapping*

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

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., $Z^o = Z$, $\alpha = 1$). Assuming for the sake of simplicity that $0 \in Z$, $z_{\rm c} = 0$, $\Omega = \max_{z\in Z}\|z\|_2$ and $\Theta = \frac{1}{2}\Omega^2$. 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

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

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

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

It is easily seen (see, e.g., [3]) that here

$$\alpha = 1, \ \ z_{\rm c} = [1/N; ...; 1/N], \ \ \Theta \leqslant \ln(N)$$

(the latter inequality becomes equality when Z contains a vertex of $\mathcal{D}_N$), and thus $\Omega \leqslant \sqrt{2\ln N}$. The prox-function is

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

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

$$(P(z,\xi))_j = \left(\sum_{i=1}^N z_i \exp\{-\xi_i\}\right)^{-1} z_j \exp\{-\xi_j\}.$$

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 $\lambda_1(a) \geqslant ... \geqslant \lambda_N(a)$ 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 distance-generating function is the matrix entropy

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

so that $Z^o = \{z \in Z : z \succ 0\}$ and $\Omega'(z) = \ln(z)$. This setup, similarly to the Simplex one, results in $\alpha = 1$, $z_c = N^{-1}I_N$, $\Theta = \ln N$ and $\Omega = \sqrt{2 \ln N}$ [2]. When $Z = \mathcal{S}$, it is relatively easy to compute the prox-mapping (see [2, 6]); 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 $P(z, \xi) = \mathcal{H}(a - \xi)$.

3. Stochastic Mirror-Prox algorithm.

3.1. Mirror-Prox algorithm with erroneous information. We are about to present the Mirror-Prox algorithm proposed in [6]. 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 $\widehat{F}(z) \in \mathcal{E}$ of $F(z)$. The t-step Mirror-Prox algorithm as applied to (1.2) is as follows:

ALGORITHM 3.1.

1. *Initialization: Choose $r_0 \in Z^o$ and stepsizes $\gamma_\tau > 0$, $1 \leqslant \tau \leqslant t$.*
2. *Step τ, $\tau = 1, 2, ..., t$: Given $r_{\tau-1} \in Z^o$, set*

$$\left\{ \begin{array}{rcl} w_\tau & = & P(r_{\tau-1}, \gamma_\tau \widehat{F}(r_{\tau-1})), \\ r_\tau & = & P(r_{\tau-1}, \gamma_\tau \widehat{F}(w_\tau)) \end{array} \right. \tag{3.1}$$

   . When $\tau < t$, loop to step $t + 1$.
3. *At step t, output*

$$\widehat{z}_t = \left[\sum_{\tau=1}^t \gamma_\tau\right]^{-1} \sum_{\tau=1}^t \gamma_\tau w_\tau. \tag{3.2}$$

The preliminary technical result on the outlined algorithm is as follows.

THEOREM 3.2. *Consider t-step algorithm 3.1 as applied to a v.i.* (1.2) *with a monotone operator F satisfying (1.4). 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*

$$y_\tau = P(y_{\tau-1}, \gamma_\tau \Delta_\tau), \quad y_0 = r_0. \tag{3.3}$$

Assume that

$$\gamma_\tau \leqslant \frac{\alpha}{\sqrt{3}L}, \tag{3.4}$$

Then

$$\mathrm{Err}_{\mathrm{vi}}(\widehat{z}_t) \leqslant \left(\sum_{\tau=1}^t \gamma_\tau\right)^{-1} \Gamma(t), \tag{3.5}$$

where $\mathrm{Err}_{\mathrm{vi}}(\widehat{z}_t)$ *is defined in* (1.3),

$$\begin{aligned}\Gamma(t) = 2\Theta(r_0) &+ \sum_{\tau=1}^t \frac{3\gamma_\tau^2}{2\alpha}\left[M^2 + (\epsilon_{r_{\tau-1}} + \epsilon_{w_\tau})^2 + \frac{\epsilon_{w_\tau}^2}{3}\right] \\ &+ \sum_{\tau=1}^t \langle \gamma_\tau \Delta_\tau, w_\tau - y_{\tau-1}\rangle\end{aligned} \tag{3.6}$$

and $\Theta(\cdot)$ *is defined by* (2.3).

Finally, when (1.2) *is a Nash v.i., one can replace* $\mathrm{Err}_{\mathrm{vi}}(\widehat{z}_t)$ *in* (3.5) *with* $\mathrm{Err}_{\mathrm{N}}(\widehat{z}_t)$.

3.2. Main result. From now on, we focus on the case when Algorithm 3.1 solves monotone v.i. (1.2), and the corresponding monotone operator F is represented by a stochastic oracle. Specifically, at the i-th 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. We refer to this specific implementation of Algorithm 3.1 as to *Stocastic Mirror Prox* (SMP) algorithm.

In the sequel, we impose on the SO in question the following assumption, slightly milder than (1.6):

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

$$\begin{array}{ll}(a) & \|\mathbf{E}\{\Xi(z,\zeta_i) - F(z)\}\|_* \leqslant \mu \\ (b) & \mathbf{E}\left\{\|\Xi(z,\zeta_i) - F(z)\|_*^2\right\} \leqslant M^2.\end{array} \tag{3.7}$$

In some cases, we augment Assumption I by the following

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

$$\mathbf{E}\left\{\exp\{\|\Xi(z,\zeta_i) - F(z)\|_*^2/M^2\}\right\} \leqslant \exp\{1\}. \tag{3.8}$$

Note that Assumption II implies (3.7.b), *since*

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

by the Jensen inequality.

REMARK 3.3. *Observe that that the accuracy of Algorithm 3.1 (cf. (3.6)) depends in the same way on the "size" of perturbation* $\epsilon_z = \|\widehat{F}(z) - F(z)\|_*$ *and the bound* M *of (1.4) on the variation of the non-Lipschitz component of* F. *This is why, to simplify*

the presentation, we decided to use the same bound M for the scale of perturbation $\Xi(z,\zeta_i)-F(z)$ in (3.7), (3.8).

REMARK 3.4. *From now on, we assume that the starting point r_0 in Algorithm 3.1 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 $\gamma_\tau \equiv \gamma$, $1 \leqslant \tau \leqslant t$, where t is a fixed in advance number of iterations of the algorithm.*

Our main result is as follows:

THEOREM 3.5. *Let v.i. (1.2) with monotone operator F satisfying (1.4) be solved by t-step Algorithm 3.1 using a SO, and let the stepsizes $\gamma_\tau \equiv \gamma$, $1 \leqslant \tau \leqslant t$, satisfy $0 < \gamma \leqslant \frac{\alpha}{\sqrt{3}L}$, see (1.4). Then*

(i) *Under Assumption I, one has*

$$\mathbf{E}\left\{\mathrm{Err}_{\mathrm{vi}}(\widehat{z}_t)\right\} \leqslant K_0(t) \equiv \left[\frac{\alpha\Omega^2}{t\gamma} + \frac{21M^2\gamma}{2\alpha}\right] + 2\mu\Omega, \tag{3.9}$$

where M is the constant from (1.4) and Ω is given by (2.3).

(ii) *Under Assumptions I, II, one has, in addition to (3.9), for any $\Lambda > 0$,*

$$\mathrm{Prob}\left\{\mathrm{Err}_{\mathrm{vi}}(\widehat{z}_t) > K_0(t) + \Lambda K_1(t)\right\} \leqslant \exp\{-\Lambda^2/3\} + \exp\{-\Lambda t\}, \tag{3.10}$$

where

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

In the case of a Nash v.i., $\mathrm{Err}_{\mathrm{vi}}(\cdot)$ in (3.9), (3.10) can be replaced with $\mathrm{Err}_{\mathrm{N}}(\cdot)$.

When optimizing the bound (3.9) in γ, we get the following

COROLLARY 3.6. *In the situation of Theorem 3.5, let the stepsizes $\gamma_\tau \equiv \gamma$ be chosen according to*

$$\gamma = \min\left[\frac{\alpha}{\sqrt{3}L}, \frac{\alpha\Omega}{M}\sqrt{\frac{2}{21t}}\right]. \tag{3.11}$$

Then under Assumption I one has

$$\mathbf{E}\left\{\mathrm{Err}_{\mathrm{vi}}(\widehat{z}_t\right\} \leqslant K_0^*(t) \equiv \max\left[\tfrac{7}{4}\tfrac{\Omega^2 L}{t} + 7\tfrac{\Omega M}{\sqrt{t}}\right] + 2\mu\Omega,. \tag{3.12}$$

(see (2.3)). Under Assumptions I, II, one has, in addition to (3.12), for any $\Lambda > 0$,

$$\mathrm{Prob}\left\{\mathrm{Err}_{\mathrm{vi}}(\widehat{z}_t) > K_0^*(t) + \Lambda K_1^*(t)\right\} \leqslant \exp\{-\Lambda^2/3\} + \exp\{-\Lambda t\} \tag{3.13}$$

with

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

In the case of a Nash v.i., $\mathrm{Err}_{\mathrm{vi}}(\cdot)$ in (3.12), (3.13) can be replaced with $\mathrm{Err}_{\mathrm{N}}(\cdot)$.

3.3. Comparison with Robust Mirror SA Algorithm. Consider the case of a Nash s.v.i. with operator F satisfying (1.4) with $L = 0$, and let the SO be unbiased (i.e., $\mu = 0$). In this case, the bound (3.12) reads

$$\mathbf{E}\left\{\mathrm{Err}_{\mathrm{N}}(\widehat{z}_t\right\} \leqslant \frac{7\Omega M}{\sqrt{t}}, \tag{3.14}$$

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 (3.14) looks very much like the efficiency estimate

$$\mathbf{E}\left\{\mathrm{Err}_{\mathrm{N}}(\tilde{z}_t)\right\} \leqslant O(1)\frac{\Omega\overline{M}}{\sqrt{t}} \tag{3.15}$$

(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 [3][1)]. In the latter estimate, Ω is exactly the same as in (3.14), 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 $M \leqslant 2\overline{M}$, 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 (3.14) never is worse, and sometimes can be much better than the SA bound (3.15). 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 3.1 with the description

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

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 Approximation: Stochastic composite minimization.

4.1. Problem description. Consider the optimization problem as follows (cf. [5]):

$$\min_{x\in X}\phi(x) := \Phi(\phi_1(x),...,\phi_m(x)), \tag{4.1}$$

where

[1)] In this reference, only the Minimization and the Saddle Point problems are considered. However, the results of [3] can be easily extended to s.v.i.'s.

1. $X \subset \mathcal{X}$ is a convex compact; the embedding space $\mathcal{X}$ is equipped with a norm $\|\cdot\|_x$, and X – with a distance-generating function $\omega_x(x)$ with certain parameters $\alpha_x, \Theta_x, \Omega_x$ w.r.t. the norm $\|\cdot\|_x$;
2. $\phi_\ell(x) : X \to \mathcal{E}_\ell$, $1 \leqslant \ell \leqslant m$, are Lipschitz continuous mappings taking values in Euclidean spaces $\mathcal{E}_\ell$ equipped with norms (not necessarily the Euclidean ones) $\|\cdot\|_{(\ell)}$ with conjugates $\|\cdot\|_{(\ell,*)}$ and with closed convex cones K_ℓ. We suppose that ϕ_ℓ are K_ℓ-convex, i.e. for any $x, x' \in X,\ \lambda \in [0,1]$,

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

where the notation $a \leqslant_K b \Leftrightarrow b \geqslant_K a$ means that $b - a \in K$.
In addition to these structural restrictions, we assume that for all $v, v' \in X,\ h \in \mathcal{X}$,

$$\begin{array}{ll} (a) & \|[\phi'_\ell(v) - \phi'_\ell(v')]h\|_{(\ell)} \leqslant [L_x\|v - v'\|_x + M_x]\|h\|_x \\ (b) & \|[\phi'_\ell(v)]h\|_{(\ell)} \leqslant [L_x\Omega_x + M_x]\|h\|_x \end{array} \tag{4.2}$$

for certain selections $\phi'_\ell(x) \in \partial^{K_\ell}\phi_\ell(x)$, $x \in X$[2)] and certain nonnegative constants L_x and M_x.
3. Functions $\phi_\ell(\cdot)$ are represented by an unbiased SO. At i-th 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 $\mathcal{E}_\ell$, $1 \leqslant \ell \leqslant m$ ($\{\zeta_i\}$ are i.i.d. random vectors) such that for any $x \in X$ and $i = 1, 2, ...,$

$$\begin{array}{ll} (a) & \mathbf{E}\left\{f_\ell(x, \zeta_i)\right\} = \phi_\ell(x),\ 1 \leqslant \ell \leqslant m \\ (b) & \mathbf{E}\left\{\max\limits_{1 \leqslant \ell \leqslant m} \|f_\ell(x, \zeta_i) - \phi_\ell(x)\|^2_{(\ell)}\right\} \leqslant M_x^2\Omega_x^2; \\ (c) & \mathbf{E}\left\{\mathbf{G}_\ell(x, \zeta_i)\right\} = \phi'_\ell(x),\ 1 \leqslant \ell \leqslant m, \\ (d) & \mathbf{E}\left\{\max\limits_{\substack{h \in \mathcal{X} \\ \|h\|_x \leqslant 1}} \|[\mathbf{G}_\ell(x, \zeta_i) - \phi'_\ell(x)]h\|^2_{(\ell)}\right\} \leqslant M_x^2,\ 1 \leqslant \ell \leqslant m. \end{array} \tag{4.3}$$

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

$$\Phi(u_1, ..., u_m) = \max_{y \in Y}\left\{\sum_{\ell=1}^m \langle u_\ell, \mathbf{A}_\ell y + b_\ell\rangle_{\mathcal{E}_\ell} - \Phi_*(y)\right\} \tag{4.4}$$

for $u_\ell \in \mathcal{E}_\ell$, $1 \leqslant \ell \leqslant m$. Here
   (a) $Y \subset \mathcal{Y}$ is a convex compact set containing the origin; the embedding Euclidean space $\mathcal{Y}$ is equipped with a norm $\|\cdot\|_y$, and Y - with a distance-generating function $\omega_y(y)$ with parameters $\alpha_y, \Theta_y, \Omega_y$ w.r.t. the norm $\|\cdot\|_y$;
   (b) 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; here K_ℓ^* is the cone dual to K_ℓ;

[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 $\phi(u) \geqslant_K \phi(x) + \mathbf{P}(u - x)$ 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 $\frac{\partial\phi(x)}{\partial x} \in \partial^K\phi(x)$.

(c) $\Phi_*(y)$ is a given convex function on Y such that

$$\|\Phi'_*(y) - \Phi'_*(y')\|_{y,*} \leqslant L_y\|y - y'\|_y + M_y \tag{4.5}$$

for certain selection $\Phi'_*(z) \in \partial\Phi_*(y)$, $y \in Y$.

Example: Stochastic Matrix Minimax problem (SMMP). For $1 \leqslant \ell \leqslant m$, 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 the spectral norms $|\cdot|_\infty$, and let K_ℓ be the cone $\mathbf{S}^{p_\ell}_+$ of symmetric positive semidefinite $p_\ell \times p_\ell$ matrices. Consider the problem

$$\min_{x\in X} \max_{1\leqslant j\leqslant k} \lambda_{\max}\left(\sum_{\ell=1}^m P_{j\ell}^T \phi_\ell(x) P_{j\ell}\right), \tag{P}$$

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 $p \times q$ 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 (4.1), (4.4) with

$$\begin{aligned}
\Phi(u) &:= \max_{1\leqslant j\leqslant k} \lambda_{\max}\left(\sum_{\ell=1}^m P_{j\ell} u_\ell P_{j\ell}^T\right) \\
&= \max_{y=\mathrm{Diag}\{y_1,...,y_k\}\in Y} \sum_{j=1}^k \mathrm{Tr}\left(\sum_{\ell=1}^m P_{j\ell}^T u_\ell P_{j\ell} y_k\right) \\
&= \max_{y=\mathrm{Diag}\{y_1,...,y_k\}\in Y} \sum_{\ell=1}^m \mathrm{Tr}\left(u_\ell\left[\sum_{j=1}^k P_{j\ell}^T y_k P_{j\ell}\right]\right) \\
&= \max_{y=\mathrm{Diag}\{y_1,...,y_k\}\in Y} \sum_{\ell=1}^m \langle u_\ell, \mathbf{A}_\ell y\rangle_F
\end{aligned}$$

(we put $\mathbf{A}_\ell y = \sum_{j=1}^k P_{j\ell}^T y_j P_{j\ell}$). The set Y is the spectahedron in the space $\mathbf{S}^{\mathbf{q}}$ of symmetric block-diagonal matrices with k diagonal blocks of the sizes $q_j \times q_j$, $1 \leqslant j \leqslant k$. When equipping Y with the spectahedron setup, we get $\alpha_y = 1$, $\Theta_y = \ln(\sum_{j=1}^k q_j)$ and $\Omega_y = \sqrt{2\ln(\sum_{j=1}^k q_j)}$, see Section 2.2.

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

$$\min_{x\in X}\left[\max_{1\leqslant\ell\leqslant m} \lambda_{\max}(\phi_\ell(x))\right]. \tag{4.6}$$

If, in addition, $p_j = q_j = 1$ for all j, we arrive at the usual ("scalar") minimax problem

$$\min_{x\in X}\left[\max_{1\leqslant\ell\leqslant m} \phi_\ell(x)\right] \tag{4.7}$$

with convex real-valued functions ϕ_ℓ.

Observe that in the case of (4.4), the optimization problem (4.1) is nothing but the primal problem associated with the saddle point problem

$$\min_{x\in X}\max_{y\in Y}\left[\phi(x,y)=\sum_{\ell=1}^m\langle\phi_\ell(x),\mathbf{A}_\ell y+b_\ell\rangle_{\mathcal{E}_\ell}-\Phi_*(y)\right] \tag{4.8}$$

and the cost function in the latter problem is Lipschitz continuous and convex-concave due to the 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 and the monotone mapping

$$F(z)\equiv F(x,y)=\left[\sum_{\ell=1}^m[\phi_\ell'(x)]^*[\mathbf{A}_\ell y+b_\ell];\ -\sum_{\ell=1}^m\mathbf{A}_\ell^*\phi_\ell(x)+\Phi_*'(y)\right]. \tag{4.9}$$

The advantage of the v.i. reformulation of (4.1) is that F is linear in $\phi_\ell(\cdot)$, so that the initial unbiased SO for ϕ_ℓ induces an unbiased stochastic oracle for F, specifically, the oracle

$$\Xi(x,y,\zeta_i)=\left[\sum_{\ell=1}^m\mathbf{G}_\ell^*(x,\zeta_i)[\mathbf{A}_\ell y+b_\ell];\ -\sum_{\ell=1}^m\mathbf{A}_\ell^*f_\ell(x,\zeta_i)+\Phi_*'(y)\right]. \tag{4.10}$$

We are about to use this oracle in order to solve the stochastic composite minimization problem (4.1) by the SMP algorithm.

4.2. Setup for the SMP as applied to (4.9)**.** In retrospect, the setup for SMP we are about to present is a kind of the best – resulting in the best possible efficiency estimate (3.12) – we can build from the entities participating in the description of the problem (4.1). Specifically, we equip the space $\mathcal{E}=\mathcal{X}\times\mathcal{Y}$ with the norm

$$\|[(x,y)\|\equiv\sqrt{\|x\|_x^2/\Omega_x^2+\|y\|_y^2/\Omega_y^2};$$

the conjugate norm clearly is

$$\|(\xi,\eta)\|_*=\sqrt{\Omega_x^2\|\xi\|_{x,*}^2+\Omega_y^2\|\eta\|_{y,*}^2}.$$

Finally, we equip $Z=X\times Y$ with the distance-generating function

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

The SMP-related properties of our setup are summarized in the following

LEMMA 4.1. *Let*

$$\mathcal{A}=\max_{y\in\mathcal{Y}:\|y\|_y\leqslant1}\sum_{\ell=1}^m\|\mathbf{A}_\ell y\|_{(\ell,*)},\ \mathcal{B}=\sum_{\ell=1}^m\|b_\ell\|_{(\ell,*)}. \tag{4.11}$$

(i) *The parameters of the just defined distance-generating function ω w.r.t. the just defined norm $\|\cdot\|$ are $\alpha=1$, $\Theta=1$, $\Omega=\sqrt{2}$.*

(ii) *One has*

$$\forall(z,z'\in Z):\|F(z)-F(z')\|_*\leqslant L\|z-z'\|+M, \tag{4.12}$$

where

$$L = 5\mathcal{A}\Omega_x\Omega_y\left[\Omega_x L_x + M_x\right] + \mathcal{B}\Omega_x^2 L_x + \Omega_y^2 L_y$$
$$M = \left[2\mathcal{A}\Omega_y + \|b\|_1\right]\Omega_x M_x + \Omega_y M_y$$

Besides this,

$$\forall(z \in Z, i) : \mathbf{E}\left\{\Xi(z,\zeta_i)\right\} = F(z); \quad \mathbf{E}\left\{\|\Xi(z,\zeta_i) - F(z)\|_*^2\right\} \leqslant M^2. \tag{4.13}$$

Furethermore, if relations (4.3.*b*,*d*) *are strengthened to*

$$\begin{array}{l}\mathbf{E}\left\{\exp\left\{\max\limits_{1\leqslant\ell\leqslant m}\|f_\ell(x,\zeta_i) - \phi_\ell(x)\|_{(\ell)}^2/(\Omega_x M)^2\right\}\right\} \leqslant \exp\{1\},\\ \mathbf{E}\left\{\exp\left\{\max\limits_{\substack{h\in\mathcal{X},\\ \|h\|_x\leqslant 1}}\|[\mathbf{G}_\ell(x) - \phi_\ell'(x)]h\|_{(\ell)}^2/M^2\right\}\right\} \leqslant \exp\{1\},\ 1 \leqslant \ell \leqslant m,\end{array} \tag{4.14}$$

then

$$\mathbf{E}\left\{\exp\{\|\Xi(z,\zeta_i) - F(z)\|_*^2/M^2\}\right\} \leqslant \exp\{1\}. \tag{4.15}$$

Combining Lemma 4.1 with Corollary (3.6) we get explicit efficiency estimates for the SMP algorithm as applied to the Stochastic composite minimization problem (4.1).

4.3. Application to Stochastic Semidefinite Feasibility problem. Assume we are interested to solve a feasible system of matrix inequalities

$$\psi_\ell(x) \preceq 0,\ \ell = 1,...,m\ \&\ x \in X, \tag{4.16}$$

where $m > 1$, $X \subset \mathcal{X}$ is as in the description of the Stochastic composite problem, and $\psi_\ell(\cdot)$ take values in the spaces $\mathcal{E}_\ell = \mathbf{S}^{p_\ell}$ of symmetric $p_\ell \times p_\ell$ matrices. We equip $\mathcal{E}_\ell$ with the Frobenius inner product, the semidefinite cone $K_\ell = \mathbf{S}_+^{p_\ell}$ and the spectral norm $\|\cdot\|_{(\ell)} = |\cdot|_\infty$ (recall that $|A|_\infty$ is the maximal singular value of matrix A). We assume that ψ_ℓ are Lipschitz continuous and $K_\ell = \mathbf{S}_+^{p_\ell}$-convex functions on X such that for all $x, x' \in X$ and for all ℓ one has

$$\begin{array}{l}\max\limits_{h\in\mathcal{X},\,\|h\|_x\leqslant 1}|[\psi_\ell'(x) - \psi_\ell'(x')]h|_\infty \leqslant L_\ell\|x - x'\|_{(\ell)} + M_\ell,\\ \max\limits_{h\in\mathcal{X},\,\|h\|_x\leqslant 1}|\psi_\ell'(x)h|_\infty \leqslant L_\ell\Omega_x + M_\ell\end{array} \tag{4.17}$$

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.

We assume that $\psi_\ell(\cdot)$ are represented by an SO which at i-th 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 $\mathcal{E}_\ell$ such that for all $x \in X$ it holds

$$\begin{array}{ll}(a) & \mathbf{E}\left\{\widehat{f}_\ell(x,\zeta_i)\right\} = \psi_\ell(x),\ \ \mathbf{E}\left\{\widehat{\mathbf{G}}_\ell(x,\zeta_i)\right\} = \psi_\ell'(x),\ 1 \leqslant \ell \leqslant m\\ (b) & \mathbf{E}\left\{\max\limits_{1\leqslant\ell\leqslant m}|\widehat{f}_\ell(x,\zeta_i) - \psi_\ell(x)|_\infty^2/(\Omega_x M_\ell)^2\right\} \leqslant 1\\ (c) & \mathbf{E}\left\{\max\limits_{\substack{h\in\mathcal{X},\\ \|h\|_x\leqslant 1}}|[\widehat{\mathbf{G}}_\ell(x,\zeta_i) - \psi_\ell'(x)]h|_\infty^2/M_\ell^2\right\} \leqslant 1,\ 1 \leqslant \ell \leqslant m.\end{array} \tag{4.18}$$

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

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

$$\mu = \max_{1 \leqslant \ell \leqslant m} \mu_\ell, \quad \beta_\ell = \frac{\mu}{\mu_\ell}, \phi_\ell(\cdot) = \beta_\ell \psi_\ell(\cdot), \quad L_x = \frac{\mu\sqrt{t}}{\Omega_x}, \quad M_x = \mu. \tag{4.19}$$

Note that by construction $\beta_\ell \geqslant 1$ and $L_x/L_\ell \geqslant \beta_\ell$, $M_x/M_\ell \geqslant \beta_\ell$ for all ℓ, so that the functions ϕ_ℓ satisfy (4.2) 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

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

By (4.18), this oracle satisfies (4.3).

B. We then build the Stochastic Matrix Minimax problem

$$\min_{x \in X} \max_{1 \leqslant \ell \leqslant m} \lambda_{\max}(\phi_i(x)), \tag{4.20}$$

associated with the just defined $\phi_1, ..., \phi_m$, that is, the Stochastic composite problem (4.1) associated with $\phi_1, ..., \phi_m$ and the outer function

$$\begin{aligned}\Phi(u_1, ..., u_m) = \max_{1 \leqslant \ell \leqslant m} \lambda_{\max}(u_\ell) &= \max_{y \in Y} \sum_{\ell=1}^m \langle u_\ell, y_\ell \rangle_F, \\ Y = \{y = \mathrm{Diag}\{y_1, ..., y_m\} \in \mathcal{Y} = \mathbf{S}^{p_1} \times ... \times \mathbf{S}^{p_m} &: y \succeq 0, \ \mathrm{Tr}(y) = 1\} \\ \subset \mathcal{Y} &= \mathbf{S}^{p_1} \times ... \times \mathbf{S}^{p_m}.\end{aligned}$$

Thus in the notation from (4.4) we have $\mathbf{A}_\ell y = y_\ell$, $b_\ell = 0$, $\Phi_* \equiv 0$. Hence $L_x = M_x = 0$, and Y is a spectahedron. We equip $\mathcal{Y}$ and Y with the Spectahedron setup, arriving at

$$\alpha_y = 1, \quad \Theta_y = \ln \sum_{\ell=1}^m p_\ell, \quad \Omega_y = \sqrt{2 \ln \sum_{\ell=1}^m p_\ell}.$$

C. We have specified all entities participating in the description of the Stochastic composite problem. It is immediately seen that these entities satisfy all conditions of Section 4.1. We can now solve the resulting Stochastic composite problem by t-step SMP algorithm with the setup presented in Section 4.2. The corresponding convex-concave saddle point problem is

$$\min_{x \in X} \max_{y \in Y} \sum_{\ell=1}^m \beta_\ell \langle \psi_\ell(x), y_\ell \rangle_F;$$

with the monotone operator and SO, respectively,

$$\begin{aligned}F(z) \equiv F(x, y) &= \left[\sum_{\ell=1}^m \beta_\ell [\psi_\ell'(x)]^* y_\ell; -\mathrm{Diag}\left\{\alpha_1 \psi_1(x), ..., \alpha_m \psi_m(x)\right\}\right], \\ \Xi((x, y), \zeta) &= \left[\sum_{\ell=1}^m \beta_\ell \widehat{\mathbf{G}}_\ell^*(x, \zeta) y_\ell; -\mathrm{Diag}\left\{\alpha_1 \widehat{f}_1(x, \zeta), ..., \alpha_m \widehat{f}_m(x, \zeta)\right\}\right].\end{aligned}$$

Combining Lemma 4.1, Corollary 3.6 and taking into account the origin of the quantities L_x, M_x, and that $\mathcal{A} = 1$, $\mathcal{B} = 0$ [3], we arrive at the following result:

[3] See (4.11) and note that we are in the case when $b_\ell = 0$ and $\|\cdot\|_{(\ell,*)}$ is the trace norm; thus, $\sum_{\ell=1}^m \|\mathbf{A}_\ell y\|_{(\ell,*)} = \sum_{\ell=1}^m |y_\ell|_1 = |y|_1 = \|y\|_y$.

PROPOSITION 4.2. *With the outlined construction, the resulting s.v.i. reads*

$$\text{find } z_* \in Z = X \times Y : \langle F(z), z - z_* \rangle \geqslant 0 \ \ \forall z \in Z, \tag{4.21}$$

for the monotone operator F which satisfies (1.4) *with*

$$L = 10 \left[\ln \sum_{\ell=1}^{m} p_\ell \right]^{\frac{1}{2}} \Omega_x \mu(\sqrt{t}+1), \ M = 4 \left[\ln \sum_{\ell=1}^{m} p_\ell \right]^{\frac{1}{2}} \Omega_x \mu;$$

Beside this, the resulting SO for F satisfies (4.13) *with the just defined value of M.*

Let now

$$\gamma = \left(10 \left[3 \ln \sum_{\ell=1}^{m} p_\ell \right]^{\frac{1}{2}} \Omega_x \mu(\sqrt{t}+1) \right)^{-1}, \ 1 \leqslant \tau \leqslant t.$$

When applying to (4.21) *the t-step SMP algorithm with the constant stepsizes $\gamma_\tau \equiv \gamma$ (cf.* (3.11) *and note that we are in the situation $\alpha = \Theta = 1$), we get an approximate solution $\widehat{z}_t = (\widehat{x}_t, \widehat{y}_t)$ such that*

$$\mathbf{E} \left\{ \max_{1 \leqslant \ell \leqslant m} \beta_\ell \lambda_{\max}(\psi_\ell(\widehat{x}_t)) \right\} \leqslant 80 \frac{\Omega_x \left[\ln \sum_{\ell=1}^{m} p_\ell \right]^{\frac{1}{2}} \mu}{\sqrt{t}} \tag{4.22}$$

(cf. (3.12) *and take into account that we are in the case of $\Omega = \sqrt{2}$, while the optimal value in* (4.20) *is nonpositive, since* (4.16) *is feasible).*

Furthermore, if assumptions (4.18.*b*,*c*) *are strengthened to*

$$\mathbf{E} \left\{ \max_{1 \leqslant \ell \leqslant m} \exp\{ |\widehat{f}_\ell(x, \zeta_i) - \psi_\ell(x)|_\infty^2 / (\Omega_x M_\ell)^2 \} \right\} \leqslant \exp\{1\},$$
$$\mathbf{E} \left\{ \exp\{ \max_{h \in \mathcal{X}, \, \|h\|_x \leqslant 1} |[\widehat{\mathbf{G}}_\ell(x, \zeta_i) - \psi'_\ell(x)]h|_\infty^2 / M_\ell^2 \} \right\} \leqslant \exp\{1\}, \ \ 1 \leqslant \ell \leqslant m,$$

then, in addition to (4.22), *we have for any $\Lambda > 0$:*

$$\begin{aligned} \mathrm{Prob} \Bigg\{ & \max_{1 \leqslant \ell \leqslant m} \beta_\ell \lambda_{\max}(\psi_\ell(\widehat{x}_t)) > 80 \frac{\Omega_x \left[\ln \sum_{\ell=1}^{m} p_\ell \right]^{\frac{1}{2}} \mu}{\sqrt{t}} + \Lambda \frac{15 \left[\ln \sum_{\ell=1}^{m} p_\ell \right]^{\frac{1}{2}} \mu}{\sqrt{t}} \Bigg\} \\ & \leqslant \exp\{-\Lambda^2/3\} + \exp\{-\Lambda t\}. \end{aligned}$$

Discussion. Imagine that instead of solving the system of matrix inequalities (4.16), 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

$$\begin{aligned} \mathbf{E} \left\{ \psi_\ell(\widehat{x}_t^\ell) \right\} & \leqslant O(1) \left[\ln(p_\ell + 1) \right]^{1/2} \Omega_x \left[\frac{\Omega_x L_\ell}{t} + \frac{M_\ell}{\sqrt{t}} \right] = O(1) \left[\ln(p_\ell + 1) \right]^{1/2} \Omega_x \frac{\mu_\ell}{\sqrt{t}} \\ & = O(1) \left[\ln(p_\ell + 1) \right]^{1/2} \beta_\ell^{-1} \frac{\Omega_x \mu}{\sqrt{t}}, \end{aligned}$$

(recall that the matrix inequality in question is feasible), where $\widehat{x}_t^\ell$ is the resulting approximate solution. Looking at (4.22), we see that the expected accuracy of the SMP as applied, in the aforementioned manner, to (4.16) is only by a logarithmic in $\sum_\ell p_\ell$ factor worse:

$$\mathbf{E}\left\{\psi_\ell(\widehat{x}_t)\right\} \leqslant O(1)\left[\ln\sum_{\ell=1}^m p_\ell\right]^{1/2}\beta_\ell^{-1}\frac{\Omega_x\mu}{\sqrt{t}} = O(1)\left[\ln\sum_{\ell=1}^m p_\ell\right]^{1/2}\frac{\Omega_x\mu_\ell}{\sqrt{t}}. \tag{4.23}$$

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 ($L_\ell = 0$). In this case, (4.23) reads

$$\mathbf{E}\left\{\psi_\ell(\widehat{x}_t)\right\} \leqslant \begin{cases} O(1)\left[\ln\sum_{\ell=1}^m p_\ell\right]^{1/2}\frac{\Omega_x^2 L_\ell}{t}, & \psi_\ell \text{ is easy,} \\ O(1)\left[\ln\sum_{\ell=1}^m p_\ell\right]^{1/2}\frac{\Omega_x M_\ell}{\sqrt{t}}, & \psi_\ell \text{ is difficult.} \end{cases}$$

In other words, the violations of the easy and the difficult constraints in (4.16) 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

$$\begin{array}{rcl} \mathrm{Opt} & = & \min\limits_{x\in X} f(x) := \lambda_{\max}(A_0 + x_1A_1 + ... + x_nA_n), \\ X & = & \{x \in \mathbf{R}^n : x \geqslant 0, \sum_{i=1}^n x_i = 1\}, \end{array} \tag{4.24}$$

where $A_0, A_1, ..., A_n$, $n > 1$, are given symmetric matrices with common block-diagonal structure $(p_1, ..., p_m)$. I.e., all A_j are block-diagonal with diagonal blocks A_j^ℓ of sizes $p_\ell \times p_\ell$, $1 \leqslant \ell \leqslant m$. We denote

$$p^{(\kappa)} = \sum_{\ell=1}^m p_\ell^\kappa, \;\; \kappa = 1, 2, 3; \;\; p^{\max} = \max_\ell p_\ell.$$

Setting

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

we represent (4.24) as a particular case of the Matrix Minimax problem (4.6), 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; these latter methods 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 from [6] to the saddle point reformulation (4.8) of our specific Matrix Minimax problem, i.e., to the saddle point problem

$$\begin{array}{c}\min\limits_{x\in X}\max\limits_{y\in Y}\langle y, A_0+\sum_{j=1}^n x_jA_j\rangle_F,\\ Y=\left\{y=\mathrm{Diag}\{y_1,...,y_m\}: y_\ell\in\mathbf{S}^{p_\ell}_+,\,1\leqslant\ell\leqslant m,\,\mathrm{Tr}(Y)=1\right\}.\end{array}\tag{4.25}$$

The accuracy of the approximate solution $\tilde{x}_t$ of the (deterministic) Mirror Prox algorithm is [6, Example 2]

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

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 the matrix $A_0+\sum_{j=1}^n x_jA_j$ and the vector $[\mathrm{Tr}(YA_1);...;\mathrm{Tr}(YA_n)]$;
2. to compute the eigenvalue decomposition of y.

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

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

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 (4.10)!). Observe that the monotone operator associated with the saddle point problem (4.25) is

$$F(x,y)=\left[\underbrace{[\mathrm{Tr}(yA_1);...;\mathrm{Tr}(yA_n)]}_{F^x(x,y)};\ \underbrace{-A_0-\sum_{j=1}^n x_jA_j}_{F^y(x,y)}\right].\tag{4.26}$$

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

$$\Xi^y=A_0+A_\jmath;\tag{4.27}$$

2. we compute the quantities $\nu_\ell=\mathrm{Tr}(y_\ell)$, $1\leqslant\ell\leqslant m$. Since $y\in Y$, we have $\nu_\ell\geqslant 0$ 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

$$\Xi^x=[\mathrm{Tr}(A_1^\imath\bar{y}_\imath);...;\mathrm{Tr}(A_n^\imath\bar{y}_\imath)],\ \bar{y}_\imath=(\mathrm{Tr}(y_\imath))^{-1}y_\imath.\tag{4.28}$$

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 $[0,1]$. Given x,y and η, the value of this function 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 i-t call to the oracle Ξ_k, $z=(x,y)\in Z=X\times Y$ being the input, the oracle returns the vector

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

where $\zeta_i=[\eta_{i1};...;\eta_{ik}]$ and $\{\eta_{is}\}_{1\leqslant i,\,1\leqslant s\leqslant k}$ 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 the saddle point problem (4.25) with the stochastic oracle Ξ_k (k being the first parameter of our construction) specify a Nash s.v.i. on the domain $Z=X\times Y$. Let us 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 $\mathcal{Y}=\mathbf{S}^{p_1}\times...\times\mathbf{S}^{p_m}$ with the Spectahedron setup (see Section 2.2). 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 distance-generating function $\omega(\cdot)$ 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 4.3. *Let $n\geqslant 3$, $p^{(1)}\geqslant 3$. Then*

(i) *The parameters of the just defined distance-generating function ω w.r.t. the just defined norm $\|\cdot\|$ are $\alpha=1$, $\Theta=1$, $\Omega=\sqrt{2}$.*

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

$$\|F(z)-F(z')\|_*\leqslant L\|z-z'\|,\quad L=2\ln(n)+4\ln(p^{(1)}). \tag{4.29}$$

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

$$\begin{array}{ll}(a) & \mathbf{E}\left\{\Xi_k(z,\zeta_i)\right\}=F(z);\\ (b) & \mathbf{E}\left\{\exp\{\|\Xi(z,\zeta_i)-F(z)\|_*^2/M^2\}\right\}\leqslant\exp\{1\},\\ & M=27[\ln(n)+\ln(p^{(1)})]A_\infty/\sqrt{k}.\end{array} \tag{4.30}$$

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] Juditsky, A. Lan, G., Nemirovski, A., Shapiro, A., Stochastic Approximation Approach to Stochastic Programming,
`http://www.optimization-online.org/DB_HTML/2007/09/1787.html`

[4] 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

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

[6] 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.
[7] Lu, Z., Nemirovski A., Monteiro, R. "Large-Scale Semidefinite Programming via Saddle Point Mirror-Prox Algorithm", *Math. Progr.*, **109** (2007), 211-237.
[8] 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
[9] Nesterov, Yu. "Smooth minimization of non-smooth functions", *Math. Progr.*, **103** (2005), 127-152.
[10] Nesterov, Yu. "Excessive gap technique in nonsmooth convex minimization", *SIAM J. Optim.*, **16** (2005), 235-249.
[11] Nesterov, Yu. "Dual extrapolation and its applications to solving variational inequalities and related problems", *Math. Progr.* **109** (2007) 319-344.

5. Appendix.

5.1. Proof of Theorem 3.2. We start with the following simple observation: if r_e is a solution to (2.2), then $\partial_Z\omega(r_e)$ contains $-e$ and thus is nonempty, so that $r_e \in Z^o$. Moreover, one has

$$\langle \omega'(r_e) - e, u - r_e\rangle \geqslant 0 \ \forall u \in Z. \tag{5.1}$$

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 the function 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 $g(0) \geqslant 0$, which is exactly (5.1).

At least the first statement of the following Lemma is well-known:

LEMMA 5.1. *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,*

$$\|P(z, \zeta) - P(z, \eta)\| \leqslant \alpha^{-1}\|\zeta - \eta\|_* \quad \forall \zeta, \eta \in \mathcal{E}. \tag{5.2}$$

Besides this,

$$\begin{array}{lll} (a) & \forall (u \in Z) : V(P(z, \zeta), u) & \leqslant V(z, u) + \langle \zeta, u - P(z, \zeta)\rangle - V(z, P(z, \zeta)) \\ (b) & & \leqslant V(z, u) + \langle \zeta, u - z\rangle + \frac{\|\zeta\|_*^2}{2\alpha}. \end{array} \tag{5.3}$$

Proof.

Let $v \in P(z, \zeta)$, $w \in P(z, \eta)$. As $V'_u(z, u) = \omega'(u) - \omega'(z)$, invoking 5.1, we have $v, w \in Z^o$ and

$$\langle \omega'(v) - \omega'(z) + \zeta, v - u\rangle \leqslant 0 \quad \forall u \in Z. \tag{5.4}$$

$$\langle \omega'(w) - \omega'(z) + \eta, w - u\rangle \leqslant 0 \quad \forall u \in Z. \tag{5.5}$$

Setting $u = w$ in (5.4) and $u = v$ in (5.5), we get

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

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

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

and (5.2) follows. This relation, as a byproduct, implies that $P(z, \cdot)$ is single-valued.

To prove (5.3), let $v = P(z, \zeta)$. We have

$$\begin{aligned} V(v,u) - V(z,u) &= [\omega(u) - \langle \omega'(v), u - v\rangle - \omega(v)] - [\omega(u) - \langle \omega'(z), u - z\rangle - \omega(z)] \\ &= \langle \omega'(v) - \omega'(z) + \zeta, v - u\rangle + \langle \zeta, u - v\rangle - [\omega(v) - \langle \omega'(z), v - z\rangle - \omega(z)] \\ \text{(due to (5.4))} \quad &\leqslant \langle \zeta, u - v\rangle - V(z,v), \end{aligned}$$

as required in (a) of (5.3). The bound (b) of (5.3) is obtained from (5.3) using the Young inequality:

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

Indeed, observe that by definition, $V(z, \cdot)$ is strongly convex with parameter α, and $V(z,v) \geqslant \frac{\alpha}{2}\|z - v\|^2$, so that

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

□

We have the following simple corollary of Lemma 5.1:

COROLLARY 5.2. *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

$$(\forall u \in Z): \quad \langle -\sum_{\tau=1}^t \xi_\tau, u\rangle \leqslant V(y_0, u) + \sum_{\tau=1}^t \zeta_\tau, \tag{5.6}$$

with

$$|\zeta_\tau| \leqslant r\|\xi_\tau\|_* \ \text{(here } r = \max_{u \in Z}\|u\|); \quad \zeta_\tau \leqslant -\langle \xi_\tau, y_{\tau-1}\rangle + \frac{\|\xi_\tau\|_*^2}{2\alpha}. \tag{5.7}$$

Proof. Using the bound (*b*) of (5.3) with $\zeta = \xi_t$ and $z = y_{t-1}$ (so that $y_t = P(y_{t-1}, \xi_t)$ we obtain for any $u \in Z$:

$$V(y_t, u) - V(y_{t-1}, u) - \langle \xi_t, u\rangle \leqslant -\langle \xi_t, y_t\rangle - V(y_{t-1}, y_t) \equiv \zeta_t.$$

Note that

$$\zeta_t = \max_{v \in Z}[-\langle \xi_t, v\rangle - V(y_{t-1}, v)],$$

so that

$$-r\|\xi_t\|_* \leqslant -\langle \xi_t, y_{t-1}\rangle \leqslant \zeta_t \leqslant r\|\xi_t\|_*.$$

Further, due to the strong convexity of V,

$$\zeta_t = -\langle \xi_t, y_{t-1}\rangle + [-\langle \xi_t, y_t - y_{t-1}\rangle - V(y_{t-1}, y_t)] \leqslant -\langle \xi_t, y_{t-1}\rangle + \frac{\|\xi_t\|_*^2}{2\alpha}.$$

When summing up from $\tau = 1$ to $\tau = t$ we arrive at the corollary. □

We also need the following result.

LEMMA 5.3. *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

$$\begin{array}{ll} (a) & \|w - r_+\| \leqslant \alpha^{-1}\|\zeta - \eta\|_* \\ (b) & V(r_+, u) - V(z, u) \leqslant \langle \eta,\, u - w\rangle + \frac{\|\zeta-\eta\|_*^2}{2\alpha} - \frac{\alpha}{2}\|w - z\|^2. \end{array} \tag{5.8}$$

Proof. (a): this is nothing but (5.2).

(b): Using (a) of (5.3) in Lemma 5.1 we can write for $u = r_+$:

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

This results in

$$V(z, r_+) \geqslant V(w, r_+) + V(z, w) + \langle \zeta, w - r_+\rangle. \tag{5.9}$$

Using (5.3) with η substituted for ζ we get

$$\begin{aligned} V(r_+, u) &\leqslant V(z, u) + \langle \eta, u - r_+\rangle - V(z, r_+) \\ &= V(z, u) + \langle \eta, u - w\rangle + \langle \eta, w - r_+\rangle - V(z, r_+) \\ [\text{by (5.9)}] \quad &\leqslant V(z, u) + \langle \eta, u - w\rangle + \langle \eta - \zeta, w - r_+\rangle - V(z, w) - V(w, r_+) \\ &\leqslant V(z, u) + \langle \eta, u - w\rangle + \langle \eta - \zeta, w - r_+\rangle - \frac{\alpha}{2}[\|w - z\|^2 + \|w - r_+\|^2], \end{aligned}$$

due to the strong convexity of V. To conclude the bound (b) of (5.8) it suffices to note that by the Young inequality,

$$\langle \eta - \zeta, w - r_+\rangle \leqslant \frac{\|\eta - \zeta\|_*^2}{2\alpha} + \frac{\alpha}{2}\|w - r_+\|^2.$$

□

We are able now to prove Theorem 3.2. By (1.4) we have that

$$\begin{aligned} \|\widehat{F}(w_\tau) - \widehat{F}(r_{\tau-1})\|_*^2 &\leqslant (L\|r_{\tau-1} - w_\tau\| + M + \epsilon_{r_{\tau-1}} + \epsilon_{w_\tau})^2 \\ &\leqslant 3L^2\|w_\tau - r_{\tau-1}\|^2 + 3M^2 + 3(\epsilon_{r_{\tau-1}} + \epsilon_{w_\tau})^2. \end{aligned} \tag{5.10}$$

Let us now apply Lemma 5.3 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 $u \in Z$

$$\begin{aligned} &\langle \gamma_\tau \widehat{F}(w_\tau), w_\tau - u\rangle + V(r_\tau, u) - V(r_{\tau-1}, u) \\ &\leqslant \frac{\gamma_\tau^2}{2\alpha}\|\widehat{F}(w_\tau) - \widehat{F}(r_{\tau-1})\|^2 - \frac{\alpha}{2}\|w_\tau - r_{\tau-1}\|^2 \\ [\text{by (5.10)}] \quad &\leqslant \frac{3\gamma_\tau^2 L^2}{2\alpha}\left[\|w_\tau - r_{\tau-1}\|^2 + M^2 + (\epsilon_{r_{\tau-1}} + \epsilon_{w_\tau})^2\right] - \frac{\alpha}{2}\|w_\tau - r_{\tau-1}\|^2 \\ [\text{by (3.4)}] \quad &\leqslant \frac{3\gamma_\tau^2}{2\alpha}\left[M^2 + (\epsilon_{r_{\tau-1}} + \epsilon_{w_\tau})^2\right] \end{aligned}$$

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

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

Hence, for all $u \in Z$,

$$\begin{aligned}
&\sum_{\tau=1}^{t}\langle\gamma_\tau F(w_\tau), w_\tau - u\rangle \\
&\leqslant \Theta(r_0) + \sum_{\tau=1}^{t}\frac{3\gamma_\tau^2}{2\alpha}\left[M^2 + (\epsilon_{r_{\tau-1}} + \epsilon_{w_\tau})^2\right] + \sum_{\tau=1}^{t}\langle\gamma_\tau\Delta_\tau, w_\tau - u\rangle \\
&= \Theta(r_0) + \sum_{\tau=1}^{t}\frac{3\gamma_\tau^2}{2\alpha}\left[M^2 + (\epsilon_{r_{\tau-1}} + \epsilon_{w_\tau})^2\right] + \sum_{\tau=1}^{t}\langle\gamma_\tau\Delta_\tau, w_\tau - y_{\tau-1}\rangle \\
&\quad + \sum_{\tau=1}^{t}\langle\gamma_\tau\Delta_\tau, y_{\tau-1} - u\rangle,
\end{aligned} \tag{5.11}$$

where y_τ are given by (3.3). Since the sequences $\{y_\tau\}$, $\{\xi_\tau = \gamma_\tau\Delta_\tau\}$ satisfy the premise of Corollary 5.2, we have

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

and thus (5.11) implies that for any $u \in Z$

$$\begin{aligned}
\sum_{\tau=1}^{t}\langle\gamma_\tau F(w_\tau), w_\tau - u\rangle &\leqslant 2\Theta(r_0) + \sum_{\tau=1}^{t}\langle\gamma_\tau\Delta_\tau, w_\tau - y_{\tau-1}\rangle \\
&\quad + \sum_{\tau=1}^{t}\frac{3\gamma_\tau^2}{2\alpha}\left[M^2 + (\epsilon_{r_{\tau-1}} + \epsilon_{w_\tau})^2 + \frac{\epsilon_{w_\tau}^2}{3}\right]
\end{aligned} \tag{5.12}$$

To complete the proof of (3.5) in the general case, note that since F is monotone, (5.12) implies that for all $u \in Z$,

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

where

$$\Gamma(t) = 2\Theta(r_0) + \sum_{\tau=1}^{t}\frac{3\gamma_\tau^2}{2\alpha}\left[M^2 + (\epsilon_{r_{\tau-1}} + \epsilon_{w_\tau})^2 + \frac{\epsilon_{w_\tau}^2}{3}\right] + \sum_{\tau=1}^{t}\langle\gamma_\tau\Delta_\tau, w_\tau - y_{\tau-1}\rangle$$

(cf. (3.6)), whence

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

When taking the supremum over $u \in Z$, we arrive at (3.5).

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 (5.12):

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

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

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

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

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

or, which is the same,

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

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

$$\mathrm{Err}_{\mathrm{N}}(\widehat{z}_t) \leqslant \left[\sum_{\tau=1}^{t} \gamma_\tau \right]^{-1} \Gamma(t). \qquad \blacksquare$$

5.2. Proof of Theorem 3.5. In what follows, we use the notation from Theorem 3.2. By this theorem, in the case of constant stepsizes $\gamma_\tau \equiv \gamma$ we have

$$\mathrm{Err}_{\mathrm{vi}}(\widehat{z}_t) \leqslant [t\gamma]^{-1} \Gamma(t), \tag{5.13}$$

where

$$\begin{aligned} \Gamma(t) &= 2\Theta + \frac{3\gamma^2}{2\alpha} \sum_{\tau=1}^{t} \left[M^2 + (\epsilon_{r_{\tau-1}} + \epsilon_{w_\tau})^2 + \frac{\epsilon_{w_\tau}^2}{3} \right] + \gamma \sum_{\tau=1}^{t} \langle \Delta_\tau, w_\tau - y_{\tau-1} \rangle \\ &\leqslant 2\Theta + \frac{7\gamma^2}{2\alpha} \sum_{\tau=1}^{t} \left[M^2 + \epsilon_{r_{\tau-1}}^2 + \epsilon_{w_\tau}^2 \right] + \gamma \sum_{\tau=1}^{t} \langle \Delta_\tau, w_\tau - y_{\tau-1} \rangle. \end{aligned} \tag{5.14}$$

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

Note that by description of the algorithm $r_{\tau-1}$ is a deterministic function of $\zeta^{N(\tau-1)}$ and w_τ is a deterministic function of $\zeta^{M(\tau)}$ for certain increasing sequences of integers $\{M(\tau)\}$, $\{N(\tau)\}$ such that $N(\tau-1) < M(\tau) < N(\tau)$. Therefore $\epsilon_{r_{\tau-1}}$ is a deterministic function of $\zeta^{N(\tau-1)+1}$, and ϵ_{w_τ} and Δ_τ are deterministic functions of

$\zeta^{M(\tau)+1}$. Denoting by $\mathbf{E}_i$ the expectation w.r.t. ζ_i, we conclude that under assumption I we have

$$\mathbf{E}_{N(\tau-1)+1}\left\{\epsilon^2_{r_{\tau-1}}\right\} \leqslant M^2,\ \mathbf{E}_{M(\tau)+1}\left\{\epsilon^2_{w_\tau}\right\} \leqslant M^2,\ \|\mathbf{E}_{M(\tau)+1}\left\{\Delta_\tau\right\}\|_* \leqslant \mu, \tag{5.15}$$

and under assumption II, in addition,

$$\begin{array}{l}\mathbf{E}_{N(\tau-1)+1}\left\{\exp\{\epsilon^2_{r_{\tau-1}}M^{-2}\}\right\} \leqslant \exp\{1\},\\ \mathbf{E}_{M(\tau)+1}\left\{\exp\{\epsilon^2_{w_\tau}M^{-2}\}\right\} \leqslant \exp\{1\}.\end{array} \tag{5.16}$$

Now, let

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

We conclude by (5.15) that

$$\mathbf{E}\left\{\Gamma_0(t)\right\} \leqslant \frac{21\gamma^2M^2t}{2\alpha}. \tag{5.17}$$

Further, $y_{\tau-1}$ clearly is a deterministic function of $\zeta^{M(\tau-1)+1}$, whence $w_\tau - y_{\tau-1}$ is a deterministic function of $\zeta^{M(\tau)}$. Therefore

$$\begin{array}{rl}\mathbf{E}_{M(\tau)+1}\left\{\langle\Delta_\tau, w_\tau - y_{\tau-1}\rangle\right\} &= \langle\mathbf{E}_{M(\tau)+1}\left\{\Delta_\tau\right\}, w_\tau - y_{\tau-1}\rangle\\ &\leqslant \mu\|w_\tau - y_{\tau-1}\| \leqslant 2\mu\Omega,\end{array} \tag{5.18}$$

where the concluding inequality follows from the fact that Z is contained in the $\|\cdot\|$-ball of radius $\Omega = \sqrt{2\Theta/\alpha}$ centered at z_c, see (2.5). From (5.18) it follows that

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

Combining the latter relation, (5.13), (5.14) and (5.17), we arrive at (3.9). (i) is proved.

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

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

we get

$$\Gamma_0(t) = \frac{7\gamma^2M^2}{2\alpha}\left[t + J_t\right]. \tag{5.19}$$

At the same time, we can write

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

where $\xi_j \geqslant 0$ is a deterministic function of $\zeta^{I(j)}$ for certain increasing sequence of integers $\{I(j)\}$. Moreover, when denoting by $\mathbf{E}_j$ conditional expectation over $\zeta^{I(j)}$, $\zeta^{I(j)+1}$..., $\zeta^{I(j)-1}$ being fixed, we have

$$\mathbf{E}_j\left\{\exp\{\xi_j\}\right\} \leqslant \exp\{1\},$$

see (5.16). It follows that

$$\mathbf{E}\left\{\exp\{\sum_{j=1}^{k+1}\xi_j\}\right\} = \mathbf{E}\left\{\mathbf{E}_{k+1}\left\{\exp\{\sum_{j=1}^{k}\xi_j\}\exp\{\xi_{k+1}\}\right\}\right\}$$

$$= \mathbf{E}\left\{\exp\{\sum_{j=1}^{k}\xi_j\}\mathbf{E}_{k+1}\left\{\exp\{\xi_{k+1}\}\right\}\right\} \leqslant \exp\{1\}\mathbf{E}\left\{\exp\{\sum_{j=1}^{k}\xi_j\}\right\}. \tag{5.20}$$

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

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

Along with (5.19) it implies that

$$\forall \Lambda \geqslant 0: \mathrm{Prob}\left\{\Gamma_0(t) > \frac{21\gamma^2M^2t}{2\alpha} + \Lambda\frac{7\gamma^2M^2t}{2\alpha}\right\} \leqslant \exp\{-\Lambda t\}. \tag{5.21}$$

Let now $\xi_\tau = \langle \Delta_\tau, w_\tau - y_{\tau-1}\rangle$. Recall that $w_\tau - y_{\tau+1}$ is a deterministic function of $\zeta^{M(\tau)}$. Besides this, we have seen that $\|w_\tau - y_{\tau-1}\| \leqslant D \equiv 2\Omega$. Taking into account (5.15), (5.16), we get

$$\begin{array}{ll}(a) & \mathbf{E}_{M(\tau)+1}\{\xi_\tau\} \leqslant \rho \equiv \mu D,\\ (b) & \mathbf{E}_{M(\tau)+1}\left\{\exp\{\xi_\tau^2R^{-2}\}\right\} \leqslant \exp\{1\}, \ \text{ with } R = MD.\end{array} \tag{5.22}$$

Observe that $\exp\{x\} \leqslant x + \exp\{9x^2/16\}$ for all x. Thus (5.22.b) implies for $0 \leqslant s \leqslant \frac{4}{3R}$

$$\mathbf{E}_{M(\tau)+1}\{\exp\{s\xi_\tau\}\} \leqslant s\rho + \exp\{9s^2R^2/16\} \leqslant \exp\{s\rho + 9s^2R^2/16\}. \tag{5.23}$$

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

$$\mathbf{E}_{M(\tau)+1}\{\exp\{s\xi_\tau\}\} \leqslant \exp\{3s^2R^2/8\}\mathbf{E}_{M(\tau)+1}\left\{\exp\left\{\frac{2\xi_\tau^2}{3R^2}\right\}\right\} \leqslant \exp\left\{\frac{3s^2R^2}{8} + \frac{2}{3}\right\}.$$

When $s \geqslant \frac{4}{3R}$, the latter quantity is $\leqslant 3s^2R^2/4$, which combines with (5.23) to imply that for $s \geqslant 0$,

$$\mathbf{E}_{M(\tau)+1}\{\exp\{s\xi_\tau\}\} \leqslant \exp\{s\rho + 3s^2R^2/4\}. \tag{5.24}$$

Acting as in (5.20), we derive from (5.24) that

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

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

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

Finally, we arrive at

$$\mathrm{Prob}\left\{\gamma\sum_{\tau=1}^{t}\langle\Delta_\tau, w_\tau - y_{\tau-1}\rangle > 2\gamma\left[\mu t + \Lambda M\sqrt{t}\right]\Omega\right\} \leqslant \exp\{-\Lambda^2/3\}. \tag{5.25}$$

for all $\Lambda > 0$. Combining (5.13), (5.14), (5.21) and (5.25), we get (3.10). ∎

5.3. Proof of Lemma 4.1.

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

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

Thus, $\omega(\cdot)$ is strongly convex on Z, modulus $\alpha = 1$, w.r.t. the norm $\|\cdot\|$. Further, the minimizer of $\omega(\cdot)$ on Z clearly is $z_{\mathrm{c}} = (x_{\mathrm{c}}, y_{\mathrm{c}})$, and

$$\Theta = \frac{1}{\alpha_x \Omega_x^2} \Theta_x + \frac{1}{\alpha_y \Omega_y^2} \Theta_y = 1,$$

so that $\Theta = 1$, whence $\Omega = \sqrt{2\Theta/\alpha} = \sqrt{2}$.

Proof of (ii). $\mathbf{1}^0$. Let $z = (x, y)$ and $z' = (x', y')$ with $z, z' \in Z$. Observe that $\|y - y'\|_y \leqslant 2\Omega_y$ and thus

$$\|y'\|_y \leqslant 2\Omega_y \tag{5.26}$$

due to $0 \in Y$.

On the other hand, we have from (4.9) $F(z') - F(z) = [\Delta_x; \Delta_y]$, where

$$\begin{aligned}\Delta_x &= \sum_{\ell=1}^m [\phi'_\ell(x') - \phi'_\ell(x)]^* [\mathbf{A}_\ell^T y' + b_\ell] + \sum_{\ell=1}^m [\phi'_\ell(x)]^* \mathbf{A}_\ell [y' - y], \\ \Delta_y &= -\sum_{\ell=1}^m \mathbf{A}_\ell^* [\phi_\ell(x) - \phi_\ell(x')] + \Phi'_*(y') - \Phi'_*(y).\end{aligned}$$

We have

$$\begin{aligned}\|\Delta_x\|_{x,*} &= \max_{h \in \mathcal{X}\ \|h\|_x \leqslant 1} \langle h, \sum_{\ell=1}^m \Big[[\phi'_\ell(x') - \phi'_\ell(x)]^* [\mathbf{A}_\ell^T y' + b_\ell] + [\phi'_\ell(x)]^* \mathbf{A}_\ell [y' - y] \Big] \rangle_{\mathcal{X}} \\ &\leqslant \sum_{\ell=1}^m \left[\max_{\substack{h \in \mathcal{X} \\ \|h\|_x \leqslant 1}} \langle h, [\phi'_\ell(x') - \phi'_\ell(x)]^* [\mathbf{A}_\ell^T y' + b_\ell] \rangle_{\mathcal{X}} + \max_{h \in \mathcal{X}, \|h\|_x \leqslant 1} \langle h, [\phi'_\ell(x)]^* \mathbf{A}_\ell [y' - y] \rangle_{\mathcal{X}} \right] \\ &= \sum_{\ell=1}^m \left[\max_{\substack{h \in \mathcal{X} \\ \|h\|_x \leqslant 1}} \langle [\phi'_\ell(x') - \phi'_\ell(x)] h, \mathbf{A}_\ell^T y' + b_\ell \rangle_{\mathcal{X}} + \max_{h \in \mathcal{X}, \|h\|_x \leqslant 1} \langle [\phi'_\ell(x)] h, \mathbf{A}_\ell [y' - y] \rangle_{\mathcal{X}} \right] \\ &\leqslant \sum_{\ell=1}^m \Bigg[\max_{h \in \mathcal{X}\ \|h\|_x \leqslant 1} \|[\phi'_\ell(x') - \phi'_\ell(x)] h\|_{(\ell)} \|\mathbf{A}_\ell y' + b_\ell\|_{(\ell,*)} \\ &\quad + \max_{h \in \mathcal{X}\ \|h\|_x \leqslant 1} \|\phi'_\ell(x) h\|_{(\ell)} \|\mathbf{A}_\ell [y' - y]\|_{(\ell,*)} \Bigg].\end{aligned}$$

Then by (4.2),

$$\begin{aligned}|\Delta_x\|_{x,*} &\leqslant \sum_{\ell=1}^m \Big[[L_x \|x - x'\|_x + M_x] [\|\mathbf{A}_\ell y'\|_{(\ell,*)} + \|b_\ell\|_{(\ell,*)}] + [L_x \Omega_x + M_x] \|\mathbf{A}_\ell [y - y']\|_{(\ell,*)} \\ &= [L_x \|x - x'\|_x + M_x] \sum_{\ell=1}^m [\|\mathbf{A}_\ell y'\|_{(\ell,*)} + \|b_\ell\|_{(\ell,*)}] + [L_x \Omega_x + M_x] \sum_{\ell=1}^m \|\mathbf{A}_\ell [y - y']\|_{(\ell,*)} \\ &\leqslant [L_x \|x - x'\|_x + M_x] [\mathcal{A} \|y'\|_y + \mathcal{B}] + [L_x \Omega_x + M_x] \mathcal{A} \|y - y'\|_y,\end{aligned}$$

by definition of $\mathcal{A}$ and $\mathcal{B}$. Next, due to (5.26) we get by definition of $\|\cdot\|$

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

what implies

$$(a):\ \|\Delta_x\|_{x,*} \leqslant [\Omega_x[2\mathcal{A}\Omega_y + \mathcal{B}]L_x + 2\mathcal{A}\Omega_y[L_x\Omega_x + M_x]]\,\|z-z'\| + [2\mathcal{A}\Omega_y + \mathcal{B}]M_x$$

Further,

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

by (4.2.b) and (4.5). Now

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

and we come to

$$(b):\ \|\Delta_y\|_{y,*} \leqslant [\Omega_x\mathcal{A}[L_x\Omega_x + M_x] + \Omega_y L_y]\,\|z-z'\| + M_y.$$

From (a) and (b) it follows that

$$\begin{aligned}\|F(z)-F(z')\|_* &\leqslant \Omega_x\|\Delta_x\|_{x,*} + \Omega_y\|\Delta_y\|_{y,*}\\ &\leqslant \left[\Omega_x^2[2\mathcal{A}\Omega_y + \mathcal{B}]L_x + 3\mathcal{A}\Omega_x\Omega_y[L_x\Omega_x + M_x] + L_y\Omega_y^2\right]\|z-z'\|\\ &\quad +\Omega_x[2\mathcal{A}\Omega_y + \mathcal{B}]M_x + \Omega_y M_y.\end{aligned}$$

We have justified (4.12)

2^0. Let us verify (4.13). The first relation in (4.13) is readily given by (4.3.a,c). Let us fix $z=(x,y)\in Z$ and i, and let

$$\begin{aligned}\Delta &= F(z) - \Xi(z,\zeta_i)\\ &= \underbrace{[\sum_{\ell=1}^m[\phi_\ell'(x) - \mathbf{G}_\ell(x,\zeta_i)]^*\overbrace{[\mathbf{A}_\ell y + b_\ell]}^{\psi_\ell}}_{\Delta_x}; \underbrace{-\sum_{\ell=1}^m\mathbf{A}_\ell^*[\phi_\ell(x) - f_\ell(x,\zeta_i)]}_{\Delta_y}.\end{aligned}\tag{5.27}$$

As we have seen,

$$\sum_{\ell=1}^m\|\psi_\ell\|_{(\ell,*)} \leqslant 2\mathcal{A}\Omega_y + \mathcal{B}\tag{5.28}$$

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

$$\begin{aligned}\|\sum_{\ell=1}^m \mathbf{A}_\ell^* u_\ell\|_{y,*} &= \max_{\eta\in\mathcal{Y},\,\|\eta\|_y\leqslant 1} \langle \sum_{\ell=1}^m \mathbf{A}_\ell^* u_\ell, \eta\rangle_{\mathcal{Y}} = \max_{\eta\in\mathcal{Y},\,\|\eta\|_y\leqslant 1} \langle \sum_{\ell=1}^m u_\ell, \mathbf{A}_\ell\eta\rangle_{\mathcal{Y}} \\ &\leqslant \max_{\eta\in\mathcal{Y},\,\|\eta\|_y\leqslant 1} \left[\sum_{1\leqslant\ell\leqslant m} \|u_\ell\|_{(\ell)}\|\mathbf{A}_\ell\eta\|_{(\ell,*)}\right] \\ &\leqslant \max_{\eta\in\mathcal{Y},\,\|\eta\|_y\leqslant 1} \left[\max_{1\leqslant\ell\leqslant m}\|u_\ell\|_{(\ell)}\right] \sum_{1\leqslant\ell\leqslant m}\|\mathbf{A}_\ell\eta\|_{(\ell,*)} = \mathcal{A}\max_{1\leqslant\ell\leqslant m}\|u_\ell\|_{(\ell)}. \end{aligned} \tag{5.29}$$

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

$$\|\Delta_y\|_{y,*} = \|\sum_{\ell=1}^m \mathbf{A}_\ell^*[\phi_\ell(x) - f_\ell(x,\zeta)]\|_{y,*} \leqslant \mathcal{A}\underbrace{\max_{1\leqslant\ell\leqslant m}\|\phi_\ell(x) - f_\ell(x,\zeta)\|_{(\ell)}}_{\xi=\xi(\zeta_i)}. \tag{5.30}$$

Further,

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

Invoking (5.28), we conclude that

$$\|\Delta_x\|_{x,*} \leqslant \sum_{\ell=1}^m \rho_\ell\xi_\ell, \tag{5.31}$$

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

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

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

$$p(\xi) \leqslant \Omega_x M_x,\ p(\xi_\ell) \leqslant M_x$$

by (4.3.*b*,*d*). Now by (5.30), (5.31),

$$\left[\mathbf{E}\left\{\|\Delta\|_*^2\right\}\right]^{\frac{1}{2}} = \left[\mathbf{E}\left\{\Omega_x^2\|\Delta_x\|_{x,*}^2 + \Omega_y^2\|\Delta_y\|_{y,*}^2\right\}\right]^{\frac{1}{2}}$$

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

and the latter quantity is $\leqslant M$, see (4.12). We have established the second relation in (4.13).

3^0. It remains to prove that in the case of (4.14), relation (4.15) takes place. To this end, one can repeat word by word the reasoning from item 2^0 with the function $p_e(\eta)=\inf\left\{t>0:\mathbf{E}\left\{\exp\{\eta^2/t^2\}\right\}\leqslant\exp\{1\}\right\}$ 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 4.3. Item (i) can be verified exactly as in the case of Lemma 4.1; 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

$$\begin{array}{l}\|(x,\,y)\|=\sqrt{\|x\|_1^2/(2\ln(n))+|y|_1^2/(4\ln(p^{(1)}))},\\ \|(\xi,\,\eta)\|_*=\sqrt{2\ln(n)\|\xi\|_\infty^2+4\ln(p^{(1)})|\eta|_\infty^2}.\end{array}\tag{5.32}$$

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

$$F(z)-F(z')=\left[\Delta_x=[\mathrm{Tr}((y-y')A_1);...;\mathrm{Tr}((y-y')A_n)];\ \Delta_y=-\sum_{j=1}^n(x_j-x'_j)A_j\right].$$

whence

$$\begin{aligned}
\|\Delta_x\|_\infty&\leqslant|y-y'|_1\max_{1\leqslant j\leqslant n}|A_j|_\infty\leqslant\sqrt{2\ln(n)}A_\infty\|z-z'\|,\\
|\Delta_y|_\infty&\leqslant\|x-x'\|_\infty\max_{1\leqslant j\leqslant n}|A_j|_\infty\leqslant2\sqrt{\ln(p^{(1)})}A_\infty\|z-z'\|,
\end{aligned}$$

and

$$\|(\Delta_x,\,\Delta_y)\|_*\leqslant[2\ln(n)+4\ln(p^{(1)})]\|z-z'\|,$$

as required in (4.29). Further, relation (4.30.a) is clear from the construction of Ξ_k. To prove (4.30.b), observe that when $(x,y)\in Z$, we have (see (4.27), (4.28))

$$\|\Xi^x(x,y,\eta)\|_\infty\leqslant|\bar y_i|\max_{1\leqslant j\leqslant n}|A_j^i|_\infty\leqslant A_\infty,$$

and, since $F^x(x,y)=\mathbf{E}\left\{\Xi^x(x,y,\zeta\right\}$,

$$\|\Xi^x(x,y,\eta)-F^x(x,y)\|_\infty\leqslant2A_\infty.\tag{5.33}$$

Clearly,

$$|\Xi^y(x,y,\eta)-F^y(x,y)|_\infty=|A_\jmath-\sum_{j=1}^nx_jA_j|_\infty\leqslant2A_\infty.\tag{5.34}$$

Applying [4, Theorem 2.1(iii), Example 3.2, Lemma 1], we derive from (5.33) and (5.34) that for every $(x, y) \in Z$ and every $i = 1, 2, ...$ it holds

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

and

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

Combining the latter bounds with (5.32) we conclude (4.30.*b*). ∎