PROX-METHOD WITH RATE OF CONVERGENCE $O(1/T)$ FOR VARIATIONAL INEQUALITIES WITH LIPSCHITZ CONTINUOUS MONOTONE OPERATORS AND SMOOTH CONVEX-CONCAVE SADDLE POINT PROBLEMS

ARKADI NEMIROVSKI*

Abstract. We propose a prox-type method with efficiency estimate $O(\epsilon^{-1})$ for approximating saddle points of convex-concave $\mathrm{C}^{1,1}$ functions and solutions of variational inequalities with monotone Lipschitz continuous operators. Application examples include matrix games, eigenvalue minimization and computing Lovasz capacity number of a graph and are illustrated by numerical experiments with large-scale matrix games and Lovasz capacity problems.

Key words. saddle point problem, variational inequality, extragradient method, prox-method, ergodic convergence

AMS subject classifications. 90C25, 90C47

1. Introduction. This paper is inspired by recent paper of Nesterov [13] where a new method for minimizing a nonsmooth Lipschitz continuous function f over a convex compact finite-dimensional set X is proposed. The characteristic feature of Nesterov's method is that under favourable circumstances it exhibits *(nearly) dimension-independent $O(1/t)$-rate of convergence*: $f(x_t) - \min\limits_X f \leq O(1/t)$, where x_t is the approximate solution built after t iterations. This is in sharp contrast with the results of Information-Based Complexity Theory which state in particular (see [11]) that for a "black-box-oriented" method (one which operates with the values and subgradients of f only, without access to the "structure" of the objective), the number of function's evaluations required to build an ϵ-solution when minimizing a Lipschitz continuous, with constant 1, function over an n-dimensional unit Euclidean ball, cannot be less than $O(1/\epsilon^2)$, provided that $n \geq 1/\epsilon^2$. The explanation of the arising "contradiction" is that Nesterov's method is *not* black-box-oriented; specifically, it is assumed that the objective function f is given as a cost function of the first player in a specific convex-concave game:

$$f(x) = \max_{y \in Y} \phi(x, y), \quad \phi(x, y) = g(x) + x^T A y + h^T y, \tag{1.1}$$

where Y is a convex compact set and g is a $\mathrm{C}^{1,1}$ (i.e., with Lipschitz continuous gradient) convex function on X [1)]. When solving the problem, we are given the structure of the objective, specifically, know X and Y, and are able (a) to compute the value and the gradient of g at a point, and (b) to multiply a vector by A and A^T. The result of Nesterov states that if X and Y are simple enough (e.g., are unit Euclidean balls), then it is possible to minimize the objective (1.1) with accuracy ϵ in $O(1)\frac{L\|A\|}{\epsilon}$ steps, with single computation of g, g', two matrix-vector multiplications (one by A and one by A^T) and $O(\dim X + \dim Y)$ additional arithmetic operations per step; here L is the Lipschitz constant of g' with respect to the standard Euclidean

*Faculty of Industrial Engineering and Management, Technion – Israel Institute of Technology, Technion City, Haifa 32000, Israel (`nemirovs@ie.technion.ac.il`)

[1)] In fact, Nesterov allows to replace the linear in y component $h^T y$ with an arbitrary concave function $h(y)$; this, however, makes no difference, since the redefinition $y \leftarrow y^+ = (y, t)$, $Y \leftarrow Y^+ = \{(y, t) : \min\limits_{y' \in Y} h(y') \leq t \leq h(y)\}$ allows to make $h(\cdot)$ linear.

norm, and $\|A\|$ is the standard matrix norm of A. Nesterov's method is based on *smoothing* – approximating f by a function with Lipschitz continuous gradient (the possibility of "computationally cheap" smoothing is given by representation (1.1)) and minimizing the approximation by a method with the rate of convergence $O(1/t^2)$.

In this paper, we propose another method for convex optimization with nearly dimension-independent, under favourable circumstances, $O(1/t)$-rate of convergence. Our "favourable circumstances" are essentially the same as in [13], specifically, we assume that the objective function is given as the cost function of the first player in a game:

$$f(x) = \max_{y \in Y} \phi(x, y) \tag{1.2}$$

with $\mathrm{C}^{1,1}$ convex-concave function ϕ; besides, both X and Y, same as in Nesterov's construction, should be "simple enough" convex compact sets. In spite of similarity in the setup and in the resulting complexity bound, our method is completely different from the one of Nesterov: instead of using smoothing and advanced techniques for minimizing $\mathrm{C}^{1,1}$ functions, we directly apply to the underlying saddle point problem a simple prox-type method. Specifically, we reduce the problem of minimizing objective (1.2), i.e., the problem of approximating a saddle point of ϕ on $X \times Y$, to solving the associated variational inequality (v.i.):

$$\begin{array}{c} \text{find } z_* = (x_*, y_*) \in X \times Y : \langle \Phi(z_*), z - z_* \rangle \geq 0 \, \forall z \in X \times Y, \\ \Phi(x, y) = \left[\begin{array}{c} \frac{\partial}{\partial x} \phi(x, y) \\ -\frac{\partial}{\partial y} \phi(x, y) \end{array} \right] \end{array} \tag{1.3}$$

Note that since ϕ is convex-concave and $\mathrm{C}^{1,1}$, the operator Φ is monotone and Lipschitz continuous:

$$\|\Phi(z) - \Phi(z')\|_* \leq L_{\|\cdot\|}[\Phi] \|z - z'\| \quad \forall z, z' \in Z \equiv X \times Y,$$

where $\|\cdot\|$ is a norm and $\|\cdot\|_*$ is the conjugate norm. We solve (1.3) by a prox-method:

$$z_t = \text{ solution to v.i. on } Z \text{ with the operator } \Phi_t(z) = \gamma \Phi(z) + \omega'(z) - \omega'(z_{t-1}), \tag{1.4}$$

where the "stepsize" γ is positive and $\omega(z)$ is a C^1 strongly convex function on Z:

$$\langle \omega'(z) - \omega'(z'), z - z' \rangle \geq \alpha \|z - z'\|^2 \quad \forall z, z' \in Z \qquad [\alpha > 0]$$

It is easily seen that the prox-method converges ergodically to the set of solutions of (1.3): setting $(x^t, y^t) \equiv z^t = \frac{1}{t} \sum\limits_{\tau=1}^{t} z_{\tau-1}$, one has

$$\begin{array}{c} \max\limits_{y \in Y} \phi(x^t, y) - \min\limits_{x \in X} \phi(x, y^t) \leq O(1) \frac{\Theta(z_0)}{\gamma \alpha t}, \\ \Theta(z_0) = \max\limits_{z \in Z} \left[\omega(z) - \omega(z_0) - \langle z - z_0, \omega'(z_0) \rangle \right]; \end{array} \tag{1.5}$$

note that (1.5) even does not exploit the fact that Φ is Lipschitz continuous. The rate of convergence in (1.5) is exactly what we are looking for; the difficulty, however, is that method (1.4) is "conceptual" rather than implementable: a step requires solving a nontrivial variational inequality. Our central observation is that *in the case of Lipschitz continuous Φ, a step of the prox-method is easily implementable, provided* *that the stepsize γ is chosen properly.* Specifically, when $\gamma \leq \frac{\alpha}{\sqrt{2}L_{\|\cdot\|}[\Phi]}$, the *prox-mapping*

$$z \mapsto \mathcal{P}_t(z) = \operatorname*{argmin}_{z' \in Z} \left[\omega(z') + \langle \gamma\Phi(z) - \omega'(z_{t-1}), z'\rangle\right] : Z \to Z$$

turns out to be a $\frac{1}{\sqrt{2}}$-contraction: $\|\mathcal{P}_t(z) - \mathcal{P}_t(z')\| \leq \frac{1}{\sqrt{2}}\|z - z'\|$ for all $z, z' \in Z$, and z_t as given by (1.4) is exactly the fixed point of this contraction. It follows that *with a properly chosen stepsize, the prox-method becomes implementable: for all computational purposes, its step requires a small number of fixed point iterations* $z \mapsto \mathcal{P}_t(z)$, i.e., a small number (in fact, just two!) of computations of $\Phi(\cdot)$ and solvings auxiliary problems of the form

$$\min_{z \in Z} \left[\omega(z) + \langle \xi, z\rangle\right]. \tag{1.6}$$

As a result, *one can approximate within accuracy ϵ a saddle point of a* $\mathrm{C}^{1,1}$ *convex-concave function* (or, more generally, a solution to a v.i. with monotone Lipschitz continuous operator) *at the cost* (# of computations of the associated monotone operator and solvings auxiliary problems (1.6)) *which is inverse proportional to ϵ.*

Note that both "building blocks" of our construction – the ergodic convergence properties of a prox-method (see [3, 10, 12, 4, 6, 16] and references therein) and the contraction properties of the prox-mapping – seem to be well-studied. Surprisingly, the aforementioned complexity result, to the best of our knowledge, is new even in the standard Euclidean case (the one where $\omega(z) = \frac{1}{2}\langle z, z\rangle$). In this case, our scheme results in the *extragradient* method

$$z_{t-1}^+ = \pi_Z\left(z_{t-1} - \gamma\Phi(z_{t-1})\right),\ z_t = \pi_Z\left(z_{t-1} - \gamma\Phi(z_{t-1}^+)\right) \qquad \left[\pi_Z(z) = \operatorname*{argmin}_{z' \in Z} \langle z - z', z - z'\rangle\right]$$

was proposed, from a different perspective, by G. Korpelevich as early as in 1976, see [7, 8]; here again the efficiency estimate $O(1/t)$ for the ergodic version $z^t = \frac{1}{t}\sum\limits_{\tau=1}^{t} z_{\tau-1}^+$ seems to be new.

The rest of the paper is organized as follows. In Section 2 we formulate the problem of interest – a variational inequality with Lipschitz continuous monotone operator – and investigate the possibility to solve the problem by a "conceptual" prox-method. In Section 3 we present "implementable" versions of the method and carry out the complexity analysis of the resulting algorithms. In Section 4, we present some extensions and modifications of the method. In Section 5, we discuss a number of generic applications of our algorithms, specifically, solving matrix games, eigenvalue minimization and computing Lovasz capacity number of a graph. In particular, we demonstrate that our techniques allow to compute the Lovasz capacity number Θ of an n-node graph within accuracy $\epsilon\Theta$, $\epsilon \in (0, 1]$, in no more than $O(n^4\epsilon^{-1}\sqrt{\ln(n)}\ln(n/\epsilon))$ arithmetic operations[2)]. In concluding Section 6, we present encouraging numerical results for matrix games with sparse matrices of sizes varying to 20,000×20,000 and for Lovasz capacity problems on random and Hamming graphs (the largest graph we consider has 1024 nodes and 197120 arcs).

2. The problem and the conceptual prox-method.

[2)] Note that if both the graph and its complement have $O(n^2)$ arcs, then the arithmetic cost of a *single* iteration of an interior point method as applied to computing Lovasz capacity is $O(n^6)$.

2.1. The problem. Let Z be a convex compact set in Euclidean space E with inner product $\langle\cdot,\cdot\rangle$, $\|\cdot\|$ be a norm on E (not necessarily the one associated with the inner product), and $F : Z \to E$ be a Lipschitz continuous monotone mapping:

$$\begin{array}{lll}\forall(z,z'\in Z): & \|F(z)-F(z')\|_*\leq L\|z-z'\| & (a)\\ \forall(z,z'\in Z): & \langle F(z)-F(z'),z-z'\rangle\geq 0 & (b)\end{array}\tag{2.1}$$

where $\|\cdot\|_*$ is the norm conjugate to $\|\cdot\|$:

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

We are interested to approximate a solution to the variational inequality (v.i.) associated with Z, F, i.e., a point $z_*\in Z$ such that

$$\langle F(z), z_*-z\rangle\leq 0\quad\forall z\in Z$$

(note that since F is single-valued and continuous on Z, the latter relation is equivalent to $\langle F(z_*), z-z_*\rangle\geq 0$ for all $z\in Z$, which is the standard definition of a (strong) solution to the v.i. associated with Z, F).

2.2. Prox-mapping – preliminaries. Let us fix a continuously differentiable function $\omega(z) : Z\to\mathbf{R}$ which is strongly convex:

$$\langle\omega'(z)-\omega'(w),z-w\rangle\geq\alpha\|z-w\|^2\quad\forall z,w\in Z\quad[\alpha>0]\tag{2.2}$$

and is such that one can easily solve problems of the form

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

Let also

$$\Omega(\xi)=\max_{z\in Z}[\langle\xi,z\rangle-\omega(z)]$$

be the Legendre transformation of $\omega\big|_Z$, and let

$$\begin{array}{rcl}\Omega_u(\xi) & = & \Omega(\xi)-\langle\xi,u\rangle : E\to\mathbf{R},\\ H_u(z) & = & \Omega_u(\omega'(z)) : Z\to\mathbf{R}.\end{array}\tag{2.3}$$

Let us fix $\gamma>0$. Given γ and $z\in Z$, let us define the *prox-mapping*

$$P_z : E\to Z:\quad P_z(\xi)=\operatorname*{argmin}_{w\in Z}\left[\omega(w)+\langle w,\xi-\omega'(z)\rangle\right].$$

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

LEMMA 2.1. *The mapping P_z is Lipschitz continuous, specifically,*

$$\|P_z(\xi)-P_z(\eta)\|\leq\alpha^{-1}\|\xi-\eta\|_*\quad\forall\xi,\eta\in E.\tag{2.4}$$

Besides this,

$$\forall(u\in Z): H_u(P_z(\xi))-H_u(z)\leq\langle\xi,u-P_z(\xi)\rangle+\underbrace{[\omega(z)+\langle\omega'(z),P_z(\xi)-z\rangle-\omega(P_z(\xi))]}_{\leq 0}.\tag{2.5}$$

Proof. Setting $v = P_z(\xi)$, $w = P_z(\eta)$, we have

$$\langle \omega'(v) - \omega'(z) + \xi, v - u \rangle \le 0 \quad \forall u \in Z. \tag{2.6}$$

$$\langle \omega'(w) - \omega'(z) + \eta, w - u \rangle \le 0 \quad \forall u \in Z. \tag{2.7}$$

Setting $u = w$ in (2.6) and $u = v$ in (2.7), we get

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

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

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

and (2.4) follows.

To prove (2.5), let $v = P_z(\xi)$. We have

$$\begin{array}{rl}
& H_u(P_z(\xi)) - H_u(z) \\
= & \Omega_u(\omega'(v)) - \Omega_u(\omega'(z)) \\
= & \Omega(\omega'(v)) - \langle u, \omega'(v) \rangle - \Omega(\omega'(z)) + \langle u, \omega'(z) \rangle \\
= & \langle \omega'(v), v \rangle - \omega(v) - \langle u, \omega'(v) \rangle - \langle \omega'(z), z \rangle + \omega(z) + \langle u, \omega'(z) \rangle \\
= & \langle \omega'(v) - \omega'(z) + \xi, v - u \rangle + [\omega(z) + \langle \omega'(z), v - z \rangle - \omega(v)] + \langle \xi, u - v \rangle \\
\le & \langle \xi, u - v \rangle + [\omega(z) + \langle \omega'(z), v - z \rangle - \omega(v)]
\end{array}$$

(we have used (2.6)), as required in (2.5). □

2.3. The prox-method. The *prox-method* with general prox-term ("with Bregman distances") for convex minimization and solving variational inequalities was investigated in many papers (see [4, 16, 6] and references therein). As applied to the v.i. associated with Z, F, the (slightly modified, see below) method in its conceptual form is as follows:

CPM [Conceptual Prox-Method]:
Initialization. Choose starting point $z_0 \in Z$.
Step t, $t = 1, 2, \ldots$: Given z_{t-1}, check whether

$$P_{z_{t-1}}(F(z_{t-1})) = z_{t-1}. \tag{2.8}$$

If it is the case, claim that z_{t-1} is the solution to the variational inequality associated with Z, F and terminate. Otherwise choose $\gamma_t > 0$ and a point $w_t \in Z$ such that

$$\begin{array}{l}\langle w_t - P_{z_{t-1}}(\gamma_t F(w_t)), \gamma_t F(w_t) \rangle \\ + \left[\omega(z_{t-1}) + \langle \omega'(z_{t-1}), P_{z_{t-1}}(\gamma_t F(w_t)) - z_{t-1} \rangle - \omega(P_{z_{t-1}}(\gamma_t F(w_t)))\right] \le 0.\end{array} \tag{2.9}$$

Set

$$\begin{array}{rcl}
z_t & = & P_{z_{t-1}}(\gamma_t F(w_t)), \\
z^t & = & \left(\sum\limits_{\tau=1}^t \gamma_\tau\right)^{-1} \sum\limits_{\tau=1}^t \gamma_\tau w_\tau
\end{array}$$

and pass to step $t + 1$.
Note: z^t is the approximate solution built in course of t steps.

Note that the standard way to ensure (2.9) is to define w_t as the solution to the v.i. with the monotone operator $F_{t-1}(z) = \omega'(z) - \omega'(z_{t-1}) + \gamma_t F(z)$, which, as it is immediately seen, results in $w_t = P_{z_{t-1}}(\gamma_t F(w_t))$ and thus makes the left hand side in (2.9) nonpositive. The resulting procedure takes the form

$$z_t = \text{ solution to the v.i. given by } Z, F_{t-1}, \tag{2.10}$$

and this is exactly what is called "prox-method with Bregman distances" and what is considered in the aforementioned references (primarily in the case when F is the (sub)gradient mapping generated by a convex function). Note that rule (2.10) is not "directly implementable", since it requires to solve a non-trivial v.i., and thus the "standard" prox-method is conceptual only, same as the above "algorithm".

The convergence properties of CPM are summarized in the following statement (which is very similar to numerous results on ergodic convergence of gradient-type algorithms for convex minimization and v.i.'s with monotone operators, see, e.g., [3, 10, 11, 12, 16]):

PROPOSITION 2.2. *Consider a "relaxed" version of CPM, where condition* (2.9) *is weakened to*

$$\begin{array}{l}\langle w_t - P_{z_{t-1}}(\gamma_t F(w_t)), \gamma_t F(w_t)\rangle + \omega(z_{t-1}) \\ +\langle \omega'(z_{t-1}), P_{z_{t-1}}(\gamma_t F(w_t)) - z_{t-1}\rangle - \omega(P_{z_{t-1}}(\gamma_t F(w_t))) \le \epsilon_t.\end{array} \tag{2.11}$$

(i) *If the relaxed algorithm terminates at certain step N according to rule* (2.8), *then z_{N-1} is a solution to the variational inequality:*

$$\langle F(z_{N-1}), u - z_{N-1}\rangle \ge 0 \quad \forall u \in Z. \tag{2.12}$$

(ii) *If the relaxed algorithm does not terminate in course of N steps, then*

$$\begin{array}{rcl}\epsilon(z^N) & \equiv & \max\limits_{u\in Z}\langle F(u), z^N - u\rangle \le \dfrac{\Theta(z_0)+\sum_{t=1}^N \epsilon_t}{\sum_{t=1}^N \gamma_t}, \\ \Theta(z_0) & = & \max\limits_{z\in Z}\left[\omega(z) - \omega(z_0) - \langle \omega'(z_0), z - z_0\rangle\right].\end{array} \tag{2.13}$$

If, in addition, the variational inequality in question is associated with a convex-concave saddle point problem, i.e.,

- $E = E_x \times E_y$ *for Euclidean spaces* E_x, E_y,
- $Z = X \times Y$ *for convex compacts* $X \subset E_x$, $Y \subset E_y$,
- $F(z) \equiv F(x,y) = \begin{bmatrix} f'_x(x,y) \\ -f'_y(x,y)\end{bmatrix}$ *for a continuously differentiable function* $f(x,y) : X \times Y \to \mathbf{R}$ *which is convex in* $x \in X$ *and is concave in* $y \in Y$,

then

$$\begin{array}{rcl}\epsilon_f(z^N) & \equiv & \left[\max\limits_{y\in Y} f(x^N, y) - \min\limits_{x\in X}\max\limits_{y\in Y} f(x,y)\right] + \left[\max\limits_{y\in Y}\min\limits_{x\in X} f(x,y) - \min\limits_{x\in X} f(x, y^N)\right] \\ & \le & \dfrac{\Theta(z_0)+\sum_{t=1}^N \epsilon_t}{\sum_{t=1}^N \gamma_t},\end{array} \tag{2.14}$$

with $\Theta(z_0)$ *given by* (2.13).

Proof. (i) is evident, since $P_z(\xi) = z$ means exactly that $\langle \xi, u - z\rangle \geq 0$ for all $u \in Z$ by definition of $P_z(\cdot)$.

To prove (ii), observe that $z_t = P_{z_{t-1}}(\gamma_t F(w_t))$, whence for every $u \in Z$ and every $t \leq N$ one has

$$\begin{array}{rcl} H_u(z_t) - H_u(z_{t-1}) & \leq & \langle \gamma_t F(w_t), u - z_t\rangle + [\omega(z_{t-1}) + \langle \omega'(z_{t-1}), z_t - z_{t-1}\rangle - \omega(z_t)] \\ & & \hfill [\text{by (2.5)}] \\ & \leq & \langle \gamma_t F(w_t), u - w_t\rangle + \epsilon_t, \\ & & \hfill [\text{by (2.11)}] \end{array}.$$

Thus,

$$\langle \gamma_t F(w_t), w_t - u\rangle \leq H_u(z_{t-1}) - H_u(z_t) + \epsilon_t. \tag{2.15}$$

Summing up the resulting inequalities over $t = 1, ..., N$, we get

$$\begin{array}{rcl} \forall (u \in Z): & & \sum\limits_{t=1}^N \gamma_t \langle F(w_t), w_t - u\rangle \\ & \leq & H_u(z_0) - H_u(z_N) + \sum\limits_{t=1}^N \epsilon_t \\ & = & [\langle z_0 - u, \omega'(z_0)\rangle - \omega(z_0)] + \underbrace{[\langle u - z_N, \omega'(z_N)\rangle + \omega(z_N)]}_{\leq \omega(u)} \\ & \leq & \omega(u) - \langle \omega'(z_0), u - z_0\rangle - \omega(z_0) + \sum\limits_{t=1}^N \epsilon_t \leq \Theta(z_0) + \sum\limits_{t=1}^N \epsilon_t. \end{array} \tag{2.16}$$

Now, to get (2.13), note that by monotonicity of F one has

$$\langle F(w_t), w_t - u\rangle \geq \langle F(u), w_t - u\rangle,$$

and thus (2.16) implies that

$$\left(\sum_{t=1}^N \gamma_t\right) \langle F(u), z^N - u\rangle = \sum_{t=1}^N \gamma_t \langle F(u), w_t - u\rangle \leq \Theta(z_0) + \sum_{t=1}^N \epsilon_t \quad \forall u \in Z,$$

whence

$$\epsilon(z^N) = \max_{u\in Z} \langle F(u), z^N - u\rangle \leq \frac{\Theta(z_0) + \sum\limits_{t=1}^N \epsilon_t}{\sum\limits_{t=1}^N \gamma_t},$$

as required in (2.13).

Now assume that the variational inequality in question comes from convex-concave saddle point problem. Setting $\lambda_t = \gamma_t / \sum\limits_{\tau=1}^N \gamma_\tau$, (2.16) gives the first inequality in the following computation (where $z_t = (x_t, y_t)$ is the trajectory of the method and

$u=(x,y)\in Z=X\times Y$):

$$\begin{array}{rcl}
\frac{\Theta(z_0)+\sum\limits_{t=1}^N\epsilon_t}{\sum\limits_{t=1}^N\gamma_t} & \geq & \sum\limits_{t=1}^N \lambda_t\langle F(z_t),z_t-u\rangle\\
& = & \sum\limits_{t=1}^N \lambda_t[\langle f'_x(x_t,y_t),x_t-x\rangle+\langle f'_y(x_t,y_t),y-y_t\rangle]\\
& \geq & \sum\limits_{t=1}^N \lambda_t\left[[f(x_t,y_t)-f(x,y_t)]+[f(x_t,y)-f(x_t,y_t)]\right]\\
& & \qquad\qquad [\text{since } f \text{ is convex in } x \text{ and concave in } y]\\
& = & \sum\limits_{t=1}^N \lambda_t\left[f(x_t,y)-f(x,y_t)\right]\\
& \geq & f(\sum\limits_{t=1}^N\lambda_t x_t,y)-f(x,\sum\limits_{t=1}^N\lambda_t y_t)\\
& = & f(x^N,y)-f(x,y^N)
\end{array}$$

The resulting inequality is valid for all $x\in X$, $y\in Y$, whence

$$\begin{array}{rcl}
\max\limits_{x\in X,y\in Y}\left[f(x^N,y)-f(x,y^N)\right] & = & \Big[[\max\limits_{y\in Y}f(x^N,y)-\min\limits_{x\in X}\max\limits_{y\in Y}f(x,y)]\\
& & +[\max\limits_{y\in Y}\min\limits_{x\in X}f(x,y)-\min\limits_{x\in X}f(x,y^N)]\Big]\\
& \leq & \frac{\Theta(z_0)+\sum\limits_{t=1}^N\epsilon_t}{\sum\limits_{t=1}^N\gamma_t},
\end{array}$$

as claimed (note that $\max\limits_{y\in Y}\min\limits_{x\in X}f(x,y)=\min\limits_{x\in X}\max\limits_{y\in Y}f(x,y)$, since X,Y are convex compacts and f is convex-concave). □

REMARK 2.3. *From now on, we assume that the starting point z_0 for the conceptual prox-method is chosen as the minimizer of $\omega(\cdot)$ on Z. With this choice of z_0, one has*

$$\Theta(z_0)\leq V[\omega]\equiv\max_{z\in Z}\omega(z)-\min_{z\in Z}\omega(z).$$

3. From conceptual method to implementable algorithm. Our main observation is extremely simple:

($*$) *Let F satisfy* (2.1)*, and let $\gamma\leq\frac{\alpha}{\sqrt{2}L}$. Then, for every $z\in Z$, the mapping*

$$\mathcal{P}_z^\gamma: Z\to Z:\mathcal{P}_z^\gamma(w)=P_z(\gamma F(w))$$

is a contraction:

$$\|\mathcal{P}_z^\gamma(w)-\mathcal{P}_z(w')\|\leq\frac{1}{\sqrt{2}}\|w-w'\|\quad\forall w,w'\in Z. \tag{3.1}$$

Indeed, the mapping in question is the superposition of the Lipschitz continuous, with constant γL, mapping $w\mapsto\gamma F(w):(Z,\|\cdot\|)\to(E,\|\cdot\|_*)$ (see (2.1)) and Lipschitz continuous with constant α^{-1} (by Lemma 2.1) mapping $\xi\mapsto P_z(\xi):(E,\|\cdot\|_*)\to(Z,\|\cdot\|)$.

Observation ($*$) suggests the following implementation of step t in the conceptual prox-method:

["Basic implementation"] *Given a point $z_{t-1} \in Z$ which does not satisfy* (2.8), *set*

$$\gamma_t = \gamma \equiv \frac{\alpha}{\sqrt{2}L}, \tag{3.2}$$

build, starting with $w_{t,0} \equiv z_{t-1}$, the iterates

$$w_{t,s} = P_{z_{t-1}}(\gamma F(w_{t,s-1})), s = 1, 2, ... \tag{3.3}$$

until the condition

$$\begin{array}{r}\langle \gamma F(w_{t,s-1}), w_{t,s-1} - w_{t,s} \rangle + \omega(z_{t-1}) \\ +\langle \omega'(z_{t-1}), w_{t,s} - z_{t-1} \rangle - \omega(w_{t,s}) \leq 0\end{array} \tag{3.4}$$

is met, let it happen when $s = s_t$, and set

$$w_t = w_{t,s_t - 1} \quad [\Rightarrow z_t = w_{t,s_t}].$$

Note that with the Basic implementation, we do ensure (2.11) with $\epsilon_t = 0$ (this requirement is nothing but (3.4)). On the other hand, the sequence $\{w_{t-1,s}\}_{s=0}^{\infty}$ rapidly converges to the fixed point of the mapping $\mathcal{P}^{\gamma}_{z_{t-1}}$:

$$\|w_{t,s+1} - w_{t,s}\| \leq \frac{1}{\sqrt{2}} \|w_{t,s} - w_{t,s-1}\|, \; s = 1, 2, ...$$

We are in the situation when $z_{t-1} \equiv w_{t,0}$ is not a fixed point of $\mathcal{P}^1_{z_{t-1}}$, whence, as it is immediately seen, $w_{t,0}$ is not a fixed point of the mapping $\mathcal{P}^{\gamma}_{z_{t-1}}$; it follows that the first term in the left hand side of (3.4) rapidly converges to zero (as a geometric progression with the ratio $\sqrt{1/2}$), while the second term, as it is immediately seen, is at least $\frac{\alpha}{16}\|w_{t,1} - w_{t,0}\|^2$. Thus,

> (!) *For all computational purposes, the number of "inner iterations"* (3.3) *at a step t in the Basic implementation of the prox-method can be treated as a moderate constant C, while the number $N(\epsilon)$ of steps sufficient to make the right hand side in* (2.13), (2.14) $\leq \epsilon$ *can be bounded from above as*
>
> $$N(\epsilon) \leq \frac{2LV[\omega]}{\alpha\epsilon} \tag{3.5}$$

We are about to prove that in fact the number of "inner iterations" can be made as small as 2. Note that in the first version of the paper and in the underlying preprint, this possibility of "two-step implementation" was established only for the case of $\mathrm{C}^{1,1}$ function $\omega(\cdot)$; since then, inspired by the latest related results of Yu. Nesterov [14], we succeeded to extend this result onto the general case.

LEMMA 3.1. *Let a nonempty set $U \subset Z$ be convex and closed, and let $z \in Z$. Let ξ, η be two points from E, and let $\gamma > 0$. Consider the points*

$$\begin{array}{rcl} w & = & \underset{y \in U}{\operatorname{argmin}} \, [\langle \gamma\xi - \omega'(z), y \rangle + \omega(y)] \\ z_+ & = & \underset{y \in U}{\operatorname{argmin}} \, [\langle \gamma\eta - \omega'(z), y \rangle + \omega(y)] \end{array} \tag{3.6}$$

Then for all $u \in U$ one has

$$
(3.7)\quad
\begin{array}{ll}
(a) & \|w - z_+\| \le \alpha^{-1}\gamma\|\xi - \eta\|_* \\
(b) & \langle \gamma\eta, w - u\rangle \le H_u(z) - H_u(z_+) \\
 & \qquad + \underbrace{\langle \gamma\eta, w - z_+\rangle + [\omega(z) + \langle \omega'(z), z_+ - z\rangle - \omega(z_+)]}_{\delta}, \\
(c) & \delta \le \underbrace{\langle \gamma(\eta - \xi), w - z_+\rangle + [\omega(z) + \langle \omega'(z), w - z\rangle + \langle \omega'(w), z_+ - w\rangle - \omega(z_+)]}_{\epsilon} \\
(d) & \epsilon \le \alpha^{-1}\gamma^2\|\xi - \eta\|_*^2 - \frac{\alpha}{2}\left[\|w - z\|^2 + \|w - z_+\|^2\right]
\end{array}
$$

Proof. (3.7.a): The verification is completely similar to the one of (2.4) (which is nothing but the particular case of (3.7.a) corresponding to the case $U = Z$). For all $u \in U$ one has

$$
(3.8)\qquad
\begin{array}{lrcl}
(a) & \langle \gamma\xi + \omega'(w) - \omega'(z), w - u\rangle & \le & 0 \\
(b) & \langle \gamma\eta + \omega'(z_+) - \omega'(z), z_+ - u\rangle & \le & 0
\end{array}
$$

Applying (3.8.a) with $u = z_+$ and (3.8.b) with $u = w$, we get

$$
\begin{array}{rcl}
\gamma\langle \xi, w - z_+\rangle & \le & \langle \omega'(z) - \omega'(w), w - z_+\rangle \\
\gamma\langle \eta, z_+ - w\rangle & \le & \langle \omega'(z) - \omega'(z_+), z_+ - w\rangle
\end{array}
$$

Summing up these inequalities, we get

$$\gamma\langle \xi - \eta, w - z_+\rangle \le \langle \omega'(z_+) - \omega'(w), w - z_+\rangle,$$

whence

$$-\gamma\|\xi - \eta\|_*\|w - z_+\| \le \gamma\langle \xi - \eta, w - z_+\rangle \le \langle \omega'(z_+) - \omega'(w), w - z_+\rangle \le -\alpha\|w - z_+\|^2,$$

and we arrive at (3.7.a).

(3.7.b): Let $\phi(u) = [\langle \gamma\eta, w - u\rangle + H_u(z_+) - H_u(z)]$; an immediate computation demonstrates that

$$\phi(z_+) = \langle \gamma\eta, w - z_+\rangle + [\omega(z) + \langle \omega'(z), z_+ - z\rangle - \omega(z_+)],$$

so that (3.7.b) is exactly the claim that $\phi(u) \le \phi(z_+)$ for all $u \in U$. To see that this claim is true, note that $\phi(u) = -\langle \gamma\eta + \omega'(z_+) - \omega'(z), u\rangle + \text{const}$, so that (3.8.$b$) is a sufficient condition for $z_+ \in U$ to maximize $\phi(\cdot)$ on U.

(3.7.c): For $u \in U$ we have

$$
\begin{array}{l}
\delta \equiv \langle \gamma\eta, w - z_+\rangle + [\omega(z) + \langle \omega'(z), z_+ - z\rangle - \omega(z_+)] \\
= \langle \gamma(\eta - \xi), w - z_+\rangle + \langle \gamma\xi, w - z_+\rangle + [\omega(z) + \langle \omega'(z), z_+ - z\rangle - \omega(z_+)] \\
\le \langle \gamma(\eta - \xi), w - z_+\rangle + \langle \omega'(z) - \omega'(w), w - z_+\rangle + [\omega(z) + \langle \omega'(z), z_+ - z\rangle - \omega(z_+)] \\
\hfill [(3.8.a) \text{ with } u = z_+] \\
= \langle \gamma(\eta - \xi), w - z_+\rangle + [\omega(z) + \langle \omega'(z), w - z\rangle + \langle \omega'(w), z_+ - w\rangle - \omega(z_+)]
\end{array}
$$

as required in (3.7.c).

(3.7.d): For $u \in U$ we have

$$
\begin{array}{l}
\epsilon \equiv \langle \gamma(\eta - \xi), w - z_+\rangle + [\omega(z) + \langle \omega'(z), w - z\rangle + \langle \omega'(w), z_+ - w\rangle - \omega(z_+)] \\
\le \gamma\|\eta - \xi\|_*\|w - z_+\| + [\omega(z) + \langle \omega'(z), w - z\rangle - \omega(w)] \\
+ [\omega(w) + \langle \omega'(w), z_+ - w\rangle - \omega(z_+)] \le \alpha^{-1}\gamma^2\|\eta - \xi\|_*^2 - \frac{\alpha}{2}\left[\|w - z\|^2 + \|w - z_+\|^2\right]
\end{array}
$$

where the concluding inequality is given by (3.7.a) and by strong convexity of $\omega(\cdot)$. □

THEOREM 3.2. *Assume that F satisfies* (2.1). *Then the Basic Implementation ensures relation* (3.4) *(and thus –* (2.11) *with $\epsilon_t = 0$) in no more than two inner iterations. Thus, with Basic Implementation, the error bounds* (2.13) *and* (2.14) *become, respectively,*

$$\epsilon(z^N) \leq \frac{\sqrt{2}\Theta(z_0)L}{\alpha N} \tag{3.9}$$

and

$$\epsilon_f(z^N) \leq \frac{\sqrt{2}\Theta(z_0)L}{\alpha N} \tag{3.10}$$

while the "computational price" of z^N does not exceed $2N$ computations of $F(\cdot)$ and solving $2N$ auxiliary problems (1.6).

Proof. Let (2.1) be the case. All we need to verify is that if, for a given t, (3.4) is not met with $s_t = 1$, it definitely is met with $s_t = 2$. Indeed, let us set $z = z_{t-1}$, $\xi = F(z)$, $w = w_{t,1}$, $\eta = F(w)$, $z_+ = w_{t,2}$ and $\gamma = \gamma_t$; let also $U = Z$. Comparing the description of Basic Implementation and (3.6), we see that our (z, w, z_+, γ, U) are exactly as required by premise of Lemma 3.1, so that by this Lemma we have

$$\begin{array}{l}
\langle \gamma F(w_{t,1}), w_{t,1} - w_{t,2}\rangle + [\omega(z_{t-1}) + \langle \omega'(z_{t-1}), w_{t,2} - z_{t-1}\rangle - \omega(w_{t,2})] \\
\qquad \equiv \langle \gamma F(w), w - z_+\rangle + [\omega(z) + \langle \omega'(z), z_+ - z\rangle - \omega(z_+)] \\
\qquad \leq \alpha^{-1}\gamma^2 \|F(z) - F(w)\|_*^2 - \frac{\alpha}{2}\left[\|w - z\|^2 + \|w - z_+\|^2\right].
\end{array}$$

The concluding quantity in this chain, by (2.1), is $\leq \left[\alpha^{-1}\gamma^2 L^2 - \frac{\alpha}{2}\right]\|z-w\|^2$, which is just zero by (3.2). Thus, condition (3.4) indeed is met after at most 2 inner iterations. □

4. Extensions.

The case of non-Lipschitzean monotone operator. In fact, we can handle monotone operators of various degrees of continuity:

THEOREM 4.1. *Assume that F is monotone and satisfies the relation*

$$\|F(z) - F(z')\|_* \leq L\|z - z'\|^\sigma \quad \forall z, z' \in Z, \tag{4.1}$$

where $L < \infty$ and $\sigma \in [0, 1]$. Note that the case of $\sigma = 0$ corresponds to bounded monotone operator.

Consider algorithm

$$\begin{array}{rcl}
z_{t-1} \mapsto w_t &=& P_{z_{t-1}}(\gamma_t F(z_{t-1})) \equiv \operatorname*{argmin}_{u\in Z} \left[\langle \gamma_t F(z_{t-1}) - \omega'(z_{t-1}), u\rangle + \omega(u)\right] \\
w_t \mapsto z_t &=& P_{z_{t-1}}(\gamma_t F(w_t)) \equiv \operatorname*{argmin}_{u\in Z} \left[\langle \gamma_t F(w_t) - \omega'(z_{t-1}), u\rangle + \omega(u)\right]
\end{array} \tag{4.2}$$

with the stepsizes

$$\gamma_t = 0.7L^{-1}\left(\frac{\Theta[\omega]}{t}\right)^{\frac{1-\sigma}{2}} \alpha^{\frac{1+\sigma}{2}}, \quad \Theta[\omega] = \max_{z_0\in Z}\Theta(z_0), \tag{4.3}$$

and the following rule for generating approximate solutions:

$$z^t = \left(\sum_{\tau=\rfloor t/2\lfloor}^{t} \gamma_\tau\right)^{-1} \sum_{\tau=\rfloor t/2\lfloor}^{t} \gamma_\tau w_\tau.$$

Then for every N one has

$$\epsilon(z^N) \le O(1)L\left(\frac{\Theta[\omega]}{\alpha N}\right)^{\frac{1+\sigma}{2}} \tag{4.4}$$

with an absolute constant $O(1)$. In the saddle point case, $\epsilon(z^N)$ in this bound can be replaced with $\epsilon_f(z^N)$.

Proof. Applying Lemma 3.1 with $U = Z$, we conclude that Algorithm (4.2) guarantees (2.11) with

$$\begin{array}{rcl}\epsilon_t &=& \alpha^{-1}\gamma_t^2\|F(z_{t-1}) - F(w_t)\|_*^2 - \frac{\alpha}{2}\left[\|w_t - z_{t-1}\|^2 + \|w_t - z_t\|^2\right]\\ &\le& \alpha^{-1}\gamma_t^2 L^2\|z_{t-1} - w_t\|^{2\sigma} - \frac{\alpha}{2}\|w_t - z_{t-1}\|^2\\ &\le& \sup\limits_{d\ge 0}\left[\alpha^{-1}\gamma_t^2 L^2 d^\sigma - \frac{\alpha}{2}d\right],\end{array}$$

and it is immediately seen that in the case of (4.3), the latter quantity is $\le \frac{\Theta[\omega]}{t}$. Applying (2.13) to the segment of iterations from $\lfloor N/2\rfloor$ to N, we get

$$\epsilon(z^N) \le \frac{\Theta[\omega]\left(1 + \sum\limits_{t=\rfloor N/2\lfloor}^{N} t^{-1}\right)}{\sum\limits_{t=\rfloor N/2\lfloor}^{N}\gamma_t} \le O(1)L\left(\frac{\Theta[\omega]}{\alpha N}\right)^{\frac{1+\sigma}{2}}.$$

□

"Bundle" version of Algorithm (4.2). In this version, one keeps in memory and utilizes not only the latest value of F, but past values as well. The algorithm is as follows:

A) At the beginning of step t we have in our disposal previous iterates $z_0, z_1, ..., z_{t-1}$ $w_1, w_2, ..., w_{t-1}$ (the set of all these iterates is denoted by W_{t-1}) along with a set

$$U_{t-1} = \left\{u \in Z : h_j^{t-1}(u) \ge 0, i \in J_{t-1}\right\},$$

which intersects the relative interior of Z. Here $h_j^{t-1}(\cdot)$ are combinations, with nonnegative coefficients, of the affine functions

$$\psi_v(u) \equiv \langle F(v), v - u\rangle,\ v \in W_{t-1}.$$

B) At step t, we set

$$\begin{array}{rcl}w_t &=& \operatorname*{argmin}\limits_{y\in U_{t-1}}\left[\langle\gamma_t F(z_{t-1}) - \omega'(z_{t-1}), y\rangle + \omega(y)\right]\\ z_t &=& \operatorname*{argmin}\limits_{y\in U_{t-1}}\left[\langle\gamma_t F(w_t) - \omega'(z_{t-1}), y\rangle + \omega(y)\right]\\ \epsilon_t &=& \langle\gamma_t(F(w_t) - F(z_{t-1})), w_t - z_t\rangle\\ && +\left[\omega(z_{t-1}) + \langle\omega'(z_{t-1}), w_t - z_{t-1}\rangle + \langle\omega'(w_t), z_t - w_t\rangle - \omega(z_t)\right]\end{array} \tag{4.5}$$

C) Applying Lemma 3.1 to $U = U_{t-1}$, $z = z_{t-1}$, $\xi = F(z_{t-1})$, $\eta = F(w_t)$, $\gamma = \gamma_t$, we get

$$\begin{array}{l}\max\limits_{u\in U_{t-1}}\left[\langle\gamma_t F(w_t), w_t - u\rangle - H_u(z_{t-1}) + H_u(z_t)\right] \le \epsilon_t\\ \qquad \le \alpha^{-1}\gamma_t^2\|F(z_{t-1}) - F(w_t)\|_*^2 - \frac{\alpha}{2}\left[\|w_t - z_{t-1}\|^2 + \|w_t - z_t\|^2\right].\end{array} \tag{4.6}$$

Recalling the structure of U_{t-1}, there exist nonnegative weights $\{\lambda_v^t\}_{v\in W_{t-1}}$ such that

$$\max_{u\in Z}\left[\langle \gamma_t F(w_t), w_t-u\rangle - H_u(z_{t-1}) + H_u(z_t) + \sum_{v\in W_{t-1}} \lambda_v^t \langle F(v), v-u\rangle\right] \le \epsilon_t,$$

whence

$$\begin{array}{c}\forall u\in Z: \langle \gamma_t F(w_t), w_t-u\rangle + \sum\limits_{v\in W_{t-1}} \lambda_v^t\langle F(v), v-u\rangle \le H_u(z_{t-1}) - H_u(z_t) + \epsilon_t,\\ \epsilon_t \le \alpha^{-1}\gamma_t^2\|F(z_{t-1}) - F(w_t)\|_*^2 - \frac{\alpha}{2}\left[\|w_t - z_{t-1}\|^2 + \|w_t - z_t\|^2\right].\end{array} \tag{4.7}$$

We compute the weights $\{\lambda_v^t\}_{v\in W_{t-1}}$, set

$$\Gamma_t = \gamma_t + \sum_{v\in W_{t-1}} \lambda_v^t, \tag{4.8}$$

choose U_t in accordance with the above restrictions and pass to step $t+1$.

D) The approximate solution z^t obtained in course of t steps is

$$z^t = \left(\sum_{\tau=\lfloor t/2\rfloor}^{t} \Gamma_\tau\right)^{-1} \sum_{\tau=\lfloor t/2\rfloor}^{t}\left[\gamma_\tau w_\tau + \sum_{v\in W_{\tau-1}} \lambda_v^\tau v\right].$$

THEOREM 4.2. *Let F be monotone. For the outlined algorithm, one has*

$$\epsilon(z^N) \le \underbrace{\left(\frac{\sum\limits_{t=\lfloor N/2\rfloor}^{N}\gamma_t}{\sum\limits_{t=\lfloor N/2\rfloor}^{N}\Gamma_t}\right)}_{\chi\le 1} \frac{\Theta[\omega] + \sum\limits_{\tau=\lfloor N/2\rfloor}^{N}\epsilon_t}{\sum\limits_{t=\lfloor N/2\rfloor}^{N}\gamma_t}. \tag{4.9}$$

In particular, if F satisfies (4.1) *and γ_t are chosen according to* (4.3)*, bound* (4.9) *becomes*

$$\epsilon(z^t) \le O(1)\chi L\left(\frac{\Theta[\omega]}{\alpha t}\right)^{\frac{1+\sigma}{2}} \tag{4.10}$$

In the case when F comes from a game with convex-concave cost function $f(x,y)$, the error measure $\epsilon(\cdot)$ in the above bounds can be replaced with $\epsilon_f(\cdot)$.

Proof. Error bound (4.9) can be derived from (4.7) exactly in the same fashion as the error bounds in Proposition 2.2 were derived from (2.15). Relation $\chi\le 1$ follows from $\Gamma_t \ge \gamma_t$ (see (4.8) and note that $\lambda_v^t \ge 0$). Same as in the proof of Theorem 4.1, in the case of (4.1), (4.3) we have $\epsilon_t \le t^{-1}\Theta[\omega]$, and therefore (4.9) implies (4.10). □

Note that the bundle version of Algorithm (4.2) is more computationally demanding than its prototype; as a compensation, it has a better theoretical efficiency estimate (since $\chi \le 1$).

5. Examples. We are about to list a number of interesting particular cases of the above construction.

Euclidean setup. In this case, $\omega(z) = \frac{1}{2}\langle z, z\rangle \equiv \frac{1}{2}\|z\|_2^2$ and $\|\cdot\| = \|\cdot\|_2$, which results in $\alpha = 1$, $\Theta[\omega] = \frac{1}{2}\max\limits_{z',z''\in Z}\|z' - z''\|_2^2$. The Basic Implementation here becomes

$$
\begin{array}{rcll}
w_t &=& \Pi_Z(z_{t-1} - \gamma F(z_{t-1})), & (a)\\
z_t &=& \Pi_Z(z_{t-1} - \gamma F(w_t)), & (b)\\
z^t &=& \frac{1}{t}\sum\limits_{\tau=1}^{t} w_t & (c)
\end{array}
\tag{5.1}
$$

where

- $\Pi_Z(u) = \operatorname*{argmin}\limits_{v\in Z}\|u - v\|_2$ is the projector onto Z,
- $\gamma = \frac{1}{\sqrt{2}L}$, $L \geq \sup\limits_{z,z'\in Z, z\neq z'}\frac{\|F(z)-F(z')\|_2}{\|z-z'\|_2}$.

The efficiency estimates (3.9), (3.10) become, respectively,

$$
\epsilon(z^N) \leq \frac{\sqrt{2}LD^2[Z]}{N}, \quad \epsilon_f(z^N) \leq \frac{\sqrt{2}LD^2[Z]}{N}, \tag{5.2}
$$

where $D[Z] = \max\limits_{z,z'\in Z}\|z - z'\|_2$ is the $\|\cdot\|_2$-diameter of Z.

Note that in order for the method to be implementable, the projector $\Pi_Z(\cdot)$ onto Z should be easy to compute, that is, Z should be "simple", like Euclidean ball, a box, the standard simplex $\Delta_n = \{x \in \mathbf{R}^n : x \geq 0, \sum\limits_x = 1\}$. A less trivial example is "matrix box" defined as follows:

- E is the space $\mathbf{S}^{(d_1,\dots,d_m)}$ of block-diagonal symmetric matrices x with $d_i \times d_i$ diagonal blocks $x[i]$, $i = 1, ..., m$, equipped with the Frobenius inner product $\langle x, y\rangle = \mathrm{Tr}(xy)$;
- $Z = \{x : |x[i]|_\infty \leq a_i,\ i = 1, ..., m\}$, where $|u|_\infty$ is the usual spectral norm (i.e., the maximal singular value) of a matrix.

Method (5.1.$a - b$) is nothing but the *extragradient method* for variational inequalities proposed by G. Korpelevich in 1976, (see [7, 8]; for recent results, see [15] and references therein); in particular, she proved the convergence of z_t to a solution of the VI. Bounds (5.2) for the ergodic version of this method are, to the best of our knowledge, new.

Note that the rate of convergence exhibited by the prox-method is, in a sense, optimal. Specifically, given $D, L > 0$, consider the case when Z is the direct product of two n-dimensional Euclidean balls X, Y of radii D and $f(x, y) = y^T(Ax - b)$ for a symmetric operator A with operator norm not exceeding L, and b is such that the operator equation $Ax = b$ has a solution in X. The corresponding saddle point problem $\max\limits_{y\in Y} f(x, y) = D\|Ax - b\|_2$ is associated with an affine Lipschitz continuous, with constant L with respect to $\|\cdot\| = \|\cdot\|_2$, monotone operator. Let $\mathcal{B}$ be an arbitrary "first order" method for solving saddle point problems of the outlined type, i.e., a method which is given b, X, Y in advance, while having no direct access to A; however, $\mathcal{B}$ has an access to an "oracle" which, given on input a vector x, returns the vector Ax. It turns out [9] that whenever $N \leq \frac{1}{4}n$ and $\mathcal{B}$ is a first order method which is allowed to make no more than N calls to the oracle, there is a problem in the family such that the result z obtained by $\mathcal{B}$ as applied to the problem satisfies the relations $\epsilon(z) \geq O(1)\frac{LD^2}{N}$, $\epsilon_f(z) \geq O(1)\frac{LD^2}{N}$. Thus, when Z is the direct product of two n-dimensional Euclidean balls and the dimension is large ($n > N$), bounds (5.2) cannot be improved by more than an absolute constant factor.

Spectahedron setup. In this case,

- $E = \mathbf{S}^{(d_1,...,d_m)}$,
- $Z = \{z \in E : z \succeq 0, \mathrm{Tr}(z) = 1, z[i] \preceq a_i I_{d_i}, \ i = 1, ..., m\}$,
- $\omega(\cdot)$ is the "regularized matrix entropy"

$$\omega(z) = \sum_{i=1}^{n} (\lambda_i(z) + n^{-1}\delta) \ln(\lambda_i(z) + n^{-1}\delta), \qquad n = \sum_{\ell=1}^{m} k_\ell, \tag{5.3}$$

  where $\delta \in (0, 1]$ is a once for ever fixed regularization parameter (say, $\delta = 10^{-16}$) and $\lambda_j(z)$, $j = 1, ..., n \equiv d_1 + ... + d_m$, are the eigenvalues of $z \in E$;
- $\|\cdot\|$ is the norm $|z|_1 = \sum_j |\lambda_j(z)|$ (so that $\|\cdot\|_*$ is the usual matrix norm $|z|_\infty$).

Note that in the case of $d_i = 1$, $i = 1, ..., m$, Z is just the truncated simplex $\{x \in \mathbf{R}^m : 0 \le x_i \le b_i, \sum_i x_i = 1\}$.

It is known (see, e.g., [2]) that for the Spectahedron setup one has

$$\alpha \ge \frac{1}{2}, \ \Theta[\omega] \le 4\ln(n/\delta) = O(1)\ln n \tag{5.4}$$

so that the efficiency estimates (2.13), (2.14) for the Basic implementation become, respectively,

$$\epsilon(z^N) \le O(1)\frac{L\ln(n)}{N}, \quad \epsilon_f(z^N) \le O(1)\frac{L\ln(n)}{N}, \tag{5.5}$$

where L is an a priori upper bound on the "$1, \infty$" Lipschitz constant

$$L_* = \sup_{z,z' \in Z, z \ne z'} \frac{|F(z) - F(z')|_\infty}{|z - z'|_1}$$

of the operator F.

Note that basically the only computation at a step of our prox-method, aside of computing F at a point, is computing $P_z(\xi)$ for given z and ξ, that is, solving the optimization problem of the form

$$\min_{z \in Z} \{\omega(z) + \langle p, z\rangle\}. \tag{5.6}$$

It is easily seen that solving the latter problem reduces to

1) computing the eigenvalue decompositions $p[i] = V_i \mathrm{Diag}\{\pi_{i1}, ..., \pi_{id_i}\} V_i^T$ of the diagonal blocks $p[i]$ of p;

2) finding the solution ξ^* to the problem

$$\min_\xi \left\{ \sum_{i=1}^{m}\sum_{j=1}^{d_i} \left[(\xi_{ij} + n\delta^{-1})\ln(\xi_{ij} + n\delta^{-1}) + \pi_{id_i}\xi_{ij}\right] : 0 \le \xi_{ij} \le a_i, \sum_{i=1}^{m}\sum_{j=1}^{d_i} \xi_{ij} = 1 \right\}; \tag{5.7}$$

3) recovering the optimal solution z^* to (5.6) as $z^*[i] = V_i \mathrm{Diag}\left\{\xi^*_{i1}, ..., \xi^*_{id_i}\right\} V_i^T$, $i = 1, ..., m$.

Now, 3) is easy. 2) also is easy: $\xi^*_{ij} = f_{ij}(\lambda_*)$, where

$$f_{ij}(\lambda) = \Pi_{[0,a_i]}(\underbrace{\exp\{\lambda - \pi_{ij}\} - n^{-1}\delta}_{s_{ij}(\lambda)})$$

is the closest to $s_{ij}(\lambda)$ point in the segment $[0, a_i]$, and λ_* is the root of the equation

$$\sum_i \sum_j f_{ij}(\lambda) = 1. \tag{5.8}$$

This equation can be easily solved by bisection; as a result, it takes just $O(n)$ operations to implement 2) within machine precision. Note also that when δ is less than machine zero, from computer's viewpoint $f_{ij}(\lambda) = \min\left[a_i, \exp\{\lambda - \pi_{ij}\}\right]$; if, in addition, there is no truncation (i.e., $a_i = 1$ for all i), computing ξ^*_{ij} requires no bisection, since within machine precision one has

$$\xi^*_{ij} = \frac{\exp\{-\pi_{ij}\}}{\sum\limits_r \sum\limits_s \exp\{-\pi_{rs}\}}.$$ [3)]

We see that the only "nontrivial" operation in solving (5.6) is 1). This operation is "numerically tractable", when the sizes d_i of blocks in $z \in Z$ are within few hundreds, and is easy, when these sizes are small integers; in particular, with $d_i = O(1)$ solving (5.6) within machine precision costs just $O(d_1 + ... + d_m)$ operations.

Mixed setups. In some applications (e.g., in those we are about to consider), E is the direct product of Euclidean spaces E_k, $k = 1, ..., K$, Z is the direct product of "standard" sets $Z_k \subset E_k$ (simplexes, boxes, spectahedrons,...), and we know what are "good setup parameters" $\omega_k(\cdot), \|\cdot\|_k$ for every one of the factors; the question is how to "assemble" these entities into a "reasonably good" setup for a variational inequality on $Z = Z_1 \times ... \times Z_K$. When answering this question, we assume that we know the quantities $\Theta_k = \Theta[\omega_k]$, the parameters α_k of strong convexity of $\omega_k(\cdot)\big|_{Z_k}$ w.r.t. $\|\cdot\|_k$, and upper bounds L_{pq} on "partial Lipschitz constants"

$$L^*_{pq} = \max_{z,z' \in Z}\left\{\frac{\|F_p(z) - F_p(z')\|^*_p}{\|z_q - z'_q]\|_q} : z_\ell = z'_\ell, \ell \neq q, z_q \neq z'_q\right\}$$

of the monotone operator F in question. Here u_p is the natural projection of u on E_p, and $\|\cdot\|^*_k$ is the norm on E_k which is conjugate to $\|\cdot\|_k$. To simplify our considerations, assume that $L_{pq} = L_{qp}$ [4)]. Let us look at "assemblings" of the form

$$\omega(z) = \sum_{k=1}^K \gamma_k \omega_k(z_k), \quad \|z\| = \sqrt{\sum_{k=1}^K \mu_k^2 \|z_k\|_k^2} \quad \left[\Leftrightarrow \|z\|_* = \sqrt{\sum_{k=1}^K \mu_k^{-2} [\|z_k\|^*_k]^2}\right] \tag{5.9}$$

where $\mu_k > 0, \gamma_k > 0$ are parameters of the construction. Note that auxiliary problems (1.6) associated with $Z = Z_1 \times ... \times Z_K$ and $\omega(\cdot)$ given by (5.9) are easy, provided that this is the case for similar problems associated with every pair $Z_k, \omega_k(\cdot)$. Further, we can easily express the (natural bounds on the) quantities $\Theta[\omega]$, α, and L (the

---

[3)] This explains the computational advantages of regularized entropy with δ of order of machine zero as compared to other functions ensuring (5.4), like $\sum\limits_i \lambda_i^p(x)$ with appropriately chosen p, specifically, $p = 1 + O(1)/ln(n)$; with the latter choice of ω, computing $P_z(\cdot)$ would always require a bisection to solve the corresponding version of (5.8).

[4)] Note that in the saddle point case ($Z = X \times Y$, Z_k are direct factors in either X, or Y, and $F_k(z) = \epsilon_k(\nabla f(z))_k$ with $\epsilon_k = \pm 1$), which is the case of our primary interest, one has $L^*_{k\ell} = L^*_{\ell k}$, so that the assumption $L_{k\ell} = L_{\ell k}$ is fully justified.

Lipschitz constant L of F w.r.t. $\|\cdot\|$) in terms of μ, γ; a straightforward computation demonstrates that

$$(5.10)\quad \Theta[\omega] \leq \widetilde{\Theta} = \sum_k \gamma_k \Theta[\omega_k],\ \ \alpha \geq \widetilde{\alpha} = \min_k \frac{\gamma_k \alpha_k}{\mu_k^2},\ \ L \leq \widetilde{L} = |\ \left[\mu_k^{-1}\mu_\ell^{-1} L_{k\ell}\right]|_\infty.$$

Now, what matters for the complexity bound of our method, the setup being given by $\omega(\cdot)$, $\|\cdot\|$, is the quantity $\alpha^{-1}\Theta L$ (the less it is, the better), and it is natural to look for the assembling which results in the smallest possible upper bound $\widetilde{\alpha}^{-1}\widetilde{\Theta}\widetilde{L}$ on this quantity. This problem can be easily solved; an optimal solution is given by (5.9) with the parameters γ_k, μ_k given by

$$(5.11)\qquad M_{k\ell} := L_{k\ell}\sqrt{\frac{\Theta_k \Theta_\ell}{\alpha_k \alpha_\ell}},\ \ \sigma_k := \frac{\sum_\ell M_{k\ell}}{\sum_{p,\ell} M_{p\ell}},\ \ \gamma_k = \frac{\sigma_k}{\Theta_k},\ \ \mu_k = \sqrt{\gamma_k \alpha_k}.$$

For the resulting assembling, one has

$$(5.12)\qquad \widetilde{\alpha} = \widetilde{\Theta} = 1, \quad \widetilde{L} = \sum_{k,\ell} L_{k\ell}\sqrt{\frac{\Theta_k \Theta_\ell}{\alpha_k \alpha_\ell}}$$ [5)]

By Theorem 3.2, the efficiency estimate for the resulting algorithm is

$$(5.13)\qquad \epsilon(z^N) \leq \frac{\sqrt{2}\sum_{k,\ell} L_{k\ell}\sqrt{\frac{\Theta_k\Theta_\ell}{\alpha_k\alpha_\ell}}}{N}$$

(in the saddle point case, $\epsilon(\cdot)$ can be replaced with $\epsilon_f(z^N)$).

Example 1: Matrix Game. Assume we are interested to find a saddle point of a bilinear function $x^T A y$, $x \in \mathbf{R}^p$, $y \in \mathbf{R}^q$, on the product Z of two standard simplexes $\Delta_p = \{x \geq 0 : \sum_i x_i = 1\}$ and $\Delta_q = \{y \geq 0 : \sum_j y_j = 1\}$. Consider the prox-method with mixed setup corresponding to Spectahedron setups for the factors. The monotone operator in question is $F(x,y) = \begin{bmatrix} Ay \\ -A^T x \end{bmatrix}$, so that $L_{11} = L_{22} = 0$, while $L_{12} = L_{21}$ is the Lipschitz constant of the mapping $x \mapsto A^T x$ considered as a mapping from $(\mathbf{R}^p, \|\cdot\|_1)$ to $(\mathbf{R}^q, \|\cdot\|_\infty \equiv \|\cdot\|_1^*)$, so that $L_{12} = L_{21} = \max_{i,j}|A_{ij}|$. Applying (5.4), (5.13), we see that the algorithm yielded by the mixed setup obeys the efficiency estimate

$$(5.14)\qquad \epsilon_f(z^N) \leq O(1)\frac{\max_{i,j}|A_{ij}|\sqrt{\ln(p+1)\ln(q+1)}}{N}$$

(recall that we treat the entropy regularization parameter δ as an absolute constant). In the Basic implementation of the resulting method, effort per step is dominated by two computations of $F(\cdot)$, that is, by 4 matrix-vector multiplications (two – by A and two – by A^T).

[5)] The fact that (5.11) implies that $\widetilde{\alpha} = \widetilde{\Theta} = 1$ follows immediately from (5.10). To verify that $\lambda \equiv \sum_{k,\ell} L_{k\ell}\sqrt{\frac{\Theta_k\Theta_\ell}{\alpha_k\alpha_\ell}} \equiv \sum_{k,\ell} M_{k\ell} \geq \widetilde{L} \equiv |\left[\mu_k^{-1}\mu_\ell^{-1}L_{k\ell}\right]|_\infty \equiv |\left[M_{k\ell}/\sqrt{\sigma_k\sigma_\ell}\right]|_\infty$, note that the matrix $\lambda \mathrm{Diag}\{\sigma_1, ..., \sigma_m\} - [M_{k\ell}]$ is diagonal-dominated and therefore is $\succeq 0$, whence $\lambda I_m - \left[M_{k\ell}/\sqrt{\sigma_k\sigma_\ell}\right] \succeq 0$ as well.

Example 2: Semidefinite Programming. Matrix game considered in Example 1 can be interpreted as follows: we are given p *diagonal* matrices A_i (the diagonal entries of A_i form i-th row in the game matrix A); the cost function of the first player is $\overline{f}(x) = \max_{y\in\Delta_q} x^T Ay$, or, which is the same, is the largest eigenvalue of the matrix $\sum_i x_i A_i$, so that the first player is looking for a convex combination of the matrices A_i which has the smallest possible maximum eigenvalue. Now consider the latter problem in the case of symmetric matrices $A_i \in \mathbf{S}^{(d_1,...,d_m)}$. This problem again can be posed as a saddle point problem for the bilinear function $f(x,y) = \mathrm{Tr}(y\sum_{i=1}^{p} x_i A_i)$, with x running through the simplex Δ_p and y running through the spectahedron $\Sigma^d = \{y \in \mathbf{S}^{(d_1,...,d_m)} : y \succeq 0, \mathrm{Tr}(y) = 1\}$. Indeed, for $B \in \mathbf{S}^{(d_1,...,d_m)}$, the quantity $\max_{y\in\Sigma^d} \mathrm{Tr}(yB)$ is exactly the maximum eigenvalue $\lambda_{\max}(B)$ of B. Thus, finding a saddle point of $f(x,y)$ on the direct product of $\Delta_p \times \Sigma^d$ is exactly the semidefinite program of minimizing the maximum eigenvalue of a convex combination of matrices $A_1, ..., A_p$. Equipping both factors with Spectahedron setup, we get $\Theta_1 = O(1)\ln(p+1)$, $\Theta_2 = O(1)\ln(q+1)$, $q = d_1 + ... + d_m$, $\alpha_1 = \alpha_2 = O(1)$. The associated monotone operator is $F(x,y) = \begin{bmatrix} (\mathrm{Tr}(A_1 y), ..., \mathrm{Tr}(A_p y))^T \\ -\sum_i x_i A_i \end{bmatrix}$, so that $L_{11} = L_{22} = 0$ and $L_{12} = L_{21}$ is the Lipschitz constant of the mapping $x \mapsto \sum_{i=1}^{p} x_i A_i$ considered as a mapping from $(\mathbf{R}^p, \|\cdot\|_1)$ to $(\mathbf{S}^{(d_1,...,d_m)}, |\cdot|_\infty \equiv |\cdot|_1^*)$, so that $L_{12} = L_{21} = \max_{1\le i\le p} |A_i|_\infty$. Applying (5.4), (5.13), we see that the algorithm yielded by the mixed setup obeys the efficiency estimate

$$\epsilon_f(z^N) \le O(1)\frac{\max_{1\le i\le p} |A_i|_\infty \sqrt{\ln(p+1)\ln(q+1)}}{N}. \tag{5.15}$$

In the Basic implementation of the resulting method, effort per step is dominated by two computations of $F(\cdot)$ and by two eigenvalue decompositions of matrices from $\mathbf{S}^{(d_1,...,d_m)}$.

Example 3. Computing Lovasz capacity. The Lovasz capacity number ϑ of a graph with vertices $1, ..., n$ and an m-element set of arcs V is, by definition, the optimal value in the optimization problem

$$\min_{x\in X} \lambda_{\max}(d+x) = \min_{x\in X}\max_{y\in\Sigma_n} f(x,y), \quad f(x,y) = \mathrm{Tr}([d+x]y), \tag{5.16}$$

where $\Sigma_n = \{x \in \mathbf{S}^n : x \succeq 0, \mathrm{Tr}(x) = 1\}$, $d = \left[d_{ij} = \begin{cases} 0, & (i,j) \in V \\ 1, & (i,j) \notin V \end{cases}\right]$, and X is the set of all symmetric $n \times n$ matrices $x = [x_{ij}]$ such that $x_{ij} = 0$ for $(i,j) \notin V$. Note that for an optimal x the matrix $\vartheta I - [d+x]$ is positive semidefinite, so that nonzero entries in x satisfy the bound $|x_{ij}| \le \vartheta - 1$. It follows that if μ is a valid a priori upper bound on ϑ, then

$$\vartheta = \min_{x\in X_\mu}\max_{y\in\Sigma_n} \mathrm{Tr}([d+x]y), \quad X_s = \{x \in X : |x_{ij}| \le s-1\}. \tag{5.17}$$

Consequently, we can approximate ϑ by our prox-method as applied to the (variational inequality associated with the) saddle point problem (5.17). Equipping X_μ with

Euclidean setup, and Σ_n with the Spectahedron one, we have $\alpha_1 = 1$, $\alpha_2 \geq 1/2$, $\Theta_1 = \frac{1}{2} \max\limits_{x', x'' \in X_\mu} \|x' - x''\|_2^2 = 4m\mu^2$, where m is the number of arcs in the graph, and $\Theta_2 = O(1)\ln(n+1)$. The monotone operator in question is $F(x,y) = \begin{bmatrix} y \\ -x-d \end{bmatrix}$, so that $L_{11} = L_{22} = 0$, while $L_{12} = L_{21}$ is the Lipschitz constant of the mapping $x \mapsto x$ considered as the mapping from $(\mathbf{S}^n, |\cdot|_2)$ to $(\mathbf{S}^n, |\cdot|_\infty \equiv |\cdot|_1^*)$, that is, $L_{12} = L_{21} = 1$. Applying (5.4), (5.13), we see that the algorithm yielded by the mixed setup obeys the efficiency estimate

$$\epsilon_f(z^N) \leq O(1)\frac{\mu\sqrt{m\ln(n)}}{N}. \tag{5.18}$$

For Basic implementation, the effort per step is dominated by the necessity to find eigenvalue decompositions of two matrices from $\mathbf{S}^n$, which requires $O(n^3)$ operations. Note that $\epsilon_f(x^N, y^N)$ is the difference of the quantities $\max\limits_{y\in\Sigma_n} \mathrm{Tr}([d+x^N]y) = \lambda_{\max}(d+x^N)$ (which is an upper bound on ϑ) and the easily computable quantity $\min\limits_{x\in X_\mu} \mathrm{Tr}([d+x]y^N)$ (which is a lower bound on ϑ). Computing these bounds, we localize the Lovasz capacity ϑ in a segment of the length $\epsilon_f(x^N, y^N)$.

We always can take $\mu = n$; with this choice, bound (5.18) implies that in order to approximate ϑ within absolute accuracy ϵ, it suffices to run $O\left(n\sqrt{m\ln(n)}\epsilon^{-1}\right)$ steps of the algorithm. Since the arithmetic cost of a step is $O(n^3)$, we arrive at the overall complexity $O\left(n^4\sqrt{m\ln(n)}\epsilon^{-1}\right) \leq O\left(n^5\sqrt{\ln(n)}\epsilon^{-1}\right)$ operations. Note that a slight modification of the outlined construction (see Section 6) allows to approximate ϑ within relative accuracy ϵ (i.e., within absolute accuracy $\epsilon\vartheta$) in $O\left(n^3\sqrt{m\ln(n)}\epsilon^{-1}\right) \leq O\left(n^4\sqrt{\ln(n)}\epsilon^{-1}\right)$ operations. It is instructive to compare these complexity estimates with those for Interior Point methods: for an n-node graph with no specific structure and the number arcs in both the graph and its complement of order of n^2, the arithmetic cost of a *single* iteration in an IP method is $O(n^6)$[6]. Thus, when approximating Lovasz capacity with fixed absolute (relative) accuracy, the prox method outperforms the IP ones by factors of order of $O\left(n/\sqrt{\ln n}\right)$, (respectively, $O\left(n^2/\sqrt{\ln n}\right)$).

In the examples above, we have dealt with optimization problems which can be reformulated as saddle point problems for *bi-affine* functions. An application which goes beyond this specific situation is minimization of the maximum of smooth (with Lipschitz continuous gradients) convex functions. Indeed, the problem

$$\min_{x\in X} \max_{1\leq \ell \leq m} f_\ell(x)$$

is nothing but the saddle point problem

$$\min_{x\in X} \max_{y\in\Delta_m} f(x,y), \quad f(x,y) = \sum_{\ell=1}^m y_\ell f_\ell(x).$$

[6] Indeed, under our assumptions both the semidefinite program $\min\limits_{\theta, x\in X}\{\theta : \theta I \succeq d + x\}$ defining ϑ and the dual of this program have $O(n^2)$ design variables, so that the cost of a Newton step is $O(n^6)$.

When f_ℓ are smooth on X, the convex-concave function $f(x,y)$ possesses Lipschitz continuous gradient, and the associated saddle point problem can be solved by our prox-method.

6. Numerical illustration. Below we present numerical results obtained with our prox-method on matrix games (Example 1) and Lovasz capacity problem (Example 3), with the setups described in the previous section. Our implementation was, essentially, the Basic one with the only modification: instead of using all the time fixed "theoretically safe" stepsizes $\gamma_t = \bar{gamma} \equiv \frac{\alpha}{\sqrt{2}L}$, a simple policy for *on-line adjusting* the stepsizes was used. Specifically, whenever at a step of the method the termination condition (3.4) was met after at most 2 inner iterations, the starting value of γ at the next step was taken 1.2 times larger than the final value of γ used at the previous step. On the other hand, when in course of a step the termination condition was not met during the first 3 inner iterations, the current value of γ at every subsequent inner iteration was reduced according to $\gamma \mapsto \max[\gamma/2, \bar{\gamma}]$. The regularization parameter δ in (5.3) was set to 1.e-16.

Matrix games. We dealt with "square" matrix games $\min\limits_{x\in\Delta_p}\max\limits_{y\in\Delta_p} f_A(x,y)$, $f_A(x,y) = x^TAy$ with sparse matrices A generated as follows. We first chose at random cells ij with nonzero entries A_{ij}, with a given probability κ to choose a particular cell. In the chosen cells, the values of A_{ij} were picked at random from the uniform distribution on $[-1,1]$. Periodically we measured the "actual accuracy"

$$\epsilon_f(z^t) = \max_{y\in\Delta_q}(x^t)^TAy - \min_{x\in\Delta_p} x^TAy^t = \max_{1\le j\le q}(A^Tx^t)_j - \min_{1\le i\le p}(Ay^t)_i$$

of current approximate solution $z^t = (x^t, y^t)$.

Note that with $|A_{ij}| \le 1$, the efficiency estimate (5.14) implies that

$$N(\epsilon) \equiv \min\{N : \epsilon_f(x^N, y^N) \le \epsilon\} \le \frac{16\ln(2p)}{\epsilon}. \tag{6.1}$$

The numerical results are presented in Table 1. We see that qualitatively speaking, the actual performance of the algorithm obeys the complexity bound (6.1) (see the values of the products $\#F\cdot\epsilon$). At the same time, the "empirical" constants C in the bound $\epsilon_f(x^t,y^t) \approx \frac{C}{t}$ are essentially better than the constant $16\ln(2p)$ in (6.1); moreover, these constants seem to decrease as p grows. This phenomenon is in full accordance with surprisingly nice behaviour of our simple on-line adjustment of the stepsizes: the "theoretically safe" value of them is about 0.125, while the averaged, over t, empirical value varies from ≈ 3 for the smallest example to ≈ 400 for the largest one. Note that even those "large" values of γ_t still allow to keep the average number of computations of F per step well below 3. Of course, we do not pretend to consider our experiments as conclusive – the nice picture we observe perhaps reflects specific features of randomly generated game problems.

Computing Lovasz capacity. In order to work with "meaningful" upper bound μ on ϑ, the execution was split into subsequent *stages*. At every stage, we applied the method to the saddle point problem (5.17) with the upper bound on the Lovasz capacity ϑ, coming from the previous stage, in the role of μ (at the very first stage, μ was set to the number n of nodes). A stage was terminated when the current upper bound on ϑ became $< \mu/2$. It is easily seen that with this implementation of the prox-method, the total number of steps to approximate ϑ within *relative* accuracy ϵ is, up to logarithmic factors, $O(1)n^3\sqrt{m} \le O(1)n^4$, where m is the number of arcs.

p / κ		$t=1$	$t=32$	$t=64$	$t=128$	$t=256$	$t=512$	$t=1024$	$t=2048$	CPU sec
100	ϵ	2.5e-1	3.7e-2	1.6e-2	7.6e-3	3.7e-3	1.8e-3	8.6e-4	4.3e-4	12
1.0	$\frac{\epsilon}{\epsilon_1}$	1.0	1.4e-1	6.2e-2	3.0e-2	1.5e-2	7.0e-3	3.4e-3	1.7e-3	
	$\#F$	2	72	144	296	592	1184	2376	4752	
	$\#F\cdot\epsilon$	5.1e-1	2.6	2.3	2.3	2.2	2.1	2.0	2.0	
	$\bar{\gamma}$	0.5	2.7	3.0	3.2	3.2	3.3	3.3	3.3	
500	ϵ	7.4e-2	1.4e-2	5.1e-3	2.2e-3	1.0e-3	4.8e-4	2.4e-4	1.2e-4	62
0.2	$\frac{\epsilon}{\epsilon_1}$	1.0	1.9e-1	7.0e-2	3.0e-2	1.4e-2	6.5e-3	3.2e-3	1.6e-3	
	$\#F$	2	68	145	293	589	1185	2373	4753	
	$\#F\cdot\epsilon$	1.5e-1	9.7e-1	7.5e-1	6.5e-1	6.0e-1	5.7e-1	5.7e-1	5.7e-1	
	$\bar{\gamma}$	0.5	8.3	11.5	13.0	13.9	14.3	14.5	14.6	
1000	ϵ	4.0e-2	9.2e-3	2.9e-3	1.2e-3	5.5e-4	2.7e-4	1.3e-4	6.5e-5	144
0.1	$\frac{\epsilon}{\epsilon_1}$	1.0	2.3e-1	7.3e-2	2.9e-2	1.4e-2	6.8e-3	3.3e-3	1.6e-3	
	$\#F$	2	68	140	292	588	1180	2372	4748	
	$\#F\cdot\epsilon$	8.0e-2	6.3e-1	4.1e-1	3.4e-1	3.2e-1	3.2e-1	3.2e-1	3.1e-1	
	$\bar{\gamma}$	0.5	14.3	23.4	27.2	28.4	29.2	29.6	29.9	
10,000	ϵ	3.3e-3	2.9e-3	7.9e-4	2.7e-4	9.5e-5	3.8e-5	1.5e-5	6.6e-6	1607
5.0e-3	$\frac{\epsilon}{\epsilon_1}$	1.0	8.8e-1	2.4e-1	8.2e-2	2.9e-2	1.2e-2	4.6e-3	2.0e-3	
	$\#F$	2	64	138	285	578	1169	2355	4732	
	$\#F\cdot\epsilon$	6.6e-3	1.9e-1	1.1e-1	7.7e-2	5.5e-2	4.5e-2	3.5e-2	3.1e-2	
	$\bar{\gamma}$	0.5	26.6	106.4	135.6	181.3	241.3	289.4	335.9	
20,000	ϵ	1.8e-3	1.8e-3	4.9e-4	1.5e-4	7.8e-5	2.9e-5	1.2e-5	5.3e-6	3566
2.5e-3	$\frac{\epsilon}{\epsilon_1}$	1.0	9.7e-1	2.7e-1	8.5e-2	4.3e-2	1.6e-2	6.9e-3	2.9e-3	
	$\#F$	2	64	139	288	578	1166	2342	4704	
	$\#F\cdot\epsilon$	3.6e-3	1.1e-1	6.7e-2	4.4e-2	4.5e-2	3.4e-2	2.9e-2	2.5e-2	
	$\bar{\gamma}$	0.5	26.6	175.0	256.3	251.2	333.4	383.4	445.7	

TABLE 6.1

Experiments with matrix games. $p = \dim x = \dim y$; κ – density of nonzeros in the game matrix; $\epsilon = \epsilon_f(x^t, y^t)$; $\epsilon_1 = \epsilon_f(x_0, y_0)$; $\#F$ – total number of computations of $F(\cdot)$ in course of steps $1, ..., t$; $\bar{\gamma} = t^{-1}\sum_{\tau=1}^{t}\gamma_\tau$. MATLAB code was run on Pentium IV 1.3 GHz PC.

Note also that every pair $(x, y) \in X \times Y$ (see (5.16)) produces a lower and an upper bound on ϑ. The upper bound, of course, is merely $\lambda_{\max}(d + x)$, while the lower bound is obtained as follows. As we remember, the x-component of a saddle point of $f(\cdot,\cdot)$ on $X \times \Sigma_n$ is a matrix from $X_{\vartheta-1}$, whence

$$\begin{array}{rcl}\vartheta & = & \min\limits_{x'\in X_{\vartheta-1}} \max\limits_{y'\in\Sigma_n} \mathrm{Tr}([d+x']y') \geq \min\limits_{x'\in X_{\vartheta-1}} \mathrm{Tr}([d+x']y) \\ & = & \mathrm{Tr}(dy) - (\vartheta - 1)\sum\limits_{(i,j)\in V}|y_{ij}| \\ \Rightarrow \vartheta & \geq & \dfrac{\mathrm{Tr}(dy)+\sum\limits_{(i,j)\in V}|y_{ij}|}{1+\sum\limits_{(i,j)\in V}|y_{ij}|}.\end{array}$$

Running the method, we computed the outlined upper and lower bounds on ϑ given by iterates (x_t, y_t) and approximate solutions (x^t, y^t), thus getting current "best observed" bounds. Computations were terminated when the difference between the "best observed so far" upper and lower bounds on ϑ became < 1. We have run the method on 6 randomly generated graphs and 6 Hamming graphs; the results are presented in Tables 2, 3.

REFERENCES

	Sizes of graph (n,m)					
	(50,616)	(100,2459)	(200,4918)	(300,11148)	(400,20006)	(500,62230)
$[\underline{\vartheta},\overline{\vartheta}]$	$[7.47, 8.46]$	$[10.08, 11.07]$	$[27.46, 28.42]$	$[34.17, 35.15]$	$[39.34, 40.32]$	[22.19,23.17]
$\#F$	527	738	1003	3647	2067	1867
CPU sec	6	52	515	6608	9580	5215

Table 6.2

Experiments with Lovasz capacity, random graphs. n is the number of nodes, m is the number of arcs; $[\underline{\vartheta},\overline{\vartheta}]$ is the resulting localizer of ϑ; $\#F$ is the total number of eigenvalue decompositions of $n\times n$ matrices in course of running the method.

	Graph and sizes (n,m)					
	$H_5^3(1)$ (243,1215)	$H_5^3(1,2)$ (243,6075)	$H_6^3(1)$ (729,4374)	$H_6^3(1,2)$ (729,26244)	$H_{10}^2(1)$ (1024,5120)	$H_{10}^2(1,2,3,4)$ (1024,197120)
$[\underline{\vartheta},\overline{\vartheta}]$	$[80.59, 81.37]$	$[17.49, 18.13]$	$[242.58, 243.33]$	$[48.25, 49.16]$	$[511.05, 512.02]$	$[17.37, 18.29]$
$\#F$	424	458	616	683	1663	1444
CPU sec	96	104	8752	8145	30154	20084

Table 6.3

Experiments with Lovasz capacity, Hamming graphs. Nodes in $H_d^q(i_1,i_2,...)$ are all d-letter words in q-element alphabet; two nodes are linked by an arc if and only if the Hamming distance between the corresponding words (# of positions where the words differ from each other) belongs to the list $i_1,i_2,...$

[1] Beck, A., Teboulle, M., "Mirror Descent and Nonlinear Projected Subgradient Methods for Convex Optimization" – *OR Letters* **31**, (2003), 167-175.

[2] Ben-Tal, A., Nemirovski, A. (2003) "Non-Euclidean Restricted Memory Level Method for Large-Scale Convex Optimization" – submitted to *Mathematical Programming*.

[3] Bruck, R., "On weak convergence of an ergodic iteration for the solution of variational inequalities with monotone operators in Hilbert space" – *J Math. Anal. Appl.* **v.** 61 (1977), No. 1.

[4] Chen, G., Teboulle, M., "Convergence analysis of a proximal-like minimization algorithm using Bregman functions" – *SIAM Journal on Optimization* **v.** 3 (1993), 538-543.

[5] Cohen, G., "Auxiliary problem principle and decomposition of optimization problems", *Journal of Optimization Theory and Applications* **v.** (1980), 277-305.

[6] Kiwiel, K., "Proximal minimization methods with generalized Bregman distances" – *SIAM Journal on Control and Optimization* **v.** 35 (1997), 1142-1168.

[7] Korpelevich, G., "The extragradient method for finding saddle points and other problems" – *Ekonomika i Matematicheskie Metody* **v.** 12 (1976), 747-756 (in Russian; English translation in *Matekon*).

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

[9] Nemirovskii, A., "Information-based complexity of linear operator equations" - *Journal of Complexity*, **v.** 8 (1992), 153-175.

[10] Nemirovski, A, Yudin, D., "Cesari convergence of the gradient method for approximation saddle points of convex-concave functions" – *Doklady AN SSSR* **v.** 239 (1978), 1056-1059 (in Russian; English translation in *Soviet Math. Doklady*).

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

[12] Nemirovski, A., "Efficient methods for solving variational inequalities" – *Ekonomika i Matem. Metody* **v.** 17 (1981), 344-359 (in Russian; English translation in *Matekon*).

[13] Nesterov, Yu. (2003), "Smooth minimization of nonsmooth functions" – CORE Discussion Paper 2003/12, February 2003.

[14] Nesterov, Yu. (2003), "Dual extrapolation and its applications for solving variational inequalities and related problems" – CORE Discussion Paper 2003/68, September 2003. http://www.core.ucl.ac.be/services/COREdp03.html

[15] Noor, M. A., "New extragradient-type methods for general variational inequalities" – *Journal*

of Math. Anal. Appl. **v.** 277 (2003), 379–394.

[16] Teboulle, M., “Convergence of proximal-like algorithms” – *SIAM Journal on Optimization* **v.** 7 (1997), 1069-1083.