Nonsmooth/Nonconvex Mechanics

Modeling, Analysis and Numerical Methods

NONLINEAR RESCALING IN DISCRETE MINIMAX

Roman A. Polyak
Department of SEOR and Mathematical Sciences Department,
George Mason University,
Fairfax VA 22030,
U.S.A.

Igor Griva
Department of SEOR,
George Mason University,
Fairfax VA 22030,
U.S.A.

Jarek Sobieszczanski-Sobieski
NASA Langley Research Center,
Hampton VA 23681,
U.S.A.

Dedicated to the memory of Professor P.D. Panagiotopoulos.

Abstract We present a general Nonlinear Rescaling (NR) methods for discrete minimax problem. The fundamental difference between the NR approach and the smoothing technique consists of using the Lagrange multipliers as the main driving force to improve the convergence rate and the numerical stability.

In contrast to the smoothing technique the NR methods converge to the primal-dual solution under a fixed scaling parameter.

It allows to avoid the ill-conditioning and at the same time improves the convergence rate. In particular, under the standard second order optimality condition the NR method converges with Q-linear rate when the scaling parameter is fixed, but small enough.

KLUWER ACADEMIC PUBLISHERS
DORDRECHT / BOSTON / LONDON

Moreover, if along with Lagrange multipliers update one decreases the scaling parameter from step to step then the NR method convereges with Q-superlinear rate.

We present two numerical realizations of the general NR method: Newton's NR and the Primal-Dual NR methods.

The obtained numerical results strongly corroborate the theory. In particular, we systematically observed the so-called "hot start" phenomenon, when prom some point on only one Newton's step is enough for the Lagrange multipliers update.

1. INTRODUCTION

A number of important technical problems which arise in structural optimization, synthesis of filters, antenna design etc. (see [Ben-Tal and Nemirovsky, 1998]) leads to well known discrete minimax problem

$$\begin{aligned} x \in X^* &= Argmin\{F(x) \mid x \in \mathbb{R}^n\} \\ &= Argmin\{\max_{1 \le i \le m} f_i(x) | x \in \mathbb{R}^n\}, \end{aligned} \tag{12.1}$$

where $f_i : \mathbb{R}^n \to \mathbb{R}$are convex and smooth enough.

Along with nonsmooth optimization methods (see [Demyanov and Malozemov, 1974],[Kiwiel, 1985],[Lemarechal, 1989],[Shor, 1998] and references in it) the smoothing technique has been used for the discrete minimax since the early 70s [Polyak, 1971], (see also [Bertsekas, 1982],[Charalambous, 1977], [Sobieszczanski-Sobieski, 1992] and references in it). It has become very popular lately due to the growing interest to the smoothing technique for the complementarity problems and constrained optimization (see [Auslender et al., 1997], [Chen and Mangasarian, 1995] and references in it).

The smoothing technique employs the smooth monotone increasing strictly convex function $\psi : \mathbb{R} \to \mathbb{R}$to transform (12.1) into a sequence unconstrained optimization problems

$$x(\mu) = argmin\{S(x, \mu) = \mu \sum_{i=1}^{m} \psi(\mu^{-1} f_i(x)) \mid x \in \mathbb{R}^n\}, \tag{12.2}$$

where $S(x, \mu)$ is a smooth approximation for the function $F(x)$.

The solution x^* for the original problem one obtains as

$$x^* = \lim_{\mu \to 0} x(\mu). \tag{12.3}$$

The smoothing technique is a penalty type approach with a smooth penalty function, so it is in fact a Sequential Unconstrained Optimization Technique (SUMT) (see [Fiacco and McCormick, 1990]) type method with all the advantages and disadvantages, that are typical for SUMT method. Along with some

very important properties of the primal trajectory $x(\mu)$ for some transformations ψ (see [Nesterov and Nemirovsky, 1993]) the smoothing method is rather slow. When the scaling parameter $\mu > 0$ is small enough, the Hessian $\nabla^2_{xx} S(x, \mu)$ becomes ill conditioned and the area where Newton's method for the problem (12.2) is "well defined" [see [Smale, 1986]) shrinks to a point.

In this paper we consider an alternative to the smoothing technique approach, which is based on the Nonlinear Rescaling (NR) methodology (see [Polyak, 1988] - [Polyak, 1999] and references therein).

The NR approach consists of using ψ to transform the original minimax problem into an equivalent one. The transformation is scaled by positive scaling parameter or by a vector of positive scaling parameters. The Classical Lagrangian for the equivalent problem is the main tool in the NR methods.

The NR method consists of finding the minimizer of the Lagrangian for the equivalent problem and updating the Lagrange multipliers, using the minimizer.

The scaling parameter or the vector of scaling parameters can be fixed or one can update it from step to step. In this paper we restrict ourself to one scaling parameter.

Our first contribution is the convergence proof of the general NR method under the fixed scaling parameter. It turns out that for a wide class of transformations the NR multipliers method converges under the standard second order optimality condition with Q-linear rate when the scaling parameter is fixed but small enough. It allows not only to avoid the ill conditioning but also to improve substantially the convergence rate and the numerical stability.

This remains true if instead of exact minimizer one uses its approximation. We have pointed out the conditions for a such approximation, which allows to retain the convergence rate.

If one decreases the scaling parameter from step to step like in the smoothing methods then the NR multipliers method converges with Q-superlinear rate instead of arithmetic rate, as it takes place in smoothing methods.

We introduced two numerical realizations of the NR method. The first is based on Newton's method for primal minimization followed by Lagrange multipliers update. The second uses Newton's method for solving primal-dual system, which combines the optimal condition for the primal minimizer with the system for the Lagrange multipliers update. This is our second contribution.

The numerical realizations have been implemented. The correspondent MATLAB codes were used for solving large enough minimax problems.

We compare the numerical results with results obtained by the smoothing technique as well as results obtained by using the NR approach for the constrained optimization problems, which are equivalent to the discrete minimax. In both cases the NR multipliers method for discrete minimax produced much better results. This is our third contribution.

The obtained numerical results show that NR method allows to solve a wide class of discrete minimax problems with up to 10 digits of accuracy using only few Lagrange multipliers update. We observed that the number of Newton's steps required for the Lagrange multipliers update systematically decreases from one update to another. From some point on (the "hot start") only one Newton's step is enough for the Lagrange multipliers update. Moreover the total number of Newton's steps is almost independent on the size of the problems.

The paper is organized as follows. In the next section we state the problem and describe the basic assumptions. We also discuss the main motivations for the NR approach to the minimax problem in Section 3.

We consider the equivalent problem, the correspondent Lagrangian and describe the NR multipliers method in Section 4.

In Section 5 we prove convergence and estimate the rate of convergence of the NR multipliers method.

We consider the numerical realization of the NR method in Section 6.

In Section 7 we describe the numerical results.

We conclude the paper with some remarks concerning the future research.

2. PROBLEM FORMULATION AND BASIC ASSUMPTIONS

The discrete minimax problem consists of finding

$$x^* \in X^* = Argmin\{F(x) \mid x \in \mathbb{R}^n\} \neq \emptyset, \tag{12.4}$$

where $F(x) = \max\limits_{1 \leq i \leq m} f_i(x)$ and $f_i : \mathbb{R}^n \to \mathbb{R}^1, i = 1, \ldots, m$ are smooth and convex.

We assume that X^* is bounded. This implies that for any $x \in \mathbb{R}^n$ and for any direction $z \in \mathbb{R}^n$ $\lim_{t \to \infty} F(x + tz) = \infty$.

Without loss of generality we can assume that $F(x^*) = 0$. So for any $c > 0$ the set $\Omega = \{x : F(x) \leq c\}$ is bounded and $f_i(x^*) < c, i = 1, ..., m$.

Therefore there exists $\lambda^* \in \mathbb{R}^m_+$ such that the following Karush-Kuhn-Tucker (KKT's) conditions for the discrete minimax hold true.

$$\nabla_x L(x^*, \lambda^*) = \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) = 0 \tag{12.5}$$

$$\lambda_i^* f_i(x^*) = 0, \; i = 1, \ldots, m \tag{12.6}$$

$$\lambda^* \in S_m = \{\lambda \in \mathbb{R}^m_+ : \sum_{i=1}^{m} \lambda_i = 1\}, \tag{12.7}$$

where $L(x,\lambda) = \sum_{i=1}^{m} \lambda_i f_i(x)$ is the Lagrangian for the original problem 12.4.

Let $I^* = \{i : f_i(x^*) = F(x^*)\} = \{1,\dots,r\}$ is the active set.

Let also $f(x) = (f_i(x), i = 1,\dots,m)$ and $f_{(r)}(x) = (f_i(x), i = 1,\dots,r)$. Respectively $\nabla f(x) = J(f(x))$ and $\nabla f_{(r)}(x) = J(f_{(r)}(x))$ are their Jacobians.

We will say that the pair (x^*,λ^*) satisfies the second order optimality conditions, if

$$(\nabla^2_{xx} L(x^*,\lambda^*)y,y) \geq \rho(y,y),\ \rho > 0,\ \forall y : \nabla f_{(r)}(x^*)y = 0, \tag{12.8}$$

$$f_i(x^*) < 0 = F(x^*),\ i = r+1,\dots,m, \tag{12.9}$$

$$rank \nabla f_{(r)}(x^*) = r, \tag{12.10}$$

$$\lambda^* \in S_m \text{ and } \lambda^*_i > 0,\ i = 1,\dots,r \tag{12.11}$$

holds true.

We complete this section with the Debreu type theorem which will be used later.

Let $A : \mathbb{R}^n \to \mathbb{R}^n$ be a symmetric matrix, B be an $r \times n$ matrix and $\Lambda = diag(\lambda_i)_{i=1}^r : \mathbb{R}^n \to \mathbb{R}^n$ be a diagonal matrix with positive elements. If $(Ay,y) \geq \rho_0(y,y)$, $\forall y : By = 0$ then there exists $\mu_0 > 0$, such that for any $0 < \rho < \rho_0$ we have

$$((A + \mu B^T \Lambda B)x,x) \geq \rho(x,x),\ \forall x \in \mathbb{R}^n \tag{12.12}$$

as far as $0 < \mu \leq \mu_0$.

3. SMOOTHING TECHNIQUE IN DISCRETE MINIMAX

The smoothing technique consists of replacing $F(x) = \max\limits_{1\leq i\leq m} f_i(x)$ by a smooth approximation and using this approximation in the framework of SUMT.

We introduce a class Ψ of smoothing transformations, which satisfies the following properties:

P1 $\psi(0) = 0, \quad \psi'(0) = 1,$

P2 $\psi'(t) > 0,$

P3 $\psi''(t) > 0,$

P4 $\lim\limits_{t\to-\infty} \psi'(t) = 0,$

Before we describe the general smoothing method let us introduce a few transformations $\psi \in \Psi$.

1. Exponential $\psi(t) = e^t - 1$,
2. Logarithmic $\psi_2(t) = -\ln(1-t)$,
3. Hyperbolic $\psi_3(t) = \frac{t}{1-t}$,
4. Log-Sigmoid $\psi_4(t) = 2\ln 0.5(1+e^t)$.

One can verify the properties P1-P4 directly for transformations ψ_1 - ψ_4.

Moreover it is easy to see that for the transformations ψ_1 - ψ_4 property P4 can be strengthen, i.e. for any $a < 0$ there exists $b > 0$ the following inequality

P5 $\psi'(\mu^{-1}a) \leq \mu b$

is true as soon as $0 < \mu \leq \mu_0$ and μ_0 is small enough.

We define the smoothing function $S : \mathbb{R}^n \times \mathbb{R}_{++} \to \mathbb{R}$ by formula

$$S(x,\mu) = \mu \sum_{i=1}^{m} \psi(\mu^{-1} f_i(x)),$$

where $\mu > 0$ is a scaling parameter.

Due to the convexity of ψ and all $f_i(x)$ the smoothing function $S(x,\mu)$ is convex in x for any $\mu > 0$. Also $S(x,\mu)$ is as smooth as ψ and f_i.

One can find an approximation for x^* by solving the unconstrained optimization problem

$$x(\mu) = argmin\{S(x,\mu) \mid x \in \mathbb{R}^n\}. \tag{12.13}$$

It turns out that $\lim_{\mu \to 0} x(\mu) = x^* \in X^*$. The existence of $x(\mu)$ follows from the boundness of X^* and the properties P2 - P4.

As we mentioned already the boundness of $X^* = \{x : f_i(x) \leq 0, i = 1, \ldots, m\}$ leads to the boundness of $\Omega = \{x : F(x) \leq c\}$ for any $c > 0$ (see [Fiacco and McCormick, 1990]). It means that the recession cone of the set Ω (see [Auslender et al., 1997])

$$\Omega_\infty = \{y : \exists t_k \to \infty,\ x_k \in \Omega \text{ with } y = \frac{x_k}{t_k}\}$$

is empty.

Therefore for any given $x \in \Omega$ and $z \neq 0$ there exists i_0 and $\bar{t} > 0$: $(\nabla f_{i_0}(x + \bar{t}z), z) > 0$. Using the convexity of f_{i_0} we obtain

$$f_{i_0}(x + tz) - f_{i_0}(x + \bar{t}z) \geq (t - \bar{t})(\nabla f_{i_0}(x + \bar{t}z), z).$$

Therefore $\lim_{t \to \infty} f_{i_0}(x + tz) = \infty$.

Using the convexity $\psi(t)$ we obtain

$$\psi(\mu^{-1}f_{i_0}(x+tz)) - \psi(\mu^{-1}f_{i_0}(x+\bar{t}z)) \geq$$

$$\mu^{-1}\psi'\left(\mu^{-1}f_{i_0}(x+\bar{t}z)\right)\left(f_{i_0}(x+tz) - f_{i_0}(x+\bar{t}z)\right)$$

So keeping in mind P2 we have

$$\lim_{t\to\infty}\psi(\mu^{-1}f_{i_0}(x+tz)) = \infty.$$

Invoking P4 we conclude that

$$\lim_{t\to\infty} S(x+tz,\mu) = \lim_{t\to\infty}\mu\sum_{i=1}^{m}\psi(\mu^{-1}f_i(x+tz)) = \infty$$

for any $\mu > 0$.

Therefore $x(\mu)$ exists, i.e.

$$\nabla_x S(x(\mu),\mu) = \sum_{i=1}^{m}\psi'\left(\mu^{-1}f_i(x(\mu))\right)\nabla f_i(x(\mu)) = 0. \tag{12.14}$$

Moreover the primal trajectory $\{x(\mu)\}_{\mu=\mu_0}^{0}$ is bounded. Taking into account $F(x(\mu)) \geq 0$ and P1, we obtain $\pi(x(\mu)) = \sum_{i=1}^{m}\psi'\left(\mu^{-1}f_i(x(\mu))\right) \geq 1$.

Let us consider the vector of the Lagrange multipliers

$$\lambda(\mu) = (\lambda_i(\mu) = \psi'\left(\mu^{-1}f_i(x(\mu))\right)\pi^{-1}(x(\mu)),\ i = 1,\ldots,m \tag{12.15}$$

The dual trajectory $\{\lambda(\mu)\}_{\mu=\mu_0>0}^{0}$ is bounded because $\lambda(\mu) \in S_m = \{\lambda \in \mathbf{R}_+^m : \sum\lambda_i = 1\}$. Without loss of generality we can assume that

$$\bar{x} = \lim_{\mu\to 0} x(\mu) \quad \text{and} \quad \bar{\lambda} = \lim_{\mu\to 0}\lambda(\mu).$$

Then for $i \in I_-(\bar{x}) = \{i : f_i(\bar{x}) < 0\}$ due to P4 we obtain $\bar{\lambda}_i = 0$. Therefore by passing both side of the system (12.14) to the limit we obtain

$$\nabla_x S(\bar{x},0) = \sum_{i\in I_0(\bar{x})}\bar{\lambda}_i\nabla f_i(\bar{x}) = 0,$$

where $I_0(\bar{x}) = \{i : f_i(\bar{x}) = 0\} = \{1,\ldots,r\}$.

In fact, assuming that $I_+(\bar{x}) = \{i : f_i(\bar{x}) > 0\} \neq \emptyset$,i.e. there exists at least one index $i_0 : f_{i_0}(\bar{x}) > 0$ we obtain $\lim_{\mu\to 0}\mu^{-1}f_{i_0}(x(\mu)) = \infty$. Due to P2 and P3 we have $\lim_{\mu\to 0}\psi\left(\mu^{-1}f_{i_0}(x(\mu))\right) = \infty$. On the other hand due to P4 for any $i \in I_-(\bar{x})$ we have

$$\lim_{\mu\to 0}\psi'\left(\mu^{-1}f_i(x(\mu))\right) = 0,$$

and

$$\lim_{\mu \to 0} \psi\left(\mu^{-1} f_i(x(\mu))\right) = \overline{f}_i > -\infty,\ i \in I_-(\overline{x}).$$

Therefore for μ small enough

$$\begin{aligned} S(x(\mu), \mu) &= \mu \sum_{i \in I_-(x(\mu))} \psi(\mu^{-1} f_i(x(\mu)) \\ &+ \mu \sum_{i \in I_+(x(\mu))} \psi(\mu^{-1} f_i(x(\mu)) > 0. \end{aligned} \tag{12.16}$$

On the other hand due to P1 - P2 we have

$$S(x(\mu), \mu) \leq S(x^*, \mu) = \mu \sum_{i=r+1}^{m} \psi(\mu^{-1} f_i(x^*)) < 0$$

The contradiction allows to conclude that $I_+(\overline{x}) = \emptyset$ and

$$F(\overline{x}) = \max_{1 \leq i \leq m} f_i(\overline{x}) = \max_{1 \leq i \leq r} f_i(\overline{x}) = 0,$$

i.e. the pair $(\overline{x}, \overline{\lambda})$ satisfies the KKT's condition (12.5)-(12.7), therefore $(\overline{x}, \overline{\lambda}) = (x^*, \lambda^*)$.

Moreover, if the second order optimality conditions (12.8)-(12.11) are satisfied, then using arguments similar to those in [Polyak, 1988] (see Lemma 2) one can prove the following lemma for any $\psi \in \Psi$.

Lemma 12.1 *If $f_i(x) \in C^2$ and the standard second order optimality conditions (12.8)-(12.11) are satisfied then for any transformation $\psi \in \Psi$ there exists a small enough $\mu_0 > 0$ such that*

1. *the estimate*

$$\|x(\mu) - x^*\| = \mathcal{O}(\mu),\ \|\lambda(\mu) - \lambda^*\| = \mathcal{O}(\mu) \tag{12.17}$$

   holds for any $0 < \mu \leq \mu_0$.

2. *the smoothing function $S(x, \mu)$ is strongly convex in the neighborhood of $x(\mu)$.*

Now we will consider the smoothing function $S(x, \mu)$, its gradient $\nabla_x S(x, \mu)$ and Hessian $\nabla^2_{xx} S(x, \mu)$ in the neighborhood of $x(\mu)$ for $\mu > 0$ small enough.

First of all using the Lipschitz condition for $f_i(x)$, $i = 1, \ldots, r$ in the neighborhood of x^* and 12.17 we obtain

$$\mu^{-1} f_i(x(\mu)) = \mu^{-1} \left(f_i(x(\mu)) - f_i(x^*)\right) = \mathcal{O}(1),\ i = 1, \ldots, m \tag{12.18}$$

for any $0 < \mu \leq \mu_0$.

Therefore in view of the smoothness f_i and $\psi \in \Psi$ and taking into account P2-P3, there is $0 < a < b$ and $0 < c < d$ such that

$$a \leq \psi' \left(\mu^{-1} f_i(x(\mu))\right) \leq b, \; i = 1, \ldots, r, \; \forall \mu \in (0, \mu_0] \tag{12.19}$$

and

$$c \leq \psi'' \left(\mu^{-1} f_i(x(\mu))\right) \leq d, \; i = 1, \ldots, r, \; \forall \mu \in (0, \mu_0] \tag{12.20}$$

Also using P4 we obtain $\lim_{\mu \to 0} \psi'(\mu^{-1} f_i(x(\mu))) = 0, i \in I_-(x^*)$. Therefore for small enough $\mu > 0$ we have

$$\nabla_x S(x(\mu), \mu) \approx \sum_{i=1}^{r} \psi'(\mu^{-1} f_i(x(\mu))) \nabla f_i(x(\mu)).$$

Hence for the Hessian $\nabla^2_{xx} S(x, \mu)$ we obtain

$$\nabla^2_{xx} S(x(\mu), \mu) \approx \sum_{i=1}^{r} \psi' \left(\mu^{-1} f_i(x(\mu))\right) \nabla^2 f_i(x(\mu)) +$$

$$\mu^{-1} \sum_{i=1}^{r} \psi'' \left(\mu^{-1} f_i(x(\mu))\right) \nabla f_i(x(\mu)) \nabla f_i^T(x(\mu)) =$$

$$\pi(x(\mu)) \sum_{i=1}^{r} \lambda_i(\mu) \nabla^2 f_i(x(\mu)) +$$

$$\mu^{-1} \nabla f_{(r)}(x(\mu)) \Psi'' \left(\mu^{-1} f(x(\mu))\right) \nabla f_{(r)}^T(x(\mu)) =$$

$$\pi(\cdot) \nabla^2_{xx} L(\cdot) + \mu^{-1} \nabla f_{(r)}(\cdot) \Psi''(\cdot) \nabla f_{(r)}^T(\cdot)$$

or

$$\nabla^2_{xx} S(\cdot, \mu) \approx \tag{12.21}$$

$$\pi(\cdot) \left[\nabla^2_{xx} L(\cdot) + (\mu \pi(\cdot))^{-1} \left(\nabla f_{(r)}(\cdot) \Psi''(\cdot) \nabla f_{(r)}^T(\cdot)\right)\right], \tag{12.22}$$

where $\pi(\cdot) = \pi(x(\mu)) = \sum_{i=1}^{m} \psi' \left(f_i(\cdot)\right)$ and $\Psi''(\cdot) = diag \left(\psi''(\mu^{-1} f_i(\cdot))\right)_{i=1}^{r}$.

It follows from (12.19) that there is $0 < a' < b'$ that $a' \leq \pi(x(\mu)) \leq b'$. Therefore using the second order optimality condition (12.8)-(12.11) and Debreu inequality (12.12) with $A = \nabla_{xx} L$, $B = \nabla f_{(r)}^T$, $\Lambda = \Psi''$ we can find $\rho > 0$ that

$$mineigenval\, \nabla_{xx}^2 S(x(\mu), \mu) \geq \rho > 0, \forall 0 < \mu \leq \mu_0.$$

On the other hand due to the second term in (12.22) for small enough $\mu > 0$ we have

$$maxeigenval\, \nabla_{xx}^2 S(x(\mu), \mu) = \mathcal{O}(\mu^{-1}).$$

Therefore

$$cond \nabla_{xx}^2 S(x(\mu), \mu) = \frac{mineigenval \nabla_{xx}^2 S(x(\mu), \mu)}{maxeigenval \nabla_{xx}^2 S(x(\mu), \mu)} = \mathcal{O}(\mu)$$

and

$$\lim_{\mu \to 0} cond \nabla_{xx}^2 S(x(\mu), \mu) = 0.$$

Hence the area around $x(\mu)$ where Newton's method for solving $\nabla_x S(x, \mu) = 0$ is well defined (see [Smale, 1986]) shrinks to a point when $\mu \to 0$.

In the next section we will consider the NR multipliers method, which allows to eliminate the mentioned drawbacks. The NR method converges under the fixed $\mu > 0$, just due to the Lagrange multipliers update. Therefore the area where Newton's method is "well defined" does not shrink to a point Moreover, under the second order optimality condition instead of estimation (12.17) the rate of convergence is Q-linear and the ratio can be made as small as one needs by adjusting the scaling parameter $\mu > 0$.

4. NONLINEAR RESCALING METHOD

First, we transform the original problem into an equivalent one using one of the transformation $\psi \in \Psi$. The transformation is scaled by a scaling parameter, i.e. instead of original problem (12.1) we consider an equivalent problem

$$x^* \in X^* = Argmin\{F_\mu(x) = \mu \max_{1 \leq i \leq m} \psi(\mu^{-1} f_i(x)) \,|\, x \in \mathbb{R}^n\}. \quad (12.23)$$

Our main tool is the Classical Lagrangian for the equivalent problem $\mathcal{L} : \mathbb{R}^n \times S_m \times \mathbb{R}_{++} \to \mathbb{R}$ which is defined by formula

$$\mathcal{L}(x, \lambda, \mu) = \mu \sum_{i=1}^m \lambda_i \psi(\mu^{-1} f_i(x)), \text{ where } \lambda \in S_m = \{\lambda \in \mathbb{R}_+^m : \sum_{i=1}^m \lambda_i = 1\}.$$

Before we will describe the NR multipliers method we would like to mention a few important properties of the Lagrangian $\mathcal{L}(x, \lambda, \mu)$ at the KKT's pair (x^*, λ^*). For any $\mu > 0$ we have

1. $\mathcal{L}(x^*, \lambda^*, \mu) = \mu \sum_{i=1}^m \lambda_i^* \psi(\mu^{-1} f_i(x^*)) = F(x^*) = 0.$
2. $\nabla_x \mathcal{L}(x^*, \lambda^*, \mu) = \sum_{i=1}^m \lambda_i^* \psi'(\mu^{-1} f_i(x^*)) \nabla f_i(x^*) =$
   $= \sum_{i=1}^m \lambda_i^* \nabla f_i(x^*) = \nabla_x L(x^*, \lambda^*) = 0$
3. $\nabla_{xx}^2 \mathcal{L}(x^*, \lambda^*, \mu) = \nabla_{xx}^2 L(x^*, \lambda^*) + \mu^{-1} \psi''(0) \nabla f_{(r)}(x^*) \Lambda_{(r)}^* \nabla f_{(r)}^T(x^*),$

where $\Lambda_{(r)}^* = diag(\lambda_i^*)_{i=1}^r$.

The properties $1^0 - 3^0$ follow directly from P1 and the complementarity condition.

The Lagrangian $\mathcal{L}(x, \lambda, \mu)$ is convex in $x \in \mathbb{R}^n$ for any $\lambda \in S_m$, it is strictly or strongly convex in x if at least one of $f_i(x)$ is strictly or strongly convex and the corresponding $\lambda_i > 0$.

The Lagrangian $\mathcal{L}(x, \lambda, \mu)$ is as smooth as $f_i(x)$ and ψ. For $\lambda = \lambda^*$ and any $\mu > 0$ it is an exact smooth approximation for the nonsmooth function $F(x)$ at $x = x^*$, i.e. for any $\mu > 0$

$$x^* = argmin\{\mathcal{L}(x, \lambda^*, \mu) | x \in \mathbb{R}^n\}, \tag{12.24}$$

Moreover, if none of f_i is convex but the standard second order optimality conditions (12.8)-(12.11) are satisfied then due to 3^0 and the Debreu theorem the Lagrangian $\mathcal{L}(x, \lambda, \mu)$ is strongly convex in x if $\mu > 0$ is small enough and (12.24) holds.

The unconstrained minimization of $\mathcal{L}(x, \lambda, \mu)$ in x followed by the Lagrange multipliers update leads to NR multipliers method.

Let $\lambda^0 \in S_m$ is a positive vector and $\mu > 0$. Let us assume that the pair $(x^s, \lambda^s) \in \mathbb{R}^n \times S_m$ have been found already. We find the next approximation (x^{s+1}, λ^{s+1}) by the following formulas:

$$x^{s+1} = argmin\{\mathcal{L}(x, \lambda^s, \mu) | x \in \mathbb{R}^n\}, \tag{12.25}$$

$$\hat{\lambda}^{s+1} = \psi'(\mu^{-1} f_i(x^{s+1})) \lambda_i^s, \quad \pi_{s+1} = \sum_{i=1}^m \hat{\lambda}_i^{s+1} \tag{12.26}$$

$$\lambda_i^{s+1} = \hat{\lambda}_i^{s+1} \pi_{s+1}^{-1}, \quad i = 1, \ldots, m. \tag{12.27}$$

Due to the properties P2-P4 the method 12.25-12.27 is well defined and the vectors of the Lagrange multipliers λ^s, $s \geq 1$ remains positive if $\lambda^0 \in \mathbb{R}_{++}^m$.

Also for x^{s+1} we have

$$\nabla \mathcal{L}(x^{s+1}, \lambda^s, \mu) = \sum_{i=1}^{m} \lambda_i^s \psi'(\mu^{-1} f_i(x^{s+1})) \nabla f_i(x^{s+1}) =$$

$$\pi_{s+1} \sum_{i=1}^{m} \lambda^{s+1} \nabla f_i(x^{s+1}) = \pi_{s+1} L(x^{s+1}, \lambda^{s+1}) = 0 \tag{12.28}$$

or

$$x^{s+1} = \arg \min_{x \in \mathbb{R}^n} L(x, \lambda^{s+1}). \tag{12.29}$$

We can avoid the normalization procedure 12.27 by introducing shifts for the functions $f_i(x)$. In other words on the step $s \geq 1$ for each $1 \leq i \leq m$ we can introduce a shift t_i^s that

$$\lambda_i^{s+1} = \psi'(\mu^{-1}(f_i(x^{s+1}) + t_i^{s+1}))\lambda_i^s$$

or

$$\psi'(\mu^{-1}(f_i(x^{s+1}) + t_i^{s+1})) = \frac{\lambda_i^{s+1}}{\lambda_i^s}.$$

Due to P3 there exists a unique inverse function ψ'^{-1}, i.e.

$$f_i(x^{s+1}) = \mu \psi'^{-1}(\frac{\lambda_i^{s+1}}{\lambda_i^s}) - t_i^{s+1}. \tag{12.30}$$

Therefore shifts t_i^{s+1} can be uniquely defined by

$$t_i^{s+1} = \mu \psi^*(\frac{\lambda_i^{s+1}}{\lambda_i^s}) - f_i(x^{s+1}),$$

where ψ^* is Fenchel transform of $\psi \in \Psi$. From (12.29) we have

$$d(\lambda^{s+1}) = \min_{x \in \mathbb{R}^n} L(x, \lambda^{s+1}) = L(x^{s+1}, \lambda^{s+1}),$$

where $d(\lambda) = \min_{x \in \mathbb{R}^n} L(x\, \lambda)$ is the dual function. Also

$$f(x^{s+1}) \in \partial d(\lambda^{s+1}),$$

where $\partial d(\lambda)$ is the differential of $d(\lambda)$. Therefore from (12.30) we obtain

$$0 \in \partial d(\lambda^{s+1}) - \sum_{i=1}^{m} \left(\mu \psi^*(\frac{\lambda_i^{s+1}}{\lambda_i^s}) - t_i^{s+1} \right) e_i, \tag{12.31}$$

where $e_i = (0, \dots, 1, \dots 0)$.

The system (12.31) is the optimality condition for the vector

$$\lambda^{s+1} = argmax\{d(\lambda) - \sum_{i=1}^{s} \lambda_i^s \left[\mu\psi^*(\frac{\lambda_i}{\lambda_i^s}) - t_i^{s+1}\frac{\lambda_i}{\lambda_i^s}\right] \mid \lambda \in \mathbb{R}^m_{++}\}. \tag{12.32}$$

Therefore the method (12.25)-(12.27) is equivalent to the Prox-type method (12.32).

5. CONVERGENCE OF THE NR METHOD

In this section we will prove the convergence and estimate the rate of convergence of the NR method (12.25)-(12.27) under the standard second order optimality conditions.

Theorem 12.1 *If the second order optimality conditions (12.8)-(12.11) are satisfied and $f_i \in C^2$, then for any positive vector of Lagrange multipliers $\lambda \in S_m$ and any $0 < \mu < \mu_0$, where $\mu_0 > 0$ is small enough the following statements are true*

1. *there exist $\hat{x}$ and $\hat{t}$ such that*

$$\hat{x} = argmin\{\sum_{i=1}^{m} \lambda_i \psi(\mu^{-1}(f_i(x) + \hat{t}_i)) \mid x \in \mathbb{R}^n\} :$$

$$\sum_{i=1}^{m} \lambda_i \psi'(\mu^{-1}(f_i(\hat{x}) + \hat{t}_i))\nabla f_i(\hat{x}) = \sum_{i=1}^{m} \hat{\lambda}_i \nabla f_i(\hat{x}) = 0,$$

where $\hat{\lambda} = \left(\hat{\lambda}_i \psi'(\mu^{-1}(f_i(\hat{x}) + \hat{t}_i)), i = 1, \dots, m\right) \in S_m$.

2. *for the pair $\hat{x}$ and $\hat{\lambda}$ the following estimates hold*

$$\|\hat{x} - x^*\| \le c\mu\|\lambda - \lambda^*\|, \ \|\hat{\lambda} - \lambda^*\| \le c\mu\|\lambda - \lambda^*\|$$

where $c > 0$ is independent on $\mu > 0$.

3. *the Lagrangian $\mathcal{L}(x, \lambda, \mu)$ is strongly convex in the neighborhood of $\hat{x}$.*

The proof is along the lines of the proof of the Theorem 1 in [Polyak, 1988]. We will only point out the main steps.

We consider the Lagrange multipliers vector

$$\hat{\lambda}(x, \lambda, t, \mu) = \left(\lambda_i \psi'(\mu^{-1}(f_i(x) + t_i)), i = 1, \dots, m\right) = \hat{\lambda}(\cdot). \tag{12.33}$$

Then $\hat{\lambda}(x^*, \lambda^*, 0, \mu) = \lambda^*$ for any $\mu > 0$ or λ^* is a fixed point of the map: $\lambda \to \hat{\lambda}(x^*, \lambda, 0, \mu)$.

Let $h(x, \lambda, t, \mu) = h(\cdot) = \sum_{i=r+1}^{m} \hat{\lambda}_i(\cdot) \nabla f_i(\cdot)$, then $h(x, \lambda, t, \mu)$ is smooth. Also $h(x^*, \lambda^*, 0, \mu) = 0$ and $h(x, \lambda, t, \mu)$ is continuous in the neighborhood of $(x^*, \lambda^*, 0, \mu)$ together with its derivatives in x, i.e. $h'(x^*, \lambda^*, 0, \mu) = 0 \in \mathbb{R}^n$.

We consider the following map $\Phi(x, \hat{\lambda}, t, \lambda, \mu) : \mathbb{R}^{n+3m+1} \to \mathbb{R}^{n+r+1}$ in the neighborhood of $(x^*, \lambda^*, 0^n, 0)$ defined by formula

$$\Phi(x, \hat{\lambda}, t, \lambda, \mu) = \begin{pmatrix} \sum_{i=1}^{r} \hat{\lambda}_i \nabla f_i(x) + h(x, \lambda, t, \mu) \\ f_i(x) + t - \mu \psi^{*\prime}\left(\frac{\hat{\lambda}_i}{\lambda_i}\right), \ i = 1, \dots, r \\ \sum_{i=1}^{r} \hat{\lambda}_i + \sum_{i=r+1}^{m} \hat{\lambda}_i(x, \lambda, t, \mu) = 1 \end{pmatrix} \tag{12.34}$$

Then in view of KKT's condition, $F(x^*) = f_i(x^*) = 0$, $i = 1, \dots, r$ and $\psi^{*\prime}(1) = 0$ we obtain $\Phi(x^*, \lambda^*, 0, \lambda^*, \mu) = 0 \in \mathbb{R}^{n+r+1}$. Also

$$\nabla_{x \hat{\lambda}_{(r)} t_{(r)}} \Phi(x^*, \hat{\lambda}^*, 0, \lambda^*, \mu) = \begin{pmatrix} \nabla^2_{xx} L & \nabla f^T_{(r)} & 0 \\ \nabla f_{(r)} & 0 & e^T_{(r)} \\ 0 & e_{(r)} & 0 \end{pmatrix}, \tag{12.35}$$

where $e_{(r)} = (1, \dots, 1) \in \mathbb{R}^r$, $\hat{\lambda}_{(r)} = (\hat{\lambda}_i, i = 1, \dots, r)$, $t_{(r)} = (t_i, i = 1, \dots, r)$.

The matrix $\nabla_{x \lambda_{(r)} t_{(r)}} \Phi$ is nonsingular. In fact, consider a vector $w = (y, v, \tau)$, $y \in \mathbb{R}^n$, $v \in \mathbb{R}^r$ and $\tau \in \mathbb{R}$. Then $\nabla_{x \lambda_{(r)} t_{(r)}} \Phi w = 0$ implies

$$\nabla_{xx} L y + \nabla f^T_{(r)} v = 0 \tag{12.36}$$

$$\nabla f_{(r)} y + \tau e_{(r)} = 0 \tag{12.37}$$

$$(e_{(r)}, v) = 0 \tag{12.38}$$

By multiplying the second system by $\lambda^*_{(r)}$ we obtain $(\lambda^*_{(r)} \nabla f_{(r)} y) + \tau (e_{(r)}, \lambda^*_{(r)}) = 0$. So in view of $\sum_{i=1}^{r} \lambda_i^* = (e_{(r)}, \lambda^*_{(r)}) = 1$ we have $(\sum_{i=1}^{r} \lambda_i^* \nabla f_i(x^*), y) + \tau = 0$.

Taking into account KKT's condition $\sum_{i=1}^{r} \lambda_i^* \nabla f_i(x^*) = 0$ we obtain $\tau = 0$. Therefore from (12.37) we have

$$\nabla f_{(r)} y = 0. \tag{12.39}$$

By multiplying (12.36) by y we obtain

$$(\nabla^2_{xx} Ly, y) + (\nabla f_{(r)} y, v) = 0.$$

Hence we have $(\nabla^2_{xx} Ly, y) = 0$ for $\forall y : \nabla f_{(r)} y = 0$. Invoking (12.8) we obtain $y = 0$. Then from (12.36) we have $\nabla f_{(r)} v = 0$, which together with (12.10) leads to $v = 0$. In other words $\nabla_{x\hat{\lambda}_{(r)}t_{(r)}} \Phi w = 0 \Rightarrow w = 0$ or $\nabla_{x\hat{\lambda}_{(r)}t_{(r)}} \Phi$ is nonsingular in the neighborhood of $z^* = (x^*, \lambda^*_{(r)}, 0, \lambda^*_{(r)}, 0)$. Since $f_i(x) \in C^2$ the implicit function theorem (see[Bertsekas, 1982]) suggests that in this neighborhood of z^* there exists smooth vector functions $\hat{x}(\cdot) = \hat{x}(\lambda, \mu), \hat{\lambda}(\cdot) = \hat{\lambda}(\lambda, \mu), \hat{t}(\cdot) = \hat{t}(\lambda, \mu)$ such that

$$\Phi\left(x(\cdot), \hat{\lambda}(\cdot), t(\cdot), \lambda, \mu\right) = \Phi(\cdot) = 0$$

or

$$\sum_{i=1}^{r} \hat{\lambda}_i(\cdot) \nabla f_i(\hat{x}(\cdot)) = 0$$

$$\hat{\lambda}_i(\cdot) = \psi'\left(\mu^{-1}(f_i(\cdot) + \hat{t}(\cdot))\right) \lambda_i, \; i = 1, \ldots, r$$

and $\sum_{i=1}^{r} \lambda_i(\cdot) = 1$.

By differentiating the identity $\Phi(\cdot) = 0$ with respect $\lambda_{(r)}$ we obtain

$$\nabla_{x\lambda_{(r)}t_{(r)}} \Phi(\cdot) w(\cdot) + \nabla_{\lambda_{(r)}} \Phi(\cdot) = 0,$$

where $w(\cdot) = \left(\nabla_{\lambda_{(r)}} x(\lambda, \mu), \nabla_{\lambda_{(r)}} \hat{\lambda}_{(r)}(\lambda, \mu), \nabla_{\lambda_{(r)}} t(\lambda, \mu)\right)$. Therefore

$$w(\cdot) = \left(\nabla_{x\lambda_{(r)}t_{(r)}} \Phi(\cdot)\right)^{-1} \nabla_{\lambda_{(r)}} \Phi(\cdot).$$

Since $\nabla_{x\hat{\lambda}_{(r)}t_{(r)}} \Phi(\cdot)$ is nonsingular in the neighborhood of λ^* there exists such $c_1 > 0$ that

$$\|\nabla_{x\hat{\lambda}_{(r)}t_{(r)}} \Phi^{-1}(\cdot)\| \leq c_1.$$

Also there is such $c_2 > 0$ that

$$\|\nabla_{\lambda_{(r)}} \Phi(\cdot)\| \leq c_2 \mu.$$

Therefore for $c = c_1 c_2$ we have

$$\|w(\cdot)\| \leq c\mu. \tag{12.40}$$

Since $x(\lambda^*, \mu) = x^*$, $\lambda_{(r)}(\lambda^*, \mu) = \lambda^*_{(r)}$ from (12.40) using consideration similar to those in [Polyak, 1988] we obtain

$$\|\hat{x}(\lambda, \mu) - x^*\| c\mu \|\lambda_{(r)} - \lambda^*_{(r)}\|, \tag{12.41}$$

$$\|\hat{\lambda}_{(r)}(\lambda, \mu) - \lambda^*\| c\mu \|\lambda_{(r)} - \lambda^*_{(r)}\|, \tag{12.42}$$

Also from (12.26) and P5 follows

$$\hat{\lambda}_i(\cdot) = \lambda_i \psi' \left(\mu^{-1} f_i(x(\cdot))\right) \leq b\lambda_i \mu, \; i = r+1, \ldots, m.$$

Therefore we can rewrite (12.42)) as follows

$$\|\hat{x} - x^*\| \leq c\mu \|\lambda - \lambda^*\|, \; \|\hat{\lambda} - \lambda^*\| \leq c\mu \|\lambda - \lambda^*\|, \tag{12.43}$$

The strong convexity of $\mathcal{L}(x, \lambda, \mu)$ in the neighborhood of (x^*, λ^*) follows directly from the formula for $\nabla^2_{xx}\mathcal{L}(x, \lambda, \mu)$, estimation (12.43) and Debreu inequality (12.12).

The method (12.25)-(12.27) is a theoretical one because it requires finding $\hat{x} = \min\limits_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda, \mu)$. It turns out we can find approximation for $\hat{x}$, which holds the estimation (12.43) as long as the standard second order optimality conditions (12.8)-(12.11) are satisfied and $\mu > 0$ is small enough.

Instead of finding $\hat{x} = \min\limits_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda, \mu)$. we consider

$$\overline{x} : \|\nabla_x \mathcal{L}(x, \lambda, \mu)\| \leq \tau\mu \|\overline{\lambda}(\overline{x}, \lambda, \mu) - \lambda\|, \tag{12.44}$$

where $\lambda(x, \lambda, \mu) = \lambda(\cdot) = \left(\lambda_i(\cdot) = \psi'(\mu^{-1} f_i(x))\lambda_i, i = 1, \ldots, m\right)$,
$\pi(x, \lambda, \mu) = \sum\limits_{i=1}^{m} \psi'(\mu^{-1} f_i(x))\lambda_i$ and
$\overline{\lambda}(\overline{x}, \lambda, \mu) = \left(\lambda_i(\overline{x}, \lambda, \mu)\pi^{-1}(\overline{x}, \lambda, \mu), \; i = 1, \ldots, m\right)$.

Then the following proposition, which is similar to the Proposition 2 in [Polyak, 1988] takes place.

Proposition 12.1 *If the second order optimality condition are satisfied and the Hessians $\nabla^2_{xx} f_i(x)$, $i = 1, \ldots, m$ satisfy the Lipschitz conditions then for any $\mu > 0$ small enough and any positive vector $\lambda \in S_m$ the following estimation holds true*

$$\|\overline{x} - x^*\| \leq c(1+\tau)\mu \|\lambda - \lambda^*\|, \; \|\overline{\lambda} - \lambda^*\| \leq c(1+\tau)\mu \|\lambda - \lambda^*\|. \tag{12.45}$$

The estimation (12.45) can be proven using considerations similar to those in [Polyak and Tretyakov, 1974] and [Polyak, 1999].

The possibility to replace the exact minimum $\hat{x}$ by $\overline{x}$ provides the stopping criteria at each step of method (12.25)-(12.27). It allows to consider a numerical

realization of the multipliers method, which requires finite number of iterations at each step.

In the following section, however, we will consider stopping criteria which is based on the converging to zero sequence of positive numbers and the primal-dual gap. We used this criteria in our calculations which are presented in Section 7.

6. NUMERICAL REALIZATION OF THE NR ALGORITHM

We consider two numerical realizations of the NR method for discrete minimax.

In the first realization we use Newton's method with step length for minimization of the Lagrangian $\mathcal{L}(x, \lambda, \mu)$ in primal space followed by the Lagrange multipliers update.

In the second realization we use Newton's method for solving the primal-dual system of equations, which consists of the KKT's equations and formulas for the Lagrange multipliers update followed by the normalization of the Lagrange multipliers.

As a stopping criteria we use the primal-dual gap

$$\Delta(x, \lambda) = F(x) - d(\lambda).$$

For any $x \in \mathbb{R}^n$ and $\lambda \in S_m$ we have $\Delta(x, \lambda) \geq 0$ and

$$\Delta(x, \lambda) = 0 \quad \text{iff} \quad x = x^*, \lambda = \lambda^*.$$

Newton's NR method for discrete minimax consists of using Newton's method with step size for minimization $\mathcal{L}(x, \lambda^s, \mu)$ in x up to the point when $\|\nabla_x \mathcal{L}(x, \lambda^s, \mu)\|$ is rather small and then update the Lagrange multipliers using the approximation for x^{s+1} in the formulas (12.26)-(12.27). In particular, we can use the formula (12.44) as a stopping criteria at each step. Another way consists of choosing a positive monotone decreasing sequence $\{\delta_s\}_{s=0}^{\infty}$: $\lim_{s \to \infty} \delta_s = 0$ to control the value $\|\nabla_x \mathcal{L}(\cdot)\|$.

In the following algorithm we choose the scaling parameter $\mu > 0$ and parameter $\delta > 0$ small enough and decrease them linearly using parameters $0 < \gamma < 1$ and $0 < \kappa < 1$ as ratios.

NR Algorithm:

input

An accuracy parameter $\epsilon > 0$

Primal $x^0 \in \mathbb{R}^n$, dual $\lambda^0 = (1, \ldots, 1) \in \mathbb{R}^m$

Initial scaling parameter $\mu > 0$, initial accuracy $\delta > 0$

and two parameter $0 < \gamma < 1$ and $0 < \kappa < 1$

begin

$x := x^0$, $\lambda := \lambda^0$, $F := \max\limits_{1\le i\le m} f_i(x^0)$
while $\Delta(x,\lambda) > \epsilon$ do
  begin
    while $\|\nabla \mathcal{L}(x,\lambda,\mu)\| > \delta$ do
      begin
      find $\Delta x : \nabla^2_{xx}\mathcal{L}(x,\lambda,\mu)\Delta x = -\nabla_x \mathcal{L}(x,\lambda,\mu)$
      $t := 1$;
       while $\mathcal{L}(x + t\Delta x, \lambda, \mu) - \mathcal{L}(x,\lambda,\mu) >$
             $0.33t(\nabla_x \mathcal{L}(x,\lambda,\mu), \Delta x)$
        do $t := t/2$;
      $x := x + t\Delta x$;
      end
    $\pi = \sum\limits_{i=1}^{m} \psi'(\mu^{-1} f_i(x))\lambda_i$;
    $\lambda_i = \lambda_i \psi'(\mu^{-1} f_i(x))\pi^{-1}$, $i = 1, \ldots, m$
    $F := \max\limits_{1\le i\le m} f_i(x)$, $D := \sum_{i=1}^{m} \lambda_i f_i(x)$, $\Delta(x,\lambda) := F - D$
    $\delta := \delta\gamma$, $\mu := \mu\kappa$
  end
end
output x, λ, F

Now we describe Primal-Dual NR method for discrete minimax.
We consider the KKT's equations

$$\nabla_x L(\hat{x}, \hat{\lambda}) = \sum_{i=1}^{m} \hat{\lambda}_i \nabla f_i(\hat{x}) = 0 \tag{12.46}$$

together with the formulas for the Lagrange multipliers update

$$\hat{\lambda}_i = \psi'\left(\mu^{-1} f_i(\hat{x})\right) \lambda_i \tag{12.47}$$

Let's linearize the system (12.46)-(12.47) with regard $\hat{x}$ and $\hat{\lambda}$, i.e. let's assume $\hat{x} = x + \Delta x$, $\hat{\lambda} = \lambda + \Delta\lambda$. Then

$$\sum_{i=1}^{m} \hat{\lambda}_i \nabla f_i(\hat{x}) = \sum_{i=1}^{m} (\lambda_i + \Delta\lambda_i) \nabla f_i(x + \Delta x) =$$

$$\sum_{i=1}^{m} (\lambda_i + \Delta\lambda_i)(\nabla f_i(x) + \nabla^2 f_i(x)\Delta x) =$$

$$\sum_{i=1}^{m} \lambda_i \nabla f_i(x) + \sum_{i=1}^{m} \Delta\lambda_i \nabla f_i(x) +$$

$$\sum_{i=1}^{m} \lambda_i \nabla^2 f_i(x)\Delta x + \sum \Delta\lambda_i \nabla f_i(x)\Delta x. \qquad (12.48)$$

By ignoring second order terms we obtain

$$(\sum \lambda_i \nabla^2 f_i(x))\Delta x + \sum_{i=1}^{m} \Delta\lambda_i \nabla f_i(x) = -\sum_{i=1}^{m} \lambda_i \nabla f_i(x) \qquad (12.49)$$

Further, the system (12.47) we can rewrite as follows:

$$\lambda_i + \Delta\lambda_i = \psi'\left(\mu^{-1}(f_i(x + \Delta x))\right)\lambda_i =$$

$$\psi'\left((\mu^{-1} f_i(x) + \mu^{-1}\nabla f_i(x)\Delta x\right)\lambda_i.$$

By ignoring second order terms in the right hand side we obtain

$$\lambda_i + \Delta\lambda_i = \lambda_i \psi'\left(\mu^{-1} f_i(x)\right) + \mu^{-1}\lambda_i \psi''\left(\mu^{-1} f_i(x)\right)\nabla f_i(x)\Delta x$$

or

$$-\mu^{-1}\lambda_i \psi''\left(\mu^{-1} f_i(x)\right)\nabla f_i(x)\Delta x + \Delta\lambda_i =$$

$$\lambda_i\left(\psi'(\mu^{-1} f_i(x)) - 1\right), i = 1, \ldots, m. \qquad (12.50)$$

Combining (12.49) and (12.50) we obtain

$$\nabla^2_{xx} L(x, \lambda)\Delta x + \nabla f^T(x)\Delta\lambda = -\nabla_x L(x, \lambda) \qquad (12.51)$$

$$-\mu^{-1}\Lambda\Psi''\left(\mu^{-1} f(x)\right)\nabla f(x)\Delta x + \Delta\lambda = \left(\Psi'(\mu^{-1} f(x)) - I\right)\lambda. \qquad (12.52)$$

By introducing the primal-dual corrector $\Delta y = (\Delta x, \Delta\lambda)$ and the dual predictor $\lambda' = \left(\lambda_i' = \psi'(\mu^{-1} f_i(x))\lambda_i,\ i = 1, \dots, m\right)$ we can rewrite the system (12.51)-(12.52) as follows

$$M\Delta y = \begin{bmatrix} -\nabla_x L(x,\lambda) \\ \lambda' - \lambda \end{bmatrix} \tag{12.53}$$

where

$$M = \begin{bmatrix} \nabla_x^2 L & \nabla f^T \\ -\mu^{-1}\Lambda\Psi''\nabla f & I \end{bmatrix}$$

and $\Psi' = diag\left(\psi'(\mu^{-1} f_i(x))\right)_{i=1}^m$, $\Lambda = diag(\lambda_i)$,
$\Psi'' = diag\left(\psi''(\mu^{-1} f_i(x))\right)_{i=1}^m$. >From (12.52) we have

$$\Delta\lambda = \mu^{-1}\Lambda\Psi''\nabla f\Delta x + (\lambda' - \lambda) \tag{12.54}$$

After we substitute $\Delta\lambda$ in (12.51) we obtain the following system

$$\nabla_{xx}^2 L\Delta x + \nabla f^T\left(\mu^{-1}\Lambda\Psi''\nabla f\Delta x + (\lambda' - \lambda)\right) = -\nabla_x L, \tag{12.55}$$

or

$$\left(\nabla_{xx}^2 L + \mu^{-1}\nabla f^T\Lambda\Psi''\nabla f\right)\Delta x = -\nabla_x L - \nabla f^T(\lambda' - \lambda) = \tag{12.56}$$

$$-\nabla_x L(x,\lambda'). \tag{12.57}$$

>From the system (12.57) we find the primal corrector Δx. Then from (12.54) find the dual corrector $\Delta\lambda$ and $\hat{\lambda} = \lambda + \Delta\lambda$. The next approximation for the Lagrange multipliers vector is $\lambda := \hat{\lambda}(\sum_{i=1}^m \hat{\lambda}_i)^{-1}$.

We can view the primal-dual method as dual-primal predictor-corrector. First we predict the dual $\lambda' := \lambda\psi'(\mu^{-1} f(x))$, then solve the system (12.57) to find the primal corrector Δx and then we find the dual corrector $\Delta\lambda$ from (12.54) and normalize the dual approximation.

The primal-dual method is fast and numerically stable in the neighborhood of (x^*, λ^*). To make the NR method converge globally we can combine it with Newton's NR or with a smoothing method, using technique similar to those described in [Melman and Polyak, 1996].

Newton's NR method has been implemented and the MATLAB based code was applied for two different sets of discrete minimax problems.

The first set is random generated Quadratic minimax problems, i.e. a problems type (12.1) with

$$f_i(x) = 0.5x^T Q_i x + q_i^T x + q_{i_0}, i = 1, \dots, m$$

where $Q_i = Q_i^T : \mathbb{R}^n \to \mathbb{R}^n$ positive definite matrices, and $q_i \in \mathbb{R}^n$, $q_{i_0} \in \mathbb{R}$.

The second set is Chebichev center problems, i.e. for a given set of points $\{y_i \in \mathbb{R}^n, i = 1, \ldots, m\}$ one wants to find the Chebichev center

$$x^* = argmin\{\max_{1 \le i \le m} \|x - y_i\|^2 \mid x \in \mathbb{R}^n\}$$

In other words we want to find the center x^* of a sphere with minimum radius, which can cover the set $\{y_i\}_{i=1}^m$.

For the first set of problems we used both versions of Newton's NR method with a fixed scaling parameter, which we update from step to step. There is a few observations following from the obtained results.

1. For all problems we observed the so-called "hot start" phenomenon (see [Polyak, 1992] [Melman and Polyak, 1996]), when very few and from some point only one Newton's step is enough for the Lagrange multipliers update.

2. The number of Lagrange multipliers update and the total number of Newton's steps is practically independent on the size of the problem.

3. All problem have been solved with final duality gap 10^{-9}.

We compared the obtained results by Newton's NR method with smoothing technique. For the smoothing method with the same nonlinear rescaling function ψ and $\mu_0 = 0.1$, $\gamma = 0.1$, 0.2 it requires much more Newton's steps to achieve the same accuracy because after each scaling parameter update the old approximation does not belong to the area where Newton's method is "well" defined (see[Smale, 1986]). Therefore after each scaling parameter update it requires some effort to get back to Newton's area (see Tables 12.4, 12.5.

In case of NR method it is possible to eliminate this effect because the area where Newton's method is "well" defined remains stable (see Table 12.6).

Finally we would like to mention the problem (12.1) is equivalent to

$$\min z = x_{n+1} \tag{12.58}$$

s.t.

$$f_i(x) - x_{n+1} \le 0, \; i = 1, \ldots, m \tag{12.59}$$

It turns out that replacing (12.1) by a constrained optimization problems (12.58)-(12.59) leads to substantial increase of the total number of Newton's steps although for the problem (12.58)-(12.59) we also applied the NR method.

In this case we also observed the "hot start" phenomenon, which is typical for NR methods, but the total number of Newton's steps is almost ten times

it	$\|g\|/n$	gap	# of steps
0	4.271734e+09	6.105779e+04	0
1	4.407985e-01	1.769816e-01	21
2	1.247075e-02	3.321384e-03	12
3	3.608019e-04	5.814294e-05	1
4	5.087897e-05	3.158259e-06	1
5	1.774342e-05	2.369148e-07	1
6	9.149201e-06	1.979549e-08	1
7	1.133337e-06	1.513192e-09	1
Total number of Newton's steps			38

Table 12.1 Quadratic Minimax. Newton's NR method. $n = 500$, $m = 300$, $r = 100$, $\mu = 0.1 = const$

more than in case when Newton's NR method was applied to discrete minimax problem directly.

It reflects "degeneracy" phenomenon which is due to the extension of the primal space. Newton's method being applied in the framework of NR technique for the problem (12.58)-(12.59) turns out to be substantially less efficient because the corresponding system of linear equations are far from being as stable as the corresponding system when NR technique is applied to the discrete minimax problems directly.

7. NUMERICAL RESULTS

The first three tables (12.1), (12.2) and (12.3) present the results for three different random generated problems with the same number of variables $n = 500$ and the same number $m = 300$ of functions $f_i(x)$ but different number r of active functions. We applied Newton's NR method to solve these problems.

The next rwo tables (Table 12.4 , Table 12.5) present results obtained by using smoothing technique for the same random generated minimax problem with $n = 300$, $m = 200$ and $r = 100$. We use different strategies for the scaling parameter update, but the total number of Newton's step is about the same in both examples.

Table 12.6 shows the performance of NR Newton's method for this problem.

it	$\|g\|/n$	gap	# of steps
0	3.987928e+10	1.312341e+03	0
1	1.563947e+00	4.702257e-01	26
2	2.398495e-02	4.586030e-03	10
3	1.480449e-04	4.008626e-05	2
4	4.998604e-05	2.391677e-07	1
5	8.392672e-06	9.477171e-10	1
Total number of Newton's steps			40

Table 12.2 Quadratic Minimax. Newton's NR method. $n = 500$, $m = 300$, $r = 10$, $\mu = 0.1 = const$

it	$\|g\|/n$	gap	# of steps
0	6.123844e+11	4.900217e+05	0
1	4.055045e-01	1.066249e-02	22
2	3.830370e-03	3.198765e-04	17
3	1.167304e-03	2.949390e-05	2
4	5.054220e-04	4.937376e-06	1
5	1.774485e-04	8.228653e-07	1
6	2.817522e-05	* 1.266776e-07	1
7	5.063400e-05	1.846078e-08	1
8	1.996417e-05	2.019193e-09	1
Total number of Newton's steps			46

Table 12.3 Quadratic Minimax. Newton's NR method. $n = 500$, $m = 300$, $r = 280$, $\mu = 0.1 = const$

it	$\|g\|/n$	gap	# of steps
0	3.8761290+10	5.544589e+04	0
1	6.645551e-02	9.767213e-02	18
2	3.209086e-02	8.295744e-03	13
3	6.623261e-02	9.155524e-04	23
4	9.243970e-02	9.878368e-05	32
5	5.570453e-02	4.583108e-06	14
6	8.603320e-02	1.017187e-06	43
7	5.195266e-02	6.913358e-08	8
8	3.564213e-02	7.167045e-09	20
Total number of Newton's steps			171

Table 12.4 Smoothing method. $n = 300$, $m = 200$, $r = 100$, $\mu_0 = 0.1$, $\gamma = 0.1$

it	$\|g\|/n$	gap	# of steps
0	3.8761290+10	5.544589e+04	0
1	6.645551e-02	9.767213e-02	18
2	8.060323e-02	1.849595e-02	6
3	2.669111e-02	3.396737e-03	6
4	6.872043e-02	3.271389e-04	9
5	4.925083e-02	8.335165e-05	18
6	8.247744e-02	2.303061e-05	18
7	9.735537e-02	5.135152e-07	12
8	3.791857e-02	8.134346e-07	21
9	8.744086e-02	4.204158e-08	14
10	6.969266e-02	1.661993e-08	19
11	7.516014e-02	2.520661e-09	16
Total number of Newton's steps			157

Table 12.5 Smoothing method. $n = 300$, $m = 200$, $r = 100$, $\mu_0 = 0.1$, $\gamma = 0.2$

it	$\|g\|/n$	gap	# of steps
0	3.8761290+10	5.544589e+04	0
1	5.372130e-01	9.724620e-02	20
2	1.104875e-02	1.305847e-03	14
3	1.219049e-03	1.446832e-05	2
4	3.675979e-04	7.240549e-07	1
5	1.818994e-04	1.547403e-07	1
6	7.641389e-05	4.647747e-08	1
7	8.322390e-06	2.197183e-09	1
Total number of Newton's steps			40

Table 12.6 Newton's NR method. $n = 300$, $m = 200$, $r = 100$, $\mu_0 = 0.1$, $\gamma = 0.66$

it	$\|g\|/n$	gap	# of steps
0	2.0298734+11	6.736350e+05	0
1	2.980716e-01	6.716260e-02	23
2	4.644735e-03	9.155294e-04	16
3	8.997086e-04	9.092558e-06	2
4	1.176992e-04	2.206625e-06	1
5	3.137098e-05	8.531576e-08	1
6	4.392721e-05	1.163497e-08	1
7	5.515149e-05	6.744794e-09	1
Total number of Newton's steps			45

Table 12.7 Nonlinear Rescaling. $n = 1000$, $m = 500$, $r = 300$, $\mu_0 = 0.1$, $\gamma = 0.66$

Table 12.7 represents the performance of NR method for the randon generated problem with $n = 1000$. The results reflect the fact that the number of Newton's steps is independent on the dimension of the problem.

it	$\|g\|/n$	gap	constr violat	# of steps
1	1.912858e+01	1.922016e+01	5.632464e-02	164
2	9.531485e-01	4.491116e-01	7.120833e-02	131
3	7.724585e-02	1.912551e-04	5.936578e-02	42
4	9.783521e-04	2.316859e-05	6.524854e-03	11
5	2.033553e-03	2.149152e-05	6.106230e-03	2
6	6.512459e-05	1.472351e-05	8.429840e-04	2
7	2.501971e-04	6.816172e-06	1.345928e-04	1
8	1.725497e-04	5.279747e-07	2.446371e-04	1
9	5.914055e-05	7.340055e-07	8.454459e-07	1
10	1.311045e-05	1.156677e-07	7.921827e-08	1
11	8.005334e-06	1.963896e-08	1.163732e-08	1
12	3.563255e-06	6.889967e-10	1.429145e-09	1
Total number of Newton's steps				358

Table 12.8 Constrained minimization. $n = 300$, $m = 200$, $r = 100$, $\mu = 0.1 = const$

Table 12.8 shows the performance of NR method applied to equivalent constrained optimization problem. The chosen problem is the same as in Table 12.4, Table 12.5 and Table 12.6. and Table 12.5.

Finally in Tables 12.9, 12.10, 12.11 we present results obtained with Newton's NR method for three problems of finding Chebichev center for $m = 200$ and $m = 1000$ points in $\mathbb{R}^2$ and one Chebichev center for $m = 10$ points in $\mathbb{R}^{50}$.

The next tables show the performance of NR algorithm for Chebichev center problems in $\mathbb{R}^n$. The number of points is m.

8. CONCLUDING REMARKS

The NR approach for discrete minimax produced results, which are in full compliance with the outlined theory. In particular, we systematically observed the so-called "hot start" phenomenon, which has been predicted in several papers where NR approach was applied for constrained optimization [Polyak, 1992] [Melman and Polyak, 1996].

Due to the "hot start" from some point on it is possible to reduce substantially the number of Newton's steps per Lagrange multiplier update. Moreover, from some point on only one Newton's step is enough to update the Lagrange multi-

it	dual value	gap	# of steps
0	3.360553e+01	5.289076e+01	0
1	1.094691e+01	1.261813e+01	4
2	2.167676e+01	1.079517e-01	6
3	2.173044e+01	4.695135e-03	2
4	2.173333e+01	1.054259e-03	1
5	2.173397e+01	2.969259e-04	1
6	2.173415e+01	8.990378e-05	1
7	2.173421e+01	2.781590e-05	1
8	2.173422e+01	8.666529e-06	1
9	2.173423e+01	2.706392e-06	1
10	2.173423e+01	8.457703e-07	1
11	2.173423e+01	2.643709e-07	1
12	2.173423e+01	8.264301e-08	1
13	2.173423e+01	2.583498e-08	1
14	2.173423e+01	8.076324e-09	1
Total number of Newton's steps			23

Table 12.9 Chebichev center. $n = 2$, $m = 200$, $\mu = 1 = const$

it	dual value	gap	# of steps
0	3.041008e+01	6.559907e+01	0
1	1.181194e+01	1.274004e+01	4
2	2.330307e+01	4.434190e-01	4
3	2.353763e+01	1.663498e-01	2
4	2.360860e+01	8.056320e-02	1
5	2.364110e+01	4.206331e-02	1
6	2.365769e+01	2.206217e-02	1
7	2.366610e+01	1.137728e-02	1
8	2.367015e+01	5.705734e-03	1
9	2.367195e+01	2.777420e-03	1
10	2.367268e+01	1.323817e-03	1
11	2.367296e+01	6.231645e-04	1
12	2.367307e+01	2.906277e-04	1
13	2.367310e+01	1.343539e-04	1
14	2.367312e+01	6.160661e-05	1
15	2.367312e+01	2.805813e-05	1
16	2.367313e+01	1.271270e-05	1
17	2.367313e+01	5.738289e-06	1
18	2.367313e+01	2.583294e-06	1
19	2.367313e+01	1.160820e-06	1
20	2.367313e+01	5.209635e-07	1
21	2.367313e+01	2.336032e-07	1
22	2.367313e+01	1.046899e-07	1
23	2.367313e+01	4.690010e-08	1
24	2.367313e+01	2.100626e-08	1
25	2.367313e+01	9.407504e-09	1
Total number of Newton's steps			32

Table 12.10 Chebichev center. $n = 2, m = 1000, \mu = 1 = const$

it	dual value	gap	# of steps
0	9.026055e+02	1.493950e+02	0
1	2.073758e+02	4.081056e+01	64
2	2.192080e+02	1.163328e-01	5
3	2.192081e+02	8.438290e-04	3
4	2.192081e+02	6.499606e-06	1
5	2.192081e+02	6.426757e-07	1
6	2.192081e+02	6.929753e-09	1
Total number of Newton's steps			75

Table 12.11 Chebichev center. $n = 50, m = 10, \mu = 1 = const$

pliers. This phenomenon allows to improve the numerical stability and obtain results with high accuracy.

Still a number of issues require further attention.

First, we have to understand better the efficiency of the primal-dual method for the discrete minimax.

Global convergence of the NR type methods in the absence of the standard second order optimality condition is the second issue.

Also it is important to characterize the "hot start" phenomenon, i.e. to understand better when "hot start" occurs. It would allow to combine the smoothing technique in the initial phase of the process with Newton's NR or Primal-Dual method in the final phase .

Using a vector scaling parameters, one parameter for each functions f_i is another line of research, which we are panning to pursue along with wide numerical experiments.

Acknowledgments

The first author, was partially supported by NSF Grant DMS-9705672. The work of the second author has been artially supported by NASA Grant NAG-1-1929.

References

Auslender, A., Cominetti, R., and Haddou, M. (1997). Asymptotic analysis for penalty and barrier method in convex and linear programming, *Mathematics of Operations Research* 22:43-62.

Ben-Tal, A., and Nemirovsky, A. (1998). *Convex optimization in engineering: Modeling analysis, algorithms,* Technion, Israel.

Bertsekas, D. (1982). *Constrained optimization and Lagrange multipliers methods,* Academic Press, New York.

Charalambous, C. (1977). Nonlinear least P-th optimization and nonlinear programming, *Mathematical Programming* 12:195-225.

Chen, C., and Mangasarian, O. (1995). Smoothing methods for convex inequalities and linear complementarity problems, *Mathematical Programming* 71:51-69.

Fiacco, A., and McCormick, G. (1990). *Nonlinear programming. Sequential unconstrained minimization techniques,* SIAM Classic in Applied Mathematics, SIAM Philadelphia, PA.

Demyanov, V., and Malozemov, V. (1974). *Introduction to minimax,* John Wiley NY.

Kiwiel, K. (1985). *Methods of descent for nondifferentiable optimization,* Lectures Notes in Mathematics, Springer-Verlay, Berlin, 1133:362.

Lemarechal, C. (1989). Nondifferentiable optimization, in: , G. Nemhauser, A. Rinnooy Kan, M. Todd, eds., Handbooks in Operations Research and Management Science, 1:529-572.

Melman, A., and Polyak, R. (1996). The Newton modified barrier method for QP problems, *Annals of Operations Research,* 62:465-519.

Nesterov, Yu., and Nemirovsky, A. (1993). *Interior-Point Polynomial Algorithms in Convex Optimization,* SIAM Studies in Applied Mathematics, Philadelphia, 1993.

Polyak, B., and Tretyakov, N. (1974). The method of penalty estimates for conditional extremum problems *Computational Mathematics and Mathematical Physics* 13:42-58.

Polyak, R. (1971). On the best convex Chebichev approximation, *Soviet Mathematics Doklady,* 200(5).

Polyak, R. (1988). Smooth optimization methods for minimax problems, *SIAM Journal Control and Optimization,* 26(6).

Polyak, R. (1992). Modified barrier functions, *Mathematical Programming* 54:177-222.

Polyak, R., and Teboulle, M. (1997). Nonlinear rescaling and proximal-like methods in convex optimization, *Mathematical Programming* 76:265-284.

Polyak, R. (1999). *Log-sigmoid multipliers method in constrained optimization,* Research Report, Department of SEOR & Mathematical Sciences Department, GMU, 1-39.

Shor, N. (1998). *Nondifferentiable optimization and Polynomial Problems,* Kluwer Academic Publishers, Boston.

Smale, S. (1986). Newton's method estimates from data at one point, in R.E. Ewing et al. eds., *The merging of Disciplines in Pure, Applied and Computational Mathematics,* Springer, New York-Berlin, 185-196.

Sobieszczanski-Sobieski, J. (1992). A technique for locating function-roots and satisfying equality constraints in optimization, *Structural Optimization* 4:241-243.