Working Paper

A Partial Regularization Method for Saddle Point Seeking

Andrzej Ruszczyński

WP-94-20
March 1994

IIASA International Institute for Applied Systems Analysis □ A-2361 Laxenburg □ Austria
Telephone: +43 2236 715210 □ Telex: 079 137 iiasa a □ Telefax: +43 2236 71313

A Partial Regularization Method for Saddle Point Seeking

Andrzej Ruszczyński

WP-94-20
March 1994

International Institute for Applied Systems Analysis □ A-2361 Laxenburg □ Austria
Telephone: +43 2236 715210 □ Telex: 079 137 iiasa a □ Telefax: +43 2236 71313

Abstract

This article generalizes the Nash equilibrium approach to linear programming to the saddle point problem. The problem is shown to be equivalent to a non-zero sum game in which objectives of the players are obtained by partial regularization of the original function. Based on that, a solution method is developed in which the players improve their decisions while anticipating the steps of their opponents. Strong convergence of the method is proved and application to convex optimization is discussed.

Key words: Saddle point, regularization, augmented Lagrangian, decomposition.

1. Introduction

Let $L : R^n \times R^m \to R$ be a finite convex-concave function and let $X \subset R^n$ and $Y \subset R^m$ be closed convex sets. The objective of this paper is to develop a method for finding a *saddle point* of L over $X \times Y$, i.e., a point $(\hat{x}, \hat{y}) \in X \times Y$ such that

$$L(\hat{x}, y) \leq L(\hat{x}, \hat{y}) \leq L(x, \hat{y}), \quad \forall x \in X, \ \forall y \in Y. \tag{1.1}$$

This is one of fundamental problems of convex programming and game theory (for a thorough treatment of the theory of saddle functions we refer the reader to [8]). There were many attempts to develop saddle point seeking procedures; the simplest algorithm (see, e.g., [1]) has the form

$$\begin{aligned} x^{k+1} &= \Pi_X \left(x^k - \tau_k L_x(x^k, y^k) \right), \\ y^{k+1} &= \Pi_X \left(y^k + \tau_k L_y(x^k, y^k) \right), \quad k = 1, 2, \ldots, \end{aligned}$$

where $L_x(x^k, y^k)$ and $L_y(x^k, y^k)$ are some subgradients of L at (x^k, y^k) with respect to x and y, and $\Pi_X(\cdot)$ and $\Pi_Y(\cdot)$ denote orthogonal projections on X and Y, respectively. Such methods are convergent only under special conditions (like strict convexity-concavity) and with special stepsizes for primal and dual updates: $\tau_k \to 0$, $\sum_{k=0}^{\infty} \tau_k = \infty$ (cf. [7]).

One possibility to overcome these difficulties is the use of the *proximal point method* [6, 10]. Its idea is to replace (1.1) by a sequence of saddle-point problems for regularized functions

$$\Lambda_k(\xi, \eta) = L(\xi, \eta) + \frac{\rho}{2} \|\xi - x^k\|^2 - \frac{\rho}{2} \|\eta - y^k\|^2. \tag{1.2}$$

A saddle point (ξ^k, η^k) of Λ_k is substituted for (x^{k+1}, y^{k+1}) at the next iteration, etc. A variation of this approach is the *alternating direction method* [3, 2].

We are going to develop an iterative method for (1.1) which does not have saddle-point subproblems. The key idea, which generalizes and simplifies the concept used for linear programming in a recent work [4] of Kallio and ours, is to replace the regularized function (1.2) by two convex-concave functions: a primal and a dual one, and to make steps in x and in y using subgradients of these functions. We shall develop the basic concept in section 2, and in section 3 we describe the method. Next, in section 4 we prove its strong convergence to a saddle point of L. Finally, in section 5 we discuss the application of this approach to some convex optimization problems of special structure.

For a convex set $X \subset R^n$, the cone of feasible directions at $x \in X$ is denoted by $K_X(x) = \{d \in R^n : \exists(\tau > 0) \ x + \tau d \in X\}$. The conjugate (negative of the polar) of a cone $K \subset R^n$ is defined to be $K^* = \{d \in R^n : \forall(x \in K) \ \langle d, x \rangle \geq 0\}$. For a convex-concave function $L : R^n \times R^m \to R$ we use $\partial_x L(x, y)$ and $\partial_y L(x, y)$ to denote its subdifferentials with respect to x and y. Elements of these subdifferentials (subgradients) will be denoted by $L_x(x, y)$ and $L_y(x, y)$.

2. The game

Let us define a non-cooperative game with two players: **P** and **D**. The objective of **P** is to minimize in the variables $x \in X$ the *regularized primal function*:

$$P(x,y) = \max_{\eta \in Y} \left[L(x,\eta) - \frac{\rho}{2}\|\eta - y\|^2 \right], \tag{2.1}$$

where $\rho > 0$ is some parameter. The objective of **D** is to maximize with respect to the variables $y \in Y$ the *regularized dual function*:

$$D(x,y) = \min_{\xi \in X} \left[L(\xi,y) + \frac{\rho}{2}\|\xi - x\|^2 \right]. \tag{2.2}$$

A *Nash equilibrium* of the game is defined as a point $(\hat{x},\hat{y}) \in X \times Y$ such that

$$\hat{x} \in \arg\min_{x \in X} P(x,\hat{y}), \tag{2.3}$$

and

$$\hat{y} \in \arg\max_{y \in Y} D(\hat{x},y). \tag{2.4}$$

We define the proximal mappings $\xi(x,y)$ and $\eta(x,y)$ as the solutions of the subproblems in (2.2) and (2.1). We also introduce the error functions

$$E(x,y) = \max_{g \in \partial_x L(\xi,y)} \langle g, x - \xi \rangle - \min_{h \in \partial_y L(x,\eta)} \langle h, y - \eta \rangle,$$

and

$$\Delta(x,y) = \|\xi - x\|^2 + \|\eta - y\|^2,$$

where $\xi = \xi(x,y)$ and $\eta = \eta(x,y)$. They satisfy the following relations.

Lemma 1. *For all $x \in X$ and $y \in Y$,*

$$\rho\Delta(x,y) \le E(x,y) \le L(x,\eta(x,y)) - L(\xi(x,y),y).$$

Proof. By the definition of $\xi = \xi(x,y)$, there exists a subgradient $L_x(\xi,y)$ such that

$$L_x(\xi,y) + \rho(\xi - x) \in K_X^*(\xi). \tag{2.5}$$

As $x - \xi \in K_X(\xi)$, we have

$$L(x,y) - L(\xi,y) \ge \max_{g \in \partial_x L(\xi,y)} \langle g, x - \xi \rangle \ge \langle L_x(\xi,y), x - \xi \rangle \ge \rho\|\xi - x\|^2.$$

In a symmetric way, from the definition of $\eta = \eta(x,y)$ it follows that

$$L(x,y) - L(x,\eta) \le \min_{h \in \partial_y L(x,\eta)} \langle h, y - \eta \rangle \le \langle L_y(x,\eta), y - \eta \rangle \le -\rho\|\eta - y\|^2.$$

Subtracting the last two inequalities, we obtain the required result. □

We can now prove the equivalence of (1.1) and our game.

Theorem 1. *The following three statements are equivalent:*

(a) $(\hat{x}, \hat{y})$ *is a Nash equilibrium of the game (2.3)-(2.4);*

(b) $E(\hat{x}, \hat{y})) = 0$*;*

(c) $(\hat{x}, \hat{y})$ *is a saddle point of* L *over* $X \times Y$*.*

Proof. We denote $\hat{\xi} = \xi(\hat{x}, \hat{y})$ and $\hat{\eta} = \eta(\hat{x}, \hat{y})$.

(a)⇒(b). Since $\rho > 0$, the function $\eta(x, y)$ is continuous. Therefore $\partial_x P(x, y) = \partial_x L(x, \eta(x, y))$. Using this equality in the optimality conditions for (2.3), we deduce that there exists a subgradient $L_x(\hat{x}, \hat{\eta}) \in K_X^*(\hat{x})$. Thus

$$L(\hat{\xi}, \hat{\eta}) - L(\hat{x}, \hat{\eta}) \geq \langle L_x(\hat{x}, \hat{\eta}), \hat{\xi} - \hat{x} \rangle \geq 0.$$

Analogously, optimality conditions for (2.4) yield $-L_y(\hat{\xi}, \hat{y}) \in K_Y^*(\hat{y})$ for some subgradient $L_y(\hat{\xi}, \hat{y})$, so

$$L(\hat{\xi}, \hat{y}) - L(\hat{\xi}, \hat{\eta}) \geq 0.$$

By Lemma 1,

$$L(\hat{x}, \hat{\eta}) - L(\hat{\xi}, \hat{y}) \geq E(\hat{x}, \hat{y}).$$

Adding the last three inequalities, we obtain (b).

(b)⇒(c). Lemma 1 implies that $\Delta(\hat{x}, \hat{y}) = 0$, so $\hat{\xi} = \hat{x}$ and $\hat{\eta} = \hat{y}$. By (2.5), $L_x(\hat{x}, \hat{y}) \in K_X^*(\hat{x})$ for some $L_x(\hat{x}, \hat{y})$. This is equivalent to the right inequality in (1.1). Similarly, $-L_y(\hat{x}, \hat{y}) \in K_Y^*(\hat{y})$ for some $L_y(\hat{x}, \hat{y})$, which completes the proof of (c).

(c)⇒(a). The left inequality in (1.1) implies

$$L(\hat{x}, \hat{y}) = \max_{\eta \in Y} L(\hat{x}, \eta) = \max_{\eta \in Y} \left[L(\hat{x}, \eta) - \frac{\rho}{2} \|\eta - \hat{y}\|^2 \right] = P(\hat{x}, \hat{y}).$$

On the other hand, for every $x \in X$, from the right inequality in (1.1) we get

$$L(\hat{x}, \hat{y}) \leq L(x, \hat{y}) \leq \max_{\eta \in Y} \left[L(x, \eta) - \frac{\rho}{2} \|\eta - \hat{y}\|^2 \right] = P(x, \hat{y}).$$

Consequently, $P(\hat{x}, \hat{y}) \leq P(x, \hat{y})$ for all $x \in X$. In the same manner we prove $D(\hat{x}, \hat{y}) \geq D(\hat{x}, y)$ for all $y \in Y$. □

3. The method

Let us now describe in detail a method for finding a saddle point of L. It is, in fact, an algorithm for solving the game (2.3)-(2.4). It can also be interpreted as a method in which both players try to predict the moves of their opponents to calculate the best response.

Initialization. Choose $x^0 \in X$, $y^0 \in Y$ and $\gamma \in (0,2)$. Set $k = 0$.

Prediction. Calculate $\eta^k = \eta(x^k, y^k)$ and $\xi^k = \xi(x^k, y^k)$.

Stopping test. If $E_k = E(x^k, y^k) = 0$, then stop.

Direction finding. Find subgradients $L_x(x^k, \eta^k)$ and $L_y(\xi^k, y^k)$ and define

$$\begin{aligned} d_x^k &= \Pi_{C_X^k}\left(-L_x(x^k, \eta^k)\right), \\ d_y^k &= \Pi_{C_Y^k}\left(L_y(\xi^k, y^k)\right), \end{aligned}$$

where C_X^k and C_Y^k are closed convex cones such that $C_X^k \supset K_X(x^k)$ and $C_Y^k \supset K_Y(y^k)$.

Stepsize calculation. Determine

$$\tau_k = \frac{\gamma E_k}{\|d^k\|^2}. \tag{3.1}$$

Step. Update the points

$$\begin{aligned} x^{k+1} &= \Pi_X\left(x^k + \tau_k d_x^k\right), \\ y^{k+1} &= \Pi_Y\left(y^k + \tau_k d_y^k\right), \end{aligned}$$

increase k by one and go to Prediction.

Our method resembles in some way the *extragradient method* of [5], but our prediction step uses proximal operators, not just a linear Jacobi step. Owing to that, we can solve nonsmooth problems. We also have a constructive stepsize rule, although calculation of directions and stepsizes is somewhat unusual. Still, the use of $C_X^k = \operatorname{cl} K_X(x^k)$, $C_Y^k = \operatorname{cl} K_Y(y^k)$ and of (3.1) is easy in some clases of problems (like polyhedral ones) and yields larger stepsizes. If such choices are not implementable, we may set $C_X^k = R^n$ and $C_Y^k = R^m$ and replace E_k with $L(x^k, \eta^k) - L(\xi^k, y^k)$ or $\rho\Delta(x^k, y^k)$ (see the remarks after the proof of convergence).

4. Convergence

To avoid obscuring the main idea, we shall now prove convergence of the method in its basic form, presented in the previous section. Various modifications and extensions will be discussed after the proof.

Theorem 2. *Assume that a saddle point of L on $X \times Y$ exists. Then the method generates a sequence $\left\{(x^k, y^k)\right\}_{k=0}^{\infty}$ convergent to a saddle point of L on $X \times Y$.*

Proof. Let (x^*, y^*) be a saddle point of L on $X \times Y$. We define

$$W_k = \|x^k - x^*\|^2 + \|y^k - y^*\|^2. \tag{4.1}$$

Since the projection on X is non-expansive,

$$\begin{aligned} \|x^{k+1} - x^*\|^2 &\leq \|x^k + \tau_k d_x^k - x^*\|^2 \\ &= \|x^k - x^*\|^2 + 2\tau_k \langle d_x^k, x^k - x^* \rangle + \tau_k^2 \|d_x^k\|^2. \end{aligned} \tag{4.2}$$

Using the formula $h = \Pi_C(h) + \Pi_{-C^*}(h)$, which holds for any closed convex cone C (cf. [12]), with $h = -L_x(x^k, \eta^k)$ and $C = C_X^k$, we obtain

$$-L_x(x^k, \eta^k) = d_x^k + \Pi_{-(C_X^k)^*}\left(L_x(x^k, \eta^k)\right).$$

Multiplying both sides of this equation by $x^* - x^k \in K_X(x^k) \subset C_X^k$ we get the inequality

$$\langle d_x^k, x^* - x^k \rangle \geq \langle L_x(x^k, \eta^k), x^k - x^* \rangle \geq L(x^k, \eta^k) - L(x^*, \eta^k).$$

Substituting the above estimate into (4.2) yields

$$\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 - 2\tau_k \left[L(x^k, \eta^k) - L(x^*, \eta^k)\right] + \tau_k^2 \|d_x^k\|^2.$$

Likewise, by obvious symmetry, we obtain

$$\|y^{k+1} - y^*\|^2 \leq \|y^k - y^*\|^2 + 2\tau_k \left[L(\xi^k, y^k) - L(\xi^k, y^*)\right] + \tau_k^2 \|d_y^k\|^2.$$

Adding the last two inequalities we conclude that

$$W_{k+1} \leq W_k - 2\tau_k \left[L(x^k, \eta^k) - L(x^*, \eta^k) - L(\xi^k, y^k) + L(\xi^k, y^*)\right] + \tau_k^2 \|d^k\|^2. \tag{4.3}$$

The saddle point conditions imply that $L(x^*, \eta^k) \leq L(\xi^k, y^*)$. By Lemma 1, $L(x^k, \eta^k) - L(\xi^k, y^k) \geq E_k$. Therefore (4.3) can be rewritten as follows:

$$W_{k+1} \leq W_k - 2\tau_k E_k + \tau_k^2 \|d^k\|^2. \tag{4.4}$$

Substituting (3.1) we get

$$W_{k+1} \leq W_k - \frac{\gamma(2-\gamma)E_k^2}{\|d^k\|^2}. \tag{4.5}$$

Thus the sequence $\{W_k\}$ is non-increasing and

$$\lim_{k \to \infty} \frac{E_k^2}{\|d^k\|^2} = 0. \tag{4.6}$$

Since W_k is bounded, the sequence $\{(x^k, y^k)\}$ has an accumulation point $(\hat{x}, \hat{y})$. Thus $\{d^k\}$ is bounded and, by (4.6), $\lim_{k \to \infty} E_k = 0$. Therefore $E(\hat{x}, \hat{y}) = 0$. By Theorem 1, $(\hat{x}, \hat{y})$ is a saddle point of L and we can use it instead of (x^*, y^*) in (4.1). Then, from (4.5) we see that the distance to $(\hat{x}, \hat{y})$ is non-increasing. Consequently, $(\hat{x}, \hat{y})$ is the only accumulation point of the sequence $\{(x^k, y^k)\}$. □

It is clear from the proof that we may replace the stepsize rule (3.1) with a more flexible requirement,

$$\frac{\lambda \rho \Delta_k}{\|d^k\|^2} \leq \tau_k \leq \frac{\gamma(L(x^k, \eta^k) - L(\xi^k, y^k))}{\|d^k\|^2},$$

with $\Delta_k = \Delta(x^k, y^k)$ and $0 < \lambda \leq \gamma < 2$. Indeed, (4.3) implies

$$W_{k+1} \leq W_k - \frac{\lambda(2-\gamma)\rho^2\Delta_k^2}{\|d^k\|^2}. \tag{4.7}$$

The rest of the proof is the same, but with Δ_k instead of E_k. We can also have iteration-dependent parameters $\rho_k > 0$ and $0 < \lambda_k \leq \gamma_k < 2$, provided that $\sum_{k=0}^{\infty} \lambda_k(2-\gamma_k)\rho_k^2 = \infty$, because (4.7) still implies $\liminf_{k\to\infty} \Delta_k = 0$.

5. Application to decomposable problems

Let us consider a convex programming problem of the form

$$\min \sum_{j=1}^{n} f_j(x_j) \tag{5.1}$$

$$\sum_{j=1}^{n} g_{ij}(x_j) \leq b_i, \ i = 1, \ldots, m, \tag{5.2}$$

$$x_j \in X_j, \ j = 1, \ldots, n. \tag{5.3}$$

We assume that the functions f_j and g_{ij} are convex and the sets X_j are convex and closed. As usual, we introduce multipliers $y \in R_+^m$ and the Lagrangian

$$L(x, y) = \sum_{j=1}^{n} f_j(x_j) + \sum_{i=1}^{m} y_i \left(\sum_{j=1}^{n} g_{ij}(x_j) - b_i \right).$$

Under the constraint qualification condition (see, e.g., [8]), problem (5.1)-(5.3) is equivalent to finding a saddle point of L on the product of $X = X_1 \times \cdots \times X_n$ and $Y = R_+^m$. Our method, when applied to this problem, takes a rather simple form.

Indeed, the prediction step in the dual variables can be carried out analytically, separately for each constraint:

$$\eta_i^k = \max \left(0, \frac{1}{\rho} \left(\sum_{j=1}^{n} g_{ij}(x_j^k) - b_i \right) + y_i^k \right), \ i = 1, \ldots, m. \tag{5.4}$$

The resulting regularized primal function (2.1) is the augmented Lagrangian (cf. [9]) for (5.1)-(5.3):

$$P(x, y) = \sum_{j=1}^{n} f_j(x_j) + \frac{\rho}{2} \sum_{i=1}^{m} \left[\max \left(0, \frac{1}{\rho} \left(\sum_{j=1}^{n} g_{ij}(x_j) - b_i \right) + y_i \right) \right]^2 - \frac{\rho}{2} \sum_{i=1}^{m} (y_i)^2.$$

Consequently, the update of primal variables is a projected subgradient step for the augmented Lagrangian function. It is clearly decomposable. Note that in a related work [11] of ours, we used here a whole sequence of nonlinear Jacobi-type steps.

The prediction step in primal variables is decomposable into subproblems

$$\min_{\xi_j \in X_j} \left[f_j(\xi_j) + \sum_{i=1}^{m} y_i^k g_{ij}(\xi_j) + \frac{\rho}{2} \|\xi_j - x_j^k\|^2 \right], \ j = 1, \ldots, n. \tag{5.5}$$

Their results ξ_j^k are then used in the dual update, which is just an under-relaxed step of the multiplier method, very similar to (5.4):

$$y_i^{k+1} = \max \left(0, \frac{\tau_k}{\rho} \left(\sum_{j=1}^{n} g_{ij}(\xi_j^k) - b_i \right) + y_i^k \right), \ i = 1, \ldots, m.$$

In some cases, subproblems (5.5) can be quite easy to solve. The simplest example is the standard linear programming problem with $f_j(x_j) = c_j x_j$, $g_{ij}(x_j) = a_{ij} x_j$ and $X_j = [l_j, u_j]$. Then (5.5) has a closed-form solution, which can be calculated in parallel for each $j = 1, \ldots, n$. It is worth noting that the regularized dual function $D(x, y)$ becomes the augmented Lagrangian function for the dual problem. Properties of our method in the case of linear programming are analyzed in detail in [4], with limit properties of the stepsizes τ_k, with the analysis of the rate of convergence, and with some numerical results. In fact, the highly encouraging properties discovered in [4] motivated the research reported in the present paper.

References

[1] K.J. Arrow, L. Hurwicz and H. Uzawa, *Studies in Linear and Nonlinear Programming* (Stanford University Press, Stanford, 1958).

[2] J. Eckstein and D.P. Bertsekas, "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Mathematical Programming* 55 (1992) 293-318.

[3] D. Gabay, "Application de la méthode des multiplicateurs aux inéquations variationelles," in: M. Fortin and R. Glowinski (eds.), *Méthodes de Lagrangien Augmenté* (Dunod, Paris, 1982) pp. 279-307.

[4] M. Kallio and A. Ruszczyński, "Parallel solution of linear programs via Nash equilibria," working paper WP-94-15, IIASA, Laxenburg, 1994.

[5] G.M. Korpelevich, "The extragradient method for finding saddle points and other problems," *Ekonomika i Matematicheskie Metody* 12 (1976) 747-756.

[6] B. Martinet, "Regularisation d'inéquations variationelles par approximations successives," *Rev. Francaise Inf. Rech. Oper.* 4 (1970) 154-159.

[7] A.S. Nemirovski and D.B. Yudin, "Cesaro convergence of the gradient method for approximation of saddle points of convex-concave functions," *Doklady AN SSSR* 239 (1978) 1056-1059.

[8] R.T. Rockafellar, *Convex Analysis* (Princeton University Press, Princeton, 1970).

[9] R.T. Rockafellar, "Augmented Lagrangians and applications of the proximal point algorithm in convex programming," *Mathematics of Operations Research* 1 (1976) 97-116.

[10] R.T. Rockafellar, "Monotone operators and the proximal point algorithm," *SIAM J. Control and Optimization* 14 (1976) 977-898.

[11] A. Ruszczyński, "Augmented Lagrangian decomposition for sparse convex optimization," working paper WP-92-75, IIASA, Laxenburg, 1992 (to appear in *Mathematics of Operations Research*).

[12] A. Wierzbicki and S. Kurcyusz, "Projection on a cone, penalty functionals and duality theory for problems with inequality constraints in Hilbert space," *SIAM J. Control and Optimization* 15 (1977) 25-56.