Ergodic Results and Bounds on the Optimal Value in Subgradient Optimization

Torbjörn Larsson[1], *Michael Patriksson*[2], *and Ann-Brith Strömberg*[1]

[1] Linköping Institute of Technology, S-581 83 Linköping, Sweden
[2] University of Washington, Seattle, WA 98195-4350, USA

1 Introduction

Subgradient methods for nonsmooth optimization have rendered considerable popularity, mainly in the context of Lagrangean relaxation; this is due to their simplicity and proven practical usefulness. We consider the minimization of a nonsmooth convex function over a convex and compact set, using a (conditional) subgradient optimization scheme with divergent series step lengths. We show that such a scheme may be augmented by a convergent lower bounding procedure based on the computation of an ergodic (averaged) sequence of subgradients and of affine functions underestimating the objective. The sequence of lower bounds enables a judgement of the quality of the solutions obtained, and can thus be used in a termination criterion. Numerical experience from an application of the bounding procedure within a subgradient scheme for a nonsmooth formulation of the linear multicommodity network flow problem is presented.

2 An Ergodic Result in Subgradient Optimization

Let the function $f : \Re^n \mapsto \Re$ be nonsmooth and convex and thus everywhere continuous and subdifferentiable, and let $X \subset \Re^n$ be a nonempty, convex, and compact set. The program considered is

$$[\mathrm{P}] \qquad f^* = \min_{\mathbf{x} \in X} f(\mathbf{x}),$$

with the nonempty, convex, and compact solution set X^*. The *subdifferential* of f at $\mathbf{x} \in \Re^n$ is (e.g., Hiriart-Urruty and Lemaréchal, 1993, Definition VI:1.2.1) the nonempty, convex, and compact set

$$\partial f(\mathbf{x}) = \left\{ \boldsymbol{\gamma} \in \Re^n \;\middle|\; f(\mathbf{y}) \geq f(\mathbf{x}) + \boldsymbol{\gamma}^T(\mathbf{y} - \mathbf{x}), \quad \forall \mathbf{y} \in \Re^n \right\}, \tag{1}$$

the elements of which are called *subgradients*. The *normal cone* of the set X at some $\mathbf{x} \in X$ is (ibid., Definition III:5.2.3) the nonempty, closed, and convex set

$$N_X(\mathbf{x}) = \left\{ \boldsymbol{\nu} \in \Re^n \;\middle|\; \boldsymbol{\nu}^T(\mathbf{y} - \mathbf{x}) \leq 0, \quad \forall \mathbf{y} \in X \right\}.$$

An optimality condition for the program [P] is given in the next proposition (ibid., Theorem VII:1.1.1 and Condition VII:(1.1.3)).

Proposition 1. $\mathbf{x} \in X^*$ *if and only if* $\partial f(\mathbf{x}) \cap -N_X(\mathbf{x}) \neq \emptyset$.

For any $\mathbf{x} \in X^*$, the nonempty set $\partial f(\mathbf{x}) \cap -N_X(\mathbf{x})$ is convex and compact.

We consider *subgradient optimization methods* (e.g., Polyak, 1969, or Shor, 1985) for solving the program [P]. These methods are conceptually simple, in general easy to implement, and have shown to be practically useful. They have therefore been frequently applied, particularly in connection with Lagrangean duality (e.g., Fisher, 1985, and Beasley, 1993).

Given an initial iterate $\mathbf{x}^0 \in X$, the traditional subgradient optimization method for solving the program [P] is given by the iteration formula

$$\mathbf{x}^{t+1} = P_X(\mathbf{x}^t - \alpha_t \boldsymbol{\gamma}(\mathbf{x}^t)), \quad t = 0, 1, \dots, \tag{2}$$

where $\boldsymbol{\gamma}(\mathbf{x}^t) \in \partial f(\mathbf{x}^t)$, $\alpha_t > 0$ is the step length, and $P_S(\mathbf{v}) = \arg\min_{\mathbf{u} \in S} \|\mathbf{u} - \mathbf{v}\|$ denotes the Euclidean projection of a vector $\mathbf{v} \in \Re^n$ onto a nonempty, closed, and convex set $S \subseteq \Re^n$.

Even though subgradient optimization methods are proven convergent, the objective value can not be guaranteed to descend in each iteration. Further, in general $\{P_{-N_X(\mathbf{x}^t)}(\boldsymbol{\gamma}(\mathbf{x}^t)) - \boldsymbol{\gamma}(\mathbf{x}^t)\} \not\to \mathbf{0}$, and even if $\mathbf{x}^t \in X^*$ it is quite likely that $\boldsymbol{\gamma}(\mathbf{x}^t) \notin -N_X(\mathbf{x}^t)$. Hence, the progress of subgradient optimization methods can *not* be monitored in terms of the approximate fulfilment of optimality conditions (e.g., that of Proposition 1).

In contrast, termination criteria in descent methods (for smooth or nonsmooth minimization) are often based on the approximate fulfilment of optimality conditions. Further, such methods enable the calculation of supplementary and useful information, for example the convergent lower bound on the optimal value in the Frank–Wolfe method; these bounds are calculated using linearizations of the objective.

In a straightforward attempt to augment the method (2) with a similar lower bounding technique one would compute the values

$$h_t^* = f(\mathbf{x}^t) + \min_{\mathbf{x} \in X} \boldsymbol{\gamma}(\mathbf{x}^t)^T \left(\mathbf{x} - \mathbf{x}^t\right), \quad t = 0, 1, \dots. \tag{3}$$

Clearly, for all t, $h_t^* \in (-\infty, f^*]$. However, in general, $\{h_t^*\} \not\to f^*$, and even in the (unlikely) case when $\mathbf{x}^t \in X^*$, $h_t^* < f^*$ generally holds; hence, the sequence $\{h_t^*\}$ does *not* possess the desired convergence properties.

A frequently used technique for inducing convergence properties that an original sequence is lacking is to construct an auxiliary, *ergodic* sequence from the original one, that is, a sequence with elements that are weighted averages of those of the original sequence. Typical applications of such averaging processes are the smoothing of oscillations in a nonconvergent sequence and the filtering out of stochastic noise (see, e.g., Kornai and Lipták, 1965; Dunn, 1973; and Powell and Sheffi, 1982).

We will show that an ergodic sequence of subgradients can be used to calculate a convergent sequence of lower bounds on f^*. This result is derived for the conditional subgradient optimization method (Larsson et al., 1996), which is a generalization of the method (2).

The *conditional* subdifferential of f with respect to X at $\mathbf{x} \in X$ (Dem'janov and Šomesova, 1978) is the nonempty, closed, and convex set

$$\partial^X f(\mathbf{x}) = \{ \boldsymbol{\gamma} \in \Re^n \mid f(\mathbf{y}) \geq f(\mathbf{x}) + \boldsymbol{\gamma}^T(\mathbf{y} - \mathbf{x}), \forall \mathbf{y} \in X \} = \partial f(\mathbf{x}) + N_X(\mathbf{x}).$$

Clearly $\partial^X f(\mathbf{x}) \supseteq \partial f(\mathbf{x})$, $\mathbf{x} \in X$, and $\mathbf{x} \in X^*$ if and only if $\mathbf{0} \in \partial^X f(\mathbf{x})$ (ibid.).

Given an initial iterate $\mathbf{x}^0 \in X$, the *conditional subgradient optimization method* (Larsson et al., 1996) is defined by the iteration formula

$$\mathbf{x}^{t+1} = P_X(\mathbf{x}^t - \alpha_t \boldsymbol{\gamma}^X(\mathbf{x}^t)), \quad t = 0, 1, \ldots, \tag{4}$$

where $\alpha_t > 0$ and $\boldsymbol{\gamma}^X(\mathbf{x}^t) = \boldsymbol{\gamma}(\mathbf{x}^t) + \boldsymbol{\nu}(\mathbf{x}^t)$, with $\boldsymbol{\gamma}(\mathbf{x}^t) \in \partial f(\mathbf{x}^t)$ and $\boldsymbol{\nu}(\mathbf{x}^t) \in N_X(\mathbf{x}^t)$. Clearly, if $\{\boldsymbol{\nu}(\mathbf{x}^t)\} = \{\mathbf{0}\}$, then the method (4) reduces to the method (2). Choosing $\boldsymbol{\nu}(\mathbf{x}^t) = P_{N_X(\mathbf{x}^t)}(-\boldsymbol{\gamma}(\mathbf{x}^t))$ results (Larsson et al., 1996, Lemma 3.2) in the special case of the method (4) that is called the *subgradient projection method*. We consider step lengths satisfying the (divergent series) conditions

$$\alpha_t > 0,\ \forall t, \quad \{\alpha_t\} \to 0, \quad \textstyle\sum_{t=0}^{\infty} \alpha_t = \infty, \quad \text{and} \quad \sum_{t=0}^{\infty} \alpha_t^2 < \infty. \tag{5}$$

The below convergence result for the sequence $\{\mathbf{x}^t\}$ follows from Theorem 2.7 in Larsson et al. (1996).

Proposition 2. *Let the method (4)–(5) be applied to the program* [P]. *If the sequence $\{\boldsymbol{\nu}(\mathbf{x}^t)\}$ is bounded, then $\{\mathbf{x}^t\} \to \mathbf{x}^\infty \in X^*$.*

Boundedness of the sequence $\{\boldsymbol{\nu}(\mathbf{x}^t)\}$ may be ensured by appropriate choices of $\boldsymbol{\gamma}^X(\mathbf{x}^t)$ in the method (4). Clearly, the sequence $\{\boldsymbol{\gamma}(\mathbf{x}^t)\}$ is bounded since f is continuous and $\{\mathbf{x}^t\} \subseteq X$ is bounded, and if subgradient projection is used, then $\|\boldsymbol{\nu}(\mathbf{x}^t)\| \leq \|\boldsymbol{\gamma}(\mathbf{x}^t)\|$, so that also the sequence $\{\boldsymbol{\nu}(\mathbf{x}^t)\}$ is bounded.

Definition 3. *(Ergodic sequence of subgradients.)* Let the method (4) be applied to the program [P]. The ergodic sequence $\{\mathbf{g}^t\}$ of subgradients is defined by

$$\mathbf{g}^t = \left(\textstyle\sum_{s=0}^{t-1} \alpha_s\right)^{-1} \sum_{s=0}^{t-1} \alpha_s \boldsymbol{\gamma}(\mathbf{x}^s), \quad t = 1, 2, \ldots.$$

Clearly, each vector $\mathbf{g}^t$ is a convex combination of the subgradients $\{\boldsymbol{\gamma}(\mathbf{x}^s)\}_{s=0}^{t-1}$; hence, the sequence $\{\mathbf{g}^t\}$ is bounded. Note that it is calculated beside the sequence $\{\mathbf{x}^t\}$ (and does not affect the course of the method (4)), and that it can be computed recursively.

The below convergence result for the sequence $\{\mathbf{g}^t\}$ follows from Theorem 3.4 in Larsson et al. (1995).

Lemma 4. *Let the method (4)–(5) be applied to the program* [P] *and let the sequence $\{\mathbf{g}^t\}$ be given by Definition 3. Suppose that the sequence $\{\boldsymbol{\nu}(\mathbf{x}^s)\}$ is bounded. Then, $\{P_{\partial f(\mathbf{x}^\infty) \cap -N_X(\mathbf{x}^\infty)}(\mathbf{g}^t) - \mathbf{g}^t\} \to \mathbf{0}$.*

3 A Lower Bounding Procedure

The ergodic sequence of subgradients will now be used to define a convergent sequence of lower bounds on the optimal value of the program [P]. With this purpose in mind we define the *affine* functions $f_t : \Re^n \mapsto \Re$, $t = 1, 2, \ldots$, as

$$f_t(\mathbf{x}) = \left(\sum_{s=0}^{t-1} \alpha_s\right)^{-1} \sum_{s=0}^{t-1} \alpha_s \left(f(\mathbf{x}^s) + \boldsymbol{\gamma}(\mathbf{x}^s)^T(\mathbf{x} - \mathbf{x}^s)\right), \quad \mathbf{x} \in \Re^n. \qquad (6)$$

By the definition (1), and since, for all s, $\boldsymbol{\gamma}(\mathbf{x}^s) \in \partial f(\mathbf{x}^s)$, $f_t(\mathbf{x}) \le f(\mathbf{x})$ holds for all $\mathbf{x} \in \Re^n$ and all t, that is, each function f_t is an affine *underestimate* of the objective f, and therefore, for all t, $\min_{\mathbf{x}\in X} f_t(\mathbf{x}) \le f^*$.

Definition 5. *(Sequence of lower bounds.)* Let the method (4) be applied to the program [P] and let, for all t, $f_t(\mathbf{x})$ be given by the definition (6). The sequence $\{f_t^*\}$ of lower bounds on f^* is defined by

$$f_t^* = \min_{\mathbf{x}\in X} f_t(\mathbf{x}), \quad t = 1, 2, \ldots.$$

The next lemma establishes that, in the limit, the sequence $\{f_t\}$ of affine functions provides *supporting hyperplanes* to $\texttt{epi}f$ at $(\mathbf{x}^\infty, f(\mathbf{x}^\infty))$; it follows from Lemma 2.4 in Larsson et al. (1995) and is used to show that the sequence of lower bounds converges to the optimal value, which is our main result.

Lemma 6. *Let the method (4)–(5) be applied to the program* [P] *and let the sequence* $\{f_t(\mathbf{x}^\infty)\}$ *be given by the definition (6). If the sequence* $\{\boldsymbol{\nu}(\mathbf{x}^s)\}$ *is bounded, then* $\{f_t(\mathbf{x}^\infty)\} \to f^*$.

Theorem 7. *Let the method (4)–(5) be applied to the program* [P] *and let the sequence* $\{f_t^*\}$ *be given by Definition 5. If the sequence* $\{\boldsymbol{\nu}(\mathbf{x}^s)\}$ *is bounded, then* $f_t^* \in (-\infty, f^*]$ *for all* t, *and* $\{f_t^*\} \to f^*$.

Proof. From Definition 5 and the definitions (6) and (1) follow that $f_t^* \le f^*$ for all t. Further, since X is bounded, f_t^* is finite for all t.

By Definitions 5 and 3, $f_t^* = f_t(\mathbf{x}^\infty) + \min_{\mathbf{x}\in X} \left(\mathbf{g}^t\right)^T (\mathbf{x} - \mathbf{x}^\infty)$. Let, for all t, $\mathbf{y}^t \in X$ solve this minimization problem, so that $\left(\mathbf{g}^t\right)^T (\mathbf{y}^t - \mathbf{x}^\infty) = f_t^* - f_t(\mathbf{x}^\infty)$, and let $\mathcal{T}$ be a subsequence such that $\lim_{t\in\mathcal{T}, t\to\infty} (\mathbf{g}^t)^T(\mathbf{y}^t - \mathbf{x}^\infty) = \liminf_{t\to\infty} (\mathbf{g}^t)^T(\mathbf{y}^t - \mathbf{x}^\infty)$. Since $\{\mathbf{g}^t\}$ and $\{\mathbf{y}^t\}$ are bounded, there exists a subsequence $\tilde{\mathcal{T}} \subseteq \mathcal{T}$ such that $\{\mathbf{y}^t\}_{t\in\tilde{\mathcal{T}}} \to \tilde{\mathbf{y}}$ and $\{\mathbf{g}^t\}_{t\in\tilde{\mathcal{T}}} \to \tilde{\mathbf{g}}$. Now,

$$0 \le \tilde{\mathbf{g}}^T \left(\tilde{\mathbf{y}} - \mathbf{x}^\infty\right) = \liminf_{t\to\infty} \left(\mathbf{g}^t\right)^T \left(\mathbf{y}^t - \mathbf{x}^\infty\right) \le \limsup_{t\to\infty} \left(f_t^* - f_t(\mathbf{x}^\infty)\right) \le 0,$$

since, by Lemma 4, $\tilde{\mathbf{g}} \in -N_X(\mathbf{x}^\infty)$, and $\tilde{\mathbf{y}} \in X$, and since, for all t, $f_t^* \le f_t(\mathbf{x}^\infty)$. Hence, $\{f_t^* - f_t(\mathbf{x}^\infty)\} \to 0$, and by Lemma 6, $\{f_t(\mathbf{x}^\infty)\} \to f^*$, thus implying that $\{f_t^*\} \to f^*$. □

Defining $\texttt{UBD}_t = \min_{0\le s\le t} f(\mathbf{x}^s)$ and $\texttt{LBD}_t = \max_{1\le s\le t} f_s^*$, for all t, Proposition 2 and Theorem 7 yield that, for any $\varepsilon > 0$ there is an iteration $t(\varepsilon)$ such that $\texttt{UBD}_{t(\varepsilon)} - \texttt{LBD}_{t(\varepsilon)} \le \varepsilon$ (implying that $\texttt{UBD}_{t(\varepsilon)} \le f^* + \varepsilon$); hence, this condition may be used as a termination criterion for the method (4)–(5).

4 Application to Multicommodity Network Flows

To illustrate the practical usefulness of the lower bounding procedure we here present numerical results from an application to a subgradient scheme for a nonsmooth formulation of the minimum cost multicommodity network flow problem.

Let $\mathcal{N}$ and $\mathcal{A}$ be the sets of nodes and arcs, respectively, of a directed network. Further, let $\mathcal{K}$ be a set of flow commodities, and let $\mathbf{d}_k$ be the demand vector and $\mathbf{x}_k$ the vector of arc flows for commodity $k \in \mathcal{K}$. Associated with the commodity flows are vectors of unit transportation costs $\mathbf{c}_k$, $k \in \mathcal{K}$, and the total arc flows are upper bounded by $\mathbf{b}$. Letting $\mathbf{A}$ denote the node-arc incidence matrix for the network, the problem of minimizing the total cost is given by

$$f^* = \min \left\{ \sum_{k\in\mathcal{K}} \mathbf{c}_k^T \mathbf{x}_k \,\middle|\, \mathbf{A}\mathbf{x}_k = \mathbf{d}_k,\ k \in \mathcal{K}; \sum_{k\in\mathcal{K}} \mathbf{x}_k \le \mathbf{b};\ \mathbf{x}_k \ge \mathbf{0},\ k \in \mathcal{K} \right\}.$$

Kennington and Shalaby (1977) employ a subgradient optimization scheme which is based on allocations of the capacities of the total arc flows to separate commodities. Letting $Y = \{ \mathbf{y} \in \Re_+^{|\mathcal{A}|\cdot|\mathcal{K}|} \mid \sum_{k\in\mathcal{K}} \mathbf{y}_k = \mathbf{b} \}$ be the set of feasible commodity capacity allocations, the equivalent nonsmooth program

$$f^* = \min_{\mathbf{y}\in Y} f(\mathbf{y}) = \min_{\mathbf{y}\in Y} \left\{ \sum_{k\in\mathcal{K}} \min \left\{ \mathbf{c}_k^T \mathbf{x}_k \,\middle|\, \mathbf{A}\mathbf{x}_k = \mathbf{d}_k; \quad 0 \le \mathbf{x}_k \le \mathbf{y}_k \right\} \right\}$$

is obtained. The objective $f : \Re_+^{|\mathcal{A}|\cdot|\mathcal{K}|} \mapsto \Re$ is convex and piecewise linear; it is evaluated by solving $|\mathcal{K}|$ single-commodity network flow problems, and a subgradient of f at $\mathbf{y}$ is given by the optimal values of the dual variables for the constraints $\mathbf{x}_k \le \mathbf{y}_k$, $k \in \mathcal{K}$. In Larsson et al. (1996) this subgradient scheme was improved through the use of projected subgradients; see this reference for a more detailed description of the algorithm.

The calculation of the lower bound f_t^* amounts to the minimization of a linear objective over the set Y, and separates into $|\mathcal{A}|$ continuous knapsack problems.

The test problems have a transportation network structure and are randomly generated in accordance with the principles in Kennington and Shalaby (1977). We generated 10 instances, each having 10 sources, 10 sinks, and 10 commodities. The supply/demand for each source/sink and commodity is in the range 0–300, arc capacities in the range 100–400, and arc costs in the range 0–100. All data is uniformly distributed.

In order to receive a good starting point for the generation of the sequence of lower bounds, we used step lengths according to Polyak (1969) during the first 100 iterations. For each instance and iteration, the bounds $f(\mathbf{x}^t)$ and f_t^* were computed; then we recorded the values $\mathtt{UBD}_t$ and $\mathtt{LBD}_t$, normalized by the optimal value (which was computed in advance) of the instance. For each iteration t we computed the average ($\mathtt{U}_t$ and $\mathtt{L}_t$, respectively) of these normalized values over the 10 instances. As a means for evaluating the effect of invoking the ergodic process we also recorded averages ($\mathtt{H}_t$) of normalized values of the nonconvergent bounds $\max_{0\le s\le t} h_s^*$, $t = 0, 1, \ldots$, where h_s^* is given by the definition (3).

The results presented in Table 1 confirm that, in this application, the introduction of the ergodic process is an effective means for obtaining a judgement of the quality of the iterates produced by the subgradient scheme.

Table 1. Convergence of normalized bounds

t	$\mathtt{U}_t$	$\mathtt{L}_t$	$\mathtt{H}_t$	$\frac{\mathtt{U}_t-\mathtt{L}_t}{\mathtt{L}_t}$ (%)	$\frac{\mathtt{U}_t-\mathtt{H}_t}{\mathtt{H}_t}$ (%)
20	1.0331		0.8298		24.50
60	1.0075		0.9268		8.71
100	1.0023	0.9072	0.9547	10.48	4.99
105	1.0020	0.9842	0.9567	1.81	4.74
110	1.0018	0.9899	0.9567	1.20	4.71
120	1.0017	0.9940	0.9591	0.77	4.44
160	1.0014	0.9964	0.9605	0.50	4.26
200	1.0011	0.9972	0.9616	0.39	4.11

Acknowledgement. The research was supported financially by grants (No. 94-292 and 282-93-1195) from the Swedish Research Council for Engineering Sciences (TFR).

References

Beasley, J.E.: Lagrangean relaxation, in: "Modern Heuristic Techniques for Combinatorial Problems." C.R. Reeves, ed., Blackwell, Oxford (1993) pp. 243–303

Dem'janov, V.F., and Šomesova, V.K.: Conditional subdifferentials of convex functions. Soviet Mathematics Doklady **19** (1978) 1181–1185

Dunn, J.C.: On recursive averaging processes and Hilbert space extensions of the contraction mapping principle. Journal of the Franklin Institute **295** (1973) 117–133

Fisher, M.L.: An applications oriented guide to Lagrangian relaxation. Interfaces **15** (1985) 10–21

Hiriart-Urruty, J.-B., and Lemaréchal, C.: "Convex Analysis and Minimization Algorithms I: Fundamentals." Springer, Berlin (1993)

Kennington, J., and Shalaby, M.: An effective subgradient procedure for minimal cost multicommodity flow problems. Management Science **23** (1977) 994–1004

Kornai, J., and Lipták, T.: Two level planning. Econometrica **33** (1965) 141–169

Larsson, T., Patriksson, M., and Strömberg, A.-B.: Ergodic results in subgradient optimization. Report LiTH-MAT-R-95-21 (1995), Department of Mathematics, Linköping Institute of Technology, Linköping, Sweden. Also in: Proceedings of the International School of Mathematics "G. Stampacchia" 21st Workshop on Nonlinear Optimization and Applications, held June 13–21, 1995, in Erice, Italy, Plenum, New York (to appear)

Larsson, T., Patriksson, M., and Strömberg, A.-B.: Conditional subgradient optimization—theory and applications. European Journal of Operational Research **88** (1996) 382–403

Polyak, B.T.: Minimization of unsmooth functionals. USSR Computational Mathematics and Mathematical Physics **9** (1969) 14–29

Powell, W.B., and Sheffi, Y.: The convergence of equilibrium algorithms with predetermined step sizes. Transportation Science **16** (1982) 45–55

Shor, N.Z.: "Minimization Methods for Non-Differentiable Functions." Springer, Berlin (1985)