Conjugate subgradient method revisited

E. A. Nurminski*

*Institute for Automation and Control Processes FEB RAS, nurmi@dvo.ru

Almost 40 years ago P. Wolfe in one of the pioneering papers on convex nondifferentiable optimization [1] suggested a conjugate subgradient algorithm similar to the well-established conjugate gradient (CG) method of Hesten-Stiefel [2] for solving quadratic optimization problems. However, unlike the famous predecessor, Wolfe's method remained practically unused. Here we provide certain justification for this and suggest a modification of this algorithm to overcome its problems.

The following basic problem of convex nondifferentiable optimization

$$\min_{x\in E} f(x) = f_\star = f(x^\star), \tag{1}$$

where E is a finite-dimensional euclidian space with the inner product xy and corresponding norm $\|x\| = \sqrt{xx}$ is considered. It is assumed that solution $x^\star$ of (1) exists. The main intention is to introduce conjugate gradient-like algorithm, prove it convergence and present some promising results of numerical experiments.

Typically computational algorithms for (1) rely on the use of subgradient oracles which provide at arbitrary point x the value of objective function $f(x)$ and some subgradient g from subdifferental set $\partial f(x)$. The simplest algorithms for (1) are subgradient methods of the kind

$$x^{k+1} = x^k - \lambda_k g^k,\ g^k \in \partial f(x^k), \quad k = 0, 1, \dots \tag{2}$$

which were under intensive study since 1960's. It was shown that (2) converges under very mild conditions for step-size λ_k satisfying "divergence series" condition $\sum_k \lambda_k = \infty,\ \lambda_k \to +0$. However numerical experiments and theoretical analysis demonstrated that this step-size rule results as a rule in slow convergence and further development went along the lines of quasi-newton ideas [3], space dilatation [4], proximal-type, bundle and level methods [7, 6], etc. On the other hand subgradient algorithms still demonstrated quite satisfactory performance under special conditions, f.i. when $f_\star$ in (1) is known or well-estimated so special step-size control rules can be engaged.

1 CSGrad-algorithm

Similar to the subgradient algorithm (2) a conjugate subgradient (CSG) algorithm generate the sequence of points

$$x^{k+1} = x^k - \lambda_k z^k,\ g^k \in \partial f(x^k), \quad k = 0, 1, \dots \tag{3}$$

alongside with the corresponding sequence of $g^k \in \partial f(x^k),\ k = 0, 1, \dots$. The set of subgradients $\{g^s, m \le s \le n\}$ (bundle) will be denoted as $G(m, s)$ and the difference between (2) and (3) is that z^k is determined as a solution of the problem

$$\min_{z\in G(m_k,k)} \|z\|^2 = \|z^k\|^2,$$

where $m_k \le k$ is a certain restart moment, preceeding k. The precise rules for defining m_k are given below.

The corresponding algorithm can be generally stated in the following form.

Initialization Set $g^0 \in \partial f(x^0)$, $G(0,0) = co\{g^0\}$, $m_0 = 0$. Set iteration counters to zero $k = t = 0$.

The main t-th iteration of the algorithm consists of the following 2 steps:

Step 1. If $t - k \ge N$ set $k = t$ (restart). Solve the least distance problem

$$\min_{z\in G(k,t)} \|z\|^2 = \|z^{t,k}\|^2.$$

If $\|z^{t,k}\| \le \delta_k$, increment $k : k \to k+1$ set $m_k = t$, clear $G(k,t) = \{g^t\},\ g^t \in \partial f(x^t)$ and repeat **Step 1.** Otherwise continue with **Step 2.**

Step 2. Solve line-search problem

$$\min_\lambda f(x^t - \lambda z^{t,k}) = f(x^t - \lambda_t z^{t,k}) = f(x^{t+1})$$

and pick $g^{t+1} \in \partial f(x^{t+1})$ such that $g^{t+1} z^{t,k} = 0$. Complement the set $G(k,t)$ with g^{t+1}: $G(k, t+1) = co\{G(k,t), g^{t+1}\}$, increment iteration counter $t \to t+1$ and continue with **Step 1.**

It should be noted that this algorithm reminds dual form of bundle methods [5], but we disregard linearization errors to simplify direction-finding problem. From our experience it does not hinder too much the performance of the algorithm. We also make use of strong convexity of objective function to establish some useful properties of minimizing sequence. As to the convergence of the algorithm, the following result can be obtained.

Theorem 1 *Let f is a strongly convex function with bounded level sets $L_C = \{x : f(x) \leq C\}$. Then the sequence of $\{x^k\}$ with x^k generated by (3) converges to a unique solution of (1).*

Strong convexity seems to be unnecessary, at least for nondifferentiable problems with unique solution, but so far is required for technical reasons.

The major difference with [1] is that in the presented algorithm the size of the bundle is restricted by some preselected constant N. It should be noted that Wolfe's algorithm accumulates the bundle of subgradients until certain conditions are satisfied. Resulting estimates for the number of accumulated subgradients are quite high and it causes memory problems. In our experience it also hinders convergence of the algorithm due to presence in the bundle of obsolete subgradients.

1.1 Connection with conjugate gradient algorithm

It is interesting to remark, that by augmenting this algorithmic idea with line-search in directions $z^{m,0}, z^{m,1}, \ldots, z^{m,s}$ we obtain for strongly convex quadratic functions exactly classic conjugate gradient algorithm of [2]. Indeed, consider the problem P_n:

$$\min \frac{1}{2}\sum_{i=1}^{n}\lambda_i^2\|g^i\|^2, \quad \sum_{i=1}^{n}\lambda_i = 1, \quad \lambda_i \geq 0, \quad i = 1, 2, \ldots, n.$$

If $g^i \geq 0,\ i = 1, 2, \ldots, n$ are mutually orthogonal, then solutions of P_n and P_{n+1} are mutually conjugate.

To demonstrate it consider the solution of P_{n+1} which is defined by the system of equations

$$\lambda_i\|g^i\|^2 + \theta = 0,\ i = 1, 2, \ldots, n+1,$$
$$\theta\sum_{i=1}^{n+1}\|g^i\|^{-2} = -1.$$

It follows that $\lambda_j = \|g^j\|^{-2}(\sum_{i=1}^{n+1}\|g^i\|^{-2})^{-1},\ j = 1, 2, \ldots, n+1$ and non-negativity constraints are fulfilled automatically.

Denote $\sigma_{n+1} = \sum_{i=1}^{n+1}\|g^i\|^{-2} = \sigma_n + \|g^{n+1}\|^{-2}$. Then

$$z^{n+1} = \sum_{i=1}^{n+1}\lambda_i g^i = \sum_{i=1}^{n}\lambda_i g^i + \lambda_{n+1}g^{n+1} =$$
$$\sum_{i=1}^{n}\|g^i\|^{-2}(\sigma_n + \|g^{n+1}\|^{-2})^{-1}g^i +$$
$$\|g^{n+1}\|^{-2}(\sigma_n + \|g^{n+1}\|^{-2})^{-1}g^{n+1} =$$
$$(\sigma_n + \|g^{n+1}\|^{-2})^{-1}(g^{n+1} + \|g^{n+1}\|^{-2}\sum_{i=1}^{n}\|g^i\|^{-2}g^i) =$$
$$\theta_{n+1}(g^{n+1} + \|g^n\|^{-2}\|g^{n+1}\|^2(\|g^n\|^2\sum_{i=1}^{n}\|g^i\|^{-2}g^i)) =$$
$$\theta_n(g^{n+1} + \|g^n\|^{-2}\|g^{n+1}\|^2 z^n) = \theta_n(g^{n+1} + \mu_{n+1}z^n),$$

which differs from conjugate gradient update schema [2] only by scaling factor θ_n.

2 Numerical experiments

To demonstrate computational efficiency of CSGrad results of numerical experiments with two well-known test problems from [11]:

Convex piece-wise quadratic function `maxquad`.

$$f(x) = \max_{1\leq k\leq 5}\phi_k(x),$$

where $\phi_k(x) = xA_kx - b^kx$, $A^{(k)}, k = 1, 2, \ldots, 5$ – symmetric positive definite matrices 10×10 such that for $i, j = 1, 2, \ldots, 10$ hold

$$A_{ij}^{(k)} = \begin{cases} \exp(\min(i,j)/\max(i,j))cos(ij)\sin(k), & i \neq j, \\ i|\sin(k)|/10 + \sum\limits_{l=1,2,\ldots,10, l\neq i}|A_{il}^{(k)}|, & i = j, \end{cases}$$
$$b_i^k = \exp(i/k)\sin(ik), \quad i = 1, 2, \ldots, 10, \quad k = 1, 2, \ldots, 5.$$

Convex piece-wise linear function `tr48`. The objective function of this problem is

$$f(x) = -\left(\sum_{i=1}^{n}s_ix_i + \sum_{j=1}^{n}d_j\min_{i=1,2,\ldots,n}(a_{ij} - x_i)\right)$$

with $n = 48$. The data for this problem and `octave` code for computing function value and subgradient can be found on [13].

References

[1] P. Wolfe *A method of conjugate subgradients for minimizing nondifferentiable functions,* Mathematical Programming Studies, v. 3, Nondifferentiable Optimization, 1975, 145-173.

[2] Hesten M.R., Stiefel E. *Methods of conjugate gradients for solving linear systems* Journal of Research of the National Bureau of Standards 49 (6), 1952, Research Paper 2379 409–436.

[3] C. Lemarechal *An extension of Davidon methods to non-differentiable problems*. Mathematical Programming Study. 1975. Vol. 3. P. 95-109.

[4] N. Z. Shor, K. C. Kiwiel and A. Ruszcayǹski *Minimization methods for non-differentiable functions*. Springer-Verlag, 1985.

[5] J.-B. Hiriart-Urruty and C. Lemarechal *Convex analysis and minimization algorithms II. Advanced theory and bundle methods*. Springer-Verlag, 1993.

[6] U. Brannlund, K. C. Kiwiel and P. O. Lindberg *A descent proximal level bundle method for convex nondifferentiable optimization*. Operations Research Letters. 1995. Vol. 17(3). P. 121-126.

[7] C. Lemarechal, A. Nemirovskii and Ju. Nesterov *New variants of bundle methods*. Mathematical Programming. 1995. Vol. 3. P. 111-147.

[8] E. A. Nurminski *Numerical methods of convex optimization*. Nauka, 1991.

[9] E. A. Nurminski *Envelope step-size control for iterative algorithms based on Fejer processes with attractants*. Optimization Methods and Software. 2010. Vol. 25(1). P .97-108.

[10] E. A. Nurminski *Fejer algorithms*. Computational Mathematics, 2011. Vol. 41(5).

[11] C. Lemarechal and R. Mifflin *Nonsmooth optimization*. Oxford, Pergamon Press, 1978.

[12] C. Lemarechal *Numerical experiments in nonsmooth optimization* Progress in nondifferentiable optimization. International Institute for Applied System Analysis, 1982.

[13] *OptimiZone site*. `http:elis.dvo.ru`.