A Survey on subgradient methods

Qianqian Ma
maqq@bu.edu

Abstract

This survey paper considers subgradient methods for nondifferentiable optimization problem with convex objective function. Though compared with other optimization methods, the convergence of subgradient method is relatively slow. There are also some superiorities for subgradient method. The main advantage of subgradient method is simplicity. It can use any subgradient to solve the optimization problem, and no line search for step rule is involved. As a result, the subgradient method is easily to be implemented. Besides, the convergence proof is also relatively simple. Moreover, another important advantage of subgradient method is its robustness. This can be clearly seen from the study of stochastic subgradient method.

1. Introduction

Subgradient methods are the principal methods used in convex nondierentiable optimization problems.This type of optimization arises in many applications, as well as in the context of duality, and various general solution strategies such as penalty function methods, regularization methods, and decomposition methods.

The basic subgradient method is similar to the ordinary gradient method for differentiable functions, but with several notable exceptions. First, the subgradient method applies directly to nondifferentiable functions.Second, the step lengths are not chosen via a line search, as in the ordinary gradient method. Actually,in the most common cases, the step lengths are fixed ahead of time. Third, unlike the ordinary gradient method, the subgradient method is not a descent method, the function value can increase.

The basic subgradient method can be readily extended to solve constrained problem. Besides, the subgradient methods can also be extended to handle problems with errors, which may come from measurements, uncertainty, or the computation is intractable.This is the stochastic subgradient method.

The subgradient method is relatively slower than Newton's method, but is much simpler and can be applied to a far wider variety of problems. By combining the subgradient method with primal or dual decomposition techniques, it is possible to develop a simple distributed algorithm for a certain problem.

Subgradient methods were first introduced in the Soviet Union in the middle sixties by N. Z. Shor. Since then, they have been extensively studied, and in general two major classes of subgradient methods have been developed: descent-based methods and nondescent methods.

The descent-based subgradient methods are based on the principal idea of the function descent, which lies in the framework of gradient-type minimization. Nondescent subgradient methods are based on the idea of the distance decrease (distance from the set of minima), and their implementation is simpler than that of descent-based methods. For nondescent subgradient methods, the early work of Ermoliev [Erm66] and Polyak [Pol67] was particularly influential. Due to their simple implementation, the nondescent subgradient methods have drawn a lot of attention, and the literature on these methods is very rich. An extensive treatment of these subgradient methods can be found in the textbooks by Dem'yanov and Vasil'ev [2], Shor [6], Minmox[4], Polyak [5] , Hiriart-Urruty and Lemarechal [3] Shor [7], and Bertsekas [1]. Besides, there are also some recent research based on subgradient methods have been developed, such as incremental subgradient methods [9] and primal-dual dubgradient methods [8].

2. Basic subgradient method

First we will consider the unconstrained case. For this part, we aim to solve the following problem:

$$minimize \quad f : \mathbb{R}^n \to \mathbb{R}, \tag{1}$$

which is convex and has domain $\mathbb{R}^n$.

2.1. Update rules

The definition of subgradient

A subgradient of f at x is any vector g that satisfies the inequality

$$f(y) \geq f(x) + g^T(y - x), \quad \forall y. \tag{2}$$

Update rules

The subgradient method uses the simple iteration

$$x^{(k+1)} = x^k - \alpha_k g^{(k)}, \tag{3}$$

where $\partial f(x^{(k)})$ denotes the subdifferential of f at $x^{(k)}$. Thus, at each iteration of the subgradient method, we take a step in the direction of a negative subgradient.When f is differentiable, the only possible choice for $g^{(k)}$ is $\nabla f(x)$,and the subgradient method then reduces to the gradient method. The condition that $g^{(k)}$be a subgradient of f at $x^{(k)}$:

$$g^{(k)} \in \partial f(x^{(k)}), \tag{4}$$

where $\partial f(x^{(k)})$ denotes the subdifferential of f at $x^{(k)}$.

Subgradient method is not a descent method, it can happen that $-g^{(k)}$ is not a descent direction for f at $x^(k)$, in such case, we have $f(x^{(k+1)} > f(x^{(k)}))$.In other words, an iteration of the subgradient method can increase the objective function.

To keep track of the best point found so far, we set

$$f_{best}^{(k)} = min\{f_{best}^{(k-1)}, f(x^{(k)})\}, \tag{5}$$

$$i_{best}^{(k)} = k \quad if \quad f(x^{(k)}) = f_{best}^{(k)}. \tag{6}$$

Then we have

$$f_{best}^{(k)} = min\{f(x^{(1)}), ..., f(x^{(k)})\}. \tag{7}$$

i.e., the best objective value found in k iterations. Since $f_{best}^{(k)}$ is decreasing, it has a limit (which can be $-\infty$).

2.2. Step size rules

In the subgradient method the step size selection is very different from the standard gradient method. Many different types of step size rules are used. Here I want to introduce five basic step size rules.

- Constant step size. $\alpha_k = \alpha$ is a positive constant, independent of k
- Constant step length. $\alpha_k = h/\|g^{(k)}\|_2$, where $h > 0$. This means that $\|x^{k+1} - x^k\|_2 = h$.
- Square summable but not summable. The step satisfies

$$\alpha_k \geq 0, \quad \sum_{k=1}^{\infty} \alpha_k^2 < \infty, \quad \sum_{k=1}^{\infty} \alpha_k = \infty. \tag{8}$$

  One typical example is $\alpha k = a/(b + k)$, where $a > 0$ and $b > 0$.

- Nonsummable diminishing.The step size satisfy

$$\alpha_k \geq 0, \quad \lim_{k\to\infty} \alpha_k = 0, \quad \sum_{k=1}^{\infty} \alpha_k = \infty. \tag{9}$$

  Step sizes that satisfy this condition are called diminishing step size rules. A typical example is $\alpha_k = a/\sqrt{k}$, where $a > 0$.

- Nonsummable diminishing step lengths. The step sizes are chosen as $\alpha_k = \gamma_k/\|g^{(k)}\|_2$, where

$$\gamma_k \geq 0, \quad \lim_{k\to\infty} \gamma_k = 0, \quad \sum_{k=1}^{\infty} \gamma_k = \infty. \tag{10}$$

The most important feature of these choices is that these step sizes are determined before the algorithm is run; they do not depend on any data computed during the algorithm.

2.3. Convergence Analysis

Next, a proof of some typical convergence results for the subgradient method will be provided.

Assumptions

We assume that there is a minimizer of f, say x^*. We also make one other assumption on f: assume that the norm of the subgradients is bounded, i.e., there is a G such that $\|g^{(k)}\|_2 \leq G$for all k. This will be the case if, for example, f satisfies the Lipschitz condition

$$|f(u) - f(v)| \leq G\|u - v\|_2, \tag{11}$$

for all u, v, because then $\|g\|_2 \leq G$ for any $g \in \partial(x)$, and any x.

Proof

For the standard gradient descent method, the convergence proof is based on the function value decreasing at each step. In the subgradient method, the key quantity is not the function value (which often increases); it is the Euclidean distance to the optimal set.

As x^* is a point that minimizes f, i.e., it is an arbitrary optimal point. We have

$$\begin{aligned} &\|x^{(k+1)} - x^*\|_2^2 \\ &= \|x^{(k)} - \alpha_k g^{(k)} - x^*\|_2^2 \\ &= \|x^{(k)} - x^*\|_2^2 - 2\alpha_k g^{(k)T}(x^{(k)} - x^*) + \alpha_k^2\|g^{(k)}\|_2^2 \\ &\leq \|x^{(k)} - x^*\|_2^2 - 2\alpha_k(f(x^{(k)}) - f^*) + \alpha_k^2\|g^{(k)}\|_2^2, \end{aligned} \tag{12}$$

where $f^* = f(x^*)$. The last inequality follows from the definition of subgradient, which gives

$$f(x^*) \geq f(x^{(k)}) + g^{(k)T}(x^* - x^{(k)}).$$

Applying the inequality above recursively, we have

$$\|x^{(k+1)} - x^*\|_2^2 \leq \|x^{(1)} - x^*\|_2^2 - 2\sum_{i=1}^{k} \alpha_i (f(x^{(i)}) - f^*) + \sum_{i=1}^{k} \alpha_i^2 \|g^{(i)}\|_2^2, \tag{13}$$

using $\|x^{(k+1)} - x^*\|_2^2 \geq 0$ and $\|x^{(1)} - x^*\|_2^2 \leq R$, we have

$$2\sum_{i=1}^{k} \alpha_i (f(x^{(i)}) \leq R^2 + \sum_{i=1}^{k} \alpha_i^2 \|g^{(i)}\|_2^2.$$

Combining this with

$$\sum_{i=1}^{k} \alpha_i (f(x^{(i)} - f^*) \geq (\sum_{i=1}^{k} \alpha_i) \min_{i=1,\ldots,k} f(x^{(i)} - f^*) = (\sum_{i=1}^{k} \alpha_i)(f_{best}^{(k)} - f^*). \tag{14}$$

we have the inequality

$$f_{best}^{(k)} - f^* = \min_{i=1,\ldots,k} f(x^{(i)} - f^*) \leq \frac{R^2 + \sum_{i=1}^{k} \alpha_i^2 \|g^{(i)}\|_2^2}{2\sum_{i=1}^{k} \alpha_i}. \tag{15}$$

Finally, using the assumption $\|g^{(k)}\|_2 \leq G$, we obtain the basic inequality

$$f_{best}^{(k)} - f^* \leq \frac{R^2 + G^2 \sum_{i=1}^{k} \alpha_i^2}{2\sum_{i=1}^{k} \alpha_i}. \tag{16}$$

From this inequality we can read off various convergence results.

Convergence Results

- Constant step size. When $\alpha_k = \alpha$, we have

$$f_{best}^{(k)} - f^* \leq \frac{R^2 + G^2\alpha^2 k}{2\alpha k} \tag{17}$$

  the righthand side converges to $G^2\alpha/2$ as $k \to \infty$. Thus, for the subgradient method with fixed step size α, $f_{best}^{(k)}$ converges to within $G^2\alpha/2$ of optimal. We also find that $f(x^{(k)}) - f^* \leq G^2\alpha$ within at most $R^2/(G^2\alpha^2)$ steps.

- Constant step length. With $\alpha_k = \gamma/\|g^{(k)}\|_2$, the inequality (2) becomes

$$f_{best}^{(k)} - f^* \leq \frac{R^2 + \gamma^2 k}{2\sum_{i=1}^{k} \alpha_i} \leq \frac{R^2 + \gamma^2 k}{2\gamma k/G} \tag{18}$$

  using $\alpha_i \geq \gamma/G$. The righthand converges to $G\gamma/2$ of optimal.

- Square summable but not summable. Now suppose

$$\|\alpha\|_2^2 = \sum_{k=1}^{\infty} \alpha_k^2 < \infty, \qquad \sum_{k=1}^{\infty} \alpha_k = \infty \tag{19}$$

  Then we have

$$f_{best}^{(k)} - f^* \leq \frac{R^2 + G^2\|\alpha\|_2^2}{2\sum_{i=1}^{k} \alpha_i} \tag{20}$$

  which converges to zero as $k \to \infty$, since the numerator converges to $R^2 + G^2\|\alpha\|_2^2$, and the denominator grows without bound. Thus, the subgradient method converges (in the sense $f_{best}^{(k)} - f^*$).

- Diminishing step size rule. If the sequence α_k converges to zero and is nonsummable, then the righthand side of the inequality 8 converges to zero, which implies the subgradient method converges. To show this, let $\epsilon > 0$. Then there exists an integer N_1 such that $\alpha_i \leq \epsilon/G^2$ for all $i > N_1$. There also exists an integer N_2 such that

$$\sum_{i=1}^{N_2} \alpha_i \geq \frac{1}{\epsilon}\left(R^2 + G^2 \sum_{i=1}^{N_2} \alpha_i^2\right), \tag{21}$$

In summary, for constant step size and constant step length, the subgradient algorithm is guaranteed to converge to within some range of the optimal value, i.e., we have

$$\lim_{k\to\infty} f_{best}^{(k)} - f^* < \epsilon, \tag{22}$$

where f^* denotes the optimal value of the problem. The number ϵ is a function of the step size parameter h, and decrease with it.

For the diminishing step size and step length rules (therefore also the square summable but not summable step size rule), the algorithm is guaranteed to converge to the optimal value, i.e., we have

$$\lim_{k\to\infty} f_{best}^{(k)} = f^*. \tag{23}$$

Its remarkable that such a simple algorithm can be used to minimize any convex function for which you can compute a subgradient at each point.

2.4. Algorithm

The algorithm of subgradient method for unconstrained problem can be summarized as:

Algorithm 1

1. Set $k := 1$ and choose an infinite sequence of positive step size value α_k.
2. Compute a subgradient $g^{(k)} \in \partial f(x^k)$.
3. Update $x^{(k+1)} = x^k - \alpha_k g^{(k)}$.
4. If algorithm has not converged, then set $k := k + 1$ and go to step 2.

3. Projected subgradient method

One important extension of the subgradient method is the projected subgradient method, which solves the following constrained convex optimization problem

$$\begin{aligned} &minimize \quad f(x) \\ &subject \quad to \quad x \in C, \end{aligned} \tag{24}$$

where C is a convex set.

3.1. Update rules

The update rule of the projected subgradient method is given by

$$x^{(k+1)} = P(x^{(k)} - \alpha_k g^{(k)}), \tag{25}$$

where P is (Euclidean) projection on C, and $g^{(k)}$ is any subgradient of f at $x^{(k)}$.

The step size rules of the projected subgradient method can also adopt the step rules given in the basic subgradient method part.

3.2. Convergence Analysis

The convergence proofs for the subgradient method can be readily extended to handle the projected subgradient method.

Let $z^{(k+1)} = x^{(k)} - \alpha_k g^{(k)}$, i.e., a standard subgradient update, before the projection back onto C. Then we have

$$\begin{aligned} &\|z^{(k+1)} - x^*\|_2^2 \\ &= \|x^{(k)} - \alpha_k g^{(k)} - x^*\|_2^2 \\ &= \|x^{(k)} - x^*\|_2^2 - 2\alpha_k g^{(k)T}(x^{(k)} - x^*) + \alpha_k^2 \|g^{(k)}\|_2^2 \\ &\leq \|x^{(k)} - x^*\|_2^2 - 2\alpha_k (f(x^{(k)}) - f^*) + \alpha_k^2 \|g^{(k)}\|_2^2. \end{aligned} \tag{26}$$

Besides, when we project a point onto C, we move closer to every point in C, and in particular, any optimal point, i.e.,

$$\|x^{(k+1)} - x^*\|_2 = \|P(z^{(k+1)}) - x^*\|_2 \leq \|z^{(k+1)} - x^*\|_2. \tag{27}$$

Combine with the inequality (13), we get

$$\begin{aligned} \|x^{(k+1)} - x^*\|_2^2 \leq &\|x^{(k)} - x^*\|_2^2 - 2\alpha_k (f(x^{(k)}) - f^*) \\ &+ \alpha_k^2 \|g^{(k)}\|_2^2. \end{aligned} \tag{28}$$

The remaining proof proceeds exactly as in the ordinary subgradient method.

Finally, we can get a similar convergence results as basic subgradient method.

3.3. Algorithm

The algorithm of projected subgradient method for constrained problem can be summarized as:

Algorithm 2

1. Set $k := 1$ and choose an infinite sequence of positive step size value α_k.
2. Compute a subgradient $g^{(k)} \in \partial f(x^k)$.
3. Update $x^{(k+1)} = P(x^k - \alpha_k g^{(k)})$.
4. If algorithm has not converged, then set $k := k + 1$ and go to step 2.

4. Stochastic subgradient method

When the subgradient of the objective function is difficult to compute exactly due to various reasons such as errors in measurements or intractability in the computation, we can use a noisy estimate of the subgradient for optimization. With a proper choice of the step size, we can guarantee the convergence with probability 1 or even stronger conclusions. Stochastic methods also apply when the objective function itself is difficult to compute exactly.

4.1. Update rules:

The definition of noisy unbiased subgradient

Random vector $\tilde{g} \in \mathbb{R}^n$ is a noisy unbiased subgradient for $f : \mathbb{R}^n \to \mathbb{R}$ at x, if Random vector $\tilde{g} \in R^n$ is a noisy unbiased subgradient for

$$f(z) \geq f(x) + (\mathbf{E}\tilde{g})^T (z - x) \tag{29}$$

$i.e., g = \mathbf{E}\tilde{g} \in \partial f(x)$

The noise can represent error in computing , measurement noise, Monte Carlo sampling error, etc

Update rules

Stochastic subgradient method is the subgradient method, using noisy unbiased subgradients:

$$x^{(k+1)} = x^{(k)} - \alpha_k \tilde{g}^{(k)}, \tag{30}$$

where $x^{(k)}$ is kth iterate, $\tilde{g}$ is any noisy unbiased subgradient of (convex) f at $x^{(k)}$, $\alpha_k > 0$ is the kth step size. Just like the basic subgradient method, here we also define $f_{best}^{(k)} = min\{f(x^{(1)}), ..., f(x^{(k)})\}$ to keep track of the best found point and corresponding function value.

4.2. Convergence Analysis

A very basic convergence result for the stochastic subgradient method will be given. First, the assumptions will be given as following:

- $f^* = inf_x f(x) > -\infty$, with $f(x^*) = f^*$
- $\mathbf{E}\|g^{(k)}\|_2^2 \leq G^2$ for all k
- $\mathbf{E}\|x^{(1)} - x^*\|_2^2 \leq R^2$

- step sizes are square-summable but not summable

It can be proved that

$$\mathbf{E}_{best}^{(k)} - f^* \leq \frac{R^2 + G^2 \parallel \alpha \parallel_2^2}{2\sum_{i=1}^{k} \alpha_i} \tag{31}$$

we can get the convergence in expectation:

$$\lim_{k\to\infty} \mathbf{E}_{best}^{(k)} \geq f^* \tag{32}$$

for various step size rules such as square-summable but not summable sequence $(e.g. \alpha_k = 1/k)$, and not summable diminishing sequence $(e.g. \alpha_k = 1/\sqrt{k})$.

Using Markovs inequality, we obtain the convergence in probability, i.e., for any $\epsilon > 0$,

$$\lim_{k\to\infty} Prob(f_{best}^{(k)} \geq f^* + \epsilon) = 0$$

More sophisticated methods can be used to show almost sure convergence.

4.3. Algorithm

Algorithm 3

1. Set $k := 1$ and choose an infinite sequence of positive step size value α_k.
2. Compute a noisy subgradient $\mathbf{E}\tilde{g}^{(k)} \in \partial f(x^k)$.
3. Update $x^{(k+1)} = x^k - \alpha_k \tilde{g}^{(k)}$.
4. If algorithm has not converged, then set $k := k + 1$ and go to step 2.

5. Recent research

Though subgradient method has been proposed for several decades, there are also many recent research are proceeding based on the scheme of subgradient method. Here I want to introduce two interesting recent research papers [9, 8] on subgradient methods.

The first paper [9] investigates incremental subgradient methods for nondifferentiable optimization. It mainly investigates a class of subgradient methods for minimizing a convex function that consists of the sum of a large number of component functions. The idea is to perform the subgradient iteration incrementally, by sequentially taking steps along the subgradients of the component functions, with intermediate adjustment of the variables after processingeac h component function. By randomizing the order of selection of component functions for iteration, the convergence rate is substantially improved.

The second paper [8] investigates primal-dual subgradient methods for convex problems. The proposed primal-dual subgradient schemes can always generate a feasible approximation to the optimum of an appropriately formulated dual problem, so that these methods can have reliable stopping criterion. The main difference from classical approach is that it can produce two control sequences, so the boundedness of the sequence of primal test points can be guaranteed even in the case of unbounded feasible set.

6. Conclusion

This paper surveys the subgradient methods nondifferentiable minimization problem.The background of subgradient methods has been introduced. Three major subgradient methods, i.e., basic subgradient methods, projected subgradients, and stochastic subgardient methods, have been investigated. The detailed convergence analysis and corresponding algorithm of the three methods have been provided. Beside, some recent research based on subgradient methods have been introduced.

References

[1] Bertsekas, D. P., Nonlinear Programming, (2nd edition), Athena Scientic, Belmont, MA, 1999.

[2] Dem'yanov, V. F., and Vasil'ev, L. V., Nondierentiable Optimization, Optimization Software, N.Y., 1985.

[3] Hiriart-Urruty, J.-B., and Lemarechal, C., Convex Analysis and Minimization Algorithms, Vols. I and II, Springer-Verlag, Berlin and N.Y., 1993.

[4] Minoux, M., Mathematical Programming: Theory and Algorithms, J. Wiley, N.Y., 1986.

[5] Polyak, B. T., Introduction to Optimization, Optimization Software Inc., N.Y., 1987.

[6] Shor, N. Z., Minimization Methods for Nondierentiable Functions, Springer-Verlag, Berlin, 1985.

[7] Solodov, M. V., "Incremental Gradient Algorithms with Stepsizes Bounded Away From Zero," *Computational Opt. and Appl.*, Vol. 11, 1998, pp. 28-35.

[8] Y.Nesterov. "Primal-dual subgradient methods for convex problems". *Math. Program. Ser.* B,120 :221259,2009.

[9] A.Nedic and D.Bertskas. "Incremental subgradient methods for nondifferentiable optimization". *SIAM J. on Optimization*, 12 :109- 138,2001.