Optimization

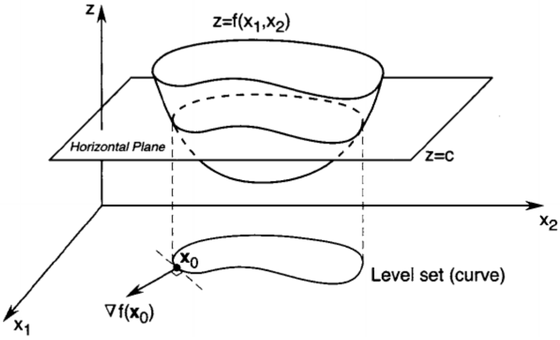
Màster de Fonaments de Ciència de Dades

Gerard Gómez

Lecture VIII Subgradient methods

Background. Gradient methods

Recall that **gradient methods** (such as steepest descent or conjugate gradient), are used for the computation of an extremum of a unconstrained minimization of a **continuously differentiable function**.



Background. Gradient methods

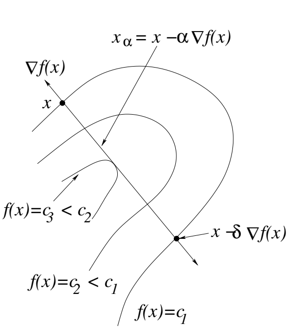
Gradient methods are based in the following equation:

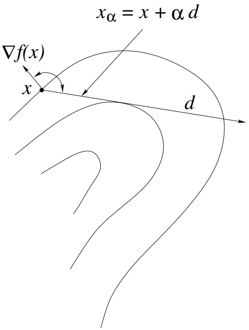
$$x_\alpha = x - \alpha \nabla f(x), \quad \alpha \geq 0,$$

that can be generalised to:

$$x_\alpha = x + \alpha d, \quad \alpha \geq 0,$$

where $\alpha \in \mathbb{R}$ is the **stepsize** and the **descent direction**,$d \in \mathbb{R}^n$ makes an angle with $\nabla f(x)$ greater than 90^o $(\nabla f(x)^T d < 0)$.



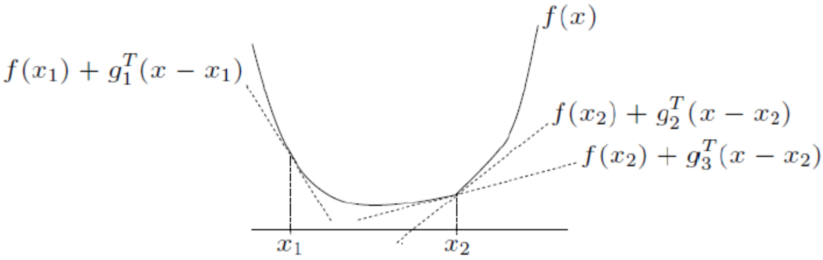


Background. Differential properties of convex functions

A **subgradient** of a convex function f at a point $x \in \mathbb{R}^n$, is any vector $g \in \mathbb{R}^n$ such that

$$f(y) \geq f(x) + g^T(y - x),$$

for every $y \in \mathbb{R}^n$.


$f(x)$
$f(x_1) + g_1^T(x - x_1)$
$f(x_2) + g_2^T(x - x_2)$
$f(x_2) + g_3^T(x - x_2)$
x_1
x_2

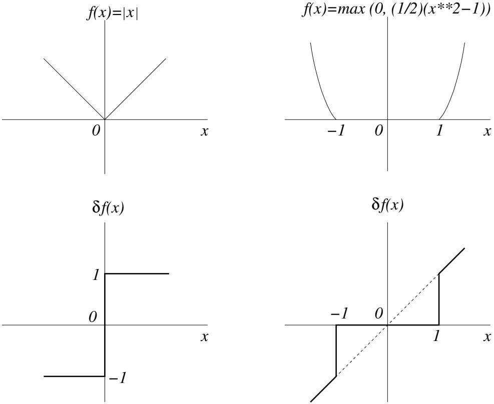
The **subgradient** of a **convex function**, is related to the ordinary gradient, in the case of differentiable convex functions, and to the directional derivatives in the more general case.

Subgradients of a convex function $f(x)$

For a convex function f it is possible that, at some point x

1. No vector $g \in \mathbb{R}^n$ satisfying $f(y) \geq f(x) + g^T(y-x)$ exists.
2. There is a unique vector $g \in \mathbb{R}^n$ satisfying $f(y) \geq f(x) + g^T(y-x)$
3. There is more than one vector $g \in \mathbb{R}^n$ satisfying $f(y) \geq f(x) + g^T(y-x)$

The set of all subgradients of a convex function f at x is denoted by $\partial f(x)$; this set is also called **subdifferential** of f



Background. Differential properties of convex functions

Some basic properties of subgradients are:

- The subdifferential of f, is a closed convex set (recall that f is convex).
- The set $\partial f(x)$ contains a single vector $g \in \mathbb{R}^n$ if and only if the convex function f is differentiable in the ordinary sense at x.
- If f is differentiable in the ordinary sense at x, then $g = \nabla f(x)$, that is

$$g_j = \frac{\partial f(x)}{\partial x_j}, \quad j = 1, ..., n.$$

- x^* is a minimizer of a convex f if and only if f is is subdifferentiable at x^* and

$$0 \in \partial f(x^*)$$

Subgradients can be characterized by the directional derivatives, according to the following theorem:

Theorem

A vector $g \in \mathbb{R}^n$ is a subgradient of a convex function f at a point x where $f(x)$ is finite if and only if

$$D^+ f(x; z) \geq g^T z,$$

for every direction z.

Calculus of subgradients

Some properties.

- Nonnegative scaling
$$\text{if } \alpha \geq 0 \quad \text{then} \quad \partial(\alpha f)(x) = \alpha \partial f(x).$$
- If $f = f_1 + ... + f_m$, with all the f_i convex, then
$$\partial f(x) = \partial f_1(x) + ... + \partial f_m(x).$$
- Affine transformation of domain. Suppose f is convex, and let $h(x) = f(Ax + b)$, then
$$\partial h(x) = A^T \partial f(Ax + b).$$
- Pointwise max. Suppose $f_1, ..., f_m$ are convex, and let $f(x) = \max_{i=1,\ldots,m} f_i(x)$, then
$$\partial f(x) = \text{convex hull } \{\partial f_i(x) \mid f_i(x) = f(x)\}.$$

Calculus of subgradients

Example 1.

Consider

$$f(x) = \max_{i=1,\dots,m} (a_i^T x + b_i)$$

- Let $f_i(x) = a_i^T x + b_i$, then $\partial f_i(x) = \{a_i\}$.
- Let

  $$\mathcal{K}(x) = \left\{ j \mid a_j^T x + b_j = \max_{i=1,\dots,m} (a_i^T x + b_i) \right\}$$

  then

  $$\partial f(x) = \mathrm{conv} \bigcup_{j \in \mathcal{K}(x)} \{a_j\}.$$
- In particular, when $\mathcal{K}(x) = \{k\}$ we have $\partial f(x) = \{a_k\}$.

Calculus of subgradients

Example 2.

Consider the case $f = f_1 + ... + f_n$, for instance:

$$f(x) = \|x\|_1 = |x_1| + ... + |x_n|$$

Then

$$\begin{aligned}
\partial f(x) &= \partial f_1(x) + ... + \partial f_m(x) \\
&= \{g \mid g_i = 1 \text{ if } x_i > 0,\ g_i = -1 \text{ if } x_i < 0,\ g_i \in [-1, 1] \text{ if } x_i = 0, \} \\
&= sign(x).
\end{aligned}$$

The subgradient method

- The **subgradient method** is a simple algorithm for minimizing a **non-differentiable convex function**.
- The method looks very much like the ordinary gradient method for differentiable functions, except that:
  - The step lengths are not chosen via a line search, as in the ordinary gradient method. In the most common cases, **the step lengths are fixed ahead of time**.
  - Unlike the ordinary gradient method, the subgradient method **is not a descent method**; the function value can (and often does) increase.
- The subgradient method is readily extended to **handle problems with constraints.**

Subgradient methods. Advantages and disadvantages

- Subgradient methods are **first-order methods**. Newton and interior-point methods are second-order methods and are much faster than subgradient methods.

- Subgradient methods do have some advantages over interior-point and Newton methods. They can be immediately applied to a far wider variety of problems than interior-point or Newton methods.

- The memory requirement of subgradient methods can be much smaller than an interior-point or Newton method, which means it **can be used for extremely large problems** for which interior-point or Newton methods cannot be used.

The subgradient method

- Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function.
- To minimize f, the **subgradient method** uses the iteration

  $$x^{k+1} = x^k - \alpha_k g^k,$$

  where x^k is the k-th iterate, g^k is any subgradient of f at x^k, and $\alpha_k > 0$ is the k-th step size.
- At each iteration of the method we take a step in the direction of a negative subgradient.
- As we have already said, a subgradient of f at x is any vector g that satisfies the inequality

  $$f(y) \geq f(x) + g^T(y - x), \quad \forall y$$
- When f is differentiable, the only possible choice for g_k is $\nabla f(x_k)$, and the subgradient method then reduces to the gradient method, except for the choice of step size α_k.

The subgradient method

- Since the subgradient method is not a descent method, it is common to keep track of the best point found so far, i.e., the one with smallest function value. At each step, we set

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

  and set

  $$i_{best}^{k} = k \quad \text{if} \quad f(x^k) = f_{best}^{k},$$

  so, x^k is the best point found so far.

- Then we have

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

  so f_{best}^{k} is the best objective value found in k iterations.

- Since f_{best}^{k} is decreasing, it has a limit (which can be $-\infty$).

- In a descent method there is no need to do this because the current point is always the best one so far.

Step size rules

Several different types of step size rules are used:

- **Constant step size**. $\alpha_k = h$ is a constant, independent of k.
- **Constant step length.** $\alpha_k = h/\|g^k\|_2$. Since $x^{k+1} = x^k - \alpha_k g^k$, this means that $\|x^{k+1} - x^k\|_2 = h$.
- **Square summable but not summable.** The step sizes satisfy

  $$\sum_{k=1}^{\infty} \alpha_k^2 < \infty, \qquad \sum_{k=1}^{\infty} \alpha_k = \infty.$$

  A typical example is $\alpha_k = a/(b+k)$, whith $a > 0$ and $b \geq 0$.
- **Nonsummable diminishing.** The step sizes satisfy

  $$\lim_{k \to \infty} \alpha_k = 0, \qquad \sum_{k=1}^{\infty} \alpha_k = \infty.$$

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

Convergence results

There are several results on convergence of the subgradient method.

- For **constant step size** and **constant step length**, the subgradient algorithm is guaranteed to **converge to within some range** of the optimal value:

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

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

  This implies that, in these cases, the subgradient method finds an ϵ-**suboptimal point within a finite number of steps**.

- For the **diminishing step size rule** and also the **square summable but not summable step size rule**, the algorithm is guaranteed to **converge to the optimal value**:

  $$\lim_{k\to\infty} f(x^k) = f^*.$$

- When the function f **is differentiable**, the subgradient method with **constant step size** yields **convergence** to the optimal value, **provided the step h is small enough**.

Convergence proof

We assume that:

- There is a minimizer of f, say x^*.
- The norm of the subgradients is bounded, i.e., there is a G such that

  $$\|g\|_2 \leq G \quad \text{for any} \quad g \in \partial f(x) \quad \text{and for any} \quad x,$$

  this is equivalent to assume that f is Lipschitz continuous with constant $G > 0$

  $$|f(x) - f(y)| \leq G\|x - y\|_2, \quad \forall x, y$$

  (see next slide for the proof).
- $$\|x^1 - x^*\|_2 \leq R$$
- In fact, some variants of the subgradient method work even when this assumption doesn't hold.

For the 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.**

Lipschitz vs bounded equivalence proof

Proof.

- Assume $\|g\|_2 \leq G$ for any subgradient at any point. Take $g_x \in \partial f(x)$, $g_y \in \partial f(y)$. Then

$$f(x) \geq f(y) + g_y^T(x-y) \quad \Rightarrow \quad f(x) - f(y) \geq g_y^T(x-y)$$

$$f(y) \geq f(x) + g_x^T(y-x) \quad \Rightarrow \quad f(y) - f(x) \geq g_x^T(y-x)$$

$$\Rightarrow \quad f(x) - f(y) \leq g_x^T(x-y)$$

  so

$$g_x^T(x-y) \geq f(x) - f(y) \geq g_y^T(x-y)$$

  and by the Cauchy-Schwarz inequality

$$G\|x-y\|_2 \geq |f(x) - f(y)| \geq -G\|x-y\|_2$$

- Assume $\|g\|_2 > G$ for some $g \in \partial f(x)$. Take $y = x + g/\|g\|_2$

$$f(y) \geq f(x) + g^T(y-x) = f(x) + \|g\|_2 > f(x) + G.$$

The basic inequality

If x^* is an optimal point, then

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

whith $f^*=f(x^*)$ and where we have used the definition of subgradient

$$f(x^*)\ge f(x^k)+(g^k)^T(x^*-x^k) \quad \Rightarrow \quad f(x^k)-f(x^*)\le (g^k)^T(x^k-x^*).$$

Applying the inequality above recursively, we have

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

Using $\|x^{k+1}-x^*\|_2^2\ge 0$, we have

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

Combining this with

$$\sum_{i=1}^k \alpha_i(f(x^i)-f^*) \ge \left(\sum_{i=1}^k \alpha_i\right)\min_{i=1,\dots,k}(f(x^i)-f^*) = \left(\sum_{i=1}^k \alpha_i\right)(f_{best}^k-f^*),$$

we get

The basic inequality (cont.)

$$f_{best}^k - f^* \leq \frac{\|x^1 - x^*\|_2^2 + \sum_{i=1}^k \alpha_i^2 \|g^i\|_2^2}{2\sum_{i=1}^k \alpha_i}.$$

Finally, using the assumption $\|g^k\|_2 \leq G$, we obtain the **basic inequality**

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

Since x^* is any minimizer of f, we can state that

$$f_{best}^k - f^* \leq \frac{dist(x^1, X^*)^2 + G^2 \sum_{i=1}^k \alpha_i^2}{2\sum_{i=1}^k \alpha_i}.$$

where X^* denotes the optimal set, and $dist(x^1, X^*)$ is the (Euclidean) distance of x^1 to the optimal set which is assumed to be bounded by R.

Convergence for constant step size h

- If $\alpha_k = h$, we have

  $$f_{best}^k - f^* \leq \frac{dist(x^1, X^*)^2 + G^2 h^2 k}{2hk} = \frac{dist(x^1, X^*)^2}{2hk} + \frac{G^2 h}{2},$$

  and the righthand side converges to $G^2 h/2$ as $k \to \infty$.

- Thus, for the subgradient method with fixed step size h, f_{best}^k converges within $G^2 h/2$ of optimal.

- We can also say that:

  $$f(x^k) - f^* \leq G^2 h$$

  within a finite number of steps.

Convergence for constant step length $\alpha_k = h/\|g^k\|_2$

- If $\alpha_k = h/\|g^k\|_2$, then the basic inequality becomes

$$f_{best}^k - f^* \leq \frac{dist(x^1, X^*)^2 + h^2 k}{2\sum_{i=1}^k \alpha_i}.$$

- By assmpution, we have $\alpha_k \geq g/G$. Applying this to the denominator of the above inequality gives

$$f_{best}^k - f^* \leq \frac{dist(x^1, X^*)^2 + h^2 k}{2hk/G} = \frac{G\ dist(x^1, X^*)^2}{2hk} + \frac{Gh}{2}.$$

- The righthand side converges to $Gh/2$ as $k \to \infty$, so in this case the subgradient method converges to within $Gh/2$ of optimal.

Square summable but not summable

Now suppose

$$\sum_{k=1}^{\infty} \alpha_k^2 < \infty, \qquad \sum_{k=1}^{\infty} \alpha_k = \infty.$$

Then, we have

$$f_{best}^k - f^* \leq \frac{dist(x^1, X^*)^2 + G^2 \|\alpha\|_2^2}{2 \sum_{i=1}^k \alpha_i},$$

which converges to zero as $k \to \infty$.

In other words, in this case the subgradient method converges:

$$f_{best}^k \to f^*$$

Nonsummable diminishing

If the sequence $\{\alpha_k\}$ converges to zero and is nonsummable

$$\lim_{k\to\infty} \alpha_k = 0, \qquad \sum_{k=1}^{\infty} \alpha_k = \infty,$$

then the right hand side of

$$f_{best}^k - f^* \leq \frac{\|x^1 - x^*\|_2^2 + G^2 \sum_{i=1}^k \alpha_i^2}{2 \sum_{i=1}^k \alpha_i}$$

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

$$\sum_{i=1}^k \alpha_i \geq \frac{1}{\epsilon} \left(\|x^1 - x^*\|_2^2 + G^2 \sum_{i=1}^{N_1} \alpha_i^2 \right) \quad \text{for all} \quad k > N_2.$$

Nonsummable diminishing (cont.)

Let $N = \max(N_1, N_2)$. Then for all $k > N$, we have

$$\begin{aligned}
\min_{i=1,\dots,k}(f(x^i) - f^*) \quad &\le \quad \frac{\|x^1 - x^*\|_2^2 + G^2 \sum_{i=1}^{N_1} \alpha_i^2}{2\sum_{i=1}^{k} \alpha_i} + \frac{G^2 \sum_{i=N_1+1}^{k} \alpha_i^2}{2\sum_{i=1}^{N_1} \alpha_i + 2\sum_{i=N_1+1}^{k} \alpha_i} \\
&\le \quad \frac{\|x^1 - x^*\|_2^2 + G^2 \sum_{i=1}^{k} \alpha_i^2}{(2/\epsilon)\left(\|x^1 - x^*\|_2^2 + G^2 \sum_{i=1}^{N_1} \alpha_i^2\right)} + \frac{G^2 \sum_{i=N_1+1}^{k} (\epsilon/G^2)\alpha_i}{2\sum_{i=N_1+1}^{k} \alpha_i} \\
&\le \quad \epsilon/2 + \epsilon/2 \;=\; \epsilon.
\end{aligned}$$

Stopping criterion

- Terminating when

$$\frac{R^2 + G^2 \sum_{i=1}^{k} \alpha_i^2}{2 \sum_{i=1}^{k} \alpha_i} \leq \epsilon$$

  which is really, really, slow

- Do an optimal choice of α_i to achieve

$$\frac{R^2 + G^2 \sum_{i=1}^{k} \alpha_i^2}{2 \sum_{i=1}^{k} \alpha_i} \leq \epsilon$$

  for smallest k, for instance:

$$\alpha_i = \frac{R/G}{\sqrt{k}}$$

  In this case, the minimum number of steps required is (see also Polyak's step length)

$$k = \left(\frac{RG}{\epsilon}\right)^2$$

- The truth: there really isn't a good stopping criterion for the subgradient method...

The projected subgradient method

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

$$\begin{cases} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathcal{C} \end{cases}$$

where $\mathcal{C}$ is a convex set.

The projected subgradient method is given by

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

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

All the step size rules described for the subgradient method can be used here, with similar convergence results.

The projected subgradient method. Convergence

The convergence proofs for the subgradient method are readily extended to handle the projected subgradient method. Let

$$z^{k+1} = x^k - \alpha_k g^k,$$

this is, a standard subgradient update before the projection back onto $\mathcal{C}$. As in the subgradient method, we have

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

Now, since when we project a point onto $\mathcal{C}$, we move closer to every point in $\mathcal{C}$ and $x^{k+1} = P(z^{k+1})$, we observe that

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

Combining this with the inequality above we get

$$\|x^{k+1} - x^*\|_2^2 \leq \|x^k - x^*\|_2^2 - 2\alpha_k (f(x^k) - f^*) + \alpha_k^2 \|g^k\|_2^2$$

and the proof proceeds exactly as in the ordinary subgradient method.

An alternative projected subgradient method

In some cases, the projected subgradient update can be set in an alternative way.

When $\mathcal{C}$ is affine, i.e., $\mathcal{C} = \{x \mid Ax = b\}$, where $A \in \mathbb{R}^{n \times m}$ is fat $(m < n)$ and full rank, the projection operator is affine, and given by

$$P(z) = z - A^T(AA^T)^{-1}(Az - b).$$

In this case, we can simplify the subgradient update to

$$x^{k+1} = x^k - \alpha_k(I - A^T(AA^T)^{-1}A)g^k$$

where we have used $Ax^k = b$.

An alternative projected subgradient method. Numerical example

Consider the least l_1-norm problem

$$\begin{cases} \text{minimize} & \|x\|_1 \\ \text{subject to} & Ax = b \end{cases}$$

whith $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times m}$ and $b \in \mathbb{R}^m$. This problem can also be solved using linear programming.

Assume that A is fat ($m < n$) and full rank, $rank(A) = m$.

As we have already seen, the subgradient of the objective function at x is given by $g = sign(x)$, where $g_i = -1$ if $x_i < 0$, $g_i = 0$ if $x_i = 0$ and $g_i = +1$ if $x_i > 0$

Thus, the projected subgradient update is

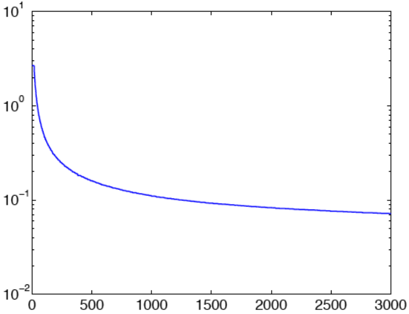
$$x^{k+1} = x^k - \alpha_k (I - A^T (AA^T)^{-1} A) \, sign(x^k).$$

An alternative projected subgradient method. Numerical example

Consider the above problem with $n = 1000$ and $m = 50$, with randomly generated A and b. We use the least-norm solution as the starting point:

$$x^1 = A^T(AA^T)^{-1}b$$

The value of $f^* \approx 3.2$ is computed using linear programming.
The figure shows the progress of the projected subgradient method, $f_{best}^k - f^*$ vs k, with the Polyak estimated step size rule $\gamma_k = 100/k$ (to be explained later).



Piecewise linear minimization

Consider the following problem:

$$\text{minimize} \quad f(x) = \max_{i=1,\ldots,m} (a_i^T x + b_i).$$

As we have already seen, finding a subgradient of f is easy: given x, we first find an index j for which

$$a_j^T x + b_j = \max_{i=1,\ldots,m} (a_i^T x + b_i).$$

Then, we can take as subgradient $g = a_j$, and $G = \max_{i=1,\ldots,m} \|a_i\|_2$.

The subgradient method update has the form

$$x^{k+1} = x^k - \alpha_k a_j,$$

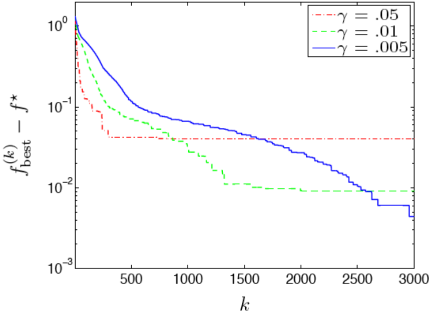
where j is chosen so that: $a_j^T x + b_j = \max_{i=1,\ldots,m}(a_i^T x + b_i)$.

Note that to apply the subgradient method, all we need is a way to evaluate $\max_{i=1,\ldots,m}(a_i^T x + b_i)$.

Piecewise linear minimization. Example

$$\text{minimize} \quad f(x) = \max_{i=1,\ldots,m} (a_i^T x + b_i),$$

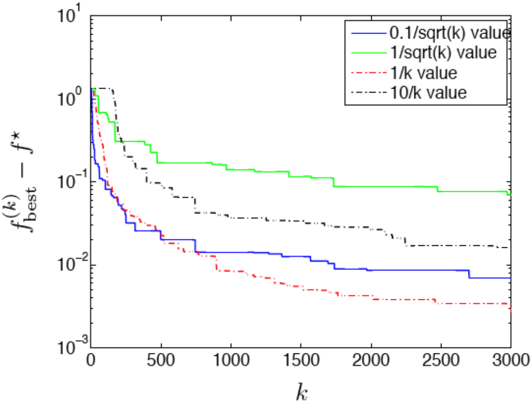
with $m = 100$ terms and $n = 20$ variables: $a_i, b_i, x \in \mathbb{R}^{20}$. Take the values of a_i and b_i randomly. There is no simple way to find a justifiable value for R (a value of R for which we can prove that $\|x^k - x*\|_2 \leq R$), and the value $R = 10$ has been used (in the numerical example: $f^* \approx 1.1$ and $R = 0.91$)



Results for constant step length $\gamma = 0.05, 0.01, 0.005$. Larger γ gives faster convergence but larger final suboptimality.

Piecewise linear minimization. Example

The subgradient method is very slow!



Results for dimishing step rules $\alpha_k = 0.1/\sqrt{k}$, $1/\sqrt{k}$, and square summable step size rules $\alpha_k = 1/k$, $10/k$

Polyak's step length

Polyak suggests a step size that can be used when the optimal value f^* is known, and is in some sense optimal.

One can think that f^* is rarely known, but we will see that's not the case.

The step size is

$$\alpha_k = \frac{f(x^k) - f^*}{\|g^k\|_2^2}.$$

To motivate this step size, imagine that that

$$f(x^k) - \alpha g^k) \approx f(x^k) + (g^k)^T \left(x^k - \alpha g^k - x^k \right) = f(x^k) - \alpha (g^k)^T g^k$$

This would be the case if α were small, and $g^k = \nabla f(x^k)$.

Replacing the lefthand side with f^* and solving for α gives the step length above.

We can give another simple motivation for the above step length. The subgradient method starts from the basic inequality:

$$\|x^{k+1} - x^*\|_2^2 \leq \|x^k - x^*\|_2^2 - 2\alpha_k (f(x^k) - f^*) + \alpha_k^2 \|g^k\|_2^2.$$

The step size minimizes the righthand side.

Polyak's step length

To analyze convergence, we substitute the value of the step size into the **basic inequality**

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

to get

$$2\sum_{i=1}^{k}\frac{(f(x^i)-f^*)^2}{\|g^i\|_2^2}\le R^2+\sum_{i=1}^{k}\frac{(f(x^i)-f^*)^2}{\|g^i\|_2^2}$$

so

$$\sum_{i=1}^{k}\frac{(f(x^i)-f^*)^2}{\|g^i\|_2^2}\le R^2$$

Using $\|g^k\|_2^2\le G$ we get

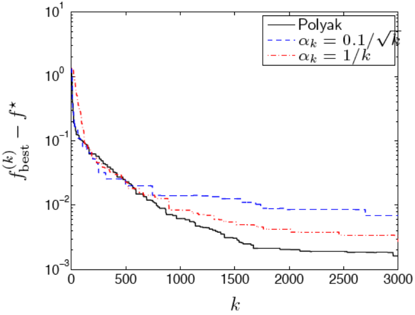
$$\sum_{i=1}^{k}(f(x^i)-f^*)^2\le R^2G^2$$

We conclude that $f(x^k)\rightarrow f^*$. The number of steps needed before we can guarantee suboptimality ϵ is $k=(RG/\epsilon)^2$.

Polyak's step length. Example

The next figure shows the progress of the subgradient method with Polyak's step size for the same piece-wise linear example.

Of course this isn't fair, since we don't know f^* before solving the problem; but even with this unfair advantage in choosing step lengths, the subgradient method is pretty slow.



Polyak's step size choice with estimated f^*

The basic idea is to estimate the optimal value f^*, as $f_{best}^k - \gamma_k$, where $\gamma_k > 0$, $\gamma_k \to 0$ and $\sum_{k=1}^{\infty} \gamma_k = \infty$. This gives as step size

$$\alpha_k = \frac{f(x^k) - f_{best}^k + \gamma_k}{\|g^k\|_2^2}.$$

In this case we have $f_{best} \to f^*$.
To show this, we substitute α_i into the basic inequality to get

$$\begin{aligned} R^2 &\geq \sum_{i=1}^{k} \left(2\alpha_i(f(x^i) - f^*) - \alpha_i^2 \|g^k\|_2^2\right) \\ &= \sum_{i=1}^{k} \frac{2(f(x^i) - f_{best}^i + \gamma_i)(f(x^i) - f^*) - (f(x^i) - f_{best}^i + \gamma_i)^2}{\|g^k\|_2^2} \\ &= \sum_{i=1}^{k} \frac{(f(x^i) - f_{best}^i + \gamma_i)\left((f(x^i) - f^*) + (f_{best}^i - f^*) - \gamma_i)\right)}{\|g^k\|_2^2} \end{aligned}$$

Polyak's step size choice with estimated f^* (cont.)

Now we can prove convergence. Suppose $f^i_{best} - f^* \geq \epsilon > 0$. Then for $i = 1, ..., k$, $f(x^i) - f^* \geq \epsilon$. Find N for which $\gamma_i \leq \epsilon$ for $i \geq N$. This implies the second term in the numerator is at least ϵ:

$$(f(x^i) - f^*) + (f^i_{best} - f^*) - \gamma_i \geq \epsilon.$$

In particular, it is positive. It follows the terms in the sum above for $i \geq N$ are positive. Let S denote the sum above, up to $i = N - 1$. (We assume $k \geq N$.) We then have

$$\sum_{i=1}^{k} \frac{(f(x^i) - f^i_{best} + \gamma_i)\left((f(x^i) - f^*) + (f^i_{best} - f^*) - \gamma_i)\right)}{\|g^k\|_2^2} \leq R^2 - S.$$

We get a lower bound on the lefthand side using

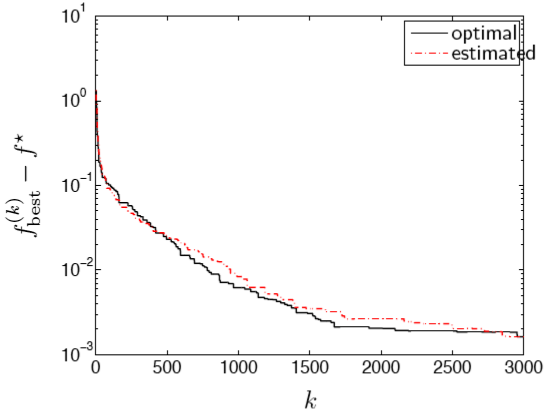
$$f(x^i) - f^i_{best} + \gamma_i \geq \gamma_i$$

along with the inequality above and $\|g^k\|_2^2 \leq G$ get

$$\frac{\epsilon}{G^2} \sum_{i=N}^{k} \leq R^2 - S.$$

Since the lefthand side converges to ∞ and righthand side doesn't depend on k, we see that k cannot be too large.

Polyak's step length. Example



Value of $f^{i}_{best} - f^{*}$ versus iteration number k, for the subgradient method with Polyak's step size (solid black line) and the estimated optimal step size (dashed red line).

Finding a point in the intersection of convex sets

Suppose we want to find a point in

$$C = C_1 \cap \dots \cap C_m$$

where $C_1, \dots, C_m \subseteq \mathbb{R}^n$ are closed and convex, and we assume that C is nonempty. We can do this by minimizing the function

$$f(x) = \max\{dist(x, C_1), \dots, dist(x, C_m)\},$$

which is convex, and has minimum value $f^* = 0$ (since C is nonempty).

Let us see how to find a subgradient g of f at x.

- If $f(x) = 0$, we can take $g = 0$ (which in any case means we are done).
- Otherwise find an index j such that $dist(x, C_j) = f(x)$, i.e., find a set that has maximum distance to x. A subgradient of f is

  $$g = \nabla dist(x, C_j) = \frac{x - P_{C_j}(x)}{\|x - P_{C_j}(x)\|_2},$$

  where P_{C_j} is Euclidean projection onto C_j. Note that $\|g\|_2 = 1$, so we can take $G = 1$.

Finding a point in the intersection of convex sets

The subgradient algorithm update, with step size rule

$$\alpha_k = \frac{f(x^k) - f_{best}^k + \gamma_k}{\|g^k\|_2^2}.$$

and assuming that the index j is one for which x^k has maximum distance to C_j, is given by

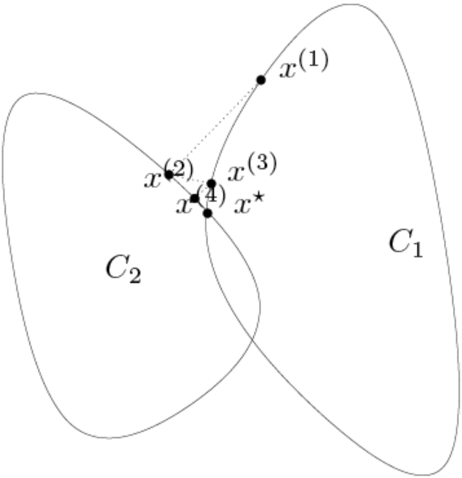
$$\begin{aligned} x^{k+1} &= x^k - \alpha_k g^k \\ &= x^k - f(x^k)\frac{x^k - P_{C_j}(x^k)}{\|x^k - P_{C_j}(x^k)\|_2} \\ &= P_{C_j}(x^k) \end{aligned}$$

We have used: $\|g^k\|_2 = 1$, $f^* = 0$, and

$$f(x^k) = dist(x, C_j) = \|x^k - P_{C_j}(x^k)\|_2$$

The algorithm is very simple: at each step, we simply project the current point onto the farthest set. This is an extension of the alternating projections algorithm. (When there are just two sets, then at each step you project the current point onto the other set. Thus the projections simply alternate.)

Alternating projections. Example



First few iterations of the gradient method that, eventually, converge to a point $x^* \in C_1 \cap C_2$

Subgradient method for inequality constrained optimization

The subgradient algorithm can be extended to solve the inequality constrained problem

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, ..., m, \end{array}$$

where f_i are convex. The algorithm takes the same form:

$$x^{k+1} = x^k - \alpha_k g^k$$

where $\alpha_k > 0$ is a step size, and g^k is a subgradient of the objective or one of the constraint functions at x^k. More specifically, we take

$$g^k \in \begin{cases} \partial f_0(x^k) & f_i(x^k) \leq 0, \quad i = 1, ..., m \\ \partial f_j(x^k) & f_j(x^k) > 0. \end{cases}$$

In other words:

- if the current point is feasible, we use an objective subgradient, as if the problem were unconstrained,
- if the current point is infeasible, we choose any violated constraint, and use a subgradient of the associated constraint function.

Subgradient method for inequality constrained optimization

As in the basic subgradient method, we keep track of the best (feasible) point found so far:

$$f_{best}^{k} = \min\{f_0(x^i) \mid x^i \text{ feasible}, \ i = 1, \ldots, k\}$$

(If none of the points $x^1, \ldots, x^k$ is feasible, then $f_{best}^{k} = \infty$.

We assume that:

- the problem is strictly feasible: there is some point x^{sf} with $f_i(x^{sf}) < 0$, $i = 1, \ldots, m$
- the problem has an optimal point x^*
- there are numbers R and G with $\|x^1 - x^*\|_2 \leq R$, $\|x^{sf} - x^*\|_2 \leq R$ and $\|g^k\|_2 \leq G$ for all k.

We will proof convergence of the generalized subgradient method using diminishing nonsummable α_k. (Similar results can be obtained for other step size rules.)

We claim that $f_{best}^{k} \to f^*$ as $k \to \infty$. This implies in particular that we obtain a feasible iterate within some finite number of steps.

Subgradient method for inequality constrained optimization. Convergence

- Assume that $f_{best}^k \to f^*$ does not occur. Then there exists some $\epsilon > 0$ so that $f_{best}^k \geq f^* + \epsilon$ for all k, which in turn means that $f(x^k) \geq f^* + \epsilon$ for all k for which x^k) is feasible. We'll show this leads to a contradiction.

- We first find a point $\tilde{x}$ and positive number μ that satisfy

$$f_0(\tilde{x}) \leq f^* + \epsilon/2, \quad f_1(\tilde{x}) \leq -\mu, ..., f_m(\tilde{x}) \leq -\mu.$$

  Such a point is $(\epsilon/2)$-suboptimal, and also satisfies the constraints with a margin of μ.

- We will take $\tilde{x} = (1-\theta)x^* + \theta x^{sf}$, where $\theta \in (0,1)$. We have

$$f_0(\tilde{x}) \leq (1-\theta)f^* + \theta f_0(x^{sf}),$$

  so, if we choose $\theta = \min\{1, (\epsilon/2)/(f_0(x^{sf}) - f^*)\}$, we have $f_0(\tilde{x}) \leq f^* + \epsilon/2$.

- We have

$$f_i(\tilde{x}) \leq (1-\theta)f_i(x^*) + \theta f_i(x^{sf}) \leq \theta f_i(x^{sf}),$$

  so, we can take

$$\mu = -\theta \min_i f_i(x^{sf}).$$

Subgradient method for inequality constrained optimization. Convergence (cont.)

- Consider any index $i \in \{1, ..., k\}$ for which x^i is feasible. Then we have $g^i \in \partial f_0(x^i)$ and also $f_0(x^i) \geq f^* + \epsilon$.
  Since x^* is $(\epsilon/2)$-suboptimal, we have $f_0(x^i) - f_0(\tilde{x}) \geq \epsilon/2$. Therefore

$$\begin{aligned} \|x^{i+1} - \tilde{x}\|_2^2 &= \|x^i - \tilde{x}\|_2^2 - 2\alpha_i (g^i)^T (x^i - \tilde{x}) + \alpha_i^2 \|g^k\|_2^2 \\ &\leq \|x^i - \tilde{x}\|_2^2 - 2\alpha_i (f_0(x^i) - f_0(\tilde{x})) + \alpha_i^2 \|g^k\|_2^2 \\ &\leq \|x^i - \tilde{x}\|_2^2 - \alpha_i \epsilon + \alpha_i^2 \|g^k\|_2^2. \end{aligned}$$

  In the second line here we use the usual subgradient inequality

$$f_0(\tilde{x}) \geq f_0(x^i) + (g^i)^T (\tilde{x} - x^i)$$

- Now suppose that $i \in \{1, ..., k\}$ is such that x^i is infeasible, and that $g^i \in \partial f_p(x^i)$ where $f_p(x^i) > 0$. Since $f_p(\tilde{x}) \leq -\mu$, we have $f_p(x^i) - f_p(\tilde{x}) \geq \mu$. Therefore

$$\begin{aligned} \|x^{i+1} - \tilde{x}\|_2^2 &= \|x^i - \tilde{x}\|_2^2 - 2\alpha_i (g^i)^T (x^i - \tilde{x}) + \alpha_i^2 \|g^k\|_2^2 \\ &\leq \|x^i - \tilde{x}\|_2^2 - 2\alpha_i (f_p(x^i) - f_p(\tilde{x})) + \alpha_i^2 \|g^k\|_2^2 \\ &\leq \|x^i - \tilde{x}\|_2^2 - 2\alpha_i \mu + \alpha_i^2 \|g^k\|_2^2. \end{aligned}$$

Subgradient method for inequality constrained optimization. Convergence (cont.)

- $$\|x^{i+1} - \tilde{x}\|_2^2 \leq \|x^i - \tilde{x}\|_2^2 - \alpha_i \delta + \alpha_i^2 \|g^k\|_2^2,$$

  where $\delta = \min\{\epsilon, 2\mu\}$. Applying this inequality recursively for $i = 1, ..., k$, we get

  $$\|x^{k+1} - \tilde{x}\|_2^2 \leq \|x^1 - \tilde{x}\|_2^2 - \delta \sum_{i=1}^{k} \alpha_i + \sum_{i=1}^{k} \alpha_i^2 \|g^k\|_2^2,$$

- It follows that

  $$\delta \sum_{i=1}^{k} \alpha_i \leq R^2 + G^2 \sum_{i=1}^{k} \alpha_i^2,$$

  which cannot hold for large k since

  $$\frac{R^2 + G^2 \sum_{i=1}^{k} \alpha_i^2}{\sum_{i=1}^{k} \alpha_i}$$

  converges to zero as $k \to \infty$.

Subgradient method for inequality constrained optimization. Numerical example

Consider the linear problem

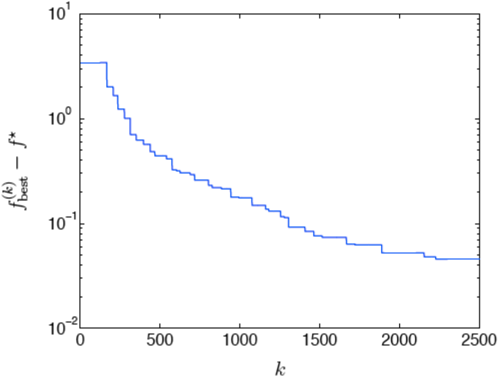
$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & a_i^T x \leq b_i, \quad i = 1, ..., m, \end{array}$$

with $x \in \mathbb{R}^n$.

The objective and constraint functions are affine, and so have only one subgradient, independent of x. For the objective function we have $g = c$, and for the i-th constraint we have $g_i = a_i$.

We solve the problem with $n = 20$ and $m = 200$ using the subgradient method. The value of $f^* \approx -3.4$ is obtained by other means. The next figure shows progress of the subgradient method, which uses the square summable step size with $\alpha_k = 1/k$ for the optimality update. The objective value only changes for the iterations when x(k) is feasible.

Subgradient method for inequality constrained optimization. Numerical example



Value of$f_{best}^{k} - f^{*}$ versus the iteration number k, using the square summable step size with $\alpha_k = 1/k$ for the optimality update.