Linear search methods

Alexandre Gramfort

Master 2 Data Science, Univ. Paris Saclay
Optimisation for Data Science

Table of Contents

Table of Contents

Why line search?

Descent algorithm reads:

$$x_{k+1} = x_k + t_k d_k\,,\;\; t_k \geq 0$$

where d_k is a descent direction ($\exists t_k > 0$ s.t. $f(x_{k+1}) < f(x_k)$).
In the case of gradient descent one uses:

$$d_k = -\nabla f(x_k)$$

and if f has a Lipschitz continuous gradient with constant L then one can use $t_k = \frac{1}{L}$.
Problem: L is a global quantity (does not depend on x_k) and can be unknown.
Objective: Derive strategies to estimate "good enough" t_k (optimal step can be really costly in non-quadratic case).

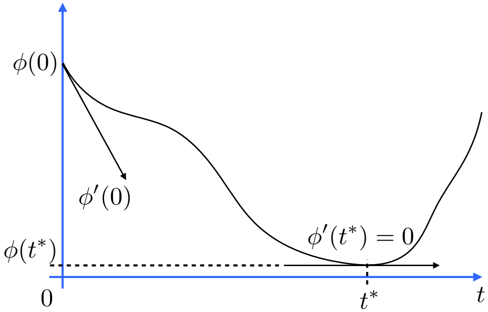
Why line search?

Let $\phi(t) = f(x_k + td_k)$

Objective: find $t > 0$ such that $\phi(t) \leq \phi(0)$

For f is smooth, the optimal step size t^* is caracterized by:

$$\begin{cases} \phi'(t^*) = 0 & \text{(is a minimum)} \\ \phi(t) \geq \phi(t^*) \text{ for } 0 \leq t \leq t^* & \text{(decreases objective)} \end{cases}$$



Why line search?

Let

$$\phi(t) = f(x_k + td_k)$$

Objective: find $t > 0$ such that $\phi(t) \leq \phi(0)$

Exercise: Show that with $d_k = -\nabla f(x_k)$ and optimal step size $d_{k+1}^T d_k = 0$.

Security interval

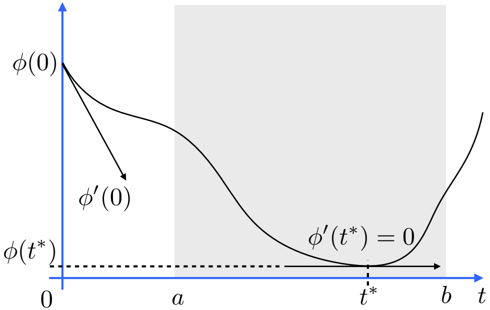
Definition (Security interval)

$[a, b]$ is a security interval if one can classify t values as:

- If $t < a$ then t is too small
- If $a \leq t \leq b$ then t is ok
- If $t > b$ then t is too big

Problem: How to translate these conditions from values of ϕ?
Problem: How to define a and b.

Security interval



Basic algorithm

Start from $[\alpha, \beta]$ with $[a, b] \subset [\alpha, \beta]$, e.g., $\alpha = 0$ and β large (always exists if f is coercive).

Basic algorithm

Start from $[\alpha, \beta]$ with $[a, b] \subset [\alpha, \beta]$, e.g., $\alpha = 0$ and β large (always exists if f is coercive).

Definition

f is coercive if

$$\lim_{\|x\| \to \infty} f(x) = +\infty$$

1. Choose t in $[\alpha, \beta]$
2. If t is too small then set $\alpha = t$ and go back to 1.
3. If t is too big then set $\beta = t$ and go back to 1.
4. If t is ok then stop

Problem: How to translate the "too small", "too big" and "ok" from values of ϕ?

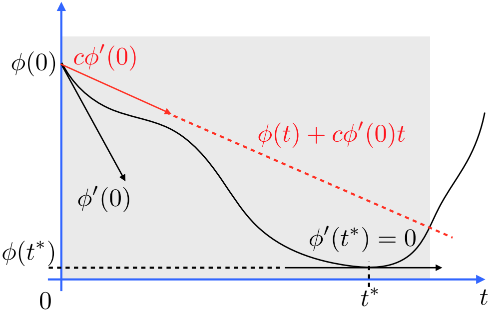
Table of Contents

Armijo's rule

Set $\alpha = 0$ and fix $0 < c < 1$.

Definition (Armijo's rule)

1. If $\phi(t) > \phi(0) + c\phi'(0)t$, then t is too big
2. If $\phi(t) \leq \phi(0) + c\phi'(0)t$, then ok



Armijo's rule

Set $\alpha = 0$ and fix $0 < c < 1$.

Definition (Armijo's rule)

1. If $\phi(t) > \phi(0) + c\phi'(0)t$, then t is too big
2. If $\phi(t) \leq \phi(0) + c\phi'(0)t$, then ok

Problem: As $\alpha = 0$, t is never considered too small. So Armijo is not heavily used in practice.

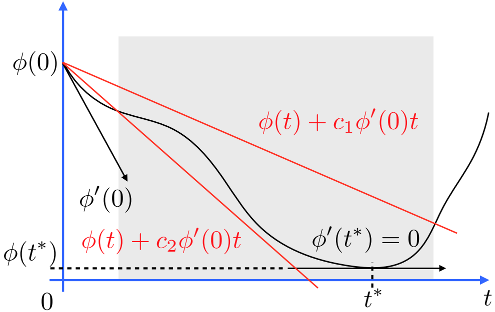
Note: You have function `scalar_search_armijo` in `scipy/optimize/linesearch.py` but it does more (cubic interpolation, backtracking).

Goldstein's rule

Goldstein is Armijo with an extra inequality. Let $0 < c_1 < c_2 < 1$.

Definition (Goldstein's rule)

1. If $\phi(t) < \phi(0) + c_2\phi'(0)t$, then t is too small
2. If $\phi(t) > \phi(0) + c_1\phi'(0)t$, then t is too big
3. If $\phi(0) + c_1\phi'(0)t \geq \phi(t) \geq \phi(0) + c_2\phi'(0)t$, then ok



Goldstein's rule

c_2 should be chosen such that t^* in the quadratic case is in the security interval.
In the quadratic case:

$$\phi(t) = \frac{1}{2}at^2 + \phi'(0)t + \phi(0), a > 0$$

and t^* satisfies $\phi'(t^*) = 0$, so $t^* = -\frac{\phi'(0)}{a}$ and so

$$\phi(t^*) = \frac{\phi'(0)}{2}t^* + \phi(0)$$

which means that one should have $c_2 \geq \frac{1}{2}$.

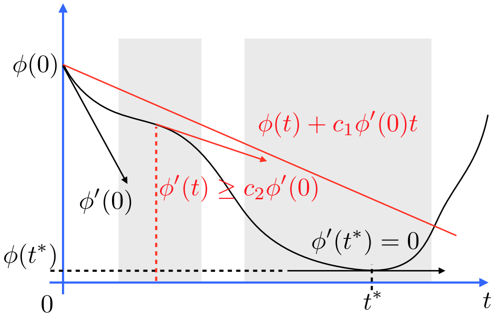
Common values used in practice are $c_1 = 0.1$ and $c_2 = 0.7$.

Wolfe's rule

Requires $\phi'(t) = d_k^\top \nabla f(x_k + t d_k)$ (in theory more costly).

Definition: Wolfe's rule (with $0 < c_1 < c_2 < 1$)

1. If $\phi(t) > \phi(0) + c_1 \phi'(0) t$, then t is too big (like Goldstein)
2. If $\phi(t) \leq \phi(0) + c_1 \phi'(0) t$, and $\phi'(t) < c_2 \phi'(0)$ then t is too small
3. If $\phi(t) \leq \phi(0) + c_1 \phi'(0) t$, and $\phi'(t) \geq c_2 \phi'(0)$, then ok



Wolfe's rule

Requires $\phi'(t) = d_k^\top \nabla f(x_k + td_k)$ (in theory more costly).

Definition: Wolfe's rule (with $0 < c_1 < c_2 < 1$)

1. If $\phi(t) > \phi(0) + c_1\phi'(0)t$, then t is too big (like Goldstein)
2. If $\phi(t) \leq \phi(0) + c_1\phi'(0)t$, and $\phi'(t) < c_2\phi'(0)$ then t is too small
3. If $\phi(t) \leq \phi(0) + c_1\phi'(0)t$, and $\phi'(t) \geq c_2\phi'(0)$, then ok

Note: The idea is to guarantee that t is not too small by requiring that the gradient is increased enough.

Strong Wolfe's rule

Requires $\phi'(t) = d_k^\top \nabla f(x_k + td_k)$ (in theory more costly).

Definition: Strong Wolfe's rule (with $0 < c_1 < c_2 < 1$)

1. If $\phi(t) > \phi(0) + c_1\phi'(0)t$, then t is too big (like Goldstein)
2. If $\phi(t) \leq \phi(0) + c_1\phi'(0)t$, and $|\phi'(t)| > c_2|\phi'(0)|$ then t is too small
3. If $\phi(t) \leq \phi(0) + c_1\phi'(0)t$, and $|\phi'(t)| \leq c_2|\phi'(0)|$, then ok

Note: This is implemented in `scipy.optimize.line_search`.

Table of Contents

Existence of steps that satisfy Wolfe conditions

Proposition (Existence)

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ *is continuously differentiable.*

Let d_k *be a descent direction at* x_k*, and assume that* f *is bounded below along the ray* $\{x_k + td_k | t > 0\}$.

Then if $0 < c_1 < c_2 < 1$*, there exist intervals of step lengths satisfying the Wolfe conditions and the strong Wolfe conditions.*

Existence of steps that satisfy Wolfe conditions

Proposition (Existence)

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ *is continuously differentiable.*

Let d_k *be a descent direction at* x_k*, and assume that* f *is bounded below along the ray* $\{x_k + td_k | t > 0\}$.

Then if $0 < c_1 < c_2 < 1$, *there exist intervals of step lengths satisfying the Wolfe conditions and the strong Wolfe conditions.*

Take home message: One can always find a good step size for a smooth and bounded below function.

Proof of existence

Since $\phi(t) = f(x_k + td_k)$ is bounded below for all $t > 0$ and since $0 < c_1 < 1$, the line $l(t) = f(x_k) + tc_1 \nabla f(x_k)^\top d_k$ must intersect the graph of ϕ at least once.
Let $t' > 0$ be the smallest intersecting value of t, that is,

$$f(x_k + t'd_k) = f(x_k) + t'c_1 \nabla f_k^\top d_k \ .$$

The sufficient decrease condition (Armijo) clearly holds for all $t \leq t'$.

Proof of existence

Since $\phi(t) = f(x_k + td_k)$ is bounded below for all $t > 0$ and since $0 < c_1 < 1$, the line $l(t) = f(x_k) + tc_1 \nabla f(x_k)^\top d_k$ must intersect the graph of ϕ at least once.
Let $t' > 0$ be the smallest intersecting value of t, that is,

$$f(x_k + t'd_k) = f(x_k) + t'c_1 \nabla f_k^\top d_k \quad .$$

The sufficient decrease condition (Armijo) clearly holds for all $t \leq t'$.

By the mean value theorem, there exists $t'' \in (0, t')$ such that

$$f(x_k + t'd_k) - f(x_k) = t' \nabla f(x_k + t''d_k)^\top d_k \quad .$$

By combining both, we obtain:

$$\nabla f(x_k + t''d_k)^\top d_k = c_1 \nabla f(x_k)^\top d_k > c_2 \nabla f(x_k)^\top d_k \quad ,$$

since $c_1 < c_2$ and $\nabla f(x_k)^\top d_k < 0$.

Proof of existence

This implies that t'' satisfies the Wolfe conditions and since $t'' < t'$, the inequalities in the 2 Wolfe conditions hold strictly.

By the smoothness assumption on f, there is an interval around t'' for which the Wolfe conditions hold.

Moreover, since $\nabla f(x_k + t'' d_k)^\top d_k$ (left-hand side in last equation) is negative, the strong Wolfe conditions hold in the same interval.

Proof of existence

This implies that t'' satisfies the Wolfe conditions and since $t'' < t'$, the inequalities in the 2 Wolfe conditions hold strictly.

By the smoothness assumption on f, there is an interval around t'' for which the Wolfe conditions hold.

Moreover, since $\nabla f(x_k + t'' d_k)^\top d_k$ (left-hand side in last equation) is negative, the strong Wolfe conditions hold in the same interval.

Take home message: One can always find a good step size for a smooth and bounded below function but it can take some time to find it...

Convergence of line search methods

Theorem (Zoutendijk)

Consider any iteration of the form $x_{k+1} = x_k + t_k d_k$*, where* d_k *is a descent direction* ($\cos\theta_k = -\frac{d_k^\top \nabla f(x_k)}{\|d_k\|\|\nabla f(x_k)\|} > 0$) *and* t_k *satisfies the Wolfe conditions.*

Suppose that f *is bounded below in* $\mathbb{R}^n$ *and that* f *is continuously differentiable in an open set* $\mathcal{N}$ *containing the level set* $\mathcal{L} = \{x : f(x) \leq f(x_0)\}$*, where* x_0 *is the starting point of the iteration.*

Assume also that the gradient ∇f *is Lipschitz continuous on* $\mathcal{N}$*, that is, there exists a constant* $L > 0$ *such that:*

$$\|\nabla f(x) - \nabla f(x')\| \leq L\|x - x'\|, \forall x, x' \in cN.$$

Then:

$$\sum_{k\geq 0}^{\infty} \cos^2\theta_k \|\nabla f(x_k)\|^2 < \infty$$

Proof of Zoutendijk's theorem

Wolfe's condition (second) implies:

$$(\nabla f(x_{k+1}) - \nabla f(x_k))^\top d_k \geq (c_2 - 1)\nabla f(x_k)^T d_k$$

Lipschitz condition implies:

$$(\nabla f(x_{k+1}) - \nabla f(x_k))^\top d_k \leq t_k L \|d_k\|^2$$

Combining the 2 we obtain:

$$t_k \geq \frac{c_2 - 1}{L} \frac{\nabla f(x_k)^\top d_k}{\|d_k\|^2}$$

Substituting this inequality into the first Wolfe condition we get:

$$f(x_{k+1}) \leq f(x_k) - c_1 \frac{1 - c_2}{L} \frac{(\nabla f(x_k)^\top d_k)^2}{\|d_k\|^2}$$

Proof of Zoutendijk's theorem

Which by the definion of θ_k is equivalent to:

$$f(x_{k+1}) \leq f(x_k) - c\|\nabla f(x_k)\|^2 \cos^2 \theta_k$$

where $c = c_1 \frac{1-c_2}{L}$.

Summing over k leads to:

$$f(x_{k+1}) \leq f(x_0) - c \sum_{k=0}^{k} \|\nabla f(x_k)\|^2 \cos^2 \theta_k$$

And since f is bounded below leads to:

$$\sum_{k=0}^{\infty} \|\nabla f(x_k)\|^2 \cos^2 \theta_k < \infty$$

Consequence of Zoutendijk's theorem

A direct consequence is that:

$$\|\nabla f(x_k)\|^2 \cos^2 \theta_k \to 0$$

So if θ_k is never too close to 90°:

$$\exists \delta > 0 \text{ s.t. } \cos \theta_k \geq \delta$$

Then x_k converges to a stationary point:

$$\|\nabla f(x_k)\| \to 0$$

Consequence of Zoutendijk's theorem

A direct consequence is that:

$$\|\nabla f(x_k)\|^2 \cos^2 \theta_k \rightarrow 0$$

So if θ_k is never too close to $90°$:

$$\exists \delta > 0 \text{ s.t. } \cos \theta_k \geq \delta$$

Then x_k converges to a stationary point:

$$\|\nabla f(x_k)\| \rightarrow 0$$

Take home message: $\|\nabla f(x_k)\|$ converges to zero, provided that search directions are never too close to orthogonality with gradient. So gradient descent with line search using Wolfe's conditions always converges to a stationary point ! (no need convexity but Lipschitz gradient)

Table of Contents

Reducing security interval

First search for starting interval or first value of t ($\alpha = 0$).

1. If t is Ok then stop
2. If t is too big then set $\beta = t$ and ok.
3. If t is too small, then set t to ct with $c > 1$ and back to 1.

Reducing the interval

Multiple strategies

1. Dichotomy. Try $t = (\alpha + \beta)/2$ and then work with $[\alpha, t]$ or $[t, \beta]$
2. Polynomial approximation of ϕ, e.g., cubic approximation.

Cubic approximation

Cubic approximation is compatible with Wolfe's method which also needs ϕ'. Take 2 values t_0 and t_1 (for example α and β). Define the third order polynomial p such that:

- $p(t_0) = \phi(t_0)$
- $p(t_1) = \phi(t_1)$
- $p'(t_0) = \phi'(t_0)$
- $p'(t_1) = \phi'(t_1)$

Then propose for t the minimum of the polynomial. If it does not provide a valid t you can fallback to dichotomy.

$\rightarrow$ Demo on notebook

References

- Wright and Nocedal, Numerical Optimization, 1999, Springer, Chapter 3.