Convergence of the Iterates of Descent Methods for Analytic Cost Functions.

R. Mahony
Department of Engineering,
Australian National University, 0200,
Australia.

Work presented coauthored with
Pierre-Antoine Absil and Ben Andrews.

Unconstrained Optimization Problem

Cost function

$$\phi : \mathbb{R}^n \longrightarrow \mathbb{R}$$

For most of the classical convergence proofs for numerical descent methods there are very few constraints placed on the nature of the cost.

Mostly a cost function comes with significant structure.

Typical properties are smoothness. Sometimes convexity and non-degeneracy.

Optimization problem is to compute

$$\arg\min_{x \in \mathbb{R}^n} \phi(x)$$

Descent Algorithms

I will consider three paradigms of descent algorithms

1. Continuous-time descent flows

2. Line-search descent methods

3. Trust region methods

Rather than enter into the technical details of the convergence proofs, I will try and provide an overview flavour of the classical and recent convergence results for these methods and show how the work I am presenting fits into the scheme.

Continuous-time descent flows

Descent Flows:

1. Compute a C_0 vector field $V(x)$ on $\mathbb{R}^n$ with the descent property

   $$d\phi V(x) \leq -\epsilon |V(x)||d\phi| \qquad \text{Angle Condition}$$

   for $\epsilon > 0$ a small positive constant.

2. Compute the solution to the ODE

   $$\dot{x}(t) = V(x(t)), \quad x(0) = x_0.$$

Line Search Descent Methods

For each iterate:

1. Compute a descent direction v_k

   $$\langle \nabla\phi(x_k), v_k\rangle \leq -\epsilon|v_k||\nabla\phi(x_k)| \qquad \text{Angle Condition}$$

   based on the best available information. (usually first or second order local derivatives).

2. Compute a step length α_k that will ensure decrease of the cost.

3. Compute the new iterate

   $$x_{k+1} = x_k + \alpha_k v_k$$

Trust Region Methods

For each iterate:

1. Compute a local quadratic model of the cost at the present iterate.

$$m_k(p) = \phi(x_k) + d\phi(x_k) + \frac{1}{2}p^T B_k p$$

where $p = x_{k+1} - x_k$.

2. Compute a trust region $\Delta_k > 0$

$$T_k = \{x_{k+1} \,|\, ||x_{k+1} - x_k|| \leq \Delta_k\}$$

3. Compute a new iterate $x_{k+1} \in T_k$ that ensures sufficient decrease of the of the model $m_k(x_{k+1} - x_k)$.

Continuous time flows
ω-limit set convergence.

- Cost ϕ has compact sub-level sets
- Cost ϕ and vector field $V(x)$ are sufficiently smooth.

Then the ω-limit set of $x(t)$ is a C^1 compact set in $\mathbb{R}^n$ on which

$$d\phi(x) = 0$$

holds.

Non-convergence of descent methods.

The potential problems with convergence of descent methods was understood as early as numerical methods were formalised.

Curry 1944 gave the following counter example.

> Let $G(x, y) = 0$ on the unit circle and $G(x, y) > 0$ elsewhere. Outside the unit circle let the surface have a spiral gully making infinitely many turns about the circle. The path C will evidently follow the gully and have all points of the circle as limit points of a sequence.

Weak convergence of numerical methods

Classical convergence results were proved in the sixties and seventies for line-search descent methods and the eighties for trust region methods.

- Line search descent methods:

1. For line-search descent algorithms with angle condition and sufficient decrease conditions on step-selection

$$\lim_{k \to \infty} ||\nabla \phi(x_k)|| = 0$$

2. For conjugate gradient methods with sufficient decrease conditions on step-selection

$$\lim \inf_{k \to \infty} ||\nabla \phi(x_k)|| = 0$$

Trust Region methods.

1: For relative decrease of function with respect to model strictly positive. That is

$$\rho_k = \frac{\phi(x_k) - \phi(x_{k+1})}{m_k(0) - m(x_{k+1} - x_k)} > 0$$

and a sufficient decrease condition

$$m(0) - m_k(x_{k+1} - x_k) \geq \eta_3 |\nabla \phi(x_k)| \min \left(\Delta_k, \frac{|\nabla \phi(x_k)|}{||B_k||} \right)$$

Then

$$\lim \inf_{k \to \infty} ||\nabla \phi(x_k)|| = 0$$

2: For relative decrease bounded away from zero $\rho_k \geq \eta > 0$ and a sufficient decrease condition. Then

$$\lim_{k \to \infty} ||\nabla \phi(x_k)|| = 0$$

Observations line search algorithms

1. All line search methods ensure some form of descent condition

$\phi(x_{k+1}) - \phi(x_k) \leq \eta_1 \langle \nabla\phi(x_k), v_k \rangle,$ Descent (Armijo) Condition

$\langle \nabla\phi(x_{k+1}), v_k \rangle \leq \eta_2 \langle \nabla\phi(x_k), v_k \rangle,$ Curvature (Wolfe) Condition

2. Most line search convergence results require some sort of angle condition

$\langle \nabla\phi(x_k), v_k \rangle \leq -\epsilon |v_k| |\nabla\phi(x_k)|$ Angle Condition

3. Best convergence that is obtained is

$$\lim ||\nabla\phi(x_k)|| = 0$$

Observations trust region methods

1. All Trust region methods apply some sort of model decrease condition

$$m(0) - m_k(x_{k+1} - x_k) \geq \eta_3 |\nabla \phi(x_k)| \min \left(\Delta_k, \frac{|\nabla \phi(x_k)|}{||B_k||} \right)$$

Based on minimum decrease obtained by the Cauchy point

2. All Trust region methods apply some sort of relative measure of decrease

$$\rho_k = \frac{f(x_k) - f(x_{k+1})}{m(x_{k+1} - x_k) - m(0)} \geq \eta \geq 0$$

3. Best convergence that is obtained is

$$\lim ||\nabla \phi(x_k)|| = 0$$

What more is possible

- Numerically it is better to know that an algorithm will converge to a single point rather than the weak $\lim |\nabla\phi| = 0$ condition.

- In practice, descent algorithms do converge to single points. Counter examples are very few and far between.

- Most cost function have more structure than we have assumed above.

The standard convergence results are based on the weakest set of conditions on cost functions that are encountered in practice.

Single-limit-point convergence results for descent flows.

- Cost ϕ has compact sub-level sets

- Cost ϕ is twice differentiable.

- On locally minimizing sets the function is Morse-Bott.

  That is the Hessian of the function $D^2\phi$ is non-degenerate on the normal space to the level set of the locally minimizing set N,

  $$D^2\phi_x\Big|_{N_x^\perp} > 0$$

Then $x(t) \to x_\infty$.

Single-limit-point Convergence results for numerical methods

- Strong convexity of ϕ implies single-limit-point convergence to global minima

1. Byrd and Nocedal (1989) for the BFGS algorithm with bounds on the condition number of B_k.

2. Burachik *et al.* (1995) for the steepest descent method.

3. Kiwiel and Murty (1996) for the steepest descent method for quasi-convex cost functions.

4. Iusem (2002) for the projected gradient algorithm.

- Hessian $D^2\phi$ is positive definite at an accumulation point implies single-limit-point convergence to the accumulation point

  1. Classical for Newton and quasi-Newton methods.
  2. Moré and Sorensen (1983) for approximate trust-region methods (using nearly exact update steps).
  3. Conn *et al.* (1993) show the same result holds for a class of trust-region methods that ensure a fraction of Cauchy decrease.

- Local minimum of ϕ is isolated implies that there exists a small basin of attraction around the point for which one obtains single-limit-convergence.

  1. Bersekes (1995) for a class of line-search descent methods.
  2. Dunn (1983, 1987) for a class of line-search descent along with some additional growth conditions on the function ϕ.

What is the problem with these conditions

1. Hessian conditions and local isolation of critical points require a significant amount of *a-priori* local knowledge about the level sets of the cost.

2. Convexity requirements are strong conditions to require of cost functions.

The goal of this presentation is to show how a result from the study of analytic varieties can be applied to provide a simple additional condition on the cost

$$\phi(x) \quad \text{is analytic}$$

that does not suffer from the first point, and provides a similarly wide class of applications as the second point.

Łojasiewicz's inequality

Due originally to Łojasiewicz's 1965 in order to characterise the nature of level sets of analytic functions (analytic varieties).

Lemma 1 *Let* ϕ *be a real analytic function on a neighbourhood of* x^* *in* $\mathbb{R}^n$ *such that* $\nabla\phi(x_*) = 0$*. Then there are constants* $c > 0$ *and* $\mu \in (0, 1)$ *such that*

$$\|\nabla\phi(x)\| \geq c|\phi(x) - \phi(x^*)|^\mu \tag{1}$$

in some neighbourhood U *of* x^**.*

Łojasiewicz theorem background

In 1984 (20 years after the original result) Łojasiewicz proved a corollary to his lemma relating to the convergence of gradient descent flows of analytic functions. It was not his main research focus and was published in an obscure workshop in Italy, published in French.

It was picked up by some people working in dynamical evolution of surfaces in the early nineties and used to show convergence of pinching and separation behaviour of surfaces under curvature flows.

The importance of the result for its own sake was recognised in the late nineties by the analytic geometry community and there are now a dozen works where the result is proved in all sorts of manners.

One of the more focused expositions is Kurdyka, Mostowski and Parusinski 2000.

Łojasiewicz theorem

Let ϕ be a real analytic function and let $x(t)$ be a C^1 curve in $\mathbb{R}^n$, with $\dot{x}(t) = \frac{dx}{dt}(t)$ denoting its time derivative. Assume that there exists a $\delta > 0$ and a real τ such that for $t > \tau$, $x(t)$ satisfies the angle condition

$$\frac{d\phi(x(t))}{dt} \equiv \langle \nabla\phi(x(t)), \dot{x}(t) \rangle \leq -\delta \|\nabla\phi(x(t))\| \|\dot{x}(t)\| \tag{2}$$

and a weak decrease condition

$$\left[\frac{d}{dt}\phi(x(t)) = 0\right] \Rightarrow [\dot{x}(t) = 0]\,. \tag{3}$$

Then, either $\lim_{t\to+\infty} \|x(t)\| = \infty$, or there exits $x^* \in \mathbb{R}^n$ such that $\lim_{t\to+\infty} = x^*$.

This theorem is deliberately phrased in a similar manner to a Lyapunov stability result.

Proof

Assume that $\|x(t)\| \not\to +\infty$.

Then $x(t)$ has an accumulation point x^* in $\mathbb{R}^n$.

It follows from (2) that $\phi(x(t))$ is non-increasing.

$$\phi(x(t)) \downarrow \phi(x^*).$$

Case (i)

There exists a $t_1 > \tau$ such that $\phi(x(t_1)) = \phi(x^*)$.

It follows that

$$\phi(x(t)) = \phi(x^*), \qquad \frac{d}{dt}\phi(x(t)) = 0$$

for all $t \geq t_1$.

The weak decrease condition ensures $x(t) = x^*$ for all $t \geq t_1$.

The weak decrease condition prevents endless wandering in the critical set.

Case (ii)

Assume without loss of generality that $\phi(x^*) = 0$.

Łojasiewicz's inequality implies

$$\frac{d\phi(x(t))}{dt} \leq -\delta\|\nabla\phi(x(t))\|\|\dot{x}(t)\| \leq -\delta c|\phi(x(t))|^{\mu}\|\dot{x}(t)\| \tag{4}$$

holds in a neighbourhood of x^*.

Thus,

$$c_1\frac{d(\phi(x(t)))^{1-\mu}}{dt} \leq -\|\dot{x}(t)\| \tag{5}$$

For $t_1 < t_2$ integrate this differential equation

$$L_{12} := \int_{t_1}^{t_2} \|\dot{x}(t)\| dt \le c_1((\phi(x(t_1)))^{1-\mu} - (\phi(x(t_2)))^{1-\mu}). \qquad (6)$$

Evaluate the limit as $t_2 \to \infty$

$$L_{1\infty} \le c_1(\phi(x(t_1)))^{1-\mu}$$

But $L_{1\infty}$ is the length of the path $x(t)$ from t_1 to infinite time. If this length is finite then it follows that

$$\lim x(t) = x_\infty = x_*$$

Mexican Hat counter example

Consider the cost

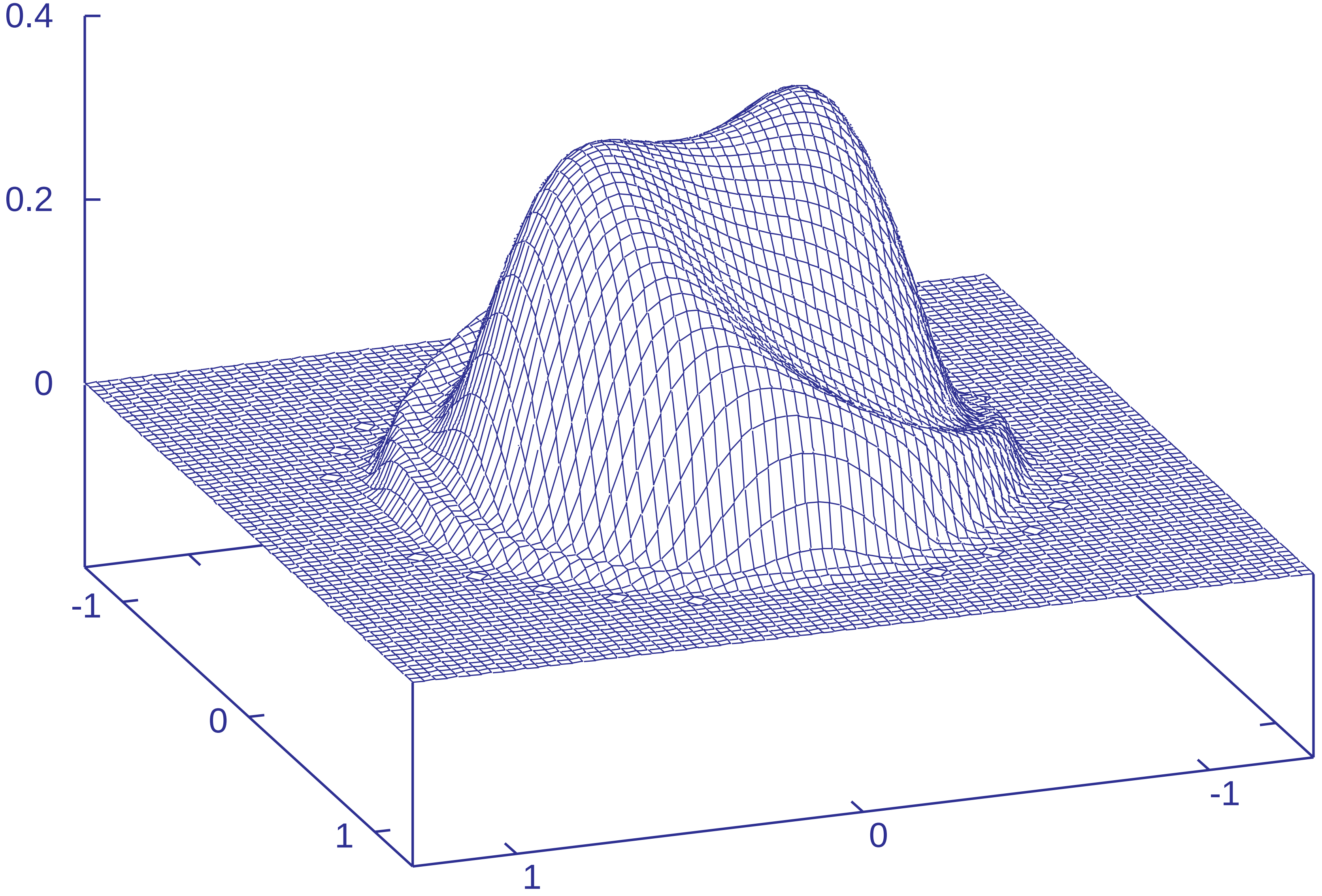
$$f(r,\theta) := \begin{cases} e^{-\frac{1}{1-r^2}}\left[1 - \frac{4r^4}{4r^4+(1-r^2)^4}\sin\left(\theta - \frac{1}{1-r^2}\right)\right] & \text{if } r < 1, \\ 0 & \text{if } r \geq 1, \end{cases} \tag{7}$$

where (r,θ) denote polar coordinates in $\mathbb{R}^2$.

The solution $(r(t),\theta(t))$ of the gradient descent flow (expressed in polar coordinates) satisfies

$$\theta(t) = \frac{1}{1-r(t)^2}. \tag{8}$$

Mexican Hat



Strong Descent Conditions

Strong sufficient-decrease condition:

$$\phi(x_k) - \phi(x_{k+1}) \geq \sigma \|\nabla \phi(x_k)\| \|x_{k+1} - x_k\| \tag{9}$$

for all k and for some $\sigma > 0$.

This condition is satisfied under Armijo's condition along with an angle condition. It also accommodates the framework of trust-region methods.

Weak decrease condition:

$$\left[\phi(x_{k+1}) = \phi(x_k)\right] \Rightarrow \left[x_{k+1} = x_k\right] \tag{10}$$

Together, these two conditions are termed the Strong descent conditions.

Single point convergence of numerical line descent methods

Let $\phi : \mathbb{R}^n \mapsto \mathbb{R}$ be an analytic cost function. Let the sequence $\{x_k\}_{k=1,2,\dots}$ satisfy the strong descent conditions. Then, either

$$\lim_{k\to\infty} \|x_k\| = +\infty,$$

or there exists a single point $x^* \in \mathbb{R}^n$ such that

$$\lim_{k\to\infty} x_k = x^*.$$

The proof is a technical adaptation of proof for continuous-case.

The key difference from classical proofs is that we use a total bound on the length of the path rather than a local bound on update steps close to the accumulation point.

Trust region results

If the Cauchy decrease condition hold (standard condition)

$$m(0) - m_k(x_{k+1} - x_k) \geq \eta_3 |\nabla\phi(x_k)| \min\left(\Delta_k, \frac{|\nabla\phi(x_k)|}{||B_k||}\right)$$

and

$$B_k > 0, \qquad \mu(B_k) := ||B_k||_2 ||B_k^{-1}||_2 \leq \kappa_2$$

then the strong descent conditions hold. (Strong descent conditions subsume the usual relative descent conditions).

As a consequence either

$$\lim_{k\to\infty} \|x_k\| = +\infty,$$

or there exists a single point $x^* \in \mathbb{R}^n$ such that

$$\lim_{k\to\infty} x_k = x^*.$$

Conclusions

- Classical convergence results for descent algorithms make very weak assumptions on the cost.
- There is considerable benefit to be gained from looking at properties cost, however, it is difficult to characterise global properties of the cost function that lead to local convergence properties.
- Analyticity of the cost function is one of the few global properties of a cost function that has strong local implications.
- The convergence results presented provide a practical tool in numerical descent method analysis for a wide class of costs of considerable interest.

Work presented was coauthored by

Pierre-Antoine Absil

School of Computational Science and Information Technology,
Florida State University,
Tallahassee FL 32306-4120, USA.

Ben Andrews

Center for Mathematical Analysis,
Institute of Advanced Studies,
Australian National University,
ACT, 0200, Australia