6th Traditional Summer School for Young Researchers: "Control. Information. Optimization"

Giuseppe Carlo Calafiore

Dipartimento di Automatica e Informatica
Politecnico di Torino – ITALY

Moscow, June 22-29

LECTURE 2

First-order and gradient algorithms

When I consider what people generally want in calculating, I found that it always is a number. I also observed that every number is composed of units, and that any number may be divided into units. Moreover, I found that every number which may be expressed from one to ten, surpasses the preceding by one unit: afterwards the ten is doubled or tripled just as before the units were: thus arise twenty, thirty, etc. until a hundred: then the hundred is doubled and tripled in the same manner as the units and the tens, up to a thousand; so forth to the utmost limit of numeration.

Muhammad ibn Musa al-Khwarizmi, (780-850 CE)

Outline

Introduction

- We illustrate first-order iterative techniques (algorithms) for *solving numerically*, up to a given accuracy, some class of unconstrained minimization problems.
- These methods share a common general structure: A at each iteration $k = 0, 1, \ldots$, the candidate point is *updated* to a new point x_{k+1}. Then, a *stopping criterion* is checked. If yes, then the current point is returned as a numerical solution (to accuracy ϵ) of the problem; otherwise we set $k \leftarrow k + 1$, and iterate the process.
- A typical update rule takes the form of a simple recursion

$$x_{k+1} = x_k + s_k v_k, \tag{1}$$

where the scalar $s_k > 0$ is called the *stepsize*, and $v_k \in \mathbb{R}^n$ is the update (or search) *direction*.

- The meaning of the above update rule is that from the current point x_k we move away along direction v_k, and the length of the move is dictated by the stepsize s_k.

Working hypotheses

- In the rest of this lecture, we shall make the standing assumption that $f_0 : \mathbb{R}^n \to \mathbb{R}$ is a *closed* function, that is all the sublevel sets $S_\alpha = \{x : f_0(x) \leq \alpha\}$, $\alpha \in \mathbb{R}$, are closed sets.
- Further, we assume that f_0 is bounded below, and that it attains its (global) minimum value f_0^* at some point $x^* \in \mathrm{dom}\, f_0$.
- Given a point $x_0 \in \mathrm{dom}\, f_0$, we define S_0 as the sublevel set

$$S_0 \doteq \{x : \ f_0(x) \leq f_0(x_0)\}.$$

Technical Preliminaries

Technical Preliminaries

Gradient Lipschitz continuity.

- A function $f_0 : \mathbb{R}^n \to \mathbb{R}$ is said to be Lipschitz continuous on a domain $S \subseteq \mathbb{R}^n$, if there exist a constant $R > 0$ (possibly depending on S) such that

$$|f_0(x) - f_0(y)| \leq R\|x - y\|_2, \quad \forall x, y \in S.$$

- A differentiable function $f_0 : \mathbb{R}^n \to \mathbb{R}$ is said to have Lipschitz continuous gradient on S if there exist a constant $L > 0$ (possibly depending on S) such that

$$\|\nabla f_0(x) - \nabla f_0(y)\|_2 \leq L\|x - y\|_2, \quad \forall x, y \in S. \tag{2}$$

- Intuitively, f_0 has Lipschitz continuous gradient if its gradient "does not vary too fast." Indeed, if f_0 is twice differentiable, the above condition is equivalent to a bound on the Hessian of f_0:

$$\nabla^2 f(x) \preceq LI, \quad \forall x \in S.$$

Technical Preliminaries

The following proposition summarizes some useful implications of gradient Lipschitz continuity:

1. If $f_0 : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable, then (2) holds if and only if f_0 has bounded Hessian on S, that is

$$\|\nabla^2 f_0(x)\|_{\mathrm{F}} \leq L, \quad \forall x \in S.$$

2. If f_0 is continuously differentiable, then (2) implies that

$$|f_0(x) - f_0(y) - \nabla f_0(y)^\top (x - y)| \leq \frac{L}{2}\|x - y\|_2^2, \quad \forall x, y \in S. \tag{3}$$

3. If f_0 is continuously differentiable and *convex*, then (2) implies that

$$0 \leq f_0(x) - f_0(y) - \nabla f_0(y)^\top (x - y) \leq \frac{L}{2}\|x - y\|_2^2, \quad \forall x, y \in S,$$

and that the following inequality holds $\forall x, y \in S$:

$$\frac{1}{L}\|\nabla f_0(x) - \nabla f_0(y)\|_2^2 \leq (\nabla f_0(x) - \nabla f_0(y))^\top (x - y) \leq L\|x - y\|_2^2.$$

Technical Preliminaries

Quadratic upper bound.

Inequality (3) implies that, for any given $y \in S$, $f_0(x)$ is upper bounded by a (strongly) convex quadratic function:

$$f_0(x) \leq f_0(y) + \nabla f_0(y)^\top (x - y) + \frac{L}{2}\|x - y\|_2^2, \quad \forall x, y \in S, \tag{4}$$

where the quadratic upper bound function is defined as

$$f_{\text{up}}(x) \doteq f_0(y) + \nabla f_0(y)^\top (x - y) + \frac{L}{2}\|x - y\|_2^2. \tag{5}$$

Technical Preliminaries

Implications on the unconstrained minimum.

- Let $x^* \in \operatorname{dom} f_0$ be a global unconstrained minimizer of f_0. Then,

$$x^* \in \arg\min_{x \in S_0} f_0(x),$$

and, if f_0 is differentiable, the unconstrained optimality condition requires that $\nabla f_0(x^*) = 0$.

- If further f_0 has Lipschitz continuous gradient on S_0, then it holds that

$$\frac{1}{2L}\|\nabla f_0(x)\|_2^2 \leq f_0(x) - f_0^* \leq \frac{L}{2}\|x - x^*\|_2^2, \quad \forall x \in S_0. \tag{6}$$

- The bound on the right in (6) is readily obtained by evaluating (4) at $y = x^*$, and recalling that $\nabla f_0(x^*) = 0$. The bound on the left in (6) is instead obtained by first evaluating (4) at $x = y - \frac{1}{L}\nabla f_0(y)$, which yields

$$f_0(x) \leq f_0(y) - \frac{1}{2L}\|\nabla f_0(y)\|_2^2, \quad \forall y \in S_0.$$

Then, since $f_0^* \leq f_0(x)$, $\forall x \in \operatorname{dom} f_0$, this last inequality also implies $f_0^* \leq f_0(y) - \frac{1}{2L}\|\nabla f_0(y)\|_2^2$, $\forall y \in S_0$, which is the desired bound.

Technical Preliminaries

Lipschitz constant for functions with compact sublevel sets.

- If f_0 is twice continuously differentiable, and the sublevel set $S_0 = \{x : f_0(x) \leq f_0(x_0)\}$ is compact, then f_0 has Lipschitz continuous gradient on S_0.
- This is due to the fact that the Hessian is continuous, hence $\|\nabla^2 f_0(x)\|_{\mathrm{F}}$ is continuous and, from the Weierstrass theorem, it attains a maximum over the compact set S_0. Therefore, we have that f_0 has Lipschitz continuous gradient on S_0, and a suitable Lipschitz constant is

$$L = \max_{x \in S_0} \|\nabla^2 f_0(x)\|_{\mathrm{F}}.$$

- Compactness of the sublevel set S_0 is guaranteed, for instance, if f_0 is coercive, or when f_0 is strongly convex.

Technical Preliminaries

Strong convexity and its implications.

- A function $f_0 : \mathbb{R}^n \to \mathbb{R}$ is said to be strongly convex if there exist $m > 0$ such that

$$f_0(x) - \frac{m}{2}\|x\|_2^2$$

is convex.

- From this definition it also follows that if f_0 is strongly convex and twice differentiable, then

$$\nabla f_0(x) \succeq mI, \quad \forall x \in \mathrm{dom}\, f_0.$$

- We know that a differentiable function f is convex if and only if

$$\forall x, y \in \mathrm{dom}\, f,\ f(y) \geq f(x) + \nabla f(x)^\top (y - x), \tag{7}$$

which means that the linear function $f(x) + \nabla f(x)^\top (y - x)$ is a global lower bound on $f(y)$.

Technical Preliminaries

Strong convexity and its implications (quadratic lower bound).

- Applying (7) to $f(x) = f_0(x) - \frac{m}{2}\|x\|_2^2$, we have that a differentiable f_0 is strongly convex if and only if

$$\forall x, y \in \mathrm{dom}\, f_0,\ f_0(y) \geq f_0(x) + \nabla f_0(x)^\top (y - x) + \frac{m}{2}\|y - x\|_2^2, \qquad (8)$$

- This means geometrically that at any $x \in \mathrm{dom}\, f_0$, there is a convex quadratic function

$$f_{\mathrm{low}}(y) \doteq f_0(x) + \nabla f_0(x)^\top (y - x) + \frac{m}{2}\|y - x\|_2^2$$

that bounds from below the graph of f_0, that is such that $f_0(y) \geq f_{\mathrm{low}}(y)$, for all $y \in \mathrm{dom}\, f_0$.

Technical Preliminaries

Strong convexity and its implications (quadratic upper bound). We show that if f_0 is strongly convex and twice continuously differentiable, then f_0 has Lipschitz continuous gradient over S_0, hence it can be upper bounded by a quadratic function.

- For any $x_0 \in \mathrm{dom}\, f_0$, the level set $S_0 = \{y : f_0(y) \leq f_0(x_0)\}$ is contained in a regular ellipsoid and hence it is bounded. To see this fact, consider the strong convexity inequality (8):

$$y \in S_0 \;\Rightarrow\; 0 \geq f_0(y) - f_0(x_0) \geq \nabla f_0(x_0)^\top (y - x_0) + \frac{m}{2}\|y - x_0\|_2^2,$$

  where the region of y satisfying the inequality $\nabla f_0(x_0)^\top (y - x_0) + \frac{m}{2}\|y - x_0\|_2^2 \leq 0$ is a bounded ellipsoid.
- $\nabla^2 f_0(x)$ continuous $\Rightarrow$ bounded over bounded regions, hence $\exists M > 0$ (possibly depending on x_0) such that $\nabla^2 f_0(x) \preceq MI, \quad \forall x \in S_0$.
- This implies that f_0 has Lipschitz continuous gradient over S_0 (with Lipschitz constant M), hence it admits a quadratic upper bound

$$f_0(x) \leq f_0(x) + \nabla f_0(x)^\top (y - x) + \frac{M}{2}\|y - x\|_2^2, \quad \forall x, y \in S_0.$$

Technical Preliminaries

Bounds on the optimality gap.

- For a strongly convex and twice differentiable function f_0 it holds that

$$mI \preceq \nabla^2 f_0(x) \preceq MI, \quad \forall x \in S_0,$$

and that f_0 is upper and lower bounded by two convex quadratic functions, as follows

$$f_{\text{low}}(y) \leq f_0(y), \quad \forall x, y \in \text{dom}\, f_0$$
$$f_0(y) \leq f_{\text{up}}(y), \quad \forall x, y \in S_0,$$

where

$$f_{\text{low}}(y) = f_0(x) + \nabla f_0(x)^\top (y - x) + \frac{m}{2}\|y - x\|_2^2, \tag{9}$$

$$f_{\text{up}}(y) = f_0(x) + \nabla f_0(x)^\top (y - x) + \frac{M}{2}\|y - x\|_2^2. \tag{10}$$

- From these two inequalities we can derive key bounds on the gap between the value of f_0 at any point $x \in \text{dom}\, f_0$ and the global unconstrained minimum value f_0^*.

Technical Preliminaries

Bounds on the optimality gap.

- Let the minimum f_0^* over $\mathrm{dom}\, f_0$ be attained at some $x^* \in \mathrm{dom}\, f_0$ (such minimizer is *unique*, since f_0 is strongly convex). It must be $x^* \in S_0$.
- Writing the inequality $f_{\mathrm{low}}(y) \leq f_0(y)$ for $x = x^*$ we obtain (since $\nabla f_0(x^*) = 0$)

$$f_0(y) \geq f_0^* + \frac{m}{2}\|y - x^*\|_2^2, \quad \forall y \in S_0. \tag{11}$$

- Further, $f_0(y) \geq f_{\mathrm{low}}(y)\ \forall y \in S_0$, implies that $f_0(y) \geq \min_z f_{\mathrm{low}}(z)$.
- Setting the gradient of $f_{\mathrm{low}}(z)$ to zero yields the minimizer $z^* = x - \frac{1}{m}\nabla f_0(x)$, hence

$$\begin{aligned} f_0(y) &\geq \min_z f_{\mathrm{low}}(z) = f_{\mathrm{low}}(z^*) \\ &= f_0(x) - \frac{1}{2m}\|\nabla f_0(x)\|_2^2, \quad \forall x, y \in S_0. \end{aligned}$$

- Since this last inequality holds for all y, it also holds for $y = x^*$, thus

$$f_0^* \geq f_0(x) - \frac{1}{2m}\|\nabla f_0(x)\|_2^2, \quad \forall x \in S_0. \tag{12}$$

Technical Preliminaries

Bounds on the optimality gap.

- Putting together (11) and (12), we obtain

$$\frac{m}{2}\|x - x^*\|_2^2 \leq f_0(x) - f_0^* \leq \frac{1}{2m}\|\nabla f_0(x)\|_2^2, \quad \forall x \in S_0, \tag{13}$$

- This shows that the gap $f_0(x) - f_0^*$ (as well as the distance from x to the minimizer x^*) is upper bounded by the norm of the gradient of f_0 at x.
- Also, we obtain the "swapped" inequality

$$\frac{1}{2M}\|\nabla f_0(x)\|_2^2 \leq f_0(x) - f_0^* \leq \frac{M}{2}\|x - x^*\|_2^2, \quad \forall x \in S_0. \tag{14}$$

First-order descent methods for unconstrained minimization

First-order descent methods

- We start by discussing a simple class of first-order methods where the iterates are of the form (1), and the search direction v_k is computed on the basis of the gradient of f_0 at x_k.
- Consider a point $x_k \in \operatorname{dom} f_0$ and a direction $v_k \in \mathbb{R}^n$. Using the first-order Taylor series expansion for f_0, we have that

$$f_0(x_k + sv_k) \simeq f_0(x_k) + s\nabla f_0(x_k)^\top v_k, \quad \text{for } s \to 0.$$

- The local rate of variation of f_0, in the neighborhood of x_k and along direction v_k, is thus given by

$$\delta_k \doteq \lim_{s \to 0} \frac{f_0(x_k + sv_k) - f_0(x_k)}{s} = \nabla f_0(x_k)^\top v_k.$$

First-order descent methods

Descent directions.

- δ_k is positive whenever $\nabla f_0(x_k)^\top v_k > 0$, that is for directions v_k that form a positive inner product with the gradient $\nabla f_0(x_k)$.
- Contrary, directions v_k for which $\nabla f_0(x_k)^\top v_k < 0$ are called decrease (or *descent*) directions. The reason for this is that if the new point x_{k+1} is chosen according to (1) as $x_{k+1} = x_k + sv_k$, then

$$f_0(x_{k+1}) < f_0(x_k), \quad \text{for sufficiently small } s > 0.$$

- From the Cauchy–Schwartz inequality we have that

$$-\|\nabla f_0(x_k)\|_2 \|v_k\|_2 \leq \nabla f_0(x_k)^\top v_k \leq \|\nabla f_0(x_k)\|_2 \|v_k\|_2,$$

hence δ_k is minimal over all v_k with $\|v_k\|_2 = 1$, for

$$v_k = -\frac{\nabla f_0(x_k)}{\|\nabla f_0(x_k)\|_2}, \tag{15}$$

i.e., when v_k points in the direction of the negative gradient.

- The direction v_k in (15) is thus called the *steepest descent* direction, with respect to the standard Euclidean norm.

First-order descent methods

A descent scheme.

Require: $f_0 : \mathbb{R}^n \to \mathbb{R}$ differentiable, $x_0 \in \operatorname{dom} f_0$, $\epsilon > 0$

1: Set $k = 0$
2: Determine a descent direction v_k
3: Determine step length $s_k > 0$
4: Update: $x_{k+1} = x_k + s_k v_k$
5: If accuracy ϵ is attained, exit and return x_k, else let $k \leftarrow k + 1$ and goto 2.

First-order descent methods

Stepsize selection.

- Consider the restriction of f_0 along the direction v_k:

$$\phi(s) \doteq f_0(x_k + sv_k), \quad s \geq 0.$$

  Clearly, ϕ is a function of the scalar variable s, and $\phi(0) = f_0(x_k)$. Choosing a suitable stepsize amounts to finding $s > 0$ such that $\phi(s) < \phi(0)$.

- A natural approach would then be to compute s that *minimizes* ϕ, that is

$$s^* = \arg\min_{s \geq 0} \phi(s).$$

  This method is called *exact line search*, and provides a stepsize s^* with the best possible decrease.

- However, finding s^* requires solving a univariate (and generically non-convex) optimization problem, which may be computationally demanding. For this reason, exact line search is rarely used in practical algorithms.

- A more practical alternative consists in searching for an s value guaranteeing a sufficient *rate* of decrease in ϕ.

First-order descent methods

Stepsize selection.

- Consider the tangent line to ϕ at 0:

$$\phi(s) \simeq \ell(s) \doteq \phi(0) + s\delta_k, \quad \delta_k \doteq \nabla f_0(x_k)^\top v_k < 0, \; s \geq 0.$$

- ℓ is a linear function with negative slope δ_k. Now, for $\alpha \in (0, 1)$, it holds that the line

$$\bar{\ell}(s) \doteq \phi(0) + s(\alpha\delta_k), \quad s \geq 0$$

lies above $\ell(s)$, hence it also lies above $\phi(s)$, at least for small $s > 0$.

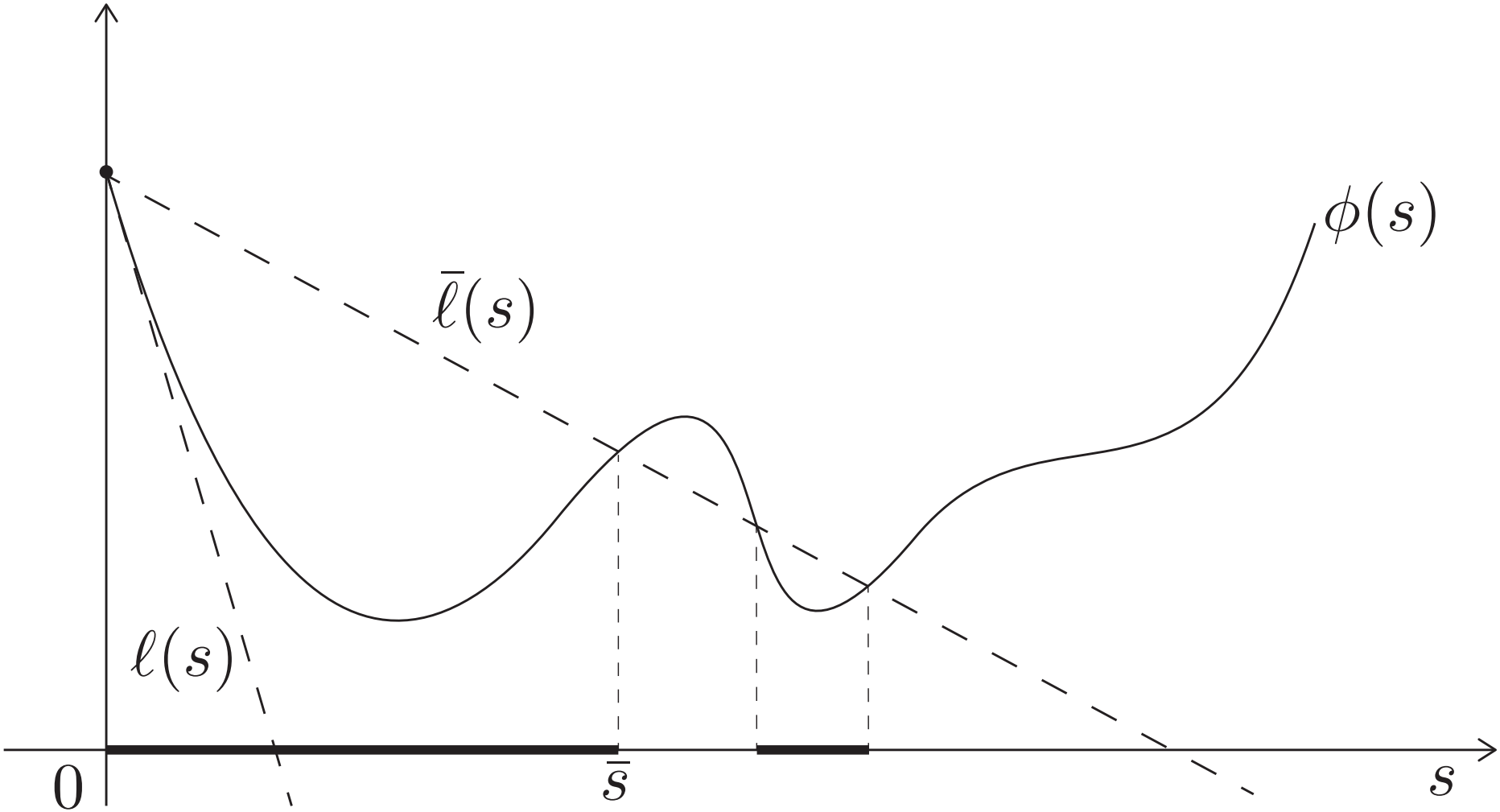
- Since ϕ is bounded below while $\bar{\ell}$ is unbounded below, there must exist a point where $\phi(s)$ and $\bar{\ell}(s)$ cross; let $\bar{s}$ be the smallest of such points. All values of s for which $\phi(s) \leq \bar{\ell}(s)$ provide a sufficient rate of decrease, given by the slope $\alpha\delta_k$ of $\bar{\ell}$.

- This rate condition is known as the *Armijo condition*, stating that the valid stepsizes must satisfy

$$\phi(s) \leq \phi(0) + s(\alpha\delta_k)$$

First-order descent methods

Stepsize selection.



First-order descent methods

Backtracking.

- The Armijo condition is clearly satisfied by all $s \in (0, \bar{s})$, hence this condition alone is still not sufficient to insure that the stepsize is not chosen too small (which is necessary for convergence of the method).
- A usual practice amounts to employing a so-called *backtracking* approach, whereby an initial value of s is fixed to some constant value s_{init} (typically, $s_{\text{init}} = 1$), and then the value of s is iteratively decreased at a fixed rate $\beta \in (0, 1)$, until the Armijo condition is met.

Require: f_0 differentiable, $\alpha \in (0, 1)$, $\beta \in (0, 1)$, $x_k \in \operatorname{dom} f_0$, v_k a descent direction, s_{init} a positive constant (typically, $s_{\text{init}} = 1$)

1: Set $s = s_{\text{init}}$, $\delta_k = \nabla f_0(x_k)^\top v_k$
2: If $f_0(x_k + sv_k) \leq f(x_k) + s\alpha\delta_k$, then return $s_k = s$
3: Else let $s \leftarrow \beta s$ and goto 2

The Gradient Method

The gradient method

- We analyze more closely the convergence properties of the descent scheme for the most common case where the descent direction is simply the negative gradient (that is, the direction of steepest local descent). We take henceforth

$$v_k = -\nabla f_0(x_k).$$

- Very little can actually be said about the properties of the gradient descent algorithm, unless we make some additional assumptions on the regularity of the objective function.

- More precisely, we shall assume that f_0 has Lipschitz continuous gradient on S_0, that is, there exist a positive constant L such that

$$\|\nabla f_0(x) - \nabla f_0(y)\|_2 \leq L\|x - y\|_2, \quad \forall x, y \in S_0.$$

The gradient method

Lower bound on stepsize.

- Let x_k be the current point in a gradient descent algorithm, and let

$$x = x_k - s\nabla f_0(x_k).$$

- Evaluating f_0 and $f_{\rm up}$ in (5) at x we obtain the restrictions of these functions along the direction $v_k = -\nabla f_0(x_k)$:

$$\begin{aligned} \phi(s) &= f_0(x_k - s\nabla f_0(x_k)) \\ \phi_{\rm up}(s) &= f_0(x_k) - s\|\nabla f_0(x_k)\|_2^2 + s^2\frac{L}{2}\|\nabla f_0(x_k)\|_2^2 \end{aligned}$$

  where (4) clearly implies that

$$\phi(s) \leq \phi_{\rm up}(s), \quad s \geq 0.$$

- $\phi(s)$ and $\phi_{\rm up}(s)$ have the same tangent line at $s = 0$, which is given by

$$\ell(s) \doteq f_0(x_k) - s\|\nabla f_0(x_k)\|_2^2.$$

The gradient method

Lower bound on stepsize.

- Then, the line

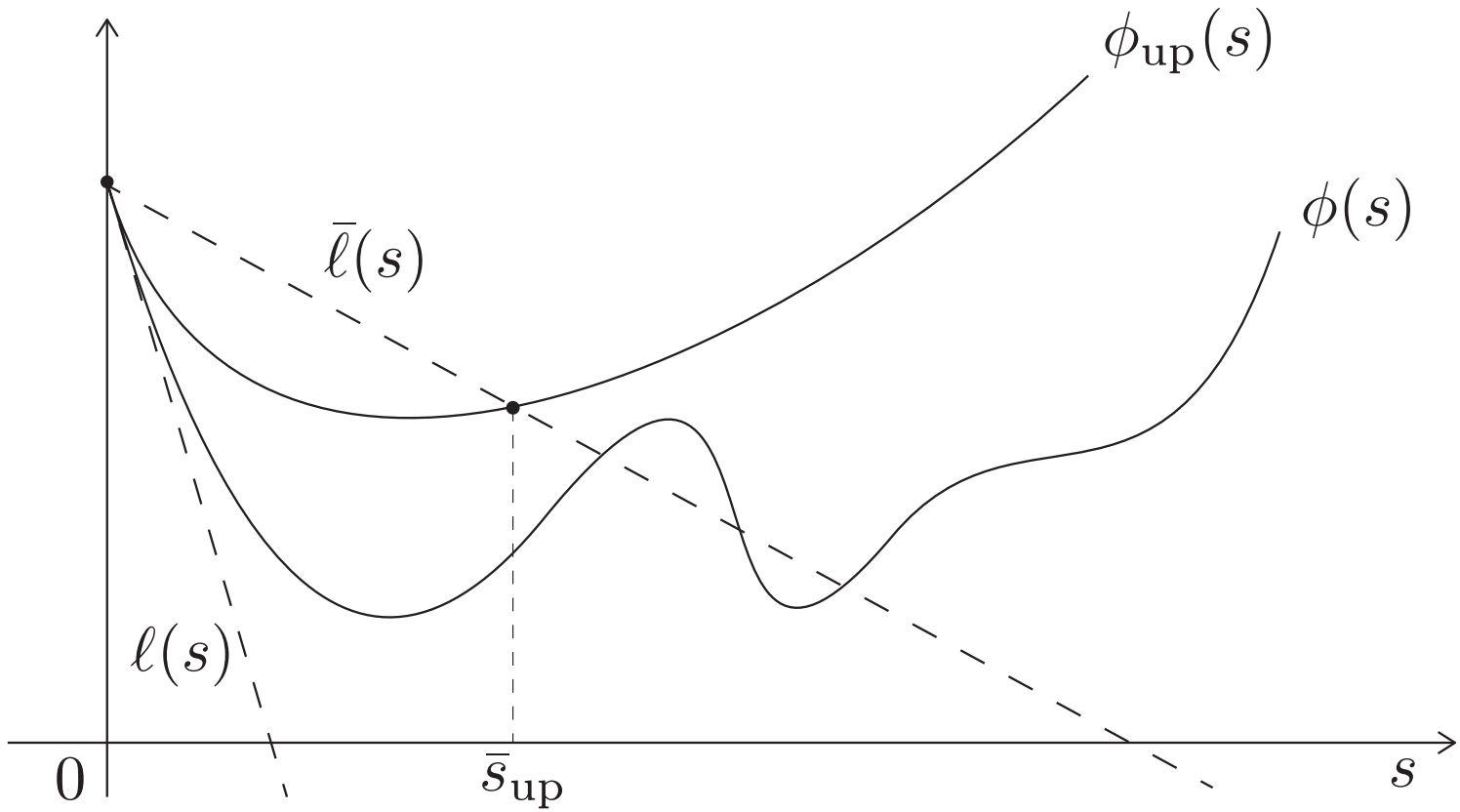
$$\bar{\ell}(s) \doteq f_0(x_k) - s\,\alpha\|\nabla f_0(x_k)\|_2^2, \quad \alpha \in (0, 1).$$

intercepts the upper bound function $\phi_{\rm up}(s)$ at the point

$$\bar{s}_{\rm up} = \frac{2}{L}(1-\alpha). \tag{16}$$

- It is then clear that a *constant* stepsize $s_k = \bar{s}_{\rm up}$ would satisfy the Armijo condition at each iteration of the algorithm.
- Note, however, that one would need to know the numerical value of the Lipschitz constant L (or an upper bound on it), in order to implement such a stepsize in practice. This can be avoided by using backtracking.

Lower bound on stepsize



For a function with Lipschitz continuous gradient there is a constant stepsize $s = \bar{s}_{\text{up}}$ that satisfies the Armijio condition.

The gradient method

Lower bound on stepsize.

- Suppose we initialize the backtracking procedure with $s = s_{\text{init}}$. Then, either this initial value satisfies the Armijo condition, or its is iteratively reduced until it does.
- The iterative reduction certainly stops at a value $s \geq \beta \bar{s}_{\text{up}}$, hence backtracking guarantees that

$$s_k \geq \min(s_{\text{init}}, \beta \bar{s}_{\text{up}}) \doteq s_{\text{lb}} \tag{17}$$

- To summarize, we have that, for both constant stepsizes and stepsizes computed according to backtracking line search, there exist a constant $s_{\text{lb}} > 0$ such that

$$s_k \geq s_{\text{lb}}, \quad \forall k = 0, 1, \ldots \tag{18}$$

The gradient method

Convergence to a stationary point.

- Consider a gradient descent algorithm

$$x_{k+1} = x_k - s_k \nabla f_0(x_k),$$

with stepsizes s_k computed via backtracking line search (or constant stepsizes), satisfying the Armijo condition

$$f_0(x_{k+1}) \leq f_0(x_k) - s_k \alpha \|\nabla f_0(x_k)\|_2^2. \tag{19}$$

- Then,

$$\begin{aligned} f_0(x_k) - f_0(x_{k+1}) &\geq s_k \alpha \|\nabla f_0(x_k)\|_2^2 \\ \text{[using (18)]} &\geq s_{\text{lb}} \alpha \|\nabla f_0(x_k)\|_2^2, \quad \forall k = 0, 1, \ldots \end{aligned}$$

Summing these inequalities from $0, 1, \ldots, k$, we obtain

$$s_{\text{lb}} \alpha \sum_{i=0}^{k} \|\nabla f_0(x_i)\|_2^2 \leq f_0(x_0) - f_0(x_{k+1}) \leq f_0(x_0) - f_0^*.$$

The gradient method

Convergence to a stationary point.

- Since the summation on the left is bounded by a constant as $k \to \infty$, we conclude that it must be

$$\lim_{k\to\infty} \|\nabla f_0(x_k)\|_2 = 0.$$

- This means that the algorithm converges to a *stationary point* of f_0, that is to a point where the gradient of f_0 is zero. Notice that such a point is not necessarily a local minimum of f_0.
- Further, by noticing that

$$\sum_{i=0}^{k} \|\nabla f_0(x_i)\|_2^2 \geq (k+1) \min_{i=0,\dots,k} \|\nabla f_0(x_i)\|_2^2,$$

we obtain from the previous inequality that

$$g_k^* \leq \frac{1}{\sqrt{k+1}} \frac{1}{\sqrt{s_{\rm lb}\alpha}} \sqrt{f_0(x_0) - f_0^*}, \tag{20}$$

where we defined $g_k^* \doteq \min_{i=0,\dots,k} \|\nabla f_0(x_i)\|_2$.

The gradient method

Convergence to a stationary point.

- This means that the sequence g_k^* of minimal gradient norms decreases at a *rate* given by the square-root of the number of iterations k.
- The stopping criterion is then typically set as

$$\|\nabla f_0(x_k)\|_2 \leq \epsilon,$$

and, using (20), we obtain that this exit condition is achieved in at most

$$k_{\max} = \left\lceil \frac{1}{\epsilon^2} \frac{f_0(x_0) - f_0^*}{s_{\mathrm{lb}} \alpha} \right\rceil$$

iterations.

The gradient method – for convex functions

- For convex f_0, x^* is a (global) minimizer if and only if $\nabla f_0(x^*) = 0$. Therefore, the gradient algorithm converges to a global minimum point. We next analyze at which *rate* this convergence is reached.
- Consider the decrease in objective function obtained at one step of the gradient algorithm (we consider, for simplicity in the proofs, the backtracking parameter to be fixed at $\alpha = 1/2$): from (19) we have

$$f_0(x_{k+1}) \leq f_0(x_k) - s_k \alpha \|\nabla f_0(x_k)\|_2^2. \tag{21}$$

Since f_0 is convex, it holds that

$$f_0(y) \geq f_0(x) + \nabla f_0(x)^\top (y - x), \quad \forall x, y \in \operatorname{dom} f_0,$$

which, for $y = x^*$, gives

$$f_0(x) \leq f_0^* + \nabla f_0(x)^\top (x - x^*), \quad \forall x \in \operatorname{dom} f_0.$$

The gradient method – for convex functions

- Substituting this into (21), we obtain

$$\begin{aligned} f_0(x_{k+1}) &\leq f_0(x_k) - s_k\alpha\|\nabla f_0(x_k)\|_2^2 \\ [\text{letting } \alpha = 1/2] &\leq f_0^* + \nabla f_0(x_k)^\top (x_k - x^*) - \frac{s_k}{2}\|\nabla f_0(x_k)\|_2^2 \\ &= f_0^* + \frac{1}{2s_k}\left(\|x_k - x^*\|_2^2 - \|x_k - x^* - s_k\nabla f_0(x_k)\|_2^2\right) \\ &= f_0^* + \frac{1}{2s_k}\left(\|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2\right) \\ [\text{since } s_k \geq s_{\text{lb}}] &\leq f_0^* + \frac{1}{2s_{\text{lb}}}\left(\|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2\right). \end{aligned}$$

- Considering this inequality for $k = 0, 1, \ldots$, we have

$$\begin{aligned} f_0(x_1) - f_0^* &\leq \frac{1}{2s_{\text{lb}}}\left(\|x_0 - x^*\|_2^2 - \|x_1 - x^*\|_2^2\right) \\ f_0(x_2) - f_0^* &\leq \frac{1}{2s_{\text{lb}}}\left(\|x_1 - x^*\|_2^2 - \|x_2 - x^*\|_2^2\right) \\ f_0(x_3) - f_0^* &\leq \frac{1}{2s_{\text{lb}}}\left(\|x_2 - x^*\|_2^2 - \|x_3 - x^*\|_2^2\right) \cdots \end{aligned}$$

The gradient method – for convex functions

- Hence, summing the first k of these inequalities, we have that

$$\begin{aligned}\sum_{i=1}^{k}(f_0(x_i)-f_0^*) &\leq \frac{1}{2s_{\rm lb}}\left(\|x_0-x^*\|_2^2-\|x_k-x^*\|_2^2\right)\\ &\leq \frac{1}{2s_{\rm lb}}\|x_0-x^*\|_2^2.\end{aligned}$$

- Now, since the sequence $f_0(x_k)-f_0^*$ is non-increasing with respect to k, its value is no larger than the average of the previous values of the sequence, that is

$$f_0(x_k)-f_0^* \leq \frac{1}{k}\sum_{i=1}^{k}(f_0(x_i)-f_0^*) \leq \frac{1}{2s_{\rm lb}k}\|x_0-x^*\|_2^2,$$

- This proves that $f_0(x_k)\to f_0^*$ at least at a rate which is inversely proportional to k. We then achieve an accuracy ϵ' on the objective function, i.e., $f_0(x_k)-f_0^*\leq\epsilon'$ in at most

$$k_{\max}=\left\lceil\frac{\|x_0-x^*\|_2^2}{2\epsilon' s_{\rm lb}}\right\rceil$$

iterations.

The gradient method – for strongly convex functions

Improved convergence rate can be obtained on strongly convex functions.

- Consider again the objective decrease in one iterate guaranteed by (21), where for simplicity we set $\alpha = 1/2$:

$$\begin{aligned} f_0(x_{k+1}) &\leq f_0(x_k) - s_k \alpha \|\nabla f_0(x_k)\|_2^2 \\ [\text{for } \alpha = 1/2] \quad &= \quad f_0(x_k) - \frac{s_k}{2}\|\nabla f_0(x_k)\|_2^2 \\ [\text{since } s_k \geq s_{\text{lb}}] \quad &\leq \quad f_0(x_k) - \frac{s_{\text{lb}}}{2}\|\nabla f_0(x_k)\|_2^2. \end{aligned}$$

- Subtracting f_0^* on both sides of this inequality, and using the bound previously derived in (12), we have that

$$\begin{aligned} f_0(x_{k+1}) - f_0^* &\leq (f_0(x_k) - f_0^*) - \frac{s_{\text{lb}}}{2}\|\nabla f_0(x_k)\|_2^2 \\ &\leq (f_0(x_k) - f_0^*) - 2m\frac{s_{\text{lb}}}{2}(f_0(x_k) - f_0^*) \\ &= (1 - m s_{\text{lb}})(f_0(x_k) - f_0^*). \end{aligned}$$

- Now, we recall from (16), (17) that, for $\alpha = 1/2$,

$$m s_{\text{lb}} = m \min(s_{\text{init}}, \beta \bar{s}_{\text{up}}) = \min(m s_{\text{init}}, \beta(m/M)).$$

The gradient method – for strongly convex functions

- Since $\beta < 1$ and $m/M \leq 1$, we have that $ms_{\text{lb}} < 1$, hence

$$(f_0(x_{k+1}) - f_0^*) \leq c(f_0(x_k) - f_0^*),$$

where $c = 1 - ms_{\text{lb}} \in (0, 1)$.

- Applying recursively the last inequality from 0 to k, we obtain

$$f_0(x_k) - f_0^* \leq c^k(f_0(x_0) - f_0^*), \tag{22}$$

which proves that convergence happens at a geometric rate.

- An accuracy ϵ' is achieved on the objective function, i.e., $f_0(x_k) - f_0^* \leq \epsilon'$ in at most

$$k_{\max} = \left\lceil \frac{\log(1/\epsilon') + d_0}{\log(1/c)} \right\rceil$$

iterations.

The gradient method – for strongly convex functions

- Further, from (22), together with Eq. (13) and (14) we have that

$$\frac{m}{2}\|x_k - x^*\|_2^2 \leq f_0(x_k) - f_0^* \leq c^k(f_0(x_0) - f_0^*) \leq c^k \frac{M}{2}\|x_0 - x^*\|_2^2,$$

hence

$$\|x_k - x^*\|_2 \leq c^{k/2}\sqrt{\frac{M}{m}}\|x_0 - x^*\|_2,$$

which provides an upper bound on the rate of convergence of x_k to x^*.

- Similarly, from the left inequality in (14) and the right inequality in (13) we obtain that

$$\frac{1}{2M}\|\nabla f_0(x_k)\|_2^2 \leq f_0(x_k) - f_0^* \leq c^k(f_0(x_0) - f_0^*) \leq c^k \frac{1}{2m}\|\nabla f_0(x_0)\|_2^2,$$

hence

$$\|\nabla f_0(x_k)\|_2 \leq c^{k/2}\sqrt{\frac{M}{m}}\|\nabla f_0(x_0)\|_2,$$

which provides an upper bound on the rate of convergence of the gradient to zero.

The gradient method – for strongly convex functions

Stopping criteria.

- From (13) we may derive useful stopping criteria in terms of the accuracy on the objective function value and on the minimizer.
- If at some iteration k we verify in a gradient algorithm that the condition $\|\nabla f_0(x_k)\| \leq \epsilon$ is met, then we can conclude that

$$f_0(x_k) - f_0^* \leq \frac{1}{2m}\|\nabla f_0(x_k)\|_2^2 \leq \frac{\epsilon^2}{2m}$$

  i.e., the current objective value $f_0(x_k)$ is $\epsilon' = \epsilon^2/(2m)$ close to the minimum.
- Similarly, from (13) we have that

$$\|x_k - x^*\|_2 \leq \frac{1}{m}\|\nabla f_0(x_k)\|_2 \leq \frac{\epsilon}{m}$$

  i.e., the current point x_k is $\epsilon'' = \epsilon/m$ close to the global minimizer x^*.

The gradient method

Example: convex quadratic form

- $f_0(x) = \frac{1}{2} x^\top H x$, $H \succ 0$.
- $\nabla f_0(x) = Hx$; descent direction $v_k = -\nabla f_0(x)$.
- Iterates of the form $x_{k+1} = x_k - s_k \nabla f_0(x_k) = x_k - s_k H x_k$.
- Exact line search: s minimizes

$$\begin{aligned} \phi(s) &= f_0(x_k - sHx_k) = \frac{1}{2}(x_k - sHx_k)^\top H(x_k - sHx_k) \\ &= \frac{1}{2}\left(c_2 s^2 - 2c_1 s + c_0\right), \end{aligned}$$

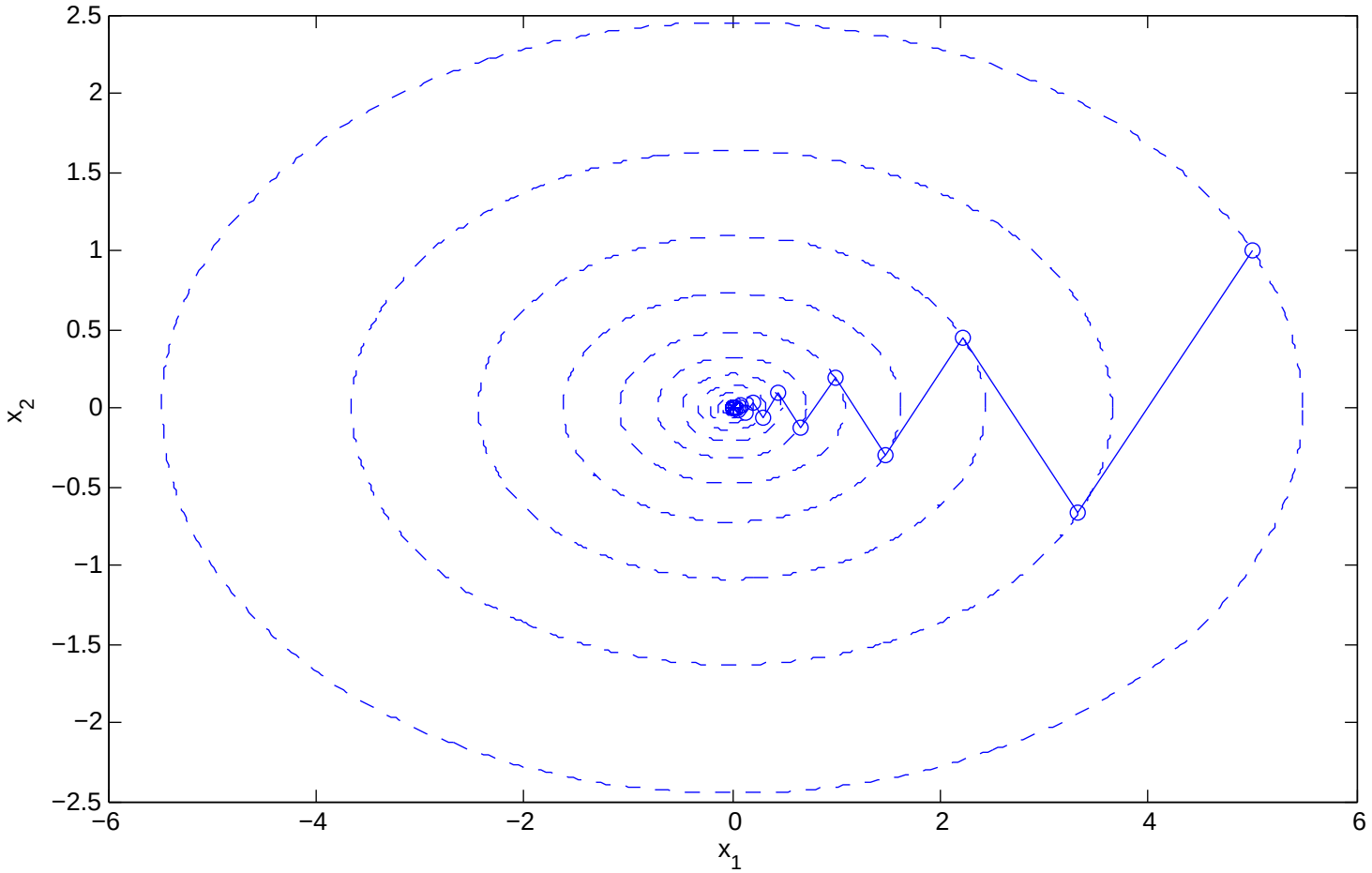
where $c_i \doteq x_k^\top H^{i+1} x_k$, $i = 0, 1, 2$. Optimal stepsize satisfies $\phi'(s) = c_2 s - c_1 = 0$, hence

$$s_k = \frac{c_1}{c_2} = \frac{x_k^\top H^2 x_k}{x_k^\top H^3 x_k}.$$

The gradient method

Example: convex quadratic form

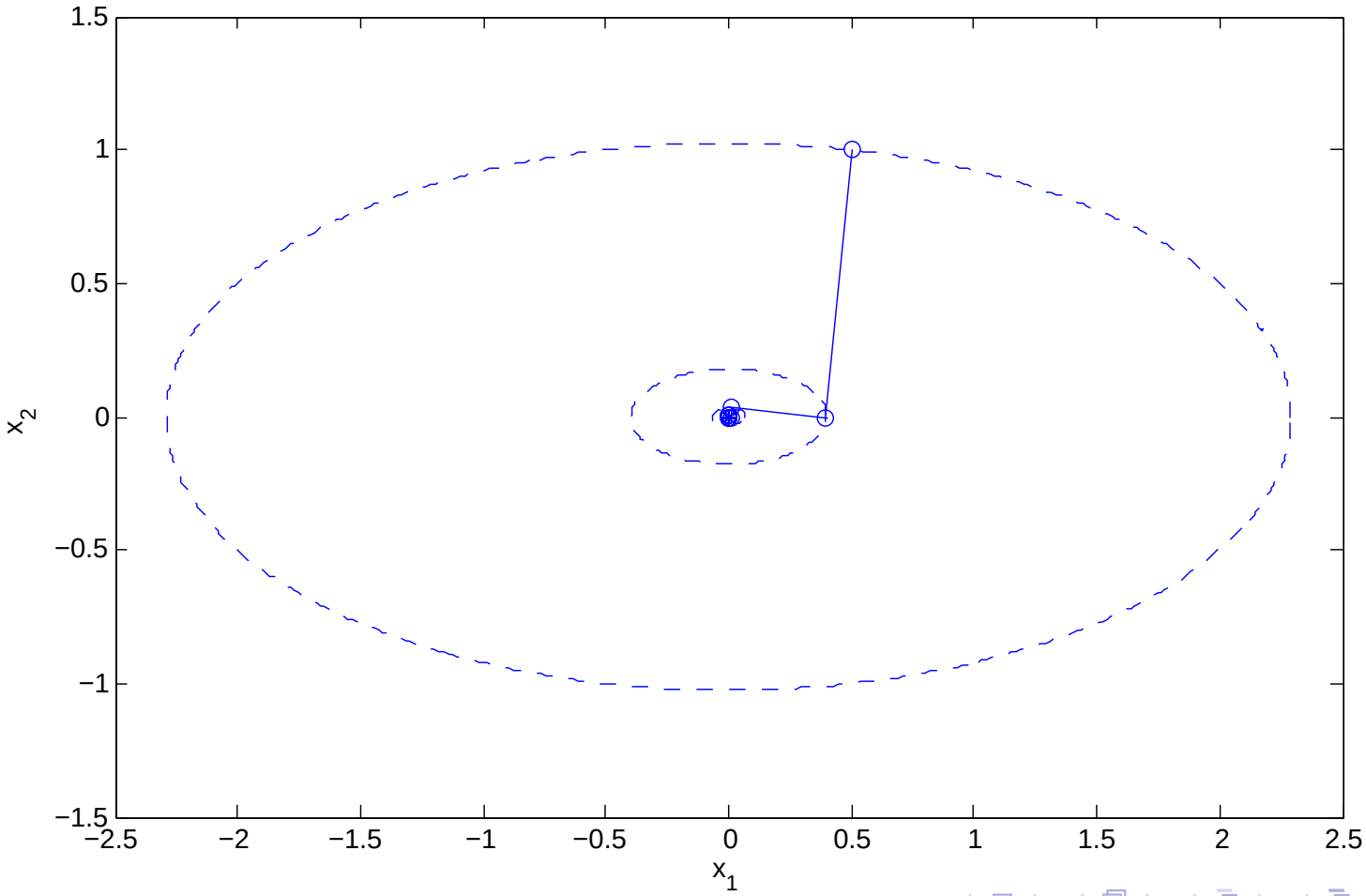
- Example 1: $H = \mathrm{diag}\,(1, \gamma)$, $\gamma = 5$, $x_0 = [1 \;\; \gamma]^\top$
- Zig-zagging convergence in 40 iterations to $\|\nabla f_0\| \leq 10^{-6}$.



The gradient method

Example: convex quadratic form

- Example 2: $H = \mathrm{diag}\,(1, \gamma)$, $\gamma = 5$, $x_0 = [0.5 \;\; 1]^\top$
- fast convergence in 10 iterations to $\|\nabla f_0\| \leq 10^{-6}$.



The gradient method

Example: non-differentiable objective (From L. Vandenberghe – EE236C)

- Objective:

$$f_0(x) = \begin{cases} \sqrt{x_1^2 + \gamma x_2^2} & \text{if } |x_2| \leq x_1 \\ \dfrac{x_1 + \gamma |x_2|}{\sqrt{1+\gamma}} & \text{if } |x_2| > x_1 \end{cases}$$

- Gradient (defined for $|x_2| \neq x_1$):

$$\nabla f_0(x) = \begin{cases} \frac{1}{\sqrt{x_1^2+\gamma x_2^2}} \begin{bmatrix} x_1 \\ \gamma x_2 \end{bmatrix} & \text{if } |x_2| < x_1 \\ \frac{1}{\sqrt{1+\gamma}} \begin{bmatrix} 1 \\ \gamma \mathrm{sgn}(x_2) \end{bmatrix} & \text{if } |x_2| > x_1 \end{cases}$$

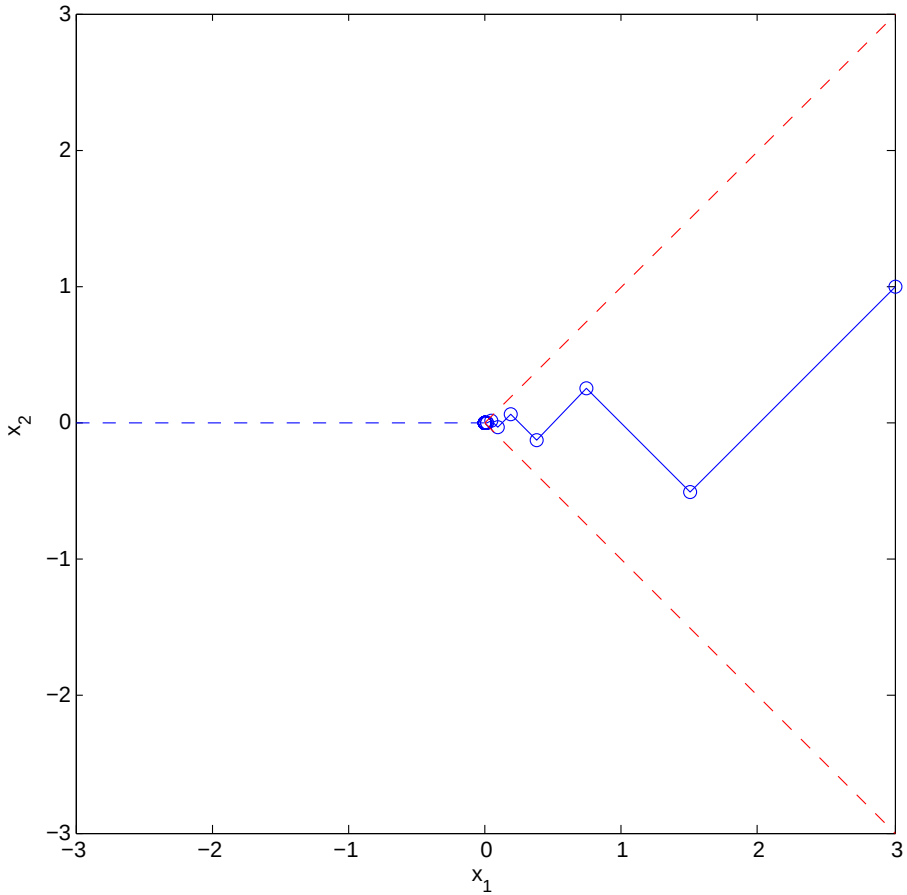
- Optimal stepsizes, in the region $|x_2| < x_1$

$$s_k = \sqrt{x_k^\top H x_k} \frac{x_k^\top H^2 x_k}{x_k^\top H^3 x_k}$$

The gradient method

Example: non-differentiable objective

- $\gamma = 3$, $x_0 = [\gamma \;\; 1]^\top$
- converges to zero, which is not optimal!



The gradient method

Advantages

- Computationally inexpensive iterations;
- Does not require second derivatives (Hessian).

Disadvantages

- Convergence may be slow;
- Does not work on non-differentiable problems.