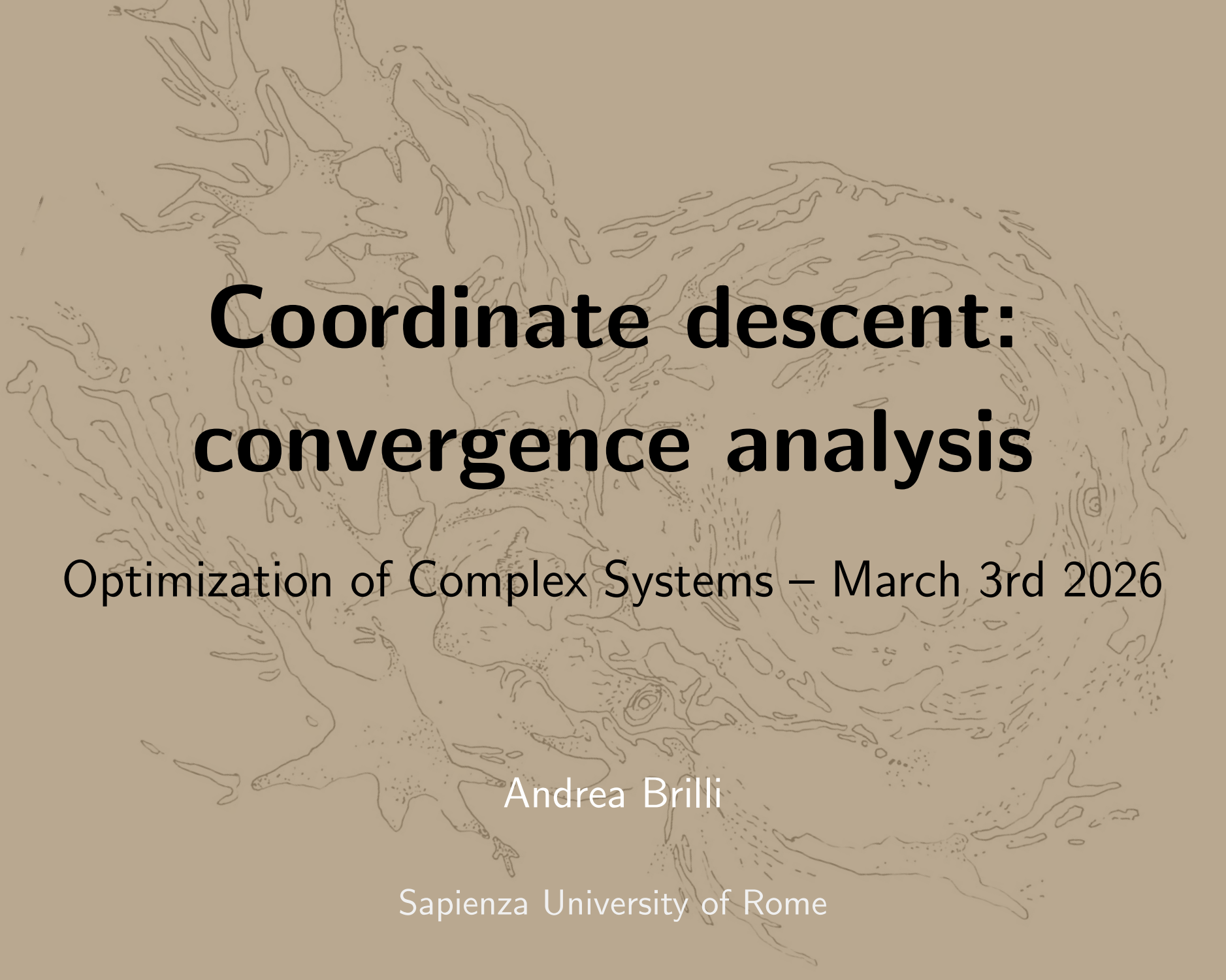


Coordinate descent: convergence analysis

Optimization of Complex Systems – March 3rd 2026

Andrea Brilli

Sapienza University of Rome

Coordinate search: pseudocode

Input: $x_0 \in \mathbb{R}^n$, $\alpha_0 > 0$, $\alpha_{\min} > 0$, $\theta \in (0,1)$

$k \leftarrow 0$, $x \leftarrow x_0$, $\alpha_k \leftarrow \alpha_0$
while $\alpha \geq \alpha_{\min}$ **do**
    $k \leftarrow k+1$
    Find $\bar{d} \in \mathcal{D} = \{\pm e_1, \ldots, \pm e_n\}$ s.t. $f(x_k + \alpha_k \bar{d}) < f(x_k)$
    if such $\bar{d}$ exists **then**
        $x_{k+1} \leftarrow x_k + \alpha_k \bar{d}$, $\alpha_{k+1} \leftarrow \alpha_k$ (success: keep step size)
    else
        $\alpha_{k+1} \leftarrow \theta \alpha_k$ (failure: reduce step size)
    end if
end while
return x_k

Stopping criterion: $\alpha_k < \alpha_{\min}$ (step size becomes negligible)

Convergence analysis: setup

Assumption:

(A1) The sub-level set $\mathcal{L}(x_0) = \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ is **compact**.

Convergence analysis: setup

Assumption:

(A1) The sub-level set $\mathcal{L}(x_0) = \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ is **compact**.

Key observations:

1. The sequence $\{f(x_k)\}$ is non-increasing: $f(x_{k+1}) \leq f(x_k)$
2. All iterates remain in $\mathcal{L}(x_0)$
3. Step size α_k either stays constant or decreases: $\alpha_{k+1} \in \{\alpha_k, \theta\alpha_k\}$

Why do we need $\alpha_k \to 0$?

Recall the key insight from descent direction theory:

- If d is a descent direction at x, then for **small enough** $\alpha > 0$, we have

$$f(x + \alpha d) < f(x)$$

- The phrase "small enough" is crucial — how small must α be?
- Near a minimum point, the gradient becomes smaller, so the improvement from taking a step also becomes smaller
- To guarantee improvement, we must allow α to become arbitrarily small

Question: In the coordinate descent algorithm, does the step-size α_k actually converge to zero?

Setting the stage

Algorithm: Coordinate descent with step-size adjustment

- Directions: $\mathcal{D} = \{\pm e_1, \pm e_2, \ldots, \pm e_n\}$ (coordinate directions)
- At iteration k: try to find $d \in \mathcal{D}$ such that $f(x_k + \alpha_k d) < f(x_k)$
- **Success:** accept the step, $x_{k+1} = x_k + \alpha_k d$, keep $\alpha_{k+1} = \alpha_k$
- **Failure:** reject all moves, $x_{k+1} = x_k$, reduce $\alpha_{k+1} = \theta \alpha_k$ with $\theta \in (0, 1)$

Key property: The step-size is **never increased**

$$\alpha_{k+1} \leq \alpha_k, \qquad \forall k$$

The main result

Theorem (Convergence of step-sizes)

Under Assumption A1 (compact level sets), the sequence of step-sizes $\{\alpha_k\}$ generated by the coordinate descent algorithm satisfies

$$\lim_{k \to \infty} \alpha_k = 0$$

Strategy: Proof by contradiction

We will assume $\alpha_k \not\to 0$ and derive a contradiction by showing that:

1. The iterates must lie on a finite mesh
2. The algorithm must eventually cycle
3. This cycling contradicts the step-size reduction mechanism

Step 1: Properties of $\{\alpha_k\}$

The sequence $\{\alpha_k\}$ is a non-increasing sequence of positive scalars, therefore it admits a unique limit point $\bar{\alpha} \geq 0$.

Proof.

By construction:

- $\alpha_0 > 0$ is chosen initially
- At each iteration, either $\alpha_{k+1} = \alpha_k$ (success) or $\alpha_{k+1} = \theta\alpha_k < \alpha_k$ (failure)
- Thus $\alpha_{k+1} \leq \alpha_k$ for all k, and $\alpha_k > 0$ for all k

Therefore $\{\alpha_k\}$ is a non-increasing sequence bounded below by 0, hence convergent to some $\bar{\alpha} \geq 0$. □

Now: Suppose by contradiction that $\bar{\alpha} > 0$

Step 2: Constant step-size and successful iterations

- If $\lim_{k\to\infty} \alpha_k = \bar{\alpha} > 0$, then by the instructions of the algorithm there exists $\bar{k}$ such that

$$\alpha_k = \bar{\alpha}, \qquad \forall k \geq \bar{k}$$

- If $\alpha_k = \bar{\alpha}$ for all $k \geq \bar{k}$, then all iterations $k \geq \bar{k}$ must be successful, or the step-size would be reduced.

  Consequence: For all $k \geq \bar{k}$, there exists $d \in \mathcal{D}$ such that

$$x_{k+1} = x_k + \bar{\alpha} d, \qquad f(x_{k+1}) < f(x_k)$$

Step 4: Sub-level set and mesh

- Since the sequence $\{f(x_k)\}$ is monotonically nonincreasing, all the iterates $\{x_k\}$ lie within the sub-level set $\mathcal{L}(x_0) = \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$.
  (In fact, we have $f(x_k) \leq f(x_0)$ for all k)
  By Assumption A1: The sub-level set $\mathcal{L}(x_0)$ is compact (closed and bounded).
- For $k \geq \bar{k}$, all iterates lie on a discrete mesh contained in $\mathcal{L}(x_0)$.

Proof.

For $k \geq \bar{k}$, we have $x_{k+1} = x_k + \bar{\alpha} d_k$ where $d_k \in \mathcal{D} = \{\pm e_1, \ldots, \pm e_n\}$.
Starting from $x_{\bar{k}}$, each coordinate can only change by integer multiples of $\bar{\alpha}$:

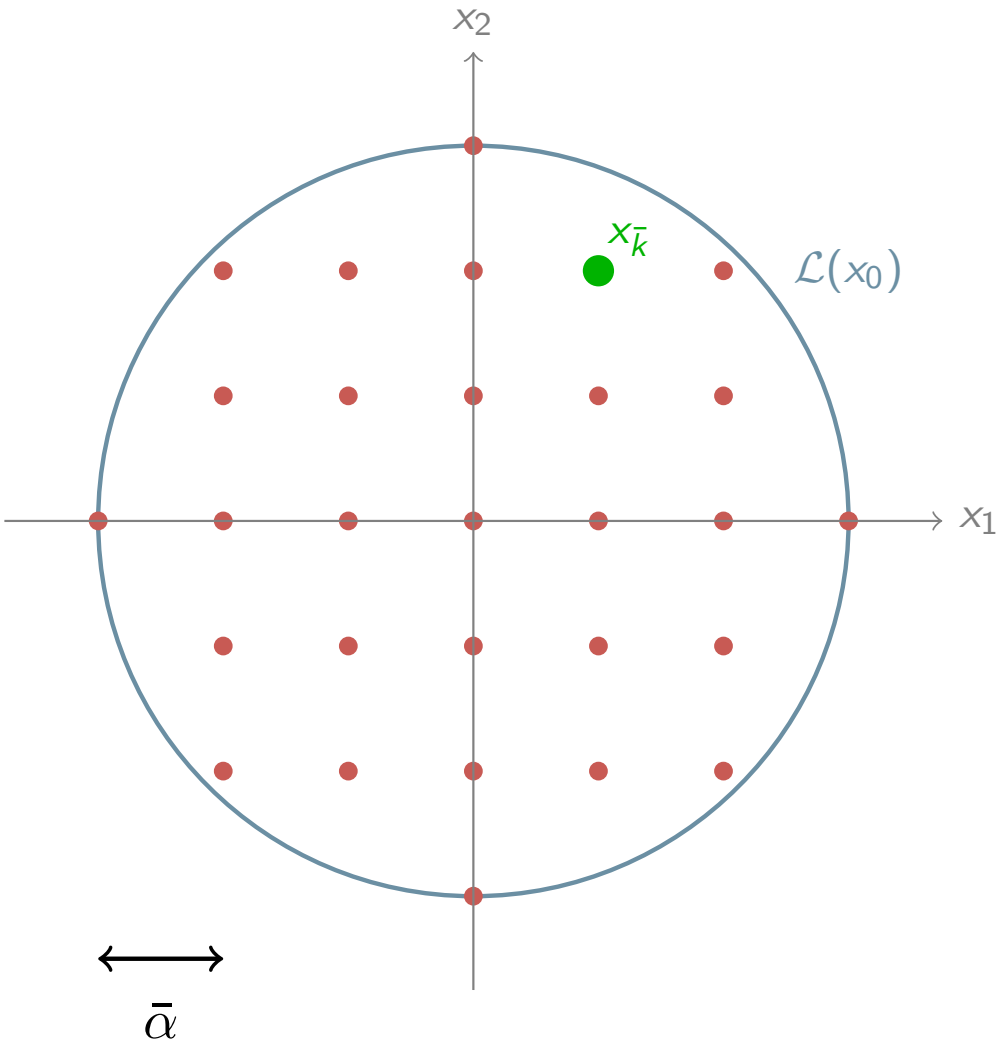
$$(x_k)_i = (x_{\bar{k}})_i + \bar{\alpha} m_i, \qquad m_i \in \mathbb{Z}, \quad i = 1, \ldots, n$$

Define the mesh:

$$\mathcal{M} = \left\{ x_{\bar{k}} + \bar{\alpha} \sum_{i=1}^{n} m_i e_i : m_i \in \mathbb{Z} \right\} \cap \mathcal{L}(x_0)$$

Then $x_k \in \mathcal{M}$ for all $k \geq \bar{k}$. □

Visualizing the mesh



The iterates $\{x_k\}_{k \geq \bar{k}}$ can only visit the red mesh points inside the sub-level set.

Step 6: Finiteness of the mesh

Because the mesh lies within the bounded sub-level set $\mathcal{L}(x_0)$, the set $\mathcal{M}$ contains only finitely many points.

Therefore, the sequence $\{x_k\}_{k \geq \bar{k}}$ is an infinite sequence of points contained in finite set. It follows that at least one point must be visited twice, i.e. there exist $k_1 > k_2 > \bar{k}$ such that $x_{k_2} = x_{k_1}$, which contradicts the fact that $f(x_{k+1}) < f(x_k)$ for all $k > \bar{k}$. (recall all iterations after $\bar{k}$ are successful)

Finally, the latter contradiction implies that $\{\alpha_k\} \to 0$.

Interpretation and consequences

What does this result tell us?

- The algorithm is **forced** to search at increasingly finer scales as it proceeds
- This is necessary because near a minimum, the "safe" step-size for guaranteed decrease becomes smaller and smaller
- The compactness assumption ensures we cannot escape to infinity — we must stay in a bounded region
- The finite mesh structure makes the cycling argument rigorous

Key insight: The combination of

- monotonic decrease in function values
- confinement to a compact level set
- discrete mesh structure with fixed spacing

creates an impossible situation unless the step-size shrinks to zero.

Note: for the convergence analysis of the step-size we only assumed the compactness of the sub-level set, we did not use continuity or differentiablity!

A bit of analysis (mean value theorem)

Theorem (Mean value theorem in $\mathbb{R}$)

Let $g : \mathbb{R} \to \mathbb{R}$ be continuous on $[a, b]$ and differentiable on (a, b). Then there exists $\xi \in (a, b)$ such that

$$g(b) - g(a) = g'(\xi)(b - a).$$

Definition (Segment in $\mathbb{R}^n$)

Given $a, b \in \mathbb{R}^n$, the segment with endpoints a and b is

$$[a, b] \stackrel{\triangle}{=} \{x \in \mathbb{R}^n : \ x = ta + (1 - t)b, \ \ t \in [0, 1]\},$$

Theorem (Mean value theorem in $\mathbb{R}^n$)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuous on $[a, b]$ and differentiable on (a, b). Then there exists $\xi \in (a, b)$ such that

$$f(b) - f(a) = \nabla f(\xi)^\top (b - a).$$

Convergence to stationary points

Assume (A1) and consider the coordinate search method (compass search) with step-size sequence $\{\alpha_k\}$. We want to show

$$\liminf_{k\to\infty} \|\nabla f(x_k)\| = 0.$$

We follow the route that assumes ∇f is continuous.

Stationarity theorem (continuous gradient)

Theorem

Assume (A1): the level set

$$L(x_0) = \{x \in \mathbb{R}^n : \ f(x) \leq f(x_0)\}$$

is compact. If f is continuously differentiable, then

$$\liminf_{k \to \infty} \|\nabla f(x_k)\| = 0.$$

Equivalently, at least one limit point of $\{x_k\}$ is stationary.

Proof idea.

- $\{x_k\} \subset L(x_0)$, hence it has limit points (compactness).
- There exists an infinite set of indices where the step-size is reduced, hence along that subsequence $\alpha_k \to 0$.
- At such indices, all coordinate trials are *unsuccessful*; apply the mean value theorem on the segments joining x_k to the trial points.

Stationarity: extracting a subsequence

Setup. By compactness of $L(x_0)$ and monotonicity of the method,

$$x_k \in L(x_0) \qquad \forall k.$$

Moreover, since the method reduces the step-size by a factor $\theta \in (0,1)$ whenever an iteration is unsuccessful, there exists an infinite index set $\mathcal{K}_1 \subseteq \{0,1,2,\dots\}$ such that

$$k \in \mathcal{K}_1 \implies \alpha_{k+1} = \theta\,\alpha_k,$$

hence $\{\alpha_k\}_{k\in\mathcal{K}_1} \to 0$.

Since $L(x_0)$ is compact, $\{x_k\}_{k\in\mathcal{K}_1}$ admits limit points (**NOTE: it might not be unique!**). Let us restrict to any given (accumulation) limit point of the sequence, that is, let us consider an infinite subsequence $\mathcal{K}_2 \subseteq \mathcal{K}_1$ such that

$$\lim_{k\to\infty,\ k\in\mathcal{K}_2} x_k = \bar{x}, \qquad \lim_{k\to\infty,\ k\in\mathcal{K}_2} \alpha_k = 0.$$

Unsuccessful trials imply two inequalities

Fix any $i \in \{1, \ldots, n\}$. For every $k \in \mathcal{K}_2$, iteration k is unsuccessful, hence

$$\begin{aligned} f(x_k + \alpha_k e_i) &\geq f(x_k), \\ f(x_k - \alpha_k e_i) &\geq f(x_k). \end{aligned}$$

Apply the mean value theorem in $\mathbb{R}^n$ on the segments $[x_k,\ x_k \pm \alpha_k e_i]$: there exists $t^u_{k,i} \in (0,1)$ and $t^v_{k,i} \in (0,1)$, with

$$u_{k,i} \stackrel{\triangle}{=} x_k + t^u_{k,i}\alpha_k e_i \qquad v_{k,i} \stackrel{\triangle}{=} x_k - t^v_{k,i}\alpha_k e_i$$

such that

$$f(x_k + \alpha_k e_i) - f(x_k) = \nabla f(u_{k,i})^\top (\alpha_k e_i) = \alpha_k \nabla f(u_{k,i})^\top e_i.$$

$$f(x_k) - f(x_k - \alpha_k e_i) = \nabla f(v_{k,i})^\top (-\alpha_k e_i) = -\alpha_k \nabla f(v_{k,i})^\top e_i,$$

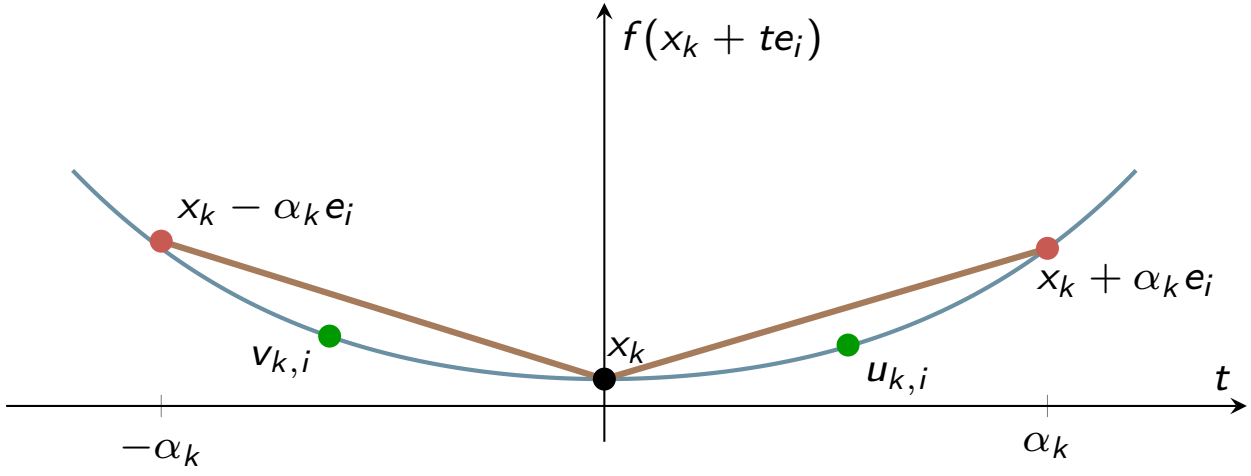
Visualization

Fix an index $i \in \{1, \ldots, n\}$ and a point x_k where the algorithm *fails*, i.e.,

$$f(x_k \pm \alpha_k e_i) \;\geq\; f(x_k).$$

Define the 1D restriction (along the i-th coordinate) as a function of t

$$f(x_k + t\, e_i), \qquad t \in \mathbb{R}.$$



Taking limits: the i-th component must vanish

From the unsuccessful-trial inequalities and the MVT identities:

$$f(x_k + \alpha_k e_i) - f(x_k) = \alpha_k \nabla f(u_{k,i})^\top e_i \ \geq \ 0,$$

$$f(x_k - \alpha_k e_i) - f(x_k) = -\alpha_k \nabla f(v_{k,i})^\top e_i \ \geq \ 0.$$

Since $\alpha_k > 0$, we can divide by α_k to obtain

$$\nabla f(u_{k,i})^\top e_i \geq 0, \qquad -\nabla f(v_{k,i})^\top e_i \geq 0.$$

Since $x_k \to \bar{x}$ and $\alpha_k \to 0$ along $\mathcal{K}_2$, we have (see the visualization in the next slide)

$$u_{k,i} \to \bar{x}, \qquad v_{k,i} \to \bar{x} \qquad (k \to \infty, \ k \in \mathcal{K}_2).$$

By continuity of ∇f,

$$\lim_{k \to \infty, \ k \in \mathcal{K}_2} \nabla f(u_{k,i})^\top e_i = \nabla f(\bar{x})^\top e_i \geq 0,$$

$$\lim_{k \to \infty, \ k \in \mathcal{K}_2} -\nabla f(v_{k,i})^\top e_i = -\nabla f(\bar{x})^\top e_i \geq 0.$$

Therefore,

$$\nabla f(\bar{x})^\top e_i = 0.$$

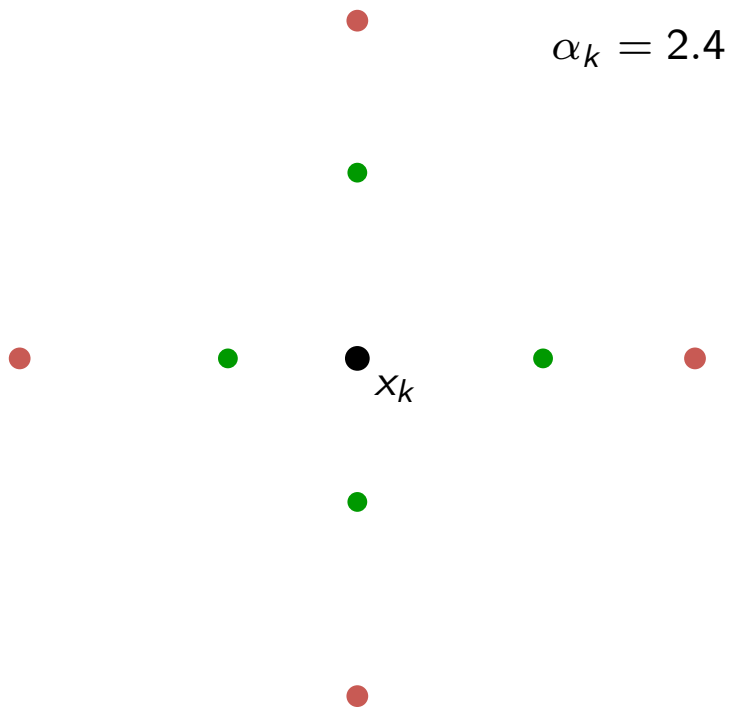
Visualization

At a failed iteration k we have *all* coordinate trials unsuccessful:

$$f(x_k \pm \alpha_k e_i) \geq f(x_k), \qquad i = 1, \ldots, n.$$

As $\alpha_k \to 0$, the trial points $x_k \pm \alpha_k e_i$ converge to x_k for each coordinate.

The trial points (•) and the mean-value points (•) collpase toward the center x_k.



Visualization

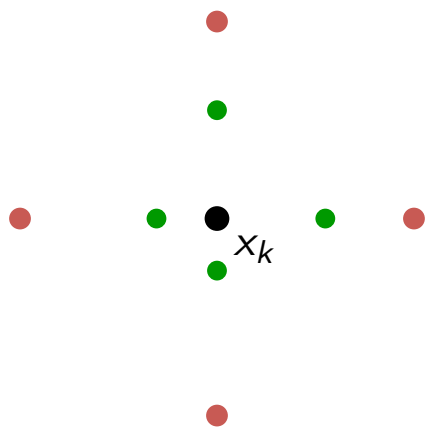
At a failed iteration k we have *all* coordinate trials unsuccessful:

$$f(x_k \pm \alpha_k e_i) \geq f(x_k), \qquad i = 1, \ldots, n.$$

As $\alpha_k \to 0$, the trial points $x_k \pm \alpha_k e_i$ converge to x_k for each coordinate.

The trial points (•) and the mean-value points (•) collpase toward the center x_k.

$\alpha_k = 1.4$



Visualization

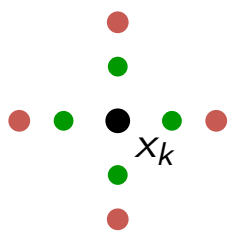
At a failed iteration k we have *all* coordinate trials unsuccessful:

$$f(x_k \pm \alpha_k e_i) \geq f(x_k), \qquad i = 1, \ldots, n.$$

As $\alpha_k \to 0$, the trial points $x_k \pm \alpha_k e_i$ converge to x_k for each coordinate.

The trial points (•) and the mean-value points (•) collpase toward the center x_k.

$$\alpha_k = 0.7$$



Visualization

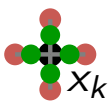
At a failed iteration k we have *all* coordinate trials unsuccessful:

$$f(x_k \pm \alpha_k e_i) \geq f(x_k), \qquad i = 1, \ldots, n.$$

As $\alpha_k \to 0$, the trial points $x_k \pm \alpha_k e_i$ converge to x_k for each coordinate.

The trial points (•) and the mean-value points (•) collpase toward the center x_k.

$$\alpha_k = 0.25$$



Conclusion: $\bar{x}$ is stationary

The previous argument holds for every $i = 1, \ldots, n$, hence

$$\nabla f(\bar{x})^\top e_i = 0, \qquad i = 1, \ldots, n.$$

Since $\{e_1, \ldots, e_n\}$ is a basis of $\mathbb{R}^n$, it follows that

$$\nabla f(\bar{x}) = 0.$$

Thus at least one limit point $\bar{x}$ of $\{x_k\}$ is stationary, and in particular

$$\liminf_{k \to \infty} \|\nabla f(x_k)\| = 0.$$

Recall: Lipschitz continuous gradient and descent lemma

Definition (Lipschitz continuous gradient)

∇f is Lipschitz continuous with constant $L > 0$ on $A \subseteq \mathbb{R}^n$ when

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \qquad \forall\, x, y \in A.$$

Theorem (Descent Lemma)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable with Lipschitz continuous gradient with constant $L > 0$. Then, for all $x, y \in \mathbb{R}^n$,

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|y - x\|^2.$$

Proof: descent lemma

Proof. Write $f(y) - f(x)$ as the integral of the directional derivative along $[x, y]$:

$$f(y) - f(x) = \int_0^1 \nabla f(x + t(y - x))^\top (y - x) \, \mathrm{d}t.$$

Subtract and add $\nabla f(x)^\top (y - x)$:

$$f(y) - f(x) - \nabla f(x)^\top (y - x) = \int_0^1 (\nabla f(x + t(y - x)) - \nabla f(x))^\top (y - x) \, \mathrm{d}t.$$

By Cauchy–Schwarz and Lipschitz continuity of ∇f:

$$\leq \int_0^1 \|\nabla f(x + t(y - x)) - \nabla f(x)\| \, \|y - x\| \, \mathrm{d}t \leq \int_0^1 Lt \|y - x\|^2 \, \mathrm{d}t = \frac{L}{2} \|y - x\|^2. \quad \square$$

From the descent lemma to the algorithm

Apply the descent lemma with $y = x_k + \alpha_k d$ (trial point) and $x = x_k$:

$$f(x_k + \alpha_k d) \leq f(x_k) + \alpha_k \nabla f(x_k)^\top d + \frac{L}{2}\alpha_k^2 \|d\|^2.$$

Question: What happens when the trial point fails *unsuccessful*, i.e., $f(x_k + \alpha_k d) \geq f(x_k)$?

From the descent lemma to the algorithm

Apply the descent lemma with $y = x_k + \alpha_k d$ (trial point) and $x = x_k$:

$$f(x_k + \alpha_k d) \leq f(x_k) + \alpha_k \nabla f(x_k)^\top d + \frac{L}{2}\alpha_k^2 \|d\|^2.$$

Question: What happens when the trial point fails *unsuccessful*, i.e., $f(x_k + \alpha_k d) \geq f(x_k)$?

Combining the two inequalities:

$$f(x_k) \leq f(x_k + \alpha_k d) \leq f(x_k) + \alpha_k \nabla f(x_k)^\top d + \frac{L}{2}\alpha_k^2 \|d\|^2.$$

Rearranging (using $\|d\| = 1$ for coordinate directions and dividing by $\alpha_k > 0$):

$$0 \leq \nabla f(x_k)^\top d + \frac{L}{2}\alpha_k,$$

$$\boxed{-\nabla f(x_k)^\top d \leq \frac{L}{2}\alpha_k.}$$

This inequality holds for *every* direction d where the trial fails.

What if all directions fail?

At an unsuccessful iteration, *all* $2n$ coordinate directions fail:

$$-\nabla f(x_k)^\top d \leq \frac{L}{2}\alpha_k, \qquad \forall\, d \in \mathcal{D} = \{\pm e_1, \ldots, \pm e_n\}.$$

Question: Can we deduce something about $\|\nabla f(x_k)\|$ from these $2n$ inequalities?

Recall (Class 1): We introduced *positive spanning sets* $\mathcal{D}$ with the property:

For any vector $v \in \mathbb{R}^n$, there exists at least one direction $d \in \mathcal{D}$ that is well-aligned with v, i.e., $v^\top d > 0$.

What if all directions fail?

At an unsuccessful iteration, *all* $2n$ coordinate directions fail:

$$-\nabla f(x_k)^\top d \leq \frac{L}{2}\alpha_k, \qquad \forall\, d \in \mathcal{D} = \{\pm e_1, \ldots, \pm e_n\}.$$

Question: Can we deduce something about $\|\nabla f(x_k)\|$ from these $2n$ inequalities?

Recall (Class 1): We introduced *positive spanning sets* $\mathcal{D}$ with the property:

For any vector $v \in \mathbb{R}^n$, there exists at least one direction $d \in \mathcal{D}$ that is well-aligned with v, i.e., $v^\top d > 0$.

Key observation: Apply this to $v = -\nabla f(x_k)$ (the anti-gradient, the direction of steepest *ascent* of $-f$).

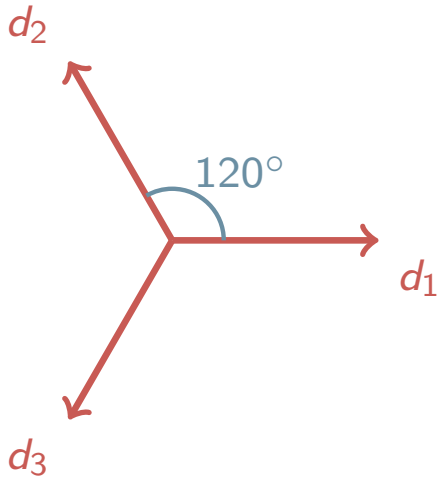
Since $\mathcal{D}$ is a positive spanning set, there exists at least one $d^* \in \mathcal{D}$ such that

$$-\nabla f(x_k)^\top d^* > 0.$$

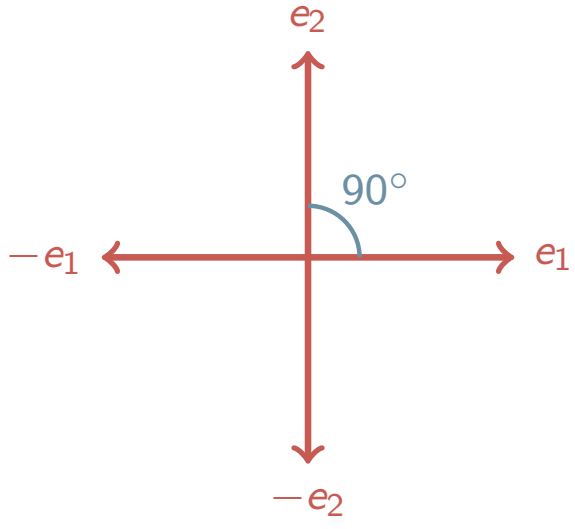
But how large can this inner product be?

Alignment = Angle

Consider two positive spanning sets in $\mathbb{R}^2$:



$\mathcal{D}_1$: $n+1$ directions



$\mathcal{D}_2$: coordinate directions

The alignment is related to the angle between the different directions. Whenever we pick a vector $v \in \mathbb{R}^n$, we are interested in the direction $d \in \mathcal{D}$ which has the best alignment with v.

Cosine Measure

For a given vector v and a set of directions $\mathcal{D}$, define the *best alignment*:

$$\phi(v, \mathcal{D}) = \max_{d \in \mathcal{D}} \cos\theta(v, d) = \max_{d \in \mathcal{D}} \frac{v^\top d}{\|v\| \, \|d\|}.$$

Question: for $v = -\nabla f(x_k)$, what is the *worst* possible value of $\phi(-\nabla f(x_k), \mathcal{D})$?

That is, for which vector v is $\phi(v, \mathcal{D})$ the *smallest*?

Cosine Measure

For a given vector v and a set of directions $\mathcal{D}$, define the *best alignment*:

$$\phi(v, \mathcal{D}) = \max_{d \in \mathcal{D}} \cos\theta(v, d) = \max_{d \in \mathcal{D}} \frac{v^\top d}{\|v\| \, \|d\|}.$$

Question: for $v = -\nabla f(x_k)$, what is the *worst* possible value of $\phi(-\nabla f(x_k), \mathcal{D})$?

That is, for which vector v is $\phi(v, \mathcal{D})$ the *smallest*?

Definition (Cosine measure)

The cosine measure of a finite set $\mathcal{D}$ is

$$\kappa(\mathcal{D}) = \min_{v \in \mathbb{R}^n,\, v \neq 0} \max_{d \in \mathcal{D}} \frac{v^\top d}{\|v\| \, \|d\|} = \min_{v \in \mathbb{R}^n,\, v \neq 0} \phi(v, \mathcal{D}).$$

Interpretation:

- $\phi(v, \mathcal{D})$ measures how well $\mathcal{D}$ aligns with a *given* vector v
- $\kappa(\mathcal{D})$ measures the *worst-case* alignment over *all possible* vectors v
- $\kappa(\mathcal{D}) > 0$ if and only if $\mathcal{D}$ is a positive spanning set

Cosine measure of coordinate directions

Let $\mathcal{D} = \{\pm e_1, \ldots, \pm e_n\}$. For any $v \in \mathbb{R}^n$ with $\|v\| = 1$:

$$\max_{d \in \mathcal{D}} \frac{v^\top d}{\|d\|} = \max_{i=1,\ldots,n} |v^\top e_i| = \max_{i=1,\ldots,n} |v_i|.$$

Hence

$$\kappa(\mathcal{D}) = \min_{\substack{v \in \mathbb{R}^n \\ \|v\|=1}} \max_{i=1,\ldots,n} |v_i|.$$

Suppose $\max_i |v_i| < 1/\sqrt{n}$ for some unit vector v. Then

$$1 = \|v\|^2 = \sum_{i=1}^{n} v_i^2 < n \cdot \frac{1}{n} = 1,$$

a contradiction. Hence $\max_i |v_i| \geq 1/\sqrt{n}$ for all unit v, giving $\kappa(\mathcal{D}) \geq 1/\sqrt{n}$.

Equality is attained at $v^* = (1/\sqrt{n}, \ldots, 1/\sqrt{n})^\top$: $\max_i |v_i^*| = 1/\sqrt{n}$.

Therefore

$$\kappa(\mathcal{D}) = \frac{1}{\sqrt{n}}$$

.

Bounding the gradient at unsuccessful iterations

Putting it together. Since $\kappa(\mathcal{D}) = 1/\sqrt{n}$, the definition of cosine measure guarantees: for every x_k with $\nabla f(x_k) \neq 0$, there exists $d^* \in \mathcal{D}$ such that

$$\frac{-\nabla f(x_k)^\top d^*}{\|\nabla f(x_k)\| \, \|d^*\|} \geq \frac{1}{\sqrt{n}},$$

i.e.,

$$-\nabla f(x_k)^\top d^* \geq \frac{1}{\sqrt{n}} \|\nabla f(x_k)\|.$$

At an unsuccessful iteration k (all trials fail), the bound derived from the descent lemma holds for *every* $d \in \mathcal{D}$, including d^*:

$$-\nabla f(x_k)^\top d^* \leq \frac{L}{2} \alpha_k.$$

Combining the two inequalities, for all unsuccessful iterations k:

$$\frac{1}{\sqrt{n}} \|\nabla f(x_k)\| \leq -\nabla f(x_k)^\top d^* \leq \frac{L}{2} \alpha_k,$$

$$\boxed{\|\nabla f(x_k)\| \leq \frac{\sqrt{n}\, L}{2} \alpha_k.}$$

Stationarity theorem (Lipschitz gradient)

Theorem

Assume (A1) and that ∇f is Lipschitz continuous with constant $L > 0$. Then

$$\liminf_{k \to \infty} \|\nabla f(x_k)\| = 0.$$

Proof. We have already shown that $\alpha_k \to 0$ (step-size convergence theorem).

At every unsuccessful iteration k, the gradient bound gives

$$\|\nabla f(x_k)\| \leq \frac{\sqrt{n}\, L}{2} \alpha_k.$$

There exist infinitely many unsuccessful iterations (they occur whenever the step-size is reduced), forming the infinite index set $\mathcal{K}_1$. For every $k \in \mathcal{K}_1$:

$$0 \leq \|\nabla f(x_k)\| \leq \frac{\sqrt{n}\, L}{2} \alpha_k \xrightarrow{k \in \mathcal{K}_1} 0.$$

What did we prove so far?

We proved convergence to stationary points using *two different approaches*:

1. **Mean Value Theorem approach** (continuous gradient)
   - Uses compactness + continuity to extract convergent subsequences
   - Shows that limit points must be stationary

2. **Lipschitz + Cosine Measure approach** (Lipschitz gradient)
   - Direct bound: $\|\nabla f(x_k)\| \leq \frac{\sqrt{n}L}{2}\alpha_k$ at unsuccessful iterations
   - No subsequence extraction needed

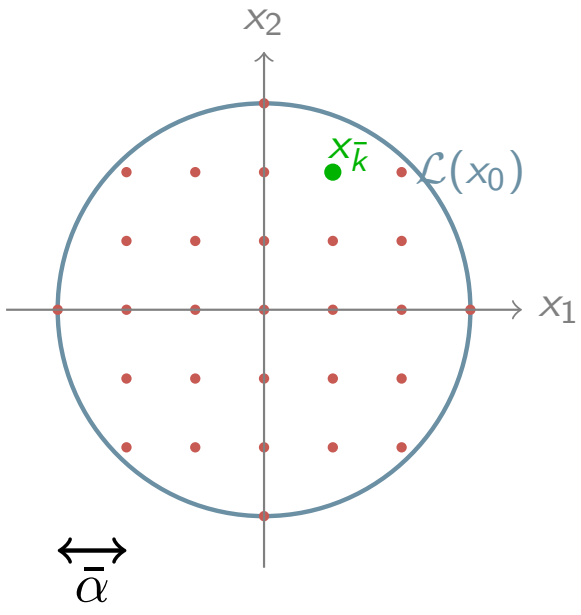
Common ingredient: Both proofs rely crucially on

$$\boxed{\lim_{k\to\infty} \alpha_k = 0}$$

Question: What guarantees $\alpha_k \to 0$ in general?

The discrete mesh argument (coordinate search)

Recall: For coordinate directions $\mathcal{D} = \{\pm e_1, \ldots, \pm e_n\}$ with fixed $\bar{\alpha} > 0$:

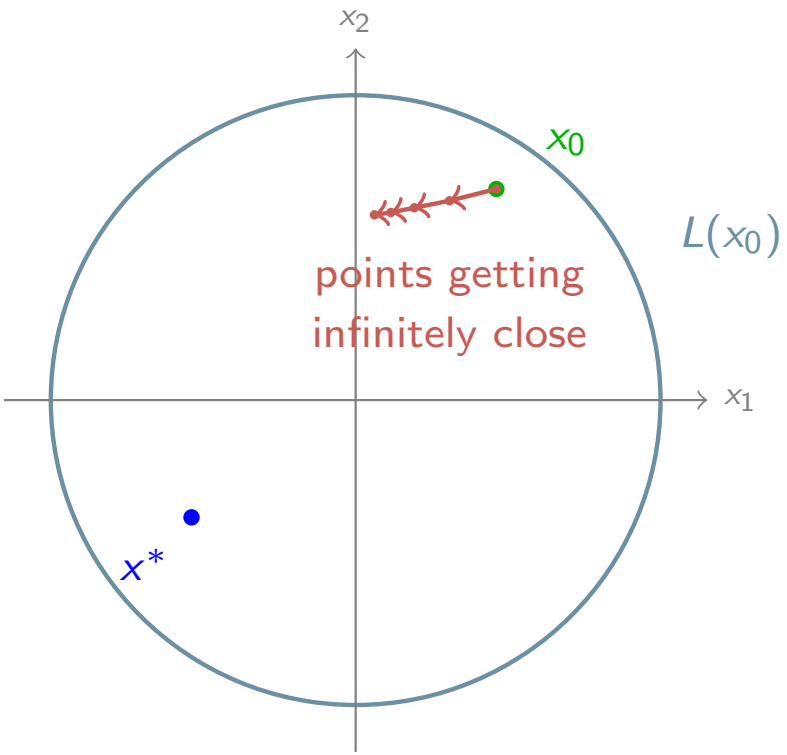


- Iterates lie on a *discrete mesh* with spacing $\bar{\alpha}$
- Finite number of points in compact level set $\Rightarrow$ algorithm must cycle
- Cycling contradicts monotone decrease $\Rightarrow$ $\bar{\alpha}$ cannot be fixed $\Rightarrow$ $\alpha_k \to 0$

What if directions change at every iteration?

Motivation: Exploit information from function evaluations to choose better directions

Problem: The discrete mesh structure is lost!

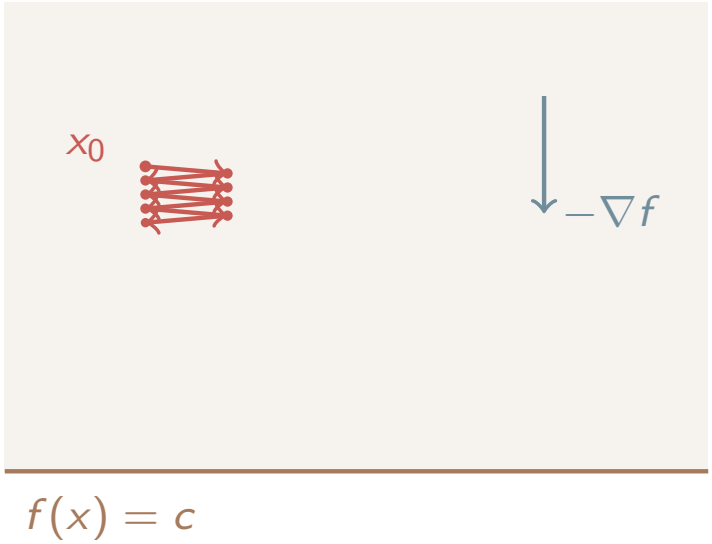


Without a mesh, points can get arbitrarily close without contradicting descent!

Example: linear function, varying directions

Setup: Consider $f(x_1, x_2) = -x_1 - x_2$ (linear, unbounded below on $\mathbb{R}^2$)

Restrict to level set $L(x_0) = \{(x_1, x_2) : x_1 + x_2 \geq -1\}$



8 steps of fixed length α, net descent $\to 0$

Each step reduces f (moves toward $-\nabla f$ direction), but steps become infinitesimally small!

Two critical issues

When directions vary, the mesh-based argument fails. Why?

1. **Points can be arbitrarily close**
   - No minimum spacing between consecutive iterates
   - Sequence can converge without exhausting a finite set

2. **Function decrease can be arbitrarily small**
   - Accepting $f(x_{k+1}) < f(x_k)$ allows infinitesimal improvements
   - Infinite descent does not imply $\alpha_k \to 0$

Solution: *Globalization strategies* ensure $\alpha_k \to 0$ by enforcing:

- **Sufficient distance** between iterates, OR
- **Sufficient decrease** in function value

Globalization strategies

Goal: Guarantee $\alpha_k \to 0$ when directions can vary

1. **Sufficient distance** (project onto artificial mesh)
   - Force iterates onto discrete grid (Mesh Adative Direct Search)
   - Recently proposed: rejection balls around visited points [state-of-the-art, 2025]
   - Not covered in this course (available as individual project topic)

2. **Sufficient decrease** (strengthen acceptance criterion)
   - Standard approach in derivative-free optimization
   - Replace $f(x_{k+1}) < f(x_k)$ with stronger condition:

$$\boxed{f(x_{k+1}) \leq f(x_k) - \rho(\alpha_k)}$$

     where $\rho : \mathbb{R}_+ \to \mathbb{R}_+$ is a *forcing function*
   - We focus on this approach

The forcing function $\rho(\alpha)$

Acceptance criterion:

$$f(x_{k+1}) \leq f(x_k) - \rho(\alpha_k)$$

Why must ρ depend on α_k?

The forcing function $\rho(\alpha)$

Acceptance criterion:

$$f(x_{k+1}) \leq f(x_k) - \rho(\alpha_k)$$

Why must ρ depend on α_k?

- Near a minimum, $\|\nabla f(x_k)\|$ is small $\Rightarrow$ achievable decrease is small
- If $\rho(\alpha_k)$ stays bounded away from zero, no trial point will be accepted
- Need $\rho(\alpha_k) \to 0$ as $\alpha_k \to 0$ to allow acceptance near minima

The forcing function $\rho(\alpha)$

Acceptance criterion:

$$f(x_{k+1}) \leq f(x_k) - \rho(\alpha_k)$$

Why must ρ depend on α_k?

- Near a minimum, $\|\nabla f(x_k)\|$ is small $\Rightarrow$ achievable decrease is small
- If $\rho(\alpha_k)$ stays bounded away from zero, no trial point will be accepted
- Need $\rho(\alpha_k) \to 0$ as $\alpha_k \to 0$ to allow acceptance near minima

Standard choice: quadratic forcing function

$$\rho(\alpha) = \gamma\alpha^2, \qquad \gamma > 0$$

The forcing function $\rho(\alpha)$

Acceptance criterion:

$$f(x_{k+1}) \leq f(x_k) - \rho(\alpha_k)$$

Why must ρ depend on α_k?

- Near a minimum, $\|\nabla f(x_k)\|$ is small $\Rightarrow$ achievable decrease is small
- If $\rho(\alpha_k)$ stays bounded away from zero, no trial point will be accepted
- Need $\rho(\alpha_k) \to 0$ as $\alpha_k \to 0$ to allow acceptance near minima

Standard choice: quadratic forcing function

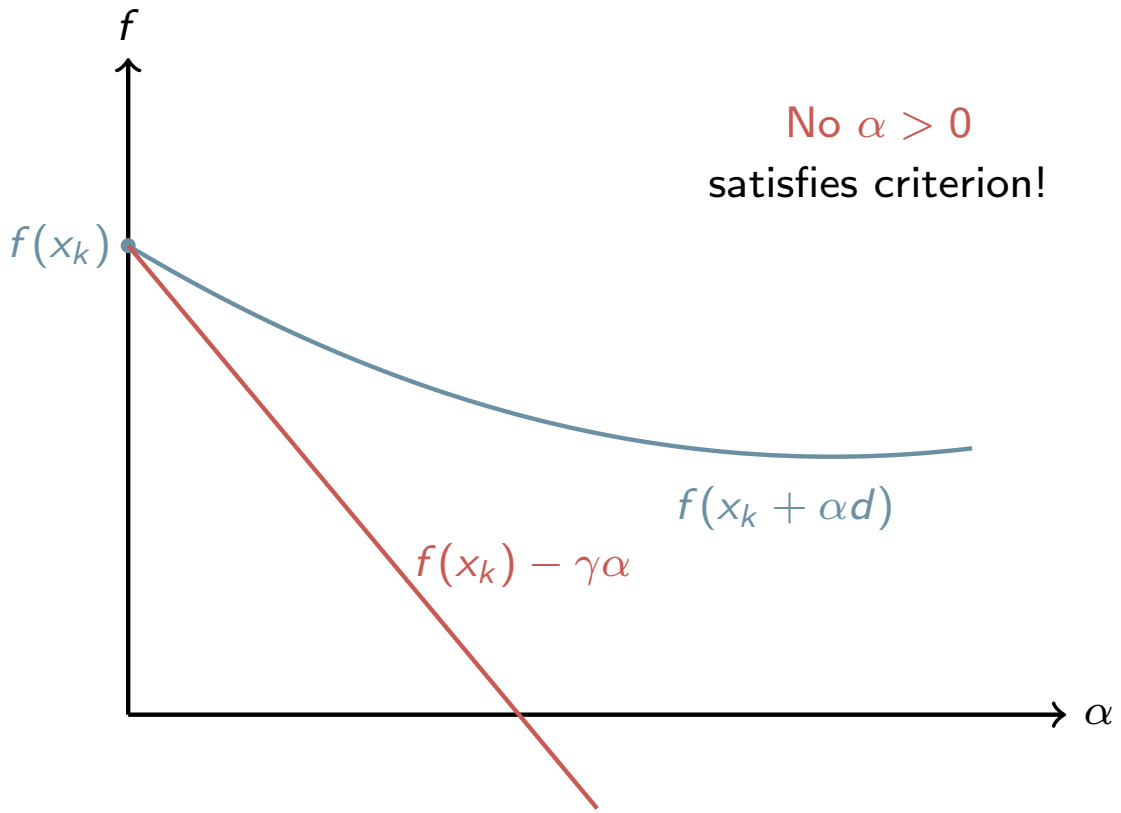
$$\rho(\alpha) = \gamma\alpha^2, \qquad \gamma > 0$$

Question asked in previous class: Why α^2? Why not linear $\rho(\alpha) = \gamma\alpha$?

Why not a linear forcing function?

Consider 1D search along descent direction d at x_k: minimize $\phi(\alpha) = f(x_k + \alpha d)$

Case 1: Linear forcing function $\rho(\alpha) = \gamma\alpha$

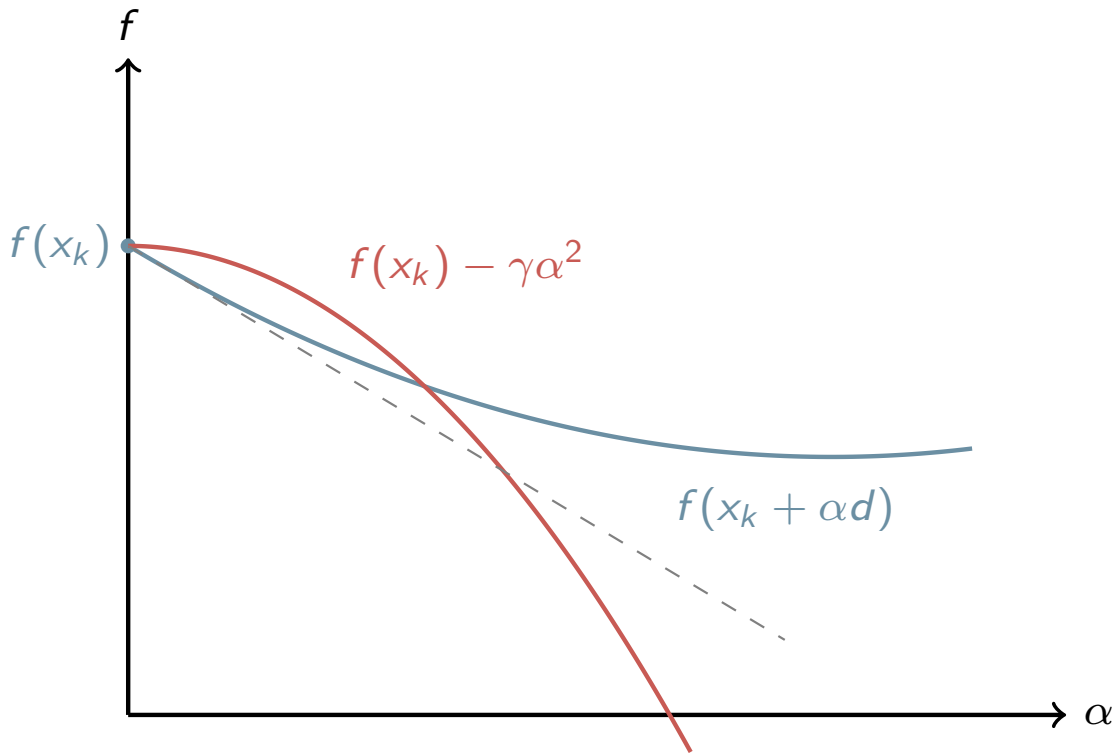


If γ is too large, $f(x_k) - \gamma\alpha$ decreases *faster* than $\phi(\alpha)$ initially

$\Rightarrow$ **no step-size accepted**, even though d is a descent direction!

Why quadratic forcing function works

Case 2: Quadratic forcing function $\rho(\alpha) = \gamma\alpha^2$



Key property: $\rho'(0) = 0$ (horizontal tangent at origin)

For small α, $\phi(\alpha) \approx f(x_k) + \nabla f(x_k)^\top(\alpha d) < f(x_k) - \gamma\alpha^2$

$\Rightarrow$ **Acceptance guaranteed for sufficiently small** α along descent direction!

Direct Search

Input: $x_0 \in \mathbb{R}^n$, $\alpha_0 > 0$, $\alpha_{\min} > 0$, $\theta \in (0,1)$, $\gamma > 0$

1: $k \leftarrow 0$
2: **while** $\alpha_k \geq \alpha_{\min}$ **do**
3: $\quad k \leftarrow k+1$
4: $\quad$ **Generate** a set of poll directions $\mathcal{D}_k$
5: $\quad$ **if** $d \in \mathcal{D}_k$ exists such that $f(x_k + \alpha_k d) \leq f(x_k) - \gamma\alpha_k^2$ **then**
6: $\quad\quad x_{k+1} \leftarrow x_k + \alpha_k d$, $\alpha_{k+1} \leftarrow \alpha_k$
7: $\quad$ **else**
8: $\quad\quad x_{k+1} \leftarrow x_k$, $\alpha_{k+1} \leftarrow \theta\alpha_k$
9: $\quad$ **end if**
10: **end while**
11: **return** x_k

Key differences from coordinate search:

- Directions d_k can vary (not restricted to $\{\pm e_i\}$)
- Acceptance uses *sufficient decrease* instead of simple descent

Does sufficient decrease ensure $\alpha_k \to 0$?

Question: With varying directions and sufficient decrease criterion, do we recover $\alpha_k \to 0$?

Does sufficient decrease ensure $\alpha_k \to 0$?

Question: With varying directions and sufficient decrease criterion, do we recover $\alpha_k \to 0$?

Intuition:

- If $\alpha_k \not\to 0$, then $\alpha_k \geq \bar{\alpha} > 0$ for infinitely many k
- Sufficient decrease: $f(x_{k+1}) \leq f(x_k) - \gamma\bar{\alpha}^2$ infinitely often
- This gives $f(x_k) \to -\infty$, contradicting boundedness of f on compact level set
- Therefore $\alpha_k \to 0$

Next step: Prove convergence to stationary points under sufficient decrease