Descent Algorithms, Line Search

Lecturer: Pradeep Ravikumar
Co-instructor: Aarti Singh

Convex Optimization 10-725/36-725

Unconstrained Minimization

$$x^* \in \arg\min_x f(x)$$

- To get to the optimal solution x^*, we typically use **iterative** algorithms
  - Compute sequence of iterates x_k that (hopefully) converge to x^* at a fast rate
  - x_{k+1} is some (simple) function of f, previous iterates

Two Classes of Iterative Algorithms

- Descent + Line Search Algorithms

Iteratively find directions p_k,
and (approximately) solve for $\min_{\alpha>0} f(x_k + \alpha\, p_k)$

- Trust Region Algorithms

Iteratively solve $\min_p m_k(x_k + p)$
where $x_k + p$ lies in some "trust region"

for some approx. $m_k(\cdot)$ to the function $f(\cdot)$,
that is accurate in trust region

Descent Algorithms

Pick direction p_k such that

$$f(x_k + \alpha p_k) < f(x_k),$$

for some $\alpha > 0$.

- Many choices of such directions p_k
  - Gradient Descent
  - Conjugate Gradient
  - Newton
  - …

Which is the direction that is the "steepest" of them all?

- By Taylor's Theorem:

$$f(x_k + \alpha p) = f(x_k) + \alpha p^T \nabla f_k + \tfrac{1}{2}\alpha^2 p^T \nabla^2 f(x_k + tp)p, \qquad \text{for some } t \in (0, \alpha)$$

-

Which is the direction that is the "steepest" of them all?

- By Taylor's Theorem:

$$f(x_k + \alpha p) = f(x_k) + \alpha p^T \nabla f_k + \tfrac{1}{2}\alpha^2 p^T \nabla^2 f(x_k + tp)p, \qquad \text{for some } t \in (0, \alpha)$$

- Rate of change of f along direction p:

$$\lim_{\alpha \to 0} \frac{f(x_k + \alpha p) - f(x_k)}{\alpha}$$
$$= p^T \nabla f_k$$

Which is the direction that is the "steepest" of them all?

- By Taylor's Theorem:

$$f(x_k + \alpha p) = f(x_k) + \alpha p^T \nabla f_k + \tfrac{1}{2}\alpha^2 p^T \nabla^2 f(x_k + tp)p, \qquad \text{for some } t \in (0, \alpha)$$

- Rate of change of f along direction p:

$$\lim_{\alpha \to 0} \frac{f(x_k + \alpha p) - f(x_k)}{\alpha}$$
$$= p^T \nabla f_k$$

- Unit direction p with most rapid decrease:

$$\min_p \; p^T \nabla f_k, \qquad \text{subject to } \|p\| = 1.$$

Which is the direction that is the "steepest" of them all?

- By Taylor's Theorem:

$$f(x_k + \alpha p) = f(x_k) + \alpha p^T \nabla f_k + \frac{1}{2}\alpha^2 p^T \nabla^2 f(x_k + tp)p, \qquad \text{for some } t \in (0, \alpha)$$

- Rate of change of f along direction p:

$$\lim_{\alpha \to 0} \frac{f(x_k + \alpha p) - f(x_k)}{\alpha}$$
$$= p^T \nabla f_k$$

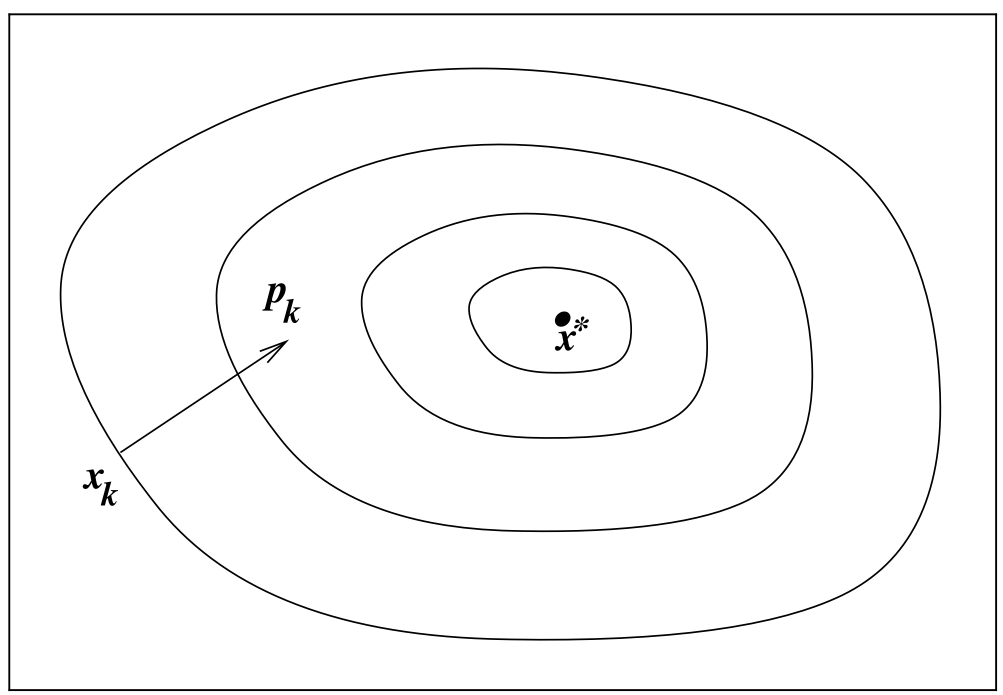
- Unit direction p with most rapid decrease:

$$p = -\nabla f_k / \|\nabla f_k\|$$

Steepest Descent is Gradient Descent

- Iteratively descend in direction:

$$p = -\nabla f_k / \|\nabla f_k\|$$



- Will study in depth in next class

Can we characterize “descent” directions?

Iteratively find directions p_k,
and (approximately) solve for $\min_{\alpha>0} f(x_k + \alpha\, p_k)$

- Taylor’s Theorem:

$$f(x_k + \epsilon p_k) = f(x_k) + \epsilon p_k^T \nabla f_k + O(\epsilon^2).$$

Can we characterize "descent" directions?

Iteratively find directions p_k,
and (approximately) solve for $\min_{\alpha > 0} f(x_k + \alpha\, p_k)$

- Taylor's Theorem:

$$f(x_k + \epsilon p_k) = f(x_k) + \epsilon p_k^T \nabla f_k + O(\epsilon^2).$$

- Suppose angle between p_k and \grad f_k is \theta_k, and cos(\theta_k) < 0 i.e. angle is strictly less than 90 degrees

$$\Rightarrow \quad p_k^T \nabla f_k = \|p_k\| \, \|\nabla f_k\| \cos\theta_k < 0.$$

$$\Rightarrow \quad f(x_k + \epsilon p_k) < f(x_k)$$

Can we characterize "descent" directions?

Iteratively find directions p_k,
and (approximately) solve for $\min_{\alpha>0} f(x_k + \alpha\, p_k)$

- Taylor's Theorem:

$$f(x_k + \epsilon p_k) = f(x_k) + \epsilon p_k^T \nabla f_k + O(\epsilon^2).$$

- Suppose angle between p_k and \grad f_k is \theta_k, and cos(\theta_k) < 0 i.e. angle is strictly less than 90 degrees

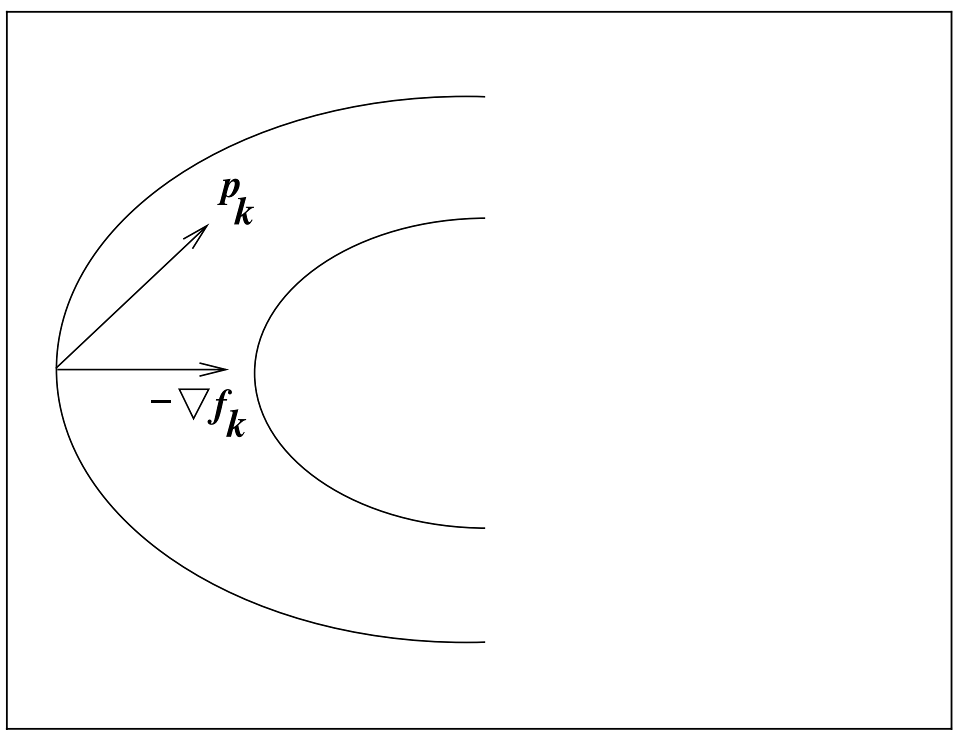
$$\Longrightarrow \quad p_k^T \nabla f_k = \|p_k\| \, \|\nabla f_k\| \cos\theta_k < 0.$$

$$\Longrightarrow \quad f(x_k + \epsilon p_k) < f(x_k)$$

- Any "downhill" direction is a descent direction

Can we characterize “descent” directions?

Iteratively find directions p_k,
and (approximately) solve for $\min_{\alpha>0} f(x_k + \alpha\, p_k)$



Downhill direction p_k

Step-size Selection

- Iterates: $x_{k+1} = x_k - \alpha_k \, p_k$
- Suppose we have a strategy to iteratively pick the descent directions p_k (e.g. steepest i.e. negative gradient)
- How to pick the **step-size \alpha_k**?

Line Search

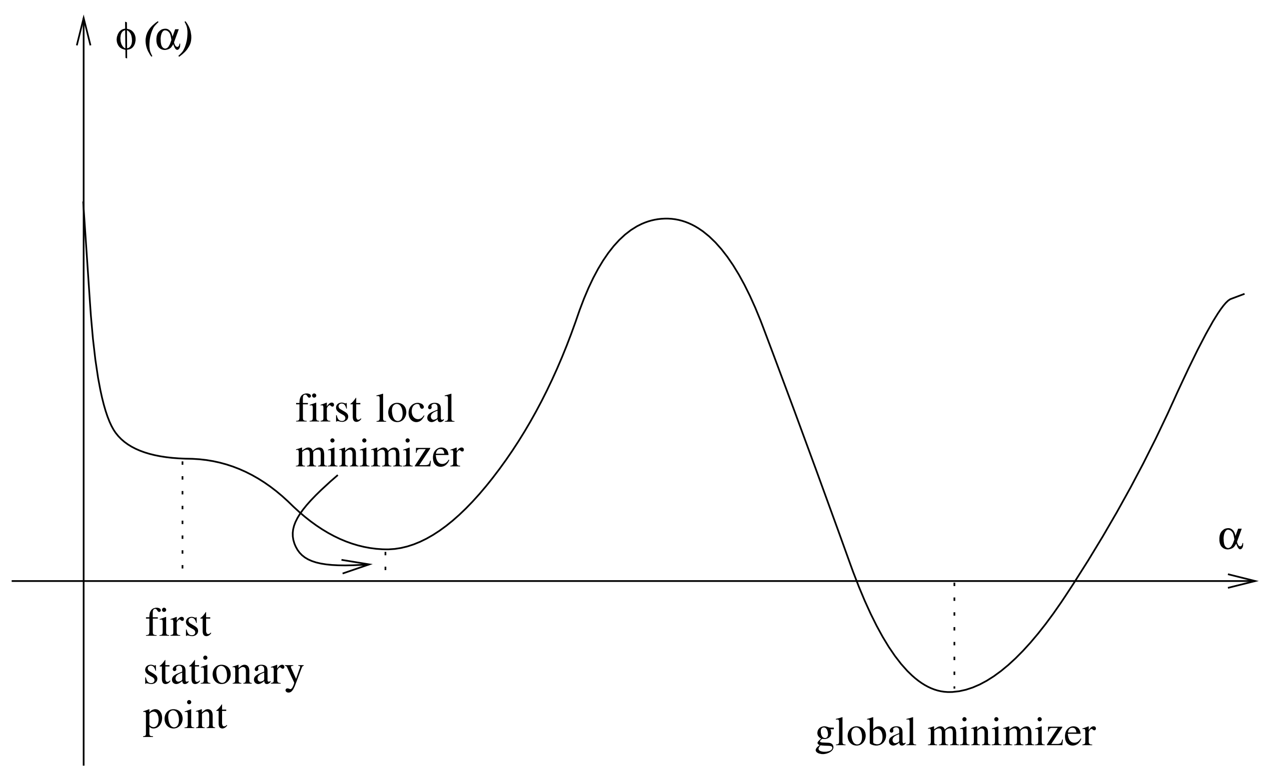
- Picking the step-size reduces to a one-dimensional optimization also called “line search”

$$\text{Let } \phi(\alpha) = f(x_k + \alpha\, p_k), \quad \alpha > 0.$$

$$\text{Line Search: } \min_{\alpha > 0} \phi(\alpha).$$

Exact Line Search

Solve for global minimum: $\min_{\alpha>0} \phi(\alpha)$.



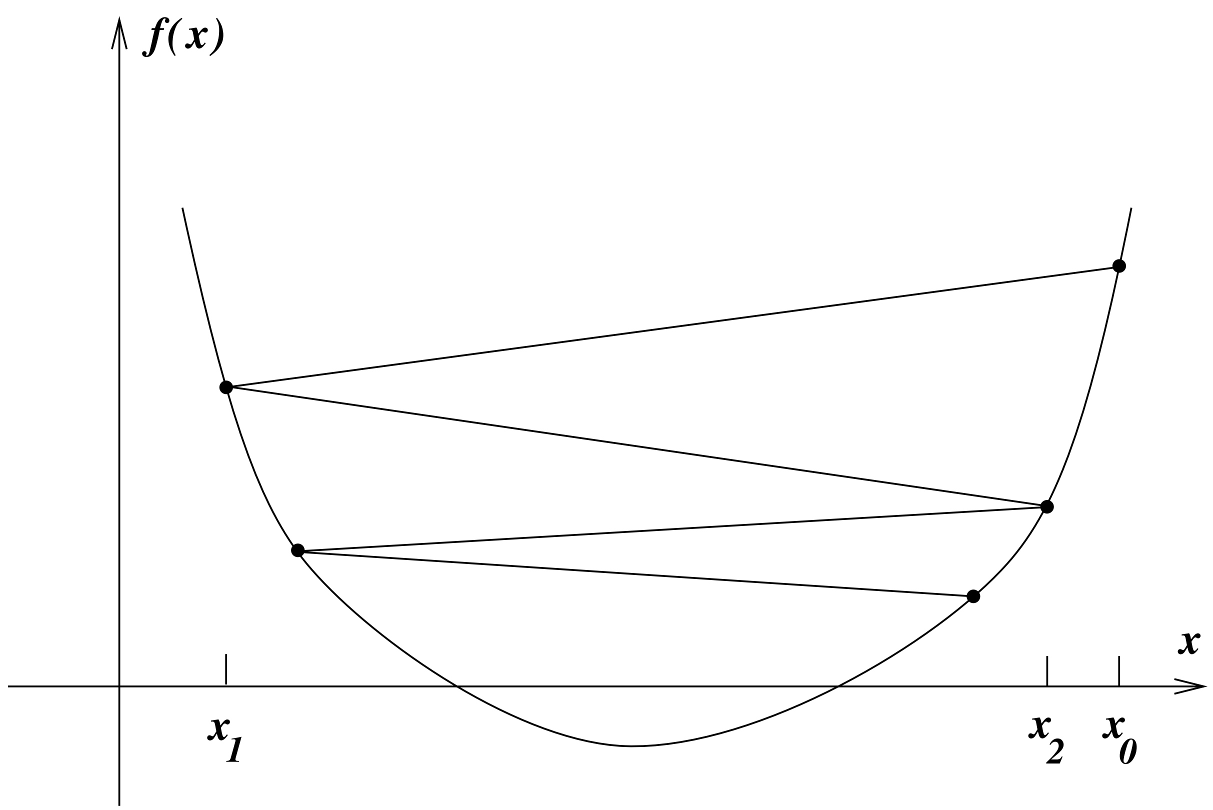
- One-dimensional non-convex optimization problem
- Might be too expensive

Inexact line search

- Solve for the optimization min_{alpha > 0} \phi(alpha) approximately and cheaply
- Question: is it sufficient to obtain an alpha that strictly?

Inexact line search

- Question: is it sufficient to obtain an alpha that strictly?

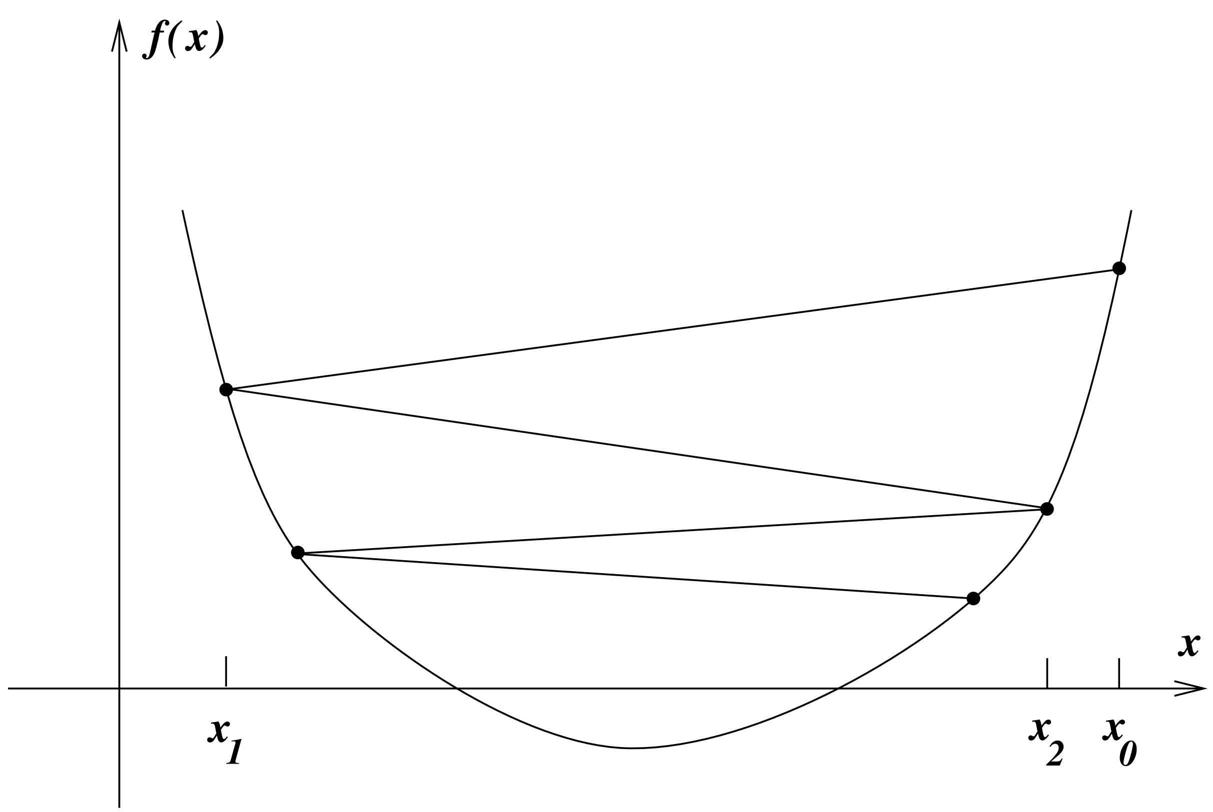


Min. function value: -1

Consider iterates x_k
s.t. f(x_k) = 5/k

Inexact line search

- Question: is it sufficient to obtain an alpha that strictly?



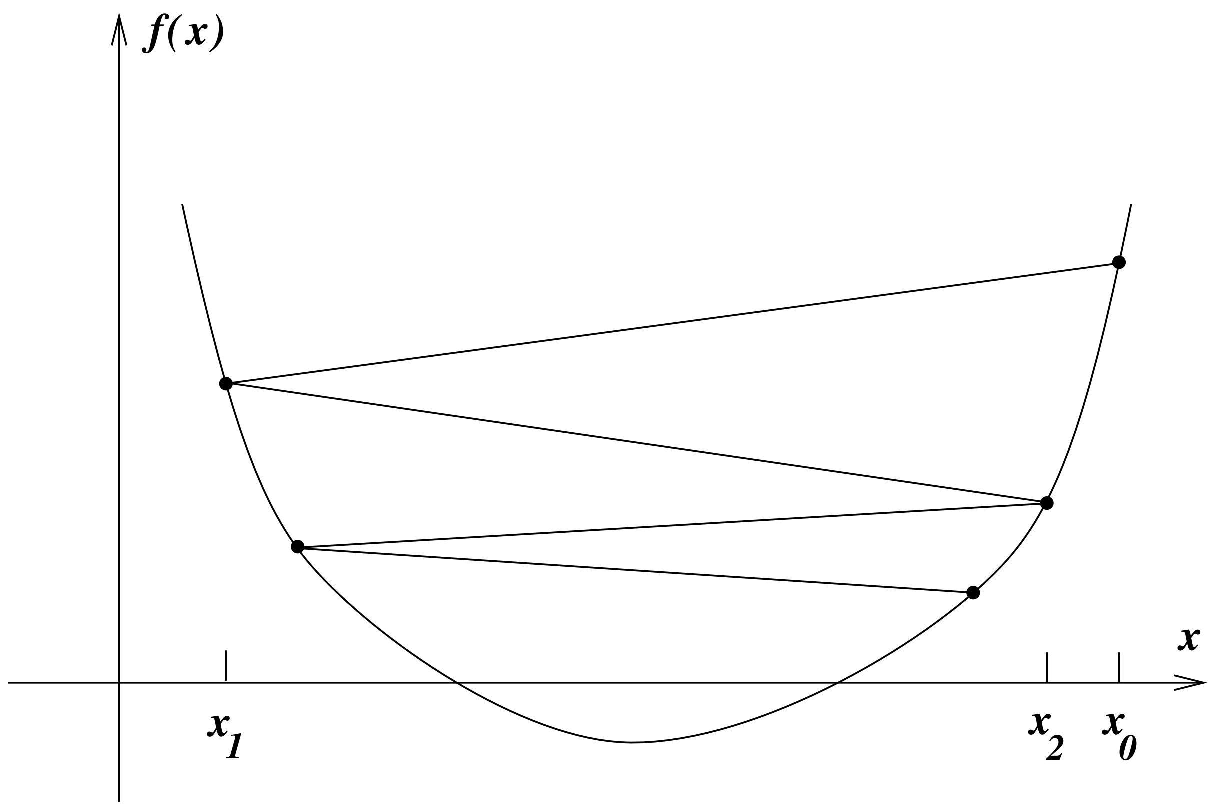
Min. function value: -1

Consider iterates x_k s.t. f(x_k) = 5/k

Each iterate results in strict function decrease

Inexact line search

- Question: is it sufficient to obtain an alpha that strictly?



Consider iterates x_k
s.t. f(x_k) = 5/k

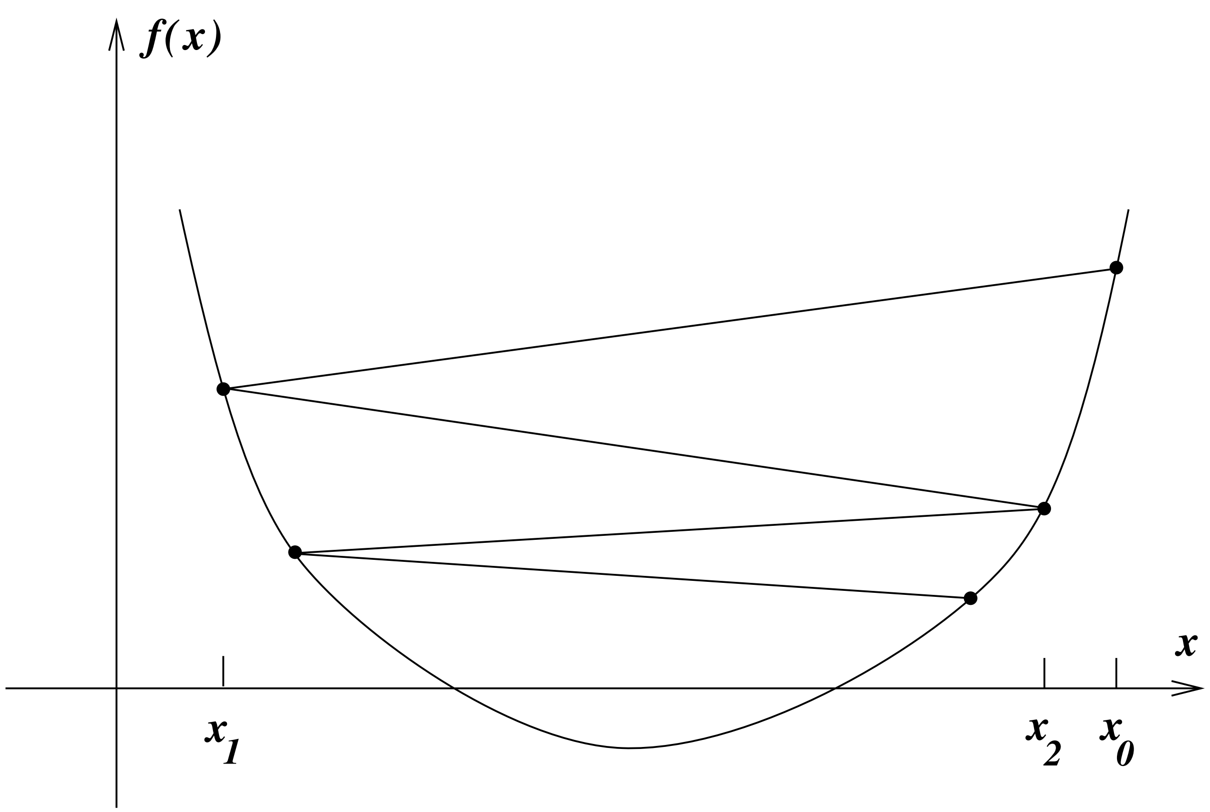
Each iterate results in
strict function decrease

But f(x_k) converges to zero,
which is greater than
min. value which is -1

Inexact line search

- Solve for the optimization min_{alpha > 0} \phi(alpha) approximately and cheaply

- Question: is it sufficient to obtain an alpha that strictly?



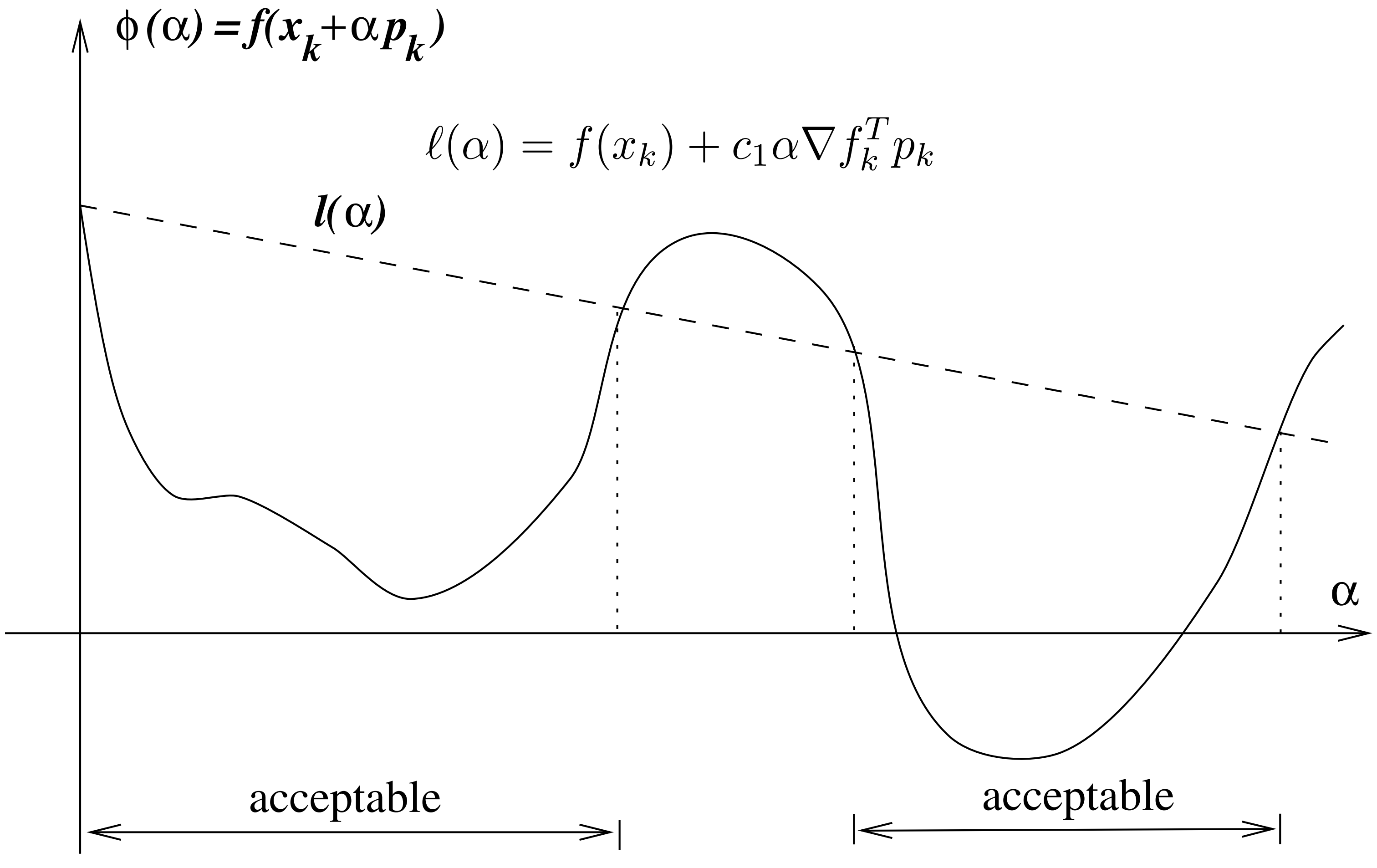
Answer: No

We need “sufficient” decrease

Armijo Condition

$$f(x_k + \alpha p_k) \leq f(x_k) + c_1 \alpha \nabla f_k^T p_k$$

for some constant c_1 in (0,1)



Backtracking Line Search with Armijo Condition

Procedure (Backtracking Line Search).

Choose $\bar{\alpha} > 0$, $\rho, c \in (0, 1)$; set $\alpha \leftarrow \bar{\alpha}$;

repeat until $f(x_k + \alpha p_k) \leq f(x_k) + c\alpha \nabla f_k^T p_k$

$\alpha \leftarrow \rho\alpha$;

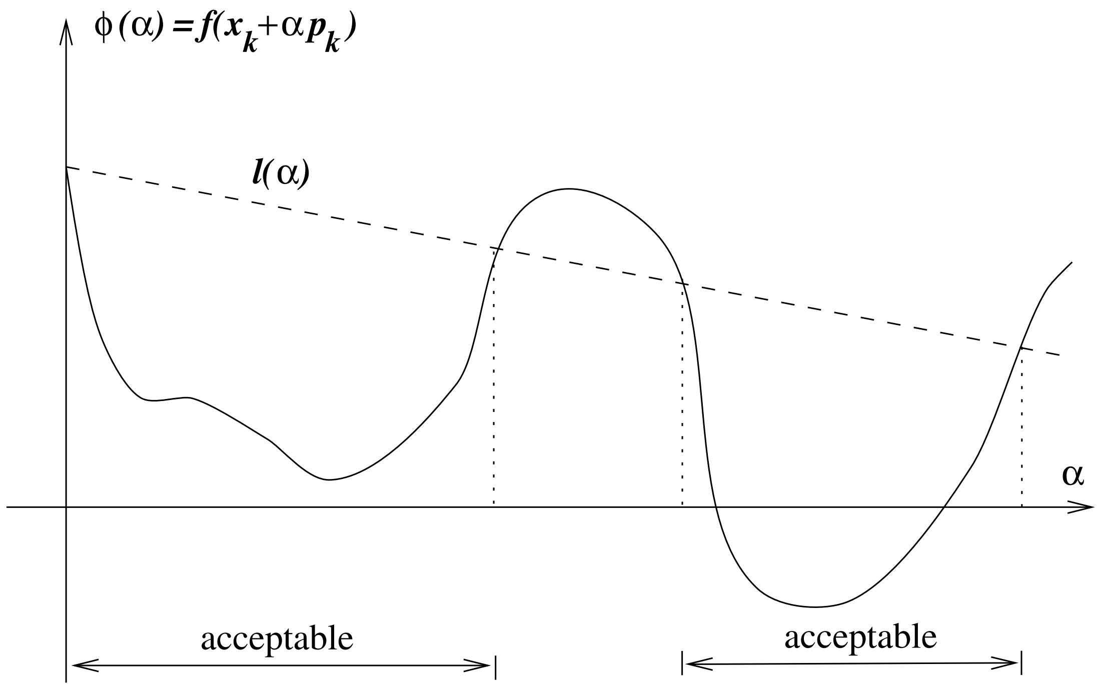
end (repeat)

Terminate with $\alpha_k = \alpha$.

- Start from a large step-size, and keep reducing by constant factor till it satisfies Armijo condition
- Typically can show similar theoretical results for this backtracking search as for exact line search
- Loosely: the step-sizes are small enough, but not too small: since a step-size that is a factor \rho larger violates the sufficient decrease condition

Decrease Condition not Sufficient

- Just a sufficient decrease (Armijo condition) is typically not sufficient

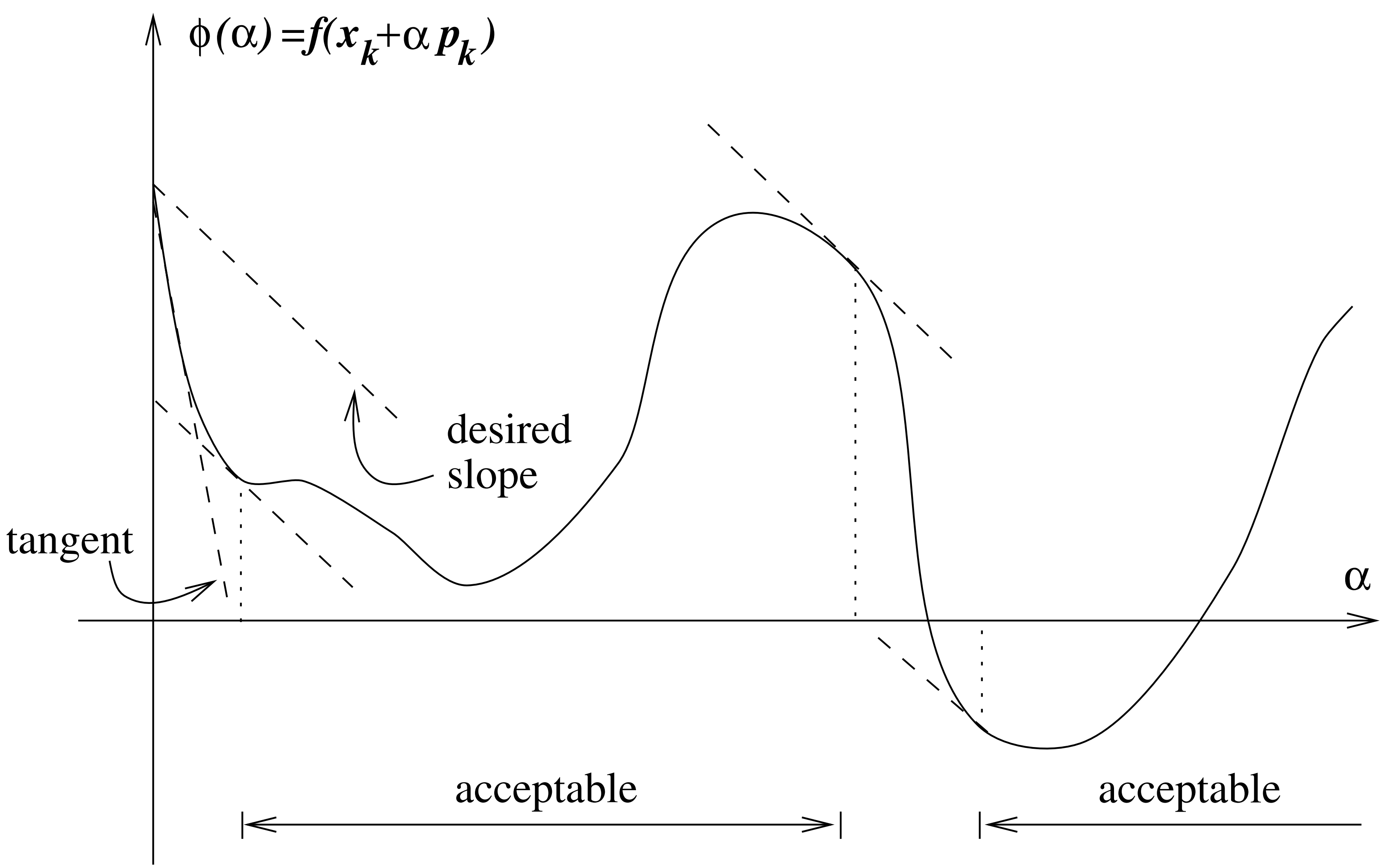


- We can see that by noting very small values of alpha also satisfy the Armijo condition
  - Backtracking partially addresses this by starting from large step-sizes and checking Armijo condition
  - But is there some other condition that we can add to Armijo?

Curvature Condition

$$\nabla f(x_k + \alpha_k p_k)^T p_k \geq c_2 \nabla f_k^T p_k,$$

for some constant $c_2 \in (c_1, 1)$



- Loosely, this says that the slope at alpha_k should be larger than at alpha = zero
- since slope at zero is negative, this entails that the slope be flatter e.g. closer to local/ global minimum

Wolfe Conditions

$$f(x_k + \alpha_k p_k) \leq f(x_k) + c_1 \alpha_k \nabla f_k^T p_k,$$

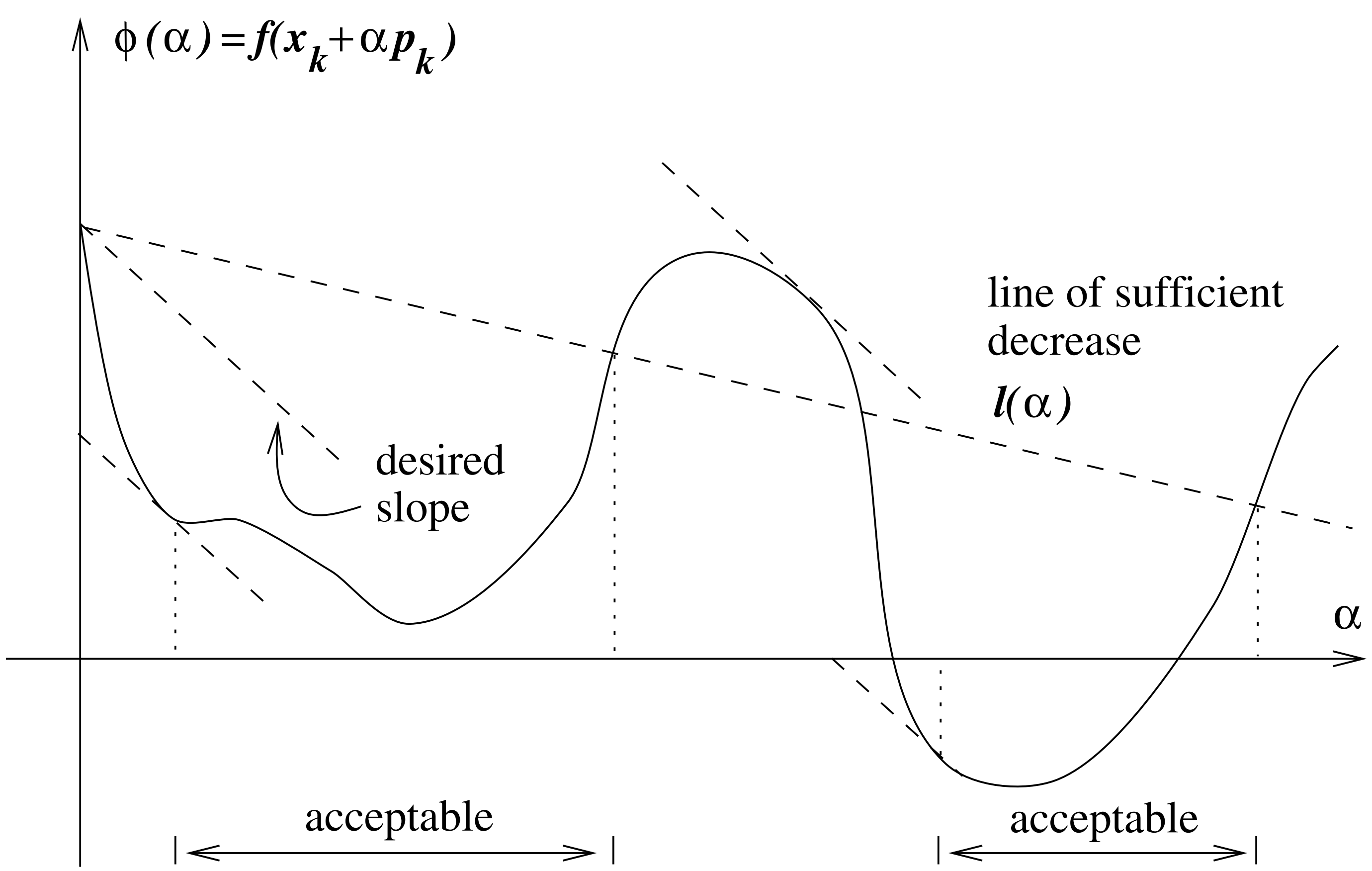
$$\nabla f(x_k + \alpha_k p_k)^T p_k \geq c_2 \nabla f_k^T p_k,$$

- Armijo and curvature conditions together
- Can show that **there always exist alpha** that satisfies Wolfe conditions
- Can provide **unified convergence analyses** for any step-size selection algorithm that satisfies Wolfe conditions

Wolfe Conditions

$$f(x_k + \alpha_k p_k) \leq f(x_k) + c_1 \alpha_k \nabla f_k^T p_k,$$

$$\nabla f(x_k + \alpha_k p_k)^T p_k \geq c_2 \nabla f_k^T p_k,$$



Zoutendijk Theorem

Loosely, for sufficiently well-behaved functions f, any descent algorithm with line search satisfying Wolfe conditions, satisfies:

$$\sum_{k \geq 0} \cos^2 \theta_k \; \|\nabla f_k\|^2 < \infty.$$

- Implies: $\cos^2 \theta_k \|\nabla f_k\|^2 \to 0.$
- If $\cos \theta_k \geq \delta > 0, \quad \text{for all } k.$

$$\Rightarrow \quad \lim_{k \to \infty} \|\nabla f_k\| = 0.$$

Strong Wolfe Conditions

$$f(x_k + \alpha_k p_k) \leq f(x_k) + c_1 \alpha_k \nabla f_k^T p_k,$$
$$|\nabla f(x_k + \alpha_k p_k)^T p_k| \leq c_2 |\nabla f_k^T p_k|,$$

- Improves curvature condition:
  - Rules out positive slopes i.e. strictly asks for flatter slope at alpha_k than at zero so that hopefully around local minimum of line search optimization problem

(Strong) Wolfe Condition Algorithms

- Armijo condition ensures sufficient decrease
- Curvature condition ensures that step-size is not too small (otherwise won't make enough progress)
- Backtracking algorithm introduced earlier finesses need for curvature condition by starting from large step-size and iteratively reducing step-size
    - not guaranteed to satisfy Wolfe conditions per se
- Algorithms targeted to satisfying Wolfe conditions tricky to code, even trickier to analyze