# Smooth Unconstrained Optimization

Antonio Frangioni

Department of Computer Science
University of Pisa
https://www.di.unipi.it/~frangio
mailto:frangio@di.unipi.it

Optimization Methods and Game Theory
Master in Artificial Intelligence and Data Engineering
University of Pisa

A.Y. 2022/23

## Outline

A Bird's Eye on Unconstrained Optimization

Local Optimization Algorithms

Gradient (steepest descent) methods

Gradient method for Quadratic Functions

Gradient method with fixed stepsize

Gradient method and (almost) exact Line Search

Inexact Line Search: Armijo-Wolfe

Wrap up & References

Solutions

## Optimization Problems

- $f : \mathbb{R}^n \to \mathbb{R}$ any function: objective function of

  $(P) \quad f_* = \min\{\, f(\, x\,) \;:\; x \in \mathbb{R}^n \,\}$ unconstrained optimization problem

- "min" w.l.o.g.: $\min\{\, f(\, x\,) \,:\, x \in \mathbb{R}^n \,\} = -\max\{\, -f(\, x\,) \,:\, x \in \mathbb{R}^n \,\}$, (but $\min\{\, f(\, x\,)\,\} \neq \max\{\, f(\, x\,)\,\}$, often rather different problems)

- $f_* = \nu(\, P\,)$ optimal value (unique if $\exists$, which it may not)

## Optimization Problems

- $f : \mathbb{R}^n \to \mathbb{R}$ any function: objective function of

  $(P) \quad f_* = \min\{\, f(\,x\,) \;:\; x \in \mathbb{R}^n \,\}$ unconstrained optimization problem

- "min" w.l.o.g.: $\min\{\, f(\,x\,) \,:\, x \in \mathbb{R}^n \,\} = -\max\{\, -f(\,x\,) \,:\, x \in \mathbb{R}^n \,\}$, (but $\min\{\, f(\,x\,)\,\} \neq \max\{\, f(\,x\,)\,\}$, often rather different problems)

- $f_* = \nu(\,P\,)$ optimal value (unique if $\exists$, which it may not)

- In fact, the problem is $\quad (P) \quad x_* \in \operatorname{argmin}\{\, f(\,x\,) \;:\; x \in \mathbb{R}^n \,\}$

- $x_*$ s.t. $f_* = f(\,x_*\,) \leq f(\,x\,) \;\; \forall x \in \mathbb{R}^n$ optimal solution (if $\exists$, which it may not)

## Optimization Problems

- $f : \mathbb{R}^n \to \mathbb{R}$ any function: objective function of

  (P) $\quad f_* = \min\{ f(x) \ : \ x \in \mathbb{R}^n \}$ unconstrained optimization problem

- "min" w.l.o.g.: $\min\{ f(x) \ : \ x \in \mathbb{R}^n \} = - \max\{ -f(x) \ : \ x \in \mathbb{R}^n \}$, (but $\min\{ f(x) \} \neq \max\{ f(x) \}$, often rather different problems)

- $f_* = \nu(P)$ optimal value (unique if $\exists$, which it may not)

- In fact, the problem is (P) $\quad x_* \in \operatorname{argmin} \{ f(x) \ : \ x \in \mathbb{R}^n \}$

- $x_*$ s.t. $f_* = f(x_*) \leq f(x) \ \forall x \in \mathbb{R}^n$ optimal solution (if $\exists$, which it may not)

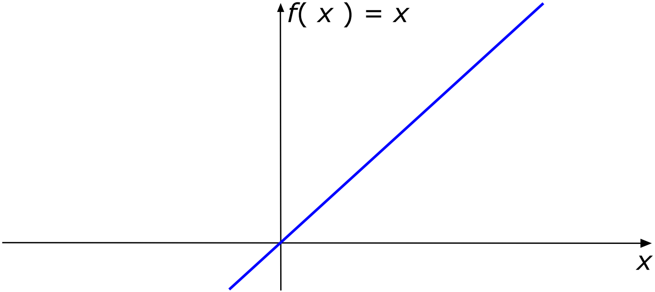
- $x_*$ may not be unique: $\exists\, x' \neq x_* \in X_* \subseteq \mathbb{R}^n$ set of optimal solutions happens but we don't care (mostly): all optimal solutions equivalent "in the eyes of $f$"

- Impossible in general: $f$ non computable function, ...

- Let's start "easy": $n = 1$, an (efficient, pointwise) oracle for $f$ available

## Still, optimization may be impossible I

- Impossible if $f$ has no minimum



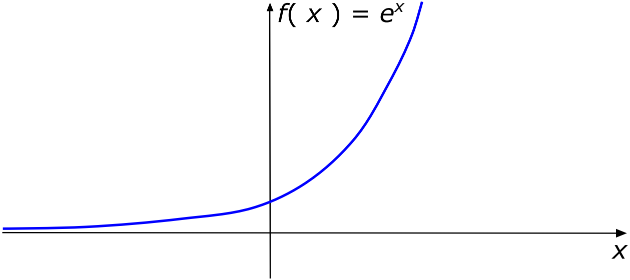
- $(P)$ unbounded (below): $\nu( P ) = -\infty$, $X_* = \emptyset$

- Solving $(P)$ actually (at least) two different things:
  - finding $x_*$ and proving it is optimal (how??)
  - constructively proving $f$ unbounded below (how??)

- Hardly ever happens in ML (since $\mathcal{L}( w ) \geq 0$ and $R( w ) \geq 0$)

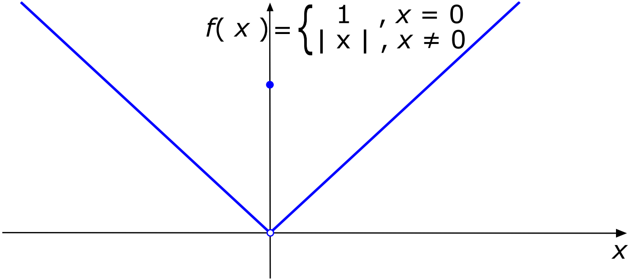
# Still, optimization may be impossible II

- Impossible if $f_* > -\infty$ but $\nexists\, x_* \;\equiv\; X_* = \emptyset$



# Still, optimization may be impossible II

- Impossible if $f_* > -\infty$ but $\nexists\, x_* \;\equiv\; X_* = \emptyset$



- However, plenty of $\varepsilon$-approximate solutions ($\varepsilon$-minima)

$$f(x_\varepsilon) \leq f_* + \varepsilon \quad \forall\, \varepsilon > 0$$

- On computers "$x \in \mathbb{R}$" actually is "$x \in \mathbb{Q}$" with up to 16 digits precision $\Longrightarrow$ approximation errors unavoidable anyway
- Exact algebraic computation possible but usually too slow
- ML actually going the opposite way (float, half, small integer weights . . . )
- Anyway, finding the exact $x_*$ impossible in general [4, p. 408]

## Optimization need be approximate

- Absolute gap: $A(x) = f(x) - f_* \; (\geq 0)$

- Relative gap: $R(x) = (f(x) - f_*) / |f_*| = A(x) / |f_*| \; (\geq 0)$

- Why $R(x)$? Because $\forall\, \alpha > 0 \quad (P) \;\equiv\; (P_\alpha) \;\; \min\{\, \alpha f(x) : \dots \}$
  same $x^*$, $\nu(P_\alpha) = \alpha \nu(P) \implies R(x)$ same, $A(x)$ not

- (Approximately) solve $(P)$: fix $\varepsilon$, find $x$ s.t. $A(x) \leq \varepsilon$ / $R(x) \leq \varepsilon$

- Issue: computing $A(x) / R(x)$ requires $f_*$ which is typically unkown

- Could argue this is "the issue" in optimization: compute (an estimate of) $f_*$

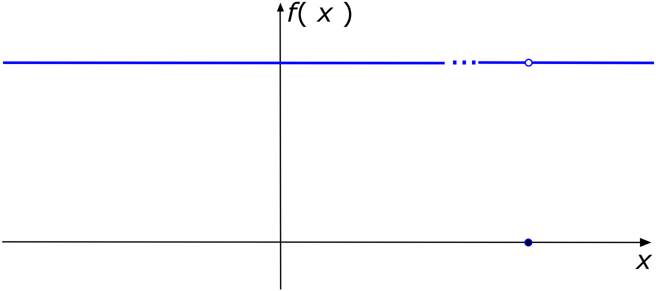
- Sometimes $\approx$ known in ML ($f_* \approx 0$ in NN) but not always (not in SVM)

- A real issue only if true optimum $x_*$ needed, hence typically not in ML

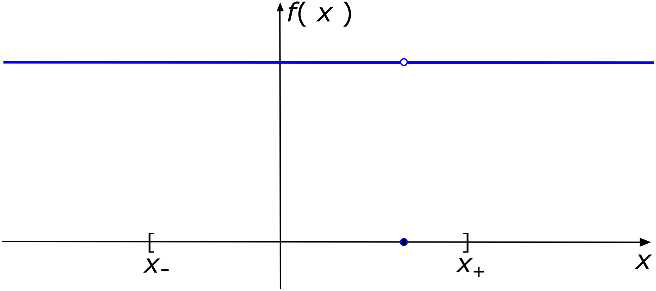
**Exercise:** $R(x)$ ill-defined if $f_* = 0$, propose solutions & justify them (change $f_*$)

# Even approximate, optimization is still impossible

- Impossible because isolated minima can be anywhere



- Impossible because isolated minima can be anywhere



- Does it help restricting to $x \in X = [\,x_- \,,\, x_+\,]$ $(-\infty < x_- < x_+ < +\infty)$?

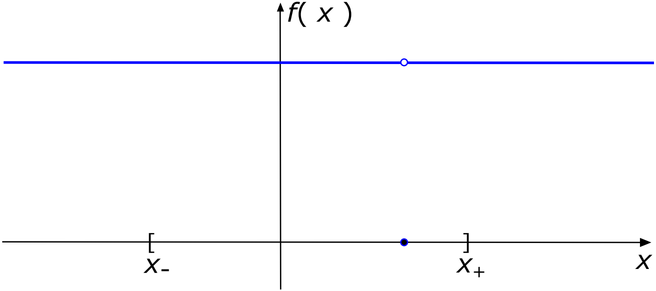
- No: still uncountably many points to try

**Exercise:** w.l.o.g. $x_- = 0$ and $x_+ = 1$: why?

**Exercise:** What could go wrong with $X = (\,x_- \,,\, x_+\,)$? Does it really matter?

# Even approximate, optimization is still impossible

- Impossible because isolated minima can be anywhere



- Does it help restricting to $x \in X = [\, x_- \, , \, x_+ \,]$ $(-\infty < x_- < x_+ < +\infty)$?

- No: still uncountably many points to try

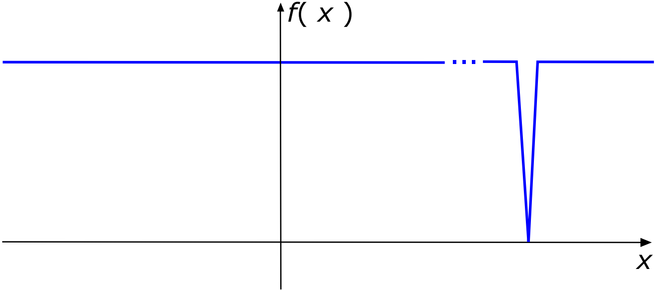
**Exercise:** w.l.o.g. $x_- = 0$ and $x_+ = 1$: why?

**Exercise:** What could go wrong with $X = (\, x_- \, , \, x_+ \,)$? Does it really matter?

- Is it because $f$ "jumps"?

## Even approximate, optimization is still impossible

- Impossible because isolated minima can be anywhere



- Does it help restricting to $x \in X = [\, x_- \,,\, x_+ \,]$ $(-\infty < x_- < x_+ < +\infty)$?

- No: still uncountably many points to try

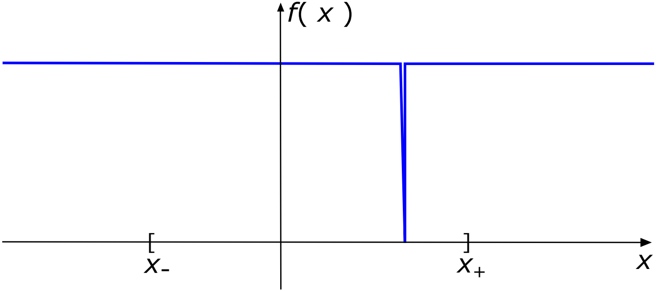
**Exercise:** w.l.o.g. $x_- = 0$ and $x_+ = 1$: why?

**Exercise:** What could go wrong with $X = (\, x_- \,,\, x_+ \,)$? Does it really matter?

- Is it because $f$ "jumps"? No, $f$ can have isolated ↓ spikes anywhere

# Even approximate, optimization is still impossible

- Impossible because isolated minima can be anywhere



- Does it help restricting to $x \in X = [\, x_- \,,\, x_+ \,]$ $(-\infty < x_- < x_+ < +\infty)$?

- No: still uncountably many points to try

**Exercise:** w.l.o.g. $x_- = 0$ and $x_+ = 1$: why?

**Exercise:** What could go wrong with $X = (\, x_- \,,\, x_+ \,)$? Does it really matter?

- Is it because $f$ "jumps"? No, $f$ can have isolated ↓ spikes anywhere

- ...even on $X = [\, x_- \,,\, x_+ \,]$ as spikes can be aribtrarily narrow

## Making optimization at least possible

- Impose $x \in [\, x_- \,,\, x_+ \,]$ with $D \,= x_+ - x_- < \infty$ (finite diameter)
- Impose spikes can't be arbitrarily narrow $\equiv$ $f$ cannot change too fast $\equiv$ $f$ Lipschitz continuous (L-c) on $X$ [6, p. 624]: $\exists L > 0$ s.t.
  $$|\, f(\, x\,) - f(\, z\,)\,| \leq L|\, x - z\,| \qquad \forall x\,,\, z \in X$$
- $f$ L-c $\implies$ $f$ does not "jump" (continuous)
- $f$ L-c $\implies$ one $\varepsilon$-optimum can be found with $O(\, LD \,/\, \varepsilon\,)$ evaluations: uniformly sample $X$ with step $2\varepsilon/L$ [4, p. 411] (**check**)
- Bad news: no algorithm can work in less than $\Omega(\, LD \,/\, \varepsilon\,)$ [4, p. 413]: # steps inversely proportional to accuracy, just not doable for "small" $\varepsilon$
- No free lunch theorem says "all algorithms equally bad" [8], i.e., "if an algorithm is very good in some cases it has to be very bad in others"
- $L$ unknown and difficult to estimate but algorithms actually need it
- Even very dramatically worse if $n \gg 1$, will see shortly

## Local optimality

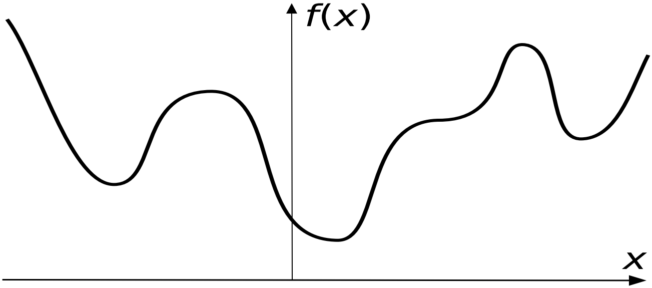
- Even if I should stumble in $x_*$, how do I recognize it?
- Turns out this is "the really difficult thing" (a.k.a. knowing $f_*$)
- Have to resort to weaker condition: $x_*$ is local minimum if it solves

$$\min\{\, f(\,x\,) \;:\; x \in X(\,x_*\,,\,\varepsilon\,) = [\,x_* - \varepsilon\,,\,x_* + \varepsilon\,]\,\} \qquad \text{for some } \varepsilon > 0$$

- Stronger notion: strict local minimum if $f(\,x_*\,) < f(\,z\,)\;\; \forall\, z \in X(\,x_*\,,\,\varepsilon\,) \setminus \{\,x_*\,\}$
- Why useful? Because "near $x_*$, $f$ typically has a predictable shape"
- $f$ (strictly) unimodal on $X$ if:
  - has minimum $x_* \in X$
  - it is (strictly) decreasing in $[\,x_-\,,\,x_*\,]$ and increasing in $[\,x_*\,,\,x_+\,]$
- $f$ unimodal in $[\,x_-\,,\,x_+\,] \implies$ can find $\varepsilon$-minimum in $O(\,LD\,/\,\log(\,\varepsilon\,)\,)$
  $\equiv$ exponentially faster (will see)

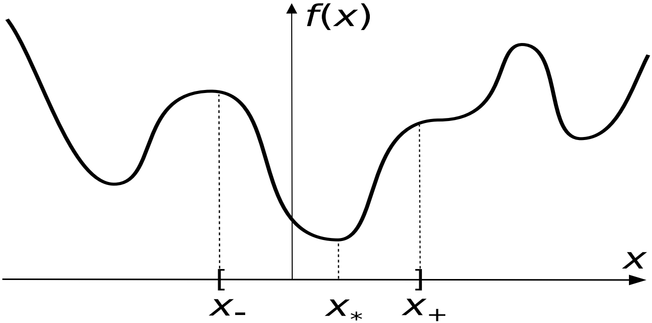
## Attraction Basins

- Most functions are not unimodal (although some are)



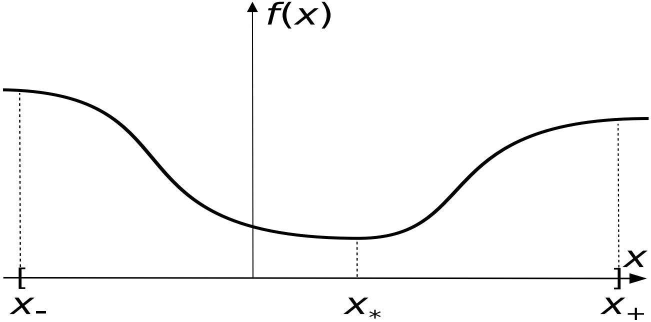
## Attraction Basins

- Most functions are not unimodal (although some are)



- But they are if you focus on the attraction basin of $x_*$ and

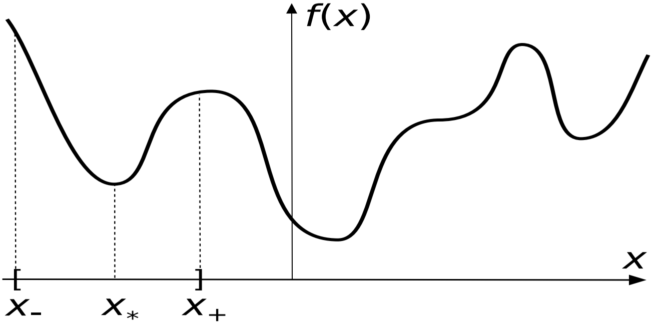
- Most functions are not unimodal (although some are)



- But they are if you focus on the attraction basin of $x_*$ and restrict there

# Attraction Basins

- Most functions are not unimodal (although some are)

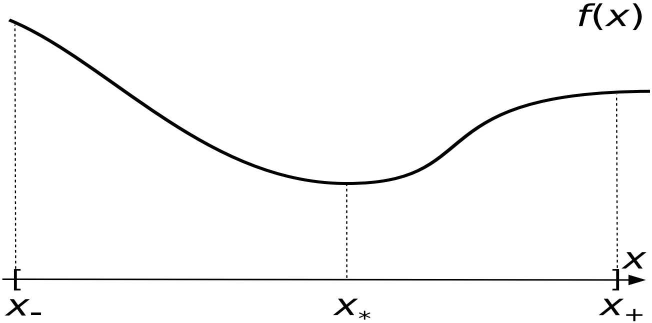


- But they are if you focus on the attraction basin of $x_*$ and restrict there

- Unfortunately, this is true for every local optimum

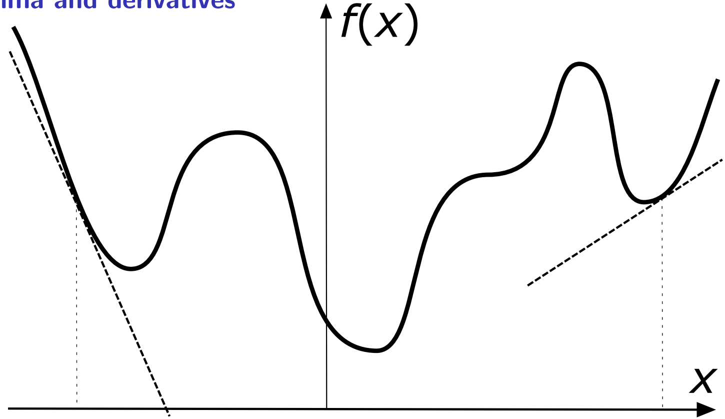
## Attraction Basins

- Most functions are not unimodal (although some are)



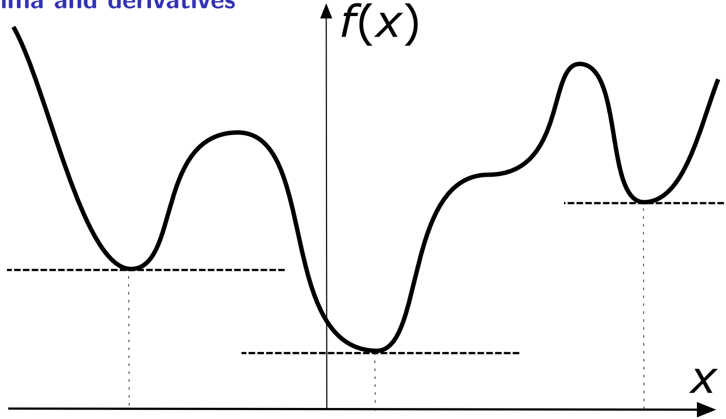
- But they are if you focus on the attraction basin of $x_*$ and restrict there
- Unfortunately, this is true for every local optimum
- All local optima "look the same", comprised the global one
- Yet, this makes it finding some local optimum a lot easier
- Finding the right (global) one another matter entirely

# Local minima and derivatives



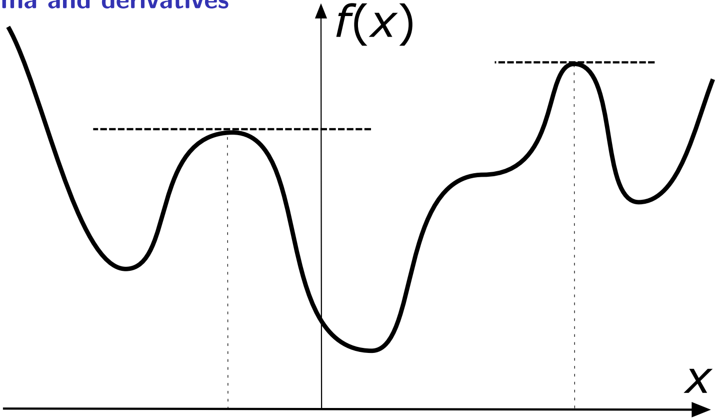
- If $f'(x) < 0$ or $f'(x) > 0$, $x$ clearly cannot be a local minimum

# Local minima and derivatives



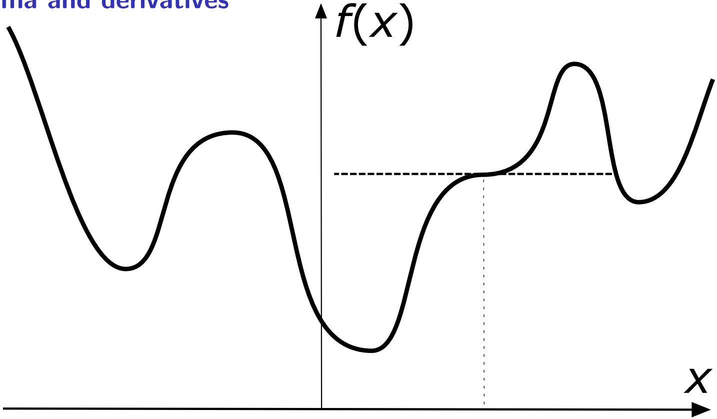
- If $f'(x) < 0$ or $f'(x) > 0$, $x$ clearly cannot be a local minimum
- Hence, $f'(x) = 0$ in all local minima (hence in the global one as well)

# Local minima and derivatives



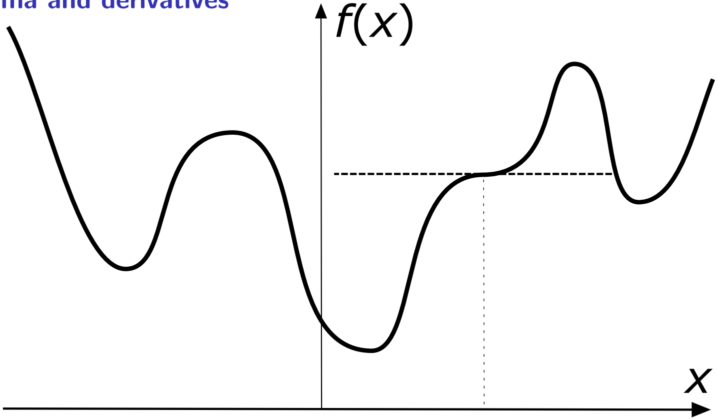
- If $f'(x) < 0$ or $f'(x) > 0$, $x$ clearly cannot be a local minimum
- Hence, $f'(x) = 0$ in all local minima (hence in the global one as well)
- However, $f'(x) = 0$ also in local (hence global) maxima

# Local minima and derivatives



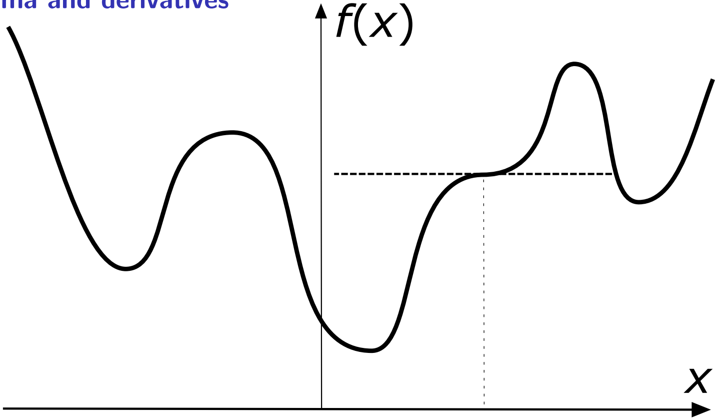
- If $f'(x) < 0$ or $f'(x) > 0$, $x$ clearly cannot be a local minimum
- Hence, $f'(x) = 0$ in all local minima (hence in the global one as well)
- However, $f'(x) = 0$ also in local (hence global) maxima and in saddle points

# Local minima and derivatives



- If $f'(x) < 0$ or $f'(x) > 0$, $x$ clearly cannot be a local minimum
- Hence, $f'(x) = 0$ in all local minima (hence in the global one as well)
- However, $f'(x) = 0$ also in local (hence global) maxima and in saddle points
- To tell them apart one could look at $f'' = [\, f' \,]'$, but this is typically not done

## Local minima and derivatives



- If $f'(x) < 0$ or $f'(x) > 0$, $x$ clearly cannot be a local minimum
- Hence, $f'(x) = 0$ in all local minima (hence in the global one as well)
- However, $f'(x) = 0$ also in local (hence global) maxima and in saddle points
- To tell them apart one could look at $f'' = [\,f'\,]'$, but this is typically not done
- All in all, local optimization $\approx f'(x) = 0$ (stationary point) $+$ crossed fingers

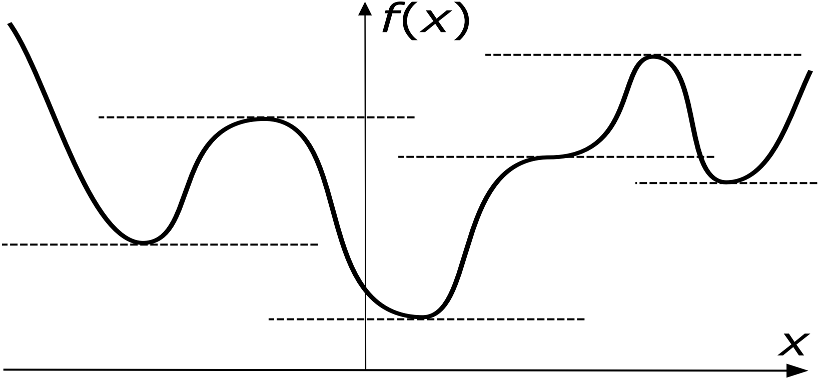
# How about global optimization?

- What does this all tells about global optimization?

# How about global optimization?

- What does this all tells about global optimization?

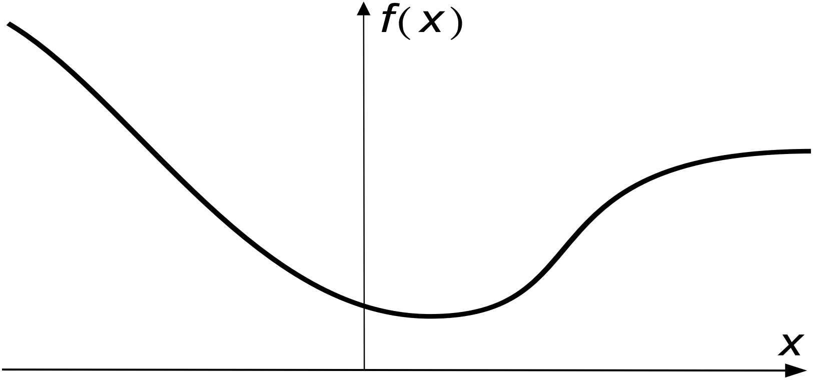
  Sadly, not much at all, unless strong assumptions are made



## How about global optimization?

- What does this all tells about global optimization?

  Sadly, not much at all, unless strong assumptions are made

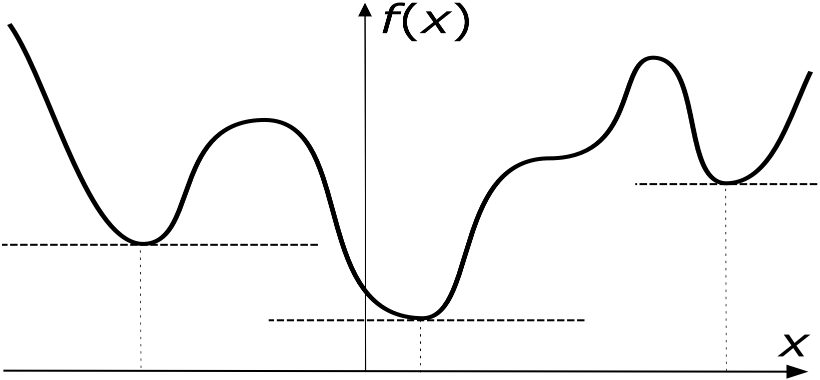


- The obvious one would be unimodal, but not easy to verify/construct

## How about global optimization?

- What does this all tells about global optimization?

  Sadly, not much at all, unless strong assumptions are made

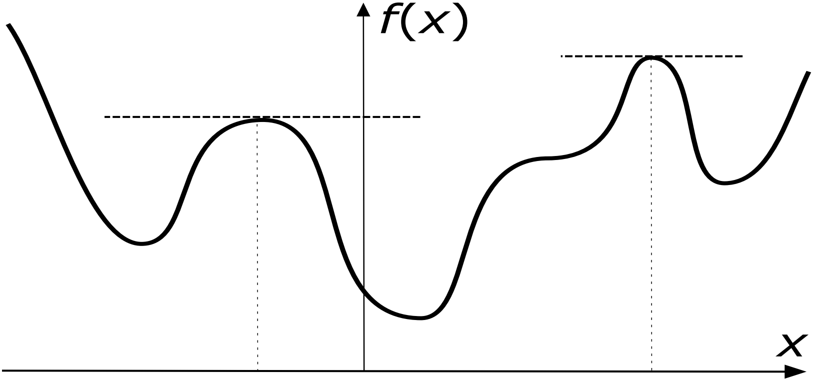


- Intuitively: $f$ has local not global minima

# How about global optimization?

- What does this all tells about global optimization?

  Sadly, not much at all, unless strong assumptions are made

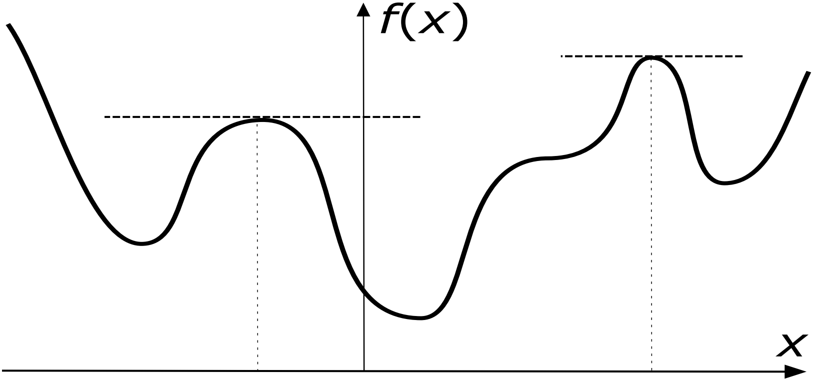


- Intuitively: $f$ has local not global minima $\Longrightarrow$ has local maxima

## How about global optimization?

- What does this all tells about global optimization?

  Sadly, not much at all, unless strong assumptions are made

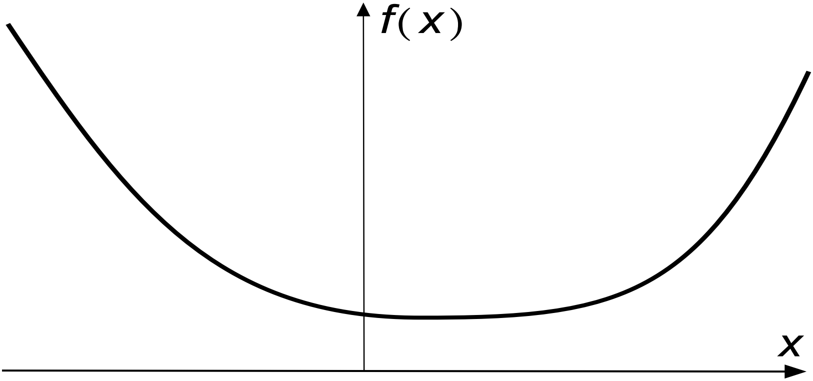


- Intuitively: $f$ has local not global minima $\Longrightarrow$ has local maxima
- Avoid it: stationary point $\Longrightarrow$ local minima $\equiv$ $f'(x) = 0 \Longrightarrow f''(x) \geq 0$

## How about global optimization?

- What does this all tells about global optimization?

  Sadly, not much at all, unless strong assumptions are made



- Intuitively: $f$ has local not global minima $\Longrightarrow$ has local maxima
- Avoid it: stationary point $\Longrightarrow$ local minima $\equiv$ $f'(x) = 0 \Longrightarrow f''(x) \geq 0$
- Sufficient condition: $f''(x) \geq 0 \; \forall x \in \mathbb{R} \Longrightarrow f$ convex

## A very quick recall about convexity

- Convex $\simeq$ $f'$ is monotone nondecreasing $\simeq$ $f'' \geq 0$
- Not really because convex $\not\Rightarrow$ $C^1$ (even less $C^2$)
- Some functions are convex + some operations preserve convexity
  $\Longrightarrow$ the convex world is relatively large
  $\Longrightarrow$ can construct complicated (multivariate) convex functions/sets
- Plenty of theory [3] and software [11]
- Many ML models are purposely constructed convex (SVM) so that (global) optimization is easy
- Some are not (NN), but global optimality not really an issue
- "If you have the choice, choose convex or be content with local minima"
- What if you don't and really need the global optimum?

## An aside: why convex and not unimodal?

- $f$ unimodal $\iff$ quasiconvex [2, Ex. 3.57] $\equiv$
  $\alpha f(x) + (1-\alpha)f(z) \leq \max\{ f(x), f(z) \}$ (??)

- $f$ quasiconvex $\iff$ $\forall$ nonempty sublevel set $S(f, l) = \{ x : f(x) \leq l \}$ is a (possibly, infinite) convex set [2, Th. 3.5.2]

**Exercise:** Prove: $f$ convex $\implies$ $f$ quasiconvex, $\impliedby$ not true

- Issue: algebra of quasiconvex (not convex) functions "weaker"

- $f$ quasiconvex, $\delta \in \mathbb{R}_+$ $\implies$ $\delta f$ quasiconvex true

- But $f$, $g$ quasiconvex $\implies$ $f + g$ quasiconvex false

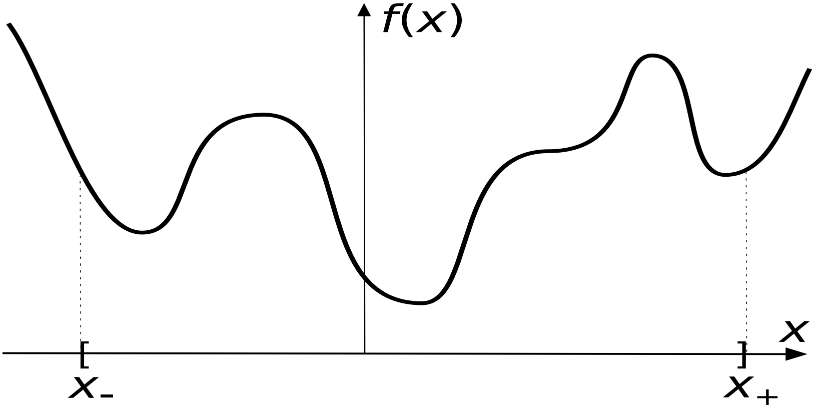
**Exercise:** Prove the two statements above

- No (or much weaker) Disciplined QuasiConvex Programming [11], $f$ "naturally" quasiconvex unlikely

- Does not mean impossible, you may be lucky, in fact NN $\approx$ quasiconvex

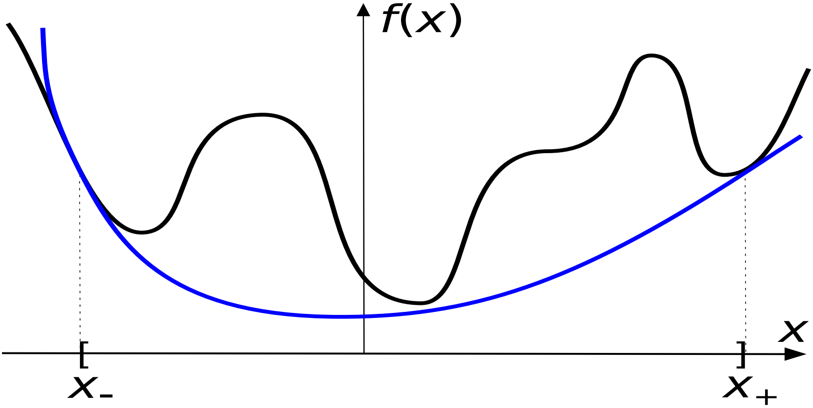
## What if you really want/need a global optima?

- Sift through all $X = [\, x_- \, , \, x_+ \,]$, which must be finite, but using a clever guide



## What if you really want/need a global optima?

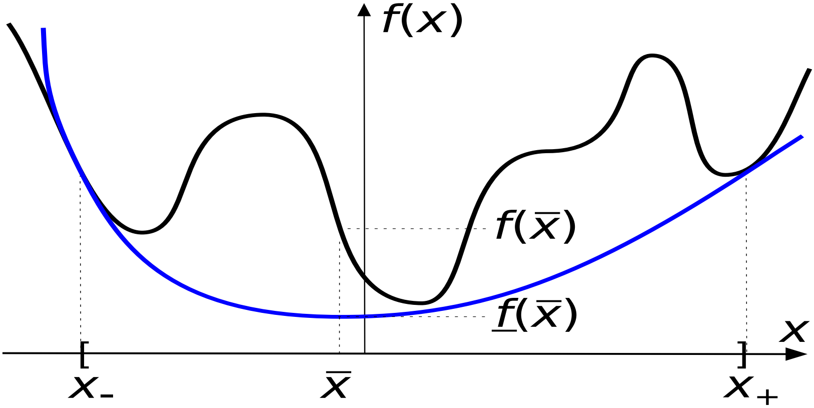
- Sift through all $X = [\, x_- \,,\, x_+ \,]$, which must be finite, but using a clever guide



- Convex lower approximation $\underline{f}$ of nonconvex $f$ on $X$

## What if you really want/need a global optima?

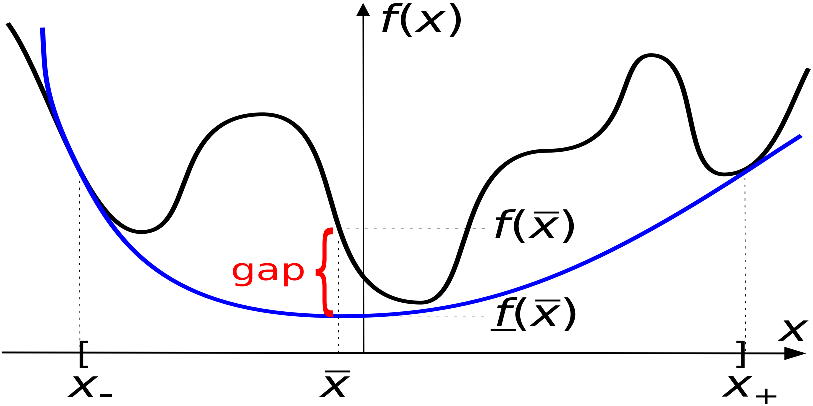
- Sift through all $X = [\, x_- \,,\, x_+ \,]$, which must be finite, but using a clever guide



- Convex lower approximation $\underline{f}$ of nonconvex $f$ on $X$
- "Easily" find local $\equiv$ global minimum $\bar{x}$, giving $\underline{f}(\,\bar{x}\,) \leq f_* \leq f(\,\bar{x}\,)$

# What if you really want/need a global optima?

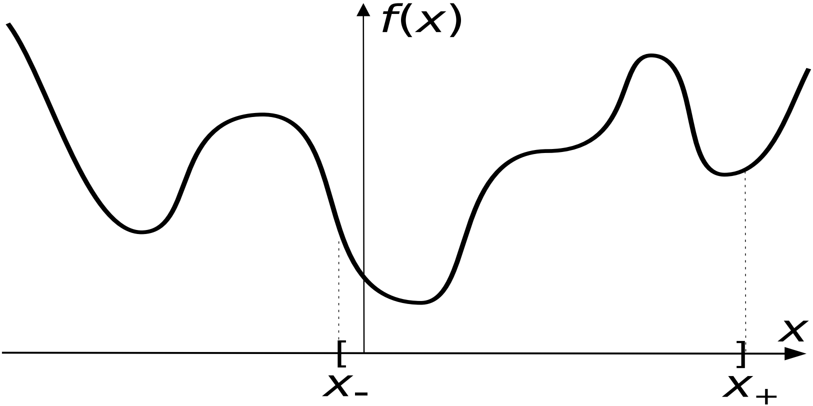
- Sift through all $X = [\, x_- \, , \, x_+ \,]$, which must be finite, but using a clever guide



- Convex lower approximation $\underline{f}$ of nonconvex $f$ on $X$
- “Easily” find local $\equiv$ global minimum $\bar{x}$, giving $\underline{f}(\,\bar{x}\,) \leq f_* \leq f(\,\bar{x}\,)$
- If gap $f(\,\bar{x}\,) - \underline{f}(\,\bar{x}\,)$ too large,

## What if you really want/need a global optima?

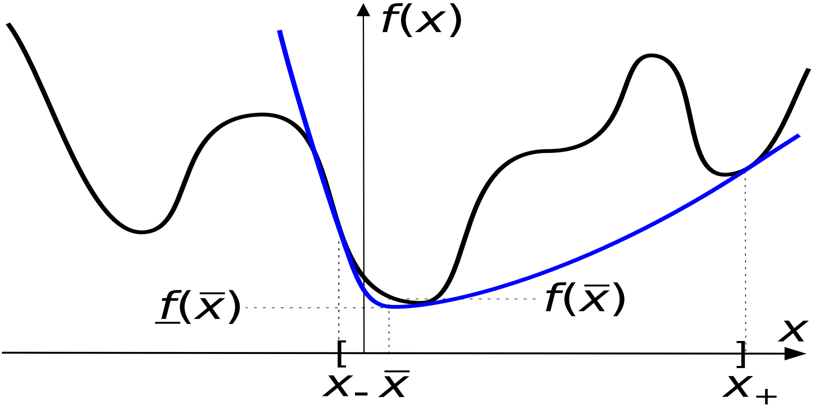
- Sift through all $X = [\, x_- \, , \, x_+ \,]$, which must be finite, but using a clever guide



- Convex lower approximation $\underline{f}$ of nonconvex $f$ on $X$
- “Easily” find local $\equiv$ global minimum $\bar{x}$, giving $\underline{f}(\,\bar{x}\,) \leq f_* \leq f(\,\bar{x}\,)$
- If gap $f(\,\bar{x}\,) - \underline{f}(\,\bar{x}\,)$ too large, partition $X$ and iterate

## What if you really want/need a global optima?

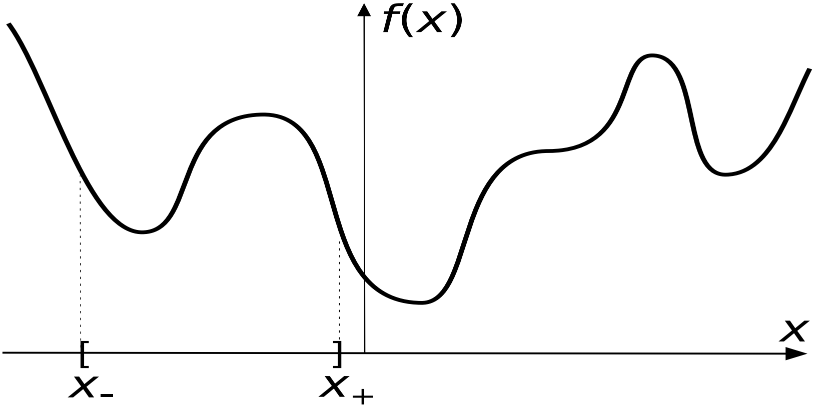
- Sift through all $X = [\, x_- \,,\, x_+ \,]$, which must be finite, but using a clever guide



- Convex lower approximation $\underline{f}$ of nonconvex $f$ on $X$
- "Easily" find local $\equiv$ global minimum $\bar{x}$, giving $\underline{f}(\,\bar{x}\,) \le f_* \le f(\,\bar{x}\,)$
- If gap $f(\,\bar{x}\,) - \underline{f}(\,\bar{x}\,)$ too large, partition $X$ and iterate
- $\underline{f}$ depends on partition, smaller partition (hopefully) $\Longrightarrow$ better gap

## What if you really want/need a global optima?

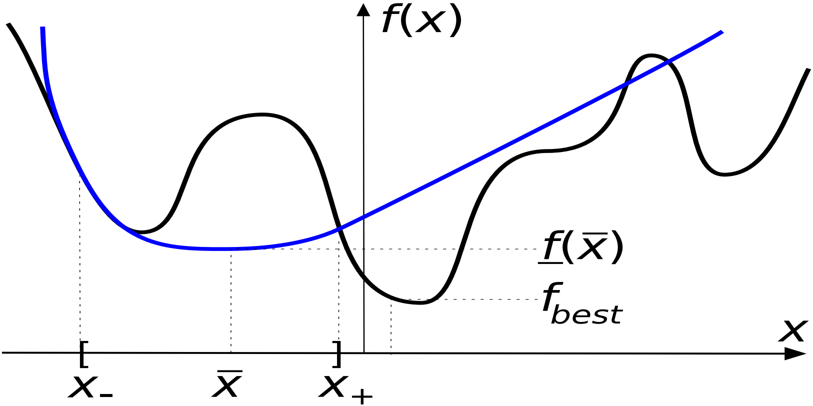
- Sift through all $X = [\, x_- \,,\, x_+ \,]$, which must be finite, but using a clever guide



- Convex lower approximation $\underline{f}$ of nonconvex $f$ on $X$
- “Easily” find local $\equiv$ global minimum $\bar{x}$, giving $\underline{f}(\,\bar{x}\,) \leq f_* \leq f(\,\bar{x}\,)$
- If gap $f(\,\bar{x}\,) - \underline{f}(\,\bar{x}\,)$ too large, partition $X$ and iterate
- If on some partition

## What if you really want/need a global optima?

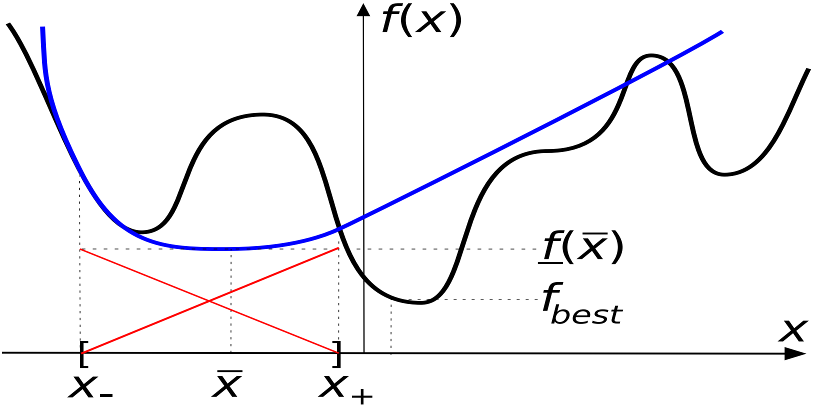
- Sift through all $X = [\, x_- \, , \, x_+ \,]$, which must be finite, but using a clever guide



- Convex lower approximation $\underline{f}$ of nonconvex $f$ on $X$
- "Easily" find local $\equiv$ global minimum $\bar{x}$, giving $\underline{f}(\,\bar{x}\,) \le f_* \le f(\,\bar{x}\,)$
- If gap $f(\,\bar{x}\,) - \underline{f}(\,\bar{x}\,)$ too large, partition $X$ and iterate
- If on some partition $\underline{f}(\,\bar{x}\,) \ge$ best $f$-value so far,

## What if you really want/need a global optima?

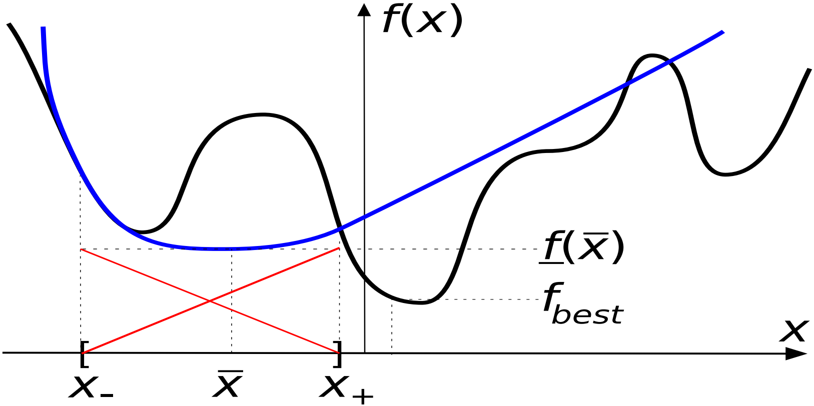
- Sift through all $X = [\, x_- \, , \, x_+ \,]$, which must be finite, but using a clever guide



- Convex lower approximation $\underline{f}$ of nonconvex $f$ on $X$
- “Easily” find local $\equiv$ global minimum $\bar{x}$, giving $\underline{f}(\,\bar{x}\,) \leq f_* \leq f(\,\bar{x}\,)$
- If gap $f(\,\bar{x}\,) - \underline{f}(\,\bar{x}\,)$ too large, partition $X$ and iterate
- If on some partition $\underline{f}(\,\bar{x}\,) \geq$ best $f$-value so far, partition killed for good

# What if you really want/need a global optima?

- Sift through all $X = [\, x_- \, , \, x_+ \,]$, which must be finite, but using a clever guide



- Convex lower approximation $\underline{f}$ of nonconvex $f$ on $X$
- "Easily" find local $\equiv$ global minimum $\bar{x}$, giving $\underline{f}(\,\bar{x}\,) \leq f_* \leq f(\,\bar{x}\,)$
- If gap $f(\,\bar{x}\,) - \underline{f}(\,\bar{x}\,)$ too large, partition $X$ and iterate
- If on some partition $\underline{f}(\,\bar{x}\,) \geq$ best $f$-value so far, partition killed for good
- Still $O(\,1\,/\,\varepsilon\,)$ worst-case (keep dicing and slicing $X$ until pieces "very small")

# And when you have more than one variable . . .

- In general $f : \mathbb{R}^n \to \mathbb{R}$, i.e., $f(x_1, x_2, \ldots, x_n) = f(x)$ with $x = [x_i]_{i=1}^n = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^n$ and $n > 1$
- $n$ can be smallish (2, 3, 100), largish ($10^4$, $10^5$) or heinously large ($10^9$, $10^{11}$)
- $\mathbb{R}^n = \mathbb{R} \times \mathbb{R} \times \ldots \mathbb{R}$, Cartesian product of $\mathbb{R}$ $n$ times $\Longrightarrow$ "exponentially larger than $\mathbb{R}$"

## And when you have more than one variable . . .

- In general $f : \mathbb{R}^n \to \mathbb{R}$, i.e., $f(\, x_1 \,, \, x_2 \,, \ldots, \, x_n \,) = f(\, x \,)$ with $x = [\, x_i \,]_{i=1}^{n} = [\, x_1 \,, \, x_2 \,, \ldots, \, x_n \,] \in \mathbb{R}^n$ and $n > 1$
- $n$ can be smallish (2, 3, 100), largish ($10^4$, $10^5$) or heinously large ($10^9$, $10^{11}$)
- $\mathbb{R}^n = \mathbb{R} \times \mathbb{R} \times \ldots \mathbb{R}$, Cartesian product of $\mathbb{R}$ $n$ times $\Longrightarrow$ "exponentially larger than $\mathbb{R}$"

> "Space is big. Really big. You just won't believe how vastly, hugely, mind-bogglingly big it is." [16]

## And when you have more than one variable ...

- In general $f : \mathbb{R}^n \to \mathbb{R}$, i.e., $f(x_1, x_2, \dots, x_n) = f(x)$ with $x = [x_i]_{i=1}^n = [x_1, x_2, \dots, x_n] \in \mathbb{R}^n$ and $n > 1$
- $n$ can be smallish (2, 3, 100), largish ($10^4$, $10^5$) or heinously large ($10^9$, $10^{11}$)
- $\mathbb{R}^n = \mathbb{R} \times \mathbb{R} \times \dots \mathbb{R}$, Cartesian product of $\mathbb{R}$ $n$ times $\Longrightarrow$ "exponentially larger than $\mathbb{R}$"

> "The vector space $\mathbb{R}^n$ is big. Really big. You just won't believe how vastly, hugely, mind-bogglingly big it is." [16]

## And when you have more than one variable . . .

- In general $f : \mathbb{R}^n \to \mathbb{R}$, i.e., $f(x_1, x_2, \dots, x_n) = f(x)$ with $x = [x_i]_{i=1}^n = [x_1, x_2, \dots, x_n] \in \mathbb{R}^n$ and $n > 1$
- $n$ can be smallish (2, 3, 100), largish ($10^4$, $10^5$) or heinously large ($10^9$, $10^{11}$)
- $\mathbb{R}^n = \mathbb{R} \times \mathbb{R} \times \dots \mathbb{R}$, Cartesian product of $\mathbb{R}$ $n$ times $\Longrightarrow$ "exponentially larger than $\mathbb{R}$"

  > "The vector space $\mathbb{R}^n$ is big. Really big. You just won't believe how vastly, hugely, mind-bogglingly big it is." [16]

- Assume we can even luckily restrict to a "small" $x \in X \subset \mathbb{R}^n$ (say, a "box")
- $I = [x_-, x_+]$, $X = I \times I \times \dots I$ hypercube (if intervals $\neq$, hyperrectangle)
- $B = \{0, 1\}$, $\#B = 2$, $H = B \times B \times \dots B$ (binary hypercube), $\#H = 2^n$
- There is "a lot more space to look at" $\Longrightarrow$ much more difficult
- Relevant for ML (hyperparameter tuning)

## . . . global optimization is exponentially more costly

- Of course need $f$ L-c (exact definition later)
- Very bad news: no algorithm can work in less than $\Omega(\,(\,LD\,/\,\varepsilon\,)^n\,)$ [4, p. 413] (although proof uses adversarial function, not typical in ML)
- Curse of dimensionality: not really doable unless $n = 3/5/10$ tops
- Can make it in $O(\,(\,LD\,/\,\varepsilon\,)^n\,)$, multidimensional grid with small enough step: the standard approach to hyperparameter optimization (but $D$, $L$ unknown)
- If $f$ analytic, clever (spatial) B&B can give ≈ global optimum
- If $f$ “black-box” (typically $\Longrightarrow$ no derivatives), many effective heuristics can give good (not provably optimal) solutions [12]
- In both cases, complexity grows “fast” in practice as $n$ grows
- Finding good global solutions hard in practice . . .

## . . . global optimization is exponentially more costly

- Of course need $f$ L-c (exact definition later)
- Very bad news: no algorithm can work in less than $\Omega(\,(\,LD\,/\,\varepsilon\,)^n\,)$ [4, p. 413] (although proof uses adversarial function, not typical in ML)
- Curse of dimensionality: not really doable unless $n = 3/5/10$ tops
- Can make it in $O(\,(\,LD\,/\,\varepsilon\,)^n\,)$, multidimensional grid with small enough step: the standard approach to hyperparameter optimization (but $D$, $L$ unknown)
- If $f$ analytic, clever (spatial) B&B can give ≈ global optimum
- If $f$ "black-box" (typically $\Longrightarrow$ no derivatives), many effective heuristics can give good (not provably optimal) solutions [12]
- In both cases, complexity grows "fast" in practice as $n$ grows
- Finding good global solutions hard in practice . . .
  unless $f$ convex $\Longrightarrow$ global $\equiv$ local

## So, is global optimization at all possible?

- Not for very-large-scale: exponential in both $1 / \varepsilon$ and $n$

- However, in practice it depends on:
  - "how much nonconvex" $f$ really is
  - how good $\underline{f}$ is as a lower approximation of $f$

- Clever approach: carefully choose your nonconvexities, e.g., integer variables

## So, is global optimization at all possible?

- Not for very-large-scale: exponential in both $1 / \varepsilon$ and $n$
- However, in practice it depends on:
  - "how much nonconvex" $f$ really is
  - how good $\underline{f}$ is as a lower approximation of $f$
- Clever approach: carefully choose your nonconvexities, e.g., integer variables
- Mixed-Integer Linear Programs: all is "trivial" when integer fixed/relaxed

## So, is global optimization at all possible?

- Not for very-large-scale: exponential in both $1 / \varepsilon$ and $n$
- However, in practice it depends on:
  - "how much nonconvex" $f$ really is
  - how good $\underline{f}$ is as a lower approximation of $f$
- Clever approach: carefully choose your nonconvexities, e.g., integer variables
- Mixed-Integer Nonlinear Convex Programs: still "easy" (less so numerically)

## So, is global optimization at all possible?

- Not for very-large-scale: exponential in both $1 / \varepsilon$ and $n$

- However, in practice it depends on:
  - "how much nonconvex" $f$ really is
  - how good $\underline{f}$ is as a lower approximation of $f$

- Clever approach: carefully choose your nonconvexities, e.g., integer variables

- Mixed-Integer Nonlinear Convex Programs: still "easy" (less so numerically) $\not\Rightarrow$ always efficient, $\underline{f}$ often "bad" $\equiv$ bounds weak $\Longrightarrow$ exponential

## So, is global optimization at all possible?

- Not for very-large-scale: exponential in both $1 / \varepsilon$ and $n$
- However, in practice it depends on:
  - "how much nonconvex" $f$ really is
  - how good $\underline{f}$ is as a lower approximation of $f$
- Clever approach: carefully choose your nonconvexities, e.g., integer variables
- Mixed-Integer Nonlinear Convex Programs: still "easy" (less so numerically) $\not\Longrightarrow$ always efficient, $\underline{f}$ often "bad" $\equiv$ bounds weak $\Longrightarrow$ exponential
- (Mixed-Integer) Nonlinear Nonconvex Programs: finding any $\underline{f}$ complex
  - rewrite the expression of $f$ in terms of unary/binary functions
  - apply specific convexification formulæ for each function
- Good news: implemented in available, well-engineered solvers and immensely less inefficient in practice than blind search
- Yet, immensely less efficient in practice than local optimization
- Not our business here, but look [10] to sate your unsavoury curiosity

## Outline

A Bird's Eye on Unconstrained Optimization

Local Optimization Algorithms

Gradient (steepest descent) methods

Gradient method for Quadratic Functions

Gradient method with fixed stepsize

Gradient method and (almost) exact Line Search

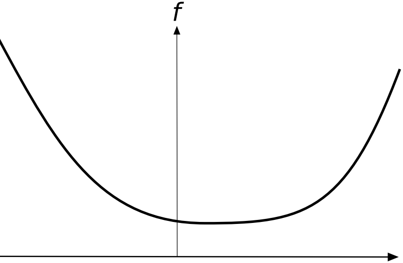
Inexact Line Search: Armijo-Wolfe

Wrap up & References

Solutions

# (Convex, differentiable) local optimization to the rescue

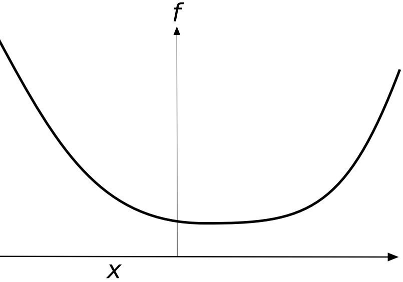
- $f$ differentiable at $x$ local minimum $\implies \nabla f(x) = 0 \equiv$ stationary point
- First-order model of $f$ at $x$: $L_x(z) = \langle \nabla f(x), z - x \rangle + f(x)$



- $f \in C^1$ convex $\iff$

# (Convex, differentiable) local optimization to the rescue

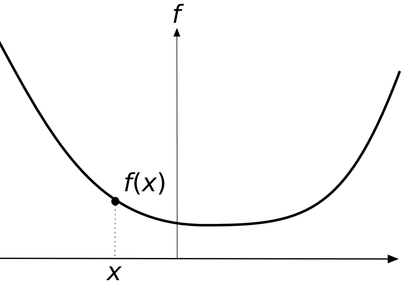
- $f$ differentiable at $x$ local minimum $\implies \nabla f(x) = 0 \equiv$ stationary point
- First-order model of $f$ at $x$: $L_x(z) = \langle \nabla f(x), z - x \rangle + f(x)$



- $f \in C^1$ convex $\iff$

## (Convex, differentiable) local optimization to the rescue

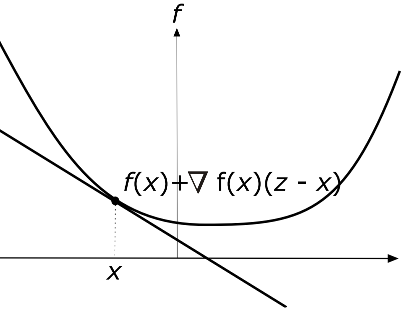
- $f$ differentiable at $x$ local minimum $\implies \nabla f(x) = 0 \equiv$ stationary point
- First-order model of $f$ at $x$: $L_x(z) = \langle \nabla f(x), z - x \rangle + f(x)$



- $f \in C^1$ convex $\iff$ $L_x(z) = f(x)$

# (Convex, differentiable) local optimization to the rescue

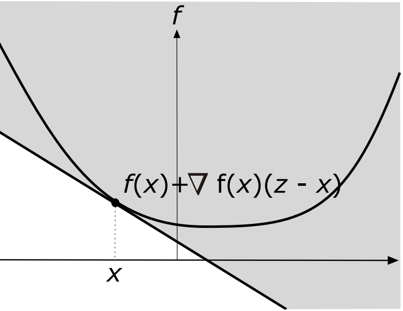
- $f$ differentiable at $x$ local minimum $\Longrightarrow$ $\nabla f(x) = 0 \equiv$ stationary point
- First-order model of $f$ at $x$: $L_x(z) = \langle \nabla f(x), z - x \rangle + f(x)$



- $f \in C^1$ convex $\Longleftrightarrow$
  $L_x(z) = f(x) + \langle \nabla f(x), z - x \rangle \leq f(z)$

# (Convex, differentiable) local optimization to the rescue

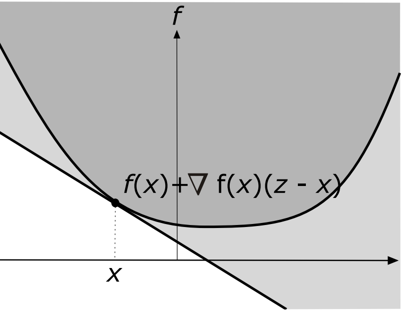
- $f$ differentiable at $x$ local minimum $\Longrightarrow$ $\nabla f(x) = 0 \equiv$ stationary point
- First-order model of $f$ at $x$: $L_x(z) = \langle \nabla f(x), z - x \rangle + f(x)$



- $f \in C^1$ convex $\Longleftrightarrow$
  $L_x(z) = f(x) + \langle \nabla f(x), z - x \rangle \leq f(z)$
- Geometrically: the epigraph is an half-space

## (Convex, differentiable) local optimization to the rescue

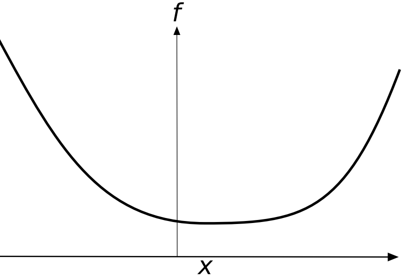
- $f$ differentiable at $x$ local minimum $\Longrightarrow \nabla f(x) = 0 \equiv$ stationary point
- First-order model of $f$ at $x$: $L_x(z) = \langle \nabla f(x), z - x \rangle + f(x)$



- $f \in C^1$ convex $\Longleftrightarrow$ $L_x(z) = f(x) + \langle \nabla f(x), z - x \rangle \leq f(z)$
- Geometrically: the epigraph is an half-space that contains that of $f$ ($epi(L_x) \supseteq epi(f)$)

# (Convex, differentiable) local optimization to the rescue

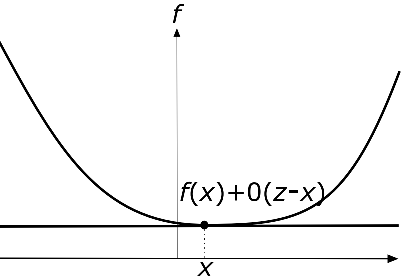
- $f$ differentiable at $x$ local minimum $\implies \nabla f(x) = 0 \equiv$ stationary point
- First-order model of $f$ at $x$: $L_x(z) = \langle \nabla f(x), z - x \rangle + f(x)$



- $f \in C^1$ convex $\iff$ $L_x(z) = f(x) + \langle \nabla f(x), z - x \rangle \leq f(z)$
- Geometrically: the epigraph is an half-space that contains that of $f$ ($epi(L_x) \supseteq epi(f)$)
- $\nabla f(x) = 0$

## (Convex, differentiable) local optimization to the rescue

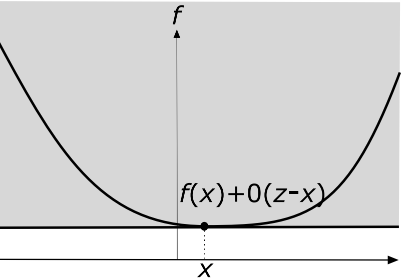
- $f$ differentiable at $x$ local minimum $\Longrightarrow \nabla f(x) = 0 \equiv$ stationary point
- First-order model of $f$ at $x$: $L_x(z) = \langle \nabla f(x), z - x \rangle + f(x)$



- $f \in C^1$ convex $\Longleftrightarrow$ $L_x(z) = f(x) + \langle \nabla f(x), z - x \rangle \leq f(z)$
- Geometrically: the epigraph is an half-space that contains that of $f$ ($epi(L_x) \supseteq epi(f)$)
- $\nabla f(x) = 0 \Longrightarrow$

## (Convex, differentiable) local optimization to the rescue

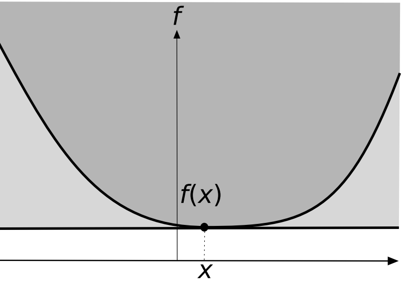
- $f$ differentiable at $x$ local minimum $\Longrightarrow$ $\nabla f(x) = 0 \equiv$ stationary point
- First-order model of $f$ at $x$: $L_x(z) = \langle \nabla f(x), z - x \rangle + f(x)$



- $f \in C^1$ convex $\Longleftrightarrow$ $L_x(z) = f(x) + \langle \nabla f(x), z - x \rangle \leq f(z)$
- Geometrically: the epigraph is an half-space that contains that of $f$ ($epi(L_x) \supseteq epi(f)$)
- $\nabla f(x) = 0 \Longrightarrow$

## (Convex, differentiable) local optimization to the rescue

- $f$ differentiable at $x$ local minimum $\Longrightarrow$ $\nabla f(x) = 0 \equiv$ stationary point
- First-order model of $f$ at $x$: $L_x(z) = \langle \nabla f(x), z - x \rangle + f(x)$



- $f \in C^1$ convex $\Longleftrightarrow$ $L_x(z) = f(x) + \langle \nabla f(x), z - x \rangle \leq f(z)$
- Geometrically: the epigraph is an half-space that contains that of $f$ ($epi(L_x) \supseteq epi(f)$)
- $\nabla f(x) = 0 \Longrightarrow f(z) \geq f(x) \;\; \forall z \in \mathbb{R}^n$ $\equiv$ $x$ global minimum

- $f \in C^2$: $f$ convex $\equiv$ $\nabla^2 f(x) \succeq 0 \quad \forall x \in \mathbb{R}^n$
- Second-order model = first-order model + second-order term (= better)

$$Q_x(z) = L_x(z) + \tfrac{1}{2}(z - x)^T \nabla^2 f(x)(z - x)$$

  a (non-homogeneous) quadratic function $\Longrightarrow$ simple
- $f \in C^2$ with $\nabla^2 f \succeq \tau I$ with $\tau > 0$ the best case for optimization

# Unconstrained local optimization

- Local optimization much better:

  most (but not all) convergence results are dimension-independent

  $\equiv$ iterations count does not explicitly depends on $n$

- If it does, the dependence is not exponential

- Does not mean all local algorithms are fast:
  - complexity may still be $O(\,1\,/\,\varepsilon\,)$ or worse (but also a lot better)
  - cost of $f$ / derivatives computation necessarily increases with $n$: for large $n$ (as in ML), even $O(\,n^2\,)$ may be too much (will see)
  - some dependency on $n$ may be hidden in $O(\cdot)$ constants

- Yet, large-scale local optimization is doable, all that ML needs

- Convex $\implies$ local $\equiv$ global

- How do we actually find a local minimum?

# Towards (unconstrained) local optimization algorithms

- We aim at $\nabla f(x) = 0 \equiv$ stationary point $\Longleftarrow$ (local) minimum
- (Hint of) the proof, because theorems' proofs breed algorithms
- By contradiction: $x$ local minimum but $\nabla f(x) \neq 0$
- Prove $x$ not local minimum not straightforward ($\nexists \equiv \forall\, /$):

  $\forall \varepsilon > 0$ "small enough" $\exists\, z \in \mathcal{B}(x, \varepsilon)$ s.t. $f(z) < f(x)$

  $\equiv$ have to construct $\infty$-ly many $z$ better then $x$ arbitrarily close to it
- Luckily all the $z$ can be taken along a single $d \in \mathbb{R}^n$: $z = x + \alpha d$, $\alpha > 0$
- Can choose $d$, use "best" one: steepest descent direction at $x$

  $\equiv$ $d$ with $\| d \| = 1$ s.t. $\frac{\partial f}{\partial d}(x)$ is most negative

  $\equiv$ the (normalised) anti-gradient $-\nabla f(x)$ $(/ \| \nabla f(x) \|)$

**Exercise:** prove $-\nabla f(x) / \| \nabla f(x) \|$ is the steepest descent direction at $x$

**Exercise:** Why are we insisting that $\| d \| = 1$? Discuss

- $\nabla f(x) = 0 \implies \exists\, \alpha > 0$ s.t. $f(z = x - \alpha \nabla f(x)) < f(x) \rightsquigarrow$ algorithm

## Local Optimization Algorithms

- Iterative procedures: start from initial guess $x^0$, some process $x^i \rightsquigarrow x^{i+1}$ $\Longrightarrow$ a sequence $\{\, x^i \,\}$ that should "go towards an optimal solution"

- The natural way: $\{\, f^i = f(\, x^i \,) \,\} \to f_* \;\equiv\; \{\, x^i \,\}$ minimizing sequence

- Note that $\{\, f^i \,\} \to -\infty \implies f_* = -\infty \implies$ minimizing sequence

- $f$ not convex $\implies$ optimal $\rightsquigarrow$ stationary point ($\nabla f(\, x_* \,) = 0$)

- Other crucial sequence $\{\, e^i = \|\, g^i = \nabla f(\, x^i \,) \,\| \,\} \to 0$ $\not\Longrightarrow$ $\{\, x^i \,\}$ minimising sequence, but our best proxy

- Two general forms of the process: $x^{i+1} \leftarrow x^i + \alpha^i d^i$, but
  - line search: first choose $d^i \in \mathbb{R}^n$ (direction), then choose $\alpha^i \in \mathbb{R}$ (stepsize $\equiv$ "learning rate" in ML-speak)
  - trust region: first choose $\alpha^i$ (trust radius), then choose $d^i$

- Crucial concept: model $f^i \approx f$ used to construct $x^{i+1}$ from $x^i$

## Outline

A Bird's Eye on Unconstrained Optimization

Local Optimization Algorithms

Gradient (steepest descent) methods

Gradient method for Quadratic Functions

Gradient method with fixed stepsize

Gradient method and (almost) exact Line Search

Inexact Line Search: Armijo-Wolfe

Wrap up & References

Solutions

## First-order model $\equiv$ gradient method

- Simplest model: first-order one $L^i(\,x\,) = L_{x^i}(\,x\,) = f(\,x^i\,) + \nabla f(\,x^i\,)(\,x - x^i\,)$
- Idea: $x^{i+1} \in \operatorname{argmin}\{\, L^i(\,x\,) \;:\; x \in \mathbb{R}^n \,\}$

## First-order model $\equiv$ gradient method

- Simplest model: first-order one $L^i(\,x\,) = L_{x^i}(\,x\,) = f(\,x^i\,) + \nabla f(\,x^i\,)(\,x - x^i\,)$

- Idea: $x^{i+1} \in \text{argmin}\{\, L^i(\,x\,) \,:\, x \in \mathbb{R}^n \,\} = \emptyset$: $L^i$ unbounded below on $\mathbb{R}^n$

- $x^{i+1}$ would be very ($\infty$-ly far) from $x^i$: bad idea!
  $L^i$ "good model" only in $\mathcal{B}(\,x^i\,,\,\varepsilon\,)$ (which $\varepsilon$?? unknown, but "small")
  $\Longrightarrow$ $x^{i+1}$ "good" $\equiv$ $f(\,x^{i+1}\,) < f(\,x^i\,)$ only for "small enough" $\alpha^i$

- Then (very) short steps $\alpha^i$, but in which direction? Natural choice:
  $$d_i = \text{argmin}\,\{\, \tfrac{\partial f}{\partial d}(\,x\,)\, [:\ \|\, d\,\| = 1\,]\,\} = -\nabla f(\,x^i\,)\,[\,/\,\|\,\nabla f(\,x^i\,)\,\|\,] =$$
  steepest descent direction $\equiv$ largest $\Delta f^i = f^i - f^{i+1}$ for infinitesimal $\alpha^i$

- Non-normalised $d^i = -g^i = -\nabla f(\,x^i\,)$ only changes $\alpha^i$, in fact better (will see)

```
procedure $x = SDQ\,(\,f\,,\,x\,,\,\varepsilon\,)$
  while( $\|\,\nabla f(\,x\,)\,\| > \varepsilon$ ) do
    $d \leftarrow -\nabla f(\,x\,)$; $\alpha \leftarrow$ stepsize( $f\,,\,x\,,\,d$ ); $x \leftarrow x + \alpha d$;
```

- stepsize($\cdot$) obviously the crucial part: how to choose it?

## How stepsize$(\cdot)$ should look like

- $\alpha^i$ must be "small", $f(\, x^{i+1}\,) \gg f(\, x^i\,)$ possible if "too large" $\alpha^i$: too long steps very bad, the algorithm may even diverge

**Exercise:** Find a too large (fixed) stepsize for $f(\, x\,) = x^2 \,/\, 2$

- Very short steps bad either: $f$ decreases, but only very little
- Example: $f(\, x\,) = -x$ , $d^i = -f'(\, x\,) = 1$ , $x^0 = 0$ , $\alpha^i = 1 \,/\, i^2 \to 0 \implies$ $x^i = -f^i = \sum_{k=1}^{i} 1 \,/\, k^2$ , $\{\, f^i \,\} \to \pi^2 \,/\, 6 \;\approx\; 1.645 \;\gg\; -\infty = f_*$ [19]

**Exercise:** Make the example work even if $f_* > -\infty$

- $\alpha^i \to 0$ possible, just "not too fast": $\alpha^i = 1 \,/\, i \implies \{\, f^i \,\} \to -\infty$ [19]
- All in all, have to avoid two opposite problems:
  - (S) Scylla: $\alpha^i$ not too large to avoid $f(\, x^{i+1}\,) > f(\, x^i\,)$
  - (C) Charybdis: $\alpha^i$ not too small to avoid stalling far from $f_*$
- Dynamic (diminishing) stepsize $\alpha^i$ ($\to 0$) avoids (S) but may hit (C)
- Fixed stepsize $\alpha^i = \bar{\alpha} > 0 \implies \sum_{i=1}^{\infty} \alpha^i = +\infty$ avoids (C) but may hit (S)

## Outline

A Bird's Eye on Unconstrained Optimization

Local Optimization Algorithms

Gradient (steepest descent) methods

Gradient method for Quadratic Functions

Gradient method with fixed stepsize

Gradient method and (almost) exact Line Search

Inexact Line Search: Armijo-Wolfe

Wrap up & References

Solutions

## Gradient method for quadratic functions

- $f(x) = \frac{1}{2}x^T Qx + qx$ , $\nabla f(x) = g = Qx + q$
- Crucial tool: $\varphi_{x,d}(\alpha) = f(x + \alpha d) : \mathbb{R} \to \mathbb{R}$ tomography of $f$ from $x$ along $d$
- Optimal stepsize: stepsize( $f$ , $x$ , $d$ ) $=$ argmin$\{ \varphi_{x,d}(\alpha) \,:\, \alpha \geq 0 \} =$ closed formula $\alpha \leftarrow \| g \|^2 / ( g^T Qg )$ (**check**)
- $O(n^2)$ like computing $g \approx$ cost of 1 iteration, but can be made $O(n)$
- Streamlining a general algorithm: adapting it to a special case

```
procedure $x = SDQ\,(Q, q, x, \varepsilon)$
  $g \leftarrow Qx + q$;
  for( ; ; )
    $\delta \leftarrow \langle g, g \rangle$; if( $\sqrt{\delta} \leq \varepsilon$ ) break;
    $v \leftarrow Qg$; $\alpha \leftarrow \delta / \langle g, v \rangle$; $x \leftarrow x - \alpha d$; $g \leftarrow g - \alpha v$;
```

- Equivalent to general form (**check**) in exact arithmethic
- Only one $O(n^2)$ operation per iteration, roughly half the cost
- $f(x)$ value not needed, but also $O(n)$ (**check**)

## Analysis of the gradient method for quadratic functions

- Does the algorithm always work for every input? (correctness)

**Exercise:** something can go wrong with the $\alpha$-formula: what does it mean? Improve the pseudo-code to take that occurrence into account.

**Exercise:** what happens if $Q \not\succeq 0$? Does the (improved) code need be fixed?

**Exercise:** could you draw similar conclusions for a generic $f(x)$? discuss

**Exercise:** discuss how to change the code to solve $\max\{ f(x) \}$ instead

- Does the algorithm finitely terminate however fixed $\varepsilon > 0$?

- Is the algorithm ran with $\varepsilon = 0$ globally convergent, i.e., does $\{ x^i \}$ converge to a local (global) optimum?

- If $\{ x^i \}$ does converge, "how fast"? (efficiency)

# Gradient method for quadratic functions: efficiency if $Q \succ 0$

- Optimal stepsize $\Longrightarrow$ $g^{i+1} \perp g^i$ (**check**), and this is true for any $f$
- $Q \succ 0$: some rather tedious algebra [5, Lm. 8.6.1] gives
$$A(\,x^{i+1}\,) = \left(1 - \frac{\|\,g^i\,\|^4}{((g^i)^T Q g^i)((g^i)^T Q^{-1} g^i)}\right) A(\,x^i\,) \quad (\textbf{check})$$
- Easy to derive rough estimate using condition number $\kappa = \lambda_1 \,/\, \lambda_n\,(\geq 1)$ of $Q$
$$\frac{\|\,x\,\|^4}{(x^T Q x)(x^T Q^{-1} x)} \geq \frac{\lambda^n}{\lambda^1} = \frac{1}{\kappa} \;\; (\textbf{check}) \;\; \Longrightarrow \;\; A(\,x^{i+1}\,) \leq \left(1 - \frac{1}{\kappa}\right) A(\,x^i\,)$$
- Error decreases as negative exponential: $A(\,x^i\,) \leq r^i A(\,x^0\,)$, $r = 1 - 1\,/\,\kappa < 1$
- Kantorovich inequality [5, 8.6.(34)] gives better estimate of rate of convergence
$$\frac{\|\,x\,\|^4}{(x^T Q x)(x^T Q^{-1} x)} \geq \frac{4\lambda^1\lambda^n}{(\lambda^1 + \lambda^n)^2} \;\; \Longrightarrow \;\; r \leq \left(\frac{\lambda^1 - \lambda^n}{\lambda^1 + \lambda^n}\right)^2 = \left(\frac{\kappa - 1}{\kappa + 1}\right)^2$$
- Quite better: $\kappa = 1000 \Longrightarrow 1 - 1\,/\,k = 0.999$, $[\,(\kappa - 1)\,/\,(\kappa + 1)\,]^2 = 0.996$
- Is it only abstract theory? Let's see in practice

## Gradient method for quadratic functions: efficiency if $Q \succ 0$

- Optimal stepsize $\Longrightarrow$ $g^{i+1} \perp g^{i}$ (**check**), and this is true for any $f$
- $Q \succ 0$: some rather tedious algebra [5, Lm. 8.6.1] gives
$$A(\, x^{i+1}\,) = \left( 1 - \frac{\|\, g^{i} \,\|^4}{((g^{i})^T Q g^{i})((g^{i})^T Q^{-1} g^{i})} \right) A(\, x^{i}\,) \quad (\textbf{check})$$
- Easy to derive rough estimate using condition number $\kappa = \lambda_1 \,/\, \lambda_n \, (\geq 1)$ of $Q$
$$\frac{\|\, x \,\|^4}{(x^T Q x)(x^T Q^{-1} x)} \geq \frac{\lambda^n}{\lambda^1} = \frac{1}{\kappa} \;\; (\textbf{check}) \;\; \Longrightarrow \;\; A(\, x^{i+1}\,) \leq \left( 1 - \frac{1}{\kappa} \right) A(\, x^{i}\,)$$
- Error decreases as negative exponential: $A(\, x^{i}\,) \leq r^{i} A(\, x^{0}\,)$, $r = 1 - 1 \,/\, \kappa < 1$
- Kantorovich inequality [5, 8.6.(34)] gives better estimate of rate of convergence
$$\frac{\|\, x \,\|^4}{(x^T Q x)(x^T Q^{-1} x)} \geq \frac{4\lambda^1 \lambda^n}{(\lambda^1 + \lambda^n)^2} \;\; \Longrightarrow \;\; r \leq \left( \frac{\lambda^1 - \lambda^n}{\lambda^1 + \lambda^n} \right)^2 = \left( \frac{\kappa - 1}{\kappa + 1} \right)^2$$
- Quite better: $\kappa = 1000 \Longrightarrow 1 - 1 \,/\, k = 0.999$, $[\,(\kappa - 1) \,/\, (\kappa + 1)\,]^2 = 0.996$
- Is it only abstract theory? Let's see in practice
- One of many different convergence types, let's name names

## Measuring Algorithms' Speed: Convergence Rate [6, p. 619]

- Crucial sequences: $\{ x^i \}$ / $\{ d^i = | x^i - x_* | \}$ [iterates / distance from $x_*$]
  $\{ f^i = f( x^i ) \}$ / $\{ a^i = A( x^i ) \}$ / $\{ r^i = R( x^i ) \}$ [$f$-values / A/R gaps]
- Converge: $\{ f^i \} \to f_* \approx\equiv \{ a^i \} \to 0 \equiv \{ r^i \} \to 0 \Longleftarrow \{ d^i \} \to 0$ ($\not\Longrightarrow$)

**Exercise:** Discuss why $\{ f^i \} \to f_*$ is only $\approx\equiv$ to $\{ a^i \} \to 0$ and why the $\not\Longrightarrow$

- But how rapidly does it ("in the tail")? Rate/order of convergence

$$\lim_{i\to\infty} \left[ \frac{f^{i+1} - f_*}{( f^i - f_* )^p} = \frac{a^{i+1}}{( a^i )^p} \approx \frac{r^{i+1}}{( r^i )^p} \right] = r$$

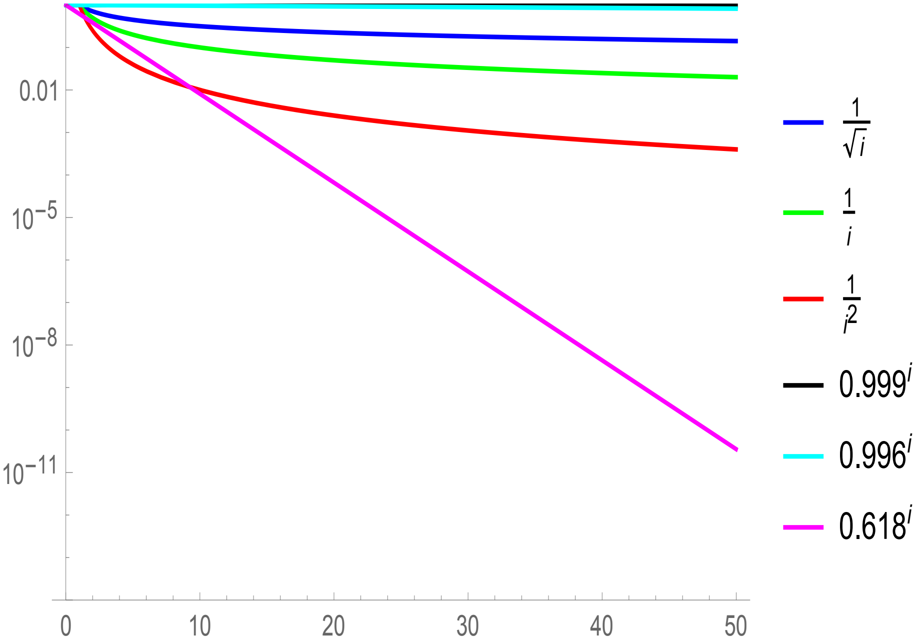
[ $x^p \to 0$ faster than $x \to 0$ when $p > 1$ **(check)**

- $p = 1$ , $r = 1$ $\equiv$ sublinear: important examples

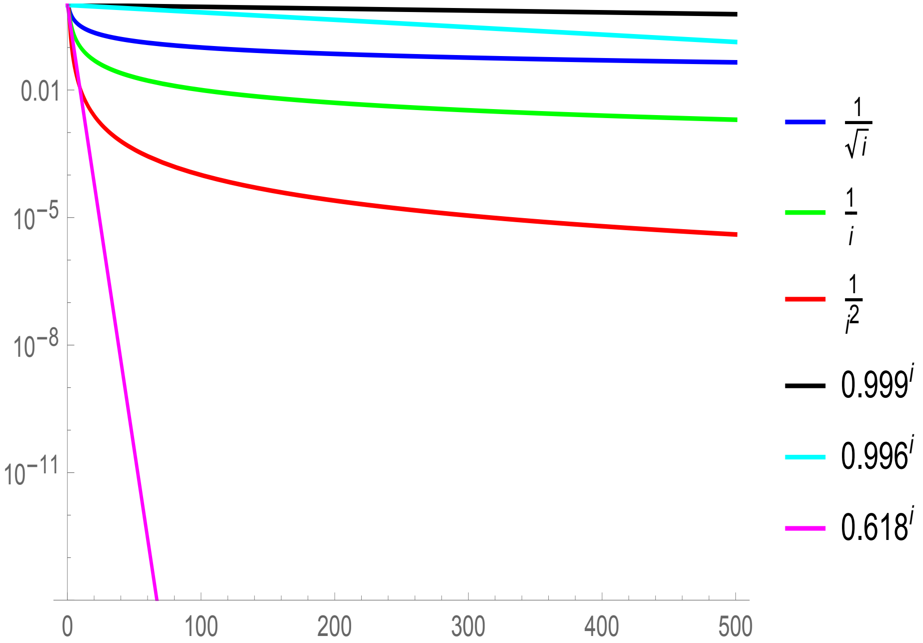
| | | | |
|---|---|---|---|
| error | $O( 1 / i )$ | $O( 1 / i^2 )$ | $O( 1 / \sqrt{i} )$ |
| $i$ | $O( 1 / \varepsilon )$ (bad) | $O( 1 / \sqrt{\varepsilon} )$ (a bit better) | $O( 1 / \varepsilon^2 )$ (horrible) |

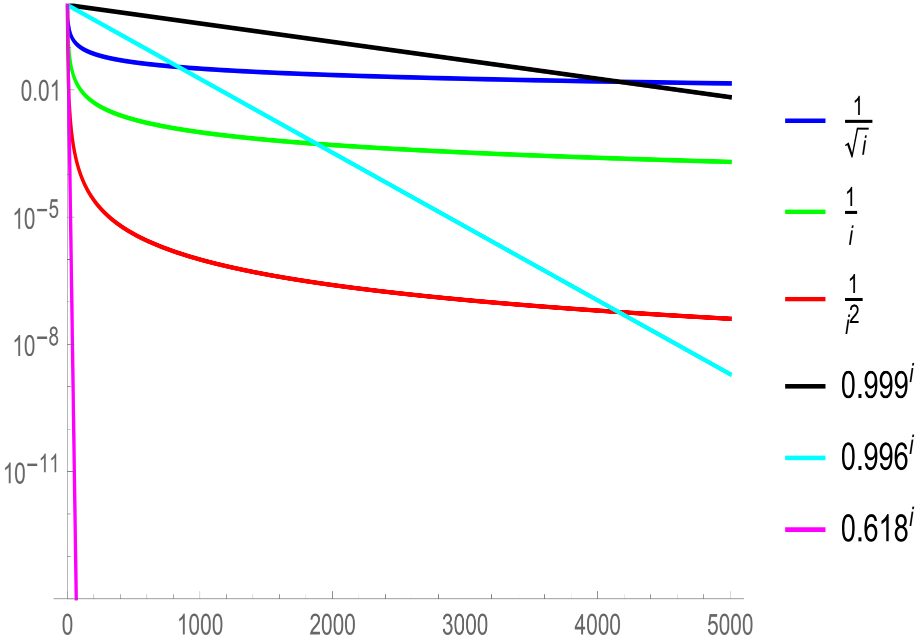
- $p = 1$ , $r < 1$ $\equiv$ linear: $r^i \Longrightarrow i \in O( \log( 1/\varepsilon ) )$ (good unless $r \approx 1$)
- $p = 2$ , $r > 0$ $\equiv$ quadratic (!!!): $\approx 1 / 2^{2^i} \Longrightarrow i \in O( \log( \log( 1/\varepsilon ) ) )$
  in practice $O( 1 )$ (correct digits double at each iteration)
- $p \in ( 1 , 2 )$ $\equiv$ $p = 1$ , $r = 0$ $\equiv$ superlinear (!): "something in the middle"

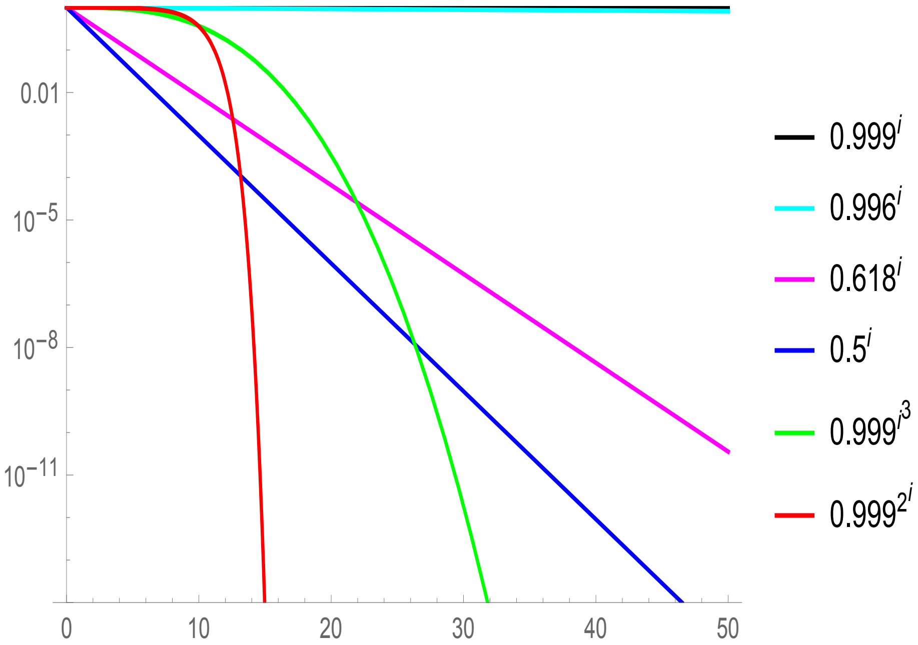
## Convergence Rates Pictorially



## Convergence Rates Pictorially




0.01
$10^{-5}$
$10^{-8}$
$10^{-11}$
0
1000
2000
3000
4000
5000
$\frac{1}{\sqrt{i}}$
$\frac{1}{i}$
$\frac{1}{i^2}$
$0.999^i$
$0.996^i$
$0.618^i$


0.01
$10^{-5}$
$10^{-8}$
$10^{-11}$
0
10
20
30
40
50
$0.999^i$
$0.996^i$
$0.618^i$
$0.5^i$
$0.999^{i^3}$
$0.999^{2^i}$

## When linear convergence may not be enough

- Convergence fast if $\lambda^1 \approx \lambda^n$ (one iteration for $\| x \|^2$), rather slow if $\lambda^1 \gg \lambda^n$: $\kappa = \lambda^1 / \lambda^n \to \infty$ ($Q$ ill conditioned) $\Longrightarrow$ $r \to 1$ $\Longrightarrow$ slow in practice

- $g^{i+1} \perp g^i$ + level sets very elongated $\Longrightarrow$ lots of "zig-zags" $\Longrightarrow$ slow

- Ex.: $\kappa = 1000 \Longrightarrow r \approx 0.996 \Longrightarrow r / ( 1 - r ) \approx 250$
  $f( x^1 ) - f_* = 1$, $\varepsilon = 10^{-6} \Longrightarrow k \geq 3450$ for $n = 2$

## When linear convergence may not be enough

- Convergence fast if $\lambda^1 \approx \lambda^n$ (one iteration for $\| x \|^2$), rather slow if $\lambda^1 \gg \lambda^n$:
  $\kappa = \lambda^1 / \lambda^n \to \infty$ ($Q$ ill conditioned) $\implies$ $r \to 1$ $\implies$ slow in practice

- $g^{i+1} \perp g^i$ + level sets very elongated $\implies$ lots of "zig-zags" $\implies$ slow

- Ex.: $\kappa = 1000 \implies r \approx 0.996 \implies r / ( 1 - r ) \approx 250$
  $f( x^1 ) - f_* = 1$, $\varepsilon = 10^{-6} \implies k \geq 3450$ for $n = 2$ . . . but also for $n = 10^8$

- Note: with coarser formula $r = 0.999 \equiv r / ( 1 - r ) \approx 1000 \implies k \geq 13800$

- In other words: $0.996^{10} \approx 0.96071$ $\qquad$ $0.999^{10} \approx 0.99004$

## When linear convergence may not be enough

- Convergence fast if $\lambda^1 \approx \lambda^n$ (one iteration for $\| x \|^2$), rather slow if $\lambda^1 \gg \lambda^n$: $\kappa = \lambda^1 / \lambda^n \to \infty$ ($Q$ ill conditioned) $\implies$ $r \to 1$ $\implies$ slow in practice

- $g^{i+1} \perp g^i$ + level sets very elongated $\implies$ lots of "zig-zags" $\implies$ slow

- Ex.: $\kappa = 1000 \implies r \approx 0.996 \implies r / ( 1 - r ) \approx 250$ $f( x^1 ) - f_* = 1$, $\varepsilon = 10^{-6} \implies k \geq 3450$ for $n = 2$ . . . but also for $n = 10^8$

- Note: with coarser formula $r = 0.999 \equiv r / ( 1 - r ) \approx 1000 \implies k \geq 13800$

- In other words: $0.996^{100} \approx 0.66978$ $\qquad 0.999^{100} \approx 0.90479$

## When linear convergence may not be enough

- Convergence fast if $\lambda^1 \approx \lambda^n$ (one iteration for $\| x \|^2$), rather slow if $\lambda^1 \gg \lambda^n$:
  $\kappa = \lambda^1 / \lambda^n \to \infty$ ($Q$ ill conditioned) $\Longrightarrow$ $r \to 1$ $\Longrightarrow$ slow in practice

- $g^{i+1} \perp g^i$ + level sets very elongated $\Longrightarrow$ lots of "zig-zags" $\Longrightarrow$ slow

- Ex.: $\kappa = 1000 \Longrightarrow r \approx 0.996 \Longrightarrow r / ( 1 - r ) \approx 250$
  $f( x^1 ) - f_* = 1$, $\varepsilon = 10^{-6} \Longrightarrow k \geq 3450$ for $n = 2$ . . . but also for $n = 10^8$

- Note: with coarser formula $r = 0.999 \equiv r / ( 1 - r ) \approx 1000 \Longrightarrow k \geq 13800$

- In other words: $0.996^{1000} \approx 0.01816$ $\qquad$ $0.999^{1000} \approx 0.36769$

## When linear convergence may not be enough

- Convergence fast if $\lambda^1 \approx \lambda^n$ (one iteration for $\|x\|^2$), rather slow if $\lambda^1 \gg \lambda^n$: $\kappa = \lambda^1 / \lambda^n \to \infty$ ($Q$ ill conditioned) $\Longrightarrow$ $r \to 1$ $\Longrightarrow$ slow in practice

- $g^{i+1} \perp g^i$ + level sets very elongated $\Longrightarrow$ lots of "zig-zags" $\Longrightarrow$ slow

- Ex.: $\kappa = 1000 \Longrightarrow r \approx 0.996 \Longrightarrow r / (1 - r) \approx 250$
  $f(x^1) - f_* = 1$, $\varepsilon = 10^{-6} \Longrightarrow k \geq 3450$ for $n = 2$ ... but also for $n = 10^8$

- Note: with coarser formula $r = 0.999 \equiv r / (1 - r) \approx 1000 \Longrightarrow k \geq 13800$

- In other words: $0.996^{2000} \approx 0.00033$ $\qquad$ $0.999^{2000} \approx 0.13520$

## When linear convergence may not be enough

- Convergence fast if $\lambda^1 \approx \lambda^n$ (one iteration for $\|\, x \,\|^2$), rather slow if $\lambda^1 \gg \lambda^n$:
  $\kappa = \lambda^1 \,/\, \lambda^n \to \infty$ ($Q$ ill conditioned) $\implies$ $r \to 1$ $\implies$ slow in practice

- $g^{i+1} \perp g^i$ + level sets very elongated $\implies$ lots of "zig-zags" $\implies$ slow

- Ex.: $\kappa = 1000 \implies r \approx 0.996 \implies r \,/\, (\, 1 - r \,) \approx 250$
  $f(\, x^1 \,) - f_* = 1$, $\varepsilon = 10^{-6} \implies k \geq 3450$ for $n = 2$ . . . but also for $n = 10^8$

- Note: with coarser formula $r = 0.999 \equiv r \,/\, (\, 1 - r \,) \approx 1000 \implies k \geq 13800$

- In other words: $0.996^{2000} \approx 0.00033$ $\qquad$ $0.999^{2000} \approx 0.13520$

- More bad news, "hidden dependency":
  $\lambda^1$ and $\lambda^n$ may depend on $n$, $\kappa$ may grow as $n \to \infty$

- More bad news: the behaviour in practice is close to the bound

- Even more bad news: $\lambda^n = 0 \;\equiv\; \kappa = \infty$ happens

## What if $\lambda_n = 0$?

- $\lambda_n = 0 \implies$ not converging?

## What if $\lambda_n = 0$?

- $\lambda_n = 0 \implies$ not converging? no, just can't prove it this way
- In fact we can prove it converges (will do soon for general $f$)
- In fact we can prove much more (in a more general setting):

  $\alpha = 1 / \lambda_1 \implies f(x^i) - f_* \leq 2\lambda_1 \| x^0 - x_* \|^2 / (i-1)$ [1, Theorem 3.3]

## What if $\lambda_n = 0$?

- $\lambda_n = 0 \implies$ not converging? no, just can't prove it this way

- In fact we can prove it converges (will do soon for general $f$)

- In fact we can prove much more (in a more general setting):
  $\alpha = 1 / \lambda_1 \implies f(x^i) - f_* \leq 2\lambda_1 \| x^0 - x_* \|^2 / (i - 1)$ [1, Theorem 3.3]

- Is it good news? Only partly. Because complexity is $k \geq 2\lambda_1 d^1 / \varepsilon$

- $O(1/\varepsilon)$ vs. $O(\log(1/\varepsilon))$: exponentially slower (but still dimension-free)

- One further digit of accuracy $\equiv$ 10 times more iterations
  $\implies$ typically unfeasible to get more than 1e−3 / 1e−4 accuracy

- The result cannot be improved (in general, will see)

- Is it bad? Rather. Can it be worse? Yes (e.g., if $f \notin C^1$)

- If $\lambda_n > 0$, can we do better than $O(\log(1/\varepsilon))$? Yes (in general, will see)

## What if $\lambda_n = 0$?

- $\lambda_n = 0 \implies$ not converging? no, just can't prove it this way
- In fact we can prove it converges (will do soon for general $f$)
- In fact we can prove much more (in a more general setting):
  $\alpha = 1 / \lambda_1 \implies f(x^i) - f_* \leq 2\lambda_1 \| x^0 - x_* \|^2 / (i - 1)$ [1, Theorem 3.3]
- Is it good news? Only partly. Because complexity is $k \geq 2\lambda_1 d^1 / \varepsilon$
- $O(1/\varepsilon)$ vs. $O(\log(1/\varepsilon))$: exponentially slower (but still dimension-free)
- One further digit of accuracy $\equiv$ 10 times more iterations
  $\implies$ typically unfeasible to get more than 1e−3 / 1e−4 accuracy
- The result cannot be improved (in general, will see)
- Is it bad? Rather. Can it be worse? Yes (e.g., if $f \notin C^1$)
- If $\lambda_n > 0$, can we do better than $O(\log(1/\varepsilon))$? Yes (in general, will see)
- Most of these results would apply to general $f$, but exact $\alpha$ unavailable

## Outline

A Bird's Eye on Unconstrained Optimization

Local Optimization Algorithms

Gradient (steepest descent) methods

Gradient method for Quadratic Functions

Gradient method with fixed stepsize

Gradient method and (almost) exact Line Search

Inexact Line Search: Armijo-Wolfe

Wrap up & References

Solutions

## Fixed stepsize

- Opposite alternative to optimal $\alpha$: fixed stepsize $\alpha^i = \bar{\alpha}$
- "Like a marriage in a catholic country": only one choice, better be good
- Important note: $\{\, x^i \,\} \rightarrow x_*$ (finite) $\Longrightarrow \{\, \| \, x^{i+1} - x^i \, \| = \alpha^i d^i \,\} \rightarrow 0$ (necessary but not sufficient (**check**))
- Using $d^i = -\nabla f(\, x^i \,) \,/\, \| \, \nabla f(\, x^i \,) \, \| \;\equiv\; \| \, d^i \, \| = 1$ would necessarily $\alpha^i \searrow 0$, which is not possible with fixed stepsize
- Luckily $d^i = -\nabla f(\, x^i \,)$ (not normalised): $\{\, \| \, x^{i+1} - x^i \, \| \,\} \rightarrow 0 \;\Longleftarrow$ $\{\, \nabla f(\, x^i \,) \,\} \rightarrow 0$, which is precisely what we want (stationary point)
- Beware (S): $\exists\, \bar{\alpha} > 0$ s.t. $f(\, x^{i+1} \,) = f(\, x^i - \bar{\alpha} \nabla f(\, x^i \,) \,) < f(\, x^i \,) \quad \forall\, i$?
- Intuitively: "if $f$ varies $\infty$-ly rapidly", then only "$\infty$-ly short $\alpha^i$" are possible
- Crucial to bound how rapidly $f$ (in fact, $\nabla f$) changes

## Recall of Lipschitz continuity, $L$-smoothness

- $f$ (globally) L-c (constant $L$): $|\, f(x) - f(z) \,| \leq L \| \, x - z \, \| \quad \forall\, x \,,\, z$

- L-c $\equiv$ boundedness of $\nabla f$:
  - $f \in C^1$, $\sup\{ \, \| \, \nabla f(x) \, \| \, \} = L < \infty \implies f$ L-c (constant $L$)
    easy to prove out of Mean Value Theorem [9, Th. 5.4.5] (**check**)
  - vice-versa, $f$ (globally) L-c (constant $L$) $\implies \| \, \nabla f(x) \, \| \leq L$
    easy to prove "from prime principles" (**check**)

- $f$ $L$-smooth on $X$ $\equiv$ $\nabla f$ L-c on $X$ (constant $L$):
  $$\| \, \nabla f(x) - \nabla f(z) \, \| \leq L \| \, x - z \, \| \;\; \forall\, x \,,\, z \in X$$

- $\nabla^2 f$ the "gradient" (Jacobian) of $\nabla f$: $\nabla f$ L-c $\equiv$ $\nabla^2 f$ bounded

- $f \in C^2 \implies f$ $L$-smooth $\equiv$
  $$-LI \preceq \nabla^2 f(x) \preceq LI \; \forall\, x \;\equiv\; \max\{ \, |\, \lambda^1 \,| \,,\, |\, \lambda^n \,| \, \} \leq L$$

- $f \in C^2$ convex: $L$-smooth $\equiv [\, 0 \preceq \,] \, \nabla^2 f(x) \preceq LI \; \equiv \; [\, 0 \leq \lambda^n \leq \,] \, \lambda^1 \leq L$

## Fixed stepsize with $L$-smoothness: first steps

- Technical result: $\varphi'_{x,-\nabla f(x)}(\alpha) = \frac{\partial f}{\partial d}(x + \alpha d) = \langle \nabla f(x + \alpha d), d \rangle$

**Exercise:** prove "by prime principles" (definition of $\varphi'$)

**Exercise:** prove using the chain rule in $\mathbb{R}^n$: $f : \mathbb{R}^m \to \mathbb{R}^k$, $g : \mathbb{R}^n \to \mathbb{R}^m$

$$h(x) = f(g(x)) : \mathbb{R}^n \to \mathbb{R}^k \implies Jh(x) = Jf(g(x)) \cdot Jg(x)$$

(note that $Jf \in \mathbb{R}^{k \times m}$, $Jg \in \mathbb{R}^{m \times n}$, in fact $Jh \in \mathbb{R}^{k \times m} \cdot \mathbb{R}^{m \times n} = \mathbb{R}^{k \times n}$)

- Consequence: $\varphi'_{x,-\nabla f(x)}(0) = \langle \nabla f(x), -\nabla f(x) \rangle = -\| \nabla f(x) \|^2 < 0$
  "the farther $\nabla f(x)$ is from 0, the steeper $\varphi'(0)$ is"
- Intuition: $L$-smoothness $\implies$ $\varphi'(\cdot)$ cannot change too rapidly $\implies$
  $\exists$ fixed minimal step $\bar{\alpha}$ s.t. $\varphi'(\bar{\alpha}) = 0 \implies \varphi'(\alpha) < 0 \; \forall \alpha \in [0, \bar{\alpha})$
- Recall $d = -\nabla f(x)$ not normalised (fixed step $\not\Longrightarrow$ fixed movement)
- Turning the intuition into a proof requires some work

## The crucial bound

- $f$ $L$-smooth $\Longrightarrow$ $\varphi$ is $[\, L\|\, d\,\|^2\,]$-smooth

**Exercise:** Prove this “from prime principles”

- $d = -\nabla f(\, x\,) \;\Longrightarrow\; \varphi'(\, 0\,) = -\|\, \nabla f(\, x\,)\,\|^2 = -\|\, d\,\|^2$

- $\varphi$ $[\, L\|\, d\,\|^2\,]$-smooth $\;\Longrightarrow\;$ $\varphi'(\, \alpha\,) \leq \varphi'(\, 0\,) + L\|\, d\,\|^2\alpha = \|\, \nabla f(\, x\,)\,\|^2(L\alpha - 1)$
  $\Longrightarrow\;$ $\varphi'(\, \alpha\,) \leq 0 \;\; \forall\, \alpha \in [\, 0\,,\, \bar{\alpha} = 1\,/\,L\,)$ $\;\Longrightarrow\;$ $1\,/\,L$ (fixed) proposed stepsize

- Issue: evaluate $\varphi(\, 0\,) - \varphi(\, 1\,/\,L\,)$

- Intuition: worst case for $\varphi'$ is linear $\varphi'(\, \alpha\,) \approx L\alpha - 1$
  $\Longrightarrow$ worst case for $\varphi$ is quadratic $\varphi(\, \alpha\,) \approx (\uparrow = \text{derivative})$ $L\alpha^2/2 - \alpha$

**Exercise:** prove using the fundamental theorem of calculus:

- $\varphi =$ function having the worst-case derivative

- Final bound: $\varphi(\, \alpha\,) \leq \varphi(\, 0\,) + \|\, \nabla f(\, x\,)\,\|^2[\, L\alpha^2/2 - \alpha\,]$

## The final complexity bound

- $\bar{\alpha} = 1 / L \implies L\bar{\alpha}^2/2 - \bar{\alpha} = 1 / 2L \implies$
  $\varphi^i(\alpha^i) - \varphi(0) = f(x^{i+1}) - f(x^i) \leq -\| \nabla f(x^i) \|^2 / 2L \implies$
  $a^{i+1} = f(x^{i+1}) - f_* \leq (a^i = f(x^i) - f_*) - \| \nabla f(x^i) \|^2 / 2L$

- Can't do better if you trust the quadratic bound (which you should not)

- Can prove: $a^i \leq 2L\| x^1 - x_* \|^2 / (i - 1) \implies i \geq O(LD^2 / \varepsilon)$ [1, Th. 3.3]

- Good news: dimension independent ($n$ not there) $\implies$ very-large-scale

- Bad news: still $O(1 / \varepsilon)$, high accuracy would be costly

- $O(1 / \varepsilon)$ not tight: $O(1 / \sqrt{\varepsilon})$ possible for $f$ $L$-smooth (will see)

- But $O(1 / \sqrt{\varepsilon})$ tight: $\exists f$ $L$-smooth s.t. $\forall$ algorithm (and large $n$)
  $a^i \geq O(LD^2 / i^2) \implies i \in \Omega(1 / \sqrt{\varepsilon})$ [1, Th. 3.14]

- Algorithms can only go so far with "nasty" problems

- How can we make the problem "nicer"?

## Recall: stronger forms of convexity

- $f$ convex $\equiv \forall\, x \in \mathbb{R}^n$

  $\alpha f(x) + (1-\alpha) f(z) \geq f(\alpha x + (1-\alpha) z) \quad \forall\, \alpha \in [0, 1], \; z \neq x \in \mathbb{R}^n$

  $f \in C^1 \equiv f(z) \geq f(x) + \langle \nabla f(x), z - x \rangle \quad \forall\, z \in \mathbb{R}^n$

  $f \in C^2 \equiv \nabla^2 f(x) \succeq 0$

- $f$ strictly convex $\equiv \alpha f(x) + (1-\alpha) f(z) > f(\alpha x + (1-\alpha) z) \equiv$

  $f(z) > f(x) + \langle \nabla f(x), z - x \rangle \; [f \in C^1] \equiv \nabla^2 f(x) \succ 0 \; [f \in C^2]$

- Quadratic with $\lambda_n > 0$ even more: "grows at least as fast as $\lambda_n \| x \|^2$"

- $f$ strongly convex modulus $\tau > 0$ ($\tau$-convex) $\equiv f(x) - \frac{\tau}{2} \| x \|^2$ convex

  $\equiv \alpha f(x) + (1-\alpha) f(z) \geq f(\alpha x + (1-\alpha) z) + \frac{\tau}{2} \alpha (1-\alpha) \| z - x \|^2$

  $\equiv f(z) \geq f(x) + \langle \nabla f(x), z - x \rangle + \frac{\tau}{2} \| z - x \|^2$

  $\equiv \nabla^2 f(x) \succeq \tau I \; [\succ 0]$ **(check)**

- $f \in C^2$, $L$-smooth and $\tau$-convex $\equiv \tau I \preceq \nabla^2 f \preceq L I \equiv [\, 0 < ] \, \tau \leq \lambda^n \leq \lambda^1 \leq L$

  $\equiv$ eigenvalues of $\nabla^2 f$ bounded above and away from 0

**Exercise:** prove: $f \in C^1$ strictly/strongly convex has unique minimum if any

## Fixed stepsize: convergence rate with strong convexity

- Want to prove: with a proper choice of $\alpha$, the distance to $x_*$ decreases ("fast")
- $z - x_* = x - \alpha \nabla f(x) - x_* = x - \alpha \nabla f(x) - x_* + \alpha \nabla f(x_*)$ $(\nabla f(x_*) = 0)$
  $= (x - x_*) - \alpha(\nabla f(x) - \nabla f(x_*))$
- Mean Value Theorem [9, Th. 5.4.5] on $\nabla f \implies \exists\, w \in [x_*, x]$ s.t.
  $\nabla f(x) - \nabla f(x_*) = \nabla f^2(w)(x - x_*) \implies$
  $z - x_* = (x - x_*) - \alpha \nabla f^2(w)(x - x_*) = (I - \alpha \nabla f^2(w))(x - x_*) \implies$
  $\| z - x_* \| \leq \| I - \alpha \nabla f^2(w) \| \| x - x_* \|$
- With $\alpha = 2 / (L + \tau)$ $(1 / L \leq \alpha < 2 / L)$ converges linearly:
  $\| x^{k+1} - x_* \| \leq r^k \| x^1 - x_* \|$ with $r = (L - \tau) / (L + \tau) < 1$
- $\kappa = L / \tau \geq \lambda_1 / \lambda_n$ $[\geq 1]$ worst-case condition number of $\nabla^2 f$
  $r = (\kappa - 1) / (\kappa + 1) < 1$ (**check**)
- A "small" difference in $f$ makes a big difference in convergence
  $\implies$ properties of $f$ more important than the algorithm
- Let's see how it works in practice

## Recall: eigenvalues & matrix norms [6, A1][13, 15]

- $(\,\lambda\,,\,v\,)$ eigenvalue/eigenvector pair (eep) for $Q$ ($Qv = \lambda v$):
  - $(\,c\lambda\,,\,v\,)$ eep for $cQ$, $c \in \mathbb{R}$ $\quad[\ (cQ)v = c(Qv) = (c\lambda)v\ ]$
  - $(\,1{+}\lambda\,,\,v\,)$ eep for $I{+}Q$ $\quad[\ (I+Q)v = v + Qv = (1+\lambda)v\ ]$
  - $(\,\lambda^2\,,\,v\,)$ eep for $Q^2 = QQ$ $\quad[\ (QQ)v = Q(\lambda v) = \lambda^2 v\ ]$, extends to $Q^k$
- $\|\,Q\,\| = \|\,Q\,\|_2 = \sqrt{\lambda_1(\,Q^TQ\,)} = \max\{\,\|\,Qd\,\|\,/\,\|\,d\,\|\ :\ d \neq 0\,\}$
  Euclidean matrix norm induced by the Euclidean vector norm $\|\cdot\|_2$ ($\exists$ others)
- Consequence: $\|\,Qd\,\| \leq \|\,Q\,\|\|\,d\,\|\ \forall\, d \in \mathbb{R}^n$
- $Q$ symmetric $\Longrightarrow \|\,Q\,\| = \max\{\,|\,\lambda_1(\,Q\,)\,|\,,\,|\,\lambda_n(\,Q\,)\,|\,\}$ (**check**)
- $Q \succeq 0$ (symmetric) $\Longrightarrow \|\,Q\,\| = \lambda_1(\,Q\,) \Longrightarrow \|\,Qv\,\| \leq \lambda_1(\,Q\,)\|\,v\,\|\ \ \forall\, v \in \mathbb{R}^n$;
- $\|\,Q\,\|_2 \leq \|\,Q\,\|_F = \sqrt{\sum_i \sum_j Q_{ij}^2}$ (Frobenius Norm)

## The choice of $\alpha$: the proof

- Want $r = \| I - \alpha \nabla f^2(w) \| =$
  $\max\{ | 1 - \alpha\lambda_1(\nabla f^2(w)) | , | 1 - \alpha\lambda_n(\nabla f^2(w)) | \} < 1$ (**check**)
- The smaller $\gamma$, the faster the convergence: choose $\alpha$ to minimize $\gamma$
- $\alpha = 1 / \lambda_1$ works if $\lambda_n > 0$ $(1 - \lambda_n / \lambda_1 < 1)$, but not optimal
- When $1 - \alpha\lambda_n \geq 1 - \alpha\lambda_1 \geq 0$, increasing $\alpha$ decreases the max
- When $0 \leq \alpha\lambda_n - 1 \leq \alpha\lambda_1 - 1$, decreasing $\alpha$ decreases the max
- The optimal $\alpha$ must be s.t. $1 - \alpha\lambda_n > 0$ and $1 - \alpha\lambda_1 < 0 \Longrightarrow$
  $$r = \max\{ -1 + \alpha\lambda_1 , 1 - \alpha\lambda_n \}$$
- $\lambda_1$ , $\lambda_n$ unknown in general but $L \geq \lambda_1$ , $\tau \leq \lambda_n$ $\Longrightarrow$
  $$r \leq \bar{r} = \max\{ -1 + \alpha L , 1 - \alpha\tau \}$$ (**check**)
- If one term ↑ the other ↓ so they must be equal $\equiv$ $\alpha = 2 / (L + \tau)$ (**check**)
  $\bar{r} = (L - \tau) / (L + \tau) = (\bar{\kappa} - 1) / (\bar{\kappa} + 1) < 1$ , with $\bar{\kappa} = L / \tau \geq 1$

## Outline

A Bird's Eye on Unconstrained Optimization

Local Optimization Algorithms

Gradient (steepest descent) methods

Gradient method for Quadratic Functions

Gradient method with fixed stepsize

Gradient method and (almost) exact Line Search

Inexact Line Search: Armijo-Wolfe

Wrap up & References

Solutions

# Gradient method and "exact" Line Search

- Fixed stepsize direction-efficient: one function evaluation per direction
- Not necessarily direction-effective: the improvement $f^i - f^{i+1}$ may be far smaller than what the direction would actually allow
- $\alpha \approx 1 / L$ best worst-case step, in practice some $\alpha^i \neq \alpha$ may be way better
- Strategy at the "opposite extreme": (Exact) Line Search (LS)
  $\alpha^i \in \mathrm{argmin}\{\, \varphi^i(\,\alpha\,) = \varphi_{x^i,d^i}(\,\alpha\,) = f(\,x^i + \alpha d^i\,) \,:\, \alpha \geq 0\,\}$
- Clearly too hard if global optimum needed $\Longrightarrow$ approximate local minimum in some attraction basin on the right of 0 $\equiv$ find $\alpha > 0$ s.t. $(\varphi^i)'(\,\alpha\,) \approx 0$

**Exercise:** prove: $\alpha^i$ local minimum $\Longrightarrow \nabla f(\,x^{i+1}\,) \perp \nabla f(\,x^i\,)$

- Only "easy" $\equiv$ closed formula if $(\varphi^i)'$ low-degree polynomial, otherwise iterative method for minimising univariate $\varphi$
- More function evaluations per iteration, but (possibly) fewer iterations
- Of course, "something in the middle" may be best (will see)

## Basic Line Search: Dichotomic Search

- $\varphi'$ continuous + intermediate value theorem [9, Th. 2.2.10] $\Longrightarrow$ $\varphi'(\,\alpha_-\,) < 0 \;\wedge\; \varphi'(\,\alpha_+\,) > 0 \;\Longrightarrow\; \exists\,\alpha \in [\,\alpha_-\,,\,\alpha_+\,]$ s.t. $\varphi'(\,\alpha\,) = 0$

- Theorems breed algorithms: dichotomic search

```
procedure $\alpha = DS\,(\,\varphi\,,\,\alpha_-\,,\,\alpha_+\,,\,\varepsilon\,)$
  do forever                 // invariant: $\varphi'(\,\alpha_-\,) < -\varepsilon$, $\varphi'(\,\alpha_+\,) > \varepsilon$
      $\alpha \leftarrow$ in_middle_of( $\alpha_-\,,\,\alpha_+$ ); compute $f'(\,\alpha\,)$;
      if( $|\,\varphi'(\,\alpha\,)\,| \leq \varepsilon$ ) then break;
      if( $\varphi'(\,\alpha\,) < 0$ )   then $\alpha_- \leftarrow \alpha$;   else $\alpha_+ \leftarrow \alpha$;
```

- in_middle_of( $\alpha_-\,,\,\alpha_+$ ) $= (\alpha_+ + \alpha_-)\,/\,2 \;\Longrightarrow$ linear convergence with $r = 0.5$

- If the assumption is not satisfied:

```
  $\Delta\alpha \leftarrow 1$;                              // or whatever value $> 0$
  while( $\varphi'(\,\alpha_+\,) \leq -\varepsilon$ ) do
      $\alpha_+ \leftarrow \alpha_+ + \Delta\alpha$; $\Delta\alpha \leftarrow 2\Delta\alpha$;      // or whatever factor $> 1$
```

- Works in practice for all “reasonable” $\varphi$, e.g. coercive: $\lim_{|\,\alpha\,| \rightarrow \infty} \varphi(\,\alpha\,) = \infty$

- If $\varphi_* = -\infty$, $\alpha_\pm$ may $\rightarrow \pm\infty$ “proving” unboundedness ($\varphi(\,\alpha_\pm\,) \rightarrow -\infty$) but how do you stop? (need a “finite $-\infty$”)

## Improving the dichotomic search: interpolation [7, § 2.4]

- Choosing $\alpha$ "right in the middle" just the dumbest possible approach: much better if $\alpha$ is close to $\alpha_*$ (ideally, $\alpha = \alpha_*$ would stop in one iteration)
- One knows a lot about $\varphi$: $\varphi(\,\alpha_-\,)$, $\varphi(\,\alpha_+\,)$, $\varphi'(\,\alpha_+\,)$, $\varphi'(\,\alpha_-\,)$, let's usethat
- Powerful general idea: construct a model of $\varphi$ based on known information
- Quadratic interpolation: $a\alpha^2 + b\alpha + c$ that "agrees" with $\varphi$ at $\alpha_+$ , $\alpha_-$
- Three parameters, four conditions, something's gotta give (three cases)
- One way: $2a\alpha_+ + b = \varphi'(\,\alpha_+\,)$, $2a\alpha_- + b = \varphi'(\,\alpha_-\,)$ $\Longrightarrow$

$$a = \frac{\varphi'(\,\alpha_+\,) - \varphi'(\,\alpha_-\,)}{2(\alpha_+ - \alpha_-)} \qquad , \qquad b = \frac{\alpha_+\varphi'(\,\alpha_-\,) - \alpha_-\varphi'(\,\alpha_+\,)}{\alpha_+ - \alpha_-}$$

- Minimum solves $2a\alpha + b = 0$ ($c$ irrelevant) $\equiv$

$$\alpha = \frac{\alpha_-\varphi'(\,\alpha_+\,) - \alpha_+\varphi'(\,\alpha_-\,)}{\varphi'(\,\alpha_+\,) - \varphi'(\,\alpha_-\,)}$$

"method of false position"
a.k.a. "secant formula"

always in the middle between $\alpha_+$ and $\alpha_-$ (**check**)

**Exercise:** develop the other cases of quadratic interpolation and discuss them

## Dichotomic search convergence: theory & practice

- Quadratic interpolation has superlinear convergence if started "close enough": [7, Th. 2.4.1] $f \in C^3$, $f'(\alpha_*) = 0$ and $f''(\alpha_*) \neq 0 \implies$ $\exists\, \delta > 0$ s.t. $\alpha^0 \in [\alpha_* - \delta\,,\, \alpha_* + \delta\,] \implies \{\alpha^i\} \to \alpha_*$ with $1 < p \approx 1.618 < 2$
- Four conditions $\implies$ can fit a cubic polynomial and use its minima
- Rather tedious to write down, analyse and implement [7, § 2.4.2][6, p. 57], but pays off: cubic interpolation has quadratic convergence ($p = 2$) "in the tail"
- Very general issue: the model is an estimate $\implies$ wrong $\implies$ bad choices
- Here the model can be "very skewed" $\implies$ very short steps $\implies$ slow $\varphi'(\alpha_+) \gg -\varphi'(\alpha_-) \implies \alpha \approx \alpha_-$ , $\varphi'(\alpha_+) \ll -\varphi'(\alpha_-) \implies \alpha \approx \alpha_+$
- General remedy: never completely trust the model $\equiv$ regularise, stabilise, ...
- In this case: minimum guaranteed decrease $\sigma \leq 0.5$ (safeguard)
  $$\alpha \leftarrow \max\{\, \alpha_- + \sigma(\alpha_+ - \alpha_-)\,,\, \min\{\, \alpha_+ - \sigma(\alpha_+ - \alpha_-)\,,\, \alpha \,\}\,\}$$
- Linear convergence with $r = 1 - \sigma$ "far from $\alpha_*$", superlinear "in the tail"

- Standard stopping criterion is $|\,\varphi'_{x,d}(\,\alpha\,)\,| \leq \varepsilon'$, can it be used?
- Recall: $\varphi'_{x,-\nabla f(\,x\,)}(\,0\,) = \langle\,\nabla f(\,x\,)\,,\,-\nabla f(\,x\,)\,\rangle = -\|\,\nabla f(\,x\,)\,\|^2 < 0$
  "the farther $\nabla f(\,x\,)$ is from 0, the steeper $\varphi'(\,0\,)$ is"

## Gradient method and "Exact" LS

- $f \in C^1 \implies \varphi \in C^1 \implies \varphi'(\alpha)$ computed anyway by the LS $\implies |\varphi'(\alpha)| \leq \varepsilon'$ reasonable stopping condition

- $\varepsilon' = 0$ (exact) in general not possible, how to choose it?

- Anyway, "outer" stopping criterion is approximate: $\| \nabla f(x^i) \| \leq \varepsilon$

- Good news: the algorithm "works" with $\varepsilon' := \varepsilon \| \nabla f(x^i) \|$ without any $L$-smoothness assumption

- only (approximate) stationary point of $\varphi$ needed $\implies$ $f$ convex/unimodal not needed (but you get what you pay for)

- Bad news: the LS should become more accurate as the algorithm proceeds down to $\varepsilon' = \varepsilon^2$ (rather high accuracy)

- Good news: the LS can be very approximate "far from $x_*$"

- Good news: usually works well in practice with arbitrary fixed $\varepsilon'$

## What does convergence exactly means?

- "The gradient method with $\varepsilon' = \varepsilon \| \nabla f( x^i ) \|$ works": meaning what?

- What is simple to prove: $\{ x^i \} \to x \implies \| \nabla f( x ) \| \leq \varepsilon$

  "if it converges, then it does at an (approximate) stationary point"

  **Proof:** $\{ x^i \} \to x$ and $| \varphi'( \alpha^i ) | \leq \varepsilon' \; \forall \, i \implies$

  $$\lim_{i \to \infty} | \langle \nabla f( x^i ) , \nabla f( x^{i+1} ) \rangle | = \langle \nabla f( x ) , \nabla f( x ) \rangle \leq \varepsilon'$$

  $$\implies \| \nabla f( x ) \| \leq \varepsilon \quad \textbf{(check)}$$

**Exercise:** This does not imply $\exists \, h$ s.t. $\| \nabla f( x^h ) \| \leq \varepsilon$ (finite termination) but almost: discuss how to get it

- Proving $\{ x^i \} \to x$ nontrivial in the first place

- Stronger & more general convergence result $\exists$, but they require either conditions on $f$ / $\nabla f$ (other than $C^1$) or exact LS [5, p. 206, p. 234]

- However, general gist: the approach should be expected to work

**Exercise:** what if one would want $\varepsilon = 0$ ("asymptotic convergence")?

## Important notes on the stopping criterion

- The stopping criterion one would want: $A(x^i) \leq \varepsilon$ / $R(x^i) \leq \varepsilon$
- Issue: $f_*$ unknown (most often), cannot be used on-line
- Would need lower bound $\underline{f} \leq f_*$, tight at least towards termination
- Good estimates of $f_*$ "hard" to get, in general no good $\underline{f}$ available
- $\| \nabla f(x^i) \|$ only a "proxy" of $A(x^i)$, choosing $\varepsilon$ non obvious

**Exercise:** assume we know $x_* \in \mathcal{B}(x^i, \delta)$ (which we don't) and $f$ convex, find the stopping tolerance ensuring $a^i \leq \varepsilon$

**Exercise:** prove: if $f$ is $\tau-$convex, then $\| \nabla f(x^i) \| \leq \sqrt{2\tau\varepsilon} \implies a^i \leq \varepsilon$

- Sometimes "relative" stopping condition $\| \nabla f(x^i) \| \leq \varepsilon \| \nabla f(x^0) \|$: scale invariant + clearer what "$\varepsilon =$ 1e-4" means
- Sometimes $\| \nabla f \|$ has a "physical" meaning that can be used
- Sometimes you don't really care if $A(x^i)$ / $R(x^i)$ "small" (ML)

## Efficiency of general gradient

- Good / bad news: efficiency is $\approx$ the same as quadratic $f$

- $f \in C^2$, $x_*$ local minimum s.t. $\nabla^2 f(x_*) \succ 0$, exact LS [6, Th. 3.4]

  $\{x^i\} \to x_* \implies$ for large enough $k$ $\{f^i\}_{i \geq k} \to f_*$ linearly, with

  $r = (\,(\lambda^1 - \lambda^n) / (\lambda^1 + \lambda^n)\,)^2$, $\lambda_1$ and $\lambda_n$ those of $\nabla^2 f(x_*)$

- In "the tail" of the convergence process $f(x) \approx Q_{x_*}(x)$ "very closely"

  $\implies$ convergence $\approx$ the same as for $Q_{x_*}$

- Crucial properties only need to hold in $\mathcal{B}(x_*, \delta)$ provided $\{x^i\} \to x_*$, proving it not obvious although usually happens in practice, anyway exact LS most often (but not always) impossible

- Result can be extended to inexact LS with $r \approx (1 - \lambda^n/\lambda^1)$ (worse), "$\approx$" depending on LS parameters [5, p. 240]

- (More) inexact LS worsens convergence rate but requires less $f$-calls, and this shows up in practice $\equiv$ nontrivial trade-off

# Outline

A Bird's Eye on Unconstrained Optimization

Local Optimization Algorithms

Gradient (steepest descent) methods

Gradient method for Quadratic Functions

Gradient method with fixed stepsize

Gradient method and (almost) exact Line Search

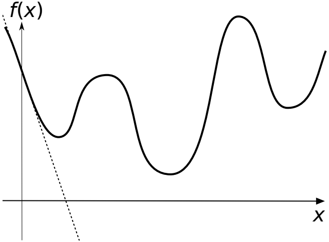
Inexact Line Search: Armijo-Wolfe

Wrap up & References

Solutions

# Inexact Line Search: avoiding Scylla = Armijo

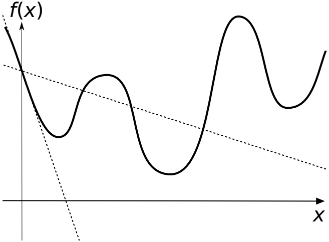
- If FS works, then "any rough LS" also should, provided "$f^i$ decreases enough"



- Armijo condition: $0 < m_1 < (\ll) 1$

# Inexact Line Search: avoiding Scylla = Armijo

- If FS works, then "any rough LS" also should, provided "$f^i$ decreases enough"

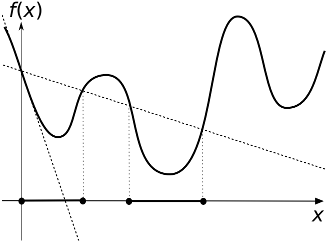


- Armijo condition: $0 < m_1 < (\ll) 1$

  (A) $\quad \varphi(\alpha) \leq \varphi(0) + m_1 \alpha \varphi'(0)$

## Inexact Line Search: avoiding Scylla = Armijo

- If FS works, then "any rough LS" also should, provided "$f^i$ decreases enough"



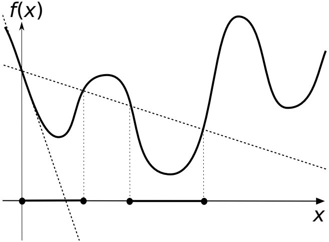
- Armijo condition: $0 < m_1 < (\ll) 1$

  (A) $\quad \varphi(\alpha) \leq \varphi(0) + m_1 \alpha \varphi'(0)$

- Charybdis looms: $\alpha \searrow 0$ satisfies (A)

# Inexact Line Search: avoiding Scylla = Armijo

- If FS works, then “any rough LS” also should, provided “$f^i$ decreases enough”



- Armijo condition: $0 < m_1 < (\ll) 1$

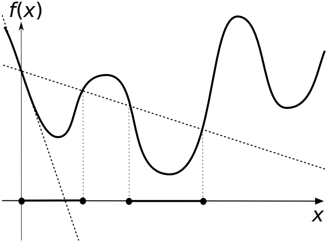
  (A) $\quad \varphi(\,\alpha\,) \leq \varphi(\,0\,) + m_1 \alpha \varphi'(\,0\,)$

- Charybdis looms: $\alpha \searrow 0$ satisfies (A)

- But if we avoid Charybdis then we “converge”

# Inexact Line Search: avoiding Scylla = Armijo

- If FS works, then "any rough LS" also should, provided "$f^i$ decreases enough"



- Armijo condition: $0 < m_1 < (\ll) 1$

  (A) $\quad \varphi(\alpha) \leq \varphi(0) + m_1 \alpha \varphi'(0)$

- Charybdis looms: $\alpha \searrow 0$ satisfies (A)

- But if we avoid Charybdis then we "converge"

- $\alpha^i \geq \bar{\alpha} > 0$ and (A) holds $\forall\, i \implies$ either $\{ f^i \} \to -\infty$ or $\{ \| \nabla f(x^i) \| \} \to 0$

  Proof: $-\varphi_i'(0) = \| \nabla f(x^i) \|^2 \geq \varepsilon > 0$ and (A) hold $\forall i \implies$

  $$f^{i+1} \leq f^i + m_1 \alpha^i \varphi_i'(0) \leq f^i - m_1 \bar{\alpha} \varepsilon \implies$$

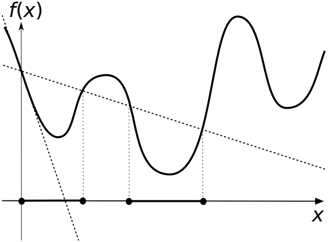
  $$f^i \leq f^0 - m_1 \bar{\alpha} \varepsilon i \implies \{ f^i \} \to -\infty$$

- Don't even need $\alpha^i \geq \bar{\alpha} > 0$, just $\sum_{i=1}^{\infty} \alpha^i = \infty$ ($\alpha^i \to 0$ "slow enough")

- But how do we ensure that $\alpha^i$ does not get "too small"?

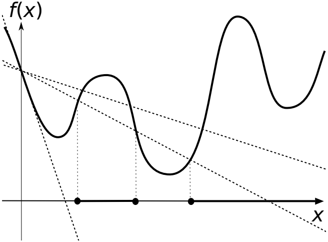
# Inexact Line Search: Avoiding Charybdis = Wolfe

- Need add some further Charybdis-avoiding condition to (A)



# Inexact Line Search: Avoiding Charybdis = Wolfe

- Need add some further Charybdis-avoiding condition to (A)

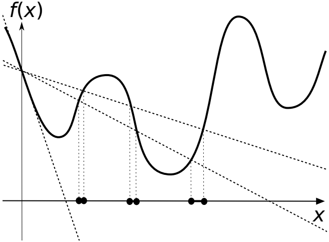


- Goldstein condition: $m_1 < m_2 < 1$

  (G) $\varphi(\,\alpha\,) \geq \varphi(\,0\,) + m_2 \alpha \varphi'(\,0\,)$

# Inexact Line Search: Avoiding Charybdis = Wolfe

- Need add some further Charybdis-avoiding condition to (A)



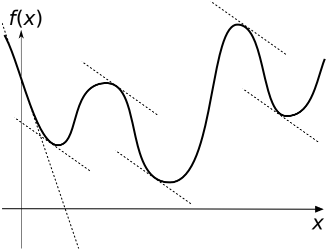
- Goldstein condition: $m_1 < m_2 < 1$

  (G) $\quad \varphi(\,\alpha\,) \geq \varphi(\,0\,) + m_2 \alpha \varphi'(\,0\,)$

- Issue: (A) $\cap$ (G) can exclude all local minima

# Inexact Line Search: Avoiding Charybdis $=$ Wolfe

- Need add some further Charybdis-avoiding condition to (A)



- Goldstein condition: $m_1 < m_2 < 1$

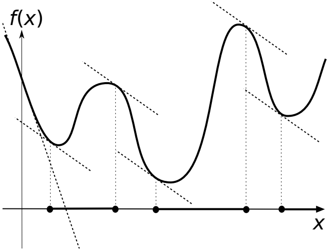
  (G) $\quad \varphi(\,\alpha\,) \geq \varphi(\,0\,) + m_2 \alpha \varphi'(\,0\,)$

- Issue: (A) $\cap$ (G) can exclude all local minima

- Wolfe condition: $m_1 < m_3 < 1$

# Inexact Line Search: Avoiding Charybdis = Wolfe

- Need add some further Charybdis-avoiding condition to (A)



- Goldstein condition: $m_1 < m_2 < 1$

  (G) $\quad \varphi(\,\alpha\,) \geq \varphi(\,0\,) + m_2 \alpha \varphi'(\,0\,)$

- Issue: (A) $\cap$ (G) can exclude all local minima

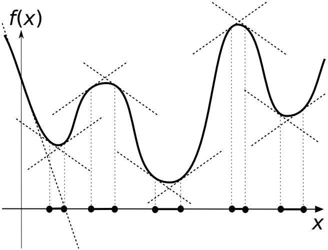
- Wolfe condition: $m_1 < m_3 < 1$

  (W) $\quad \varphi'(\,\alpha\,) \geq m_3 \varphi'(\,0\,)$

- "The derivative has to be a bit closer to 0" (but can be $\gg 0$)

## Inexact Line Search: Avoiding Charybdis = Wolfe

- Need add some further Charybdis-avoiding condition to (A)



- Goldstein condition: $m_1 < m_2 < 1$

  (G) $\quad \varphi(\,\alpha\,) \geq \varphi(\,0\,) + m_2 \alpha \varphi'(\,0\,)$

- Issue: (A) ∩ (G) can exclude all local minima

- Wolfe condition: $m_1 < m_3 < 1$

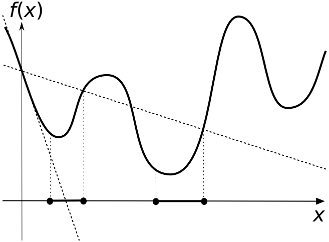
  (W) $\quad \varphi'(\,\alpha\,) \geq m_3 \varphi'(\,0\,)$

- "The derivative has to be a bit closer to 0" (but can be $\gg 0$)

- Strong Wolfe: (W') $\quad |\,\varphi'(\,\alpha\,)\,| \leq m_3 |\,\varphi'(\,0\,)\,| = -m_3 \varphi'(\,0\,) \quad [\implies \text{(W)}\,]$

## Inexact Line Search: Avoiding Charybdis = Wolfe

- Need add some further Charybdis-avoiding condition to (A)



- Goldstein condition: $m_1 < m_2 < 1$

  (G) $\quad \varphi(\,\alpha\,) \geq \varphi(\,0\,) + m_2 \alpha \varphi'(\,0\,)$

- Issue: (A) $\cap$ (G) can exclude all local minima

- Wolfe condition: $m_1 < m_3 < 1$

  (W) $\quad \varphi'(\,\alpha\,) \geq m_3 \varphi'(\,0\,)$

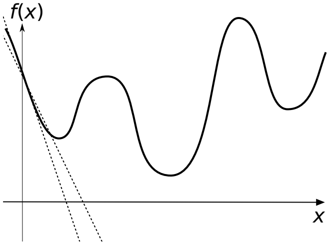
- "The derivative has to be a bit closer to 0" (but can be $\gg 0$)

- Strong Wolfe: (W') $\quad |\,\varphi'(\,\alpha\,)\,| \leq m_3 |\,\varphi'(\,0\,)\,| = -m_3 \varphi'(\,0\,) \quad [\implies \text{(W)}\,]$

- (A) $\cap$ (W) captures all local minima (& maxima)

# Inexact Line Search: Avoiding Charybdis = Wolfe

- Need add some further Charybdis-avoiding condition to (A)



- Goldstein condition: $m_1 < m_2 < 1$

  (G) $\quad \varphi(\, \alpha \,) \geq \varphi(\, 0 \,) + m_2 \alpha \varphi'(\, 0 \,)$

- Issue: (A) $\cap$ (G) can exclude all local minima

- Wolfe condition: $m_1 < m_3 < 1$

  (W) $\quad \varphi'(\, \alpha \,) \geq m_3 \varphi'(\, 0 \,)$

- "The derivative has to be a bit closer to 0" (but can be $\gg 0$)

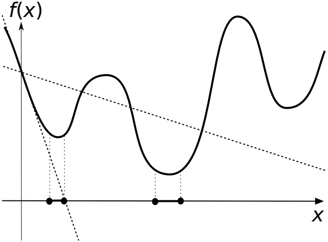
- Strong Wolfe: (W') $\quad |\, \varphi'(\, \alpha \,) \,| \leq m_3 |\, \varphi'(\, 0 \,) \,| = -m_3 \varphi'(\, 0 \,) \quad [\implies \text{(W)}\,]$

- (A) $\cap$ (W) captures all local minima (& maxima)

  unless $m_1$ too close to 1 (that's why usually $m_1 \approx 0.0001$)

# Inexact Line Search: Avoiding Charybdis = Wolfe

- Need add some further Charybdis-avoiding condition to (A)



- Goldstein condition: $m_1 < m_2 < 1$

  (G) $\varphi(\alpha) \geq \varphi(0) + m_2 \alpha \varphi'(0)$

- Issue: (A) $\cap$ (G) can exclude all local minima

- Wolfe condition: $m_1 < m_3 < 1$

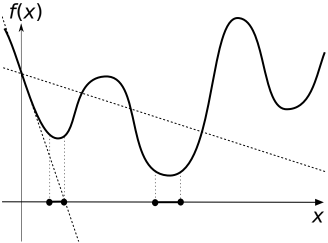
  (W) $\varphi'(\alpha) \geq m_3 \varphi'(0)$

- "The derivative has to be a bit closer to 0" (but can be $\gg 0$)
- Strong Wolfe: (W') $|\varphi'(\alpha)| \leq m_3 |\varphi'(0)| = -m_3 \varphi'(0)$ [ $\Longrightarrow$ (W) ]
- (A) $\cap$ (W) captures all local minima (& maxima)

  unless $m_1$ too close to 1 (that's why usually $m_1 \approx 0.0001$)
- (A) $\cap$ (W') ensures $\varphi'(\alpha) \not\gg 0$, should do away with some local maxima

## Inexact Line Search: Avoiding Charybdis = Wolfe

- Need add some further Charybdis-avoiding condition to (A)



- Goldstein condition: $m_1 < m_2 < 1$

  (G) $\varphi(\alpha) \geq \varphi(0) + m_2 \alpha \varphi'(0)$

- Issue: (A) $\cap$ (G) can exclude all local minima

- Wolfe condition: $m_1 < m_3 < 1$

  (W) $\varphi'(\alpha) \geq m_3 \varphi'(0)$

- "The derivative has to be a bit closer to 0" (but can be $\gg 0$)

- Strong Wolfe: (W') $|\varphi'(\alpha)| \leq m_3 |\varphi'(0)| = -m_3 \varphi'(0)$ [ $\Longrightarrow$ (W) ]

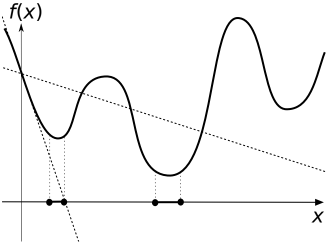
- (A) $\cap$ (W) captures all local minima (& maxima)
  unless $m_1$ too close to 1 (that's why usually $m_1 \approx 0.0001$)

- (A) $\cap$ (W') ensures $\varphi'(\alpha) \not\gg 0$, should do away with some local maxima

- But do such points always $\exists$?

# Inexact Line Search: Avoiding Charybdis = Wolfe

- Need add some further Charybdis-avoiding condition to (A)



- Goldstein condition: $m_1 < m_2 < 1$
  (G) $\varphi(\,\alpha\,) \geq \varphi(\,0\,) + m_2 \alpha \varphi'(\,0\,)$
- Issue: (A) ∩ (G) can exclude all local minima
- Wolfe condition: $m_1 < m_3 < 1$
  (W) $\varphi'(\,\alpha\,) \geq m_3 \varphi'(\,0\,)$
- "The derivative has to be a bit closer to 0" (but can be $\gg 0$)
- Strong Wolfe: (W') $|\,\varphi'(\,\alpha\,)\,| \leq m_3 |\,\varphi'(\,0\,)\,| = -m_3 \varphi'(\,0\,)$ $[\implies$ (W) $]$
- (A) ∩ (W) captures all local minima (& maxima)
  unless $m_1$ too close to 1 (that's why usually $m_1 \approx 0.0001$)
- (A) ∩ (W') ensures $\varphi'(\,\alpha\,) \not\gg 0$, should do away with some local maxima
- But do such points always ∃? Of course they do

# Armijo-Wolfe LS is always possible: the proof

- $\varphi \in C^1 \wedge \varphi(\,\alpha\,)$ bounded below for $\alpha \geq 0 \implies \exists\,\alpha$ s.t. (A) $\cap$ (W) holds without any need for $L$-smoothness

## Armijo-Wolfe LS is always possible: the proof

- $\varphi \in C^1 \wedge \varphi(\alpha)$ bounded below for $\alpha \geq 0 \implies \exists\, \alpha$ s.t. (A) $\cap$ (W) holds without any need for $L$-smoothness
- Rolle's theorem [9, Th. 2.3.8]: $f : \mathbb{R} \to \mathbb{R} \in C^0$ on $[\, a\, ,\, b\,]$, $\in C^1$ on $(\, a\, ,\, b\,)$, s.t. $f(\, a\,) = f(\, b\,) \implies \exists\, c \in (\, a\, ,\, b\,)$ s.t. $f'(\, c\,) = 0$
- Twisted first-order model of $\varphi$ (in 0): $l(\,\alpha\,) = \varphi(\,0\,) + m_1 \alpha \varphi'(\,0\,)$
- $d(\,\alpha\,) = l(\,\alpha\,) - \varphi(\,\alpha\,)$ distance between $l$ and $\varphi$: $d(\,0\,) = 0$, $d'(\,\alpha\,) = m_1 \varphi'(\,0\,) - \varphi'(\,\alpha\,)$, $d'(\,0\,) = (\, m_1 - 1\,)\varphi'(\,0\,) > 0$
- $\nexists\, \bar{\alpha} > 0$ s.t. $d(\,\bar{\alpha}\,) = 0 \implies \varphi$ unbounded below (**check**)
- Smallest $\bar{\alpha} > 0$ s.t. $d(\,\bar{\alpha}\,) = 0$: (A) is satisfied $\forall \alpha \in (\,0\,,\,\bar{\alpha}\,]$ (**check**)
- Rolle's theorem: $\exists\, \alpha' \in (\,0\,,\,\bar{\alpha}\,)$ s.t. $d'(\,\alpha'\,) = 0 \equiv m_1 \varphi'(\,0\,) = \varphi'(\,\alpha'\,)$ $\implies m_3 \varphi'(\,0\,) < m_1 \varphi'(\,0\,) = \varphi'(\,\alpha'\,)$ $[m_3 > m_1]$ $\implies$ (W) holds in $\alpha'$
- $\alpha'$ $\exists$, but how do I actually find it?

## Armijo-Wolfe LS in practice

- $m_1$ small enough s.t. local minima are not cut $\Longrightarrow$ just go for the local minima and stop whenever (A) $\cap$ (W) / (W') holds $\equiv$ dichotomic LS (with interpolation ...) + new stopping criterion

- Hard to say if $m_1$ is small enough, although $m_1 = 0.0001$ most often is

- Specialized LS can be constructed for the odd case it is not [6, Algorithm 3.5], some more logic for the nasty cases

- An even simpler version: "backtracking" LS = only check (A)

```
procedure $\alpha = BLS\,(\,\varphi\,,\,\alpha\,,\,m_1\,,\,\tau\,)$        // $\tau < 1$
  while( $\varphi(\,\alpha\,) > \varphi(\,0\,) + m_1\alpha\varphi'(\,0\,)$ ) do $\alpha \leftarrow \tau\,\alpha$;
```

- Recall: $\exists\bar{\alpha}^i > 0$ s.t. (A) is satisfied $\forall\alpha \in (\,0\,,\,\bar{\alpha}^i\,]$

- Assume $\alpha = 1$ (input): BLS produces $\alpha \geq \tau^{h_i}$ with $h_i \geq \min\{\,k\,:\,\tau^k \leq \bar{\alpha}^i\,\}$

- If $\bar{\alpha}^i \geq \bar{\alpha} > 0\;\forall\,i$, then $\exists\,h$ s.t. $\alpha \geq \tau^h\;\forall\,i \Longrightarrow$ convergence

- Need conditions on $f$ (not surprising ones) to get this

## Convergence with Armijo-Wolfe / Backtracking LS

- Recall: $f$ $L$-smooth $\Longrightarrow$ $\varphi$ is $[\, L\| \, d \, \|^2 \,]$-smooth
- Recall: $\exists$ smallest $0 < \alpha' < \bar{\alpha}$ s.t.
  - (A) holds $\forall \alpha \in (\, 0 \,, \, \bar{\alpha} \,]$, and
  - $\varphi'(\, \alpha' \,) = m_3 \varphi'(\, 0 \,) > \varphi'(\, 0 \,)$
- Recall: $-\varphi'(\, 0 \,) = \| \, d \, \|^2 = \| \, \nabla f(\, x \,) \, \|^2$
- $\varphi$ is $[\, L\| \, d \, \|^2 \,]$-smooth $\Longrightarrow$ $\alpha'$ (and therefore $\bar{\alpha}$) "large":

  $L\| \, d \, \|^2(\, \alpha' - 0 \,) \geq \varphi'(\, \alpha' \,) - \varphi'(\, 0 \,) > (1 - m_3)(\, -\varphi'(\, 0 \,)) = (1 - m_3)\| \, d \, \|^2$

  $\Longrightarrow [\, \bar{\alpha} > \,] \, \alpha' > (1 - m_3) \, / \, L$
- Note: $(1 - m_3) \, / \, L < 1 \, / \, L$ but "of the same order of magnitude"
- Gradient method with AWLS or BLS converges
- $f$ also $\tau$-convex $\Longrightarrow$ convergence linear with $r \approx (\, 1 - \tau \, / \, L \,)$

  "$\approx$" depending on $m_1$, $m_3$ [5, p. 240]
- This may be rather slow, we need something better

## Outline

A Bird's Eye on Unconstrained Optimization

Local Optimization Algorithms

Gradient (steepest descent) methods

Gradient method for Quadratic Functions

Gradient method with fixed stepsize

Gradient method and (almost) exact Line Search

Inexact Line Search: Armijo-Wolfe

Wrap up & References

Solutions

# Wrap up

- Gradient (descent direction) + “reasonable” step = convergence
- “Reasonable” step = Goldilocks’ step = avoid Scylla and Charybdis
- Most natural approach = line search, but exact almost always impossible
- Different practical inexact line searches, up to “no search at all” $\Longrightarrow$ different trade-offs between cost and convergence speed
- Convergence of gradient methods can be from quite bad to horrible, in practice as well as in theory, unless $\nabla f(\, x_* \,)$ “very well conditioned”
- Something better sorely needed: “make $\nabla f(\, x_* \,)$ well conditioned”

## Wrap up

- Gradient (descent direction) + “reasonable” step = convergence

- “Reasonable” step = Goldilocks’ step = avoid Scylla and Charybdis

- Most natural approach = line search, but exact almost always impossible

- Different practical inexact line searches, up to “no search at all”
  $\Longrightarrow$ different trade-offs between cost and convergence speed

- Convergence of gradient methods can be from quite bad to horrible,
  in practice as well as in theory, unless $\nabla f(\, x_* \,)$ “very well conditioned”

- Something better sorely needed: “make $\nabla f(\, x_* \,)$ well conditioned”

- “But the conditioning of $\nabla f(\, x_* \,)$ is a property of the space, master!”

## Wrap up

- Gradient (descent direction) + “reasonable” step = convergence
- “Reasonable” step = Goldilocks’ step = avoid Scylla and Charybdis
- Most natural approach = line search, but exact almost always impossible
- Different practical inexact line searches, up to “no search at all” $\Longrightarrow$ different trade-offs between cost and convergence speed
- Convergence of gradient methods can be from quite bad to horrible, in practice as well as in theory, unless $\nabla f(x_*)$ “very well conditioned”
- Something better sorely needed: “make $\nabla f(x_*)$ well conditioned”
- “But the speed of light is a property of the space, master!”

# Wrap up

- Gradient (descent direction) + "reasonable" step = convergence
- "Reasonable" step = Goldilocks' step = avoid Scylla and Charybdis
- Most natural approach = line search, but exact almost always impossible
- Different practical inexact line searches, up to "no search at all"
  $\Longrightarrow$ different trade-offs between cost and convergence speed
- Convergence of gradient methods can be from quite bad to horrible,
  in practice as well as in theory, unless $\nabla f(x_*)$ "very well conditioned"
- Something better sorely needed: "make $\nabla f(x_*)$ well conditioned"
- "But the speed of light is a property of the space, master!"
  "OK, so let's just change the space!" [14]

## References I

[1] S. Bubeck *Convex Optimization: Algorithms and Complexity*, arXiv:1405.4980v2, https://arxiv.org/abs/1405.4980, 2015

[2] M.S. Bazaraa, H.D. Sherali, C.M. Shetty *Nonlinear Programming: Theory and Algorithms*, John Wiley & Sons, 2006.

[3] S. Boyd, L. Vandenberghe *Convex Optimization*, https://web.stanford.edu/~boyd/cvxbook Cambridge University Press, 2008

[4] P. Hansen, B. Jaumard "Lipschitz Optimization" in *Handbook of Global Optimization – Nonconvex optimization and its applications*, R. Horst and P.M. Pardalos (Eds.), Chapter 8, 407–494, Springer, 1995.

[5] D.G. Luenberger, Y. Ye *Linear and Nonlinear Programming*, Springer International Series in Operations Research & Management Science, 2008

[6] J. Nocedal, S.J. Wright, *Numerical Optimization – second edition*, Springer Series in Operations Research and Financial Engineering, 2006

[7] W. Sun, Y.-X. Yuan, *Optimization Theory and Methods – Nonlinear Programming*, Springer Optimization and Its Applications, 2006

[8] L. Serafino *Optimizing Without Derivatives: What Does the No Free Lunch Theorem Actually Say?* Notices of the AMS 61(7):750–755, 2014 https://www.ams.org/notices/201407/rnoti-p750.pdf

[9] W.F. Trench, *Introduction to Real Analysis* http://ramanujan.math.trinity.edu/wtrench/texts/TRENCH_REAL_ANALYSIS.PDF Free Hyperlinked Edition 2.04, December 2013

[10] CommaLab: http://commalab.di.unipi.it/courses

[11] CVX: http://cvxr.com

[12] DFL: http://www.iasi.cnr.it/~liuzzi/DFL

[13] Wikipedia – Eigenvalues and Eigenvectors https://en.wikipedia.org/wiki/Eigenvalues_and_eigenvectors

[14] Wikipedia – Islands of Space https://en.wikipedia.org/wiki/Islands_of_Space

[15] Wikipedia – Matrix Norm
https://en.wikipedia.org/wiki/Matrix_norm

[16] Wikipedia – The Hitchhiker's Guide to the Galaxy
https://en.wikipedia.org/wiki/The_Hitchhiker's_Guide_to_the_Galaxy

[17] Wikipedia – Integral https://en.wikipedia.org/wiki/Integral

[18] Wikipedia – Rolle's Theorem
https://en.wikipedia.org/wiki/Rolle's_theorem

[19] Wikipedia – Series (mathematics)
https://en.wikipedia.org/wiki/Series_(mathematics)

# Outline

A Bird's Eye on Unconstrained Optimization

Local Optimization Algorithms

Gradient (steepest descent) methods

Gradient method for Quadratic Functions

Gradient method with fixed stepsize

Gradient method and (almost) exact Line Search

Inexact Line Search: Armijo-Wolfe

Wrap up & References

Solutions

## Solutions I

- Use $\max\{\,|\,f_*\,|\,,\;1\,\}$ instead; this corresponds to $\min\{\,f(\,x\,)+1\,\}$ **[back]**

- One could consider the function on the transformed space $f(\,z\,) = f(\,x_- + z(\,x_+ - x_-\,)\,)$; clearly, $z \in [\,0\,,\,1\,] \implies x \in [\,x_-\,,\,x_+\,]$ **[back]**

- Of course, the issue is that the optimal solution would be in one of the two extremes of the interval; consider e.g. $f(\,x\,) = x$ or $f(\,x\,) = -x$. It obviously has no real impact as in the previous cases: $x_- + \varepsilon$ or $x_+ - \varepsilon$ are then $\varepsilon$-optimal solutions for any $\varepsilon > 0$ **[back]**

- Let $x_*$ be any optimal solution in $X$; by definition it belongs to (at least) one interval $[\,x_i\,,\,x_{i+1}\,]$, with $x_{i+1} - x_i \leq 2\varepsilon\,/\,L$. Assume that $x_* - x_i \leq x_{i+1} - x_*$ (the other case is analogous); then $x_* - x_i \leq \varepsilon\,/\,L$. Hence, L-c gives $f(\,x_i\,) - f(\,x_*\,) \leq L|\,x_i - x_*\,| \leq \varepsilon$ **[back]**

## Solutions II

- Take $x$ s.t. $f(x) \leq l$, $z$ s.t. $f(z) \leq l$, and any $\alpha \in [0, 1]$: then, by convexity $f(\alpha x + (1-\alpha)z) \leq \alpha f(x) + (1-\alpha)f(z) \leq \alpha l + (1-\alpha)l = l$, i.e., $\alpha x + (1-\alpha)z \in S(f, l) \implies S(f, l)$ is a (possibly, infinite) convex set
  On the other hand, consider the "downward spike function centred at $c$", i.e., $s_c(x) = \min\{ |x - c|, 1 \}$. Clearly, $s_c$ is quasiconvex: in fact, $S(f, l) = \emptyset$ if $l < 0$, $S(f, l) = [c - l, c + l]$ if $0 \leq l < 1$, and $S(f, l) = \mathbb{R}$ if $l \geq 1$.
  However, $s_0$ is not convex: in fact,
  $(1/2)s_0(0) + (1/2)s_0(2) = 1/2 < 1 = s_0((1/2)0 + (1/2)2) = s_0(1)$ [**back**]

- $S(\delta f, l) = \{ x : \delta f(x) \leq l \} = \{ x : \delta f(x) \leq l / \delta \} = S(f, l / \delta)$: since the latter is an interval (convex set), the former also is
  To prove $\not\Longleftarrow$ consider $f(x) = s_{-1}(x) + s_1(x)$ (cf. previous exercise). Clearly, $f(-1) = f(1) = 0$ but $f(x) > 0$ for all other values of $x$, i.e., $S(f, 0) = \{ -1, 1 \}$ is not an interval [**back**]

## Solutions III

- We know that $\frac{\partial f}{\partial d}(x) = \langle \nabla f(x) , d \rangle = \| \nabla f(x) \| \| d \| \cos(\theta) =$ $= \| \nabla f(x) \| \cos(\theta)$ (as $\| d \| = 1$). Clearly, this number is minimum when $\cos(\theta)$ is, i.e., $\theta = \pi \equiv \cos(\theta) = -1$. This corresponds to $d$ being collinear to $\nabla f(x)$ with opposite direction, i.e., $d = -\nabla f(x) / \| \nabla f(x) \|$ **[back]**

- Because $\frac{\partial f}{\partial \beta d} = \beta \frac{\partial f}{\partial d}$, hence $\| d \| \to \infty \implies \frac{\partial f}{\partial d} \to -\infty$ (with right $d$) **[back]**

- $f(x) = x^2 / 2 \implies f'(x) = x \implies x^{i+1} \leftarrow x^i - \alpha x^i = x^i(1 - \alpha) \implies$ $f(x^{i+1}) = f(x^i(1 - \alpha)) = (1 - \alpha)^2 f(x^i)$, hence $f(x^k) r^k f(x^0)$ for $r = (1 - \alpha)^2$. If $\alpha > 2$ then $r = (1 - \alpha)^2 > 1$ and $\{ f^i = f(x^i) \} \to +\infty$ (exponentially fast), unless $x^i = 0$ **[back]**

- Take $\alpha^i = 1 / 2i^2 \leq 1 / 2$, $f(x) = x^2 / 2 - x$, $f'(x) = x - 1 \implies$ $x^{i+1} \leftarrow x^i - \alpha^i(x^i - 1) = x^i(1 - \alpha^i) + \alpha^i \leq x^i + \alpha^i$. Take $x^1 = 0$ to get $x^{i+1} \leq \sum_{k=1}^{i} \alpha^k$; thus, $x^i \leq \sum_{k=1}^{\infty} \alpha^k = \pi^2 / 12 \approx 0.8225 < 1 = x_*$ for all $i$, i.e., the algorithm stalls (long) before reaching the optimal solution **[back]**

- $\varphi(\alpha) = f(x - \alpha g) - f(x) = \frac{1}{2}(x - \alpha g)^T Q(x - \alpha g) + q(x - \alpha g) - f(x) =$
  $= \frac{1}{2}\alpha^2 g^T Qg - \alpha[g^T Qx + qg] = \frac{1}{2}\alpha^2 g^T Qg - \alpha\| g \|$ $(g = Qx + q)$
  $\implies \varphi'(\alpha) = 0 \equiv \alpha g^T Qg = \| g \|$ **[back]**

- With $g = Qx + q$, we have $x^+ = x - \alpha g$ and $g^+ = Qx^+ + q =$
  $= Q(x - \alpha g) + q = Qx + q - \alpha Qg = (I - \alpha Q)g$. Hence we can use $v = Qg$ for computing both $\alpha$, as $g^T Qg = \langle g , v \rangle$, and $g^+ = g - \alpha v$ **[back]**

- $f(x) = \frac{1}{2}x^T Qx + \langle q , x \rangle = \frac{1}{2}(x^T Qx + 2\langle q , x \rangle) =$
  $= \frac{1}{2}(\langle x , Qx + q \rangle + \langle q , x \rangle) = \frac{1}{2}(\langle x , g \rangle + \langle q , x \rangle) = \frac{1}{2}\langle x , g + q \rangle$ **[back]**

- The issue clearly is $g^T Qg = 0$ (very small), which means that $\varphi_{x,-g}$ is (almost) linear, and therefore $f$ is unbounded below. One should therefore add a line
  **if**( $g^T Qg \leq \sigma$ ) **then break**;
  for a "very small" $\sigma$, but also add a proper way for the algorithm to signal that the returned $x$ is not optimal, e.g., by also returning a "status code" **[back]**

- Having added the extra check above, the code just works: if $g^T Qg < 0$ then $(-)g$ is direction where $\varphi$ has negative curvature, which still implies $f$ is unbounded below. Note that this is not guaranteed to happen [**back**]

- In a word: no. If $f(x)$ is a "black box", i.e., one can only evaluate it (and the gradient) but has no clue about how this is done, it's impossible to declare that $f_* = -\infty$. Indeed, even for $n = 1$, $f$ may decrease "for a very long time" but then abruptly start increasing again, and there is no way of knowing whether or not this will eventually happen. Thus, proving unboundedness is harder than proving (local) optimality, for which $\nabla f(x) = 0$ (approximately) suffices, unless one has more control on the function's properties such as knowing its exact algebraic form (and even this may not be enough) [**back**]

- Because $g^T Qg < 0$, the step $\alpha$ will be negative, which means one is going in direction $g$ rather than $-g$. The algorithm remains the same, except that the extra check above has to become $g^T Qg \geq -\delta$ [**back**]

- For quadratic $f$ the proof is just easy algebra: $g^i = Q(x^i - x_*) = Qx^i + q$, $\alpha^i = \| g^i \|^2 / [(g^i)^T Q g^i]$, $g^{i+1} = Qx^{i+1} + q = Q(x^i - \alpha^i g^i) + q =$ $= (I - \alpha^i Q)g^i \implies \langle g^{i+1} , g^i \rangle = \| g^i \|^2 - \alpha^i [ (g^i)^T Q g^i ] = 0$
  We will see later the proof that the result holds in general **[back]**

- $Q$ nonsingular $\implies x^i - x_* = Q^{-1} g^i \implies$
  $a^i = \frac{1}{2}(x^i - x_*)^T Q(x^i - x_*) = \frac{1}{2}(g^i)^T Q^{-1} g^i \implies$
  $a^{i+1} = \frac{1}{2}(x^{i+1} - x_*)^T Q(x^{i+1} - x_*) = \frac{1}{2}(x^i - \alpha^i g^i - x_*)^T g^{i+1} = \frac{1}{2}(x^i - x_*)^T g^{i+1}$
  [ using $\langle g^{i+1} , g^i \rangle = 0$ ] $= \frac{1}{2}(x^i - x_*)^T Q(x^i - \alpha^i g^i - x_*)$
  $= \frac{1}{2}(x^i - x_*)^T Q(x^i - x_*) - \frac{1}{2}\alpha^i (x^i - x_*)^T Q g^i = a^i - \frac{1}{2}\alpha^i \| g^i \|^2$
  [ using $Q(x^i - x_*) = g^i$ ] $= a^i - \frac{1}{2}\| g^i \|^4 / (g^i)^T Q g^i$
  $= a^i - \frac{\| g^i \|^4 \qquad \frac{1}{2}(g^i)^T Q^{-1} g^i}{((g^i)^T Q g^i)((g^i)^T Q^{-1} g^i)} = a^i \left( 1 - \frac{\| g^i \|^4}{((g^i)^T Q g^i)((g^i)^T Q^{-1} g^i)} \right)$ **[back]**

- Recall $1 \,/\, \lambda^n \geq \ldots \geq 1 \,/\, \lambda^1 \; > \; 0$ eigenvalues of $Q^{-1}$; from the usual $\lambda^n \| \, x \, \|^2 \leq x^T Q x \leq \lambda^1 \| \, x \, \|^2$ (applied to $Q^{-1}$ as well) one has $\| \, g \, \|^2 \,/\, g^T Q g \geq 1 \,/\, \lambda_1$ and $\| \, g \, \|^2 \,/\, g^T Q^{-1} g \geq 1 \,/\, [1 \,/\, \lambda_n]$ **[back]**

- If $f_* = -\infty$, $f_i \to -\infty$ is OK (minimising sequence) but $a^i = a^{i+1} = \infty$ and therefore their ratio is not well-defined. Since $f$ is continuous, $\{ \, d^i \, \} \to 0 \implies \{ \, a^i \, \} \to 0$, but the converse need not happen in general: say, $\{ \, x^{2i} \, \} \to x'_*$ and $\{ \, x^{2i+1} \, \} \to x''_*$ with $x'_* \neq x''_*$ optimal solutions **[back]**

- Simply, $\lim_{x \to 0} x^p \,/\, x = \lim_{x \to 0} x^{p-1} = 0$: the numerator goes to 0 faster than the denominator **[back]**

- $\alpha_- \varphi'( \, \alpha_+ \, ) - \alpha_+ \varphi'( \, \alpha_- \, ) = \alpha_- \varphi'( \, \alpha_+ \, ) - \alpha_+ \varphi'( \, \alpha_- \, ) + \alpha_- \varphi'( \, \alpha_- \, ) - \alpha_- \varphi'( \, \alpha_- \, ) = \alpha_- ( \, \varphi'( \, \alpha_+ \, ) - \varphi'( \, \alpha_- \, ) \, ) - \varphi'( \, \alpha_- \, )( \, \alpha_+ - \alpha_- \, )$. Divide by $\varphi'( \, \alpha_+ \, ) - \varphi'( \, \alpha_- \, )$ to get $\alpha = \alpha_+ + \beta( \, \alpha_+ - \alpha_- \, )$ with $0 \leq \beta = -\varphi'( \, \alpha_- \, ) \,/\, ( \, \varphi'( \, \alpha_+ \, ) - \varphi'( \, \alpha_- \, ) \, ) \leq 1$; it is then plain to see that $\alpha_- \leq \alpha \leq \alpha_+$ **[back]**

## Solutions VIII

- A full development would not be didactical. The four conditions are $a\alpha_+^2 + b\alpha_+ + c = \varphi(\,\alpha_+\,)$, $a\alpha_-^2 + b\alpha_- + c = \varphi(\,\alpha_-\,)$, $2a\alpha_+ + b = \varphi'(\,\alpha_+\,)$, $2a\alpha_- + b = \varphi'(\,\alpha_-\,)$; each three of them give a linear system with three equations in the three unknowns $a$, $b$, $c$ that gives (not necessarily) different solutions (mind the special cases) and therefore quadratic models **[back]**

- Seen already: $f(\,x\,) = -x$, $f'(\,x\,) = -1$, $\alpha^i = 1\,/\,i$: $x^{i+1} - x^i = \alpha^i = 1\,/\,i \to 0$ as $i \to \infty$, but $x^i \to \infty$ **[back]**

- By the mean value theorem, $f \in C^1 \implies \forall\, x\,,\, z\, \exists\, w$ in the segment with extremes $x$ and $z$ such that $f(\,z\,) - f(\,x\,) = \langle\, \nabla f(\,w\,)\,,\, z - x\,\rangle \leq$ $\leq \|\, \nabla f(\,w\,)\,\| \|\, z - x\,\| \leq L\|\, z - x\,\|$ (directly by the definition of $L$) **[back]**

- $g = \nabla f(\,x\,)$, $d = g\,/\,\|\,g\,\|$; $[\,0 \leq\,]\,\|\,g\,\| = \langle\, g\,,\, g\,\rangle\,/\,\|\,g\,\| = \langle\, g\,,\, d\,\rangle = \frac{\partial f}{\partial d}(\,x\,) =$ $= \lim_{t\to 0}\,(\,f(\,x + td\,) - f(\,x\,)\,)\,/\,t = \lim_{t\to 0}\,|\,(\,f(\,x + td\,) - f(\,x\,)\,)\,/\,t\,| =$ $\lim_{t\to 0}\,|\,f(\,x + td\,) - f(\,x\,)\,|\,/\,|\,t\,|$. Since $f$ is L-c, $|\,f(\,x + td\,) - f(\,x\,)\,| \leq L|\,t\,|$; hence, $\|\,\nabla f(\,x\,)\,\| \leq L$ **[back]**

- $\varphi'(\alpha) = \lim_{t\to 0}(\varphi(\alpha + t) - \varphi(\alpha)) / t =$
  $= \lim_{t\to 0}(f(x + (\alpha + t)d) - f(x + \alpha d)) / t =$
  $= \lim_{t\to 0}(f([x + \alpha d] + td) - f(x + \alpha d)) / t = \frac{\partial f}{\partial d}(x + \alpha d) =$
  (definition of directional derivative) $= \langle \nabla f(x + \alpha d), d \rangle$ **[back]**

- $f : \mathbb{R}^n \to \mathbb{R}$, $Jf(x) = \nabla f(x) : \mathbb{R}^n \to \mathbb{R}^n$
  $g(t) = x + td : \mathbb{R} \to \mathbb{R}^n$, $Jg(t) = d : \mathbb{R} \to \mathbb{R}^n$
  $h(t) = f(x + td) = f(g(t)) : \mathbb{R} \to \mathbb{R}$
  $Jh(t) = h'(t) = Jf(g(t)) \cdot Jg(t) = \langle \nabla f(x + td), d \rangle$ **[back]**

- $|\varphi'(\alpha) - \varphi'(\beta)| = |\langle \nabla f(x + \alpha d), d \rangle - \langle \nabla f(x + \beta d), d \rangle| =$
  $= |\langle \nabla f(x + \alpha d) - \nabla f(x + \beta d), d \rangle| \leq$
  $\leq \| \nabla f(x + \alpha d) - \nabla f(x + \beta d) \| \| d \| \leq$ ($L$-smoothnes)
  $\leq [L\| (x + \alpha d) - (x + \beta d) \|] \| d \| = [L\| d \|^2] | \alpha - \beta |$ **[back]**

- Recall: “the derivative is the inverse of the integral”
  $f : \mathbb{R} \to \mathbb{R}$, $f \in C^0$, $F$ antiderivative of $f$ if $F'(x) = f(x)\ \forall\, x \in \mathbb{R}$
  Fundamental theorem of calculus [17] (only the needed direction):
  $F$ antiderivative of $f \implies \int_{x_-}^{x_+} f(x)dx = F(x_+) - F(x_-)\ \ \forall\, x_- \leq x_+$
  Integration is monotone [17]: $f(x) \geq g(x)\ \forall\, x \in [\, x_- \, , \, x_+ \,] \implies$
  $$F(x) \ = \ \int_{x_-}^{x_+} f(x)dx \geq G(x) \ = \ \int_{x_-}^{x_+} g(x)dx$$
  $\varphi'(\alpha) \leq \|\nabla f(x)\|^2(L\alpha - 1) \implies \varphi(\alpha) - \varphi(0) = \int_0^\alpha \varphi'(\beta)d\beta \leq$
  $\leq \|\nabla f(x)\|^2 \int_0^\alpha [\, L\beta - 1\,]d\beta = \|\nabla f(x)\|^2(\, A(\alpha) - A(0)\,)$
  where $A(\alpha) = L\alpha^2/2 - \alpha$ antiderivative of $L\alpha - 1$
  All in all, $\varphi(\alpha) - \varphi(0) \leq \|\nabla f(x)\|^2[\, L\alpha^2/2 - \alpha\,]$ **[back]**

- $g(x) = f(x) - \frac{\tau}{2}\|x\|^2$ , $\nabla^2 g(x) = \nabla^2 f(x) - \frac{\tau}{2} I$ ,
  $\lambda_n(\nabla^2 g(x)) = \lambda_n(\nabla^2 f(x)) - \tau$ **[back]**

- $\exists\, x_*$ minimum of $f \implies \nabla f(\, x_* \,) = 0$. Pick any $z \neq x_*$: for strictly convex, $f(\, z \,) > f(\, x_* \,) + \langle\, \nabla f(\, x_* \,) \,,\, z - x_* \,\rangle = f(\, x_* \,)$, for strongly convex, $f(\, z \,) \geq$ $\geq f(\, x_* \,) + \langle\, \nabla f(\, x_* \,) \,,\, z - x_* \,\rangle + \frac{\tau}{2} \|\, z - x_* \,\|^2 > f(\, x_* \,)$ (as $\|\, z - x_* \,\| > 0$) In fact, "$f \in C^1$" not needed: result true for nondifferentiable functions, easy to prove when we'll get there [**back**]

- Divide numerator and denominator by $\tau$, use the definition of $\kappa$ [**back**]

- $\|\, Q \,\| = \sqrt{\lambda_1(\, Q^T Q \,)} = \sqrt{\lambda_1(\, Q^2 \,)}$. The eigenvalues of $Q^2$ are the square of those of $Q$, hence their square root is the absolute value of those of $Q$. Clearly, the largest of the absolute values is either that of maximum eigenvalue or that of the minimum eigenvalue. [**back**]

- Recall in general $\|\, Q \,\| = \max\{\, |\, \lambda_1 \,| \,,\, |\, \lambda_n \,| \,\}$ (only simplifies if $\succeq 0$ / $\preceq 0$), and $(\, \lambda \,,\, v \,)$ eep of $Q \implies (\, 1 + c\lambda \,,\, v \,)$ eep of $I + cQ$ [**back**]

- Assuming $\alpha > 0$ is chosen so that $-1 + \alpha L \geq 0$ and $1 - \alpha\tau$ one has
  $L \geq \lambda_1 > 0 \implies -1 + \alpha\lambda_1 \leq -1 + \alpha L$
  $0 < \tau \leq \lambda_n \implies 1 - \alpha\lambda_n \leq 1 - \alpha\tau$ [**back**]

- $-1 + 2L \,/\, (L + \tau) = (-L - \tau + 2L) \,/\, (L + \tau) = (L - \tau) \,/\, (L + \tau)$
  $1 - 2\tau \,/\, (L + \tau) = \quad (L + \tau - 2\tau) \,/\, (L + \tau) = (L - \tau) \,/\, (L + \tau)$ [**back**]

- $\varphi'(\,\alpha\,) = \lim_{t\to 0}(\,\varphi(\,\alpha + t\,) - \varphi(\,\alpha\,)\,) \,/\, t =$
  $= \lim_{t\to 0}(\,f(\,x + (\alpha + t)d\,) - f(\,x + \alpha d)\,) \,/\, t =$
  $= \lim_{t\to 0}(\,f(\,[\,x + \alpha d\,] + td\,) - f(\,x + \alpha d\,)\,) \,/\, t = \frac{\partial f}{\partial d}(\,x + \alpha d\,) =$
  (definition of directional derivative) $= \langle\, \nabla f(\,x + \alpha d\,)\,,\, d\,\rangle$ [**back**]

- $f : \mathbb{R}^n \to \mathbb{R}$, $Jf(\,x\,) = \nabla f(\,x\,) : \mathbb{R}^n \to \mathbb{R}^n$
  $g(\,t\,) = x + td : \mathbb{R} \to \mathbb{R}^n$, $Jg(\,t\,) = d : \mathbb{R} \to \mathbb{R}^n$
  $h(\,t\,) = f(\,x + td\,) = f(\,g(\,t\,)\,) : \mathbb{R} \to \mathbb{R}$
  $Jh(\,t\,) = h'(\,t\,) = Jf(\,g(\,t\,)) \cdot Jg(\,t\,) = \langle\, \nabla f(\,x + td\,)\,,\, d\,\rangle$ [**back**]

- This is the announced generalisation of the result proven for quadratic $f$ $\alpha^i$ is optimal for the LS $\iff \varphi'(\alpha^i) = 0 \equiv \langle \nabla f(x^i + \alpha^i d^i), d^i \rangle = 0$ $\equiv \langle \nabla f(x^{i+1}), -\nabla f(x^i) \rangle = 0 \equiv \nabla f(x^{i+1}) \perp \nabla f(x^i)$ [**back**]

- By definition of limit, $\forall\, \delta > 0\, \exists\, h$ s.t. $\| \nabla f(x^h) \| \leq \varepsilon + \delta$; just use some $\bar{\varepsilon} < \varepsilon$ as close as you want (anyway, numerical accuracy is limited) [**back**]

- The question hardly has practical sense, since $\varepsilon$ need be (a lot) greater than the machine precision anyway: using `double` (machine precision ≈ `1e-16`), any $\varepsilon \ll$ `1e-12` is likely to be impractical. Yet, infinite-precision computation is in principle possible (albeit slow), although one cannot expect to get a solution with 0 accuracy in finite time and have to be content to finitely achieving any arbitrary accuracy. For that, it would not be right to just set $\varepsilon = 0$, as then even the first LS may never terminate. The obvious solution is to run the algorithm infinitely many times, at the $h$-th call using some fixed $\varepsilon^h > 0$, but having $\varepsilon^h \to 0$ as $h \to \infty$. Of course, one then have to use the last iteration of the $h$-th call as the starting point of the $h+1$-th call, which is all the

## Solutions XIV

information one needs to carry forward (unlike other approaches we'll see, the gradient method has "no memory" beyond the current iterate $x^i$) **[back]**

- Convexity implies $f_* = f(\, x_* \,) \geq f(, x^i \,) + \langle\, \nabla f(\, x^i \,) \,,\, x_* - x^i \,\rangle$, i.e., $\langle\, \nabla f(, x^i \,) \,,\, x_* - x^i \,\rangle \geq f(\, x^i \,) - f_*[\, \geq 0 \,]$, hence $\|\, \nabla f(\, x^i \,) \,\| \delta \geq$ $\geq \|\, \nabla f(\, x^i \,) \,\| \|\, x_* - x^i \,\| \geq |\, \langle\, \nabla f(\, x^i \,) \,,\, x_* - x^i \,\rangle \,| \geq a^i$, which finally gives $\|\, \nabla f(\, x^i \,) \,\| \leq \varepsilon \,/\, \delta \implies a^i \leq \varepsilon$ **[back]**

- $g^i = \nabla f(\, x^i \,)$, $f_* = f(\, x_* \,) \geq f(\, x^i \,) + \langle\, g^i \,,\, x_* - x^i \,\rangle + \frac{\tau}{2} \|\, x_* - x^i \,\|^2 \equiv$ $-h(\, x_* \,) = -\langle\, g^i \,,\, x_* - x^i \,\rangle - \frac{\tau}{2} \|\, x_* - x^i \,\|^2 \geq f(\, x^i \,) - f_* = a^i$. Since we don't know $x_*$, we need to overestimate the LHS, i.e., to compute $\max\{\, -h(\, x \,) \,\} = -\min\{\, h(\, x \,) \,\}$. As usual, putting $\nabla h(\, \bar{x} \,) = 0$ gives $g^i + \tau(\, \bar{x} - x^i \,) = 0 \;\equiv\; \bar{x} - x^i = -g^i \,/\, \tau$, whence $-h(\, \bar{x} \,) = \|\, g^i \,\|^2 \,/\, (2\tau)$. All in all, $\|\, g^i \,\| \leq \sqrt{2\tau\varepsilon} \implies \varepsilon \geq \|\, g^i \,\|^2 \,/\, (2\tau) \geq a^i$ **[back]**

- Since $d'(\,0\,) > 0$ and $d'(\,\alpha\,) = 0$ never happens for $\alpha > 0$, $d'(\,\alpha\,) > 0 \; \forall \alpha > 0$; in fact, $\exists\, \bar{\alpha} > 0$ s.t. $d'(\,\bar{\alpha}\,) < 0 \implies \exists\, \alpha \in (\,0\,,\,\bar{\alpha}\,)$ s.t. $d'(\,\alpha\,) = 0$ by the Intermediate Value Theorem [9, Th. 2.2.10], since $d'(\cdot) \in C^0$. Hence $d(\cdot)$ is increasing $\forall\, \alpha \geq 0$: since $d(\,0\,) = 0$, $d(\,\alpha\,) \geq 0 \;\equiv\; l(\,\alpha\,) \geq \varphi(\,\alpha\,) \; \forall \alpha \geq 0$. Since $l(\,\alpha\,) \rightarrow -\infty$ as $\alpha \rightarrow \infty$, the same must happen to $\varphi(\,\alpha\,)$ [**back**]

- Again Intermediate Value Theorem, and $d(\cdot) \in C^0$: if there was some $\alpha < \bar{\alpha}$ s.t. $d(\,\bar{\alpha}\,) < 0$, then there should be a further $\alpha' \in (\,0\,,\,\alpha\,)$ s.t. $d(\,\alpha'\,) = 0$, contradicting the assumption that $\bar{\alpha}$ is the smallest [**back**]