Smooth Unconstrained Multivariate Optimization

Antonio Frangioni

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

Computational Mathematics for Learning and Data Analysis
Master in Computer Science – University of Pisa

A.Y. 2024/25

Outline

- A way to see the algorithm: a model $f^i \approx f$ is used to construct x^{i+1} from x^i
- Simplest model: first-order one $L^i(x) = L_{x^i}(x) = f(x^i) + \nabla f(x^i)(x - x^i)$
- Idea: $x^{i+1} \in \operatorname{argmin}\{ L^i(x) : x \in \mathbb{R}^n \}$

- A way to see the algorithm: a model $f^i \approx f$ is used to construct x^{i+1} from x^i
- Simplest model: first-order one $L^i(x) = L_{x^i}(x) = f(x^i) + \nabla f(x^i)(x - x^i)$
- Idea: $x^{i+1} \in \operatorname{argmin}\{ L^i(x) : x \in \mathbb{R}^n \} = \emptyset$: L^i unbounded below on $\mathbb{R}^n$
- Anyway never blindly trust a model: L^i only "good" in $\mathcal{B}(x^i, \varepsilon)$ ("small" ε) $\implies$ x^{i+1} "good" $\equiv$ $f(x^{i+1}) < f(x^i)$ only for "small enough" α^i
- $d^i = -\nabla f(x)$ (steepest) descent direction $\implies$ (almost) same algorithm

```
procedure x = SDQ ( f , x , ε )
  while( ‖ ∇f( x ) ‖ > ε ) do
    d ← −∇f( x ); α ← stepsize( f , x , d ); x ← x + αd;
```

- stepsize$(\cdot)$ obviously the crucial part: how to choose it?
- $f(x^{i+1}) \gg f(x^i)$ for "large" α^i $\implies$ too long steps bad, algorithm diverges

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

- stepsize(f , x , d) $\in$ argmin$\{ \varphi_{x,d}(\alpha) \}$ impossible if global optimum needed
- Restricting to attraction basin on the right of 0 helps, but even exact local minimum impossible in general, only approximate one feasible
- Recall $\varphi \in C^1$: $\varphi'(\alpha) = \langle \nabla f(x + \alpha d) , d \rangle$, ∇f computed anyway: reasonable stopping criterion for "Line Search": $| \varphi'(\alpha) | \leq \varepsilon'$, but $\varepsilon' = 0$ (exact) in general not possible, how to choose it?
- Good news: the algorithm "works" without any L-smoothness assumption, with $\varepsilon' := \varepsilon \| \nabla f(x^i) \|$, ε that of the "outer" stopping condition
- Only (approximate) stationary point of φ needed $\Longrightarrow$ 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_*" + usually works well in practice with arbitrary fixed ε'

- "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 $| \varphi'(\alpha^i) | \leq \varepsilon' \; \forall \, i \implies$

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

  $\implies \| \nabla f(x) \| \leq \varepsilon$ (**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 / ∇f (other than C^1) or exact LS [4, p. 206, p. 234]
- However, general gist: the approach should be expected to work

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

Recall the stopping criterion issue

- The stopping criterion one would want: $A(x^i) \leq \varepsilon$ or $R(x^i) \leq \varepsilon$
- 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 ε 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 $a^i \leq \varepsilon$

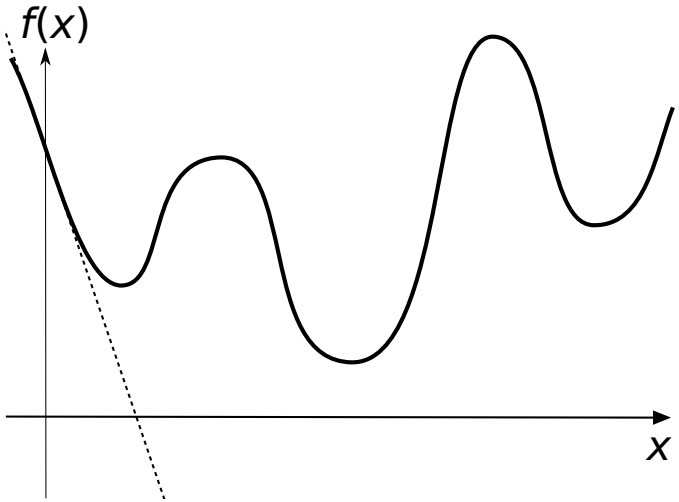
Exercise: prove: if f is $\tau-$convex, then $\|\nabla f(x^i)\| \leq \sqrt{2\tau\varepsilon} \implies a^i \leq \varepsilon$

- 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
- In learning you don't really care if $A(x^i)$ or $R(x^i)$ "very small" (but in optimization you do, because $f(x)$ can be real money)

Outline

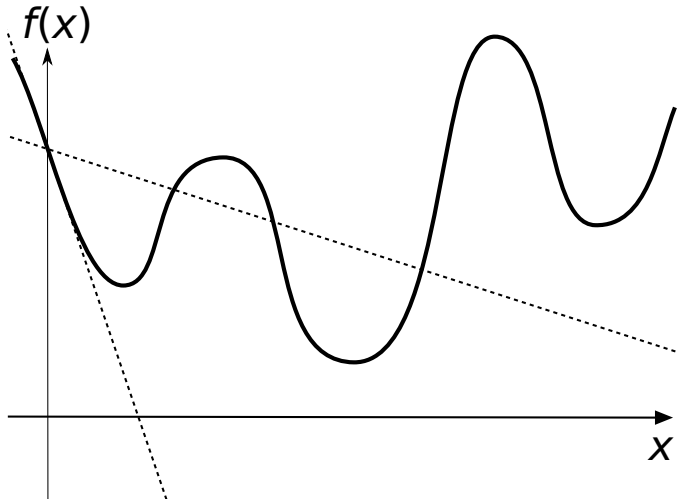
Inexact Line Search: Armijo

- Exact LS not needed $\Longrightarrow$ just need that "f^i decreases enough"



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

- Exact LS not needed $\Longrightarrow$ just need that "f^i decreases enough"

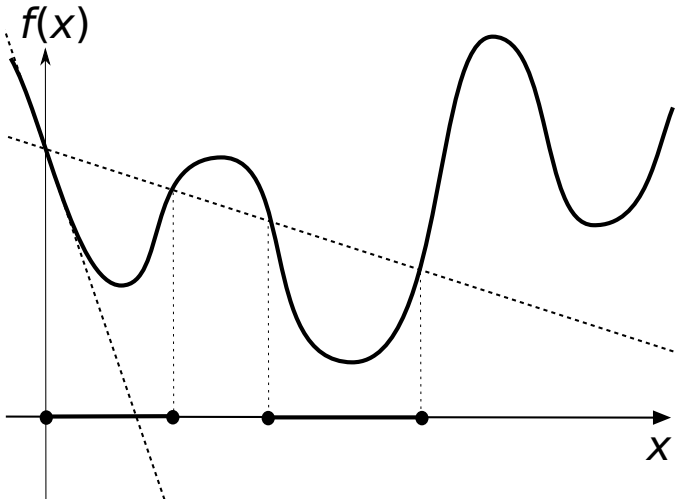


- Armijo condition: $0 < m_1 < (\ll) 1$

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

- $m_1 (\ll 1)$ of the descent promised by φ'

- Exact LS not needed $\Longrightarrow$ just need that "f^i decreases enough"

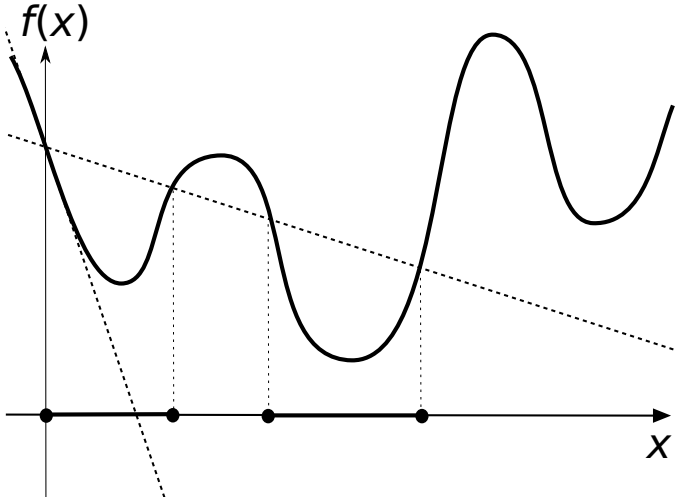


- Armijo condition: $0 < m_1 < (\ll) 1$

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

- $m_1 (\ll 1)$ of the descent promised by φ'
- Seems simple: $\alpha \searrow 0$ satisfies (A) (cf. theorem)

- Exact LS not needed $\Longrightarrow$ just need that "f^i decreases enough"



- Armijo condition: $0 < m_1 < (\ll) 1$
  (A) $\quad \varphi(\alpha) \leq \varphi(0) + m_1 \alpha \varphi'(0)$
- $m_1 (\ll 1)$ of the descent promised by φ'
- Seems simple: $\alpha \searrow 0$ satisfies (A) (cf. theorem)
  bad idea! too short steps can be dangerous!

- Example: $f(x) = -x$, $d^i = -f'(x) = 1$, $x^0 = 0$, $\alpha^i = 1/i^2 \to 0 \Longrightarrow$
  $x^i = -f^i = \sum_{k=1}^{i} 1/k^2$, $\{f^i\} \to \pi^2/6 \approx 1.645 \gg -\infty = f_*$ [15]

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

Exercise: is it possible to detect $f_* = -\infty$ as for quadratic $f(x)$? discuss

- $\alpha^i \to 0$ possible, just "not too fast": $\alpha^i = 1/i \Longrightarrow \{f^i\} \to -\infty$ [15]

Armijo + "not disaster" just works

- $\alpha^i \to 0$ "too fast" dangerous, must avoid it
- Simple form: keep it bounded away from 0 (but arbitrarily small)
- $\alpha^i \geq \bar{\alpha} > 0$ and (A) holds $\forall\, i \implies$ either $\{ f^i \} \to -\infty$ or $\{ \| \nabla f(x^i) \| \} \to 0$
- All accumulation points of $\{ x^i \}$ (if any) are stationary

Armijo + "not disaster" just works

- $\alpha^i \to 0$ "too fast" dangerous, must avoid it
- Simple form: keep it bounded away from 0 (but arbitrarily small)
- $\alpha^i \geq \bar{\alpha} > 0$ and (A) holds $\forall\, i \implies$ either $\{ f^i \} \to -\infty$ or $\{ \| \nabla f(x^i) \| \} \to 0$
- All accumulation points of $\{ x^i \}$ (if any) are stationary
- The proof, for the once (simple and informative):

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

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

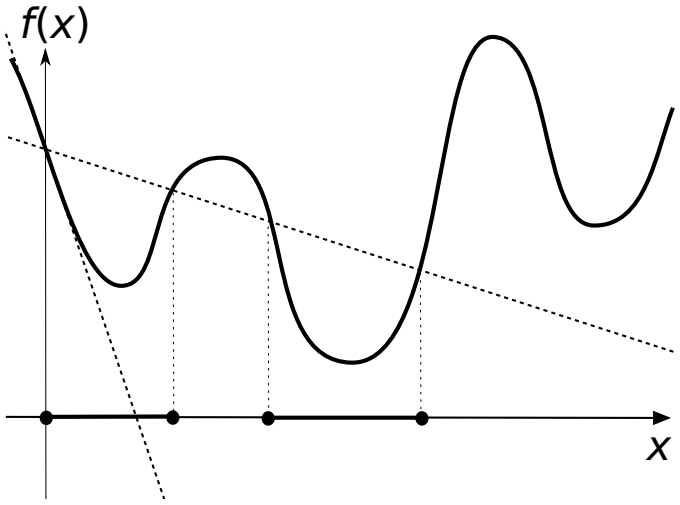
$$f^i \leq f^0 - m_1 \bar{\alpha} \varepsilon i \implies \{ f^i \} \to -\infty$$

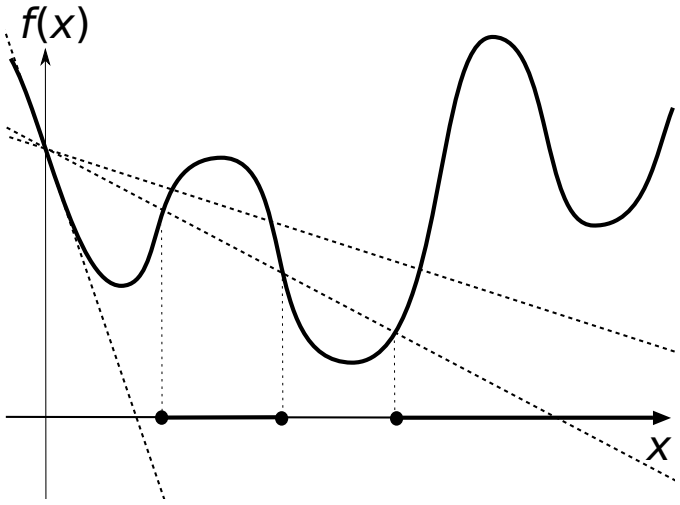
- $\alpha^i \to 0$ "too fast" dangerous, must avoid it
- Simple form: keep it bounded away from 0 (but arbitrarily small)
- $\alpha^i \geq \bar{\alpha} > 0$ and (A) holds $\forall\, i \implies$ either $\{ f^i \} \to -\infty$ or $\{ \| \nabla f(x^i) \| \} \to 0$
- All accumulation points of $\{ x^i \}$ (if any) are stationary
- The proof, for the once (simple and informative):

$$-\varphi_i'(0) = \| \nabla f(x^i) \|^2 \geq \varepsilon > 0 \text{ and (A) hold } \forall i \implies$$
$$f^{i+1} \leq f^i + m_1 \alpha^i \varphi_i'(0) \leq f^i - m_1 \bar{\alpha} \varepsilon \implies$$
$$f^i \leq f^0 - m_1 \bar{\alpha} \varepsilon i \implies \{ f^i \} \to -\infty$$

- 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 α^i does not get "too small"?
- Need add some further condition to (A)

Inexact Line Search: (strong) Wolfe

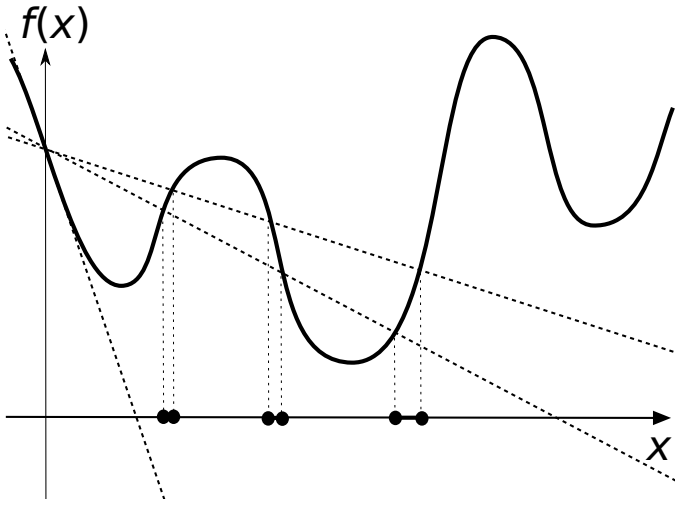




- Goldstein condition: $m_1 < m_2 < 1$

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

Inexact Line Search: (strong) Wolfe

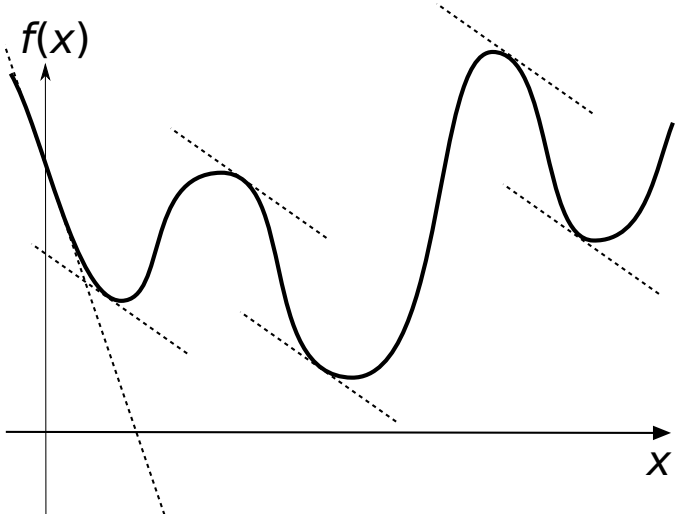


- Goldstein condition: $m_1 < m_2 < 1$

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

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

Inexact Line Search: (strong) Wolfe

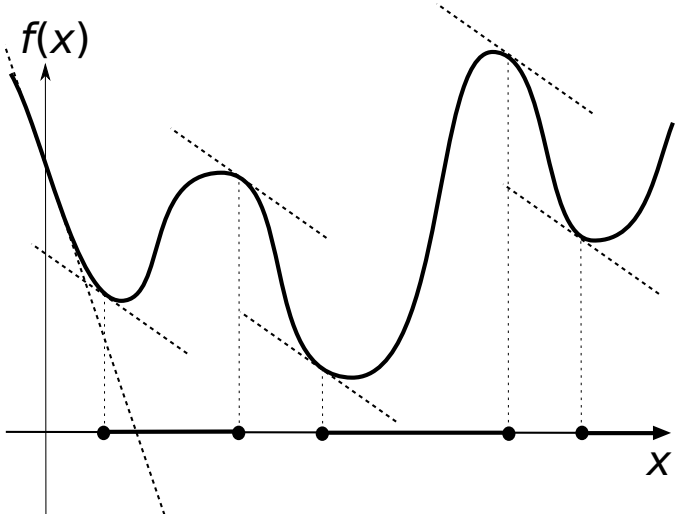


- Goldstein condition: $m_1 < m_2 < 1$

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

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

- Wolfe condition: $m_1 < m_3 < 1$



- Goldstein condition: $m_1 < m_2 < 1$

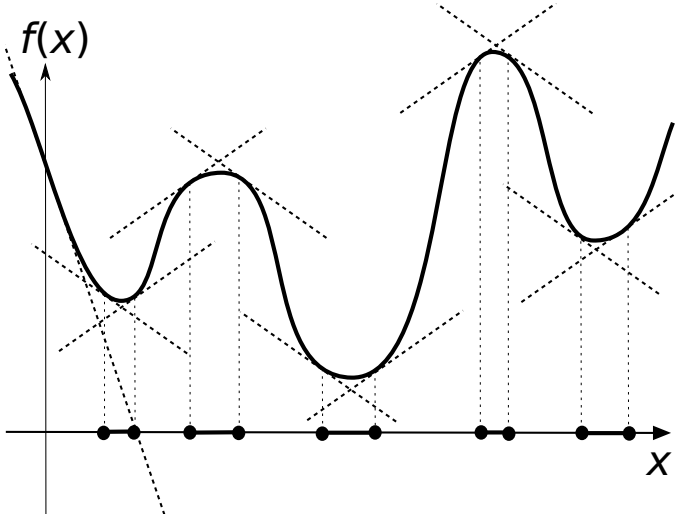
  (G) $\varphi(\alpha) \geq \varphi(0) + m_2 \alpha \varphi'(0)$

- 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 $\gg 0$)



- Goldstein condition: $m_1 < m_2 < 1$

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

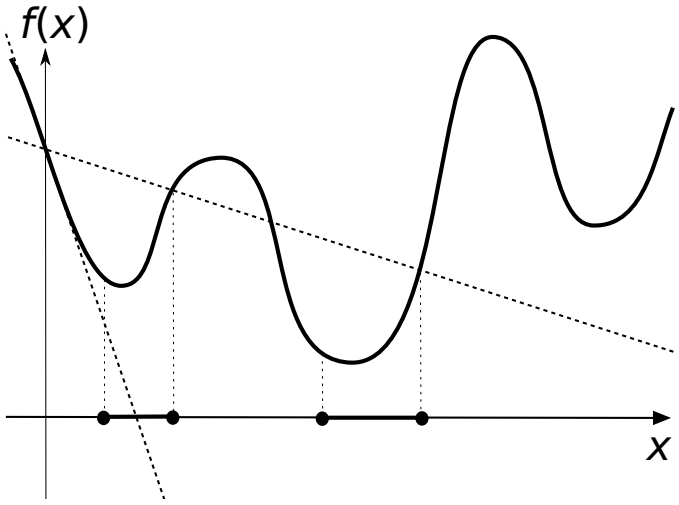
- Issue: (A) $\cap$ (G) can exclude all local minima

- Wolfe condition: $m_1 < m_3 < 1$

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

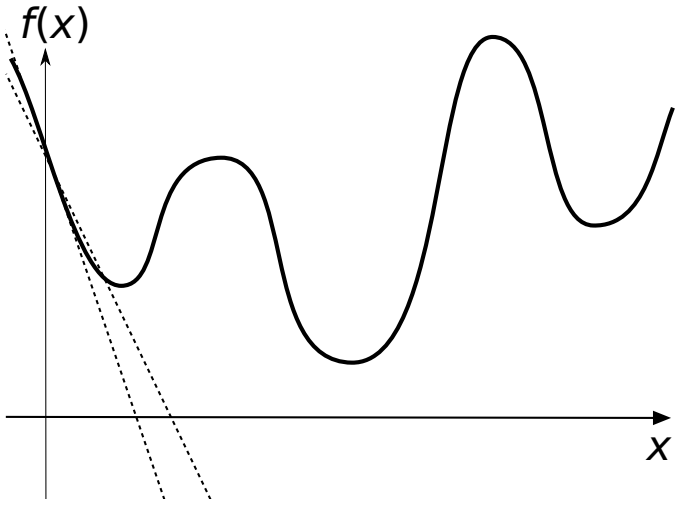
- "The derivative has to be a bit closer to 0" (but can be $\gg 0$)

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



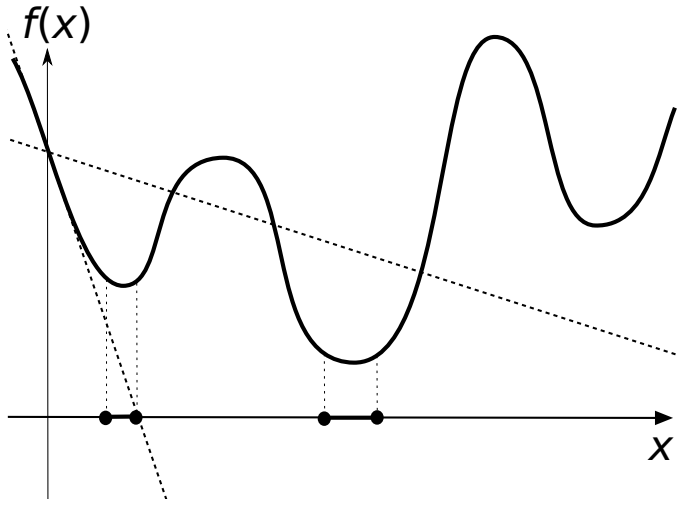
- Goldstein condition: $m_1 < m_2 < 1$
  (G) $\varphi(\alpha) \geq \varphi(0) + m_2 \alpha \varphi'(0)$
- Issue: (A) $\cap$ (G) can exclude all local minima
- Wolfe condition: $m_1 < m_3 < 1$
  (W) $\varphi'(\alpha) \geq m_3 \varphi'(0)$

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

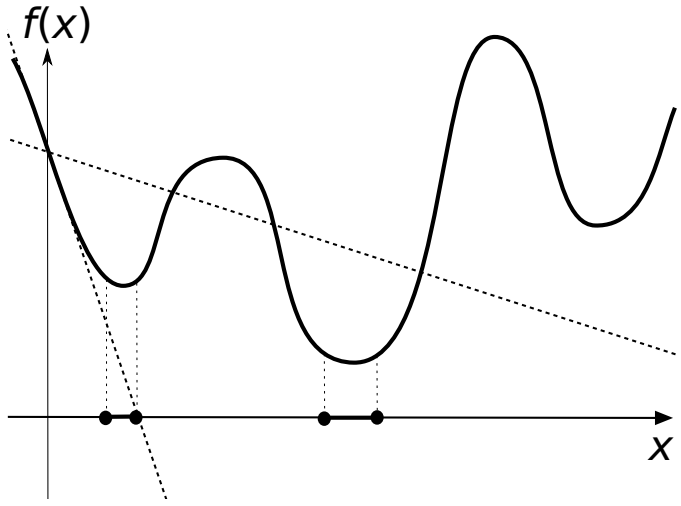


- Goldstein condition: $m_1 < m_2 < 1$
  (G) $\varphi(\alpha) \geq \varphi(0) + m_2 \alpha \varphi'(0)$
- Issue: (A) $\cap$ (G) can exclude all local minima
- Wolfe condition: $m_1 < m_3 < 1$
  (W) $\varphi'(\alpha) \geq m_3 \varphi'(0)$

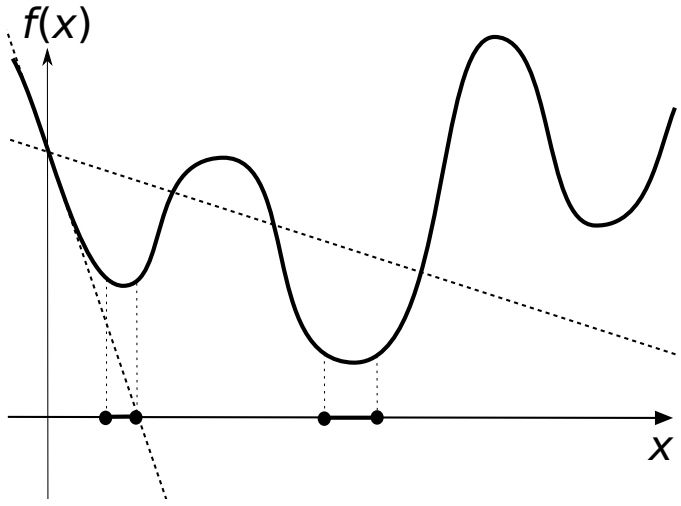
- "The derivative has to be a bit closer to 0" (but can be $\gg 0$)
- Strong Wolfe: (W') $|\varphi'(\alpha)| \leq m_3 |\varphi'(0)| = -m_3 \varphi'(0)$ [$\Longrightarrow$ (W)]
- (A) $\cap$ (W) captures all local minima (& maxima)
  unless m_1 too close to 1 (that's why usually $m_1 \approx 0.0001$)



- 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



- 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 $\exists$?



- 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 $\exists$? Of course they do

- $\varphi \in C^1 \wedge \varphi(\alpha)$ bounded below for $\alpha \geq 0 \implies \exists\, \alpha$ s.t. (A) $\cap$ (W) holds

- $\varphi \in C^1 \wedge \varphi(\alpha)$ bounded below for $\alpha \geq 0 \implies \exists\, \alpha$ s.t. (A) $\cap$ (W) holds
- Rolle's theorem [6, 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 φ (in 0): $l(\alpha) = \varphi(0) + m_1 \alpha \varphi'(0)$
- $d(\alpha) = l(\alpha) - \varphi(\alpha)$ distance between l and φ: $d(0) = 0$, $d'(\alpha) = m_1 \varphi'(0) - \varphi'(\alpha)$, $d'(0) = (m_1 - 1)\varphi'(0) > 0$
- $\nexists\, \bar{\alpha} > 0$ 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 α'
- α' $\exists$, but how do I actually find it?

- 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
  $\equiv$ any univariate optimization seen in deck 2 + 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
  [5, Algorithm 3.5], some more logic for the nasty cases
- An even simpler version: "backtracking" LS = only check (A)

procedure $\alpha = BLS\,(\,\varphi\,,\,\alpha\,,\,m_1\,,\,\tau\,)$ $\qquad$ // $\tau < 1$
  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. $\alpha \geq \tau^h\ \forall\, i \implies$ convergence
- Need conditions on f (not surprising ones) to get this

- f (globally) L-c (constant L): $|f(x)-f(z)| \leq L\|x-z\| \quad \forall x, z$
- L-c $\equiv$ boundedness of ∇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 [6, 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 ∇f: ∇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 $\equiv 0 \preceq \nabla^2 f(x) \preceq LI \equiv 0 \leq \lambda^n \leq \lambda^1 \leq L$

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

Exercise: prove "by prime principles" (definition of φ')

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"

- f L-smooth $\implies$ φ is $[L\| d \|^2]$-smooth

Exercise: Prove this "from prime principles"

- Intuitively: $\varphi'(0)$ starts large and can only decrease slowly $\implies$
  the stepsize cannot become too small

- Recall: $\exists$ smallest $0 < \alpha' < \bar{\alpha}$ s.t.
  - (A) holds $\forall \alpha \in (0, \bar{\alpha}]$, and
  - $\varphi'(\alpha') = m_3 \varphi'(0) > \varphi'(0)$

- φ is $[L\|d\|^2]$-smooth $\implies$ α' (and therefore $\bar{\alpha}$) "large":
  $L\|d\|^2(\alpha' - 0) \geq \varphi'(\alpha') - \varphi'(0) > (1 - m_3)(-\varphi'(0)) = (1 - m_3)\|d\|^2$
  $\implies [\bar{\alpha} >] \alpha' > (1 - m_3)/L$

- Note: $(1 - m_3)/L < 1/L$ but "of the same order of magnitude"

- Gradient method with AWLS or BLS converges; but how fast?

- In the quadratic case it depends on having or not directions of null curvature

- Need appropriare generalisation of the concept

- Recall: 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 $\tau > 0$ (τ-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 τ-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

Overall: efficiency of gradient method for general functions

- 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 [5, Th. 3.4]

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

  $r = ((\lambda^1 - \lambda^n)/(\lambda^1 + \lambda^n))^2$, λ_1 and λ_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 with inexact LS is "a bit" worse: $r \approx (1 - \lambda^n / \lambda^1)$

  "$\approx$" depending on m_1, m_3 [4, p. 240]
- (More) inexact LS worsens convergence rate but requires less f-calls, and this shows up in practice $\equiv$ nontrivial trade-off

Outline

- Fixed stepsize $\alpha^i = \bar{\alpha}$: simple, inexpensive but rigid
  "like a marriage in a catholic country": only one choice, better be good

- Important note: $\{ x^i \} \to x_*$ (finite) $\implies \{ \| x^{i+1} - x^i \| = \alpha^i d^i \} \to 0$
  (necessary but not sufficient (**check**))

- Using $d^i = -\nabla f(x^i) \,/\, \| \nabla f(x^i) \| \;\equiv\; \| d^i \| = 1$ would necessarily imply $\alpha^i \searrow 0$, which is not possible with fixed stepsize

- Luckily $d^i = -\nabla f(x^i)$ (not normalised): $\{ \| x^{i+1} - x^i \| \} \to 0 \iff \{ \nabla f(x^i) \} \to 0$, which is precisely what we want (stationary point)

- Still have to show: $\exists\, \bar{\alpha} > 0$ 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 ∞-ly rapidly", then only "∞-ly short α^i" are possible

- Crucial to bound how rapidly f (in fact, ∇f) changes $\equiv$ L-smoothness

- Recall: f L-smooth $\implies$ φ 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: $-\varphi'(\, 0 \,) = \| d \|^2 = \| \nabla f(\, x \,) \|^2$
- 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

Mathematically speaking: The crucial bound

- $d = -\nabla f(x) \implies \varphi'(0) = -\|\nabla f(x)\|^2 = -\|d\|^2$
- φ $[L\|d\|^2]$-smooth $\implies \varphi'(\alpha) \leq \varphi'(0) + L\|d\|^2\alpha = \|\nabla f(x)\|^2(L\alpha - 1)$
  $\implies \varphi'(\alpha) \leq 0 \ \forall \alpha \in [0, \bar{\alpha} = 1/L) \implies 1/L$ (fixed) proposed stepsize
- Issue: evaluate $\varphi(0) - \varphi(1/L)$
- Intuition: worst case for φ' is linear $\varphi'(\alpha) \approx L\alpha - 1$
  $\implies$ worst case for φ 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

- $\bar{\alpha} = 1 / L \implies L\bar{\alpha}^2/2 - \bar{\alpha} = 1 / 2L \implies$
  $\varphi^i(\alpha^i) - \varphi(0) = f(x^{i+1}) - f(x^i) \le -\| \nabla f(x^i) \|^2 / 2L$
- Can't do better if you trust the quadratic bound (which you should not)
- Immediately gives the estimate of the error decrease:
  $a^{i+1} = f(x^{i+1}) - f_* \le (a^i = f(x^i) - f_*) - \| \nabla f(x^i) \|^2 / 2L$
- Bad news: $a^{i+1} \le a^i - \Delta^i$ rather than $a^{i+1} \le r a^i \implies$ sublinear convergence
- Can prove: $a^i \le 2L\| x^1 - x_* \|^2 / (i - 1) \implies i \ge O(LD^2 / \varepsilon)$ [1, Th. 3.3]
  as in quadratic case with $\lambda_n = 0$, in fact precisely the same result
- $O(1 / \varepsilon)$ 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)
  $a^i \ge O(LD^2 / i^2) \implies i \in \Omega(1 / \sqrt{\varepsilon})$ [1, Th. 3.14]

The final complexity bound

- $\bar{\alpha} = 1/L \implies L\bar{\alpha}^2/2 - \bar{\alpha} = 1/2L \implies$
  $\varphi^i(\alpha^i) - \varphi(0) = f(x^{i+1}) - f(x^i) \le -\|\nabla f(x^i)\|^2/2L$
- Can't do better if you trust the quadratic bound (which you should not)
- Immediately gives the estimate of the error decrease:
  $a^{i+1} = f(x^{i+1}) - f_* \le (a^i = f(x^i) - f_*) - \|\nabla f(x^i)\|^2/2L$
- Bad news: $a^{i+1} \le a^i - \Delta^i$ rather than $a^{i+1} \le r a^i \implies$ sublinear convergence
- Can prove: $a^i \le 2L\|x^1 - x_*\|^2/(i-1) \implies i \ge O(LD^2/\varepsilon)$ [1, Th. 3.3]
  as in quadratic case with $\lambda_n = 0$, in fact precisely the same result
- $O(1/\varepsilon)$ 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)
  $a^i \ge O(LD^2/i^2) \implies i \in \Omega(1/\sqrt{\varepsilon})$ [1, Th. 3.14]
- Algorithms can only go so far with "nasty" problems
- Is it different "if $\lambda_n > 0$" $\equiv$ τ-convex?

The final complexity bound

- $\bar{\alpha} = 1/L \implies L\bar{\alpha}^2/2 - \bar{\alpha} = 1/2L \implies$
  $\varphi^i(\alpha^i) - \varphi(0) = f(x^{i+1}) - f(x^i) \le -\|\nabla f(x^i)\|^2/2L$
- Can't do better if you trust the quadratic bound (which you should not)
- Immediately gives the estimate of the error decrease:
  $a^{i+1} = f(x^{i+1}) - f_* \le (a^i = f(x^i) - f_*) - \|\nabla f(x^i)\|^2/2L$
- Bad news: $a^{i+1} \le a^i - \Delta^i$ rather than $a^{i+1} \le r a^i \implies$ sublinear convergence
- Can prove: $a^i \le 2L\|x^1 - x_*\|^2/(i-1) \implies i \ge O(LD^2/\varepsilon)$ [1, Th. 3.3]
  as in quadratic case with $\lambda_n = 0$, in fact precisely the same result
- $O(1/\varepsilon)$ 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)
  $a^i \ge O(LD^2/i^2) \implies i \in \Omega(1/\sqrt{\varepsilon})$ [1, Th. 3.14]
- Algorithms can only go so far with "nasty" problems
- Is it different "if $\lambda_n > 0$" $\equiv$ τ-convex? You bet!

- Want to prove: with a proper choice of α, 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 [6, Th. 5.4.5] on $\nabla f \implies \exists\, w \in [x, x_*]$ s.t.
  $\nabla f(x) - \nabla f(x_*) = \nabla f^2(w)(x - x_*) \implies$
  $z - x_* = (x - x_*) - \alpha\nabla f^2(w)(x - x_*) = (I - \alpha\nabla f^2(w))(x - x_*) \implies$
  $\|z - x_*\| \leq \|I - \alpha\nabla f^2(w)\|\|x - x_*\| \implies$ minimize $\|I - \alpha\nabla f^2(w)\|$
- 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**)

- Want to prove: with a proper choice of α, 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 [6, Th. 5.4.5] on $\nabla f \implies \exists\, w \in [x, x_*]$ s.t.
  $\nabla f(x) - \nabla f(x_*) = \nabla f^2(w)(x - x_*) \implies$
  $z - x_* = (x - x_*) - \alpha\nabla f^2(w)(x - x_*) = (I - \alpha\nabla f^2(w))(x - x_*) \implies$
  $\|z - x_*\| \le \|I - \alpha\nabla f^2(w)\|\|x - x_*\| \implies$ minimize $\|I - \alpha\nabla f^2(w)\|$
- With $\alpha = 2/(L + \tau)$ $(1/L \le \alpha < 2/L)$ converges linearly:
  $\|x^{k+1} - x_*\| \le r^k\|x^1 - x_*\|$ with $r = (L - \tau)/(L + \tau) < 1$
- $\kappa = L/\tau \ge \lambda_1/\lambda_n \ge 1$ 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
- All this may be rather slow, we need something better

Mathematically speaking: Eigenvalues & matrix norms [5, A1][8, 12] 20

- (λ, v) eigenvalue/eigenvector pair (eep) for Q ($Qv = \lambda v$):
  - $(c\lambda, v)$ eep for cQ, $c \in \mathbb{R}$ [$(cQ)v = c(Qv) = (c\lambda)v$]
  - $(1+\lambda, v)$ eep for $I+Q$ [$(I+Q)v = v + Qv = (1+\lambda)v$]
  - (λ^2, v) eep for $Q^2 = QQ$ [$(QQ)v = Q(\lambda v) = \lambda^2 v$], extends to Q^k

- $\| Q \| = \| Q \|_2 = \sqrt{\lambda_1(Q^T Q)} = \max\{ \| Qd \| / \| d \| : d \neq 0 \}$
  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)

Mathematically speaking: The choice of α

- 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 γ, the faster the convergence: choose α to minimize γ
- $\alpha = 1/\lambda_1$ 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 α decreases the max
- When $0 \leq \alpha\lambda_n - 1 \leq \alpha\lambda_1 - 1$, decreasing α decreases the max
- The optimal α must be s.t. $1 - \alpha\lambda_n > 0$ and $1 - \alpha\lambda_1 < 0 \Longrightarrow$
  $r = \max\{ -1 + \alpha\lambda_1 , 1 - \alpha\lambda_n \}$
- λ_1 , λ_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 $\uparrow$ the other $\downarrow$ 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

∞-ly many possible directions

- Outstanding assumption so far: $d^i = -\nabla f(x^i)$: really needed?
- Crucial convergence arguments:
  1. $\varphi_i'(0) = -\|\nabla f(x^i)\|^2$: "far from x_* the derivative is very negative"
  2. "you can get a non-vanishing fraction of the descent promised by $\varphi_i'(0)$"
     $\equiv$ "exact" LS or Armijo or FS + L-smooth $\implies$ α_i does not $\to 0$ "too fast"

  $\implies$ "significant decrease at each step unless $\|\nabla f(x^i)\| \to 0$"
- 2. does not really depend on the chosen direction, and

  $\exists$ many other directions that ensure 1. holds (within some factor)

∞-ly many possible directions

- Outstanding assumption so far: $d^i = -\nabla f(x^i)$: really needed?
- Crucial convergence arguments:
  1. $\varphi_i'(0) = -\|\nabla f(x^i)\|^2$: "far from x_* the derivative is very negative"
  2. "you can get a non-vanishing fraction of the descent promised by $\varphi_i'(0)$"
     $\equiv$ "exact" LS or Armijo or FS + L-smooth $\implies$ α_i does not $\to 0$ "too fast"

  $\implies$ "significant decrease at each step unless $\|\nabla f(x^i)\| \to 0$"
- 2. does not really depend on the chosen direction, and
  $\exists$ many other directions that ensure 1. holds (within some factor)
- The (parodied) twisted gradient algorithm: "$d^i = -\nabla f(x^i)$ rotated by $\pi/4$"
  $\equiv$ $d^i = R(-\nabla f(x^i))$, rotation matrix R [13]
- Gives $\varphi_i'(0) = -\|\nabla f(x^i)\|^2 \cos(\pi/4) < 0$ (**check**)
  $\implies$ convergence proofs carry forward largely unchanged
- Not just $\pi/4$: θ not too close to $\pi/2$ $\equiv$ $\cos(\theta)$ "not too small"
- ∞-ly many feasible θ and ∞-ly many $\neq d$ for each θ $\equiv$ ∞-ly many R

- Descent direction $\equiv \frac{\partial f}{\partial d^i}(x^i) < 0 \equiv \langle d^i, \nabla f(x^i) \rangle < 0 \equiv \cos(\theta^i) < 0$
  $\equiv$ "d^i points roughly in the same direction as $-\nabla f(x^i)$"

- There is a whole half space of descent directions $\implies$ a lot of flexibility

- Zoutendijk's Theorem [5, Th. 3.2]: $f \in C^1$, f L-smooth, $f_* > -\infty$,
  (A) $\cap$ (W) $\implies$ $\sum_{i=1}^{\infty} \cos^2(\theta^i) \| \nabla f(x^i) \|^2 < \infty$

- Consequence: $\sum_{i=1}^{\infty} \cos^2(\theta^i) = \infty \implies \{ \| \nabla f(x^i) \| \} \to 0$
  $\equiv$ d^i does not get $\perp \nabla f(x^i)$ "too fast" $\implies$ convergence

- Simple case: $\cos(\theta^i) \geq \bar{\theta} > 0$ (bounded away from 0),
  gradient method just the obvious case, $cos^2(\theta^i) = 1$

- Very many d^i to choose from, but which d^i is better than $-\nabla f$?

- Not clear if you only look to first-order model $\implies$ have to look farther

Outline

- Want a better direction = faster convergence? Use a better model!
- Next better model to linear ($\equiv$ gradient): quadratic
- $\nabla^2 f(x^i) \succ 0 \implies \exists$ minimum of second-order model $Q_{x^i}(z) \implies$ Newton's direction $d^i = -[\nabla^2 f(x^i)]^{-1} \nabla f(x^i)$ (**check**)
- No problem with the step here, $\alpha^i = 1$ (the minimum $\exists$)
  $\implies$ Newton's method: $x^{i+1} = x^i + d^i$ (just do step $\alpha^i = 1$ along d^i)
- Nonlinear equation interpretation: want to solve $\nabla f(x) = 0$, write $\nabla f(x) \approx \nabla f(x^i) + \nabla^2 f(x^i)(x - x^i)$ and solve linear equation instead
- We know Newton's not globally convergent $\implies$ has to be globalised
- "Easy" as $\nabla^2 f(x^i) \succ 0 \implies [\nabla^2 f(x^i)]^{-1} \succ 0 \implies d^i$ is of descent:
  $\langle \nabla f(x^i), d^i \rangle = -\nabla f(x^i)^T [\nabla^2 f(x^i)]^{-1} \nabla f(x^i) < 0$
  (but < 0 is not enough, we need it to be "negative enough")

- Globalised Newton's: simply add AWLS / BLS with $\alpha^0 = 1$
- Theorem 1: $f \in C^2$ L-smooth and τ-convex $\implies$ $\cos(\theta^i) \leq -\tau / L\,[< 0]$
  $\implies$ global convergence (via Zoutendijk)
- Theorem 2: $f \in C^3$, $\nabla f(x_*) = 0$, $\nabla^2 f(x_*) \succ 0$ $\implies$ $\exists\, \delta > 0$ s.t.
  $x^0 \in \mathcal{B}(x_*, \delta)$ $\implies$ "pure" Newton's ($\alpha^i = 1$) $\{x^i\} \to x_*$ quadratically
- Theorem 3: If $\{x^i\} \to x_*$, $\exists\, h$ s.t. $\alpha^i = 1$ satisfies (A) for all $i \geq h$
  (requires $m_1 \leq 1/2$, $m_1 > 1/2$ cuts away minimum when f quadratic)
- "Global phase" (α^i varies) + quadratically convergent "pure Newton's phase"
- Pure Newton's phase ends in $O(1)$ (≈ 6) iterations in practice
- If $\nabla^2 f$ M-smooth then global phase also "$O(1)$" [2, (9.40)]:
  $O(M^2 L^2 (f(x^0) - f_*) / \tau^5)$ (??, but quite fast in practice)

- Theorem 1, two technical steps using $\nabla^2 f(x^i)d^i = -\nabla f(x^i)$:
  - $\langle \nabla f(x^i), d^i \rangle = -(d^i)^T \nabla^2 f(x^i) d^i \leq -\tau \| d^i \|^2$
  - $\| \nabla f(x^i) \| = \| \nabla^2 f(x^i) d^i \| \leq \| \nabla^2 f(x^i) \| \| d^i \| \leq L \| d^i \|$

  $\Longrightarrow \cos(\theta^i) = \langle \nabla f(x^i), d^i \rangle / (\| \nabla f(x^i) \| \| d^i \|) \leq -\tau / L$
- Theorem 2: basically same proof as for $n = 1$
- Theorem 3 (sketch): $\{x^i\} \to x_* \implies \| \nabla f(x^i) \| \to 0 \implies \| d^i \| \to 0$

$$
\begin{aligned}
f(x^i + d^i) &= f(x^i) + \langle \nabla f(x^i), d^i \rangle + \tfrac{1}{2}(d^i)^T [\nabla^2 f(x^i)] d^i + R(d^i) \\
&= f(x^i) - \nabla f(x^i)^T [\nabla^2 f(x^i)]^{-1} \nabla f(x^i) \\
&\qquad + \tfrac{1}{2} \nabla f(x^i)^T [\nabla^2 f(x^i)]^{-1} \nabla f(x^i) + R(d^i) \\
&= f(x^i) - \tfrac{1}{2} \nabla f(x^i)^T [\nabla^2 f(x^i)]^{-1} \nabla f(x^i) + R(d^i) \\
&= f(x^i) + \tfrac{1}{2} \langle \nabla f(x^i), d^i \rangle + R(d^i)
\end{aligned}
$$

$\varphi'_{x^i, d^i}(0) = \langle \nabla f(x^i), d^i \rangle \to 0$ as $d^i \to 0$, but $R(d^i) \to 0$ faster

$\Longrightarrow$ eventually $R()$ negligible $\Longrightarrow$ eventually (A) holds with $m_1 < 1/2$

Exercise: complete the sketch of the proof of Theorem 3

Interpretation: Newton = Gradient + Space dilation

- Interesting interpretation: Newton $\equiv$ Gradient in a twisted space
- Holds for $f(x) = \frac{1}{2}x^T Q x + qx$, $d = -x - Q^{-1}q \implies$
  $\nabla f(x + d) = 0$: Newton ends in one iteration
- Trick seen already: $Q \succ 0 \implies Q = RR$, R nonsingular since Q is
- Bijective change of variable: $z = Rx \equiv x = R^{-1}z$
- $h(z) = f(R^{-1}z) = \frac{1}{2}z^T I z + qR^{-1}z$, $\nabla h(z) = z + R^{-1}q$:
  "in z-space, $\nabla^2 f(x^i)$ looks like I" $\implies$ gradient is fast
- In fact: $g = -\nabla h(z) = -z - R^{-1}q \implies \nabla h(z + g) = 0$ (**check**)
- Translate g from z-space to x-space:
  $R^{-1}g = R^{-1}(-z - R^{-1}q) = -x - Q^{-1}q = d$
- $z = Rx$ not the only choice, $z \approx Rx$ ("very $\approx$") works (will see)

- Newton's method $\equiv$ space dilation: a linear map making $\nabla^2 f$ "simple"
- Must it necessarily be $\nabla^2 f(x^i)^{-1}$? No, especially if $\nabla^2 f(x^i)^{-1} \not\succeq 0$
- $d^i \leftarrow -[H^i]^{-1}\nabla f(x^i)$, $\tau I \preceq H^i \preceq LI$, (A) $\cap$ (W) $\Longrightarrow$ global convergence
  (rewrite Theorem 1 with H^i in place of $\nabla^2 f(x^i)$)
- $\nabla^2 f \not\succ 0$: choose "small" ε^i s.t. $H^i = \nabla^2 f(x^i) + \varepsilon^i I \succ 0$
- Any $\varepsilon^i > -\lambda^n$ works ($\lambda^n < 0$), but numerical issues:
  any double $\leq$ 1e-16 "is 0" (1e-16 very optimistic, at least 1e-12)
- Algorithmic issues: $\lambda^n(\nabla^2 f(x^i) + \varepsilon I)$ "very small" $\Longrightarrow$ axes of $S(Q_{x^i}, \cdot)$
  "very elongated" $\Longrightarrow$ "x^{i+1} very far from x^i", not good for a local model
- Simple form: $\varepsilon = \max\{0, \delta - \lambda^n\}$ for appropriately chosen smallish δ
  (1e-8? 1e-4? 1e-12? hard to say in general)

- Turns out $\varepsilon = \max\{ 0 , \delta - \lambda^n \}$ solves $\min\{ \| H - \nabla^2 f(x^i) \|_2 \ : \ H \succeq \delta I \}$
- Can use $\neq$ norms: to solve $\min\{ \| H - \nabla^2 f(x^i) \|_F \ : \ H \succeq \delta I \}$
  - compute spectral decomposition $\nabla^2 f(x^i) = H \Lambda H^T$
  - $H^i = H \bar{\Lambda} H^T$ with $\bar{\lambda}^i = \max\{ \lambda^i , \delta \}$
- In both cases, if $\{ x^i \} \to x_*$ with $\nabla^2 f(x_*) \succeq \delta I \implies \varepsilon^i = 0 \equiv$ $H^i = \nabla^2 f(x^i)$ eventually $\implies$ quadratic convergence in the tail
- In both cases, $O(n^3)$; say, compute λ^n + Cholesky factorization $H^i = L^i (L^i)^T$, L^i triangular (fastest and more stable way)
- Can modify factorization on the fly (diagonal $< 0 \implies$ increase ε) [5, p. 52+]
- Whatever you do, $O(n^3)$ too much for large-scale ($n = 10^{4+}$): something way cheaper needed, $O(n^2)$ or less

A different approach: Trust Region (sketch)

- Two general forms of the process: always $x^{i+1} \leftarrow x^i + \alpha^i d^i$, but
  - line search: first choose $d^i \in \mathbb{R}^n$ (direction), then choose $\alpha^i \in \mathbb{R}$ (stepsize)
  - trust region: first choose α^i (trust radius), then choose d^i
- $\nabla^2 f(x^i) \not\succ 0 \implies \exists$ negative curvature direction along which f decreases

A different approach: Trust Region (sketch)

- Two general forms of the process: always $x^{i+1} \leftarrow x^i + \alpha^i d^i$, but
  - line search: first choose $d^i \in \mathbb{R}^n$ (direction), then choose $\alpha^i \in \mathbb{R}$ (stepsize)
  - trust region: first choose α^i (trust radius), then choose d^i
- $\nabla^2 f(x^i) \not\succ 0 \implies \exists$ negative curvature direction along which f decreases
  $\equiv$ exactly what we want when minimizing f, why excluding them?
- How? $Q^i(z)$ has no minimum ... on $\mathbb{R}^n$, but it does on a compact set
- $\mathbb{R}^n \supset \mathcal{T}^i =$ (compact) trust region around x^i "where Q_{x^i} can be trusted"
  $x^{i+1} \in \operatorname{argmin}\{ Q^i(z) : z \in \mathcal{T}^i \}$ a constrained problem
- Even worse: it is $\mathcal{NP}$-hard even for simple $\mathcal{T}$ like $\mathcal{B}_1(x^i, r)$ or $\mathcal{B}_\infty(x^i, r)$

A different approach: Trust Region (sketch)

- Two general forms of the process: always $x^{i+1} \leftarrow x^i + \alpha^i d^i$, but
  - line search: first choose $d^i \in \mathbb{R}^n$ (direction), then choose $\alpha^i \in \mathbb{R}$ (stepsize)
  - trust region: first choose α^i (trust radius), then choose d^i
- $\nabla^2 f(x^i) \not\succ 0 \implies \exists$ negative curvature direction along which f decreases
  $\equiv$ exactly what we want when minimizing f, why excluding them?
- How? $Q^i(z)$ has no minimum ... on $\mathbb{R}^n$, but it does on a compact set
- $\mathbb{R}^n \supset \mathcal{T}^i =$ (compact) trust region around x^i "where Q_{x^i} can be trusted"
  $x^{i+1} \in \operatorname{argmin}\{ Q^i(z) : z \in \mathcal{T}^i \}$ a constrained problem
- Even worse: it is $\mathcal{NP}$-hard even for simple $\mathcal{T}$ like $\mathcal{B}_1(x^i, r)$ or $\mathcal{B}_\infty(x^i, r)$
  ... but not for $\mathcal{B}_2(x^i, r)$: "round balls are simpler than kinky balls" [3]
- An optimization problem with quadratic constraints
- Which r?

- Can use any $H^i \approx \nabla^2 f(x^i)$, not necessarily $\succ 0$
- x^{i+1} optimal $\equiv$ $x^{i+1} = x^i + d^i$ and $\exists\, \lambda^i \geq 0$ s.t.

$[H^i + \lambda^i I]d^i = -\nabla f(x^i)$	[linear]	Karush-
$H^i + \lambda^i I \succeq 0$	[semidefinite]	Khun-
$\lambda^i(r - \| d^i \|) = 0$	[nonlinear]	Tucker

- $\lambda > 0 \implies$ like in line search with $\varepsilon^i = \lambda$ (but here λ unknown)
- $\| d^i \| < r \implies \lambda^i = 0 \implies$ normal Newton step ($\mathcal{T}$ has no effect)
- $\{x^i\} \to x_* \implies \| d^i \| \to 0 \implies$ eventually $\lambda^i = 0 \implies$ quadratic convergence in the tail
- Plenty of smart ways to find λ, x^{i+1} or approximate them (just as well), [5, §4.1], but matrix factorizations may be needed $\implies O(n^3)$ again
- LS: first d^i, then α^i; TR: first r ($\approx \alpha^i$), then d^i. Ultimately, similar
- In both cases, properly choose $H^i \approx \nabla^2 f(x^i)$ to reduce the cost crucial

- The space is big

- The space of H^i that give "fast convergence" is big
- Superlinear convergence if "H^i looks like $\nabla^2 f(x^i)$ along d^i" [5, Th. 3.6]
  $\lim_{i\to\infty} \| (H^i - \nabla^2 f(x^i)) d^i \| / \| d^i \| = 0$ (don't care elsewhere)
- General derivation of Quasi-Newton methods:
  $$m^i(x) = \langle \nabla f(x^i), x - x^i \rangle + \tfrac{1}{2}(x - x^i)^T H^i (x - x^i) \quad , \quad x^{i+1} = x^i + \alpha^i d^i$$
- Having computed x^{i+1} and $\nabla f(x^{i+1})$, new model
  $$m^{i+1}(x) = \langle \nabla f(x^{i+1}), x - x^{i+1} \rangle + \tfrac{1}{2}(x - x^{i+1})^T H^{i+1} (x - x^{i+1})$$
- Nice properties we would like H^{i+1} to have:
  i) $H^{i+1} \succ 0$ (the new model is strongly convex)
  ii) $\nabla m^{i+1}(x^i) = \nabla f(x^i)$ (the new model agrees with old information)
  iii) $\| H^{i+1} - H^i \|$ "small" (the new model is not too different)
- ii) $\equiv$ $H^{i+1}(x^{i+1} - x^i) = \nabla f(x^{i+1}) - \nabla f(x^i)$ "secant equation"

- Depending on choices at iteration i, i) $\cap$ ii) may not be possible
- Notation: $s^i = x^{i+1} - x^i = \alpha^i d^i$, $y^i = \nabla f(x^{i+1}) - \nabla f(x^i)$ (fixed)
  secant equation $\equiv$ (S) $H^{i+1} s^i = y^i$ (**check**)
- (S) $\implies \langle s^i, y^i \rangle = (s^i)^T H^{i+1} s^i$, i) $\cap$ ii) $\implies \langle s^i, y^i \rangle > 0$
  "curvature condition" (C) (most often written $\rho^i = 1 / \langle s^i, y^i \rangle > 0$)
- s^i need be properly chosen at iteration i for things to work at $i+1$
- Quasi-Newton $\implies$ d^i fixed, but s^i also depends on α^i which is "free"
- Very good news: (W) $\implies$ (C)
  Proof: $\varphi'(\alpha^i) = \langle \nabla f(x^{i+1}), d^i \rangle \geq m_3 \varphi'(0) = m_3 \langle \nabla f(x^i), d^i \rangle \implies$
  $\langle \nabla f(x^{i+1}) - \nabla f(x^i), d^i \rangle = \langle y^i, d^i \rangle \geq (m_3 - 1)\varphi'(0) > 0$
  but $s^i = \alpha^i d^i$ and $\alpha^i > 0 \implies \langle y^i, s^i \rangle = \alpha^i \langle y^i, d^i \rangle > 0$
- Assuming an AWLS, (C) can always be satisfied

- i) $\cup$ i)) $\cup$ iii) $\equiv$ $H^{i+1} = \text{argmin} \{ \| H - H^i \| : (\text{S}), H \succeq 0 \}$
- Appropriate "$\| \cdot \|$" [5, p. 138]: Davidon-Fletcher-Powell formula

  (DFP) $H^{i+1} = (I - \rho^i y^i (s^i)^T) H^i (I - \rho^i s^i (y^i)^T) + \rho^i y^i (y^i)^T$
- H^{i+1} = rank-two correction of H^i, $O(n^2)$ to produce H^{i+1} out of H^i
- Actually need $B^{i+1} = [H^{i+1}]^{-1}$: Sherman-Morrison-Woodbury formula [16]

  (SMW) $[A + ab^T]^{-1} = A^{-1} - A^{-1} ab^T A^{-1} / (1 - b^T A^{-1} a)$

  $\Longrightarrow$ (DFP^{-1}) $B^{i+1} = B^i + \rho^i s^i (s^i)^T - B^i y^i (y^i)^T B^i / (y^i)^T B^i y^i$

  $\Longrightarrow$ $O(n^2)$ per iteration, just matrix-vector products, no inverse
- This $\approx$ learning $\nabla^2 f$ out of samples of ∇f (learning2optimize)
- Quite efficient, but can do better

- Write (S) for B^{i+1}: $s^i = B^{i+1} y^i \implies B^{i+1} = \operatorname{argmin} \{ \| B - B^i \| : \dots \}$
  everything is symmetric, just $B \longleftrightarrow H$ and $s \longleftrightarrow y$
- Broyden-Fletcher-Goldfarb-Shanno formulæ [5, p. 139], still $O(n^2)$:

  (BFGS) $H^{i+1} = H^i + \rho^i y^i (y^i)^T - H^i s^i (s^i)^T H^i / (s^i)^T H^i s^i$

  (BFGS) $B^{i+1} = (I - \rho^i s^i (y^i)^T) B^i (I - \rho^i y^i (s^i)^T) + \rho^i s^i (s^i)^T$

  $= B^i + \rho^i [(1 + \rho^i (y^i)^T B^i y^i) s^i (s^i)^T - (B^i y^i (s^i)^T + s^i (y^i)^T B^i)]$
- Broyden family [5, § 6.3]: $H^{i+1} = \beta H^{i+1}_{\text{DFP}} + (1 - \beta) H^{i+1}_{\text{BFGS}}$, still $O(n^2)$.
- Surely satisfies (S) and $H^{i+1} \succeq 0$ if $\beta \in [0, 1]$ (but $\exists$ feasible $\beta \notin [0, 1]$)
- Flexible, good compromise between iteration cost and convergence speed, convergence theory available [5, § 6.4] (not exactly trivial)
- Important choice: B^0. Obvious solution $B^0 = \delta I$, but which δ?
  Alternative: $B^0 =$ finite difference $\approx [\nabla^2 f(x^0)]^{-1}$

Exercise: Discuss how to compute a "finite difference" and how much does it cost

- For very large n even $O(n^2)$ is way too much
- $O(n)$ new information per iteration $\nabla f(x^i)$: only keep/use last $k \ll n$
- Limited-memory BFGS (L-BFGS): just unfold the last k iterations

$$B^{i+1} = (V^i)^T B^i V + \rho^i s^i (s^i)^T \quad \text{with } V^k = I - \rho^i y^i (s^i)^T \quad \equiv$$

$$\begin{aligned} B^{i+1} = {} & (V^{i-k} V^{i-k+1} \dots V^i)^T B^{i-k} (V^{i-k} V^{i-k+1} \dots V^i) + \\ & + \rho^{i-k} (V^{i-k+1} \dots V^i)^T s^{i-k} (s^{i-k})^T (V^{i-k+1} \dots V^i) + \\ & + \rho^{i-k+1} (V^{i-k+2} \dots V^i)^T s^{i-k+1} (s^{i-k+1})^T (V^{i-k+2} \dots V^i) + \\ & + \dots + \rho^i s^i (s^i)^T \end{aligned}$$

- Memory/time cost per iteration is $O(kn)$ [5, Algorithm 7.4], but trade-off: convergence worsens as $k \searrow$ (k large $\approx$ Newton but k small $\approx$ gradient)
- Funny tidbit: can choose B^{i-k} arbitrarily anew at each i, but of course it need be sparse, e.g., $B^{i-k} = \gamma^i I$ with $\gamma^i = \langle s^i, y^{i-1} \rangle / \| y^{i-1} \|^2$
- Just one of many possible large-scale quasi-Newton variants [5, Chapter 7]

Outline

- Twisting $\equiv$ $d^i = H^i(-\nabla f(x^i))$ is at least $O(n^2)$ by definition
  (not even counting forming H^i) unless H^i "very special" $\equiv$ rather dirty tricks
- Cheaper alternative: deflecting $\equiv$ $d^i = -\nabla f(x^i) + v^i$, $O(n)$ by definition
- But how to choose v^i in the whole of $\mathbb{R}^n$ (cheaply)?
- Simple idea: $v^i = \beta^i d^{i-1}$, direction at previous iteration scaled by some β^i (?)
- If $v^0 = 0$, then $d^i = -[\sum_{h=1}^{i} \gamma^h \nabla f(x^h)]$ for some γ^i: (opposite of)
  aggregated of all past gradients $\equiv$ "history" of computation ($\approx H^i$ in BFGS)
- For twisting, easy to ensure $\varphi'_{x^i,d^i}(0) < 0$ (just $H^i \succeq 0$)
  nontrivial to choose β^i that does the same (crucial . . . or not?)
- Will clearly happen as $\beta^i \to 0$ (**check**), but then $d^i \to -\nabla f(x^i) \implies$ slow
- Need better ideas, but we know one already

Outline

procedure $x = NCG\,(\,f\,,\,x\,,\,\varepsilon\,)$
  $\nabla f^{-} = 0$;
  while($\|\,\nabla f(\,x\,)\,\| > \varepsilon$) **do**
    if($\nabla f^{-} = 0$) **then** $d \leftarrow -\nabla f(\,x\,)$;
      else { $\beta = \langle$ right deflection value $\rangle$; $d \leftarrow -\nabla f(\,x\,) + \beta d^{-}$; }
    $\alpha \leftarrow \mathrm{LS}(\,f\,,\,x\,,\,d\,)$; $x \leftarrow x + \alpha d$; $d^{-} \leftarrow d$; $\nabla f^{-} \leftarrow \nabla f(\,x\,)$;

- Many $\neq \beta$-formulæ, all $\equiv$ for quadratic f but not so here
  1. Fletcher-Reeves: $\beta^{i}_{FR} = \|\,\nabla f(\,x^{i}\,)\,\|^{2} \,/\, \|\,\nabla f(\,x^{i-1}\,)\,\|^{2}$
  2. Polak-Ribière: $\beta^{i}_{PR} = \langle\,\nabla f(\,x^{i}\,) - \nabla f(\,x^{i-1}\,)\,,\,\nabla f(\,x^{i}\,)\,\rangle \,/\, \|\,\nabla f(\,x^{i-1}\,)\,\|^{2}$
  3. Hestenes-Stiefel:
     $\beta^{i}_{HS} = \langle\,\nabla f(\,x^{i}\,) - \nabla f(\,x^{i-1}\,)\,,\,\nabla f(\,x^{i}\,)\,\rangle \,/\, \langle\,\nabla f(\,x^{i}\,) - \nabla f(\,x^{i-1}\,)\,,\,d^{i-1}\,\rangle$
  4. Dai-Yuan: $\beta^{i}_{DY} = \|\,\nabla f(\,x^{i}\,)\,\|^{2} \,/\, \langle\,\nabla f(\,x^{i}\,) - \nabla f(\,x^{i-1}\,)\,,\,d^{i-1}\,\rangle$
- f quadratic + exact LS $\Longrightarrow$ quadratic CG $\equiv$ n iterations (exact arithmetic)
  $\ll n$ if clustered eigenvalues . . . (e.g., properly preconditioned)
- LS only exact if possible (quadratic f, . . .), otherwise AWLS

- Convergence nontrivial, depends a lot on β-formula + conditions
- F-R requires $m_1 < m_2 < 1/2$ for (A) $\cap$ (W') to work
- (A) $\cap$ (W') $\not\Longrightarrow$ d^i of P-R is of descent, unless $\beta^i_{PR+} = \max\{\, \beta^i_{PR} \,,\, 0 \,\}$
  similar $\beta^i_{HS+} = \max\{\, \beta^i_{HS} \,,\, 0 \,\}$ useful for H-S
- The above is a restart: from time to time, take "plain" $-\nabla f$
- Turns out restarts are a good idea, especially for F-R:
  $\| \nabla f(x^i) \| \ll \| d^i \| \iff \cos(\theta^i) \approx 0 \equiv \nabla f(x^i) \approx\perp d^i$
  $\Longrightarrow x^{i+1} \approx x^i \Longrightarrow \cos(\theta^{i+1}) \approx 0$
  $\Longrightarrow$ one bad step leads to many bad steps, restarting cures this
- In fact, restarts help a lot in proving convergence [5, p. 127], but almost a trick: the deflection "asymptotically vanish" and the gradient does all the work
- Typical restart after n steps, not very nice when n large (or small)
- Unrestarted P-R (not using β^i_{PR+}) does not converge for some f [5, Th. 5.8]

- Efficiency n-step quadratic [5, (5.51)]: n CG steps $\approx$ 1 Newton step

  $$\| x^{i+n} - x_* \| \leq r \| x^i - x_* \|^2$$

- Makes sense: "close to x_*, $f(\cdot) \approx Q_{x_*}(\cdot)$" +

  "in n steps the CG exactly solves a quadratic function"

- Not very nice when n large

- Interesting relationships with quasi-Newton methods [5, §7.2], hybrid versions ...

- Variants surprisingly $\neq$ in practice; P-R or D-Y often better but varies a lot

- All in all: powerful approach, but not easy to manage

Outline

"Poorman's conjugate gradient": Heavy Ball Gradient

- A "slightly" different process: $x^{i+1} \leftarrow x^i - \alpha^i \nabla f(x^i) + \beta^i (x^i - x^{i-1})$
  - $\beta^i (x^i - x^{i-1}) =$ "momentum term", keep x^{i+1} going in same direction
  - while $-\nabla f(x^i)$ "force" steering the trajectory towards x_* (x^i "heavy")
- Large "momentum" $\beta^i \implies$ less "zig-zags" $\implies$ better convergence
- Hard to ensure $f(x^{i+1}) < f(x^i)$, in fact not a f-descent algorithm
- But with appropriate α^i, β^i, a(n $\approx$) linear d-descent one:
  $d^{i+1} = \| x^{i+1} - x_* \| \approx\leq r \| x^i - x_* \| = d^i$ with $r = (\sqrt{\kappa} - 1) / (\sqrt{\kappa} + 1)$
- Optimal rate [1, Th. 3.15], can't do better (except "$\approx$", we'll see why)
- $\kappa = L / \tau = 1000 \implies$ this $r \approx 0.938$, gradient $r \approx 0.996$: may seem small, but $0.996^{100} = 0.6698$, $0.938^{100} = 0.0016$, and it can show in practice
- Geared towards $\alpha^i = \alpha$, $\beta^i = \beta$ constants
  (easy, inexpensive, but rigid: need to choose well)
- Requires specific (complicated) analysis, but main ideas seen already

Mathematically speaking: Heavy Ball analysis I

- All starts from weird-ish two-terms recurrence definition of Heavy Ball:

$$\begin{bmatrix} x^{i+1} - x_* \\ x^i - x_* \end{bmatrix} = \begin{bmatrix} x^i + \beta^i (x^i - x^{i-1}) - \alpha^i (\nabla f(x^i) - \nabla f(x_*)) - x_* \\ x^i - x_* \end{bmatrix}$$

- Mean Value Theorem [6, Th. 5.4.5] applied to $\nabla f(\cdot) \implies \exists\, w^i \in [x_*, x^i]$ s.t $\nabla f(x^i) - \nabla f(x_*) = \nabla^2 f(w^i)(x^i - x_*) \implies$

$$\begin{bmatrix} x^{i+1} - x_* \\ x^i - x_* \end{bmatrix} = \begin{bmatrix} (x^i - x_*) - \alpha^i \nabla^2 f(w^i)(x^i - x_*) + \beta^i (x^i - x^{i-1}) \\ x^i - x_* \end{bmatrix} =$$

$$= \begin{bmatrix} [I - \alpha^i \nabla^2 f(w^i)](x^i - x_*) + \beta^i (x^i - x^{i-1}) + \beta^i x_* - \beta^i x_* \\ x^i - x_* \end{bmatrix} =$$

$$= \begin{bmatrix} [I - \alpha^i \nabla^2 f(w^i) + \beta I](x^i - x_*) - \beta^i (x^{i-1} - x_*) \\ x^i - x_* \end{bmatrix} =$$

$$= \begin{bmatrix} (1 + \beta^i) I - \alpha^i \nabla^2 f(w^i) & -\beta^i I \\ I & 0 \end{bmatrix} \begin{bmatrix} x^i - x_* \\ x^{i-1} - x_* \end{bmatrix}$$

- If we could find α^i , β^i such that

$$\| C^i \| = \left\| \begin{bmatrix} (1 + \beta^i) I - \alpha^i D^i & -\beta^i I \\ I & 0 \end{bmatrix} \right\| < 1 \;, \quad D^i = \nabla^2 f(w^i)$$

we would be done, but it's not that simple: $\| C^i \| > 1$

Mathematically speaking: Heavy Ball analysis II (complicated)

- C^i not symmetric, $\| C^i \| \geq \rho(C^i) =$ spectral radius $= \max_j \{ | \lambda_j(C^i) | \}$
  (careful: $\lambda_j(C^i)$ can be complex, $| \cdot |$ not the usual absolute value)

- $\rho(C^i) = \max_{j=1,\ldots,n} \{ \rho(C_j) \}$ with
$$C_j = \begin{bmatrix} 1 + \beta^i - \alpha^i \lambda_j(D) & -\beta^i \\ 1 & 0 \end{bmatrix} \in \mathbb{R}^{2\times 2}$$ (**check**) [nontrivial]

- Result: $\beta^i = \max\{ | 1 - \sqrt{\alpha^i \tau} | , | 1 - \sqrt{\alpha^i L} | \}^2 \implies$ [**extremely** tedious]
  $\rho(C^i) \leq \sqrt{\beta^i} = \max\{ | 1 - \sqrt{\alpha^i \tau} | , | 1 - \sqrt{\alpha^i L} | \}$ (**check**)

- $\alpha = 4 / (\sqrt{L} + \sqrt{\tau})^2 \implies \sqrt{\beta} = (\sqrt{L} - \sqrt{\tau}) / (\sqrt{L} + \sqrt{\tau}) < 1$
  $1 / L \leq \alpha \leq 4 / L$, growing as L / τ does, $\beta \leq 1$ (**check**)

- $r = \sqrt{\beta} = (\sqrt{\kappa} - 1) / (\sqrt{\kappa} + 1)$ optimal rate [1, Th. 3.15]

- This would be if we could prove linear convergence with $r = \sqrt{\beta}$,
  which is almost true but not quite

Mathematically speaking: Heavy Ball analysis III (++complicated) 44

- Simplifying assumption: f quadratic $\implies$ $\nabla^2 f$ costant $\implies$ C^i costant

  $$\left\| \begin{bmatrix} x^{i+1} - x_* \\ x^i - x_* \end{bmatrix} \right\| \leq \| C^i \| \left\| \begin{bmatrix} x^1 - x_* \\ x^0 - x_* \end{bmatrix} \right\| \quad (i\text{-th power, by recursion})$$

- Gelfand's formula [9]: $\rho(C) = \lim_{i \to \infty} \| C^i \|^{1/i}$ (er . . . eh?) $\implies$

  $\forall \varepsilon > 0 \; \exists h$ s.t. $\rho(C) - \varepsilon \leq \| C^i \|^{1/i} \leq \rho(C) + \varepsilon \;\; \forall i \geq h \implies$

  $\| C^i \| \leq (\rho(C) + \varepsilon)^i \implies$ converges linearly if $\rho(C) + \varepsilon < 1$

- ε arbitrary small provided h "large": "sooner or later it starts converging" (but it may not at the beginning)

- The larger h, the more the convergence rate is closer to $\rho(C)$: quasi-linear convergence with rate $\rho(C)$

- Can be proven for general L-smooth τ-convex f, we'll live to fight another day

- Works well in practice provided you find right α and β (nontrivial)

- For non-convex f converges if $\beta \in [0, 1)$, $\alpha \in (0, 2(1-\beta)/L)$ [7, p. 168] (β "free" but $\alpha \to 0$ as $\beta \to 1$, and $2/L$ already rather small to start with)

Heavy Ball Gradient: what if $\tau = 0$?

- Can prove $O(1/i)$ error, not better than gradient
  (although better in practice with properly chosen α, β)
- "Accelerated Gradient" has better theoretical convergence:

```
procedure y = ACCG( f, x, ε )
  x_ ← x; γ ← 1;
  do {   // warning: black magic ahead
  γ_ ← γ; γ ← ( √(4γ² + γ⁴) − γ² )/2; β ← γ( 1/γ_ − 1 );
  y ← x + β( x − x_ ); g ← ∇f( y ); x_ ← x; x ← y − (1/L)g;
  } while( ‖g‖ > ε );
```

  $\approx$ HB, except ∇f computed after momentum but before descent

- Optimal [1, Th. 3.14] $O(LD^2/\sqrt{\varepsilon})$ for L-smooth only [1, Th. 3.19],
  optimal linear $r = (\sqrt{\kappa}-1)/(\sqrt{\kappa}+1)$ if also τ-convex [1, Th. 3.18]
- Non-monotone but can be made so (two f computations per iteration)
- Complex theory, algorithm constructed to optimize worst-case behaviour
- In practice consistently slowish: carefully crafted to attain
  a given convergence speed, gets what it is constructed for

Outline

- Descent direction (e.g., gradient) + "reasonable" step = convergence
- Different practical inexact line searches, up to "no search at all"
- Convergence from quite bad to horrible depending on conditioning of $\nabla f(x_*)$

- Descent direction (e.g., gradient) + "reasonable" step = convergence
- Different practical inexact line searches, up to "no search at all"
- Convergence from quite bad to horrible depending on conditioning of $\nabla f(x_*)$
- "But the conditioning of $\nabla f(x_*)$ is a property of the space, master!"

- Descent direction (e.g., gradient) + "reasonable" step = convergence
- Different practical inexact line searches, up to "no search at all"
- Convergence from quite bad to horrible depending on conditioning of $\nabla f(x_*)$
- "But the speed of light is a property of the space, master!"

- Descent direction (e.g., gradient) + "reasonable" step = convergence
- Different practical inexact line searches, up to "no search at all"
- Convergence from quite bad to horrible depending on conditioning of $\nabla f(x_*)$
- "But the speed of light is a property of the space, master!"
  "OK, so let's just change the space!" [10]
- Thanks goodness, you can go (much) faster than light (warp drive)

- Descent direction (e.g., gradient) + "reasonable" step = convergence
- Different practical inexact line searches, up to "no search at all"
- Convergence from quite bad to horrible depending on conditioning of $\nabla f(x_*)$
- "But the speed of light is a property of the space, master!"
  "OK, so let's just change the space!" [10]
- Thanks goodness, you can go (much) faster than gradient (Newton's method)

- Descent direction (e.g., gradient) + "reasonable" step = convergence
- Different practical inexact line searches, up to "no search at all"
- Convergence from quite bad to horrible depending on conditioning of $\nabla f(x_*)$
- "But the speed of light is a property of the space, master!"
  "OK, so let's just change the space!" [10]
- Thanks goodness, you can go (much) faster than gradient (Newton's method)
- Second-order methods have vastly better convergence ($\nearrow$ quadratic),
  but $\nabla^2 f$ has to $\exists$, be continuous, and you have to use it
- Although, you can use $\nabla^2 f$ without ever computing it
- First-and-a-half-order methods provide interesting trade-offs
- A lot of details need be considered, numerical aspects nontrivial
- There is only so much you can get with first-order methods, but
  do not complain, as even a smooth ∇f is not granted

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

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

[3] E. de Klerk "The complexity of optimizing over a simplex, hypercube or sphere: a short survey" *Central European Journal of Operations Research* 16: 111–125, 2008 https://link.springer.com/content/pdf/10.1007/s10100-007-0052-9.pdf

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

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

[6] 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

[7] P. Ochs, *Local Convergence of the Heavy-Ball Method and iPiano for Non-convex Optimization* J. Optim. Theory Appl. 177, 153–180, 2018
https://doi.org/10.1007/s10957-018-1272-y

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

[9] Wikipedia – Gelfand's formula
https://en.wikipedia.org/wiki/Spectral_radius#Gelfand's_formula

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

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

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

[13] Wikipedia – Rotation matrix
https://en.wikipedia.org/wiki/Rotation_matrix

[14] Wikipedia – Permutation matrix
https://en.wikipedia.org/wiki/Permutation_matrix

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

[16] Wikipedia – Sherman-Morrison Formula
https://en.wikipedia.org/wiki/Sherman-Morrison_formula

[17] Wikipedia – Matrix Similarity
https://en.wikipedia.org/wiki/Matrix_similarity

Outline

- $f(x) = x^2/2 \implies f'(x) = x \implies x^{i+1} \leftarrow x^i - \alpha x^i = x^i(1-\alpha) \implies f(x^{i+1}) = f(x^i(1-\alpha)) = (1-\alpha)^2 f(x^i)$, 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**]

- 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**]

- 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 [11] (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 [11]: $f(x) \geq g(x)\ \forall x \in [x_-, x_+] \implies$
  $F(x) = \int_{x_-}^{x_+} f(x)dx \geq G(x) = \int_{x_-}^{x_+} g(x)dx$
  $\varphi'(\alpha) \leq \|\nabla f(x)\|^2(L\alpha - 1) \implies \varphi(\alpha) - \varphi(0) = \int_0^\alpha \varphi'(\beta)d\beta \leq$

$\leq \| \nabla f(x) \|^2 \int_0^\alpha [L\beta - 1] d\beta = \| \nabla f(x) \|^2 (A(\alpha) - A(0))$
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 τ, use the definition of κ [**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 (λ , 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**]

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

- $g^i = \nabla f(x^i)$, $f_* = \min\{f(x)\} \geq f(x^i) + \min\{q^i(x)\}$, with $q^i(x) = \\ = \langle g^i, x - x^i \rangle + \frac{\tau}{2}\|x - x^i\|^2 \implies \nabla q^i(\bar{x}) = 0 \equiv g^i + \tau(\bar{x} - x^i) = 0 \equiv \\ \equiv \bar{x} - x^i = -g^i/\tau \implies g^i(\bar{x}) = \langle g^i, -g^i/\tau \rangle + \frac{\tau}{2}\|-g^i/\tau\|^2 = \\ = -\|g^i\|^2/2\tau = -\|\nabla f(x^i)\|^2/2\tau$ **[back]**

- $\lim_{i\to\infty} \nabla f(x^i) = \lim_{i\to\infty} \nabla f(x^{i+1}) = \nabla f(x)$ since $f \in C^1 \equiv \nabla f \in C^0$; then, $\langle \nabla f(x^i), \nabla f(x^{i+1}) \rangle \leq \varepsilon \|\nabla f(x^i)\| \implies$ (take the limit) $\|\nabla f(x)\|^2 \leq \varepsilon \|\nabla f(x)\|$ **[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 ε need be (a lot) greater than the machine precision anyway: using `double` (machine precision $\approx$ `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 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 $\| \nabla f(x^i) \| \leq \varepsilon / \delta \implies a^i \leq \varepsilon$ (looks familiar?) [**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 [6, 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 $l(\,\alpha\,) \to -\infty$ as $\alpha \to \infty$, 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**]

- $\varphi'_{x,d}(\alpha) = \langle d, \nabla f(x+\alpha d)\rangle$, hence $\varphi'_{x,d}(0) = \|d\|\|\nabla f(x)\|\cos(\theta)$. If d where $-\nabla f(x)$ then $\theta = \pi$, since it's rotated by further 45 degrees $(\pi/2)$, then either $\theta = 3\pi/4$ or $\theta = 5\pi/4$; in either case, $\cos(\theta) = -\sqrt{2}/2 = -\cos(\pi/4)$, hence $\varphi'_{x,d}(0) = -\|\nabla f(x)\|^2\cos(\pi/4)$ [**back**]

- $Q^i(d) = f(x^i) + \langle \nabla f(x^i), d\rangle + \frac{1}{2}d^T\nabla^2 f(x^i)d$, $\nabla Q^i(d^i) = 0 \equiv$ $\equiv \nabla f(x^i) + \nabla^2 f(x^i)d^i \equiv d^i = -[\nabla^2 f(x^i)]^{-1}\nabla f(x^i)$ [**back**]

- As in Theorem 1,a $-\langle \nabla f(x^i), d^i\rangle = (d^i)^T\nabla^2 f(x^i)d^i \geq \tau\|d^i\|^2$. By Taylor's theorem, $\lim_{d\to 0} R(d)/\|d\|^2 = 0 \equiv \forall \varepsilon > 0\ \exists h$ s.t. $R(d^i) \leq \varepsilon\|d^i\|^2$ $\forall i \geq h$. Thus, $R(d^i) \leq \varepsilon\|d^i\|^2 \leq (-\varepsilon/\tau)\langle \nabla f(x^i), d^i\rangle$ Hence, $f(x^i+d^i) - f(x^i) = \frac{1}{2}\langle \nabla f(x^i), d^i\rangle + R(d^i) \leq (\frac{1}{2} - \varepsilon/\tau)\langle \nabla f(x^i), d^i\rangle =$ $(\frac{1}{2} - \varepsilon/\tau)\varphi'_{x^i,d^i}(0)$ eventually holds for all large enough i however chosen ε. Thus, the Armijo condition $f(x^i+d^i) - f(x^i) \leq m_1\varphi'_{x^i,d^i}(0)$ will eventually hold at every iteration however chosen $m_1 < 1/2$
  This uses τ-convexity, that is required for global convergence, but one could rather assume the milder $\nabla^2 f(x_*) \succ 0$ (required anyway for quadratic

convergence) and use $\lambda_n(\nabla^2 f(x_*)) > 0$ in place of τ at the cost of complicating the argument somewhat [**back**]

- Obvious (we've seen it happening), but: $g = -z - R^{-1}q$, $z + g = -R^{-1}q$, $\nabla h(z+g) = (-R^{-1}q) + R^{-1}q = 0$ [**back**]

- ii) $\equiv \nabla m^{i+1}(x) = \nabla f(x^{i+1}) + H^{i+1}(x - x^{i+1}) \implies$ $\nabla m^{i+1}(x^i) = \nabla f(x^i) \equiv \nabla f(x^{i+1}) + H^{i+1}(x^i - x^{i+1}) = \nabla f(x^i) \equiv$ $\nabla f(x^{i+1}) - \nabla f(x^i) = H^{i+1}(x^{i+1} - x^i) \equiv$ (S) [**back**]

- For any $f(\cdot)$, a finite difference approximation of the derivative $f'(x)$ can be computed as $(f(x+\varepsilon) - f(x))/\varepsilon$ for some appropriately chosen "small" ε. Hence, this also holds for $\nabla^2 f(\cdot)$, which is the Jacobian of $\nabla f(\cdot)$. A finite difference approximation of the i-th column of $\nabla^2 f(x)$ can be computed as $(\nabla f(x+\varepsilon u^i) - \nabla f(x))/\varepsilon$, u^i as usual the i-the vector of the canonical basis; in other words, $[x + \varepsilon u^i]_h = x_h$ for all $h \neq i$, while $[x + \varepsilon u^i]_i = x_i + \varepsilon$. Computing this H^0 costs $n+1$ gradient computations (i.e., as many gradient computations as iterations, and n can be large), plus it needs be inverted /

factorised which is in general $O(\,n^3\,)$. Finding the appropriate numerical value for ε is nontrivial either: too large and the approximation will be bad, too small and the numerical errors in the computation of $\nabla f(\cdot)$ will be so large that the noise overwhelms the signal and the approximation will be bad too [**back**]

- $\nabla f(\cdot) \in C^0$ and d^{i-1} fixed $\implies \lim_{\beta\to 0}\varphi'_{x^i,d^i(\,\beta\,)}(\,0\,) = \lim_{\beta\to 0}\langle\, \nabla f(\,x^i\,)\,,\, -\nabla f(\,x^i\,) + \beta^i d^{i-1}\,\rangle = -\|\,\nabla f(\,x^i\,)\,\|^2 < 0$ (otherwise the algorithm would have stopped already) [**back**]

- The first step is diagonalization of the upper-left block. $A = (\,1+\beta\,)I - \alpha D$ has eigenvalues $\lambda'_i = (\,1+\beta\,)I - \lambda_i$ and spectral decomposition $A = H\Lambda' H^T$ (λ_i, H_i those of D); thus,

$$C' = \begin{bmatrix} H & 0 \\ 0 & H \end{bmatrix}\begin{bmatrix} A & -\beta I \\ I & 0 \end{bmatrix}\begin{bmatrix} H^T & 0 \\ 0 & H^T \end{bmatrix} = \begin{bmatrix} (\,1+\beta\,)I - \alpha\Lambda & -\beta I \\ I & 0 \end{bmatrix}$$

$H^T = H^{-1} \implies C'$ similar to $C \implies$ has the same eigenvalues [17]
Now, $C' \rightsquigarrow C''$ 2×2 block diagonal by exchanges of rows and columns

$$C'' = PC'P^T = \begin{bmatrix} C_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & C_n \end{bmatrix}, \quad C_i = \begin{bmatrix} 1+\beta-\alpha\lambda_i & -\beta \\ 1 & 0 \end{bmatrix} \in \mathbb{R}^{2\times 2}$$

P permutation matrix $\implies P^T = P^{-1}$ [14] $\implies$ C'' similar to C' $\implies$ eigenvalues of C the union of those of C_i [**back**]

- The eigenvalues of C_i are the roots of the characteristic polynomial $p(\lambda) = \det(C_i - \lambda I) = \lambda^2 + (1+\beta-\alpha\lambda_i)\lambda + \beta$. These are extremely tedious (but possible) to compute and write down, the use of a symbolic system is advised (see, e.g., the screenshot below). Once this is done, it is easy (with the symbolic system) to check that the largest of the two eigenvalues is always $\leq \sqrt{\beta}$ if $\beta \geq (1-\sqrt{\alpha\lambda_i})^2$ (and $\leq$ something > 1, so wo don't care)

```
In[62]:= Eigenvalues[{{1 + b - a, -b}, {1, 0}}]

Out[62]= {1/2 (1 - a - Sqrt[(-1 + a - b)^2 - 4 b] + b), 1/2 (1 - a + Sqrt[(-1 + a - b)^2 - 4 b] + b)}

In[60]:= Reduce[Abs[1/2 (1 - a - Sqrt[(-1 + a - b)^2 - 4 b] + b)] <= Sqrt[b] && a >= 0 && b >= 0, {a, b}]

Out[60]= (a == 0 && b == 1) || (a > 0 && 1 - 2 Sqrt[a] + a <= b <= 1 + 2 Sqrt[a] + a) ||
  (0 <= a <= 1 && (0 <= b < 1 - 2 Sqrt[a] + a || b > 1 + 2 Sqrt[a] + a)) ||
  (a > 1 && b > 1 + 2 Sqrt[a] + a) || (a == 1 && b == 0) || (0 <= a < 1 && b == 0)

In[61]:= Reduce[Abs[1/2 (1 - a + Sqrt[(-1 + a - b)^2 - 4 b] + b)] <= Sqrt[b] && a >= 0 && b >= 0, {a, b}]

Out[61]= (a == 0 && b == 1) || (a > 0 && 1 - 2 Sqrt[a] + a <= b <= 1 + 2 Sqrt[a] + a) ||
  (a > 1 && (0 <= b < 1 - 2 Sqrt[a] + a || b == 0)) || (a == 1 && b == 0)
```

Hence, $\beta = \max_{i=1,\ldots,n}\{ (1 - \sqrt{\alpha\lambda_i})^2 \} \implies$
$\rho(C) \leq \sqrt{\beta} = \max\{ |1 - \sqrt{\alpha^i \tau}|, |1 - \sqrt{\alpha^i L}| \}$ [**back**]

- Since $0 < \tau \leq L$, $\alpha = 4 / (\sqrt{L} + \sqrt{\tau})^2 \leq 4 / (\sqrt{L})^2 = 4 / L$. On the other direction, $\alpha = 4 / (\sqrt{L} + \sqrt{\tau})^2 \geq 4 / (\sqrt{L} + \sqrt{L})^2 = 4 / (2\sqrt{L})^2 = L$. Note that $\tau \to 0$ (very elongated level sets) $\implies \alpha \to 4 / L$, while $\tau = L$ (perfectly circular level sets) $\implies \alpha = 1 / L$: the step is longer the more elongated are the level sets. For the rest, $\sqrt{\beta} = \max\{ |1 - \sqrt{\alpha\tau}|, |1 - \sqrt{\alpha L}| \} =$

$$= \max\left\{ \left| 1 - \sqrt{4\tau / (\sqrt{L} + \sqrt{\tau})^2} \right|, \left| 1 - \sqrt{4L / (\sqrt{L} + \sqrt{\tau})^2} \right| \right\} =$$

$$= \max\{ |(\sqrt{L} + \sqrt{\tau} - 2\sqrt{\tau}) / (\sqrt{L} + \sqrt{\tau})|, \\ |(\sqrt{L} + \sqrt{\tau} - 2\sqrt{L}) / (\sqrt{L} + \sqrt{\tau})| \} =$$

$$= \max\{ |(\sqrt{L} - \sqrt{\tau}) / (\sqrt{L} + \sqrt{\tau})|, |(\sqrt{\tau} - \sqrt{L}) / (\sqrt{L} + \sqrt{\tau})| \} =$$

$$= (\sqrt{L} - \sqrt{\tau}) / (\sqrt{L} + \sqrt{\tau}) \leq 1 \implies \beta \leq 1$$ [**back**]