Unconstrained optimization I
Gradient-type methods

Antonio Frangioni

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

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

Outline

Unconstrained optimization

Gradient method for quadratic functions

Gradient method for general functions

Exact Line Search: first-order approaches

Exact Line Search: second-order approaches

Exact Line Search: zeroth-order approaches

Inexact Line Search: Armijo-Wolfe

Really inexact Line Search: fixed stepsize

Optimization algorithms

- Iterative procedures (doh!)
- Start from initial guess x^0, some process $x^i \rightsquigarrow x^{i+1}$
- Want the sequence $\{ x^i \}$ to "go towards an optimal solution"
- Actually three different forms:
  - (strong) $\{ x^i \} \to x_*$: the whole sequence converges to an optimal solution
  - (weaker) all accumulation points of $\{ x^i \}$ (if any) are optimal solutions
  - (weakest) at least one accumulation point of $\{ x^i \}$ (if any) is optimal
- X compact helps (accumulation points always $\exists$), but here $X = \mathbb{R}^n$
- f not convex $\implies$ "optimal $\longrightarrow$ stationary point"
- Two general forms of the process:
  - line search: first choose $d^i \in \mathbb{R}^n$ (direction), then choose $\alpha^i \in \mathbb{R}$ (stepsize) s.t. $x^{i+1} \leftarrow x^i + \alpha^i d^i$
  - trust region: first choose α^i (trust radius), then choose d^i
- In ML, α^i is often called "learning rate"
- Crucial concept: model of f used to construct next iterate

Outline

First example of line search: gradient method

- Simplest idea: my model is linear
- Best linear model of f at x^i: $f^i(x) = f(x^i) + \nabla f(x^i)(x - x^i)$
- $x^{i+1} \leftarrow \operatorname{argmin}\{ f^i(x) : x \in \mathbb{R}^n \}$
- Except, of course, argmin is empty: f^i is unbounded below on $\mathbb{R}^n$
- "Go infinitely much along the steepest descent direction $d^i = -\nabla f(x^i)$"
- But this clearly is trusting the model too much: $f(x) \gg f^i(x)$ far from x^i
- As you move along d^i, ∇f changes; soon $\frac{\partial f}{\partial d^i}$ will no longer be negative
- Beware too long steps, as f will (probably) start growing after a while
- Too short steps are bad either: f will decrease, but only too little
- The best step ever: $\alpha^i \in \operatorname{argmin}\{ f(x^i + \alpha d^i) : \alpha \geq 0 \} \equiv$ exact line search (doh!)
- Then, $x^{i+1} \leftarrow x^i + \alpha^i d^i$
- Exact line search is "difficult" in general, let's start simple

Exercise: prove $\alpha^i > 0$

Gradient method for quadratic functions

- Couldn't be simpler than $f(x) = \frac{1}{2} x^T Q x + qx$
- Think $Q \succeq 0$ as otherwise f is surely unbounded below
- x_* solves $Qx = -q$ (if $\exists$), so this is linear algebra
- Inverting/factorizing Q is $O(n^3)$ in practice, can we do better?
- $d^i = -\nabla f(x^i) = -Qx^i - q$ ($O(n^2)$ to compute)
- Good news: line search is easy, $\alpha^i = \|d^i\|^2 / ((d^i)^T Q d^i) \Longrightarrow$

```
procedure x = SDQ ( Q , q , x , ε )  {
   while( ‖∇f(x)‖ > ε ) do  {
         d ← −∇f(x); α ← ‖d‖²/( dᵀQd ); x ← x + αd;
   }
}
```

Exercise: prove the formula for α^i

Exercise: there is a glaring numerical problem in that procedure, fix it

Exercise: something can go wrong with that formula: what does it mean? Improve the code to take that occurrence into account.

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

Gradient method: convergence

- "The gradient method works": what does this mean?
- Asymptotic analysis: $\varepsilon = 0 \Longrightarrow \{ x^i \}$ is/contains a minimizing sequence
- Fundamental relationship: $\langle \nabla f(x^i), \nabla f(x^{i+1}) \rangle = 0$
  Proof: $\langle d^i, \nabla f(x^{i+1}) \rangle = \frac{\partial f}{\partial d^i}(x^{i+1})$, but x^{i+1} local minimum along d^i
- $\{ x^i \} \to x \Longrightarrow \nabla f(x) = 0$
  Proof: $\lim_{i \to \infty} \langle \nabla f(x^i), \nabla f(x^{i+1}) \rangle = 0 = \langle \nabla f(x), \nabla f(x) \rangle$ (**why?**)
- Any subsequence that converges does so at a stationary point (weaker)
- Do (sub)sequence(s) converge? X compact would help, but $X = \mathbb{R}^n$
- $\varepsilon > 0 \Longrightarrow$ finitely terminates (**why?**), "no convergence required"

Exercise: prove that if $Q \succ 0$, then $\{ x^i \} \to x_*$, unique optimum

Gradient method: efficiency

- "The gradient method is (not) fast": what does this mean?
- How rapidly $\| x^i - x_* \|$ decreases ... hard; $\{ x^i \}$ may not converge different subsequences $\rightarrow$ different optima (which x_*?)
- Typically, how rapidly $\| f(x^i) - f_* \|$ decreases (eventually, it has to)
- Rate/order of convergence:

$$\lim_{i \to \infty} \frac{f(x^{i+1}) - f_*}{(\, f(x^i) - f_* \,)^p} = R$$

$p = 1, R = 1 \Longrightarrow$ sublinear convergence	$1/i \quad , \quad 1/i^2 \ldots$
$p = 1, R < 1 \Longrightarrow$ linear convergence	$\gamma^i \quad , \quad \gamma < 1$
$p = 1, R = 0 \Longrightarrow$ superlinear (!) convergence	$\gamma^{i^2} \quad , \quad \gamma < 1$
$p = 2, R > 0 \Longrightarrow$ quadratic (!!!) convergence	$\gamma^{2^i} \quad , \quad \gamma < 1$

- Linear convergence: "in the tail", $f(x^{i+1}) - f_* \approx R(\, f(x^i) - f_* \,) \Longrightarrow$
  $f(x^i) - f_* \leq (\, f(x^1) - f_* \,) R^i$, "as fast as a negative exponential"
- $f(x^i) - f_* \leq \varepsilon$ for $i \geq \log(\, (\, f(x^1) - f_* \,)/\varepsilon \,) \,/\, \log(1/R)$
- $O(\log(1/\varepsilon))$ [good!], but the constant $\rightarrow \infty$ as $R \rightarrow 1$ [bad!]

Gradient method: efficiency

- Analysis is not obvious, have to use property of x_* (unknown)
- In this case, nifty trick:

$$f_*(x) = \tfrac{1}{2}(x - x_*)^T Q(x - x_*) = f(x) + \tfrac{1}{2}x_*^T Q x_* = f(x) - f_*$$

"the error at x is the distance between x and x_* in $\|\cdot\|_Q$"

Exercise: check the above formula (hint: remember $Qx_* + q = 0$)

- One can then prove that if $Q \succ 0$ then

$$f_*(x^{i+1}) = \left(1 - \frac{\|d^i\|^4}{((d^i)^T Q d^i)((d^i)^T Q^{-1} d^i)}\right) f_*(x^i)$$

"the error decreases by exactly a constant factor at each iteration"

- Making sense of the above bound requires a bit of work

Exercise: check the above formula (hint: for $y^i = x^i - x_*$, $d^i = Qy^i$)

Gradient method: efficiency (cont.d)

- Recall a few facts:
  - $\Lambda(Q) = \lambda^1 \geq \ldots \geq \lambda^n > 0$ eigenvalues of $Q \Longrightarrow$
    $\Lambda(Q^{-1}) = 1/\lambda^n \geq \ldots \geq 1/\lambda^1 > 0$ eigenvalues of Q^{-1}
  - $\lambda^n \| x \|^2 \leq x^T Q x \leq \lambda^1 \| x \|^2 \quad \forall x \in \mathbb{R}^n$

- Hence, $\| x \|^2 / x^T Q x \geq 1/\lambda^1 \quad , \quad \| x \|^2 / x^T Q^{-1} x \geq \lambda^n$ (**check**) $\Longrightarrow$

$$\forall x \in \mathbb{R}^n \quad \frac{\|x\|^4}{(x^T Q x)(x^T Q^{-1} x)} \geq \frac{\lambda^n}{\lambda^1}$$

- A better estimate is possible (technical, just believe it):

$$\forall x \in \mathbb{R}^n \quad \frac{\|x\|^4}{(x^T Q x)(x^T Q^{-1} x)} \geq \frac{4\lambda^1 \lambda^n}{(\lambda^1 + \lambda^n)^2}$$

- A bit better: with $\lambda^1 = 1000\lambda^n$

$$\frac{\lambda^n}{\lambda^1} = 0.001 < \frac{4\lambda^1 \lambda^n}{(\lambda^1 + \lambda^n)^2} \approx 0.004$$

Gradient method: efficiency (wrap up)

- All in all:
$$f(x^{i+1}) - f_* \leq \left(\frac{\lambda^1 - \lambda^n}{\lambda^1 + \lambda^n} \right)^2 (f(x^i) - f_*)$$
"the prototype of all linear convergence results"
- Good news: the bound is dimension independent $\equiv$ does not depend on $n \Longrightarrow$ holds the same for very-large-scale
- Bad news: the bound depends badly on conditioning of Q
- Example: $\lambda^1 = 1000\lambda^n \Longrightarrow R \approx 0.996 \equiv 1/\log(1/R) \approx 576$
Note: with coarser formula $R = 0.999 \equiv 1/\log(1/R) \approx 2.301$
- With $f(x^1) - f_* = 1$, $\varepsilon = 10^{-6}$ requires ≈ 3500 iterations even for $n = 2$

Gradient method: efficiency (wrap up)

- All in all:
$$f(x^{i+1}) - f_* \leq \left(\frac{\lambda^1 - \lambda^n}{\lambda^1 + \lambda^n} \right)^2 (f(x^i) - f_*)$$
"the prototype of all linear convergence results"
- Good news: the bound is dimension independent $\equiv$ does not depend on $n \Longrightarrow$ holds the same for very-large-scale
- Bad news: the bound depends badly on conditioning of Q
- Example: $\lambda^1 = 1000\lambda^n \Longrightarrow R \approx 0.996 \equiv 1/\log(1/R) \approx 576$
Note: with coarser formula $R = 0.999 \equiv 1/\log(1/R) \approx 2.301$
- With $f(x^1) - f_* = 1$, $\varepsilon = 10^{-6}$ requires ≈ 3500 iterations even for $n = 2$
. . . but also for $n = 10^8$
- Dimension independence is liked a lot in ML, but R may $\to 1$ as n grows
- More bad news: the behaviour in practice is close to the bound
- Intuitively, the algorithm zig-zags a lot when level sets are very elongated

En passant: the stopping criterion

- The stopping criterion is not what one would want, which is

  $f(x^i) - f_* = \varepsilon_A \leq \varepsilon$ (absolute error) or $\varepsilon_A / |f_*| = \varepsilon_R \leq \varepsilon$ (relative error)

  (more or less alternative version has $|f(x^i)|$ at the denominator)

Exercise: the definition of ε_R has a glaring numerical problem, fix it

Exercise: explain exactly why ε_R is "better" than ε_A

- Except, f_* is unknown (most often) and cannot be used on-line
- Need a lower bound $\underline{f} \leq f_*$, tight at least towards termination
- Estimating f_* could be considered "the true problem"
- Often $\underline{f}$ not there, hence $\| \nabla f(x^i) \| \leq \varepsilon$ the only workable alternative
- But the relationship between the two "ε" is far from obvious
- Sometimes $\| \nabla f(x) \|$ has a "physical" meaning that can be used

Exercise: for $X = \mathcal{B}(0, r)$ and f convex, estimate ε_A when $\| \nabla f(x^i) \| \leq \varepsilon$

Outline

Gradient method: non-quadratic case

- What happens when f is a general nonlinear function?
- Good news: convergence is the same (never used "f quadratic")
- Condition $\langle \nabla f(x^i), \nabla f(x^{i+1}) \rangle = 0$ holds at local minima (but also at local maxima and saddle points), so convexity not crucial
- Good/bad news: efficiency is basically the same.
- $f \in C^2$, x_* local minimum such that $\nabla^2 f(x_*) = Q \succ 0$; if $\{ x^i \} \to x_*$, then $\{ f(x^i) \} \to f(x_*)$ linearly with the same R as in the quadratic case (depending on λ_1 and λ_n of Q)
- In the tail of the convergence process $f \approx$ its second-order model, so convergence is $\approx$ the same
- Fundamental issue: exact line search is "difficult"
- Algebraic solution (compute $f'(x - \alpha \nabla f(x))$, find its roots) possible only in a limited set of cases
- Has to algorithmically search along the line for the right α^i (doh!)

Outline

Line Search: first-order approaches

- For $\varphi(\alpha) = f(x^i + \alpha d^i) : \mathbb{R} \to \mathbb{R}$, $\varphi'(\alpha) = \langle \nabla f(x^i + \alpha d^i), d \rangle$

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

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

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

- Find α^i s.t. $\varphi'(\alpha^i) = 0$
- ∇f continuous $\implies \varphi'$ continuous (**why?**)
- α^i must exist if $\exists\, \bar{\alpha}$ s.t. $\varphi'(\bar{\alpha}) > 0$

Exercise: prove this (hint: use the intermediate value theorem)

- Obvious solution:

```
ᾱ ← 1;                          // or whatever value
while( φ'(ᾱ) > 0 ) do
    ᾱ ← 2ᾱ;                     // or whatever factor
```

- Will work in practice for all "reasonable" function
- Works if φ coercive: $\lim_{\alpha \to \infty} \varphi(\alpha) = \infty$ (ex. f strongly convex)

Exercise: construct an example where $\bar{\alpha}$ exists but it is not found

Line Search: Bisection method

- Pretty darn obvious:

```
procedure α = LSBM ( φ' , α , ε )  {
   α_- ← 0; α_+ ← α;
   while( true ) do  {
         α ← (α_+ + α_-)/2; v ← φ'( α );
         if(  |v| ≤ ε  ) then break;
         if(  v < 0  )  then α_- ← α;
                        else α_+ ← α;
   }
}
```

- Asymptotic convergence: $\varepsilon = 0$, $\{ \alpha^k \}$ infinite sequence
- $\{ \alpha^k \} \subset [0 , \bar{\alpha}] \Longrightarrow \exists$ convergent subsequence to α_* (**why?**)
- $\alpha_* \in [\alpha_-^k , \alpha_+^k] \ \forall k, \ \alpha_+^k - \alpha_-^k = \bar{\alpha} 2^{-k} \Longrightarrow \{ \alpha^k \} \to \alpha_*$ (**why?**)
  $\Longrightarrow \{ \varphi'(\alpha^k) \} \to \varphi'(\alpha_*) = 0$ (**why?**) $\Longrightarrow$ finitely terminate for $\varepsilon > 0$

Exercise: prove: φ' locally Lipschitz at $\alpha_* \Longrightarrow \{ \varphi'(\alpha^k) \} \to 0$ linearly (R?)

Exercise: construct counter-example (φ' not locally Lipschitz)

Exercise: suggest assumptions for φ' locally Lipschitz $\Longrightarrow$ linear convergence

Improving the bisection method: interpolation

- Choosing α^{k+1} "right in the middle" just the dumbest possible approach
- One knows a lot about φ: $\varphi(\alpha_-)$, $\varphi(\alpha_+)$, $\varphi'(\alpha_+)$, $\varphi'(\alpha_-)$
  (need be computed, but usually $\approx$ free if one computes φ')
- Quadratic interpolation: $a\alpha^2 + b\alpha + c$ that "agrees" with φ at α_+, α_-
- Three parameters, four conditions, something's gotta give (three cases)
- Example: $2a\alpha_+ + b = \varphi'(\alpha_+)$, $2a\alpha_- + b = \varphi'(\alpha_-) \Longrightarrow$

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

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

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

  a convex combination between α_+ and α_- (**check**)

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

Improving the bisection method: more interpolation

- It can be proven (long and complicated) that, if $\varphi \in C^3$, then quadratic interpolation has convergence of order $1 < p < 2$ (superlinear)
- For instance, the previous formula (a.k.a. "method of false position" or "secant formula") has $p = (1 + \sqrt{5})/2 \approx 1.618$

Exercise: propose a simple modification that guarantees (linear) convergence even if $\varphi \notin C^3$ while changing "as little as possible" the "normal run"

- Four conditions $\Longrightarrow$ can fit a cubic polynomial and use its minima
- Rather tedious to write down, analyse and implement
- Theoretically pays: cubic interpolation has quadratic convergence ($p = 2$)
- Seems to work pretty well in practice

Exercise (not for the faint of heart): develop cubic interpolation

Outline

Line Search: second-order approaches

- More derivatives $\Longrightarrow$ same information with less points
- $f \in C^2 \Longrightarrow \varphi''(\alpha) = d^T \nabla^2 f(x + \alpha d) d$ $\exists$ and continuous (**why?**)

Exercise: prove this using the chain rule

- Computing $\nabla^2 f \Longrightarrow$ quadratic convergence with only one point
- Newton's method (tangent method): first-order Taylor of φ' at α^k

  $\varphi'(\alpha) \approx \varphi'(\alpha^k) + \varphi''(\alpha^k)(\alpha - \alpha^k)$, solve $\varphi'(\alpha) = 0 \Longrightarrow$

  $$\alpha = \alpha^k - \varphi'(\alpha^k) / \varphi''(\alpha^k)$$

- This is clearly second-order approximation of φ

```
procedure α = LSNM ( φ' , φ'' , α , ε )   {
   while( | φ'( α ) | > ε ) do
         α ← α − φ'( α ) / φ''( α );
}
```

- Fantastically simple
- Extremely good convergence (under appropriate conditions)
- Clearly numerically delicate: what if $\varphi''(\alpha) \approx 0$?

Analysis of Newton's method

- Theoretical analysis of Newton's method instructive
- If $\varphi \in C^3$, $\varphi'(\alpha_*) = 0$ and $\varphi''(\alpha_*) \neq 0$, then $\exists \delta > 0$ s.t. if Newton's method starts at $\alpha \in [\alpha_* - \delta, \alpha_* + \delta]$, then $\{\alpha^k\} \to \alpha_*$ with $p = 2$

  Proof: the iteration gives

  $$\alpha^{k+1} - \alpha_* = \alpha^k - \alpha_* - (\varphi'(\alpha^k) - \varphi'(\alpha_*)) / \varphi''(\alpha^k)$$
  $$= [\varphi'(\alpha^k) - \varphi'(\alpha_*) + \varphi''(\alpha^k)(\alpha^k - \alpha_*)] / \varphi''(\alpha^k)$$

  For some $\beta \in [\alpha^k, \alpha^*]$, Taylor gives

  $$\varphi'(\alpha_*) = \varphi'(\alpha^k) + \varphi''(\alpha^k)(\alpha^k - \alpha_*) + \varphi'''(\beta)(\alpha^k - \alpha_*)^2/2$$
  $$\implies \quad \alpha^{k+1} - \alpha_* = [\varphi'''(\beta) / 2\varphi''(\alpha^k)](\alpha^k - \alpha_*)^2$$

  $\exists \delta > 0$ s.t. $\varphi''(\alpha) \geq k_2 > 0$ (**why?**) and $|\varphi'''(\beta)| \leq k_1 < \infty$ (**why?**)

  for $\alpha, \beta \in [\alpha_* - \delta, \alpha_* + \delta] \implies |\alpha^{k+1} - \alpha_*| \leq [k_1 / 2k_2](\alpha^k - \alpha_*)^2$

  $k_1(\alpha^k - \alpha_*)/2k_2 \leq 1 \implies |\alpha^{k+1} - \alpha_*| < |\alpha^k - \alpha_*| \implies$

  $\{\alpha^k\} \to \alpha_*$, and the convergence is quadratic
- Convergence only if $|\alpha^{k+1} - \alpha_*|$ small enough
- Nontrivial to ensure in practice

Outline

Line Search: zeroth-order approaches

- Computing $\nabla f / \nabla^2 f$ can be costly ($d^T \nabla^2 f \, d$ is $O(n^2)$ already)
- Only use φ values: less derivatives $\Longrightarrow$ more points
- Golden ratio search: assuming $\varphi(\,0\,) \leq \varphi(\,\alpha\,)$

```
procedure α = LSGRM ( φ , α , ε )  {
  α_- ← 0; α_+ ← α; α'_- ← 0.382 α; α'_+ = 0.618 α;
  while( α_+ − α_- ≤ ε ) do
      if(  φ( α'_- ) > φ( α'_+ )  )
          then { α_- ← α'_-; α'_- ← α ← α'_+; α'_+ ← 0.618(α_+ − α_-); }
          else { α_+ ← α'_+; α'_+ ← α ← α'_-; α'_- ← 0.382(α_+ − α_-); }
}
```

- $0.618 \approx (\sqrt{5}-1)/2$ (golden ratio), $0.382 = 1 - 0.618$
- Property: $0.618 \approx r = (1-r)/r \approx 0.382/0.618$, i.e., $r : 1 = (1-r) : r$
- Can compute only one $\varphi(\,\alpha\,)$ per iteration
- Can do slightly better by using $r^k = F_{n-k}/F_{n-k+1}$ (Fibonacci sequence)

Exercise: picture out graphically how it works

Exercise: analyse asymptotic and finite convergence of the approach

Gradient method and (inexact) line search

- Is $|\varphi'(\alpha^i)| \leq \varepsilon$ enough for convergence? It depends on ε (of course)
- Trick: $d^i = -\nabla f(x^i)/\|\nabla f(x^i)\| \Longrightarrow \|d^i\| = 1,\ \varphi'(0) = -\|\nabla f(x^i)\|$
- $|\varphi'(\alpha^i)| = |\langle d^i, \nabla f(x^{i+1})\rangle| = |\langle \nabla f(x^i)/\|\nabla f(x^i)\|, \nabla f(x^{i+1})\rangle|$
- $\{x^i\} \to x \longrightarrow \lim_{i\to\infty} |\langle \nabla f(x^i)/\|\nabla f(x^i)\|, \nabla f(x^{i+1})\rangle| =$
  $\langle \nabla f(x)/\|\nabla f(x)\|, \nabla f(x)\rangle = \|\nabla f(x)\| \leq \varepsilon$ (note: $\|\nabla f(x^i)\| > \varepsilon$)
- $\varepsilon > 0$ and $\{x^i\} \to x \Longrightarrow$ for finite i, x^i is approximate stationary point
- Note: with $d^i := -\nabla f(x^i)$, $\varepsilon := \varepsilon\|\nabla f(x^i)\|$
- Other assumptions on f needed to ensure $\{x^i\} \to x$ ($\mathbb{R}^n$ not compact)
- Simple one: f coercive $\equiv \lim_{\|x\|\to\infty} f(x) = +\infty$
- f continuous $\Longrightarrow$ f coercive $\Longleftrightarrow$ $S(f, v)$ compact $\forall\, v$

Exercise: prove "f coercive (+ what else neded) $\Longrightarrow$ algorithm finitely stops

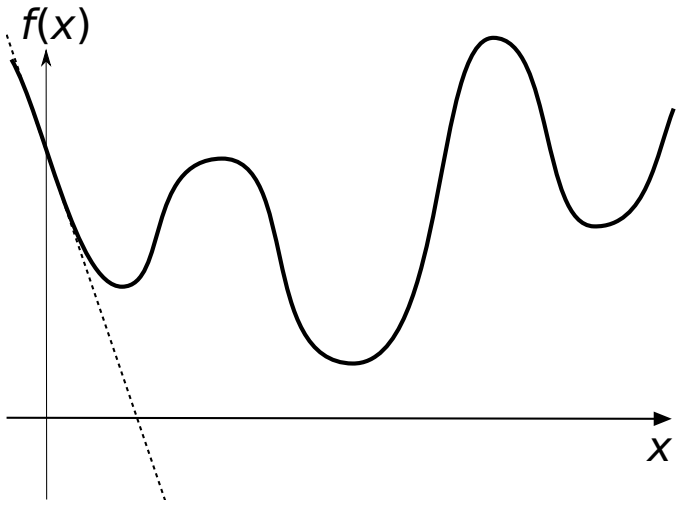
Exercise: discuss how to get asymptotic convergence ($\varepsilon = 0$)

- Do we really need a close approximation to $\nabla f(x) = 0$?

Outline

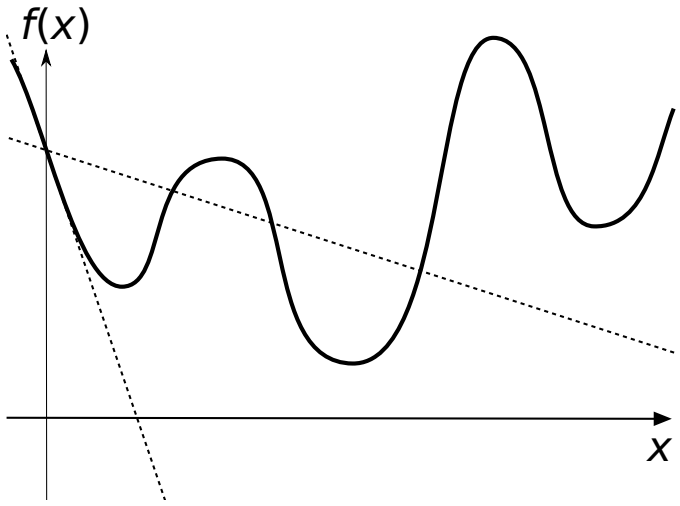
Gradient method and (really) inexact line search

- Don't need to get a local minimum, "just decrease enough"
  - Armijo condition: $0 < m_1 < 1$,



Gradient method and (really) inexact line search

- Don't need to get a local minimum, "just decrease enough"

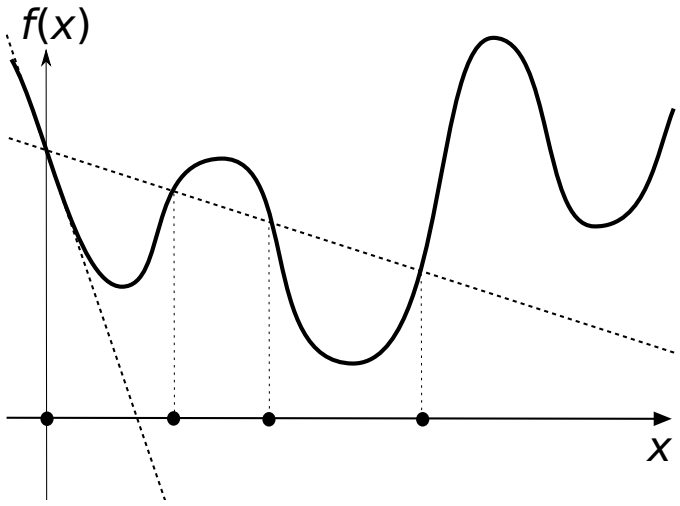


- Armijo condition: $0 < m_1 < 1$,

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

Gradient method and (really) inexact line search

- Don't need to get a local minimum, "just decrease enough"



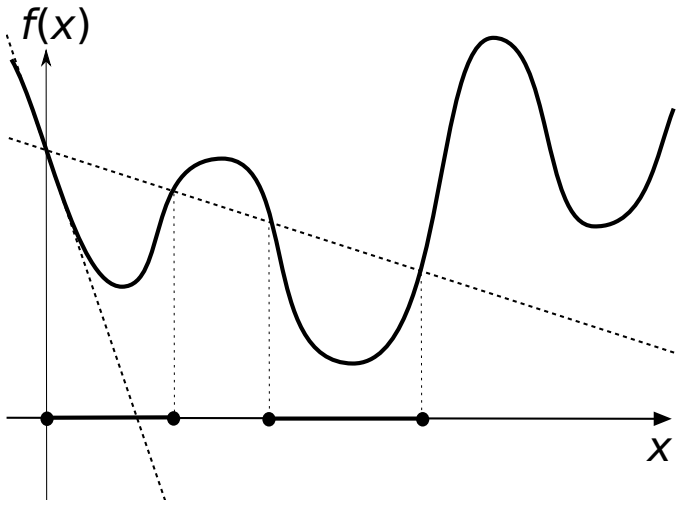
- Armijo condition: $0 < m_1 < 1$,

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

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

Gradient method and (really) inexact line search

- Don't need to get a local minimum, "just decrease enough"



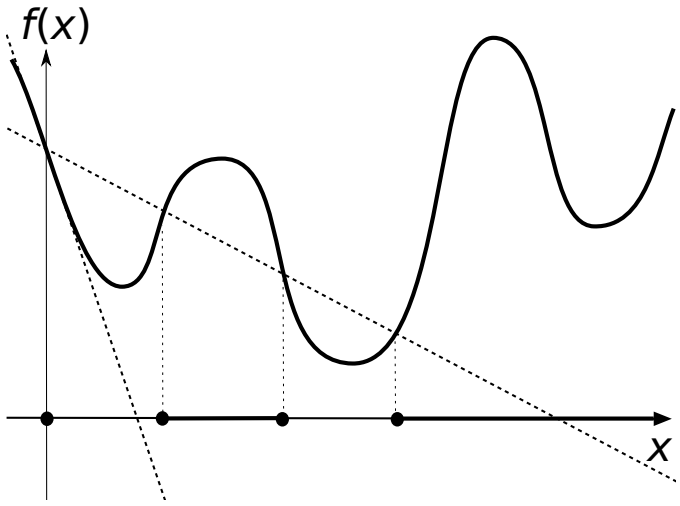
- Armijo condition: $0 < m_1 < 1$,

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

- $m_1 (\ll 1)$ of the descent promised by φ'
- Issue: arbitrarily short steps satisfy (A)

Gradient method and (really) inexact line search

- Don't need to get a local minimum, "just decrease enough"



- Armijo condition: $0 < m_1 < 1$,

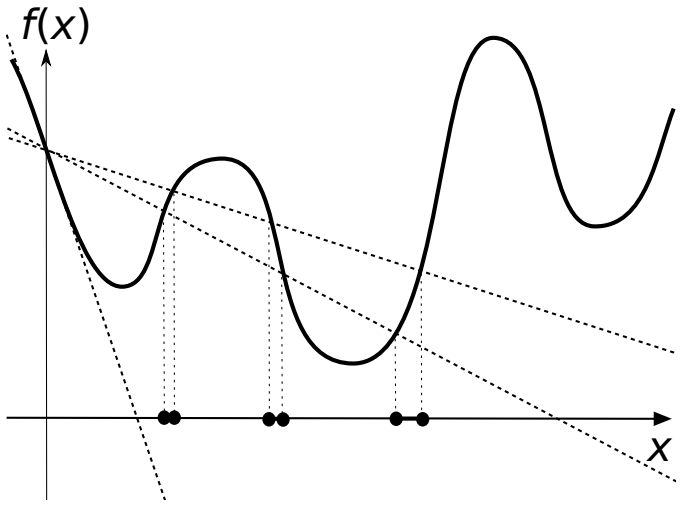
  (A) $\quad \varphi(\alpha) \le \varphi(0) + m_1 \alpha \varphi'(0)$

- $m_1 (\ll 1)$ of the descent promised by φ'
- Issue: arbitrarily short steps satisfy (A)
- Goldstein condition: $m_1 < m_2 < 1$,

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

Gradient method and (really) inexact line search

- Don't need to get a local minimum, "just decrease enough"



- Armijo condition: $0 < m_1 < 1$,

  (A) $\varphi(\alpha) \le \varphi(0) + m_1 \alpha \varphi'(0)$

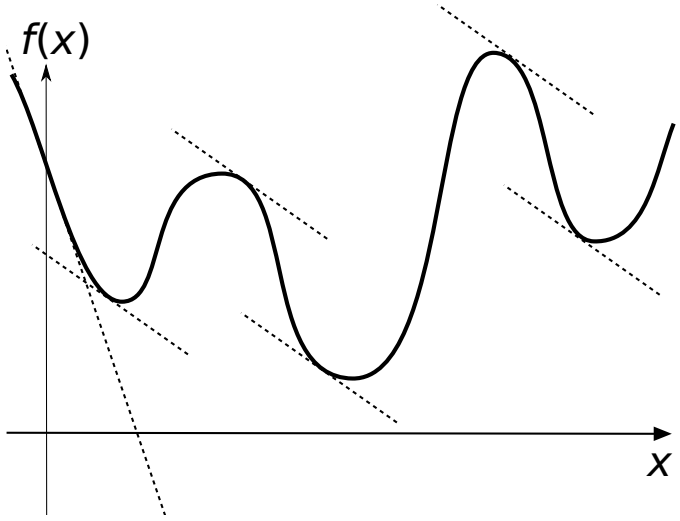
- $m_1 (\ll 1)$ of the descent promised by φ'
- Issue: arbitrarily short steps satisfy (A)
- Goldstein condition: $m_1 < m_2 < 1$,

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

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

Gradient method and (really) inexact line search

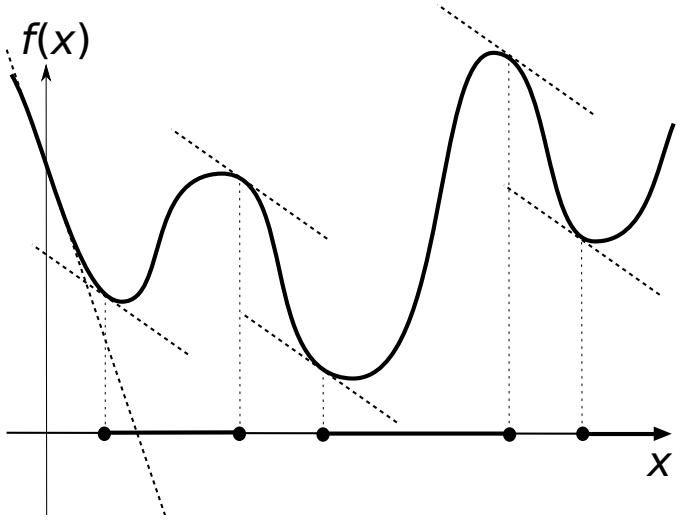
- Don't need to get a local minimum, "just decrease enough"



- Armijo condition: $0 < m_1 < 1$,
  (A) $\varphi(\alpha) \le \varphi(0) + m_1 \alpha \varphi'(0)$
- $m_1 (\ll 1)$ of the descent promised by φ'
- Issue: arbitrarily short steps satisfy (A)
- Goldstein condition: $m_1 < m_2 < 1$,
  (G) $\varphi(\alpha) \ge \varphi(0) + m_2 \alpha \varphi'(0)$
- Issue: (A) $\cap$ (G) can easily exclude all local minima
- Wolfe condition: $m_1 < m_3 < 1$, (W) $\varphi'(\alpha) \ge m_3 \varphi'(0)$

Gradient method and (really) inexact line search

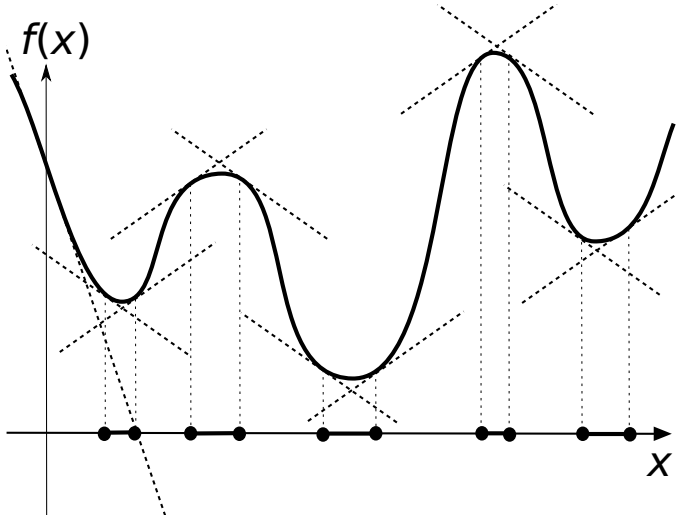
- Don't need to get a local minimum, "just decrease enough"



- Armijo condition: $0 < m_1 < 1$,
  (A) $\varphi(\alpha) \le \varphi(0) + m_1 \alpha \varphi'(0)$
- $m_1 (\ll 1)$ of the descent promised by φ'
- Issue: arbitrarily short steps satisfy (A)
- Goldstein condition: $m_1 < m_2 < 1$,
  (G) $\varphi(\alpha) \ge \varphi(0) + m_2 \alpha \varphi'(0)$
- Issue: (A) $\cap$ (G) can easily exclude all local minima
- Wolfe condition: $m_1 < m_3 < 1$, (W) $\varphi'(\alpha) \ge m_3 \varphi'(0)$
  "the curvature has to be a bit closer to 0" (but can be $\gg 0$)

Gradient method and (really) inexact line search

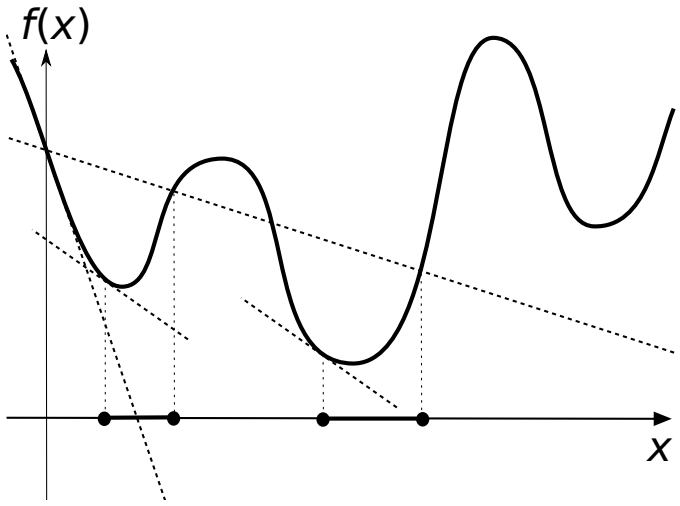
- Don't need to get a local minimum, "just decrease enough"



- Armijo condition: $0 < m_1 < 1$,
  (A) $\quad \varphi(\alpha) \le \varphi(0) + m_1 \alpha \varphi'(0)$
- $m_1 (\ll 1)$ of the descent promised by φ'
- Issue: arbitrarily short steps satisfy (A)
- Goldstein condition: $m_1 < m_2 < 1$,
  (G) $\quad \varphi(\alpha) \ge \varphi(0) + m_2 \alpha \varphi'(0)$
- Issue: (A) $\cap$ (G) can easily exclude all local minima
- Wolfe condition: $m_1 < m_3 < 1$, (W) $\quad \varphi'(\alpha) \ge m_3 \varphi'(0)$
  "the curvature has to be a bit closer to 0" (but can be $\gg 0$)
- Strong Wolfe: (W') $\quad |\varphi'(\alpha)| \le m_3 |\varphi'(0)| = -m_3 \varphi'(0)$
  cannot be $\gg 0$, but still captures all local minima (and maxima)
- Clearly, (W') $\Longrightarrow$ (W)

Gradient method and (really) inexact line search

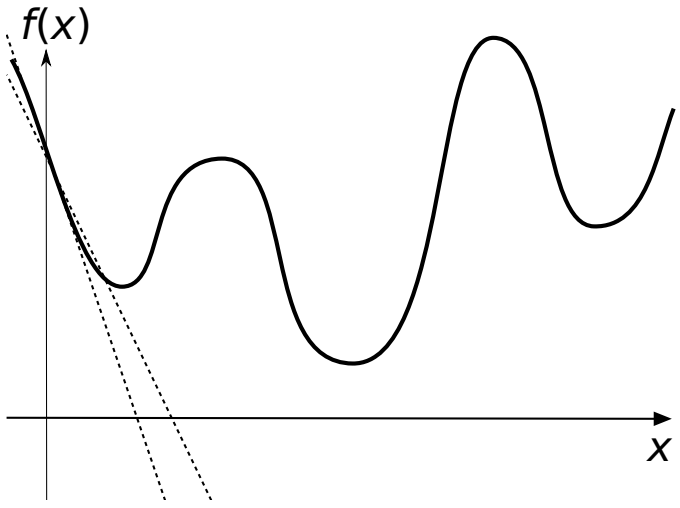
- Don't need to get a local minimum, "just decrease enough"



- Armijo condition: $0 < m_1 < 1$,
  (A) $\varphi(\alpha) \le \varphi(0) + m_1 \alpha \varphi'(0)$
- $m_1 (\ll 1)$ of the descent promised by φ'
- Issue: arbitrarily short steps satisfy (A)
- Goldstein condition: $m_1 < m_2 < 1$,
  (G) $\varphi(\alpha) \ge \varphi(0) + m_2 \alpha \varphi'(0)$
- Issue: (A) $\cap$ (G) can easily exclude all local minima
- Wolfe condition: $m_1 < m_3 < 1$, (W) $\varphi'(\alpha) \ge m_3 \varphi'(0)$
  "the curvature has to be a bit closer to 0" (but can be $\gg 0$)
- Strong Wolfe: (W') $|\varphi'(\alpha)| \le m_3 |\varphi'(0)| = -m_3 \varphi'(0)$
  cannot be $\gg 0$, but still captures all local minima (and maxima)
- Clearly, (W') $\Longrightarrow$ (W)
- (A) $\cap$ (W) / (W') typically captures all local minima

Gradient method and (really) inexact line search

- Don't need to get a local minimum, "just decrease enough"



- Armijo condition: $0 < m_1 < 1$,
  (A) $\quad \varphi(\alpha) \leq \varphi(0) + m_1 \alpha \varphi'(0)$
- $m_1 (\ll 1)$ of the descent promised by φ'
- Issue: arbitrarily short steps satisfy (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 easily exclude all local minima
- Wolfe condition: $m_1 < m_3 < 1$, (W) $\quad \varphi'(\alpha) \geq m_3 \varphi'(0)$
  "the curvature has to be a bit closer to 0" (but can be $\gg 0$)
- Strong Wolfe: (W') $\quad |\varphi'(\alpha)| \leq m_3 |\varphi'(0)| = -m_3 \varphi'(0)$
  cannot be $\gg 0$, but still captures all local minima (and maxima)
- Clearly, (W') $\Longrightarrow$ (W)
- (A) $\cap$ (W) / (W') typically captures all local minima
  unless m_1 too close to 1 (that's why $m_1 \approx 0.0001$)

Armijo-Wolfe line search

- $\varphi \in C^1 \wedge \varphi(\alpha)$ bounded below for $\alpha \geq 0 \Longrightarrow \exists\, \alpha$ s.t. (A) $\cap$ (W') holds

  Proof: $l(\alpha) = \varphi(0) + m_1 \alpha \varphi'(0)$, $d(\alpha) = l(\alpha) - \varphi(\alpha) \Longrightarrow$

  $d(0) = 0$, $d'(0) = (m_1 - 1)\varphi'(0) > 0$ $(m_1 < 1)$

  $\nexists \bar{\alpha} > 0$ s.t. $d(\bar{\alpha}) = 0 \Longrightarrow \varphi$ unbounded below (**why?**)

  Smallest $\bar{\alpha} > 0$ s.t. $d(\bar{\alpha}) = 0$: (A) is satisfied $\forall \alpha \in (0, \bar{\alpha}]$ (**why?**)

  Rolle's theorem: $d'(\bar{\alpha}) < 0 \Longrightarrow \varphi'(\bar{\alpha}) > m_1 \varphi'(0)\, (> m_3 \varphi'(0) > \varphi'(0))$

  Intermediate value theorem (on φ'): $\exists\, \alpha' \in (0, \bar{\alpha})$ s.t. $\varphi'(\alpha') = m_3 \varphi'(0)$

  $\Longrightarrow$ (W') also holds in α'

- But how do I actually find such a point?
- 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
- Hard to say if m_1 is small enough, although $m_1 = 0.0001$ most often is
- Specialized line search can be constructed for the odd case it is not
- Basic idea: find an interval $[\underline{\alpha}, \bar{\alpha}]$ that surely contains points satisfying (A) $\cap$ (W) / (W') (cf. proof above), restrict the search there inside

Exercise (not for the faint of heart): develop specialized line search

Convergence with Armijo-Wolfe line search

- ∇f Lipschitz continuous $\wedge$ (A) $\cap$ (W) always hold $\Longrightarrow$
  either f unbounded below or $\{ \| \nabla f(x^i) \| \} \to 0$
  Proof: (W) $\Longrightarrow \varphi'(\alpha^i) - \varphi'(0) \geq (1 - m_3)(-\varphi'(0)) =$
  ∇f Lipschitz $\Longrightarrow \varphi'$ Lipschitz and L does not depend on x^i (**check**)
  $\Longrightarrow \alpha^i \geq (1 - m_3)(-\varphi'(0))/L$ (**check: where has $|\cdot|$ gone?**)
  $-\varphi'(0) = \| \nabla f(x^i) \| \geq \varepsilon > 0 \Longrightarrow \alpha^i \geq \delta > 0$
  (A) $\Longrightarrow f(x^{i+1}) \leq f(x^i) - m_1 \alpha^i \| \nabla f(x^i) \| \leq f(x^i) - m_1 \delta \varepsilon \Longrightarrow$
  $\{ f(x^i) \} \to -\infty$ (or $\{ \| \nabla f(x^i) \| \} \to 0$)
- Usual stuff: $\{ x^i \} \to x_* \Longrightarrow x_*$ a stationary point
- Hence, the algorithm finitely terminates with $\varepsilon > 0$
- Insight from the proof: (W) (+ Lipschitz) serve to ensure that
  $\alpha^k \geq c \| \nabla f(x^i) \|$ for some $c > 0$
- Can we get the same in a simpler way?

Backtracking line search

- Backtracking line search:

```
procedure α = BLS ( φ , φ' , α , m1 , τ )  {
  while( φ( α ) > φ( 0 ) + m1 α φ'( 0 ) ) do α ← τ α;
  }
```

- ∇f Lipschitz $\Longrightarrow$ gradient method with BLS works

  Proof: for simplicity, $\alpha = 1$ (input). Remember the proof:

  $\exists\, \bar{\alpha}$ s.t. (A) holds $\forall \alpha \in (0, \bar{\alpha}]$ and $\varphi'(\bar{\alpha}) > m_1 \varphi'(0) > \varphi'(0) \Longrightarrow$

  $L(\bar{\alpha} - 0) \geq \varphi'(\bar{\alpha}) - \varphi'(0) > (1 - m_1)(-\varphi'(0)) \Longrightarrow$

  $\bar{\alpha} > (1 - m_1) \| \nabla f(x^i) \| / L$ (same as before)

  $\| \nabla f(x^i) \| > \varepsilon \ \ \forall i \Longrightarrow \bar{\alpha} > \delta > 0 \ \ \forall i$

  $h = \min\{ k \,:\, \tau^{-k} \leq \delta \} \Longrightarrow \alpha^i \geq \tau^{-h} > 0 \ \ \forall\, i \Longrightarrow$

  $f(x^{i+1}) \leq f(x^i) - m_1 \tau^{-h} \varepsilon \Longrightarrow \{ f(x^i) \} \to -\infty$ or ↯

- Now, $\{ x^i \} \to x \Longrightarrow x$ stationary blah blah
- Fundamental trick: α^i can $\searrow 0$, but only as fast as $\| \nabla f(x^i) \|$
- Would be simpler if $\alpha^i \geq \delta > 0$ for good

Exercise: remove assumption $\alpha = 1$ (input)

Outline

Line Search: really really inexact

- $\ldots \equiv$ no line search at all $\equiv$ fixed step size
- Recall ∇f Lipschitz $\Longrightarrow f(y) \leq f(x) + \nabla f(x)(\, y - x\,) + \frac{L}{2}\|y - x\|^2$
- $y := x^{i+1}$, $x := x^i$, $y - x := -\alpha \nabla f(x^i) \Longrightarrow$
  $$f(x^{i+1}) - f(x^i) \leq (\, L\alpha^2/2 - \alpha\,)\|\, \nabla f(x^i)\,\|^2 \text{ (\textbf{check})}$$
- Powerful idea: find α that provides best worst-case improvement
- $v(\alpha) = L\alpha^2/2 - \alpha$, $v'(\alpha) = L\alpha - 1 = 0 \Longrightarrow \alpha_* = 1/L$, $v(\alpha_*) = -1/2L$
- All in all: $f(x^{i+1}) - f(x^i) \leq -\|\, \nabla f(x^i)\,\|^2/(2L)$
- Can't do better if you trust the quadratic upper estimate
  (which of course must not be trusted)
- In fact, $\alpha^i = 1/L$ terrible in practice $\Longrightarrow$ use the previous methods
- Enticing because simple and inexpensive
- Selecting the parameters that lead to best performances for a model a very powerful idea in general

Fixed stepsize: convergence rate

- Once you have convergence, you can talk efficiency (easier with α fixed)
- Already know the error decreases, but how fast?
  $\Delta^{i+1} := f(x^{i+1}) - f(x_*) \leq (\,\Delta^i := f(x^i) - f(x_*)\,) - \|\,\nabla f(x^i)\,\|^2/(2L)$
- x^i "any" and $f(x_*) \leq f(x^{i+1}) \Longrightarrow f(x_*) - f(x) \leq -\|\,\nabla f(x)\,\|^2/(2L)$
- f convex $\Longrightarrow \nabla f(x)(x - x_*) \geq f(x) - f(x_*) \geq \|\,\nabla f(x)\,\|^2/(2L) \;\; \forall\, x$
- This proves $r^i := \|\,x^i - x_*\,\|$ decreases:
  $(r^{i+1})^2 = \|\,x^{i+1} - x_*\,\|^2 = \|\,x^i - x_* - \nabla f(x^i)/L\,\|^2 =$
  $= \|\,x^i - x_*\,\|^2 - 2\nabla f(x^i)(x^i - x_*)/L + \|\,\nabla f(x^i)/L\,\|^2 \leq \|\,x^i - x_*\,\|^2 = (r^i)^2$
- Hence, at the very least $\{\,x^i\,\} \to x_*$ (no problem here)
- Technical step: $\|\,\nabla f(x^i)\,\| \geq (r^i/r^1)\|\,\nabla f(x^i)\,\| \geq$ [Cauchy-Swartz]
  $|\,\nabla f(x^i)(x^i - x_*)\,| \geq (f(x^i) - f(x_*))/r^1$ [convexity] $= \Delta^i/r^1$
- Conclusion: $\Delta^{i+1} \leq \Delta^i - \|\,\nabla f(x^i)\,\|^2/(2L) \leq \Delta^i - (\Delta^i)^2/(2(r^1)^2 L) =$
  $= \Delta^i(\,1 - \Delta^i/(2(r^1)^2 L)\,)$

  not linear convergence as R is not constant $\equiv$ sublinear

Fixed stepsize: convergence rate (cont'd)

- What does this mean, exactly?
- $\Delta^{i+1} \leq \Delta^i - (\Delta^i)^2/(2(r^1)^2 L)$: divide by $\Delta^{i+1}\Delta^i \Longrightarrow$

$$\frac{1}{\Delta^i} \leq \frac{1}{\Delta^{i+1}} - \frac{\Delta^i}{\Delta^{i+1} 2(r^1)^2 L} \quad \Longrightarrow \quad \frac{1}{\Delta^{i+1}} \geq \frac{1}{\Delta^i} + \frac{1}{2(r^1)^2 L} \quad \textbf{(why?)}$$

$\equiv 1/\Delta$ grows by a constant at each $i \Longrightarrow 1/\Delta^{i+1} \geq 1/\Delta^1 + i/(2(r^1)^2 L)$
$\Longrightarrow \Delta^{i+1} \leq 2\Delta^1 (r^1)^2 L/(\,2(r^1)^2 L + i\Delta^1\,)$

- Error decreases as $O(\,1\,/\,i\,) \Longrightarrow O(\,1\,/\,\varepsilon\,)$ iterations (**check details**)
- Exponentially worse than $O(\,1\,/\,\log(\,\varepsilon\,)\,)$
- However, this is unfair: we used Q nonsingular $\equiv \lambda_n > 0$
- Does it make a difference? You bet

Fixed stepsize: convergence rate with strong convexity

- Basically strong convexity
- Eigenvalues bounded both above and below: $u\,I \preceq \nabla f^2(x) \preceq L\,I$, $u > 0$
- Taylor $\implies f(\,x\,) \geq f(\,x^i\,) + \nabla f(x^i)(\,x - x^i\,) + u\|\,x - x^i\,\|^2/2$ (**why?**)
- Minimize on x both sides independently

  $\implies f(\,x_*\,) \geq f(\,x^i\,) - \|\,\nabla f(x^i)\,\|^2/(2u)$ (**check**)

  $\implies \|\,\nabla f(x^i)\,\|^2 \geq 2u(\,f(\,x^i\,) - f(\,x_*\,)\,)$
- Put in $f(x^{i+1}) - f(x_*) \leq f(x^i) - f(x_*) - \|\,\nabla f(x^i)\,\|^2/(2L) \implies$

  $f(x^{i+1}) - f(x_*) \leq (\,f(x^i) - f(x_*)\,)(\,1 - u/L\,)$
- $\approx$ with exact step, funnily same as with coarse estimate, i.e., much worse
- A "small" difference in f makes a big difference in convergence
- Properties of f even more important than the algorithm
- $O(\,1/\varepsilon\,)$ not the best for not strongly convex, can be $O(\,1/\sqrt{\varepsilon}\,)$ better, but still much worse than $O(\,1\,/\,\log(\,\varepsilon\,)\,)$
- Hence better algorithms do count, we'll work towards that
- However, $O(\,1/\sqrt{\varepsilon}\,)$ is tight: can't do better without strong convexity
- Algorithms can only get so far with nasty problems

Wrap up

- Gradient (descent direction) + line search = convergence
- Line search by no means have to be exact

Wrap up

- Gradient (descent direction) + line search = convergence
- Line search by no means have to be exact . . . but not too coarse either
- Many different practical line searches, up to "no search at all"
- Convergence of gradient methods can be from quite bad to horrible

Wrap up

- Gradient (descent direction) + line search = convergence
- Line search by no means have to be exact . . . but not too coarse either
- Many different practical line searches, up to "no search at all"
- Convergence of gradient methods can be from quite bad to horrible
  . . . in practice as well as in theory
- Something better sorely needed