Convex Optimization

Prof. Nati Srebro

Lecture 3:
Steepest Descent
and Gradient Descent

Reading: Boyd and Vandenberghe 9.1-9.4
Alternative reading: Bertsekas Nonlinear Programming 1.2-1.3
Optional reading on other linesearch and stepsize strategies:
Nocedal and Wright 3.1-3.2,parts of 3.3,3.5

Unconstrained Optimization Problems

$$\min_{x \in \mathbb{R}^n} f(x)$$

$$f: \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$$

- We will assume (for now) that we are given some starting point $x^{(0)} \in dom(f)$ (i.e. with $f(x^{(0)}) < \infty$)

- Optimize given $x^{(0)}$ and access to a 1st order oracle
$$x \mapsto f(x), \nabla f(x)$$

Center of Mass Method

- Requires keeping track of polyhedral with increasing number of facets—$O(nk) = O\left(n^2 \log \frac{1}{\epsilon}\right)$ memory
- Requires computing center of mass
  - Equivalent to integrating—harder then optimizing!
  - Can approximate well enough in randomized $poly(n)$ time

- Reasonable number of iterations / grad evals:

$$O\left(n \log {}^{1}/_{\epsilon}\right)$$

  …but horrible runtime

(Generic) Descent Method

Init $x^{(0)} \in dom(f)$

Iterate $x^{(k+1)} \leftarrow x^{(k)} + t^{(k)}\Delta x^{(k)}$

stepsize $t \in \mathbb{R}$

direction $\Delta x \in \mathbb{R}$

- Def: Δx is a descent direction iff $\langle \nabla f(x), \Delta x \rangle < 0$

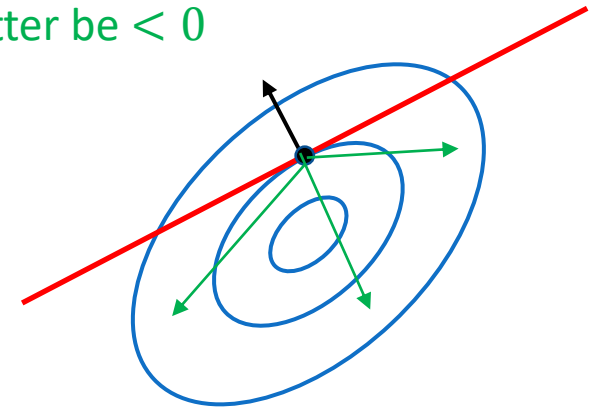
- Recall:

$$f(x^+) = f(x + t\Delta x) \geq f(x) + t\langle \nabla f(x), \Delta x \rangle$$

Better be < 0

- Cutting plane view:

$$x^* \in \{x | \langle \nabla f(x^{(k)}), x - x^{(k)} \rangle < 0\}$$

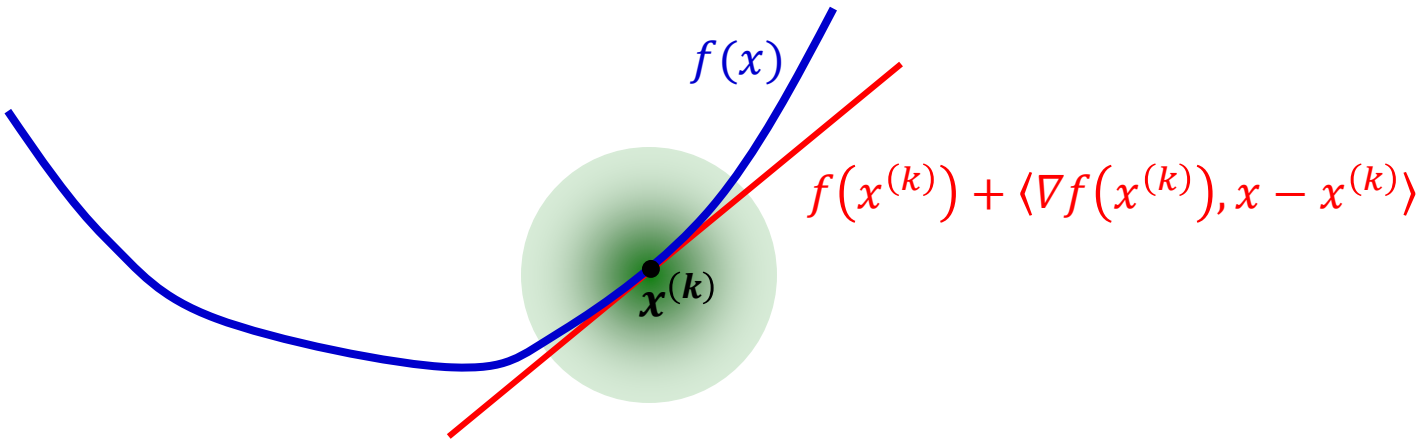


Which Descent Direction?

$$x^{(k+1)} \leftarrow \arg\min f\left(x^{(k)}\right) + \langle \nabla f\left(x^{(k)}\right), x - x^{(k)} \rangle + \frac{\alpha}{2}\left\|x - x^{(k)}\right\|^2$$

1st order approx of $f(x)$

Only valid near $x^{(k)}$



Which Descent Direction?

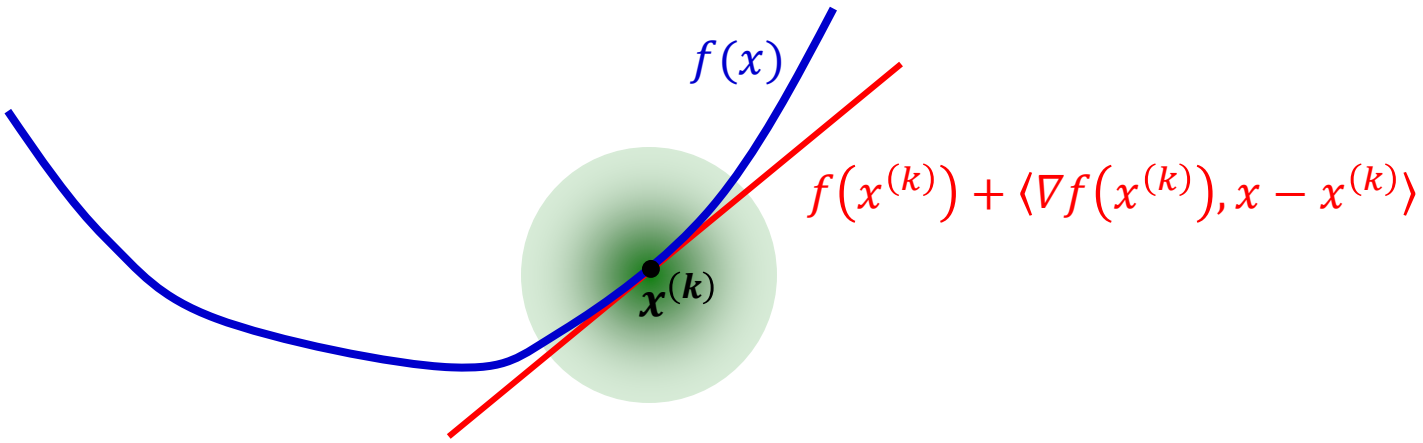
$$x^{(k+1)} \leftarrow \arg\min f\left(x^{(k)}\right) + \langle \nabla f\left(x^{(k)}\right), x - x^{(k)} \rangle + \frac{\alpha}{2}\left\| x - x^{(k)} \right\|^2$$

$$= \arg\min_{\|\Delta x\|^2 = 1} f\left(x^{(k)}\right) + t\langle \nabla f\left(x^{(k)}\right), \Delta x \rangle + \frac{\alpha}{2} t^2 \|\Delta x\|^2$$

$x = x^{(k)} + t\Delta x$

$$t = \frac{-\langle \nabla f(x), \Delta x \rangle}{\alpha}$$

$$\Delta x = -\arg\max_{\|v\|=1} \langle \nabla f\left(x^{(k)}\right), v \rangle$$



Direction of Steepest Descent

$$\Delta x = -\arg\max_{\|v\|=1} \langle \nabla f\left(x^{(k)}\right), v \rangle$$

- Choice of norm $\|\cdot\|$ is crucial!
- Using the Euclidean norm: $\|x\|_2 = \sqrt{\sum_i x[i]^2}$:

  $\Delta x \propto -\nabla f\left(x^{(k)}\right)$ (as vectors)

- Depends on choice of basis!
- Choice of norm (eg basis for Euclidan norm) relates the primal x space to the dual space of gradients ∇f

Gradient Descent
(Steepest Descent w.r.t Euclidean Norm)

$$\Delta x = -\nabla f\left(x^{(k)}\right)$$

Init	$x^{(0)} \in dom(f)$
Iterate	$x^{(k+1)} \leftarrow x^{(k)} - t^{(k)} \nabla f\left(x^{(k)}\right)$

Reminder: We are violating here the distinction between the primal space and the dual gradient space—we are implicitly linking them by matching representations w.r.t. a chosen basis

Note: Δx is not normalized (i.e. we don't require $\|\Delta x\|_2 = 1$). This just changes the meaning of t.

How do we choose the stepsize $t^{(k)}$?

Setting the Stepsize
Option 1: Exact Linesearch

$$t^{(k)} \leftarrow \arg\min_{t \in \mathbb{R}} f\left(x^{(k)} + t\Delta x^{(k)}\right)$$

- This is a convex one-dimensional problem
  ➔ can use bisection!

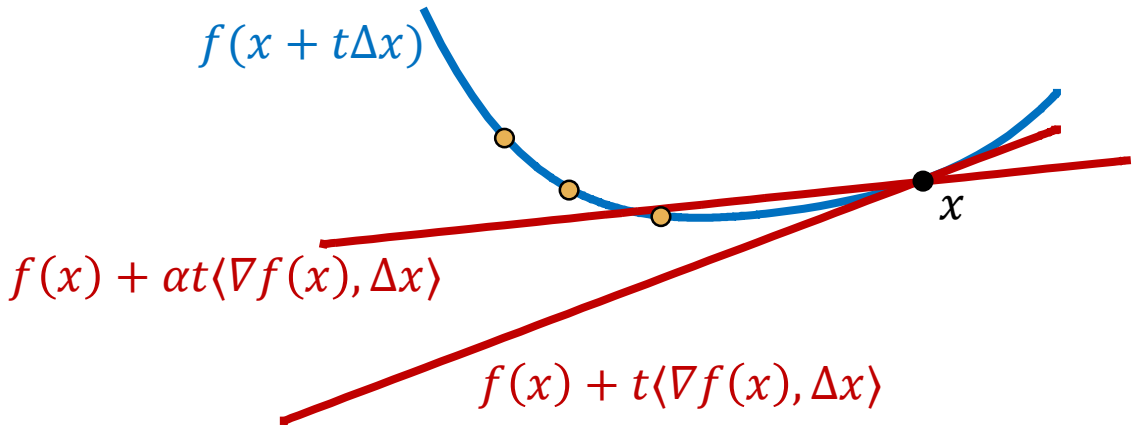
- But to what accuracy?

- Outer loop (updating $x^{(k)}$) and inner loop (optimizing $t^{(k)}$): when do we stop inner loop and iterate outer loop?

Option 2: Backtracking Linesearch (Armijo's Rule)

- Parameters: $0 < \alpha < \frac{1}{2}$ and $0 < \beta < 1$
- Input: initial point x
  subgradient $\nabla f(x)$ at initial point
  direction/vector Δx
  evaluation oracle for $f(\cdot)$
- Output: stepsize t

Init $t \leftarrow 1$
Until $f(x + t\Delta x) < f(x) + \alpha \cdot t\langle \nabla f(x), \Delta x \rangle$
  Set $t \leftarrow \beta \cdot t$



Gradient Descent

Init $x^{(0)} \in dom(f)$

Iterate $\Delta x^{(k)} = -\nabla f\left(x^{(k)}\right)$

Set $t^{(k)}$ by backtracking linesearch

$x^{(k+1)} \leftarrow x^{(k)} + t^{(k)} \Delta x^{(k)}$

Stopping condition?

Runtime analysis? How many iterations?

Smoothness and Strong Convexity

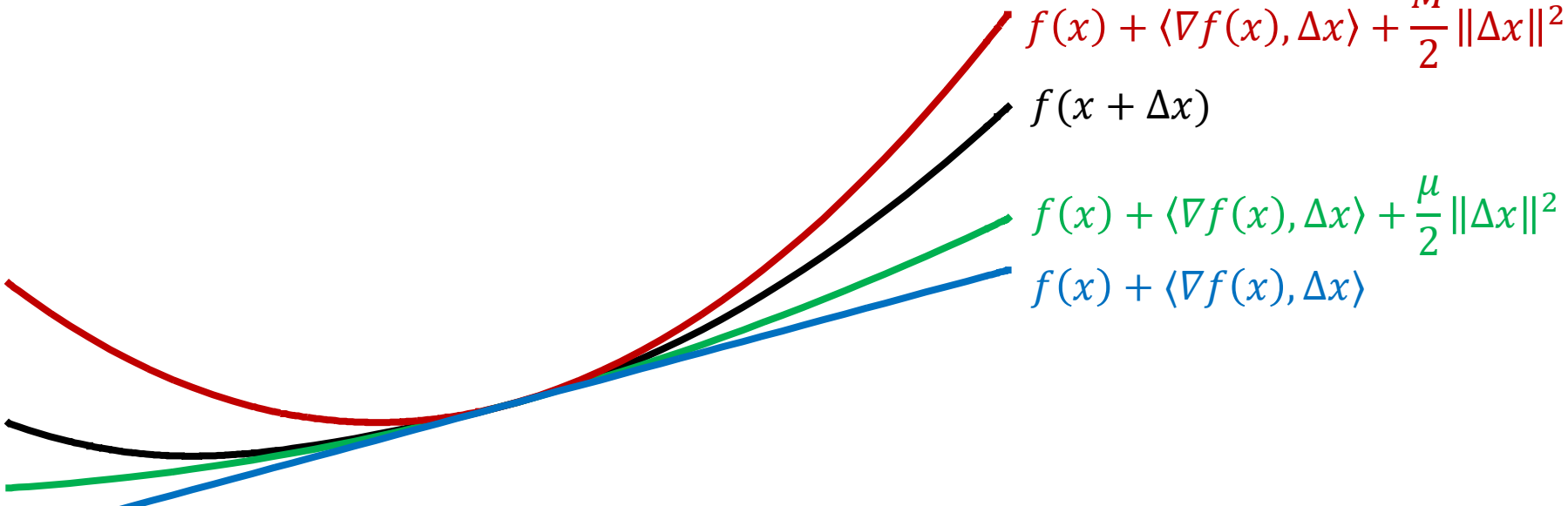
Def: f is μ-strongly convex

Def: f is M-smooth

$$f(x) + \langle \nabla f(x), \Delta x \rangle + \frac{\boldsymbol{\mu}}{2} \|\Delta x\|_2^2 \leq f(x + \Delta x) \leq f(x) + \langle \nabla f(x), \Delta x \rangle + \frac{\boldsymbol{M}}{2} \|\Delta x\|_2^2$$

Can be viewed as a condition on the directional 2nd derivatives

$$\boldsymbol{\mu} \leq f_v''(x) = \frac{\partial^2}{\partial t^2} f(x + tv) = v^\top \nabla^2 f(x) v \leq \boldsymbol{M} \quad \text{(for } \|v\|_2 = 1\text{)}$$



Smoothness and Strong Convexity

Def: f is μ-strongly convex

Def: f is M-smooth

$$f(x) + \langle \nabla f(x), \Delta x \rangle + \frac{\boldsymbol{\mu}}{2} \|\Delta x\|_2^2 \leq f(x + \Delta x) \leq f(x) + \langle \nabla f(x), \Delta x \rangle + \frac{\boldsymbol{M}}{2} \|\Delta x\|_2^2$$

Can be viewed as a condition on the directional 2nd derivatives

$$\boldsymbol{\mu} \leq f_v''(x) = \frac{\partial^2}{\partial t^2} f(x + tv) = v^\top \nabla^2 f(x) v \leq \boldsymbol{M} \quad \text{(for } \|v\|_2 = 1\text{)}$$

And as condition on eigenvalues of Hessian:

$$\boldsymbol{\mu} \leq \lambda_{\min}(\nabla^2 f(x)), \lambda_{\max}(\nabla^2 f(x)) \leq \boldsymbol{M}$$

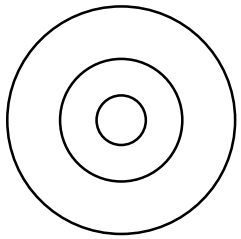
$$\boldsymbol{\mu} I \preccurlyeq \nabla^2 f(x) \preccurlyeq \boldsymbol{M} I$$

$$\kappa = \frac{\boldsymbol{M}}{\boldsymbol{\mu}} = \frac{\max_x \boldsymbol{\lambda_{\max}}\left(\boldsymbol{\nabla^2 f(x)}\right)}{\min_x \boldsymbol{\lambda_{\min}}\left(\boldsymbol{\nabla^2 f(x)}\right)}$$

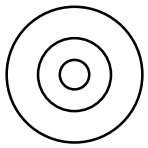
Examples

$$f(x) = \frac{1}{2} x^\top H x + b^\top x$$

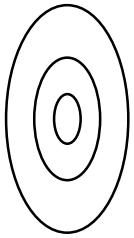
$H = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
$\mu = 1, M = 1$



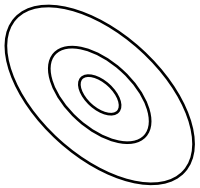
$H = \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix}$
$\mu = 5, M = 5$



$H = \begin{bmatrix} 5 & 0 \\ 0 & 1 \end{bmatrix}$
$\mu = 1, M = 5$



$H = \begin{bmatrix} 3 & 2 \\ 2 & 3 \end{bmatrix}$
$\mu = 1, M = 5$



Strong Convexity and Sub-optimality

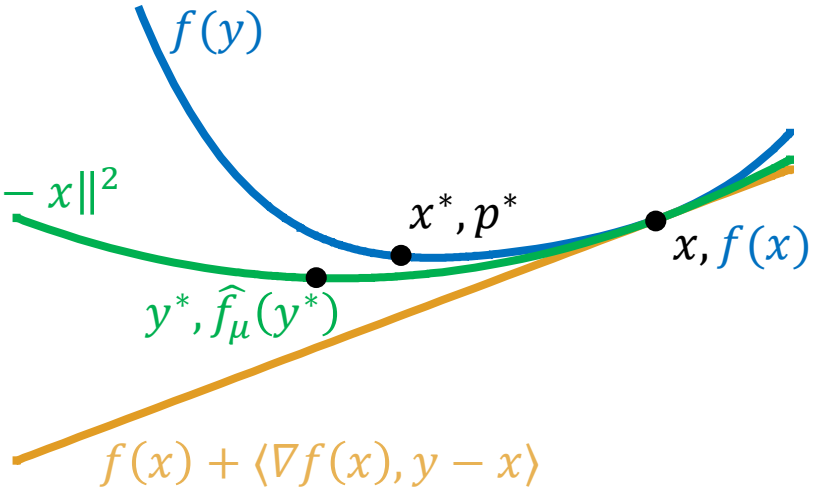
- Assume f is μ-strongly-convex (i.e. $\forall_x \mu I \preccurlyeq \nabla^2 f(x)$)

$$y^* = x - \frac{1}{\mu}\nabla f(x)$$

$$p^* = f(x^*)$$

$$\geq \min_y \underbrace{f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|_2^2}_{\widehat{f}_\mu(y)} = \underbrace{f(x) - \frac{1}{2\mu}\|\nabla f(x)\|_2^2}_{\widehat{f}_\mu(y^*)}$$

$$\widehat{f}_\mu(y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2$$



Strong Convexity and Sub-optimality

- Assume f is μ-strongly-convex (i.e. $\forall_x \mu I \preccurlyeq \nabla^2 f(x)$)

$$p^* = f(x^*) \geq f(x) + \langle \nabla f(x), x^* - x \rangle + \frac{\mu}{2}\|x^* - x\|_2^2$$

$$\geq \min_y f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|_2^2 = f(x) - \frac{1}{2\mu}\|\nabla f(x)\|_2^2$$

$$\rightarrow \quad f(x) - p^* \leq \frac{1}{2\mu}\|\nabla f(x)\|_2^2$$

- Stopping condition to ensure ϵ-suboptimality:

$$\|\nabla f(x)\| \leq \sqrt{2\mu\epsilon}$$

Runtime Analysis in terms of $\kappa = \frac{M}{\mu}$

- Assuming strong convexity AND smoothness $\forall_x \ \mu I \leq \nabla^2 f(x) \leq MI$
- Using exact linesearch:

$$\begin{aligned} f(x^+) &= \min_t f\left(x - t\nabla f(x)\right) \\ &\leq \min_t f(x) + \langle \nabla f(x), -t\nabla f(x)\rangle + \frac{M}{2}\|t\nabla f(x)\|_2^2 \\ &= \min_t f(x) - t\|\nabla f(x)\|^2 + t^2\frac{M}{2}\|\nabla f(x)\|_2^2 = f(x) - \frac{1}{2M}\|\nabla f(x)\|_2^2 \\ &\leq f(x) - \frac{2\mu}{2M}(f(x) - p^*) \end{aligned}$$

➔ $f\left(x^{(k+1)}\right) - p^* \leq f\left(x^{(k)}\right) - p^* - \frac{1}{\kappa}\left(f\left(x^{(k)}\right) - p^*\right) = \left(1 - \frac{1}{\kappa}\right)\left(f\left(x^{(k)}\right) - p^*\right)$

➔ $$f\left(x^{(k)}\right) - p^* \leq \left(1 - \frac{1}{\kappa}\right)^k \left(f\left(x^{(0)}\right) - p^*\right)$$

➔ Number of iterations required for $f\left(x^{(k)}\right) \leq p^* + \epsilon$:

$$k \leq \frac{1}{\log\left(\frac{\kappa}{\kappa-1}\right)} \log\left(\frac{f\left(x^{(0)}\right) - p^*}{\epsilon}\right)$$

Gradient Descent for Smooth and Strongly Convex Objectives

- Theorem: If f is μ-strongly convex and M-smooth on

  $S^0 = \{x | f(x) \leq f(x^{(0)})\}$ (i.e. $\forall_{f(x) \leq f(x^{(0)})} \mu I \preccurlyeq \nabla^2 f(x) \preccurlyeq MI$),

  then Gradient Descent with exact lineasearch with at most

$$k \leq \frac{1}{-\log\left(1 - \frac{1}{\kappa}\right)} \log\left(\frac{f(x^{(0)}) - p^*}{\epsilon}\right) \approx \kappa \cdot \log\left(\frac{f(x^{(0)}) - p^*}{\epsilon}\right)$$

  iterations ensures $f(x^{(k)}) \leq p^* + \epsilon$

Using Backtracking Lineasearch

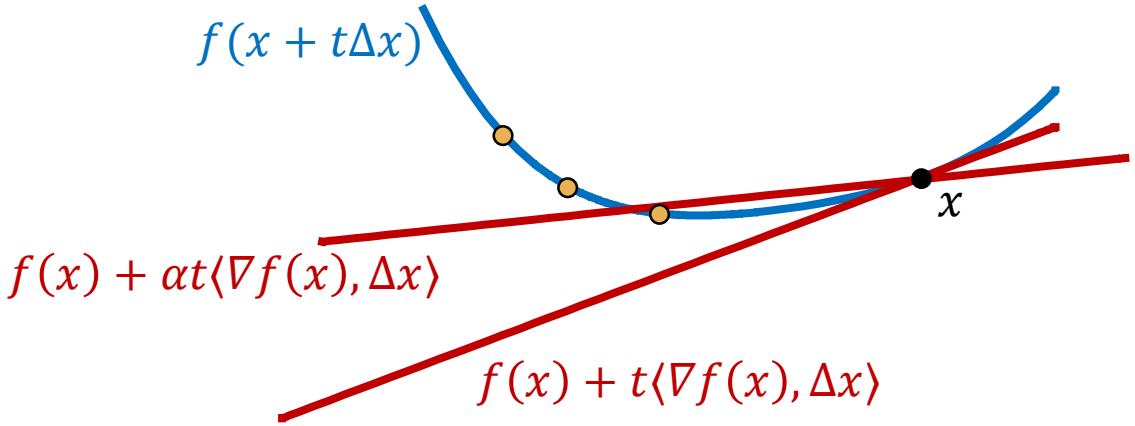
Init $t \leftarrow 1$
Until $f(x + t\Delta x) \leq f(x) + \alpha \cdot t\langle \nabla f(x), \Delta x \rangle = f(x) - t\alpha \|\nabla f(x)\|_2^2$
Set $t \leftarrow \beta \cdot t$

- Claim: If f is M-smooth, then any $t < \frac{1}{M}$ satisfies Armijo for any $\alpha < 1/2$

  Proof: $f\left(x - t\nabla f(x)\right) \leq f(x) + \langle \nabla f(x), -t\nabla f(x) \rangle + \frac{M}{2}\|t\nabla f(x)\|_2^2$

  $= f(x) - \left(1 - \frac{M}{2}t\right) t\|\nabla f(x)\|_2^2 \leq f(x) - t\frac{1}{2}\|\nabla f(x)\|_2^2 \leq f(x) - t\alpha\|\nabla f(x)\|_2^2$

- Conclusion: we either use $t = 1$ or $t > \beta/M$



Using Backtracking Lineasearch

Init $t \leftarrow 1$
Until $f(x + t\Delta x) \leq f(x) + \alpha \cdot t\langle \nabla f(x), \Delta x \rangle = f(x) - t\alpha \|\nabla f(x)\|_2^2$
Set $t \leftarrow \beta \cdot t$

- Claim: If f is M-smooth, then any $t < \frac{1}{M}$ satisfies Armijo for any $\alpha < 1/2$

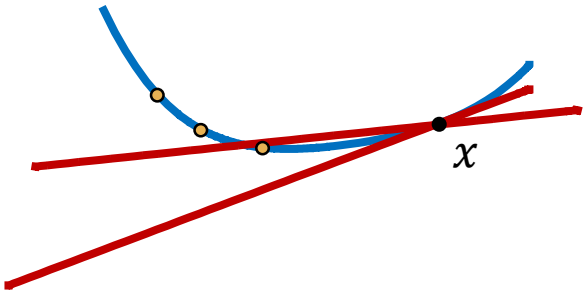
  Proof: $f\left(x - t\nabla f(x)\right) \leq f(x) + \langle f(x), -t\nabla f(x)\rangle + \frac{M}{2}\|t\nabla f(x)\|_2^2$

  $= f(x) - \left(1 - \frac{M}{2}t\right) t\|\nabla f(x)\|_2^2 \leq f(x) - t\frac{1}{2}\|\nabla f(x)\|_2^2 \leq f(x) - t\alpha\|\nabla f(x)\|_2^2$

- Conclusion: we either use $t = 1$ or $t > \beta/M$

➔ $f\left(x - t\nabla f(x)\right) \leq f(x) - \min\left(1, \frac{\beta}{M}\right)\alpha\|\nabla f(x)\|_2^2$

$$\#\text{iter } k \leq \frac{1}{-\log\left(1 - 2\alpha \min\left(\mu, \frac{\beta}{\kappa}\right)\right)} \log\left(\frac{f\left(x^{(0)}\right) - p^*}{\epsilon}\right)$$



- How many inner iterations?

Since $t < 1/M$ always OK, at most $\left\lceil \frac{\log\frac{1}{M}}{\log \beta} \right\rceil$ per outer iteration

Gradient Descent with Backtracking Linesearch

Init	$x^{(0)} \in dom(f)$
Iterate	$\Delta x^{(k)} = -\nabla f\left(x^{(k)}\right)$
	Stop if $\left\|\nabla f\left(x^{(k)}\right)\right\|_2^2 \leq 2\mu\epsilon$
	Set $t^{(k)}$ by backtracking linesearch with params α, β
	$x^{(k+1)} \leftarrow x^{(k)} + t^{(k)}\Delta x^{(k)}$

If f is $\boldsymbol{\mu}$-strongly-convex and $\boldsymbol{M}$-smooth on $\left\{f(x) \leq f\left(x^{(0)}\right)\right\}$, #func evals :

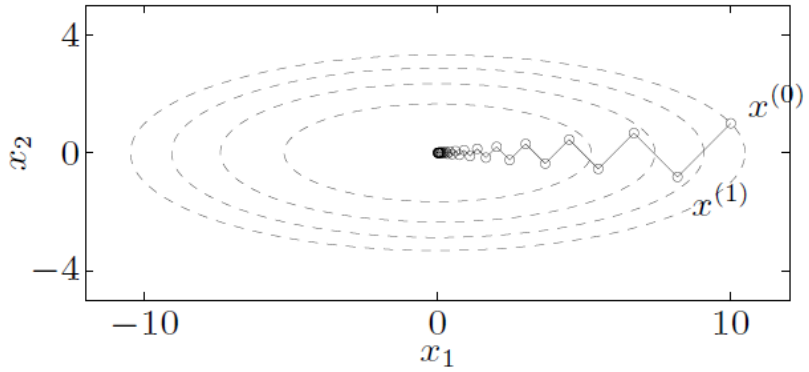
$$k = O\left(\left\lceil\frac{\log M}{\log\frac{1}{\beta}}\right\rceil\frac{1}{\alpha}\max\left(\frac{1}{M},\frac{1}{\beta}\right)\boldsymbol{\kappa}\log\left(\frac{f\left(x^{(0)}\right)-p^*}{\epsilon}\right)\right)$$

$\boldsymbol{\kappa} = {}^{\boldsymbol{M}}/_{\boldsymbol{\mu}}$

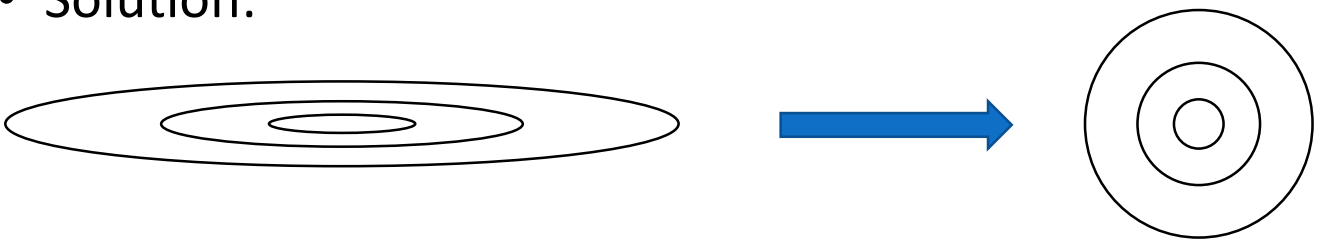
and additional runtime: $O\left(\boldsymbol{n}\left\lceil\frac{\log M}{\log\frac{1}{\beta}}\right\rceil\frac{1}{\alpha}\max\left(\frac{1}{M},\frac{1}{\beta}\right)\boldsymbol{\kappa}\log\left(\frac{f\left(x^{(0)}\right)-p^*}{\epsilon}\right)\right)$

Is dependence on κ real?

- Yes! Even for quadratic, even with exact linesearch.



- Solution:



$$f(x) = \frac{1}{2} x^\top H x$$

$$\tilde{f}(\tilde{x}) = f\left(H^{-1/2}\tilde{x}\right) = \frac{1}{2}\tilde{x}^\top I \tilde{x}$$

$$\tilde{x} = H^{1/2}x$$

- Another way to view this: change of basis

Change of variables / basis

$$\tilde{f}(\tilde{x}) = f\left(H^{-1/2}\tilde{x}\right) \quad \tilde{x} \leftarrow H^{1/2}x$$

- GD direction:

$$\Delta\tilde{x} = -\nabla\tilde{f}(\tilde{x}) = -H^{-1/2}\nabla f\left(H^{-1/2}\tilde{x}\right)$$

- What's the update in the original basis?

$$\tilde{x}^{+} \leftarrow \tilde{x} + t\Delta\tilde{x} \qquad \equiv \qquad x \leftarrow x + t\Delta x$$

$$\Delta x = H^{-1/2}\Delta\tilde{x} = -H^{-1/2}H^{-1/2}\nabla f\left(H^{-1/2}\tilde{x}\right) = -H^{-1}\nabla f(x)$$

- Alternative view: steepest descent w.r.t. $\|x\|_H$

Pre-Conditioned Gradient Descent

Init	$x^{(0)} \in dom(f)$
	$\boldsymbol{H} = \nabla^2 f\left(x^{(0)}\right)$
Iterate	$\Delta x^{(k)} = -\boldsymbol{H}^{-1} \nabla f\left(x^{(k)}\right)$
	Set $t^{(k)}$ by backtracking linesearch with params α, β
	$x^{(k+1)} \leftarrow x^{(k)} + t^{(k)} \Delta x^{(k)}$