# Unconstrained minimization

A supplementary note to Chapter 9 of *Convex Optimization* by S. Boyd and L. Vandenberghe

Optimization Lab.

IE department
Seoul National University

2nd December 2009

## Unconstrained minimization

Consider

$$\min \quad f(x) \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex and twice continuously differentiable (on an open domain).

**Assumption**

*There exists an optimal point $x^*$ such that $p^* = f(x^*) = \inf_x f(x)$.*

Since $f$ is differentiable and convex, a point $x^*$ is optimal if and only if

$$\nabla f(x^*) = 0. \tag{2}$$

Thus, solving the unconstrained minimization problem (1) is the same as finding a solution of (2), which is a set of $n$ equations in the $n$ variables $x_1, \ldots, x_n$.

## Unconstrained minimization(*cont'd*)

- We can find a solution of (1)
  - by either analytically solving equation (2), or
  - using an iterative algorithm.
- An iterative algorithm computes a sequence of points $x^{(0)}, x^{(1)}, \cdots \in \mathrm{dom} f$ with

$$f(x^{(k)}) \to p^* \quad \text{as } k \to \infty.$$

- The iterative algorithms normally require a suitable starting point $x^{(0)}$ such that
  - $x^{(0)} \in \mathrm{dom} f$, and
  - $S = \{x \in \mathrm{dom} f | f(x) \leq f(x^{(0)})\}$ is closed.

## Examples: Quadratic min. and least-squares

### Example (General convex quadratic minimization problem)

$$\min \quad \tfrac{1}{2}x^T P x + q^T x + r, \tag{3}$$

*where* $P \in \mathbb{S}^n_+, q \in \mathbb{R}^n$, *and* $r \in \mathbb{R}$.

- *When* $P \succ 0$, $x^* = -P^{-1}q$.
- *Otherwise, any* $x^*$ *satisfying* $Px^* = -q$ *is an optimal solution.*
- *If* $Px = -q$ *does not have a solution, (3) is unbounded below.*

### Example (Least-square problem)

$$\min \quad \|Ax - b\|_2^2 = x^T(A^T A)x - 2(A^T b)^T x + b^T b. \tag{4}$$

*The optimality conditions* $A^T A x^* = A^T b$ *are called the normal equations of the least-square problem.*

## Examples: Unconstrained geometric programming

**Example (Unconstrained geometric program in convex form)**

$$\min \quad f(x) = \log(\textstyle\sum_{i=1}^{m} exp(a_i^T x + b_i)). \tag{5}$$

*The optimality condition is*

$$\nabla f(x^*) = \frac{1}{\sum_{i=1}^{m} exp(a_i^T x + b_i)} \sum_{i=1}^{m} exp(a_i^T x + b_i) a_i = 0.$$

- There may be no analytical solution in general. Then we must resort to an iterative algorithm.

## Examples: Analytic center of linear inequality and linear matrix inequality

### Example (Logarithmic barrier $f(x)$ for $a_i^T x \leq b_i$)

$$f(x) = -\sum_{i=1}^{m} \log(b_i - a_i^T x), \;\; \mathit{dom} f = \{x | a_i^T x < b_i, \; i = 1, \ldots, m\}.$$

*The solution of the problem* $\min \; f(x)$ *is called the analytic center of the inequalities. Domain* $\mathit{dom} f = \{x : a_i^T x < b_i, \; i = 1, \ldots, m\}$*. If initial point* $x^{(0)}$ *is in the domain,* $S = \{x : f(x) \leq f(x^{(0)})\}$ *is closed. For* $S$ *is contained in the union of the closed sets* $\{x : b_i - a_i^T x \geq \delta\}$ $(\subseteq \mathit{dom} f)$ *for some* $\delta > 0$*.*

### Example (Logarithmic barrier $f(x)$ for LMI $F(x) \succeq 0$)

$$f(x) = \log\det F(x)^{-1}, \;\; \mathit{dom} f = \{x | F(x) = x_0 F_0 + x_1 F_1 + \cdots + x_n F_n \succ 0\}$$

*The solution of the problem* $\min \; f(x)$ *is called the analytic center of the LMI.*

## Strong convexity and implications

In much of this chapter, we rely on the following stronger assumption.

**Definition**

*A function $f$ is strongly convex on $S$ if there exists an $m > 0$ such that*

$$\nabla^2 f(x) \succeq mI$$

*for all $x \in S$.*

Suppose $f$ is strongly convex on $S$. Then, since

$$f(y) = f(x) + \nabla f(x)^T(y-x) + \frac{1}{2}(y-x)^T \nabla^2 f(z)(y-x) \text{ for some } z \in [x,y],$$

we have

$$f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{m}{2}\|y-x\|_2^2, \ \forall\, x, y \in S. \quad (6)$$

When $m = 0$, it reduces to the first order condition for convexity.

## Strong convexity and implications: Upper bound on $f(x) - p^*$

- Right hand side of (6), convex quadratic function of $y$, is minimized at $\tilde{y} = x - \frac{1}{m}\nabla f(x)$.

$$\begin{aligned} f(y) &\geq f(x) + \nabla f(x)^T(y-x) + \frac{m}{2}\|y-x\|_2^2 \\ &\geq f(x) + \nabla f(x)^T(\tilde{y}-x) + \frac{m}{2}\|\tilde{y}-x\|_2^2 \\ &= f(x) - \frac{1}{2m}\|\nabla f(x)\|_2^2. \end{aligned}$$

Taking $y = x^*$, we get:

**Theorem**

*Suboptimality of the point $x$, $f(x) - p^* \leq \frac{1}{2m}\|\nabla f(x)\|_2^2$.*

Hence if gradient is small enough, then the point is nearly optimal:

$$\|\nabla f(x)\|_2 \leq (2m\epsilon)^{1/2} \Rightarrow f(x) - p^* \leq \epsilon.$$

## Strong convexity and implications: Upper bound on $\|x - x^*\|_2$

- From (6) with $y = x^*$, for any $x$

$$\begin{aligned} p^* = f(x^*) \quad &\geq f(x) + \nabla f(x)^T (x^* - x) + \frac{m}{2}\|x^* - x\|_2^2 \\ &\geq f(x) - \|\nabla f(x)\|_2 \|x^* - x\|_2 + \frac{m}{2}\|x^* - x\|_2^2. \end{aligned}$$

- Since $f(x) \geq p^*$, $\|\nabla f(x)\|_2 \|x^* - x\|_2 \geq \frac{m}{2}\|x^* - x\|_2^2$.

**Theorem**

$\|x^* - x\|_2 \leq \frac{2}{m}\|\nabla f(x)\|_2.$

This implies optimal point $x^*$ is unique.

## Strong convexity and implications: Lower bound on $f(x) - p^*$

- (6) implies the sublevel sets contained in $S$ are bounded, so in particular, $S$ is bounded. (If we let $x = x^*$, $f(y) \geq p^* + \frac{m}{2}\|y - x^*\|^2$. Thus if $f(y) \leq \alpha \leq f(x^{(0)})$, $\|y - x^*\|^2 \leq$ some constant.)
- Then, the maximum eigenvalue of $\nabla^2 f(x)$, which is a continuous function of $x$ on the compact set $S$, achieves its maximum $M$ on $S$.
- This means that $\nabla^2 f(x) \preceq MI$ for all $x \in S$.

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{M}{2}\|y - x\|_2^2, \ \forall\, x, y \in S. \qquad (7)$$

**Theorem**

$\frac{1}{2M}\|\nabla f(x)\|_2^2 \leq f(x) - p^*$.

**Proof** Similar to the proof of lower bound. □

## Strong convexity and implications: Condition number of $\nabla^2 f(x)$

**Definition**

*The condition number of $\nabla^2 f(x)$ is the ratio of its largest eigenvalue to its smallest eigenvalue.*

From the strong convexity, $mI \preceq \nabla^2 f(x) \preceq MI, \forall\, x \in S$, the condition number of $\nabla^2 f(x)$ is bounded by $\frac{M}{m}$.

# Strong convexity and implications: Condition number of convex sets

## Definition

- *The width of a convex set* $C$*, in the direction* $q$*,* $\|q\|_2 = 1$*, as*

$$W(C, q) = \sup_{z \in C} q^T z - \inf_{z \in C} q^T z.$$

- *The minimum width and the maximum width of* $C$ *are given by*

$$W_{\min} := \inf_{\|q\|_2 = 1} W(C, q), \quad W_{\max} := \sup_{\|q\|_2 = 1} W(C, q)$$

- *The condition number of* $C$ *is* $cond(C) = \frac{W_{\max}^2}{W_{\min}^2}$.

## Strong convexity and implications: Condition number of $\alpha$-sublevel sets

Suppose $mI \preceq \nabla^2 f(x) \preceq MI$ and $C_\alpha := \{x | f(x) \le \alpha\}$ where $p^* < \alpha \le f(x^{(0)})$.

- From (6) and (7) with $x = x^*$, we get

$$p^* + (m/2)\|y - x^*\|^2 \le f(y) \le p^* + (M/2)\|y - x^*\|_2^2.$$

- This implies $B_{\text{inner}} \subseteq C_\alpha \subseteq B_{\text{output}}$ where

$$B_{\text{inner}} := \{y | \|y - x^*\|_2 \le (2(\alpha - p^*)/M)^{1/2}$$
$$B_{\text{outer}} := \{y | \|y - x^*\|_2 \le (2(\alpha - p^*)/m)^{1/2}$$

  For $y \in B_{\text{inner}} \Rightarrow f(y) \le p^* + \frac{M}{2}\|y - x^*\|_2^2 \le \alpha$; and $f(y) \le \alpha \Rightarrow p^* + (m/2)\|y - x^*\|^2 \le \alpha \Rightarrow y \in B_{\text{outer}}$.

- Thus, min width of $C_\alpha \ge (2(\alpha - p^*/M)^{1/2}$ and max width of $C_\alpha \le (2(\alpha - p^*)/m)^{1/2}$ and hence $\text{cond}(C_\alpha) \le \frac{M}{m}$.

## Iterative algorithms and descent method

In iterative algorithms,

- we generate a minimizing sequence $x^{(k)}$, $k = 1, 2, \ldots$

$$x^{(k+1)} = x^{(k)} + t^{(k)}\Delta x^{(k)}, \ \ t^{(k)} > 0,$$

- where, $\Delta x^{(k)}$ is called *search direction* at iteration $k$, and
- $t^{(k)}$ *step size* or *step length* at iteration $k$.

In descent method,

- sequence $x^{(k)}$, $k = 1, 2, \ldots$ satisfies

$$f(x^{(k+1)}) < f(x^{(k)}),$$

- which implies for all $k$, $x^{(k)} \in S$, where $S$ is the initial sublevel set.

# Iterative algorithms and descent method

**Proposition**

*If $\Delta x^{(k)}$ is a search direction for a descent method,*

$$\nabla f(x^{(k)})^T \Delta x^{(k)} < 0.$$

**Proof** Since $f$ is a convex function,

$$f(x^{(k+1)}) \geq f(x^{(k)}) + t^{(k)} \nabla f(x^{(k)})^T \Delta x^{(k)}.$$

By assumption $f(x^{(k+1)}) - f(x^{(k)}) < 0$, and hence

$$t^{(k)} \nabla f(x^{(k)})^T \Delta x^{(k)} < 0.$$

Since $t^{(k)} > 0$,

$$\nabla f(x^{(k)})^T \Delta x^{(k)} < 0. \ \square$$

## General descent method

**Algorithm**

**given** *a starting point* $x \in \mathbf{dom} f$.

**repeat**

1. *Determine a descent direction* $\Delta x$.
2. *Line search. Choose a step size* $t > 0$.
3. *Update.* $x := x + t\Delta x$.

**until** *stopping criterion is satisfied.*

## Exact line search

In *exact line search*,

- $t$ is chosen to minimize $f$ along the ray $\{x + t\Delta x | t \geq 0\}$:

$$t = \text{argmin}_{s \geq 0} f(x + s\Delta x). \tag{8}$$

- An exact line search is used when the computation (8) is marginal to computation of the search direction itself.

**Remark**

*Most line searches used in practice are inexact: the step length is chosen to approximately minimize $f$ along the ray $\{x + t\Delta x | t \geq 0\}$, or to reduce $f$ enough.*

## Backtracking line search

**Algorithm**

given *descent direction* $\Delta x$ *for* $f$ *at* $x \in \mathbf{dom} f$, $\alpha \in (0, 0.5), \beta \in (0, 1)$.

$t := 1.$

while $f(x + t\Delta x) > f(x) + \alpha t \nabla f(x)^T \Delta x$,

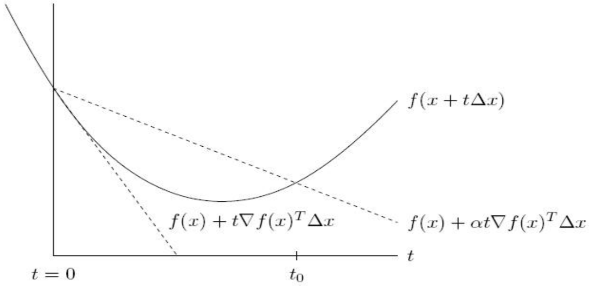
$t := \beta t.$

Since $\Delta x$ is a descent direction, we have $\nabla f(x)^T \Delta x < 0$. Thus, for small enough $t$ we have

$$f(x + t\Delta x) \approx f(x) + t\nabla f(x)^T \Delta x < f(x) + \alpha t \nabla f(x)^T \Delta x,$$

which implies the backtracking line search eventually terminates.

# Backtracking line search(*cont'd*)




- The backtracking exit inequality $f(x + t\Delta x) \leq f(x) + \alpha t \nabla f(x)^T \Delta x$ holds for $t \geq 0$ in an interval $(0, t_0]$.
- It follows that the backtracking line search stops with a step length $t$ that satisfies

$$t = 1, \quad \text{or } t \in (\beta t_0, t_0] \Rightarrow t \geq \min\{1, \beta t_0\}.$$

A natural choice for search direction is the negative gradient $\Delta x = -\nabla f(x)$, most-decreasing direction of $f$ at $x$.

**Algorithm (Gradient descent method)**

**given** *a starting point* $x \in \mathbf{dom} f$.

**repeat**

1. $\Delta x = -\nabla f(x)$.
2. *Line search. Choose a step size* $t > 0$ *via exact or backtracking.*
3. *Update.* $x := x + t\Delta x$.

**until** *stopping criterion is satisfied. (usually,* $\|\nabla f(x)\|_2 \leq \eta (> 0)$.*)*

## Convergence analysis

- Assume $f$ is strongly convex on $S$ and hence $\exists$ $m$ and $M$ s.t. $mI \preceq \nabla^2 f(x) \preceq MI$ $\forall x \in S$.
- Define $\tilde{f} : \mathbb{R} \to \mathbb{R}$ by $\tilde{f}(t) = f(x - t\nabla f(x))$.
- From $f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{M}{2}\|y-x\|_2^2$ with $y = x - t\nabla f(x)$,

$$\tilde{f}(t) \leq f(x) - t\|\nabla f(x)\|_2^2 + \frac{Mt^2}{2}\|\nabla f(x)\|_2^2.$$

## Analysis for exact line search

Suppose the exact line search is used, and let $t^*$ be the minimizer of $\tilde{f}$.

- $f(x) - t\|\nabla f(x)\|_2^2 + \frac{Mt^2}{2}\|\nabla f(x)\|_2^2$ is minimized at $t = \frac{1}{M}$ and has minimum value $f(x) - \frac{1}{2M}\|\nabla f(x)\|_2^2$.
- Thus,

$$f(x - t^*\nabla f(x)) \leq f(x) - \frac{1}{2M}\|\nabla f(x)\|_2^2.$$

- Subtracting $p^*$ from both sides and combining with

$$\|\nabla f(x)\|_2^2 \geq 2m(f(x) - p^*),$$

  we have

$$f(x - t^*\nabla f(x)) - p^* \leq (1 - \frac{m}{M})(f(x) - p^*).$$

- It implies $f(x^{(k)}) - p^* \leq (1 - \frac{m}{M})^k(f(x^{(0)}) - p^*)$, and hence $f(x^{(k)})$ converges to $p^*$ as $k \to \infty$.

## Analysis for exact line search(*cont'd*)

Consider $f(x^{(k)}) - p^* \leq (1 - \frac{m}{M})^k (f(x^{(0)} - p^*)$,

- To obtain $f(x^{(k)}) - p^* \leq \epsilon$,

$$
\begin{aligned}
& (1 - \tfrac{m}{M})^k (f(x^{(0)}) - p^*) \leq \epsilon \\
\Leftrightarrow \quad & (1 - \tfrac{m}{M})^k \leq \tfrac{\epsilon}{f(x^{(0)}) - p^*} \\
\Leftrightarrow \quad & k \leq \frac{\log \frac{\epsilon}{f(x^{(0)}) - p^*}}{\log\left(1 - \frac{m}{M}\right)} = \frac{\log \frac{f(x^{(0)}) - p^*}{\epsilon}}{-\log\left(1 - \frac{m}{M}\right)}
\end{aligned}
$$

- The numerator implies that the number of iterations depends on how good the initial point is, and what the final required accuracy is.
- The denominator implies that the number of iterations depends on the condition number, $M/m$ of $\nabla^2 f(x)$. (Note $-\log(1 - m/M) \approx m/M$.)

## Analysis for backtracking line search

Suppose the backtracking line search is used.

**Lemma**

*If $0 \leq t \leq 1/M$ and $\alpha < 1/2$, then $\tilde{f}(t) \leq f(x) - \alpha t \|\nabla f(x)\|_2^2$.*

**Proof** Since $0 \leq t \leq 1/M$, $-t + \frac{Mt^2}{2} \leq -t/2$. Then, for $0 \leq t \leq 1/M$ and $\alpha < 1/2$,

$$\begin{aligned} \tilde{f}(t) &\leq f(x) - t\|\nabla f(x)\|_2^2 + \tfrac{Mt^2}{2}\|\nabla f(x)\|_2^2 \\ &\leq f(x) - \tfrac{t}{2}\|\nabla f(x)\|_2^2 \\ &\leq f(x) - \alpha t\|\nabla f(x)\|_2^2. \ \square \end{aligned}$$

Thus, when we use backtracking line search with $t_0 := 1$, line search terminates with either $t = 1$ or $t \geq \beta/M$.

## Steepest descent direction

From first-order Taylor approximation of $f(x+v)$ around $x$,

$$f(x+v) \approx f(x) + \nabla f(x)^T v.$$

directional derivative $\nabla f(x)^T v$ gives an approximate change in $f$ for a small $v$, a descent direction if $\nabla f(x)^T v < 0$.

**Definition (Normalized steepest descent direction)**

$$\Delta x_{nsd} := argmin\{\nabla f(x)^T v \mid \|v\| = 1\}.$$

A search direction of unit norm giving largest decrease in the linear approximation of $f$.

We use as search direction an unnormalized steepest descent direction:

$$\Delta x_{\mathrm{sd}} = \|\nabla f(x)\|_* \Delta x_{\mathrm{nsd}},$$

where, $\|\cdot\|_*$ is dual norm of $\|\cdot\|$: $\|x\|_* = \max\{x^T y : \|y\| = 1\}$. (For instance, dual norms of $\|\cdot\|_2$, $\|\cdot\|_P$, and $\|\cdot\|_1$ are resp., $\|\cdot\|_2$, $\|\cdot\|_{P^{-1}}$, and $\|\cdot\|_\infty$.)
Also from definition,

$$\nabla f(x)^T \Delta x_{\mathrm{nsd}} = -\|\nabla f(x)\|_*^2.$$

**Algorithm (Steepest descent method)**

**given** *a starting point* $x \in \mathbf{dom} f$.

**repeat**

1. *Compute steepest descent direction* $\Delta x_{sd}$.
2. *Line search. Choose a step size* $t > 0$ *via backtracking or exact line search.*
3. *Update.* $x := x + t\Delta x_{sd}$.

**until** *stopping criterion is satisfied.*

## Steepest descent for various norms

- When $\|\cdot\|_2$ is used, $\Delta x_{\mathrm{sd}} = -\nabla f(x)$.
- When a quadratic norm, $\|z\|_P = (z^T P z)^{1/2} = \|P^{1/2} z\|_2,\ P \in \mathbb{S}^n_{++}$ is used,

$$\Delta x_{\mathrm{nsd}} = -\left(\nabla f(x)^T P^{-1} \nabla f(x)\right)^{-1/2} P^{-1} \nabla f(x), \tag{9}$$

- For $l_1$-norm,

$$\Delta x_{\mathrm{nsd}} = \mathrm{argmin}\{\nabla f(x)^T v \,|\, \|v\|_1 \leq 1\}.$$

Let $i$ be any index for which $\|\nabla f(x)\|_\infty = |(\nabla f(x))_i|$. Then, a normalized steepest descent direction for the $l_1$-norm is given by

$$\Delta x_{\mathrm{nsd}} = -\mathrm{sign}\left(\frac{\partial f(x)}{\partial x_i}\right) e_i, \tag{10}$$

where $e_i$ is the $i$th vector of standard basis.

## Convergence analysis

We assume $f$ is strongly convex on the initial sublevel set $S$, and hence $\nabla^2 f(x) \preceq MI$. Then,

$$\begin{aligned} f(x + t\Delta x_{\text{sd}}) &\leq f(x) + t\nabla f(x)^T \Delta x_{\text{sd}} + \frac{M\|x_{\text{sd}}\|_2^2}{2} t^2 \\ &\leq f(x) + t\nabla f(x)^T \Delta x_{\text{sd}} + \frac{M\|x_{\text{sd}}\|_*^2}{2\gamma^2} t^2 \\ &= f(x) - t\|\nabla f(x)\|_*^2 + \frac{M}{2\gamma^2} t^2 \|\nabla f(x)\|_*^2. \end{aligned}$$

where $\gamma \in (0, 1]$ and $\|x\|_* \geq \gamma \|x\|_2$ for all $x$.

- Note that the upper bound $f(x) - t\|\nabla f(x)\|_*^2 + \frac{M}{2\gamma^2} t^2 \|\nabla f(x)\|_*^2$ is minimized at $\hat{t} = \gamma^2 / M$.

## Convergence analysis(*cont'd*)

When backtracking line search is used,

- since $\alpha < 1/2$ and $\nabla f(x)^T \Delta x_{\text{sd}} = -\|\nabla f(x)\|_*^2$,

$$f(x + \hat{t}\Delta x_{\text{sd}}) \leq f(x) - \frac{\gamma^2}{2M}\|\nabla f(x)\|^2 \leq f(x) + \frac{\alpha\gamma^2}{M}\nabla f(x)^T \Delta x_{\text{sd}}$$

  satisfies the exit condition for backtracking line search.

- Thus, line search returns a step size $t \geq \min\{1, \beta\gamma^2/M\}$, and we have

$$\begin{aligned} f(x + t\Delta x_{\text{sd}}) &\leq f(x) - \alpha t\|\nabla f(x)\|^2 (\text{Line search exit criterion}) \\ &\leq f(x) - \alpha \min\{1, \beta\gamma^2/M\}\|\nabla f(x)\|^2 \\ &\leq f(x) - \alpha\gamma^2 \min\{1, \beta\gamma^2/M\}\|\nabla f(x)\|_2^2 \end{aligned}$$

## Convergence analysis(*cont'd*)

- This implies that

$$f(x + t\Delta x_{\text{sd}}) - p^* \le f(x) - p^* - \alpha\gamma^2 \min\{1, \beta\gamma^2/M\}\|\nabla f(x)\|_2^2$$

  But, from $f(x) - p^* \le \frac{1}{2m}\|\nabla f(x)\|_2^2$, or $-\|\nabla f(x)\|_2^2 \le -2m(f(x) - p^*)$, we get

$$f(x + t\Delta x_{\text{sd}}) - p^* \le c(f(x) - p^*),$$

  where $c = 1 - 2m\alpha\gamma^2 \min\{1, \beta\gamma^2/M\} < 1$.
- Hence $f(x^{(k)}) - p^* \le c^k(f(x^{(0)}) - p^*)$.

**Definition (Newton step)**

*For $x \in \mathbf{dom} f$, the vector*

$$\Delta x_{nt} := -\nabla^2 f(x)^{-1} \nabla f(x)$$

*is called the Newton step for $f$ at $x$.*

- If $\nabla^2 f(x) \succ 0$,

  $$\nabla f(x)^T \Delta x_{nt} = -\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) < 0,$$

  unless $\nabla f(x) = 0$.
- This implies that the Newton step is a descent direction.

## Some interpretations

- Consider the second-order Taylor approximation $\hat{f}$ of $f$ at $x$ is

$$\hat{f}(v) := f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v,$$

which is a convex quadratic function of $v$.
Then $\hat{f}$ is minimized when $v = \Delta x_{nt}$ as we have $\nabla \hat{f}(\Delta x_{nt}) = 0$.

- Newton step is also the steepest descent direction at $x$ for the quadratic norm defined by $\nabla^2 f(x)$,

$$\|u\|_{\nabla^2 f(x} = (u^T \nabla^2 f(x) u)^{\frac{1}{2}}.$$

- Linearizing optimality condition $\nabla f(x^*) = 0$ around $x$, we get

$$\nabla f(x+v) \approx \nabla f(x) + \nabla^2 f(x) v = 0.$$

Thus $x + \Delta x_{\mathrm{nt}}$ is the solution of the linear approximation of optimality condition.

### Algorithm (Newton's method)

given *a starting point $x \in \mathbf{dom} f$, tolerance $\epsilon > 0$.*

repeat

1. *Compute the Newton step and decrement:* $\Delta x_{nt} := -\nabla^2 f(x)^{-1} \nabla f(x)$; $\lambda^2 := \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)$.
2. *Stopping criterion. quit if* $\lambda^2/2 \leq \epsilon$.
3. *Line search. Choose a step size* $t > 0$ *via backtracking line search.*
4. *Update.* $x := x + t\Delta x_{sd}$.

## Convergence analysis

We assume that

(i) $f$ is twice continuously differentiable,

(ii) strongly convex with constants $m$ and $M$, i.e.,

$$mI \preceq \nabla^2 f(x) \preceq MI \quad \text{for } x \in S, \text{ and}$$

(iii) the Hessian of $f$ is *Lipschitz continuous* on $S$ with constant $L$, i.e.,

$$\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \leq L\|x - y\|_2, \ \forall x, y \in S.$$

Note that $L = 0$ is valid for a quadratic function. Thus, $L$ measures how well $f$ can be approximated by a quadratic model. Intuition suggests that Newton's method will work very well for a small $L$.

## Outline of convergence proof

We can prove that there are numbers $0 < \eta \leq m^2/L$ and $\gamma > 0$ such that

(i) if $\|\nabla f(x^{(k)})\|_2 \geq \eta$, then $f(x^{(k+1)}) - f(x^{(k)}) \leq -\gamma$, and

(ii) if $\|\nabla f(x^{(k)})\|_2 < \eta$, then the backtracking line search selects $t^{(k)} = 1$, and $\frac{L}{2m^2}\|\nabla f(x^{(k+1)})\|_2 \leq \left(\frac{L}{2m^2}\|\nabla f(x^{(k)})\|_2\right)^2$.

- From (i), the number of steps satisfying $\|\nabla f(x^{(k)})\|_2 \geq \eta$ cannot exceed $\frac{f(x^{(0)}) - p^*}{\gamma}$ since $f$ decreases by at least $\gamma$ at each iteration.
- From (ii), if $\|\nabla f(x^{(k)})\|_2 < \eta$, then $\|\nabla f(x^{(k+1)})\|_2 \leq \frac{L}{2m^2}\|\nabla f(x^{(k)})\|_2^2 \leq \frac{L}{2m^2}\eta^2$ which is $\leq \eta$ since $\eta \leq m^2/L$.

## Outline of convergence proof(*cont'd*)

- Thus once $\|\nabla f(x^{(k)})\|_2 < \eta$, then $\|\nabla f(x^{(l)})\|_2 < \eta$ and

$$\frac{L}{2m^2}\|\nabla f(x^{(l+1)})\|_2 \leq (\frac{L}{2m^2}\|\nabla f(x^{(l)})\|_2)^2, \ \forall l \geq k,$$

  called *quadratic convergence*.

- Applying this inequality recursively,

$$\frac{L}{2m^2}\|\nabla f(x^{(l)})\|_2 \leq \left(\frac{L}{2m^2}\|\nabla f(x^{(k)})\|_2\right)^{2^{l-k}} \leq \left(\frac{1}{2}\right)^{2^{l-k}}.$$

- and hence

$$f(x^{(l)}) - p^* \leq \frac{1}{2m}\|\nabla f(x^{(l)})\|_2^2 \leq \frac{2m^3}{L^2}\left(\frac{1}{2}\right)^{2^{l-k+1}}.$$

# Outline of convergence proof(*cont'd*)

The iterations in Newton's method fall into two stages:

- *damped Newton* phase where $\|\nabla f(x)\|_2 > \eta$ and algorithm can choose $t < 1$, and
- *pure Newton* phase where $\|\nabla f(x)\|_2 \leq \eta$ and hence algorithm choose full step size, $t = 1$.

From the previous observations, the number of iterations

- from damped Newton phase is $\leq (f(x^{(0)}) - p^*)/\gamma$, and
- from pure Newton phase, is given by $\epsilon \leq \frac{2m^3}{L^2}(\frac{1}{2})^{2^{l-k+1}}$, and hence bounded by

$$\log_2 \log_2(\epsilon_0/\epsilon), \text{ where } \epsilon_0 = 2m^3/L^2.$$

Thus, total number of iterations until $f(x) - p^* \leq \epsilon$ is bounded by

$$(f(x^{(0)}) - p^*)/\gamma + \log_2 \log_2(\epsilon_0/\epsilon) \approx (f(x^{(0)}) - p^*)/\gamma + 6.$$

# Homework

9.1, 9.3, 9.5, 9.7, 9.10

Additional Problems

1. Verify (9) and (10).

2. Verify that dual norms of $\|\cdot\|_2$, $\|\cdot\|_P$, and $\|\cdot\|_1$ are resp., $\|\cdot\|_2$, $\|\cdot\|_{P^{-1}}$, and $\|\cdot\|_\infty$.

3. Newton step is the steepest descent direction at $x$ for the quadratic norm defined by $\nabla^2 f(x)$.