# Review for the first Midterm 

October 30, 2015

The first midterm covers all the materials on the slides, up to the end of unconstrained problem (least square problems). NO MATLAB implementation related questions, but details, stability and convergence properties of some algorithms may be asked. If you don't have enough time, the review notes and the practice problems (on a separate files) should be more than enough for preparing the first midterm. Good luck.

## Contents

## 1 Introduction

### 1.1 Convex sets and Convex functions

The concepts of convex sets and convex functions play an essential role in optimization. Convex optimization problems are almost as "easy" as linear optimization problems (thought no simplex method) and the same time have the special structure to be adapted to many efficient algorithms.


Figure 1: The first (northwest) set is convex and others are not.
A set $\Omega$ is convex if for any $x, y \in \Omega$, the line segment $[x, y]$ is in $\Omega$, or

$$
\begin{equation*}
(1-\lambda) x+\lambda y \in \Omega \tag{1}
\end{equation*}
$$

for any $\lambda \in[0,1]$ (see Figure ??). A function $f$, defined on a convex set $\Omega$, is convex if

$$
\begin{equation*}
f((1-\lambda) x+\lambda y) \leq(1-\lambda) f(x)+\lambda f(y) . \tag{2}
\end{equation*}
$$

This condition is geometrically interpreted in Figure ??.
There are a lot of equivalent criteria that characterize convex functions.


Figure 2: Geometric interpretation of a function $f$ is convex: the graph of the function $f$ on the interval between $x$ and $y$ always lies below the straight line connecting $(x, f(x))$ and $(y, f(y))$.

Theorem 1.1. For a smooth function $f(x)$ of a single variable $x \in \mathbb{R}$, the following are equivalent
(a) The function $f$ is convex in the sense of (??).
(b) The first order derivative $f^{\prime}$ is increasing.
(c) The second order derivative $f^{\prime \prime}$ is nonnegative.

Similar when $f$ is a function of multiple variables $x \in \mathbb{R}^{n}$, we have the following.
Theorem 1.2. For a smooth function $f(x)$ for multiple variable $x \in \mathbb{R}^{n}$, similarly the following are equivalent
(a) The function $f$ is convex in the sense of (??)
(b) The gradient $\nabla f$ is monotone, i.e.

$$
(\nabla f(y)-\nabla f(x), y-x) \geq 0, \quad \forall x, y \in \Omega
$$

(c) The Hessian matrix $\nabla^{2} f$ is nonnegative definite, i.e. $p^{t} \nabla^{2} f(x) p \geq 0$ for any $p \in \mathbb{R}^{n}$ and any $x \in \Omega$.

These conditions can be used to show that a function is convex.
Example 1.1. Show that $f(x)=-\ln x$ is convex on $(0, \infty)$.
Since $f^{\prime}(x)=-1 / x$ and $f^{\prime \prime}(x)=1 / x^{2}>0$ on $(0, \infty), f(x)$ is convex on $(0, \infty)$.

Example 1.2. Show that the function $f(x, y)=x^{4} / 2+x^{2} y^{2}+y^{4} / 2$ is convex on $\mathbb{R}^{2}$.
The gradient and Hessian matrix of $f$ are

$$
\nabla f(x, y)=\binom{2 x^{3}+2 x y^{2}}{2 x^{2} y+2 y^{3}}, \quad \nabla^{2} f(x, y)=\left(\begin{array}{cc}
6 x^{2}+2 y^{2} & 4 x y \\
4 x y & 2 x^{2}+6 y^{2}
\end{array}\right)
$$

It is easy to see that $\operatorname{tr}\left(\nabla^{2} f(x, y)\right)=8\left(x^{2}+y^{2}\right) \geq 0$ and $\operatorname{det}\left(\nabla^{2} f(x, y)\right)=12\left(x^{4}+y^{4}+2 x^{2} y^{2}\right)=12\left(x^{2}+\right.$ $\left.y^{2}\right)^{2} \geq 0$. The characteristic equation is

$$
\operatorname{det}\left(\lambda I-\nabla^{2} f(x, y)\right)=\lambda^{2}-\lambda \operatorname{tr}\left(\nabla^{2} f(x, y)\right)+\operatorname{det}\left(\nabla^{2} f(x, y)\right)=0
$$

and the roots are given by

$$
\lambda_{ \pm}=\frac{\operatorname{tr}\left(\nabla^{2} f(x, y)\right) \pm \sqrt{\operatorname{tr}\left(\nabla^{2} f(x, y)\right)^{2}-4 \operatorname{det}\left(\nabla^{2} f(x, y)\right)}}{2} .
$$

Since the discriminant

$$
\operatorname{tr}\left(\nabla^{2} f(x, y)\right)^{2}-4 \operatorname{det}\left(\nabla^{2} f(x, y)\right)=16\left(x^{4}+2 x^{2} y^{2}+y^{4}\right)=16\left(x^{2}+y^{2}\right)^{2} \geq 0
$$

there are two real root. Since $\lambda_{+}$is nonnegative, from the relation $\lambda_{+} \lambda_{-}=\operatorname{det}\left(\nabla^{2} f(x, y)\right)^{2}-4 \operatorname{det}\left(\nabla^{2} f(x, y)\right) \geq$ $0, \lambda_{-} \geq 0$. Therefore, $f$ is convex on $\mathbb{R}^{2}$.

Many times the function $f$ is convex only on part of the domain, which is exactly the region where the second order derivative (or Hessian matrix) is nonnegative (nonnegative definite).

Example 1.3. Find the largest connected domain (one if there are more than one) on which $f(x)=e^{-x^{2} / 2}$ is convex.

Since $f^{\prime}(x)=-x e^{-x^{2} / 2}$ and $f^{\prime \prime}(x)=\left(x^{2}-1\right) e^{-x^{2} / 2}, f$ is convex for on the domain where $f^{\prime \prime}(x) \geq 0$, i.e.,

$$
\left(x^{2}-1\right) e^{-x^{2} / 2} \geq 0
$$

or $|x| \geq 1$. Therefore the last connected domain on which $f$ is convex is $[1, \infty)$ (or $(-\infty,-1]$ ).

There are other important non-smooth convex functions, one of which is in the form of norms of a vector for the absolute value for scalars.

Example 1.4. Show that $f(x)=\|x\|_{p}=\left(\left|x_{1}\right|^{p}+\cdots+\left|x_{n}\right|^{p}\right)^{1 / p}$ is convex when $p \geq 1$.
We can show this using the definition of the convexity and the triangle inequality of norms as follow:

$$
\begin{align*}
f((1-\lambda) x+\lambda y) & =\|(1-\lambda) x+\lambda y\|_{p} & & \\
& \leq\|(1-\lambda) x\|_{p}+\|\lambda y\|_{p} & & \text { (Triangle inequality } \left.\|x+y\|_{p} \leq\|x\|_{p}+\|y\|_{p}\right) \\
& =(1-\lambda)\|x\|_{p}+\lambda\|y\|_{p} & & \text { (Homogeneity of norm }\|\lambda x\|=|\lambda|\|x\|) \\
& =(1-\lambda) f(x)+\lambda f(y) . & & \tag{3}
\end{align*}
$$

We can show some functions are convex, building on other convex functions.
Example 1.5. Show that if $f$ and $g$ are convex on $\Omega$, so is $h(x)=\max [f(x), g(x)]$.
For any $x, y \in \Omega, \lambda \in[0,1]$, without loss of generality, we can assume that

$$
h((1-\lambda) x+\lambda y)=\max [f((1-\lambda) x+\lambda y), g((1-\lambda) x+\lambda y)]=f((1-\lambda) x+\lambda y)
$$

Therefore,

$$
\begin{align*}
h((1-\lambda) x+\lambda y) & =f((1-\lambda) x+\lambda y) \\
& \leq(1-\lambda) f(x)+\lambda f(y) \quad(f \text { is convex }) \\
& \leq(1-\lambda) \max [f(x), g(x)]+\lambda \max [f(y), g(y)] \\
& =(1-\lambda) h(x)+\lambda h(x) . \quad \text { (Definition of } h) \tag{4}
\end{align*}
$$

This proves that $h$ is convex.
More convex functions in the rest of this class.

### 1.2 Graphic method for the (simple) optimization problems

For simple problems (mainly for two variables), a lot of problems can be solved just graphic method, by plotting the feasible region (the region that all the constraints are satisfied and then the level sets (or contours) of the objective function. Depending on the problem is maximization or minimization, the problem is solved by moving these contours.

Example 1.6. Solve the following problem using graphic method.


Figure 3: Graphic method for Example 1.6.
The constraint $x_{1}+x_{2}=2$ is just a straight line. The contour line for the objective function is all the points $\left(x_{1}, x_{2}\right)$ such that $\sqrt{\left(x_{1}-2\right)^{2}+\left(x_{2}-2\right)^{2}}=c$ for the same constant $c$. The minimizer is the point $\left(x_{1}^{*}, x_{2}^{*}\right)=(1,1)$, the point where the contour line is tangent with the straight line. The minimal value is $\sqrt{1+1}=\sqrt{2}$.

Example 1.7. Find the minimizer of the following problem by graphic method.

$$
\begin{array}{ll}
\underset{x}{\operatorname{minimize}} & f\left(x_{1}, x_{2}\right)=\sqrt{\left(x_{1}-2\right)^{2}+\left(x_{2}-2\right)^{2}} \\
\text { subject to } & x_{1} \leq 0.75 \\
& x_{2} \leq x_{1}
\end{array}
$$

Since the constraints are now inequalities, each inequality corresponds to a region (instead of a line for equality constraint). Determine first the boundary of the region (by taking the equality in the constraint), and then which side of the region is feasible. The second step is not always so straightforward, you can take some special points or the asymptotic region to help you select the right region. For example, the points


Figure 4: Graphic method for Example 1.7.
$x_{1} \rightarrow+\infty$ and $x_{2} \rightarrow-\infty$ is definitely in the feasible region for the constraint $x_{2} \leq x_{1}$. This implies that we should choose the lower right part of the region. Finally, take the intersection of all feasible regions corresponding to one single constraint.

For this problem, the minimizer is $\left(x_{1}^{*}, x_{2}^{*}\right)=(0.75,0.75)$.
The graphic method gives a lot of insights for the constrained optimization in the later of the semester, and helps the understanding of how the conditions for local minimizer/maximizer should be modified.

### 1.3 Rate of convergence

The rate of convergence is used to measure how fast a particular algorithm approaches its optimal. It can be measured either as a quotient (Q-convergence) such that

$$
\frac{\left|x_{n+1}-x^{*}\right|}{\left|x_{n}-x^{*}\right|} \leq r
$$

for all $n$ large or as a power

$$
\left|x_{n}-x^{*}\right|^{1 / n} \leq r .
$$

Here $r$ is called the rate constant. In both case. we need $r<1$ for convergence. If such a number $r$ goes to zero, as $n$ increases, it is called $Q$-superlinear, i.e.,

$$
\lim _{n \rightarrow \infty} \frac{\left|x_{n+1}-x^{*}\right|}{\left|x_{n}-x^{*}\right|}=0
$$

An even faster convergence is possible (as in the Newton's method) is

$$
\left|x_{n+1}-x^{*}\right| \leq C\left|x_{n}-x^{*}\right|^{2}
$$

for some nonnegative number $C$.

Example 1.8. When Newton's method is applied to $f(x)=x^{4}$ with the starting point $x_{0}=1$. Find the recursive formula, the solution at each iteration and the convergence rate.

Since $f^{\prime}(x)=4 x^{3}$ and $f^{\prime \prime}(x)=12 x^{2}$, the Newton's method is

$$
x_{n+1}=x_{n}-\frac{f^{\prime}\left(x_{n}\right)}{f^{\prime \prime}\left(x_{n}\right)}=x_{n}-\frac{1}{3} x_{n}=\frac{2}{3} x_{n}
$$

The solution at each iteration is given by

$$
x_{n}=\frac{2}{3} x_{n-1}=\left(\frac{2}{3}\right)^{2} x_{n-2}=\cdots=\left(\frac{2}{3}\right)^{n} .
$$

Therefore, it converges to

$$
x^{*}=\lim _{n \rightarrow \infty}\left(\frac{2}{3}\right)^{n}=0
$$

Since

$$
\frac{\left|x_{n+1}-x^{*}\right|}{\left|x_{n}-x^{*}\right|}=\frac{2}{3}=r
$$

the convergence is Q -linear (also R -linear) and the rate constant is $2 / 3$.

## 2 Unconstrained Optimization

### 2.1 Necessary and sufficient conditions for smooth functions

Theorem 2.1 (Necessary conditions). If $x^{*}$ is optimal, then

- 1st-order necessary condition (NC1): $\nabla f\left(x^{*}\right)=0$
- 2nd-order necessary condition (NC2): the Hessian $\nabla^{2} f\left(x^{*}\right)$ is positive definite

These conditions are necessary, in the sense that if they are violated, we can find a nearby point $x$ with smaller function value, hence it can not be optimal. If $\nabla f\left(x^{*}\right) \neq 0$, then we can find a point with smaller function value along the negative gradient direction $p=-\nabla f\left(x^{*}\right)$. The tool for the proof is Taylor's Expansion, and the technique part is that which version to use and how to use it. You should remember of the intuition, but forget about the technique part.

Similarly, if $\nabla f\left(x^{*}\right)=0$ but $\nabla^{2} f\left(x^{*}\right)$ is NOT non-negative definite, i.e., there exists $p(\neq 0) \in \mathbb{R}^{n}$, such that $p^{t} A p<0$. The the function value is also going to decreasing along the direction $p$ (or $\phi(t)=f\left(x^{*}+t p\right)$ is decreasing on $[0, \delta)$ for $\delta$ small).
Theorem 2.2 (Sufficient condition (SC2)). If $x^{*}$ is such that $\nabla f\left(x^{*}\right)=0$ and $\nabla^{2} f\left(x^{*}\right)$ is positive definite, then $x^{*}$ is a local minimum.

If the Hessian matrix $\nabla^{2} f\left(x^{*}\right)$ is only nonnegative definite, then we need higher order derivatives to get a conclusion.

Theorem 2.3 (Sufficient condition with higher order derivatives(for a function of one single variable)). If at a point $x^{*}$, we have

$$
f^{\prime}\left(x^{*}\right)=0, \quad f^{\prime \prime}\left(x^{*}\right)=0, \quad \cdots, \quad f^{(n-1)}\left(x^{*}\right)=0, \quad f^{(n)}\left(x^{*}\right) \neq 0
$$

If $n$ is odd, then $x^{*}$ is neither a local minimizer nor a local maximizer (called saddle point); if $n$ is even and $f^{(n)}\left(x^{*}\right)>0$, then $x^{*}$ is a local minimizer; if $n$ is even and $f^{(n)}\left(x^{*}\right)<0$, then $x^{*}$ is a local maximizer.
Example 2.1. Find all the points $x^{*}$ such that $f^{\prime}\left(x^{*}\right)=0$ for the following function, determine it is a local minimizer, a local maximizer or a saddle point.

$$
\begin{gather*}
f(x)=x^{4}+4 x^{3}+6 x^{2}+4 x  \tag{5a}\\
f(0)=0, f^{\prime}(x)=(x+1) x^{2}(x-1)^{3}  \tag{5b}\\
f(x)=x+\sin (x) \tag{5c}
\end{gather*}
$$

In general, we have to find all the minimizers, compare their function value and obtain the global maximizer. There are different ways to show a given minimizer $x^{*}$ is a global minimizer. If $\nabla f\left(x^{*}\right)=0$ has only one solution, then it must be the global one, otherwise we need convexity.
Theorem 2.4. If $f$ is convex, then any local minimizer $x^{*}$ is a global minimizer.
Example 2.2. Find the maximum likelihood estimator $\lambda$ from the sample points $t_{i}$, by maximizer the function

$$
L=\prod_{i=1}^{m} \lambda e^{-\lambda t_{i}}
$$

Show that the estimate you find is the global minimizer, by showing that $-\ln L$ is convex (or $\ln L$ is convex).

### 2.2 Piecewise-smooth function

If the function $f$ is only piecewise smooth, we can consider the function on the interval on which it is smooth, and the local minimizer and then the global minimizer.
Example 2.3. Find the minimizer of $f(x)=\min (|x|-1,0)$.
The function $f(x)=0$ if and only if $|x|-1 \leq 0$ or $|x|<1$. Therefore,

$$
f(x)= \begin{cases}x-1, & x \geq 1 \\ 0, & -1 \leq x \geq 1 \\ -x-1, & x \leq-1\end{cases}
$$

Since $f^{\prime}(x)=1$ on $x \geq 1$, it is increasing and the minimizer is obtained at $x=1$ with minimal value $f(1)=0$. Similarly $f(x)$ is decreasing on $x \leq-1$ and has a minimizer at $x=-1$ with minimal value $f(-1)=0$. Therefore, the minimizer is the interval $[-1,1]$ with minimal value 0 . Obviously, the minimizers are not strict.
Example 2.4. Find the minimizer of

$$
f(x)=\left|x-x_{1}\right|+\left|x-x_{1}\right|+\cdots+\left|x-x_{m}\right|, \quad x_{1}<x_{2}<\cdots<x_{m} .
$$

### 2.3 Algorithms

You should be to produce one or two iteration of Steepest Descent method or Newton's method, but not the details of Conjugate Gradient method and forget about the formulas in Quasi-Newton method. Some problems may test your understanding of these methods: the advantage and disadvantages (under some conditions), the convergence rate, and the stability.

### 2.4 Least Square problem

The least square problem are originated from data fitting. We can actually get "solution" to overdetermined problems, using the least square approach.

Example 2.5. Consider the system of linear equations

$$
\begin{aligned}
x & =4 \\
y & =6 \\
x+y & =2 .
\end{aligned}
$$

1. Write the system in the form $A x=b$. Is the system consistent?
2. Find the point $\left(x^{*}, y^{*}\right)$ that minimizes the error $\|A x-b\|_{2}^{2}$. Calculate $A x-b$ and the error $\|A x-b\|_{2}^{2}$.
3. Now find the point on the line $y=3$ that minimizes the error $\|A x-b\|_{2}^{2}$. What is the error $\|A x-b\|_{2}^{2}$ for this solution? Compare to the error $\|A x-b\|$ from the previous part.
4. Sketch the linear equations. Mark both optimal solutions on your sketch.
5. Was there anything special about the value $A x-b$ from part (b)? If so, explain.

## 3 MATLAB practice

1. List of commands in Linear Algebra

| Commands | Description |
| :--- | :--- |
| eig(A) | The eigenvalue if $A$ |
| $[\mathrm{~V}, \mathrm{~L}]=\operatorname{eig}(\mathrm{A})$ | The eigenvalue (diagonal elements of V$)$ and <br> the corresponding eigenvectors (columns of L) of $A$ |
| norm $(\mathrm{x}, 1), \operatorname{norm}(\mathrm{x}, 2), \operatorname{norm}\left(\mathrm{x},{ }^{\prime}\right.$ 'inf ' $)$ | $\\|x\\|_{1},\\|x\\|_{2},\\|x\\|_{\infty}$ of the vector $x$ |
| trace (A), det (A) | Trace and Determinant of $A$ |
| $\mathrm{x}=\mathrm{A} \backslash \mathrm{b}$ | $x=A^{-1} b$ or the solution $x$ to the equation $A x=b$ |
| $[\mathrm{~L}, \mathrm{U}, \mathrm{P}]=\operatorname{lu}(\mathrm{A})$ | The LU decomposition of a matrix $A, A=L U$ |
| $[\mathrm{Q}, \mathrm{R}]=\operatorname{qr}(\mathrm{A})$ | The QR factorization of a matrix $A, A=Q R$ |
| $[\mathrm{U}, \mathrm{S}, \mathrm{V}]=\operatorname{svd}(\mathrm{A})$ | The svd decomposition of a matrix $A, A=U S V^{\prime}$ |

2. Condition number $\kappa(A)$ of a matrix $A$. Get a random 10 -by-10 matrix $A$ by $A=r a n d(10)$. The built-in command to compute the condition number of a matrix $A$ is cond(A). Can you find any relation (equality, or inequality) between the following numbers?
```
cond(A)^2, cond(A'*A), cond(A*A)
```

