Convex Optimisation

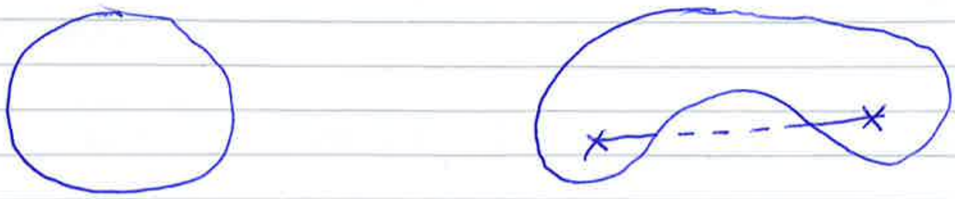
In the last lecture, you studied general nonlinear optimisation problems, and conditions for local optima. The theory is particularly nice for minimisation of convex functions over convex sets. In this case, local minima are also global minima. Moreover, the algorithms we will study in the next lecture are guaranteed to converge to global optima in such problems.

Def. : Convex sets :

A set $C \subseteq \mathbb{R}^n$ is said to be convex if $\forall \; \underline{x}_1, \underline{x}_2 \in C$ and $\forall \; \alpha \in [0,1]$,

$$\alpha \underline{x}_1 + (1-\alpha)\underline{x}_2 \in C.$$

In words, the line segment joining any two points in the set lies wholly within the set.



Convex Not convex

Def. : Convex functions

Let $C \subseteq \mathbb{R}^n$ be a convex set.

A function $f : C \rightarrow \mathbb{R}$ is said to be convex if, for all $\underline{x}_0, \underline{x}_1 \in C$, and all $\alpha \in [0, 1]$, we have

$$f\left((1-\alpha)\underline{x}_0 + \alpha \underline{x}_1\right) \leq (1-\alpha) f(\underline{x}_0) + \alpha f(\underline{x}_1)$$

Remarks

1. For notational convenience, we will let x_α denote $(1-\alpha) x_0 + \alpha x_1$. Note that $x_\alpha = x_0$ if $\alpha = 0$ and $x_\alpha = x_1$ if $\alpha = 1$

2. In studying convex functions, it is often convenient to allow $+\infty$ as a value. In the definition above, if we extend f to all of $\mathbb{R}^n$ by setting $f(\underline{x}) = +\infty$ if $\underline{x} \notin C$, then the definition goes through unchanged.

3. We don't allow $-\infty$ as a value because there are no interesting convex functions that take the value $-\infty$. If f is convex, and $f(\underline{x}) = -\infty$ for some $\underline{x}$, then $f(\underline{x}) = -\infty$ for all $\underline{x}$.

4. The set on which f is finite is called its domain

$$\mathrm{dom}(f) = \{ \underline{x} : f(\underline{x}) \neq \pm\infty \}$$

If f is a convex function. $\mathrm{dom}(f)$ is a convex set.

Def. : Concave functions

A function $f : \mathbb{R}^n \to \mathbb{R} \cup \{-\infty\}$ is said to be concave if $-f$ is convex.

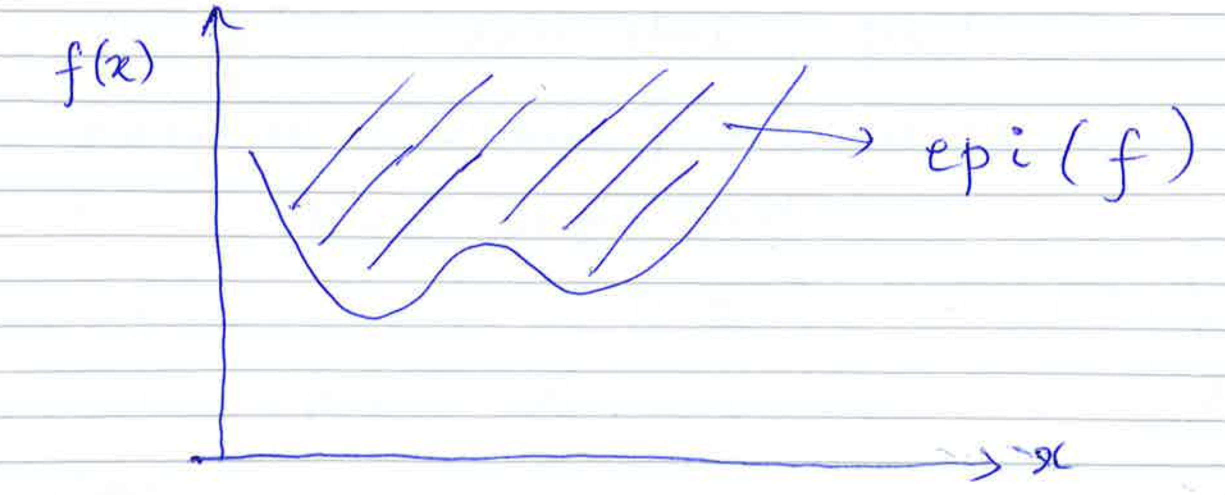
Note that, if f is a concave function, then $\text{dom}(f)$ is a convex set.

We now give an alternative characterisation of convex functions in terms of epigraphs, namely the set of points that lie on or above its graph.

Def : Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty, -\infty\}$ be any function. Its epigraph is the subset of $\mathbb{R}^n \times \mathbb{R}$ defined as

$$\text{epi}(f) = \{(\underline{x}, y) \in \mathbb{R}^n \times \mathbb{R} : y \geq f(\underline{x})\}$$

If f is identically $+\infty$, its epigraph is the empty set. If it is identically $-\infty$, its epigraph is all of $\mathbb{R}^n \times \mathbb{R}$.



Lemma 1 : A function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is convex iff epi(f) is a convex subset of $\mathbb{R}^{n+1}$.

Examples of convex functions

1. $f(x) = e^{\beta x}, \; x \in \mathbb{R}, \; \beta \in \mathbb{R}$.

2. $f(x) = \begin{cases} -\log x, & x > 0 \\ +\infty, & x \leq 0 \end{cases}$.

3. $f(x) = x^{2n}, \; n \in \mathbb{N}, \; x \in \mathbb{R}$

4. $f(x) = \begin{cases} x^{-\beta}, & x > 0 \\ +\infty, & x \leq 0 \end{cases}$, where $\beta > 0$

~~5. $f(\underline{x}) = A\underline{x} + \underline{b}, \; \underline{x} \in \mathbb{R}^n, \; \underline{b} \in \mathbb{R}^m$~~
~~$A \in \mathbb{R}^{m \times n}$~~

~~Such an f is called an affine function~~

5. $f(\underline{x}) = \underline{a}^T \underline{x} + b, \; \underline{a}, \underline{x} \in \mathbb{R}^n, \; b \in \mathbb{R}$

- such an f is called an affine function

6. $f(\underline{x}) = \underline{x}^T A \underline{x}, \; \underline{x} \in \mathbb{R}^n,$
$A \in \mathbb{R}^{n \times n}$, symmetric, psd

Exercise : Show f is convex

Hint : First show that

$$f(\underline{x}_\alpha) - (1-\alpha) f(\underline{x}_0) - \alpha f(\underline{x}_1)$$
$$= -\alpha(1-\alpha) (\underline{x}_0 - \underline{x}_1)^T A (\underline{x}_0 - \underline{x}_1)$$

Convexity is preserved by certain operations. Two important examples are linear combinations with positive coefficients, and maximisation.

Lemma 2

a) Suppose $f_1, \dots, f_k : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ are convex functions, and $c_1, \dots, c_k > 0$ are positive constants. Then

$$g = c_1 f_1 + c_2 f_2 + \dots + c_k f_k \quad \text{is convex}$$

b) Suppose f_i, $i \in \mathcal{I}$, are convex functions, where $\mathcal{I}$ is a (possibly infinite) index set. Let $f = \max_{i \in \mathcal{I}} f_i$. Then f is convex.

Proof:

(a) is straightforward from the defn. of convexity.

(b) The epigraph of f is given by

$$\mathrm{epi}(f) = \bigcap_{i \in \mathcal{I}} \mathrm{epi}(f_i)$$

Now, for each i, $\mathrm{epi}(f_i)$ is a convex set. Moreover, the intersection of convex sets is convex. Hence, $\mathrm{epi}(f)$ is a convex set, which implies that f is a convex function.

There is a well-known characterisation of convexity for twice differentiable functions, with which you may already be familiar

Lemma 3 : Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable, with Hessian (second derivative) matrix denoted $D^2 f$. Then, f is convex iff $D^2 f(\underline{x})$ is psd for all $\underline{x} \in \mathbb{R}^n$.

The proof is rather involved, so we omit it. We saw one special case of this as an example, namely $f(\underline{x}) = \underline{x}^T A \underline{x}$, for which $D^2 f(\underline{x}) = 2A$.

Another special case is $f(\underline{x}) = \underline{c}^T \underline{x} + b$, for which $D^2 f(\underline{x}) = 0$, the zero matrix.

Note that the latter function, $f(\underline{x}) = \underline{c}^T \underline{x} + b$ is both convex & concave.

With this background, we are now ready to look at convex optimisation problems.

Unconstrained Optimisation

Theorem : Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable. Then, $\underline{x}^*$ is a *global minimiser* of f iff $\nabla f(\underline{x}^*) = \underline{0}$

Remark : Note that the theorem doesn't claim such an $\underline{x}^*$ always exists, or that it is unique. f could be unbounded below, e.g. $f(x) = x$, $x \in \mathbb{R}$. Or the minimiser might not be unique, e.g., $f(x) = 0 \ \forall \ x \in \mathbb{R}$

Proof :

Suppose first that $\nabla f(\underline{x}^*) \neq \underline{0}$, and let $\underline{y} = (\nabla f(\underline{x}^*))^T$. (Recall that $\nabla f(\underline{x}^*)$ is a row vector of length n, so $\underline{y} \in \mathbb{R}^n$.)

Now,
$$\begin{aligned} f(\underline{x}^* - \varepsilon \underline{y}) &= f(\underline{x}^*) - \varepsilon \nabla f(\underline{x}^*)\underline{y} + o(\varepsilon) \\ &= f(\underline{x}^*) - \varepsilon\, \underline{y}^T \underline{y} + o(\varepsilon) \\ &= f(\underline{x}^*) - \varepsilon \cdot \|\underline{y}\|^2 + o(\varepsilon) \end{aligned}$$

Here, $\|\underline{y}\|$ denotes the Euclidean norm of $\underline{y} = (\nabla f(\underline{x}^*))^T$, which is strictly positive by the assumption that $\nabla f(\underline{x}^*) \neq \underline{0}$

Hence, $f(\underline{x}^* - \varepsilon \underline{y}) < f(\underline{x}^*)$, so $\underline{x}^*$ is not a global minimiser.

Proof (cont'd.)

Conversely, suppose $\underline{x}^*$ is not a global minimiser. Then,

$$\exists\ \underline{y} \in \mathbb{R}^n : f(\underline{y}) < f(\underline{x}^*)$$

Now, for $\varepsilon \in [0,1]$, we have by the convexity of f that

$$f(\underline{x}^* + \varepsilon(\underline{y} - \underline{x}^*)) \le (1-\varepsilon) f(\underline{x}^*) + \varepsilon f(\underline{y})$$

Hence, $\forall\ \varepsilon \in [0,1)$,

$$\frac{f(\underline{x}^* + \varepsilon(\underline{y} - \underline{x}^*)) - f(\underline{x}^*)}{\varepsilon} \le \frac{\varepsilon(f(\underline{y}) - f(\underline{x}^*))}{\varepsilon}$$

$$= f(\underline{y}) - f(\underline{x}^*)$$

The RHS is < 0 by the choice of $\underline{y}$.

Taking limits on the LHS as $\varepsilon \to 0$, we get $\nabla f(\underline{x}^*) \cdot (\underline{y} - \underline{x}^*)$.

If the product of these vectors is to be strictly negative, then $\nabla f(\underline{x}^*)$ can't be zero.

Thus we have shown that, if $\underline{x}^*$ is not a global minimiser, then $\nabla f(\underline{x}^*) \neq \underline{0}$.

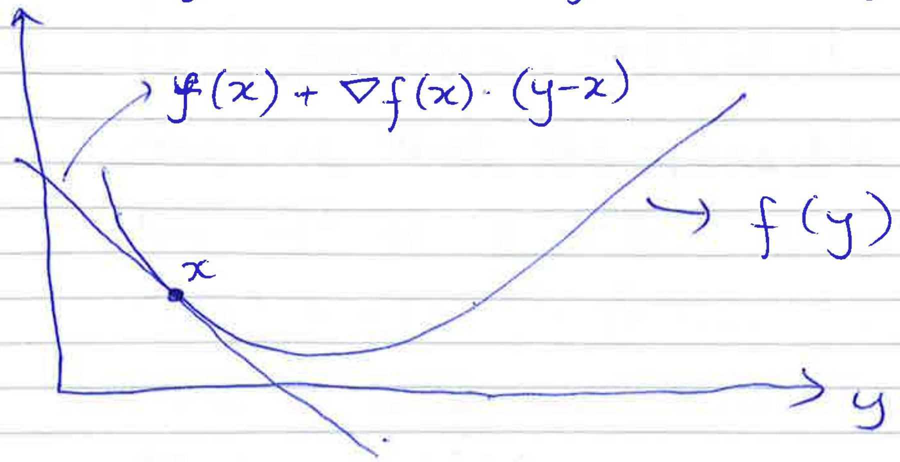
#

Remarks : The intuition behind the proof is not hard, and not very different from the one-dimensional case. The idea is that, if $f : \mathbb{R}^n \to \mathbb{R}$, then the directional derivative of f at $\underline{x}$, in the direction $\underline{y}$, is given by $\nabla f(\underline{x}) \cdot \underline{y}$. So, if $\nabla f(\underline{x})$ is non-zero, there must be some direction $\underline{y}$ in which $\nabla f(\underline{x}) \underline{y}$ is non-zero.

If it is negative, then $\underline{y}$ is a direction of decrease of f, contradicting minimality of $\underline{x}$. If it is positive, then $\underline{y}$ is a direction of increase and $-\underline{y}$ a direction of decrease. This part of the theorem (that $\nabla f(\underline{x}^*) = 0$ is necessary) does not require convexity.

The converse very much requires and uses convexity, which guarantees that

$$f(\underline{x} + \underline{y}) \geq f(\underline{x}) + \nabla f(\underline{x}) \underline{y} \quad \forall\ \underline{x}, \underline{y} \in \mathbb{R}^n$$



Constrained Optimisation

A convex optimisation problem in standard form is as follows:

$$\min \; f_0(\underline{x}) \;,\quad \underline{x} \in \mathbb{R}^n$$

subject to $f_i(\underline{x}) \leq 0 \,, \quad i = 1, \ldots, m$

$$A\underline{x} = \underline{b} \;,\quad A \in \mathbb{R}^{p \times n} \,,\; \underline{b} \in \mathbb{R}^p$$

where f_0 & f_i, $i = 1, \ldots, m$ are convex functions from $\mathbb{R}^n$ to $\mathbb{R}$.

Note that there are both (m) inequality and (p) equality constraints.

Remarks

1. If f_0 is a concave function to be maximised, it can be put in standard form by considering $-f_0$.
2. If we have a constraint $f_i(x) \leq c_i$, consider $g_i(x) = f_i(x) - c_i$.
3. If we have a constraint $f_i(x) \geq 0$ and f_i is concave, replace it with $g_i = -f_i$
4. Observe that the feasible set, namely the set of $\underline{x} \in \mathbb{R}^n$ satisfying the constraints, is (either empty or) convex. This is why we need the equality constraints to be affine.

The Lagrangian associated with the above convex optimisation problem is the function $L : \mathbb{R}^n \times \mathbb{R}^m_+ \times \mathbb{R}^p$ defined as

$$L(\underline{x}, \underline{\lambda}, \underline{\nu}) = f_0(\underline{x}) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i (A\underline{x} - \underline{b})_i$$

Note that we require $\lambda_i \geqslant 0$ but ν_i are unconstrained.

Dual problem

The Lagrangian dual function $g : \mathbb{R}^m_+ \times \mathbb{R}^p \rightarrow \mathbb{R} \cup \{-\infty\}$ is defined as

$$g(\underline{\lambda}, \underline{\nu}) = \min_{\underline{x} \in \mathbb{R}^n} L(\underline{x}, \underline{\lambda}, \underline{\nu})$$

Remark : For fixed $\underline{x}$, note that $L(\underline{x}, \underline{\lambda}, \underline{\nu})$ is an affine, hence concave, function of $(\underline{\lambda}, \underline{\nu})$. Therefore $g(\underline{\lambda}, \underline{\nu})$, being the minimum of concave functions, is also a concave function of $(\underline{\lambda}, \underline{\nu})$.

The dual problem is

$$\max_{\underline{\lambda}, \underline{\nu}} \; g(\underline{\lambda}, \underline{\nu})$$

where the max. is taken over all $\underline{\lambda} \in \mathbb{R}^m_+$ and $\underline{\nu} \in \mathbb{R}^p$.

What is the relation between the primal & dual problems? In the case of LP problems, we know that they have the same value. This is also "usually" the case for convex optimisation, but requires some extra technical assumptions called "constraint qualifications."

In the following, we denote the value of the primal solution by p^* (which could be $-\infty$ if the problem is unbounded, and is defined to be $+\infty$ if it is infeasible), and the value of the dual solution by d^* (which is defined to be $-\infty$ if it is infeasible).

<u>Weak Duality Theorem</u>: $d^* \leq p^*$

<u>Proof</u>: If the primal is infeasible, then $p^* = +\infty$ and there is nothing to prove. Hence, consider an $\underline{x}$ which is feasible, i.e., such that $A\underline{x} = \underline{b}$ and $f_i(\underline{x}) \leq 0 \ \forall i$.

Then, for all $\lambda_i \geq 0$, $\lambda_i f_i(\underline{x}) \leq 0$, so

$$L(\underline{x}, \underline{\lambda}, \underline{\nu}) = f_0(\underline{x}) + \sum \lambda_i f_i(\underline{x}) + \underline{\nu}^T(A\underline{x} - \underline{b})$$
$$\leq f_0(\underline{x})$$

$$\therefore \ g(\underline{\lambda}, \underline{\nu}) \leq L(\underline{x}, \underline{\lambda}, \underline{\nu}) \leq f_0(\underline{x})$$

Proof (contd.)

Thus, the relation

$$g(\underline{\lambda},\underline{\nu}) \le f_0(\underline{x})$$

holds for all dual-feasible $\underline{\lambda}, \underline{\nu}$ (i.e., $\underline{\lambda} \in \mathbb{R}^m_+$, $\underline{\nu} \in \mathbb{R}^p$) and all primal-feasible $\underline{x}$ (i.e. $f_i(\underline{x}) \le 0 \ \forall i$ and $A\underline{x} = \underline{b}$).

Taking the min. over $\underline{x}$ in the RHS gives p^*, taking the max. over $(\underline{\lambda},\underline{\nu})$ in the LHS gives d^*.

$$\therefore \quad d^* \le p^*$$

$\#$

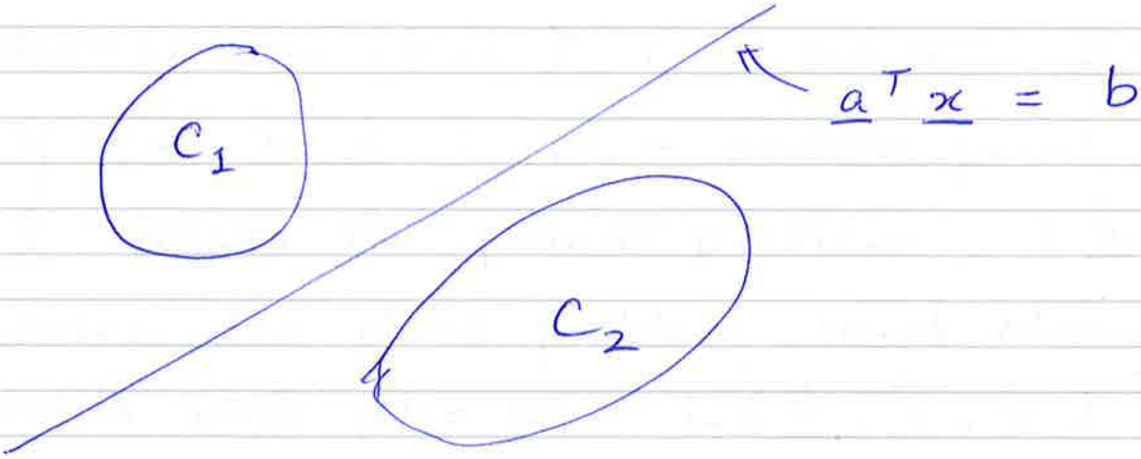
Strong duality theorem : Suppose that $\exists\ \underline{x} \in \mathbb{R}^n$ satisfying Slater's condition (strict feasibility) $f_i(\underline{x}) < 0$, $i = 1, \dots, m$, $A\underline{x} = \underline{b}$. Then, $d^* = p^*$.

The proof is quite difficult & we will skip it. If you wish to follow it in the book, then it uses the separating hyperplane theorem, stated below.

Separating hyperplane theorem :

Suppose C_1 and C_2 are disjoint convex subsets of $\mathbb{R}^n$. Then, $\exists\ \underline{a} \in \mathbb{R}^n,\ b \in \mathbb{R}$, $\underline{a} \neq \underline{0}$ such that

$$\begin{aligned} \underline{a}^T \underline{x} - b &\geq 0 \quad \forall\ \underline{x} \in C_1 \\ &\leq 0 \quad \forall\ \underline{x} \in C_2 \end{aligned}$$



We end this section with one more characterisation of primal & dual solutions in terms of saddle points of the Lagrangian.

Saddle point theorem :

We say that $(\hat{\underline{x}}, \hat{\underline{\lambda}}, \hat{\underline{\nu}})$ is a saddle point of the Lagrangian $L(\underline{x}, \underline{\lambda}, \underline{\nu})$ if

$$L(\hat{\underline{x}}, \underline{\lambda}, \underline{\nu}) \leq L(\hat{\underline{x}}, \hat{\underline{\lambda}}, \hat{\underline{\nu}}) \leq L(\underline{x}, \hat{\underline{\lambda}}, \hat{\underline{\nu}})$$

for all primal feasible $\underline{x}$ (i.e., such that $A\underline{x} = \underline{b}$ & $f_i(\underline{x}) \leq 0$, $i = 1, \dots, m$) and all dual-feasible $(\underline{\lambda}, \underline{\nu})$ (i.e. $\underline{\lambda} \in \mathbb{R}^m_+$, $\underline{\nu} \in \mathbb{R}^P$).

If $(\hat{\underline{x}}, \hat{\underline{\lambda}}, \hat{\underline{\nu}})$ is a saddle point, then $\hat{\underline{x}}$ solves the primal problem, $(\hat{\underline{\lambda}}, \hat{\underline{\nu}})$ solves the dual problem, and strong duality holds with $d^* = p^* = L(\hat{\underline{x}}, \hat{\underline{\lambda}}, \hat{\underline{\nu}})$.

Proof : If $\hat{\underline{x}}$ is infeasible, then either $f_i(\hat{\underline{x}}) > 0$ for some i, or $(A\hat{x} - b)_i \neq 0$ for some i. Then, we can choose $\lambda_i > 0$ or $\nu_i \in \mathbb{R}$ such as to make $L(\hat{\underline{x}}, \underline{\lambda}, \underline{\nu})$ arbitrarily large. So the inequalities can only hold with the latter two terms being $+\infty$, and the primal problem is infeasible.

Similarly, if $(\hat{\underline{\lambda}}, \hat{\underline{\nu}})$ is infeasible, then the dual problem is infeasible, and the first two terms in the inequalities must be $-\infty$.

From now, we ignore these uninteresting cases.

Proof (contd.)

So, we consider $(\underline{\hat{x}}, \underline{\hat{\lambda}}, \underline{\hat{\nu}})$ feasible.

Recall that

$$L(\underline{x}, \underline{\lambda}, \underline{\nu}) = f_0(\underline{x}) + \sum \lambda_i f_i(\underline{x}) + \underline{\nu}^T(A\underline{x} - \underline{b})$$

So we have

$$\begin{aligned} f_0(\underline{\hat{x}}) = L(\underline{\hat{x}}, \underline{0}, \underline{0}) &\leq L(\underline{\hat{x}}, \underline{\hat{\lambda}}, \underline{\hat{\nu}}) \\ &\leq L(\underline{x}, \underline{\hat{\lambda}}, \underline{\hat{\nu}}) \\ &= f_0(\underline{x}) + \sum_i \hat{\lambda}_i f_i(\underline{x}) + \underline{\hat{\nu}}^T(A\underline{x} - \underline{b}) \\ &\leq f_0(\underline{x}) \quad \forall \text{ feasible } \underline{x} \end{aligned}$$

Hence, $\underline{\hat{x}}$ solves the primal problem, and $p^* = f_0(\underline{\hat{x}}) \leq L(\underline{\hat{x}}, \underline{\hat{\lambda}}, \underline{\hat{\nu}})$. — ①

Next, we have for all $(\underline{\lambda}, \underline{\nu}) \in \mathbb{R}^m_+ \times \mathbb{R}^p$ that

$$\begin{aligned} g(\underline{\lambda}, \underline{\nu}) \leq L(\underline{\hat{x}}, \underline{\lambda}, \underline{\nu}) &\leq L(\underline{\hat{x}}, \underline{\hat{\lambda}}, \underline{\hat{\nu}}) \\ &= g(\underline{\hat{\lambda}}, \underline{\hat{\nu}}) \end{aligned}$$

Hence, $(\underline{\hat{\lambda}}, \underline{\hat{\nu}})$ solves the dual problem, and $d^* = g(\hat{\lambda}, \hat{\nu}) = L(\underline{\hat{x}}, \underline{\hat{\lambda}}, \underline{\hat{\nu}})$ — ②

From eqs. ① & ②, we get

$$p^* \leq L(\underline{\hat{x}}, \underline{\hat{\lambda}}, \underline{\hat{\nu}}) = d^*$$

But $d^* \leq p^*$ by weak duality. So we must have $p^* = d^* = L(\underline{\hat{x}}, \underline{\hat{\lambda}}, \underline{\hat{\nu}})$.

<u>Theorem</u> (Sufficiency of KKT conditions):

Suppose $\underline{\hat{x}}$ is primal feasible, $\underline{\hat{\lambda}} \in \mathbb{R}^m_+$, $\underline{\hat{\nu}} \in \mathbb{R}^p$, and $(\underline{\hat{x}}, \underline{\hat{\lambda}}, \underline{\hat{\nu}})$ satisfy the KKT conditions

$$\nabla f_0(\underline{\hat{x}}) + \sum_i \hat{\lambda}_i \nabla f_i(\underline{\hat{x}}) + \underline{\hat{\nu}}^T(A\underline{\hat{x}} - \underline{b}) = \underline{0},$$

$$\hat{\lambda}_i = 0 \text{ if } f_i(\underline{\hat{x}}) < 0 \quad \text{(complementary slackness)}.$$

Then $(\underline{\hat{x}}, \underline{\hat{\lambda}}, \underline{\hat{\nu}})$ is a saddle point of L.

<u>Proof</u>:

By complementary slackness, $\sum \hat{\lambda}_i f_i(\hat{x}) = 0$, and so $L(\underline{\hat{x}}, \underline{\hat{\lambda}}, \underline{\hat{\nu}}) = f_0(\underline{\hat{x}})$

But $L(\underline{\hat{x}}, \underline{\lambda}, \underline{\nu}) = f_0(\underline{\hat{x}}) + \sum \lambda_i f_i(\underline{\hat{x}}) + \underline{\nu}^T(A\underline{\hat{x}} - \underline{b})$

The second term on the RHS is ≤ 0 because $\lambda_i \geq 0 \ \forall i$ & $f_i(\underline{\hat{x}}) \leq 0 \ \forall i$. The third term is zero because $A\underline{\hat{x}} = \underline{b}$. Thus, $\forall (\underline{\lambda}, \underline{\nu}) \in \mathbb{R}^m_+ \times \mathbb{R}^p$, we have

$$L(\underline{\hat{x}}, \underline{\lambda}, \underline{\nu}) \leq f_0(\hat{x}) = L(\underline{\hat{x}}, \underline{\hat{\lambda}}, \underline{\hat{\nu}}).$$

This proves the first inequality in the defn. of a saddle point.

For the second, observe that for fixed $\underline{\hat{\lambda}}, \underline{\hat{\nu}}$, $L(\underline{x}, \underline{\hat{\lambda}}, \hat{\nu})$ is a convex function of $\underline{x}$, & the KKT condition is exactly the condition for it to have a global minimum at $\underline{\hat{x}}$.

//

Algorithms : Unconstrained problems

So far, we have looked at conditions guaranteeing optimality, such as the KKT conditions or the saddle point theorem. However, these can usually only be applied in very small problems (1-dimensional or small number of dimensions), or if the objective function and constraints take very specific forms (e.g. quadratic cost function and affine constraints). For general, medium-sized problems (100s - 1000s of dimensions), are there efficient algorithms to approximate the solution?

For convex optimisation problems, the answer is usually yes, and it is possible to provide guarantees on algorithm performance.

For general nonlinear optimisation problems, guarantees are rarely available, but the same algorithms often work well in practice, if the functions are not too "badly behaved".

We will study a few of these algorithms

Unconstrained problems

Given a convex function $f : \mathbb{R}^n \to \mathbb{R}$, we want an algorithm to determine

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

What kind of algorithm we use depends very much on what kind of "information" is available to it.

There are three things that we would ideally like to be able to do.

Given $\underline{x} \in \mathbb{R}^n$, evaluate

(i) $f(\underline{x})$, (ii) $\nabla f(\underline{x})$, (iii) $D^2 f(\underline{x})$

If even (i) is too computationally expensive, the problem is hopeless. So we assume we can always do (i).

The complexity increases steeply as we go from (i) to (ii) to (iii), as the object we want to compute goes from being a number to a length n vector to an $n \times n$ matrix.

The algorithms we study will generally use (ii) as well, and some will use (iii).

Basic idea : Descent methods

The basic idea is very simple.

1 Start at some $\underline{x}^{(0)} \in \mathbb{R}^n$

2. When at $\underline{x}^{(k)}$,

2(i) Pick a descent direction $\Delta \underline{x}^{(k)}$

2(ii) Decide the distance t to move in this direction

End up at $\underline{x}^{(k+1)} = \underline{x}^{(k)} + t \cdot \Delta \underline{x}^{(k)}$

3. Check a termination condition.

If it is satisfied, Stop.

If not, repeat step 2.

Remarks

1. By "descent direction", we mean that a small enough move will decrease the objective function, i.e.

$$f(\underline{x}^k + \varepsilon \Delta \underline{x}^k) < f(\underline{x}^k)$$

for $\varepsilon > 0$ sufficiently small

2. The term descent "direction" is used somewhat loosely : $\Delta \underline{x}^{(k)}$ need not be a unit vector

We will now mainly discuss how to choose $\Delta \underline{x}^{(k)}$ and t, and be vague about the termination condition.

Line Search

Given a direction $\Delta x^{(k)}$, we first address how to choose t, i.e., how to decide how far to move in this direction

t is usually called the step size.

Observe that once $\Delta x^{(k)}$ is chosen, we are faced with a 1-dim. problem:

$$\cancel{\min_{t \in \mathbb{R}} f(x^{(k)} + t\Delta x^{(k)}}$$

$$\min_{t \in \mathbb{R}} f(x^{(k)} + t \cdot \Delta x^{(k)})$$

Moreover, it is easy to check that this is a convex optimisation problem

It is often not hard to solve this "exactly" (to any required precision).

If we do this in each iteration step, the algorithm is said to be a <u>descent algorithm with exact line search</u>

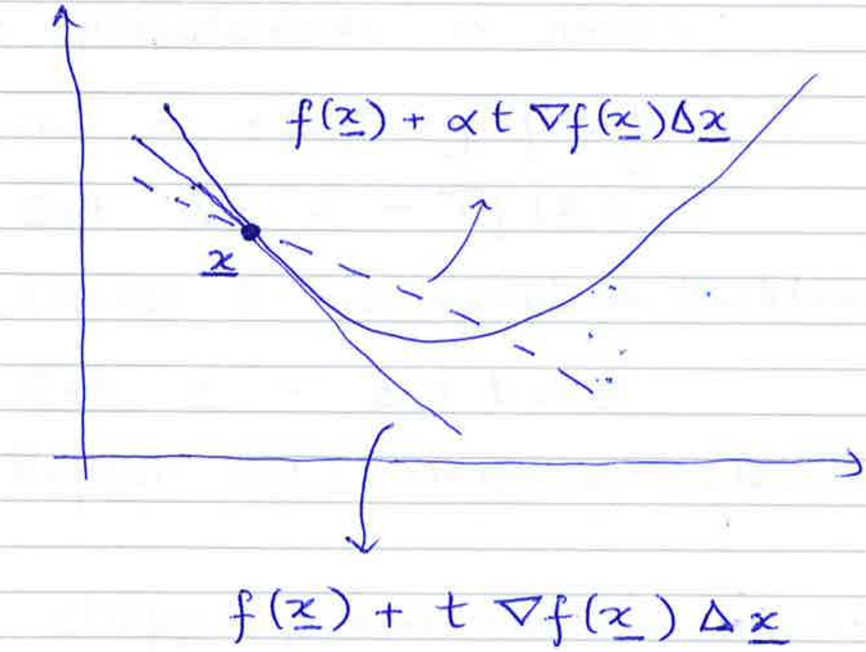
However, if function evaluations are too computationally expensive, then we might not want to do exact line search.

We just take a "suitable" step in direction $\Delta x^{(k)}$ and move to the next iteration

Backtracking line search

This is an alternative algorithm to exact line search, and seeks "enough improvement" in direction $\Delta x^{(k)}$ before moving on. The pseudocode is as follows:

```
1.  Pick two parameters $\alpha \in (0, \frac{1}{2})$, $\beta \in (0,1)$
    Given $\underline{x}$ and $\Delta \underline{x}$
2.  Set $t = 1$
3.  While $f(\underline{x} + t\Delta \underline{x}) > f(\underline{x}) + \alpha t \nabla f(\underline{x}) \Delta \underline{x}$
        set $t = \beta t$
4.  End
```



$f(\underline{x}) + \alpha t \nabla f(\underline{x}) \Delta \underline{x}$
$\underline{x}$
$f(\underline{x}) + t \nabla f(\underline{x}) \Delta \underline{x}$

Note that this method relies on being able to evaluate ∇f

Gradient descent method

Now we turn to how to choose the direction $\Delta \underline{x}^{(k)}$. A natural choice is the negative of the gradient. Why?

$$f(\underline{x} - \varepsilon \nabla f(\underline{x})^T)$$
$$= f(\underline{x}) - \varepsilon \nabla f(\underline{x}) \cdot \nabla f(\underline{x})^T + o(\varepsilon)$$
$$= f(\underline{x}) - \varepsilon \|\nabla f(\underline{x})\|^2 + o(\varepsilon)$$
$$< f(\underline{x}) \text{ for small enough } \varepsilon.$$

Thus $-\nabla f(\underline{x})^T$ is always a descent direction.

The pseudocode is now:

```
Given a starting point $\underline{x}$,
1.  Set $\Delta \underline{x} = -\nabla f(\underline{x})^T$.
2.  Choose $t$ via exact or backtracking line search
3.  Set $\underline{x} = \underline{x} + t \Delta \underline{x}$
4   Repeat until stopping criterion is satisfied.
```

The stopping criterion is usually of the form $\|\nabla f(\underline{x})\| \leq \eta$, where $\eta > 0$ is a suitably chosen small positive constant.

Steepest descent

The first-order Taylor expansion of f around a point $\underline{x} \in \mathbb{R}^n$ is

$$f(\underline{x}+\underline{v}) \approx f(\underline{x}) + \nabla f(\underline{x}) \cdot \underline{v}$$

Here, $\nabla f(\underline{x}) \cdot \frac{\underline{v}}{\|\underline{v}\|}$ is the directional derivative in the direction $\underline{v}/\|\underline{v}\|$.

If it is negative, we have a descent direction. By making it as negative as possible, we get the steepest descent direction

So we consider the problem

$$\min \ \nabla f(\underline{x}) \underline{v} \quad \text{subject to } \|\underline{v}\| = 1$$

The solution depends on the choice of norm, $\|\underline{v}\|$. If we use the standard Euclidean norm, then we recover gradient descent. Some other choices are:

1. Let $P \in \mathbb{R}^{n \times n}$ be a positive definite matrix. Define $\|\underline{z}\|_P = (\underline{z}^T P \underline{z})^{1/2} = \|P^{1/2}\underline{z}\|$

This leads to a steepest descent direction

$$\Delta \underline{x}_{SD} = -P^{-1}(\nabla f(\underline{x}))^T$$

2. l_1 norm : $\|\underline{z}\| = |z_1| + |z_2| + \dots + |z_n|$.

Steepest descent now corresponds to choosing the co-ordinate which yields the steepest descent.

Remarks

Gradient descent can be slow in "ill-conditioned" problems. For example, consider the quadratic objective function

$$f(\underline{x}) = \frac{1}{2}\,\underline{x}^T A\,\underline{x} + \underline{b}^T\underline{x} + c$$

where $A \in \mathbb{R}^{n \times n}$ is positive definite.

The solution can be obtained directly (by differentiating) as $\underline{x}^* = -A^{-1}\underline{b}$, but suppose we try to solve it by gradient descent. The speed of convergence will be dictated by the "condition number" of A, the ratio of its largest to smallest eigenvalue (ranked in absolute value). The bigger the condition number, the slower the convergence.

A good choice of P can help speed up convergence. In this case, $P = A$ is the perfect choice. In general, it won't be so easy to find the perfect P! But a good choice might be available.

Newton's method

$$\Delta \underline{x} = -\left(D^2 f(\underline{x})\right)^{-1} \nabla f(\underline{x})^T$$

Just as steepest descent was motivated by a first-order Taylor expansion around $\underline{x}$, this is ~~a~~ motivated by a second-order expansion

$$f(\underline{x}+\underline{v}) \simeq f(\underline{x}) + \nabla f(\underline{x}) \cdot \underline{v} + \frac{1}{2} \underline{v}^T D^2 f(\underline{x}) \underline{v}$$

This is a quadratic function of $\underline{v}$, and the matrix $D^2 f(\underline{x})$ is positive definite if f is strictly convex.

The Newton step is the exact solution to this quadratic optimisation problem.

Remarks

1. In practice, you don't invert the matrix $D^2 f(\underline{x})$. Instead, you solve the system of linear eqns. $D^2 f(\underline{x}) \cdot \Delta \underline{x} = -\nabla f(\underline{x})^T$.
2. The Newton method is usually very fast in terms of number of iterations, but each iteration could be quite expensive.