Notes on optimization

October 17, 2019

1 Optimization Problems

We consider the following problem:

$$\text{Minimize } f(\mathbf{x}) \text{ subject to } \mathbf{x} \in S, \tag{1}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $S \subseteq \mathbb{R}^n$.

Example: $f(\mathbf{x}) = \mathbf{c}^T\mathbf{x}$ and $S = \{\mathbf{x} \in \mathbb{R}^n : A\mathbf{x} = \mathbf{b}, \mathbf{x} \geq 0\}$ – Linear Programming.

Local versus Global Optima: $\mathbf{x}^*$ is a *global minimum* if it is an actual minimizer in (1).

$\mathbf{x}^*$ is a *local minimum* if there exists $\delta > 0$ such that $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x} \in B(\mathbf{x}^*) \cap S$, where $B(\mathbf{x}, \delta) = \{\mathbf{y} : |\mathbf{y} - \mathbf{x}| \leq \delta\}$ is the *ball* of radius δ, centred at $\mathbf{x}$.

See Diagram 1 at the end of these notes.

If $S = \emptyset$ then we say that the problem is *unconstrained*, otherwise it is *constrained*.

2 Convex sets and functions

2.1 Convex Functions

A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be *convex* if

$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda)f(\mathbf{y}).$$

See Diagram 2 at the end of these notes.

Examples of convex functions:

F1 A linear function $f(\mathbf{x}) = \mathbf{a}^T\mathbf{x}$ is convex.

F2 If $n = 1$ then f is convex iff

$$f(y) \geq f(x) + f'(x)(y - x) \text{ for all } x, y. \tag{2}$$

Proof. Suppose first that f is convex. Then for $0 < \lambda \leq 1$,

$$f(x + \lambda(y - x)) \leq (1 - \lambda)f(x) + \lambda f(y).$$

Thus, putting $h = \lambda(y - x)$ we have

$$f(y) \geq f(x) + \frac{f((x + h) - f(x))}{h}(y - x).$$

Taking the limit as $\lambda \to 0$ implies (2).

Now suppose that (2) holds. Choose $x \neq y$ and $0 \leq \lambda \leq 1$ and let $z = \lambda x + (1 - \lambda)y$. Then we have

$$f(x) \geq f(z) + f'(z)(x - z) \text{ and } f(y) \geq f(z) + f'(z)(y - z).$$

Multiplying the first inequality by λ and the second by $1 - \lambda$ and adding proves that

$$\lambda f(x) + (1 - \lambda)f(y) \geq f(z).$$

□

F3 If $n \geq 1$ then f is convex iff $f(\mathbf{y}) \geq f(\mathbf{x}) + (\nabla f(\mathbf{x}))^T(\mathbf{y} - \mathbf{x})$ for all $\mathbf{x}, \mathbf{y}$.
Apply F2 to the function $h(t) = f(t\mathbf{x} + (1 - t)\mathbf{y})$.

F4 A $n = 1$ and f is twice differentiable then f is convex iff $f''(z) \geq 0$ for all $z \in \mathbb{R}$.

Proof. Taylor's theorem implies that

$$f(y) = f(x) + f'(x)(y - x) + \frac{1}{2}f''(z)(y - x)^2 \text{ where } z \in [x, y].$$

We now just apply (2). □

F5 It follows from F4 that e^{ax} is convex for any $a \in \mathbb{R}$.

F6 x^a is convex on $\mathbb{R}_+$ for $a \geq 1$ or $a \leq 0$. x^a is concave for $0 \leq_a \leq 1$.
Here f is *concave* iff $-f$ is convex.

F7 Suppose that A is a symmetric $n \times n$ positive semi-definite matrix. Then $Q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$ is convex.
By positive semi-definite we mean that $Q(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$.
We have

$$\begin{aligned}
&Q(\lambda\mathbf{x} + (1 - \lambda)\mathbf{y}) - \lambda Q(\mathbf{x}) - (1 - \lambda)Q(\mathbf{y}) \\
=&\lambda^2 Q(\mathbf{x}) + (1 - \lambda)^2 Q(\mathbf{y}) + 2\lambda(1 - \lambda)\mathbf{x}^T A\mathbf{y} - \lambda Q(\mathbf{x}) - (1 - \lambda)Q(\mathbf{y}) \\
=& - \lambda(1 - \lambda)Q(\mathbf{y} - \mathbf{x}) \leq 0.
\end{aligned}$$

F8 If $n \geq 1$ then f is convex iff $\nabla^2 F = \left[\frac{\partial f^2}{dx_i dx_j}\right]$ is positive semi-definite for all $\mathbf{x}$.
Apply F7 to the function $h(t) = f(\mathbf{x} + t\mathbf{d})$ for all $\mathbf{x}, \mathbf{d} \in \mathbb{R}^n$.

Operations on convex functions

E1 If f, g are convex, then $f + g$ is convex.

E2 If $\lambda > 0$ and f is convex, then λf is convex.

E3 If f, g are convex then $h = \max\{f, g\}$ is convex.

Proof.

$$\begin{aligned} h(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) &= \max\{f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}), g(\lambda\mathbf{x} + (1-\lambda)\mathbf{y})\} \\ &\leq \max\{\lambda f(\mathbf{x}) + (1-\lambda)f(\mathbf{y}), \lambda g(\mathbf{x}) + (1-\lambda)g(\mathbf{y})\} \\ &\leq \lambda \max\{f(\mathbf{x}), g(\mathbf{x})\} + (1-\lambda)\max\{f(\mathbf{y}), g(\mathbf{y})\} \\ &= \lambda h(\mathbf{x}) + (1-\lambda)h(\mathbf{y}). \end{aligned}$$

□

Jensen's Inequality
If f is convex and $\mathbf{a}_i \in \mathbb{R}^n, \lambda_i \in \mathbb{R}_+, 1 \leq i \leq m$ and $\lambda_1 + \lambda_2 + \cdots + \lambda_m = 1$ then

$$f\left(\sum_{i=1}^{m} \lambda_i \mathbf{a}_i\right) \leq \sum_{i=1}^{m} f(\lambda_i \mathbf{a}_i).$$

The proof is by induction on m. $m = 2$ is from the definition of convexity and then we use

$$\sum_{i=1}^{m} \lambda_i \mathbf{a}_i = \lambda_m \mathbf{a}_m + (1-\lambda_m) \sum_{i=1}^{m-1} \frac{\lambda_i}{1-\lambda_m} \mathbf{a}_i.$$

Application: Arithmetic versus geometric mean.
Suppose that $a_1, a_2, \ldots, a_m \in \mathbb{R}_+$. Then

$$\frac{a_1 + a_2 + \cdots + a_m}{m} \geq (a_1 a_2 \cdots a_m)^{1/m}. \tag{3}$$

$-\log(x)$ is a convex function for $x \geq 0$. So, applying (3),

$$-\log\left(\sum_{i=1}^{m} \lambda_i \mathbf{a}_i\right) \leq \sum_{i=1}^{m} -\log(\lambda_i \mathbf{a}_i).$$

Now let $\lambda_i = 1/m$ for $i = 1, 2, \ldots, m$.

2.2 Convex Sets

A set $S \subseteq \mathbb{R}^n$ is said to be *convex* if $\mathbf{x}, \mathbf{y} \in S$ then the *line segment*

$$L(\mathbf{x}, \mathbf{y}) = \{\lambda \mathbf{x} + (1 - \lambda)\mathbf{y} \in S : 0 \leq \lambda \leq 1\}.$$

See Diagram 3 at the end of these notes.

Examples of convex sets:

C1 $S = \left\{\mathbf{x} : \mathbf{a}^T\mathbf{x} = 1\right\}$. $\mathbf{x}, \mathbf{y} \in S$ implies that

$$\mathbf{a}^T(\lambda \mathbf{x} + (1 - \lambda)\mathbf{y}) = \lambda \mathbf{a}^T\mathbf{x} + (1 - \lambda)\mathbf{a}^T\mathbf{y} = \lambda + (1 - \lambda) = 1.$$

C2 $S = \left\{\mathbf{x} : \mathbf{a}^T\mathbf{x} \leq 1\right\}$. Proof similar to C1.

C3 $S = B(0, \delta)$: $\mathbf{x}, \mathbf{y} \in S$ implies that

$$|\lambda \mathbf{x} + (1 - \lambda)\mathbf{y}| \leq |\lambda \mathbf{x}| + |(1 - \lambda)\mathbf{y}| \leq \lambda\delta + (1 - \lambda)\delta = \delta.$$

C4 If f is convex, then the *level set* $\{\mathbf{x} : f(\mathbf{x}) \leq 0\}$ is convex.
$f(\mathbf{x}), f(\mathbf{y}) \leq 0$ implies that $f(\lambda \mathbf{x} + (1 - \lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1 - \lambda)f(\mathbf{y}) \leq 0$.

Operations on convex sets:

O1 S convex and $\mathbf{x} \in \mathbb{R}^n$ implies that $\mathbf{x} + S = \{\mathbf{x} + \mathbf{y} : \mathbf{y} \in S\}$ is convex.

O2 S, T convex implies that $A = S \cap T$ is convex. $\mathbf{x}, \mathbf{y} \in A$ implies that $\mathbf{x}, \mathbf{y} \in S$ and so $L = L(\mathbf{x}, \mathbf{y}) \subseteq S$. Similarly, $L \subseteq T$ and so $L \subseteq S \cap T$.

O3 Using induction we see that if $S_i, 1 \leq i \leq k$ are convex then so is $\bigcap_{i=1}^k S_i$.

O4 If S, T are convex sets and $\alpha, \beta \in \mathbb{R}$ then $\alpha S + \beta T = \{\alpha \mathbf{x} + \beta \mathbf{y}\}$ is convex.
If $\mathbf{z}_i = \alpha \mathbf{x}_i + \beta \mathbf{y}_i \in T, i = 1, 2$ then

$$\lambda \mathbf{z}_1 + (1 - \lambda)\mathbf{z}_2 = \alpha(\lambda \mathbf{x}_1 + (1 - \lambda)\mathbf{x}_2) + \beta(\lambda \mathbf{y}_1 + (1 - \lambda)\mathbf{y}_2) \in T.$$

It follows from C1,C2 and O3 that an affine subspace $\{\mathbf{x} : A\mathbf{x} = \mathbf{b}\}$ and a halfspace $\{\mathbf{x} : A\mathbf{x} \leq \mathbf{b}\}$ are convex for any matrix A any vector $\mathbf{b}$.

We now prove something that implies the importance of the above notions. Most optimization algorithms can only find local minima. We do however have the following theorem:

Theorem 2.1. *Let f, S both be convex in* (1). *Then if $\mathbf{x}^*$ is a local minimum, it also a global minimum.*

Proof.

See Diagram 4 at the end of these notes.

Let δ be such that $\mathbf{x}^*$ minimises f in $B(\mathbf{x}^*, \delta) \cap S$ and suppose that $\mathbf{x} \in S \setminus B(\mathbf{x}^*, \delta)$. Let $\mathbf{z} = \lambda \mathbf{x}^* + (1-\lambda)\mathbf{y}$ be the point on $L(\mathbf{x}^*, \mathbf{y})$ at distance δ from $\mathbf{x}^*$. Note that $\mathbf{x} \in S$ by convexity of S. Then by the convexity of f we have

$$f(\mathbf{x}^*) \leq f(\mathbf{x}) \leq \lambda f(\mathbf{x}^*) + (1-\lambda) f(\mathbf{x})$$

and this implies that $f(\mathbf{x}^*) \leq f(\mathbf{x})$. □

The following shows the relationship between convex sets and functions.

Lemma 2.2. *let $f_1, f_2, \ldots, f_m$ be convex functions on $\mathbb{R}^n$. Let $\mathbf{b} \in \mathbb{R}^m$ and let*

$$S = \{\mathbf{x} \in \mathbb{R}^n : f_i(\mathbf{x}) \leq b_i, i = 1, 2, \ldots, m\}.$$

Then S is convex.

Proof. It follows from O3 that we can consider the case $m = 1$ only and drop the subscript. Suppose now that $\mathbf{x}, \mathbf{y} \in S$ i.e. $f(\mathbf{x}), f(\mathbf{y}) \leq b$. Then for $0 \leq \lambda \leq 1$

$$f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) \leq \lambda b + (1-\lambda) b = b.$$

So, $\lambda \mathbf{x} + (1-\lambda)\mathbf{y} \in S$. □

3 Algorithms

3.1 Line search – $n = 1$

Here we consider the simpler problem of minimising a convex (more generally *unimodal*) function $f : \mathbb{R} \to \mathbb{R}$.

See Diagram 5 at the end of these notes.

We assume that we are given a_0, a_1 such that $a_0 \leq x^* \leq a_1$ where x^* minimises f. This is not a significant assumption. We can start with $a_0 = 0$ and then consider the sequences $\zeta_i = f(2^i), \xi_i = f(-2^i)$ until we find $\zeta_{i-1} \leq \min\{\zeta_0, \zeta_i\}$ (resp. $\xi_{i-1} \leq \min\{\xi_0, \xi_i\}$). Then we know that $x^* \in [\zeta_0, \zeta_i]$ (resp. $x^* \in [\xi_0, \xi_i]$).

Assume then that we have an interval $[a_0, a_1]$ of uncertainty for x^*. Furthermore, we will have evaluated f at two points in this interval, two points inside the interval at $a_2 = a_0 + \alpha^2(a_1 - a_0)$ and $a_3 = a_0 + \alpha(a_1 - a_0)$ respectively. We will determine α shortly. And at each iteration we make one new function evaluation and decrease the interval of uncertainty by a factor α. There are two possibilities:

(i) $f(a_2) \leq f(a_3)$. This implies that $x^* \in [a_0, a_3]$. So, we evaluate $f(a_0 + \alpha^2(a_3 - a_0))$ and make the changes $a_i \to a_i'$:

$$a_0' \leftarrow a_0,\ a_1' \leftarrow a_3,\ a_2' \leftarrow a_0 + \alpha^2(a_3 - a_0),\ a_3' \leftarrow a_2.$$

(ii) $f(a_2) > f(a_3)$. This implies that $x^* \in [a_2, a_1]$. So, we evaluate $f(a_0+)$ and make the changes $a_i \to a_i'$:

$$a_0' \leftarrow a_2,\ a_1' \leftarrow a_1,\ a_2' \leftarrow a_3,\ a_3' \leftarrow a_2 + \alpha^2(a_1 - a_0).$$

In case (i) we see that $a_1' - a_0' = a_3 - a_0 = \alpha(a_1 - a_0)$ and so the interval has shrunk by the required amount. Next we see that $a_2' - a_0' = \alpha^2(a_3 - a_0) = \alpha^2(a_1' - a_0)$. Furthermore, $a_3' - a_0' = a_2 - a_0 = \alpha^2(a_1 - a_0) = \alpha(a_1' - a_0')$.

In case (ii) we see that $a_1' - a_0' = a_1 - a_2 = a_1 - (a_0 + \alpha^2(a_1 - a_0)) = (1 - \alpha^2)(a_1 - a_0)$. So, shrink by α in this case we choose α to satisfy $1 - \alpha^2 = \alpha$. This gives us

$$\alpha = \frac{\sqrt{5} - 1}{2} \text{ – \textit{the golden ratio.}}$$

Next we see that $a_2' - a_0' = a_3 - a_2 = (\alpha - \alpha^2)(a_1 - a_0) = \frac{\alpha - \alpha^2}{\alpha}(a_1' - a_0') = (1 - \alpha)(a_1' - a_0') = \alpha^2(a_1' - a_0')$. Finally, we have $a_3' - a_0' = a_2 + \alpha^2(a_1 - a_0) - a_2 = \alpha^2(a_1 - a_0) = \alpha(a_1' - a_0')$.

Thus to achieve an accuracy within δ of x^* we need to take t steps, where $\alpha^t D \leq \delta$ where D is our initial uncertainty.

3.2 Gradient Descent

See Diagram 6 at the end of these notes.

Here we consider the unconstrained problem. At a point $\mathbf{x} \in \mathbb{R}^n$, if we move a small distance h in direction $\mathbf{d}$ then we have

$$f(\mathbf{x} + h\mathbf{d}/|\mathbf{d}|) = f(\mathbf{x}) + h(\nabla f)^T \frac{\mathbf{d}}{|\mathbf{d}|} + O(h^2) \geq f(\mathbf{x}) - h|\nabla f| + O(h^2).$$

Thus, at least infinitessimally, the best direction is $-\nabla f$. So, for us, the steepest algorithm will follow a sequence of points $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_k, \ldots$, where

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k).$$

Then we have

$$\begin{aligned} |\mathbf{x}_{k+1} - \mathbf{x}^*|^2 &= |\mathbf{x}_k - \mathbf{x}^*|^2 - 2\alpha_k \nabla f(\mathbf{x}_k)^T(\mathbf{x}_k - \mathbf{x}^*) + \alpha_k^2 |\nabla f(\mathbf{x}_k)^2| \\ &\leq |\mathbf{x}_k - \mathbf{x}^*|^2 - 2\alpha_k(f(\mathbf{x}_k) - f(\mathbf{x}^*)) + \alpha_k^2 |\nabla f(\mathbf{x}_k)|^2. \end{aligned} \tag{4}$$

The inequality comes from F3.

Applying (4) repeatedly we get

$$|\mathbf{x}_k - \mathbf{x}^*|^2 \leq |\mathbf{x}_0 - \mathbf{x}^*|^2 - 2\sum_{i=1}^{k} \alpha_i (f(\mathbf{x}_i) - f(\mathbf{x}^*)) + \sum_{i=1}^{K} \alpha_i^2 |\nabla f(\mathbf{x}_k)|^2. \tag{5}$$

Putting $R = |\mathbf{x}_0 - \mathbf{x}^*|$, we see from (5) that

$$2\sum_{i=1}^{k} \alpha_i (f(\mathbf{x}_i) - f(\mathbf{x}^*)) \leq R^2 + \sum_{i=1}^{K} \alpha_i^2 |\nabla f(\mathbf{x}_k)|^2. \tag{6}$$

On the other hand,

$$\sum_{i=1}^{k} \alpha_i (f(\mathbf{x}_i) - f(\mathbf{x}^*)) \geq \left(\sum_{i=1}^{k} \alpha_i\right) \min \{f(\mathbf{x}_k) - f(\mathbf{x}^*) : i \in [k]\} = \left(\sum_{i=1}^{k} \alpha_i\right) (f(\mathbf{x}_{min} - f(\mathbf{x}^*)), \tag{7}$$

where $f(\mathbf{x}_{min}) = \min \{f(\mathbf{x}_i) : i \in [k]\}$.

Combining (6) and (7) we get

$$f(\mathbf{x}_{min}) - f(\mathbf{x}^*) \leq \frac{R^2 + G^2 \sum_{i=1}^{k} \alpha_i^2}{2\sum_{i=1}^{k} \alpha_i},$$

where $G = \max \{|\nabla f(\mathbf{x}_i)| : i \in [\kappa]\}$.

So, if we choose α_k so that $\sum_{i=1}^{\infty} \alpha_i = \infty$ and $\sum_{i=1}^{\infty} \alpha_i^2 = O(1)$ then

$$|f(\mathbf{x}_{min}) - f(\mathbf{x}^*)| \to 0 \text{ as } k \to \infty.$$

As an example, we could let $\alpha_i = 1/i$.

4 Separating Hyperplane

See Diagram 7 at the end of these notes.

Theorem 4.1. *Let C be a convex set in $\mathbb{R}^n$ and suppose $\mathbf{x} \notin C$. Then there exists $\mathbf{0} \neq \mathbf{a} \in \mathbb{R}^n$ and $b \in \mathbb{R}$ such that (i) $\mathbf{a}^T\mathbf{x} \geq b$ and (ii) $C \subseteq \left\{\mathbf{y} \in \mathbb{R}^n : \mathbf{a}^T\mathbf{y} \leq b\right\}$.*

Proof.
Case 1: C is closed.
Let $\mathbf{z}$ be the closest point in C to $\mathbf{x}$. Let $\mathbf{a} = \mathbf{x} - \mathbf{z} \neq \mathbf{0}$ and $b = (\mathbf{x} - \mathbf{z})^T\mathbf{z}$. Then

$$\mathbf{a}^T\mathbf{x} - b = (\mathbf{x} - \mathbf{z})^T\mathbf{x} - (\mathbf{x} - \mathbf{z})^T\mathbf{z} = |\mathbf{x} - \mathbf{z}|^2 > 0.$$

This verifies (i). Suppose (ii) fails and there exists $\mathbf{y} \in C$ such that $\mathbf{a}^T\mathbf{y} > b$. Let $\mathbf{w} \in C$ be the closest point to $\mathbf{x}$ on the line segment $L(\mathbf{y}, \mathbf{z}) \subseteq C$. The triangle formed by $\mathbf{x}, \mathbf{w}, \mathbf{z}$ has a right angle at $\mathbf{w}$ and an acute angle at $\mathbf{z}$. This implies that $|\mathbf{x} - \mathbf{w}| < |\mathbf{x} - \mathbf{z}|$, a contradiction.

Case 2: $\mathbf{x} \notin \bar{C}$**.**
We observe that $\bar{C} \supseteq C$ and is convex (exercise). We can thus apply Case 1, with $\bar{C}$ replacing C.

Case 3: $\mathbf{x} \in \bar{C} \setminus C$**.** Every ball $B(\mathbf{x}, \delta)$ contains a point of $\mathbb{R}^n \setminus \bar{C}$ that is distinct from $\mathbf{x}$. Choose a sequence $\mathbf{x}_n, \notin \bar{C}, n \geq 1$ that tends to $\mathbf{x}$. For each $\mathbf{x}_n$, let $\mathbf{a}_n, b_n = \mathbf{a}_n^T \mathbf{z}_n$ define a hyperplane that separates $\mathbf{x}_n$ from $\bar{C}$, as in Case 2. We can assume that $|\mathbf{a}_n| = 1$ (scaling) and that b_n is in some bounded set and so there must be a convergent subsequence of $(\mathbf{a}_n, b_n), n \geq 1$ that converges to $(\mathbf{a}, b), |\mathbf{a}| = 1$. Assume that we re-label so that this subsequence is $(\mathbf{a}_n), n \geq 1$. Then for $\mathbf{y} \in \bar{C}$ we have $\mathbf{a}_n^T \mathbf{y} \leq b_n$ for all n. Taking limits we see that $\mathbf{a}^T \mathbf{y} \leq b$. Furthermore, for $\mathbf{y} \notin \bar{C}$ we see that for large enough n, $\mathbf{a}_n^T \mathbf{y} > b_n$. taking limits we see that $\mathbf{a}^T \mathbf{y} \leq b$. □

Corollary 4.2. *Suppose that $S, T \subseteq \mathbb{R}^n$ are convex and that $S \cap T = \emptyset$. Then there exists $\mathbf{a}, b$ such that $\mathbf{a}^T \mathbf{x} \leq b$ for all $\mathbf{x} \in S$ and $\mathbf{a}^T \mathbf{x} \geq b$ for all $\mathbf{x} \in T$.*

Proof. Let $W = S + (-1)T$. Then $\mathbf{0} \notin W$ and applying Theorem 4.1 we see that there exists $\mathbf{a}$ such that $\mathbf{a}^T \mathbf{z} \leq 0$ for all $\mathbf{z} \in W$. Now put

$$b = \frac{1}{2} \left(\sup_{\mathbf{x} \in S} \mathbf{a}^T \mathbf{x} + \inf_{\mathbf{x} \in T} \mathbf{a}^T \mathbf{x} \right).$$

□

Corollary 4.3 (Farkas Lemma)**.** *For an $m \times n$ matrix and $\mathbf{b} \in \mathbb{R}^m$, exactly one of the following holds:*

(i) There exists $\mathbf{x} \in \mathbb{R}^n$ such that $\mathbf{x} \geq \mathbf{0}, A\mathbf{x} = \mathbf{b}$.

(ii) There exists $\mathbf{u} \in \mathbb{R}^m$ such that $\mathbf{u}^T A \geq \mathbf{0}$ and $\mathbf{u}^T \mathbf{b} < 0$.

Proof. We cannot have both (i), (ii) holding. For then we have

$$0 \leq \mathbf{u}^T A\mathbf{x} = \mathbf{u}^T \mathbf{b} < 0.$$

Suppose then that (i) fails to hold. Let $S = \{\mathbf{y} : \mathbf{y} = A\mathbf{x} \text{ for some } \mathbf{x} \geq \mathbf{0}\}$. Then $\mathbf{b} \notin S$ and since S is closed there exists $\boldsymbol{\alpha}, \beta$ such that (a) $\boldsymbol{\alpha}^T \mathbf{b} \leq \beta$ and (b) $\boldsymbol{\alpha}^T A\mathbf{x} \geq \beta$ for all $\mathbf{x} \geq 0$. This implies that $\boldsymbol{\alpha}^T(\mathbf{b} - A\mathbf{x}) \leq 0$ for all $\mathbf{x} \geq 0$. This then implies that $\mathbf{u} = \boldsymbol{\alpha}$ satisfies (ii). □

4.1 Convex Hulls

See Diagram 8 at the end of these notes.

Given a set $S \subseteq \mathbb{R}^n$, we let

$$conv(S) = \left\{ \sum_{i \in I} \lambda_i \mathbf{x}_i : (i)\ |I| < \infty,\ (ii)\ \sum_{i \in I} \lambda_i = 1,\ (iii)\ \lambda_i > 0, i \in I,\ (iv)\ \mathbf{x}_i \in S, i \in I \right\}.$$

Clearly $S \subseteq conv(S)$, since we can take $|I| = 1$.

Lemma 4.4. *$conv(S)$ is a convex set.*

Proof. Let $\mathbf{x} = \sum_{i \in I} \lambda_i \mathbf{x}_i, \mathbf{y} = \sum_{j \in J} \mu_j \mathbf{y}_j \in conv(S)$. Let $K = I \cup J$ and put $\lambda_i = 0, i \in J \setminus I$ and $\mu_j = 0, j \in I \setminus J$. Then for $0 \leq \alpha \leq 1$ we see that

$$\alpha \mathbf{x} + (1-\alpha)\mathbf{y} = \sum_{i \in K} (\alpha \lambda_1 + (1-\alpha)\mu_i)\mathbf{x}_i \text{ and } \sum_{i \in K} (\alpha \lambda_1 + (1-\alpha)\mu_i) = 1$$

implying that $\alpha \mathbf{x} + (1-\alpha)\mathbf{y} \in conv(S)$ i.e. $conv(S)$ is convex. □

Lemma 4.5. *If S is convex, then $S = conv(S)$.*

Proof. Exercise. □

Corollary 4.6. *$conv(conv(S)) = conv(S)$ for all $S \subseteq \mathbb{R}^n$.*

Proof. Exercise. □

4.1.1 Extreme Points

A point $\mathbf{x}$ of a convex set S is said to be an *extreme point* if **THERE DO NOT EXIST** $\mathbf{y}, \mathbf{z} \in S$ such that $\mathbf{x} \in L(\mathbf{y}, \mathbf{z})$. We let $ext(S)$ denote the set of extreme points of S.

EX1 If $n = 1$ and $S = [a, b]$ then $ext(S) = \{a, b\}$.

EX2 If $S = B(0, 1)$ then $ext(S) = \{\mathbf{x} : |\mathbf{x}| = 1\}$.

EX3 If $S = \{\mathbf{x} : A\mathbf{x} = \mathbf{b}\}$ is the set of solutions to a set of linear equations, then $ext(S) = \emptyset$.

Theorem 4.7. *Let S be a closed, bounded convex set. Then $S = conv(ext(S))$.*

Proof. We prove this by induction on the dimension n. For $n = 1$ the result is trivial, since then S must be an interval $[a, b]$.

Inductively assume the result for dimensions less than n. Clearly, $S \supseteq T = conv(ext(S))$ and suppose there exists $\mathbf{x} \in S \setminus T$. Let $\mathbf{z}$ be the closest point of T to $\mathbf{x}$ and let $H = \left\{\mathbf{y} : \mathbf{a}^T\mathbf{y} = b\right\}$ be the hyperplane defined in Theorem 4.1. Let $b^* = \max\left\{\mathbf{a}^T\mathbf{y} : \mathbf{y} \in S\right\}$. We have $b^* < \infty$ since S is bounded. Let $H^* = \left\{\mathbf{y} : \mathbf{a}^T\mathbf{y} = b^*\right\}$ and let $S^* = S \cap H^*$.

We observe that if $\mathbf{w}$ is a vertex of S^* then it is also a vertex of S. For if $\mathbf{w} = \lambda\mathbf{w}_1 + (1 - \lambda)\mathbf{w}_2, \mathbf{w}_1, \mathbf{w}_2 \in S, 0 < \lambda < 1$ then we have

$$b^* = \mathbf{a}^T\mathbf{w} = \lambda\mathbf{a}^T\mathbf{w}_1 + (1-\lambda)\mathbf{a}^T\mathbf{w}_2 \leq \lambda b^* + (1-\lambda)b^* = b^*.$$

This implies that $\mathbf{a}^T\mathbf{w}_1 = \mathbf{a}^T\mathbf{w}_2 = b^*$ and so $\mathbf{w}_1, \mathbf{w}_2 \in S^*$, contradiction.

Now consider the point $\mathbf{w}$ on the half-line from $\mathbf{z}$ through $\mathbf{x}$ that lies in S^* i.e

$$\mathbf{w} = \mathbf{z} + \frac{b^* - b}{\mathbf{a}^T\mathbf{x} - b}(\mathbf{x} - \mathbf{z}).$$

Now by induction, we can write $\mathbf{w} = \sum_{i=1}^k \lambda_i\mathbf{w}_i$ where $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k$ are extreme points of S^* and hence of S. Also, $\mathbf{x} = \mu\mathbf{w} + (1-\mu)\mathbf{z}$ for some $0 < \mu \leq 1$ and so $\mathbf{x} \in ext(S)$. □

The following is sometimes useful.

Lemma 4.8. *Suppose that S is a closed bounded convex set and that f is a convex function. The f achieves its* maximum *at an extreme point.*

Proof. Suppose the maximum occurs at $\mathbf{x} = \lambda_1\mathbf{x}_1 + \cdots + \lambda_k\mathbf{x}_k$ where $0 \leq \lambda_1, \ldots, \lambda_k \leq 1$ and $\lambda_1 + \cdots + \lambda_k = 1$ and $\mathbf{x}_1, \ldots, \mathbf{x}_k \in ext(S)$. Then by Jensen's inequality we have $f(\mathbf{x}) \leq \lambda_1 f(\mathbf{x}_1) + \cdots + \lambda_k f(\mathbf{x}_k) \leq \max\{f(\mathbf{x}_i) : 1 \leq i \leq k\}$. □

This explains why the solutions to linear programs occur at extreme points.

5 Lagrangean Duality

See Diagram 9 at the end of these notes.

Here we consider the *primal problem*

$$\text{Minimize } f(\mathbf{x}) \text{ subject to } g_i(\mathbf{x}) \leq 0, i = 1, 2, \ldots, m, \tag{8}$$

where $f, g_1, g_2, \ldots, g_m$ are convex functions on $\mathbb{R}^n$.

The Lagrangean

$$L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \sum_{i=1}^m \lambda_i g(\mathbf{x}).$$

The *dual problem* is

$$\text{Maximize } \phi(\boldsymbol{\lambda}) \text{ subject to } \boldsymbol{\lambda} \geq 0 \text{ where } \phi(\boldsymbol{\lambda}) = \min_{\mathbf{x} \in \mathbb{R}^n} L(\mathbf{x}, \boldsymbol{\lambda}). \tag{9}$$

We note that ϕ is a concave function. It is the minimum of a collection of convex (actually linear) functions of $\boldsymbol{\lambda}$ – see E3.

D1 :Linear programming. Let $f(\mathbf{x}) = \mathbf{c}^T\mathbf{x}$ and $g_i(\mathbf{x}) = -\mathbf{a}_i^T\mathbf{x} + b_i$ for $i = 1, 2, \ldots, m$. Then

$$L(\mathbf{x}, \boldsymbol{\lambda}) = \left(\mathbf{c}^T - \boldsymbol{\lambda}^T A\right)\mathbf{x} + \mathbf{b}^T\boldsymbol{\lambda} \text{ where } A \text{ has rows } \mathbf{a}_1, \ldots, \mathbf{a}_m.$$

It follows that $A\boldsymbol{\lambda} \neq \mathbf{c}$ implies that $\phi(\boldsymbol{\lambda}) = -\infty$. So the dual problem is

$$\text{Minimize } \mathbf{b}^T\boldsymbol{\lambda} \text{ subject to } A^T\boldsymbol{\lambda} = \mathbf{c}.$$

Weak Duality: If $\boldsymbol{\lambda}$ is feasible for (9) and $\mathbf{x}$ is feasible for (8) then $f(\mathbf{x}) \geq \phi(\boldsymbol{\lambda})$.

$$\phi(\boldsymbol{\lambda}) \leq L(\mathbf{x}, \boldsymbol{\lambda}) \leq f(\mathbf{x}) \text{ since } \lambda_i \geq 0, g_i(\mathbf{x}) \leq 0,\ i = 1, 2, \ldots, m. \tag{10}$$

Now note that $\phi(\boldsymbol{\lambda}) = -\infty$, unless $\mathbf{c}^T = \boldsymbol{\lambda}^T A$, since $\mathbf{x}$ is unconstrained in the definition of ϕ. And if $\mathbf{c}^T = \boldsymbol{\lambda}^T A$ then $\phi(\boldsymbol{\lambda}) = \mathbf{b}^T\boldsymbol{\lambda}$. So, the dual problem is to
Maximize $\mathbf{b}^T\boldsymbol{\lambda}$ subject to $\mathbf{c}^T = \boldsymbol{\lambda}^T A$ and $\boldsymbol{\lambda} \geq 0$, i.e. the LP dual.

Strong Duality: We give a sufficient condition *Slater's Constraint Condition* for tightness in (10).

Theorem 5.1. *Suppose that there exists a point $\mathbf{x}^*$ such that $g_i(\mathbf{x}^*) < 0, i = 1, 2, \ldots, m$. Then*

$$\max_{\boldsymbol{\lambda} \geq \mathbf{0}} \phi(\lambda) = \min_{\mathbf{x}: g_i(\mathbf{x}) \leq 0, i \in [m]} f(\mathbf{x}).$$

Proof. Let

$$\mathcal{A} = \{\mathbf{u}, t) : \exists \mathbf{x} \in \mathbb{R}^n, g_i(\mathbf{x}) \leq u_i, i = 1, 2, \ldots, m \text{ and } f(\mathbf{x}) \leq t\}.$$
$$\mathcal{B} = \left\{(0, s) \in \mathbb{R}^{m+1} : s < f^*\right\} \text{ where } f^* = \min_{\mathbf{x}: g_i(\mathbf{x}) \leq 0, i \in [m]} f(\mathbf{x}).$$

Now $\mathcal{A} \cap \mathcal{B} = \emptyset$ and so from Corollary 4.2 there exists $\boldsymbol{\lambda}, \gamma, b$ such that $(\boldsymbol{\lambda}, \gamma) \neq \mathbf{0}$ and

$$b \leq \min\left\{\boldsymbol{\lambda}^T\mathbf{u} + \gamma t : (\mathbf{u}, t) \in \mathcal{A}\right\}. \tag{11}$$
$$b \geq \max\left\{\boldsymbol{\lambda}^T\mathbf{u} + \gamma t : (\mathbf{u}, t) \in \mathcal{B}\right\}. \tag{12}$$

We deduce from (11) that $\boldsymbol{\lambda} \geq 0$ and $\bar{g} \geq 0$. If $\gamma < 0$ or $\lambda_i < 0$ for some i then the minimum in (11) is $-\infty$. We deduce from (12) that $\gamma t < b$ for all $t < f^*$ and so $\gamma f^* \leq b$. And from (11) that

$$\gamma f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) \geq b \geq \gamma f^* \qquad \text{for all } \mathbf{x} \in \mathbb{R}^n. \tag{13}$$

If $\gamma > 0$ then we can divide (13) by γ and see that $L(\mathbf{x}, \boldsymbol{\lambda}) \geq f^*$, and together with weak duality, we see that $L(\mathbf{x}, \boldsymbol{\lambda}) = f^*$.

If $\gamma = 0$ then substituting $\mathbf{x}^*$ into (13) we see that $\sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}^*) \geq 0$ which then implies that $\boldsymbol{\lambda} = 0$, contradiction. $\square$

6 Conditions for a minimum: First Order Condition

6.1 Unconstrained problem

We discuss necessary conditons for $\mathbf{a}$ to be a (local) minimum. (We are not assuming that f is convex.) We will assume that our functions are differentiable. Then Taylor's Theorem

$$f(\mathbf{a}+\mathbf{h}) = f(a) + (\nabla f(\mathbf{a}))^T\mathbf{h} + o(|\mathbf{h}|)$$

implies that

$$\nabla f(\mathbf{a}) = 0 \tag{14}$$

is a necessary condition for $\mathbf{a}$ to be a local minimum. Otherwise,

$$f(\mathbf{a} - t\nabla f(\mathbf{a})) \leq f(\mathbf{a}) - t|\nabla f(\mathbf{a})|^2/2$$

for small $t > 0$.

Of course (14) is not sufficient in general, $\mathbf{a}$ could be a local maximum. Generally spealking, one has to look at second order conditions to distinguish between local minima and local maxima.

However,

Lemma 6.1. *If f is convex then* (14) *is also a sufficient condition.*

Proof. This follows directly from F3. □

6.2 Constrained problem

We will consider Problem (8), but we will not assume convexity, only differentiability. The condition corresponding to (14) is the *Karush-Kuhn-Tucker* or KKT condition. Assume that $f, g_1, g_2, \ldots, g_m$ are differentiable. Then (subject to some *regularity conditions*, a necessary condition for $\mathbf{a}$ to be a local minimum (or maximum) to Problem (8) is that there exists $\boldsymbol{\lambda}$ such that

$$g_i(\mathbf{a}) \leq 0, \qquad 1 \leq i \leq m.$$

$$\lambda_i \geq 0 \qquad 1 \leq i \leq m. \tag{15}$$

$$\nabla f(\mathbf{a}) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\mathbf{a}) = 0. \tag{16}$$

$$\lambda_i g_i(\mathbf{a}) = 0, \qquad 1 \leq i \leq m. \qquad \text{Complementary Slackness} \tag{17}$$

The second condition says that only *active* constraints ($g_i(\mathbf{a}) = 0$) are involved in the first condition.

One deals with $g_i(\mathbf{x}) \geq 0$ via $-g_i(\mathbf{x}) \leq 0$ (and $\lambda_i \leq 0$) and $g_i(\mathbf{x}) = 0$ by $g_i(\mathbf{x}) \geq 0$ and $-g_i(\mathbf{x}) \leq 0$ (and λ_i not constrined to be non-negative or non-positive).

In the convex case, we will see that (16), (15) and (17) are sufficient for a global minimum.

6.2.1 Heuristic Justification of KKT conditions

See Diagram 10 at the end of these notes.

Suppose that $\mathbf{a}$ is a local minimum and assume w.l.o.g. that $g_i(\mathbf{a}) = 0$ for $i = 1, 2, \ldots, m$. Then (heuristically) Taylor's theorem implies that if (i) $\mathbf{h}^T \nabla g_i(\mathbf{a}) \leq 0, i = 1, 2, \ldots, m$ then (ii) we should have $\mathbf{h}^T \nabla f(\mathbf{a}) \geq 0$. (The heuristic argument is that (i) holds then we should have (iii) $\mathbf{a} + \mathbf{h}$ feasible for small $\mathbf{h}$ and then we should have (ii) since we are at a local minimum. You need a regularity condition to ensure that (ii) implies (iii).)

Applying Corollary 4.3 we see that the KKT conditions hold. We let A have *columns* $\nabla g_i(\mathbf{a}), i = 1, 2, \ldots, m$. Then the KKT conditions are $A\boldsymbol{\lambda} = -\nabla f(\mathbf{a})$.

Convex case: Suppose now that $f, g_1, \ldots, g_m$ are all convex functions and that $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ satisfies the KKT conditions. Now $\boldsymbol{\lambda}^* \geq 0$ implies that $\phi(\mathbf{x}) = L(\mathbf{x}, \boldsymbol{\lambda}^*)$ is a convex function of $\mathbf{x}$. Equation (16) and Lemma 6.1 implies that $\mathbf{x}^*$ minimises ϕ. But then for any feasible $\mathbf{x}$ we have

$$f(\mathbf{x}^*) = \phi(\mathbf{x}^*) \leq \phi(\mathbf{x}) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i^* g_i(\mathbf{x}) \leq f(\mathbf{x}).$$

For much more on this subject see Convex Optimization, by Boyd and Vendenberghe

Diagram 1

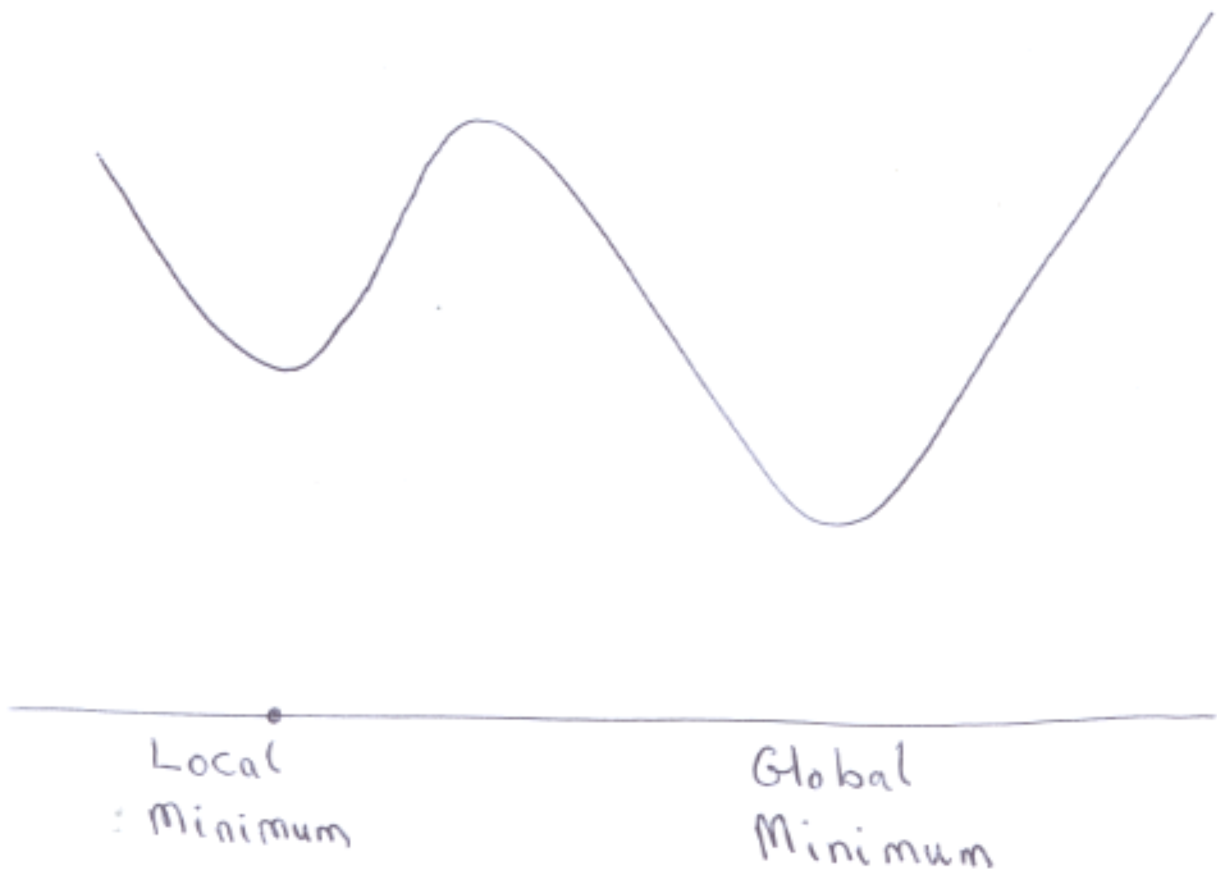
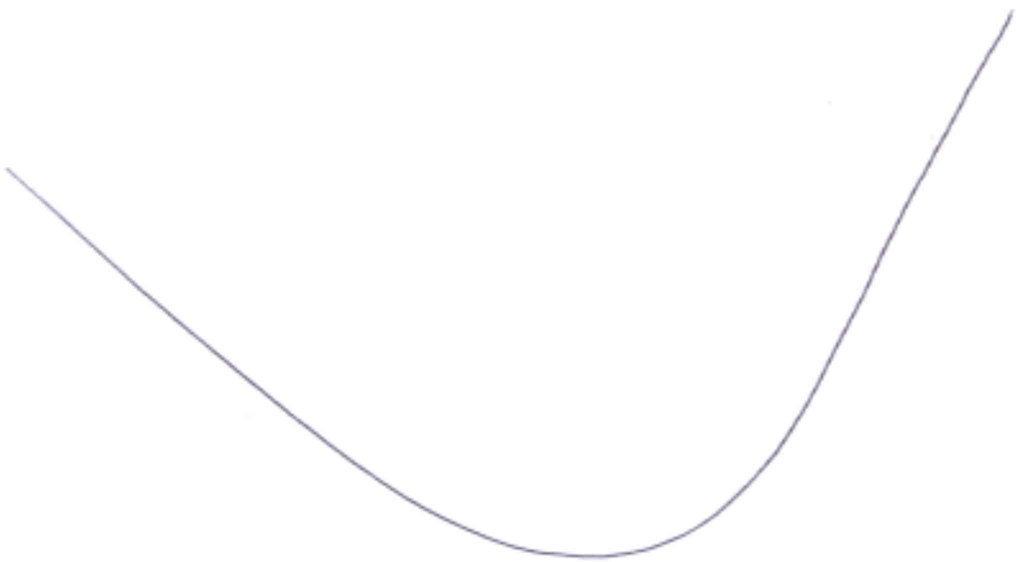
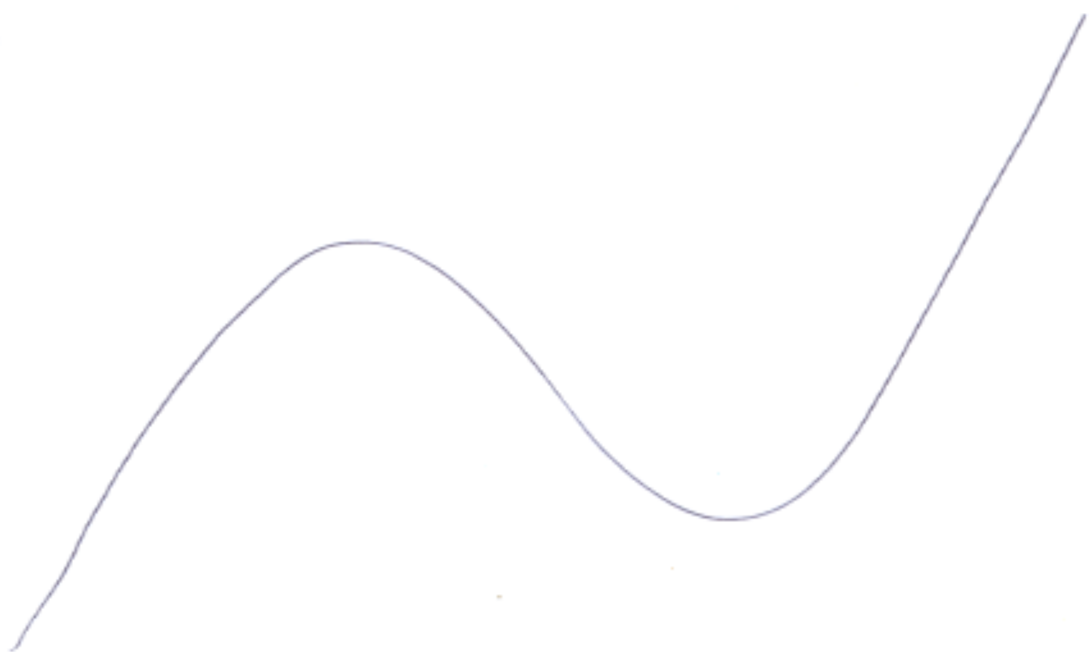


Diagram 2

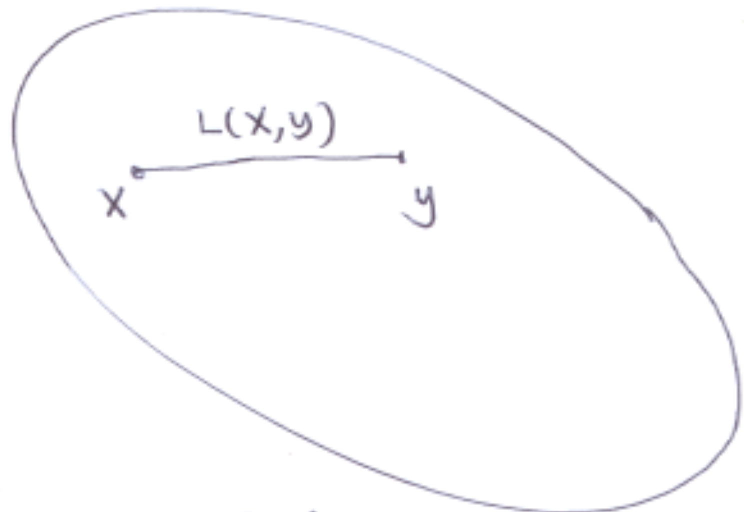


Convex Function

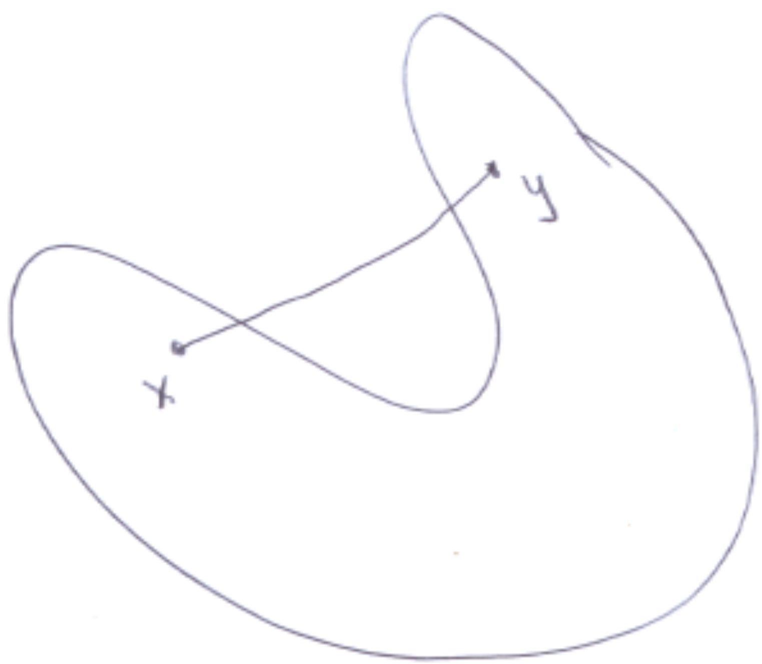


Non-Convex Function

Diagram 3



Convex Set



Non-Convex Set

Diagram 4

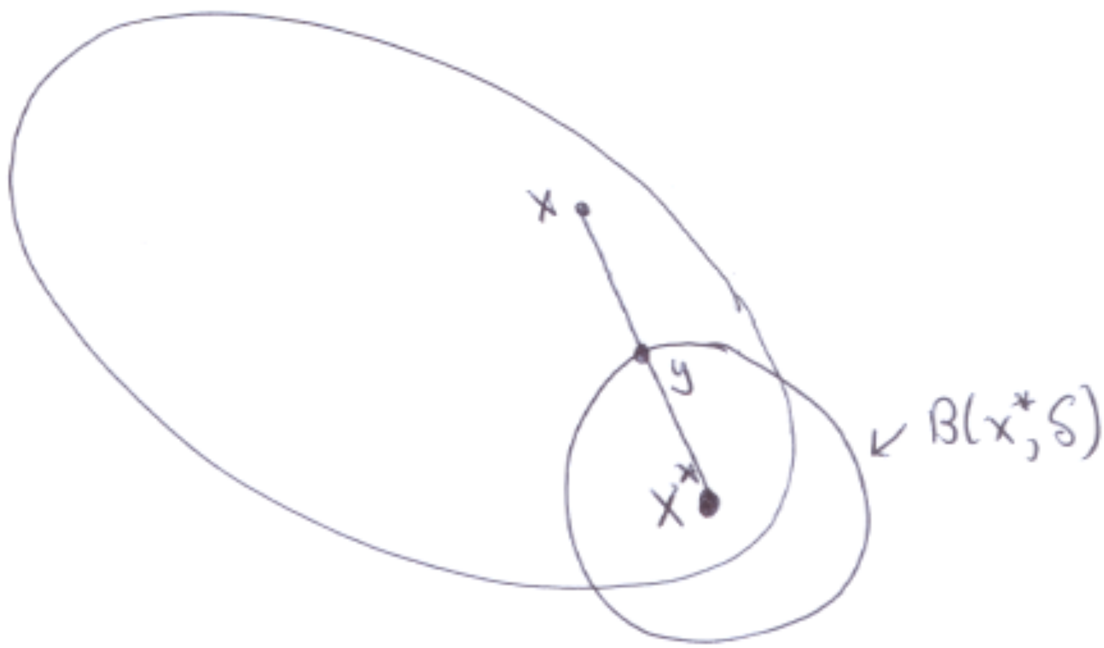


Diagram 5

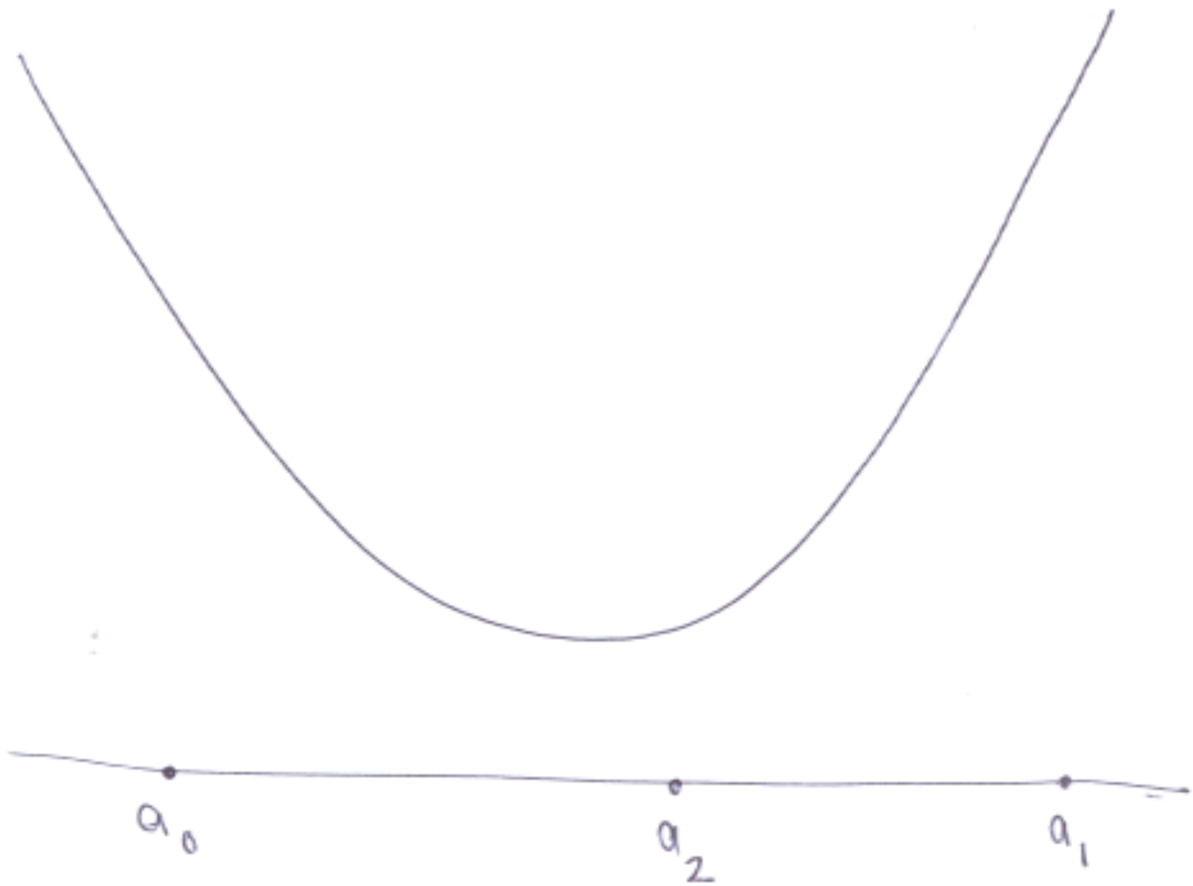


Diagram 6

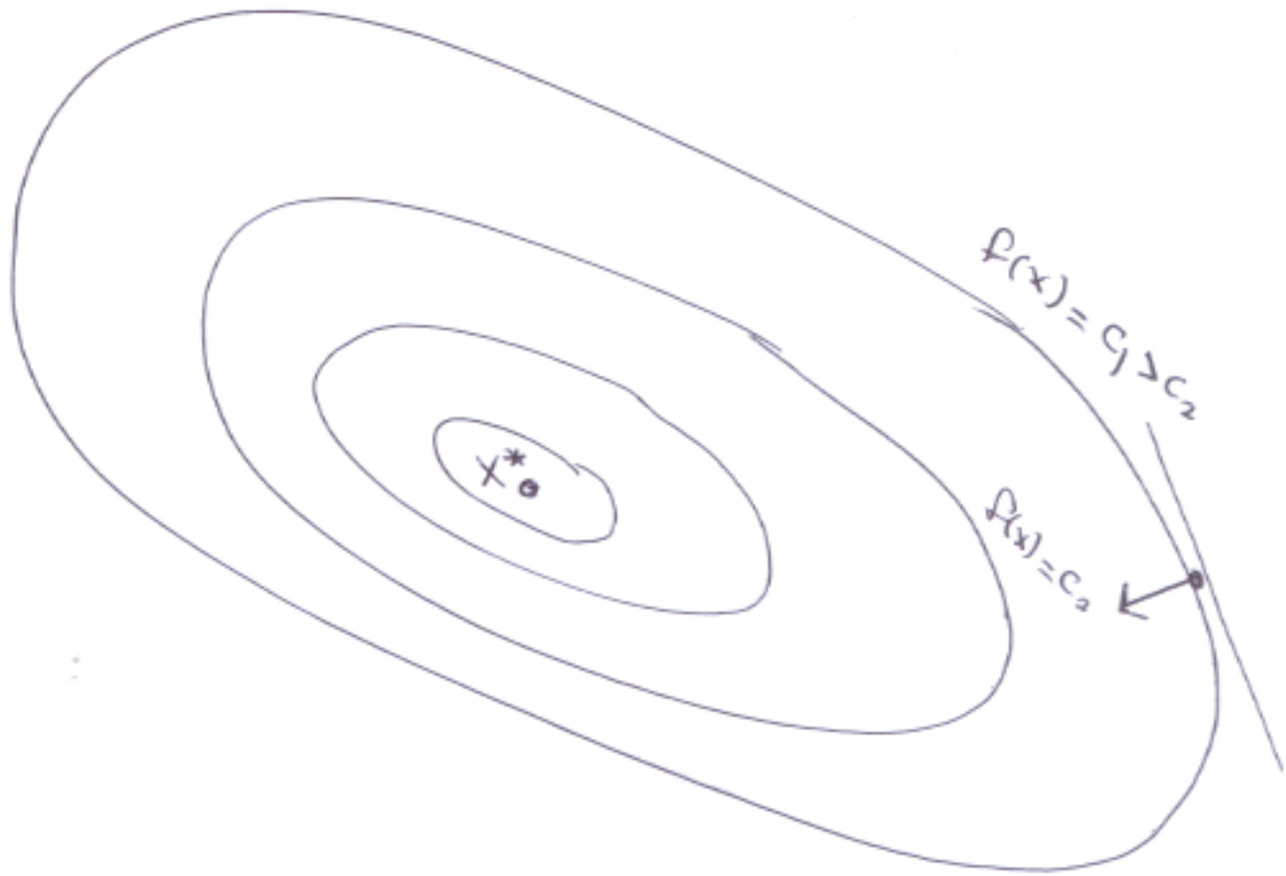


Diagram 7

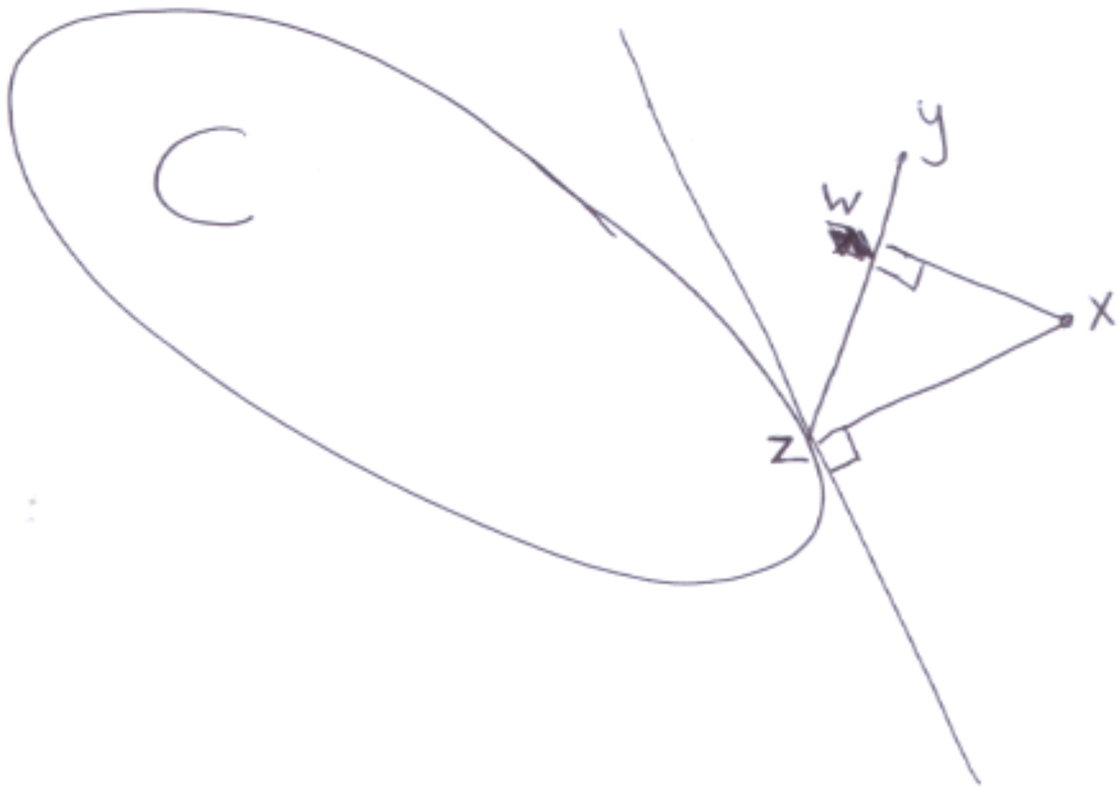


Diagram 8

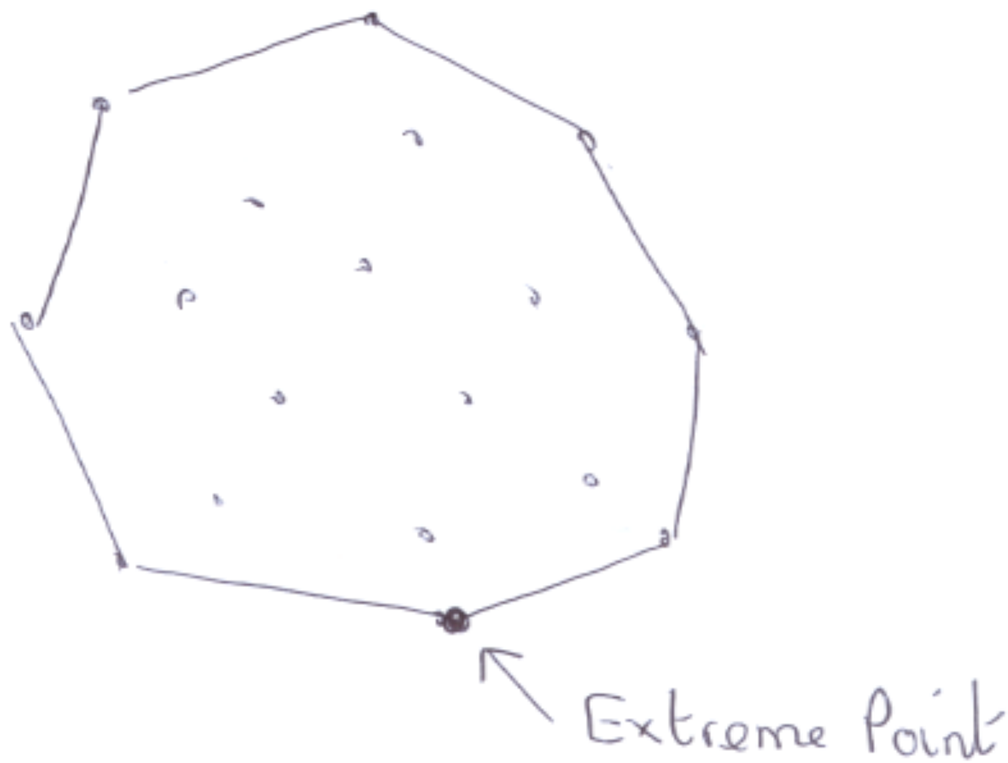


Diagram 9

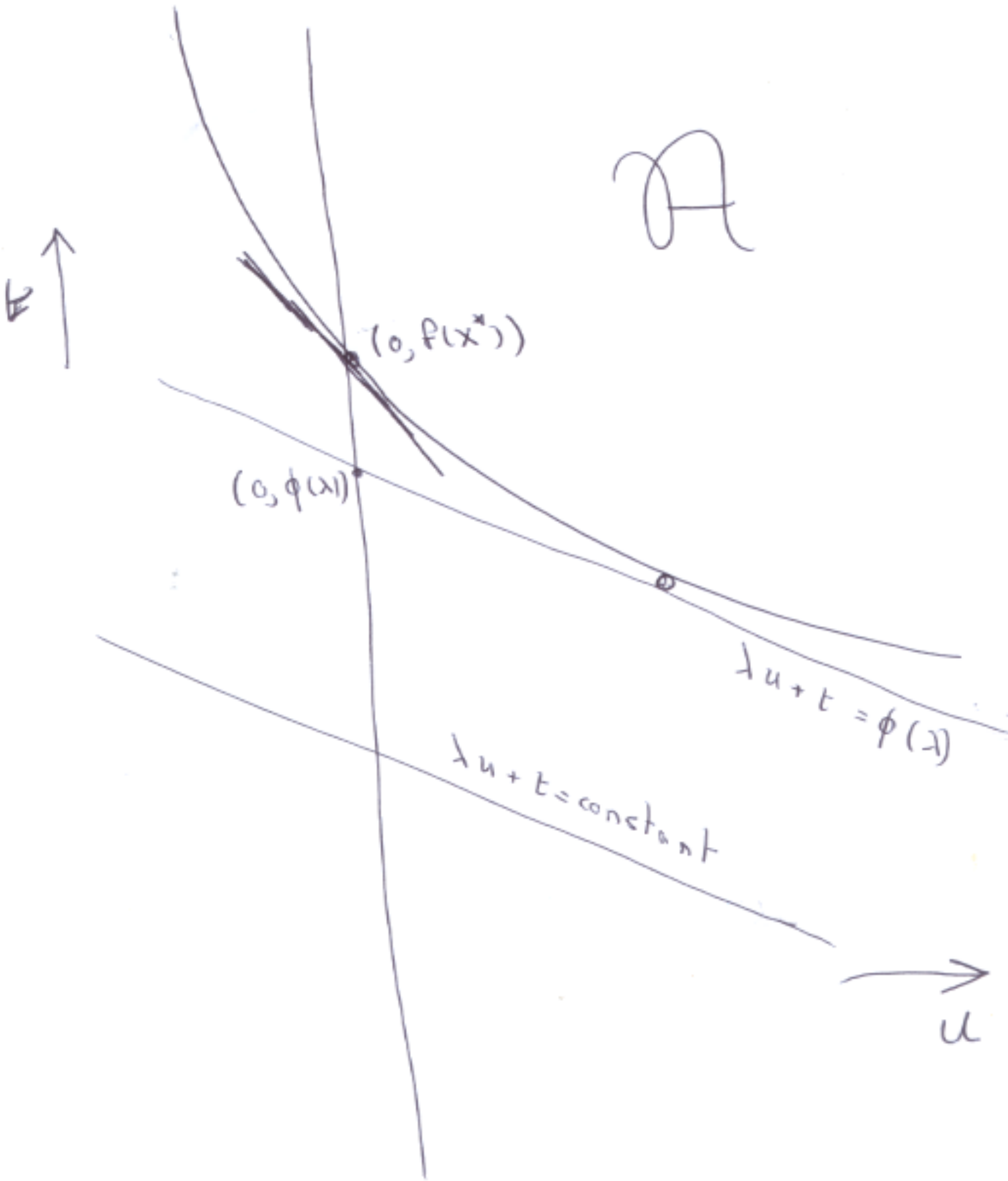


Diagram 10