Theory of Constrained Optimization. Week 1

Note Title 1/9/2022

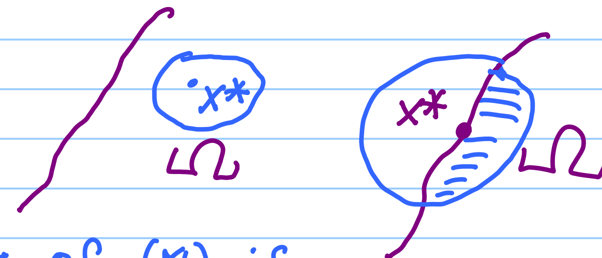
$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{s.t.} \quad \begin{cases} c_i(x) = 0, & i \in \mathcal{E} \leftarrow \text{equality constraints} \\ c_i(x) \geq 0, & i \in \mathcal{I} \leftarrow \text{inequality constraints} \end{cases}$$

$\uparrow$ objective function

$(\ast)$

Feasible set $\Omega := \{x \in \mathbb{R}^n \,;\, c_i(x) = 0,\ i \in \mathcal{E} \,;\, c_i(x) \geq 0,\ i \in \mathcal{I}\}$

$(\ast)$ can be rewritten as $\displaystyle\min_{x \in \Omega} f(x)$



<u>Def</u>: $x^* \in \Omega$ is a local solution/minimizer of $(\ast)$ if
$\exists$ $\underbrace{\text{a neighborhood } N \text{ of } x^*}_{\equiv \text{ an open ball around } x^*}$ s.t. $f(x) \geq f(x^*)$ $\forall x \in N \cap \Omega$.

["local solution" $\rightsquigarrow$ "strict local solution" when "$\geq$" $\rightsquigarrow$ "$>$"]

For simplicity, assume the objective function f and all constraint functions c_i, $i \in \mathcal{E} \cup \mathcal{I}$, are C^2 smooth.

For <u>unconstrained</u> optimization problem (i.e. $\Omega = \mathbb{R}^n$),

Necessary conditions: Local unconstrained minimizers have $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ positive semidefinite.

Sufficient conditions: Any point x^* at which $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive definite is a strong local minimizer of f.

$\underset{\text{strict}}{\overset{\text{strong}}{\parallel}}$

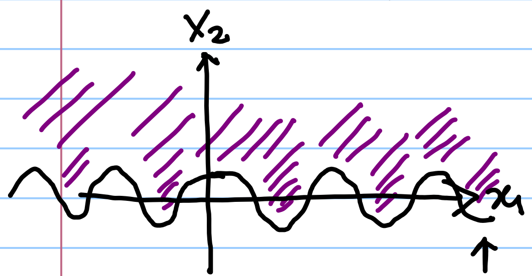
[Please study the proofs, they use the quadratic Taylor approximation in tandem with the spectral theorem.]

[Some subtle details will be found in the HW.]

Example: $\min\ (x_2+100)^2 + 0.01\, x_1^2 = [x_1,\ x_2-(-100)] \begin{bmatrix} 1/100 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2-(-100) \end{bmatrix}$

s.t. $x_2 - \cos x_1 \geq 0$.

a strictly convex quadratic, with a unique global minimizer at $(0,-100)$.



Ω is clearly not a convex set.

Solution:
local minimizers at the boundary of Ω, more specifically, at

$(x_1,x_2) = (k\pi, \underbrace{\cos(k\pi)}_{=-1})$, $k=\pm1,\pm3,\pm5,\cdots$

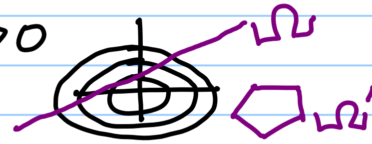
global minimizers at

$(x_1,x_2) = (\pm\pi, -1)$

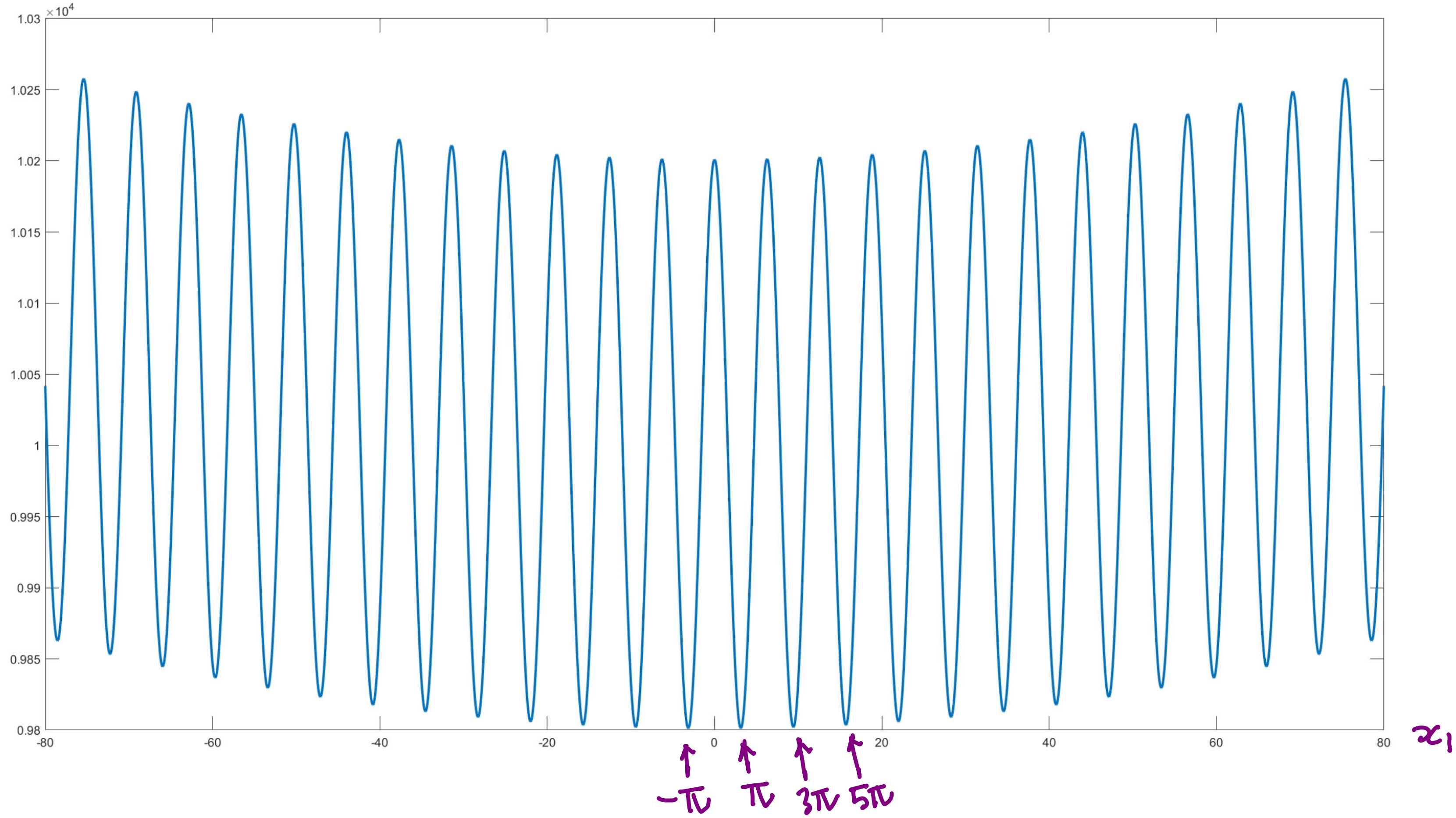
$$\boxed{\begin{array}{l} \text{Aside: If } f:\mathbb{R}^n \to \mathbb{R} \text{ is a convex function} \\ \qquad \Omega \subseteq \mathbb{R}^n \text{ is a convex set} \\ \quad \text{then} \\ \qquad f:\Omega \to \mathbb{R} \text{ is also convex.} \\ \text{In this case, the solution set of } \min_{x\in\Omega} f(x) \\ \text{is also a convex set.} \end{array}}$$

e.g. $f(x_1,x_2) = \lambda_1 x_1^2 + \lambda_2 x_2^2$, $\lambda_1, \lambda_2 > 0$

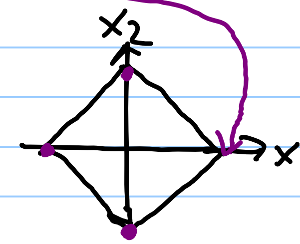
Ω = a line $\leftarrow$ convex



Plot of $f(x_1, \cos(x_1))$, $x_1 \in [-80, 80]$



Note: That we assume $c_i(x)$ to be C^2 smooth does not mean the boundary of Ω cannot have "Kinks".



E.g. $\Omega = \{x \in \mathbb{R}^2 : \|x\|_1 = |x_1| + |x_2| \leq 1\}$

can be rewritten as

$$\{x \in \mathbb{R}^2 : \ x_1 + x_2 \leq 1, \ x_1 - x_2 \leq 1, \ -x_1 + x_2 \leq 1, \ -x_1 - x_2 \leq 1\}$$

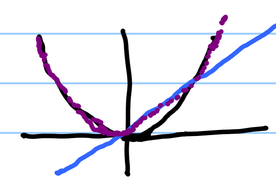
$$\parallel$$

$$\{x \in \mathbb{R}^2 : \ \underbrace{1 - x_1 - x_2 \geq 0}_{c_1(x)}, \ \underbrace{1 - x_1 + x_2 \geq 0}_{c_2(x)}, \ \underbrace{1 + x_1 - x_2 \geq 0}_{c_3(x)}, \ \underbrace{1 + x_1 + x_2 \geq 0}_{c_4(x)}\}$$

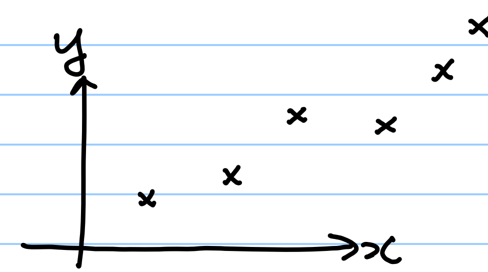
more interestingly, nonsmooth, unconstrained optimization problems can sometimes be reformulated as smooth constrained problems.

Here is a toy example: $\min_{x \in \mathbb{R}} f(x) = \max(x^2, x)$ ← non-smooth

can be recast as $\min_{t \in \mathbb{R}} t$ s.t. $t \geq x$, $t \geq x^2$



A more serious example : robust L^1 regression



A basic machine learning problem :

> Given data (x_i, y_i) $i = 1, \cdots, m$,
> find a function $y = f(x)$ that
> "best explains the data".

i.e. Find f s.t. $y_i \approx f(x_i)$ $\forall i$.

But this is meaningless if allow f to be any function. (why?)

Quite often, one may restrict attention to

$$f(x) = \alpha_0 + \alpha_1 x \qquad \text{or} \qquad f(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2$$

The least square method :

find $\alpha_0, \alpha_1, \alpha_2$ so that $\sum_{i=1}^{m} (\alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 - y_i)^2$ is minimized.

This is not only a smooth, unconstrained optimization problem in 3 variables, but also have a simple closed-form solution because the objective is a quadratic. Note:

$$\sum_{i=1}^{m} (\alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 - y_i)^2 = \Bigg\| \underbrace{\begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ & \vdots & \\ 1 & x_m & x_m^2 \end{bmatrix}}_{=A} \underbrace{\begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{bmatrix}}_{=\alpha} - \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}}_{=b} \Bigg\|_2^2$$

$$= \| A\alpha - b \|_2^2$$

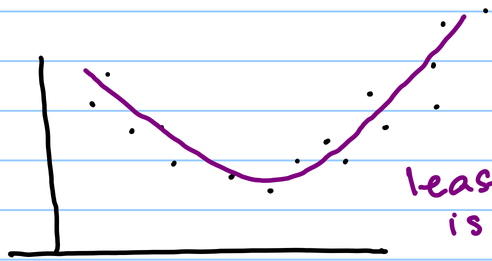
$$= \alpha^T \underbrace{A^T A}_{\text{positive definite (as long as rank}(A)=3)} \alpha - 2 b^T A \alpha + b^T b$$

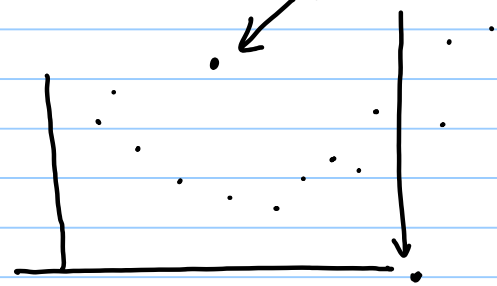
positive definite (as long as rank(A)=3)

$$\boxed{\alpha_{min} = (A^T A)^{-1} (A^T b).}$$

The celebrated least square method is widely used. The problem is that when just a few of the data values y_i are very erronous, the "square" may become very unforgiving.

Idea: change $\hat{}2$ to $|\cdot|$.





Least L^1 regression solves:

$$\min_{\alpha_0,\alpha_1,\alpha_2} \sum_{i=1}^{m} \left| \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 - y_i \right|$$

Now, an ingenious trick: the unconstrained, but nonsmooth optimization problem above is equivalent to:

$$\min_{\substack{\alpha_0,\alpha_1,\alpha_2\\ \gamma_1,\dots,\gamma_m}} \gamma_1 + \dots + \gamma_m \quad \text{s.t.} \quad -\gamma_i \le \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 - y_i \le \gamma_i \quad i=1,\dots,m$$

$$\gamma_1,\dots,\gamma_m \ge 0$$

Note: all the objective and constraint functions cannot possibly be any smoother, they are LINEAR functions!

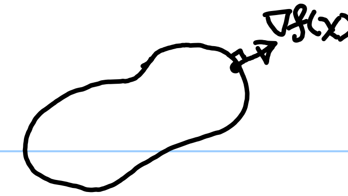
The latter constraint optimization problem is an example of LINEAR PROGRAM (LP).

Definition:

The active set $\mathcal{A}(x)$ *at any feasible* x *consists of the equality constraint indices from* $\mathcal{E}$ *together with the indices of the inequality constraints* i *for which* $c_i(x) = 0$; *that is,*

$$\mathcal{A}(x) = \mathcal{E} \cup \{i \in \mathcal{I} \mid c_i(x) = 0\}.$$

At a feasible point x, the inequality constraint $i \in \mathcal{I}$ is said to be *active* if $c_i(x) = 0$ and *inactive* if the strict inequality $c_i(x) > 0$ is satisfied.



You may recall from your multivariate calculus course that:

$$\boxed{\nabla f(x) \text{ is orthogonal to the level surface } \{y \in \mathbb{R}^n : f(y) = f(x)\} \text{ at } x.}$$

or curve?

$$\{y \in \mathbb{R}^n : f(y) = f(x)\} = f^{-1}(f(x))$$

If you find this confusing, it is because it is!

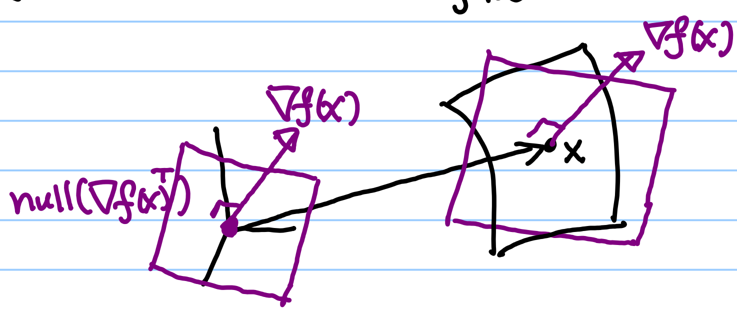
Recall:

$$\underbrace{f(y)}_{\text{nonlinear}} \approx \underbrace{f(x) + \nabla f(x)^T (y-x)}_{\text{linear}} \quad , \quad y \approx x$$

Expect:

$$\underbrace{f^{-1}(f(x))}_{\text{curved}} \approx \underbrace{\{y : \nabla f(x)^T (y-x) = 0\}}_{\text{flat}} = x + \underbrace{\text{null}(\nabla f(x)^T)}_{\text{"hyperplane"}} \quad \text{near } x$$

$$\begin{cases} (n-1)\text{-dimensional} & \text{if } \nabla f(x) \neq \vec{0} \\ n\text{-dimensional} & \text{if } \nabla f(x) = \vec{0} \end{cases}$$



The implicit function theorem tells us that the level set $f^{-1}(f(x))$, in a neighborhood of x, is indeed a C^1 hypersurface (i.e. the graph of a C^1 function $h : \underset{\text{open}}{\mathcal{U}} \subset \mathbb{R}^{n-1} \to \mathbb{R}$) when $\nabla f(x) \neq \vec{0}$.

In this case, the hyperplane $x + \text{null}(\nabla f(x)^T)$ is tangent to the hypersurface $f^{-1}(f(x))$.

Also $\nabla f(x) \perp$ the tangent plane.

What if $\nabla f(x) = 0$? The local linear approximation loses its power in telling you how $f^{-1}(f(x))$ looks like near x!

Consider $f : \mathbb{R}^n \to \mathbb{R}^1$ $\quad f(x) = \sum_{i=1}^{n} x_i^2$, $\quad f^{-1}(\overset{\overset{0}{\shortparallel}}{f(\vec{0})}) = \{\vec{0}\}$ ← 0-dimension

Consider $f : \mathbb{R}^n \to \mathbb{R}^1$ $\quad f(x) = 0$, $\quad f^{-1}(f(\text{any pt})) = \mathbb{R}^n$ ← n dim.

In each case, the level set is not a hypersurface!

$\nabla f(x) = \vec{0}$.

more generally,
Consider $f: \mathbb{R}^n \to \mathbb{R}^1$ $\quad f(x) = \sum_{i=1}^{k} x_i^2 = [x_1, \dots, x_n] \begin{bmatrix} \overbrace{1 \ddots 1}^{k} & \\ & 0 \ddots 0 \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$

$$f^{-1}(f(\vec{0})) = \left\{ \begin{bmatrix} 0 \\ \vdots \\ 0 \\ x_{k+1} \\ \vdots \\ x_n \end{bmatrix} : x_{k+1}, \dots, x_n \in \mathbb{R} \right\} \leftarrow (n-k) \text{ dimensional}$$

This simple example tells us that when $\nabla f(x) = \vec{0}$, the level set $f^{-1}(f(x))$ can have any dimension from 0 to n!

Never forget that
$$\boxed{\begin{aligned} &f'(x, p) \\ &\quad \parallel \\ &\frac{d}{dt} f(x + tp)\Big|_{t=0} = \nabla f(x)^T p \end{aligned}}$$
$\leftarrow$ directional derivative of f at x in the direction p

If $\nabla f(x)^T p > 0$, $\quad f(x + tp) > f(x)$ for small $t > 0$ (p is an ascent dir.)

$\nabla f(x)^T p < 0$, $\quad f(x + tp) < f(x)$ for small $t < 0$

$\nabla f(x)^T p = 0$, $\quad f(x + tp) \approx f(x)$ for small t (p is a descent dir.)

Half-spaces :

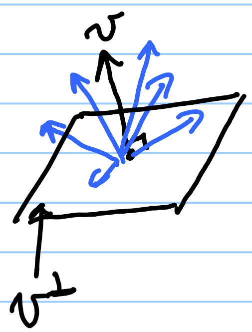
If $v \in \mathbb{R}^n$, $v \neq \vec{0}$

$$v^{\perp} = \{ d \in \mathbb{R}^n : v^T d = 0 \}$$ ← (n-1) dimesional, a hyperplane

$\{ d \in \mathbb{R}^n : v^T d \geq 0 \}$ ← a closed half space

||

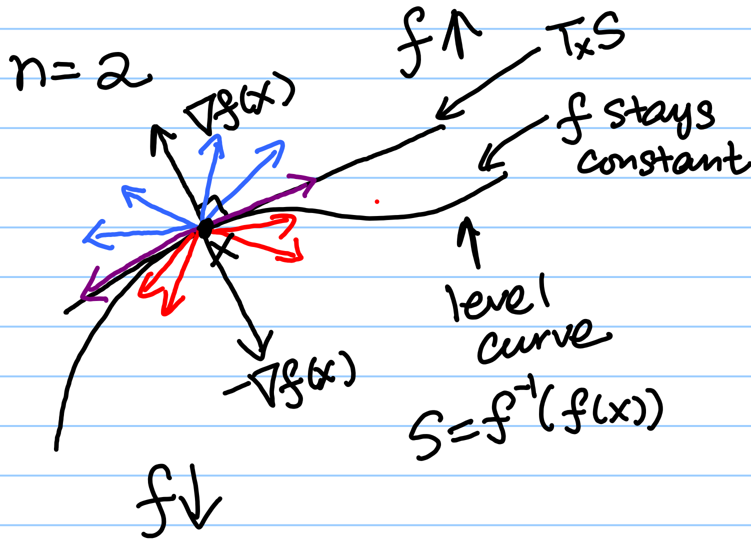
all the points on the same side of $v^{\perp}$ as v, including points in $v^{\perp}$.

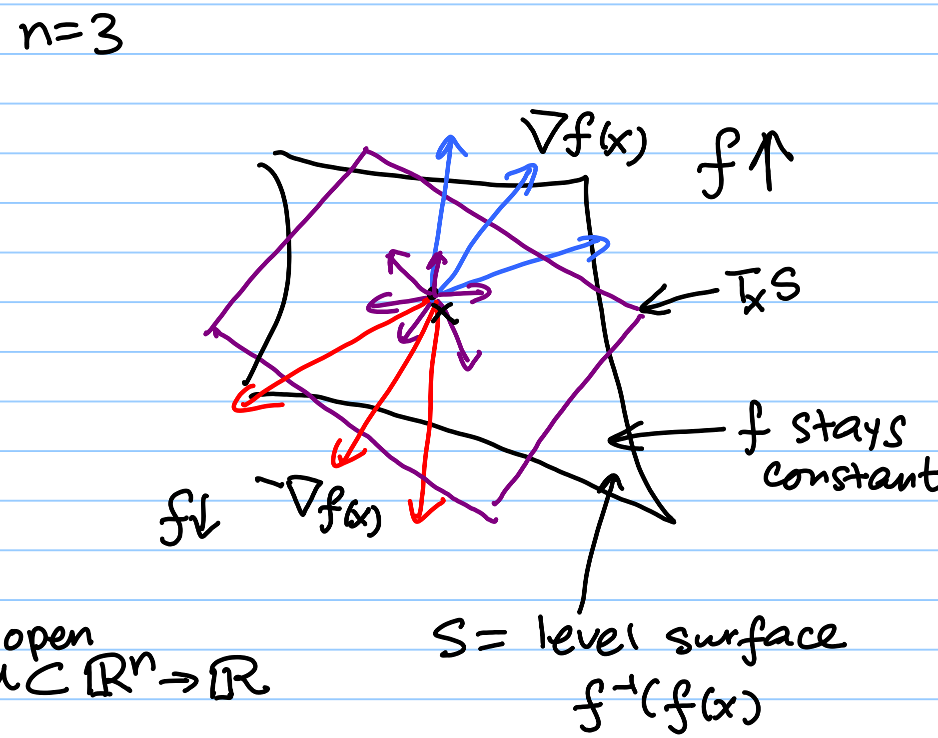


Similarly, $\{ d \in \mathbb{R}^n : v^T d > 0 \}$ ← an open half space

||

all the points on the same side of $v^{\perp}$ as v, excluding points in $v^{\perp}$.





Local picture of $f : \mathcal{U} \overset{\text{open}}{\subset} \mathbb{R}^n \to \mathbb{R}$

@ $x \in \mathcal{U}$

with $\nabla f(x) \neq \vec{0}$.

This is why if x is a local mimimizer or maximizer of f, then it is necessary that $\nabla f(x) = \vec{0}$, for otherwise $\nabla f(x) \neq \vec{0}$ and we can choose

$$p = \nabla f(x)$$

to make

$$\nabla f(x)^T p = \nabla f(x)^T \nabla f(x) = \|\nabla f(x)\|_2^2 > 0,$$

meaning that in the $+\nabla f(x)$ direction f increases.

Similarly, in the $-\nabla f(x)$ direction, f decreases.

To conclude: $\boxed{\begin{array}{l} \nabla f(x) \neq \vec{0} \implies x \text{ cannot be a local min or max.} \\ \text{(Fermat's theorem is just the contrapositive of the above.)} \end{array}}$

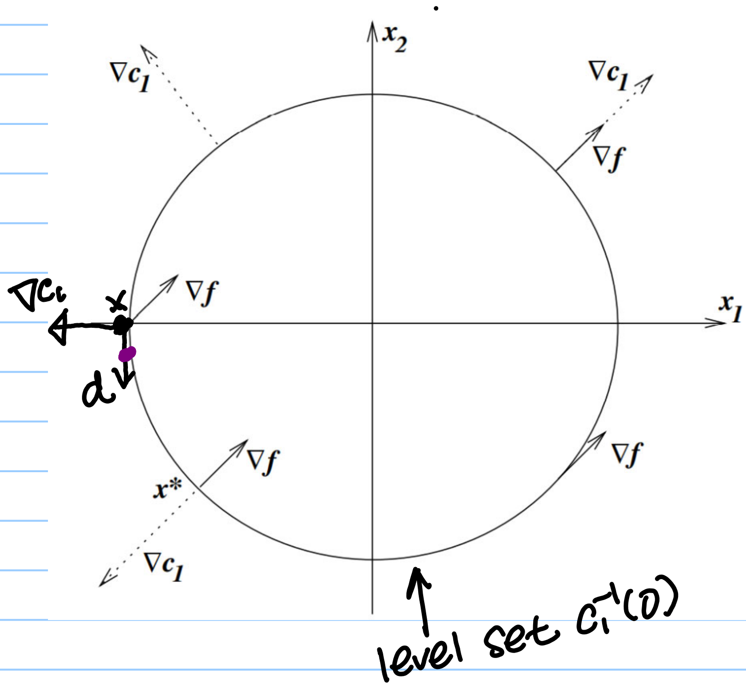
The fact that $\frac{d}{dt} f(x - t\nabla f(x))\Big|_{t=0} = -\nabla f(x)^T \cdot \nabla f(x) < 0$ when $\nabla f(x) \neq \emptyset$ is the foundation of the gradient descent method for unconstrained optimization.

What if there are constraints?

Example:

$$\min \underbrace{x_1 + x_2}_{f(x)} \quad \text{s.t.} \quad \underbrace{x_1^2 + x_2^2 - 2}_{c_1(x)} = 0$$

$$\nabla f(x) = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \neq \vec{0}, \quad \nabla c_1(x) = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix} = 2x \neq \vec{0}$$



Q: If x is a solution, what condition must it satisfy?

Note:

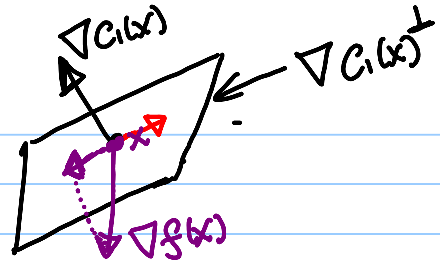
If $\exists\, d \in \mathbb{R}^2$ s.t. $\nabla c_1(x)^T d = 0$ ← means $c_1(x + \varepsilon d) \approx c_1(x)$

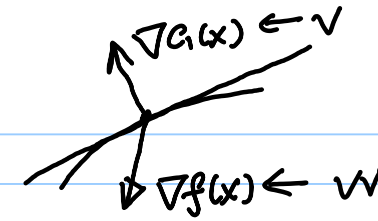
and

$\nabla f(x)^T d < 0$ ← means $f(x + \varepsilon d) < f(x)$

then we can decrease the value of f by moving to a nearby point within $c^{-1}(c(x))$.

If x is a solution, it should be that $\underbrace{\nexists\, d \text{ s.t. } \nabla c_1(x)^T d = 0 \text{ \& } \nabla f(x)^T d < 0}_{\text{inconvenient to work with......}}$



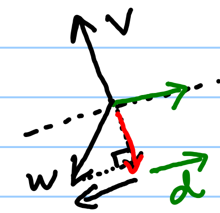


Note: $\nexists\, d$ s.t. $v^T d = 0$ and $w^T d < 0$ $\quad$ (Assume $v \neq \vec{0}$)

$\iff \exists\, \lambda \in \mathbb{R}$ s.t. $w = \lambda v$.

($\Leftarrow$) If $w = \lambda v$, then $v^T d = 0 \Rightarrow w^T d = \lambda v^T d = 0$

($\Rightarrow$) If v and w are not parallel,



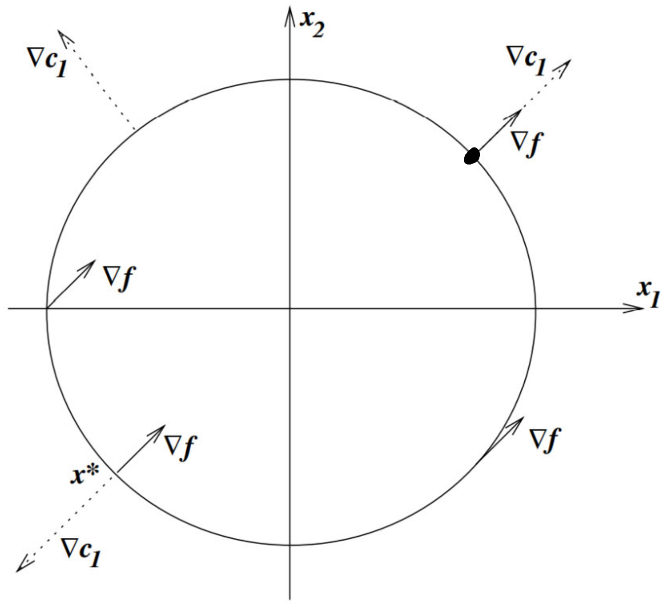
let $d = -\left(w - \dfrac{w^T v}{v^T v} \cdot v\right) \neq \vec{0}$

and $v^T d = 0$, $w^T d < 0$. ∎

Conclusion: If x^* is a solution, it should be that $\exists\, \lambda_1^* \in \mathbb{R}$

$$\nabla f(x^*) = \lambda_1^* \nabla c_1(x^*)$$

Lagrange multiplier



$\begin{bmatrix}1\\1\end{bmatrix} = \lambda \cdot 2\begin{bmatrix}x_1\\x_2\end{bmatrix}$ and $c_1(x)=0$ can only be satisfied @ $x=\begin{bmatrix}1\\1\end{bmatrix}$ or $\begin{bmatrix}-1\\-1\end{bmatrix}$

$\lambda=\frac{1}{2}$ $\quad$ $\lambda=-\frac{1}{2}$

It is clear that the only minimizer is

$$x^*=\begin{bmatrix}-1\\-1\end{bmatrix}. \quad (\lambda^*=-\tfrac{1}{2}.)$$

By introducing the Lagragian function

$$\mathcal{L}(x,\lambda_1)=f(x)-\lambda_1 c_1(x).$$

Noting that $\quad \nabla_x \mathcal{L}(x,\lambda_1)=\nabla f(x)-\lambda_1\nabla c_1(x)$

$$\frac{\partial}{\partial\lambda_1}\mathcal{L}(x,\lambda_1)=-c_1(x)$$

(n+1)(nonlinear) equations in (n+1) variables

The necessary condition can be written as

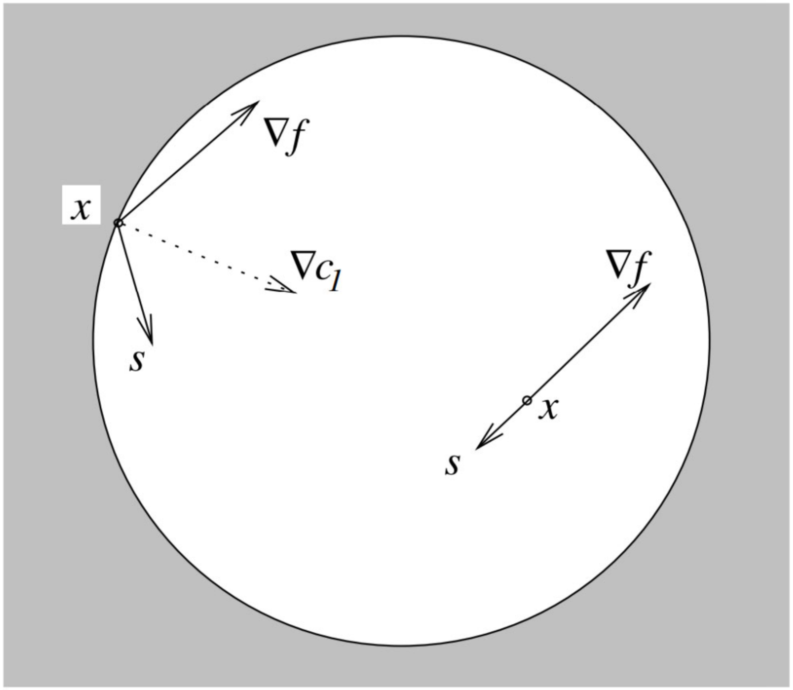
$$\boxed{\nabla_x\mathcal{L}(x,\lambda_1)=\vec{0}\ ,\ c_1(x)=0}$$

Notice that if we change the constraint $x_1^2+x_2^2-2=0$ to $\overbrace{2-x_1^2-x_2^2}^{\text{new } c_1}=0$, the problem does not change, but

$$\nabla c_1(x) = -2\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad x^* = \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \quad \lambda_1^* = 1/2 .$$

(sign flipped) (same as before) (sign flipped)

Now, let's modify the example to $\min \underbrace{x_1+x_2}_{f(x)}$ s.t. $\underbrace{2-x_1^2-x_2^2}_{c_1(x)} \geq 0$.



What is a necessary condition for x to be a solution?

(I) If $c_1(x) > 0$ and x is a solution, then it is necessary that

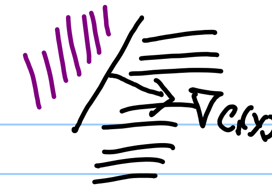
$$\nabla f(x) = \vec{0}. \quad \text{(why?)}$$

(II) If $c_1(x) = 0$ and x is a solution, then it is necessary that

$$\nexists\, d \in \mathbb{R}^n \text{ s.t. } \nabla c_1(x)^T d \geq 0 \text{ and } \nabla f(x)^T d < 0$$

(why?)

It should be evident that:



$$\{d : \nabla c_1(x)^T d \geq 0\} \cap \{d : \nabla f(x)^T d < 0\} = \emptyset$$

$\{d : \nabla c_1(x)^T d \geq 0\}$ ← a closed half space

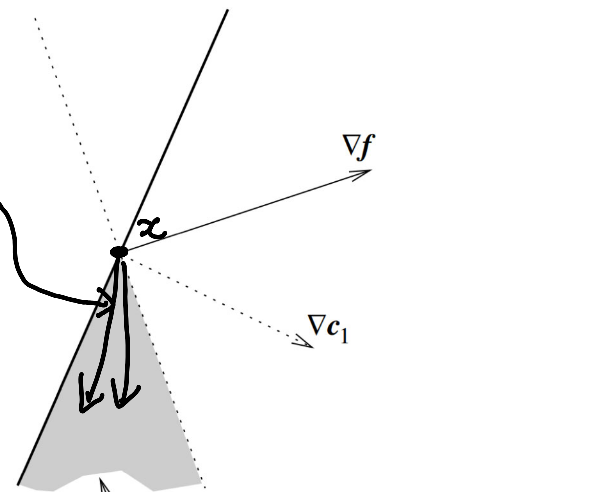
$\{d : \nabla f(x)^T d < 0\}$ ← an open half space

$\Longleftrightarrow$ $\nabla f(x)$ and $\nabla c_1(x)$ points in the same direction

i.e. $\nabla f(x) = \lambda_1 \nabla c_1(x)$ for some $\lambda_1 \geq 0$

If $\nabla f(x)$ and $\nabla c_1(x)$ do not point in the same direction, then

$$\{d : \nabla c_1(x)^T d \geq 0\} \cap \{d : \nabla f(x)^T d < 0\} \neq \emptyset \text{ is a cone.}$$



Any d in this cone is a good search direction, to first order

For any d in this cone,

$$c_1(x + \varepsilon d) > 0$$

$$f(x + \varepsilon d) < f(x),$$

meaning that x cannot be a solution.

To conclude: A necessary condition for optimality is

$$c_1(x^*) > 0 \;\&\; \nabla f(x^*) = \vec{0} \quad \text{or} \quad c_1(x^*) = 0 \;\&\; \nabla f(x^*) = \lambda_1^* \nabla c_1(x^*)$$
$$\lambda_1^* \geq 0$$

Solution to the problem:

$x^* = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$, $\lambda_1^* = \frac{1}{2}$. the inequality constraint is active at this solution. (i.e. $c_1(x^*) = 0$)

(there is no solution that satisfies $c_1(x^*) > 0$ as $\nabla f(x^*) = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \neq \vec{0}$)

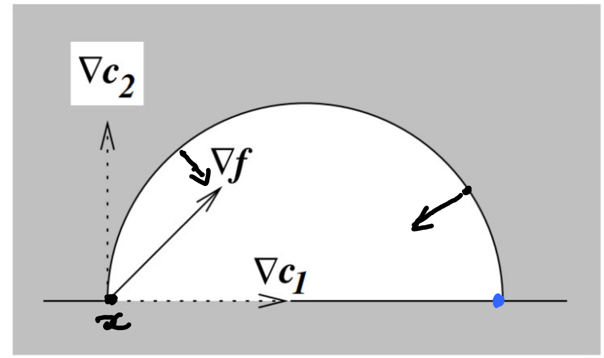
Finally, there is a cute way to rewrite the optimality condition above:

$$\boxed{\begin{gathered} \nabla_x \mathcal{L}(x^*, \lambda_1^*) = 0 \quad \text{for some } \lambda_1^* \geq 0 \\ \text{and} \\ \lambda_1^* c_1(x^*) = 0. \end{gathered}}$$

called a complimentarity condition

λ_1^* can only be strictly positive when the corresponding constraint is active.

Let's add one more constraint:



$$\min \underbrace{x_1+x_2}_{=f(x)} \quad \text{s.t.} \quad \underbrace{2-x_1^2-x_2^2}_{=c_1(x)} \geq 0, \quad \underbrace{x_2}_{=c_2(x)} \geq 0$$

Just like before, if x is a solution, then

(I) If $c_1(x)>0$, $c_2(x)>0$, then $\nabla f(x)=\vec{0}$ ← impossible since $\nabla f(x)=\begin{bmatrix}1\\1\end{bmatrix}$

(II) If $c_1(x)=0$, $c_2(x)>0$, then
$\nexists\, d$ s.t. $\nabla c_1(x)^T d \geq 0$ and $\nabla f(x)^T d<0$ $\iff$ $\overset{=\begin{bmatrix}1\\1\end{bmatrix}}{\nabla f(x)} = \lambda_1 \overset{=-2\begin{bmatrix}x_1\\x_2\end{bmatrix}}{\nabla c_1(x)}$, $\lambda_1 \geq 0$
impossible

(III) If $c_1(x)>0$, $c_2(x)=0$, then
$\nexists\, d$ s.t. $\nabla c_2(x)^T d \geq 0$ and $\nabla f(x)^T d<0$ $\iff$ $\overset{=\begin{bmatrix}1\\1\end{bmatrix}}{\nabla f(x)} = \lambda_2 \overset{=\begin{bmatrix}0\\1\end{bmatrix}}{\nabla c_2(x)}$
impossible

(IV) If $c_1(x)=0$, $c_2(x)=0$, then
$\nexists\, d$ s.t. $\nabla c_1(x)^T d \geq 0$, $\nabla c_2(x)^T d \geq 0$ and $\nabla f(x)^T d<0$.

$c_1(x)=0$, $c_2(x)=0$ $\iff$ $x=\begin{bmatrix}-\sqrt{2}\\0\end{bmatrix}$ or $x=\begin{bmatrix}\sqrt{2}\\0\end{bmatrix}$

$\nexists\, d$ s.t. $\nabla c_1(x)^T d \geq 0$, $\nabla c_2(x)^T d \geq 0$ and $\nabla f(x)^T d < 0$.

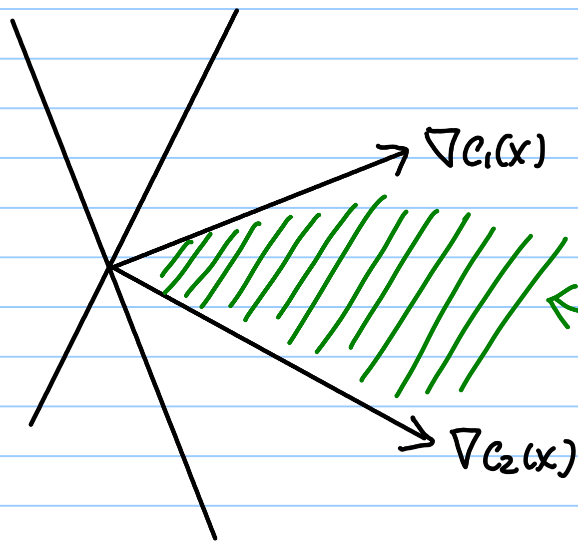
$$\Longleftrightarrow$$

$$(\star) \quad \{d : \nabla c_1(x)^T d \geq 0\} \cap \{d : \nabla c_2(x)^T d \geq 0\} \cap \{d : \nabla f(x)^T d < 0\} = \emptyset$$

$$\Longleftrightarrow$$

can you convince yourself why it should be true from the figure below?

$$\boxed{\nabla f(x) = \lambda_1 \nabla c_1(x) + \lambda_2 \nabla c_2(x) \text{ for some } \lambda_1, \lambda_2 \geq 0}$$



$\{\lambda_1 \nabla c_1(x) + \lambda_2 \nabla c_2(x) : \lambda_1, \lambda_2 \geq 0\}$

$\nabla f(x)$ is in this cone exactly when the optimality condition $(\star)$ is satisfied.

Armed with this condition, we can see that $x^* = \begin{bmatrix} -\sqrt{2} \\ 0 \end{bmatrix}$ is the only minimizer.

[$\nabla f(x^*) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $\nabla c_1(x^*) = \begin{bmatrix} 2\sqrt{2} \\ 0 \end{bmatrix}$, $\nabla c_2(x^*) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, $\nabla f(x^*) = \frac{1}{2\sqrt{2}} \nabla c_1(x^*) + 1 \cdot \nabla c_2(x^*)$.]

Exercise: Show $\begin{bmatrix} \sqrt{2} \\ 0 \end{bmatrix}$ does not satisfy the optimality condition.

more generally, given $b_1, \cdots, b_m \in \mathbb{R}^n$, $g \in \mathbb{R}^n$

consider the cone $K = \{ \lambda_1 b_1 + \cdots + \lambda_m b_m : \lambda_i \geq 0 \}$

Theorem: $\bigcap_{i=1}^{m} \{d : b_i^T d \geq 0\} \cap \{d : g^T d \geq 0\} = \emptyset$

$\Longleftrightarrow$ $g \in K$

($\Leftarrow$) If $g = \sum \lambda_i b_i$, $\lambda_i \geq 0$

and $\exists d$ st. $b_i^T d \geq 0$ $\forall i$

then $g^T d = \sum \underbrace{\lambda_i}_{\geq 0} \underbrace{b_i^T d}_{\geq 0} \geq 0$. So $\bigcap_i \{d : b_i^T d \geq 0\} \cap \{d : g^T d \geq 0\} = \emptyset$.

($\Rightarrow$) Less obvious. We shall prove a slightly more general version of it called Farkas' lemma.