Project

- project groups
  - groups of 3-4
  - if needed, feel free to use "**Search for Teammates!**" feature on Piazza (pinned)
  - send me an email (mvasconcelos@eng.ucsd.edu) stating who are the group **members** (please use your official UCSD name) as soon as you know it, with deadline Tuesday, 1/18

- project proposal
  - due Tuesday, 2/1 @ 11:59pm
  - one–page maximum stating:
    - problem
    - data you will use
    - draft of proposed solution (can be updated later)
    - experiments you will run (can be updated later)
    - references (you can use an additional page for this)

ECE 271B – Winter 2022

Optimization

Disclaimer:
This class will be recorded
and made available to students asynchronously.

Manuela Vasconcelos

ECE Department, UCSD

Optimization

- many engineering problems boil down to **optimization**
- goal: find **maximum** or **minimum** of a function

- **Definition:** given functions f, $g_i, i = 1, \ldots, r$ and $h_i, i = 1, \ldots, m$ defined on some domain $\Omega \in \mathbb{R}^n$

$$\begin{array}{ll} \min\limits_{\mathbf{w}} & f(\mathbf{w}), \mathbf{w} \in \Omega \\ \text{subject to} & g_i(\mathbf{w}) \leq 0, \forall i \\ & h_i(\mathbf{w}) = 0, \forall i \end{array}$$

- $f(\mathbf{w})$: **cost**; h_i (equality), g_i (inequality): **constraints**
- for compactness, we write $g(\mathbf{w}) \leq 0$ instead of $g_i(\mathbf{w}) \leq 0, \forall i$ and similarly $h(\mathbf{w}) = 0$
- note that $g(\mathbf{w}) \geq 0 \Leftrightarrow -g(\mathbf{w}) \leq 0$ (no need for ≥ 0)

Optimization

- **note**: maximizing $f(\mathbf{w})$ is the same as minimizing $-f(\mathbf{w})$, so this definition also works for maximization

- the feasible region is the region where $f(\cdot)$ is defined and all constraints hold

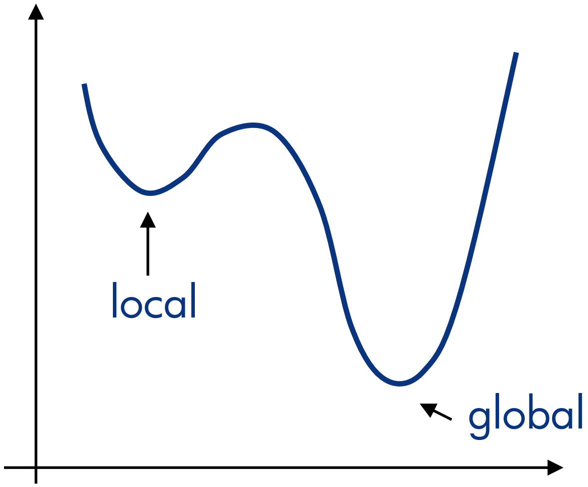
$$\boxed{\Re = \{\mathbf{w} \in \Omega \mid g(\mathbf{w}) \leq 0, h(\mathbf{w}) = 0\}}$$

- $\mathbf{w}^*$ is a global minimum of $f(\mathbf{w})$ if

$$f(\mathbf{w}) \geq f(\mathbf{w}^*), \forall\, \mathbf{w} \in \Omega$$

- $\mathbf{w}^*$ is a local minimum of $f(\mathbf{w})$ if

$$\exists \varepsilon > 0 \ \text{ s.t. } \ \|\mathbf{w} - \mathbf{w}^*\| < \varepsilon \Rightarrow f(\mathbf{w}) \geq f(\mathbf{w}^*)$$



Derivative

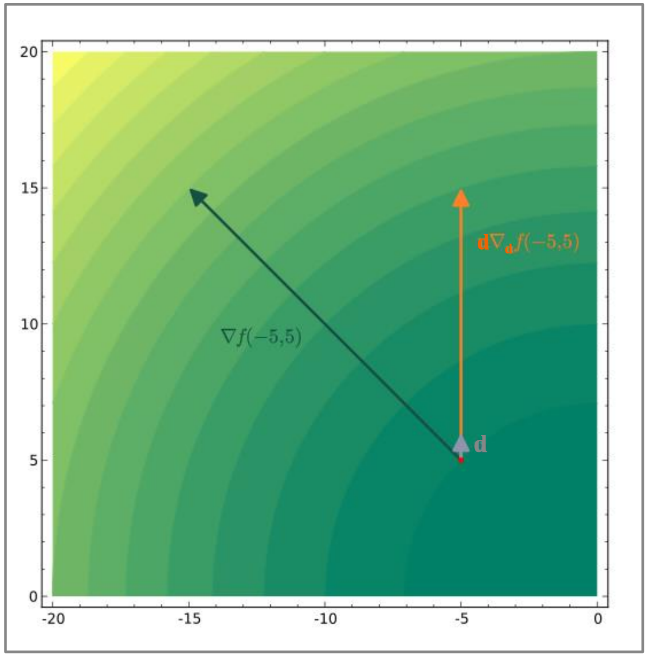
- a function $f(w)$ is **differentiable** if it has derivatives for all w
- the **derivative** at point w is defined as

$$\frac{\partial f}{\partial w} = \lim_{\alpha \to 0} \frac{f(w+\alpha) - f(w)}{\alpha}$$

  - note that the magnitude of the derivative is a **measure** how much the function is growing at point w

- for a **multivariate function** $f(\mathbf{w}), \mathbf{w} \in \mathbb{R}^n$
  - the problem is **more complex** because we can compute the derivative in <u>many</u> directions
  - e.g. contour plot of

$$\boxed{f(\mathbf{w}) = \|\mathbf{w}\|^2 = w_1^2 + w_2^2}$$

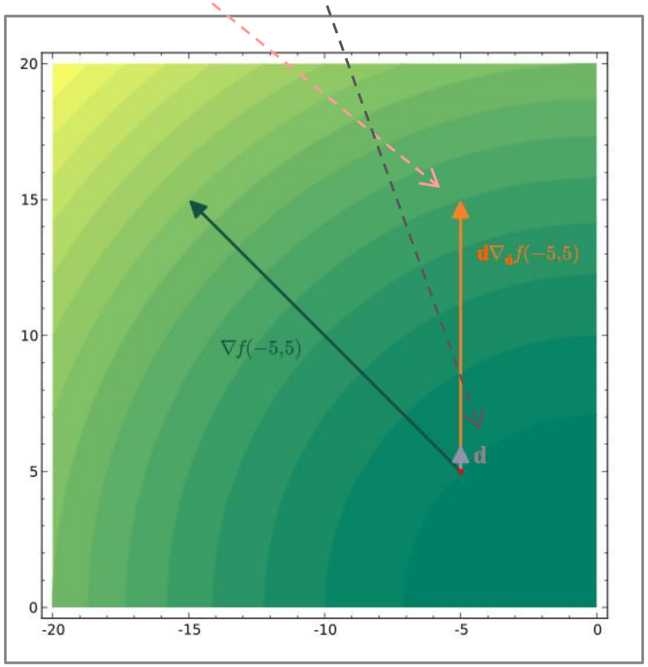


Directional Derivative

- the directional derivative of $f(\mathbf{w})$ at $\mathbf{w}$, along direction $\mathbf{d}$ is

$$D_{\mathbf{d}} f(\mathbf{w}) = \lim_{\alpha \to 0} \frac{f(\mathbf{w} + \alpha \mathbf{d}) - f(\mathbf{w})}{\alpha}$$

  - (note that we are assuming that $\mathbf{d}$ is a unit vector $\|\mathbf{d}\| = 1$, otherwise we have to divide by $\|\mathbf{d}\|$)
  - this **measures** how much the **function grows** if we give an infinitesimal step along $\mathbf{d}$



- from Taylor series expansion of $f(\mathbf{w})$,

$$f(\mathbf{w} + \alpha \mathbf{d}) = f(\mathbf{w}) + \alpha\, \mathbf{d}^T \nabla f(\mathbf{w}) + O(\alpha^2)$$

where

$$\nabla f(\mathbf{z}) = \left(\frac{\partial f}{\partial w_0}(\mathbf{z}), \cdots, \frac{\partial f}{\partial w_{n-1}}(\mathbf{z}) \right)^T$$

is the gradient of a function $f(\mathbf{w})$ at $\mathbf{z}$

The Gradient

$$f(\mathbf{w} + \alpha \mathbf{d}) = f(\mathbf{w}) + \alpha\, \mathbf{d}^T \nabla f(\mathbf{w}) + O(\alpha^2)$$

$$f(\mathbf{w} + \alpha \mathbf{d}) - f(\mathbf{w}) = \alpha\, \mathbf{d}^T \nabla f(\mathbf{w}) + O(\alpha^2)$$

- it follows that

$$D_{\mathbf{d}} f(\mathbf{w}) = \lim_{\alpha \to 0} \frac{f(\mathbf{w} + \alpha \mathbf{d}) - f(\mathbf{w})}{\alpha}$$

can be written as

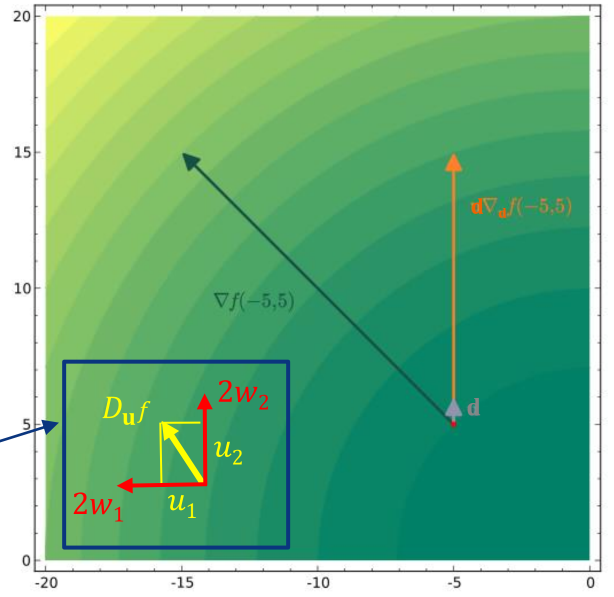
dot–product of the gradient with the direction vector

$$D_{\mathbf{d}} f(\mathbf{w}) = \mathbf{d}^T \nabla f(\mathbf{w}) = \sum_i d_i \frac{\partial f(\mathbf{w})}{\partial w_i}$$

- note that **each partial derivative** is a <u>**function**</u>

- the **gradient** is a set of n <u>basis functions</u> (the **partial derivatives**) that you can use to reconstruct the derivative along <u>**any**</u> direction

$$f(\mathbf{w}) = w_1^2 + w_2^2$$
$$\nabla f(\mathbf{w}) = (2w_1, 2w_2)$$



The Gradient

- an important consequence is that

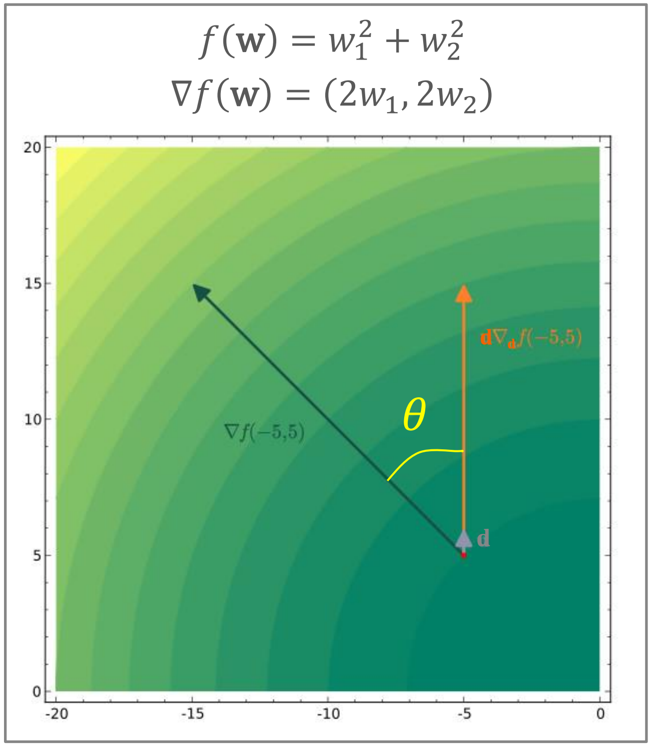
$$D_{\mathbf{d}} f(\mathbf{w}) = \mathbf{d}^T \nabla f(\mathbf{w}) = \|\mathbf{d}\| \|\nabla f(\mathbf{w})\| \cos\theta$$
$$= \|\nabla f(\mathbf{w})\| \cos\theta$$

  - this implies that the direction of **maximum** derivative $\mathbf{d}_o$ is that of the gradient $(\theta = 0)$

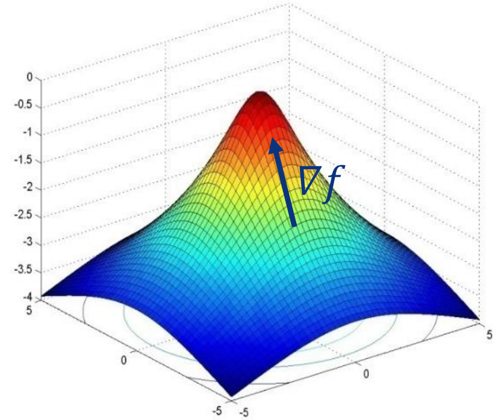
$$\boxed{\mathbf{d}_o = \arg\max_{\mathbf{d}} D_{\mathbf{d}} f(\mathbf{w}) = \frac{\nabla f(\mathbf{w})}{\|\nabla f(\mathbf{w})\|}}$$

  - the derivative along this direction is

$$D_{\mathbf{d}_o} f(\mathbf{w}) = \max_{\mathbf{d}} D_{\mathbf{d}} f(\mathbf{w}) = \|\nabla f(\mathbf{w})\|$$

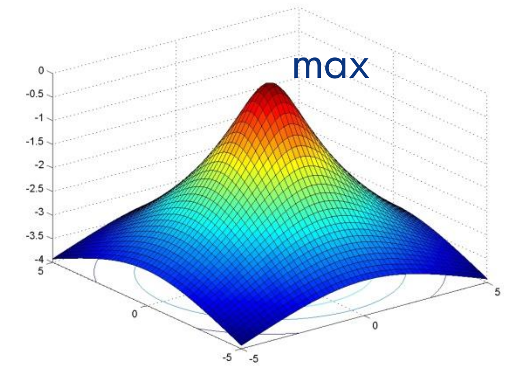


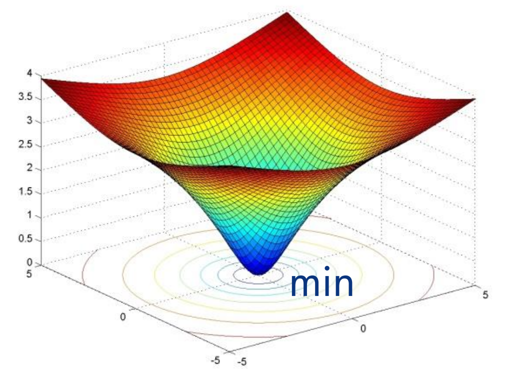
- in **summary**
  - the direction of the gradient is that of **steepest growth** of the function
  - the magnitude of the gradient is a measure how much the function is growing at point $\mathbf{w}$ (in that direction)

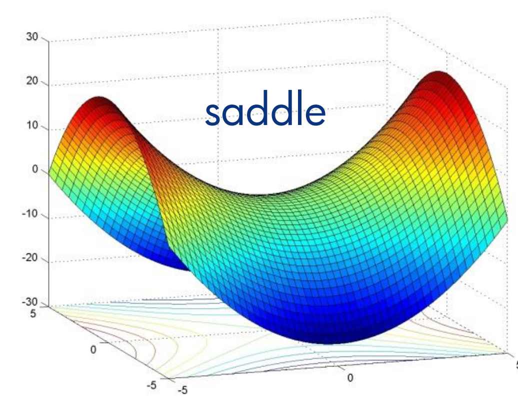


The Gradient

- note that if $\nabla f = 0$
  - there is <u>no</u> direction of growth
  - also $-\nabla f = 0$, and there is <u>no</u> direction of decrease
  - we are either at a local minimum or maximum or “saddle” point
- conversely, at local min or max or saddle point
  - no direction of growth or decrease
  - $\nabla f = 0$
- this shows that we have a **critical point** if and only if $\nabla f = 0$
- to determine which type, we need second–order conditions







The Hessian

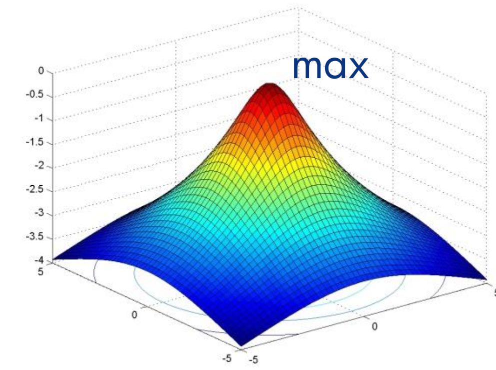
- if $\nabla f = 0$, by Taylor series,

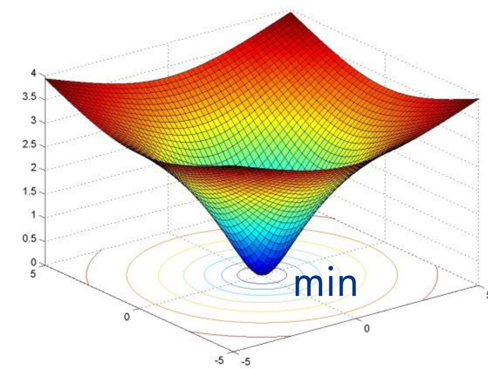
$$f(\mathbf{w} + \alpha\mathbf{d}) = f(\mathbf{w}) + \underbrace{\alpha\mathbf{d}^T\nabla f(\mathbf{w})}_{0} + \frac{\alpha^2}{2}\mathbf{d}^T\nabla^2 f(\mathbf{w})\mathbf{d} + O(\alpha^3)$$

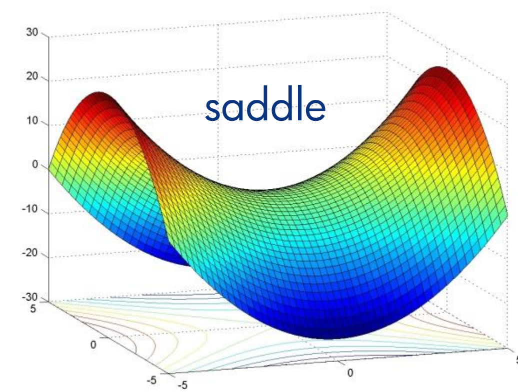
and

$$\frac{f(\mathbf{w} + \alpha\mathbf{d}) - f(\mathbf{w})}{\alpha^2} = \frac{1}{2}\mathbf{d}^T\nabla^2 f(\boldsymbol{w})\mathbf{d} + O(\alpha)$$

- pick α such that $O(\alpha) \ll |\mathbf{d}^T\nabla^2 f\ \mathbf{d}|, \forall\mathbf{d} \neq \mathbf{0}$
  - maximum at $\mathbf{w}$ if and only if $\mathbf{d}^T\nabla^2 f\mathbf{d} \leq 0, \forall\mathbf{d} \neq \mathbf{0}$
  - minimum at $\mathbf{w}$ if and only if $\mathbf{d}^T\nabla^2 f\mathbf{d} \geq 0, \forall\mathbf{d} \neq \mathbf{0}$
  - saddle, otherwise
- this proves the following theorems







Minima Conditions (Unconstrained)

- **Theorem:** Let $f(\mathbf{w})$ be continuously differentiable. $\mathbf{w}^*$ is a **local minimum** of $f(\mathbf{w})$ if and only if
  - f has zero gradient at $\mathbf{w}^*$

$$\boxed{\nabla f(\mathbf{w}^*) = 0}$$

  - and the Hessian of f at $\mathbf{w}^*$ is positive−semidefinite

$$\boxed{\mathbf{d}^T \nabla^2 f(\mathbf{w}^*)\mathbf{d} \geq 0, \forall \mathbf{d} \in \mathbb{R}^n}$$

where

$$\nabla^2 f(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_0^2}(\mathbf{x}) & \cdots & \frac{\partial^2 f}{\partial x_0 \partial x_{n-1}}(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_{n-1} \partial x_0}(\mathbf{x}) & \cdots & \frac{\partial^2 f}{\partial x_{n-1}^2}(\mathbf{x}) \end{bmatrix}$$

Maxima Conditions (Unconstrained)

- **Theorem:** Let $f(\mathbf{w})$ be continuously differentiable. $\mathbf{w}^*$ is a local maximum of $f(\mathbf{w})$ if and only if
  - f has zero gradient at $\mathbf{w}^*$

$$\boxed{\nabla f(\mathbf{w}^*) = 0}$$

  - and the Hessian of f at $\mathbf{w}^*$ is negative−semidefinite

$$\boxed{\mathbf{d}^T \nabla^2 f(\mathbf{w}^*)\mathbf{d} \leq 0, \forall \mathbf{d} \in \mathbb{R}^n}$$

where

$$\nabla^2 f(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_0^2}(\mathbf{x}) & \cdots & \frac{\partial^2 f}{\partial x_0 \partial x_{n-1}}(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_{n-1} \partial x_0}(\mathbf{x}) & \cdots & \frac{\partial^2 f}{\partial x_{n-1}^2}(\mathbf{x}) \end{bmatrix}$$

Example

- consider the functions

$$f(\mathbf{x}) = x_1 + x_2 \qquad\qquad h(\mathbf{x}) = x_1^2 + x_2^2$$

- the gradients are

$$\nabla f(\mathbf{x}) = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad\qquad \nabla h(\mathbf{x}) = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix}$$

- f has no minima or maxima

- h has a critical point at the origin $\mathbf{x} = (0,0)$ and, since the Hessian is positive–definite

$$\nabla^2 h(\mathbf{x}) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix},$$

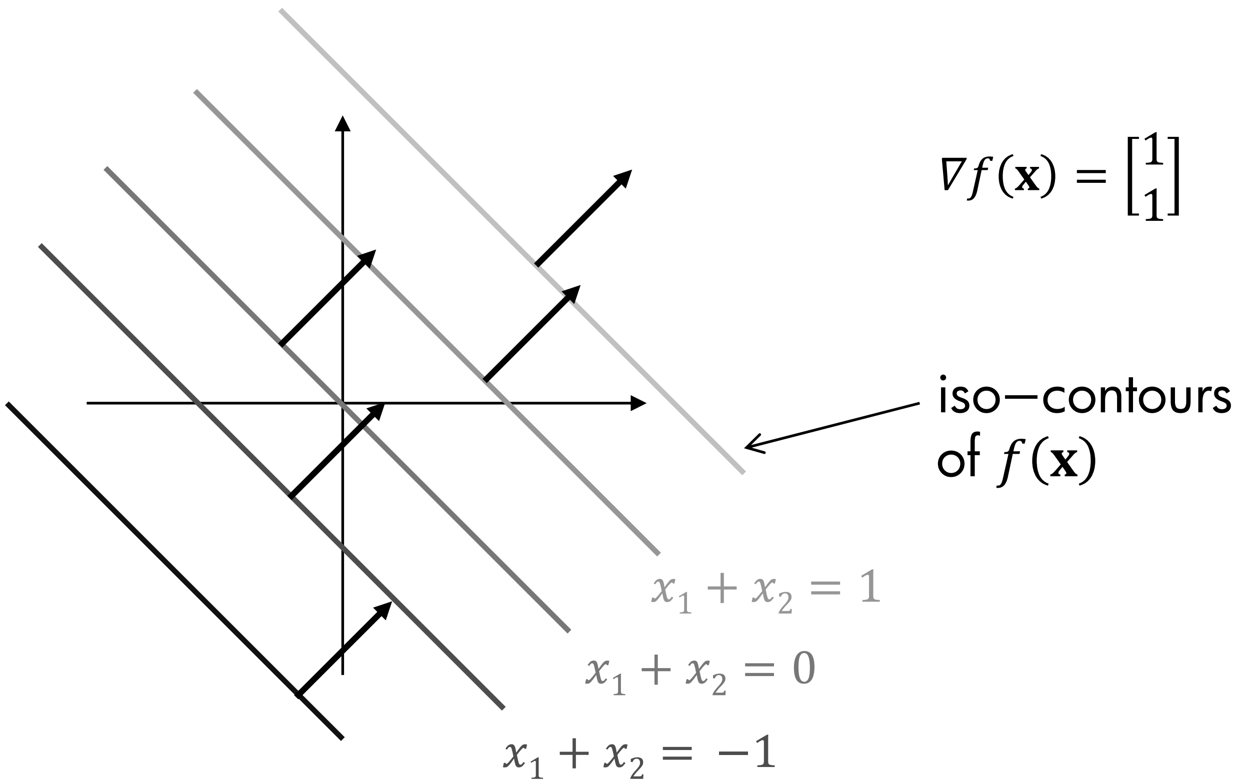
  this is a minimum

Example (cont)

- makes sense because

$$f(\mathbf{x}) = x_1 + x_2$$

is a **plane**, gradient is constant



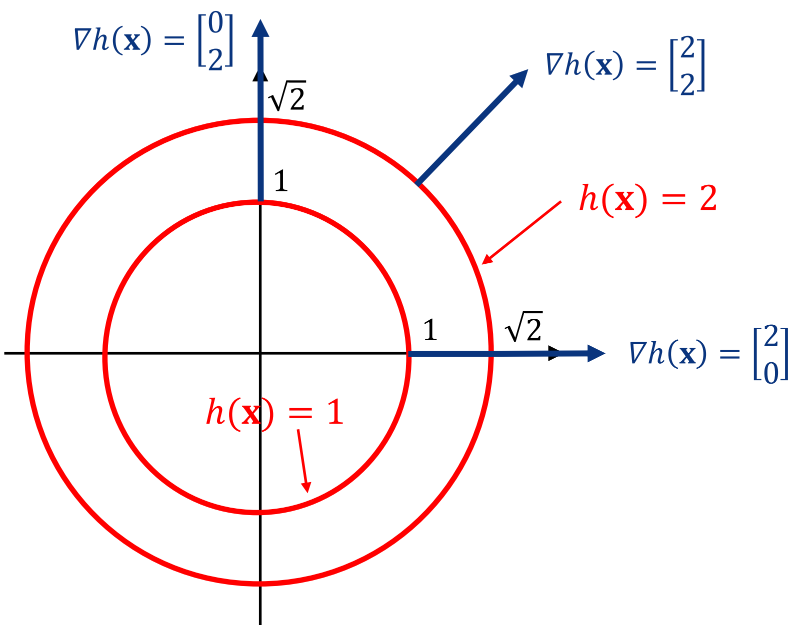
Example (cont)

- makes sense because

$$h(\mathbf{x}) = x_1^2 + x_2^2$$

  is a **quadratic**, positive everywhere but the origin

- note how gradient points towards largest increase



$$\nabla h(\mathbf{x}) = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix}$$

Convex Functions

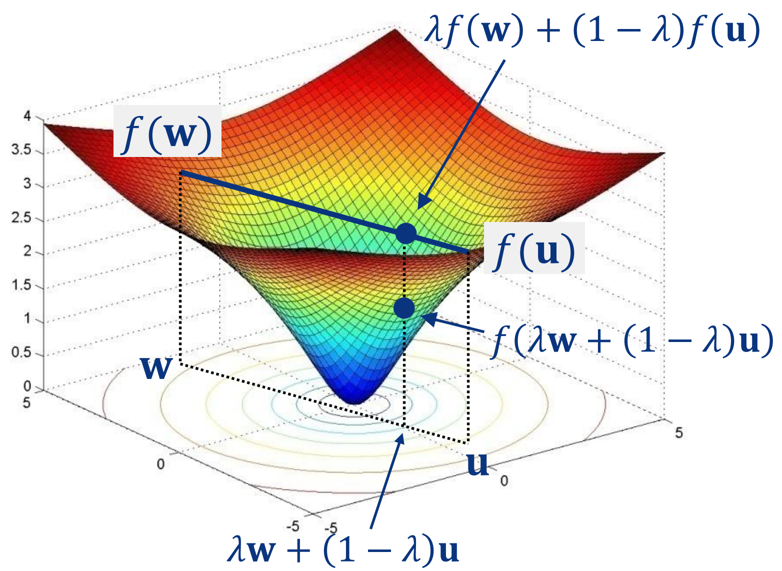
- **Definition:** $f(\mathbf{w})$ is convex if $\forall\ \mathbf{w}, \mathbf{u} \in \Omega$ and $\lambda \in [0,1]$

$$f(\lambda\mathbf{w} + (1-\lambda)\mathbf{u}) \leq \lambda f(\mathbf{w}) + (1-\lambda) f(\mathbf{u})$$

- **Theorem:** $f(\mathbf{w})$ is convex if and only if its Hessian is positive−definite for all $\mathbf{w}$

$$\boxed{\mathbf{y}^T \nabla^2 f(\mathbf{w}) \mathbf{y} \geq 0, \forall\ \mathbf{y} \in \Omega}$$

- **Proof:**
  - requires some intermediate results that we will not cover
  - we will skip it



Concave Functions

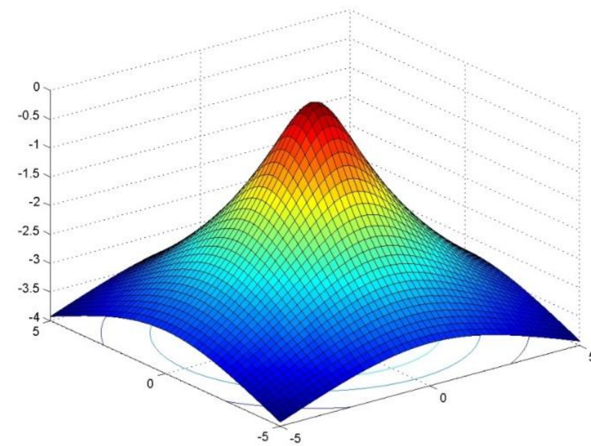
- **Definition:** $f(\mathbf{w})$ is concave if $\forall\ \mathbf{w}, \mathbf{u} \in \Omega$ and $\lambda \in [0,1]$

$$f(\lambda \mathbf{w} + (1-\lambda)\mathbf{u}) \geq \lambda f(\mathbf{w}) + (1-\lambda) f(\mathbf{u})$$

- **Theorem:** $f(\mathbf{w})$ is concave if and only if its Hessian is negative−definite for all $\mathbf{w}$

$$\boxed{\mathbf{y}^T \nabla^2 f(\mathbf{w}) \mathbf{y} \leq 0, \forall\ \mathbf{y} \in \Omega}$$

- **Proof:**
  - $-f(\mathbf{w})$ is convex
  - by previous theorem, Hessian of $-f(\mathbf{w})$ is positive−definite
  - Hessian of $f(\mathbf{w})$ is negative−definite ■



Convex Functions

- **Theorem:** If $f(\mathbf{w})$ is convex, any local minimum $\mathbf{w}^*$ is also a global minimum.

- **Proof:**

  $\mathbf{w}^*$ is a global minimum of $f(\mathbf{w})$ if $f(\mathbf{w}) \geq f(\mathbf{w}^*), \forall\, \mathbf{w} \in \Omega$

  - we need to show that, $f(\mathbf{w}^*) \leq \; f(\mathbf{u}), \forall\, \mathbf{u}$,
  - for $\forall\, \mathbf{u}$ and $\lambda \in [0,1]$: $\|\mathbf{w}^* - [\lambda \mathbf{w}^* + (1-\lambda)\mathbf{u}]\| = (1-\lambda)\|\mathbf{w}^* - \mathbf{u}\|$
  - and, making λ arbitrarily close to 1, we can make

  $$\|\mathbf{w}^* - [\lambda \mathbf{w}^* + (1-\lambda)\mathbf{u}]\| \leq \varepsilon, \forall \varepsilon > 0$$

  $\mathbf{w}^*$ is a local minimum of $f(\mathbf{w})$ if $\exists \varepsilon > 0$ s.t. $\|\mathbf{w} - \mathbf{w}^*\| < \varepsilon \Rightarrow f(\mathbf{w}) \geq f(\mathbf{w}^*)$

  - since $\mathbf{w}^*$ is local minimum, it follows that $f(\mathbf{w}^*) \leq f(\lambda \mathbf{w}^* + (1-\lambda)\mathbf{u})$ and, by convexity, that $f(\mathbf{w}^*) \leq \lambda f(\mathbf{w}^*) + (1-\lambda) f(\mathbf{u})$

  $f(\mathbf{w})$ is convex if $\forall\, \mathbf{w}, \mathbf{u} \in \Omega$ and $\lambda \in [0,1]$
  $f(\lambda \mathbf{w} + (1-\lambda)\mathbf{u}) \leq \lambda f(\mathbf{w}) + (1-\lambda) f(\mathbf{u})$

  - or $(1-\lambda) f(\mathbf{w}^*) \leq (1-\lambda) f(\mathbf{u})$
  - and $f(\mathbf{w}^*) \leq f(\mathbf{u})$ ∎

Constrained Optimization

- in summary:
  - we know what are conditions for **unconstrained** max and min
  - we like **convex** functions (find a minima, it will be global minimum)

- what about **optimization with constraints**?
- a few definitions to start with

- **Definition:** An inequality $g_i(\mathbf{w}) \leq 0$ is **active** if $g_i(\mathbf{w}) = 0$, otherwise is **inactive**

- inequalities can be expressed as equalities by introduction of **slack variables**

$$g_i(\mathbf{w}) \leq 0 \quad \Leftrightarrow \quad g_i(\mathbf{w}) + \xi_i = 0 \quad \text{and} \quad \xi_i \geq 0$$

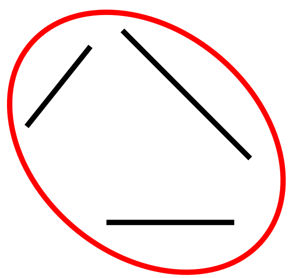
Convex Optimization

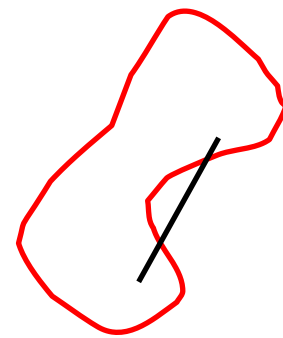
- **Definition:** A set Ω is convex if

  $$\forall\, \mathbf{w}, \mathbf{u} \in \Omega \text{ and } \lambda \in [0,1] \text{ then } \lambda\mathbf{w} + (1-\lambda)\mathbf{u} \in \Omega$$

- "a line between any two points in Ω is also in Ω"

convex





not convex

- **Definition:** An optimization problem where the set Ω, the cost f and all constraints g and h are convex is said to be **convex**

- **note: linear** constraints $g(x) = Ax + b$ are <u>always</u> convex (zero Hessian)

Constrained Optimization

- we will consider **general** (not only convex) constrained optimization problems, start by the case with **only equalities**

- **Theorem:** Consider the problem

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{subject to} \quad h(\mathbf{x}) = 0$$

$h_i, i = 1, \dots, m$
$h_i(\mathbf{x}) = 0, \forall i$

where the constraint gradients $\nabla h_i(\mathbf{x}^*)$ are linearly independent. Then, $\mathbf{x}^*$ is a solution if and only if there exits a unique vector $\boldsymbol{\lambda}$ such that

gradient condition

i) $\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla h_i(\mathbf{x}^*) = 0$

"constraint gradients & Hessians"

Hessian condition

ii) $\mathbf{y}^T[\nabla^2 f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla^2 h_i(\mathbf{x}^*)]\mathbf{y} \geq 0, \forall \mathbf{y} \text{ s.t. } \nabla h(\mathbf{x}^*)^T \mathbf{y} = 0$

i) $\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla h_i(\mathbf{x}^*) = 0$

ii) $\mathbf{y}^T[\nabla^2 f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla^2 h_i(\mathbf{x}^*)]\mathbf{y} \geq 0, \forall \mathbf{y} \text{ s.t. } \nabla h(\mathbf{x}^*)^T \mathbf{y} = 0$

Alternative Formulation

- stating the conditions through the Lagrangian

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

- the theorem can be compactly written as

i) $\nabla_{\mathbf{x}} L(\mathbf{x}^*, \boldsymbol{\lambda}^*) = 0$

ii) $\nabla_{\boldsymbol{\lambda}} L(\mathbf{x}^*, \boldsymbol{\lambda}^*) = 0$ → this just means that $h_i(\mathbf{x}) = 0, \forall i$

iii) $\mathbf{y}^T \nabla^2_{\mathbf{xx}} L(\mathbf{x}^*, \boldsymbol{\lambda}^*)\, \mathbf{y} \geq 0, \forall \mathbf{y} \text{ s.t. } \nabla h(\mathbf{x}^*)^T \mathbf{y} = 0$

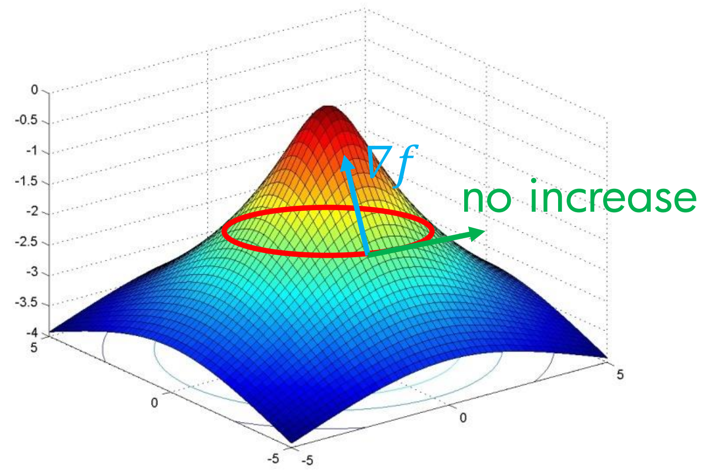
- the entries of λ are referred to as Lagrange multipliers

The Gradient (Revisited)

- recall that derivative of f along $\mathbf{d}$ is

$$\boxed{\lim_{\alpha \to 0} \frac{f(\mathbf{w} + \alpha \mathbf{d}) - f(\mathbf{w})}{\alpha} = \mathbf{d}^T \nabla f(\mathbf{w}) = \|\mathbf{d}\| \|\nabla f(\mathbf{w})\| \cos\big(\mathbf{d}, \nabla f(\mathbf{w})\big)}$$

- this means that
  - greatest increase when $\mathbf{d} \parallel \nabla f$
  - no increase when $\mathbf{d} \perp \nabla f$ since there is no increase when $\mathbf{d}$ is tangent to iso−contour $f(\mathbf{x}) = k$
  - the gradient is perpendicular to the tangent of the iso−contour



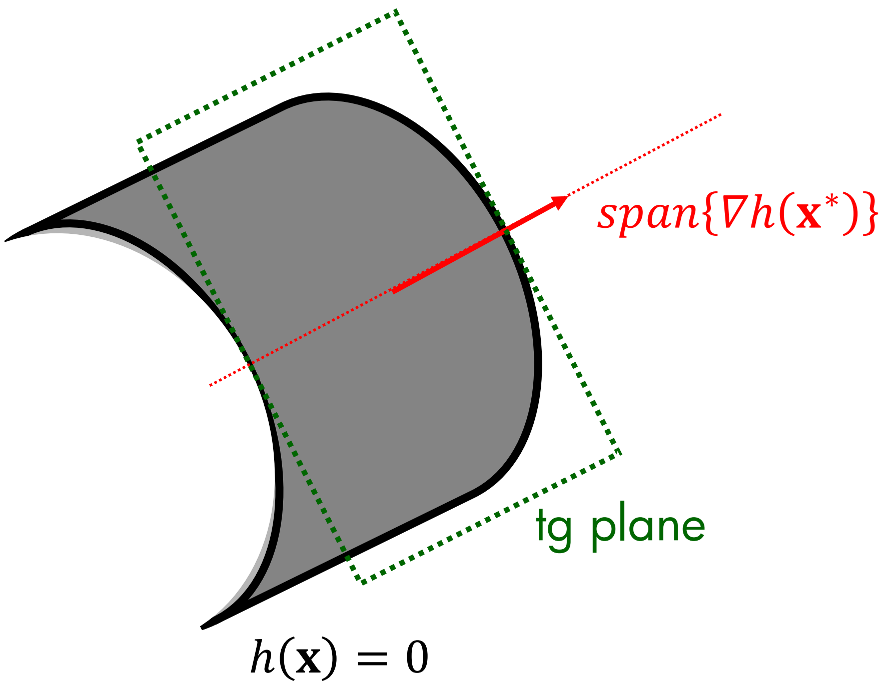
- allows <u>geometric interpretation</u> of the Lagrangian conditions

i) $\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla h_i(\mathbf{x}^*) = 0$

ii) $\mathbf{y}^T[\nabla^2 f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla^2 h_i(\mathbf{x}^*)]\mathbf{y} \geq 0, \forall \mathbf{y}$ s.t. $\nabla h(\mathbf{x}^*)^T \mathbf{y} = 0$

Lagrangian Optimization

- <u>geometric interpretation</u>:
  - since $h(\mathbf{x}) = 0$ is an iso–contour of $h(\mathbf{x})$, $\nabla h(\mathbf{x}^*)$ is perpendicular to the iso–contour
  - $\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla h_i(\mathbf{x}^*) = 0$ says that $\nabla f(\mathbf{x}^*) \in span\{\nabla h_i(\mathbf{x}^*)\}$
  - i.e. ∇f $\perp$ to tangent space of the constraint surface $h(\mathbf{x}) = 0$
  - **intuitively**
    - direction of largest increase of f is $\perp$ to constraint surface
    - the gradient is zero along the constraint
    - no way to give an infinitesimal gradient step, without violating the constraint
    - it is impossible to increase f and still satisfy the constraint



Example

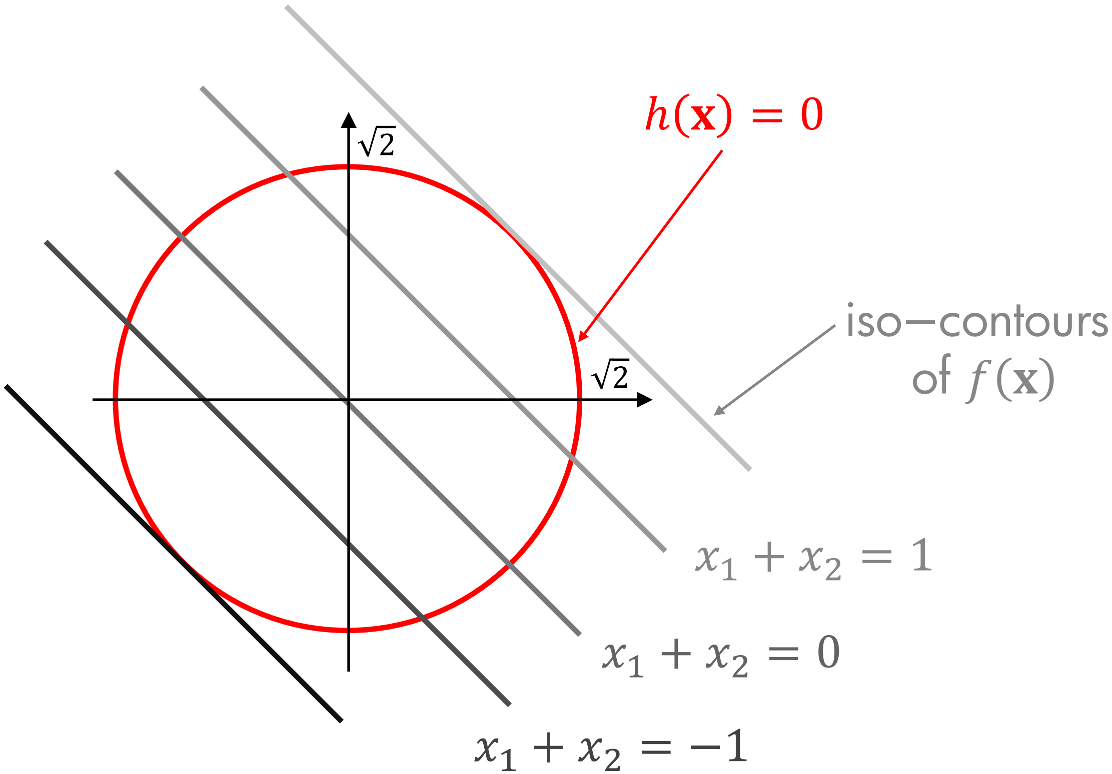
- consider the problem

$$\boxed{\min x_1 + x_2 \ \text{ subject to } \ {x_1}^2 + {x_2}^2 = 2}$$

- it leads to the following picture

$$f(\mathbf{x}) = x_1 + x_2$$

$$h(\mathbf{x}) = x_1^2 + x_2^2 - 2$$



$$\nabla f(\mathbf{x}) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

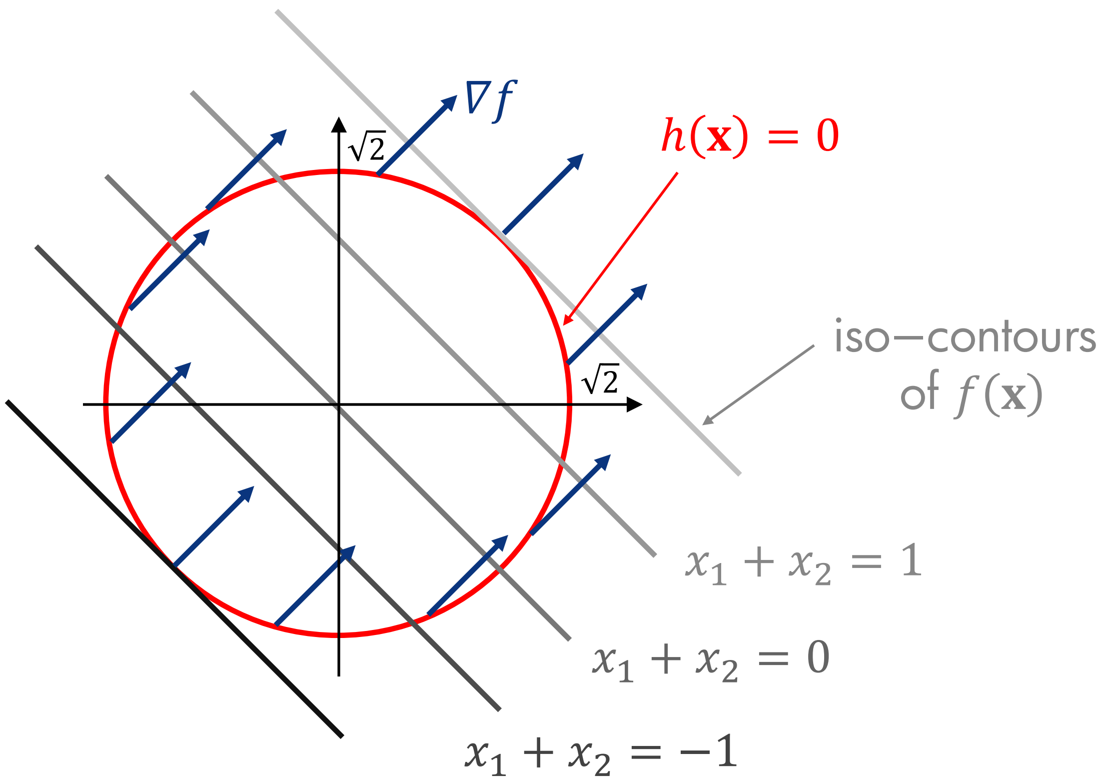
$$\nabla h(\mathbf{x}) = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix}$$

Example (cont)

- consider the problem

$$\boxed{\min x_1 + x_2 \ \text{subject to} \ {x_1}^2 + {x_2}^2 = 2}$$

- ∇f ⊥ to the iso−contours of f $(x_1 + x_2 = k)$



$$\nabla f(\mathbf{x}) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

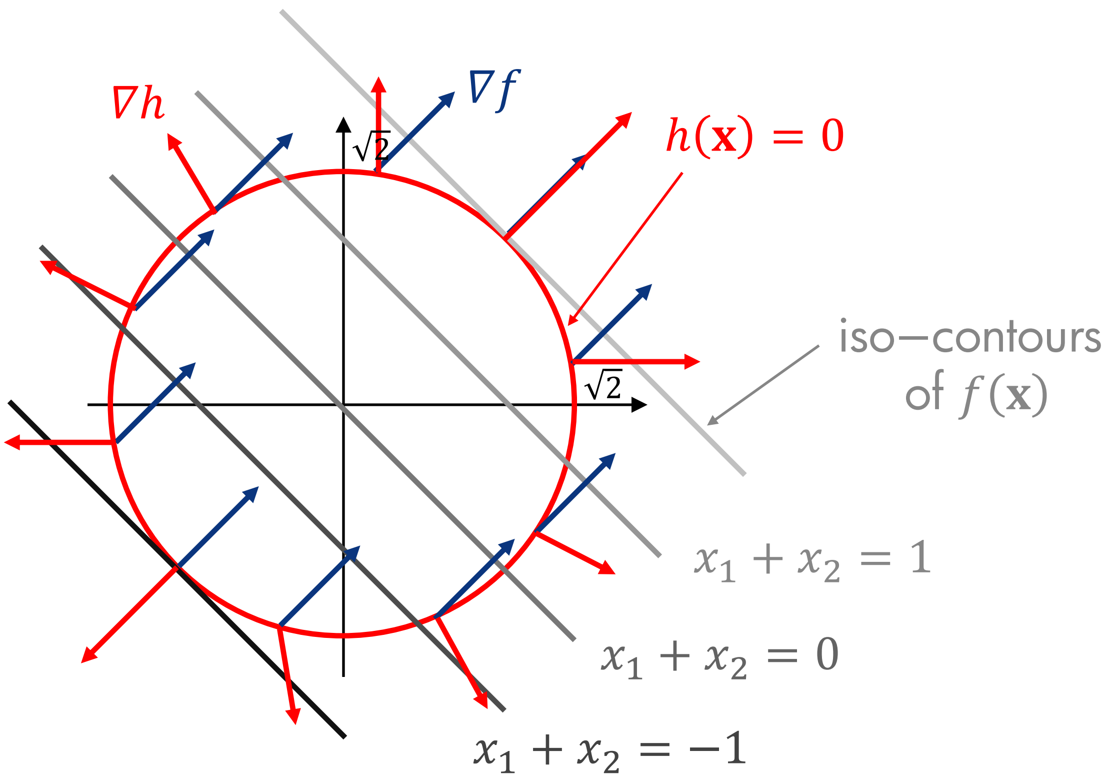
$$\nabla h(\mathbf{x}) = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix}$$

Example (cont)

- consider the problem

$$\boxed{\min x_1 + x_2 \text{ subject to } {x_1}^2 + {x_2}^2 = 2}$$

- $\nabla h \perp$ to the iso−contour of h $({x_1}^2 + {x_2}^2 - 2 = 0)$



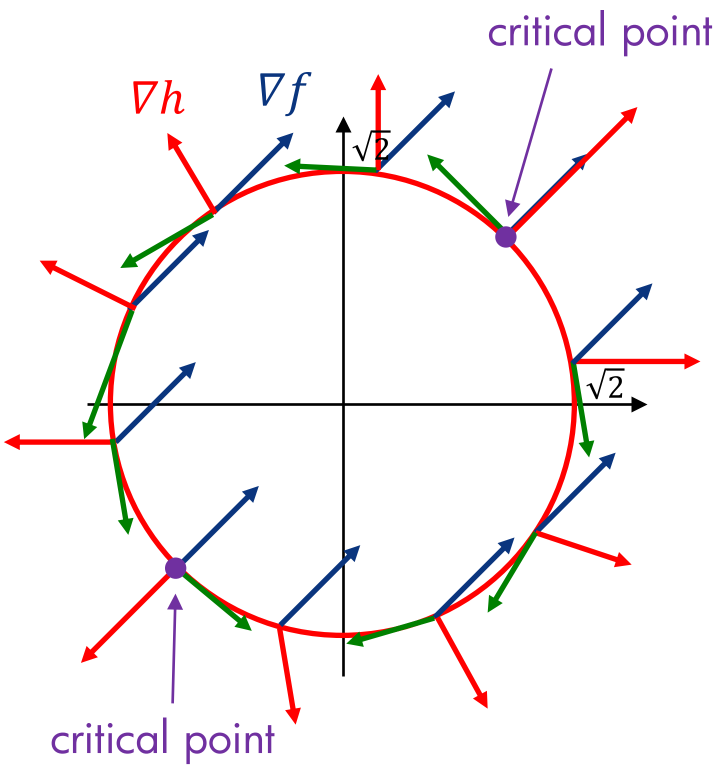
$$\nabla f(\mathbf{x}) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\nabla h(\mathbf{x}) = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix}$$

Example (cont)

- recall that derivative along $\mathbf{d}$ is

$$\lim_{\alpha \to 0} \frac{f(\mathbf{w} + \alpha \mathbf{d}) - f(\mathbf{w})}{\alpha} = \mathbf{d}^T \nabla f(\mathbf{w}) = \|\mathbf{d}\| \|\nabla f(\mathbf{w})\| \cos\big(\mathbf{d}, \nabla f(\mathbf{w})\big)$$



- moving along the tangent is descent as long as

$$\cos(tg, \nabla f) < 0$$

- i.e.

$$\pi/2 < \measuredangle(tg, \nabla f) < 3\pi/2$$

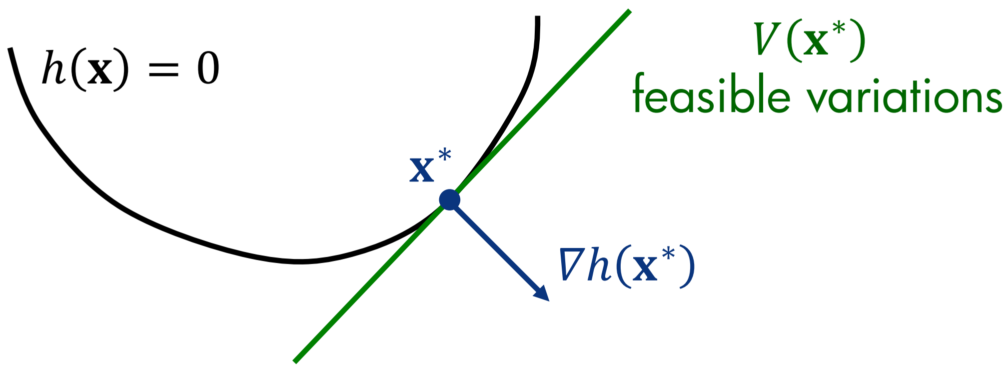
- can always find such $\mathbf{d}$ unless $\nabla f \perp tg$
- critical point when $\nabla f \parallel \nabla h$
- to find which type, we need 2nd order (as before)

Alternative View

- consider the tangent space to the iso–contour $h(\mathbf{x}) = 0$
- this is the **subspace** of first–order feasible variations

$$\boxed{V(\mathbf{x}^*) = \left\{\Delta\mathbf{x} \mid \nabla h_i^T(\mathbf{x}^*)\,\Delta\mathbf{x} = 0, \forall i\right\}}$$

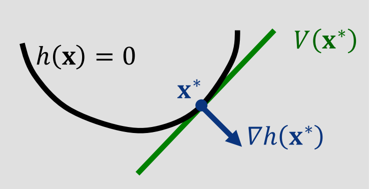
i.e. space of $\Delta\mathbf{x}$ for which a **step** $\mathbf{x} + \Delta\mathbf{x}$ satisfies the constraints $h_i(\mathbf{x})$ up to first–order approximation



$V(\mathbf{x}^*) = \{\Delta\mathbf{x} \mid \nabla h_i^T(\mathbf{x}^*)\,\Delta\mathbf{x} = 0, \forall i\}$

Feasible Variations

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla h_i(\mathbf{x}^*) = 0$$



- multiplying our first Lagrangian condition by $\Delta\mathbf{x}$

$$\nabla f^T(\mathbf{x}^*)\,\Delta\mathbf{x} + \sum_{i=1}^{m} \lambda_i \underbrace{\nabla h_i^T(\mathbf{x}^*)\,\Delta\mathbf{x}}_{\mathbf{0}} = 0$$

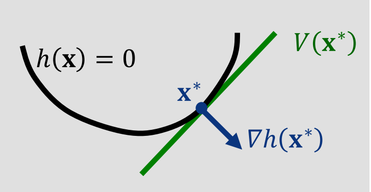
- it follows that

$$\boxed{\nabla f^T(\mathbf{x}^*)\,\Delta\mathbf{x} = 0, \forall \Delta\mathbf{x} \in V(\mathbf{x}^*)}$$

- this is a generalization of $\boxed{\nabla f(\mathbf{x}^*) = 0}$ in the **unconstrained case**
  - here, all that matters is that $\nabla f(\mathbf{x}^*)$ has no projection in $V(\mathbf{x}^*)$
  - implies that $\nabla f(\mathbf{x}^*) \perp V(\mathbf{x}^*)$ and therefore $\nabla f(\mathbf{x}^*) \parallel \nabla h(\mathbf{x}^*)$
  - **note:**
    - Hessian constraint only defined for $\mathbf{y}$ in $V(\mathbf{x}^*)$
    - **makes sense**: we cannot move anywhere else, does not really matter what Hessian is outside $V(\mathbf{x}^*)$

$V(\mathbf{x}^*) = \{\Delta\mathbf{x} \mid \nabla h_i^T(\mathbf{x}^*)\,\Delta\mathbf{x} = 0, \forall i\}$

Feasible Variations



- returning to our optimality conditions

i) $\nabla f(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i \nabla h_i(\mathbf{x}^*) = 0$

$\nabla^2_{\mathbf{xx}} L(\mathbf{x}^*, \boldsymbol{\lambda}^*)$

ii) $\mathbf{y}^T[\nabla^2 f(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i \nabla^2 h_i(\mathbf{x}^*)]\mathbf{y} \geq 0, \forall \mathbf{y}$ s.t. $\nabla h(\mathbf{x}^*)^T\mathbf{y} = 0$

- this explains the "extra stuff" in the Hessian condition
  - it restricts the Hessian constraint to $\mathbf{y}$ in $V(\mathbf{x}^*)$
  - the Lagragian <u>only</u> **has to be positive–definite in** $V(\mathbf{x}^*)$
  - makes sense: we cannot move anywhere else, does not really matter what Hessian is outside $V(\mathbf{x}^*)$

In Summary

- for a constrained optimization problem with <u>equality</u> constraints

- **Theorem:** Consider the problem

$$\boxed{\mathbf{x}^* = \arg\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{subject to} \quad h(\mathbf{x}) = 0}$$

where the constraint gradients $\nabla h_i(\mathbf{x}^*)$ are linearly independent. Then, $\mathbf{x}^*$ is a solution if and only if there exits a unique vector $\boldsymbol{\lambda}$ such that

i) $\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla h_i(\mathbf{x}^*) = 0$

ii) $\mathbf{y}^T[\nabla^2 f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla^2 h_i(\mathbf{x}^*)]\mathbf{y} \geq 0, \forall \mathbf{y} \text{ s.t. } \nabla h(\mathbf{x}^*)^T \mathbf{y} = 0$

Alternative Formulation

- stating the conditions through the Lagrangian

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

- the theorem can be **compactly** written as

i) $\nabla_{\mathbf{x}} L(\mathbf{x}^*, \boldsymbol{\lambda}^*) = 0$

ii) $\nabla_{\boldsymbol{\lambda}} L(\mathbf{x}^*, \boldsymbol{\lambda}^*) = 0$

iii) $\mathbf{y}^T \nabla^2_{\mathbf{xx}} L(\mathbf{x}^*, \boldsymbol{\lambda}^*)\, \mathbf{y} \geq 0, \forall \mathbf{y} \text{ s.t. } \nabla h(\mathbf{x}^*)^T \mathbf{y} = 0$

- the entries of $\boldsymbol{\lambda}$ are referred to as Lagrange multipliers

General Optimization

- what about problems with <u>both</u> equality and inequality constraints?

$$\begin{aligned} \min_{\mathbf{w}} \quad & f(\mathbf{w}), \mathbf{w} \in \Omega \\ \text{subject to} \quad & g_i(\mathbf{w}) \leq 0, \forall i \\ & h_i(\mathbf{w}) = 0, \forall i \end{aligned}$$

- **inequalities** can be **expressed as equalities** by introduction of slack variables

$$g_i(\mathbf{w}) \leq 0 \quad \Leftrightarrow \quad g_i(\mathbf{w}) + \xi_i = 0 \quad \text{and} \quad \xi_i \geq 0$$

- so, the solution is <u>similar</u>, but we have to figure out the values of the ξ_i
- we will talk about this later