ACM Summer School 2021 - Shape Modeling
Linear algebra & Optimization

Aditya Tatu

June 14, 2021

Optimization

Optimization problem

- Let $f : \mathbb{R}^n \to \mathbb{R}$.

max f(x)

min −f(x)

$$\min f(x)$$

$$\text{sub. to } \left.\begin{aligned} h_1(x) &= 0 \\ &\vdots \\ h_m(x) &= 0 \end{aligned}\right\} \text{Equality} \qquad \left.\begin{aligned} g_1(x) &\le 0 \\ &\vdots \\ g_p(x) &\le 0 \end{aligned}\right\} \text{Inequality constraints}$$

$$x \in \Omega \subset \mathbb{R}^n$$

- Denote $h(x) := [h_1(x) \;\ldots\; h_m(x)]^T$ and $g(x) := [g_1(x)\; g_2(x) \;\ldots\; g_p(x)]^T$, with $h : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^n \to \mathbb{R}^p$. Typically, f, g, h are assumed to be $\mathcal{C}^2$.
- The function f is called a *Cost/Objective* function. / Energy fnc.
  Constraints of the form $h(x) = 0$ and $g(x) \le 0$ are called *Functional constraints*, while constraints of the form $x \in \Omega$ are called *Set constraints*.
- Points satisfying all constraints are called *feasible points*.

Types of Minima

- Let $\Omega \subset \mathbb{R}^n$ denote the set of feasible points.
- A point $x^* \in \Omega$ is said to be a **global minima** if $\forall x \in \Omega \setminus \{x^*\}, f(x^*) \leq f(x)$.
- A point $x^* \in \Omega$ is said to be a **strict global minima** if $\forall x \in \Omega \setminus \{x^*\}, f(x^*) < f(x)$.
- A point $x^* \in \Omega$ is said to be a **local minima** if there is an $\epsilon > 0$, such that $\forall x \in \Omega \setminus \{x^*\}$, with $||x - x^*|| < \epsilon, f(x^*) \leq f(x)$.
- A point $x^* \in \Omega$ is said to be a **strict local minima** if there is an $\epsilon > 0$, such that $\forall x \in \Omega \setminus \{x^*\}$, with $||x - x^*|| < \epsilon, f(x^*) < f(x)$.

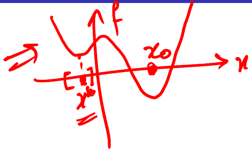
Types of Minima

- Let $\Omega \subset \mathbb{R}^n$ denote the set of feasible points.
- A point $x^* \in \Omega$ is said to be a **global minima** if $\forall x \in \Omega \setminus \{x^*\}, f(x^*) \leq f(x)$.
- A point $x^* \in \Omega$ is said to be a **strict global minima** if $\forall x \in \Omega \setminus \{x^*\}, f(x^*) < f(x)$.
- A point $x^* \in \Omega$ is said to be a **local minima** if there is an $\epsilon > 0$, such that $\forall x \in \Omega \setminus \{x^*\}$, with $||x - x^*|| < \epsilon, f(x^*) \leq f(x)$.
- A point $x^* \in \Omega$ is said to be a **strict local minima** if there is an $\epsilon > 0$, such that $\forall x \in \Omega \setminus \{x^*\}$, with $||x - x^*|| < \epsilon, f(x^*) < f(x)$.

Types of Minima

- Let $\Omega \subset \mathbb{R}^n$ denote the set of feasible points.
- A point $x^* \in \Omega$ is said to be a **global minima** if $\forall x \in \Omega \setminus \{x^*\}, f(x^*) \leq f(x)$.
- A point $x^* \in \Omega$ is said to be a **strict global minima** if $\forall x \in \Omega \setminus \{x^*\}, f(x^*) < f(x)$.
- A point $x^* \in \Omega$ is said to be a **local minima** if there is an $\epsilon > 0$, such that $\forall x \in \Omega \setminus \{x^*\}$, with $||x - x^*|| < \epsilon, f(x^*) \leq f(x)$.
- A point $x^* \in \Omega$ is said to be a **strict local minima** if there is an $\epsilon > 0$, such that $\forall x \in \Omega \setminus \{x^*\}$, with $||x - x^*|| < \epsilon, f(x^*) < f(x)$.

Types of Minima



- Let $\Omega \subset \mathbb{R}^n$ denote the set of feasible points.
- A point $x^* \in \Omega$ is said to be a **global minima** if $\forall x \in \Omega \setminus \{x^*\}, f(x^*) \leq f(x)$.
- A point $x^* \in \Omega$ is said to be a **strict global minima** if $\forall x \in \Omega \setminus \{x^*\}, f(x^*) < f(x)$.
- A point $x^* \in \Omega$ is said to be a **local minima** if there is an $\epsilon > 0$, such that $\forall x \in \Omega \setminus \{x^*\}$, with $||x - x^*|| < \epsilon, f(x^*) \leq f(x)$.

  $\{x \mid d(x, x^*) < \epsilon\}$
- A point $x^* \in \Omega$ is said to be a strict local minima if there is an $\epsilon > 0$, such that $\forall x \in \Omega \setminus \{x^*\}$, with $||x - x^*|| < \epsilon, f(x^*) < f(x)$.

global minima $\Rightarrow$ local minima

Types of Minima

- Let $\Omega \subset \mathbb{R}^n$ denote the set of feasible points.
- A point $x^* \in \Omega$ is said to be a **global minima** if $\forall x \in \Omega \setminus \{x^*\}, f(x^*) \leq f(x)$.
- A point $x^* \in \Omega$ is said to be a **strict global minima** if $\forall x \in \Omega \setminus \{x^*\}, f(x^*) < f(x)$.
- A point $x^* \in \Omega$ is said to be a **local minima** if there is an $\epsilon > 0$, such that $\forall x \in \Omega \setminus \{x^*\},$ with $||x - x^*|| < \epsilon, f(x^*) \leq f(x)$.
- A point $x^* \in \Omega$ is said to be a **strict local minima** if there is an $\epsilon > 0$, such that $\forall x \in \Omega \setminus \{x^*\},$ with $||x - x^*|| < \epsilon, f(x^*) < f(x)$.

Infimum and Supremum

- Infimum (Greatest Lower bound) of a set $A \subset \mathbb{R}$ is $\boxed{p \in \mathbb{R}}$ such that
  - $p \leq a, \forall a \in A$, and
  - if $q \in \mathbb{R}$ is such that $q \leq a, \forall a \in A$, then $q \leq p$.
- Notation: $p = \inf A$.

min $[0,1] = 0$ min $\cancel{(0,1)} \subseteq \mathbb{R}$

inf $[0,1] = 0$ inf $(0,1) = 0$

- Supremum (Smallest Upper bound) of a set $A \subset \mathbb{R}$ is $s \in \mathbb{R}$ such that
  - $a \leq s, \forall a \in A$, and
  - if $t \in \mathbb{R}$ is such that $a \leq t, \forall a \in A$, then $s \leq t$.
- Notation: $s = \sup A$.

Infimum and Supremum

- Infimum (Greatest Lower bound) of a set $A \subset \mathbb{R}$ is $p \in \mathbb{R}$ such that
  - $p \leq a, \forall a \in A$, and
  - if $q \in \mathbb{R}$ is such that $q \leq a, \forall a \in A$, then $q \leq p$.
- Notation: $p = \inf A$.
- Supremum (Smallest Upper bound) of a set $A \subset \mathbb{R}$ is $s \in \mathbb{R}$ such that
  - $a \leq s, \forall a \in A$, and
  - if $t \in \mathbb{R}$ is such that $a \leq t, \forall a \in A$, then $s \leq t$.
- Notation: $s = \sup A$.

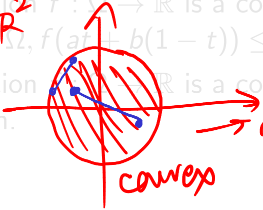
- A set $\Omega \subset \mathbb{R}^n$ is said to be a convex set if $\forall a, b \in \Omega$, $ta + (1-t)b \in \Omega, \forall t \in [0,1]$.

$\mathbb{R}^2$

$\{(x,y) \mid x^2+y^2 \leq 1\}$.

- A function $f : \Omega \to \mathbb{R}$ is a convex function if Ω is convex and $\forall a, b \in \Omega, f(at + b(1-t)) \leq tf(a) + (1-t)f(b), \forall t \in [0,1]$.
- A function $f : \Omega \to \mathbb{R}$ is a concave function if $-f$ is a convex function.

not convex

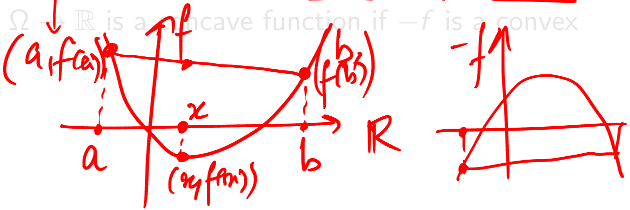


disk.

convex

Convex functions

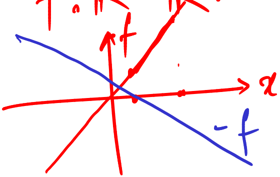
- A set $\Omega \subset \mathbb{R}^n$ is said to be a convex set if $\forall a, b \in \Omega$, $ta + (1-t)b \in \Omega, \forall t \in [0,1]$.
- A function $f : \Omega \to \mathbb{R}$ is a convex function if Ω is convex and $\forall a, b \in \Omega$, $f(at + b(1-t)) \leq tf(a) + (1-t)f(b), \forall t \in [0,1]$.
- A function $f : \Omega \to \mathbb{R}$ is a concave function if $-f$ is a convex function.



Convex functions

- A set $\Omega \subset \mathbb{R}^n$ is said to be a convex set if $\forall a, b \in \Omega$, $ta + (1-t)b \in \Omega, \forall t \in [0,1]$.
- A function $f : \Omega \to \mathbb{R}$ is a convex function if Ω is convex and $\forall a, b \in \Omega, f(at + b(1-t)) \leq tf(a) + (1-t)f(b), \forall t \in [0,1]$.
- A function $f : \Omega \to \mathbb{R}$ is a concave function if $-f$ is a convex function.

$f(x) = 2x$, $f: \mathbb{R} \to \mathbb{R}$.



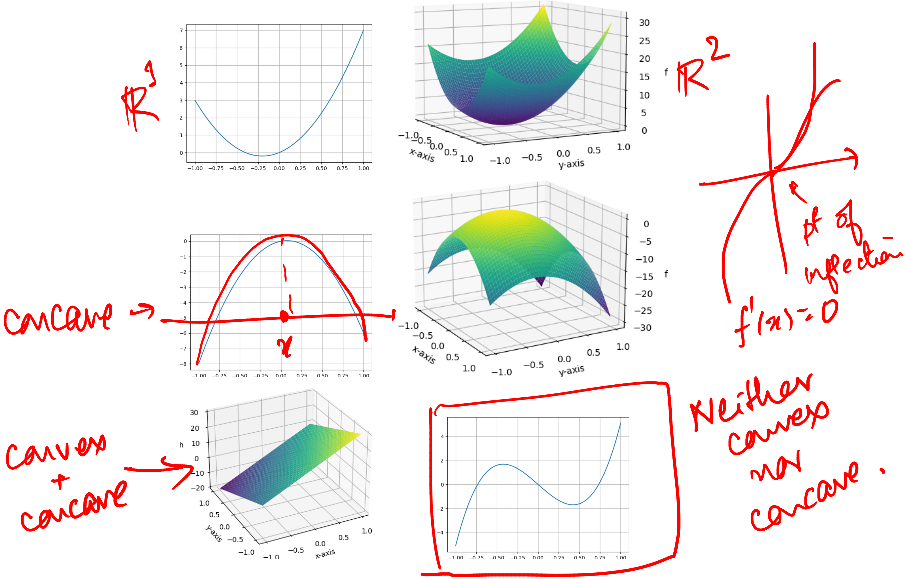


Figure: Row 1: Convex functions, Row 2: Concave functions, Row 3: (left) Convex and Concave function, (right) Neither convex nor concave

Why convexity is important?

Theorem

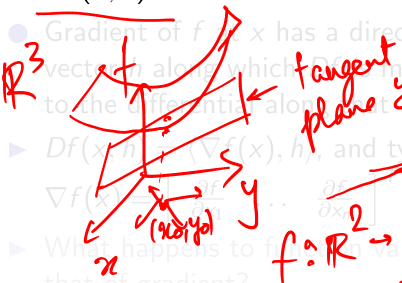
Let $f : \Omega \to \mathbb{R}$ be a convex function. Then any local minima $x^* \in \Omega$ of f is also a global minima

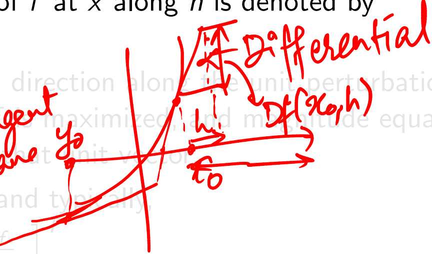
Proof.

- Let $y \in \Omega$ be any arbitrary point. Since x^* is a local minima, $f(x^*) \leq f(\alpha y + (1-\alpha)x^*)$ for small enough α
- $f(x^*) \leq f(\alpha y + (1-\alpha)x^*) \leq (1-\alpha)f(x^*) + \alpha f(y)$
- $\Rightarrow f(x^*) \leq f(y)$.

$\square$

- Differential of a function $f : \mathbb{R}^n \to \mathbb{R}$, at a point $x \in \mathbb{R}^n$ describes the change in the *linear approximation* of f at x along any perturbation of x. Denoting the perturbation by $h \in \mathbb{R}^n$, the differential of f at x along h is denoted by $Df(x, h)$.





- Gradient of f at x has a direction along the unit perturbation vector h along which Df is maximized, and magnitude equal to the differential along that unit vector.
  - $Df(x, h) = \langle \nabla f(x), h \rangle$, and typically, $\nabla f(x) = \left[\frac{\partial f}{\partial x_1} \dots \frac{\partial f}{\partial x_n} \right]$.
  - What happens to function values in directions orthogonal to that of gradient?

$f : \mathbb{R}^2 \to \mathbb{R}$

$\nabla f(x) \in \mathbb{R}^2$

Gradient of a function

- Differential of a function $f : \mathbb{R}^n \to \mathbb{R}$, at a point $x \in \mathbb{R}^n$ describes the change in the *linear approximation* of f at x along any perturbation of x. Denoting the perturbation by $h \in \mathbb{R}^n$, the differential of f at x along h is denoted by $Df(x, h)$.
- Gradient of f at x has a direction along the unit perturbation vector h along which Df is "maximized," and magnitude equal to the differential along that unit vector.
  - $Df(x, h) = \langle \nabla f(x), h \rangle$, and typically, $\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \dots & \frac{\partial f}{\partial x_n} \end{bmatrix}^T$.
  - What happens to function values in directions orthogonal to that of gradient?

- Differential of a function $f : \mathbb{R}^n \to \mathbb{R}$, at a point $x \in \mathbb{R}^n$ describes the change in the *linear approximation* of f at x along any perturbation of x. Denoting the perturbation by $h \in \mathbb{R}^n$, the differential of f at x along h is denoted by $Df(x,h)$.
- Gradient of f at x has a direction along the unit perturbation vector h along which Df is maximized, and magnitude equal to the differential along that unit vector.
  - $Df(x,h) = \langle \nabla f(x), h \rangle$, and typically, $\boxed{\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \dots & \frac{\partial f}{\partial x_n} \end{bmatrix}^T.}$
  - What happens to function values in directions orthogonal to that of gradient?

Gradient of a function

- Differential of a function $f : \mathbb{R}^n \to \mathbb{R}$, at a point $x \in \mathbb{R}^n$ describes the change in the *linear approximation* of f at x along any perturbation of x. Denoting the perturbation by $h \in \mathbb{R}^n$, the differential of f at x along h is denoted by $Df(x, h)$.
- Gradient of f at x has a direction along the unit perturbation vector h along which Df is maximized, and magnitude equal to the differential along that unit vector.
  - $Df(x, h) = \langle \nabla f(x), h \rangle$, and typically, $\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \dots & \frac{\partial f}{\partial x_n} \end{bmatrix}^T$.
  - What happens to function values in directions orthogonal to that of gradient?

Unconstrained Optimization

- Given $f : \mathbb{R}^n \to \mathbb{R}$, solve: $\arg\min_{x \in \mathbb{R}^n} f(x)$.
- Steepest Descent:
  1. Intialize $k = 0, x_k$
  2. Let $h_k = \nabla f(x_k)$
  3. $\alpha_k = \arg\min_{\alpha \in \mathbb{R}^+} f(x_k - \alpha h_k)$
  4. $x_{k+1} = x_k - \alpha_k h_k$
  5. Check Stopping criteria $[||x_k - x_{k+1}||, |f(x_k) - f(x_{k+1})|, f(x_{k+1}) > f(x_k)]$. If not, set $k = k + 1$, continue with Step 2.

$-h_k$

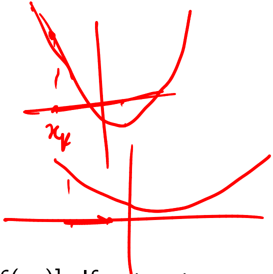
- Gradient Descent:
  1. Intialize $k = 0, x_k$
  2. Let $h_k = \nabla f(x_k)$
  3. $x_{k+1} = x_k - \alpha_k h_k$, α_k is chosen heuristically.
  4. Check stopping criteria. If not, set $k = k + 1$, continue with Step 2.
- Newton's Method: $h_k = (Hf(x_k))^{-1} \nabla f(x_k)$.

Unconstrained Optimization

- Given $f : \mathbb{R}^n \to \mathbb{R}$, solve: $\arg\min_{x \in \mathbb{R}^n} f(x)$.
- Steepest Descent:
  1. Intialize $k = 0, x_k$
  2. Let $h_k = \nabla f(x_k)$
  3. $\alpha_k = \underline{\arg\min_{\alpha \in \mathbb{R}^+} f(x_k - \alpha h_k)}$ X 1D optimization
  4. $x_{k+1} = x_k - \alpha_k h_k$
  5. Check Stopping criteria $[||x_k - x_{k+1}||, |f(x_k) - f(x_{k+1})|, f(x_{k+1}) > f(x_k)]$. If not, set $k = k + 1$, continue with Step 2.
- Gradient Descent:
  1. Intialize $k = 0, x_k$
  2. Let $h_k = \nabla f(x_k)$
  3. $x_{k+1} = \underline{x_k - \alpha_k h_k}$, $\underline{\alpha_k}$ is chosen heuristically.
  4. Check stopping criteria. If not, set $k = k + 1$, continue with Step 2.
- Newton's Method: $h_k = (Hf(x_k))^{-1} \nabla f(x_k)$.

Unconstrained Optimization

- Given $f : \mathbb{R}^n \to \mathbb{R}$, solve: $\arg\min_{x \in \mathbb{R}^n} f(x)$.



- **Steepest Descent:**
  1. Intialize $k = 0, x_k$
  2. Let $h_k = \nabla f(x_k)$
  3. $\alpha_k = \arg\min_{\alpha \in \mathbb{R}^+} f(x_k - \alpha h_k)$
  4. $x_{k+1} = x_k - \alpha_k h_k$
  5. Check Stopping criteria $[||x_k - x_{k+1}||, |f(x_k) - f(x_{k+1})|, f(x_{k+1}) > f(x_k)]$. If not, set $k = k + 1$, continue with Step 2.

$\alpha_k = 0.5$ $\boxed{\alpha_k = \frac{1}{2\|h_k\|}}$

- **Gradient Descent:**
  1. Intialize $k = 0, x_k$
  2. Let $h_k = \nabla f(x_k)$
  3. $x_{k+1} = x_k - \alpha_k h_k$, α_k is chosen heuristically.
  4. Check stopping criteria. If not, set $k = k + 1$, continue with Step 2.

Hessian

- **Newton's Method:** $h_k = (Hf(x_k))^{-1} \nabla f(x_k)$.

Constrained minimization

- Nonlinear constrained problem:

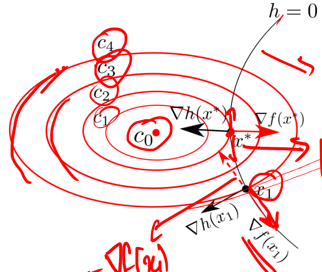
$$\begin{aligned} \min\ & f(x) \\ \text{sub. to } & h_1(x) = 0 & \qquad g_1(x) \leq 0 \\ & \vdots & \vdots \\ & h_m(x) = 0 & \qquad g_p(x) \leq 0 \\ & x \in \Omega \subset \mathbb{R}^n \end{aligned}$$

- Denote $h(x) := [h_1(x)\ h_2(x)\ \ldots\ h_m(x)]^T$ and $g(x) := [g_1(x)\ g_2(x)\ \ldots\ g_p(x)]^T$, with $h : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^n \to \mathbb{R}^p$. Typically, f, g, h are assumed to be $\mathcal{C}^2$.
- Constraints of the form $h(x) = 0$ and $g(x) \leq 0$ are called *Functional constraints*, while constraints of the form $x \in \Omega$ are called *Set constraints*.
- Points satisfying all constraints are called *feasible points*.

Equality constrained minimization

- Consider $h : \mathbb{R}^n \to \mathbb{R}$, a single equality constraint.

$h(x) = 0$



pt of minima

x_1

$\nabla h(x_1)$ T

$-\nabla f(x_1)$

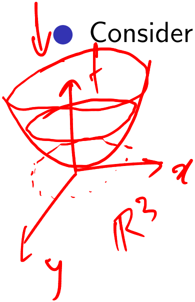
Figure: Equality constrained optimization. Assume $c_4 > c_3 > c_2 > c_1 > c_0$

$\lambda^* = \nabla h(x^*) \parallel \nabla f(x^*)$

- At x^*, $\nabla f(x^*)$ and $\nabla h(x^*)$ are linearly dependent, thus there exists a unique scalar λ^* such that $\nabla f(x^*) + \lambda^* \nabla h(x^*) = 0$.
- This unique λ^* is called **Lagrange Multiplier**.

Equality constrained minimization

- Consider $h : \mathbb{R}^n \to \mathbb{R}$, a single equality constraint.



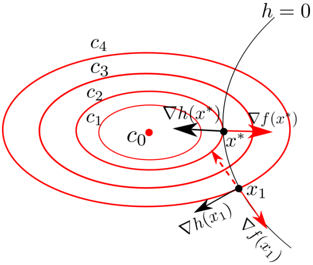


Figure: Equality constrained optimization. Assume $c_4 > c_3 > c_2 > c_1 > c_0$

- At x^*, $\nabla f(x^*)$ and $\nabla h(x^*)$ are linearly dependent, thus there exists a unique scalar λ^* such that $\nabla f(x^*) + \lambda^* \nabla h(x^*) = 0$.
- This unique λ^* is called **Lagrange Multiplier**.

Equality constrained minimization

- Consider $h : \mathbb{R}^n \to \mathbb{R}$, a single equality constraint.

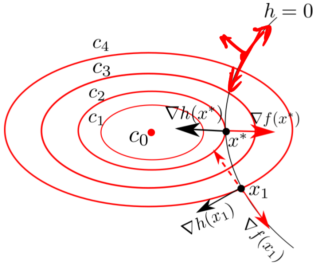


Figure: Equality constrained optimization. Assume $c_4 > c_3 > c_2 > c_1 > c_0$

- At x^*, $\nabla f(x^*)$ and $\nabla h(x^*)$ are linearly dependent, thus there exists a unique scalar λ^* such that $\nabla f(x^*) + \lambda^* \nabla h(x^*) = 0$.
- This unique λ^* is called **Lagrange Multiplier**.

- Define the Lagrangian as $L(x, \lambda) = f(x) + \lambda h(x)$.
- **First Order Necessary Condition:** If x^* is a local minima of the equality constrained minimization problem then, (a) $\nabla_x L(x^*, \lambda^*) = 0$, and (b) $\nabla_\lambda L(x*, \lambda^*) = 0$.

$$\nabla_x L(x^*, \lambda^*) = \nabla f(x^*) + \lambda^* \nabla h(x^*) = 0$$

$$\nabla_\lambda L(x^*, \lambda^*) = h(x^*) = 0$$

- Given multiple constraints $h_1, \ldots, h_m$, a point $x \in \Omega$ where $\{\nabla h_i(x), i = 1 \ldots, m\}$ are linearly independent is called a **regular point**.
- Denoting $\lambda \in \mathbb{R}^m$ as the vector of all lagrange multipliers, and the lagrangian defined as $L(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i h_i(x) = f(x) + \lambda^T h(x)$, we have

Theorem (First Order Necessary Condition(FONC))

If $x^* \in \Omega$ is a local minima, then $\nabla_x L(x^*, \lambda^*) = \nabla f(x^*) + (Jh(x^*))^T \lambda^* = 0$, and $\nabla_\lambda L(x*, \lambda^*) = h(x^*) = 0$.

Theorem (Second Order Necessary Condition(SONC))

If x^* is a local minima, and x^* is a regular point, then $\exists \lambda^* \in \mathbb{R}^m$ satisfying the FONC conditions, and $v^T \nabla_x^2 L(x^*, \lambda^*) v \geq 0, \forall v \in ker(Jh(x^*))$.

- Define the Lagrangian as $L(x, \lambda) = f(x) + \lambda h(x)$.
- **First Order Necessary Condition:** If x^* is a local minima of the equality constrained minimization problem then, (a) $\nabla_x L(x^*, \lambda^*) = 0$, and (b) $\nabla_\lambda L(x*, \lambda^*) = 0$.
- Given multiple constraints $h_1, \ldots, h_m$, a point $x \in \Omega$ where $\{\nabla h_i(x), i = 1 \ldots, m\}$ are linearly independent is called a **regular point**.
- Denoting $\lambda \in \mathbb{R}^m$ as the vector of all lagrange mulipliers, and the lagrangian defined as $L(x, \lambda) = f(x) + \sum_{i=1}^m \lambda_i h_i(x) = f(x) + \lambda^T h(x)$, we have

$h(x) = [h_1(x), \ldots, h_m(x)]^T$

Theorem (First Order Necessary Condition(FONC))

If x^* is a local minima, then $\nabla_x L(x^*, \lambda^*) = \nabla f(x^*) + (Jh(x^*))^T \lambda^* = 0$, and $\nabla_\lambda L(x*, \lambda^*) = h(x^*) = 0$.

$Hf \geq 0$

$\underbrace{[v^T]}_{1\times n} \underbrace{\left[\nabla_x^2 L\right]}_{n\times n} \underbrace{[v]}_{n\times 1} \geq 0$

Theorem (Second Order Necessary Condition(SONC))

If x^* is a local minima, and x^* is a regular point, then $\exists \lambda^* \in \mathbb{R}^m$ satisfying the FONC conditions, and $v^T \nabla_x^2 L(x^*, \lambda^*) v \geq 0, \forall v \in ker(Jh(x^*))$.

- Define the Lagrangian as $L(x, \lambda) = f(x) + \lambda h(x)$.
- **First Order Necessary Condition:** If x^* is a local minima of the equality constrained minimization problem then, (a) $\nabla_x L(x^*, \lambda^*) = 0$, and (b) $\nabla_\lambda L(x*, \lambda^*) = 0$.
- Given multiple constraints $h_1, \dots, h_m$, a point $x \in \Omega$ where $\{\nabla h_i(x), i = 1 \dots, m\}$ are linearly independent is called a **regular point**.
- Denoting $\lambda \in \mathbb{R}^m$ as the vector of all lagrange mulipliers, and the lagrangian defined as $L(x, \lambda) = f(x) + \sum_{i=1}^m \lambda_i h_i(x) = f(x) + \lambda^T h(x)$, we have

Theorem (First Order Necessary Condition(FONC))

If $x^* \in \Omega$ is a local minima, then $\nabla_x L(x^*, \lambda^*) = \nabla f(x^*) + (Jh(x^*))^T \lambda^* = 0$, and $\nabla_\lambda L(x*, \lambda^*) = h(x^*) = 0$

Jacobian

Theorem (Second Order Necessary Condition(SONC))

If x^* is a local minima, and x^* is a regular point, then $\exists \lambda^* \in \mathbb{R}^m$ satisfying the FONC conditions, and $v^T \nabla_x^2 L(x^*, \lambda^*) v \geq 0, \forall v \in ker(Jh(x^*))$.

$Jh(x^*) = \begin{bmatrix} - \nabla h_1 - \\ - \nabla h_2 - \\ \vdots \\ - \nabla h_m - \end{bmatrix}$

Equality & Inequality constraints

- For a feasible point x, the set of constraints $\{g_j \mid g_j(x) < 0\}$ are called *inactive*, while $\{g_i \mid g_i(x) = 0\}$ are called active.
- Let x be such that $h(x) = 0$ and $g(x) \leq 0$, with J being the set of indices for which $g_j(x) = 0, j \in J$. x is said to be *regular point* of the constraints if the set of vectors $\{\nabla h_i(x), 1 \leq i \leq m, \nabla g_j(x), j \in J\}$ are linearly independent.

Theorem (Karush-Kuhn-Tucker conditions (KKT))

Let x^* be a local minima for the objective f subject to $h(x) = 0$ and $g(x) \leq 0$, and assume x^* is a regular point for these constraints. Then $\exists \lambda^* \in \mathbb{R}^m$ and $\exists \mu^* \in \mathbb{R}^p$ with $\mu^* \geq 0$ such that

$$\nabla f(x^*) + (Jh(x^*))^T \lambda^* + (Jg(x^*))^T \mu^* = 0 \tag{1}$$

$$\mu^{*T} g(x^*) = 0 \tag{2}$$

Equality & Inequality constraints

- For a feasible point x, the set of constraints $\{g_j \mid g_j(x) < 0\}$ are called inactive, while $\{g_i \mid g_i(x) = 0\}$ are called active.
- Let x be such that $h(x) = 0$ and $g(x) \leq 0$, with J being the set of indices for which $g_j(x) = 0, j \in J$. x is said to be *regular point* of the constraints if the set of vectors $\{\nabla h_i(x), 1 \leq i \leq m, \nabla g_j(x), j \in J\}$ are linearly independent.

Theorem (Karush-Kuhn-Tucker conditions (KKT))

Let x^* be a local minima for the objective f subject to $h(x) = 0$ and $g(x) \leq 0$, and assume x^* is a regular point for these constraints. Then $\exists \lambda^* \in \mathbb{R}^m$ and $\exists \mu^* \in \mathbb{R}^p$ with $\mu^* \geq 0$ such that

$$\nabla f(x^*) + (Jh(x^*))^T \lambda^* + (Jg(x^*))^T \mu^* = 0 \qquad (1)$$

Complementary slackness

$$\mu^{*T} g(x^*) = 0 \qquad (2)$$

$\nabla f + \sum_{i=1}^{m} \lambda_i \nabla h_i(x) + \sum_{j \in J} \mu_j \nabla g_j(x^*) = 0$

$\mu_j = 0, j \notin J$

KKT Conditions

- Consider only inequality conditions $g := [g_1, g_2]$.
- If $\mu_2 < 0$, there exists $v \perp \nabla g_1(x_0)$ such that $\langle v, \nabla g_2(x_0)\rangle \cancel{>} < 0$. Moreover,

$\langle \nabla g_1, v\rangle = 0$

$\langle v, \nabla g_2\rangle$

$$\langle \nabla f(x_0) + (Jg(x_0))^T \mu^*, v\rangle = \langle \nabla f(x_0), v\rangle = -\mu_2 \langle \nabla g_2(x_0), v\rangle < 0$$

$\leq 0 \qquad < 0 \qquad =$

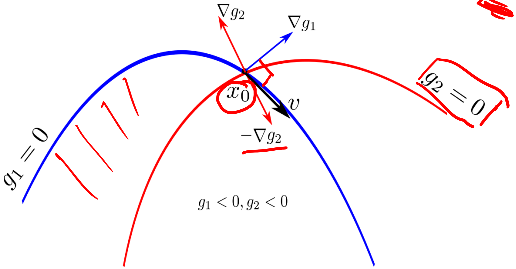


Figure: KKT Conditions

Second order conditions

Theorem (SONC)

Let $f, g, h \in \mathcal{C}^2$ and x^* is a regular point of the constraints. If x^* is a local minimum, then $\exists \lambda^* \in \mathbb{R}^m, \exists \mu^* \in \mathbb{R}^p, \mu^* \geq 0$ such that the FONC are satisfied and $\nabla_x^2 L(x^*, \lambda^*, \mu^*) = \nabla_x^2 f(x^*) + {\lambda^*}^T \nabla_x^2 h(x^*) + {\mu^*}^T \nabla_x^2 g(x^*)$ is positive semi-definite on $T_{x^*}S$.

$v \in \ker([J h \; g])$ $\langle v, \nabla h_i \rangle = 0$ $i = 1, \cdots, m$

$A \in \mathbb{R}^{m \times n}$

$x, Ax = 0 \longrightarrow$ Nullspace

Constrained Optimization Algorithms

- We know what conditions a minima satisfies, but how do we find such points?
- Consider the inequality constrained optimization problem for $f : \mathbb{R}^n \to \mathbb{R}, g : \mathbb{R}^n \to \mathbb{R}^m$,

$$\inf_{x \in X \subset \mathbb{R}^n} f(x)$$
$$g(x) \leq 0$$

  - Rewriting,

$$\inf_{x \in X} \sup_{\mu \geq 0} f(x) + \mu^T g(x)$$

- We know what conditions a minima satisfies, but how do we find such points?
- Consider the inequality constrained optimization problem for $f : \mathbb{R}^n \to \mathbb{R}, g : \mathbb{R}^n \to \mathbb{R}^m$,

$$\inf_{x \in X \subset \mathbb{R}^n} f(x)$$
$$g(x) \leq 0$$

$h(x) = 0$

$\Downarrow$

$h(x) \leq 0$ g_1

$h(x) \geq 0$

$\Downarrow$

$-h(x) \leq 0$ g_2

- Rewriting,

$$\inf_{x \in X} \sup_{\mu \geq 0} f(x) + \mu^T g(x)$$

- We know what conditions a minima satisfies, but how do we find such points?
- Consider the inequality constrained optimization problem for $f : \mathbb{R}^n \to \mathbb{R}, g : \mathbb{R}^n \to \mathbb{R}^m$,

$$\boxed{\begin{gathered}\inf_{x \in X \subset \mathbb{R}^n} f(x) \\ g(x) \leq 0\end{gathered}}$$

$x, \; \underline{g(x) < 0}$

- Rewriting,

$\mu \cdot g(x)$

Max $f(x)+\mu^T g(x)$
$\mu \geq 0$

< 0

$$\inf_{x \in X} \left[\sup_{\mu \geq 0} \left\{ f(x) + \mu^T g(x) \right\}\right]$$

∞

$\underline{\mu < 0} \rightarrow$

$f(x)$

Duality

- Let $L(x, \mu) = f(x) + \mu^T g(x)$.
  - Primal problem:

$$p^* = \inf_{x \in X} \sup_{\mu \geq 0} L(x, \mu)$$

  - If we switch the order of inf and sup, we get the dual problem:

$$d^* = \sup_{\mu \geq 0} \inf_{x \in X} L(x, \mu) = \sup_{\mu \geq 0} q(\mu),$$

    where q is called the *dual function*.
  - Weak Duality:

$$L(x, \mu) \leq \sup_{\mu \geq 0} L(x, \mu)$$

$$\inf_{x \in X} L(x, \mu) \leq \inf_{x \in X} \sup_{\mu \geq 0} L(x, \mu) = p^*$$

$$\Rightarrow \quad d^* \leq p^*$$

- When do we have $d^* = p^*$ (Strong Duality)?

Duality

- Let $L(x, \mu) = f(x) + \mu^T g(x)$.
  - Primal problem:
$$p^* = \inf_{x \in X} \sup_{\mu \geq 0} L(x, \mu)$$
  - If we switch the order of inf and sup, we get the dual problem:
$$d^* = \sup_{\mu \geq 0} \inf_{x \in X} L(x, \mu) = \sup_{\mu \geq 0} q(\mu),$$
    where q is called the *dual function*.
  - Weak Duality:
$$L(x, \mu) \leq \sup_{\mu \geq 0} L(x, \mu)$$
$$\inf_{x \in X} L(x, \mu) \leq \inf_{x \in X} \sup_{\mu \geq 0} L(x, \mu) = p^*$$
$$\Rightarrow \quad d^* \leq p^*$$
- When do we have $d^* = p^*$ (Strong Duality)?

Duality

- Let $L(x,\mu) = f(x) + \mu^T g(x)$.
  - Primal problem:

$$p^* = \inf_{x \in X} \sup_{\mu \geq 0} L(x,\mu)$$

  - If we switch the order of inf and sup, we get the dual problem:

$$d^* = \sup_{\mu \geq 0} \inf_{x \in X} L(x,\mu) = \sup_{\mu \geq 0} q(\mu),$$

    where q is called the *dual function*.
  - Weak Duality:

$$L(x,\mu) \leq \sup_{\mu \geq 0} L(x,\mu)$$

$$\inf_{x \in X} L(x,\mu) \leq \inf_{x \in X} \sup_{\mu \geq 0} L(x,\mu) = p^*$$

$$\Rightarrow \quad d^* \leq p^*$$

- When do we have $d^* = p^*$ (Strong Duality)?

Duality

- Let $L(x,\mu) = f(x) + \mu^T g(x)$.
  - Primal problem:
  $$p^* = \inf_{x \in X} \sup_{\mu \geq 0} L(x,\mu)$$
  - If we switch the order of inf and sup, we get the dual problem:
  $$d^* = \sup_{\mu \geq 0} \inf_{x \in X} L(x,\mu) = \sup_{\mu \geq 0} q(\mu),$$
  where q is called the *dual function*.
  - Weak Duality:
  $$\Rightarrow \quad L(x,\mu) \leq \sup_{\mu \geq 0} L(x,\mu)$$
  $$\forall \mu \quad \inf_{x \in X} L(x,\mu) \leq \inf_{x \in X} \sup_{\mu \geq 0} L(x,\mu) = p^*$$
  $$\Rightarrow \quad d^* \leq p^*$$
  $$\sup_{\mu} p^* = p^*$$
- When do we have $d^* = p^*$ (Strong Duality)?

$d^* = p^*$

Duality

- Let $L(x, \mu) = f(x) + \mu^T g(x)$.
  - Primal problem:

$$\boxed{p^* = \inf_{x \in X} \sup_{\mu \geq 0} L(x, \mu)}$$

  - If we switch the order of inf and sup, we get the dual problem:

Dual Algorithm

$$d^* = \sup_{\mu \geq 0} \inf_{x \in X} L(x, \mu) = \sup_{\mu \geq 0} \boxed{q(\mu),}$$

where q is called the *dual function*.

  - Weak Duality:

$$L(x, \mu) \leq \sup_{\mu \geq 0} L(x, \mu)$$

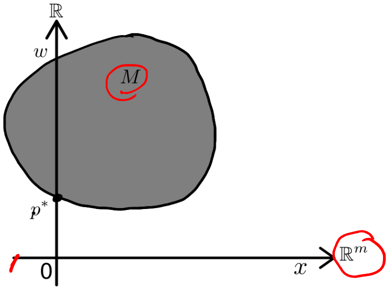
$$\inf_{x \in X} L(x, \mu) \leq \inf_{x \in X} \sup_{\mu \geq 0} L(x, \mu) = p^*$$

$$\Rightarrow \quad d^* \leq p^*$$

- When do we have $\boxed{d^* = p^*}$ (Strong Duality)?

Geometric Interpretation

- Consider a set $M \subset \mathbb{R}^{m+1}$. # of ineq. constraints
- Find the infimum of y coordinate of the points in $M \cap \{(0, w), w \in \mathbb{R}\}$.



m=1.

$$p^* = \inf_{(0,w)\in M} w$$

Figure: Note that the x axis represents $\mathbb{R}^m$

- Let $p^* = \inf_{(0,w)\in M} w$

Geometric Interpretation

- Consider a set $M \subset \mathbb{R}^{m+1}$.
- Find the infimum of y coordinate of the points in $M \cap \{(0, w), w \in \mathbb{R}\}$.

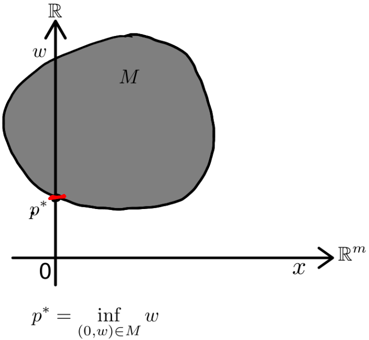


Figure: Note that the x axis represents $\mathbb{R}^m$

- Let $p^* = \inf_{(0,w) \in M} w$

Geometric Interpretation

- Consider a set $M \subset \mathbb{R}^{m+1}$.
- Find the infimum of y coordinate of the points in $M \cap \{(0, w), w \in \mathbb{R}\}$.

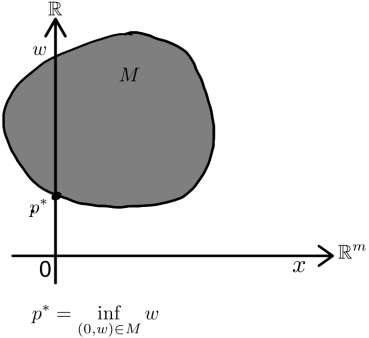
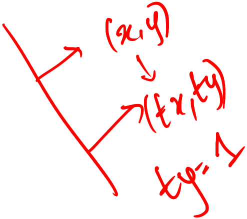


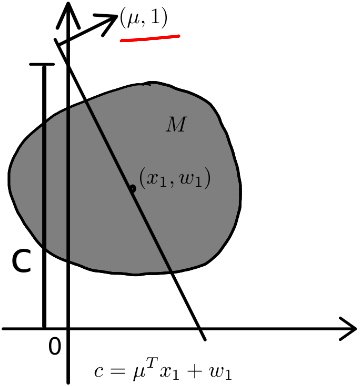
Figure: Note that the x axis represents $\mathbb{R}^m$

- Let $p^* = \inf_{(0,w)\in M} w$

line in $\mathbb{R}^2$

- Consider a non-vertical ~~hyperplane~~ in $\mathbb{R}^m$, with normal $(\mu, 1)$.
  - If this hyperplane passes through a point $(x_1, w_1) \in M$, what is the intercept?





m=1

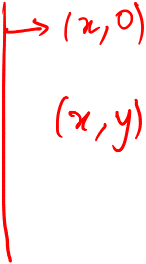


Figure: Note that the x axis represents $\mathbb{R}^m$

- Consider a non-vertical hyperplane in $\mathbb{R}^m$, with normal $(\mu, 1)$.
  - If this hyperplane passes through a point $(x_1, w_1) \in M$, what is the intercept?

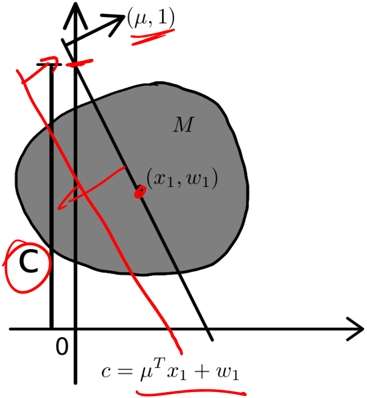


Figure: Note that the x axis represents $\mathbb{R}^m$

- For hyperplanes with same normal $(\mu, 1)$, what is the smallest intercept such that the hperplance passes through a point in M.
  - Denote the value of intercept by $q(\mu)$.

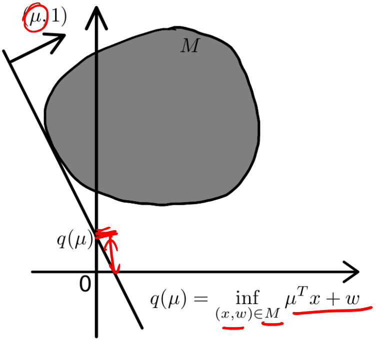


Figure: Note that the x axis represents $\mathbb{R}^m$

- Maximize the intercept by varying the normal vector $(\mu, 1), \forall \mu \in \mathbb{R}^m$.
  - Denote the maximum intercept by $d^* = \sup_{\mu \in \mathbb{R}^m} q(\mu)$.

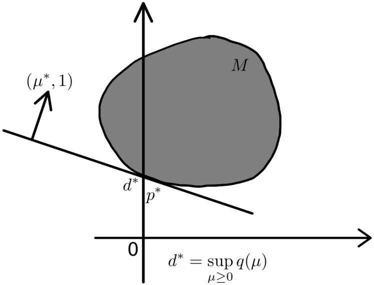


Figure: Note that the x axis represents $\mathbb{R}^m$

  - What is the relation between d^* and p^*?

- Maximize the intercept by varying the normal vector $(\mu, 1), \forall \mu \in \mathbb{R}^m$.
  - Denote the maximum intercept by $d^* = \sup_{\mu \in \mathbb{R}^m} q(\mu)$.

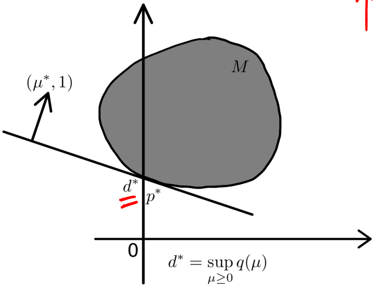


Figure: Note that the x axis represents $\mathbb{R}^m$

- What is the relation between d^* and p^*?

- Maximize the intercept by varying the normal vector $(\mu, 1), \forall \mu \in \mathbb{R}^m$.
  - Denote the maximum intercept by $d^* = \sup_{\mu \in \mathbb{R}^m} q(\mu)$.

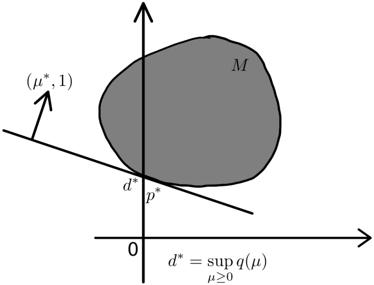


Figure: Note that the x axis represents $\mathbb{R}^m$

  - What is the relation between d^* and p^*?

- $d^* \leq p^*$.

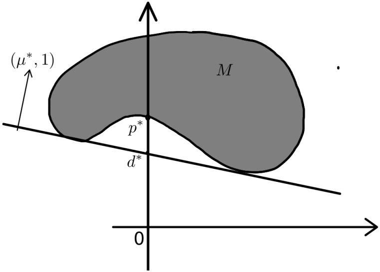


Figure: Note that the x axis represents $\mathbb{R}^m$

How is this related to optimization?

- Consider the set $\overline{M} = \{(g(x), f(x)), \forall x \in X\}$, and $M = \{(x, w) \in \mathbb{R}^{m+1} \mid \exists u \in X, g(u) \leq x, f(u) \leq w\}$.

$\inf_{x \in X} f(x)$

$g(x) \leq 0$

$g(x) \in \mathbb{R}^m$

$f(x) \in \mathbb{R}$

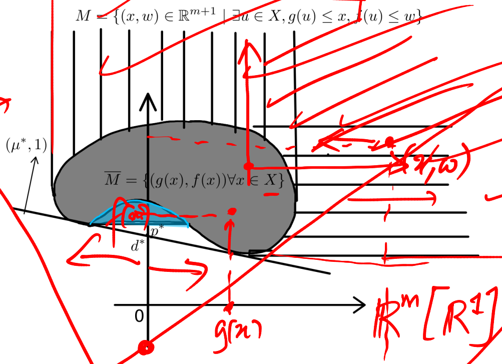


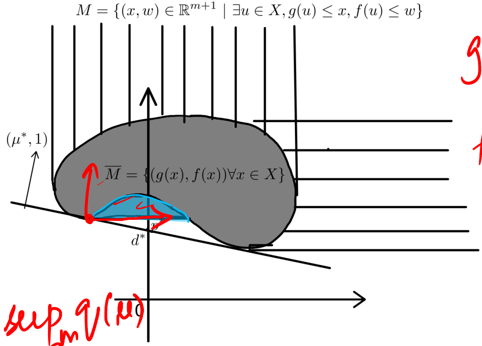
Figure: Note that the x axis represents $\mathbb{R}^m$

$q(\mu) = \inf_{(x,w) \in M} \{\mu^T x + w\} = -\infty$

- Note that for any $\mu \not\geq 0$, $q(\mu) = -\infty$.

How is this related to optimization?

- Consider the set $\overline{M} = \{(g(x), f(x)), \forall x \in X\}$, and $M = \{(x, w) \in \mathbb{R}^{m+1} \mid \exists u \in X, g(u) \leq x, f(u) \leq w\}$.



g(x) ~ Interpret as x-coord

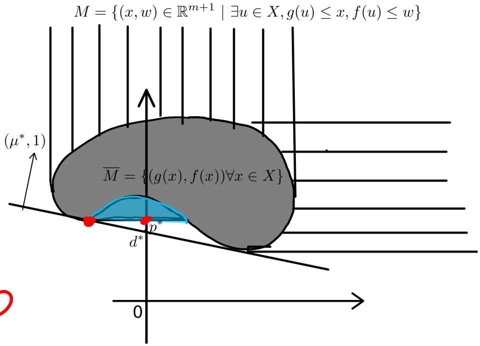
f(x) − y coordinate

p^*, d^*

$d^* = \sup_{\mu \in \mathbb{R}^m} q(\mu)$

Figure: Note that the x axis represents $\mathbb{R}^m$

- Note that for any $\mu \ngeq 0$, $q(\mu) = -\infty$.



$g(x) = 0$

Figure: Note that the x axis represents $\mathbb{R}^m$

- Note that $p^* = \inf_{(0,w)\in M} w = \inf_{x \in X, g(x) \leq 0} f(x)$.
- Also, $q(\mu) = \inf_{x \in X} f(x) + \mu^T g(x) = \inf_{x \in X} L(x, \mu)$.
- Dual problem: $d^* = \sup_{\mu \geq 0} q(\mu)$.

Primal Problem

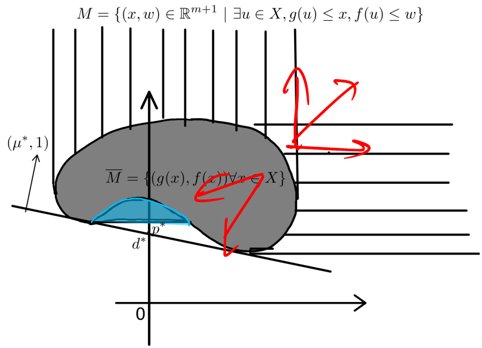


Figure: Note that the x axis represents $\mathbb{R}^m$

- Note that $p^* = \inf_{(0,w) \in M} w = \inf_{x \in X, g(x) \leq 0} f(x)$.
- Also, $q(\mu) = \inf_{x \in X} f(x) + \mu^T g(x) = \inf_{x \in X} L(x, \mu)$.
- Dual problem: $d^* = \sup_{\mu \geq 0} q(\mu)$.

$q(\mu) = \inf_{(x,w) \in M} \{\mu^T x + w\}$ $f(x)$ $g(x)$

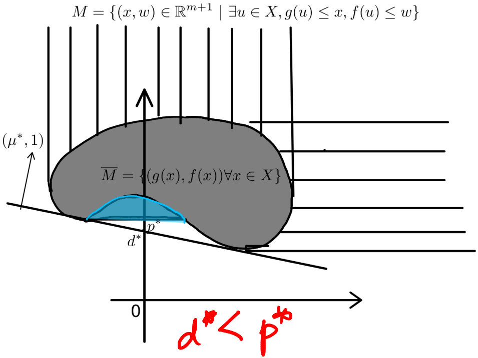


Figure: Note that the x axis represents $\mathbb{R}^m$

- Note that $p^* = \inf_{(0,w)\in M} w = \inf_{x\in X, g(x)\le 0} f(x)$.
- Also, $q(\mu) = \inf_{x\in X} f(x) + \mu^T g(x) = \inf_{x\in X} L(x,\mu)$.
- Dual problem: $d^* = \sup_{\mu \ge 0} q(\mu)$.

Strong Duality

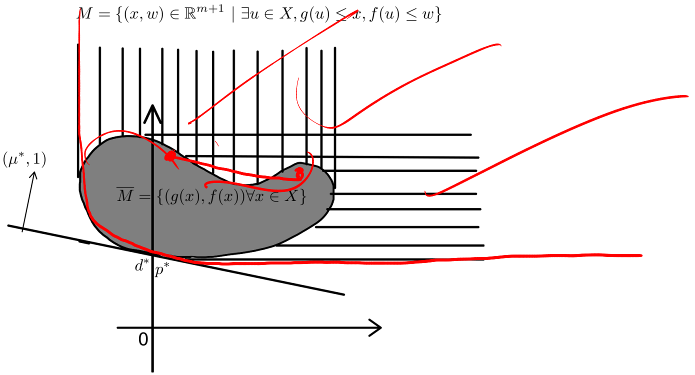


Figure: Note that the x axis represents $\mathbb{R}^m$

- Strong Duality may hold even if $\overline{M}$ is not convex.

Convex Programming Duality

- Let $f, g_1, \ldots, g_m$ be convex functions, and $X \subset \mathbb{R}^n$ be a convex set.
  - Solve $\arg\min_{x \in X, g(x) \leq 0} f(x)$.

Theorem (Convex Programming Duality)

If $\inf_{x \in X, g(x) \leq 0} f(x)$ *is finite, and there exists an* $x \in X$ *such that* $g(x) < 0$*, then* $p^* = d^*$*, and the solution to the dual problem exists.*

- Let $f, g_1, \ldots, g_m$ be convex functions, and $X \subset \mathbb{R}^n$ be a convex set.
  - Solve $\arg\min_{x \in X, g(x) \leq 0} f(x)$.

p^*

Theorem (Convex Programming Duality)

If $\inf_{x \in X, g(x) \leq 0} f(x)$ is finite, and there exists an $x \in X$ such that $g(x) < 0$, then $p^* = d^*$, and the solution to the dual problem exists.

Dual Ascent Algorithm

- $\sup_{\mu \geq 0} \inf_{x \in X} f(x) + \mu^T g(x)$
- Let (x_0, μ_0) be initial estimates of the primal and dual solution, and set $k = 0$
- Estimate for primal variable: $x_{k+1} = \arg\min_{x \in X} f(x) + \mu_k^T g(x)$
- Estimate for dual variable: $\mu_{k+1} = \arg\max_{\mu \geq 0} f(x_{k+1}) + \mu g(x_{k+1})$
  - $\mu_{k+1} = \max(0, \mu_k + \alpha_k g(x_{k+1}))$.
  - Stop, if stopping criteria satisfied, else $k = k + 1$ and repeat.

Dual Ascent Algorithm

- $\sup_{\mu \geq 0} \inf_{x \in X} f(x) + \mu^T g(x)$
- Let (x_0, μ_0) be initial estimates of the primal and dual solution, and set $k = 0$
- Estimate for primal variable: $x_{k+1} = \arg\min_{x \in X} f(x) + \mu_k^T g(x)$
- Estimate for dual variable: $\mu_{k+1} = \arg\max_{\mu \geq 0} f(x_{k+1}) + \mu g(x_{k+1})$
  - $\mu_{k+1} = \max(0, \mu_k + \alpha_k g(x_{k+1}))$.
  - Stop, if stopping criteria satisfied, else $k = k + 1$ and repeat.

Dual Ascent Algorithm

- $\sup_{\mu \geq 0} \inf_{x \in X} f(x) + \mu^T g(x)$
- Let (x_0, μ_0) be initial estimates of the primal and dual solution, and set $k = 0$
- Estimate for primal variable: $x_{k+1} = \arg\min_{x \in X} f(x) + \mu_k^T g(x)$
- Estimate for dual variable: $\mu_{k+1} = \arg\max_{\mu \geq 0} f(x_{k+1}) + \mu g(x_{k+1})$
  - $\mu_{k+1} = \max(0, \mu_k + \alpha_k g(x_{k+1}))$.
  - Stop, if stopping criteria satisfied, else $k = k + 1$ and repeat.

Dual Ascent Algorithm

- $\sup_{\mu \geq 0} \inf_{x \in X} f(x) + \mu^T g(x)$
- Let (x_0, μ_0) be initial estimates of the primal and dual solution, and set $k = 0$
- Estimate for primal variable: $x_{k+1} = \arg\min_{x \in X} f(x) + \mu_k^T g(x)$
- Estimate for dual variable: $\mu_{k+1} = \arg\max_{\mu \geq 0} f(x_{k+1}) + \mu g(x_{k+1})$
  - $\mu_{k+1} = \max(0, \mu_k + \alpha_k g(x_{k+1}))$.
  - Stop, if stopping criteria satisfied, else $k = k + 1$ and repeat.

Dual Ascent Algorithm

- $\sup_{\mu \geq 0} \inf_{x \in X} f(x) + \mu^T g(x)$
- Let (x_0, μ_0) be initial estimates of the primal and dual solution, and set $k = 0$
- Estimate for primal variable: $x_{k+1} = \arg\min_{x \in X} f(x) + \mu_k^T g(x)$
- Estimate for dual variable: $\mu_{k+1} = \arg\max_{\mu \geq 0} f(x_{k+1}) + \mu g(x_{k+1})$
  - $\mu_{k+1} = \max(0, \underbrace{\mu_k + \alpha_k g(x_{k+1})}_{\text{gradient ascent}})$.
  - Stop, if stopping criteria satisfied, else $k = k + 1$ and repeat.

Dual Ascent Algorithm

- $\sup_{\mu \geq 0} \inf_{x \in X} f(x) + \mu^T g(x)$
- Let (x_0, μ_0) be initial estimates of the primal and dual solution, and set $k = 0$
- Estimate for primal variable: $x_{k+1} = \arg\min_{x \in X} f(x) + \mu_k^T g(x)$
- Estimate for dual variable: $\mu_{k+1} = \arg\max_{\mu \geq 0} f(x_{k+1}) + \mu g(x_{k+1})$
- $\mu_{k+1} = \max(0, \mu_k + \alpha_k g(x_{k+1}))$.
- Stop, if stopping criteria satisfied, else $k = k + 1$ and repeat.

- Other variants:
  - Penalty Method.
  - Augmented Lagrangian Method (ALM)
- References:
  - Introuction to Optimization, Chong & Zak, 4th ed., Wiley.
  - Linear and Nonlinear Programming, David Luenberger, 2nd ed., Springer
  - Convex Optimization Theory, Dimitri Bertsekas, World Scientific.
  - Convex Optimization, Boyd & Vanderberghe, Cambridge University Press.

- Other variants:
  - Penalty Method.
  - Augmented Lagrangian Method (ALM) → Alternating Direction Multiplier Method (ADMM)
- References:
  - Introuction to Optimization, Chong & Zak, 4th ed., Wiley
  - Linear and Nonlinear Programming, David Luenberger, 2nd ed., Springer
  - Convex Optimization Theory, Dimitri Bertsekas, World Scientific.
  - Convex Optimization, Boyd & Vanderberghe, Cambridge University Press.

Convex Optimization

- Other variants:
  - Penalty Method.
  - Augmented Lagrangian Method (ALM)
- References:
  - Introuction to Optimization, Chong & Zak, 4th ed., Wiley.
  - Linear and Nonlinear Programming, David Luenberger, 2nd ed., Springer
  - Convex Optimization Theory, Dimitri Bertsekas, World Scientific.
  - Convex Optimization, Boyd & Vanderberghe, Cambridge University Press.

Convex Optimization

- Other variants:
  - Penalty Method.
  - Augmented Lagrangian Method (ALM)
- References:
  - Introuction to Optimization, Chong & Zak, 4th ed., Wiley.
  - Linear and Nonlinear Programming, David Luenberger, 2nd ed., Springer
  - Convex Optimization Theory, Dimitri Bertsekas, World Scientific.
  - Convex Optimization, Boyd & Vanderberghe, Cambridge University Press.

- Other variants:
  - Penalty Method.
  - Augmented Lagrangian Method (ALM)
- References:
  - Introuction to Optimization, Chong & Zak, 4th ed., Wiley.
  - Linear and Nonlinear Programming, David Luenberger, 2nd ed., Springer
  - Convex Optimization Theory, Dimitri Bertsekas, World Scientific.
  - Convex Optimization, Boyd & Vanderberghe, Cambridge University Press.

Convex Optimization

- Other variants:
  - Penalty Method.
  - Augmented Lagrangian Method (ALM)
- References:
  - Introuction to Optimization, Chong & Zak, 4th ed., Wiley.
  - Linear and Nonlinear Programming, David Luenberger, 2nd ed., Springer
  - Convex Optimization Theory, Dimitri Bertsekas, World Scientific.
  - Convex Optimization, Boyd & Vanderberghe, Cambridge University Press.

Linear Algebra

Vector Space

- A set V with two binary operations $+$ and $\cdot$, is said to be a *Vector space over* $\mathbb{F}$ if
- V is closed under $+$, i.e., $\forall x, y \in V, x + y \in V$
- V is closed under $\cdot$, i.e., $\forall x \in V, \forall a \in \mathbb{F}, a \cdot x \in V$
- Commutativity: $\forall x, y \in V, x + y = y + x$
- Associativity: $\forall x, y, z \in V, (x + y) + z = x + (y + z)$
- Additive Identity: $\exists \mathbf{0} \in V, \forall x \in V, x + \mathbf{0} = x$.
- Additive Inverse: $\forall x \in V, \exists y \in V$ such that $x + y = \mathbf{0}$.
- Multiplicative Identity: $\forall x \in V, 1 \cdot x = x$.
- Distributive: $a \cdot (x + y) = a \cdot x + a \cdot y, (a + b) \cdot x = a \cdot x + b \cdot y, \forall a, b \in \mathbb{F}, \forall x, y \in V$.

- If $(V, +, \cdot)$ satisfies these properties, V is said to be a vector space over $\mathbb{F}$.
  - Any element $x \in V$ is called a *vector*, $+$ is called vector addition and $\cdot$ is caled *scalar multiplication*.
  - For this course, assume $\mathbb{F} = \mathbb{R}$ (*Real* vector space) or $\mathbb{C}$ (*Complex* vector space).

Examples

- $\mathbb{R}^2$:
  - $x = (x_1, x_2), y = (y_1, y_2), x + y = (x_1 + y_1, x_2 + y_2)$
  - $x = (x_1, x_2), a \in \mathbb{R}, a \cdot x = (ax_1, ax_2)$

Examples

- $\mathbb{R}^2$:
  - $x = (x_1, x_2), y = (y_1, y_2), x + y = (x_1 + y_1, x_2 + y_2)$
  - $x = (x_1, x_2), a \in \mathbb{R}, a \cdot x = (ax_1, ax_2)$

Examples

- $\mathbb{R}^2$:
  - $x = (x_1, x_2), y = (y_1, y_2), x + y = (x_1 + y_1, x_2 + y_2)$
  - $x = (x_1, x_2), a \in \mathbb{R}, a \cdot x = (ax_1, ax_2)$

- $\mathbb{R}^n$ or $\mathbb{C}^n$:
  - $x = (x_1, x_2, \ldots, x_n), y = (y_1, y_2, \ldots, y_n),$
    $x + y = (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n)$
  - $x = (x_1, x_2, \ldots, x_n), a \in \mathbb{R}$ (or $\mathbb{C}$), $a \cdot x = (ax_1, ax_2, \ldots, ax_n)$
- $V =$ Set of real-valued functions defined on $[0, 1] \subset \mathbb{R}$.
  - $f, g \in V, (f + g)(x) = f(x) + g(x), x \in [0, 1]$.
  - $f \in V, a \in \mathbb{R}, (a \cdot f)(x) = af(x), x \in [0, 1]$.
- V is the set of all polynomials on one real variable (say x) with degree at most 2.
  - $p, q \in V, p + q = p_0 + q_0 + (p_1 + q_1)x + (p_2 + q_2)x^2$
  - $p \in V, a \in \mathbb{R}, a \cdot p = ap_0 + ap_1x + ap_2x^2$

- $\mathbb{R}^n$ or $\mathbb{C}^n$:
  - $x = (x_1, x_2, \ldots, x_n), y = (y_1, y_2, \ldots, y_n)$, $x + y = (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n)$
  - $x = (x_1, x_2, \ldots, x_n), a \in \mathbb{R}$ (or $\mathbb{C}$), $a \cdot x = (ax_1, ax_2, \ldots, ax_n)$
- $V =$ Set of real-valued functions defined on $[0, 1] \subset \mathbb{R}$.
  - $f, g \in V, (f + g)(x) = f(x) + g(x), x \in [0, 1]$.
  - $f \in V, a \in \mathbb{R}, (a \cdot f)(x) = af(x), x \in [0, 1]$.
- V is the set of all polynomials on one real variable (say x) with degree at most 2.
  - $p, q \in V, p + q = p_0 + q_0 + (p_1 + q_1)x + (p_2 + q_2)x^2$
  - $p \in V, a \in \mathbb{R}, a \cdot p = ap_0 + ap_1x + ap_2x^2$

- $\mathbb{R}^n$ or $\mathbb{C}^n$:
  - $x = (x_1, x_2, \ldots, x_n), y = (y_1, y_2, \ldots, y_n)$,
    $x + y = (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n)$
  - $x = (x_1, x_2, \ldots, x_n), a \in \mathbb{R}$ (or $\mathbb{C}$), $a \cdot x = (ax_1, ax_2, \ldots, ax_n)$
- $V =$ Set of real-valued functions defined on $[0, 1] \subset \mathbb{R}$.
  - $f, g \in V, (f + g)(x) = f(x) + g(x), x \in [0, 1]$.
  - $f \in V, a \in \mathbb{R}, (a \cdot f)(x) = af(x), x \in [0, 1]$.
- V is the set of all polynomials on one real variable (say x) with degree at most 2.
  - $p, q \in V, p + q = p_0 + q_0 + (p_1 + q_1)x + (p_2 + q_2)x^2$
  - $p \in V, a \in \mathbb{R}, a \cdot p = ap_0 + ap_1x + ap_2x^2$

- $\mathbb{R}^n$ or $\mathbb{C}^n$:
  - $x = (x_1, x_2, \ldots, x_n), y = (y_1, y_2, \ldots, y_n)$, $x + y = (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n)$
  - $x = (x_1, x_2, \ldots, x_n), a \in \mathbb{R}$ (or $\mathbb{C}$), $a \cdot x = (ax_1, ax_2, \ldots, ax_n)$
- $V =$ Set of real-valued functions defined on $[0, 1] \subset \mathbb{R}$.
  - $f, g \in V, (f + g)(x) = f(x) + g(x), x \in [0, 1]$.
  - $f \in V, a \in \mathbb{R}, (a \cdot f)(x) = af(x), x \in [0, 1]$.
- V is the set of all polynomials on one real variable (say x) with degree at most 2.
  - $p, q \in V, p + q = p_0 + q_0 + (p_1 + q_1)x + (p_2 + q_2)x^2$
  - $p \in V, a \in \mathbb{R}, a \cdot p = ap_0 + ap_1x + ap_2x^2$

- $\mathbb{R}^n$ or $\mathbb{C}^n$:
  - $x = (x_1, x_2, \dots, x_n), y = (y_1, y_2, \dots, y_n)$, $x + y = (x_1 + y_1, x_2 + y_2, \dots, x_n + y_n)$
  - $x = (x_1, x_2, \dots, x_n), a \in \mathbb{R}$ (or $\mathbb{C}$), $a \cdot x = (ax_1, ax_2, \dots, ax_n)$
- V = Set of real-valued functions defined on $[0, 1] \subset \mathbb{R}$.
  - $f, g \in V, (f + g)(x) = f(x) + g(x), x \in [0, 1]$.
  - $f \in V, a \in \mathbb{R}, (a \cdot f)(x) = af(x), x \in [0, 1]$.
- V is the set of all polynomials on one real variable (say x) with degree at most 2.
  - $p, q \in V, p + q = p_0 + q_0 + (p_1 + q_1)x + (p_2 + q_2)x^2$
  - $p \in V, a \in \mathbb{R}, a \cdot p = ap_0 + ap_1x + ap_2x^2$

- $\mathbb{R}^n$ or $\mathbb{C}^n$:
  - $x = (x_1, x_2, \ldots, x_n), y = (y_1, y_2, \ldots, y_n)$, $x + y = (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n)$
  - $x = (x_1, x_2, \ldots, x_n), a \in \mathbb{R}$ (or $\mathbb{C}$), $a \cdot x = (ax_1, ax_2, \ldots, ax_n)$
- $V =$ Set of real-valued functions defined on $[0, 1] \subset \mathbb{R}$.
  - $f, g \in V, (f + g)(x) = f(x) + g(x), x \in [0, 1]$.
  - $f \in V, a \in \mathbb{R}, (a \cdot f)(x) = af(x), x \in [0, 1]$.
- V is the set of all polynomials on one real variable (say x) with degree at most 2.
  - $p, q \in V, p + q = p_0 + q_0 + (p_1 + q_1)x + (p_2 + q_2)x^2$
  - $p \in V, a \in \mathbb{R}, a \cdot p = ap_0 + ap_1x + ap_2x^2$

- $\mathbb{R}^n$ or $\mathbb{C}^n$:
  - $x = (x_1, x_2, \ldots, x_n), y = (y_1, y_2, \ldots, y_n)$, $x + y = (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n)$
  - $x = (x_1, x_2, \ldots, x_n), a \in \mathbb{R}$ (or $\mathbb{C}$), $a \cdot x = (ax_1, ax_2, \ldots, ax_n)$
- $V =$ Set of real-valued functions defined on $[0, 1] \subset \mathbb{R}$.
  - $f, g \in V, (f + g)(x) = f(x) + g(x), x \in [0, 1]$.
  - $f \in V, a \in \mathbb{R}, (a \cdot f)(x) = af(x), x \in [0, 1]$.
- V is the set of all polynomials on one real variable (say x) with degree at most 2.
  - $p, q \in V, p + q = p_0 + q_0 + (p_1 + q_1)x + (p_2 + q_2)x^2$
  - $p \in V, a \in \mathbb{R}, a \cdot p = ap_0 + ap_1x + ap_2x^2$

- $\mathbb{R}^n$ or $\mathbb{C}^n$:
  - $x = (x_1, x_2, \ldots, x_n), y = (y_1, y_2, \ldots, y_n)$, $x + y = (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n)$
  - $x = (x_1, x_2, \ldots, x_n), a \in \mathbb{R}$ (or $\mathbb{C}$), $a \cdot x = (ax_1, ax_2, \ldots, ax_n)$
- $V =$ Set of real-valued functions defined on $[0, 1] \subset \mathbb{R}$.
  - $f, g \in V, (f + g)(x) = f(x) + g(x), x \in [0, 1]$.
  - $f \in V, a \in \mathbb{R}, (a \cdot f)(x) = af(x), x \in [0, 1]$.
- V is the set of all polynomials on one real variable (say x) with degree at most 2.
  - $p, q \in V, p + q = p_0 + q_0 + (p_1 + q_1)x + (p_2 + q_2)x^2$
  - $p \in V, a \in \mathbb{R}, a \cdot p = ap_0 + ap_1x + ap_2x^2$

- $\mathbb{R}^n$ or $\mathbb{C}^n$:
  - $x = (x_1, x_2, \ldots, x_n), y = (y_1, y_2, \ldots, y_n)$, $x + y = (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n)$
  - $x = (x_1, x_2, \ldots, x_n), a \in \mathbb{R}$ (or $\mathbb{C}$), $a \cdot x = (ax_1, ax_2, \ldots, ax_n)$
- $V =$ Set of real-valued functions defined on $[0, 1] \subset \mathbb{R}$.
  - $f, g \in V, (f + g)(x) = f(x) + g(x), x \in [0, 1]$.
  - $f \in V, a \in \mathbb{R}, (a \cdot f)(x) = af(x), x \in [0, 1]$.
- V is the set of all polynomials on one real variable (say x) with degree at most 2.
  - $p, q \in V, p + q = p_0 + q_0 + (p_1 + q_1)x + (p_2 + q_2)x^2$
  - $p \in V, a \in \mathbb{R}, a \cdot p = ap_0 + ap_1x + ap_2x^2$

- $V = \{f : S \to \mathbb{R}\}$, where S is a surface in $\mathbb{R}^3$.

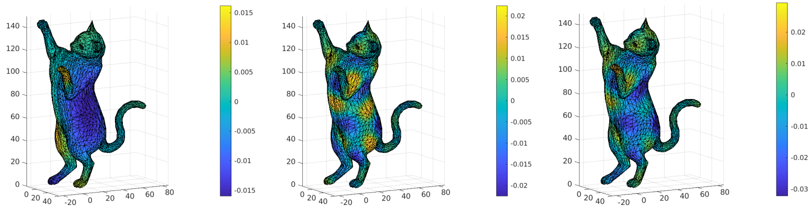


Figure: Examples of functions, say $p : S \to \mathbb{R}, q : S \to \mathbb{R}$ and $p + q$(right)

Subspace

- A subset U of a vector space V is said to be a *Subspace* if U is closed with respect to vector addition and scalar multiplication (defined from V).
  - $V = \mathbb{R}^3, U$ is any plane or line passing through the origin.
  - $V = \{f : [0,1] \to \mathbb{R}\}$,
    $U = \{g(x) = ax + b, \forall x \in [0,1], \forall a, b \in \mathbb{R}\}$.

Linear Combination & Span

- Given a set of vectors $U = \{v_1, \ldots, v_n\} \subset V$, a vector $v = \sum_{i=1}^{n} a_i v_i, a_i \in \mathbb{R}$ is said to be a **linear combination** of U.

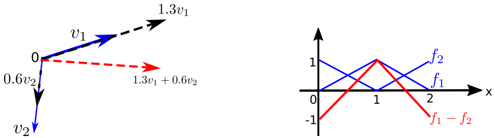


Figure: Linear Combination of two vectors in (left) $V = \mathbb{R}^2$, (right) $V = \{f : [0, 2] \to \mathbb{R}\}$

- The set of all linear combinations of $U = \{v_1, \ldots, v_n\}$ is called the **span** of U.
- If there exists a finite set U of vectors such that $span(U) = V$, then V is a **Finite Dimensional Vector space**(FDVS).

- If $v = \sum_{i=1}^{n} a_i v_i$, then v is said to be linearly dependent on $\{v_1, \ldots, v_n\}$, and if v cannot be expressed as a linear combination of $\{v_1, \ldots, v_n\}$, then we say that v is **linearly independent** of $\{v_1, \ldots, v_n\}$.
- A set of vectors $\{v_1, \ldots, v_n\}$ is said to be **linearly independent**, if none of the vectors in the set are linearly dependent on the rest, i.e.,

$$\sum_{i=1}^{n} a_i v_i = \mathbf{0} \quad \Rightarrow \quad a_i = 0, \forall i$$

- Example: Let $V = \{f : [0, 2\pi] \to \mathbb{R}\}$, $U = \{\sin t, \cos t\}$

Basis

- A basis of a vector space V is a subset $U \subset V$ such that
  1. $Span(U) = V$
  2. U is a set of linearly independent vectors.

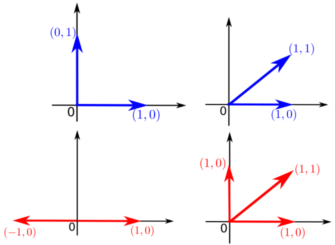


Figure: (Top row) Examples of basis of $\mathbb{R}^2$, (Bottom row) Examples of sets of vectors that are not basis.

- The number of elements in a set of basis vectors is called the **Dimension** of the vector space.
  - Examples: $\mathbb{R}^n, \mathbb{C}^n$ are n-dimensional VS, while the VS of polynomials of one variable is an infinite dimensional VS.

Basis

- A basis of a vector space V is a subset $U \subset V$ such that
  1. $Span(U) = V$
  2. U is a set of linearly independent vectors.

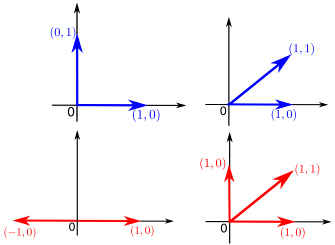


Figure: (Top row) Examples of basis of $\mathbb{R}^2$, (Bottom row) Examples of sets of vectors that are not basis.

- The number of elements in a set of basis vectors is called the **Dimension** of the vector space.
  - Examples: $\mathbb{R}^n, \mathbb{C}^n$ are n-dimensional VS, while the VS of polynomials of one variable is an infinite dimensional VS.

- Let $U = \{v_1, \dots, v_n\}$ be a basis of V. Then for any $v \in V, v = \sum_{i=1}^{n} a_i v_i$.
  - Moreover, the linear combination is unique, i.e.,

$$v = \sum_{i=1}^{n} a_i v_i = \sum_{i=1}^{n} b_i v_i \quad \Rightarrow \quad a_i = b_i, \forall i.$$

  - The coefficents written as a column vector $[a_i]_{i=1}^{n}$ is called the representation of v in the basis U. Also denoted by $[v]_U$.

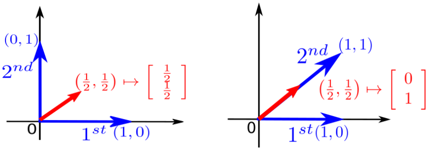


Figure: Examples of representation of vector $(\frac{1}{2}, \frac{1}{2})$ in two different basis of $\mathbb{R}^2$. Note that the order of vectors in the basis matters.

Linear Transformations

- Let $(U, +, \cdot)$ and $(V, +, \cdot)$ be two vector spaces on $\mathbb{R}$ with dimensions m and n respectively.
- A mapping, $T : U \rightarrow V$ is said to be linear if
  1. T is additive: $\forall u_1, u_2 \in U, T(u_1 + u_2) = T(u_1) + T(u_2)$.
  2. T is homogeneous: $\forall u \in U, \forall a \in \mathbb{R}, T(a \cdot u) = a \cdot T(u)$.
- Examples
- Planar rotation: $T : \mathbb{R}^2 \rightarrow \mathbb{R}^2$,
  $T(x, y) = (x \cos\theta - y \sin\theta, x \sin\theta + y \cos\theta)$.
- Matrix multiplication: Let $M \in \mathbb{R}^{m \times n}$.
  $M : \mathbb{R}^n \rightarrow \mathbb{R}^m, M : x \mapsto Mx, \forall x \in \mathbb{R}^n$.

Representation of a Linear Transformation

- Let $T : U \to V$ be a linear transformation, and let $\alpha = \{u_1, \ldots, u_n\}, \beta = \{v_1, \ldots, v_m\}$ be a chosen set of basis for U and V respectively.
  - Let $x \in U, y \in V$ such that $y = Tx$.
  - Let $x = \sum_{j=1}^{n} x_j u_j$.
  - Using properties of T, we have,

$$
\begin{aligned}
y = & Tx = T(\sum_{j=1}^{n} x_j u_j) \\
= & \sum_{j=1}^{n} x_j T(u_j)
\end{aligned}
$$

- Let $Tu_j = \sum_{i=1}^{m} c_{ij} v_i$.

$$
\begin{aligned}
y &= \sum_{j=1}^{n} x_j \sum_{i=1}^{m} c_{ij} v_i = \sum_{j=1}^{n} \sum_{i=1}^{m} x_j c_{ij} v_i \\
&= \sum_{i=1}^{m} \sum_{j=1}^{n} x_j c_{ij} v_i \\
[y]_\beta &= \sum_{j=1}^{n} c_{ij} x_j \\
\begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} &= \begin{bmatrix} c_{11} & \dots & c_{1n} \\ \vdots & \vdots & \vdots \\ c_{m1} & \dots & c_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \\
[y]_\beta &= [T]_\alpha^\beta [x]_\alpha
\end{aligned}
$$

- $[T]_\alpha^\beta$ is the matrix representing the linear transformation T in the basis α and β.

- Different representation for the same linear transformation.

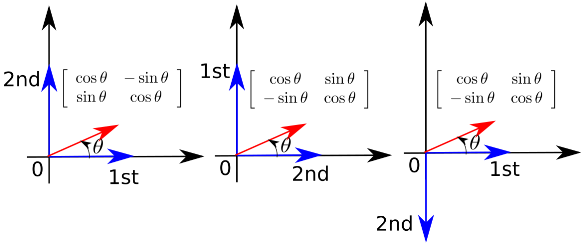


Figure: Rotation by angle θ, R_θ in an anticlockwise direction represented in three different basis in $\mathbb{R}^2$, $[R_\theta]_{\beta_i}^{\beta_i}, i = 1, 2, 3$

- The set of all linear transformations between vector spaces U and V will be denoted by $\mathcal{L}(U, V)$.
- If the domain and co-domain vector spaces are the same, i.e., $U = V$, we usually will use the term **Linear Operator**, and the set of all linear operators on U will be denoted by $\mathcal{L}(U)$.

- Different representation for the same linear transformation.

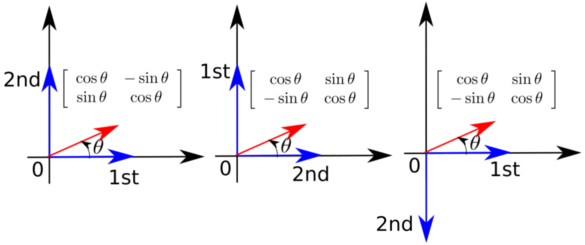


Figure: Rotation by angle θ, R_θ in an anticlockwise direction represented in three different basis in $\mathbb{R}^2$, $[R_\theta]_{\beta_i}^{\beta_i}, i = 1, 2, 3$

- The set of all linear transformations between vector spaces U and V will be denoted by $\mathcal{L}(U, V)$.
- If the domain and co-domain vector spaces are the same, i.e., $U = V$, we usually will use the term **Linear Operator**, and the set of all linear operators on U will be denoted by $\mathcal{L}(U)$.

- Different representation for the same linear transformation.

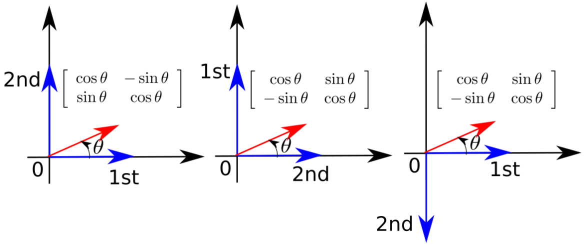


Figure: Rotation by angle θ, R_θ in an anticlockwise direction represented in three different basis in $\mathbb{R}^2$, $[R_\theta]_{\beta_i}^{\beta_i}, i = 1, 2, 3$

- The set of all linear transformations between vector spaces U and V will be denoted by $\mathcal{L}(U, V)$.
- If the domain and co-domain vector spaces are the same, i.e., $U = V$, we usually will use the term **Linear Operator**, and the set of all linear operators on U will be denoted by $\mathcal{L}(U)$.

- Relation between different representations:

$$
\begin{aligned}
[y]_{\beta_i} =& [T]_{\beta_i}^{\beta_i}[x]_{\beta_i} = M_{\beta_j}^{\beta_i}[y]_{\beta_j} = M_{\beta_j}^{\beta_i}[T]_{\beta_j}^{\beta_j}[x]_{\beta_j} \\
=& \underbrace{M_{\beta_j}^{\beta_i}[T]_{\beta_j}^{\beta_j}(M_{\beta_j}^{\beta_i})^{-1}}_{[T]_{\beta_i}^{\beta_i}}[x]_{\beta_i} \\
[T]_{\beta_i}^{\beta_i} =& M_{\beta_j}^{\beta_i}[T]_{\beta_j}^{\beta_j}(M_{\beta_j}^{\beta_i})^{-1}
\end{aligned}
$$

Eigenvectors & Eigenvalues

- Let $T : U \to U$ be a linear operator.
- If for $v \in U, v \neq \mathbf{0}, Tv = \lambda v, \lambda \in \mathbb{F}$, then we say that v is an **eigenvector** of T associated with the **eigenvalue** λ.
- For an n-dimensional vector space U, if there exists n linearly independent eigenvectors $b = \{u_1, \dots, u_n\}$ of a linear operator T, we say b is an **eigenbasis** or T-**eigenbasis** of U.
  - If the associated eigenvalues are $\{\lambda_1, \dots, \lambda_n\}$,

$$[T]_b^b = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ 0 & \dots & 0 & \lambda_n \end{bmatrix}$$

Eigenvectors & Eigenvalues

- Let $T : U \to U$ be a linear operator.
- If for $v \in U, v \neq \mathbf{0}$, $Tv = \lambda v, \lambda \in \mathbb{F}$, then we say that v is an **eigenvector** of T associated with the **eigenvalue** λ.
- For an n-dimensional vector space U, if there exists n linearly independent eigenvectors $b = \{u_1, \ldots, u_n\}$ of a linear operator T, we say b is an **eigenbasis** or T-**eigenbasis** of U.
  - If the associated eigenvalues are $\{\lambda_1, \ldots, \lambda_n\}$,

$$[T]_b^b = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ 0 & \ldots & 0 & \lambda_n \end{bmatrix}$$

Eigenvectors & Eigenvalues

- Let $T : U \to U$ be a linear operator.
- If for $v \in U, v \neq \mathbf{0}, Tv = \lambda v, \lambda \in \mathbb{F}$, then we say that v is an **eigenvector** of T associated with the **eigenvalue** λ.
- For an n-dimensional vector space U, if there exists n linearly independent eigenvectors $b = \{u_1, \dots, u_n\}$ of a linear operator T, we say b is an **eigenbasis** or T-**eigenbasis** of U.
  - If the associated eigenvalues are $\{\lambda_1, \dots, \lambda_n\}$,

$$[T]_b^b = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ 0 & \dots & 0 & \lambda_n \end{bmatrix}$$

Eigenvectors & Eigenvalues

- Let $T : U \to U$ be a linear operator.
- If for $v \in U, v \neq \mathbf{0}$, $Tv = \lambda v, \lambda \in \mathbb{F}$, then we say that v is an **eigenvector** of T associated with the **eigenvalue** λ.
- For an n-dimensional vector space U, if there exists n linearly independent eigenvectors $b = \{u_1, \ldots, u_n\}$ of a linear operator T, we say b is an **eigenbasis** or T-**eigenbasis** of U.
  - If the associated eigenvalues are $\{\lambda_1, \ldots, \lambda_n\}$,

$$[T]_b^b = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ 0 & \ldots & 0 & \lambda_n \end{bmatrix}$$

Examples

- $T : \mathbb{R}^2 \to \mathbb{R}^2$, $T = I_2$, the 2×2 identity matrix.

- $T : \mathbb{R}^2 \to \mathbb{R}^2$, $T = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$.

Examples

- $T : \mathbb{R}^2 \to \mathbb{R}^2$, $T = I_2$, the 2×2 identity matrix.

- $T : \mathbb{R}^2 \to \mathbb{R}^2$, $T = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$.

- $T : \mathbb{C}^n \to \mathbb{C}^n$, $\delta = \{\delta_1, \ldots, \delta_n\}$,
$\delta_i = \left[\begin{array}{ccccccc} 0 & \ldots & 0 & \underbrace{1}_{i^{th}\text{posn}} & 0 & \ldots & 0 \end{array}\right]^T$

$$[T]_\delta^\delta = \begin{bmatrix} c_1 & c_n & \ldots & c_2 \\ c_2 & c_1 & c_n & \ldots \\ \vdots & \vdots & \vdots & \vdots \\ c_n & c_{n-1} & \ldots & c_1 \end{bmatrix},$$

where $c_i \in \mathbb{C}, i = 1, \ldots, n$.

- Then, $b = \{[\exp(j\frac{2\pi mk}{n})]_{m=0}^{n-1}, \quad k = 0, \ldots, n-1\}$ is an eigenbasis.
- M_δ^b is called the **Discrete Fourier Transform**.

- $T : \mathbb{C}^n \to \mathbb{C}^n$, $\delta = \{\delta_1, \dots, \delta_n\}$,

$$\delta_i = \begin{bmatrix} 0 & \dots & 0 & \underbrace{1}_{i^{th}\text{posn}} & 0 & \dots & 0 \end{bmatrix}^T$$

$$[T]_\delta^\delta = \begin{bmatrix} c_1 & c_n & \dots & c_2 \\ c_2 & c_1 & c_n & \dots \\ \vdots & \vdots & \vdots & \vdots \\ c_n & c_{n-1} & \dots & c_1 \end{bmatrix},$$

where $c_i \in \mathbb{C}, i = 1, \dots, n$.

- Then, $b = \{[\exp(j\frac{2\pi mk}{n})]_{m=0}^{n-1}, \quad k = 0, \dots, n-1\}$ is an eigenbasis.
- M_δ^b is called the **Discrete Fourier Transform**.

- $T : \mathbb{C}^n \to \mathbb{C}^n$, $\delta = \{\delta_1, \ldots, \delta_n\}$,

$$\delta_i = \begin{bmatrix} 0 & \ldots & 0 & \underbrace{1}_{i^{th}\text{posn}} & 0 & \ldots & 0 \end{bmatrix}^T$$

$$[T]_\delta^\delta = \begin{bmatrix} c_1 & c_n & \ldots & c_2 \\ c_2 & c_1 & c_n & \ldots \\ \vdots & \vdots & \vdots & \vdots \\ c_n & c_{n-1} & \ldots & c_1 \end{bmatrix},$$

where $c_i \in \mathbb{C}, i = 1, \ldots, n$.

- Then, $b = \{[\exp(j\frac{2\pi mk}{n})]_{m=0}^{n-1}, \quad k = 0, \ldots, n-1\}$ is an eigenbasis.
- M_δ^b is called the **Discrete Fourier Transform**.

Inner Products & Norms

- The vector space does not have notions like angles, distance between vectors, and length of vectors defined.
- **Norm**: A norm on a vector space V is a map $||\cdot|| : V \to \mathbb{R}$ that satisfies:
  1. Positive Definiteness: $\forall x \in V, ||x|| \geq 0, \quad = 0 \Leftrightarrow x = \mathbf{0}$.
  2. Absolute homogeneity: $\forall x \in V, \forall a \in \mathbb{F}, ||ax|| = |a| \cdot ||x||$.
  3. Triangular inequality: $\forall x, y \in V, ||x + y|| \leq ||x|| + ||y||$.
- $||x||$ can be interpeted as *length* of x.
- $||x - y||$ can be interpreted as *distance* between x and y.

- The vector space does not have notions like angles, distance between vectors, and length of vectors defined.
- **Norm**: A norm on a vector space V is a map $||\cdot|| : V \to \mathbb{R}$ that satisfies:
  1. Positive Definiteness: $\forall x \in V, ||x|| \geq 0, \quad = 0 \Leftrightarrow x = \mathbf{0}$.
  2. Absolute homogeneity: $\forall x \in V, \forall a \in \mathbb{F}, ||ax|| = |a| \cdot ||x||$.
  3. Triangular inequality: $\forall x, y \in V, ||x + y|| \leq ||x|| + ||y||$.
- $||x||$ can be interpeted as *length* of x.
- $||x - y||$ can be interpreted as *distance* between x and y.

- The vector space does not have notions like angles, distance between vectors, and length of vectors defined.
- **Norm**: A norm on a vector space V is a map $||\cdot|| : V \to \mathbb{R}$ that satisfies:
  1. Positive Definiteness: $\forall x \in V, ||x|| \geq 0, \quad = 0 \Leftrightarrow x = \mathbf{0}$.
  2. Absolute homogeneity: $\forall x \in V, \forall a \in \mathbb{F}, \; ||ax|| = |a| \cdot ||x||$.
  3. Triangular inequality: $\forall x, y \in V, ||x + y|| \leq ||x|| + ||y||$.
- $||x||$ can be interpeted as *length* of x.
- $||x - y||$ can be interpreted as *distance* between x and y.

Examples

- $\mathbb{R}^n$: $||x|| = \sqrt{x_1^2 + \ldots + x_n^2} = \sqrt{x^T x}$
- $\mathbb{R}^n$: $||x|| = \sqrt{x^T G x}$, where G is a Symmetric Positive Definite(SPD)[1] matrix.
- $\mathbb{C}^n$: $||x|| = \sqrt{|x_1|^2 + \ldots + |x_n|^2} = \sqrt{x^* x}$
- $V = \{f : [0,1] \to \mathbb{R} \mid \int_0^1 |f(x)|^2 \, dx < \infty\}$, $||f|| = \sqrt{\int_0^1 |f(x)|^2 \, dx}$. This is often referred to as the vector space of **real-valued square integrable functions**, and the norm is called the L^2 norm.

[1] A matrix G is said to be positive definite if $\forall x \neq \mathbf{0}$, $x^T G x > 0$

Examples

- $\mathbb{R}^n$: $||x|| = \sqrt{x_1^2 + \ldots + x_n^2} = \sqrt{x^T x}$
- $\mathbb{R}^n$: $||x|| = \sqrt{x^T G x}$, where G is a Symmetric Positive Definite(SPD)[1] matrix.
- $\mathbb{C}^n$: $||x|| = \sqrt{|x_1|^2 + \ldots + |x_n|^2} = \sqrt{x^* x}$
- $V = \{f : [0,1] \to \mathbb{R} \mid \int_0^1 |f(x)|^2 \, dx < \infty\}$, $||f|| = \sqrt{\int_0^1 |f(x)|^2 \, dx}$. This is often referred to as the vector space of **real-valued square integrable functions**, and the norm is called the L^2 norm.

[1] A matrix G is said to be positive definite if $\forall x \neq \mathbf{0}, \; x^T G x > 0$

Examples

- $\mathbb{R}^n$: $||x|| = \sqrt{x_1^2 + \ldots + x_n^2} = \sqrt{x^T x}$
- $\mathbb{R}^n$: $||x|| = \sqrt{x^T G x}$, where G is a Symmetric Positive Definite(SPD)[1] matrix.
- $\mathbb{C}^n$: $||x|| = \sqrt{|x_1|^2 + \ldots + |x_n|^2} = \sqrt{x^* x}$
- $V = \{f : [0,1] \to \mathbb{R} \mid \int_0^1 |f(x)|^2 \, dx < \infty\}$, $||f|| = \sqrt{\int_0^1 |f(x)|^2 \, dx}$. This is often referred to as the vector space of **real-valued square integrable functions**, and the norm is called the L^2 norm.

[1] A matrix G is said to be positive definite if $\forall x \neq \mathbf{0}, \; x^T G x > 0$

Examples

- $\mathbb{R}^n$: $||x|| = \sqrt{x_1^2 + \ldots + x_n^2} = \sqrt{x^T x}$
- $\mathbb{R}^n$: $||x|| = \sqrt{x^T G x}$, where G is a Symmetric Positive Definite(SPD)[1] matrix.
- $\mathbb{C}^n$: $||x|| = \sqrt{|x_1|^2 + \ldots + |x_n|^2} = \sqrt{x^* x}$
- $V = \{f : [0,1] \to \mathbb{R} \mid \int_0^1 |f(x)|^2 \, dx < \infty\}$, $||f|| = \sqrt{\int_0^1 |f(x)|^2 \, dx}$. This is often referred to as the vector space of **real-valued square integrable functions**, and the norm is called the L^2 norm.

[1] A matrix G is said to be positive definite if $\forall x \neq \mathbf{0}, \; x^T G x > 0$

Inner Product

- **Inner Product**: An inner product on a vector space V is a map $\langle \cdot, \cdot \rangle : V \times V :\rightarrow \mathbb{C}$ (or $\mathbb{R}$ if the vector space is real) that satisfies:
  1. Positive Definiteness: $\forall x \in V, \langle x, x \rangle \geq 0, \quad = 0 \Leftrightarrow x = \mathbf{0}$.
  2. Linearity in first argument: $\forall x, y, z \in V, \forall a, b \in \mathbb{F}$, $\langle ax + by, z \rangle = a\langle x, z \rangle + b\langle y, z \rangle$.
  3. Conjugate Symmetry: $\forall x, y \in V, \langle y, x \rangle = \overline{\langle x, y \rangle}$.
- Note that $||x|| = \sqrt{\langle x, x \rangle}$ is a norm.
  - Cauchy-Schwartz inequality: $|\langle x, y \rangle| \leq ||x|| \cdot ||y||$.
  - $\theta_{xy} = \cos^{-1}\left(\frac{\langle x, y \rangle}{||x|| \cdot ||y||}\right)$ can be interpreted as angle between x and y when the vector space is real.
  - A set of vectors $\{v_1, \ldots, v_n\}$ are said to be **orthogonal** if $\langle v_i, v_j \rangle = 0, \forall i \neq j$, and **orthonormal** if additionally $||v_i|| = 1, \forall i$.

Inner Product

- **Inner Product**: An inner product on a vector space V is a map $\langle \cdot, \cdot \rangle : V \times V :\rightarrow \mathbb{C}$ (or $\mathbb{R}$ if the vector space is real) that satisfies:
  1. Positive Definiteness: $\forall x \in V, \langle x, x \rangle \geq 0, \quad = 0 \Leftrightarrow x = \mathbf{0}$.
  2. Linearity in first argument: $\forall x, y, z \in V, \forall a, b \in \mathbb{F}$, $\langle ax + by, z \rangle = a\langle x, z \rangle + b\langle y, z \rangle$.
  3. Conjugate Symmetry: $\forall x, y \in V, \langle y, x \rangle = \overline{\langle x, y \rangle}$.
- Note that $||x|| = \sqrt{\langle x, x \rangle}$ is a norm.
- Cauchy-Schwartz inequality: $|\langle x, y \rangle| \leq ||x|| \cdot ||y||$.
- $\theta_{xy} = \cos^{-1}\left(\frac{\langle x,y \rangle}{||x|| \cdot ||y||}\right)$ can be interpreted as angle between x and y when the vector space is real.
- A set of vectors $\{v_1, \ldots, v_n\}$ are said to be **orthogonal** if $\langle v_i, v_j \rangle = 0, \forall i \neq j$, and **orthonormal** if additionally $||v_i|| = 1, \forall i$.

Examples

- $\mathbb{R}^n$: $\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i = x^T y$
- $\mathbb{R}^n$: $\langle x, y \rangle = x^T G y$, where G is SPD.
- $\mathbb{C}^n$: $\langle x, y \rangle = \sum_{i=1}^{n} x_i \overline{y_i} = y^* x$
- $V = \{f : [0,1] \to \mathbb{R} \mid \int_0^1 |f(x)|^2 \, dx < \infty\}$, $\langle f, g \rangle = \int_0^1 f(x) g(x) \, dx$. This is referred to as the L^2 inner product.

Examples

- $\mathbb{R}^n$: $\langle x, y\rangle = \sum_{i=1}^n x_i y_i = x^T y$
- $\mathbb{R}^n$: $\langle x, y\rangle = x^T G y$, where G is SPD.
- $\mathbb{C}^n$: $\langle x, y\rangle = \sum_{i=1}^n x_i \overline{y_i} = y^* x$
- $V = \{f : [0,1] \to \mathbb{R} \mid \int_0^1 |f(x)|^2 \, dx < \infty\}$, $\langle f, g\rangle = \int_0^1 f(x)g(x) \, dx$. This is referred to as the L^2 inner product.

Examples

- $\mathbb{R}^n$: $\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i = x^T y$
- $\mathbb{R}^n$: $\langle x, y \rangle = x^T G y$, where G is SPD.
- $\mathbb{C}^n$: $\langle x, y \rangle = \sum_{i=1}^{n} x_i \overline{y_i} = y^* x$
- $V = \{f : [0,1] \to \mathbb{R} \mid \int_0^1 |f(x)|^2 \, dx < \infty\}$, $\langle f, g \rangle = \int_0^1 f(x)g(x) \, dx$. This is referred to as the L^2 inner product.

Examples

- $\mathbb{R}^n$: $\langle x, y\rangle = \sum_{i=1}^{n} x_i y_i = x^T y$
- $\mathbb{R}^n$: $\langle x, y\rangle = x^T G y$, where G is SPD.
- $\mathbb{C}^n$: $\langle x, y\rangle = \sum_{i=1}^{n} x_i \overline{y_i} = y^* x$
- $V = \{f : [0,1] \to \mathbb{R} \mid \int_0^1 |f(x)|^2 \, dx < \infty\}$, $\langle f, g\rangle = \int_0^1 f(x) g(x) \, dx$. This is referred to as the L^2 inner product.

Orthogonal Projection

- Let W be a subspace of a vector space V.
- For any $v \in V$, let $v_W \in W$ be such that $\langle v - v_W, w \rangle = 0, \forall w \in W$. v_W is called the **orthogonal projection** of v into W.
  - Let V be a FDVS, and let $W = span(\{w_1, \dots, w_p\})$.
  - Since $v_W \in W, v = \sum_{j=1}^{p} a_j w_j$.
  - Substituting the basis of W gives,

$$\langle v, w_i \rangle = \langle \sum_{j=1}^{p} r_j w_j, w_i \rangle$$

$$[\langle v, w_i \rangle]_{i=1}^{p} = \underbrace{[\langle w_j, w_i \rangle]_{i,j=1}^{p}}_{G-\text{Gram matrix}} \underbrace{[r_j]_{j=1}^{p}}_{r}$$

  - Note that v_W is the also the closest vector[2] in W to v.

[2] according to the norm induced by the inner product.

Orthogonal Projection

- Let W be a subspace of a vector space V.
- For any $v \in V$, let $v_W \in W$ be such that $\langle v - v_W, w \rangle = 0, \forall w \in W$. v_W is called the **orthogonal projection** of v into W.
- Let V be a FDVS, and let $W = span(\{w_1, \ldots, w_p\})$.
- Since $v_W \in W$, $v = \sum_{j=1}^{p} a_j w_j$.
- Substituting the basis of W gives,

$$\langle v, w_i \rangle = \langle \sum_{j=1}^{p} r_j w_j, w_i \rangle$$

$$[\langle v, w_i \rangle]_{i=1}^{p} = \underbrace{[\langle w_j, w_i \rangle]_{i,j=1}^{p}}_{G-\text{Gram matrix}} \underbrace{[r_j]_{j=1}^{p}}_{r}$$

- Note that v_W is the also the closest vector[2] in W to v.

[2] according to the norm induced by the inner product

Orthogonal Projection

- Let W be a subspace of a vector space V.
- For any $v \in V$, let $v_W \in W$ be such that $\langle v - v_W, w \rangle = 0, \forall w \in W$. v_W is called the **orthogonal projection** of v into W.
- Let V be a FDVS, and let $W = span(\{w_1, \ldots, w_p\})$.
- Since $v_W \in W, v = \sum_{j=1}^{p} a_j w_j$.
- Substituting the basis of W gives,

$$\langle v, w_i \rangle = \langle \sum_{j=1}^{p} r_j w_j, w_i \rangle$$

$$[\langle v, w_i \rangle]_{i=1}^{p} = \underbrace{[\langle w_j, w_i \rangle]_{i,j=1}^{p}}_{G-\text{Gram matrix}} \underbrace{[r_j]_{j=1}^{p}}_{r}$$

- Note that v_W is the also the closest vector[2] in W to v.

[2] according to the norm induced by the inner product.

Orthogonal Projection

- Let W be a subspace of a vector space V.
- For any $v \in V$, let $v_W \in W$ be such that $\langle v - v_W, w \rangle = 0, \forall w \in W$. v_W is called the **orthogonal projection** of v into W.
- Let V be a FDVS, and let $W = span(\{w_1, \dots, w_p\})$.
- Since $v_W \in W, v = \sum_{j=1}^{p} a_j w_j$.
- Substituting the basis of W gives,

$$\langle v, w_i \rangle = \langle \sum_{j=1}^{p} r_j w_j, w_i \rangle$$

$$[\langle v, w_i \rangle]_{i=1}^{p} = \underbrace{[\langle w_j, w_i \rangle]_{i,j=1}^{p}}_{G-\text{Gram matrix}} \underbrace{[r_j]_{j=1}^{p}}_{r}$$

- Note that v_W is the also the closest vector[2] in W to v.

[2] according to the norm induced by the inner product

Orthogonal Projection

- Let W be a subspace of a vector space V.
- For any $v \in V$, let $v_W \in W$ be such that $\langle v - v_W, w \rangle = 0, \forall w \in W$. v_W is called the **orthogonal projection** of v into W.
- Let V be a FDVS, and let $W = span(\{w_1, \dots, w_p\})$.
- Since $v_W \in W, v = \sum_{j=1}^{p} a_j w_j$.
- Substituting the basis of W gives,

$$\langle v, w_i \rangle = \langle \sum_{j=1}^{p} r_j w_j, w_i \rangle$$

$$[\langle v, w_i \rangle]_{i=1}^{p} = \underbrace{[\langle w_j, w_i \rangle]_{i,j=1}^{p}}_{G-\text{Gram matrix}} \underbrace{[r_j]_{j=1}^{p}}_{r}$$

- Note that v_W is the also the closest vector[2] in W to v.

[2] according to the norm induced by the inner product

Orthogonal Projection

- Let W be a subspace of a vector space V.
- For any $v \in V$, let $v_W \in W$ be such that $\langle v - v_W, w \rangle = 0, \forall w \in W$. v_W is called the **orthogonal projection** of v into W.
- Let V be a FDVS, and let $W = span(\{w_1, \ldots, w_p\})$.
- Since $v_W \in W, v = \sum_{j=1}^{p} a_j w_j$.
- Substituting the basis of W gives,

$$\langle v, w_i \rangle = \langle \sum_{j=1}^{p} r_j w_j, w_i \rangle$$

$$[\langle v, w_i \rangle]_{i=1}^{p} = \underbrace{[\langle w_j, w_i \rangle]_{i,j=1}^{p}}_{G-\text{Gram matrix}} \underbrace{[r_j]_{j=1}^{p}}_{r}$$

- Note that v_W is the also the closest vector[2] in W to v.

[2]according to the norm induced by the inner product

Example: Linear Least Squares

- Solve for $x \in \mathbb{R}^n : Ax = b, A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m, rank(A) = n$.
  - $A = \begin{bmatrix} a_1 & \dots & a_n \end{bmatrix}$
  - If $b \notin \mathcal{R}(A)$, find x such that $Ax = b_{\mathcal{R}(A)}$.
  - Note that $span(\{a_1, \dots, a_n\}) = \mathcal{R}(A)$.

$$
\begin{aligned}
\langle Ax - b, a_i \rangle =& 0, i = 1, \dots, n \\
[\langle Ax, a_i \rangle]_{i=1}^n =& [\langle b, a_i \rangle]_{i=1}^n \\
[a_i^T Ax]_{i=1}^n =& [a_i^T b]_{i=1}^n \\
(A^T A)x =& A^T b \\
x =& (A^T A)^{-1} A^T b
\end{aligned}
$$

  - $(A^T A)^{-1} A^T$ is called the **Pseudo-inverse** of A, and is denoted by $A^\dagger$. Also $b_{\mathcal{R}(A)} = AA^\dagger b$.
  - Notice that $(AA^\dagger)^T = AA^\dagger$ and $(AA^\dagger)^2 = AA^\dagger$. Conversely any matrix P such that $P^T = P$ and $P^2 = P$ is an orthogonal projection to its range.

Example: Linear Least Squares

- Solve for $x \in \mathbb{R}^n : Ax = b, A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m, rank(A) = n$.
  - $A = \begin{bmatrix} a_1 & \dots & a_n \end{bmatrix}$
  - If $b \notin \mathcal{R}(A)$, find x such that $Ax = b_{\mathcal{R}(A)}$.
  - Note that $span(\{a_1, \dots, a_n\}) = \mathcal{R}(A)$.

$$\begin{aligned}
\langle Ax - b, a_i \rangle &= 0, i = 1, \dots, n \\
[\langle Ax, a_i \rangle]_{i=1}^n &= [\langle b, a_i \rangle]_{i=1}^n \\
[a_i^T Ax]_{i=1}^n &= [a_i^T b]_{i=1}^n \\
(A^T A)x &= A^T b \\
x &= (A^T A)^{-1} A^T b
\end{aligned}$$

  - $(A^T A)^{-1} A^T$ is called the **Pseudo-inverse** of A, and is denoted by $A^\dagger$. Also $b_{\mathcal{R}(A)} = AA^\dagger b$.
  - Notice that $(AA^\dagger)^T = AA^\dagger$ and $(AA^\dagger)^2 = AA^\dagger$. Conversely any matrix P such that $P^T = P$ and $P^2 = P$ is an orthogonal projection to its range.

Example: Linear Least Squares

- Solve for $x \in \mathbb{R}^n : Ax = b, A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m, rank(A) = n$.
  - $A = \left[\begin{array}{ccc} a_1 & \dots & a_n \end{array} \right]$
  - If $b \notin \mathcal{R}(A)$, find x such that $Ax = b_{\mathcal{R}(A)}$.
  - Note that $span(\{a_1, \dots, a_n\}) = \mathcal{R}(A)$.

$$
\begin{aligned}
\langle Ax - b, a_i \rangle &= 0, i = 1, \dots, n \\
[\langle Ax, a_i \rangle]_{i=1}^n &= [\langle b, a_i \rangle]_{i=1}^n \\
[a_i^T Ax]_{i=1}^n &= [a_i^T b]_{i=1}^n \\
(A^T A)x &= A^T b \\
x &= (A^T A)^{-1} A^T b
\end{aligned}
$$

  - $(A^T A)^{-1} A^T$ is called the **Pseudo-inverse** of A, and is denoted by $A^\dagger$. Also $b_{\mathcal{R}(A)} = AA^\dagger b$.
  - Notice that $(AA^\dagger)^T = AA^\dagger$ and $(AA^\dagger)^2 = AA^\dagger$. Conversely any matrix P such that $P^T = P$ and $P^2 = P$ is an orthogonal projection to its range.

Example: Linear Least Squares

- Solve for $x \in \mathbb{R}^n : Ax = b, A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m, rank(A) = n$.
  - $A = \left[\begin{array}{ccc} a_1 & \dots & a_n \end{array} \right]$
  - If $b \notin \mathcal{R}(A)$, find x such that $Ax = b_{\mathcal{R}(A)}$.
  - Note that $span(\{a_1, \dots, a_n\}) = \mathcal{R}(A)$.

$$\begin{aligned} \langle Ax - b, a_i \rangle =& 0, i = 1, \dots, n \\ [\langle Ax, a_i \rangle]_{i=1}^n =& [\langle b, a_i \rangle]_{i=1}^n \\ [a_i^T Ax]_{i=1}^n =& [a_i^T b]_{i=1}^n \\ (A^T A)x =& A^T b \\ x =& (A^T A)^{-1} A^T b \end{aligned}$$

  - $(A^T A)^{-1} A^T$ is called the **Pseudo-inverse** of A, and is denoted by $A^\dagger$. Also $b_{\mathcal{R}(A)} = AA^\dagger b$.
  - Notice that $(AA^\dagger)^T = AA^\dagger$ and $(AA^\dagger)^2 = AA^\dagger$. Conversely any matrix P such that $P^T = P$ and $P^2 = P$ is an orthogonal projection to its range.

Example: Linear Least Squares

- Solve for $x \in \mathbb{R}^n : Ax = b, A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m, rank(A) = n$.
  - $A = \begin{bmatrix} a_1 & \dots & a_n \end{bmatrix}$
  - If $b \notin \mathcal{R}(A)$, find x such that $Ax = b_{\mathcal{R}(A)}$.
  - Note that $span(\{a_1, \dots, a_n\}) = \mathcal{R}(A)$.

$$\begin{aligned}
\langle Ax - b, a_i \rangle &= 0, i = 1, \dots, n \\
[\langle Ax, a_i \rangle]_{i=1}^n &= [\langle b, a_i \rangle]_{i=1}^n \\
[a_i^T Ax]_{i=1}^n &= [a_i^T b]_{i=1}^n \\
(A^T A)x &= A^T b \\
x &= (A^T A)^{-1} A^T b
\end{aligned}$$

  - $(A^T A)^{-1} A^T$ is called the **Pseudo-inverse** of A, and is denoted by $A^\dagger$. Also $b_{\mathcal{R}(A)} = AA^\dagger b$.
  - Notice that $(AA^\dagger)^T = AA^\dagger$ and $(AA^\dagger)^2 = AA^\dagger$. Conversely any matrix P such that $P^T = P$ and $P^2 = P$ is an orthogonal projection to its range.

Example: Linear Least Squares

- Solve for $x \in \mathbb{R}^n : Ax = b, A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m, rank(A) = n$.
  - $A = \begin{bmatrix} a_1 & \dots & a_n \end{bmatrix}$
  - If $b \notin \mathcal{R}(A)$, find x such that $Ax = b_{\mathcal{R}(A)}$.
  - Note that $span(\{a_1, \dots, a_n\}) = \mathcal{R}(A)$.

$$
\begin{aligned}
\langle Ax - b, a_i \rangle =& 0, i = 1, \dots, n \\
[\langle Ax, a_i \rangle]_{i=1}^n =& [\langle b, a_i \rangle]_{i=1}^n \\
[a_i^T Ax]_{i=1}^n =& [a_i^T b]_{i=1}^n \\
(A^T A)x =& A^T b \\
x =& (A^T A)^{-1} A^T b
\end{aligned}
$$

  - $(A^T A)^{-1} A^T$ is called the **Pseudo-inverse** of A, and is denoted by $A^\dagger$. Also $b_{\mathcal{R}(A)} = AA^\dagger b$.
  - Notice that $(AA^\dagger)^T = AA^\dagger$ and $(AA^\dagger)^2 = AA^\dagger$. Conversely any matrix P such that $P^T = P$ and $P^2 = P$ is an orthogonal projection to its range.