A Tutorial on Duality

Richard Xu

March 6, 2022

1 Motivation

inequality-constrained optimization often appear in Machine Learning Literature:

1.1 reinforcement Learning

$$
\begin{aligned}
&\max_{\pi} \Bigg[\mathbb{E}_{\tau \sim \beta} \Bigg[\sum_{t=0}^{\infty} \gamma^t \frac{\pi(a_t|s_t)}{\beta(a_t|s_t)} A^{\beta}(s_t, a_t) \Bigg] \Bigg] \\
&\text{s.t.} \quad \mathrm{KL}(\pi \| \beta) \leq \delta
\end{aligned}
\tag{1}
$$

1.2 sensitive GAN

$$
\begin{aligned}
\text{let} \quad \mathcal{L}^{D}_{\theta_D}(\mathbf{x}) &= \min_{\theta_G} \big(\mathcal{L}_{\theta_D, \theta_G}(\mathbf{x}) \big) \\
&= \min_{\theta_G} \Big(\mathbb{E}_{\mathbf{x} \sim p_{\mathbf{x}}(\mathbf{x})} [\log D_{\theta_D}(\mathbf{x})] + \mathbb{E}_{z \sim p_z(\mathbf{z})} [\log(1 - D_{\theta_D}(G_{\theta_G}(\mathbf{z})))] \Big)
\end{aligned}
\tag{2}
$$

then sensitive GAN is designed to:

$$
\begin{aligned}
&\max_{\theta_D} \big(\mathcal{L}^{D}_{\theta_D}(\mathbf{x}) \big) \\
&\text{s.t.} \quad \mathcal{L}^{D}_{\theta_D}(\mathbf{x}) \leq \mathcal{L}^{D}_{\theta_D}(G_{\theta_G}(\mathbf{z})) - \triangle(\mathbf{x}, G_{\theta_G}(\mathbf{z}))
\end{aligned}
\tag{3}
$$

1.3 Support vector machine

$$
\begin{aligned}
&\min \left(\frac{1}{2} \|\mathbf{w}\|^2 \right) \\
&\text{subject to:} \quad 1 - y_i(\mathbf{w}^T x_i + w_0) \leq 0 \quad \forall i
\end{aligned}
\tag{4}
$$

2 Optimization with inequality constraints

A constrained optimization is of the following form (ignore the equality constraints for now):

$$
\begin{aligned}
&\min f(\mathbf{x}) \\
&\text{s.t. } g_i(\mathbf{x}) \leq 0 \; \forall i \in 1, \dots, m
\end{aligned}
\tag{5}
$$

After defining $\mathbf{I}(u) = \begin{cases} 0, & \text{if } u \leq 0 \\ \infty, & \text{otherwise} \end{cases}$, i.e., a "huge step function", we can turn a constrained equation into **unconstrained** equation:

$$J(\mathbf{x}) = f(\mathbf{x}) + \sum_i \mathbf{I}[g_i(\mathbf{x})] \tag{6}$$

it words, it makes infeasible region to have prohibitively large value, i.e., ∞ making it impossible to find a **minimization** solution in infeasible region

Similarly, in **maximization**, infeasible region are assigned value of $-\infty$ making it impossible to find a maximum solution in infeasible region

$$J(\mathbf{x}) = f(\mathbf{x}) - \sum_i \mathbf{I}[g_i(\mathbf{x})] \tag{7}$$

3 An alternative objective function

Replace $\mathbf{I}[g_i(x)]$ by its lower bound $\lambda_i g_i(\mathbf{x})$, with $\lambda_i \geq 0$. Therefore $J(x) \to \mathcal{L}(x, \lambda)$:

$$\mathcal{L}(\mathbf{x}, \lambda) = f(\mathbf{x}) + \sum_i \lambda_i g_i(\mathbf{x}) \tag{8}$$

3.1 re-write the objective

since $\lambda_i g_i(\mathbf{x})$ is lower bound of $\mathbf{I}[g_i(x)]$:

$$\mathcal{L}(\mathbf{x}, \lambda) \leq J(\mathbf{x}) \tag{9}$$

we can just write:

$$J(\mathbf{x}) \equiv \max_\lambda \mathcal{L}(\mathbf{x}, \lambda) \tag{10}$$

3.1.1 verify above was the case

Think about what values does λ take when we perform $\max_\lambda \mathcal{L}(\mathbf{x}, \lambda)$?

1. for $x : g(x) < 0$:

$$\arg\max_\lambda \Big(f(\mathbf{x}) + \sum_i \lambda_i g_i(\mathbf{x}) \Big) = 0 \tag{11}$$

2. for $x : g(x) > 0$:

$$\arg\max_\lambda \Big(f(\mathbf{x}) + \sum_i \lambda_i g_i(\mathbf{x}) \Big) = \infty \tag{12}$$

3.1.2 pointless, but doable

$$
\begin{aligned}
\min_{\mathbf{x}} \max_{\lambda} \mathcal{L}(\mathbf{x}, \lambda) &= \min_{\mathbf{x}} J(\mathbf{x}) \\
&= p^*
\end{aligned}
\tag{13}
$$

In words, it means for $\mathcal{L}(\mathbf{x}, \lambda)$ we maximize λ first, then minimize $\mathbf{x}$ and we obtain $J(\mathbf{x}^*)$.

However, it is point-less to do so in that optimization order.

3.2 Swap optimization order: $\min_x$ first, then $\max_\lambda$

from Eq(13)

$$
\begin{aligned}
&\min_{\mathbf{x}} \max_{\lambda} \mathcal{L}(\mathbf{x}, \lambda) = \min_{\mathbf{x}} J(\mathbf{x}) \\
\Longrightarrow\ &\max_{\lambda} \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \lambda) \le \min_{\mathbf{x}} \max_{\lambda} \mathcal{L}(\mathbf{x}, \lambda) = \min_{\mathbf{x}} J(\mathbf{x}) \\
\Longrightarrow\ &\Big(d^* \equiv \max_{\lambda} \min_{x} \mathcal{L}(\mathbf{x}, \lambda)\Big) \le \Big(p^* \equiv \min_{\mathbf{x}} \max_{\lambda} \mathcal{L}(\mathbf{x}, \lambda) = \min_{\mathbf{x}} J(\mathbf{x})\Big) \\
\Longrightarrow\ &\big(d^* \equiv \max_{\lambda} f_{\lambda}^{(\star)}(\lambda)\big) \le p^*
\end{aligned}
\tag{14}
$$

$f_{\lambda}^{(\star)}(\lambda)$ is called dual function

3.2.1 max-min inequality

this relationship can be understood by **max-min inequality**

$$
\sup_{\lambda} \inf_{x} f(\lambda, x) \le \inf_{x} \sup_{\lambda} f(\lambda, x) \tag{15}
$$

"the greatest of all minima" is less or equal to "the least of all maxima", **proof**:

$$
\begin{aligned}
&\inf_{x} f(\lambda, x) \le f(\lambda, x), \forall x && \lambda \text{ is a constant} \\
\Longrightarrow\ &\sup_{\lambda} \inf_{x} f(\lambda, x) \le \sup_{\lambda} f(\lambda, x), \forall x && \sup_{\lambda} \text{ both sides} \\
\Longrightarrow\ &\sup_{\lambda} \inf_{x} f(\lambda, x) \le \inf_{x} \sup_{\lambda} f(\lambda, x) && \text{on RHS: } \because \inf_{x} \in \forall x
\end{aligned}
\tag{16}
$$

3.3 if strong duality holds

$$
d^* = p^* \tag{17}
$$

4 advantage of dual function

in summary, the duality procedure is to find λ^*

$$
\begin{aligned}
\lambda^* &= \arg\max_{\lambda} \big(\min_{x} \mathcal{L}(\mathbf{x}, \lambda)\big) \\
&= \arg\max_{\lambda} f_{\lambda}^{(\star)}(\lambda)
\end{aligned}
\tag{18}
$$

dual function $f_\lambda^{(\star)}(\lambda) \equiv \min_x \mathcal{L}(\mathbf{x}, \lambda)$ is concave, even when the initial problem is not convex. Because it is a point-wise (in λ) infimum of affine functions:

$$\begin{aligned} f_\lambda^{(\star)}(\lambda) \equiv \min_x \mathcal{L}(\mathbf{x}, \lambda) \triangleq \min_x \Big(f(\mathbf{x}) + \sum_i \lambda_i g_i(\mathbf{x}) \Big) \\ = f(\mathbf{x}') + \sum_i \underbrace{\lambda_i}_{x} \underbrace{g_i(\mathbf{x}')}_{m} \end{aligned} \tag{19}$$

where $g_i(\mathbf{x})$ are fixed co-efficient (m), and λ_i is the variable (x) of the line, they form "envelops" of lines, to be concave.

note also that, dual function $f_\lambda^{(\star)}(\lambda)$ can be thought as a function defined over "gradient space". It can be best visualized by plotting $f_\lambda^{(\star)}(\lambda)$ using lines defined by a finite $\{\mathbf{x}\}$, and $\mathbf{x}$ are treated like "constant line parameters"

4.1 convex-concave theorem

Let $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^m$ be compact convex sets. If $f : X \times Y \to \mathbb{R}$ is a continuous function that is convex-concave:

$$\begin{aligned} &f(\cdot, y) : X \to \mathbb{R} \text{ is convex for fixed } y \\ &f(x, \cdot) : Y \to \mathbb{R} \text{ is concave for fixed } x \end{aligned} \tag{20}$$

then:

$$\min_{x \in X} \max_{y \in Y} f(x, y) = \max_{y \in Y} \min_{x \in X} f(x, y) \tag{21}$$

5 A quick note on equality constraint: Lagrange

$$\begin{aligned} &\text{maximize } f(\mathbf{x}) \\ &\text{subject to: } g(\mathbf{x}) = 0 \end{aligned} \tag{22}$$

The problem can be transformed into finding $\mathbf{x}$ satisfying these two conditions:

$$\begin{cases} \nabla_{\mathbf{x}} f(\mathbf{x}) - \mu \nabla_{\mathbf{x}} g(\mathbf{x}) = 0 & \text{as contour line } f(\mathbf{x}) = k \text{ and } g(\mathbf{x}) \text{ share same tangent} \\ g(\mathbf{x}) = 0 & \text{original constraint} \end{cases} \tag{23}$$

conveniently, one can re-frame these two constraints as to let both partial derivatives μ and $\mathbf{x}$ of lagrange function $\mathcal{L}(\mathbf{x}, \mu)$ equal zero, where:

$$\begin{aligned} \mathcal{L}(\mathbf{x}, \mu) &= f(\mathbf{x}) - \mu g(\mathbf{x}) \\ \implies \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \mu) &= \underbrace{\nabla_{\mathbf{x}} f(\mathbf{x}) - \mu \nabla_{\mathbf{x}} g(\mathbf{x}) = 0} \\ \nabla_\mu \mathcal{L}(\mathbf{x}, \mu) &= \underbrace{g(\mathbf{x}) = 0} \end{aligned} \tag{24}$$

6 KKT condition: complementary slackness

we let

$$\begin{cases} \mathbf{x}^{(f)} & = \arg\min_{\mathbf{x}} \big(f(\mathbf{x})\big) \\ \mathbf{x}^{(g)} & = \arg\max_{x} \big(\lambda g(\mathbf{x})\big) \qquad \text{for some } \lambda > 0 \end{cases} \tag{25}$$

Note that it is the same $\mathbf{x}^{(g)}$ $\forall\lambda$. $\mathbf{x}^{(g)}$ must occur in the feasible region, i.e., $g(\mathbf{x}^{(g)}) \leq 0$

6.1 simple explanation from the view of $f(\mathbf{x})$ only

since $f(\mathbf{x})$ is convex. Then if $g(\mathbf{x}^{(f)}) < 0$, minimum of primal must be $\mathbf{x}^* = \mathbf{x}^{(f)}$ since $g(\mathbf{x})$ do not play a part. If $g(\mathbf{x}^{(f)}) > 0$, then $f(x^*)$ must occur at the boundary where $g(\mathbf{x}^*) = 0$ as it will increase inside the feasible region.

6.2 explanation from the view of $\mathcal{L}(\mathbf{x}, \lambda)$

Lemma 1 *given two functions $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$, and some domain dom$^*(\mathbf{x})$ where $f_2(\mathbf{x}) \geq f_1(\mathbf{x})$ for $\mathbf{x} \in$ dom$^*(\mathbf{x})$. Additionally, $\min f_1(\mathbf{x}), \min f_2(\mathbf{x}) \in$ dom$^*(\mathbf{x})$, then:*

$$\min(f_1(\mathbf{x})) \leq \min(f_2(\mathbf{x})) \tag{26}$$

conversely, if $f_2(\mathbf{x}) \leq f_1(\mathbf{x})$ for $\mathbf{x} \in \text{dom}^*(\mathbf{x})$, and $\min f_1(\mathbf{x}), \min f_2(\mathbf{x}) \in \text{dom}^*(\mathbf{x})$ then:

$$\min(f_1(\mathbf{x})) \geq \min(f_2(\mathbf{x})) \tag{27}$$

6.2.1 what if we increase λ

Note that as we increase λ, we are adding more weights towards $\lambda g(\mathbf{x})$ part of $f(\mathbf{x})+\lambda g(\mathbf{x})$.

Also, notice that as we increase λ, $\lambda g(\mathbf{x})$ becomes "sharper" about the point $(\mathbf{x}^*, 0)$ where $g(\mathbf{x}^*) = 0$:

$$\lambda_1 \leq \lambda_2 \implies \begin{cases} \lambda_1 g(\mathbf{x}) \leq \lambda_2 g(\mathbf{x}) & \text{when } g(\mathbf{x}) > 0 \\ \lambda_1 g(\mathbf{x}) \geq \lambda_2 g(\mathbf{x}) & \text{when } g(\mathbf{x}) \leq 0 \end{cases} \tag{28}$$

6.2.2 when $\mathbf{x}^{(f)}$ occurs inside of feasible region: $g(\mathbf{x}^{(f)}) \leq 0$

since $\mathbf{x}^{(f)}$ is aready inside the feasible region, and we know that including when $\lambda_1 = 0$:

$$\lambda_1 \leq \lambda_2 \implies f(\mathbf{x}) + \lambda_1 g(\mathbf{x}) \geq f(\mathbf{x}) + \lambda_2 g(\mathbf{x}) \tag{29}$$

apply Lemma 1, we have:

$$\lambda_1 \leq \lambda_2 \implies \min\big(f(\mathbf{x}) + \lambda_1 g(\mathbf{x})\big) \geq \min\big(f(\mathbf{x}) + \lambda_2 g(\mathbf{x})\big) \tag{30}$$

which means that $f_\lambda^{(\star)}(\lambda)$ is maximum at $\lambda = 0$ with value $= f(\mathbf{x}^{(f)})$ and it's monotonically decreasing.

6.2.3 When $\mathbf{x}^{(f)}$ occurs outside of feasible region: $g(\mathbf{x}^{(f)}) > 0$

It is still an interpolation between $\mathbf{x}^{(f)}$ and $\mathbf{x}^{(g)}$. As λ increases, the interpolated minimum $\min\big(f(\mathbf{x}) + \lambda g(\mathbf{x})\big)$ is lean towards $x^{(g)}$. There will be a λ^* such that:

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \big(f(\mathbf{x}) + \lambda^* g(\mathbf{x})\big) \quad \text{where } g(\mathbf{x}^*) = 0 \tag{31}$$

so we split the domain of $\mathbf{x}$ into two regions: $[\mathbf{x}^{(f)}, \dots, \mathbf{x}^*)$ and $[\mathbf{x}^*, \dots, \mathbf{x}^{(g)})$

1. in the region of $[\mathbf{x}^{(f)}, \dots, \mathbf{x}^*)$:

   apply Lemma 1, we have:

$$\begin{aligned} \lambda_1 \leq \lambda_2 &\implies f(\mathbf{x}) + \lambda_1 g(\mathbf{x}) \leq f(\mathbf{x}) + \lambda_2 g(\mathbf{x}) \\ &\implies \min\big(f(\mathbf{x}) + \lambda_1 g(\mathbf{x})\big) \leq \min\big(f(\mathbf{x}) + \lambda_2 g(\mathbf{x})\big) \end{aligned} \tag{32}$$

2. in the region of $[\mathbf{x}^*, \dots, \mathbf{x}^{(g)})$:

   apply Lemma 1, we have:

$$\begin{aligned} \lambda_1 \leq \lambda_2 &\implies f(\mathbf{x}) + \lambda_1 g(\mathbf{x}) \geq f(\mathbf{x}) + \lambda_2 g(\mathbf{x}) \\ &\implies \min\big(f(\mathbf{x}) + \lambda_1 g(\mathbf{x})\big) \geq \min\big(f(\mathbf{x}) + \lambda_2 g(\mathbf{x})\big) \end{aligned} \tag{33}$$

which means $f_\lambda^{(\star)}(\lambda)$ monotonically increase from $\lambda = 0$ to λ^*, and it becomes monotonically decreasing.

6.2.4 combine the two

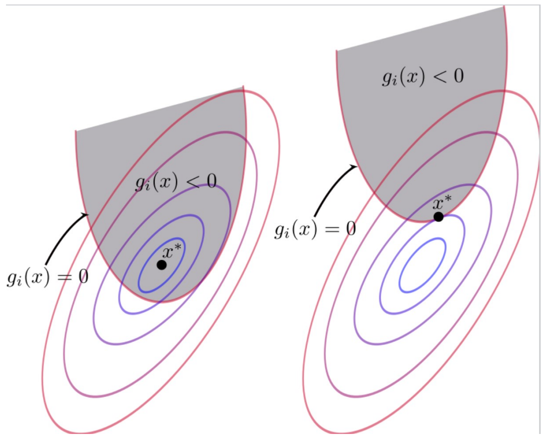
Combine the above two cases, we found either:

$$\begin{cases} \lambda_i^* = 0 & \text{when } \mathbf{x}^{(f)} \text{ is in feasible region} \\ g_i(\mathbf{x}^*) = 0 & \text{when } \mathbf{x}^{(f)} \text{ is outside of feasible region} \end{cases} \tag{34}$$

We can specify it in a single equation:

$$\lambda_i^* \; g_i(\mathbf{x}^*) = 0 \tag{35}$$

This is called **complimentary slackness**. Diagrammatically, this is illustrated from a diagram from Wikipedia:



7 summary of KKT condition

now we understood complementary slackness:

optimization problem with both equality and inequality constraints:

$$\begin{aligned} \mathbf{x}^* &= \underset{\mathbf{x}}{\operatorname{argmin}} f(\mathbf{x}) \\ &\quad \text{subject to } h_i(\mathbf{x}) = 0 \qquad \text{added for completeness} \\ &\quad \text{subject to } g_i(\mathbf{x}) \leq 0 \end{aligned} \tag{38}$$

So how do we produce duality function $\lambda^* = \arg\max_\lambda \min_x \mathcal{L}(\mathbf{x}, \lambda)$ being carried out in practice, also since we have additional equality constraint, we now have $\mathcal{L}(\mathbf{x}, \mu, \lambda)$ instead:

$$\mathcal{L}(\mathbf{x}, \mu, \lambda) = f(\mathbf{x}) + \sum_{i=1}^{m} \mu_i h_i(\mathbf{x}) + \sum_{i=1}^{n} \lambda_i g_i(\mathbf{x}) \tag{39}$$

1. obtain $f_\lambda^{(\star)}(\lambda) = \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \mu, \lambda)$ by:

(a) solve $\mathbf{x}^*$, such that:

$$\begin{aligned}
&\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^*,\mu,\lambda) = 0 \\
\Longrightarrow\ &\nabla_{\mathbf{x}}\Big(f(\mathbf{x}^*) + \sum_{i=1}^{m}\mu_i h_i(\mathbf{x}^*) + \sum_{i=1}^{n}\lambda_i g_i(\mathbf{x}^*)\Big) = 0 \\
\Longrightarrow\ &\nabla_{\mathbf{x}} f(\mathbf{x}^*) + \sum_{i=1}^{m}\mu_i \nabla_{\mathbf{x}^*} h_i(\mathbf{x}^*) + \sum_{i=1}^{n}\lambda_i \nabla_{\mathbf{x}} g_i(\mathbf{x}^*) = 0
\end{aligned} \tag{40}$$

(b) write $\mathbf{x}^*$ in terms of λ and substitute back into $\mathcal{L}(\mathbf{x}^*,\mu,\lambda)$ and obtain:

$$f_{\lambda}^{(\star)}(\lambda) = \min_{\mathbf{x}} \mathcal{L}(\mathbf{x},\mu,\lambda) \tag{41}$$

note $f_{\lambda}^{(\star)}(\lambda)$ should contain no $\mathbf{x}$

now we can $\max_{\lambda} f_{\lambda}^{(\star)}(\lambda)$ together with the complementary slackness conditions

2. to ensure **equality constraints**, we need to solve:

$$\begin{aligned}
&\nabla_{\mu}\mathcal{L}(\mathbf{x}^*,\mu,\lambda) = 0 \\
\Longrightarrow\ &\nabla_{\mu} f(\mathbf{x}^*) + \sum_{i=1}^{m} \nabla_{\mu_{\mathrm{i}}}\mu_i h_i(\mathbf{x}^*) + \sum_{i=1}^{n}\lambda_i \nabla_{\mu} g_i(\mathbf{x}^*) = 0 \\
\Longrightarrow\ &\sum_{i=1}^{m}\nabla_{\mu_{\mathrm{i}}}\mu_i h_i(\mathbf{x}^*) = 0 \quad \text{both } f(\mathbf{x}) \text{ and } g_i(\mathbf{x}) \text{ disappeared} \\
\Longrightarrow\ &\sum_{i=1}^{m} h_i(\mathbf{x}^*) = 0 \qquad \text{just the original equality condition}
\end{aligned} \tag{42}$$

3. to ensure **Inequality constraints a.k.a. complementary slackness condition**

$$\begin{aligned}
\lambda_i^* g_i(\mathbf{x}^*) = 0, &\quad \forall i \\
\lambda_i \geq 0, &\quad \forall i \\
g_i(\mathbf{x}^*) \leq 0, &\quad \forall i
\end{aligned} \tag{43}$$

the final solution for dual λ^* needs to be take account of all above equations, and let's see the classical example of solution for Support Vector Machine

Theorem 1 *For a problem with strong duality, $\mathbf{x}^*$, λ^*, μ^* satisfy KKT conditions if and only if x^*, λ^* and μ^* are primal and dual solutions.*

7.1 zero duality gap if and only if KKT condition exist

let $\mathbf{x}$ and (λ, μ) be primal and dual solutions with **zero duality gap** (i.e. strong duality holds), then let's show if and only if KKT condition exists.

7.1.1 prove necessity

We start by showing zero duality gap implies KKT condition exist

$$
\begin{aligned}
f(\mathbf{x}^*) &= f_\lambda^{(\star)}(\mu^*, \lambda^*) \quad \text{assume zero duality gap} \\
&= \min_{\mathbf{x}} \Big(f(\mathbf{x}) + \sum_{j=1}^{m} \mu_j^* h_j(\mathbf{x}) + \sum_{i=1}^{r} \lambda_i^* g_i(\mathbf{x}) \Big) \quad \text{definition of } f_\lambda^{(\star)}(\mu^*, \lambda^*) \\
&\le \Big(f(\mathbf{x}^*) + \sum_{j=1}^{m} \mu_j^* h_j(\mathbf{x}^*) + \sum_{i=1}^{r} \lambda_i^* g_i(\mathbf{x}^*) \Big) \quad \text{must less than any } \mathbf{x}, \text{ including } \mathbf{x} = \mathbf{x}^* \\
&\le f(\mathbf{x}^*) \quad h_j(\mathbf{x}^*) = 0,\ g_i(\mathbf{x}^*) \le 0
\end{aligned} \tag{44}
$$

Therefore, all inequalities above become equal, it means we can also:

1. first equality: $\mathbf{x}^*$ is the minimizer of $\mathcal{L}(\mathbf{x}, \mu^*, \lambda^*)$:

$$
\begin{aligned}
&\min_{\mathbf{x}} \Big(f(\mathbf{x}) + \sum_{j=1}^{m} \mu_j^* h_j(\mathbf{x}) + \sum_{i=1}^{r} \lambda_i^* g_i(\mathbf{x}) \Big) = f(\mathbf{x}^*) + \sum_{j=1}^{m} \mu_j^* h_j(\mathbf{x}^*) + \sum_{i=1}^{r} \lambda_i^* g_i(\mathbf{x}^*) \\
&\implies 0 \in \partial_{\mathbf{x}} \mathcal{L}(\mathbf{x}^*, \mu^*, \lambda^*) \quad 0 \in \partial_{\mathbf{x}} \quad \text{sub-gradient} \\
&\implies 0 \in \partial_{\mathbf{x}} \Big(f(\mathbf{x}^*) + \sum_{j=1}^{m} \mu_j^* h_j(\mathbf{x}^*) + \sum_{i=1}^{r} \lambda_i^* g_i(\mathbf{x}^*) \Big)
\end{aligned} \tag{45}
$$

   this shows stationary condition occur at $\mathbf{x}^*$. A side note, this precisely what we need to compute in Eq.(40)

2. second equality:

$$
\begin{aligned}
f(\mathbf{x}^*) + \sum_{j=1}^{m} \mu_j^* h_j(\mathbf{x}^*) + \sum_{i=1}^{r} \lambda_i^* g_i(\mathbf{x}^*) &= f(\mathbf{x}^*) \\
\implies \sum_{i=1}^{r} \lambda_i^* g_i(\mathbf{x}^*) &= 0
\end{aligned} \tag{46}
$$

   this means all complimentary slackness satisfy

therefore, we have shown that if there is zero duality gap, then KKT condition satisfies

7.1.2 prove sufficiency

we then show if KKT condition exist, it implies zero duality gap:

now, if there exists $\mathbf{x}^*, \mu^*, \lambda^*$ that satisfy the KKT conditions:

$$
\begin{aligned}
f_\lambda^{(\star)}(\mu, \lambda) &= f(\mathbf{x}^*) + \sum_{j=1}^{m} \mu_j^* h_j(\mathbf{x}^*) + \underbrace{\sum_{i=1}^{r} \lambda_i^* g_i(\mathbf{x}^*)}_{=0:\ \text{KKT}} \\
&= f(\mathbf{x}^*)
\end{aligned} \tag{47}
$$

8 Example through Support Vector Machine

8.1 Linear Discriminant Function (geometry)

8.1.1 motivation

this is maximum margin hyperplane, i.e., it doesn't just simply find the decision boundary for the two-class data:

$$\mathbf{x}^\top \mathbf{w} + w_0 = 0 \tag{48}$$

8.1.2 geometry of $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + w_0$ in 1D

think about the normal line $f(x) = wx + w_0$ and without w_0:

$$\begin{aligned} f(x) &= wx \\ \implies \begin{bmatrix} w & -1 \end{bmatrix} \begin{bmatrix} x \\ f(x) \end{bmatrix} &= 0 \end{aligned} \tag{49}$$

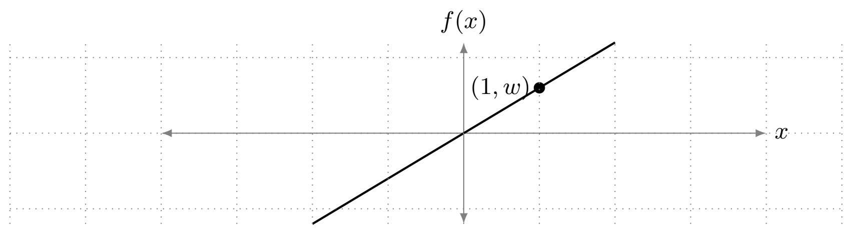


Figure 1: $f(x) = wx$

the point $(0, 0)$ satisfies the line with has normal $\begin{bmatrix} w & -1 \end{bmatrix}$
adding w_0, which we have $f(x) = wx + w_0$:

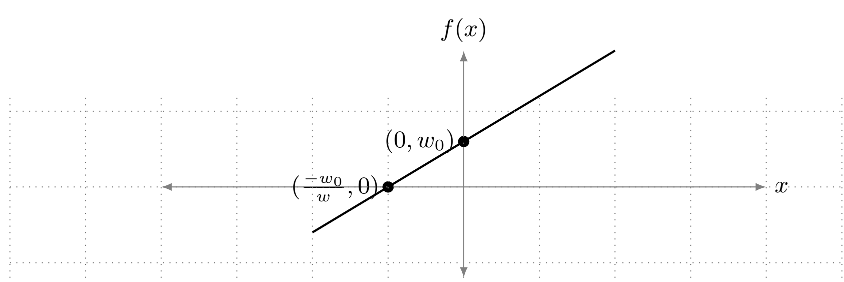


Figure 2: $f(x) = wx + w_0$

8.1.3 geometry of $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + w_0$ in 2D

in a similar fashion, we think about the 3-D hyper-plane without the w_0:

$$\begin{aligned} f(\mathbf{x}) &= \mathbf{w}^\top \mathbf{x} \\ \implies \begin{bmatrix} w_1 & w_2 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ f(x_1, x_2) \end{bmatrix} &= 0 \end{aligned} \tag{50}$$

The gradient vectors are $(w_1, w_2)^\top$. In the plane where $f(x_1, x_2) = 0$, all $(x_1, x_2) \perp (w_1, w_2)$ satisifies the line, it must include $(0, 0)$.

Ignore w_0 in both cases, in 1-D case, the **line** with normal $[w, -1]^\top$ cuts the **line** $f(x) = 0$ at the origin (just a single point). In 2-D case, the **plane** with normal $\begin{bmatrix} w_1 & w_2 & -1 \end{bmatrix}$ cuts the **plane** $f(x_1, x_2) = 0$. However, now the two 3-D planes meets in a line. $(w_1, w_2)^\top$ controls the orientation of the line in (x_1, x_2) plane (which go through the origin) where two plane meets. The point $\big(x_1, x_2, f(\mathbf{x})\big) = (0, 0, 0)$ are on this line without changing the normal.

Now by adding w_0, it is shift along the $f(\mathbf{x})$ axis.

8.1.4 arbitary dimensions

More generically, $\mathbf{w}$ controls the direction of "cutting plane" and w_0 moves this plane in the $f(\mathbf{x})$ direction. Note that w_0 does **not** change the direction of cutting plane in the $\mathbf{x}$ plane. Its only cause the cutting line to move in parallel in the $f(\mathbf{x}) = 0$ plane.

8.1.5 the margin idea

it also put data of each class behind their *margins*:

$$\begin{cases} \text{all data } \mathbf{x} \text{ having label } y = +1 \text{ is } \textbf{above} \text{ the boundary} & \mathbf{w}^\top \mathbf{x} + w_0 = 1 \\ \text{all data } \mathbf{x} \text{ having label } y = -1 \text{ is } \textbf{below} \text{ the boundary} & \mathbf{w}^\top \mathbf{x} + w_0 = -1 \end{cases} \tag{51}$$

to solve this problem, we design a linear plane that "cuts" through the middle of the decision boundry $\mathbf{x}^\top \mathbf{w} + w_0 = 0$, which will produce $y(\mathbf{x})$ having the desired effect

$$y(\mathbf{x}) = \begin{cases} \mathbf{x}^\top \mathbf{w} + w_0 & \geq 1 \quad \forall \text{ +ve data } \mathbf{x} \\ \mathbf{x}^\top \mathbf{w} + w_0 & \leq 1 \quad \forall \text{ -ve data } \mathbf{x} \end{cases} \tag{52}$$

therefore, the goal is to find $\mathbf{w}, w_0$ to make the have the **maximum margin**

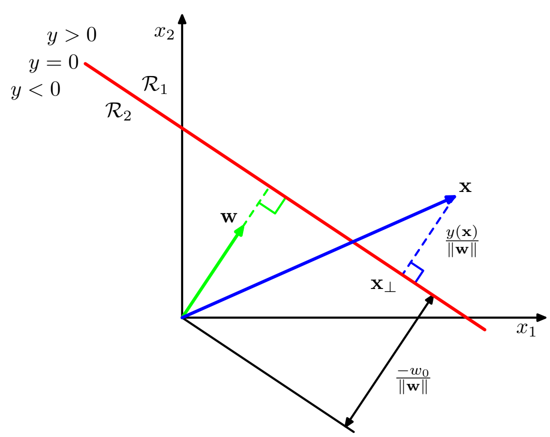
8.1.6 expression for margin

let r be the margin, i.e., perpendicular distance between arbitrary point $\mathbf{x}$ from the middle of the decision surface

Let's see how it is relate to the parameters $\mathbf{w}$ and/or w_0:

$$
\begin{aligned}
\mathbf{x} &= \mathbf{x}_\perp + r\frac{\mathbf{w}}{\|\mathbf{w}\|} \qquad \text{sum of these two vectors} \\
\implies \underbrace{\mathbf{w}^\top\mathbf{x} + w_0}_{y(\mathbf{x})} &= \mathbf{w}^\top\left(\mathbf{x}_\perp + r\frac{\mathbf{w}}{\|\mathbf{w}\|}\right) + w_0 \qquad \text{apply } (\mathbf{w}^\top \times \quad +w_0) \text{ to both sides} \\
\implies y(\mathbf{x}) &= \underbrace{\mathbf{w}^\top\mathbf{x}_\perp + w_0}_{=0} + \mathbf{w}^\top r\frac{\mathbf{w}}{\|\mathbf{w}\|} \\
\implies y(\mathbf{x}) &= r\frac{\mathbf{w}^\top\mathbf{w}}{\|\mathbf{w}\|} = r\frac{\|\mathbf{w}\|^2}{\|\mathbf{w}\|} \\
\implies r &= \frac{y(\mathbf{x})}{\|\mathbf{w}\|}
\end{aligned} \tag{53}
$$

since we want to maximize margins between $y(\mathbf{x}) = +1$ and $y(\mathbf{x}) = -1$, the margin to be maximized must be $\frac{2}{\|\mathbf{w}\|}$:



$$
\max(\text{margin})_{\mathbf{w},w_0} \implies \max\left(\frac{2}{\|\mathbf{w}\|}\right)
$$

$$
\text{subject to: } \begin{cases} \min(\mathbf{w}^T\mathbf{x}_i + w_0) = 1 & i : y_i = +1 \\ \max(\mathbf{w}^T\mathbf{x}_i + w_0) = -1 & i : y_i = -1 \end{cases}
$$

the two inequality constraints can be written as one:

$$
\implies \text{subject to: } \underbrace{y_i(\mathbf{w}^T\mathbf{x}_i + w_0)}_{\text{both need to be SAME sign}} \geq 1 \quad \forall i
$$

$$
\implies \text{subject to: } 1 - y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \leq 0 \geq 1 \quad \forall i
$$

8.1.7 primal optimization

$$\begin{aligned} &\min \left(\frac{1}{2}\|\mathbf{w}\|^2\right) \\ &\text{subject to:} \quad 1 - y_i(\mathbf{w}^T x_i + w_0) \le 0 \quad \forall i \end{aligned} \tag{54}$$

8.2 Lagrangian Dual for SVM

in primal form, there is no kernel trick to exploit. So people are motivated to solve this in its **Lagrange dual**. there is no equality constraint in this case:

$$\mathcal{L}(\underbrace{\mathbf{w}, w_0}_{\mathbf{x}}, \underbrace{\lambda}_{\text{there is no }\mu}) = \underbrace{\frac{1}{2}\|\mathbf{w}\|^2}_{f(\mathbf{x})} + \underbrace{\sum_{i=1}^{p} \mu_i h_i(\mathbf{x})}_{=0} + \sum_{i=1}^{N} \lambda_i \underbrace{[1 - y_i(\mathbf{w}^\top \mathbf{x}_i + w_0)]}_{g_i(\mathbf{x})} \tag{55}$$

to solve $\mathbf{x}'$ for $\min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \mu, \lambda)$, i.e., $\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}', \mu, \lambda) = 0$

$$\begin{aligned} \frac{\partial \mathcal{L}(\mathbf{w}, w_0, \lambda)}{\partial \mathbf{w}} &= w - \sum_{i=1}^{N} \lambda_i y_i \mathbf{x}_i = 0 \implies \mathbf{w}' = \sum_{i=1}^{N} \lambda_i y_i \mathbf{x}_i \\ \frac{\partial \mathcal{L}(\mathbf{w}, w_0, \lambda)}{\partial w_0} &= \underbrace{\sum_{i=1}^{N} \lambda_i y_i}_{\text{not a function of } w_0} = 0 \end{aligned} \tag{56}$$

8.3 write expression for $f_\lambda^{(\star)}(\lambda)$

substitute $\mathbf{x}'$ (in terms of λ), i.e.,:

$$\begin{cases} \mathbf{w}' & = \sum_{i=1}^{n} \lambda_i y_i \mathbf{x}_i \\ \sum_{i=1}^{n} \lambda_i y_i & = 0 \end{cases}$$

to $\quad \mathcal{L}(\mathbf{w}, w_0, \lambda) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{n} \lambda_i [1 - y_i(\mathbf{w}^\top \mathbf{x}_i + w_0)]$

$$\begin{aligned} \implies f_\lambda^{(\star)}(\lambda) &= \inf_{\mathbf{x}} \mathcal{L}(\mathbf{w}, w_0, \lambda) \\ &= \frac{1}{2}\Big(\sum_{i=1}^{n} \lambda_i y_i \mathbf{x}_i\Big)^\top \Big(\sum_{i=1}^{n} \lambda_i y_i \mathbf{x}_i\Big) + \sum_{i=1}^{n} \lambda_i \Big[1 - y_i\Big(\Big(\sum_{i=1}^{n} \lambda_i y_i \mathbf{x}_i\Big)^\top \mathbf{x}_i + w_0\Big)\Big] \\ &= \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j - \sum_{i=1}^{n} \lambda_i y_i \Big(\sum_{j=1}^{n} \lambda_j y_j \mathbf{x}_j^\top\Big) \mathbf{x}_i - w_0 \underbrace{\sum_{i=1}^{n} \lambda_i y_i}_{=0} + \sum_{i=1}^{n} \lambda_i \\ &= \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j \\ &\text{subject to: } \sum_{i=1}^{N} \lambda_i y_i = 0 \text{ and } \lambda_i \ge 0 \end{aligned} \tag{57}$$

8.4 The dual problem

$$\arg\max_{\lambda_1,\ldots\lambda_n} \mathcal{L}_\lambda(\lambda) = \arg\max_{\lambda_1,\ldots\lambda_n} \left(\sum_{i=1}^n \lambda_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j \right)$$
$$\text{subject to: } \sum_{i=1}^n \lambda_i y_i = 0 \text{ and } \lambda_i \geq 0 \tag{58}$$

since $\mathbf{x}_i^\top \mathbf{x}_j$ can be replaced by kernel $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)$

Use **complementary slackness:**

$$\begin{aligned}
\lambda_i^* > 0 &\implies g_i(\mathbf{w}^*, {w_0}^*) = 0 \\
&\implies 1 - y_i({\mathbf{w}^*}^\top \mathbf{x}_i + {w_0}^*) = 0 \\
&\implies y_i({\mathbf{w}^*}^\top \mathbf{x}_i + {w_0}^*) = 1 \\
&\qquad\qquad\qquad\qquad \text{i.e., } \mathbf{x}_i \text{ is support vector points} \\
\lambda_i^* = 0 &\implies g_i(w^*, {w_0}^*) < 0 \\
&\implies 1 - y_i({\mathbf{w}^*}^\top x_i + {w_0}^*) < 0 \\
&\implies y_i({\mathbf{w}^*}^\top x_i + {w_0}^*) > 1 \\
&\qquad\qquad\qquad\qquad \text{i.e., } \mathbf{x}_i \text{ is non support vector points}
\end{aligned} \tag{59}$$

8.4.1 inference

substitute a new x into the dual inference algorithm and knowing that $\mathbf{w}' = \sum_{i=1}^N \lambda_i y_i \mathbf{x}_i$ from Eq.(56):

$$y = \mathbf{w}^\top x + w_0 = \left(\sum_{i=1}^N \alpha_i y_i \mathbf{x}_i \right)^\top x + w_0 = \sum_{i=1}^n \alpha_i y_i \langle \mathbf{x}_i, \mathbf{x} \rangle + w_0 \tag{60}$$

Since there is only a few $\lambda_i > 0$, dual inference is **efficient**!

9 Farkas Lemma

9.1 application: prove Strong duality in Linear Programming

before discussion Farkas Lemma, let's first look duality in Linear Programming

$$\min_{\mathbf{x}} \left[\mathcal{C}^\top \mathbf{x} \;\middle|\; \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq 0 \right] \tag{61}$$

$$\begin{aligned}
q(\lambda) &= \inf_{\mathbf{x} \geq \mathbf{0}} \left[\mathcal{L}(\mathbf{x}, \lambda) \right] \\
&= \inf_{\mathbf{x} \geq 0} \left[\mathcal{C}^\top \mathbf{x} + \lambda^\top (\mathbf{A}\mathbf{x} - \mathbf{b}) \right] \quad \text{only equality constraint} \\
&= \inf_{\mathbf{x} \geq 0} \left[(\mathcal{C}^\top + \lambda^\top \mathbf{A})\mathbf{x} - \lambda^\top \mathbf{b} \right] \\
&= \inf_{\mathbf{x} \geq 0} \left[(\mathcal{C}^\top + \lambda^\top \mathbf{A})\mathbf{x} \right] - \lambda^\top \mathbf{b}
\end{aligned} \tag{62}$$

firstly to note that it does not appear that $\inf_{\mathbf{x}\geq 0}\big[(\mathcal{C}^\top + \lambda^\top \mathbf{A})\mathbf{x}\big]$ can be solved using dual norm nor convex conjugate. But it can be solved using heuristic methods:

$$\begin{cases} (\mathcal{C}^\top + \lambda^\top \mathbf{A}) < 0 & \implies \inf_{\mathbf{x}\geq 0}\big[(\mathcal{C}^\top + \lambda^\top \mathbf{A})\mathbf{x}\big] = -\infty \quad (\text{sub } \mathbf{x} = \infty) \\ \\ (\mathcal{C}^\top + \lambda^\top \mathbf{A}) \geq 0 & \implies \inf_{\mathbf{x}\geq 0}\big[(\mathcal{C}^\top + \lambda^\top \mathbf{A})\mathbf{x}\big] = 0 \quad (\text{sub } \mathbf{x} = 0) \end{cases} \tag{63}$$

so adding the $-\lambda^\top \mathbf{b}$ part, we have:

$$q(\lambda, \lambda) = [-\lambda^\top \mathbf{b} \mid (\mathcal{C}^\top + \lambda^\top \mathbf{A}) \geq 0] \tag{64}$$

in the above example, primal constraint $\mathbf{x} \geq 0$ is brought to the $q(\lambda)$ to help to partition regions and obtain the dual feasibility.

$$\begin{aligned} &\max_{\lambda}\big[-\lambda^\top \mathbf{b} \mid \mathcal{C}^\top + \lambda^\top \mathbf{A} \geq 0\big] \\ \text{or let } \lambda' = -\lambda : & \\ &\max_{\lambda'}\big[\lambda'^\top \mathbf{b} \mid \mathcal{C}^\top \geq \lambda'^\top \mathbf{A}\big] \end{aligned} \tag{65}$$

in summary:

$$\begin{cases} & \textbf{primal form:} \\ z^* & = \min\left(\mathcal{C}^\top \mathbf{x}\right) \\ & \text{s.t. } \mathbf{A}\mathbf{x} = \ \mathbf{b} \\ & \text{and } \mathbf{x} \geq \ \mathbf{0} \end{cases} \qquad \begin{cases} & \textbf{dual form:} \\ \tilde{z} & = \max\left(\mathbf{b}^\top \color{red}{\lambda}\right) \\ & \text{s.t. } \mathbf{A}^\top \color{red}{\lambda} \leq \ \mathcal{C} \\ & \color{red}{\lambda} \text{ is dual variable} \end{cases} \tag{66}$$

9.1.1 a "hack" solution

using dual feasibility condition $\mathcal{C}^\top \geq \lambda^\top \mathbf{A} \implies \lambda^\top \mathbf{A} \leq \mathcal{C}^\top$

$$\begin{aligned} & \lambda^\top \mathbf{A} \leq \mathcal{C}^\top \ \forall \ \lambda \\ & \lambda^\top \mathbf{A}\mathbf{x}^* \leq \mathcal{C}^\top \mathbf{x}^* \ \forall \ \lambda \quad \mathbf{x}^* \text{is optimal solution and since } \mathbf{x}^* \geq 0 \text{, no change sign} \\ \implies & \lambda^\top \underbrace{\mathbf{b}} \leq \mathcal{C}^\top \mathbf{x}^* \ \forall \lambda \ \text{ using } \mathbf{A}\mathbf{x}^* = \mathbf{b} \\ & \qquad\quad = \min_{\mathbf{x}} \big[\mathcal{C}^\top \mathbf{x} \,\big|\, \mathbf{A}\mathbf{x} = \mathbf{b},\ \mathbf{x} \geq 0\big] \\ \implies & \underbrace{\max_{\lambda}[\lambda^\top \mathbf{b}]}_{\lambda^*} \leq \min_{\mathbf{x}} \big[\mathcal{C}^\top \mathbf{x} \,\big|\, \mathbf{A}\mathbf{x} = \mathbf{b},\ \mathbf{x} \geq 0\big] \\ \implies & \max_{\lambda} \big[\lambda^\top \mathbf{b} | \color{blue}{\lambda^\top \mathbf{A} \leq \mathcal{C}^\top \ \forall \ \lambda}\big] \leq \min_{\mathbf{x}} \big[\mathcal{C}^\top \mathbf{x} \,\big|\, \mathbf{A}\mathbf{x} = \mathbf{b},\ \mathbf{x} \geq 0\big] \qquad \text{write the condition in} \end{aligned} \tag{67}$$

9.1.2 W-GAN Linear Programming Primal and Dual form

$$\begin{cases} & \textbf{primal form:} \\ z^* & = \min\left(\mathcal{C}^\top \mathbf{\Gamma}\right) \\ & \text{s.t. } \mathbf{A\Gamma} = \ \mathbf{b} \\ & \text{and } \mathbf{\Gamma} \geq \ \mathbf{0} \end{cases} \qquad \begin{cases} & \textbf{dual form:} \\ \tilde{z} & = \max\left(\mathbf{b}^\top \lambda\right) \\ & \text{s.t. } \mathbf{A}^\top \lambda \leq \ \mathcal{C} \\ & \lambda \text{ is dual variable} \end{cases} \tag{68}$$

1. $\mathbf{\Gamma} \equiv \gamma(x, y)$ acts like a vectorized joint distribution, each element ≥ 0
2. $\mathcal{C} \equiv \text{vec}(\mathbf{D}(x, y))$ acts like a vectorized cost
3. $\mathbf{b} = \begin{bmatrix} p_{\text{r}}(y) \\ p_{\text{g}}^{\theta}(x) \end{bmatrix}$

and we optimize W-GAN on the dual form

9.2 Convex and Conic combination

matrix $\mathbf{A} \in \mathbb{R}^{d \times n} \triangleq (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n)$

Definition 1 *Convex combination:*

$$C = \{\mathbf{a} \big| \mathbf{a} = \alpha_1 \mathbf{a}_1 + ... + \alpha_k \mathbf{a}_k, \alpha_1 + ... + \alpha_k = 1, \alpha_i \geq 0\} \tag{69}$$

for example $\mathbf{A} \in \mathbb{R}^{2\times 3}$*, then it looks like a painted triangle*

Definition 2 *Conic combination is:*

$$C = \{\mathbf{a} \big| \mathbf{a} = \alpha_1 \mathbf{a}_1 + ... + \alpha_k \mathbf{a}_k, \alpha_i \geq 0\} \tag{70}$$

for example $\mathbf{A} \in \mathbb{R}^{2\times 3}$*, it looks painted cone from the origin*

Lemma 2 **Farkas Lemma** *say, for a vector* $\mathbf{b}$*, there are exactly two* **mutually exclusive** *possibilities:*

1. $\mathbf{b}$ *inside the cone:*

$$\exists\, \mathbf{x} \in \mathbb{R}^n, \mathbf{x} \geq 0 \textit{ (in every dimension) } \textit{ s.t. } \mathbf{Ax} = \mathbf{b} \tag{71}$$

2. $\mathbf{b}$ *outside the cone:*

$$\begin{aligned} \nexists\, \mathbf{x} \in \mathbb{R}^n, \mathbf{x} \geq 0 \textit{ (in every dimension) } \textit{ s.t. } \mathbf{Ax} = \mathbf{b} \\ \forall\, \mathbf{x} \geq 0, \textit{ (in every dimension) } \textit{ s.t. } \mathbf{Ax} \neq \mathbf{b} \end{aligned} \tag{72}$$

these are not the most useful definitions, we use instead:

$$\exists \lambda \in \mathbb{R}^m, \textit{s.t. } \mathbf{A}^\top \lambda \leq 0 \textit{ and } \mathbf{b}^\top \lambda > 0 \tag{73}$$

if one draws a line $\lambda_\perp$ which is perpendicular to λ. Then every $\mathbf{a}_i \in \mathbf{A}$ is on one side of $\lambda_\perp$ (including $\lambda_\perp$), and $\mathbf{b}$ is on the other side. Therefore $\mathbf{b}$ must be outside of cone $\mathbf{A}$

if the case we have proven existence of $\mathbf{x}$ for case **one** (i.e., b inside the cone), $\implies$ case **two** is not possible. Therefore, alternative way of saying: $\nexists\ \lambda \in \mathbb{R}^m$, s.t. $\mathbf{A}^\top \lambda \leq 0$ and $\mathbf{b}^\top \lambda > 0$ is:

$$\forall\ \lambda \in \mathbb{R}^m : \mathbf{A}^\top \lambda \leq 0 \implies \mathbf{b}^\top \lambda \leq 0 \tag{74}$$

which is something we will use in the rest of the proof

9.3 use Farkas Lemma to prove strong duality

need to prove:

$$\begin{aligned} &\min_{\mathbf{x}} \left[\mathcal{C}^\top \mathbf{x} \,\middle|\, \mathbf{A}\mathbf{x} = \mathbf{b},\, \mathbf{x} \geq 0 \right] \\ &= \max_{\lambda} \left[\lambda^\top \mathbf{b} \,\middle|\, \lambda^\top \mathbf{A} \leq \mathcal{C}^\top\ \forall\ \lambda \right] \end{aligned} \tag{75}$$

let

$$z^* = \min_{\mathbf{x}} \left[\mathcal{C}^\top \mathbf{x} \,\middle|\, \mathbf{A}\mathbf{x} = \mathbf{b},\, \mathbf{x} \geq 0 \right] \quad \text{be min in primal occur at } \mathbf{x}^* \tag{76}$$

9.3.1 "larger" system extension

first we extend a single dimension by adding one more equation $\mathcal{C}^\top \mathbf{x} = z^* - \epsilon$ or $-\mathcal{C}^\top \mathbf{x} = -z^* + \epsilon$ (we need to use the latter for Eq.(88)).

So now the "larger" system contain both equations of the primal:

$$\begin{cases} \mathbf{A}\mathbf{x} & = \mathbf{b} \\ -\mathcal{C}^\top \mathbf{x} & = -z^* + \epsilon \qquad \iff \qquad \mathcal{C}^\top \mathbf{x} = z^* - \epsilon \end{cases} \tag{77}$$

put them in a linear system form:

$$\hat{\mathbf{A}} = \begin{bmatrix} \mathbf{A} \\ -\mathcal{C}^\top \end{bmatrix}, \quad \hat{\mathbf{b}}_\epsilon = \begin{bmatrix} \mathbf{b} \\ -z^* + \epsilon \end{bmatrix} \tag{78}$$

the reason to use $-z^* + \epsilon$ is to control which Farkas condition the "larger" system because we can then change ϵ to decide which Farkas case:

$$\begin{cases} \epsilon = 0: & \mathbf{A}\mathbf{x}^* = \mathbf{b} \quad \wedge \quad \mathcal{C}^\top \mathbf{x}^* = z^* \qquad \implies \text{Farkas 1} \\ \\ \epsilon > 0: & \nexists \mathbf{x} \text{ s.t., } (\mathbf{A}\mathbf{x} = \mathbf{b}) \wedge (\mathcal{C}^\top \mathbf{x} = z^* - \epsilon) \implies \text{Farkas 2} \end{cases} \tag{79}$$

$\mathcal{C}^\top \mathbf{x} = z^*$ only has solution $\mathbf{x} = \mathbf{x}^*$ when $\epsilon = 0$. However, when $\epsilon > 0$, there is no $\mathbf{x}$ satisify. z^* is already the minimal solution! So even $\mathbf{x}^*$ can't be feasible, let alone any other $\mathbf{x}$.

we also extend the dual variable λ by one dimension too $\alpha \in \mathbb{R}$. Note that we should not place any constraint on it:

$$\hat{\lambda} = \begin{bmatrix} \lambda \\ \alpha \end{bmatrix} \quad \text{where } \alpha \in \mathbb{R} \tag{80}$$

note that $\mathbf{x}$ does **not** extend, so it can be applied in both systems

9.3.2 what are we going to prove?

we then prove:

$$\tilde{z} = \max_{\lambda} \left[\mathbf{b}^\top \lambda \mid \mathbf{A}^\top \lambda \leq \mathcal{C} \right] > z^* - \epsilon \quad \forall \epsilon > 0 \tag{81}$$

it is obvious $\tilde{z} \in \big((z^* - \epsilon), z^*\big)$, where z^* is the primal minimum. Then by making ϵ infinitely small, we get:

$$\tilde{z} = z^* \tag{82}$$

9.3.3 Prove $\alpha > 0$ using Farkas Lemma

looking at our extension in Section (9.3.1), we have:

1. let $\epsilon = 0$

   since it's Farkas case (1), then Farkas (2) can not exist, i.e., repeating Eq.(74):

   α-condition 1:

$$\forall \hat{\lambda} : \ \hat{\mathbf{A}}^\top \hat{\lambda} \leq 0 \implies \hat{\mathbf{b}}_0^\top \hat{\lambda} \leq 0 \tag{83}$$

2. let $\epsilon > 0$, there exists **no** non-negative solution, meaning $\forall \mathbf{x}\ \hat{\mathbf{A}}\mathbf{x} \neq \hat{\mathbf{b}}_\epsilon$

   if Farkas(1) does not exist, then Farkas (2) must exist, i.e.:

   $\exists\ \hat{\lambda} : \hat{\mathbf{A}}^\top \hat{\lambda} \leq 0 \implies \mathbf{b}_\epsilon^\top \hat{\lambda} > 0$

$$\begin{aligned} 0 < \hat{\mathbf{b}}_\epsilon^\top \hat{\lambda} &= \mathbf{b}^\top \lambda + \alpha(-z^* + \epsilon) \\ &= \underbrace{\mathbf{b}^\top \lambda + \alpha(-z^*)}_{\hat{\mathbf{b}}_0^\top \hat{\lambda}} + \alpha\epsilon \\ &= \hat{\mathbf{b}}_0^\top \hat{\lambda} + \alpha\epsilon \end{aligned} \tag{84}$$

   α-condition 2: $\quad \epsilon > 0$:

$$\exists \hat{\lambda}\ \hat{\mathbf{A}}^\top \hat{\lambda} \leq 0 \implies \hat{\mathbf{b}}_0^\top \hat{\lambda} + \alpha\epsilon > 0 \tag{85}$$

3. combine both to prove $\alpha > 0$

$$\begin{cases} \alpha\text{-condition 1, } \epsilon = 0: & \forall \hat{\lambda}: \quad \hat{\mathbf{A}}^\top \hat{\lambda} \leq 0 \implies \hat{\mathbf{b}}_0^\top \hat{\lambda} \leq 0 \\ \\ \alpha\text{-condition 2, } \epsilon > 0: & \exists \hat{\lambda}: \quad \hat{\mathbf{A}}^\top \hat{\lambda} \leq 0 \implies \hat{\mathbf{b}}_0^\top \hat{\lambda} + \alpha\epsilon > 0 \end{cases} \tag{86}$$

therefore, to find a $\hat{\lambda}$ to satisfy both:

$$\hat{\mathbf{b}}_0^\top \hat{\lambda} \leq 0 \quad \text{and} \quad \hat{\mathbf{b}}_0^\top \hat{\lambda} + \alpha\epsilon > 0 \tag{87}$$

we must have $\alpha > 0$. Othewise, if we let $\alpha \leq 0 \implies \hat{\mathbf{b}}_0^\top \hat{\lambda} \geq 0$ for α-condition 2, which contradicts α-condition 1.

9.3.4 Prove $\tilde{z} > z^* - \epsilon$ using Farkas Lemma

we just proved that $\alpha > 0$, which implies by it won't change sign when dividing by α in Eq.(89).

We saw when $\epsilon > 0$, there exists no non-negative solution of $\mathbf{x}$, this $\implies$ it is Farkas case (2): meaning when $\epsilon > 0$ (i.e., Farkas (1) does not exist), then, there exist $\hat{\lambda} \equiv \begin{bmatrix} \lambda \\ \alpha \end{bmatrix}$ solution such that:

$$\begin{aligned} &\hat{\mathbf{A}}^\top \hat{\lambda} \leq 0 \quad \wedge \quad \mathbf{b}_\epsilon^\top \hat{\lambda} > 0 \\ \implies &\underbrace{\begin{bmatrix} \mathbf{A} \\ -\mathcal{C}^\top \end{bmatrix}^\top \begin{bmatrix} \lambda \\ \alpha \end{bmatrix} \leq \mathbf{0}}_{\implies \mathbf{A}^\top \lambda \leq \alpha \mathcal{C}} \quad \wedge \quad \underbrace{\begin{bmatrix} \mathbf{b} \\ -z^* + \epsilon \end{bmatrix} \begin{bmatrix} \lambda \\ \alpha \end{bmatrix} > 0}_{\implies \mathbf{b}^\top \lambda > \alpha(z^* - \epsilon)} \end{aligned} \tag{88}$$

further massage the equations:

$$\begin{aligned} \mathbf{A}^\top \lambda \leq \alpha \mathcal{C} &\implies \mathbf{A}^\top \frac{\lambda}{\alpha} \leq \mathcal{C} \quad \text{from L.H.S of Eq.(88)} \\ \mathbf{b}^\top \lambda > \alpha(z^* - \epsilon) &\implies \mathbf{b}^\top \frac{\lambda}{\alpha} > (z^* - \epsilon) \quad \text{from R.H.S of Eq.(88)} \end{aligned} \tag{89}$$

now we have: $\mathbf{A}^\top \frac{\lambda}{\alpha} \leq \mathcal{C}$ and $\mathbf{b}^\top \frac{\lambda}{\alpha} > (z^* - \epsilon)$, since any α works, we choose $\alpha = 1$:

$$\underbrace{\mathbf{A}^\top \lambda \leq \mathcal{C}}_{\text{constraint}} \quad \text{and} \quad \underbrace{\mathbf{b}^\top \lambda > (z^* - \epsilon)}_{\text{obj}} \tag{90}$$

combine the two above, we have:

$$\tilde{z} = \max_{\lambda} \left[\mathbf{b}^\top \lambda \, \middle| \mathbf{A}^\top \lambda \leq \mathcal{C} \right] > z^* - \epsilon \tag{91}$$

we can make ϵ arbitrarily small, to make $\tilde{z} = z^*$, so we have **strong** duality!

References

[1] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro, "The implicit bias of gradient descent on separable data," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 2822–2878, 2018.