A Quick Tutorial on Duality

Richard Xu

January 3, 2021

1 Optimization with inequality constraints

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

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

After defined $\mathbf{I}(u) = \begin{cases} 0, & \text{if } u \leq 0 \\ \infty, & \text{otherwise} \end{cases}$, we can turn a constrained equation using **unconstrained** equation:

$$J(x) = f(x) + \sum_i \mathbf{I}[g_i(x)] \tag{2}$$

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

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

$$J(x) = f(x) - \sum_i \mathbf{I}[g_i(x)] \tag{3}$$

2 Lower Bound constraints

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

$$\mathcal{L}(x, \lambda) = f(\mathbf{x}) + \sum_i \lambda_i g_i(\mathbf{x}) \tag{4}$$

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

$$\begin{aligned} &\mathcal{L}(x, \lambda) \leq J(\mathbf{x}) \\ \text{i.e.,} \quad &\max_\lambda \mathcal{L}(\mathbf{x}, \lambda) = J(\mathbf{x}) \end{aligned} \tag{5}$$

if we were to minimize both side for $\mathbf{x}$:

$$
\begin{aligned}
p^* &= \min_{\mathbf{x}} \max_{\lambda} \mathcal{L}(\mathbf{x}, \lambda) \\
&= \min_{\mathbf{x}} J(\mathbf{x})
\end{aligned} \tag{6}
$$

This means that for $\mathcal{L}(\mathbf{x}, \lambda)$ we maximize λ first, then minimize $\mathbf{x}$ and we obtain $J(\mathbf{x})$. However, it's point-less to do it in this order

3 swap the order: $\min_x$ first, then $\max_\lambda$

from Eq(6)

$$
\begin{aligned}
\min_{\mathbf{x}} \max_{\lambda} \mathcal{L}(\mathbf{x}, \lambda) &= \min_{\mathbf{x}} J(\mathbf{x}) \\
\implies \max_{\lambda} \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \lambda) \leq \min_{\mathbf{x}} \max_{\lambda} \mathcal{L}(\mathbf{x}, \lambda) &= \min_{\mathbf{x}} J(\mathbf{x}) \\
\implies \left(d^* \equiv \max_{\lambda} \min_{x} \mathcal{L}(\mathbf{x}, \lambda)\right) \leq \left(p^* \equiv \min_{\mathbf{x}} \max_{\lambda} \mathcal{L}(\mathbf{x}, \lambda) \right.&= \left. \min_{\mathbf{x}} J(\mathbf{x})\right)
\end{aligned} \tag{7}
$$

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

$$
\sup_{\lambda \in \Lambda} \inf_{x \in \mathcal{X}} f(\lambda, x) \leq \inf_{x \in \mathcal{X}} \sup_{\lambda \in \Lambda} f(\lambda, x) \tag{8}
$$

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

$$
\begin{aligned}
&\inf_{x} f(\lambda, x) \leq f(\lambda, x), \forall \lambda \, \forall x \\
\implies &\sup_{\lambda} \inf_{x} f(\lambda, x) \leq \sup_{\lambda} f(\lambda, x), \forall x \qquad \sup_{\lambda} \text{ both sides} \\
\implies &\sup_{\lambda} \inf_{x} f(\lambda, x) \leq \inf_{x} \sup_{\lambda} f(\lambda, x) \qquad \text{on RHS: } \because \inf_{x} \in \forall x
\end{aligned} \tag{9}
$$

if strong duality holds:

$$
d^* = p^* \tag{10}
$$

3.1 duality summary

in summary, the duality procedure is:

$$
d^* \equiv \max_{\lambda} \min_{x} \mathcal{L}(\mathbf{x}, \lambda) \tag{11}
$$

3.2 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:

$$\begin{aligned} &f(\cdot, y) : X \to \mathbb{R} \text{ is convex for fixed } y \\ &f(x, \cdot) : Y \to \mathbb{R} \text{ is concave for fixed } x \end{aligned} \tag{12}$$

then:

$$\min_{x \in X} \max_{y \in Y} f(x, y) = \max_{y \in Y} \min_{x \in X} f(x, y) \tag{13}$$

4 complementary slackness

4.1 when constraints are all satisfied: i.e., $g_i(\mathbf{x}^*) \leq 0 \; \forall i$

$$\mathcal{L}(\mathbf{x}, 0) = f(\mathbf{x}) \tag{14}$$

best λ_i occurs when:

$$\lambda_i^* = \arg\max_{\lambda_i} \mathcal{L}(x, \lambda_i) = 0 \tag{15}$$

this is because $\lambda_i \geq 0$, in case:

$$g_i(\mathbf{x}) \leq 0 \text{ and } \lambda_i > 0 \implies \lambda_i g_i(\mathbf{x}) \leq 0 \tag{16}$$

so **max** occur when $\lambda_i = 0$

4.2 When constraints are not all satisfied: $\exists_i \; g_i(\mathbf{x}^*) > 0$

$$\min_{\mathbf{x}} \max_{\lambda} \mathcal{L}(\mathbf{x}, \lambda) = \min_{\mathbf{x}} J(\mathbf{x}) \tag{17}$$

we can **maximize** $\mathcal{L}(\mathbf{x}, \lambda)$ by taking $\lambda_i \to +\infty$. We can see that way to prevent $\mathcal{L}(\mathbf{x}, \lambda)$ going to infinity is to locate new $\mathbf{x}*$ to be "sub-optimal" solution of the unconstrained solution, where:

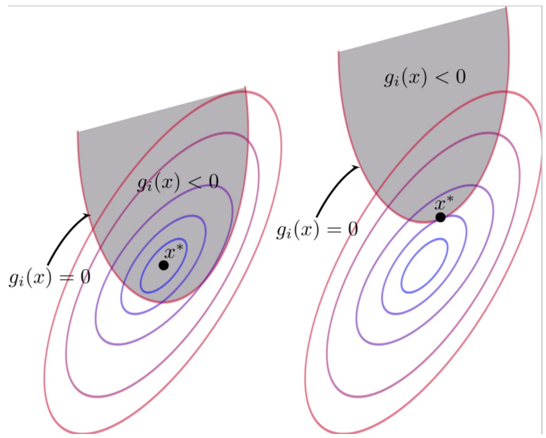
$$g_i(\mathbf{x}^*) = 0 \tag{18}$$

instead of original $\mathbf{x}^*$, optimal unconstrained solution.

The the above two cases, we found either $\lambda_i = 0$ or $g_i(\mathbf{x}) = 0$. We can specify it in a single equation:

$$\lambda_i g_i(\mathbf{x}) = 0 \tag{19}$$

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



5 summary of KKT condition

optimization problem with both equality and inequality constraints:

$$\begin{aligned} \mathbf{x}^* = \underset{\mathbf{x}}{\text{argmin}} f(\mathbf{x}) \\ \text{subject to } h_i(\mathbf{x}) = 0 \\ \text{subject to } g_i(\mathbf{x}) \leq 0 \end{aligned} \tag{20}$$

so how does duality procedure $d^* \equiv \max_\lambda \min_x \mathcal{L}(\mathbf{x}, \lambda)$ being carried out, also since we have additional equality constraint, we now have $\mathcal{L}(\mathbf{x}, \mu, \lambda)$ instead

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

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

   $$\begin{aligned} &\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \mu, \lambda) = 0 \\ \implies &\nabla_{\mathbf{x}} \Big(f(\mathbf{x}) + \sum_{i=1}^{m} \mu_i h_i(\mathbf{x}) + \sum_{i=1}^{n} \lambda_i g_i(\mathbf{x}) \Big) = 0 \\ \implies &\nabla_{\mathbf{x}} f(\mathbf{x}) + \sum_{i=1}^{m} \mu_i \nabla_{\mathbf{x}} h_i(\mathbf{x}) + \sum_{i=1}^{n} \lambda_i \nabla_{\mathbf{x}} g_i(\mathbf{x}) = 0 \end{aligned} \tag{21}$$

   (b) substitute $\mathbf{x}'$ in terms of λ into $\mathcal{L}(\mathbf{x}', \mu, \lambda)$ to obtain:

   $$\mathcal{L}_\lambda(\lambda) = \min_x \mathcal{L}(\mathbf{x}, \mu, \lambda) \tag{22}$$

   note $\mathcal{L}_\lambda(\lambda)$ should contain no $\mathbf{x}$

   then we can $\max_\lambda \mathcal{L}_\lambda(\lambda)$

2. to ensure **equality constraints**

$$
\begin{aligned}
&\nabla_\mu \mathcal{L}(\mathbf{x}, \mu, \lambda) \\
\implies &\nabla_\mu f(\mathbf{x}) + \sum_{i=1}^{m} \mu_i \nabla_\mu h_i(\mathbf{x}) + \sum_{i=1}^{n} \lambda_i \nabla_\mu g_i(\mathbf{x}) = 0 \\
\implies &\sum_{i=1}^{m} \mu_i \nabla_\mu h_i(\mathbf{x}) = 0
\end{aligned} \tag{23}
$$

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

$$
\begin{aligned}
\lambda_i g_i(\mathbf{x}) = 0, &\quad \forall i \\
\lambda_i \geq 0, &\quad \forall i \\
g_i(\mathbf{x}) \leq 0, &\quad \forall i
\end{aligned} \tag{24}
$$

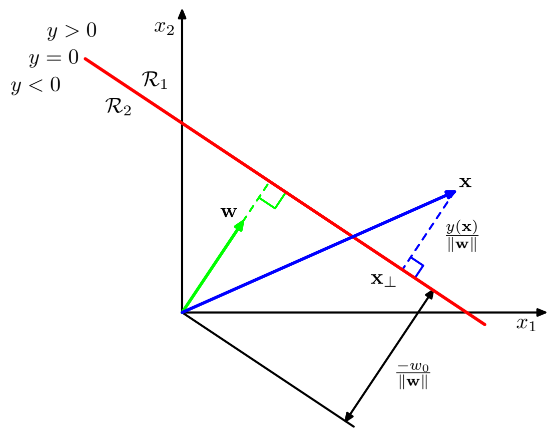
6 example through Support Vector Machine

6.1 Linear Discriminant Function (geometry)

$$
y(\mathbf{x}) = \mathbf{x}^T \mathbf{w} + w_0 \tag{25}
$$

let perpendicular distance r of arbitrary point $\mathbf{x}$ from the decision surface be r, an arbitrary $\mathbf{x}$ can be written as:

$$
\begin{aligned}
&\mathbf{x} = \mathbf{x}_\perp + r\frac{\mathbf{w}}{\|\mathbf{w}\|} \\
\implies &\underbrace{\mathbf{w}^\top \mathbf{x} + w_0}_{y(\mathbf{x})} = \mathbf{w}^\top \left(\mathbf{x}_\perp + r\frac{\mathbf{w}}{\|\mathbf{w}\|} \right) + w_0 \qquad \text{apply } (\mathbf{w}^\top \times \quad + w_0) \text{ to both sides} \\
\implies &y(\mathbf{x}) = \underbrace{\mathbf{w}^\top \mathbf{x}_\perp + w_0}_{=0} + \mathbf{w}^\top r\frac{\mathbf{w}}{\|\mathbf{w}\|} \\
\implies &y(\mathbf{x}) = r\frac{\mathbf{w}^\top \mathbf{w}}{\|\mathbf{w}\|} = r\frac{\|\mathbf{w}\|^2}{\|\mathbf{w}\|} \\
\implies &r = \frac{y(\mathbf{x})}{\|\mathbf{w}\|}
\end{aligned} \tag{26}
$$



our goal is to maximize margin r:

$$\max(\text{margin})_{\mathbf{w},w_0} = \max\left(\frac{2}{\|\mathbf{w}\|}\right)$$
$$\text{subject to: } \begin{cases} \min(\mathbf{w}^T x_i + w_0) = 1 & i : y_i = +1 \\ \max(\mathbf{w}^T x_i + w_0) = -1 & i : y_i = -1 \end{cases}$$

resulting classifier $y = \text{sign}(\mathbf{w}^T + w_0)$ can be re-written as the **primal optimization**:

$$\begin{aligned} &\min\left(\frac{1}{2}\|\mathbf{w}\|^2\right) \\ &\text{subject to: } \underbrace{y_i(\mathbf{w}^T x_i + w_0)}_{\text{both need to be SAME sign}} \geq 1 \\ &\implies 1 - y_i(\mathbf{w}^T x_i + w_0) \leq 0 \end{aligned} \tag{27}$$

6.2 Lagrangian Dual for SVM

in primal, there is no kernel trick to exploit. can be written in **Lagrange dual**. there is no equality constraint

$$\mathcal{L}(\underbrace{w, b,}_{\mathbf{x}} \underbrace{\lambda}_{\text{there is no } \mu}) = \underbrace{\frac{1}{2}\|\mathbf{w}\|^2}_{f(\mathbf{x})} + \underbrace{\sum_{i=1}^{p} \mu_i h_i(\mathbf{x})}_{=0} + \sum_{i=1}^{N} \lambda_i \underbrace{[1 - y_i(w^T x_i + w_0)]}_{g_i(\mathbf{x})} \tag{28}$$

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

$$
\begin{aligned}
\frac{\partial \mathcal{L}(w,b,\lambda)}{\partial w} &= w - \sum_{i=1}^{N} \lambda_i y_i x_i = 0 \implies w' = \sum_{i=1}^{N} \lambda_i y_i x_i \\
\frac{\partial \mathcal{L}(w,b,\lambda)}{\partial b} &= \underbrace{\sum_{i=1}^{N} \lambda_i y_i}_{\text{not a function of } b} = 0
\end{aligned} \tag{29}
$$

6.3 write expression for $\mathcal{L}_\lambda(\lambda)$

Substitute $\mathbf{x}'$, i.e., $w' = \sum_{i=1}^{n} \lambda_i y_i x_i$ and $\sum_{i=1}^{n} \lambda_i y_i = 0$ to:

$$
\mathcal{L}(w,b,\lambda) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{n} \lambda_i [1 - y_i(w^\top x_i + w_0)] \tag{30}
$$

$$
\begin{aligned}
\mathcal{L}_\lambda(\lambda) &= \inf_x \mathcal{L}(w,b,\lambda) \\
&= \frac{1}{2}\Big(\sum_{i=1}^{n} \lambda_i y_i x_i\Big)^\top \Big(\sum_{i=1}^{n} \lambda_i y_i x_i\Big) + \sum_{i=1}^{n} \lambda_i \Big[1 - y_i\Big(\Big(\sum_{i=1}^{n} \lambda_i y_i x_i\Big)^\top x_i + w_0\Big)\Big] \\
&= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j x_i^\top x_j - \sum_{i=1}^{n} \lambda_i y_i \Big(\sum_{j=1}^{n} \lambda_j y_j x_j^\top\Big) x_i - \underbrace{w_0 \sum_{i=1}^{n} \lambda_i y_i}_{=0} + \sum_{i=1}^{n} \lambda_i \\
&= \sum_{i=1}^{n} \lambda_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j x_i^\top x_j \\
&\text{subject to: } \sum_{i=1}^{N} \lambda_i y_i = 0 \text{ and } \lambda_i \geq 0
\end{aligned} \tag{31}
$$

6.4 The dual problem

$$
\begin{aligned}
\arg\max_{\lambda_1,\ldots\lambda_n} \mathcal{L}_\lambda(\lambda) &= \arg\max_{\lambda_1,\ldots\lambda_n} \left(\sum_{i=1}^{n} \lambda_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j x_i^\top x_j\right) \\
&\text{subject to: } \sum_{i=1}^{n} \lambda_i y_i = 0 \text{ and } \lambda_i \geq 0
\end{aligned} \tag{32}
$$

since $x_i^\top x_j$ can be replaced by kernel $\mathcal{K}(x_i, x_j)$

Use **complementary slackness:**

$$
\begin{aligned}
\lambda_i^* > 0 &\implies g_i(w^*, b^*) = 0 \\
&\implies 1 - y_i({w^*}^\top x_i + {w_0}^*) = 0 \\
&\implies y_i({w^*}^\top x_i + {w_0}^*) = 1 \\
&\qquad\qquad\qquad\qquad \text{i.e., } x_i \text{ is support vector points} \\
\lambda_i^* = 0 &\implies g_i(w^*, b^*) < 0 \\
&\implies 1 - y_i({w^*}^\top x_i + {w_0}^*) < 0 \\
&\implies y_i({w^*}^\top x_i + {w_0}^*) > 1 \\
&\qquad\qquad\qquad\qquad \text{i.e., } x_i \text{ is non support vector points}
\end{aligned} \tag{33}
$$

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