A Tutorial on Duality

Richard Xu

March 19, 2021

1 Motivation

inequality-constrainted optimization often appear in Machine Leanring Literatures:

1.1 reinforcement Leanring

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

1.2 sensative 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 sensative GAN is designed to:

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

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 funciton", 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 Looking at the 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 $J(x) \to \mathcal{L}(x, \lambda)$:

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

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:

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

3.2 if we were to minimize $\mathbf{x}$ on both sides

$$\begin{aligned} \min_{\mathbf{x}} \max_\lambda \mathcal{L}(\mathbf{x}, \lambda) &= \min_{\mathbf{x}} J(\mathbf{x}) \\ &= p^* \end{aligned} \tag{11}$$

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

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

from Eq(11)

$$\begin{aligned} &\min_{\mathbf{x}} \max_\lambda \mathcal{L}(\mathbf{x}, \lambda) = \min_{\mathbf{x}} J(\mathbf{x}) \\ \implies &\max_\lambda \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \lambda) \leq \min_{\mathbf{x}} \max_\lambda \mathcal{L}(\mathbf{x}, \lambda) = \min_{\mathbf{x}} J(\mathbf{x}) \\ \implies &\left(d^* \equiv \max_\lambda \min_x \mathcal{L}(\mathbf{x}, \lambda)\right) \leq \left(p^* \equiv \min_{\mathbf{x}} \max_\lambda \mathcal{L}(\mathbf{x}, \lambda) = \min_{\mathbf{x}} J(\mathbf{x})\right) \\ \implies &\left(d^* \equiv \max_\lambda f_\lambda^{(\star)}(\lambda)\right) \leq p^* \end{aligned} \tag{12}$$

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

4.1 max-min inequality

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

$$\sup_{\lambda} \inf_{x} f(\lambda, x) \leq \inf_{x} \sup_{\lambda} f(\lambda, x) \tag{13}$$

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

$$\begin{aligned}
& \inf_{x} f(\lambda, x) \leq f(\lambda, x), \forall \lambda \, \forall x \\
\Longrightarrow & \sup_{\lambda} \inf_{x} f(\lambda, x) \leq \sup_{\lambda} f(\lambda, x), \forall x \qquad \sup_{\lambda} \text{ both sides} \\
\Longrightarrow & \sup_{\lambda} \inf_{x} f(\lambda, x) \leq \inf_{x} \sup_{\lambda} f(\lambda, x) \qquad \text{on RHS: } \because \inf_{x} \in \forall x
\end{aligned} \tag{14}$$

4.2 if strong duality holds

$$d^* = p^* \tag{15}$$

5 advantage of dual function

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

$$\begin{aligned}
\lambda^* &= \arg\max_{\lambda} \big(\min_{x} \mathcal{L}(\mathbf{x}, \lambda) \big) \\
&= \arg\max_{\lambda} f_{\lambda}^{(\star)}(\lambda)
\end{aligned} \tag{16}$$

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:

$$\begin{aligned}
f_{\lambda}^{(\star)}(\lambda) \equiv \min_{x} \mathcal{L}(\mathbf{x}, \lambda) &\triangleq \min_{x} \Big(f(\mathbf{x}) + \sum_{i} \lambda_i g_i(\mathbf{x}) \Big) \\
&= f(\mathbf{x}') + \sum_{i} \underbrace{\lambda_i}_{x} \underbrace{g_i(\mathbf{x}')}_{m}
\end{aligned} \tag{17}$$

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"

5.1 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:

$$\begin{aligned} &f(\cdot, y) : X \to \mathbb{R} \text{ is convex for fixed } y \\ &f(x, \cdot) : Y \to \mathbb{R} \text{ is concave for fixed } x \end{aligned} \tag{18}$$

then:

$$\min_{x \in X} \max_{y \in Y} f(x, y) = \max_{y \in Y} \min_{x \in X} f(x, y) \tag{19}$$

this is the reason $d^* = p^*$

6 complementary slackness

by definition $\lambda_i \geq 0$, so let's see where $\lambda_i = 0$ and $\lambda_i > 0$ applies:

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

best λ_i occurs when:

$$\begin{aligned} \lambda_i^* &= \arg\max_{\lambda_i} \mathcal{L}(\mathbf{x}^*, \lambda_i) \\ &= \arg\max_{\lambda_i} f(\mathbf{x}^*) + \lambda_i g_i(\mathbf{x}) \\ &= \arg\max_{\lambda_i} \Big(\lambda_i \underbrace{g_i(\mathbf{x})}_{\leq 0}\Big) \quad \text{use the case} \quad g_i(\mathbf{x}^*) \leq 0 \\ &= 0 \end{aligned} \tag{20}$$

this is because **on the contrary** when $\lambda_i^* > 0$, then:

$$g_i(\mathbf{x}^*) \leq 0 \text{ and } \lambda_i^* > 0 \implies \lambda_i^* g_i(\mathbf{x}^*) \leq 0 \tag{21}$$

therefore, when **max** $= \lambda_i^* g_i(\mathbf{x}^*) = 0$ and **argmax** $\lambda_i^* = 0$

$$g_i(\mathbf{x}^*) \leq 0 \implies \lambda^* = 0 \tag{22}$$

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

since $g_i(\mathbf{x}^*) > 0$, one may "damagingly" **maximize** $\mathcal{L}(\mathbf{x}^*, \lambda) = f(\mathbf{x}^*) + \lambda_i g_i(\mathbf{x}^*)$ by taking $\lambda_i \to +\infty$.

We can see the way to prevent $\mathcal{L}(\mathbf{x}, \lambda)$ going to infinity is to locate new $\mathbf{x}'$ to be a "sub-optimal" solution at the contour where:

$$g_i(\mathbf{x}') = 0 \tag{23}$$

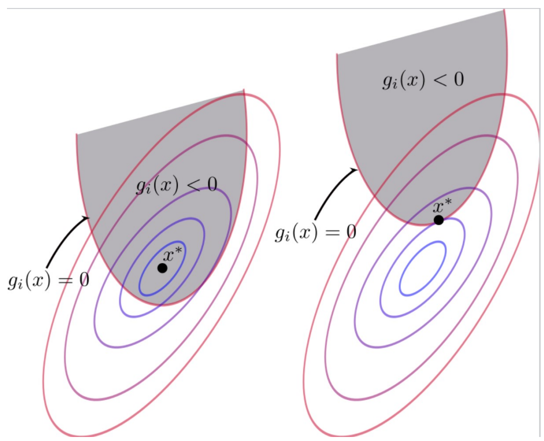
instead of original $\mathbf{x}^*$, i.e., optimal unconstrained solution $\nabla_{\mathbf{x}} f(\mathbf{x}^*) = 0$

6.3 combine the two

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

$$\lambda_i g_i(\mathbf{x}) = 0 \tag{24}$$

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



6.3.1 in summary

- primal:

$$\begin{aligned} &\min f(\mathbf{x}) \\ &\text{s.t. } g_i(\mathbf{x}) \leq 0 \ \forall i \in 1, \ldots, m \end{aligned} \tag{25}$$

- dual:

$$\begin{aligned} &\max f^{(*)}(\lambda) \\ &\text{s.t. } \lambda_i \geq 0 \ \forall i \in 1, \ldots, m \end{aligned} \tag{26}$$

- complementary slackness:

$$\lambda_i g_i(\mathbf{x}) = 0 \ \forall i \in 1, \ldots, m \tag{27}$$

6.3.2 name of slack variable

$$\text{when constraint} \begin{cases} g_i(\mathbf{x}^*) = 0 & \text{is } \textit{tight} \quad \implies \lambda_i > 0 \\ g_i(\mathbf{x}^*) \leq 0 & \text{is } \textit{slack} \quad \implies \lambda_i = 0 \end{cases} \tag{28}$$

slack variable doesn't need to be muplitication it can be addition too:

$$\begin{aligned} & g(\mathbf{x}) \leq 0 \\ \text{can be replaced by} \quad g(\mathbf{x}) + & \underbrace{\lambda}_{\text{slack varaible}} = 0 \qquad \lambda \geq 0 \end{aligned} \tag{29}$$

7 a quick note on Lagrange Cosntraint

$$\begin{aligned} &\text{maximize } f(\mathbf{x}) \\ &\text{subject to: } g(\mathbf{x}) = 0 \end{aligned} \tag{30}$$

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

$$\begin{cases} \nabla_{\mathbf{x}} f(\mathbf{x}) - \mu \nabla_{\mathbf{x}} g(\mathbf{x}) = 0 & \text{as contour line } f(\mathbf{x}) = k \text{ and } g(\mathbf{x}) \text{ share same tangent} \\ g(\mathbf{x}) = 0 & \text{original constraint} \end{cases} \tag{31}$$

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:

$$\begin{aligned} \mathcal{L}(\mathbf{x}, \mu) &= f(\mathbf{x}) - \mu g(\mathbf{x}) \\ \implies \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \mu) &= \underbrace{\nabla_{\mathbf{x}} f(\mathbf{x}) - \mu \nabla_{\mathbf{x}} g(\mathbf{x}) = 0} \\ \nabla_{\mu} \mathcal{L}(\mathbf{x}, \mu) &= \underbrace{g(\mathbf{x}) = 0} \end{aligned} \tag{32}$$

8 summary of KKT condition

optimization problem with both equality and inequality constraints:

$$\begin{aligned} \mathbf{x}^* = \underset{\mathbf{x}}{\operatorname{argmin}} f(\mathbf{x}) & \\ \text{subject to } h_i(\mathbf{x}) = 0 & \qquad \text{added for completeness} \\ \text{subject to } g_i(\mathbf{x}) \leq 0 & \end{aligned} \tag{33}$$

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

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

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

$$\begin{aligned} &\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}',\mu,\lambda)=0 \\ \Longrightarrow\ &\nabla_{\mathbf{x}}\Big(f(\mathbf{x}')+\sum_{i=1}^{m}\mu_i h_i(\mathbf{x}')+\sum_{i=1}^{n}\lambda_i g_i(\mathbf{x}')\Big)=0 \\ \Longrightarrow\ &\nabla_{\mathbf{x}}f(\mathbf{x}')+\sum_{i=1}^{m}\mu_i\nabla_{\mathbf{x}'}h_i(\mathbf{x}')+\sum_{i=1}^{n}\lambda_i\nabla_{\mathbf{x}}g_i(\mathbf{x}')=0 \end{aligned} \tag{34}$$

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

$$f_{\lambda}^{(\star)}(\lambda)=\min_{\mathbf{x}}\mathcal{L}(\mathbf{x},\mu,\lambda) \tag{35}$$

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:

$$\begin{aligned} &\nabla_{\mu}\mathcal{L}(\mathbf{x}',\mu,\lambda)=0 \\ \Longrightarrow\ &\nabla_{\mu}f(\mathbf{x}')+\sum_{i=1}^{m}\nabla_{\mu_\mathbf{i}}\mu_i h_i(\mathbf{x}')+\sum_{i=1}^{n}\lambda_i\nabla_{\mu}g_i(\mathbf{x}')=0 \\ \Longrightarrow\ &\sum_{i=1}^{m}\nabla_{\mu_\mathbf{i}}\mu_i h_i(\mathbf{x}')=0 \\ \Longrightarrow\ &\sum_{i=1}^{m}h_i(\mathbf{x}')=0 \qquad \text{just the original equality condition} \end{aligned} \tag{36}$$

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

$$\begin{aligned} \lambda_i g_i(\mathbf{x})=0, &\quad \forall i \\ \lambda_i\geq 0, &\quad \forall i \\ g_i(\mathbf{x})\leq 0, &\quad \forall i \end{aligned} \tag{37}$$

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

9 Example through Support Vector Machine

9.1 Linear Discriminant Function (geometry)

9.1.1 motivation

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

$$\mathbf{x}^{\top}\mathbf{w}+w_0=0 \tag{38}$$

9.1.2 geometry of $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + w_0$

if we think about the hyper-plane without the w_0, let's visualize it as 3D plane:

$$
\begin{aligned}
f(\mathbf{x}) &= \mathbf{w}^\top \mathbf{x} \\
\implies \begin{bmatrix} w_1 & w_2 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ f(x_1, x_2) \end{bmatrix} &= 0
\end{aligned} \tag{39}
$$

by adding w_0 "shift along the $f(\mathbf{x})$ axis" into the picture, it is a 3-D plane with normal $\begin{bmatrix} w_1 & w_2 & -1 \end{bmatrix}$ and shifted by w_0

9.1.3 the margin idea

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

$$
\begin{cases} \text{all data } \mathbf{x} \text{ having label } y = +1 \text{ is above the boundary} & \mathbf{w}^\top \mathbf{x} + w_0 = 1 \\ \text{all data } \mathbf{x} \text{ having label } y = -1 \text{ is below the boundary} & \mathbf{w}^\top \mathbf{x} + w_0 = -1 \end{cases} \tag{40}
$$

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

$$
y(\mathbf{x}) = \begin{cases} \mathbf{x}^\top \mathbf{w} + w_0 & \geq 1 \quad \forall \text{ +ve data } \mathbf{x} \\ \mathbf{x}^\top \mathbf{w} + w_0 & \leq 1 \quad \forall \text{ -ve data } \mathbf{x} \end{cases} \tag{41}
$$

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

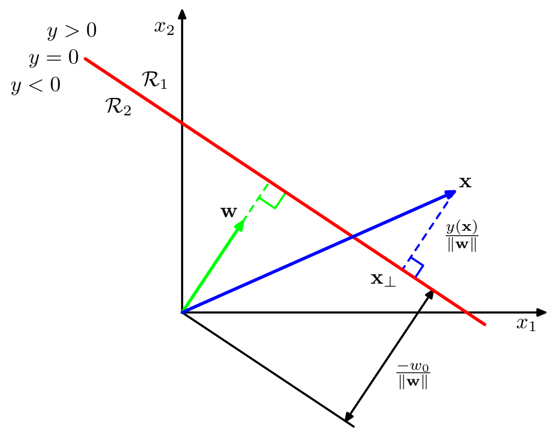
9.1.4 expression for margin

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:

$$
\begin{aligned}
\mathbf{x} &= \mathbf{x}_\perp + r \frac{\mathbf{w}}{\|\mathbf{w}\|} \qquad \text{sum of these two vectors} \\
\implies \underbrace{\mathbf{w}^\top \mathbf{x} + w_0}_{y(\mathbf{x})} &= \mathbf{w}^\top \left(\mathbf{x}_\perp + r \frac{\mathbf{w}}{\|\mathbf{w}\|} \right) + w_0 \qquad \text{apply } (\mathbf{w}^\top \times \quad + w_0) \text{ to both sides} \\
\implies y(\mathbf{x}) &= \underbrace{\mathbf{w}^\top \mathbf{x}_\perp + w_0}_{=0} + \mathbf{w}^\top r \frac{\mathbf{w}}{\|\mathbf{w}\|} \\
\implies y(\mathbf{x}) &= r \frac{\mathbf{w}^\top \mathbf{w}}{\|\mathbf{w}\|} = r \frac{\|\mathbf{w}\|^2}{\|\mathbf{w}\|} \\
\implies r &= \frac{y(\mathbf{x})}{\|\mathbf{w}\|}
\end{aligned} \tag{42}
$$

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



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

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

the two inequality constraints can be written as one:

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

$$\implies \text{subject to: } 1 - y_i(\mathbf{w}^T x_i + w_0) \leq 0$$

9.1.5 primal optimization

$$\min\left(\frac{1}{2}\|\mathbf{w}\|^2\right) \qquad \text{subject to: } \quad 1 - y_i(\mathbf{w}^T x_i + w_0) \leq 0 \quad \forall i \tag{43}$$

9.2 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:

$$\mathcal{L}(\underbrace{w, b,}_{\mathbf{x}} \underbrace{\lambda}_{\text{there is no } \mu}) = \underbrace{\frac{1}{2}\|\mathbf{w}\|^2}_{f(\mathbf{x})} + \underbrace{\sum_{i=1}^{p} \mu_i h_i(\mathbf{x})}_{=0} + \sum_{i=1}^{N} \lambda_i \underbrace{[1 - y_i(w^T x_i + w_0)]}_{g_i(\mathbf{x})} \tag{44}$$

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

$$
\begin{aligned}
\frac{\partial \mathcal{L}(w,b,\lambda)}{\partial w} &= w - \sum_{i=1}^{N} \lambda_i y_i x_i = 0 \implies w' = \sum_{i=1}^{N} \lambda_i y_i x_i \\
\frac{\partial \mathcal{L}(w,b,\lambda)}{\partial b} &= \underbrace{\sum_{i=1}^{N} \lambda_i y_i}_{\text{not a function of } b} = 0
\end{aligned} \tag{45}
$$

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

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

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

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

$$
\begin{aligned}
\implies f_{\lambda}^{(\star)}(\lambda) &= \inf_x \mathcal{L}(w,b,\lambda) \\
&= \frac{1}{2}\Big(\sum_{i=1}^{n} \lambda_i y_i x_i\Big)^\top \Big(\sum_{i=1}^{n} \lambda_i y_i x_i\Big) + \sum_{i=1}^{n} \lambda_i \Big[1 - y_i\Big(\Big(\sum_{i=1}^{n} \lambda_i y_i x_i\Big)^\top x_i + w_0\Big)\Big] \\
&= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j x_i^\top x_j - \sum_{i=1}^{n} \lambda_i y_i \Big(\sum_{j=1}^{n} \lambda_j y_j x_j^\top\Big) x_i - w_0 \underbrace{\sum_{i=1}^{n} \lambda_i y_i}_{=0} + \sum_{i=1}^{n} \lambda_i \\
&= \sum_{i=1}^{n} \lambda_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j x_i^\top x_j \\
&\text{subject to: } \sum_{i=1}^{N} \lambda_i y_i = 0 \text{ and } \lambda_i \geq 0
\end{aligned} \tag{46}
$$

9.4 The dual problem

$$
\begin{aligned}
\underset{\lambda_1,\ldots\lambda_n}{\arg\max}\, \mathcal{L}_\lambda(\lambda) = \underset{\lambda_1,\ldots\lambda_n}{\arg\max} \left(\sum_{i=1}^{n} \lambda_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j x_i^\top x_j\right) \\
\text{subject to: } \sum_{i=1}^{n} \lambda_i y_i = 0 \text{ and } \lambda_i \geq 0
\end{aligned} \tag{47}
$$

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

Use **complementary slackness:**

$$
\begin{aligned}
\lambda_i^* > 0 &\implies g_i(w^*, b^*) = 0 \\
&\implies 1 - y_i({w^*}^\top x_i + {w_0}^*) = 0 \\
&\implies y_i({w^*}^\top x_i + {w_0}^*) = 1 \\
&\qquad\qquad\qquad\qquad \text{i.e., } x_i \text{ is support vector points} \\
\lambda_i^* = 0 &\implies g_i(w^*, b^*) < 0 \\
&\implies 1 - y_i({w^*}^\top x_i + {w_0}^*) < 0 \\
&\implies y_i({w^*}^\top x_i + {w_0}^*) > 1 \\
&\qquad\qquad\qquad\qquad \text{i.e., } x_i \text{ is non support vector points}
\end{aligned} \tag{48}
$$

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