Lecture 13. Support Vector Machine I

Lecturer: Jie Wang Date: Nov 24, 2021

The major references of this lecture are [3, 2].

1 Introduction

Suppose that we are given a data set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_i^n$, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathcal{C} = \{-1, 1\}$, $i = 1, 2, \ldots, n$. Support vector machine (SVM) tries to find a linear function $f : \mathbb{R}^d \to \mathbb{R}$ in the form of

$$f(X; \mathbf{w}, b) = b + \sum_{j=1}^{d} w_j X_j,$$

such that

$$y_i = \mathbf{sign}\,(f(\mathbf{x}_i; \mathbf{w}, b)).$$

To fit the data, we need to put all the positive training instances in the positive half space and the negative training instances in the negative half space.

2 SVM for Linearly Separable Cases

2.1 Maximum Margin

To illustrate the idea of SVM, we consider a simple case where the training samples are ***linearly separable***, that is, we can find a ***hyperplane***—which separates the feature space into two ***half-spaces***: the positive halfspace and the negative halfspace—such that positive and negative data instances fall into the positive and negative halfspaces, respectively.

Definition 1. Let $\mathbf{w} \in \mathbb{R}^d$, $\mathbf{w} \neq 0$, and $b \in \mathbb{R}$. A linear classifier that takes the form of

$$f(\mathbf{x}; \mathbf{w}, b) = \langle \mathbf{w}, \mathbf{x} \rangle + b, \tag{1}$$

defines a hyperplane (its 0-level set)

$$H_f = \{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}; \mathbf{w}, b) = 0\},$$

separating the feature space into two halfspaces: the positive halfspace

$$H_f^+ = \{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}; \mathbf{w}, b) > 0\},$$

and the negative halfspace

$$H_f^- = \{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}; \mathbf{w}, b) < 0\}, .$$

Thus, linearly separable indeed assumes the existence of a hyperplane H_f (specified by a linear classifier f) such that all positive (negative) labeled data instances belong to the positive (negative) half space H_f^+ (H_f^-). In other words, the labels of the data instances share the same sign with the halfspaces they fall into. This leads to a concise definition of linearly separable.

Definition 2. A training sample is linearly separable if there exists $(\hat{\mathbf{w}}, \hat{b})$ such that

$$y_i = \mathbf{sign}\left(f(\mathbf{x}_i; \hat{\mathbf{w}}, \hat{b})\right), \forall\, i \in [n], \tag{2}$$

which is equivalent to

$$y_i f(\mathbf{x}_i; \hat{\mathbf{w}}, \hat{b}) > 0, \forall\, i \in [n], \tag{3}$$

where $[n] = \{1, \ldots, n\}$.

In this section, we assume that the training sample is linearly separable.

Assumption 1. *The training sample $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_i^n$ is linearly separable.*

However, we can find infinitely many hyperplanes such that the inequality in (3) holds. Which one shall we choose? The SVM classifier makes the decision based on the notion of ***geometric margin***.

Definition 3. Suppose that we have a data sample $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_i^n$. The ***geometric margin*** $\gamma_f(\mathbf{x}_i)$ of a linear classifier

$$f(\mathbf{x}; \mathbf{w}, b) = \langle \mathbf{w}, \mathbf{x} \rangle + b$$

at a point $\mathbf{x}_i$ is its ***signed Euclidean distance*** to the hyperplane $\{\mathbf{x} : \langle \mathbf{w}, \mathbf{x} \rangle + b = 0\}$:

$$\gamma_f(\mathbf{x}_i) = \frac{y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)}{\|\mathbf{w}\|}.$$

The ***geometric margin*** γ_f of a linear classifier f for a sample $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_i^n$ is the minimum geometric margin over the points in the sample, that is

$$\gamma_f = \min_{i \in [n]} \gamma_f(\mathbf{x}_i).$$

Remark 1. The geometric margin of a data instance to a hyperplane can be *negative*, which implies that it falls into the wrong side of the hyperplane. Given a training sample, a ***negative*** geometric margin implies that some of the data instances are ***misclassified***.

SVM looks for the hyperplane which maximizes the geometric margin, and thus it is known as the ***maximum margin classifier***. Specifically, we can model SVM by the following optimization problem:

$$\max_{\mathbf{w}, b} \gamma_f = \max_{\mathbf{w}, b} \min_{i \in [n]} \frac{y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)}{\|\mathbf{w}\|} = \max_{\mathbf{w}, b} \frac{1}{\|\mathbf{w}\|} \left(\min_{i \in [n]} y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \right). \tag{4}$$

Remark 2. The problem in (4) is challenging to solve. One obvious reason is that the variables are mixtures of continuous variables ($\mathbf{w}$,b) and discrete variables (the index i), leaving many optimization methods we are familiar with out of our options. However, ***a surprising fact*** is that, the problem in (4) is equivalent to a convex optimization problem, which can be readily solved by many popular methods.

2.2 The Convex Version

We show that we can transform the problem in (4) to a convex optimization problem.

Step 1: Reducing the search space

Recall that the problem in (4) has two sets of variables: the continuous variables $(\mathbf{w}, b)$ and the discrete variable $i \in [n]$. We can see that the domain of the problem in (4) is

$$D = D_1 \times D_2,$$

where

$$D_1 = \{(\mathbf{w}, b) : \mathbf{w} \in \mathbb{R}^d, \mathbf{w} \neq 0, b \in \mathbb{R}\} \text{ and } D_2 = \{i : i = 1, 2, \ldots, n\}.$$

Notice that, the value of the objective function in problem (4), i.e., the geometric margin γ_f, is unchanged if we multiply $(\mathbf{w}, b)$ by a *positive* scalar (why positive?), that is

$$\gamma_f = \gamma_{\lambda f}, \forall\, \gamma > 0.$$

In other words, for any $(\mathbf{w}, b) \in \mathbb{R}^{d+1}$ with $\mathbf{w} \neq 0$, all points of the ray $\{\lambda(\mathbf{w}, b) : \lambda > 0\}$ share the same value of the geometric margin. Thus, for any ray in $\mathbb{R}^{d+1}$ (except the two rays going upside and downside), we can consider only one single point of it. But which one shall we pick? Here comes the first trick in deriving SVM: we pick $(\mathbf{w}, b)$ that satisfies the constraint as follows.

$$\min_i\ y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) = 1. \tag{5}$$

This transforms the problem in (4) to

$$\begin{aligned} &\max_{\mathbf{w}, b} \frac{1}{\|\mathbf{w}\|}, \\ &\text{s.t. } \min_i\ y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) = 1. \end{aligned} \tag{6}$$

Remark 3. Notice that, what we are looking for is indeed a separating hyperplane defined by a linear classifier. However, different linear classifiers may specify the same separting hyperplanes. For example, it is easy to see that $H_f = H_{\lambda f}$ for any $\lambda > 0$. Thus, for a set of linear classifiers that define the same separating hyperplanes, we can only consider one of them. This is the geometric intuition behind the transformation from (4) to (6).

Step 2: Transforming the objective function to a convex function

In view of the problem in (6), we can see that maximizing $1/\|\mathbf{w}\|$ is equivalent to minimizing $\|\mathbf{w}\|$. Thus, we can transform (6) as follows.

$$\begin{aligned} &\min_{\mathbf{w}, b} \frac{1}{2}\|\mathbf{w}\|^2, \\ &\text{s.t. } \min_i\ y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) = 1. \end{aligned} \tag{7}$$

Remark 4. We note that, though the problems in (6) and (7) are similar to each other, the former is NOT equivalent to the latter. The key difference is that, the problem in (6) does not allow $\mathbf{w} = 0$, while the problem (7) does.

Question 1. Under which cases, the problem in (7) admits optimal solutions in the form of $(0, b)$?

Step 3: Relaxing the constraints

The constraint in problem (7) is in the form of a minimization problem, which is difficult to deal with. However, we can relax the constraint (5) by requiring that

$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1, \, \forall \, i \in [n],$$

Then, the problem in (7) changes to

$$\begin{aligned} &\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2, \\ &\text{s.t. } y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1. \end{aligned} \tag{8}$$

The problem in (8) is the commonly-seem formulation of SVM for the linearly separable data samples. Though we arrive at (8) by relaxing the constraint in (7), we can show that the problems (7) and (8) are equivalent to each other, that is, one of the constraints in (8) must hold as an equality at its optimal solution.

Question 2.

1. Show there is at least one of the constraints holds as an equality at the optimum.
2. Show there exist at least one positive **and** negative samples such that the equality holds at the optimum.
3. Can we remove the inequalities that hold strictly at the optimum without affecting the solution?

Definition 4. Given a linear classifier in the form of (1), the ***marginal hyperplanes*** are

$$H_f(1) = \{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) = 1\} \text{ and } H_f(-1) = \{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) = -1\}.$$

The ***support vectors*** are the data instances on the marginal hyperplanes, i.e.,

$$\{\mathbf{x} : |\langle \mathbf{w}, \mathbf{x} \rangle + b| = 1, \, \mathbf{x} \in \mathcal{D}\}.$$

3 SVM for Non-separable Cases

In most real applications, the training data instances are not linearly separable, that is, for any hyperplane H_f, there exists $\mathbf{x}_i \in \mathcal{D}$ such that

$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) < 0.$$

Thus, the constraints in (8) can not hold simultaneously. To address this problem, we introduce a set of nonnegative ***slack variables*** $\{\xi_i\}_{i=1}^n$ to relax the constraints as

$$y_i(\langle \mathbf{w}, \mathbf{x} \rangle + b) \geq 1 - \xi_i, \, i \in [n].$$

We can see that the value of ξ_i measures the vector $\mathbf{x}_i$'s violation of the corresponding inequality $y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1$. To limit the violations over all data instances, we add a penalty to the objective function in (8), which leads to

$$\begin{aligned} &\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i, \\ &\text{s.t. } y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \xi_i, \, \xi_i \geq 0, \, i \in [n]. \end{aligned} \tag{9}$$

The problem in (9) is a widely-used version of SVM for non-separable cases.

Question 3. For a linearly separable data sample, shall we arrive at the same separating hyperplane by solving the problems in (8) and (9), respectively?

Duality plays an important role in analyzing SVM. Besides interesting theoretical results, duality also motives many efficient algorithms for solving SVM. In this section, we introduce ***elements of Lagrangian duality***. There are several different approaches to Lagrangian duality. We follow the approach introduced in [1, 2], which are based on geometric observations.

4 The Primal Problem

We consider the problem—that is, the ***primal problem***—as follows.

$$\begin{aligned} \min_{\mathbf{x}} \ & f(\mathbf{x}) \\ \text{s.t. } & g_i(\mathbf{x}) \leq 0,\ i = 1, \ldots, m, \\ & h_i(\mathbf{x}) = 0,\ i = 1, \ldots, p, \\ & \mathbf{x} \in X, \end{aligned} \tag{10}$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $g_i : \mathbb{R}^n \to \mathbb{R}$, $i \in [m]$, and $h_i : \mathbb{R}^n \to \mathbb{R}$, $i \in [p]$, are all continuously differentiable, and $X \subseteq \mathbb{R}^n$. To simplify notations, let $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^m$ be a vector function whose i^{th} component is g_i, and $\mathbf{h} : \mathbb{R}^n \to \mathbb{R}^p$ be a vector function whose i^{th} component is h_i. Then, we can write the problem in (10) in a more compact form as follows.

$$\begin{aligned} \min_{\mathbf{x}} \ & f(\mathbf{x}) \\ \text{s.t. } & \mathbf{g}(\mathbf{x}) \leq 0, \\ & \mathbf{h}(\mathbf{x}) = 0, \\ & \mathbf{x} \in X. \end{aligned} \tag{11}$$

We denote the ***feasible set*** of (11) by

$$D_0 = \{\mathbf{x} : \mathbf{g}(\mathbf{x}) \leq 0, \mathbf{h}(\mathbf{x}) = 0, \mathbf{x} \in X\}. \tag{12}$$

Each element in D_0 is called a ***feasible solution***. The ***optimal function value*** is

$$f^* = \inf_{\mathbf{x} \in D_0} f(\mathbf{x}). \tag{13}$$

Assumption 2. Feasibility and Boundedness *The feasible set is nonempty and the objective function is bounded from below, that is,*

$$-\infty < f^* = \inf_{\mathbf{x} \in D_0} f(\mathbf{x}) < \infty.$$

Remark 5. Notice that, Assumption 2 does not assume the existence of the optimum of the problem in (11).

5 Geometric Observations

We used to analyze and/or solve optimization problems by focusing on the problem domain D_0, as the variable $\mathbf{x}$ lies in D_0, and so does the optimum we are looking for (if it exists). Surprisingly, taking the perspective of the ***constraint-cost pairs*** as $\mathbf{x}$ goes over X, that is, the subset of $\mathbb{R}^{m+p+1}$

$$S = \{(\mathbf{g}(\mathbf{x}), \mathbf{h}(\mathbf{x}), f(\mathbf{x})) : \mathbf{x} \in X\}, \tag{14}$$

brings us fresh insights. Figure 1 shows a simple example of S for problems with only one inequality constraint. Indeed, the key idea to Lagrangian duality in [1, 2] is to *interpret the primal problem (11) by the geometric properties of the set S via hyperplanes*.

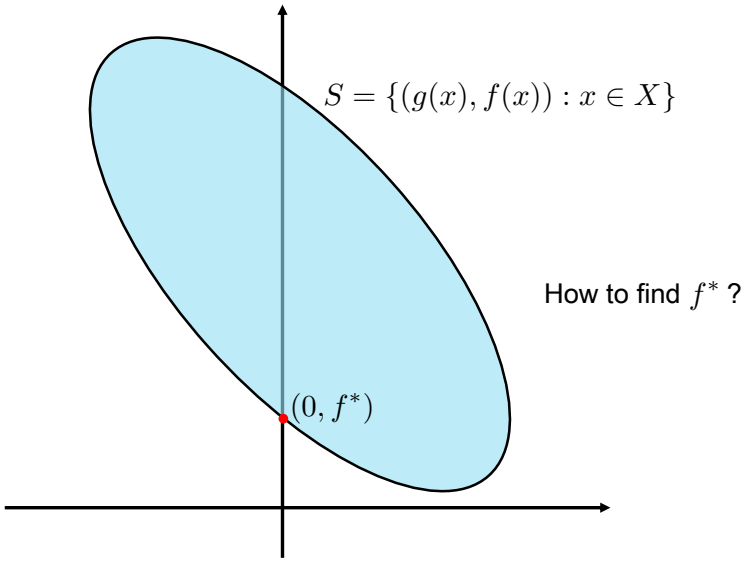

Figure 1: Illustration of the set S of the constraint-cost pairs for a simple problem with only one inequality constraint.

5.1 The Lagrangian

To illustrate the idea of Lagrangian duality from a geometric perspective, we first consider a simple problem with only one inequality constraint. We show the set S of constraint-cost pairs in Figure 1. We can see that, the optimal function value f^* of the primal problem is indeed the second component of the red dot, that is, the point with the smallest value of the second component among the points whose first components are non-positive.

Thus, a natural question arises, *instead of solving the primal problem (11), can we find the optimal function value f^* by analyzing the set S?* The answer is yes. The working horse is the (simple) hyperplanes. The linear function that specifies the hyperplanes is the so-called ***Lagrangian***.

Definition 5. Associated with the primal problem, we define the Lagrangian $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ as

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

5.2 Hyperplanes Defined by the Lagrangian

Hyperplanes can be specified by level sets of linear functions. Given a constant c, the Lagrangian defines a hyperplane in $\mathbb{R}^{m+p+1}$—where the set S of constraint-cost pairs lies in—by

$$H_L(c) = \{(\mathbf{y}, \mathbf{w}, z) : z + \langle \lambda, \mathbf{y} \rangle + \langle \mu, \mathbf{w} \rangle = c, z \in \mathbb{R}, \mathbf{y} \in \mathbb{R}^m, \mathbf{w} \in \mathbb{R}^p\}.$$

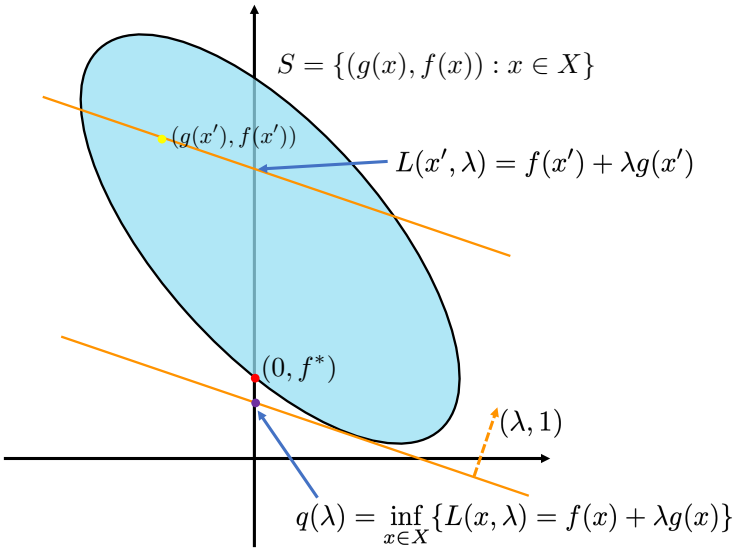

Figure 2: Illustration of the hyperplanes specified by the Lagrangian.

The normal of $H_L(c)$ is $(\lambda, \mu, 1)$, which implies that the hyperplane is ***nonvertical*** (why?).

Figure 2 shows the hyperplanes defined by the Lagrangian for a simple problem with only inequality constraints. The two hyperplanes share the same normal vector $(\lambda, 1)$. As the function value of the Lagrangian at the yellow point $(g(x'), f(x'))$ is clearly given by

$$L(x', \lambda) = f(x') + \lambda g(x'),$$

the hyperplane which go through the point $(g(x'), f(x'))$ is

$$H_L(L(x', \lambda)) = \{(y, z) : z + \lambda y = L(x', \mu) = f(x') + \lambda g(x')\}.$$

Moreover, we can see that, *the hyperplane $H_L(L(x', \lambda))$ intercepts the vertical axis $\{(0, z) : z \in \mathbb{R}\}$ at the level $L(x', \lambda)$*.

5.3 The Lagrangian Dual Function

The geometric properties we observe in Section 5.2 lead us to the fact that, for a nonvertical hyperplane, the level of interception of the vertical axis is indeed the (linear) function value that defines the hyperplane. Thus, given a vector $(\lambda, \mu, 1) \in \mathbb{R}^{m+p+1}$, if we define

$$q(\lambda, \mu) = \inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda, \mu), \tag{15}$$

the hyperplane $H_L(q(\lambda, \mu))$ intercepts the vertical axis at the level $q(\lambda, \mu)$ if it exists. Figure 2 shows a simple example where a hyperplane intercepts the vertical axis at the level $q(\lambda)$.

In general, What is the relationship between $q(\lambda, \mu)$ and f^*? In view of Figure 2, a reasonable guess would be

$$q(\lambda, \mu) \leq f^*, \forall \lambda \geq 0,$$

where $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_m)$ and $\lambda \geq 0$ is an abbreviation for $\lambda_i \geq 0$, $i \in [m]$. Indeed, the above guess is true, and we have the result as follows.

Lemma 1. *For any $\lambda \geq 0$, the following result holds:*

$$q(\lambda, \mu) \leq f^*.$$

Proof. By definition, we have

$$\begin{aligned} q(\lambda, \mu) &= \inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda, \mu) \\ &= \inf_{\mathbf{x} \in X} f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) + \sum_{i=1}^{p} \mu_i h_i(\mathbf{x}) \\ &\leq \inf_{\mathbf{x} \in D_0} f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) + \sum_{i=1}^{p} \mu_i h_i(\mathbf{x}). \end{aligned}$$

The definition of D_0 implies that, for any $\mathbf{x} \in D_0$, we have

$$g_i(\mathbf{x}) \leq 0, i \in [m], \text{ and } h_i(\mathbf{x}) = 0, i \in [p].$$

Thus, the above inequality becomes

$$q(\lambda, \mu) \leq \inf_{\mathbf{x} \in D_0} f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) + \sum_{i=1}^{p} \mu_i h_i(\mathbf{x}) \leq \inf_{\mathbf{x} \in D_0} f(\mathbf{x}) = f^*,$$

which completes the proof. □

The function $q(\lambda, \mu)$ is the so-called ***Lagrangian dual function***. The domain of q is the set for which $q(\lambda, \mu)$ is finite:

$$\mathbf{dom}\ q = \{(\lambda, \mu) : q(\lambda, \mu) > -\infty\}.$$

We can similarly define the ***dual feasible set*** by

$$D_1 = \{(\lambda, \mu) : \lambda \geq 0\} \cap \mathbf{dom}\ q = \{(\lambda, \mu) : \lambda \geq 0, q(\lambda, \mu) > -\infty\}.$$

Remark 6. We do not require that $\lambda \geq 0$ for the points in $\mathbf{dom}\,(q)$.

A surprising result is that, the Lagrangian dual function q is concave, no matter the primal problem is convex or not.

Theorem 1. *The domain of q is convex and q is concave over* $\mathbf{dom}\,(q)$.

Proof. We first show that $\mathbf{dom}\,(q)$ is convex.

Suppose that $q(\lambda_1, \mu_1)$ and $q(\lambda_2, \mu_2)$ are finite and $(\lambda_1, \mu_1) \neq (\lambda_2, \mu_2)$. Let $\theta \in [0, 1]$.

$$\begin{aligned} q(\theta\lambda_1 + (1-\theta)\lambda_2, \theta\mu_1 + (1-\theta)\mu_2) &= \inf_{\mathbf{x} \in X} L(\mathbf{x}, \theta\lambda_1 + (1-\theta)\lambda_2, \theta\mu_1 + (1-\theta)\mu_2) \\ &= \inf_{\mathbf{x} \in X} \theta L(\mathbf{x}, \lambda_1, \mu_1) + (1-\theta) L(\mathbf{x}, \lambda_2, \mu_2) \\ &\geq \inf_{\mathbf{x} \in X} \theta L(\mathbf{x}, \lambda_1, \mu_1) + \inf_{\mathbf{x} \in X} (1-\theta) L(\mathbf{x}, \lambda_2, \mu_2) \\ &= \theta q(\lambda_1, \mu_1) + (1-\theta) q(\lambda_2, \mu_2) \\ &> -\infty. \end{aligned}$$

Thus, we have $\mathbf{dom}\,(q)$ is convex.

The concavity of q can easily be seen by noting that q is the infimum of a set of linear functions of (λ, μ). □

References

[1] M. Bazaraa, H. Sherali, and C. Shetty. *Nonlinear Programming.* Wiley-Interscience, 2006.

[2] D. P. Bertsekas. *Nonlinear Programming, 3ed.* Athena Scientific, 2016.

[3] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning, 2ed.* The MIT Press, 2018.