

LECTURE 11: CONVEX OPTIMIZATION

In this lecture we first give some background about convex optimization including the KKT condition and duality. We then derive the SVM dual optimization problem.

Consider the constrained minimization problem of the form

$$\begin{aligned} & \text{minimize} && f(\mathbf{x}) \\ & \text{subject to:} && g_i(\mathbf{x}) = 0 \quad i = 1, \dots, m \leq n \\ & && h_j(\mathbf{x}) \leq 0 \quad j = 1, \dots, p. \end{aligned} \tag{1}$$

g_i 's are equality constraints and h_j 's are inequality constraints and usually they are assumed to be within the class C^2 . A point that satisfies all constraints is said to be a *feasible point*. An inequality constraint is said to be *active* at a feasible point \mathbf{x} if $h_j(\mathbf{x}) = 0$ and *inactive* if $h_j(\mathbf{x}) < 0$. Equality constraints are always active at any feasible point. To simplify notation we write $\mathbf{h} = [h_1, \dots, h_p]$ and $\mathbf{g} = [g_1, \dots, g_m]$, and the constraints now become $\mathbf{g}(\mathbf{x}) = \mathbf{0}$ and $\mathbf{h}(\mathbf{x}) \leq \mathbf{0}$.

Karush-Kuhn-Tucker (KKT) Conditions

KKT conditions (a.k.a. Kuhn-Tucker conditions) are necessary conditions for the local minimum solutions of problem (1). Let \mathbf{x}^* be a local minimum point for Problem (1) and suppose \mathbf{x}^* is a regular point for the constraints. Then there is a vector $\mu \in \mathbb{R}^m$ and a vector $\lambda \in \mathbb{R}^p$ with $\lambda \geq \mathbf{0}$ such that

$$\nabla f(\mathbf{x}^*) + \lambda^T \nabla \mathbf{h}(\mathbf{x}^*) + \mu^T \nabla \mathbf{g}(\mathbf{x}^*) = \mathbf{0} \tag{2}$$

$$\mathbf{g}(\mathbf{x}^*) = \mathbf{0} \tag{3}$$

$$\lambda_j h_j(\mathbf{x}^*) = 0 \quad (j = 1, \dots, p) \tag{4}$$

Convince yourself why the above conditions hold geometrically. It is convenient to introduce the Lagrangian associated with the problem as

$$L(\mathbf{x}, \lambda, \mu) = f(\mathbf{x}) + \lambda^T \mathbf{h}(\mathbf{x}) + \mu^T \mathbf{g}(\mathbf{x})$$

where $\mu \in \mathbb{R}^m$, $\lambda \in \mathbb{R}^p$ and $\lambda \geq \mathbf{0}$ are *Lagrange multipliers*. Note that equation (2), (3) and (4) together give a total of $n + m + p$ equations in the $n + m + p$ variables \mathbf{x}^* , λ and μ .

From now on we assume that we only have inequality constraints for simplicity. The case with equality constraints can be done in a similar way, except that μ does not have the nonnegative constraint as λ . So in our case we have the following optimization problem:

$$\min f(\mathbf{x}) \quad \text{s.t.} \quad \mathbf{h}(\mathbf{x}) \leq \mathbf{0}.$$

Weak Duality and Strong Duality

Consider the Lagrangian $L(\mathbf{x}, \lambda)$ for the above optimization problem. Then we have the following two types of dualities:

- *weak duality*: we have obviously for any $\lambda \geq \mathbf{0}$ that

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

and thus

$$\sup_{\lambda \geq 0} \inf_x L(x, \lambda) \leq \inf_x \sup_{\lambda \geq 0} L(x, \lambda).$$

Such a relation always holds.

- *strong duality*: suppose in addition there exist x^* and $\lambda^* \geq 0$ such that

$$L(x^*, \lambda) \leq L(x^*, \lambda^*) \leq L(x, \lambda^*)$$

for all feasible x and $\lambda \geq 0$. Then we have

$$\begin{aligned} \inf_x \sup_{\lambda \geq 0} L(x, \lambda) &\leq \sup_{\lambda \geq 0} L(x^*, \lambda) \\ &= L(x^*, \lambda^*) \\ &= \inf_x L(x, \lambda^*) \\ &\leq \sup_{\lambda \geq 0} \inf_x L(x, \lambda). \end{aligned}$$

Thus we have

$$\inf_x \sup_{\lambda \geq 0} L(x, \lambda) = \sup_{\lambda \geq 0} \inf_x L(x, \lambda).$$

The point (x^*, λ^*) is called the saddle point. One example is the function $L(x, \lambda) = x^2 - \lambda^2$, with saddle point $(0, 0)$ as shown in Figure 1.

Weak duality always holds, and strong duality holds if f and h_j 's are convex and there exists at least one feasible point which is an interior point. The Lagrange dual function $D(\lambda)$ is defined as

$$D(\lambda) := \inf_x L(x, \lambda) = \inf_x \left\{ f(x) + \sum_{j=1}^p \lambda_j h_j(x) \right\}$$

and we define the dual optimization problem as:

$$\max D(\lambda) \text{ s.t. } \lambda \geq 0.$$

Note that (1) $D(\lambda)$ is a concave function; (2) for any feasible λ and x we have $D(\lambda) \leq f(x)$. In fact if we define p^* to be the minimum solution of the primal optimization problem (*primal solution*), and d^* to the maximum of the dual problem $d^* = \sup_{\lambda \geq 0} D(\lambda)$ (*dual solution*). Then the weak duality says $d^* \leq p^*$. The quantity $p^* - d^*$ is known as the *duality gap*, which can be a useful criteria for convergence.

Now we illustrate this duality relationship with a simple example where we only have one inequality constraint:

$$\min f(x) \text{ s.t. } h(x) \leq 0.$$

Define $\omega(z) = \inf\{f(x) : h(x) \leq z\}$ for $z \in \mathbb{R}$. Then it is easy to observe that $\omega(z)$ is monotone on each coordinate of z . The duality can be illustrated by the fact that the primal solution p^* is the intercept of $\omega(z)$ with the vertical axis $z = 0$, and it is an upperbound of the maximum intercept with the vertical axis of all hyperplanes that lie below $\omega(\cdot)$. Such hyperplanes have the form $l_\lambda(z) = -\lambda^T z + \inf_x \{f(x) + \lambda h(x)\}$ with $\lambda \geq 0$. An example is shown in Figure 2.

SVM Dual Problem

The SVM primal problem can be written as

$$\begin{aligned} \min_{w, b, \xi} & \frac{1}{n} \sum_{i=1}^n \xi_i + \lambda \|w\|^2 \\ \text{s.t.} & y_i(\mathbf{w}^T x_i + b) \geq 1 - \xi_i; \quad \xi_i \geq 0 \quad \forall i \end{aligned}$$

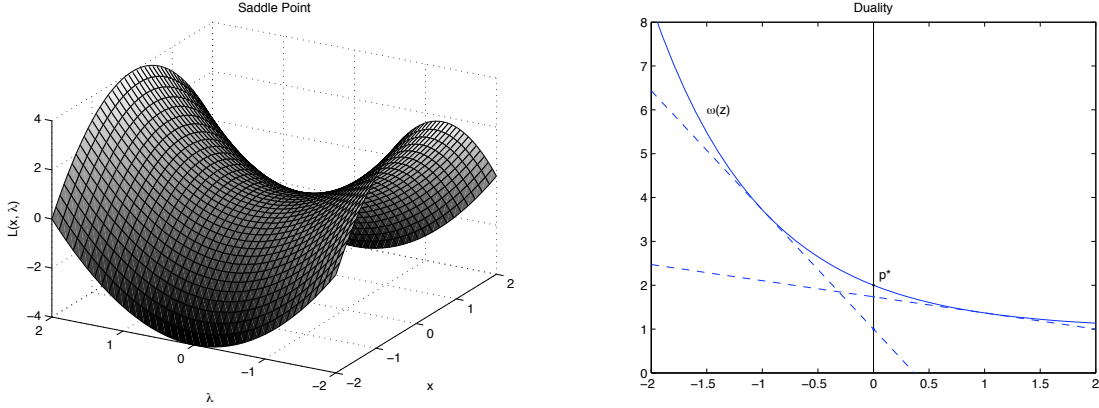


Figure 1: Left: Saddle point $(0, 0)$ of $L(x, \lambda) = x^2 - \lambda^2$; Right: Geometric interpretation of duality.

Now the Lagrangian can be written as

$$L(w, b, \xi, \alpha, \beta) = \frac{1}{n} \sum_{i=1}^n \xi_i + \lambda w^T w + \sum_{i=1}^n \alpha_i (1 - \xi_i - y_i w^T x_i - y_i b) - \sum_{i=1}^n \beta_i \xi_i$$

where the Lagrange multipliers $\alpha \geq 0$ and $\beta \geq 0$. We want to remove the primal variables w, b, ξ by maximization, i.e. set the following derivatives to zero:

$$\begin{aligned} \frac{\partial L}{\partial w} = 0 &\implies w = \frac{1}{2\lambda} \sum_{i=1}^n \alpha_i y_i x_i \\ \frac{\partial L}{\partial b} = 0 &\implies \sum_{i=1}^n \alpha_i y_i = 0 \\ \frac{\partial L}{\partial \xi} = 0 &\implies \alpha_i + \beta_i = \frac{1}{n}. \end{aligned}$$

Plugging in and we obtain the dual:

$$D(\alpha, \beta) = \sum_{i=1}^n \alpha_i - \frac{1}{4\lambda} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j.$$

Since we have $\alpha_i \geq 0$ and $\beta_i \geq 0$ and $\alpha_i + \beta_i = 1/n$, thus we have $0 \leq \alpha_i \leq 1/n$. So the dual optimization problem becomes

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^n \alpha_i - \frac{1}{4\lambda} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j \\ \text{s.t.} \quad & 0 \leq \alpha_i \leq 1/n. \end{aligned}$$

which is a quadratic programming problem. Note that due to the constraints, the dual solution is in general sparse, i.e. we have many α_i 's equal to 0. We have the following observations:

1. If $\alpha_i > 0$: we have $y_i(w^T x_i + b) = 1 - \xi_i \leq 1$. So the example is either at or on the wrong side of the margin. Such examples for $\alpha_i > 0$ are called support vectors.
2. If $\alpha_i = 0$: we have $\beta_i = 1/n$ and thus $\xi_i = 0$. So $y_i(w^T x_i + b) \geq 1$. Such examples are on the correct side of the margin.

3. If $y_i(w^T x_i + b) < 1$: we have $\xi > 0$ and thus $\beta_i = 0$ and $\alpha_i = 1/n$. So if an example causes margin error then its dual variable α_i will take at the right boundary $1/n$.
4. It is possible that for examples which are on the correct side of the margin, their α_i 's are nonzero.
5. In the objective x_i 's appear always in the form of inner product $x_i^T x_j$. So if we first map x_i into a feature vector $\phi(x_i)$, then we could replace $x_i^T x_j$ by $\langle \phi(x_i), \phi(x_j) \rangle$. This leads to the introduction of reproducing kernel Hilbert space in SVM.