Convex optimization

- General form of a convex optimization problem:

$$\min_x f(x) \text{ subject to:}$$
$$g_i(x) \leq 0, i = 1, \cdots, m$$
$$h_i(x) = 0, i = 1, \cdots, p.$$
$$x \in \mathcal{R}^n.$$

  where f is convex, $g_i(x)$ are convex, $h_i(x)$ are affine function (linear).

- Given such a formulation, the Lagrangian is a function $\mathcal{L} : \mathcal{R}^n \times \mathcal{R}^m_+ \times \mathcal{R}^p \to \mathcal{R}$ defined by:

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

Lagrangian duality

- x: primal variables, (λ, μ) are dual variables associated with constraints (also called Lagrangian multipliers).
- Notice that λ (the variables associated with the inequality constraints) has to be non-negative.
- Define the primal function as :

$$\mathcal{L}_P(x) = max_{\lambda,\mu}\mathcal{L}(x, \lambda, \mu)$$

- Define the primal problem as:

$$\min_x \mathcal{L}_P(x)$$

- A point x satisfying the constraints: $g_i(x) \leq 0$ and $h_i(x) = 0$ are called primal feasible.

Lagrangian duality

- The dual function is defined as:

$$\mathcal{L}_D(\lambda, \mu) = min_x \mathcal{L}(x, \lambda, \mu).$$

- The dual problem is defined as:

$$\max_{(\lambda,\mu)} \mathcal{L}_D(\lambda, \mu)$$

- (λ, μ) are called dual feasible if $\lambda \geq 0$.
- The primal function is convex.

Lagrangian duality

- Rewrite the primal function as:

$$\mathcal{L}_P(x) = \max_{\lambda,\mu}[f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{i=1}^{p} \mu_i h_i(x)]$$
$$= f(x) + \max_{\lambda,\mu}[\sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{i=1}^{p} \mu_i h_i(x)]$$

- At a point x, the value of the primal function can be easily evaluated.
- If x is primal feasible (satisfying the constraints), the terms $[\sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{i=1}^{p} \mu_i h_i(x)]$ automatically go to 0.
- If x is not primal feasible (not satisfying one of the constraints), i.e: for some i: $g_i(x) > 0$. Thus we can choose λ_i to be ∞ to maximize the whole function.

Lagrangian duality

- $\mathcal{L}_P(x) = \begin{cases} f(x) & \text{if x is primal feasible} \\ \infty & \text{otherwise} \end{cases}$
- Example:

$$\min_x \frac{1}{2}\|x - z\|^2 \text{ such that: } Ax = 0$$

where $z \in \mathcal{R}^n$ is a fixed point and $A \in \mathcal{R}^{m \times m}$.
Introduce multipliers $\mu \in \mathcal{R}^m$, we have the Lagrangian:

$$\mathcal{L}(x, \mu) = \frac{1}{2}\|x - z\|^2 + \langle \mu, Ax \rangle.$$

- The dual function: minimize $\mathcal{L}(x, \mu)$, so differentiate with respect to x: we get

$$x^*(\mu) = z - A^T \mu.$$

Lagrangian duality

- The dual problem takes the form:

$$\max_{\mu} -\frac{1}{2}\|A^T\mu\|^2 + \langle \mu, Az \rangle.$$

- Let μ^* be the solution the problem above, then we have the relation:

$$x^*(\mu) = z - A^T\mu^*.$$

- Also notice that the optimal value of the dual and primal problem are the same.

Lagrangian duality

- The dual function $\mathcal{L}_D(\lambda, \mu)$ is a concave function,i.e -$\mathcal{L}_D(\lambda, \mu)$ is a convex function.
- Let x^* be the minimizer of the primal problem then $\mathcal{L}_D(\lambda, \mu) \leq f(x*)$.

$$
\begin{aligned}
\mathcal{L}_D(\lambda, \mu) &= min_x \mathcal{L}(x, \lambda, \mu) \leq \mathcal{L}(x^*, \lambda, \mu) \\
&= f(x^*) + \sum_{i=1}^{m} \lambda_i g_i(x^*) + \sum_{i=1}^{m} \mu_i h_i(x^*) = f(x^*).
\end{aligned}
$$

- This shows that for any pair of (λ, μ) that are dual feasible, the dual function gives a lower bound on the optimal value of the primal problem.
- Weak duality: Optimal value of dual problem $\leq$ optimal value of the primal problem.

Lagrangian duality

- If there exists (x^*, λ^*, μ^*) such that: the dual problem obtains its maximum at (λ^*, μ^*) and the primal problem obtains its minimum at x^*, and the optimal value of the primal and dual problems are the same, then x^* is the global minimum of the original constrained optimization problem.
- Based on these results, we can look for a solution of the primal problem by first solving the dual problem to get (λ^*, μ^*) and then find primal solution x^* based on
- In order for the primal and dual problems to have the same value, they have to satisfy some constraints qualification (Slater's condition) which requires existence of a feasible primal solution x such that:$g_i(x) < 0, i = 1, 2 \cdots, m$.

Lagrangian duality

- Strong duality: Optimal value of dual problem = Optimal value of primal problem.
- If strong duality holds, $\lambda_i^* g_i(x^*) = 0, i = 1, 2, \cdots, m$. (Complementary Slackness)
- Complementary slackness implies:

$$\text{If } \lambda_i^* > 0 \rightarrow g_i(x^*) = 0.$$
$$\text{If } g_i(x^*) < 0 \rightarrow \lambda_i^* = 0.$$

- This says, if a dual variables λ_i^* is non-zero, the corresponding inequality constraint be 0 (active constraint).

Lagrangian duality

- Karush Kuhn Tucker condition: When f, g_i are convex differentiable functions, and exists (x^*, λ^*, μ^*) satisfying:
  1. Primal feasibility: $g_i(x^*) \leq 0, i = 1, \cdots, m, h_i(x^*) = 0, i = 1, \cdots, p.$
  2. Dual feasibility: $\lambda \geq 0.$
  3. Complementary slackness: $\lambda_i g_i(x^*) = 0, i = 1, \cdots, m.$
  4. Stationary: $\nabla \mathcal{L}(x^*, \lambda^*, \mu^*) = 0.$

  then x^* is the solution to the primal problem, (λ^*, μ^*) is solution to dual problem. Conversely, if strong duality holds, (x^*, λ^*, μ^*) satisfy all those conditions above.

Lagrangian duality

- Example:

$$\min_x x1^2 + x_2 \text{ such that } :$$
$$4 - 2x_1 - x_2 \leq 0, 1 - x_2 \leq 0.$$

- Introduce dual variables: λ_1, λ_2, the Lagrangian function is :

$$\mathcal{L}(x, \lambda) = x1^2 + x_2 + \lambda_1(4 - 2x_1 - x_2) + \lambda_2(1 - x_2).$$

- The dual function is : $min_x \mathcal{L}(x, \lambda)$.
- $\frac{\partial \mathcal{L}(x,\lambda)}{\partial x} = \begin{pmatrix} 2x_1 - 2\lambda_1 \\ 1 - \lambda_1 - \lambda_2 \end{pmatrix}$. Set this to 0, we get $x_1 = \lambda_1$

Lagrangian duality

- Now we can show that the dual function has the form:

$$
\begin{aligned}
\mathcal{L}_D(x,\lambda) &= min_x \mathcal{L}(x,\lambda) \\
&= \min_{x_2} \lambda_1^2 + x_2 + \lambda_1(4 - 2\lambda_1 - x_2) + \lambda_2(1 - x_2) \\
&= \min_{x_2} -\lambda_1^2 + 4\lambda_1 + \lambda_2 + x_2(1 - \lambda_1 - \lambda_2)
\end{aligned}
$$

- If $1 - \lambda_1 - \lambda_2 > 0$, $\mathcal{L}_D(x,\lambda)$ can obtain $-\infty$ value by choosing $x_2 = -\infty$. Similar when $1 - \lambda_1 - \lambda_2 < 0$.
- Since the goal is to maximize the dual function, we'd better have $1 - \lambda_1 - \lambda_2 = 0$.

Lagrangian duality

- The dual problem is:

$$\begin{aligned} max_\lambda - \lambda_1^2 + 4\lambda_1 + \lambda_2 \text{ such that } &: \\ \lambda \geq 0, 1 - \lambda_1 - \lambda_2 = 0. \end{aligned}$$

SVM nonseparable case

- SVM nonseparable formulation:

$$\min_{w,b} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} max(0, 1 - y_i(b + w^T x_i))$$

- Nonconstrained formulation:

$$\min_{w,b} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} a_i \text{ such that:}$$

$$a_i \geq 0, a_i \geq 1 - y_i(b + w^T x_i), i = 1, 2, \cdots, n.$$

$$\min_{w,b} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} a_i \text{ such that:}$$

$$-a_i \leq 0, 1 - y_i(b + w^T x_i) - a_i \leq 0, i = 1, 2, \cdots, n.$$

SVM nonseparable case

- The Lagrangian function:$\mathcal{L}(x, \lambda, \mu)$

$$\frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} a_i + \sum_{i=1}^{n} \lambda_i(1 - y_i(b + w^T x_i) - a_i) - \sum_{i=1}^{n} \mu_i a_i$$

- The dual problem: $\max_{\lambda\geq 0, \mu \geq 0} \mathcal{L}_D(\lambda, \mu)$ where $\mathcal{L}_D(\lambda, \mu) = \min_{w,b,a} \mathcal{L}(w, b, a_i, \lambda, \mu)$
- Stationary condition: (denote Y is the diagonal matrix with diagonal =y)

$$\nabla \mathcal{L}_w = w - \sum_{i=1}^{n} \lambda_i y_i x_i = w - \lambda^T Y X = w - X^T Y \lambda = 0$$

$$\nabla \mathcal{L}_b = \sum_{i=1}^{n} \lambda_i y_i = 0$$

$$\nabla \mathcal{L}_{a_i} = C - \lambda_i - \mu_i = 0.$$

SVM nonseparable case

- Plug the condition we have above into the Lagrangian, we obtain the dual function:

$$\mathcal{L}_D(\lambda, \mu) = \frac{1}{2}\|w\|_2^2 + \sum_{i=1}^{n} \lambda_i - X^T Y \lambda.$$

- Use the fact that $w = X^T Y \lambda$, the dual function becomes:

$$\sum_{i=1}^{n} \lambda_i - \frac{1}{2}\lambda^T Y^T X X^T Y \lambda.$$

- The dual problem is:

$$\max_{\lambda,\mu} \sum_{i=1}^{n} \lambda_i - \frac{1}{2}\lambda^T Y^T X X^T Y \lambda \text{ such that:}$$

$$\sum_{i=1}^{n} y_i \lambda_i = 0$$

$$\lambda \geq 0, \mu \geq 0, \lambda + \mu = C.$$

SVM nonseparable case

- Since the dual function does not have μ in it, we can simplify:

$$\max_{\lambda,\mu} \sum_{i=1}^{n} \lambda_i - \frac{1}{2}\lambda^T Y^T X X^T Y \lambda \text{ such that:}$$

$$\sum_{i=1}^{n} y_i \lambda_i = 0$$

$$0 \leq \lambda \leq C.$$

- The dual problem is also a quadratic function, with much simpler constraints, with only n unknown variables λ.
- If we can solve the dual problem, the primal solution w can be recovered as: $w = X^T Y \lambda$.

SVM nonseparable case

- Complementary slackness:

$$\begin{aligned} \lambda_i(1 - a_i - y_i(w^T x_i + b)) = 0. \\ \mu_i a_i = 0. \end{aligned}$$

- The condition translate to:

$$\begin{aligned} \lambda_i = 0 \rightarrow y_i(w^T x_i + b) \geq 1. \\ 0 < \lambda_i < C \rightarrow y_i(w^T x_i + b) = 1. \\ \lambda_i = C \rightarrow y_i(w^T x_i + b) \leq 1. \end{aligned}$$

- Notice that for those constraints i that correspond to $0 < \lambda_i < C$, those points are actually on hyperplane class 1 or -1.

SVM nonseparable case

- Calculate intercept b: Once w is obtained, we can calculate b by: find those points of class 1 that are on hyperplane 1: $\min_i w^T x_i$, for those we have:

$$w^T x_i + b = 1.$$

Similarly, for class -1 points that are on hyperplane -1: $\max_i w^T x_i$:$w^T x_i + b = 1$.