University of Illinois Fall 2018

ECE 544NA: Pattern Recognition
Lecture 7: September 20

Lecturer: Alexander Schwing Scribe: Zhipeng Hong

1 Introduction

In the last lecture we discussed the optimization problem without constraints. So how do we solve a constrained optimization problem? One obvious method is let us just use the previous procedure, and make sure that we fall into the feasible set. For example, we can use projection gradient descent, which means we take a step and project to feasible set. However, this is only practicable when the projection is easy to do, like bound constraints, where every entry of $\mathbf{w}$ is between like 0 and 1. Besides this, there is another method, called Lagrangian, which has more universality, and is more powerful to solve constrained optimization problem. In this article, we will introduce what is Lagrangian and dual program, and discuss the mathematical principle behind them.

2 Dual Program

In the following section we show that how we can use Lagrangian to get dual program, then solve constrained optimization problem. Two examples of solving Linear Program and Logistic Regression using dual program will be provided. We will discuss the properties of dual program at end.

2.1 General step of obtaining dual program

The standard form of original constrained optimization problem is listed below:

$$\begin{array}{ll} \min_{\mathbf{w}} & f_0(\mathbf{w}) \\ s.t. & f_i(\mathbf{w}) \leq 0 \quad \forall i \in \{1, ..., C_1\} \\ & h_i(\mathbf{w}) = 0 \quad \forall i \in \{1, ..., C_2\} \end{array}$$

Where f_0 is the function we want to minimize. f_i and h_i are inequality constraints and equality constraints respectively. Here we got $(C_1 + C_2)$ constraints. To put these conditions into consideration together, one idea is combining them in one expression. Then we introduce two sets of parameter, λ_i and ν_i, to be the coefficients of every inequality constraints and equality constraints. We call them Lagrange multiplier. Here we get:

$$L(\mathbf{w}, \lambda, \nu) = f_0(\mathbf{w}) + \sum_{i=1}^{C_1} \lambda_i f_i(\mathbf{w}) + \sum_{i=1}^{C_1} \nu_i h_i(\mathbf{w}) \tag{1}$$

Here $L(\mathbf{w}, \lambda, \nu)$ is Lagrangian, which combines the optimization problem and constraints together. The $\mathbf{w}$ should be in the set $\mathcal{W}$, in which we subsume all other constaints. That means some constraints are not combined in Lagrangian and will be taken care of explicitly within $\mathcal{W}$. That is because when the constraint is simple, like simply $x \geq 0$, combining it into Lagrangian and dealing with them implicitly will complicate the dual program and add some unnecessary terms. Now, we need to find the relationship between Lagrangian and origianal problem. Let $\lambda_i \geq 0$, because $f_i(\mathbf{w}) \leq 0$ and $h_i(\mathbf{w}) = 0$, we get:

$$f_0(\mathbf{w}) \geq L(\mathbf{w}, \lambda, \nu) \tag{2}$$

Thus, Lagrangian gives the lower bound of original problem. To further analyze the Lagrangian, we minimize it with respect to $\mathbf{w} \in \mathcal{W}$. $\mathcal{W}$ is larger than feasible set because here we do not consider constraints, the optimal $\mathbf{w}^*$ we get here might not satisfy $f_i(\mathbf{w}^*) \leq 0$. The further reason will be discussed in Section 3. The result of minimization will be smaller or equal than Lagrangian where $\mathbf{w}$ can be any possible value. we can call it $g(\lambda, \nu)$:

$$L(\mathbf{w}, \lambda, \nu) \geq \min_{\mathbf{w} \in \mathcal{W}} L(\mathbf{w}, \lambda, \nu) = g(\lambda, \nu) \tag{3}$$

Combine (2) and (3), we can get the relationship between $f()$ and $g()$:

$$f_0(\mathbf{w}) \geq g(\lambda, \nu) \tag{4}$$

Remember our goal is to minimize $f_0(\mathbf{w})$ in the original problem. Thus, to get the best lower bound, we should maximize $g(\lambda, \nu)$ with respect to λ and ν. We call $f_0(\mathbf{w})$ primal program, and $g(\lambda, \nu)$ dual problem. Now we have dual problem which lives in another space, but give us values that are guarantee to be smaller than minimum of primal program. The the minimization of primal problem is transformed to equivalent maximization of dual problem. Sometimes the dual program is easier to solve, thus we can get the answer of a difficult primal problem by dealing with a more solvable one.

The main idea of Lagrangian is space transformation. The primal program has parameter $\mathbf{w}$, the dual program has parameters λ_i and ν, and the Lagrangian has all of parameters above. Thus we can regard Lagrangian as a bridge between primal and dual problem, helping us transform a problem which is difficult to solve in one space to another space.

In conclusion, to get the dual program, there are 5 steps:

1. Bring primal program into standard form.
2. Assign Lagrange multipliers to a suitable set of constraints.
3. Subsume all other constrains in $\mathcal{W}$.
4. Write down the Lagrangian L.
5. Minimize Lagrangian w.r.t. primal variables s.t. $w \in \mathcal{W}$.

Once we get dual program, we can try to solve it and get the lower bound of primal program.

2.2 Two examples

2.2.1 linear program

The linear program is the most classical constrained optimization problem, with considerable realistic significance [2].
. Here we derive the dual program of it with the step described above. The original problem is:

$$\min_{\mathbf{w}} \mathbf{c}^T \mathbf{w} \qquad s.t. \quad \mathbf{A}\mathbf{w} \leq \mathbf{b}$$

We bring the constraints into standard form, i.e. $\mathbf{A}\mathbf{w} - \mathbf{b} \leq 0$, then we can get Lagrangian with Lagrange multipliers $\lambda_i \geq 0$:

$$L() = \mathbf{c}^{\mathrm{T}}\mathbf{w} + \lambda^{\mathrm{T}}(\mathbf{A}\mathbf{w} - \mathbf{b}) = (\mathbf{c} + \mathbf{A}^{\mathrm{T}}\lambda)^{\mathrm{T}}\mathbf{w} - \mathbf{b}^{\mathrm{T}}\lambda$$

Then we minimizing Lagrangian w.r.t. primal variables:

$$\min_{\mathbf{w}} L() = \begin{cases} \mathbf{b}^{\mathrm{T}}\lambda\mathbf{c} + \mathbf{A}^{\mathrm{T}} & \lambda = 0 \\ -\infty & otherwise \end{cases}$$

The dual program is:

$$\max_{\lambda \geq 0} -\mathbf{b}^{\mathrm{T}}\lambda \qquad s.t. \quad \mathbf{c} + \mathbf{A}^{\mathrm{T}}\lambda = 0$$

We can notice that if the constraint conditions is more than decision variables in linear program, i.e. the number of columns of A is larger than the number of rows of A, the number of constraints is less in dual program than that in primal program. In practical issues, this transformation can usually simplify the problem.

2.2.2 logistic regression

The method of dual program can also be applied to some optimization which do not have constraints explicitly, for example, the logistic regression with $l2$ penalty. Contrast to linear regression, the loss function of logistic regression has logarithmic term, so it is too complex to be optimized analytically. However, the method of dual program gives us a way to solve this problem. The original problem is:

$$\min_{w} \frac{C}{2} \|\mathbf{w}\|_2^2 + \sum_{(x^{(i)}, y^{(i)}) \in D} \log(1 + \exp(-y^{(i)}\mathbf{w}^T\phi(x^{(i)})))$$

We can add some constraints by introducing a new variable z. Make z always equal to a part of the expression and regard it as constraint. Then the problem can reformat to:

$$\min_{w} \frac{C}{2} \|\mathbf{w}\|_2^2 + \sum_{(x^{(i)}, y^{(i)}) \in D} \log(1 + \exp(-z))) \qquad s.t. \quad z = y^{(i)}\mathbf{w}^T\phi(x^{(i)})$$

Now we have constraints, so we can go through the same 5 steps to get the dual program. The Lagrangian is blow:

$$L() = \frac{C}{2} \|\mathbf{w}\|_2^2 + \sum_{(x^{(i)}, y^{(i)}) \in D} \lambda^{(i)} y^{(i)}\mathbf{w}^T\phi(x^{(i)}) + \sum_{(x^{(i)}, y^{(i)}) \in D} [\log(1 + \exp(-z^{(i)}) + \lambda^{(i)} z^{(i)}]$$

Here we get first two terms respected to $\mathbf{w}$ and last term respected to z. To minimize Lagrangian over primal variables, we can set the partial derivatives of $\mathbf{w}$ and z, to 0.

$$\frac{\partial L}{\partial \mathbf{w}} = 0 : \qquad C\mathbf{w} = \sum_{(\mathbf{x^{(i)}}, \mathbf{y^{(i)}}) \in \mathbf{D}} \lambda^{(\mathbf{i})} \mathbf{y}^{(\mathbf{i})} \phi(\mathbf{x}^{(\mathbf{i})})$$

$$\frac{\partial L}{\partial z^{(i)}} = 0 : \qquad \lambda^{(i)} = \frac{exp(-z^{(i)})}{1 + exp(-z^{(i)})} \quad \Rightarrow \quad \lambda^{(i)} \geq 0$$

$$z^{(i)} = log\frac{1-\lambda^{(i)}}{\lambda^{(i)}} \qquad \Rightarrow \quad \lambda^{(i)} \leq 1$$

In the process of finding the derivation of z, by analyzing the feasible value of λ, we obtain two new constraints: $0 \leq \lambda \leq 1$. Then we replace every $\mathbf{w}$ and z in Lagrangian, by expression of λ. The dual program can be expressed below:

$$\max_{w} \quad -\frac{1}{2C}\left\|\sum_{(x^{(i)},y^{(i)})\in D}\lambda^{(i)}y^{(i)}\phi(x^{(i)})\right\|_2^2 + \sum_{(x^{(i)},y^{(i)})\in D} H(\lambda^{(i)}) \qquad s.t. \quad 0 \leq \lambda \leq 1$$

Where $H(\lambda^{(i)})$ is binary entropy. Then the problem become to maximize $g()$. The first part of $g()$ is negative quadratic, and the second part is entropy which is concave function, thus maximizing both parts makes sense. Once we get the optimum λ that can maximize g, we can plug it in the expression and get the best w. Thus the logistic regression can be easily solved by this dual program.

2.3 Properties of dual program

Compared to primal program, dual program sometimes has less constraints, sometimes is easier to optimize, sometimes gives us interesting insights, sometimes gives non-trivial lower bounds. There is nothing for sure. However, the method of dual program just gives us an opportunity, to simplify a difficult constraint optimization problem. The dual program can also be used for sensitivity analysis.

Another interesting thing is that dual program is always concave, no matter what the primal problem looks like. The concavity is a nice property to have, because we can find the global maximum by searching for stationary points. Here we give the proof of this property:

Proof: The dual program can be regarded as the minimum with respect to $\mathbf{w}$ in Lagrangian when Lagrangian multiplier λ and ν are considered to be constants. So it can be expressed as follow:

$$g(\lambda,\nu) = \min\{L(\mathbf{w_1},\lambda,\nu), L(\mathbf{w_2},\lambda,\nu), ..., L(\mathbf{w_n},\lambda,\nu)\} \qquad n=\infty \tag{5}$$

Where $\mathbf{w_n}$ denotes the infinite number of possible $\mathbf{w}$ in $\mathcal{W}$ domain. For convenience, let $\gamma = (\lambda,\nu)$:

$$\begin{aligned}
g(\theta\gamma + (1-\theta)\gamma') &= \min\{L(\mathbf{w_1},\theta\gamma+(1-\theta)\gamma'), ..., L(\mathbf{w_n},\theta\gamma,(1+\theta)\gamma')\} &(6)\\
&\geq \min\{\theta L(\mathbf{w_1},\gamma) + (1-\theta)L(\mathbf{w_1},\gamma'), .., \theta L(\mathbf{w_n},\gamma) + (1-\theta)L(\mathbf{w_n},\gamma')\} &(7)\\
&\geq \theta\min\{L(\mathbf{w_1},\gamma), .., L(\mathbf{w_n},\gamma)\} + (1-\theta)\min\{L(\mathbf{w_1},\gamma'), .., L(\mathbf{w_n},\gamma')\} &(8)\\
&= \theta g(\gamma) + (1-\theta)g(\gamma') &(9)
\end{aligned}$$

By the definition of concave function, the concavity of dual program $g(\lambda,\nu)$ is proved.

In step(6) to step(7), we use the concave property of linear function. Since in $L(\mathbf{w_n},\gamma)$ the value of $\mathbf{w_n}$ has been fixed, $f_i(\mathbf{w_n})$ and $h_i(\mathbf{w_n})$ are constants, then $L()$ has linear relationship with λ,ν. The proof is more intuitive in geometric view. Think about the Lagrangian of an optimization problem with only one constraint. If we draw a graph in λ (or ν) space, whose x-axis is the Lagrangian multiplier and y-axis is the value of $L(\mathbf{w_1},\lambda)$. For any specific $\mathbf{w_n}$, we can draw a line, whose slope is defined by $\mathbf{w_n}$. In the feasible set imagine we have infinite $\mathbf{w_n}$, thus we can draw infinite lines. Since our goal is to find the $\mathbf{w_n}$ which can minimize Lagrangian, we can draw a line

perpendicular to x-axis and sweep along the x-axis, then find the interception which has smallest y value. That means for any possible value of λ, we find a specific $\mathbf{w_n}$ given by slope which can minimize Lagrangian. The line formed by those intersection points is minimized Lagrangian with respect to $\mathbf{w_n}$, i.e. the dual program. These intersection point are on the bottom $\mathbf{w_n}$ lines and continuous, which means the shape of dual program line is convex, just like half of an polygon.

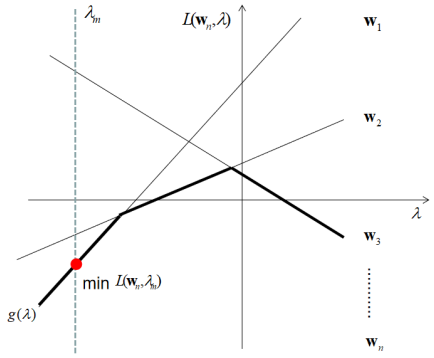

Figure 1: Convexity of Dual

This intuition can be easily extend to n-dimension. Suppose we have n constraints, we can draw Lagrangian of specific $\mathbf{w_n}$ as a hyperplane instead of a line, in $(n+1)$ dimensional space. The hyperplane of dual program will be formed by infinite small linear hyperplane, thus is convex as in one constraint case.

3 Duality

In the sections above, we discussed the dual program provide the lower bound of the primal program. That means the maximum of dual is not necessarily the minimum of primal , there might be a gap between them. Thus we cannot regard the solution of dual program as the final answer of original optimization problem. It is necessary to analyze their relationship, which we call it duality.

The ideal case is the minimum of primal is identical to the maximum of dual, which means, $f(\mathbf{w}^*) = g(\lambda^*, \nu^*)$. We call this relationship Strong Duality. However, this does not hold in general, but usually holds for convex problems. What always holds for convex and non-convex problems is Weak Duality, which means $f(\mathbf{w}^*) \geq g(\lambda^*, \nu^*)$. Although in this case we cannot solve primal directly through solving the dual, we can still get tightest nontrivial lower bound of original problem, which can be used for approximation.

Why strong duality does not hold when solving non-convex problem? To get more intuition, we can view the duality in a geometric way.

To simplify the problem, we still assume a optimization problem $f_0(\mathbf{w})$ with a single inequality constraint $f_1(\mathbf{w}) \leq \mathbf{0}$. Let the value of $f_0(\mathbf{w})$ be t, and the value of $f_1(\mathbf{w})$ be u, then we can define a feasible region $\mathcal{G}$, where every point inside is a possible combination of t and u. In the Figure 2, the region enclosed by the curve is the feasible region $\mathcal{G}$. The shape of $\mathcal{G}$ only depend on the form of $f_0(\mathbf{w})$ and $f_1(\mathbf{w})$. It would be convex when $f_0(\mathbf{w})$ and $f_1(\mathbf{w})$ are both convex, and the rage of $\mathbf{w}$ is continuous. To analyse the case of weak duality, we draw this region as a concave set. Remember the constraint is $f_1(\mathbf{w}) \leq \mathbf{0}$, thus only half of $\mathcal{G}$ on the left side of t-axis is feasible for this problem.

So where is dual program $g(\lambda)$? In this u,t-space, we can simply draw a line, whose interception is $g(\lambda)$ and slope is Lagrangian multiplier λ, since:

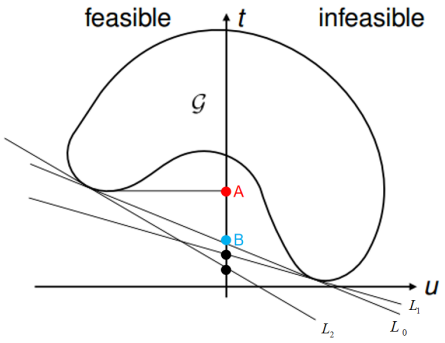

Figure 2: Geometric Interpretation of Duality

$$g(\lambda) = \min_{(u,t)\in\mathcal{G}} t + \lambda u \Rightarrow t^* = -\lambda u^* + g(\lambda) \qquad for \quad (u^*, t^*) \in \mathcal{G} = \{(f_0(\mathbf{w}), f_1(\mathbf{w}) | \mathbf{w} \in \mathcal{W}\} \quad (10)$$

Where (u^*, t^*) is a point in the feasible region, which is the minimum of Lagrangian with respect to primal variables. That mean the line has to have at least one common point with the feasible region. We can prove that the line must be tangent to the feasible region, and (u^*, t^*) is the tangent point.

Proof: Assume a point in the feasible region is blow the line $t = -\lambda u + g(\lambda)$, whose coordinate is (u', t'). Let λ be a constant, then we get $t' + \lambda u' < g(\lambda)$. According to (10), $g(\lambda)$ has to be the minimum with respect to all the points in the feasible region, which is contradict to the inequity above. Thus such point does not exist. Since the line has to have common point with the feasible region, it must be tangent line and the common point is the tangent point.

Then the problem of solving dual program, can be transform to finding a tangent line of feasible region, whose

1. slope must equal or smaller than zero. $(\lambda \geq 0)$
2. intersect is as large as possible (tightest lower bound)
3. whole body must be under the feasible region

From the figure we can notice that, the best line which satisfy all of conditions above is L_0. This line is limited by the minimum of $\mathcal{G}$ on both side of t-axis, feasible half and infeasible half. Other lines like L_1 and L_2, who does not consider either local minimum, have smaller interception than L_0. Thus they only give loose lower bound. From the line L_0, we can get the interception point B, Whose t-value is the maximum of $g(\lambda)$ and the lower bound of primal program.

However, the solution of primal program, is the smallest t-value among all points in feasible set, because t is value of $f_0(\mathbf{w})$. At this time we should not consider the right half of feasible region since there is constraint $f_1(\mathbf{w}) \leq \mathbf{0}$. Then we can get A point by simply draw a horizontal line from the lower point of left half of $\mathcal{G}$ to the vertical axis. There is gap between A and B, which means the solution of dual is smaller than the solution of primal in this case. This gap shows what is weak duality intuitively.

What is the reason of this inconformity? In the process of getting the dual program, we have to minimize Lagrangian with respect to primal variables. Remember in the Lagrangian we simply

combine the optimization and constraints together with Lagrangian multipliers, with no consideration of the inequality of these constraints. The Lagrangian does not represent whether $f_i(\mathbf{w})$ and $h_i(\mathbf{w})$ are smaller than 0, equal to 0, or something else. Thus when we minimize Lagrangian, we ignore the value of constraints actually, and may use $\mathbf{w}$ which is not feasible at all. That is why $\mathcal{W}$ is larger than feasible set as mentioned above.

If we analyse the primal program directly, we can know which part of the feasible region in u-t space is actually feasible, from the constraints. However, if the feasible region is not convex, the solution of dual program is determined by the possible local minimums on both feasible part and infeasible part of the region. Thus the answer can be different from the primal where we only use the feasible part. There are still some cases that the solution of dual is identical to primal for non-convex problem. For example, think about the concave feasible region in Figure 2. If the lowest point on the right side of t-axis is higher than lowest point on left side, we can only draw a horizontal line on left side as the solution of dual which is same as in primal. Because if we still connect two bulges as we did in Figure 2, the slope would be larger than zero. Therefore, for a non-convex problem, only when the local optimum on infeasible part is more powerful than local optimum on feasible part, the strong duality would fail.

4 Consequence of Strong Duality

Although the weak duality is more general, we still can use strong duality in many case, since convex problem is very common in real life. Suppose strong duality holds, what can we get from this property?

4.1 Complementary Slackness

We assume $\mathbf{w}^*$ is the value of $\mathbf{w}$ when $f_0()$ is minimum, and λ^* and ν^* are the value of Lagrangian multipliers when$g()$ is maximum. Since strong duality holds, the minimum and the maximum are identical, we get:

$$\begin{aligned} f_0(\mathbf{w}^*) = g(\lambda^*, \nu^*) &= \min_{\mathbf{w}\in\mathcal{W}} (f_0(\mathbf{w}) + \sum_{i=1}^{C_1} \lambda_i^* f_i(\mathbf{w}) + \sum_{i=1}^{C_2} \nu_i^* h_i(\mathbf{w})) \\ &= f_0(\mathbf{w}^*) + \sum_{i=1}^{C_1} \lambda_i^* f_i(\mathbf{w}^*) + \sum_{i=1}^{C_2} \nu_i^* h_i(\mathbf{w}^*) \\ &\Rightarrow \lambda_i^* f_i(\mathbf{w}^*) = 0 \qquad \forall i \in \{1, ..., C_1\} \end{aligned} \tag{11}$$

Consequently:

$$\begin{cases} \lambda_i^* > 0 & \Rightarrow \quad f_i(\mathbf{w}^*) = 0 \\ f_i(\mathbf{w}^*) < 0 & \Rightarrow \quad \lambda_i^* = 0 \end{cases} \tag{12}$$

This relationship is known as complementary slackness. This property suggests that one and only one between $f_i(\mathbf{w}^*)$ and λ_i^* is equal to 0. There is a more intuitive way to understand it. If we do not consider constraints, the optimum point $\mathbf{w}^*$ have already satisfy the inequity $f_i(\mathbf{w}^*) < 0$, then this inequality constraint is useless. Its Lagrangian multiplier is set to 0 to eliminate the term. On the other hand, if we know λ_i^* is lager than 0, that means this inequity constraint is effective, so we need to take this constraint into consideration. Now the original $\mathbf{w}^*$ is infeasible, we have to find another point in the feasible area of $f_i(\mathbf{w}) \leq 0$ which can best approximate the original optimum

point. Apparently such point is on the edge of feasible area, given that f_0 is convex and monotonic in that area. The expression of the edge is $f_i(\mathbf{w}) = 0$. At this time, the inequality constraint behave same as equality constraint.

4.2 Karush-Kuhn-Tucker Conditions

If strong duality holds and $\mathbf{w}, \lambda, \nu$ are optimal, then they must satisfy the Karush-Kuhn-Tucker (KKT) conditions:

- Primal feasibility: $f_i(\mathbf{w}) \leq 0 \quad \forall i \in \{1, ..., C_1\}; \qquad h_i(\mathbf{w}) = 0 \quad \forall i \in \{1, ..., C_2\}$
- Dual feasibility: $\lambda_i \geq 0 \quad \forall i \in \{1, ..., C_1$
- Complementary slackness: $\lambda_i f_i(\mathbf{w}) \geq 0 \quad \forall i \in \{1, ..., C_1\}$
- Stationarity of Lagrangian

$$\nabla f_0(\mathbf{w}) + \sum_{i=1}^{C_1} \lambda_i \nabla f_i(\mathbf{w}) + \sum_{i=1}^{C_1} \nu_i \nabla h_i(\mathbf{w}) = 0$$

The first three conditions are obvious, which give some basic properties of primal and dual programs. The forth one is the stationary expression of Minimizing Lagrangian. The derivation of $\mathbf{w}$ is set to zero,which is equal to the process of minimizing Lagrangian with respect to primal variables. This form gives us more geometric intuition of constrained optimization. Imagine the contour lines of f_0 and f_i, h_i. By setting the sum of gradient to 0, we get the point where the gradient of f_0 and gradient of f_i, h_i are on the same line and opposite direction. Thus, their contour lines are tangent. Then the Lagrangian multiplier have more intuitive meaning. It represents the ratio of the gradient of optimization problem and corresponding constraints on the optimum point. In other word, it is the sensitivity to a small change of the corresponding constraints. When $|\lambda|$ is small, $\nabla f_i(\mathbf{w})$ is larger than $\nabla f_0(\mathbf{w})$. That means when optimal point changes a little, it will deviate the constraint severely. Otherwise, when $|\lambda|$ is large, the constraint is less sensitive to the optimization problem. Thus sometimes we can sacrifice the precision of constraints a little to get more optimal value in this case.

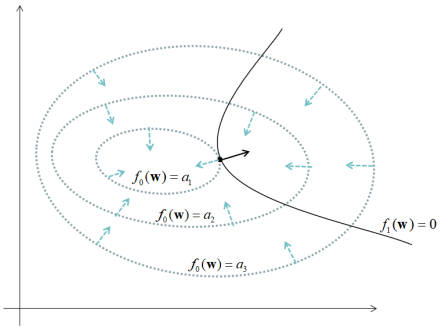

Figure 3: Stationarity of Lagrangian

Once we find the $\mathbf{w}, \lambda, \nu$ which satisfy the KKT conditions, we can get a point on the contour graph equivalently. Because the problem is convex, the shape of contour lines of f_0 and f_i, h_i are also convex, such point is the only which satisfy the conditions. According to the analysis above, this point is the optimal point.

5 Conclusion

The Lagrangian and dual program provide us a powerful method to solve challenging constrained optimization problems. Especially for convex optimization, by solving a dual program in another space, which is usually easier, we can get the answer of original problem directly. In the area of machine learning, this method is useful, not only by helping us simplify the loss function like in logistic regression as we mentioned above, but also in some algorithms like Support Vector Machine, where we consider the training data as constraints. Sometimes using dual formulation also provide opportunity to train a better deep learning model, like Generative adversarial nets (GANs) [1].
We believe the Lagrangian and dual program could spur further research , and help people make great progress in these area.

References

[1] Y. Li, A. Schwing, K.-C. Wang, and R. Zemel. Dualing gans. In *Advances in Neural Information Processing Systems*, pages 5606–5616, 2017.

[2] A. Schrijver. *Theory of linear and integer programming.* John Wiley & Sons, 1998.