EECS 553: Machine Learning (ECE) **University of Michigan**

Constrained Optimization

Winter 2023 **Clayton Scott**

Constrained optimization problems abound in machine learning. In these notes we introduce the basics of constrained optimization, with an emphasis on Lagrange multiplier theory. This will give us enough machinery to derive the support vector machine and understand several other machine learning algorithms that are formulated as constrained optimization problems.

A constrained optimization problem has the form

$$\begin{aligned} \min_{\boldsymbol{u}} \ & f(\boldsymbol{u}) \\ \text{s.t.} \ & g_i(\boldsymbol{u}) \leqslant 0, \quad i = 1, \dots, r \\ & h_j(\boldsymbol{u}) = 0, \quad j = 1, \dots, s \end{aligned}$$

where $\boldsymbol{u} \in \mathbb{R}^p$ and $f, g_1, \dots, g_r, h_1, \dots, h_s : \mathbb{R}^p \to \mathbb{R}$. As an example, the optimal soft-margin hyperplane

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

can be expressed in this form where

$$\boldsymbol{u} = [\boldsymbol{w}^T \ b \ \xi_1 \cdots \xi_n]^T,$$

$p = d + 1 + n$, $r = 2n$, and $s = 0$.

If $\boldsymbol{u}$ satisfies all of the constraints, it is said to be *feasible*. The set of all feasible points is called the *feasible set*, and is assumed to be nonempty. Also assume f is defined on the feasible set..

1 The Lagrangian

The *Lagrangian* is the function

$$L(\boldsymbol{u}, \boldsymbol{\lambda}, \boldsymbol{\nu}) := f(\boldsymbol{u}) + \sum_{i=1}^{r} \lambda_i g_i(\boldsymbol{u}) + \sum_{j=1}^{s} \nu_j h_j(\boldsymbol{u})$$

and $\boldsymbol{\lambda} = [\lambda_1, \dots, \lambda_r]^T$ and $\boldsymbol{\nu} = [\nu_1, \dots, \nu_s]^T$ are called *Lagrange multiplier* or *dual variables*.

The *Lagrange dual function* is

$$L_D(\boldsymbol{\lambda}, \boldsymbol{\nu}) := \min_{\boldsymbol{u}} L(\boldsymbol{u}, \boldsymbol{\lambda}, \boldsymbol{\nu})$$

Note 1. L_D is concave, being the point wise minimum of a family of affine functions. See Figure 1.

The *dual* optimization problem is

$$\max_{\boldsymbol{\lambda},\boldsymbol{\nu}:\lambda_i \geqslant 0} L_D(\boldsymbol{\lambda}, \boldsymbol{\nu})$$

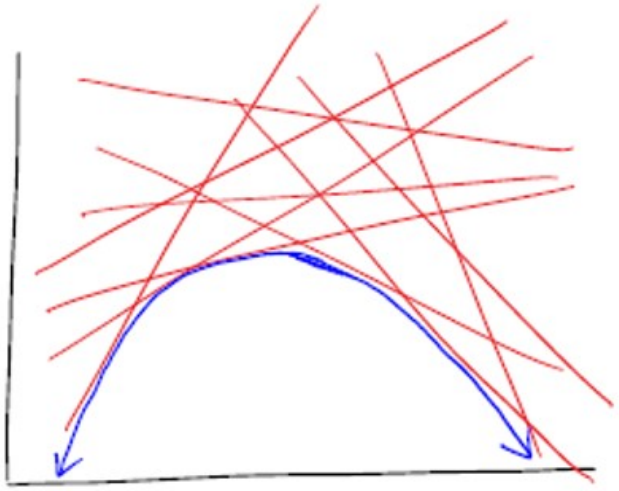

Figure 1: Illustration of $L_D(\boldsymbol{\lambda}, \boldsymbol{\nu})$

Similarly, the *primal function* is

$$L_p(\boldsymbol{u}) := \max_{\boldsymbol{\lambda}, \boldsymbol{v}: \lambda_i \geqslant 0} L(\boldsymbol{u}, \boldsymbol{\lambda}, \boldsymbol{v})$$

and the *primal optimization problem* is

$$\min_{\boldsymbol{u}} L_p(\boldsymbol{u})$$

Notice that

$$L_p(\boldsymbol{u}) = \begin{cases} f(\boldsymbol{u}) & \text{if } \boldsymbol{u} \text{ is feasible} \\ \infty & \text{otherwise} \end{cases}$$

and therefore, the primal problem and the original problem have the same solution(s) and optimal objective values, yet the primal problem is unconstrained.

2 Duality

Denote the optimal objective function values of the primal problem and dual problem by

$$p^* = \min_{\boldsymbol{u}} L_p(\boldsymbol{u}) = \min_{\boldsymbol{u}} \max_{\boldsymbol{\lambda}, \boldsymbol{\nu}: \lambda_i \geq 0} L(\boldsymbol{u}, \boldsymbol{\lambda}, \boldsymbol{\nu})$$
$$d^* = \max_{\boldsymbol{\lambda}, \boldsymbol{\nu}: \lambda_i \geq 0} L_D(\boldsymbol{u}) = \max_{\boldsymbol{\lambda}, \boldsymbol{\nu}: \lambda_i \geq 0} \min_{\boldsymbol{u}} L(\boldsymbol{u}, \boldsymbol{\lambda}, \boldsymbol{\nu}).$$

These values are related.

2.1 Weak Duality

The following property is always true, and is referred to as *weak duality*.

Proposition 1. $d^* \leq p^*$

Proof. Let $\tilde{\boldsymbol{u}}$ be feasible. Then for any $\boldsymbol{\lambda}, \boldsymbol{\nu}$ with $\lambda_i \geq 0$

$$L(\tilde{\boldsymbol{u}}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f(\tilde{\boldsymbol{u}}) + \sum_{i=1}^{r} \lambda_i g_i(\tilde{\boldsymbol{u}}) + \sum_{j=1}^{s} \nu_j h_j(\tilde{\boldsymbol{u}}) \leq f(\tilde{\boldsymbol{u}})$$

Hence

$$L_D(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \min_{\boldsymbol{u}} L(\boldsymbol{u}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \leq f(\tilde{\boldsymbol{u}}).$$

This is true for any feasible $\tilde{\boldsymbol{u}}$, so

$$L_D(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq \min_{\tilde{\boldsymbol{u}} \text{ feasible}} f(\tilde{\boldsymbol{u}}) = p^*.$$

Taking the max order $\boldsymbol{\lambda}, \boldsymbol{\nu} : \lambda_i \geq 0$, we have

$$d^* = \max_{\boldsymbol{\lambda}, \boldsymbol{\nu}: \lambda_i \geq 0} L_D(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq p^*.$$

□

The difference $p^* - d^*$ is called the *duality gap.*

2.2 Strong Duality

If $p^* = d^*$, we say *strong duality* holds. The original constrained optimization problem is said to be *convex* if f and $g_1, \ldots, g_r$ are convex functions and $h_1, \ldots, h_s$ are affine. We state the following without proof.

Theorem 1. *If the original problem is convex and a constraint qualification holds, then $p^* = d^*$.*

Two examples of constraint qualifications are

- All g_i are affine
- (strict feasibility) $\exists \boldsymbol{u}$ s.t. $h_j(\boldsymbol{u}) = 0 \ \forall j$ and $g_i(\boldsymbol{u}) < 0 \ \forall i$.

3 KKT Conditions

In this section, assume $f, g_1, \ldots, g_r, h_1, \ldots, h_s$ are differentiable. For unconstrained optimization, we know $\nabla f(\boldsymbol{u}^*) = 0$ is necessary for $\boldsymbol{u}^*$ to be a global minimizer, and sufficient if f is additionally convex. The following two results generalize these properties to constrained optimization.

Theorem 2. *(Necessity) If $p^* = d^*$, $\boldsymbol{u}^*$ is primal optimal, and $(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ is dual optimal, then the Karesh-Kuhn-Tucker (KKT) conditions hold:*

(1) $\nabla_{\boldsymbol{u}} f(\boldsymbol{u}^*) + \sum_{i=1}^{r} \lambda_i^* \nabla_{\boldsymbol{u}} g_i(\boldsymbol{u}^*) + \sum_{j=1}^{s} \nu_j^* \nabla_{\boldsymbol{u}} h_j(\boldsymbol{u}^*) = 0$

(2) $g_i(\boldsymbol{u}^*) \leq 0 \ \forall i$

(3) $h_j(\boldsymbol{u}^*) = 0 \ \forall j$

(4) $\lambda_i^* \geq 0 \ \forall i$

(5) $\lambda_i^* g_i(\boldsymbol{u}^*) = 0 \ \forall i$ *(complimentary slackness).*

Proof. (2) - (3) hold since $\boldsymbol{u}^*$ is feasible. (4) holds by definition of the dual problem. To prove (5) and (1):

$$\begin{aligned} f(\boldsymbol{u}^*) &= L_D(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) \qquad \text{[by strong duality]} \\ &= \min_{\tilde{\boldsymbol{u}}} \left(f(\tilde{\boldsymbol{u}}) + \sum_{i=1}^{r} \lambda_i^* g_i(\tilde{\boldsymbol{u}}) + \sum_{j=1}^{s} \nu_j^* h_j(\tilde{\boldsymbol{u}}) \right) \\ &\leq f(\boldsymbol{u}^*) + \sum_{i=1}^{r} \lambda_i^* g_i(\boldsymbol{u}^*) + \sum_{j=1}^{s} \nu_j^* h_j(\boldsymbol{u}^*) \\ &\leq f(\boldsymbol{u}^*) \qquad \text{[by (2) - (4)]} \end{aligned}$$

and therefore the two inequalities are equalities. Equality of the last two lines implies $\lambda_i^* g_i(\boldsymbol{u}^*) = 0 \ \forall i$. Equality of the 2nd and 3rd lines implies $\boldsymbol{u}^*$ is a minimizer of $L(\boldsymbol{u}, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ with respect to $\boldsymbol{u}$. Therefore,

$$\nabla_{\boldsymbol{u}} L(\boldsymbol{u}^*, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) = \mathbf{0},$$

which is (1). □

If $(\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}})$ satisifies the KKT conditions, we say that $(\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}})$ is a *feasible KKT point*. The preceding result shows that if $\boldsymbol{u}^*$ is primal optimal, and $(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ is dual optimal, then $(\boldsymbol{u}^*, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ is a feasible KKT point.

Theorem 3. *(Sufficiency) If the original problem is convex and $\tilde{\boldsymbol{u}}$, $\tilde{\boldsymbol{\lambda}}$, $\tilde{\boldsymbol{\nu}}$ satisfy the KKT conditions, then $\tilde{\boldsymbol{u}}$ is primal optimal, $(\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}})$ is dual optimal, and strong duality holds.*

Proof. By (2) and (3), $\tilde{\boldsymbol{u}}$ is feasible. By (4), $L(\boldsymbol{u}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\mu}})$ is convex in $\boldsymbol{u}$. By (1), $\tilde{\boldsymbol{u}}$ is a minimizer of $L(\boldsymbol{u}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}})$. Then

$$\begin{aligned} L_D(\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}}) \qquad &= L(\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}}) \\ &= f(\tilde{\boldsymbol{u}}) + \sum_{i=1}^{r} \tilde{\lambda}_i g_i(\tilde{\boldsymbol{u}}) + \sum_{j=1}^{s} \tilde{\nu}_j h_j(\tilde{\boldsymbol{u}}) \\ &= f(\tilde{\boldsymbol{u}}). \text{ [by (5) and (3)]} \end{aligned}$$

Therefore $p^* = d^*$ and the result follows. □

In conclusion, if a constrained optimization problem is differentiable, convex, and a constraint qualification holds, then the KKT conditions are necessary and sufficient for primal/dual optimality (with zero duality gap). The KKT conditions can be used to simplify and/or solve such problems.

4 Kernel Ridge Regression Revisited

When we studied kernel ridge regression (KRR), we used the matrix inversion lemma to kernelize ridge regression (RR). You may have wondered why this is valid when the feature space associated to the kernel is infinite-dimensional, as in the case of the Gaussian kernel. We will now use the above theory to rederive KRR while avoiding this issue.

Let $(\boldsymbol{x}_i, y_i)$, $i = 1, \ldots, n$ be training data for regression. For simplicity, we let's focus on KRR w/o offset (the w/ offset case is left as an exercise). We may view ridge regression as the solution of

$$\begin{aligned} &\min_{\boldsymbol{w}, \xi} \ \lambda \|\boldsymbol{w}\|^2 + \sum_{i=1}^{n} {\xi_i}^2 \\ &\text{s.t. } y_i - \boldsymbol{w}^T \boldsymbol{x}_i = \xi_i, \quad i = 1, \ldots, n. \end{aligned}$$

Denote the dual variables by $\alpha_i \in \mathbb{R}$. Notice that the above problem is convex and differentiable. Therefore, we can apply the KKT sufficiency theorem. Our goal is thus to find $(\boldsymbol{w}^*, \boldsymbol{\xi}^*, \boldsymbol{\alpha}^*)$ satisfying the KKT conditions, and by the theorem, we will know that $(\boldsymbol{w}^*, \boldsymbol{\xi}^*)$ solves the original/primal problem, and that $\boldsymbol{\alpha}^*$ solves the dual. Since this problem has no inequality constraints, we are only concerned with KKT conditions (1) and (3).

The Lagrangian is

$$L(\boldsymbol{w}, \boldsymbol{\xi}, \boldsymbol{\alpha}) = \lambda \|\boldsymbol{w}\|^2 + \sum_{i=1}^{n} {\xi_i}^2 + \sum_{i=1}^{n} \alpha_i (y_i - \boldsymbol{w}^T \boldsymbol{x}_i - \xi_i).$$

According to the first KKT condition, we require

$$\frac{\partial L}{\partial \boldsymbol{w}} = 2\lambda \boldsymbol{w} - \sum_i \alpha_i \boldsymbol{x}_i = \boldsymbol{0}$$
$$\frac{\partial L}{\partial \xi_i} = 2\xi_i - \alpha_i = 0 \qquad \forall i.$$

This leads to

$$\boldsymbol{w} = \frac{1}{2\lambda} \sum_i \alpha_i \boldsymbol{x}_i \tag{2}$$
$$\xi_i = \frac{1}{2} \alpha_i \qquad \forall i. \tag{3}$$

According to the third KKT condition, we need

$$y_i - \boldsymbol{w}^T \boldsymbol{x}_i = \xi_i \qquad \forall i.$$

Plugging (2) and (3) into this equation, and reformulating as a vector equation, we have

$$\boldsymbol{y} = \frac{1}{2\lambda} (\boldsymbol{K} + \lambda \boldsymbol{I}) \boldsymbol{\alpha}$$

where $\boldsymbol{K}$ is the gram matrix and $\boldsymbol{y} = [y_1 \ \cdots \ y_n]^T$. Therefore, the KKT conditions are satisfied by

$$\boldsymbol{\alpha}^* = 2\lambda (\boldsymbol{K} + \lambda \boldsymbol{I})^{-1} \boldsymbol{y}$$
$$\boldsymbol{w}^* = \frac{1}{2\lambda} \sum_i \alpha_i^* \boldsymbol{x}_i$$
$$\xi_i^* = \frac{1}{2} \alpha_i^*.$$

We may express

$$\begin{aligned} \boldsymbol{w}^* &= \frac{1}{2\lambda} \boldsymbol{X}^T \boldsymbol{\alpha}^* \\ &= \boldsymbol{X}^T (\boldsymbol{K} + \lambda \boldsymbol{I})^{-1} \boldsymbol{y} \end{aligned}$$

where

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix}.$$

The KRR (w/o offset) predictor is then

$$\begin{aligned} \widehat{f}(\boldsymbol{x}) &= (\boldsymbol{w}^*)^T \boldsymbol{x} \\ &= \boldsymbol{y}^T (\boldsymbol{K} + \lambda \boldsymbol{I})^{-1} \boldsymbol{k}(\boldsymbol{x}), \end{aligned}$$

where

$$\boldsymbol{k}(\boldsymbol{x}) = \begin{bmatrix} \langle \boldsymbol{x}_1, \boldsymbol{x} \rangle \\ \vdots \\ \langle \boldsymbol{x}_n, \boldsymbol{x} \rangle \end{bmatrix}.$$

This agrees with our earlier derivation but did not require the matrix inversion lemma.

Exercises

1. (★★) Prove the following property. If $\boldsymbol{u}^*$ is primal optimal, $(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ is dual optimal, and strong duality holds, then $(\boldsymbol{u}^*, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ is a saddle point of L, i.e.,

   $$L(\boldsymbol{u}^*, \boldsymbol{\lambda}, \boldsymbol{\nu}) \leq L(\boldsymbol{u}^*, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) \leq L(\boldsymbol{u}, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$$

   for all $\boldsymbol{u} \in \mathbb{R}^p, \boldsymbol{\lambda} \in \mathbb{R}^m$ with $\lambda_i \geq 0$, and $\boldsymbol{\nu} \in \mathbb{R}^n$. See Figure 2.

2. (★★) Repeat the KRR example for KRR with offset.

3. (☆☆) Use the definition of concavity to prove that for any constrained optimization problem, the dual function is concave.

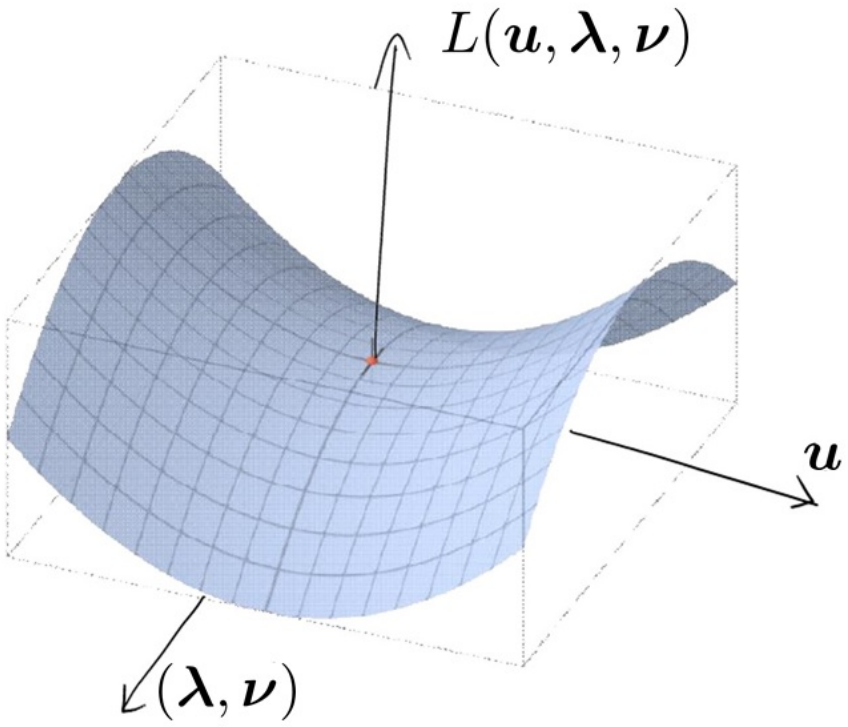

Figure 2: Illustration of a saddle point.