



# Lagrange Multipliers & the Kernel Trick

Nicholas Ruozzi

University of Texas at Dallas

# The Strategy So Far...



- Choose hypothesis space
- Construct loss function (ideally convex)
- Minimize loss to “learn” correct parameters

A mathematical detour, we'll come back to SVMs soon!

$$\min_{x \in \mathbb{R}^n} f_0(x)$$

subject to:

$$\begin{aligned} f_i(x) &\leq 0, & i &= 1, \dots, m \\ h_i(x) &= 0, & i &= 1, \dots, p \end{aligned}$$

$$\min_{x \in \mathbb{R}^n} f_0(x)$$

$f_0$  is not necessarily convex

subject to:

$$\begin{aligned} f_i(x) &\leq 0, & i &= 1, \dots, m \\ h_i(x) &= 0, & i &= 1, \dots, p \end{aligned}$$

$$\min_{x \in \mathbb{R}^n} f_0(x)$$

subject to:

$$\begin{aligned} f_i(x) &\leq 0, & i &= 1, \dots, m \\ h_i(x) &= 0, & i &= 1, \dots, p \end{aligned}$$

Constraints do not need to be linear

# Example



$$\min_{x \in \mathbb{R}^3} x_1 \log x_1 + x_2 \log x_2$$

subject to:

$$x_1 + x_2 = 1$$

$$x_1 \geq 0$$

$$x_2 \geq 0$$

# Example



$$\min_{x \in \mathbb{R}^3} x_1 \log x_1 + x_2 \log x_2$$

subject to:

$$1 - x_1 - x_2 = 0$$

$$-x_1 \leq 0$$

$$-x_2 \leq 0$$

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p \nu_i h_i(x)$$

- Incorporate constraints into a new objective function
- $\lambda \geq 0$  and  $\nu$  are vectors of **Lagrange multipliers**
- The Lagrange multipliers can be thought of as enforcing soft constraints



# Example



$$\min_{x \in \mathbb{R}^3} x_1 \log x_1 + x_2 \log x_2$$

subject to:

$$1 - x_1 - x_2 = 0$$

$$-x_1 \leq 0$$

$$-x_2 \leq 0$$

$$L(x_1, x_2, \nu_1, \lambda_1, \lambda_2)$$

$$= x_1 \log x_1 + x_2 \log x_2 + \nu_1 \cdot (1 - x_1 - x_2) - \lambda_1 x_1 - \lambda_2 x_2$$

- Construct a **dual function** by minimizing the Lagrangian over the primal variables

$$g(\lambda, \nu) = \inf_x L(x, \lambda, \nu)$$

- $g(\lambda, \nu) = -\infty$  whenever the Lagrangian is not bounded from below for a fixed  $\lambda$  and  $\nu$

# Example



$$\min_{x \in \mathbb{R}^3} x_1 \log x_1 + x_2 \log x_2$$

subject to:

$$1 - x_1 - x_2 = 0$$

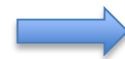
$$-x_1 \leq 0$$

$$-x_2 \leq 0$$

$$L(x_1, x_2, \nu_1, \lambda_1, \lambda_2)$$

$$= x_1 \log x_1 + x_2 \log x_2 + \nu_1 \cdot (1 - x_1 - x_2) - \lambda_1 x_1 - \lambda_2 x_2$$

$$\frac{\partial L}{\partial x_1} = \log x_1 + 1 - \nu_1 - \lambda_1 = 0$$



$$x_1 = \exp(\nu_1 + \lambda_1 - 1)$$

$$\frac{\partial L}{\partial x_2} = \log x_2 + 1 - \nu_1 - \lambda_2 = 0$$

$$x_2 = \exp(\nu_1 + \lambda_2 - 1)$$

# Example



$$\min_{x \in \mathbb{R}^3} x_1 \log x_1 + x_2 \log x_2$$

subject to:

$$1 - x_1 - x_2 = 0$$

$$-x_1 \leq 0$$

$$-x_2 \leq 0$$

$$L(x_1, x_2, \nu_1, \lambda_1, \lambda_2)$$

$$= x_1 \log x_1 + x_2 \log x_2 + \nu_1 \cdot (1 - x_1 - x_2) - \lambda_1 x_1 - \lambda_2 x_2$$

$$g(\nu_1, \lambda_1, \lambda_2)$$

$$= \exp(\nu_1 + \lambda_1 - 1) (\nu_1 + \lambda_1 - 1)$$

$$+ \exp(\nu_1 + \lambda_2 - 1) (\nu_1 + \lambda_2 - 1)$$

$$+ \nu_1 (1 - \exp(\nu_1 + \lambda_1 - 1) - \exp(\nu_1 + \lambda_2 - 1))$$

$$- \lambda_1 \exp(\nu_1 + \lambda_1 - 1) - \lambda_2 \exp(\nu_1 + \lambda_2 - 1)$$

# Example



$$\min_{x \in \mathbb{R}^3} x_1 \log x_1 + x_2 \log x_2$$

subject to:

$$\begin{aligned} 1 - x_1 - x_2 &= 0 \\ -x_1 &\leq 0 \\ -x_2 &\leq 0 \end{aligned}$$

$$\begin{aligned} L(x_1, x_2, \nu_1, \lambda_1, \lambda_2) \\ = x_1 \log x_1 + x_2 \log x_2 + \nu_1 \cdot (1 - x_1 - x_2) - \lambda_1 x_1 - \lambda_2 x_2 \end{aligned}$$

$$g(\nu_1, \lambda_1, \lambda_2) = -\exp(\nu_1 + \lambda_1 - 1) - \exp(\nu_1 + \lambda_2 - 1) + \nu_1$$

$$\min_{x \in \mathbb{R}^n} f_0(x)$$

subject to:

$$\begin{aligned} f_i(x) &\leq 0, & i &= 1, \dots, m \\ h_i(x) &= 0, & i &= 1, \dots, p \end{aligned}$$

Equivalently,

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

Why are these equivalent?

$$\min_{x \in \mathbb{R}^n} f_0(x)$$

subject to:

$$\begin{aligned} f_i(x) &\leq 0, & i &= 1, \dots, m \\ h_i(x) &= 0, & i &= 1, \dots, p \end{aligned}$$

Equivalently,

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

$$\sup_{\lambda \geq 0, \nu} \left[ f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p \nu_i h_i(x) \right] = \infty$$

whenever  $x$  violates the constraints

$$\sup_{\lambda \geq 0, \nu} g(\lambda, \nu)$$

Equivalently,

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

- The dual problem is always concave, even if the primal problem is not convex
  - For each  $x$ ,  $L(x, \lambda, \nu)$  is a linear function in  $\lambda$  and  $\nu$
  - Maximum (or supremum) of concave functions is concave!



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

- Why?
  - $g(\lambda, \nu) \leq L(x, \lambda, \nu)$  for all  $x$
  - $L(x', \lambda, \nu) \leq f_0(x')$  for any feasible  $x'$ ,  $\lambda \geq 0$ 
    - $x$  is **feasible** if it satisfies all of the constraints
  - Let  $x^*$  be the optimal solution to the primal problem and  $\lambda \geq 0$

$$g(\lambda, \nu) \leq L(x^*, \lambda, \nu) \leq f_0(x^*)$$

# Example



$$\min_{x \in \mathbb{R}^3} x_1 \log x_1 + x_2 \log x_2$$

subject to:

$$\begin{aligned} 1 - x_1 - x_2 &= 0 \\ -x_1 &\leq 0 \\ -x_2 &\leq 0 \end{aligned}$$

$$\begin{aligned} L(x_1, x_2, \nu_1, \lambda_1, \lambda_2) \\ = x_1 \log x_1 + x_2 \log x_2 + \nu_1 \cdot (1 - x_1 - x_2) - \lambda_1 x_1 - \lambda_2 x_2 \end{aligned}$$

$$g(\nu_1, \lambda_1, \lambda_2) = -\exp(\nu_1 + \lambda_1 - 1) - \exp(\nu_1 + \lambda_2 - 1) + \nu_1$$

$$\frac{\partial g}{\partial \nu_1} = -\exp(\nu_1 + \lambda_1 - 1) - \exp(\nu_1 + \lambda_2 - 1) + 1 = 0$$

$g$  is a decreasing function of  $\lambda_1$  and  $\lambda_2$ ,  
so the optimum is achieved at the boundary  $\lambda_1 = \lambda_2 = 0$

# Example



$$\min_{x \in \mathbb{R}^3} x_1 \log x_1 + x_2 \log x_2$$

subject to:

$$1 - x_1 - x_2 = 0$$

$$-x_1 \leq 0$$

$$-x_2 \leq 0$$

$$L(x_1, x_2, \nu_1, \lambda_1, \lambda_2)$$

$$= x_1 \log x_1 + x_2 \log x_2 + \nu_1 \cdot (1 - x_1 - x_2) - \lambda_1 x_1 - \lambda_2 x_2$$

$$g(\nu_1, \lambda_1, \lambda_2) = -\exp(\nu_1 + \lambda_1 - 1) - \exp(\nu_1 + \lambda_2 - 1) + \nu_1$$

$$\frac{\partial g}{\partial \nu_1} = -\exp(\nu_1 + \lambda_1 - 1) - \exp(\nu_1 + \lambda_2 - 1) + 1 = 0$$

$$-\exp(\nu_1 - 1) - \exp(\nu_1 - 1) + 1 = 0$$

$$\exp(\nu_1 - 1) = .5$$

$$\nu_1 = \log(.5) + 1$$

# More Examples



- Minimize  $x^2 + y^2$  subject to  $x + y \geq 1$
- Given a point  $z \in \mathbb{R}^n$  and a hyperplane  $w^T x + b = 0$ , find the projection of the point  $z$  onto the hyperplane

- Under certain conditions, the two optimization problems are equivalent

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

- This is called **strong duality**
- If the inequality is strict, then we say that there is a **duality gap**
  - Size of gap measured by the difference between the two sides of the inequality

# Slater's Condition



For any optimization problem of the form

$$\min_{x \in \mathbb{R}^n} f_0(x)$$

subject to:

$$\begin{aligned} f_i(x) &\leq 0, & i = 1, \dots, m \\ Ax &= b \end{aligned}$$

where  $f_0, \dots, f_m$  are **convex functions**, strong duality holds if there exists an  $x$  such that

$$\begin{aligned} f_i(x) &< 0, & i = 1, \dots, m \\ Ax &= b \end{aligned}$$

$$\min_w \frac{1}{2} \|w\|^2$$

such that

$$y_i(w^T x^{(i)} + b) \geq 1, \text{ for all } i$$

- Note that Slater's condition holds as long as the data is linearly separable

$$L(w, b, \lambda) = \frac{1}{2} w^T w + \sum_i \lambda_i (1 - y_i (w^T x^{(i)} + b))$$

Convex in  $w$ , so take derivatives to form the dual

$$\frac{\partial L}{\partial w_k} = w_k + \sum_i -\lambda_i y_i x_k^{(i)} = 0$$

$$\frac{\partial L}{\partial b} = \sum_i -\lambda_i y_i = 0$$



$$L(w, b, \lambda) = \frac{1}{2} w^T w + \sum_i \lambda_i (1 - y_i (w^T x^{(i)} + b))$$

Convex in  $w$ , so take derivatives to form the dual

$$w = \sum_i \lambda_i y_i x^{(i)}$$

$$\sum_i \lambda_i y_i = 0$$

$$\max_{\lambda \geq 0} -\frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j x^{(i)T} x^{(j)} + \sum_i \lambda_i$$

such that

$$\sum_i \lambda_i y_i = 0$$

- By strong duality, solving this problem is equivalent to solving the primal problem
  - Given the optimal  $\lambda$ , we can easily construct  $w$  ( $b$  can be found by **complementary slackness...**)

# Complementary Slackness



- Suppose that there is zero duality gap
- Let  $x^*$  be an optimum of the primal and  $(\lambda^*, \nu^*)$  be an optimum of the dual

$$\begin{aligned} f_0(x^*) &= g(\lambda^*, \nu^*) \\ &= \inf_x \left[ f_0(x) + \sum_{i=1}^m \lambda_i^* f_i(x) + \sum_{i=1}^p \nu_i^* h_i(x) \right] \\ &\leq f_0(x^*) + \sum_{i=1}^m \lambda_i^* f_i(x^*) + \sum_{i=1}^p \nu_i^* h_i(x^*) \\ &= f_0(x^*) + \sum_{i=1}^m \lambda_i^* f_i(x^*) \\ &\leq f_0(x^*) \end{aligned}$$

# Complementary Slackness



- This means that

$$\sum_{i=1}^m \lambda_i^* f_i(x^*) = 0$$

- As  $\lambda \geq 0$  and  $f_i(x_i^*) \leq 0$ , this can only happen if  $\lambda_i^* f_i(x^*) = 0$  for all  $i$
- Put another way,
  - If  $f_i(x^*) < 0$  (i.e., the constraint is not tight), then  $\lambda_i^* = 0$
  - If  $\lambda_i^* > 0$ , then  $f_i(x^*) = 0$
  - **ONLY applies when there is no duality gap**

$$\max_{\lambda \geq 0} -\frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j x^{(i)T} x^{(j)} + \sum_i \lambda_i$$

such that

$$\sum_i \lambda_i y_i = 0$$

- By complementary slackness,  $\lambda_i^* > 0$  means that  $x^{(i)}$  is a support vector (can then solve for  $b$  using  $w$ )

$$\max_{\lambda \geq 0} -\frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j x^{(i)T} x^{(j)} + \sum_i \lambda_i$$

such that

$$\sum_i \lambda_i y_i = 0$$

- Takes  $O(n^2)$  time just to evaluate the objective function
  - Active area of research to try to speed this up