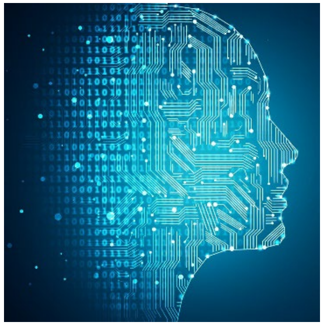
Machine Learning

Constrained Optimization

Kevin Moon (kevin.moon@usu.edu)

STAT/CS 5810/6655



Outline

1. Constrained optimization problems
2. The Lagrangian
3. Dual optimization problems
4. KKT conditions

Motivation

- Today's lecture will allow us to better understand the optimal soft-margin hyperplane which solves

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

- In particular, by converting the above constrained optimization problem to its dual, we will be able to kernelize this method. This leads to the so-called *support vector machine*.

- Constrained optimization problems are ubiquitous in machine learning.

Constrained Optimization

- A *constrained optimization problem* has the form

$$
\begin{aligned}
\min_{\boldsymbol{x} \in \mathbb{R}^d} \quad & f(\boldsymbol{x}) \\
s.t. \quad & g_i(\boldsymbol{x}) \leq 0, \quad i = 1, \ldots, m \\
& h_j(\boldsymbol{x}) = 0, \quad j = 1, \ldots, n
\end{aligned}
$$

  where $\boldsymbol{x} \in \mathbb{R}^d$.

- If $\boldsymbol{x}$ satisfies all of the constraints, it is said to be *feasible*.
- Assume f is defined at all feasible points.

Constrained Optimization

- A *constrained optimization problem* has the form

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{R}^d} \quad & f(\boldsymbol{x}) \\ s.t. \quad & g_i(\boldsymbol{x}) \leq 0, \quad i = 1, \dots, m \\ & h_j(\boldsymbol{x}) = 0, \quad j = 1, \dots, n \end{aligned}$$

where $\boldsymbol{x} \in \mathbb{R}^d$.

- A constrained optimization problem is convex if:
  1. f is <u>convex</u>
  2. g_i is <u>convex</u> $\forall i = 1, \dots, m$
  3. h_j is <u>linear/affine</u> $\forall j = 1, \dots, n$

Lagrangian

- The *Lagrangian* is

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

- $\boldsymbol{\lambda} = [\lambda_1, \dots, \lambda_m]^T$ and $\boldsymbol{\nu} = [\nu_1, \dots, \nu_n]^T$ are called *Lagrange multipliers* or *dual variables*

Dual Function

- The *Lagrangian dual function* is

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

- L_D is concave (proof in Duality.pdf)
- The dual optimization problem is

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

- The original constrained optimization problem is sometimes called the *primal optimization problem*

Rewriting the Primal

- The primal may be rewritten as

$$\min_{\boldsymbol{x}} \max_{\boldsymbol{\lambda},\boldsymbol{\nu}:\lambda_i \geq 0} L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$$

- If x is not feasible, the value of $\max_{\boldsymbol{\lambda},\boldsymbol{\nu}:\lambda_i \geq 0} L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$ is ∞.
  - Otherwise, it is $f(\boldsymbol{x})$

Weak Duality

- Denote the optimal objective function values of the primal and dual

$$
\begin{aligned}
p^* &= \min_{\boldsymbol{x}} \max_{\boldsymbol{\lambda},\boldsymbol{\nu}:\lambda_i \geq 0} L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \\
d^* &= \max_{\boldsymbol{\lambda},\boldsymbol{\nu}:\lambda_i \geq 0} \min_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = \max_{\boldsymbol{\lambda},\boldsymbol{\nu}:\lambda_i \geq 0} L_D(\boldsymbol{\lambda}, \boldsymbol{\nu})
\end{aligned}
$$

- Weak duality refers to the following fact which always holds:
- **Theorem**: $d^* \leq p^*$

Weak Duality

Proof of weak duality: Let $\tilde{\boldsymbol{x}}$ be feasible. Then for any $\boldsymbol{\lambda}, \boldsymbol{\nu}$ with $\lambda_i \geq 0$

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

Hence

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

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

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

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

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

Strong Duality

- If $p^* = d^*$, we say *strong duality* holds.
- The original unconstrained optimization problem is said to be *convex* if f and $g_1, \dots, g_m$ are convex functions and $h_1, \dots, h_n$ are affine.
- We state the following without proof.
- **Theorem:** If the original problem is convex and a constraint qualification holds, then $p^* = d^*$.
- Examples of constraint qualifications:
  - All g_i are affine
  - (Strict feasibility) $\exists \boldsymbol{x}$ s.t. $h_j(\boldsymbol{x}) = 0 \; \forall j$ and $g_i(\boldsymbol{x}) < 0 \; \forall i$

Big Picture

- For unconstrained optimization problems with differentiable objective functions, we saw that
  - $\nabla f(\boldsymbol{x}) = \mathbf{0}$ is *necessary* for $\boldsymbol{x}$ to be a global minimizer
  - If f is convex, then $\nabla f(\boldsymbol{x}) = \mathbf{0}$ is *sufficient* for $\boldsymbol{x}$ to be a global minimizer
- For constrained optimization problems with differentiable objective and constraints, a similar result holds where $\nabla f(\boldsymbol{x}) = \mathbf{0}$ is replaced by the *Karesh-Kuhn-Tucker (KKT) conditions*
- We can use these conditions to solve and understand constrained optimization problems

KKT Conditions: Necessity

- From now on assume f, g_i and h_j are all differentiable.
- **Theorem:** If $p^* = d^*$, $\boldsymbol{x}^*$ is a primal optimal, and $(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ is dual optimal, then the KKT conditions hold:
  1. $\nabla f(\boldsymbol{x}^*) + \sum_{i=1}^m \lambda_i^* \nabla g_i(\boldsymbol{x}^*) + \sum_{j=1}^n \nu_j^* \nabla h_j(\boldsymbol{x}^*) = \mathbf{0}$
  2. $g_i(\boldsymbol{x}^*) \leq 0 \ \forall i$
  3. $h_j(\boldsymbol{x}^*) = 0 \ \forall j$
  4. $\lambda_i^* \geq 0 \ \forall i$
  5. $\lambda_i^* g_i(\boldsymbol{x}^*) = 0 \ \forall i$ (complimentary slackness)

KKT Conditions: Necessity

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

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

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

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

which is (1).

KKT Conditions: Sufficiency

- **Theorem:** If the original problem is convex and $\tilde{\boldsymbol{x}}$, $\tilde{\boldsymbol{\lambda}}$, $\tilde{\boldsymbol{\nu}}$ satisfy the KKT conditions

  1. $\nabla f(\tilde{\boldsymbol{x}}) + \sum_{i=1}^{m} \tilde{\lambda}_i \nabla g_i(\tilde{\boldsymbol{x}}) + \sum_{j=1}^{n} \tilde{\nu}_j \nabla h_j(\tilde{\boldsymbol{x}}) = \mathbf{0}$
  2. $g_i(\tilde{\boldsymbol{x}}) \leq 0 \ \forall i$
  3. $h_j(\tilde{\boldsymbol{x}}) = 0 \ \forall j$
  4. $\tilde{\lambda}_i \geq 0 \ \forall i$
  5. $\tilde{\lambda}_i g_i(\tilde{\boldsymbol{x}}) = 0 \ \forall i$ (complimentarity or complementary slackness)

  then $\tilde{\boldsymbol{x}}$ is primal optimal, $(\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}})$ is dual optimal, and strong duality holds.

KKT Conditions: Sufficiency

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

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

Therefore $p^* \leq d^*$. But we know $p^* \geq d^*$ by weak duality, and so we must have $p^* = d^*$, with $\tilde{\boldsymbol{x}}$ being primal optimal, and $(\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}})$ being dual optimal.

How is this useful?

- We can use the KKT conditions to solve the primal and/or dual
- Sometimes it is easier to solve the dual than the primal (computationally or analytically)

- In particular: if $(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ is dual optimal, then any primal optimal point $\boldsymbol{x}^*$ is a solution of

$$\min_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$$

or

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

Example

Consider the following optimization problem:

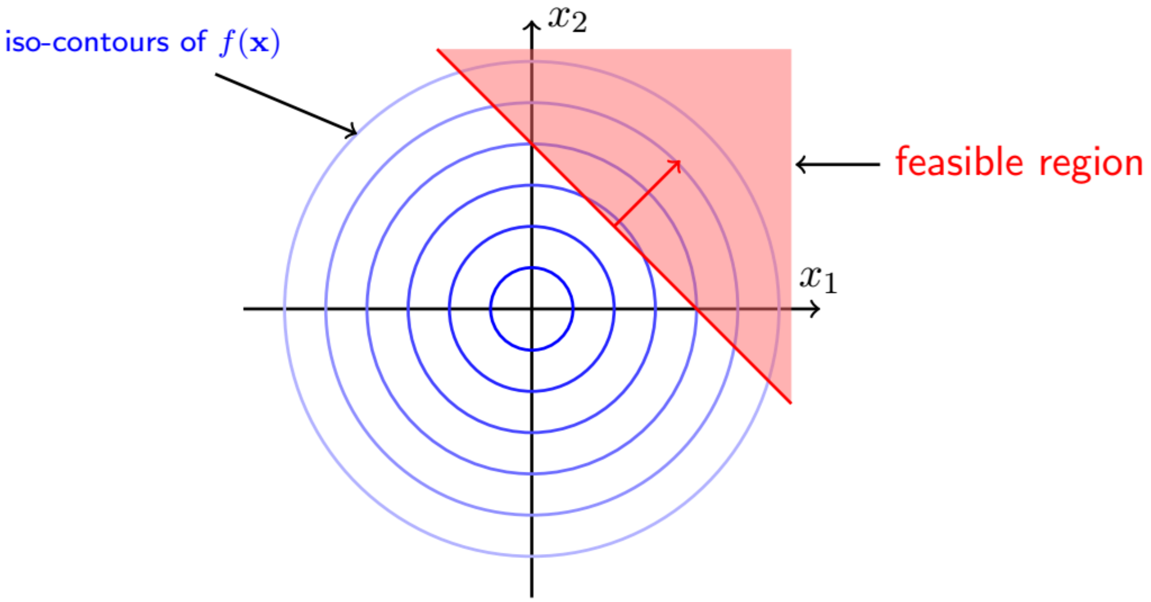
$$\begin{aligned} \underset{\boldsymbol{x} \in \mathbb{R}^2}{\text{minimize}} \quad & \frac{2}{5}(x_1^2 + x_2^2) \\ \text{subject to} \quad & 2 - x_1 - x_2 \leq 0 \end{aligned}$$

1. Write down the Lagrangian and KKT conditions
2. Solve the primal using the KKT conditions
3. Argue that strong duality holds i.e. $p^* = d^*$.
4. Write down the Lagrangian dual function and dual optimization problem
5. Solve the problem a second way, by first solving the dual problem and then inferring the primal solution from the dual solution.

Example

Consider the following optimization problem:

$$\begin{aligned} \underset{\boldsymbol{x}\in\mathbb{R}^2}{\text{minimize}} \quad & \frac{2}{5}(x_1^2 + x_2^2) \\ \text{subject to} \quad & 2 - x_1 - x_2 \leq 0 \end{aligned}$$



$$\boxed{g(\mathbf{x}) = 2 - x_1 - x_2 \leq 0}$$

Example

1. Write down the Lagrangian and KKT conditions

$$L(\boldsymbol{x}, \lambda) = \frac{2}{5}\left(x_1^2 + x_2^2\right) + \lambda(2 - x_1 - x_2)$$

1. $\frac{\partial L}{\partial x_1} = \frac{4}{5}x_1 - \lambda = 0, \frac{\partial L}{\partial x_2} = \frac{4}{5}x_2 - \lambda = 0$

2. $2 - x_1 - x_2 \leq 0$

3. N/A

4. $\lambda \geq 0$

5. $\lambda(2 - x_1 - x_2) = 0$

Example

2. Solve the primal using the KKT conditions

From the gradient conditions, if $\lambda = 0$, then $\boldsymbol{x} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. This is not feasible since we require $2 - x_1 - x_2 \leq 0$.
Thus by condition 5, this means that $2 - x_1 - x_2 = 0$.
By condition 1, $x_1 = x_2 = \frac{5}{4}\lambda$.
Therefore $2 - \frac{5}{2}\lambda = 0 \Rightarrow \lambda = \frac{4}{5}$.

$$\Rightarrow x_1 = x_2 = 1$$

Example

3. Argue that strong duality holds i.e. $p^* = d^*$.

The primal is convex and g_1 is affine.

Example

4. Write down the Lagrangian dual function and dual optimization problem

$$
\begin{aligned}
L_D(\lambda) &= \min_{\boldsymbol{x}} L(\boldsymbol{x}, \lambda) \\
&= \min_{x} \frac{2}{5}\left(x_1^2 + x_2^2\right) + \lambda(2 - x_1 - x_2) \\
&= \frac{2}{5}\left(\left(\frac{5}{4}\lambda\right)^2 + \left(\frac{5}{4}\lambda\right)^2\right) + \lambda\left(2 - \frac{5}{2}\lambda\right) \\
&= \frac{5}{4}\lambda^2 + 2\lambda - \frac{5}{2}\lambda^2 \\
&= -\frac{5}{4}\lambda^2 + 2\lambda
\end{aligned}
$$

Example

5. Solve the problem a second way, by first solving the dual problem and then inferring the primal solution from the dual solution.

Dual: $\max_{\lambda \geq 0} -\frac{5}{4}\lambda^2 + 2\lambda \Rightarrow \lambda^* = \frac{4}{5}$

To recover $\boldsymbol{x}^* = \arg\min L(\boldsymbol{x}, \lambda^*) = [1\ 1]^T$

Multinomial MLE

- Can use the KKT conditions to apply MLE to estimate the probabilities for data with a multinomial distribution.

Multinomial MLE

Let $p_1, \dots, p_k$ be a discrete pmf. N observations, $n_i = $ # of times outcome i was observed.

Then $n_1 + n_2 + \cdots + n_k = N$. What is the MLE of $\boldsymbol{\theta} = [p_1, \dots, p_k]^T$?

$$\max_{\boldsymbol{\theta}} \binom{N}{n_1! \dots n_k!} p_1^{n_1} \dots p_k^{n_k}$$
$$s.t.\ p_i \geq 0$$
$$\sum_i p_i = 1$$

Multinomial MLE

- Solution: $\hat{p}_i = \frac{n_k}{N}$

- Can solve using the Lagrange multipliers.

- Trick: Ignore inequality constraints $(p_i \geq 0)$ and show that the solution of the resulting problem satisfies the constraints anyway.

Further reading

- ESL Sections 4.5.2, 12.2.1