SPR day 04 B

Lagrange multipliers for solving equality constraints

Reading Bishop PRML Appendix E

Outline:

(1) Recipe and Example

(2) Why it works

Optimization with Equality Constraints

Given an objective function f

$$\mathbb{R}^D \rightarrow \boxed{f} \rightarrow \mathbb{R}$$

$$\theta = [\theta_1, \theta_2, \ldots \theta_D] \qquad f(\theta)$$

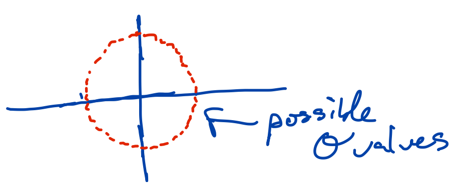
Goal: <u>Equality Constrained optimal value</u>

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \; f(\theta)$$

subject to $g(\theta) = 0$

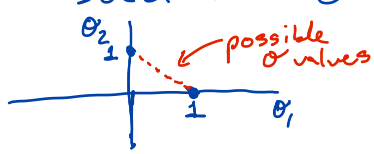
<u>Examples</u>

Find θ to maximize f, such that θ lies on <u>unit circle</u>



$$g(\theta) = 1 - \theta_1^2 - \theta_2^2$$

Find θ to maximize f, such that θ <u>sums to one</u>



$$g(\theta) = 1 - \sum_{d=1}^{D} \theta_d$$

<u>Recipe</u>

$\boxed{\text{Step 1}}$ Define objective f and constraint g

$\boxed{\text{Step 2}}$ Define expanded objective, with new "multiplier" $\lambda \neq 0$

$$\mathcal{L}(\theta, \lambda) = f(\theta) + \lambda g(\theta)$$

$\boxed{\text{Step 3}}$ calculus — Setup system of $D+1$ equations, where each partial derivative is zero

$$\frac{\partial}{\partial \theta_1}\mathcal{L} = 0, \quad \frac{\partial}{\partial \theta_2}\mathcal{L} = 0, \quad \ldots \quad \frac{\partial}{\partial \theta_D}\mathcal{L} = 0, \quad \frac{\partial}{\partial \lambda}\mathcal{L} = 0$$

$\boxed{\text{Step 4}}$ algebra — <u>Solve</u> the system of equations to find optimal values of $\theta_1^*, \theta_2^*, \ldots \theta_D^*, \lambda^*$

Example: Quadratic objective w/ linear constraint.

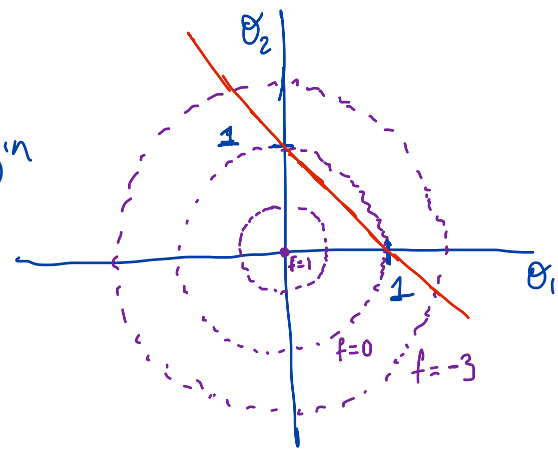
Step 1

$$\theta = [\theta_1, \theta_2]$$

$$f(\theta) = 1 - \theta_1^2 - \theta_2^2 \qquad g(\theta) = \theta_1 + \theta_2 - 1$$

line with slope -1
intercept $+1$

Intuitively,
f penalizes distance from origin



$\theta^* = [\frac{1}{2}, \frac{1}{2}]$

closest point on line to origin

Step 2

$$\mathcal{L}(\theta, \lambda) = 1 - \theta_1^2 - \theta_2^2 + \lambda(\theta_1 + \theta_2 - 1)$$

Step 3

$$0 = \frac{\partial \mathcal{L}}{\partial \theta_1} = -2\theta_1 + \lambda \qquad (1)$$

$$0 = \frac{\partial \mathcal{L}}{\partial \theta_2} = -2\theta_2 + \lambda \qquad (2)$$

$$0 = \frac{\partial \mathcal{L}}{\partial \lambda} = \theta_1 + \theta_2 - 1 \qquad (3)$$

Step 4

Using (1), we know $\lambda = 2\theta_1$

plugging in (2), we get $0 = -\cancel{2\theta_2} + \overbrace{2\theta_1}^{=\lambda}$ (2)

adding that to **2** times (3): $0 = +\cancel{2\theta_2} + 2\theta_1 - 2$ $2 \cdot (3)$

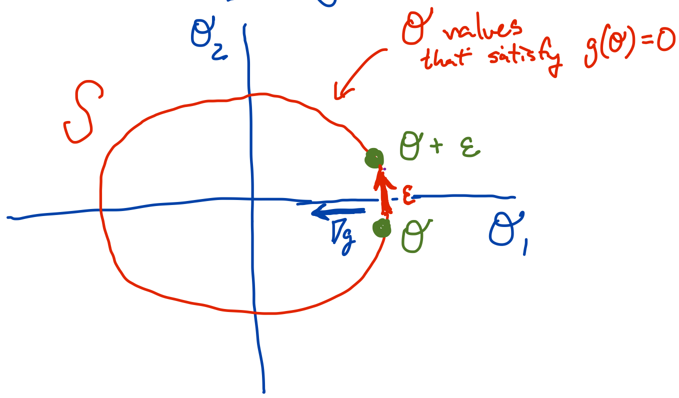
$$0 = 4\theta_1 - 2 \Rightarrow \theta_1 = \frac{1}{2}$$

$$0 = -2\theta_2 + 2(\tfrac{1}{2}) \Rightarrow \theta_2 = \frac{1}{2}$$

thus, $\theta^* = [\frac{1}{2}, \frac{1}{2}]$

Why does Lagrange multiplier recipe work?

Consider the subset S of possible θ values that satisfy the constraint



Lemma 1: Any θ that satisfies $g(\theta)=0$, will have gradient perpendicular to constraint surface.

$$\nabla_\theta g \perp \varepsilon$$

Examine two "close" points that satisfy. Call these points θ and $\theta + \varepsilon$

We know
$$g(\theta)=0 \quad (1)$$
$$g(\theta+\varepsilon)=0$$

Taylor's theorem says
$$g(\theta+\varepsilon) \approx g(\theta) + \varepsilon^T \nabla_\theta g(\theta) \quad (2)$$

Plugging (1) into (2), we have
$$0 = \varepsilon^T \nabla_\theta g(\theta)$$

Vectors have zero dot product if and only if they are perpendicular!

Thus, since vector ε goes along contour, vector ∇g must be perpendicular.

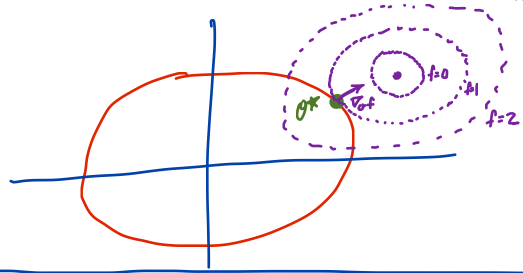
Lemma 2: At optima θ^*, the gradient of the objective f will be perpendicular to constraint surface

$$\nabla_\theta f \Big|_{\theta=\theta^*} \perp \varepsilon$$

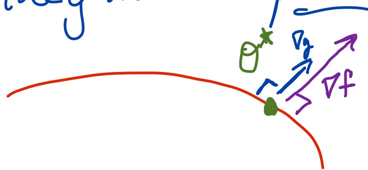
Argument: by Contradiction

Suppose ∇f was not orthogonal at θ^* then we could move locally along constraint surface S and improve f.

Thus, θ^* is not a local optima of f.



Punchline: At optima θ^*, $\nabla_\theta f$ and $\nabla_\theta g$ are orthogonal to ε, thus they must be parallel to each other.



If $\nabla f, \nabla g$ are parallel, then $\exists$ some $\lambda \neq 0$ s.t.

$$\nabla f + \lambda \nabla g = 0$$

Why it works (cont'd)?

Thus, any optimal value $\theta^\star$ must satisfy these D+1 equations

$\theta^\star$ must be optimal

$$\vec{\nabla}_f + \lambda \vec{\nabla}_g = \vec{0} \quad \rightarrow \quad \begin{bmatrix} \frac{\partial}{\partial \theta_1} f \\ \vdots \\ \frac{\partial}{\partial \theta_D} f \end{bmatrix} + \lambda \begin{bmatrix} \frac{\partial}{\partial \theta_1} g \\ \\ \frac{\partial}{\partial \theta_D} g \end{bmatrix} \begin{matrix} = 0 \\ \\ = 0 \end{matrix} \quad \begin{matrix} (1) \\ (2) \\ (D) \end{matrix}$$

$\theta^\star$ must satisfy constraints

$$g(\theta^\star) = 0 \quad \rightarrow \quad \frac{\partial}{\partial \lambda}\left[\lambda\, g(\theta)\right] = 0 \quad (D+1)$$

Can view the whole system as partial derivatives of expanded objective

$$\mathcal{L}(\theta, \lambda) = f(\theta) + \lambda g(\theta)$$

where $\theta \in \mathbb{R}^D$, $\lambda \neq 0$ are unknown, to be solved for.