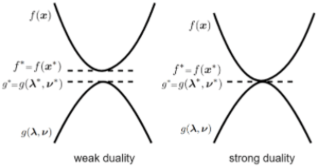
Optimization in Machine Learning

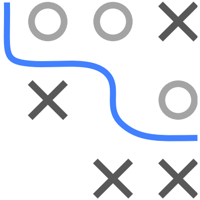
Nonlinear programs and Lagrangian



Learning goals

- Lagrangian for general constrained optimization
- Geometric intuition for Lagrangian duality
- Properties and examples

NONLINEAR CONSTRAINED OPTIMIZATION



Previous lecture: **Linear programs**

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^d} \quad & f(\mathbf{x}) := \mathbf{c}^\top \mathbf{x} \\ \text{s.t.} \quad & \mathbf{A}\mathbf{x} \leq \mathbf{b} \\ & \mathbf{G}\mathbf{x} = \mathbf{h} \end{aligned}$$

Related to its (Lagrange) dual formulation by the *Lagrangian*

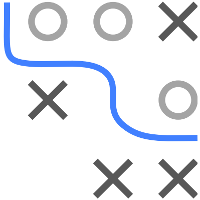
$$\mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \mathbf{c}^\top \mathbf{x} + \boldsymbol{\alpha}^\top (\mathbf{A}\mathbf{x} - \mathbf{b}) + \boldsymbol{\beta}^\top (\mathbf{G}\mathbf{x} - \mathbf{h}).$$

Weak duality: For $\boldsymbol{\alpha} \geq 0$ and $\boldsymbol{\beta}$:

$$f(\mathbf{x}^*) \geq \min_{\mathbf{x}\in\mathcal{S}} \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \geq \min_{\mathbf{x}\in\mathbb{R}^d} \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) =: g(\boldsymbol{\alpha}, \boldsymbol{\beta})$$

Recall: Implicit domain constraint in *Lagrange dual function* $g(\boldsymbol{\alpha}, \boldsymbol{\beta})$.

NONLINEAR CONSTRAINED OPTIMIZATION / 2



General form of a constraint optimization problem

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

- Functions f, g_i, h_j generally nonlinear
- Transfer the Lagrangian from linear programs:

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) := f(\mathbf{x}) + \sum_{i=1}^{k} \alpha_i g_i(\mathbf{x}) + \sum_{j=1}^{\ell} \beta_j h_j(\mathbf{x})$$

- Dual variables $\alpha_i \geq 0$ and β_i are also called *Lagrange multipliers*.

CONSTRAINED PROBLEMS: THE DIRECT WAY

Simple constrained problems can be solved in a direct way.

Example 1:

$$
\begin{array}{ll}
\min\limits_{x \in \mathbb{R}} & 2 - x^2 \\
\text{s.t.} & x - 1 = 0
\end{array}
$$

Solution: Resolve the constraint by

$$
\begin{aligned}
x - 1 &= 0 \\
x &= 1
\end{aligned}
$$

and insert it into the objective:

$$x^* = 1, \quad f(x^*) = 1$$

CONSTRAINED PROBLEMS: THE DIRECT WAY / 2

Example 2:

$$
\begin{aligned}
\min_{\mathbf{x}\in\mathbb{R}^2} \quad & -2 + x_1^2 + 2x_2^2 \\
\text{s.t.} \quad & x_1^2 + x_2^2 - 1 = 0
\end{aligned}
$$

Solution: Resolve the constraint

$$x_1^2 = 1 - x_2^2$$

and insert it into the objective

$$
\begin{aligned}
f(x_1, x_2) &= -2 + (1 - x_2^2) + 2x_2^2 \\
&= -1 + x_2^2.
\end{aligned}
$$

$\Rightarrow$ Minimum at $\mathbf{x}^* = (\pm 1, 0)^\top$. However, direct way mostly not possible.

A CLASSIC EXAMPLE: "MILKMAID PROBLEM"

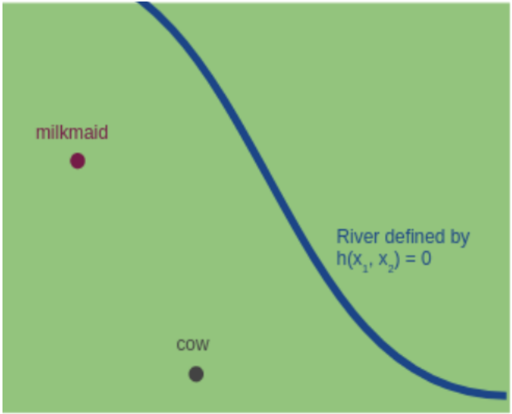
Question 1: Is there a general recipe for solving general constrained nonlinear optimization problems?
Question 2: Can we understand this recipe geometrically?
Question 3: How does this relate to the Lagrange function approach?

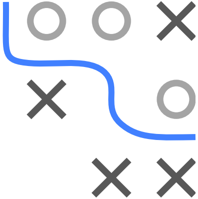
For this purpose, we consider the classical “milkmaid problem”; the following example is taken from *Steuard Jensen, An Introduction to Lagrange Multipliers* (but the example works of course equally well with a “milk man”).

- Assume a milk maid is sent to the field to get the day’s milk
- The milkmaid wants to finish her job as quickly as possible
- However, she has to clean her bucket first at the nearby river.

A CLASSIC EXAMPLE: "MILKMAID PROBLEM" / 2

Where is the best point *P* to clean her bucket?

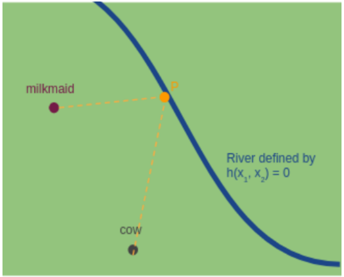




A CLASSIC EXAMPLE: "MILKMAID PROBLEM" / 3

Aim: Find point P at the river for minimum total distance $f(P)$

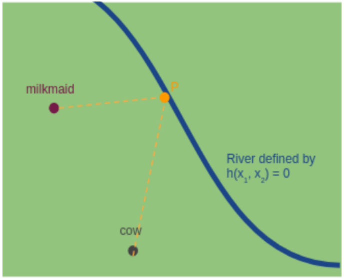
- $f(P) := d(M, P) + d(P, C)$ (d measures distance)
- $h(P) = 0$ describes the river



A CLASSIC EXAMPLE: "MILKMAID PROBLEM" / 4

Corresponding optimization problem:

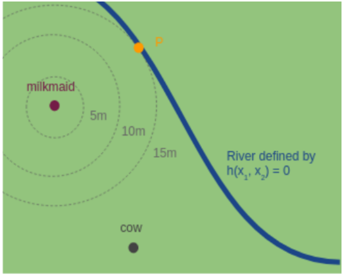
$$\begin{aligned} \min_{x_1, x_2} \quad & f(x_1, x_2) \\ \text{s.t.} \quad & h(x_1, x_2) = 0 \end{aligned}$$



A CLASSIC EXAMPLE: "MILKMAID PROBLEM" / 5

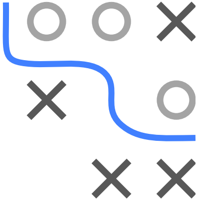
Question: How far can the milkmaid get for a fixed total distance $f(P)$?

Assume: We only care about $d(M, P)$.



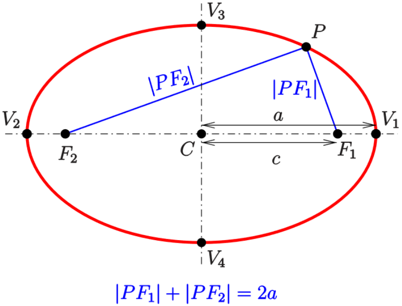
Observe: Radius of circle touching river first is the minimal distance.

A CLASSIC EXAMPLE: "MILKMAID PROBLEM" / 6



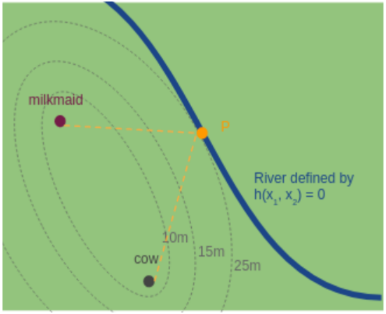
- For $f(P) = d(M, P) + d(P, C)$: Use an **ellipse**.
- **Def.:** Given two focal points F_1, F_2 and distance $2a$:

$$E = \{P \in \mathbb{R}^2 \mid d(F_1, P) + d(P, F_2) = 2a\}$$



A CLASSIC EXAMPLE: "MILKMAID PROBLEM" / 7

- Let M and C be focal points of a collection of ellipses
- Find **smallest** ellipse touching the river $h(x_1, x_2)$
- Define P as the touching point



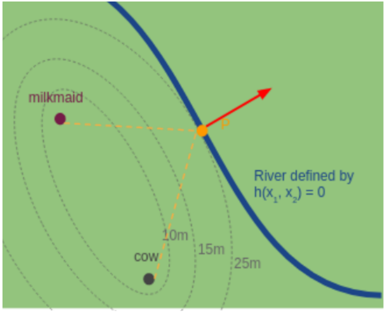
Question: How can we make sense of this mathematically?

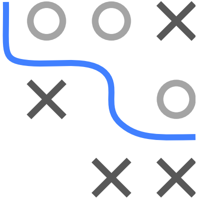
A CLASSIC EXAMPLE: "MILKMAID PROBLEM" / 8

- **Recall:** Gradient is normal (perpendicular) to contour lines
- Since contour lines of f and h touch, gradients are parallel:

$$\nabla f(P) = \beta \nabla h(P)$$

- Multiplier β is called **Lagrange multiplier**.





LAGRANGE FUNCTION

General: Solve problem with single equality constraint by:

$$\nabla f(\mathbf{x}) = \beta \nabla h(\mathbf{x})$$
$$h(\mathbf{x}) = 0$$

- **First line**: Parallel gradients | **Second line:** Constraint

Observe: Above system is equivalent to

$$\nabla \mathcal{L}(\mathbf{x}, \beta) = \mathbf{0}$$

for **Lagrange function** (or **Lagrangian**) $\mathcal{L}(\mathbf{x}, \beta) := f(\mathbf{x}) + \beta h(\mathbf{x})$

Indeed:

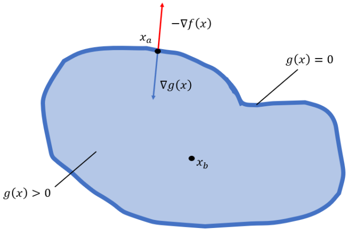
$$\begin{pmatrix} \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \beta) \\ \nabla_{\beta} \mathcal{L}(\mathbf{x}, \beta) \end{pmatrix} = \begin{pmatrix} \nabla f(\mathbf{x}) + \beta \nabla h(\mathbf{x}) \\ h(\mathbf{x}) \end{pmatrix}$$

LAGRANGE FUNCTION / 2

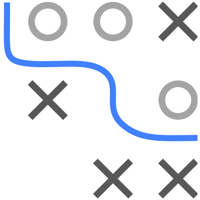
Idea can be extended to **inequality** constraints $g(\mathbf{x}) \leq 0$.

There are two possible cases for a solution:

- Solution $\mathbf{x}_b$ inside feasible set: constraint is inactive ($\alpha_b = 0$)
- Solution $\mathbf{x}_a$ on boundary $g(\mathbf{x}) = 0$: $\nabla f(\mathbf{x}_a) = \alpha_a \nabla g(\mathbf{x}_a)$ ($\alpha_a > 0$)



LAGRANGE FUNCTION AND PRIMAL PROBLEM



General constrained optimization problems:

$$
\begin{aligned}
\min_{\mathbf{x}} \quad & f(\mathbf{x}) \\
\text{s.t.} \quad & g_i(\mathbf{x}) \leq 0, \quad i = 1, \dots, k \\
& h_j(\mathbf{x}) = 0, \quad j = 1, \dots, \ell
\end{aligned}
$$

Extend Lagrangian ($\alpha_i \geq 0$, β_i Lagrange multipliers):

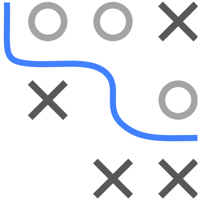
$$
\mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) := f(\mathbf{x}) + \sum_{i=1}^{k} \alpha_i g_i(\mathbf{x}) + \sum_{j=1}^{\ell} \beta_j h_j(\mathbf{x})
$$

Equivalent primal problem:

$$
\min_{\mathbf{x}} \max_{\boldsymbol{\alpha} \geq 0, \boldsymbol{\beta}} \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})
$$

Question: Why?

LAGRANGE FUNCTION AND PRIMAL PROBLEM / 2



For simplicity: Consider only single inequality constraint $g(\mathbf{x}) \leq 0$

If $\mathbf{x}$ **breaks** inequality constraint ($g(\mathbf{x}) > 0$):

$$\max_{\alpha \geq 0} \mathcal{L}(\mathbf{x}, \alpha) = \max_{\alpha \geq 0} f(\mathbf{x}) + \alpha g(\mathbf{x}) = \infty$$

If $\mathbf{x}$ **satisfies** inequality constraint ($g(\mathbf{x}) \leq 0$):

$$\max_{\alpha \geq 0} \mathcal{L}(\mathbf{x}, \alpha) = \max_{\alpha \geq 0} f(\mathbf{x}) + \alpha g(\mathbf{x}) = f(\mathbf{x})$$

Combining yields **original formulation**:

$$\min_{\mathbf{x}} \max_{\alpha \geq 0} \mathcal{L}(\mathbf{x}, \alpha) = \begin{cases} \infty & \text{if } g(\mathbf{x}) > 0 \\ \min_{\mathbf{x}} f(\mathbf{x}) & \text{if } g(\mathbf{x}) \leq 0 \end{cases}$$

Similar argument holds for equality constraints $h_j(\mathbf{x})$

EXAMPLE: LAGRANGE FUNCTION FOR QP'S

We consider quadratic programming

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) := \frac{1}{2}\mathbf{x}^\top \mathbf{Q}\mathbf{x} \\ \text{s.t.} \quad & h(\mathbf{x}) := \mathbf{C}\mathbf{x} - \mathbf{d} = \mathbf{0} \end{aligned}$$

with $\mathbf{Q} \in \mathbb{R}^{d \times d}$ symmetric, $\mathbf{C} \in \mathbb{R}^{\ell \times d}$, and $\mathbf{d} \in \mathbb{R}^{\ell}$.

Lagrange function: $\mathcal{L}(\mathbf{x}, \boldsymbol{\beta}) = \frac{1}{2}\mathbf{x}^\top \mathbf{Q}\mathbf{x} + \boldsymbol{\beta}^\top (\mathbf{C}\mathbf{x} - \mathbf{d})$

Solve

$$\begin{aligned} \nabla \mathcal{L}(\mathbf{x}, \boldsymbol{\beta}) = \begin{pmatrix} \partial \mathcal{L} / \partial \mathbf{x} \\ \partial \mathcal{L} / \partial \boldsymbol{\beta} \end{pmatrix} &= \begin{pmatrix} \mathbf{Q}\mathbf{x} + \mathbf{C}^\top \boldsymbol{\beta} \\ \mathbf{C}\mathbf{x} - \mathbf{d} \end{pmatrix} = \mathbf{0} \\ \Leftrightarrow \qquad \begin{pmatrix} \mathbf{Q} & \mathbf{C}^\top \\ \mathbf{C} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \boldsymbol{\beta} \end{pmatrix} &= \begin{pmatrix} \mathbf{0} \\ \mathbf{d} \end{pmatrix} \end{aligned}$$

Observe: Solve QP by solving a linear system

LAGRANGE DUALITY

Dual problem:

$$\max_{\boldsymbol{\alpha} \geq 0, \boldsymbol{\beta}} \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})$$

Define **Lagrange dual function** $g(\boldsymbol{\alpha}, \boldsymbol{\beta}) := \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})$

Important characteristics of the dual problem:

- **Convexity** (pointwise minimum of *affine* functions)
  - Gives methods based on dual solutions
  - Might be computationally inefficient (expensive minimizations)
- **Weak duality:**

$$f(\mathbf{x}^*) \geq g(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$$

- **Strong duality** if primal problem satisfies *Slater's condition*[1]:

$$f(\mathbf{x}^*) = g(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$$

[1] **Slater's condition**: Primal problem convex and "strictly feasible" ($\exists \mathbf{x} \forall i : g_i(\mathbf{x}) < 0$).