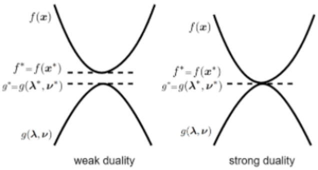
Optimization in Machine Learning

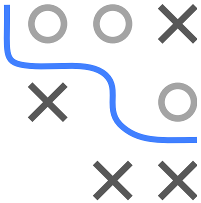
Constrained Optimization
Nonlinear programs and Lagrangian



Learning goals

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

NONLINEAR CONSTRAINED OPTIMIZATION



- Previous lecture: **Linear programs**

$$\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) := \mathbf{c}^T \mathbf{x} \quad \text{s.t. } \mathbf{A}\mathbf{x} \leq \mathbf{b}, \;\; \mathbf{G}\mathbf{x} = \mathbf{h}$$

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

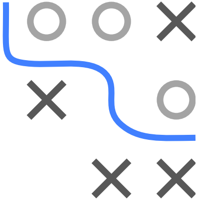
$$\mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \mathbf{c}^T \mathbf{x} + \boldsymbol{\alpha}^T (\mathbf{A}\mathbf{x} - \mathbf{b}) + \boldsymbol{\beta}^T (\mathbf{G}\mathbf{x} - \mathbf{h})$$

- **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* $g(\boldsymbol{\alpha}, \boldsymbol{\beta})$

GENERAL CONSTRAINED OPTIMIZATION



- General form of a constrained optimization problem

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

- 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:

$$\min_{x \in \mathbb{R}} 2 - x^2 \quad \text{s.t. } x - 1 = 0$$

- **Solution:** Resolve the constraint by $x - 1 = 0 \Rightarrow x = 1$ and insert it into the objective: $x^* = 1, \; f(x^*) = 1$

CONSTRAINED PROBLEMS: THE DIRECT WAY

Example 2:

$$\min_{\boldsymbol{x}\in\mathbb{R}^2} -2 + x_1^2 + 2x_2^2 \quad \text{s.t. } x_1^2 + x_2^2 - 1 = 0$$

- **Solution:** Resolve the constraint $x_1^2 = 1 - x_2^2$ and insert it into the objective

$$f(x_1, x_2) = -2 + (1 - x_2^2) + 2x_2^2 = -1 + x_2^2$$

- $\Rightarrow$ Minimum at $\boldsymbol{x}^* = (\pm 1, 0)^T$
- However, direct way mostly not possible

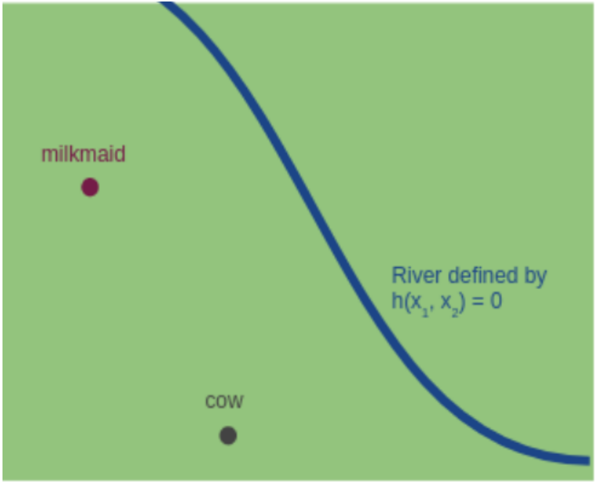
A CLASSIC EXAMPLE: MILKMAID PROBLEM

- **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”
- Assume a milkmaid 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

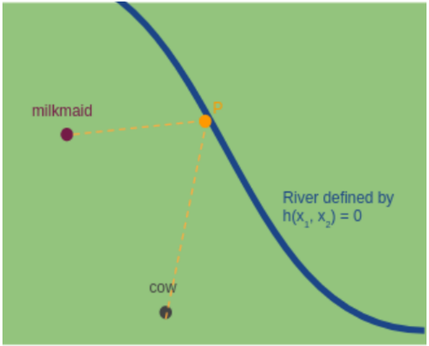
MILKMAID PROBLEM

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

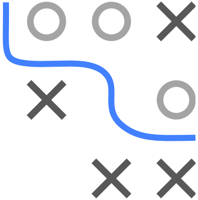


MILKMAID PROBLEM

- **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

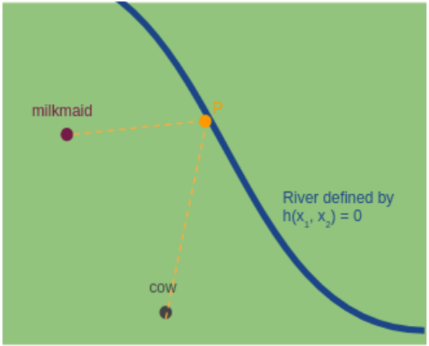


MILKMAID PROBLEM

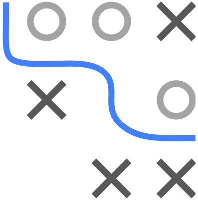


- Corresponding optimization problem:

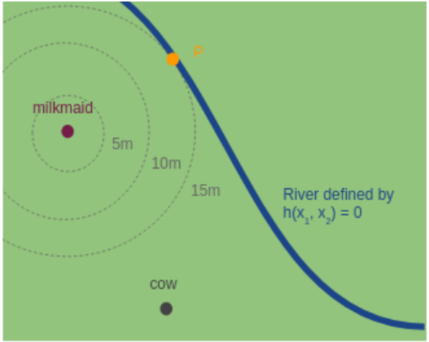
$$\min_{x_1, x_2} f(x_1, x_2) \quad \text{s.t. } h(x_1, x_2) = 0$$



MILKMAID PROBLEM



- **Question:** How far can the milkmaid get for a fixed total distance $f(P)$?
- **Assume:** We only care about $d(M, P)$

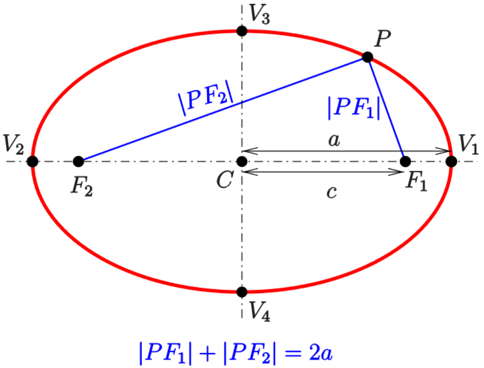


- **Observe:** Radius of circle touching river first is the minimal distance

MILKMAID PROBLEM

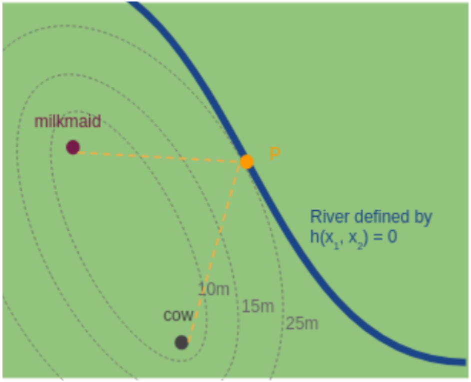
- For $f(P) = d(M, P) + d(P, C)$: Use an **ellipse**
- **Def.:** Given two focal points F_1, F_2 and distance $2a$:

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



MILKMAID PROBLEM

- 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



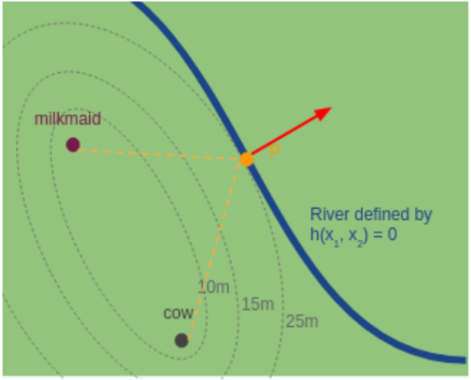
- **Question:** How can we make sense of this mathematically?

MILKMAID PROBLEM

- **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:

$$\nabla f(\boldsymbol{x}) = \beta \nabla h(\boldsymbol{x}) \quad \text{and} \quad h(\boldsymbol{x}) = 0$$

- **First line**: Parallel gradients | **Second line:** Constraint
- **Observe:** Above system is equivalent to

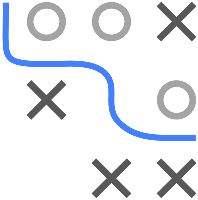
$$\nabla \mathcal{L}(\boldsymbol{x}, \beta) = \mathbf{0}$$

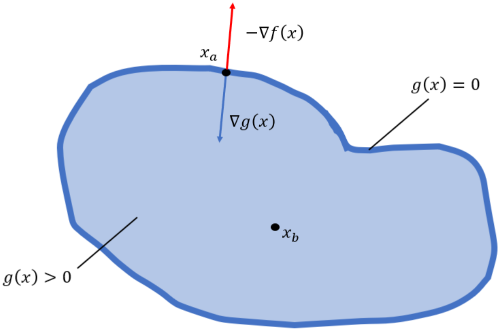
  for **Lagrange function** (or **Lagrangian**) $\mathcal{L}(\boldsymbol{x}, \beta) := f(\boldsymbol{x}) + \beta h(\boldsymbol{x})$
- Indeed:

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

LAGRANGE FUNCTION

- Idea can be extended to **inequality** constraints $g(\boldsymbol{x}) \leq 0$
- There are two possible cases for a solution:
  - Solution $\boldsymbol{x}_b$ inside feasible set: constraint is inactive ($\alpha_b = 0$)
  - Solution $\boldsymbol{x}_a$ on boundary $g(\boldsymbol{x}) = 0$: $\nabla f(\boldsymbol{x}_a) = \alpha_a \nabla g(\boldsymbol{x}_a)$ ($\alpha_a > 0$)





LAGRANGE FUNCTION AND PRIMAL PROBLEM

- General constrained optimization problems:

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

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

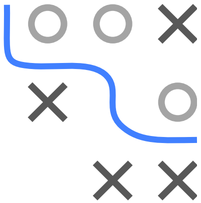
$$\mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) := f(\mathbf{x}) + \sum_{i=1}^{k} \alpha_i g_i(\mathbf{x}) + \sum_{j=1}^{\ell} \beta_j h_j(\mathbf{x})$$

- **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



- For simplicity: Consider only single inequality constraint $g(\boldsymbol{x}) \leq 0$
- If $\boldsymbol{x}$ **breaks** inequality constraint ($g(\boldsymbol{x}) > 0$):

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

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

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

- Combining yields **original formulation**:

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

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

EXAMPLE: LAGRANGE FUNCTION FOR QPS

- We consider quadratic programming

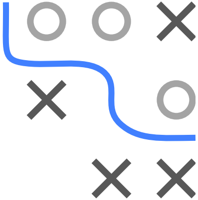
$$\min_{\mathbf{x}} f(\mathbf{x}) := \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} \quad \text{s.t. } h(\mathbf{x}) := \mathbf{C}\mathbf{x} - \mathbf{d} = \mathbf{0}$$

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

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

- **Observe:** Solve QP by solving a linear system

LAGRANGE DUALITY



- **Dual problem:**

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

- Define **Lagrange dual function** $g(\boldsymbol{\alpha}, \boldsymbol{\beta}) := \min_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})$
- 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:** $f(\boldsymbol{x}^*) \geq g(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$
- **Strong duality** if primal problem satisfies *Slater's condition*[(1)]:
  $f(\boldsymbol{x}^*) = g(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$

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