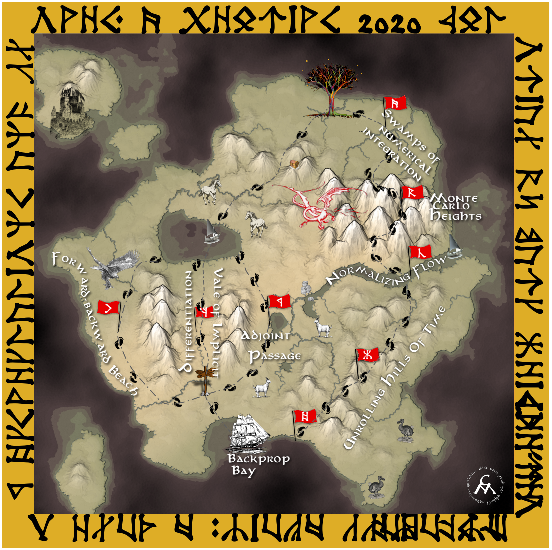
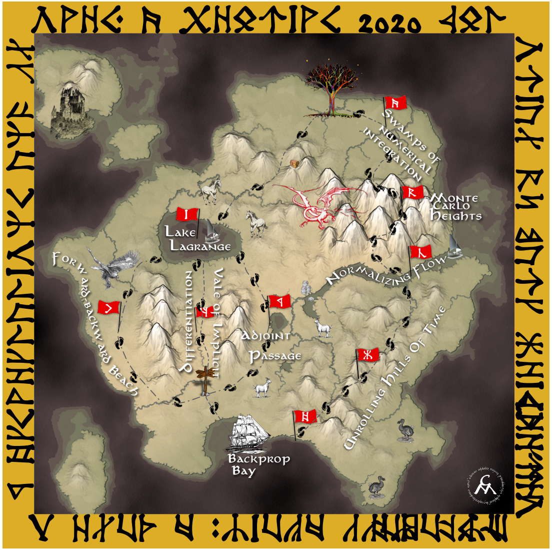

2020
Swamps of Numerical Integration
Monte Carlo Heights
Normalizing Flow
Unrolling Hills of Time
Backprop Bay
Adjoint Passage
Vale of Implicit Differentiation
Forward-backward Beach


Swamps of Numerical Integration
Monte Carlo Heights
Lake Lagrange
Normalizing Flow
Forward-backward Beach
Differentiation
Vale of Implicit
Adjoint Passage
Unrolling Hills of Time
Backprop Bay

Method of Lagrange

Cheng Soon Ong
Marc Peter Deisenroth

December 2020

Motivation

- In machine learning, we use gradients to train predictors
- For functions $f(\boldsymbol{x})$ we can directly obtain its gradient $\nabla f(\boldsymbol{x})$
- We may wish to calculate a gradient for functions with constraints
- Assume that we can calculate gradient for corresponding unconstrained problem

Key idea

How to take a gradient with respect to a constrained optimization problem.

History

- Joseph-Louis Lagrange (b. 1736) worked on calculus of variations and founded the idea
- Nonlinear programming was coined by Harold W. Kuhn and Albert W. Tucker in 1951.
- Turns out William Karush already wrote about similar conditions in his master's thesis in 1939, and was reintroduced into the literature by Akira Takayama in 1974.
- Classic backpropagation paper, by Yann LeCun in 1988, observed the relation between backpropagation and Lagrange multipliers.

Constrained optimization

Primal optimization problem

Given $f : \mathbb{R}^D \to \mathbb{R}$ and $g_i : \mathbb{R}^D \to \mathbb{R}$ for $i = 1, \ldots, m$,

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & g_i(\boldsymbol{x}) \leqslant 0 \quad \text{for all} \quad i = 1, \ldots, m \end{aligned}$$

Lagrange multipliers as relaxation

We can convert the primal optimization problem to an unconstrained problem.

$$J(\boldsymbol{x}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \mathbf{1}(g_i(\boldsymbol{x}))\,,$$

where $\mathbf{1}(z)$ is an infinite step function

$$\mathbf{1}(z) = \begin{cases} 0 & \text{if } z \leqslant 0 \\ \infty & \text{otherwise} \end{cases}.$$

Key idea

Approximate indicator function by linear function.

Lagrangian

We associate the primal optimization problem with a **Lagrangian**, by introducing Lagrange multipliers λ_i for each constraint g_i.

Lagrangian

$$\begin{aligned}\mathfrak{L}(\boldsymbol{x}, \boldsymbol{\lambda}) &= f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i g_i(\boldsymbol{x}) \\ &= f(\boldsymbol{x}) + \boldsymbol{\lambda}^\top \boldsymbol{g}(\boldsymbol{x})\end{aligned}$$

Why Lagrangian?

- $\min_{\boldsymbol{x} \in \mathbb{R}^d} \mathfrak{L}(\boldsymbol{x}, \boldsymbol{\lambda})$ is an unconstrained optimization problem for a given value of $\boldsymbol{\lambda}$
- Turns out: If solving $\min_{\boldsymbol{x} \in \mathbb{R}^d} \mathfrak{L}(\boldsymbol{x}, \boldsymbol{\lambda})$ is easy, then the overall problem is easy

Dual optimization problem

Key idea

Convert problem in $\boldsymbol{x} \in \mathbb{R}^D$, to another problem in $\boldsymbol{\lambda} \in \mathbb{R}^m$.

- Primal and dual problems are related
- The objective values at optimality related via gradients
- the dual problem (maximization over $\boldsymbol{\lambda}$) is a maximum over a set of affine functions, and hence is a concave function, even though $f(\cdot)$ and $g_i(\cdot)$ may be non-convex

Primal dual pair

The problem

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & g_i(\boldsymbol{x}) \leqslant 0 \quad \text{for all} \quad i = 1, \ldots, m \end{aligned}$$

is known as the **primal problem**, corresponding to the primal variables $\boldsymbol{x}$. The associated **Lagrangian dual problem** is given by

$$\begin{aligned} \max_{\boldsymbol{\lambda} \in \mathbb{R}^m} \quad & \mathfrak{D}(\boldsymbol{\lambda}) \\ \text{subject to} \quad & \boldsymbol{\lambda} \geqslant \mathbf{0}\,, \end{aligned}$$

where $\boldsymbol{\lambda}$ are the dual variables and $\mathfrak{D}(\boldsymbol{\lambda}) = \min_{\boldsymbol{x} \in \mathbb{R}^d} \mathfrak{L}(\boldsymbol{x}, \boldsymbol{\lambda})$.

Minimax inequality

Why Lagrangian duality is useful? Because **primal values are greater than dual values**.

Key idea

For any function with two arguments, the maximin is less than the minimax.

$$\max_{\boldsymbol{y}} \min_{\boldsymbol{x}} \varphi(\boldsymbol{x}, \boldsymbol{y}) \leqslant \min_{\boldsymbol{x}} \max_{\boldsymbol{y}} \varphi(\boldsymbol{x}, \boldsymbol{y}) \,.$$

Weak duality

Key idea

Primal values $f(\boldsymbol{x})$ are always greater than (or equal to) dual values $\mathfrak{D}(\boldsymbol{\lambda})$

We can see this by applying the minimax inequality

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} \underbrace{\max_{\boldsymbol{\lambda}\geqslant\mathbf{0}} \mathfrak{L}(\boldsymbol{x},\boldsymbol{\lambda})}_{f(\boldsymbol{x})} \geqslant \max_{\boldsymbol{\lambda}\geqslant\mathbf{0}} \underbrace{\min_{\boldsymbol{x}\in\mathbb{R}^d} \mathfrak{L}(\boldsymbol{x},\boldsymbol{\lambda})}_{\mathfrak{D}(\boldsymbol{\lambda})} .$$

Motivating the Karush Kuhn Tucker conditions

Lagrangian

$$\mathfrak{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i g_i(\boldsymbol{x})$$

Stationarity

Solve for optimal value by taking gradient w.r.t. $\boldsymbol{x}$ and set to zero

Lagrange multipliers are orthogonal to constraints

- Recall that we have one Lagrange multiplier per constraint
- If $\lambda_i = 0$, then $g_i(\boldsymbol{x})$ can be any value (has slack)
- If $g_i(\boldsymbol{x}) = 0$ (active), then λ_i has slack

Retain feasibility

Karush Kuhn Tucker conditions

stationarity

$$\mathbf{0} \in \nabla f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\boldsymbol{x})\,. \tag{1}$$

complementary slackness

$$\lambda_i g_i(\boldsymbol{x}) = 0 \quad \text{for all} \quad i = 1, \ldots, m\,. \tag{2}$$

primal feasibility

$$g_i(\boldsymbol{x}) \leqslant 0 \quad \text{for all} \quad i = 1, \ldots, m\,. \tag{3}$$

dual feasibility

$$\lambda_i \geqslant 0 \quad \text{for all} \quad i = 1, \ldots, m\,. \tag{4}$$

Constrained optimization

Key idea

Objective gradient is a conic combination of active constraint gradients

By stationarity condition

$$\nabla f(\boldsymbol{x}) = -\sum_{i=1}^{m} \lambda_i \nabla g_i(\boldsymbol{x})$$

for a given value of $\boldsymbol{\lambda}$.

KKT implies strong duality

Key idea

KKT conditions are sufficient for zero duality gap

Proof sketch:

- By weak duality, $f(\boldsymbol{x}) \geqslant \mathfrak{D}(\boldsymbol{\lambda})$. Let $\boldsymbol{x}^*$ and $\boldsymbol{\lambda}^*$ denote optimal values
- By stationarity conditions

$$\mathfrak{D}(\boldsymbol{\lambda}^*) = f(\boldsymbol{x}^*) + \sum_{i=1}^{m} \lambda_i^* g_i(\boldsymbol{x}^*)\,,$$

- By complementary slackness

$$f(\boldsymbol{x}^*) + \sum_{i=1}^{m} \lambda_i^* g_i(\boldsymbol{x}^*) = f(\boldsymbol{x}^*)\,.$$

- Hence duality gap is zero

KKT implies strong duality

Key idea

Challenge is to express geometric constraints as algebraic equations

Constraint qualification additionally needed for KKT to necessarily imply zero duality gap.

- Mangasarian Fromovitz
- Linear independence
- Fritz John
- Slater's condition

Summary

Key idea

How to take a gradient with respect to a constrained optimization problem.

- Objective gradient is a conic combination of active constraint gradients

$$\nabla f(\boldsymbol{x}) = -\sum_{i=1}^{m} \lambda_i \nabla g_i(\boldsymbol{x})$$

- Check that one of Lagrange multipliers or primal constraints is zero
- The KKT conditions are sufficient for a zero duality gap