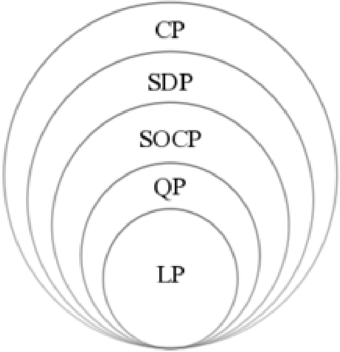
Optimization in Machine Learning

Nonlinear Programs
Solvers



Learning goals

- Sequential quadratic programming
- Penalty methods
- Barrier / interior-point methods
- Constrained optimization in R

SEQUENTIAL QUADRATIC PROGRAMMING

- For simplification, consider only equality constraints:

$$\min f(\boldsymbol{x}) \qquad \text{s.t.} \quad h(\boldsymbol{x}) = 0$$

Idea: Instead of f, optimize 2nd order Taylor approximation at point $\tilde{\mathbf{x}}$

$$\tilde{f}(\boldsymbol{x}) = f(\tilde{\mathbf{x}}) + \nabla_x f(\tilde{\mathbf{x}})^T(\boldsymbol{x} - \tilde{\mathbf{x}}) + \frac{1}{2}(\boldsymbol{x} - \tilde{\mathbf{x}})^T \nabla^2_{xx} f(\tilde{\mathbf{x}})(\boldsymbol{x} - \tilde{\mathbf{x}})$$

- h is also replaced by its linear approximation at $\tilde{\mathbf{x}}$

$$\tilde{h}(\boldsymbol{x}) = h(\tilde{\mathbf{x}}) + \nabla h(\tilde{\mathbf{x}})^T(\boldsymbol{x} - \tilde{\mathbf{x}})$$

SEQUENTIAL QUADRATIC PROGRAMMING

- With $\boldsymbol{d} := (\boldsymbol{x} - \tilde{\mathbf{x}})$ we formulate the **quadratic auxiliary problem**

$$\begin{aligned} \min_{\boldsymbol{d}} \quad & \tilde{f}(\boldsymbol{d}) := f(\tilde{\mathbf{x}}) + \boldsymbol{d}^T \nabla_x f(\tilde{\mathbf{x}}) + \frac{1}{2} \boldsymbol{d}^T \nabla^2_{xx} f(\tilde{\mathbf{x}}) \boldsymbol{d} \\ \text{s.t.} \quad & \tilde{h}(\boldsymbol{d}) := h(\tilde{\mathbf{x}}) + \nabla h(\tilde{\mathbf{x}})^T \boldsymbol{d} = 0 \end{aligned}$$

- Even if no optimality conditions can be formulated for the actual problem, KKT conditions apply at the optimum of this problem
- If $\nabla^2_{xx} f(\boldsymbol{x})$ is positive semidefinite, it is a **convex optimization problem**

SEQUENTIAL QUADRATIC PROGRAMMING

- Using the Lagrange function

$$L(\boldsymbol{d}, \beta) = \boldsymbol{d}^T \nabla_x f(\tilde{\mathbf{x}}) + \frac{1}{2} \boldsymbol{d}^T \nabla^2_{xx} f(\tilde{\mathbf{x}}) \boldsymbol{d} + \beta^T (h(\tilde{\mathbf{x}}) + \nabla h(\tilde{\mathbf{x}})^T \boldsymbol{d})$$

  we formulate the KKT conditions

  - $\nabla_{\boldsymbol{d}} L(\boldsymbol{d}, \beta) = \nabla_x f(\tilde{\mathbf{x}}) + \nabla^2_{xx} f(\tilde{\mathbf{x}}) \boldsymbol{d} + \nabla h(\tilde{\mathbf{x}})^T \beta = 0$
  - $h(\tilde{\mathbf{x}}) + \nabla h(\tilde{\mathbf{x}})^T \boldsymbol{d} = 0$

- Or in matrix notation

$$\begin{pmatrix} \nabla^2_{xx} f(\tilde{\mathbf{x}}) & \nabla h(\tilde{\mathbf{x}})^T \\ \nabla h(\tilde{\mathbf{x}}) & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{d} \\ \beta \end{pmatrix} = - \begin{pmatrix} \nabla_x f(\tilde{\mathbf{x}}) \\ h(\tilde{\mathbf{x}}) \end{pmatrix}$$

- The **quadratic subproblem** can be traced back to solving a linear system

SQP ALGORITHM

Algorithm SQP for problems with equality constraints

Select a feasible starting point $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$
while Stop criterion not fulfilled **do**
    Solve quadratic subproblem by solving the equation

$$\begin{pmatrix} \nabla^2_{xx} L(\boldsymbol{x}, \boldsymbol{\mu}) & \nabla h(\boldsymbol{x})^T \\ \nabla h(\boldsymbol{x}) & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{d} \\ \boldsymbol{\beta} \end{pmatrix} = - \begin{pmatrix} \nabla_x L(\boldsymbol{x}, \boldsymbol{\mu}) \\ h(\boldsymbol{x}) \end{pmatrix}$$

    Set $\boldsymbol{x}^{(i+1)} = \boldsymbol{x}^{(i)} + \boldsymbol{d}$
end while

PENALTY METHODS

Idea: Replace constrained problem with a sequence of unconstrained problems using a **penalty function**

- Instead of

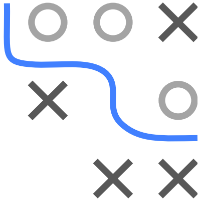
$$\min f(\boldsymbol{x}) \qquad \text{s.t.} \quad h(\boldsymbol{x}) = 0$$

  we look at the unconstrained problem

$$\min_{\boldsymbol{x}} p(\boldsymbol{x}) = f(\boldsymbol{x}) + \rho \frac{\|h(\boldsymbol{x})\|^2}{2}$$

- Under appropriate conditions, solutions for $\rho \to \infty$ converge to the solution of the original problem

BARRIER METHOD



Idea: Establish a “barrier” that penalizes if $\boldsymbol{x}$ comes too close to the edge of the feasible set $\boldsymbol{S}$

- For the problem

$$\min f(\boldsymbol{x}) \qquad \text{s.t.} \quad g(\boldsymbol{x}) \leq 0$$

a common **barrier function** is

$$B_\rho = f(\boldsymbol{x}) - \rho \sum_{i=1}^{m} \ln(-g_i(\boldsymbol{x}))$$

- Penalty term becomes larger as $\boldsymbol{x}$ approaches 0, i.e., the limit of the feasible set
- Under certain conditions, solutions of $\min B_\rho$ for $\rho \to 0$ converge to the optimum
- Also called **interior-point method**

CONSTRAINED OPTIMIZATION IN R

- `optim(..., method = "L-BFGS-B")` uses quasi-Newton methods and can handle box constraints
- `nlminb()` uses trust-region procedures and can also handle box constraints
- `constrOptim()` can be used for linear inequality constraints and is based on interior-point methods
- `nloptr` is an interface to NLopt, an open-source library for nonlinear optimization