Properties of Dual Problems

(**P**)-the primal problem (**D**)-the dual problem

- (**P**) optimal $\iff$ (**D**) optimal
- (**P**) unbounded $\Longrightarrow$ (**D**) infeasible
- (**D**) unbounded $\Longrightarrow$ (**P**) infeasible
- (**P**) infeasible $\Longrightarrow$ (**D**) infeasible or unbounded
- (**D**) infeasible $\Longrightarrow$ (**P**) infeasible or unbounded

Theorem

Assume x^* and w^* are feasible solutions to the primal and dual LP problems, respectively (either in the symmetric or asymmetric form). x^* and w^* are optimal solutions to their respective problems if and only if:

1. $(c^T - {w^*}^T A)x^* = 0 \iff (c_i - {w^*}^T a_i)x_i^* = 0$, $i = \overline{1, n}$
2. ${w^*}^T(b - Ax^*) = 0 \iff w_j^*(b_j - a^j x^*) = 0$ $j = \overline{1, m}$.

a_i is the i-th column of matrix A.

a^j is the j-th row of matrix A.

Example

Solve the following minimization problem

$$\begin{aligned} \text{minimize} \quad & 2x_1 + 3x_2 + 5x_3 + 2x_4 + 3x_5 \\ \text{subject to} \quad & x_1 + x_2 + 2x_3 + x_4 + 3x_5 \geq 4 \\ & 2x_1 - 2x_2 + 3x_3 + x_4 + x_5 \geq 3 \\ & x_i \geq 0,\ i = \overline{1,5}. \end{aligned}$$

Stochastic Gradient Descent SGD

SGD is one of the widely used numerical optimization methods in Machine Learning (ML).

First let's see how minimization problems arise in ML.

Here we consider supervised ML algorithms. Assume we have a collection of examples of the form

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\},$$

where for each $i = 1, \ldots n$ x_i is the vector of features of the observation and y_i is the label of observation.

Our aim is to find a prediction function $y = g(x)$.

In order to choose o prediction function we need to:

- choose a class of predictors;
- choose a particular predictor from the class of the predictors that is optimal.

Usually, one chooses a parametric class of functions. To choose the optimal prediction function one takes a loss function and minimizes the Expected Risk.

Assuming some probability distribution (unknown) behind (X, Y). The expected risk of predictor g is

$$Risk(g) = \mathbb{E}(\ell(Y, g(X, w))).$$

Our problem will be to solve the following problem

$$\begin{aligned} \text{minimize} \quad & Risk(g) \\ \text{subject to} \quad & \text{all } g(x, w). \end{aligned}$$

or

$$\begin{aligned} \text{minimize} \quad & Risk(g(x, w)) \\ \text{subject to} \quad & w \in \mathbb{R}^d. \end{aligned}$$

$w^* = \arg\min_{w \in \mathbb{R}^d} Risk(g(x, w))$ and $g(x, w^*)$ will be our predictor.

Approximate

$$Risk(g) \approx \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, g(x_i, w)).$$

$$Empirical\ Risk(g) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, g(x_i, w)).$$

Empirical Risk(g) is a function of w.

Now instead of minimizing the expected risk *Risk*(g) we are going to minimize *Empirical Risk*(g).

$$\begin{array}{rl} \text{minimize} & Empirical\ Risk(g) \\ \text{subject to} & w \in \mathbb{R}^d. \end{array}$$

Let's denote

$$f_i(w) = \ell(y_i, g(x_i, w)).$$

Our problem will become

$$\begin{aligned} \text{minimize} \quad & \tfrac{1}{n}\textstyle\sum_{i=1}^{n} f_i(w) \\ \text{subject to} \quad & w \in \mathbb{R}^d. \end{aligned}$$

Example

Assume we have some observations (x_i, y_i) and our aim is to predict y as a function of x: $y = g(x)$. As a class of prediction functions we take the class of quadratic functions $g(x, w) = w_0 + w_1 x$ and as loss function we take the quadratic loss function $\ell(u, v) = (u - v)^2$. In order to find the prediction function we need to solve the following minimization problem

$$\begin{aligned} \text{minimize} \quad & \tfrac{1}{n} \textstyle\sum_{i=1}^{n} (y_i - w_0 - w_1 x_i)^2 \\ \text{subject to} \quad & w \in \mathbb{R}^2. \end{aligned}$$

Examples of other loss functions:

log loss function $\ell(u, v) = \ln(1 + e^{-uv})$

hinge loss function $\ell(u, v) = \max\{0, 1 - uv\}$

Numerical minimization algorithms

- Batch Gradient Descent

  Choose initial approximation $w^{(0)}$

  $$w^{(k+1)} = w^{(k)} - \frac{\alpha_k}{n} \sum_{i=1}^{n} \nabla f_i(w^{(k)}), \quad k = 0, \ldots$$

  Choose a step size α_k using some method that we discussed above

  To stop use some stopping condition

- Stochastic Gradient Descent

  Choose initial approximation $w^{(0)}$

  For each k choose a random i_k (uniformly) from $\{1, ..., n\}$

$$w^{(k+1)} = w^{(k)} - \alpha_k \nabla f_{i_k}(w^{(k)}), \quad k = 0, \dots$$

  Choose a step size α_k.

  To stop use some stopping condition

Motivation for Stochastic Gradient Descent Methods

- Not as expensive as Batch GD
- Usually, SGD employs information more efficiently