Optimization MTH702

Proximal GD, subGD, FWGD,CoorGD

Hilal AlQuabeh

Machine Learning Department
MBZUAI

September 6, 2021

Table of Contents

Projected Gradient Descent

- The constrained optimization forces the feasible set to be $X \subseteq R^n$
- Projected Gradient Descent works well when the functions is convex and the feasible set as well (convex set have unique projection).
- **Idea** project every step $\Pi_X(y) := argmin_{x \in X} ||x - y||$.
- Projected GD step: $x_{t+1} = \Pi_X \left[x_t - \gamma \nabla f(x_t) \right]$
- Let $X \subseteq R^d$ be closed and convex, $x \in X$, $y \in R^d$. Then (i) $(x - \Pi_X(y))^T (y - \Pi_X(y)) \leq 0$
(ii) $||x - \Pi_X(y)||^2 + ||y - \Pi_X(y)||^2 \leq ||x - y||^2$

PGD Convergence Rate

- The same number of steps as gradient descent with same proofs but each step involves a projection onto X, may or may not be efficient(depends on the set).

Proximal Gradient Descent

Consider the objective function to be composed as:

$$f(x) := g(x) + h(x)$$

where g is nice function, but h is only simple (not differentiable) e.g.L1 norm or the indicator function.

- The classical GD step for minimizing g is:

$$x_{t+1} = argmin_y g(x_t) + \nabla g(x_t)^T (y - x_t) + \frac{1}{2\gamma} ||y - x_t||^2$$

For the step size = 1/L it exactly minimizes the local quadratic model of g at our current iterate x t , formed by the smoothness property with parameter L.

- Now for f = g + h, we do the exactly the same for g and h as it is:

$$x_{t+1} := argmin_y g(x_t) + \nabla g(x_t)^T (y - x_t) + \frac{1}{2\gamma} ||y - x_t||^2 + h(y)$$

$$x_{t+1} := argmin_y \frac{1}{2\gamma} ||y - (x_t - \gamma \nabla g(x_t)||^2 + h(y)$$

This is called proximal update (step).

Proximal Gradient Descent

- The proximal gradient step is also written as :

$$x_{t+1} := Prox_{h,\gamma}(x_t - \gamma \nabla g(x_t))$$

  where the proximal operator for a a given function and parameter $\gamma > 0$ is defined as: $Prox_{h,\gamma} := argmin_y \left\{ \frac{1}{2\gamma} ||y - z||^2 + h(y) \right\}$
- If h(x) $= 0$, we will have original GD.
- If $h = \Phi_x$ we will have projected GD.