10.3 Projected Gradient Descent

So far, we were concerned with finding the optimal solution of an unconstrained optimization problem. In real life, optimization problems we are likely to come across constrained optimization problems. In this section, we discuss how to solve constrained optimization problem:

$$\min_{x \in X} f(x)$$

where f is a convex function and X is a convex set.

If we wish to use gradient descent update to a point $x_t \in X$, it is possible that the iterate $x_{t+1} = x_t - \frac{\nabla f(x_t)}{L}$ may not belong to the constraint set X. In the projected gradient descent, we simply choose the point nearest to $x_t - \frac{\nabla f(x_t)}{L}$ in the set X as x_{t+1} i.e., the projection of $x_t - \frac{\nabla f(x_t)}{L}$ onto the set X.

Definition 10.3 *The projection of a point y, onto a set X is defined as*

$$\Pi_X(y) = \operatorname*{argmin}_{x \in X} \frac{1}{2}\|x - y\|_2^2.$$

Projected Gradient Descent (PGD): Given a starting point $x_0 \in X$ and step-size $\gamma > 0$, PGD works as follows until a certain stopping criterion is satisfied,

$$x_{t+1} = x_t - \gamma \Pi_X(x_t - \nabla f(x_t)), \forall t \geq 1.$$

In this lecture, for an L smooth convex function, we fix the step-size to be $\gamma = \frac{1}{L}$.

Proposition 10.4 *The following inequalities hold:*

1. *If $y \in X$, then $\Pi_X(y) = y$.*
2. *The projection onto a convex set X is non-expansive.*
   $\|\Pi_X(x) - \Pi_X(y)\|_2 \leq \|x - y\|_2$,
   $\|\Pi_X(x) - \Pi_X(y)\|_2^2 \leq \langle \Pi_X(x) - \Pi_X(y), x - y \rangle \leq \|x - y\|_2^2$.

Proof:

1. We first note that the $\frac{1}{2}\|x - y\|_2^2$ is strictly convex since $\nabla^2 f(x) = 1$. Hence the solution to the optimization problem is unique. If $y \in X$, then we have $\frac{1}{2}\|y - y\|_2^2 = 0$. Since $\frac{1}{2}\|x - y\|_2^2 \geq 0$, zero is the optimal value of the function and y is it's unique minimizer, thus giving $\Pi_X(y) = y$.
2. For any feasible x^*, the optimality conditions are given by

   $$\langle \nabla f(x^*), z - x^* \rangle \geq 0, \forall z \in X.$$

   Hence for x, y, we have

   $$\langle \Pi_X(y) - y, \Pi_X(x) - \Pi_X(y) \rangle \geq 0,$$
   $$\langle \Pi_X(x) - x, \Pi_X(y) - \Pi_X(x) \rangle \geq 0,$$

   since $\Pi_X(y), \Pi_X(x) \in X$.

Combining the above equations, we have

$$\langle \Pi_X(x) - \Pi_X(y) - (x-y), 2(\Pi_X(y) - \Pi_X(x)) \rangle \geq 0,$$

$$\langle y - x, \Pi_X(y) - \Pi_X(x) \rangle \geq \langle \Pi_X(y) - \Pi_X(x), \Pi_X(y) - \Pi_X(x) \rangle,$$

From Cauchy-Schwartz we further have

$$\langle y - x, \Pi_X(y) - \Pi_X(x) \rangle \leq \|\Pi_X(x) - \Pi_X(y)\|_2 \|x - y\|_2$$

giving us $\|\Pi_X(x) - \Pi_X(y)\|_2 \leq \|x - y\|_2$.

■

10.3.1 Interpretation

We present a few useful observations.

1. Each iterate in the PGD can be viewed as the minimizer of a quadratic approximation of the objective function, similar to the case of Gradient Descent (GD).

$$x_{t+1} = \operatorname*{argmin}_{x \in X} \{ f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{\mu}{2} \|x - x_t\|^2 \}.$$

   We also observe that the objective function is non-increasing with each iteration, $f(x_{t+1}) \leq f(x_t)$.

2. The optimal solution x^* is a fixed point of $x = \Pi_X(x - \frac{\nabla f(x)}{L})$ i.e., $x^* = \Pi_X(x^* - \frac{\nabla f(x^*)}{L})$.

   Proof:

$$\begin{aligned}
x^* \text{ is optimal} &\iff \langle \nabla f(x^*), z - x^* \rangle \geq 0, \forall z \in X \\
&\iff \langle -\frac{1}{L} \nabla f(x^*), z - x^* \rangle \leq 0, \forall z \in X \\
&\iff \langle (x^* - \frac{1}{L} \nabla f(x^*)) - x^*, z - x^* \rangle \leq 0, \forall z \in X \\
&\iff \langle x^* - (x^* - \frac{1}{L} \nabla f(x^*)), z - x^* \rangle \geq 0, \forall z \in X \\
&\iff x^* \text{ is the projection of } x^* - \frac{1}{L} \nabla f(x^*)
\end{aligned}$$

   where the last line follows because the projection of y is the minimizer of the function $\frac{1}{2}\|x - y\|_2^2$ over the set X. ■

3. We can rewrite the iterates of PGD as follows

$$x_{t+1} = x_t - \frac{1}{L} g_X(x_t).$$

   by defining $g_X(x) = L(x - x^\dagger)$, where $x^\dagger = \Pi_X(x - \frac{1}{L}\nabla f(x))$. The function $g_X(x)$ is often called the *Gradient Mapping*. Note that if the problem is unconstrained, $x^\dagger = x - \frac{1}{L}\nabla f(x)$, $g_X(x) = \nabla f(x)$, thus reducing to the usual gradient descent case.

It can be shown that the key inequalities used to show convergence in the gradient descent method follow in a similar way for the projected gradient descent as well by replacing the gradient with the gradient mapping.

10.3.2 Convergence Analysis

Recall that when showing the convergence rates for the gradient descent algorithm, we used the following properties:

(a) For a L-smooth function f, the iterates given by the gradient descent with step size $\gamma = \frac{1}{L}$ satisfy

$$f(x_{t+1}) - f(x_t) \leq -\frac{\|\nabla f(x_t)\|^2}{2L}.$$

(b) If f is aslo μ-strongly convex, we have

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{\mu L}{\mu + L}\|x - y\|^2 + \frac{1}{\mu + L}\|\nabla f(x) - \nabla f(y)\|^2.$$

Similar results hold for the projected gradient descent which are presented below. Details can be found in Section 2.2 from [NES'04].

Proposition 10.5 *For a convex and L-smooth function f, we have*

$$f(x^\dagger) - f(x) \leq -\frac{\|g_X(x)\|^2}{2L}.$$

If f is also μ-strongly convex, we have

$$\langle g_X(x), x - x^* \rangle \geq \frac{\mu}{2}\|x - x^*\|^2 + \frac{1}{2L}\|g_X(x)\|^2.$$

Remark. With these facts, we can immediately adapt previous convergence analysis for GD method to analyze PGD and obtain similar results, namely, a sublinear rate $O(1/t)$ for general smooth convex case and a linear rate $O((1-\kappa^{-1})^t)$ for the smooth strongly convex case. Moreover, if we combine the projected gradient descent with Nesterov's acceleration, we will also obtain the optimal convergence results for constained convex optimization, similar to what we have in the unconstrained case.

Theorem 10.6 *For a convex function f that is L-smooth, the iterates given by the projected gradient descent with step size $\gamma = \frac{1}{L}$ satisfy*

$$f(x_t) - f(x^*) \leq \frac{2L}{t}\|x_0 - x^*\|^2.$$

If f is further μ-strongly convex, we have

$$\|x_t - x^*\|^2 \leq (1 - \frac{\mu}{L})^t \|x_0 - x^*\|^2.$$

Proof: The proofs are similar to the gradient descent case. We present the proof for μ-strongly convex case.

$$\begin{aligned}
\|x_{t+1} - x^*\|^2 &= \|x_t - \frac{1}{L}g_X(x_t) - x^*\|^2 \\
&= \|x_t - x^*\|^2 + \frac{1}{L^2}\|g_X(x_t)\|^2 - \frac{2}{L}\langle g_X(x_t), x_t - x^* \rangle \\
&\leq \|x_t - x^*\|^2 + \frac{1}{L^2}\|g_X(x_t)\|^2 - \frac{2}{L}(\frac{\mu}{2}\|x_t - x^*\|^2 + \frac{1}{2L}\|\nabla g_X(x_t)\|^2) \\
&= (1 - \frac{\mu}{L})\|x_t - x^*\|^2 \\
&\leq (1 - \frac{\mu}{L})^t \|x_0 - x^*\|^2.
\end{aligned}$$

where we have used Proposition 10.5 to bound the inner product. ■