7.3.3 Nesterov Iteration, no proof

We can even define an *accelerated* version of the proximal gradient method. Iterations take the form:

$$\begin{aligned} \xi_{k+1} &= \Pi_\Omega \left(y_k - \alpha \nabla f(y_k) \right) \\ y_k &= \xi_k + \beta(\xi_k - \xi_{k-1}) \end{aligned} \tag{7.15}$$

Note that when $\Pi_\Omega = I$, we recover the standard Nesterov algorithm. When $\beta = 0$, we recover the proximal gradient method. This method will converge in

$$O\left(\sqrt{\frac{L}{m}} \log(1/\epsilon) \right)$$

iterations for strongly convex functions.

7.4 The Conditional Gradient ("Frank-Wolfe") Method

Often times, the computation of the projection onto the set Ω is a very expensive operation. Moreover, for many sets that arise in optimization, it is often considerably simpler to minimize a linear objective over Ω than it is to project onto this set. For example, minimizing a linear objective over the simplex simply requires extracting a maximum, whereas the Euclidean projection naively requires sorting a list of numbers. The conditional gradient method, the first variant of which was proposed by Frank and Wolfe [19], provides an effective algorithm for constrained optimization that requires only linear minimization rather than Euclidean projection.

Conditional gradient method replaces the objective in (7.3) by a linear Taylor-series approximation around the current iterate x^k, and solves the following subproblem:

$$\bar{x}^k := \arg\min_{\bar{x} \in \Omega} \; f(x^k) + \nabla f(x^k)^T (\bar{x} - x^k) = \arg\min_{\bar{x} \in \Omega} \; \nabla f(x^k)^T \bar{x}. \tag{7.16}$$

Note that the constraint set Ω is unchanged. The next iterate is obtained by steeping toward $\bar{x}^k$ from x^k, as follows

$$x^{k+1} = x^k + \alpha_k (\bar{x}^k - x^k), \quad \text{for some } \alpha_k \in (0, 1]. \tag{7.17}$$

Note that if the initial iterate x^0 is feasible (that is, $x^0 \in \Omega$), all subsequent iterates x^k, $k = 1, 2, \ldots$ are also feasible, as are all the subproblem solutions $\bar{x}^k$, $k = 0, 1, 2 \ldots$.

This method is practical when the linearized subproblem (7.16) is much easier to solve than the original problem (7.3). As we have discussed, this is the case in many applications of interest.

The original Frank-Wolfe approach made the particular choice of step length $\alpha_k = 2/(k+2)$, $k = 0, 1, 2, \ldots$. The resulting method converges at a sublinear rate, as we show now. Again assume that $\Omega \subset \mathbb{R}^n$ is a closed, bounded convex set and f is a smooth convex function. We define the *diameter* D of Ω as follows:

$$D := \max_{x, y \in \Omega} \; \|x - y\|. \tag{7.18}$$

Theorem 7.4. *Suppose that f is a convex function whose gradient is Lipschitz continuously differentiable with constant L on an open neighborhood of Ω, where Ω is a closed bounded convex set*

with diameter D, and that solution x^* to (7.3) exists. Then if algorithm (7.16)-(7.17) is applied from some $x^0 \in \Omega$ with steplength $\alpha_k = 2/(k+2)$, we have

$$f(x^k) - f(x^*) \leq \frac{2LD^2}{k+2}, \quad k = 1, 2, \ldots .$$

Proof. Setting $x = x^k$ and $y = x^{k+1} = x^k + \alpha_k(\bar{x}^k - x^k)$ in Lemma **??**, we have

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + \alpha_k \nabla f(x^k)^T(\bar{x}^k - x^k) + \frac{1}{2}\alpha_k^2 L\|\bar{x}^k - x^k\|^2 \\ &\leq f(x^k) + \alpha_k \nabla f(x^k)^T(\bar{x}^k - x^k) + \frac{1}{2}\alpha_k^2 LD^2, \end{aligned} \tag{7.19}$$

where the second inequality comes from the definition of D. For the first-order term, we have by definition of $\bar{x}^k$ in (7.16) and feasibility of x^* that

$$\nabla f(x^k)^T(\bar{x}^k - x^k) \leq \nabla f(x^k)^T(x^* - x^k) \leq f(x^*) - f(x^k).$$

By substituting this bound into both sides of (7.19) and subtracting $f(x^*)$ from both sides, we have

$$f(x^{k+1}) - f(x^*) \leq (1 - \alpha_k)[f(x^k) - f(x^*)] + \frac{1}{2}\alpha_k^2 LD^2.$$

We now demonstrate the required bound by induction. By setting $k = 0$ and substituting $\alpha_0 = 1$, we have

$$f(x^1) - f(x^*) \leq \frac{1}{2}LD^2 < \frac{2}{3}LD^2,$$

as required. For the inductive step, we suppose that the claim holds for some k, and demonstrate that it still holds for $k+1$. We have

$$\begin{aligned} f(x^{k+1}) - f(x^*) &\leq \left(1 - \frac{2}{k+2}\right)[f(x^k) - f(x^*)] + \frac{1}{2}\frac{4}{(k+2)^2}LD^2 \\ &= LD^2\left[\frac{2k}{(k+2)^2} + \frac{2}{(k+2)^2}\right] \\ &= 2LD^2\frac{(k+1)}{(k+2)^2} \\ &= 2LD^2\frac{k+1}{k+2}\frac{1}{k+2} \\ &\leq 2LD^2\frac{k+2}{k+3}\frac{1}{k+2} = \frac{2LD^2}{k+3}, \end{aligned}$$

as required. □

Note that the same result holds if we choose α_k to exactly minimize f along the line from x^k to $\bar{x}^k$; only minimal changes to the proof are needed.

Notes and References

Pointer to proof of Nesterov's method with projection.

Homework: projection and linear optimization on the simplex