DSCC/CSC 435 & ECE 412 Optimization for Machine Learning	Lecture 14
Nonconvex Optimization	
Lecturer: Jiaming Liang	*November 21, 2023*

1 Setup

We are interested in solving

$$\min\{\phi(x) := f(x) + h(x)\}. \tag{1}$$

We assume that

(A1) h is closed and convex;

(A2) there exist scalar $L \geq 0$ and a compact convex set $\Omega \supset \operatorname{dom} h$ such that f is nonconvex and differentiable on Ω, and

$$\left\|\nabla f(u) - \nabla f\left(u'\right)\right\| \leq L\left\|u - u'\right\|, \quad \forall u, u' \in \Omega.$$

It follows from (A1) and (A2) that the set of optimal solutions X_* is nonempty and compact. Second, if L satisfies (A2) then the pair $(M, m) = (L, L)$ satisfies

$$-\frac{m}{2}\left\|u - u'\right\|^2 \leq f(u) - \ell_f\left(u; u'\right) \leq \frac{M}{2}\left\|u - u'\right\|^2, \quad \forall u, u' \in \Omega. \tag{2}$$

Clearly,

$$0 \leq m \leq L, \quad 0 \leq M \leq L.$$

The interesting case is when $m \leq M$, and we say function f is m-weakly convex if

$$-\frac{m}{2}\left\|u - u'\right\|^2 \leq f(u) - \ell_f\left(u; u'\right), \quad \forall u, u' \in \Omega.$$

It is easy to check that $g(\cdot) = f(\cdot) + \frac{1}{2\lambda}\|\cdot\|^2$ is convex if $\lambda \leq 1/m$. Hence, weakly-convex functions are convexifiable.

Solving for global minima or even local minima are intractable in nonconvex optimization, so we instead solve for stationary points, i.e. $x \in \operatorname{dom} h$ satisfying

$$0 \in \nabla f(x) + \partial h(x).$$

Definition 1. *Given $\rho > 0$, a pair (v, x) is called a ρ-approximate stationary pair of problem* (1) *if*

$$v \in \nabla f(x) + \partial h(x), \quad \|v\| \leq \rho.$$

2 Projected gradient method

Algorithm 1 Projected gradient method

Input: Initial point $x_0 \in \operatorname{dom} h$, $\lambda \in (0, 1/M]$, $\rho > 0$.

for $k \geq 0$ **do**

Step 1. Compute

$$x_{k+1} = \operatorname{argmin} \left\{ \ell_f(u; x_k) + h(u) + \frac{1}{2\lambda} \|u - x_k\|^2 \right\}$$

Step 2. Compute

$$v_{k+1} = \frac{x_k - x_{k+1}}{\lambda} + \nabla f(x_{k+1}) - \nabla f(x_k),$$

if $\|v_{k+1}\| \leq \rho$, then stop.

end for

Lemma 1. *For every $k \geq 0$, we have*

$$v_{k+1} \in \nabla f(x_{k+1}) + \partial h(x_{k+1}), \quad \|v_{k+1}\| \leq \left(L + \frac{1}{\lambda} \right) \|x_{k+1} - x_k\|.$$

Proof. The optimality condition of the subproblem in Algorithm 1 is

$$0 \in \nabla f(x_k) + \partial h(x_{k+1}) + \frac{1}{\lambda} (x_{k+1} - x_k).$$

Hence, the inclusion in the lemma holds after rearrangement. The inequality follows from the definition of v_{k+1} and the fact that ∇f is L-smooth. □

Theorem 1. *For every $k \geq 1$, we have*

$$\min_{1 \leq i \leq k} \|v_i\| \leq \left(L + \frac{1}{\lambda} \right) \frac{\sqrt{2[\phi(x_0) - \phi_*]}}{\sqrt{Mk}}.$$

Proof. It follows from the subproblem in Algorithm 1 that for every $u \in \operatorname{dom} h$,

$$\begin{aligned} &\ell_f(u; x_k) + h(u) + \frac{1}{2\lambda} \|u - x_k\|^2 - \frac{1}{2\lambda} \|u - x_{k+1}\|^2 \\ &\geq \ell_f(x_{k+1}; x_k) + h(x_{k+1}) + \frac{1}{2\lambda} \|x_{k+1} - x_k\|^2 \\ &\geq \phi(x_{k+1}) - \frac{M}{2} \|x_{k+1} - x_k\|^2 + \frac{1}{2\lambda} \|x_{k+1} - x_k\|^2, \end{aligned}$$

where the second inequality is due to the second inequality of (2). Taking $u = x_k$ in the above inequality and using the fact that $\lambda \leq 1/M$, we have

$$\phi(x_k) - \phi(x_{k+1}) \geq \left(\frac{1}{\lambda} - \frac{M}{2} \right) \|x_{k+1} - x_k\|^2 \geq \frac{M}{2} \|x_{k+1} - x_k\|^2.$$

Summing the above inequality, we obtain

$$\phi(x_0) - \phi(x_k) \geq \frac{M}{2} k \min_{0 \leq i \leq k-1} \|x_{i+1} - x_i\|^2 \geq \frac{M}{2} k \left(L + \frac{1}{\lambda}\right)^{-2} \min_{1 \leq i \leq k} \|v_i\|^2 .$$

Hence, the lemma holds. □

3 Frank-Wolfe method

Recall the Frank-Wolfe method from Lecture 8.

Algorithm 2 Generalized Frank-Wolfe method

Input: Initial point $x_0 \in \operatorname{dom} h$
for $k \geq 0$ **do**
  Step 1. Compute $y_k = \operatorname{argmin}_{y \in \mathbb{R}^n} \{\langle y, \nabla f(x_k)\rangle + h(y)\}$.
  Step 2. Choose $t_k \in [0, 1]$ and set $x_{k+1} = (1 - t_k)x_k + t_k y_k$.
end for

Also recall the following results.

Definition 2. *The Wolfe gap is the function $S(x) : \operatorname{dom} h \to \mathbb{R}$ given by*

$$S(x) = \max_{y \in \mathbb{R}^n} \{\langle \nabla f(x), x - y\rangle + h(x) - h(y)\}.$$

Lemma 2. *The following statements hold:*

(a) $S(x) \geq 0$ for any $x \in \operatorname{dom} h$;

(b) $S(x_*) = 0$ if and only if $-\nabla f(x_*) \in \partial h(x_*)$, that is, if and only if x_* is a stationary point of (1).

Proof. (b) It follows from (a) that $S(x_*) = 0$ if and only if $S(x_*) \leq 0$, which is the same a

$$\langle \nabla f(x_*), x_* - x\rangle + h(x_*) - h(x) \leq 0, \quad x \in \operatorname{dom} h.$$

Rearraging the terms gives

$$h(x) \geq h(x_*) + \langle -\nabla f(x_*), x - x_*\rangle,$$

which is equivalent to $-\nabla f(x_*) \in \partial h(x_*)$. □

Lemma 3. *Let $x \in \operatorname{dom} h$ and $t \in [0, 1]$. Then, we have*

$$\phi((1 - t)x + ty) \leq \phi(x) - tS(x) + \frac{t^2 L}{2} \|y - x\|^2,$$

where $y = \operatorname{argmin}_{u \in \mathbb{R}^n} \{\langle u, \nabla f(x)\rangle + h(u)\}$.

Three stepsize rules

1) predefined diminishing stepsize:
$$\alpha_k = \frac{2}{k+2};$$

2) adaptive stepsize:
$$\beta_k = \min\left\{1, \frac{S(x_k)}{L\left\|y_k - x_k\right\|^2}\right\};$$

3) exact minimization/line search:
$$\eta_k \in \operatorname{argmin}_{t\in[0,1]} \phi\left((1-t)x_k + t y_k\right).$$

The following lemma is (P3) of PS3.

Lemma 4. *Using the adaptive or exact line search stepsizes in Frank-Wolfe, then for every $k \geq 0$, we have*
$$\phi(x_k) - \phi(x_{k+1}) \geq \frac{1}{2}\min\left\{S(x_k), \frac{S^2(x_k)}{LD^2}\right\}, \tag{3}$$
where D is the diameter of $\operatorname{dom} h$.

Now, with all the above technical results, we are ready to present the main convergence result of Frank-Wolfe method applied to nonconvex optimization.

Theorem 2. *Using the adaptive or exact line search stepsizes in Frank-Wolfe, then the following statements hold:*

(a) for every $k \geq 0$, $\phi(x_k) \geq \phi(x_{k+1})$ and $\phi(x_k) > \phi(x_{k+1})$ if x_k is not a stationary point of (1)*;*

(b) $S(x_k) \to 0$ as $k \to \infty$;

(c) for every $k \geq 1$,
$$\min_{0\leq i\leq k-1} S(x_i) \leq \max\left\{\frac{2\left(\phi(x_0) - \phi_*\right)}{k}, \frac{\sqrt{2L_f D^2\left(\phi(x_0) - \phi_*\right)}}{\sqrt{k}}\right\}$$

(d) all limit points of the sequence $\{x_k\}_{k\geq 0}$ are stationary points of problem (1)*.*

Proof. (a) The first result directly follows from Lemma 4 and the fact that $S(x_k) \geq 0$. If x_k is not a stationary point, then $S(x_k) > 0$ and hence $\phi(x_k) > \phi(x_{k+1})$ in view of (3).

(b) Since $\{\phi(x_k)\}$ is non-increasing and bounded from below, it follows that it is convergent. In particular, $\phi(x_k) > \phi(x_{k+1}) \to 0$ as $k \to \infty$. Therefore, it follows from Lemma 4 that $S(x_k) \to 0$ as $k \to \infty$.

(c) Summation of (3) over iterations gives

$$\phi(x_0) - \phi(x_k) \geq \frac{1}{2} \sum_{i=0}^{k-1} \min \left\{ S(x_i), \frac{S^2(x_i)}{LD^2} \right\} \geq \frac{k}{2} \min_{0 \leq i \leq k-1} \min \left\{ S(x_i), \frac{S^2(x_i)}{LD^2} \right\},$$

then the result holds after simple calculation.

(d) Suppose that $\bar{x}$ is a limit point of $\{x_k\}_{k \geq 0}$. Then there exists a subsequence $\left\{x_{k_j}\right\}_{j \geq 0}$ that converges to $\bar{x}$. By the definition of the Wolfe gap $S(\cdot)$, it follows that for any $x \in \operatorname{dom} h$,

$$S(x_{k_j}) \geq \left\langle \nabla f\left(x_{k_j}\right), x_{k_j} - x \right\rangle + h\left(x_{k_j}\right) - h(x).$$

Passing to the limit $j \to \infty$ and using the fact that $S\left(x_{k_j}\right) \to 0$ as $j \to \infty$, as well as the continuity of ∇f and the lower semicontinuity of h, we obtain that

$$0 \geq \langle \nabla f(\bar{x}), \bar{x} - x \rangle + h(\bar{x}) - h(x), \quad \forall x \in \operatorname{dom} h,$$

which is the same as the relation $-\nabla f(\bar{x}) \in \partial h(\bar{x})$, that is, $\bar{x}$ is a stationary point of (1). □

In the proof of part (d), we have used the fact that h is closed is equivalent to h is lower semicontinuous. We say h is lower semicontinuous at x_0 if and only if

$$\liminf_{x \to x_0} h(x) \geq h\left(x_0\right).$$