Frank-Wolfe Method

Lecturer: Jiaming Liang *February 8, 2024*

1 Frank-Wolfe method

Consider the problem $\min\{f(x) : x \in Q\}$ where f is convex and $Q \subseteq \operatorname{dom} f$ is convex and compact. We also assume f is differentiable over $\operatorname{dom} f$. One method can be employed is the projected gradient method

$$x_{k+1} = \operatorname{proj}_Q(x_k - t_k \nabla f(x_k)),$$

which is equivalent to

$$x_{k+1} = \operatorname{argmin} \left\{ \ell_f(x; x_k) + \frac{1}{2t_k}\|x - x_k\|^2 : x \in Q \right\}.$$

In this lecture, we will present an alternative approach that does not require the projection operator proj_Q. The idea is to minimize the linearization of f (without the quadratic term) over Q

$$y_k = \operatorname{argmin} \left\{ \ell_f(x; x_k) : x \in Q \right\} = \operatorname{argmin} \left\{ \langle \nabla f(x_k), x \rangle : x \in Q \right\},$$

and then take a convex combination

$$x_{k+1} = x_k + t_k(y_k - x_k), \quad t_k \in [0, 1].$$

This algorithm is called Frank-Wolfe method, a.k.a., conditional gradient method.

Algorithm 1 Frank-Wolfe method

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

This is a projection-free method since we minimize a linear function over Q. In many case, linear optimization over Q is simpler than projection onto Q.

Frank-Wolfe method satisfies an even more important property: it produces sparse iterates. More precisely, consider the situation where $Q \subset \mathbb{R}^n$ is a polytope, that is the convex hull of a finite set of points (vertices). Then Carathéodory's theorem states that any point $x \in Q \subset \mathbb{R}^n$ can be written as a convex combination of at most $n+1$ vertices of Q. On the other hand, by step 2 of Frank-Wolfe, one knows that the k-th iterate x_k can be written as a convex combination of $k+1$ vertices (assuming that x_0 is a vertex). Thanks to the dimension-free rate of convergence, we are interested in the regime where $k \ll n$, and thus we see that the iterates of Frank-Wolfe are very sparse in their vertex representation.

Let us consider the general composite opimization problem.

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

- h is closed and convex and $\operatorname{dom} h$ is compact;
- f is closed and convex, $\operatorname{dom} h \subseteq \operatorname{dom} f$, and f is L-smooth over some set $\operatorname{dom} h$, i.e.,

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\|, \quad \forall x, y \in \operatorname{dom} h;$$

- the optimal set X_* is nonempty.

It is not difficult to deduce that the last condition is implied by the first two conditions.

The three properties of Frank-Wolfe method are projection-free (prox-free), norm-free, and sparse iterates.

In the rest of the lecture, we will consider the following generalized Frank-Wolfe method.

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

2 Convergence analysis

Definition 1. *The Wolfe gap is the function* $S(x) : \operatorname{dom} f \to \mathbb{R}$ *given by*

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

Lemma 1. *The following statements hold:*

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

(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).

The above lemma gives the importance of the Wolfe gap $S(x)$, which can be used to analyze the convergence of Frank-Wolfe for nonconvex optimization.

Lemma 2. *Let $x \in \operatorname{dom} h$ and $t \in [0,1]$. Then, we have*

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

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

Proof. Let $x^+ = (1-t)x + ty$. Then, using the smoothness of f and the convexity of h, we easily show

$$\begin{aligned}
\phi(x^+) &= f(x^+) + h(x^+) \\
&\leq f(x) - t\langle \nabla f(x), x-y\rangle + \frac{t^2 L}{2}\|y-x\|^2 + h(x^+) \\
&\leq f(x) - t\langle \nabla f(x), x-y\rangle + \frac{t^2 L}{2}\|y-x\|^2 + (1-t)h(x) + th(y) \\
&= \phi(x) - t\left[\langle \nabla f(x), x-y\rangle + h(x) - h(y)\right] + \frac{t^2 L}{2}\|y-x\|^2 \\
&= \phi(x) - tS(x) + \frac{t^2 L}{2}\|y-x\|^2.
\end{aligned}$$

□

Note that so far, we do not use the convexity of f yet.

Three stepsize rules

1) predefined diminishing stepsize:

$$\alpha_k = \frac{2}{k+2};$$

2) adaptive stepsize:

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

3) exact minimization/line search:

$$\eta_k \in \operatorname{argmin}_{t\in[0,1]} \phi\left((1-t)x_k + ty_k\right).$$

The intuition of the adaptive stepsize is β_k minimizes the right-hand side of (2) w.r.t. $t \in [0,1]$ when $x = x_k$. It is clear the exact minimization rule chooses $t_k = \eta_k$ to minimize the left-hand side of (2). The underlying intuition for the first rule α_k is related to the accelerated gradient method. This is not elaborated here due to its complexity.

The following lemma shows that Wolfe gap $S(x)$ is in fact a primal-dual gap, and henc it upper bounds both primal and dual gaps.

Lemma 3. *For any* $x \in \operatorname{dom} f$*, we have*

$$S(x) = \phi(x) - \psi(\nabla f(x)),$$

where $\psi(y) := -f^*(y) - h^*(-y)$ *denotes the Lagrange dual function of* $\phi(x)$*. Moreover, using convexity of* ϕ*, we have*

$$S(x) \geq \phi(x) - \phi_*, \quad S(x) \geq \psi^* - \psi(\nabla f(x)),$$

where $\phi_* = \min_{x\in\mathbb{R}^n} \phi(x)$ *and* $\psi^* = \max_{y\in\mathbb{R}^n} \psi(y)$*.*

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

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

where we use the fact that $\langle \nabla f(x), x\rangle = f(x) + f^*(\nabla f(x))$ in the last identity (see Theorem 3(i) of Lecture 3). Using the definitions of ϕ and ψ, and weak duality, we obtain

$$S(x) = \phi(x) - \psi(\nabla f(x)) \geq \phi(x) - \psi^* \geq \phi(x) - \phi_*,$$

and

$$S(x) = \phi(x) - \psi(\nabla f(x)) \geq \phi_* - \psi(\nabla f(x)) \geq \psi^* - \psi(\nabla f(x)).$$

□

Theorem 1. *The generalized Frank-Wolfe method with any of the three stepsize rules satisfies*

$$\phi(x_k) - \phi_* \leq \frac{2LD^2}{k+1}, \quad \forall k \geq 1, \tag{3}$$

where D *is the diameter of* $\operatorname{dom} h$*.*

Proof. Using Lemma 2 with $t = t_k$ and $x = x_k$, we have

$$\phi((1 - t_k)x_k + t_k y_k) \leq \phi(x_k) - t_k S(x_k) + \frac{t_k^2 L}{2}\|y_k - x_k\|^2.$$

1) If the predefined stepsize is used, i.e., $t_k = \alpha_k$, then

$$\phi((1 - \alpha_k)x_k + \alpha_k y_k) \leq \phi(x_k) - \alpha_k S(x_k) + \frac{\alpha_k^2 L}{2}\|y_k - x_k\|^2.$$

2) If the adaptive stepsize is used, i.e., $t_k = \beta_k$, then

$$\beta_k = \operatorname{argmin}_{t\in[0,1]} \left\{ -tS\left(x_k\right) + \frac{t^2 L}{2} \left\|y_k - x_k\right\|^2 \right\},$$

and hence

$$\begin{aligned}\phi((1-\beta_k)x_k + \beta_k y_k) &\leq \phi(x_k) - \beta_k S(x_k) + \frac{\beta_k^2 L}{2}\|y_k - x_k\|^2 \\ &\leq \phi(x_k) - \alpha_k S(x_k) + \frac{\alpha_k^2 L}{2}\|y_k - x_k\|^2.\end{aligned}$$

3) If the exact minimization/line search is used, i.e., $t_k = \eta_k$, then

$$\begin{aligned}\phi((1-\eta_k)x_k + \eta_k y_k) &\leq \phi((1-\alpha_k)x_k + \alpha_k y_k) \\ &\leq \phi(x_k) - \alpha_k S(x_k) + \frac{\alpha_k^2 L}{2}\|y_k - x_k\|^2.\end{aligned}$$

In any case, we have

$$\phi(x_{k+1}) \leq \phi(x_k) - \alpha_k S(x_k) + \frac{\alpha_k^2 L}{2}\|y_k - x_k\|^2.$$

Consider a sequence of averages of $\nabla f(x_k)$ defined as $u_0 = \nabla f(x_0)$ and

$$u_{k+1} = (1-\alpha_k)u_k + \alpha_k \nabla f(x_k), \quad \forall k \geq 0.$$

Since the dual function ψ is concave,

$$\psi(u_{k+1}) \geq (1-\alpha_k)\psi(u_k) + \alpha_k \psi(\nabla f(x_k)), \quad \forall k \geq 0. \tag{4}$$

Using Lemma 3 and the above inequality, we have

$$\begin{aligned}\phi(x_{k+1}) &\leq \phi(x_k) - \alpha_k[\phi(x_k) - \psi(\nabla f(x_k))] + \frac{\alpha_k^2 L}{2}\|y_k - x_k\|^2 \\ &\overset{(4)}{\leq} (1-\alpha_k)\phi(x_k) + \psi(u_{k+1}) - (1-\alpha_k)\psi(u_k) + \frac{\alpha_k^2 L}{2}\|y_k - x_k\|^2.\end{aligned}$$

Rearranging the terms, we have

$$\phi(x_{k+1}) - \psi(u_{k+1}) \leq (1-\alpha_k)[\phi(x_k) - \psi(u_k)] + \frac{\alpha_k^2 L D^2}{2}. \tag{5}$$

Clearly, to prove (3), it suffices to show

$$\phi(x_k) - \psi(u_k) \leq \frac{2LD^2}{k+1}. \tag{6}$$

It follows from (5) and the definition of α_k that

$$\begin{aligned}\phi(x_{k+1}) - \psi(u_{k+1}) &\le (1-\alpha_k)[\phi(x_k) - \psi(u_k)] + \frac{\alpha_k^2 LD^2}{2} \\ &= \frac{k}{k+2}[\phi(x_k) - \psi(u_k)] + \frac{2LD^2}{(k+2)^2}.\end{aligned}$$

Hence,

$$\begin{aligned}(k+1)(k+2)[\phi(x_{k+1}) - \psi(u_{k+1})] &\le k(k+1)[\phi(x_k) - \psi(u_k)] + \frac{2(k+1)LD^2}{k+2} \\ &\le k(k+1)[\phi(x_k) - \psi(u_k)] + 2LD^2.\end{aligned}$$

Summing over the iterations, we have

$$k(k+1)[\phi(x_k) - \psi(u_k)] \le 2kLD^2,$$

and thus (6) holds. □

The above proof also suggests the following primal-dual Frank-Wolfe method.

Algorithm 3 Primal-dual Frank-Wolfe method

Input: Initial point $x_0 \in \operatorname{dom} h$ and $u_0 = \nabla f(x_0)$

for $k \ge 0$ **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 $x_{k+1} = (1-t_k)x_k + t_k y_k$ and $u_{k+1} = (1-\alpha_k)u_k + \alpha_k \nabla f(x_k)$.

end for

The primal-dual convergence is given by (6), which also implies (3).

3 Duality between Frank-Wolfe and mirror descent

We present a fascinating connection between Frank-Wolfe and mirror descent, that is, Frank-Wolfe applied to the dual problem is equivalent to mirror descent applied to the primal problem. We consider the following primal and dual problems.

Primal

$$\min_{x\in\mathbb{R}^n}\{\phi(x) := f(Ax) + h(x)\}$$

and dual

$$\max_{y\in C}\{\psi(y) := -h^*(-A^\top y) - f^*(y)\}.$$

Assuming h is μ-strongly convex, then h^* is smooth. We also assume that $A^\top \mathrm{dom}\, f^*$ is bounded (i.e., f is Lipschitz continuous), that is

$$R = \max_{y_1, y_2 \in \mathrm{dom}\, f^*} \|A^\top (y_1 - y_2)\|_* = \mathrm{diam}(A^\top \mathrm{dom}\, f^*). \tag{7}$$

Applying Algorithm 2 to the dual problem, we have the following dual Frank-Wolfe method.

Algorithm 4 Frank-Wolfe method for dual problem

Input: Initial point $y_0 \in \mathrm{dom}\, f^*$

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

Step 1. Compute $x_k = \mathrm{argmin}_{x \in \mathbb{R}^n} \{\langle x, A^\top y_k \rangle + h(x)\} = \nabla(h^*)(-A^\top y_k)$.

Step 2. Compute $\bar{y}_k \in \mathrm{Argmax}_{y \in C} \{\langle y, Ax_k \rangle - f^*(y)\} = \partial f(Ax_k)$.

Step 3. Choose $t_k \in [0, 1]$ and set $y_{k+1} = (1 - t_k) y_k + t_k \bar{y}_k$.

end for

Theorem 1 directly gives the following convergence result for the dual problem.

Theorem 2. *For every $k \geq 1$, we have*

$$\psi^* - \psi(y_k) \leq \frac{2R^2}{\mu(k+1)}.$$

If we add an auxiliary "primal average" of x_k in Algorithm 4, then we can prove a similar primal-dual convegence guarantee as in (6) for the dual problem.

3.1 Mirror descent

Consider the primal problem

$$\min_{x \in \mathbb{R}^n} \{\phi(x) := f(Ax) + h(x)\},$$

we present the following special mirror descent method for the primal problem.

Algorithm 5 Mirror descent for primal problem

Input: Given $y_0 \in \mathrm{dom}\, f^*$, set initial point $x_0 = \nabla(h^*)(-A^\top y_0)$ and $h'(x_0) = -A^\top y_0$.

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

Step 1. Choose $t_k \in [0, 1]$ and compute $x_{k+1} = \mathrm{argmin}_{x \in \mathbb{R}^n} \left\{ \ell_\phi(x; x_k) + \frac{1}{t_k} D_h(x, x_k) \right\}$.

Step 2. Set $h'(x_{k+1}) = (1 - t_k) h'(x_k) - t_k A^\top f'(Ax_k)$.

end for

Note that we linearize the whole primal function ϕ and use the μ-strongly convex function h as the distance generating function.

The following theorem show that the dual Frank-Wolfe method is equivalent to the above mirror descent method.

Theorem 3. *If both Algorithms 4 and 5 use the same subgradient oracle of f, i.e., $\bar{y}_k = f'(Ax_k)$ where $f'(Ax_k)$ is the one used in Step 1 of Algorithm 5, then given the same initial point $y_0 \in \operatorname{dom} f^*$, both algorithms generate same iterates $\{x_k\}$.*

Proof. It follows from Step 1 of Algorithm 5 that

$$0 \in t_k \left(A^\top f'(Ax_k) + h'(x_k) \right) + \partial h(x_{k+1}) - h'(x_k),$$

and hence that

$$0 \in -(1-t_k)h'(x_k) + t_k A^\top f'(Ax_k) + \partial h(x_{k+1}).$$

This is equivalent to

$$\partial h(x_{k+1}) \ni (1-t_k)h'(x_k) - t_k A^\top f'(Ax_k).$$

Using Theorem 3 of Lecture 3, we have

$$x_{k+1} \in \partial h^* \left((1-t_k)h'(x_k) - t_k A^\top f'(Ax_k) \right).$$

Since h is strongly convex, we know h^* is smooth and $\partial h^* = \nabla h^*$. This means x_{k+1} is unique

$$x_{k+1} = \nabla h^* \left((1-t_k)h'(x_k) - t_k A^\top f'(Ax_k) \right). \tag{8}$$

Next, we consider Algorithm 4 and prove that $-A^\top y_k$ from Algorithm 4 is equal to $h'(x_k)$ from Algorithm 5, i.e.,

$$-A^\top y_k = h'(x_k). \tag{9}$$

We prove this relation by induction. It clearly holds for $k=0$ in view of the input of Algorithm 5. Suppose (9) holds for some $k \geq 0$. Then, it follows from Step 3 of Algorithm 4 and the assumption that $\bar{y}_k = f'(Ax_k)$ that

$$-A^\top y_{k+1} = -(1-t_k)A^\top y_k - t_k A^\top \bar{y}_k = (1-t_k)h'(x_k) - t_k A^\top f'(Ax_k) = h'(x_{k+1}),$$

where the last identity is due to Step 2 of Algorithm 5. Hence, we prove (9).

Now, using Step 2 of Algorithm 5 and (9), we conclude that (8) is equivalent to

$$x_{k+1} = \nabla h^* \left(-A^\top y_{k+1} \right),$$

which agrees with Step 1 of Algorithm 4. Therefore, we finally prove that dual Frank-Wolfe and mirror descent are equivalent. □

Recall that Algorithm 3 has a primal-dual pair (x_k, u_k) and we can show primal-dual convergence (6), which also implies both primal convergence (Theorem 1) and dual convergence (Theorem 2). We prove that Algorithm 5 is the dual to Algorithm 4, hence we also want to derive a "dual" to Theorem 2, which will be a primal convergence result similar to Theorem 1. The following theorem is such a result as we show convergence of an average point.

Theorem 4. *If we choose $t_k = \alpha_k$ in Algorithm 5, then*

$$\phi(\bar{x}_k) - \phi_* + D_h(x_*, x_k) \le \frac{2R^2}{\mu(k+1)}, \tag{10}$$

where

$$\bar{x}_k = \frac{2}{k(k+1)} \sum_{i=1}^{k} i x_{i-1}.$$

Proof. Similar to the proof of mirror descent for Lemma 2 of Lecture 3, we have

$$\ell_\phi(x; x_k) + \frac{1}{\alpha_k} D_h(x, x_k) \ge \ell_\phi(x_{k+1}; x_k) + \frac{1}{\alpha_k} D_h(x_{k+1}, x_k) + \frac{1}{t_k} D_h(x, x_{k+1}).$$

Using convexity of f and the definition of Bregman divergence D_h, we have

$$\phi(x) \ge \ell_f(x; x_k) + h(x) = \ell_\phi(x; x_k) + D_h(x, x_k).$$

Combining the above two inequalities, we have

$$\phi(x) + \left(\frac{1}{\alpha_k} - 1\right) D_h(x, x_k) \ge \ell_\phi(x_{k+1}; x_k) + \frac{1}{\alpha_k} D_h(x_{k+1}, x_k) + \frac{1}{t_k} D_h(x, x_{k+1}).$$

Since h is μ-strongly convex, we know

$$D_h(x_{k+1}, x_k) \ge \frac{\mu}{2} \|x_{k+1} - x_k\|^2.$$

Thus, it follows that

$$\phi(x) + \left(\frac{1}{\alpha_k} - 1\right) D_h(x, x_k) \ge \ell_\phi(x_{k+1}; x_k) + \frac{\mu}{2\alpha_k} \|x_{k+1} - x_k\|^2 + \frac{1}{t_k} D_w(x, x_{k+1}).$$

Rearranging the terms and using the Cauchy-Schwarz inequality, we obtain

$$\phi(x_k) - \phi(x) \le \left(\frac{1}{\alpha_k} - 1\right) D_h(x, x_k) - \frac{1}{\alpha_k} D_h(x, x_{k+1}) + \|\phi'(x_k)\|_* \|x_{k+1} - x_k\| - \frac{\mu}{2\alpha_k} \|x_{k+1} - x_k\|^2. \tag{11}$$

Recalling (9) from the proof of Theorem 3, we know

$$h'(x_k) \in -A^\top \mathrm{dom}\, f^*.$$

Hence,

$$\|\phi'(x_k)\|_* = \|A^\top f'(Ax_k) + h'(x_k)\|_* \le \max_{y_1, y_2 \in \mathrm{dom}\, f^*} \|A^\top (y_1 - y_2)\|_* = R.$$

Plugging the above bound into (11), we have

$$\phi(x_k) - \phi(x) \le \left(\frac{1}{\alpha_k} - 1\right) D_h(x, x_k) - \frac{1}{\alpha_k} D_h(x, x_{k+1}) + \frac{\alpha_k R^2}{2\mu}.$$

Using the definition of α_k, we have

$$(k+1)[\phi(x_k) - \phi_*] \leq \frac{k(k+1)}{2} D_h(x_*, x_k) - \frac{(k+1)(k+2)}{2} D_h(x_*, x_{k+1}) + \frac{(k+1)R^2}{\mu(k+2)}.$$

Summing the above inequality, we obtain

$$\sum_{i=0}^{k-1} (i+1)[\phi(x_i) - \phi_*] \leq -\frac{k(k+1)}{2} D_h(x_*, x_k) + \frac{kR^2}{\mu}.$$

Finally, we conclude that (10) holds. $\square$