Insights, resources, and application of “Revisting Frank-Wolfe"

BY YALIN LIAO

1 Frank-Wolfe Algorithm

Frank-Wolfe can be applied to solve the constrained convex optimization problem

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

where the objective function f is convex and continuously differentiable and the domain D[1] is a compact convex subset of any vector space.

Why is Frank-Wolfe meaningful? Assume $x = x^{(k)}$ and we are searching for a new point $x^{(k+1)}$ such that

$$f(x^{(k+1)}) \leq f(x^{(k)})$$

For very small $r > 0, \forall s \in B(x^{(k)}, r)$,

$$f(s) \approx f(x^{(k)}) + \langle s - x, \nabla f(x^{(k)}) \rangle$$

so we switch to minimize[2]

$$\min_{s \in D} f(x^{(k)}) + \langle s - x^{(k)}, \nabla f(x^{(k)}) \rangle$$

or[3]

$$\min_{s \in D} \langle s, \nabla f(x^{(k)}) \rangle$$

which is a generalization form of

$$\min_{\|s\| \leq 1} \langle s, \nabla f(x^{(k)}) \rangle$$

whose solution $s = -\frac{\nabla f(x^{(k)})}{\|\nabla f(x^{(k)})\|}$, is exactly the descent direction, used in gradient descent algorithm. After obtaining the descent direction

$$s = \arg\min_{s \in D} \langle s, \nabla f(x^{(k)}) \rangle$$

1. The convex constraint D should be bounded otherwise the solution s may not exist in Frank-Wolfe algorithm. In R^n, a set is compact if and only if it is a closed bounded set.

2. The idea is the same as deriving gradient descent (GD). In GD algorithm, we could think the convex constraint is the unit ball, i.e., $D = \{x \mid \|x\| \leq 1\}$ and the solution of

$$\min_{\|s\| \leq 1} \langle s, \nabla f(x^{(k)}) \rangle$$

is $s = -\frac{\nabla f(x^{(k)})}{\|\nabla f(x^{(k)})\|}$. GD and Frank-Wolfe algorithms share the same idea to find the descent direction. But the update rules are slightly different.

3. Since $f(x^{(k)})$ and $-\langle x^{(k)}, \nabla f(x^{(k)}) \rangle$ are constant with respect to s, two optimization problems are equivalent.

What's the meaning of s? It is the vector in the constraint D, which is mostly similar to $-\nabla f(x^{(k)})$.

Frank-Wolfe algorithm updates the solution as follows

$$x^{(k+1)} := (1-\gamma)x^{(k)} + \gamma s, \text{for } \gamma := \frac{2}{k+2}$$

One question: can we update x as in GD?

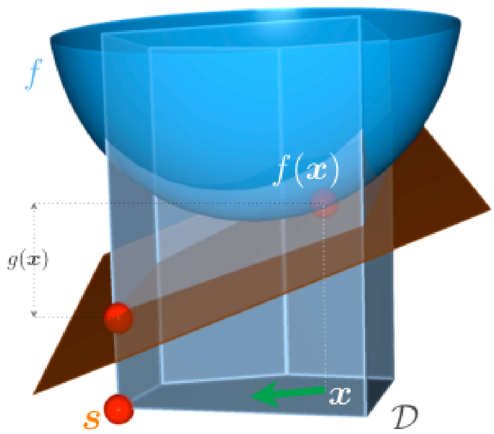
No. If $x^{(k+1)} = x^{(k)} + \eta s$ for small $\eta > 0$, then $x^{(k+1)}$ is not garanteed to be in D.

Another question: Is such s always a descent direction for $f(x)$ at $x^{(k)}$?

No. Surely, $s = \arg\min_{s \in D} \langle s, \nabla f(x^{(k)}) \rangle$ is always a descent direction for

$$L(s) = f(x^{(k)}) + \langle s - x^{(k)}, \nabla f(x^{(k)}) \rangle$$

at $x^{(k)}$. But if $s = \arg\min_{s \in D} \langle s, \nabla f(x^{(k)}) \rangle$ could be far away from $x^{(k)}$, $L(s)$ cannot approximate $f(s)$ well. So s may not be a descent direction for $f(x)$ at $x^{(k)}$. This can be seen from the figure 1 in the paper,



So in Frank-Wolfe algorithm, we should choose γ small enough so that hopefully γs is a descent direction. Maybe we can consider the following optimization problem,

$$s = \arg\min_{s \in D} \langle s, \nabla f(x^{(k)}) \rangle + \lambda \|x - x^{(k)}\|_2^2$$

to determine the descent direction where $\lambda > 0$ is a hyperparameter.

2 The Duality Gap

We denote x^* as the optimal solution, i.e.,

$$f(x^*) \le f(x) \, \forall x \in D$$

The surrogate duality gap

$$g(x) := \max_{s \in D} \langle x - s, \nabla f(x) \rangle$$

Can we interprate $g(x)$ differently?

$$
\begin{aligned}
g(x) &= \max_{s\in D} \langle x-s, \nabla f(x)\rangle \\
&= \max_{s\in D} - \langle s-x, \nabla f(x)\rangle \\
&= -\min_{s\in D} \langle s-x, \nabla f(x)\rangle \\
&= -\min_{s\in D} f(x) + \langle s-x, \nabla f(x)\rangle - f(x) \\
& \quad f(x) \,\text{is}\, a \,\text{constant in term of}\, s \\
&= f(x) - \min_{s\in D} f(x) + \langle s-x, \nabla f(x)\rangle \\
& \quad f(x) + \langle s-x, \nabla f(x)\rangle \leq f(s) \,\text{as}\, f \,\text{is convex} \\
&\geq f(x) - \min_{s\in D} f(s) \\
&= f(x) - f(x^*)
\end{aligned}
$$

So $g(x)$ is an upper bound of the gap between $f(x)$ and $f(x^*)$. Assume $x = x^{(k)}$ after k iteration for any optimization algorithm. If we know $g(x^{(k)})$ is very small, we could say $x^{(k)}$ is one optimizer even though x^* is unknown.