IE 598: Big Data Optimization **Fall 2016**

Lecture 11: Conditional Gradient (Frank-Wolfe Algorithm) - September 27

Lecturer: Niao He *Scriber: Cheng-Yang (Peter) Liu*

Overview: In this lecture we are going to discuss a new gradient-based and projection-free algorithm, called Conditional Gradient Method, also known as Frank-Wolfe Algorithm. The algorithm was initially proposed by Frank and Wolfe in 1956 for solving quadratic programming problems with linear constraints, but regained many interests and new developments in machine learning community over the last few days. In this lecture, we are going to discuss the motivation, algorithm itself, convergence analysis, lower bound, and some recent advances.

11.1 Introduction

In the previous lecture, we discussed how to solve constrained convex optimization problems

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

by Projected Gradient Descent, which requires to compute the projection on a convex set at each iteration. Recall the definition of projection of a vector y on a convex set X is defined as $\Pi_X(y) = \text{argmin}_{x\in X} \frac{1}{2}\|y-x\|_2^2$. Here are some examples.

- **Simplex:** $X = \{x \in \mathbb{R}^n : x \geq 0, \sum x_i \leq 1\}$. The projection is given by

$$\Pi_X(y) = \begin{cases} y & y \in X \\ (y-\mu_\star)_+ & y \notin X \end{cases}$$

  where μ_* is the Lagrange multiplier such that $\sum_i (y_i - \mu_*)_+ = 1$.
  The simplex domain is widely used in many applications including portfolio management, LASSO, boosting, MAP inference, density estimation, structured SVM, etc.

- **Nuclear norm ball:** $X = \{x \in \mathbb{R}^{m\times n}, \|x\|_{nuc} = \|\sigma(x)\|_1 \leq 1\}$. The projection is given by

$$\Pi_X(y) = \begin{cases} y & y \in X \\ \sum_i (\sigma_i - \mu_\star)_+ u_i v_i^T & y \notin X, \quad \text{where } y = \sum_i \sigma_i u_i v_i^T \end{cases}$$

  where μ_* is the Lagrange multiplier such that $\sum_i (\sigma_i - \mu_*)_+ = 1$.
  The nuclear norm domain is widely used in many recent applications including robust PCA, matrix completion, network estimation, etc.

- **Polyhedron:** $X = \{x \in \mathbb{R}^n : Ax \leq b\}$. The projection in this case simply reduces to solving a convex quadratic program (QP).

Apparently, there are advantages and disadvantages when using Projected Gradient Descent. The pros are:

(i) In many cases, computing the projection often admits closed-form solution or easy-to-solve subproblems, e.g. ℓ_2-ball, convex cone, halfspaces, box, simplex, etc.

(ii) Projected Gradient Descent enjoyes the same convergence rates and similar analysis as the unconstrained case.

On the other hand, the potential cons are that:

(i) For high-dimensional problems, computing the projection can be computational expensive. e.g., it requires $\mathcal{O}(n)$ for simplex domain, $\mathcal{O}(mn^2 \wedge nm^2)$ for nuclear norm domain, and iteratively solving a QP for polyhedron domains.

(ii) Full gradient updating may destroy certain structure, such as sparsity and low rank.

To avoid the expense of projection, an alternative is to use linear minimization over the convex set instead!

Definition 11.1 (Linear minimization oracle) *For a given vector y, the linear minimization oracle over a convex set X is defined as*

$$LMO_X(y) = \operatorname*{argmin}_{x \in X} y^T x.$$

Let us revisit those examples.

- **Simplex:** $X = \{x \in \mathbb{R}^n : x \geq 0, \sum x_i = 1\}$. The linear minimization oracle is given by

  $$\text{LMO}_X(y) = e_i, \quad i = \operatorname*{argmin}_{k=1,\ldots,n}(y_k)$$

  which has only one non-zero entry and is much simpler than projection.

- **Nuclear norm ball:** $X = \{x \in \mathbb{R}^{m \times n}, \|x\|_{nuc} = \|\sigma(x)\|_1 \leq 1\}$. The linear minimization oracle is given by

  $$\text{LMO}_X(y) = -u_i v_i^T, \quad i = \operatorname*{argmax}_{k=1,\ldots,n}(\sigma_k)$$

  which is a rank one matrix and requires only to calculate the top singular value and corresponding singular vectors. This can be efficiently done via power iteration.

- **Polyhedron:** $X = \{x \in \mathbb{R}^n : Ax \leq b\}$. The linear minimization reduces to solving a linear program (LP) instead of QP.

11.2 Generic Conditional Gradient Method

We consider the constrained convex optimization problem

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

where f is L-smooth and convex, set X is convex and compact. We denote the diameter of the set X as $D = \max_{x,y \in X} \|x - y\|_2$, which is finite since X is compact.

Conditional Gradient method. The algorithm works as follows. We start with an initial solution x_0 and at each iteration, compute the solution $\hat{x}_t$ which minimizes the linear approximation $f(x_t)+\nabla f(x_t)^T(x-x_t)$ of objective function. Then, we update x_t to x_{t+1} such that it is at least as good as the intermediate point $\gamma_t\hat{x}_t + (1-\gamma_t)x_t$, namely, $f(x_{t+1}) \leq f(\gamma_t\hat{x}_t + (1-\gamma_t)x_t)$.

- *Algorithm*:
  - Initialize $x_0 \in X$
  - Compute $\hat{x}_t = \text{LMO}_X(\nabla f(x_t)) = \text{argmin}_{x\in X}\langle x, \nabla f(x_t)\rangle$
  - Find $x_{t+1} \in X$, s.t. $f(x_{t+1}) \leq f(\gamma_t\hat{x}_t + (1-\gamma_t)x_t)$, with $\gamma_t = \frac{2}{t+2}$

  There are several different ways to update x_{t+1} at each iteration.
- *Updating rules*
  - fixed step size: $x_{t+1} = \gamma_t\hat{x}_t + (1-\gamma_t)x_t$
  - line search: $x_{t+1} = \text{argmin}_{0\leq\gamma\leq 1} f(\gamma\hat{x}_t + (1-\gamma)x_t)$
  - fully-corrective search: $x_{t+1} = \text{argmin}_{x\in conv\{x_0,\hat{x}_1,\ldots,\hat{x}_t)} f(x)$

  There is clearly some tradeoff between the computation cost of updating x_{t+1} vs. the reduction of objective function value. The more effort you pay for each iteration, most likely you will get more reduction, and the algorithm will converge faster.

In contrast to Projected Gradient Descent, Conditional Gradient method requires cheaper iteration cost and maintains desirable structure of the solution such as sparsity and low rank.

11.3 Convergence and Lower Bound Complexity

Theorem 11.2 (Convergence) *Let f be L-smooth and convex, X be convex compact with diameter $D > 0$. The above Conditional Gradient method satisfies*

$$f(x_{t+1}) - f(x_*) \leq \frac{2LD^2}{t+2}, \quad \forall t \geq 1$$

Proof: By convexity, we have

$$f(x) \geq f(x_t) + \nabla f(x_t)^T(x - x_t), \forall x \in X.$$

Taking minimum on both side leads to

$$f(x_*) \geq f(x_t) + \nabla f(x_t)^T(\hat{x}_t - x_t).$$

Hence, we have

$$f(x_t) - f(x_*) \leq \nabla f(x_t)^T(x_t - \hat{x}_t).$$

Let us denote $\tilde{x}_{t+1} = \gamma_t \hat{x}_t + (1-\gamma_t)x_t$ as the intermediate point. By definition of x_{t+1}, we have

$$\begin{aligned} f(x_{t+1}) - f(x_t) &\leq f(\tilde{x}_{t+1}) - f(x_*) \\ &\leq f(x_t) + \nabla f(x_t)^T(\tilde{x}_{t+1} - x_t) + \tfrac{L}{2}\|\tilde{x}_{t+1} - x_t\|^2 - f(x_*) \\ &\leq f(x_t) + \gamma_t \nabla f(x_t)^T(\hat{x}_t - x_t) + \tfrac{L\gamma_t^2}{2}\|\hat{x}_t - x_t\|_2^2 - f(x_*) \\ &\leq f(x_t) - f(x_*) - \gamma_t(f(x_t) - f(x_*)) + \tfrac{L{\gamma_t}^2}{2}D^2 \end{aligned}$$

Define $\epsilon_t = f(x_t) - f_*$, we arrive at

$$\epsilon_{t+1} \leq (1-\gamma_t)\epsilon_t + \frac{L{\gamma_t}^2}{2}D^2$$

By induction, we can show that $\epsilon_t \leq \frac{2LD^2}{t+2}$. First of all, when $t = 1$, $\gamma_0 = 1$, we have

$$\epsilon_1 \leq (1-\gamma_0)\epsilon_0 + \frac{LD^2}{2}\gamma_0^2 = \frac{LD^2}{2} \leq \frac{3}{2}LD^2.$$

Now assume it holds for $t \geq 1$, that $\epsilon_t \leq \frac{2LD^2}{t+2}$, then

$$\begin{aligned} \epsilon_{t+1} &\leq (1 - \tfrac{2}{t+2}) \cdot \tfrac{2LD^2}{t+2} + \tfrac{LD^2}{2}(\tfrac{2}{t+2})^2 \\ &= \tfrac{2LD^2(t+1)}{(t+2)^2} \\ &\leq \tfrac{2LD^2}{t+3}, \quad \text{since } (t+2)^2 \geq (t+1)(t+3). \end{aligned}$$

■

Remark. The above result implies that for smooth convex problems, Conditional Gradient method achieves an overall $O(LD^2/t)$ convergence rate. Recall that when applied with Nesterov's accelerated gradient descent, we obtain a $O(LD^2/t^2)$ rate for smooth convex problems and a linear $O((\frac{1-\sqrt{\kappa}}{1+\sqrt{\kappa}})^t)$ rate for smooth strongly convex problems. Does this imply conditional gradient is sub-optimal? Can we improve the convergence results for strongly convex problems? The following theorem on lower bound complexity suggests that the answer is no.

Theorem 11.3 (Lower Bound) *For any $x_0 \in \mathbb{R}^n$, $\quad 1 \leq t \leq \frac{n}{2} - 1$, then $\exists$ a L-smooth convex function f and a convex set X with diameter D s.t. for any algorithm $\mathcal{M}$ that only computes local gradients of f and do linear minimization over X, we have*

$$f(x_t) - f_* \geq \frac{LD^2}{8(t+1)}, \ \forall t \geq 1.$$

Proof: We construct the case when

$$f(x) = \frac{L}{2}\|x\|_2^2$$

$$X = \{x \in \mathbb{R}^n : x \geq 0, \sum x_i = D_0\}, \text{ where } D_0 = \frac{D}{\sqrt{2}}.$$

Hence, f is L-smooth and convex, X is convex compact with diameter D. Consider the optimization problem

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

the optimal solution is given by $x_* = (\frac{D_0}{n}, \ldots, \frac{D_0}{n})$ and the optimal value is $f(x_*) = \frac{LD_0^2}{2n}$.

Without loss of generality, let $x_0 = D_0 e_1$. Hence, by induction, we can see that

$$x_t = conv(D_0 e_1, D_0 e_{p_1}, \ldots, D_0 e_{p_t})$$

for some index $p_1, \ldots, p_t$. Therefore,

$$f(x_t) \geq \min_{x \in D_0 conv(e_0, e_{p_1}, \ldots, e_{p_t})} f(x) \geq \frac{LD_0^2}{2(t+1)}$$

Since $n \geq 2(t+1)$, we further have

$$\begin{aligned} f(x_t) - f(x_*) &\geq \frac{LD_0^2}{2}(\frac{1}{t+1} - \frac{1}{n}) \\ &\geq \frac{LD_0^2}{2}(\frac{1}{t+1} - \frac{1}{2(t+1)}) \\ &\geq \frac{LD_0^2}{4(t+1)} = \frac{LD^2}{8(t+1)} \end{aligned}$$

which proves the theorem. ■

Remark. The above theorem implies that

1. CG is indeed optimal among all gradient-based algorithms that only use linear minimization oracles over the domain.
2. The convergence rate $\mathcal{O}(\frac{1}{t})$ cannot be improved even assuming strong convexity.

Recent advances. We mention a few recent works on conditional gradient methods here:

1. **Linear convergence for some strongly convex problems**
   Although we show that in the worst case, we cannot improve the sublinear rate for strongly convex cases, it is shown that for specific situation (e.g., domain is given by a polyhedron) or under additional assumptions (e.g., optimum lies in the interior), linear rate of convergence can be established. See recent works [BS16] and [LJ15].
2. **Improvement over gradient oracles**
   Note that in order to achieve an ϵ-accuracy solution, the conditional gradient requires $\mathcal{O}(1/\epsilon)$ number of calls to compute the gradient. A recent algorithm, called *conditional gradient sliding*, is able to improve this to optimal complexity bounds, i.e., $\mathcal{O}(1/\sqrt{\epsilon})$ for smooth convex problems and $\mathcal{O}(\log(\frac{1}{\epsilon}))$ for smooth strongly convex problems, respectively. See details in [LY14].

References

[J13] M. Jaggi, "Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization," *ICML*, 2013.

[HJN14] Z. Harcaoui, A. Juditsky and A. Nemirovski, "Conditional gradient algorithms for norm-regularized smooth convex optimization," *Mathematical Programming*, 1–32, 2014.

[BS16] A. Beck and S. Shtern, "Linearly convergent away-step conditional gradient for non-strongly convex functions," *Mathematical Programming*, 1–27, 2016.

[LJ15] S. Lacoste-Julien and M. Jaggi, "On the Global Linear Convergence of Frank-Wolfe Optimization Variants," *NIPS*, 2015.

[GH12] D. Garber and E. Hazan, "Playing non-linear games with linear oracles," *In 54th Annual IEEE Symposium on Foundations of Computer Science*, FOCS, 2013a.

[LY14] G. Lan and Z. Yi, "Conditional gradient sliding for convex optimization," *Optimization-Online preprint*, 4605, 2014.