Indian Institute of Technology Kanpur
CS777 Topics in Learning Theory

Scribe: Afroz Alam
Instructor: Purushottam Kar
Date: March 7, 2018

LECTURE
15

Optimization Techniques - II

Introduction

intro

We continue to discuss two optimization algorithms namely projected gradient descent and Frank-Wolfe algorithm that work to minimize a convex objective function on a convex set.

Both algorithms solve an optimization problem w.r.t. $\mathcal{C}$. Algorithm 1 solves it in line 4 at the projection step and algorithm 2 solves it in line 3 to find the argmin of the inner product. The preference of an algorithm depends on how difficult it is to solve the optimization step in the given $\mathcal{C}$.

Algorithm 1: Projected Gradient Descent

Input: Convex objective function f, step lengths η_t, convex set $\mathcal{C}$
Output: $\hat{\mathbf{x}} \in \mathcal{C}$ such that $\hat{\mathbf{x}}$ is close to optimum

1: $\mathbf{x}^0 \leftarrow$ INIT
2: **for** $t = 1, 2, \ldots, T-1$ **do**
3: $\quad \mathbf{z}^{t+1} \leftarrow \mathbf{x}^t - \eta_t \cdot \nabla f(\mathbf{x}^t)$
4: $\quad \mathbf{x}^{t+1} \leftarrow \Pi_{\mathcal{C}}(\mathbf{z}^{t+1})$
5: **end for**
6: $\hat{\mathbf{x}} \leftarrow \text{SELECT}(\mathbf{x}^0, \mathbf{x}^1, \ldots, \mathbf{x}^T)$
7: **return** $\hat{\mathbf{x}}$

Now, let us consider Frank-Wolfe algorithm more closely.

Frank-Wolfe Algorithm

Example 15.1. Consider a simplified version of the SVM(dual) problem.

$$\max_{0 \le \alpha \le c} \alpha^\intercal \mathbb{1} - \frac{1}{2}\alpha^\intercal \mathbf{Q}\alpha \qquad \text{where } \mathbf{Q}_{ij} = y^i y^j k(\mathbf{x}^i, \mathbf{x}^j)$$

The step 3 of algorithm 2 solves the following

$$\max_{0 \le \mathbf{z}_i \le c} \langle g, \mathbf{z} \rangle \qquad \text{where } g = \mathbb{1} - \mathbf{Q}\alpha$$

Algorithm 2: Frank-Wolfe algorithm

Input: Convex objective function f, step lengths η_t, convex set $\mathcal{C}$
Output: $\hat{\mathbf{x}} \in \mathcal{C}$ such that $\hat{\mathbf{x}}$ is close to optimum

1: $\mathbf{x}^0 \leftarrow$ INIT
2: **for** $t = 1, 2, \ldots, T-1$ **do**
3: $\quad \mathbf{z}^{t+1} \leftarrow \arg\min_{\mathbf{z} \in \mathcal{C}} \left\langle \nabla f(\mathbf{x}^t), z \right\rangle$
4: $\quad \mathbf{x}^{t+1} \leftarrow (1 - \eta_t) \cdot \mathbf{x}^t + \eta_t \cdot \mathbf{z}^t$
5: **end for**
6: $\hat{\mathbf{x}} \leftarrow \mathbf{x}^T$
7: **return** $\hat{\mathbf{x}}$

The above optimization problem has a closed form solution

$$\mathbf{z} = \begin{cases} 0 & g_i \leq 0 \\ c & g_i > 0 \end{cases}$$

Exercise 15.1. Solve the following optimization problem:

$$\min_{\substack{b \geq \mathbf{0} \\ b \geq \lambda y + z}} b^\intercal \mathbb{1} \qquad \text{where } b \in \mathbb{R}^n, \lambda \in \mathbb{R}$$

Frank-Wolfe algorithm is useful in problems where $\mathcal{C}$ is the convex hull of a set of possibly infinite points ($\mathcal{A}$). In such conditions, the projection step is generally not possible to compute. We will soon show that this condition gives nice properties to the output of the algorithm.

$$\begin{aligned} \mathcal{C} &= \text{conv}(\mathcal{A}) \qquad \text{(convex hull)} \\ \mathcal{A} &= \{a_1, a_2, \ldots, a_n, \ldots\} \end{aligned}$$

Example 15.2. Unit ball defined w.r.t. *L1-norm*

$$\begin{aligned} \mathcal{B}_1(1) &= \{\mathbf{x} : \|\mathbf{x}\|_1 \leq 1\} \\ \mathcal{C} &= \text{conv}\{\pm e_i : e_i \in \mathbb{R}^n, e_i \text{ is one-hot}\} \end{aligned}$$

Example 15.3. Unit ball defined w.r.t. *Trace-norm*

$$\begin{aligned} \mathcal{B}_{tr}(1) &= \{\mathcal{A} : tr(\mathcal{A}) \leq 1\} \\ \mathcal{C} &= \text{conv}\{UU^\intercal : U \in \mathbb{R}^n, \|U\|_2 = 1\} \end{aligned}$$

For a generalized norm ($\|\cdot\|$) the consider the set $\mathcal{C} = \{\mathbf{x} : \|\mathbf{x}\| \leq r\}$. Then, the following optimization problem is of interest to us

$$\begin{aligned} \min_{\|\mathbf{x}\| \leq r} \langle g, \mathbf{x} \rangle &= -r \max_{\|\mathbf{x}\| \leq 1} \langle -g, \mathbf{x} \rangle \\ &= -r \underbrace{\|-g\|_*}_{\text{dual norm}} \end{aligned}$$

Dual Norm

The dual norm

Some properties of the output from Frank-Wolfe algorithm which has $\mathcal{C}$ of the above have been discussed in the following section.

Frank-Wolfe Properties

Sparsity

A consequence of $\mathcal{C} = \text{conv}(\mathcal{A})$ is that the solution of optimization problem of the form

$$\arg\min_{\mathbf{z}\in\mathcal{C}} \langle g, \mathbf{z}\rangle \in \mathcal{A}$$

Also, since $\mathbf{x}^\intercal = \text{LIN}(\mathbf{x}^0, \mathbf{z}^1, \mathbf{z}^2, \ldots, \mathbf{z}^{\intercal-1})$, we have $x^\intercal = \text{LIN}\{a : a \in \mathcal{A}\}$. After T iterations, there can only be T entries that are non-zero in the output vector therefore the following guarantees hold:

- $\mathcal{B}_1(1) : \|\mathbf{x}^\intercal\|_0 \leq T$
- $\mathcal{B}_{tr}(1) : \text{rank}(A^\intercal) \leq T$

Closed Form Updates

Optimality Gap

For convex function f, we have:

$$\begin{aligned}
& f(\mathbf{x}) \geq f(\mathbf{x}^t) + \left\langle \nabla f(\mathbf{x}^t), \mathbf{x} - \mathbf{x}^t \right\rangle \\
\implies\ & f(\mathbf{x}^*) \geq f(\mathbf{x}^t) + \left\langle \nabla f(\mathbf{x}^t), \mathbf{x}^* - \mathbf{x}^t \right\rangle \\
& \geq f(\mathbf{x}^t) + \left\langle \nabla f(\mathbf{x}^t), \mathbf{z}^{t+1} - \mathbf{x}^t \right\rangle && (\mathbf{z}^{t+1} \text{ is minimizer}) \\
\implies\ & f(\mathbf{x}^t) - f(\mathbf{x}^*) \leq \left\langle \nabla f(\mathbf{x}^t), \mathbf{x}^t - \mathbf{z}^{t+1} \right\rangle
\end{aligned}$$

This result shows the upper bound of the difference between the current and the optimal value which gives us a good stopping criterion for our algorithm.

Analysis Frank-Wolfe

Assumptions

1. f is β-smooth

$$f(X) \leq f(Y) + \langle \nabla f(Y), X - Y\rangle + \frac{\beta}{2}\|X - Y\|_2^2$$

2. f is convex

$$f(X) \geq f(Y) + \langle \nabla f(Y), X - Y\rangle$$

$$
\begin{aligned}
f(\mathbf{x}^{t+1}) &= f(\mathbf{x}^t + \eta_t(\mathbf{z}^{t+1} - \mathbf{x}^t)) \\
&\leq f(\mathbf{x}^t) + \eta_t \left\langle \nabla f(\mathbf{x}^t), \mathbf{z}^{t+1} - \mathbf{x}^t \right\rangle + \frac{\eta_t^2 \beta}{2} \left\| \mathbf{z}^{t+1} - \mathbf{x}^t \right\|_2^2 \\
&\qquad\qquad \text{(From assumption 1)} \\
&\leq f(\mathbf{x}^t) + \eta_t \left\langle \nabla f(\mathbf{x}^t), \mathbf{x}^* - \mathbf{x}^t \right\rangle + \frac{\eta_t^2 \beta}{2} \left\| \mathbf{z}^{t+1} - \mathbf{x}^t \right\|_2^2 \\
&\qquad\qquad (\mathbf{z}^{t+1} = \arg\min \left\langle \nabla f(\mathbf{x}^t), \mathbf{z} \right\rangle) \\
&\leq f(\mathbf{x}^t) + \eta_t (f(\mathbf{x}^t) - f(\mathbf{x}^*)) + \frac{\eta_t^2 \beta R^2}{2} \qquad (\mathcal{C} = \mathcal{B}_2(0, R),\ f \text{ is convex}) \\
\implies f(\mathbf{x}^{t+1}) - f(\mathbf{x}^*) &\leq (1 - \eta_t)(f(\mathbf{x}^t) - f(\mathbf{x}^*)) + \frac{\eta_t^2 \beta R^2}{2} \\
\implies f(\mathbf{x}^{\intercal}) - f(\mathbf{x}^*) &\leq \frac{\beta R^2}{2(T+2)}
\end{aligned}
$$

The above result shows that Frank-Wolfe guarantees ϵ accurate solution with $\frac{1}{\epsilon}$ sparsity.

Analysis for PGD

A general analysis strategy that we'd be following here is to construct a valid potential function(Lyapunov functions) and show that it reduces in each iteration of our algorithm. We will carry out the analysis under two different conditions. The potential function for both the cases will be:

$$\Phi_t = f(\mathbf{x}^t) - f(\mathbf{x}^*)$$

Let us denote $\left\| \mathbf{x}^t - \mathbf{x}^* \right\| = \mathcal{D}_t$ and $\nabla f(\mathbf{x}^t) = g^t$ for our analysis.

Convex with Bounded Gradients

$$
\begin{aligned}
\Phi_t &= f(\mathbf{x}^t) - f(\mathbf{x}^*) \\
&\leq (\mathbf{x}^t - \mathbf{x}^*)^\intercal g^t && \text{(Since } f \text{ is convex)} \\
&\leq \frac{1}{\eta_t}(\mathbf{x}^t - \mathbf{x}^*)^\intercal (\eta_t g^t) \\
&\leq \frac{1}{2\eta_t}\left(\mathcal{D}_t^2 + \eta_t^2 \left\|g^t\right\|^2 - \left\|\mathbf{x}^t - \eta_t g^t - \mathbf{x}^*\right\|^2\right) && \text{(cosine rule)} \\
&\leq \frac{1}{2\eta_t}\left(\mathcal{D}_t^2 + \eta_t^2 G^2 - \left\|\mathbf{z}^{t+1} - \mathbf{x}^*\right\|^2\right) \\
&\leq \frac{1}{2\eta_t}\left(\mathcal{D}_t^2 + \eta_t^2 G^2 - \mathcal{D}_{t+1}^2\right) && \text{(Projection property 2)} \\
&\leq \frac{1}{2\eta_t}(\mathcal{D}_t^2 - \mathcal{D}_{t+1}^2) + \frac{\eta_t G^2}{2} \\
\implies \sum_{t=1}^{T} \Phi_t &\leq \frac{\mathcal{D}_0^2}{2\eta} + \frac{T\eta G^2}{2} \leq \mathcal{D}_0^2 G^2 \sqrt{T} && \text{(for } \eta = \tfrac{\mathcal{D}_0}{G\sqrt{T}}\text{)} \\
\implies \frac{1}{T}\sum_{t=1}^{T} f(\mathbf{x}^t) &\leq f(\mathbf{x}^*) + \frac{\mathcal{D}_0^2 G^2}{\sqrt{T}} \\
\implies f\left(\frac{1}{T}\sum_{t=1}^{T} \mathbf{x}^t\right) &\leq f(\mathbf{x}^*) + \frac{\mathcal{D}_0^2 G^2}{\sqrt{T}} && \text{(Jensen's Inequality)}
\end{aligned}
$$

The above result shows that choosing the average models gives us an upper bound on the deviation from the optimal value. Also, choosing a model $\mathbf{x} \in \{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^T\}$ which minimizes $f(\cdot)$ also makes sense because if we can bound the average error on models, then we can bound the least error as well.

$\alpha-$strongly Convex and Bounded Gradients

$$
\begin{aligned}
\Phi_t &\leq \frac{1}{\eta_t}(\mathbf{x}^t - \mathbf{x}^*)\eta_t g^t - \frac{\alpha}{2}\mathcal{D}_t^2 \\
&\leq \frac{1}{2\eta_t}\left(\mathcal{D}_t^2 + \eta_t^2 \left\|g^t\right\|^2 - \left\|\mathbf{x}^t - \eta_t g^t - \mathbf{x}^*\right\|^2\right) - \frac{\alpha}{2}\mathcal{D}_t^2 && \text{(cosine rule)} \\
&\leq \frac{1}{2\eta_t}\left(\mathcal{D}_t^2 + \eta_t^2 G^2 - \left\|\mathbf{z}^{t+1} - \mathbf{x}^*\right\|^2\right) - \frac{\alpha}{2}\mathcal{D}_t^2 \\
&\leq \frac{1}{2\eta_t}\left(\mathcal{D}_t^2 + \eta_t^2 G^2 - \mathcal{D}_{t+1}^2\right) - \frac{\alpha}{2}\mathcal{D}_t^2 && \text{(Projection property 2)} \\
&\leq \frac{1}{2}\left(\frac{\mathcal{D}_t^2}{\eta_t} - \frac{\mathcal{D}_{t+1}^2}{\eta_t} - \alpha\mathcal{D}_t^2\right) + \frac{\eta_t G^2}{2} \\
\implies \sum \Phi_t &\leq \sum \frac{\mathcal{D}_t^2}{2}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \alpha\right) + \left(\sum \eta_t\right)\frac{G^2}{2} \\
&\leq \frac{G^2}{2}\sum \frac{1}{\alpha t} \leq \frac{G^2 \ln T}{2\alpha} && \text{(setting } \alpha = \tfrac{1}{\alpha t}\text{)} \\
\implies f(\hat{\mathbf{x}}) &\leq f(\mathbf{x}^*) + \mathcal{O}\left(\frac{\ln T}{T}\right)
\end{aligned}
$$