Indian Institute of Technology Kanpur
CS777 Topics in Learning Theory

Scribe: Jaivardhan Kapoor
Instructor: Purushottam Kar
Date: March 19, 2018

LECTURE
18

Sparse Recovery-II

1 Introduction

In the previous few lectures we have looked at standard algorithms for convex optimization such as Projected Gradient Descent and Frank Wolfe method. We analyzed their convergence in the case of varying cases of convexity, strong convexity and strong smoothness. In the last lecture we also looked at the problem of sparse recovery, and analyzed the convergence of Projected Gradient Descent method. We also introduced the concepts of restricted convexity and smoothness, and employed these in our analysis.

In this lecture, we look at some more techniques for optimization in sparse recovery problems such as Pursuit Techniques and finally Stochastic Gradient Descent. We will also analyze its convergence in the case of convexity and bounded gradients.

2 Recap

Our objective is to solve the following optimization problem

$$\min_{\mathbf{x}\in\mathcal{C}} f(\mathbf{x})$$

Here the function is of the form

$$f(\mathbf{x}) = \|A\mathbf{x} - \mathbf{b}\|_2^2$$

and the set over which we optimize, $\mathcal{C} = \mathcal{B}_0(k) = \left\{\mathbf{x} \in \mathbb{R}^d, \|\mathbf{x}\|_0 \leq k\right\} \subseteq \mathbb{R}^d$, is a set of k-sparse vectors, not necessarily convex.

2.1 Projected Gradient Descent for Sparse Recovery

We restate the updates for the PGD algorithm as described in the previous lecture. The main update equations consist of

$$\mathbf{z}^{t+1} \leftarrow \mathbf{x}^t - \eta_t \nabla f\left(\mathbf{x}^t\right)$$
$$\mathbf{x}^{t+1} \leftarrow \Pi_{\mathcal{C}}(\mathbf{z}^{t+1})$$

The second update term consists projecting onto a (possibly)nonconvex set, which may be NP hard in itself. However, in the case of $\mathcal{C}$ being a set of k-sparse vectors, the projection step is simple, as it simply consists of hard thresholding.
If $\kappa = \frac{\beta}{\alpha} < 2$, then

$$f(\mathbf{x}^T) \leq f^* + \mathcal{O}(exp(-T(2-\kappa)))$$

3 Some relaxations of the problem

3.1 Convex relaxation

In the above mentioned optimization problem, it is mentioned that the optimization is subject to a constrained k-sparse set of vectors. However, the optimization problem can be relaxed to yield a convex optimization problem using the regularization schema proposed by Ivanov. The optimization objective thus is as follows.

$$\begin{aligned} \min_{\mathbf{x}} \quad & \|A\mathbf{x} - \mathbf{b}\|_2^2 \\ \text{subject to} \quad & \|\mathbf{x}\|_1 \leq r \end{aligned}$$

It turns out that this objective is equivalent to a regularization schema proposed by Tikhonov, which is essentially LASSO regression.

$$\min_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\|_2^2 + \lambda \|\mathbf{x}\|_1$$

The interested reader is referred to Oneto et al. (2016) for these relaxations.

3.2 Nonconvex relaxation

Nonconvex relaxation of the optimization objective is done using L^p norms, where $p < 1$.In this case the norm is itself nonconvex. The objective is thus

$$\begin{aligned} \min_{\mathbf{x}} \quad & \|\mathbf{x}\|_p, p < 1 \\ \text{subject to} \quad & \|A\mathbf{x} - \mathbf{b}\|_2^2 < \epsilon \end{aligned}$$

4 Pursuit Techniques

4.1 Orthogonal Matching Pursuit(OMP)

Orthogonal Matching Pursuit(OMP), depicted in the algorithm below, is a greedy iterative algorithm for solving the sparse recovery problem. It iteratively expands an initially empty set S of indices, corresponding to the support of $\mathbf{x}$.

Alternatively, for steps in lines 6-8, we could simply perform the optimization procedure

$$\mathbf{x}^{t+1} \leftarrow \underset{supp(\mathbf{x}) \subseteq S^{t+1}}{\arg\min} \|A\mathbf{x} - \mathbf{b}\|_2^2$$

Note that the iterative updates in lines 6-8 are starkly resemble the Frank-Wolfe updates. In fact, as we shall see, they are a case of full-corrective FW algorithm.

The FW algorithm computes $\arg\max\langle \mathbf{v}, \nabla f(\mathbf{x}^t)\rangle$. In OPM, the iterative optimization objective is $\arg\max_i |\langle \mathbf{e}_i, A(A\mathbf{x}^t - \mathbf{b})\rangle| = A\mathbf{r}^t$. It then finds the convex combination of the coordinates $[i]$. This pattern of rediscovering FW in many other such optimization schemes has been noted many times over the years.

It shall be noted that the number of iterations in Projected GD is independent of the sparsity(k), whereas OMP is linearly dependent on k in the number of iterations.

Algorithm 1: Orthogonal Matching Pursuit

Input: Data $A, \mathbf{b}$
Output: $\mathbf{x} \in \mathbb{R}$

$\mathbf{x}^1 \leftarrow \mathbf{0}$,$S^1 \leftarrow \Phi$,$\mathbf{r}^1 \leftarrow \mathbf{b}$
for $t = 1, 2, \ldots, T-1$, or until convergence **do**
  $\mathbf{r}^t \leftarrow A\mathbf{x}^t - \mathbf{b}$
  Find $i \in [d] = \arg\min_i |\langle A_i, \mathbf{r}^t \rangle|$ //A_i `is the` i`-th column of A`
  $S^{t+1} \leftarrow S^t \cup \{i\}$
  $\mathbf{z}^{t+1} \leftarrow \arg\min_{\mathbf{z} \in \mathbb{R}^{|S^{t+1}|}} \|A_{S^{t+1}}\mathbf{z} - \mathbf{b}\|_2^2$ //$A_{S^{t+1}}$ `is set of columns of` A `chosen according to` S^{t+1}
  $\mathbf{x}^{t+1}_{S_{t+1}} \leftarrow \mathbf{z}^{t+1}$
  $\mathbf{x}^{t+1}_{\overline{S}_{t+1}} \leftarrow 0$ //$\overline{S}_{t+1}$ `denotes the complement of` S_{t+1}
end for
return $\mathbf{x}^\top$

Remark 18.1. The Basis Pursuit technique, $\min_{\mathbf{x}} \|\mathbf{x}\|_1$ *subject to* $A\mathbf{x} = \mathbf{b}$, is essentially equivalent to LASSO regression in the noiseless case. It is not part of the pursuit family of optimization methods.

5 Stochastic Gradient Descent

In cases of large amount of data, the objective function that has to be calculated and correspondingly its gradient for the update steps in optimization procedures scale linearly with the data. Thus, sometimes the subset of data is chosen and gradients are calculated according to that subset of data. Let us denote the parameter of choosing as θ. Then the function is of the form $F(\mathbf{x}, \theta)$. Also,

$$f(\mathbf{x}) = \mathop{\mathbb{E}}_{\theta \sim \mathcal{D}} [F(\mathbf{x}, \theta)]$$

Some examples of this formulation are given below.

Example 18.1. F is the population risk w.r.t. the logistic loss

$$F(\mathbf{x}, \theta) = log(1 + exp(-y\mathbf{x}^T\mathbf{w}))$$

where $\theta = (y, w)$. This can be interpreted as selecting data point w and corresponding label y from a set of data points denoted by $\mathcal{D}$.

Example 18.2. f is the training/empirical risk , and F is the empirical risk of a single datum.

$$f(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} F(\mathbf{x}, \theta^i)$$

Example 18.3. f is a "regularized" risk.

$$F(\mathbf{x}, \theta) = log(1 + exp(-y\mathbf{x}^T\mathbf{w})) + \frac{\lambda}{2} \|\mathbf{x}\|_2^2$$

$$f(\mathbf{x}) = \mathop{\mathbb{E}}_{\theta} [F(\mathbf{x}, \theta)]$$

We now present the algorithm for projected stochastic gradient descent(P-SGD).

5.1 Algorithm

Algorithm 2: Projected Stochastic Gradient Descent

Input: Parameter η, Distribution $\mathcal{D}$
Output: $\hat{\mathbf{x}} \in \mathbb{R}$
1: $\mathbf{x}^1 \leftarrow \mathbf{0}$
2: **for** $t = 1, 2, \ldots, T-1$ **do**
3: Sample i.i.d $\theta^t \sim \mathcal{D}$
4: $\mathbf{g}^t \leftarrow \nabla F(\mathbf{x}^t, \theta^t)$
5: $\mathbf{z}^{t+1} \leftarrow \mathbf{x}^t - \eta_t \mathbf{g}^t$
6: $\mathbf{x}^{t+1} \leftarrow \Pi_{\mathcal{C}}(\mathbf{z}^{t+1})$
7: **end for**
8: **return** $\mathbf{x}^\top$

It must be noted that the additional constraint on $\mathbf{g}^t$ is that

$$\mathbb{E}\left[\mathbf{g}^t | \mathcal{H}^t\right] = \nabla f(\mathbf{x}^t)$$

$\mathcal{H}^t$ is the history of $\mathbf{g}^t$, $\mathcal{H}^t = \{\mathbf{g}^1, \mathbf{g}^2, \ldots \mathbf{g}^{t-1}\}$. The above equation is satisfied as

$$\mathbb{E}\mathbf{g}^t = \mathop{\mathbb{E}}_{\theta}\left[\nabla F(\mathbf{x}^t, \theta)\right] = \nabla \mathop{\mathbb{E}}_{\theta}\left[F(\mathbf{x}^t, \theta)\right] = \nabla f(\mathbf{x}^t)$$

5.2 Analysis

The convergence analysis of the above algorithm is presented for convex f with bounded gradients. It is also assumed that $\mathcal{C}$ is convex. Another assumption we make is that $F(\mathbf{x}, \theta)$ is convex for all θ. Also, $\left\|\mathbf{g}^t\right\|_2 < G$.

Using convexity condition between $\mathbf{x}^t$ and $\mathbf{x}^*$, where the latter is the optimum, we get

$$f(\mathbf{x}^t) - f(\mathbf{x}^*) \leq \langle \nabla f(\mathbf{x}^t), \mathbf{x}^t - \mathbf{x}^* \rangle$$

This is not particularly useful to us, since the algorithm deals with $\mathbf{g}^t$ instead of $\nabla f(\mathbf{x}^t)$. Therefore we work with $F(\mathbf{x}, \theta)$ and $\mathbf{g}^t$, and then take expectation over δ on both sides. This will get rid of θ.

$$F(\mathbf{x}^t, \theta^t) - F(\mathbf{x}^*, \theta^t) \leq \langle \mathbf{g}^t, \mathbf{x}^t - \mathbf{x}^* \rangle$$

We now take expectation ($\mathbb{E}\left[\cdot | \mathcal{H}_t\right]$) on both sides:

$$\begin{aligned}
f(\mathbf{x}^t) - f(\mathbf{x}^*) &\leq \langle \mathbf{g}^t, \mathbf{x}^t - \mathbf{x}^* \rangle && \text{complete the squares to remove inner product} \\
&\leq \eta \frac{G^2}{2} + \frac{D_t^2}{2\eta} - \frac{\left\|\mathbf{z}^{t+1} - \mathbf{x}^*\right\|_2^2}{2\eta} && \text{Use Projection Property 0} \\
&\leq \eta \frac{G^2}{2} + \frac{D_t^2}{2\eta} - \frac{D_{t_1}^2}{2\eta}
\end{aligned}$$

We now take the expectation w.r.t $\mathcal{H}_t$ on both the sides, and sum over all t. The term $D_t^2 - D_{t+1}^2$ telescopes, giving the following result for $\eta = \frac{D_0}{G\sqrt{T}}$:

$$\frac{1}{T} \mathop{\mathbb{E}}_{\mathcal{H}_t}\left[f(\mathbf{x}^t)\right] \leq f(\mathbf{x}^*) + \frac{2G^2 D_0^2}{\sqrt{T}}$$

Since the procedure is stochastic, we need a Chernoff-like bound for the convergence of $f(\mathbf{x}^t)$-$f(\mathbf{x}^*)$. For this, we propose 3 claims:

Claim 18.1. With high probability$(1-\delta)$,

$$\sum (f(\mathbf{x}^t) - F(\mathbf{x}^t, \theta^t)) \leq \sqrt{T} log(\frac{1}{\delta})$$

Proof. Let $Y_t = f(\mathbf{x}^t) - F(\mathbf{x}^t, \theta^t)$. We see that given history $\mathcal{H}_t$, the estimator Y_t is unbiased, $\mathbb{E}[Y_t|\mathcal{H}_t] = 0$. Assume $|Y_t| \leq B$. Observe that the sequence $Y_1, Y_2, \ldots$ is a martingale difference sequence, and thus using Azuma-Hoeffding's inequality, we get

$$\mathbb{P}\left[\sum Y_t > T\epsilon\right] \leq exp(\frac{-2T\epsilon^2}{B^2})$$

Setting $\epsilon = B\sqrt{\frac{1}{T} log(\frac{1}{\delta})}$, we get the required inequality. □

Claim 18.2. With high probability$(1-\delta)$,

$$\sum (F(\mathbf{x}^*, \theta^t) - f(\mathbf{x}^*)) \leq \sqrt{T} log(\frac{1}{\delta})$$

Proof. The proof follows directly from the above result, by changing $\mathbf{x}^t$ to $\mathbf{x}^*$. A sign change is absorbed in the derivation given above. □

Claim 18.3. With high probability$(1-\delta)$,

$$\sum (F(\mathbf{x}^t, \theta^t) - F(\mathbf{x}^*, \theta^t)) \leq \sqrt{T} log(\frac{1}{\delta})$$

Proof. Proof is left as an exercise to the reader. (Hint: set η appropriately after completing the squares to remove the inner product.) □

Combining the above 3 inequalities together, we get, w.h.p$(1-\delta)$

$$\frac{1}{T} \sum (f(\mathbf{x}^t) - f(\mathbf{x}^*)) \leq \frac{1}{\sqrt{T}} log(\frac{3}{\delta})$$

which gives us a confidence bound on the convergence of the function to its optimum. A proof of Azuma-Hoeffding's inequality is give in Azuma (1967).

The subsequent lectures will deal with alternating minimization methods such as MajMin and MinMax.

References

Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Math. J. (2)*, 19(3):357–367, 1967. doi: 10.2748/tmj/1178243286. URL `https://doi.org/10.2748/tmj/1178243286`.

Luca Oneto, Sandro Ridella, and Davide Anguita. Tikhonov, ivanov and morozov regularization for support vector machine learning. *Machine Learning*, 103(1):103–136, 2016.