Indian Institute of Technology Kanpur
CS774 Optimization Techniques

Scribe: Lavisha Aggarwal
Instructor: Purushottam Kar
Date: October 20, 2016

LECTURE
19

Online Learning and Coordinate Descent

1 Introduction

In the previous lecture we studied Stochastic optmization including Stochastic gradient descent and mini-batch SGD. In this lecture we will start by understanding Online learning and move onto Coordinate descent(CD). We will briefly discuss about Block coordinate descent, Gauss Southwell rule, Cyclic coordinate descent and Random coordinate descent. In particular we will be analyzing Randomized block coordinate descent.

2 Online Learning

In Online learning, data becomes available in a sequential manner and each data point is used to update the model at each step. Wikipedia . This is opposed to the batch learning methods which learn the entire data set in a single step. The choice of the data point selected at each step is decided by an adversary, as opposed to SGD where the algorithm is allowed to choose data points from a distribution. The algorithms for online learning are of the form:

Algorithm 1: Online learning

1: **for** $t = 1, 2, \ldots, T$ **do**
2: Algorithm predicts $\mathbf{x}^t$
3: Adversary presents $f_t : \mathcal{X} \to \mathbb{R}$
4: Algorithm incurs penalty $f_t(\mathbf{x}^t)$
5: **end for**

After proposing a model $\mathbf{x}^t$, the algorithm suffers an instantaneous penalty $f_t(\mathbf{x}^t)$. The model is completely unaware of the next possible data sample while predicting $\mathbf{x}^t$. Also, at each step the adversary tries to give the worst performing instance to the model.

Definition 19.1. The Regret of the algorithm after T iterations is the total penalty incurred by the algorithm till then and is defined as: (Hazan, 2015)

$$R(T) := \sum_{t=1}^{T} f_t(\mathbf{x}^t) - \min_{\mathbf{x} \in \mathcal{C}} \sum_{t=1}^{T} f_t(\mathbf{x})$$

In online learning the data point corresponding to each f_t is not available in one go and the algorithm must make a decision before it sees the next data point.

Claim 19.1. It is trivial that $R(T) = \mathcal{O}(T)$ i.e. linear time. This can be proven by assuming f_t to be bounded i.e. $|f_t(\mathbf{x})| \leq B \ \forall \ \mathbf{x} \in \mathcal{C}$, $B \in \mathbb{R}$ and $t = 1, 2, \ldots, T$.

Claim 19.2. It would be interesting if $R(T) = o(T)$ i.e sublinear time since then the average regret would decrease with T. In general, sublinear regrets imply feasible learning and a good average performance.

Remark 19.1. Note that the adversary is not allowed to change the benchmark model $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathcal{C}} \sum_{t=1}^{T} f_t(\mathbf{x})$ in order to achieve linear and sublinear rates. For slight variations in the benchmark model, rates better than $\mathcal{O}(T)$ are still possible but not for large variations.

2.1 Online Gradient Descent

Online gradient descent is one of the simplest algorithms to perform online learning and achieves a sublinear regret. Its algorithm is as follows:

Algorithm 2: Online learning

1: **for** $t = 1, 2, \ldots, T$ **do**
2: Receive $f_t(\mathbf{x}^t)$
3: Update $\mathbf{x}^{t+1} \leftarrow \Pi_{\mathcal{C}}(\mathbf{x}^t - \eta_t \nabla f_t(\mathbf{x}^t))$
4: **end for**

For each iteration the algorithm takes a step from the previous point in the direction of the gradient of the previous cost. The point may lie outside the convex set $\mathcal{C}$ hence, it is projected back in $\mathcal{C}$. The analysis of online gradient descent is similar to that of gradient descent and it has a sublinear regret with $R(T) \leq \mathcal{O}(\sqrt{T})$.

Remark 19.2. Stochastic gradient descent can be considered as a special case of Online gradient descent where the data point is not provided by the adversary, rather chosen by the algorithm.

3 Coordinate Descent

In this section we will study about coordinate descent. Coordinate descent algorithms optimize a single coordinate of the model $\mathbf{x}$ at each step. All elements of $\mathbf{x}$ except the i^{th} coordinate are fixed and the objective function is optimized for only one variable $\mathrm{x_i}$ at each step. Consider the optimization problem:

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(x_1, x_2, \ldots, x_n)$$

For the sake of simplicity we consider an unconstrained optimization problem. The algorithm for coordinate descent is as follows:

Algorithm 3: Coordinate Descent

1: **for** $t = 1, 2, \ldots, T$ **do**
2: Choose $i_t \leftarrow A(f, \mathbf{x}^t) \in [n]$
3: $x_j^{t+1} := x_j^t, \quad j \neq i_t$
4: $x_{i_t}^{t+1} \leftarrow \arg min_v f(x_1^t, x_2^t, \ldots, x_{i-1}^t, v, x_{i+1}^t, \ldots, x_n^t)$
5: **end for**

$[n]$ refers to the set $\{1, 2, \ldots, n\}$

3.1 Block Coordinate Descent

In block coordinate descent, the different elements of $\mathbf{x}$ are clamped into blocks. The blocks are non overlapping and optimization is done with respect to a single block at a time, keeping all other blocks fixed.
We define the blocks as: $s_i \subseteq [n],\ i = 1, 2, \ldots, k$ corresponding to non overlapping subsets of $[n]$.

$$|s_i| = n_i, \quad \sum_{i=1}^{k} n_i = n$$

$$\cup s_i = [n], \quad s_i \cap s_j = \phi \quad if\ i \neq j$$

The objective function for block coordinate descent is given by:

$$\min_{\mathbf{z}_i \in \mathbb{R}^{n_i}} f(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_k)$$

where

$$\mathbf{z}_i = X_{s_i}$$

$$X_{s_i} = \begin{cases} (X_{s_i})_j = x_j & \text{if } j \in s_i \\ (X_{s_i})_j = 0 & \text{if } j \notin s_i \end{cases}$$

Remark 19.3. Alternating minimization is a special case of block coordinate descent where there are just 2 blocks i.e. $k = 2$.

3.2 Gauss Southwell Rule

The Gauss Southwell rule chooses the coordinate or the block with the largest directional gradient.

$$A(f, \mathbf{x}^t) = arg \max_{i \in [k]} ||\nabla_i f(\mathbf{x}^t)||$$

where $\nabla_i f(\mathbf{x}^t) = (\nabla f(\mathbf{x}^t))_{s_i}$.

Remark 19.4.

1. The convergence bounds for Gauss-Southwell rule were established very early. (Luo and P., 1992)

2. (Nutini et al., 2015) recently proved the superiority of Gauss-Southwell rule over random selection in certain settings.

3.3 Derivative free rules

The order for the coordinates or blocks can also be chosen as one of the following rules:

1. **Cyclic**: The order is cyclic i.e., 1, 2, 3, ..., k, 1, 2, 3, ..., k, 1, 2, ...
2. **Random**: The coordinates or blocks are chosen randomly from a fixed distribution: $A(f, \mathbf{x}^t) \sim \mathcal{D}$. For eg: Uniform(k).
3. **Random Cyclic**: Each cycle is permuted randomly : $\underbrace{\sigma_1(1), \sigma_1(2), \ldots, \sigma_1(k)}_{\text{epoch1}}, \underbrace{\sigma_2(1), \ldots, \sigma_2(k)}_{\text{epoch2}}, \ldots$

Example 19.1. Consider the unconstrained objective function :

$$\min_{\mathbf{x}\in\mathbb{R}^n} ||\mathbf{b} - A\mathbf{x}||_2^2, \qquad \mathbf{b} \in \mathbb{R}^m, A \in \mathbb{R}^{m\times n}$$

$$\Rightarrow \min_{\mathbf{x}\in\mathbb{R}^n} ||\mathbf{b} - A\mathbf{x}||_2^2 \;=\; \min_{\mathbf{x}\in\mathbb{R}^n} ||\mathbf{b} - \sum_{i=1}^{n} x_i A_i||_2^2$$

We will apply coordinate descent and optimize the objective with respect to i$^{\text{th}}$ coordinate i.e. x_i,

$$= \min_{\mathbf{x}\in\mathbb{R}^n} ||\mathbf{z}_i - x_i A_i||_2^2, \text{ where } \mathbf{z}_i = \mathbf{b} - \sum_{j\neq i} x_j A_j$$

Differentiating w.r.t x_i and equating to zero:

$$\begin{aligned} 2x_i||A_i||_2^2 - 2A_i^T\mathbf{z}_i &= 0 \\ \Rightarrow x_i^{new} &= \frac{<\mathbf{z}_i, A_i>}{||A_i||_2^2} \end{aligned}$$

If we denote residue($\mathbf{r}$) as $\mathbf{r} = \mathbf{b} - \sum_j x_j A_j$ then,

$$\begin{aligned} \mathbf{z}_i &= \mathbf{r} + A_i x_i^{old} \\ \Rightarrow\; x_i^{new} &= \frac{<\mathbf{r}, A_i>}{||A_i||_2^2} + x_i^{old} \end{aligned}$$

3.4 Randomized Block coordinate descent

In this section we perform the convergence analysis of the randomized block coordinate descent algorithm. The objective f is assumed to be convex and blockwise smooth. Consider the i$^{\text{th}}$ block for optimization.
(Nesterov, 2012) Let f be L_i-strongly smooth for block s_i,

$$f(\mathbf{x} + \mathbf{h}_{s_i}) \;\leq\; f(\mathbf{x}) + <\nabla_i f(\mathbf{x}), \mathbf{h}_{s_i}> + \frac{L_i}{2}||\mathbf{h}_{s_i}||_2^2 \tag{1}$$

$\mathbf{h}_{s_i}$ is non zero only for coordinates in block s_i and $\nabla_i f(\mathbf{x})$ is the gradient w.r.t block s_i (non zero only for block s_i)i.e. $\nabla_i f(\mathbf{x}^t) = (\nabla f(\mathbf{x}^t))_{s_i}$.

$$\mathbf{x}^{t+1} \;=\; \mathbf{x}^t + \arg\min_{\mathbf{h}_{s_i}} f(\mathbf{x} + \mathbf{h}_{s_i})$$

Instead of $f(\mathbf{x}+\mathbf{h}_{s_i})$ we minimize an upper bound,

$$\mathbf{x}^{t+1} \;=\; \mathbf{x}^t + \arg\min_{\mathbf{h}_{s_{i_t}}}\{f(\mathbf{x}^t) + < \nabla_i f(\mathbf{x}^t), \mathbf{h}_{s_{i_t}} > + \frac{L_i}{2}||\mathbf{h}_{s_{i_t}}||^2\}$$

Differentiating and equating to zero,

$$\begin{aligned}\nabla_i f(\mathbf{x}^t) + L_i \mathbf{h}_{s_{i_t}} &= 0\\ \Rightarrow\; \mathbf{h}_{s_{i_t}} &= -\frac{1}{L_{i_t}}\nabla_i f(\mathbf{x}^t)\end{aligned}$$

Hence,

$$\mathbf{x}^{t+1} \;=\; \mathbf{x}^t - \frac{1}{L_{i_t}}\nabla_i f(\mathbf{x}^t)$$

From (1),

$$\begin{aligned}f(\mathbf{x}^{t+1}) &\le f(\mathbf{x}^t) + < \nabla_i f(\mathbf{x}^t),\; \frac{-1}{L_{i_t}}\nabla_i f(\mathbf{x}^t) > + \frac{L_{i_t}}{2}||\frac{-1}{L_{i_t}}\nabla_i f(\mathbf{x}^t)||_2^2\\ \Rightarrow\; f(\mathbf{x}^{t+1}) &\le f(\mathbf{x}^t) - \frac{1}{2L_{i_t}}||\nabla_i f(\mathbf{x}^t)||_2^2\end{aligned}$$

If $\quad L = \max\{L_{i_t}\}$ then,

$$\begin{aligned}f(\mathbf{x}^{t+1}) &\le f(\mathbf{x}^t) - \frac{1}{2L}||\nabla_i f(\mathbf{x}^t)||_2^2\\ \Rightarrow\; \mathbb{E}[f(\mathbf{x}^{t+1})|\mathbf{x}^t] &\le f(\mathbf{x}^t) - \frac{1}{2kL}||\nabla f(\mathbf{x}^t)||_2^2\end{aligned} \tag{2}$$

since $\;\mathbb{E}[f(\mathbf{x}^t)|\mathbf{x}^t] = f(\mathbf{x}^t)$ and $\;\mathbb{E}[||\nabla_i f(\mathbf{x}^t)||_2^2|\mathbf{x}^t] = \frac{1}{k}\sum_i^k ||\nabla_i f(\mathbf{x}^t)||_2^2$
Let us assume

$$R = \max_{\mathbf{x}\in\mathbb{R}^n}\{||\mathbf{x}^* - \mathbf{x}||_2 : f(\mathbf{x}) \le f(\mathbf{x}^0)\}$$

and

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} f(\mathbf{x})$$

Remark 19.5. $\mathbf{x}^t \in \mathcal{B}_2(\mathbf{x}^*, R)\;\forall\; t \ge 1$ since $f(\mathbf{x}^{t+1}) \le f(\mathbf{x}^t) \le f(\mathbf{x}^0)\;\forall\; t \ge 1$

From the convexity of f and applying Cauchy Schwartz inequality,

$$\begin{aligned}f(\mathbf{x}^*) - f(\mathbf{x}^t) &\le < \nabla f(\mathbf{x}^t), \mathbf{x}^* - \mathbf{x}^t >\\ \Rightarrow\; f(\mathbf{x}^*) - f(\mathbf{x}^t) &\le ||\nabla f(\mathbf{x}^t)||_2 R\end{aligned} \tag{3}$$

From (2) and (3),

$$\mathbb{E}[f(\mathbf{x}^{t+1})|\mathbf{x}^t] \;\le\; f(\mathbf{x}^t) - \frac{1}{2kLR^2}(f(\mathbf{x}^t) - f(\mathbf{x}^*))^2$$

Subtracting f^* and taking Expectation,

$$\begin{aligned}\Rightarrow\; \mathbb{E}[f(\mathbf{x}^{t+1}) - f^*] &\le \mathbb{E}[f(\mathbf{x}^t) - f^*] - \frac{1}{2kLR^2}(\mathbb{E}[f(\mathbf{x}^t) - f^*])^2\\ \Rightarrow\; \mathbb{E}[f(\mathbf{x}^{t+1}) - f^*] &\le \mathbb{E}[f(\mathbf{x}^t) - f^*]\end{aligned} \tag{4}$$

Let the Potential function Φ_t be,

$$\Phi_t \;=\; \frac{1}{\mathbb{E}[f(\mathbf{x}^t) - f^*]}, \qquad \Phi_t \geq 0 \;\forall\; t \geq 0 \text{ since } f(\mathbf{x}^t) \geq f^*$$

Then from (4),

$$\begin{aligned}
\frac{1}{\Phi_{t+1}} &\leq \frac{1}{\Phi_t} - \frac{1}{2kLR^2}\frac{1}{\Phi_t^2} \\
\frac{1}{\Phi_{t+1}} &\leq \frac{1}{\Phi_t} - \frac{1}{2kLR^2}\frac{1}{\Phi_t\Phi_{t+1}} \\
\frac{1}{\Phi_t} - \frac{1}{\Phi_{t+1}} &\geq \frac{1}{2kLR^2}\frac{1}{\Phi_t\Phi_{t+1}} \\
\Phi_{t+1} - \Phi_t &\geq \frac{1}{2kLR^2} \\
\Phi_{t+1} - \Phi_0 &\geq \frac{T}{2kLR^2} \\
\Phi_{t+1} &\geq \Phi_0 + \frac{T}{2kLR^2} \\
\Phi_{t+1} &\geq \frac{T}{2kLR^2} \qquad\qquad (\text{since } \Phi_0 \geq 0)
\end{aligned}$$

Hence,

$$\mathbb{E}[f(\mathbf{x}^t) - f^*] \;\leq\; \frac{2kLR^2}{T}$$

Remark 19.6. Shalev-Shwartz and Zhang The SVM objective function:

$$\min_{\mathbf{w}} \frac{1}{2}||\mathbf{w}||^2 + c\sum_i max(0, 1 - y_i\mathbf{w}^T\mathbf{x}_i)$$

is equivalent to

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\boldsymbol{\alpha}^T Q\boldsymbol{\alpha} - \boldsymbol{\alpha}^T 1 = f(\boldsymbol{\alpha}), \quad Q_{ij} = y_i y_j < \mathbf{x}_i, \mathbf{x}_j >, \; 0 \leq \alpha_i \leq C \tag{5}$$

Here, α_i corresponds to a single coordinate for the dual and a single data point for the primal formulation. Hence, we can observe that doing stochastic gradient descent on primal is equivalent to doing Coordinate descent on the dual. In general, coordinate descent is just another way of carrying out stochastic gradient descent.

Remark 19.7. The popular LIBLINEAR library implements SVM and logistic regression models trained via a coordinate descent algorithm.
From (5)

$$f(\boldsymbol{\alpha} + \mathbf{h}_{\mathbf{e_i}}) = \frac{1}{2}h^2 Q_{ii} + (Q\boldsymbol{\alpha})_i h - h + g(\boldsymbol{\alpha})$$

This is a quadratic equation in one variable h, to be minimised s.t. $0 \leq \alpha_i \leq c$.
Satisfying the constraints we obtain,

$$\alpha_i^{new} = min[max(\alpha_i^{old} - \frac{z_i}{Q_{ii}}, 0), c], \quad z_i = (Q\boldsymbol{\alpha})_i - 1$$

Since $\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i$, we can write $y_i\mathbf{w}^T\mathbf{x}_i = (Q\boldsymbol{\alpha})_i$.

References

Elad Hazan. *Introduction to Online Convex Optimization* , volume 2. 2015.

Z.Q. Luo and Tseng P. On the convergence of coordinate descent method for convex dierentiable minimization . *J Optim Theory Appl*, 72(1):7, 1992.

Yu. Nesterov. Efficiency of Coordinate Descent Methods on Huge-Scale Optimization Problems . *SIAM J. Optim*, 22(2):341–362, 2012.

Julie Nutini, Mark Schmidta, Issam Laradji, Michael Friedlander, and Hoyt Koepke. Coordinate Descent Converges Faster with the Gauss-Southwell Rule Than Random Selection . *ICML-15*, pages 1632–1641, 2015.

Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization . *J. Mach. Learn. Res*, 14:567599, 2013.

Wikipedia. Online Machine Learning . *The free Encyclopedia*, 2016.