CSE599s, Spring 2014, Online Learning Lecture 13 - 05/13/2014

Per-coordinate Learning Rate and Per-round Norm

Lecturer: Brendan McMahan or Ofer Dekel *Scribe: Danyang Zhuo*

1 Recap

For $r_t(w) \geq 0, r_t(w_t) = 0$, we have

$$w_{t+1} = \operatorname*{argmin}_{\mathcal{W}} f_{1:t}(w) + r_{0:t}(w).$$

$$\text{Regret} \leq r_{0:t}(u) + \frac{1}{2}\sum_{t=1}^{T} \|g_t\|^2_{(t),*}.$$

Let $\eta_t = \frac{\sqrt{2}B}{\sqrt{\sum_{s=1}^{t}\|g_s\|^2}}$, we have the following bound:

$$\text{Regret} \leq 2\sqrt{2}B\sqrt{\sum_{s=1}^{t}\|g_s\|^2}.$$

2 Per-coordinate Learning Rate

The key observation here is that we should not use the same learning rate for different coordinates. Because the regret bound depends on the magnitude on the point we played, some coordinates might have a lot of zeros. In Generalized Linear Model,

$$f_t(w) = l(w, x_t).$$
$$p_t = \sigma(w \cdot x_t).$$

$$\nabla f_t(w_t) = l'(w \cdot x_t) \cdot x_t = g_t.$$

example: bag-of-words, each value in the vector represents whether a word exists in the text.

$$x_t = (1, 0, 0, 0, 1, 0, 0, 1).$$

$$\text{Regret}(u) = \sum_{t=1}^{T} g_t w_t - g_t u.$$

$$\text{Regret}(u) = \sum_{t=1}^{T}\sum_{i=1}^{d} g_{t,i}(w_{t,i} - u_i).$$

$$\text{Regret}(u) = \sum_{i=1}^{d}\sum_{t=1}^{T} g_{t,i}(w_{t,i} - u_i).$$

Note that $\sum_{t=1}^{T} g_{t,i}(w_{t,i} - u_i)$ is the per-coordinate regret.

We can now assign different learning rate for each coordinate.

If $\mathcal{W} = w|w_i \in [-B^i, B_i]$, we can run d gradient descent algorithm on each coordinate i with

$$\eta_{t,i} = \frac{\sqrt{2}B_i}{\sqrt{\sum_{s=1}^{t} \|g_{t,i}\|^2}}.$$

$$\text{Regret}(u) \leq 2\sqrt{2}\vec{B} \cdot \vec{g}$$

$$\text{Regret}(u) \leq 2\sqrt{2} \left\|\vec{B}\right\|_2 \cdot \|\vec{g}\|_2$$

Let $B_i = B$ for every coordinate i, we can further simply the regret to be

$$\text{Regret}(u) \leq 2\sqrt{2}\sqrt{d}B\sqrt{\sum_{t=1}^{T} \|g_t\|_2^2}$$

Here we can see that this regret bound is $\sqrt{d}$ better than the previous bound, because the previous bound's comparation point is in a ball that holds the hypercube where the per-coordinate learning is playing in.

However, if we have a feasible set of w, we may not be able to run each coordinate independently.

3 An example to show that fixed learning rate is bad

There are two kinds of loss functions for $w \in [-B, B]$.

$$\text{Game I} \qquad f_t(w) = w$$

$$\text{Game II} \qquad f_t(w) = |w - \epsilon|$$

Adversary plays Game I for T_0 rounds and then plays $T_0^{\frac{1}{3}}$ sub-problems II, each with length $T_0^{\frac{1}{3}}$.

It can be shown that with fixed learning rate, Regret $= \Omega(T^{\frac{2}{3}})$.
With per-coordinate learning rate, Regret $= O(T^{\frac{1}{2}})$.

4 Per-round Norm

Let Q_t be a positive definite matrix.

$$\|x\|_{(t)} = \left\|Q_t^{\frac{1}{2}}\right\|_2.$$

$$\|x\|_{(t),*} = \left\|Q_t^{-\frac{1}{2}}\right\|_2.$$

We assume Q_t is a diagonal matrix.

$$r_t(w) = \frac{1}{2}\left\|Q_t^{\frac{1}{2}}(w - w_t)\right\|_2^2$$

$r_{0:t}(u)$ is 1-strongly-convex with respect to $\|x\|_{(t)}$.

$$\|g_t\|^2_{(t),*} = \left\|Q_{1:t}^{-\frac{1}{2}} g_t\right\|_2^2 = \sum_{i=1}^{d} \frac{1}{\sigma_{1:t}} \|g_t\|_2^2$$

$$r_0(w) = \lambda \|w\|_1 + \frac{1}{2} \left\|Q_t^{\frac{1}{2}}(w - w_t)\right\|_2^2$$

L1 regularization can enforce a sparse solution and thus is helpful for dimension reduction. Also, it can speed up processing, w_t might have 10^9 coordinates where only 10^8 of them are non-zero.