Subgradient Methods

October 8, 2024

1 Step size choices

Step size on the k-th iteration, t_k, must satisfy the following conditions:

1. $\sum_{k=1}^{\infty} t_k^2 < \infty$
2. $\sum_{k=1}^{\infty} t_k = \infty$

i.e. step sizes should go to zero but not too fast.

2 Fixed step size

Theorem 1. *Let f be a Lipschitz continous function with constant G. For a fixed step size $t_k = t\ \forall k$, the subgradient method satisfies:*

$$\lim_{k\to\infty} f(x_{best}^{(k)}) \leq f(x^*) + \frac{G^2 t}{2}$$

We can interpret $\frac{G^2 t}{2}$ as a bias. (Recall $x_{\text{best}}^{(k)}$ is the best x found by the subgradient method after k iterations)

3 Diminishing step sizes

Theorem 2. *For diminishing step sizes, the subgradient method satisfies:*

$$\lim_{k\to\infty} f(x_{best}^{(k)}) = f(x^*)$$

We now proceed to prove this theorem. We make an assumption on f that the subgradient is bounded:

$$\forall x\ \forall g \in \partial f\ \exists G \text{ such that } ||g^{(k)}||_2 \leq G \tag{1}$$

Note that this is a *necessary condition*. If f is Lipschitz continuous, then we automatically have that f satisfies Eqn 1.

Proof. Let $f^* = f(x^*) = \min_x f(x)$ be the optimal value. Now consider:

$$\begin{aligned}
||x^{(k+1)} - x^*||_2^2 &= ||x^{(k)} - t_k g^{(k)} - x^*||_2^2 \text{ (Gradient Descent Update)} \\
&= ||x^{(k)} - x^*||_2^2 - 2t_k g^{(k)T}(x^{(k)} - x^*) + t_k^2 ||g^{(k)}||_2^2 \text{ (Expanding brackets)} \\
&\leq ||x^{(k)} - x^*||_2^2 - 2t_k(f(x^{(k)}) - f^*) + t_k^2 ||g^{(k)}||_2^2 \text{ (Definition of subgradient)}
\end{aligned}$$

We iteratively apply these steps to get:

$$||x^{(k+1)} - x^*||_2^2 \leq ||x^{(1)} - x^*||_2^2 - 2\sum_{i=1}^{k} t_i(f(x^{(i)}) - f^*) + \sum_{i=1}^{k} t_i^2 ||g^{(i)}||_2^2$$

Verifying that $||x^{(k+1)} - x^*||_2^2 \geq 0$ we have that:

$$2\sum_{i=1}^{k} t_i\Big(f(x^{(i)}) - f^*\Big) \leq ||x^{(1)} - x^*||_2^2 + \sum_{i=1}^{k} t_i^2 ||g^{(i)}||_2^2$$

Now note that

$$\sum_{i}^{k} t_i \Big(f(x^{(i)} - f^*) \Big) \geq \left(\sum_{i=1}^{k} t_i \right) \min_{i=1,\ldots k} \Big(f(x^{(i)} - f^*) \Big) \tag{2}$$

Recall the definition of $f_{\text{best}}^{(k)}$ from the subgradient algorithm:

$$\begin{aligned} f_{\text{best}}^{(k)} - f^* &= \min_{i=1,\ldots k} \left(f_{\text{best}}^{(k)} - f^* \right) \\ &\leq \frac{||x^{(1)} - x^*||_2^2 + \sum_{i=1}^{k} t_i^2 ||g^{(i)}||_2^2}{2\sum_{i=1}^{k} t_i} \quad \text{(By applying fact 2)} \end{aligned}$$

□

Note that this is why we required the conditions shown in Section 1

3.1 Special cases

1. If we have a constant step size we get that

$$\frac{||x^{(1)} - x^*||_2^2 + \sum_{i=1}^{k} t_i^2 ||g^{(i)}||_2^2}{2\sum_{i=1}^{k} t_i} \longrightarrow \frac{G^2 t}{2} \quad \text{as} \quad k \to \infty$$

2. “Square summable by not summable" (i.e. conditions in Section 1):

$$\frac{||x^{(1)} - x^*||_2^2 + \sum_{i=1}^{k} t_i^2 ||g^{(i)}||_2^2}{2\sum_{i=1}^{k} t_i} \longrightarrow 0 \quad \text{as} \quad k \to \infty$$

4 Comparison to Gradient Descent

In gradient descent we converge in $O(\frac{1}{\epsilon})$ iterations. In the subgradient method we converge in $O(\frac{1}{\epsilon}^2)$ iterations. We can see this when we assuming step size $t_i = \frac{R}{G\sqrt{k}}$:

$$f_{\text{best}}^{(k)} - f^* \leq \frac{R^2 + G^2 \sum_{i=1}^{k} t_i^2}{2\sum_{i=1}^{k} t_i} = \frac{RG}{\sqrt{k}} \leq \epsilon$$

Which means we need $O(\frac{1}{\epsilon}^2)$ iterations to converge.

5 Polyak Step size

What happens if you knew x^*? Then the optimal step size is:

$$t_k = \frac{f^{(k-1)} - f^*}{||g^{(k-1)}||_2^2}$$

An example of this is when we consider the distance to the intersection of sets problem. Let $f_i(x) = dist(x, C_i)$ be the distance to set C_i. Let $f(x) = \max f_i(x)$ be the worst case maximum distance. We want to find x^* such that $\min f(x)$ (i.e. the minimum worst case distance).

Note that here we know $f(x^*) = 0 \implies x^* \in C_1 \cap \ldots \cap C_n$.

Recall that $\partial dist(x, C) = \frac{x - P_c(x)}{||x - P_c(x)||_2^2}$. By the subgradient rule we know $\partial f = Conv[\cup \partial f_i(x)]$.

Let $g_i = \nabla f_i(x) = \frac{x - P_{C^i}(x)}{||x - P_{C^i}(x)||_2}$. We know that $||g^{(k-1)}||_2^2 = 1$. The Polyak Step size becomes:

$$t_k = f(x^{k-1})$$

Now notice that when we substitute this into the update rule:

$$x^{(k)} = x^{(k-1)} - f(x^{(k-1)})\frac{x^{(k-1)} - P_{Ci}(x)}{||x^{(k-1)} - P_{Ci}(x)||_2} = P_c(x^{(k-1)})$$

since $f(x^{(k-1)}) = \frac{x^{(k-1)} - P_{Ci}(x)}{||x^{(k-1)} - P_{Ci}(x)||_2}$

5.1 Projected Subgradient Method

The above exploration leads us to the projected subgradient method. Consider the problem $\min_x f(x)$ st $x \in C$. We then have the update rule:

$$x^{(k)} = P_C\big(x^{(k-1)} - t_k g^{(k-1)}\big)$$