Subgradient method

Axel Böhm

September 9, 2021

Smooth vs. nonsmooth

$$\min_x \; f(x)$$

f is *smooth* and convex

Smoothness means $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$:

$$\text{GD:} \quad x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$
$$f(x_k) - f^* = \mathcal{O}(\frac{1}{k})$$

if the stepsize fulfills $\alpha_k \leq 1/L$.

nonsmooth but convex: **subgradient method**

$$\lfloor \text{ pick } g_k \in \partial f(x_k) x_{k+1} = x_k - \alpha_k g_k \quad f(x_k) - f^* = \mathcal{O}(\frac{1}{\sqrt{k}})$$

Convergence statement

Theorem

f is convex, subgradients are bounded $\|g(x)\| \leq G$ for all $g(x) \in \partial f(x)$. Then,

$$f(\bar{x}_k) - f^* \leq \frac{\|x_1 - x^*\|^2 + G^2 \sum_{i=1}^k \alpha_i^2}{2 \sum_{i=1}^k \alpha_i}$$

for the averaged iterates $\bar{x}_k = \frac{\sum_{i=1}^k \alpha_i x_i}{\sum_{i=1}^k \alpha_i}$

Proof

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &\leq \|x_k - \alpha_k g_k - x^*\|^2 \\ &= \|x_k - x^*\|^2 + 2\alpha_k \langle g_k, x^* - x_k \rangle + \alpha^2 \|g_k\|^2.\end{aligned}$$

Using the subgradient ineq. $\langle g_k, x^* - x_k \rangle \leq f(x^*) - f(x_k)$ we deduce

$$\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 + 2\alpha_k (f(x^*) - f(x_k)) + \alpha_k^2 \|g_k\|^2.$$

Via the *bounded subgradient* assumption

$$2 \sum_{i=1}^{k} \alpha_i (f(x_i) - f(x^*)) + \|x_{k+1} - x^*\|^2 \leq \|x_1 - x^*\|^2 + \sum_{i=1}^{k} \alpha_i^2 G^2.$$

Using Jensens inequality

$$\sum_i \lambda_i f(x_i) \geq \sum_i f\left(\frac{\sum_i \lambda_i x_i}{\sum_i \lambda_i}\right)$$

we obtain

How to choose the stepsize?

$$f(\bar{x}_k) - f^* \leq \frac{\|x_1 - x^*\|^2 + G^2 \sum_{i=1}^k \alpha_i^2}{2\sum_{i=1}^k \alpha_i}$$

Clearly $\alpha_i = \ell_2 \; \ell_1$ leads convergence, for example $1/i$. However, $\alpha_i = \mathcal{O}(1/\sqrt{i})$ gives

$$\sum \alpha_i = (\frac{1}{\sqrt{1}} + \frac{1}{\sqrt{2}} + \cdots + \frac{1}{\sqrt{k}}) > \sqrt{k}$$

$\sum \alpha_i^2 = (\frac{1}{1} + \frac{1}{2} + \cdots + \frac{1}{k}) \approx \log(k)$

$$f(\bar{x}_k) - f^* \leq \frac{\|x_1 - x^*\|^2 + G^2 \log(k)}{2\sqrt{k}}$$

gives complexity

$$\mathcal{O}\left(\frac{\log(k)}{k}\right) =: \tilde{\mathcal{O}}\left(\frac{1}{k}\right)$$

Projected subgradient method

$$\text{(constrained setting)} \quad \min_x f(x)$$

$$x_{k+1} = P_C(x_k - \alpha_k g_k)$$

By using the fact that the projection is a contraction

$$\|P_C(x) - P_C(y)\| \leq \|x - y\|$$

we can deduce the exact same inequality as before

$$\|x_{k+1} - x^*\|^2 = \|P_C(x_k - \alpha_k g_k) - x^*\|^2$$

$$\begin{aligned}
&\leq \|x_k - \alpha_k g_k - x^*\|^2 \\
&= \|x_k - x^*\|^2 + 2\alpha_k \langle g_k, x^* - x_k \rangle + \alpha^2 \|g_k\|^2 \\
&\leq \|x_k - x^*\|^2 + 2\alpha_k (f^* - f(x_k)) + \alpha^2 \|g_k\|^2.
\end{aligned}$$

If C is bounded we can improve a bit

Polyak stepsize

Let's revisit the convergence proof of the subgradient method

$$
\begin{aligned}
\|x_{k+1} - x^*\|^2 &\leq \|x_k - \alpha_k g_k - x^*\|^2 \\
&= \|x_k - x^*\|^2 + 2\alpha_k \langle g_k, x^* - x_k \rangle + \alpha^2 \|g_k\|^2 \\
&\leq \|x_k - x^*\|^2 + 2\alpha_k (f^* - f(x_k)) + \alpha^2 \|g_k\|^2.
\end{aligned}
$$

Can we pick α_k such that the RHS is minimized?

$$
\min_{\alpha} \; \alpha^2 \|g_k\|^2 + 2\alpha_k (f^* - f(x_k))
$$

gives

$$
\alpha^* = \frac{f(x_k) - f^*}{\|g_k\|^2}
$$

$$
\|x_{k+1} - x^*\|^2 = \|x_k - x^*\|^2 - \left(\frac{f(x_k) - f^*}{\|g_k\|} \right)^2
$$

Poljak stepsize [contd]

- Requires us to know the optimal objective function value
- can be the case in certain setting: separable data, feasibility problems
- modern deep learning interpolation setting

Poljak stepsize [contd]

- Requires us to know the optimal objective function value
- can be the case in certain setting: separable data, feasibility problems
- modern deep learning interpolation setting

Figure: Interpolation / overparametrization regime