CMO: Goldstein and Wolfe optimization

Eklavya Sharma

Definition 1. *C_L^1 is the subset of C^1 functions for which*

$$\|\nabla_f(x) - \nabla_f(z)\| \le L\|x - z\|$$

This is called the Lipschitz condition.

Objective: Minimize a lower-bounded C_L^1 function $f : \mathbb{R}^d \mapsto \mathbb{R}$.

Contents

1 Goldstein and Wolfe conditions

Let u be a direction of decrease at $x^{(i)}$ (i.e. $\nabla_f(x^{(i)})^T u < 0$). Our descent algorithm will repeatedly choose a direction of descent (not necessarily steepest descent) and move in that direction with magnitude α.

Unlike the previous algorithms we saw, we'll not necessarily pick α as $\operatorname{argmin}_{\alpha>0} f(x+\alpha u)$. This is called **inexact line search**. But this doesn't mean we can pick α arbitrarily. We still have to be smart about picking α to guarantee (quick) convergence. There are 2 famous ways of picking α: by the Goldstein conditions and the Wolfe conditions.

Let $g(\alpha) = f(x^{(i)} + \alpha u)$. Therefore, $g'(0) = \nabla_f(x^{(i)})^T u < 0$. Also, g is lower bounded because f is lower-bounded.

Draw a line which passes through $(0, g(0))$ with slope $m_1 g'(0)$, where $0 < m_1 < 1$ (note that the slope is negative). Let $h_1(\alpha) = g(0) + m_1 g'(0)\alpha$ be that line. Let $t(\alpha) = h_1(\alpha) - g(\alpha)$.

Lemma 1. *t has a positive zero. Let $\overline{\alpha}_1$ be the smallest positive zero. Then t is positive in the interval $(0, \overline{\alpha}_1)$. Formally,*

$$\exists \overline{\alpha}_1 > 0, (t(\overline{\alpha}_1) = 0 \wedge (\forall \alpha \in (0, \overline{\alpha}_1), t(\alpha) > 0))$$

Proof. Let f^* be the minimum value of f.

$$h_1(\alpha) - g(\alpha) < 0 \Leftarrow h_1(\alpha) - f^* < 0 \iff \alpha > \frac{f^* - g(0)}{m_1 g'(0)} > 0$$

Therefore, there is an α for which $t(\alpha) < 0$.

Since $g \in C^1$, by Taylor series, we get that for very small positive α,

$$g(\alpha) = g(0) + g'(0)\alpha + \alpha o(1)$$

$$\implies t(\alpha) = \alpha((1 - m_1)(-g'(0)) + o(1)) > 0$$

Therefore, there is an α for which $t(\alpha) > 0$.

Since g is continuous, by the intermediate value theorem, there must be an $\overline{\alpha}_1 > 0$ for which $t(\overline{\alpha}_1) = 0$. Without loss of generality, assume that $\overline{\alpha}_1$ is the smallest positive zero of t. Since $t(\alpha) > 0$ for small positive α, $t(\alpha) > 0$ for all $\alpha \in (0, \overline{\alpha}_1)$. □

In our descent algorithm, if we choose α from the interval $(0, \overline{\alpha}_1)$, then $g(0) = h_1(0) > h_1(\alpha) > g(\alpha)$. This means that $f(x^{(i)}) > f(x^{(i)} + \alpha u)$, which is what we required.

However, the decrease may be too small, especially if α is very close to 0. To counteract this, we'll impose another condition on α. We have 2 choices here.

1.1 Goldstein condition

Let $h_2(\alpha) = g(0) + m_2 g'(0)\alpha$, where $0 < m_1 < m_2 < 1$. Therefore, $h_2 - g$ has a smallest positive zero $\overline{\alpha}_2$. Also, $\overline{\alpha}_2 < \overline{\alpha}_1$. We'll chose α from the interval $(\overline{\alpha}_2, \overline{\alpha_1})$. This is called the Goldstein condition for choosing α.

1.2 Wolfe condition

Choose an $\alpha \in (0, \overline{\alpha}_1)$ such that $g'(\alpha) \geq m_3 g'(0)$, where $m_3 \in (0, 1)$. This is called the Wolfe condition.

Theorem 2. *If $m_3 \geq m_1$, it's possible to satisfy the Wolfe condition.*

Proof. Suppose we choose $\widehat{\alpha} \in (0, \overline{\alpha}_1)$. Since g is differentiable, by mean value theorem, we get

$$\exists \alpha \in [\widehat{\alpha}, \overline{\alpha}_1], g'(\alpha)(\overline{\alpha}_1 - \widehat{\alpha}) = g(\overline{\alpha}_1) - g(\widehat{\alpha})$$

Combine the above result with $g(\widehat{\alpha}) < h_1(\widehat{\alpha})$ and $g(\overline{\alpha}) = h_1(\overline{\alpha})$ to get $g'(\alpha) > g'(0)m_1$.

If we choose $m_3 \geq m_1$, then $g'(0)m_3 \leq g'(0)m_1 < g'(\alpha)$. Therefore, the Wolfe condition is satisfied for some $\alpha \in (0, \overline{\alpha}_1)$. □

2 Convergence of Wolfe condition

$$
\begin{aligned}
& g'(\alpha) \geq m_3 g'(0) && \text{(by Wolfe condition)} \\
& \Rightarrow \nabla_f(x^{(i)} + \alpha u)^T u \geq m_3 \, \nabla_f(x^{(i)})^T u \\
& \Rightarrow \left(\nabla_f(x^{(i)} + \alpha u) - \nabla_f(x^{(i)})\right)^T u \geq -(1 - m_3)\, \nabla_f(x^{(i)})^T u \\
& && \text{(subtract } \nabla_f(x^{(i)})^T u \text{ from both sides)} \\
& \Rightarrow \left\|\nabla_f(x^{(i)} + \alpha u) - \nabla_f(x^{(i)})\right\| \geq -(1 - m_3)\, \nabla_f(x^{(i)})^T u \\
& && \text{(both sides were +ve. Apply Cauchy-Schwarz inequality)} \\
& \Rightarrow L\alpha \|u\|^2 \geq -(1 - m_3)\, \nabla_f(x^{(i)})^T u && \text{(Lipschitz condition)} \\
& \Rightarrow \alpha \geq \frac{-(1 - m_3)\, \nabla_f(x^{(i)})^T u}{L\|u\|^2}
\end{aligned}
$$

$$
\begin{aligned}
& g(\alpha) < h_1(\alpha) = g(0) + m_1 g'(0)\alpha \\
& \Rightarrow f(x^{(i+1)}) < f(x^{(i)}) + m_1 \, \nabla_f(x^{(i)})^T u \alpha \\
& \Rightarrow f(x^{(i)}) - f(x^{(i+1)}) > \frac{m_1(1 - m_3)}{L} \left(\frac{\nabla_f(x^{(i)})^T u}{\|u\|}\right)^2
\end{aligned}
$$

Let $\nabla_f(x^{(i)})^T u = -\cos\theta_i \left\|\nabla_f(x^{(i)})\right\| \|u\|$. We'll impose another constraint: we'll choose u to not just be the descent direction, but also in a way that $\cos\theta_i$ is lower-bounded by a positive constant.

$$
f(x^{(i)}) - f(x^{(i+1)}) \geq \frac{m_1(1 - m_3)}{L} \cos^2\theta_i \| \nabla_f(x^{(i)})\|^2
$$

Summing i from 0 to $T-1$, we get

$$
\forall T, f(x^{(i)}) - f^* \geq f(x^{(0)}) - f(x^{(T)}) \geq \frac{m_1(1 - m_3)}{L} \sum_{i=0}^{T-1} \cos^2\theta_i \left\|\nabla_f(x^{(i)})\right\|^2
$$

$\therefore \sum_{i=0}^{\infty} \cos^2\theta_i \left\|\nabla_f(x^{(i)})\right\|^2$ is a convergent series. So for $i \to \infty$, $\nabla_f(x^{(i)}) \to 0$.

Therefore, for $i \to \infty$, $x^{(i)}$ approaches a stationary point. Therefore, the descent algorithm which uses Wolfe condition converges to a stationary point, which would hopefully be a local minimum.

3 Alternate Characterization of C_L^1

Let $f \in C_L^1$. Let $g(\alpha) = f(x + \alpha(y - x))$. Then $g'(\alpha) = \nabla_f(x + \alpha(y - x))^T(y - x)$. Therefore, $g(0) = f(x)$, $g(1) = f(y)$ and $g'(0) = \nabla_f(x)^T(y - x)$.

$$
\int_0^1 (g'(\alpha) - g'(0)) d\alpha = f(y) - f(x) - \nabla_f(x)^T(y - x)
$$

$$
\begin{aligned}
&\left|f(y) - f(x) - \nabla_f(x)^T(y-x)\right| \\
&= \left|\int_0^1 (g'(\alpha) - g'(0))d\alpha\right| \\
&\leq \int_0^1 |g'(\alpha) - g'(0)|d\alpha \\
&= \int_0^1 \left|\left(\nabla_f(x+\alpha(y-x)) - \nabla_f(x)\right)^T(y-x)\right| d\alpha \\
&\leq \int_0^1 \|\nabla_f(x+\alpha(y-x)) - \nabla_f(x)\|\,\|y-x\|d\alpha && \text{(Cauchy-Schwarz inequality)} \\
&\leq \int_0^1 L\alpha\|y-x\|^2 d\alpha && \text{(Lipschitz condition)} \\
&= \frac{L}{2}\|y-x\|^2
\end{aligned}
$$

4 Convergence of Goldstein condition

Let $u = \nabla_f(x^{(i)})$ and $x^{(i+1)} = x^{(i)} - \alpha u$.

Let $g(\alpha) = f(x^{(i)} - \alpha u)$. Then $g'(0) = -\nabla_f(x^{(i)})^T u = -\|u\|^2$.

$h_1(\alpha) = g(0) + m_1 g'(0)\alpha = f(x^{(i)}) - \alpha m_1\|u\|^2$. Similarly $h_2(\alpha) = f(x^{(i)}) - \alpha m_2\|u\|^2$.

$$
\begin{aligned}
&h_2(\alpha) \leq g(\alpha) \leq h_1(\alpha) \\
&\Rightarrow f(x^{(i)}) - m_2\alpha\|u\|^2 \leq f(x^{(i+1)}) \leq f(x^{(i)}) - m_1\alpha\|u\|^2 \\
&\Rightarrow m_1\alpha\|u\|^2 \leq f(x^{(i)}) - f(x^{(i+1)}) \leq m_2\alpha\|u\|^2
\end{aligned}
$$

$$
\begin{aligned}
&f(x^{(i)}) - f(x^{(i+1)}) + \nabla_f(x^{(i)})^T(x^{(i+1)} - x^{(i)}) \\
&\leq m_2\alpha\|u\|^2 + \nabla_f(x^{(i)})^T(x^{(i+1)} - x^{(i)}) \\
&= m_2\alpha\|u\|^2 - \alpha\|u\|^2 \\
&= -(1-m_2)\alpha\|u\|^2
\end{aligned}
$$

Therefore, by Lipschitz condition,

$$\begin{aligned}
&(1-m_2)\alpha\|u\|^2 \le \frac{L}{2}\|x^{(i+1)}-x^{(i)}\|^2 = \frac{L\alpha^2\|u\|^2}{2} \\
&\implies \alpha \ge \frac{2(1-m_2)}{L} \\
&\implies \frac{2(1-m_2)m_1}{L}\|u\|^2 \le m_1\alpha\|u\|^2 \le f(x^{(i)}) - f(x^{(i+1)}) \\
&\implies \forall T, \frac{2(1-m_2)m_1}{L}\sum_{i=0}^{T-1}\left\|\nabla_f(x^{(i)})\right\|^2 \le f(x^{(0)}) - f(x^{(T)}) \le f(x^{(0)}) - f^* \\
&\implies \forall T, \sum_{i=0}^{T-1}\left\|\nabla_f(x^{(i)})\right\|^2 \le \frac{(f(x^{(0)})-f^*)L}{2m_1(1-m_2)}
\end{aligned}$$

$\therefore \sum_{i=0}^{\infty}\left\|\nabla_f(x^{(i)})\right\|^2$ is a convergent series. So for $i \to \infty$, $\nabla_f(x^{(i)}) \to 0$.

Therefore, for $i \to \infty$, $x^{(i)}$ approaches a stationary point. Therefore, the descent algorithm which uses Goldstein condition converges to a stationary point, which would hopefully be a local minimum.

5 Rate of convergence

When descent direction is $-\nabla_f(x^{(i)})$, for both the Wolfe condition and the Goldstein condition, the sum $\sum_{i=0}^{T-1}\left\|\nabla_f(x^{(i)})\right\|^2$ is upper-bounded. Denote the upper bound by N.

Let $\delta = \min_i \|\nabla_f(x^{(i)})\|$. Then $T\delta^2 \le N$. Therefore, $\delta \le \sqrt{\frac{N}{T}}$. This tells us how fast $x^{(i)}$ converges to a stationary point.