COMPLEXITY IN CONTINUOUS OPTIMIZATION

Course 1: January 30, 2025

Detailed proofs:

- $\|x\| = \sqrt{\sum_{i=1}^{n} x_i^2} \quad \forall i = 1..n$

- Global minimum $\Rightarrow$ Local minimum $\Rightarrow$ $\|\nabla f(\bar{x})\| = 0$

Key Lemma

If $x_k \in \mathbb{R}^n$ is such that $\|\nabla f(x_k)\| \neq 0$,

then for any $\alpha > 0$,

$$f(x_k - \alpha \nabla f(x_k)) \leq f(x_k) + \nabla f(x_k)^T (x_k - \alpha \nabla f(x_k) - x_k) + \frac{L}{2} \|x_k - \alpha \nabla f(x_k) - x_k\|^2$$

$$= f(x_k) - \alpha \|\nabla f(x_k)\|^2 + \frac{L}{2} \alpha^2 \|\nabla f(x_k)\|^2$$

$$= f(x_k) - \underbrace{\left(\alpha - \frac{L}{2}\alpha^2\right)}_{>0 \text{ when } \alpha < \frac{2}{L}} \underbrace{\|\nabla f(x_k)\|^2}_{>0}$$

Hence

$$f(x_k - \alpha \nabla f(x_k)) < f(x_k) \quad \text{when } \alpha < \frac{2}{L}$$

Iteration complexity result

Set $\alpha_k = \frac{1}{L} \quad \forall k$ (why? $\alpha - \frac{L}{2}\alpha^2$ maximized for $\alpha = \frac{1}{L}$)

Then, $\forall k$,

$$f(x_{k+1}) \leq f(x_k) - \left(\frac{1}{L} - \frac{L}{2}\left(\frac{1}{L}\right)^2\right) \|\nabla f(x_k)\|^2$$

$$= f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|^2$$

Suppose that $\exists K \geq 1$, $\|\nabla f(x_k)\| > \varepsilon \quad \forall k = 0 \dots K-1$

(false $\Rightarrow \|\nabla f(x_0)\| \leq \varepsilon$, so the method stops at the first iteration)

Then, $\forall k = 0 \dots K-1$, $f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2$

$\Leftrightarrow$ $f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}\varepsilon^2$

$\Leftrightarrow$ $\frac{1}{2L}\varepsilon^2 \leq f(x_k) - f(x_{k+1})$

Sum the inequalities:

$$\underbrace{\sum_{k=0}^{K-1} \frac{1}{2L}\varepsilon^2}_{\frac{K\varepsilon^2}{2L}} \leq \underbrace{\sum_{k=0}^{K-1} \left(f(x_k) - f(x_{k+1})\right)}_{\text{"telescopic sum"}} = f(x_0) - f(x_K)$$

$$\frac{K\varepsilon^2}{2L} \leq f(x_0) - f(x_K) \leq f(x_0) - \underbrace{f_{low}}_{\text{lower bound of } f}$$

$$\Leftrightarrow K \leq 2L\left(f(x_0) - f_{low}\right)\varepsilon^{-2}$$

We have shown that if $\|\nabla f(x_k)\| > \varepsilon \quad \forall k = 0 \dots K-1$,

then $K < 2L\left(f(x_0) - f_{low}\right)\varepsilon^{-2}$

Thus, if we run the method for $\lceil 2L\left(f(x_0) - f_{low}\right)\varepsilon^{-2} \rceil$ iterations,

Then necessarily there will be an iterate x_k such that $\|\nabla f(x_k)\| \leq \varepsilon$.

Conclusion: the algorithm will terminate in at most

$$\lceil 2L\,(f(x_0) - f_{low})\,\varepsilon^{-2} \rceil \text{ iterations.}$$

- L: Lipschitz constant (Measure of nonlinearity)
- $f(x_0) - f_{low}$: Distance (in function value) between x_0 and the lower bound

→ Bound is $O(\varepsilon^{-2})$ ($\leq C\varepsilon^{-2}$ where $C > 0$ that does not depend on ε)

→ $\left(\varepsilon \to \frac{\varepsilon}{10}\right) \Rightarrow \left(O(\varepsilon^{-2}) \to O(100\,\varepsilon^{-2})\right)$

→ The result says that $\min_{0 \leq k \leq K-1} \|\nabla f(x_k)\| \leq \varepsilon$

after at most $K = O(\varepsilon^{-2})$ iterations

Equivalently, after $K \geq 1$ iterations, we have

$$\min_{0 \leq k \leq K-1} \|\nabla f(x_k)\| \leq O\left(\frac{1}{\sqrt{K}}\right)$$

"Convergence rate"

NB: If $f(x) = \frac{1}{2}\sum_{i=1}^{N} f_i(x)^2 = \frac{1}{2}\|r(x)\|^2$

$$\|\nabla f(x_k)\| \leq \varepsilon \quad \Rightarrow \quad \begin{cases} \|\nabla f(x)\| \cdot < \varepsilon \|r(x)\| \\ \text{or} \quad \|r(x)\| < \varepsilon \end{cases}$$

where $r(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_N(x) \end{bmatrix} \in \mathbb{R}^N$

Line search

→ Want $\alpha > 0$ such that $f(x_k - \alpha \nabla f(x_k)) < f(x_k) - c\alpha\|\nabla f(x_k)\|^2$

→ Try $\alpha = 1$, then $\alpha = 0$ if $\alpha = 1$ fails, ...

① $\alpha = 1$ satisfies the condition ✓

② If $\alpha = 1$ does not satisfy the condition, then

(1) $f(x_k) - c\alpha\|\nabla f(x_k)\|^2 \leq f(x_k - \alpha \nabla f(x_k))$ for some $\alpha \in \{1, 0, \theta^2, \dots\}$

Since $f \in C^{1,1}_L$,

(2) $f(x_k - \alpha \nabla f(x_k)) \leq f(x_k) - \alpha\|\nabla f(x_k)\|^2 + \frac{L\alpha^2}{2}\|\nabla f(x_k)\|^2$

Putting (1) and (2) together,

$$f(x_k) - c\alpha\|\nabla f(x_k)\|^2 \leq f(x_k) - \alpha\|\nabla f(x_k)\|^2 + \frac{L\alpha^2}{2}\|\nabla f(x_k)\|^2$$

$$(1-c)\,\alpha\,\|\nabla f(x_k)\|^2 \leq \frac{L}{2}\,\alpha^2\,\|\nabla f(x_k)\|^2 \quad \text{)} \ \|\nabla f(x_k)\| \neq 0$$

$$\boxed{\alpha \geq \frac{2(1-c)}{L}}$$

Thus, if α doesn't satisfy the condition, then $\alpha \geq \frac{2(1-c)}{L}$

$$\alpha = \theta^j \text{ for } j \in \mathbb{N} \Rightarrow \theta^j \geq \frac{2(1-c)}{L}$$

$$\Leftrightarrow j \leq \log_\theta\left(\frac{2(1-c)}{L}\right)$$

As a result, the line search must terminate after

$j_k \leq 1 + \left\lfloor \log_\theta\left(\frac{2(1-c)}{L}\right) \right\rfloor$ tries for α and

α_k is chosen as θ^{j_k} with $\theta^{j_k} \geq \frac{2\theta(1-c)}{L}$

Iteration complexity

At every iteration,

$$f(x_{k+1}) \leq f(x_k) - c\,\alpha_k\,\|\nabla f(x_k)\|^2$$

$$\leq f(x_k) - \frac{2\theta c(1-c)}{L}\,\varepsilon^2$$

In practice: If $\alpha_{k-1} = \theta^{j_{k-1}}$, then at iteration k start line search with $\frac{\theta^{j_{k-1}}}{\theta} = \theta^{j_{k-2}}$

↳ Other conditions:

- $\alpha_k = \underset{\alpha>0}{\operatorname{argmin}}\, f(x_k - \alpha \nabla f(x_k))$ "Exact line search"
- Wolfe conditions

$$f(x_k - \alpha \nabla f(x_k)) \leq f(x_k) - c_1 \alpha \|\nabla f(x_k)\|^2$$

$$-\nabla f(x_k - \alpha \nabla f(x_k))^T \nabla f(x_k) \leq -c_2 \|\nabla f(x_k)\|^2$$

Give $O(\varepsilon^{-2})$ iteration complexity

Bad function for gradient descent (simplified version, Cartis Gould Toint 2022 "Evaluation complexity of nonconvex optimization methods")

Fixed stepsize: $\alpha_k = \alpha > 0$

Goal: Design a function so that gradient descent

$n = 1$ $\qquad x_{k+1} = x_k - \alpha_k g_k$, $g_k = \nabla f(x_k)$

takes exactly ε^{-2} iterations (assuming $\frac{1}{\varepsilon}$ integer)

Define: $\cdot\ k_\varepsilon = \varepsilon^{-2}$

$\cdot\ \forall k \in [\![0, k_\varepsilon]\!]$, $\quad g_k = -\varepsilon\left(2 - \frac{k}{k_\varepsilon}\right)$

$\cdot\ g_0 = -2\varepsilon \leq g_1 < g_2 < \dots < g_{k_\varepsilon} = -\varepsilon$

The first iteration for which $|g_k| \leq \varepsilon$ is k_ε

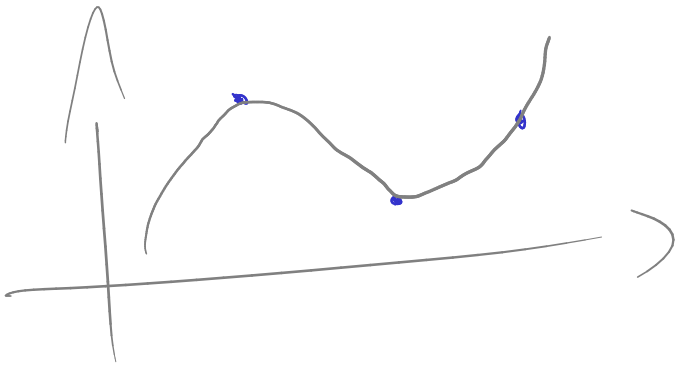
$\cdot\ x_0 = 0$, then $x_{k+1} = x_k - \alpha g_k \quad \forall k \in [\![0, k_\varepsilon - 1]\!]$

- $f_0 = 4\alpha > 0$, then $f_{k+1} = f_k - \alpha |g_k|^2$ $\forall k \in [\![0, k_\varepsilon - 1]\!]$

$\Rightarrow$ Guarantees that line search would accept α as a stepsize

Next step: Define a function $f \in C^{1,1}_L$ such that $f(x_k) = f_k$ $\forall k \in [\![0, k_\varepsilon]\!]$ and $\nabla f(x_k) = g_k$ $\forall k \in [\![0, k_\varepsilon]\!]$

$\Rightarrow$ Hermite interpolation guarantees that such a function exists



Rosenbrock function

$\forall x \in \mathbb{R}^2$, $f(x) = 100\left(x_2 - x_1^2\right)^2 + \left(1 - x_1\right)^2$

unique minimum $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$

Newton: 18 iterations, $\|\nabla f(x_k)\| \simeq 10^{-7}$

Gradient descent: 100 iterations, $\|\nabla f(x_k)\| \simeq 0.8$

Newton: Uses the second-order derivative of the function

$$x_{k+1} = x_k - \alpha_k \left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k)$$

$\nabla^2 f(x_k)$: matrix $\mathbb{R}^{n \times n}$ of derivatives

$\alpha_k = 1$ works most of the time

Typical: line search starting for 1

Bad example for Newton

→ In $\mathbb{R}^2$.

→ Newton iterates well defined

$x_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $x_{k+1} = x_k - \underbrace{H_k^{-1} g_k}_{\begin{bmatrix} 1/f_k \\ \varepsilon/f_k \end{bmatrix}}$

$f_k = 2 - k\varepsilon^2$

$g_k = \begin{pmatrix} -\varepsilon^2 f_k \\ -\varepsilon f_k \end{pmatrix}$

$H_k = \begin{bmatrix} \varepsilon^2 f_k^2 & 0 \\ 0 & f_k^2 \end{bmatrix}$

$$\|g_k\| = \sqrt{\varepsilon^4 f_k^2 + \varepsilon^2 f_k^2} = \sqrt{(\varepsilon^4 + \varepsilon^2)(2 - k\varepsilon^2)^2}$$

$$= \varepsilon \underbrace{\sqrt{(\varepsilon^2+1)}}_{>1} \underbrace{(2 - k\varepsilon^2)}_{\geq 1 \text{ when } k \leq \varepsilon^{-2}} > \varepsilon \quad \text{for } k \leq \varepsilon^{-2}$$

The function

$$\sum_{k=0}^{k_\varepsilon^{-1}} \sigma(\|x - x_k\|)\left(f_k + g_k^T(x - x_k) + \frac{1}{2}(x - x_k)^T H_k (x - x_k)\right)$$

or Hermite interpolation