OPTIMIZATION (SI 416) – LECTURE 5

Harsha Hutridurga

IIT Bombay

- Take a continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$
- Gradient descent algorithm reads as follows:

$$\begin{cases} \text{Begin with a} & x^{(0)} \in \mathbb{R}^n \\ \text{Build iterates using} & x^{(n+1)} = x^{(n)} - \delta \nabla f(x^{(n)}) \quad \text{for } n = 0, 1, 2, \ldots \end{cases}$$

- If f is strongly convex, then there is a unique global minimizer x_*
- If f is further assumed to be β-smooth, then picking $\delta \in (0, \beta^{-1})$ yields a minimizing sequence, i.e. $f(x^{(n+1)}) \leq f(x^{(n)})$
- Furthermore, we have the estimate:

$$\left\| x^{(n)} - x_* \right\| \leq \left(\frac{1}{1 + 2\delta\lambda} \right)^{\frac{n}{2}} \left\| x^{(0)} - x_* \right\| \qquad \text{for } n = 0, 1, \ldots$$

- Suppose we have a tolerance of $\varepsilon > 0$, i.e we are looking for $x^{(n)}$ which is at ε distance from x_*
- Observe that

$$\left(\frac{1}{1+2\delta\lambda}\right)^{\frac{n}{2}} \left\|x^{(0)} - x_*\right\| \leq \varepsilon \implies \left\|x^{(n)} - x_*\right\| \leq \varepsilon$$

- That is

$$n \geq \frac{2}{\ln(1+2\delta\lambda)} \ln\left(\frac{\left\|x^{(0)} - x_*\right\|}{\varepsilon}\right)$$

- Hence, for the n^{th} iterate to be ε close to x_*, we must have

$$n = \mathcal{O}(\ln(\varepsilon^{-1}))$$

- Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is such that $\nabla^2 f(x)$ is invertible for all x
- Consider the algorithm

$$\begin{cases} \text{Begin with a} & x^{(0)} \in \mathbb{R}^n \\ \text{Build iterates using} & x^{(n+1)} = x^{(n)} - \delta \left(\nabla^2 f(x^{(n)}) \right)^{-1} \nabla f(x^{(n)}) \end{cases}$$

  for $n = 0, 1, 2, \ldots$
- The parameter $\delta > 0$ to be chosen later
- Does this generate a minimizing sequence?
- Employing Taylor's theorem, we get

$$f(x^{(n+1)}) = f(x^{(n)}) - \delta \left\langle \nabla f(x^{(n)}), \left(\nabla^2 f(x^{(n)}) \right)^{-1} \nabla f(x^{(n)}) \right\rangle$$
$$+ \frac{\delta^2}{2} \left\langle \nabla^2 f(y) \left(\nabla^2 f(x^{(n)}) \right)^{-1} \nabla f(x^{(n)}), \left(\nabla^2 f(x^{(n)}) \right)^{-1} \nabla f(x^{(n)}) \right\rangle$$

- Observe that choosing $\delta \ll 1$, we can drop the term of $\mathcal{O}(\delta^2)$
- Note: Positive definite $\nabla^2 f$ will generate minimizing sequence

- Similar to the β-smoothness condition, we shall assume that

$$\left\|\nabla^2 f(x)v - \nabla^2 f(y)v\right\| \leq \gamma \|x-y\| \, \|v\| \qquad \text{for all } x, y, v \in \mathbb{R}^n$$

  for some $\gamma > 0$
- Suppose f is strongly convex, i.e. a minimizer x_* exists
- Take $\delta = 1$ in the algorithm and note that

$$\begin{aligned}
x^{(n+1)} - x_* &= x^{(n)} - x_* - \left(\nabla^2 f(x^{(n)})\right)^{-1} \nabla f(x^{(n)}) \\
&= x^{(n)} - x_* \\
&\quad - \left(\nabla^2 f(x^{(n)})\right)^{-1} \int_0^1 \nabla^2 f(x_* + \alpha(x^{(n)} - x_*))(x^{(n)} - x_*)\, \mathrm{d}\alpha \\
&= \left(\nabla^2 f(x^{(n)})\right)^{-1} \left(\nabla^2 f(x^{(n)})\right) \left(x^{(n)} - x_*\right) \\
&\quad - \left(\nabla^2 f(x^{(n)})\right)^{-1} \int_0^1 \nabla^2 f(x_* + \alpha(x^{(n)} - x_*))(x^{(n)} - x_*)\, \mathrm{d}\alpha
\end{aligned}$$

- Recall that we had

$$x^{(n+1)} - x_* = \left(\nabla^2 f(x^{(n)})\right)^{-1} \left(\int_0^1 \left(\nabla^2 f(x^{(n)}) - \nabla^2 f(x_* + \alpha(x^{(n)} - x_*))\right) \mathrm{d}\alpha\right) \left(x^{(n)} - x_*\right)$$

- Hence we deduce that

$$\left\|x^{(n+1)} - x_*\right\| \le \left\|\left(\nabla^2 f(x^{(n)})\right)^{-1}\right\| \left\|\int_0^1 \left(\nabla^2 f(x^{(n)}) - \nabla^2 f(x_* + \alpha(x^{(n)} - x_*))\right) \mathrm{d}\alpha\right\| \left\|x^{(n)} - x_*\right\|$$

$$\le \frac{\left\|x^{(n)} - x_*\right\|}{\lambda} \int_0^1 \left\|\nabla^2 f(x^{(n)}) - \nabla^2 f(x_* + \alpha(x^{(n)} - x_*))\right\| \mathrm{d}\alpha$$

thanks to strong convexity of f

- Using the smoothness assumption on the Hessian, we obtain

$$\left\| x^{(n+1)} - x_* \right\| \leq \frac{\left\| x^{(n)} - x_* \right\|}{\lambda} \int_0^1 \gamma(1-\alpha) \left\| x^{(n)} - x_* \right\| \, \mathrm{d}\alpha$$
$$= \frac{\gamma}{2\lambda} \left\| x^{(n)} - x_* \right\|^2$$

- Hence we deduce that

$$\left\| x^{(n)} - x_* \right\| \leq \left(\frac{\gamma}{2\lambda} \right)^{2^n - 1} \left\| x^{(0)} - x_* \right\|^{2^n}$$

- Observe that

$$\left\| x^{(0)} - x_* \right\| \leq \frac{\lambda}{\gamma} \implies \left\| x^{(n)} - x_* \right\| \leq \left(\frac{2\lambda}{\gamma} \right) 2^{-2^n}$$

- Suppose we have a tolerance of $\varepsilon > 0$, i.e we are looking for $x^{(n)}$ which is at ε distance from x_*
- Observe that

$$\left(\frac{2\lambda}{\gamma}\right) 2^{-2^n} \leq \varepsilon \implies \left\| x^{(n)} - x_* \right\| \leq \varepsilon$$

- That is

$$n \geq \log_2 \left(\log_2 \left(\frac{2\lambda}{\gamma\varepsilon} \right) \right)$$

- Hence, for the n^{th} iterate to be ε close to x_*, we must have

$$n = \mathcal{O}(\ln(\ln(\varepsilon^{-1})))$$

- Recall that for gradient descent, we had $n = \mathcal{O}(\ln(\varepsilon^{-1}))$

- If a sequence $\{x^{(n)}\} \subset \mathbb{R}^n$ converging to a point $x_* \in \mathbb{R}^n$, then

$$\lim_{n\to\infty} \left\| x^{(n)} - x_* \right\| = 0$$

- For a convergent sequence, we can talk about rate of convergence
  - The convergence is **linear** if there exists a $\theta \in (0,1)$ such that

  $$\left\| x^{(n+1)} - x_* \right\| \le \theta \left\| x^{(n)} - x_* \right\|$$

    for all n sufficiently large
  - The convergence is **superlinear** if

  $$\lim_{n\to\infty} \frac{\left\| x^{(n+1)} - x_* \right\|}{\left\| x^{(n)} - x_* \right\|} = 0$$

  - The convergence is **quadratic** if there exists a $C > 0$ such that

  $$\left\| x^{(n+1)} - x_* \right\| \le C \left\| x^{(n)} - x_* \right\|^2$$

    for all n sufficiently large

- Recall: For the gradient descent algorithm to minimize a strongly convex β-smooth function, we had

$$\left\| x^{(n+1)} - x_* \right\| \leq \left(\frac{1}{1 + 2\delta\lambda} \right)^{\frac{1}{2}} \left\| x^{(n)} - x_* \right\|$$

- Hence the convergence here is linear
- Recall: For the Newton's algorithm to minimize a smooth strongly convex function, we had

$$\left\| x^{(n+1)} - x_* \right\| \leq \frac{\gamma}{2\lambda} \left\| x^{(n)} - x_* \right\|^2$$

- Hence the convergence here is quadratic

- Start with an initial vector $x^{(0)} \in \mathbb{R}^n$ and a direction $p^{(0)} \in \mathbb{R}^n$
- Find the next iterate $x^{(1)}$ along the line $x^{(0)} + \alpha p^{(0)}$ with $\alpha > 0$ s.t.

$$f(x^{(1)}) \leq f(x^{(0)})$$

- At the point $x^{(1)}$, pick a new direction $p^{(1)} \in \mathbb{R}^n$
- Find the next iterate $x^{(2)}$ along the line $x^{(1)} + \alpha p^{(1)}$ with $\alpha > 0$ s.t.

$$f(x^{(2)}) \leq f(x^{(1)})$$

- General principle of line search algorithms:
  - At the current iterate $x^{(n)}$, choose a direction $p^{(n)}$
  - Pick the next iterate $x^{(n+1)}$ along the line $x^{(n)} + \alpha p^{(n)}$ with $\alpha > 0$ such that

$$f(x^{(n+1)}) \leq f(x^{(n)})$$

- At each iteration step, we may perform a one-dimensional minimization problem:

$$\min_{\alpha>0} f(x^{(n)} + \alpha p^{(n)})$$

- But, in practice, we are content with finding a candidate that comes close to solving the above one-dimensional problem
- The direction $p^{(n)}$ is referred to as the SEARCH DIRECTION
- Recall the steepest descent algorithm:

$$x^{(n+1)} = x^{(n)} - \delta \nabla f(x^{(n)})$$

- So, here the search direction at the n^{th} iteration step is

$$p^{(n)} = -\nabla f(x^{(n)})$$

- At the iterate $x^{(n)}$ and for any search direction $p^{(n)}$, we have

$$f(x^{(n)} + \alpha p^{(n)}) = f(x^{(n)}) + \alpha \left\langle \nabla f(x^{(n)}), p^{(n)} \right\rangle + \frac{\alpha^2}{2} \left\langle \nabla^2 f(x^{(n)} + s p^{(n)}) p^{(n)}, p^{(n)} \right\rangle$$

  for some $s \in (0, \alpha)$, thanks to Taylor's theorem.
- Define a function $g : [0, \infty) \to \mathbb{R}$ as follows:

$$g(\alpha) := f(x^{(n)} + \alpha p^{(n)}) \quad \text{for } \alpha \in [0, \infty).$$

- Observe that

$$g'(0) = \left\langle \nabla f(x^{(n)}), p^{(n)} \right\rangle$$

- That is, the rate of change of f at the point $x^{(n)}$ in the direction $p^{(n)}$ is given by

$$\left\langle \nabla f(x^{(n)}), p^{(n)} \right\rangle$$

- If we are interested in finding a unit direction of maximum decrease at the point $x^{(n)}$, we should understand

$$\min_{p\in\mathbb{R}^n,\, \|p\|=1} \left\langle \nabla f(x^{(n)}), p \right\rangle$$

- Recall that, if θ_n denotes the angle between $\nabla f(x^{(n)})$ and p, then

$$\left\langle \nabla f(x^{(n)}), p \right\rangle = \|p\| \left\| \nabla f(x^{(n)}) \right\| \cos(\theta_n) = \left\| \nabla f(x^{(n)}) \right\| \cos(\theta_n)$$

- So, the minimum possible value of $\left\langle \nabla f(x^{(n)}), p \right\rangle$ is obtained when

$$\cos(\theta_n) = -1$$

- Observe that the unit vector p which realises that is

$$p = -\frac{\nabla f(x^{(n)})}{\left\| \nabla f(x^{(n)}) \right\|}$$

- We have seen that steepest descent is a line search algorithm where we take the search direction

$$p^{(n)} = -\nabla f(x^{(n)})$$

- Taylor's theorem says

$$f(x^{(n)} + \alpha p^{(n)}) = f(x^{(n)}) + \alpha \left\langle \nabla f(x^{(n)}), p^{(n)} \right\rangle + \frac{\alpha^2}{2} \left\langle \nabla^2 f(x^{(n)} + s p^{(n)}) p^{(n)}, p^{(n)} \right\rangle$$

- Hence, if we take $0 < \alpha \ll 1$, and if we ensure that

$$\left\langle \nabla f(x^{(n)}), p^{(n)} \right\rangle < 0$$

then we find that $f(x^{(n+1)}) < f(x^{(n)})$
- Any such direction $p^{(n)}$ is referred to as DESCENT DIRECTION

- For any search direction p, we have by Taylor's theorem:

$$f(x^{(n)} + p) = f(x^{(n)}) + \left\langle \nabla f(x^{(n)}), p \right\rangle + \frac{1}{2} \left\langle \nabla^2 f(x^{(n)} + sp)p, p \right\rangle$$

  for some $s \in (0, 1)$.
- Let us assume that $\nabla^2 f(x^{(n)} + sp) \approx \nabla^2 f(x^{(n)})$
- Hence we obtain

$$f(x^{(n)} + p) \approx f(x^{(n)}) + \left\langle \nabla f(x^{(n)}), p \right\rangle + \frac{1}{2} \left\langle \nabla^2 f(x^{(n)})p, p \right\rangle =: F(p)$$

- Observe that F is a quadratic function in p
- If $\nabla^2 f$ is positive definite, then $F(p)$ has a unique global minimum
- Recall: that global minimizer p_* is a critical point of F, i.e.

$$\nabla F(p_*) = 0 \implies p_* = -\left(\nabla^2 f(x^{(n)})\right)^{-1} \nabla f(x^{(n)})$$

- This is the search direction in Newton's algorithm

- Newton's algorithm is also a line search algorithm
- The search direction in Newton's algorithm is

$$p^{(n)} = -\left(\nabla^2 f(x^{(n)})\right)^{-1} \nabla f(x^{(n)})$$

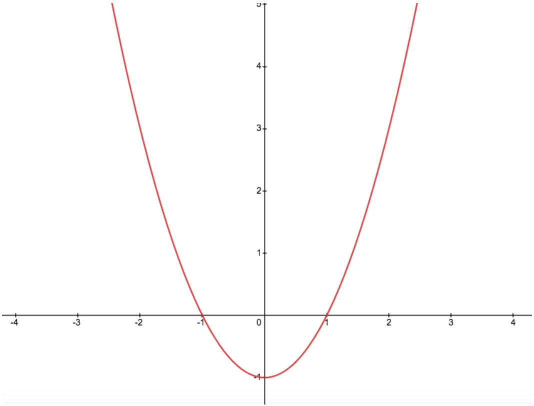
- If $\nabla^2 f$ is strictly positive definite, then

$$\left\langle \nabla f(x^{(n)}), p^{(n)} \right\rangle = -\left\langle \nabla f(x^{(n)}), \left(\nabla^2 f(x^{(n)})\right)^{-1} \nabla f(x^{(n)}) \right\rangle < 0$$

- Hence the above $p^{(n)}$ is a descent direction

- Consider the function $f : \mathbb{R} \to \mathbb{R}$ defined as follows:

$$f(x) := x^2 - 1 \qquad \text{for } x \in \mathbb{R}$$



- The point $x = 0$ is the minimizer and $f(0) = -1$

- Take $x^{(0)} = -2$ and $p^{(0)} = 1$
- Note that $f(x^{(0)}) = 3$
- Take $\alpha_0 = 2 + \sqrt{3}$ so that $x^{(1)} = x^{(0)} + \alpha_0 p^{(0)} = \sqrt{3}$
- Note that $f(x^{(1)}) = 2$
- Take $p^{(1)} = -1$
- Take $\alpha_1 = \sqrt{3} + \sqrt{2}$ so that $x^{(2)} = x^{(1)} + \alpha_1 p^{(1)} = -\sqrt{2}$
- Note that $f(x^{(2)}) = 1$
- Take $p^{(2)} = 1$
- Take $\alpha_2 = \sqrt{2} + \sqrt{\frac{5}{3}}$ so that $x^{(3)} = x^{(2)} + \alpha_2 p^{(2)} = \sqrt{\frac{5}{3}}$
- Note that $f(x^{(3)}) = \frac{2}{3}$
- Take $p^{(3)} = -1$
- Take $\alpha_3 = \sqrt{\frac{5}{3}} + \sqrt{\frac{3}{2}}$ so that $x^{(4)} = x^{(3)} + \alpha_3 p^{(3)} = -\sqrt{\frac{3}{2}}$
- Note that $f(x^{(4)}) = \frac{1}{2}$
- Observe that

$$f(x^{(0)}) > f(x^{(1)}) > f(x^{(2)}) > f(x^{(3)}) > f(x^{(4)})$$

- We can thus build a minimizing sequence $x^{(n)}$ such that

$$f(x^{(n)}) = \frac{2}{n} \qquad \text{for } n = 1, 2, \ldots$$

- But the limiting function value for this sequence is zero
- Recall that the minimum value of the objective function is -1
- This illustrates the possibility of a general line search algorithm
  - leading to insufficient reduction in f in each iteration
  - failing to converge to the minimizer of f
- The root cause for this behaviour stems from the choice of step lengths α_n in each iteration step
- Here we encounter certain sufficient decrease conditions

END OF LECTURE 5

THANK YOU FOR YOUR ATTENTION