Lecture 27 Gradient Descent

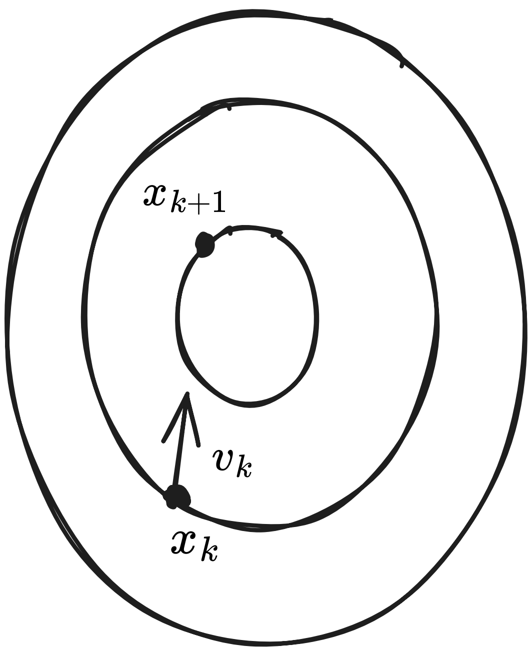
Note by Samion Suwito on 4/29/25

Last Time

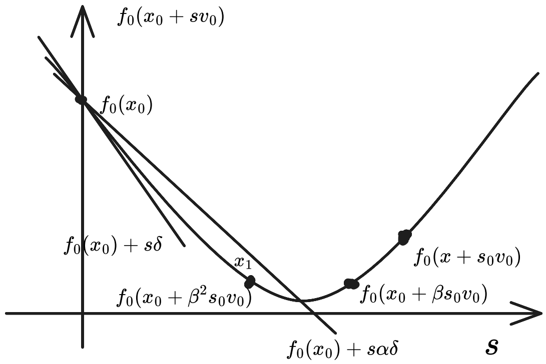
Basic Descent Algorithm For Unconstrained Optimisation

Given: $x_0 \in \operatorname{int} \operatorname{dom} f_0$,
For $k = 1, 2, 3, \cdots$

1. Determine direction v_k, step size s_k
2. $x_{k+1} = x_k + s_k v_k$



Many methods for choosing each of v_k, s_k:
Ex: $v_k = -\nabla f_0(x_k) \equiv$ gradient descent (accelerated gradient descent)
Ex: $s_k = s \quad \forall k \geq 1$ (fixed step size) or ideal line search…
Ex: **Lecture 26 Backtracking Line Search**
Given $\alpha, \beta \in (0, 1)$, $s_0 > 0$
1. Set $s = s_0, \delta_k = \nabla f_0(x_k)^\top v_k$
2. If $f_0(x_k + s v_k) \leq f_0(x_k) + \alpha s \delta_k$ (Armijo Condition), then return $s_k = s$, else $s = \beta s$ and repeat step 2.



Analysis of Gradient Descent

$v_k = -\nabla f_0(x_k), \alpha = \frac{1}{2}$ (arbitrarily)
Assume step size s_k chosen small enough so that Armijo Condition is satisfied with first order characterisation:

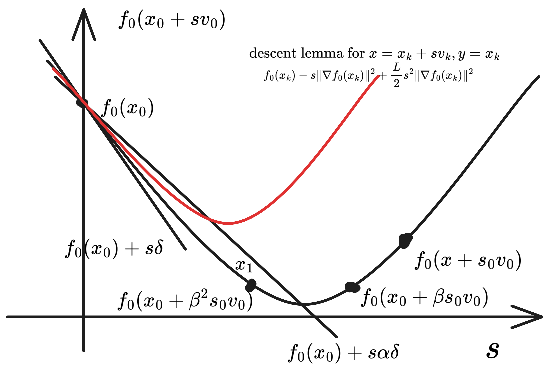
$$
\begin{aligned}
f_0(x_{k+1}) \leq f_0(x_k) + \alpha s_k \delta_k &= f_0(x_k) - \frac{s_k}{2}\|\nabla f_0(x_k)\|_2^2 \\
&\leq f_0(x^\star) + \nabla f_0(x_k)^\top(x_k - x^\star) - \frac{s_k}{2}\|\nabla f_0(x_k)\|_2^2 \\
&= f(x^\star) + \frac{1}{2s_k}(\|x_k - x^\star\|^2 - \|x_k - x^\star - s_k\nabla f_0(x \\
&= f_0(x^\star) + \frac{1}{2s_k}(\|x_k - x^\star\|^2 - \|x_{k+1} - x^\star\|^2) \\
&\implies f_0(x_{k+1}) - f(x^\star) \leq \frac{1}{2s_k}(\|x_k - x^\star\|_2^2 - \|x_{k+1} \\
&\underbrace{\leq}_{\text{if } s_k \geq s_{LB} \forall k \geq 1} \frac{1}{2s_{LB}}(\|x_k - x^\star\|_2^2 - \|x_{k+1} - x^\star\|_2^2) \\
&\implies (f_0(x_k) - f_0(x^\star)) \leq \frac{1}{k}\sum_{i=1}^{k}(f_0(x_i - f_0(x^\star)) \\
&\leq \frac{1}{2s_{LB}k}\|x_0 - x^\star\|_2^2
\end{aligned}
$$

LB stands for lower bound
If you sum over k you'll find a telescoping series that and things cancel out. You see that the suboptimality is inversely proportional to k and squared proportional to the initial sub optimality.

Question: What about step size?
Assume: $\nabla f(x)$ is L-Lipschitz: $\|\nabla f_0(x) - \nabla f_0(y)\|_2 \leq L\|x-y\|_2$ essentially saying the gradient doesn't change too fast.
(*Descent Lemma*): Under assumption of L-Lipschitz

$$f_0(x) \leq f_0(y) + \nabla f_0(y^\top)(x-y) + \frac{L}{2}\|x-y\|^2$$

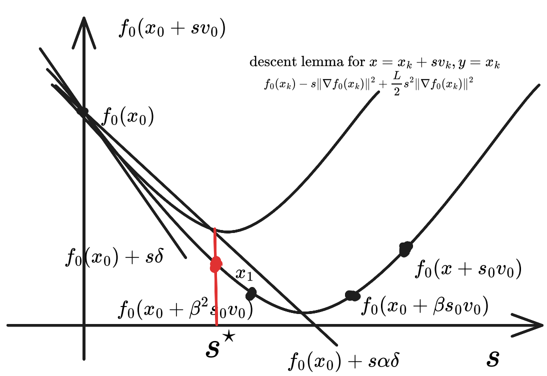
Gives this quadratic function global upper bound.



Proof: Assumption (divide by $\|x-y\|^2$) $\implies \nabla^2 f_0(x) \leq L \cdot I \quad \forall x$
$\implies \frac{L}{2}\|x\|^2 - f_0(x)$ is convex.
First order characteristic of convexity:

$$\frac{L}{2}\|x\|^2 - f_0(x) \geq \frac{L}{2}\|y\|^2 - f_0(y) + (Ly - \nabla f_0(y))^\top (x-y)$$

Since we have an upper bound on the function we can always step at the intersection



If Step size $s_k \leq \frac{1}{L} \quad \forall k$, then will always satisfy Armijo Condition.
$\frac{s^\star}{2} = \frac{L}{2}s^{\star 2} \implies s^\star = \frac{1}{L}$

- So, gradient descent with fixed step size $s = \frac{1}{L}$ ensures convergence and we moreover have the accuracy given by $f_0(x_k) - f_0(x^\star) \leq \frac{L}{2k}\|x_0 - x^\star\|^2$

- What about backtracking? $s_k \geq \min\{\frac{\beta}{L}, s_0\} =: s_{LB}$
  $\implies f_0(x_k) - f_0(x^\star) \leq \max\left\{\frac{L}{2\beta}, \frac{1}{2s_0}\right\} \frac{1}{k} \|x_0 - x^\star\|$

Roughly Speaking, need $O\left(\frac{1}{\varepsilon}\right)$ steps to get ε-suboptimality w/ gradient descent. *Remark*: $O\left(\frac{1}{\sqrt{\varepsilon}}\right)$ is in a sense best possible under the Lipschitz conditions, achieved by Nestrov's Accelerated gradient descent (choose v_k based on the current and previous gradient).

If more conditions are imposed (f_0 is strongly convex (more convex than a quadratic)), then can get better rate $O\left(\log\left(\frac{1}{\varepsilon}\right)\right)$

Second Order methods

Undamped Newton's Method

Newton's method chooses $v_k = -(\nabla^2 f_0(x_k))^{-1} \nabla f_0(x_k) =$"Newton Step"
Intuition: If f_0 was quadratic, then $x_k + v_k$ would be optimal $x^\star$.

Damped Newton's Method

Newton's method will not converge in general it can be too aggressive, can overshoot when less convex than quadratic
Take $x_{k+1} = x_k - s_k(\nabla^2 f_0(x_k))^{-1} \nabla f_0(x_k)$ where s_k is chosen by backtracking.

ε-suboptimality ensures $O\left(\log\log\left(\frac{1}{\varepsilon}\right)\right)$ under strong convexity. While it may look easier than gradient descent it is complex to compute the step itself.