March 2, 2022

In a previous lecture:

upper bound on # iters. of gradient descent that is quadratic in $1/\varepsilon$.

Today:

improved bound that is logarithmic in $1/\varepsilon$.

First, Hessians

Hessian of a function f (assume f is twice continuously differentiable).

1. If $V = \mathbb{R}^n$. Hessian at $\vec{x}$ is simply:

$$[Hf_x]_{ij} = \frac{\partial^2 f(\vec{x})}{\partial \vec{x}_i \, \partial \vec{x}_j}$$

This matrix is symmetric.

2. If V has a non-degenerate inner product, Hf_x is a linear transformation:

$$\nabla f_{x+y} = \nabla f_x + Hf_x(\vec{y}) + g(\vec{y})$$

where $g(\vec{y})$ to first order at $\vec{0}$.

$\Rightarrow$ Hessian is the Jacobian of the gradient,

3. The Hessian is a bilinear form $V \times V \to \mathbb{R}$ in a local Taylor expansion of f:

$$f(\vec{x}+\vec{y}) = f(\vec{x}) + df_x(\vec{y}) + \frac{1}{2} Hf_x(\vec{y},\vec{y}) + r(\vec{y})$$

$r(\vec{y})$ vanishes to second order at $\vec{y}=0$

<u>Definition</u> A matrix M is positive semi-definite (p.s.d)

$\lambda_{min}(M) \geq 0.$

$\Rightarrow \langle x, Mx \rangle \geq 0. \quad \forall x \in V.$

<u>Lemma</u>: If $K \subseteq V$ a closed & convex,

$f: K \to \mathbb{R}$ is convex iff Hf_x is p.s.d $\forall x \in K$.

<u>Pf:</u> $(\Leftarrow)$ Assume Hf_x is p.s.d $\forall x \in K$.

By mean-value theorem, $\exists\, \vec{z}$ on the closed line segment from $\vec{x}$ to $\vec{y}$ st.

$$f(\vec{y}) = f(\vec{x}) + \langle \nabla f_x, \vec{y} - \vec{x} \rangle + \frac{1}{2}\langle \vec{y} - \vec{x}, Hf_z(\vec{y} - \vec{x})\rangle$$

$$\therefore\ f(\vec{y}) \geq f(\vec{x}) + df_x(\vec{y}) - df_x(\vec{x}).$$

$(\Rightarrow)$ Exercise.

<u>Example:</u>

$$f(\vec{x}) = \frac{1}{2}\langle \vec{x}, P\vec{x} \rangle + \langle \vec{x}, \vec{q} \rangle + r.$$

$P \in \mathbb{R}^{n \times n}$ symmetric p.s.d.

$\vec{q} \in \mathbb{R}^n$, $r \in \mathbb{R}$.

$\nabla f_x = P\vec{x} + \vec{q}$

$Hf_x = P.$

$\Rightarrow$ second-order Taylor expansion of a quadratic is itself.

If P is positive definite ($\lambda_{min}(P) > 0$)

For optimality, we require:

$$0 = \nabla f_{\vec{x}^*} = P\vec{x}^* + \vec{q} \Rightarrow \vec{x}^* = -P^{-1}\vec{q}.$$

STRONG CONVEXITY & SMOOTHNESS

$f: K \to \mathbb{R}$ is α-strongly convex if

$\lambda_{min}(Hf_x) \geq \alpha > 0.$

& is β-smooth if $\lambda_{max}(Hf_x) \leq \beta$

$\forall x \in K.$

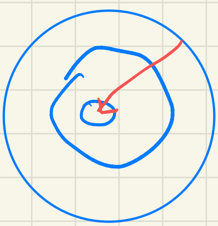
Condition number is $\kappa = \beta/\alpha \geq 1.$

$f(\vec{x}) = \frac{1}{2}\langle \vec{x}, P\vec{x} \rangle$ is $\lambda_{min}(P)$-strongly convex

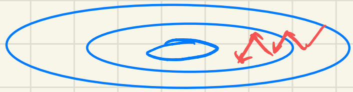
for P p.s.d. $\lambda_{max}(P)$-smooth.

$$\kappa = \frac{\lambda_{max}}{\lambda_{min}}.$$

If $\kappa = 1$.



If $\kappa >> 1$



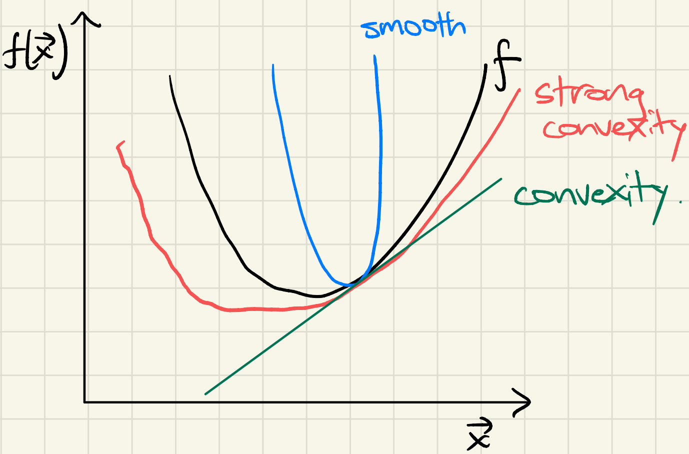
<u>Corollary</u>:

If f is α-strongly convex & β-smooth.

Then $\forall x, y \in K$

1. $$f(\vec{y}) \geq f(\vec{x}) + \langle \nabla f_x, \vec{y} - \vec{x} \rangle + \frac{1}{2}\alpha \|\vec{y} - \vec{x}\|^2.$$

2. $$f(\vec{y}) \leq f(\vec{x}) + \langle \nabla f_x, \vec{y} - \vec{x} \rangle + \frac{1}{2}\beta \|\vec{y} - \vec{x}\|^2.$$



<u>Algorithm:</u> GD w/ 'line search'

Given a feasible $\vec{x}^{(0)}$

repeat

$$\Delta\vec{x} \longleftarrow -\nabla f_x$$

$$t_{opt} \longleftarrow \underset{t\geq 0}{\operatorname{argmin}} \{ f(\vec{x}+t\Delta\vec{x}) \}.$$ # line search

$$\vec{x} \longleftarrow \vec{x} + t_{opt}\,\vec{x}.$$

until $\|\nabla f_x\|^2 \leq 2\varepsilon\alpha.$

$$f(\vec{y}) \geq f(\vec{x}) + \langle \nabla f_x, \vec{y}-\vec{x} \rangle + \frac{1}{2}\alpha\|\vec{y}-\vec{x}\|^2.$$

RHS of this inequality is minimized for

$\tilde{\vec{y}} = \vec{x} - \frac{1}{\alpha}\nabla f_x.$ Plug-in this $\tilde{\vec{y}}$ on the RHS.

$\Rightarrow$

$$f(\vec{y}) \geq f(\vec{x}) - \frac{1}{2\alpha}\|\nabla f_x\|^2.$$

$$\Rightarrow \quad f(\vec{x}) - f(\vec{x}^*) \leq \frac{1}{2\alpha}\|\nabla f_x\|^2.$$

$$f(\vec{y}) \leq f(\vec{x}) + \langle \nabla f_x, \vec{y}-\vec{x} \rangle + \frac{1}{2}\beta\|\vec{y}-\vec{x}\|^2.$$

If we set $\vec{y} = \vec{x} - t\nabla f_x$, and we set $\tilde{t} = \frac{1}{\beta}$ to minimize RHS:

$$f(\vec{x} - t\nabla f_x) \leq f(\vec{x}) - \frac{1}{2\beta}\|\nabla f_x\|^2.$$

By our line search:

$$f(\vec{x} - t_{opt}\nabla f_x) - f(\vec{x}^*) \leq f(\vec{x}) - f(\vec{x}^*) - \frac{1}{2\beta}\|\nabla f_x\|^2.$$

$$\Rightarrow \quad f(\vec{x} - t_{opt}\nabla f_x) - f(\vec{x}^*) \leq f(\vec{x}) - f(\vec{x}^*)$$

$$- \frac{\alpha}{\beta}\left(f(\vec{x}) - f(\vec{x}^*)\right).$$

$$= \left(1 - \frac{1}{\kappa}\right)\left(f(\vec{x}) - f(\vec{x}^*)\right).$$

If we have an initial bound on suboptimality:

$$f(x^{(0)}) - f(\vec{x}^*) \leq D.$$

The number of iterations is upper bounded by:

$$\kappa \ln\left(D/\varepsilon\right).$$

<u>Newton's Method</u>

$$\Delta x_{nt} = -(Hf_x^{-1})(\nabla f_x).$$