$$\phi''(0) = \lim_{t \to 0} \frac{\phi(t) - \phi(0)}{t^2} < 0$$

Pick t small enough $\frac{\phi(t) - \phi(0)}{t^2} \leq \frac{\phi''(0)}{2}$.

$$\Rightarrow \phi(t) \leq \phi(0) + \underbrace{t^2 \frac{\phi''(0)}{2}}_{<0}$$

$$\Rightarrow f(x^* + ts) < f(x^*)$$

$\Rightarrow$ Not local min. $\square$

$\boxed{\text{Starting } 9/7}$

Last time, we showed that necessarily ever local minimum of some $f(x)$ has

$$\nabla f(x^*) = 0, \qquad \nabla^2 f(x^*) \succeq 0.$$

"zero gradient" "positive semidefinite Hessian"

If $\nabla f(x^*) \neq 0$, then $x^* \leftarrow x^* - t \nabla f(x^*)$

If $\nabla^2 f(x) \not\succeq 0$, then $x^* \leftarrow x^* - t s$

$\Updownarrow$

$\exists$ negative eigenvalue

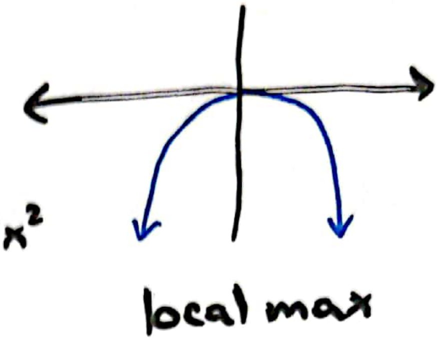
most negative eigenvector in $\nabla^2 f(x^*)$

Scanned with CamScanner

Is it sufficient to have zero grad and p.s.d. Hessian?

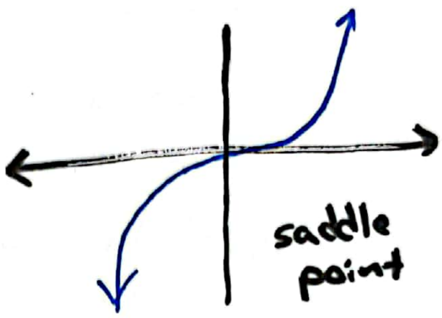
No! $f(x) = -x^4$

$f''(0) = 0$ since $f''(x) = -12x^2$



$f(x) = x^3$

$f''(0) = 0$



5. <u>Theorem (Second-Order Sufficient Condition)</u>

Suppose f is twice diff.

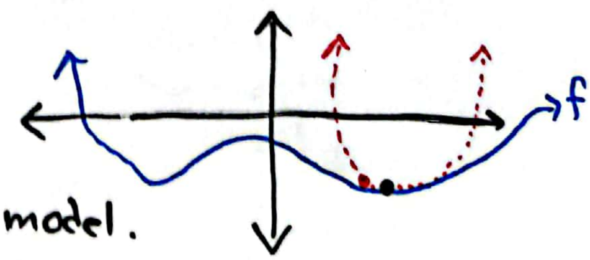
If $x^* \in \mathbb{R}^d$ satisfies $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) \succ 0$.

$\Updownarrow$

then x^* is a strict local minimum. $\forall s \neq 0 \quad s^T \nabla^2 f(x^*) s > 0$.

<u>Proof Idea.</u> Look in 1D.

Strict local min of second-order model.



Proof. Observe that $\lambda_{min}(\nabla^2 f(x^*)) > 0$ (strictly).

For some $\varepsilon > 0$, $\lambda > 0$, every $x \in B(x^*, \varepsilon)$

has $\lambda_{min}(\nabla^2 f(x)) > \lambda > 0$,

(since $\nabla^2 f(x)$ is continuous and $\lambda_{min}(\cdot)$ is a continuous.)

The Fundamental Theorem of Calculus

$$\theta(1) = \theta(0) + \int_0^1 \theta'(t)\, dt$$

and for second derivatives, we have

$$\theta(1) = \theta(0) + \theta'(0) + \int_0^1 \int_0^t \theta''(\alpha)\, d\alpha\, dt$$

$\boxed{\text{Part of HW 1}}$

For $s \in \mathbb{R}^d$, $\|s\| \leq \varepsilon$, $\theta(t) = f(x^* + ts)$.

$$\theta'(0) = \nabla f(x^*)^T s = 0$$

$$\theta''(\alpha) = s^T \nabla^2 f(x^* + \alpha s) s \geq \lambda \|s\|_2^2 \geq 0$$

For any s,

$$f(x^* + s) \geq f(x^*) + 0 + \lambda \|s\|_2^2 \underbrace{\int_0^1 \int_0^t 1\, d\alpha\, dt}_{= 1/2}$$

$$> f(x^*)$$

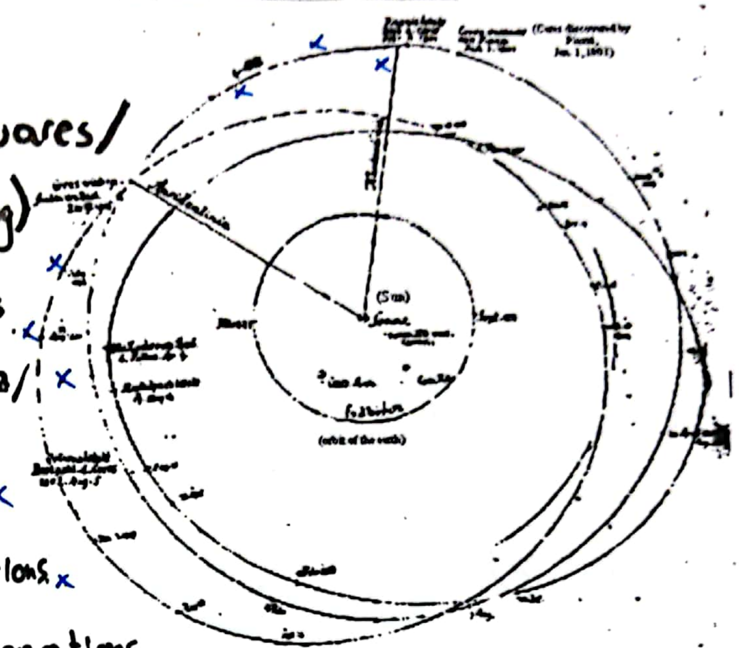
if $s \neq 0$.

$\Rightarrow$ Every nearby point is strictly worse than x^*. $\square$

Two Example Optimization Problems

Example 1 (Least Squares/ Data Fitting)

In 1801, Gauss predicts the location Ceres (asteriod/ dwarf planet).



Given several past locations.

In particular, 22 observations

$$(x_1, y_1), \ldots, (x_N, y_N), \; N = 22.$$

Model this as an ellipse (conic section)

$$(*) \quad \alpha x^2 + \beta y^2 + \gamma x y = 1 \quad \text{for some } \alpha, \beta, \gamma$$

Gauss

What is the best fitting elipse?

Minimize how much we violate (*) squared:

$$\min_{(\alpha,\beta,\gamma)\in\mathbb{R}^3} \sum_{i=1}^{N} \left(\alpha x_i^2 + \beta y_i^2 + \gamma x_i y_i - 1\right)^2$$

Gauss solved this and suceeded

More generally, this is least squares optimization

$$\min_{x \in \mathbb{R}^d} \|Ax-b\|_2^2 = \sum_{i=1}^{N} (a_i^T x - b_i)^2$$

Selecting $x = \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix}$ $b_i = 1$, $a_i = \begin{bmatrix} x_i^2 \\ y_i^2 \\ x_i y_i \end{bmatrix}$

If we want more "accurate" model, we search over

$$\alpha x_i^4 + \beta y_i^4 + \gamma x_i^3 y_i + \dots + \lambda x_i + \omega y_i = 1$$

$$\Rightarrow \quad x = \begin{bmatrix} \alpha \\ \beta \\ \gamma \\ \vdots \\ \omega \end{bmatrix}, \; b_i = 1, \; a_i = \begin{bmatrix} x_i^4 \\ y_i^4 \\ \vdots \\ y_i \end{bmatrix}.$$

HW2 will show $\|Ax-b\|_2^2$ is smooth, convex, and then solve it very effectively.

Example 2 (Logistic Regression)

Suppose you want "quick and dirty" test for if someone might have corona.

Given a patient, we know

c_1 = Patient's age

c_2 = blood pressure

c_3 = heart rate

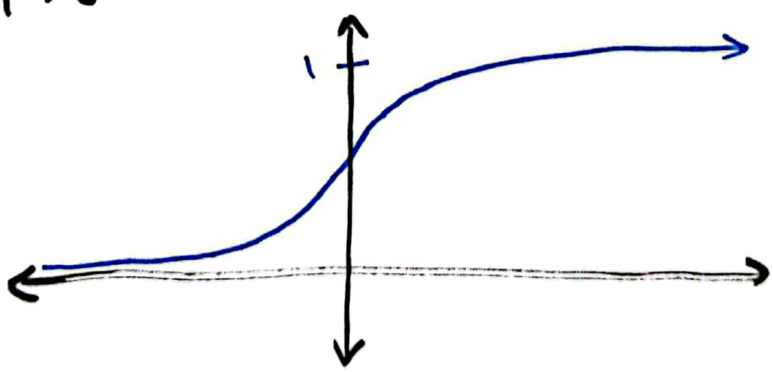
c_4 = temp

c_5 = Family history of cancer

$\vdots$

Lets build a linear model $c^T x$ and use

$$\frac{e^{c^T x}}{1+e^{c^T x}} \approx \mathbb{P}(\text{patient has corona})$$



Given past patients c_i, and outcome

$$r_i = \begin{cases} 0 & \text{if did not have corona} \\ 1 & \text{otherwise.} \end{cases}$$

$$\min_{x \in \mathbb{R}^d} \left(\sum_{\substack{i \\ \text{s.t. } r_i = 0}} \frac{e^{c_i^T x}}{1+e^{c_i^T x}} - \sum_{\substack{i \\ \text{s.t. } r_i = 1}} \frac{e^{c_i^T x}}{1+e^{c_i^T x}} \right).$$

$$\cong \sum_i \left| \frac{e^{c_i^T x}}{1+e^{c_i^T x}} - r_i \right|$$

↑ smooth but not convex.

Neural Network model could replace

$$c^T x \longleftarrow x^T \phi(\cdots \phi(A_2 \phi(A_1 c)))$$

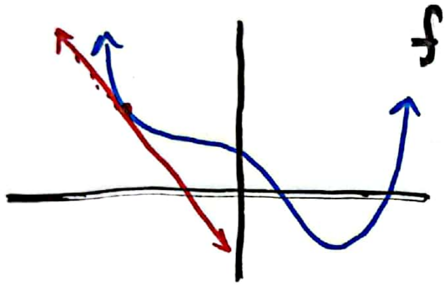
↑ elementwise $\max\{x, 0\}$

↑ nonsmooth and nonconvex

First / Second - Order Iterative Improvement

Given $x_k \in \mathbb{R}^d$, how do we find a better point x_{k+1} given $(\nabla f(x_k), \nabla^2 f(x_k))$?

"First-Order Approach"



$$f(x_k + s) \approx f(x_k) + \nabla f(x_k)^T s$$

Attempt 1

$$\min_{s \in \mathbb{R}^d} f(x_k) + \nabla f(x_k)^T s$$

$$= \begin{cases} -\infty & \text{if } \nabla f(x_k) \neq 0 \\ f(x_k) & \text{otherwise.} \end{cases}$$

Attempt 2

$$\min_{\|s\|_2 \le \varepsilon} f(x_k) + \nabla f(x_k)^T s$$

$$x_{k+1} \leftarrow x_k + \underbrace{\operatorname{argmin}\{\quad\}}_{= -\frac{\nabla f(x_k)}{\|\nabla f(x_k)\|}\varepsilon}$$

Claim: $-\varepsilon \dfrac{\nabla f(x_k)}{\|\nabla f(x_k)\|} = \underset{s}{\operatorname{argmin}} \{ \nabla f(x_k)^T s \mid \|s\|_2 \leq \varepsilon \}$

Proof. Suppose for contradiction that s' does better

$$\nabla f(x_k)^T s' < \nabla f(x_k)^T \left(\frac{-\varepsilon \nabla f(x_k)}{\|\nabla f(x_k)\|_2} \right)$$
$$= -\varepsilon \|\nabla f(x_k)\|_2$$

By Cauchy-Schwartz,

$$-\|\nabla f(x_k)\|_2 \|s'\|_2 < -\varepsilon \|\nabla f(x_k)\|_2$$

$\Rightarrow \quad \|s'\|_2 > \varepsilon .$

$\Rightarrow \quad s'$ is not feasible. $\square$

$\Rightarrow \quad x_{k+1} = x_k - \alpha_k \nabla f(x_k)$ for some α_k.

"Gradient Descent".