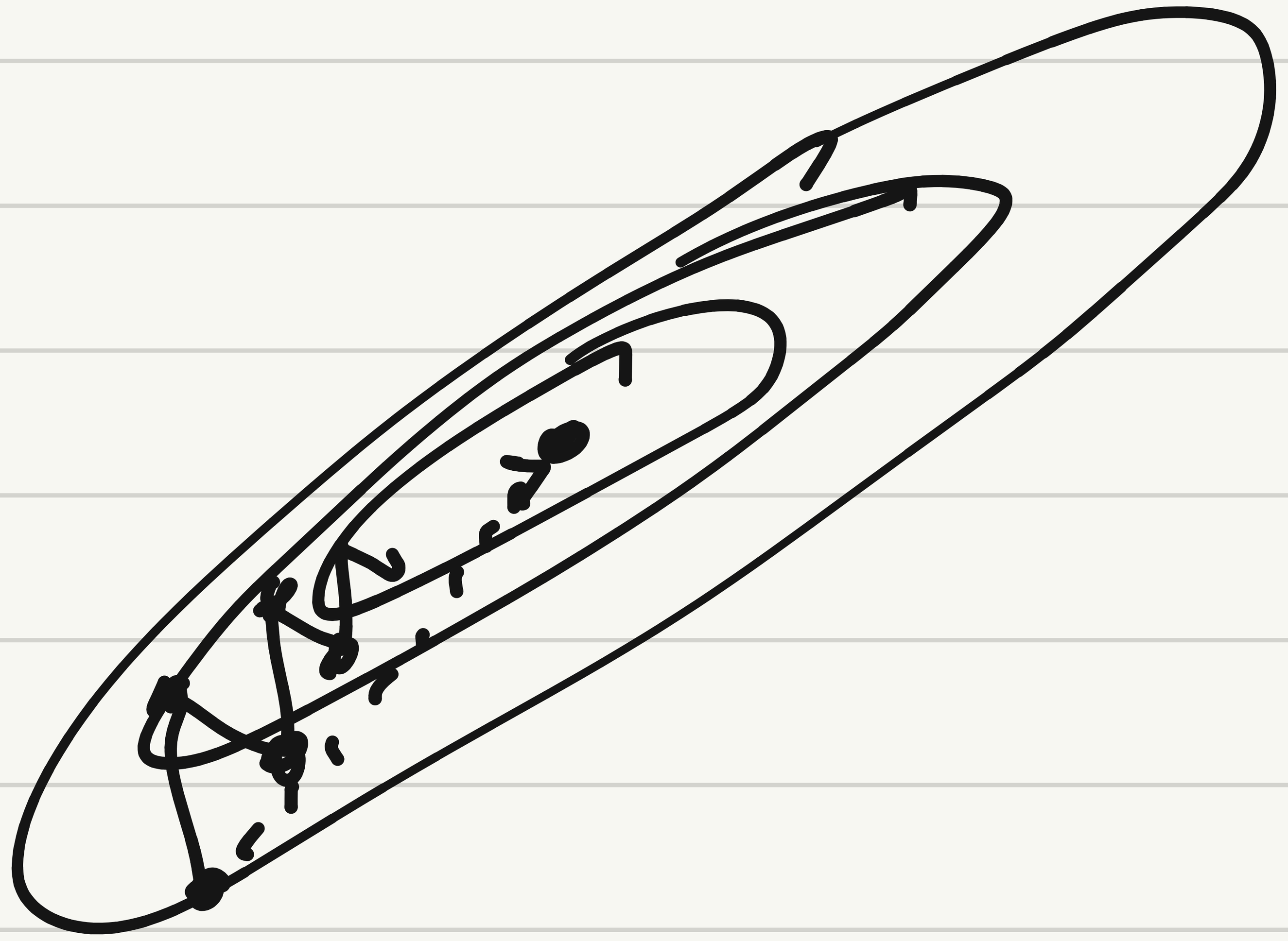
Descent methods

$x^{(0)}, x^{(1)}, x^{(2)}, \ldots \in S_0 = \{ x : f(x) \le f(x^{(0)}) \}$

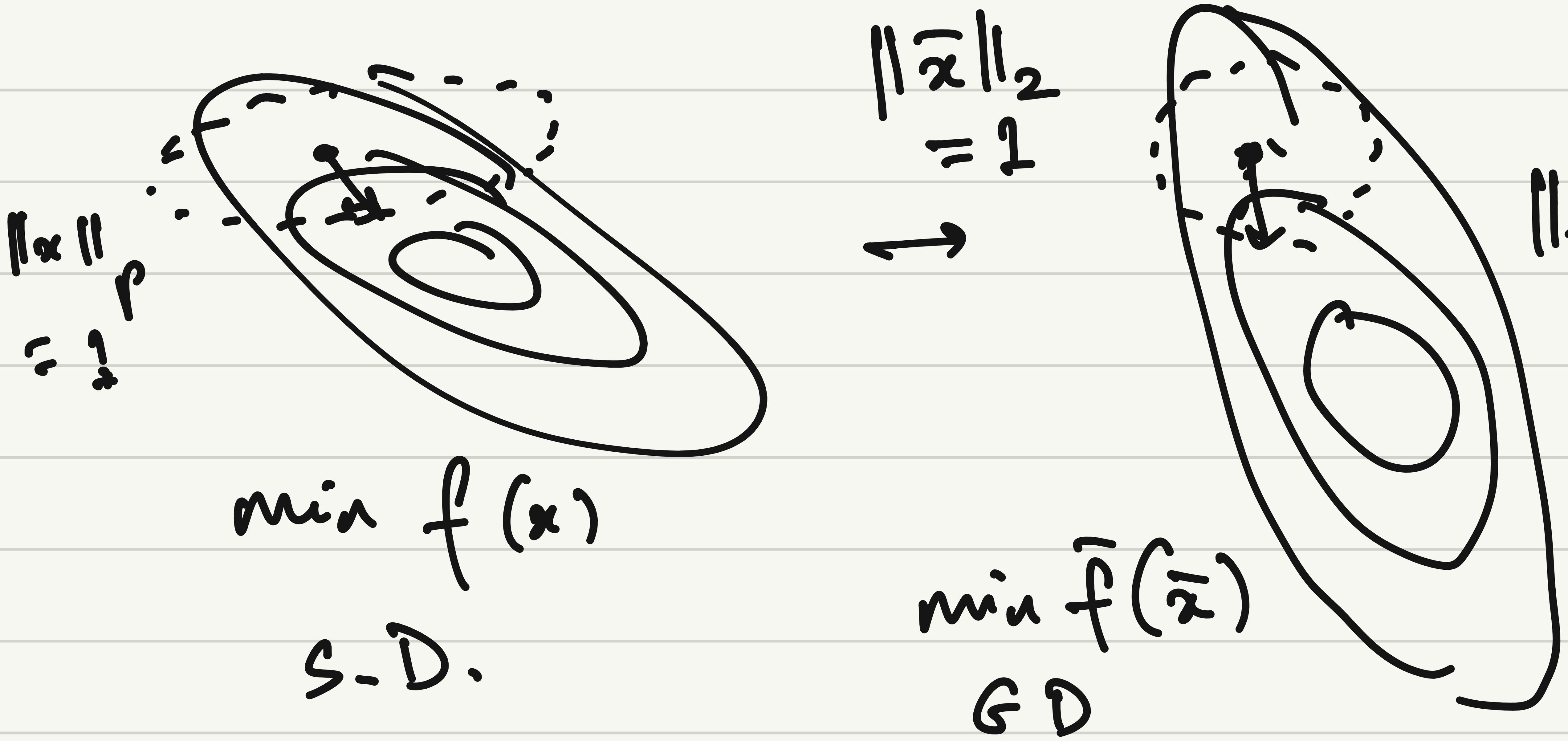


m, M s.t. $mI \preceq \nabla^2 f(x) \preceq MI$ for all $x \in S_0$

GD : $f(x^+) - p^* \le (1 - m/M)(f(x) - p^*)$

Steepest desc. w/ quadratic norm

$P \succ 0, \quad \|x\|_P = \sqrt{x^T P x} \longrightarrow d = -P^{-1} \nabla f(x)$



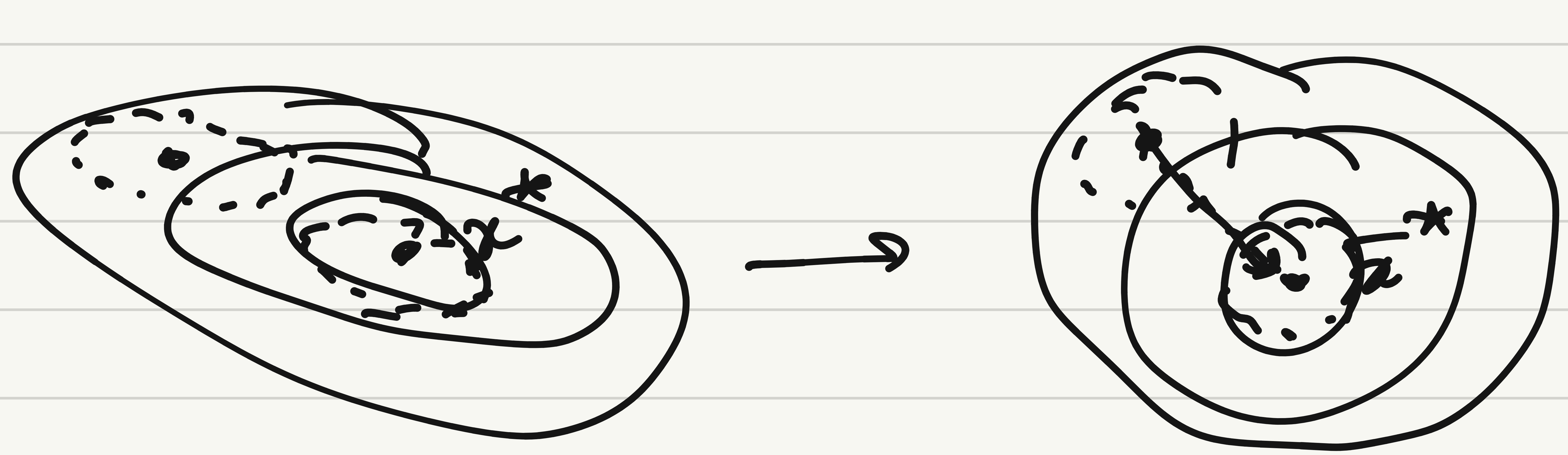
$\bar{x} = Cx$ where $C^T C = P$

$\|x\|_P = \sqrt{x^T P x} = \sqrt{\bar{x}^T \bar{x}} = \|\bar{x}\|_2$

$\bar{f}(\bar{x}) = f(C^{-1}\bar{x}) \longrightarrow \nabla \bar{f}(\bar{x}) = C^{-T} \nabla f(x)$

$\nabla^2 \bar{f}(\bar{x}) = \underbrace{C^{-T} \nabla^2 f(x) C^{-1}}$

$\underline{m, M}$



$P = \nabla^2 f(x^*)$

$C^T C = P$

At x^*, $\nabla^2 \bar{f}(\bar{x}^*) = C^{-T} \nabla^2 f(x) C^{-1}$

↑ $C^T C$ → $= I$

Steepest desc.: $(1 - m/M)$

$$mI \preceq C^{-T} \nabla^2 f(x) C^{-1} \preceq MI$$

$m, M \approx 1$

$P =$

Newton's method (interpretation 1)

Steepest desc. with $P = \nabla^2 f(x)$

If f is strongly convex $\rightarrow m > 0$, $\nabla^2 f(x) \succeq mI \Rightarrow \nabla^2 f(x) \succeq 0$ SPD

$\underline{d} = -P^{-1} \nabla f(x) = -(\nabla^2 f(x))^{-1} \nabla f(x)$ is a descent direction.

(1st order taylor series)

Interp 2. : $\vec{f}(\vec{x}) = \vec{0} \rightarrow \hat{f}(x+h) = f(x) + J(x)\cdot h = 0$, solve for h

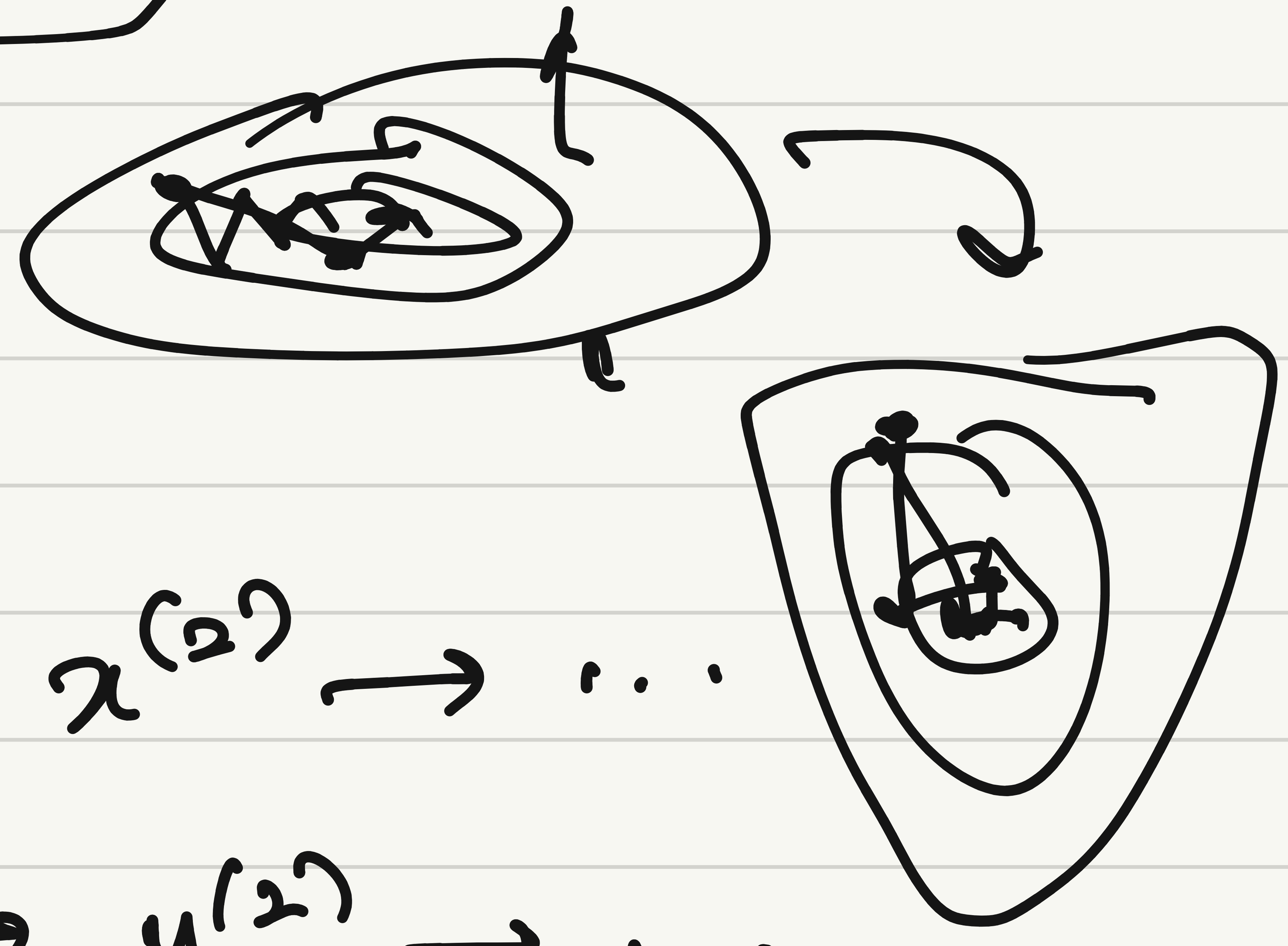
$\min f(\vec{x}) \rightarrow \hat{f}(x+h) = f(x) + \nabla f(x)^T h$

$+ \frac{1}{2} h^T \nabla^2 f(x) h$, min over h

Interp 3 : $\vec{g}(\vec{x}) = \nabla f(x) = 0$

$\rightarrow \hat{g}(x+h) = g(x) + J_g(x)\cdot h = 0$

$\nabla f(x) + \nabla^2 f(x) h = 0 \rightarrow h = -\nabla^2 f(x)^{-1} \nabla f(x)$



Properties of Newton's method

- **affine invariant** : $\min f(x)$: $x^{(0)} \rightarrow x^{(1)} \rightarrow x^{(2)} \rightarrow \dots$

$\min g(y)$: $y^{(0)} \rightarrow y^{(1)} \rightarrow y^{(2)} \rightarrow \dots$

$y = Ax$, $g(y) = f(A^{-1}y)$; $y^{(0)} = Ax^{(0)}$, $y^{(1)} = Ax^{(1)}$, $y^{(2)} = Ax^{(2)}$

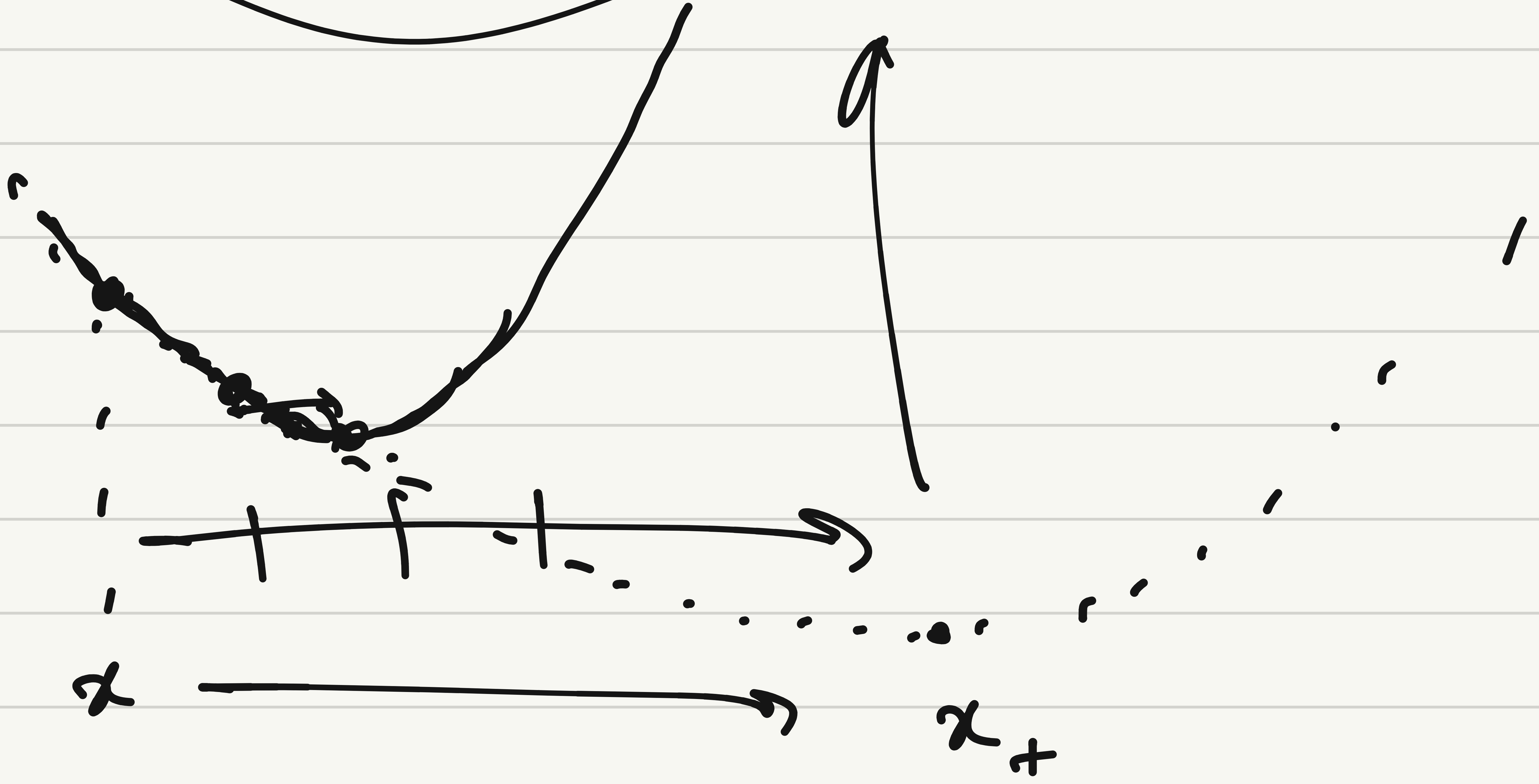
$$f\left(\begin{bmatrix} m \\ l \\ w \\ h \\ s \end{bmatrix}\right)$$

- m ← kg, g
- l, w, h ← mm, m, km
- s ← m/s, km/h, m/d, ..

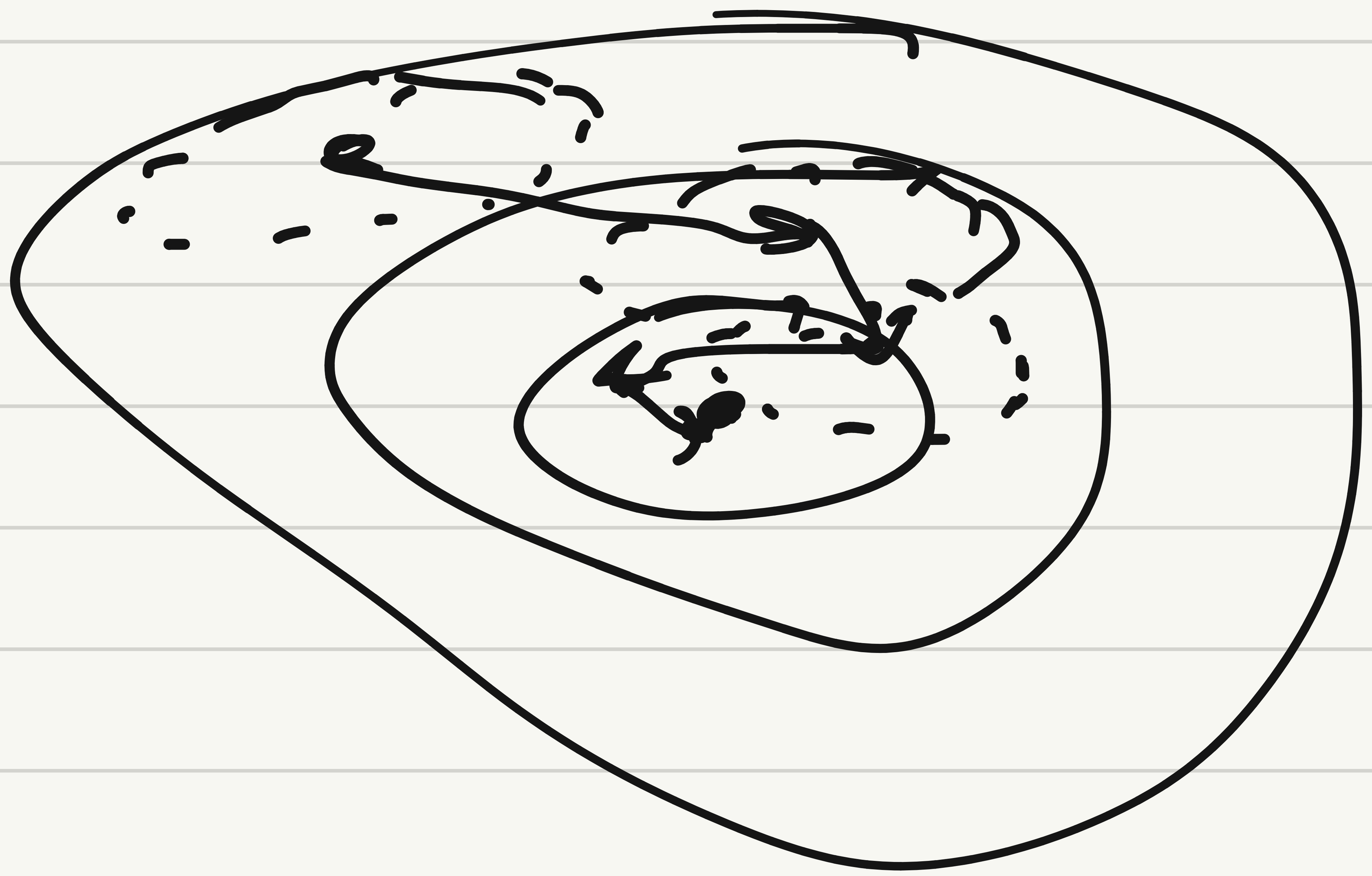
$$\begin{bmatrix} & & \\ & A & \\ & & \end{bmatrix} \begin{bmatrix} m \\ l \\ w \\ h \\ s \end{bmatrix}$$

$\Delta x = -\nabla^2 f(x)^{-1} \nabla f(x)$ is estimate of opt point?

loop: $x = x - \nabla^2 f^{-1} \nabla f$ } "pure" Newton's method



loop:
$d = -\nabla^2 f(x)^{-1} \nabla f(x)$
t = line search
$x = x + td$
} "damped" Newton's method



Convergence of Newton

Hessian bounds m, M

$\nabla^2 f(x) \not\succ 0$

$\underbrace{\nabla^2 f(x) + \mu I}$ $\succ 0$

$I \succ 0$

$\Rightarrow d = -\nabla f(x)$

f is quadratic $\Rightarrow$ $\nabla^2 f$ is const.

Assume $\nabla^2 f$ is Lipschitz continuous with const. L:

$$\| \nabla^2 f(x) - \nabla^2 f(y) \|_2 \leq L \| x - y \|_2$$

Key ideas: m, M bounds $\Rightarrow$ x "close" to x^* $\Leftrightarrow$ $f(x)$ "close" to p^*

$c_1 \| x - x^* \|^2 \leq | f(x) - p^* | \leq c_2 \| x - x^* \|^2$ $\Leftrightarrow$ $\| \nabla f(x) \|$ is "small"

1. **Damped phase:** x far from x^* $\Rightarrow$ $\|\nabla f(x)\|$ large

line search chooses $t \leq 1$

$\Rightarrow$ $f(x) - f(x^*)$ is large $\geq \gamma$

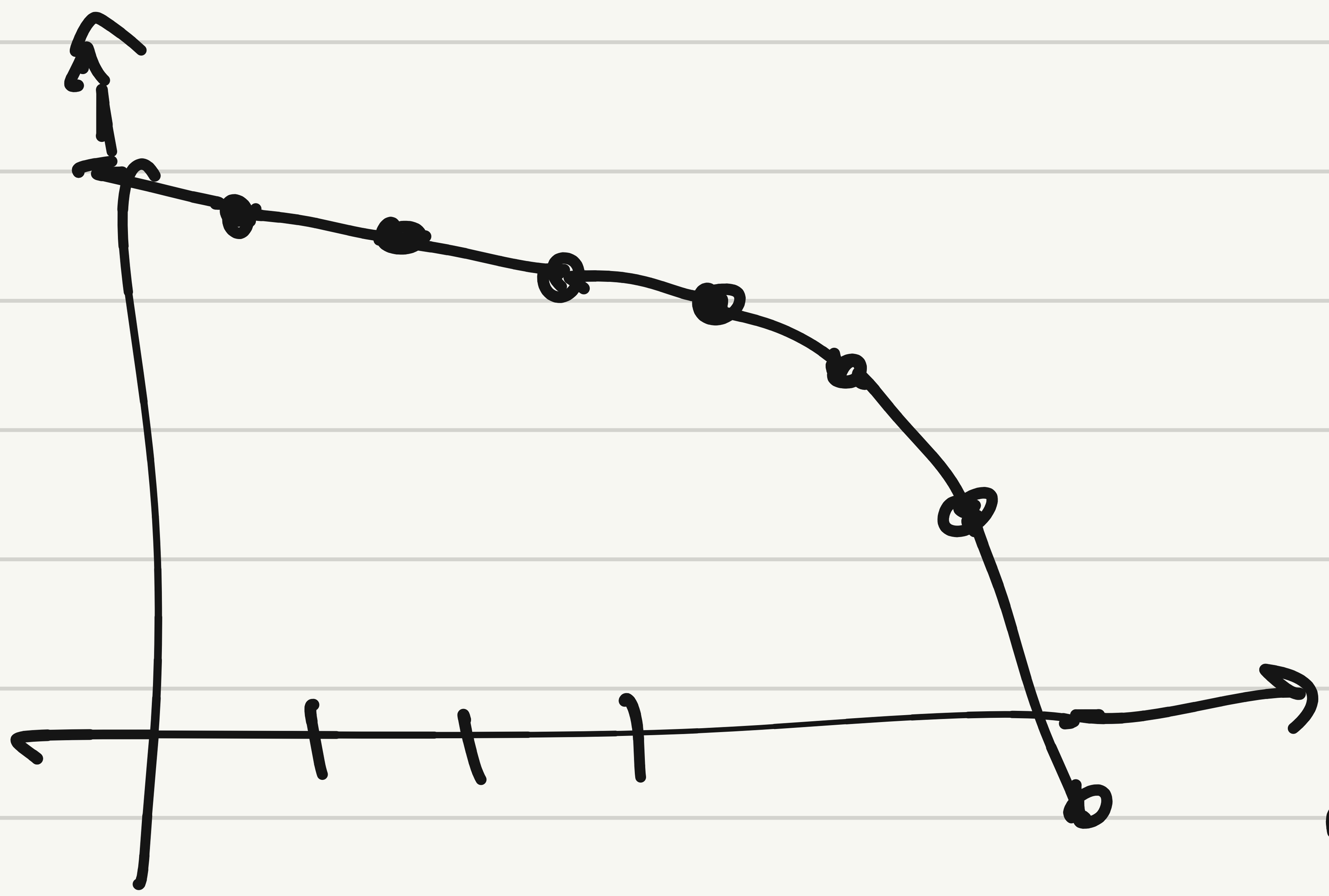
$\Rightarrow$ x gets closer to x^*

2. **Quadratically convergent phase**

x close to x^*

$\Rightarrow$ line search chooses $t = 1$

$\Rightarrow$ $\|\nabla f(x^+)\| = O(\|\nabla f(x)\|^2)$ $\rightarrow$ converge to mach. prec. in ~5-6 iters



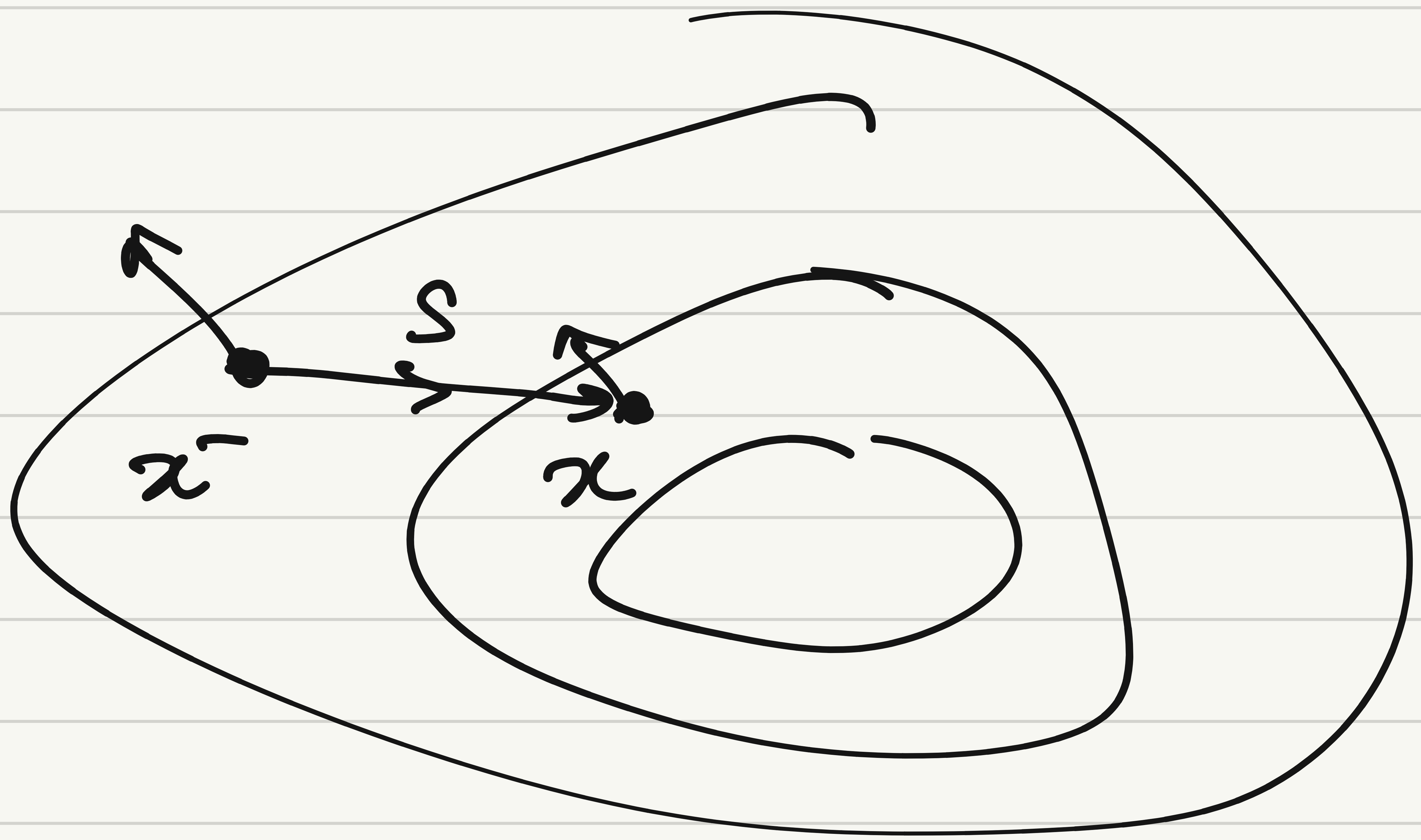
#iters

$$\text{total \# iters} \leq \frac{|f(x) - p^*|}{\gamma} + 6$$

↳ depends on m, M, L

Variations of Newton → [Quasi-Newton method]

- BFGS : Secant / Broyden's
- Gauss-Newton : $\min \|\vec{f}(\vec{x})\|_2$



$\nabla^2 f(x)$: $O(n^2)$ flops

$(\nabla^2 f(x))^{-1} \nabla f(x)$: $O(n^3)$ flops

I know $\nabla f(x^{(k-1)})$, $\nabla f(x^{(k)})$ ⟶ estimate $\nabla^2 f(x^{(k)})$

$\nabla f(x^-)$, $\nabla f(x)$

$s = x - x^-$, $y = \nabla f(x) - \nabla f(x^-)$. find B s.t. $B\underline{s} = \underline{y}$

Current guess B, find B^+ s.t. $\underline{B^+ s = y}$

- symmetric
- pos definite

Broyden's method: $B^+ = B + \underline{uv^T}$ ← not symm!

Symm. rank-1 update: $\underline{B^+ = B + \alpha u u^T}$ — not always pos. def.
(SR1) — α: scalar

$$B^+ = B + \alpha u u^T + \beta v v^T$$

"Best" choice: $u = Bs$, $v = y$

"Best": minimizes $\| B_+^{-1} - B^{-1} \|_{norm}$

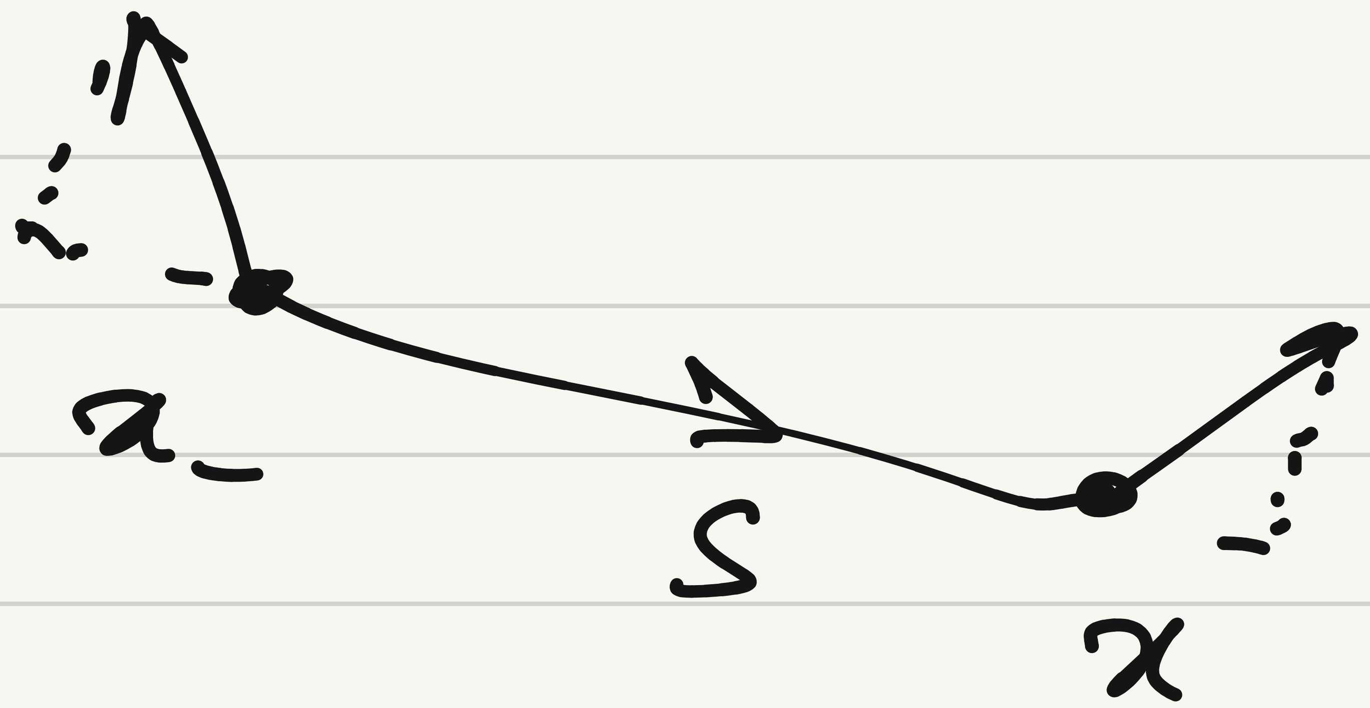
$$\Rightarrow \alpha = -\frac{1}{s^T B s}, \quad \beta = \frac{1}{y^T s}$$

$$B^+ = B - \frac{B s s^T B}{s^T B s} + \frac{y y^T}{y^T s}$$

: **BFGS update** — pos def if $\underline{y^T s > 0}$

$y^T s > 0 \Rightarrow \nabla f(x)^T s > \nabla f(x^-)^T s$

always true for convex fns!



Rate of conv. superlinear but subquadratic

Cheaper: $B = LL^T \rightarrow B^+ = L^+ L^{+T}$