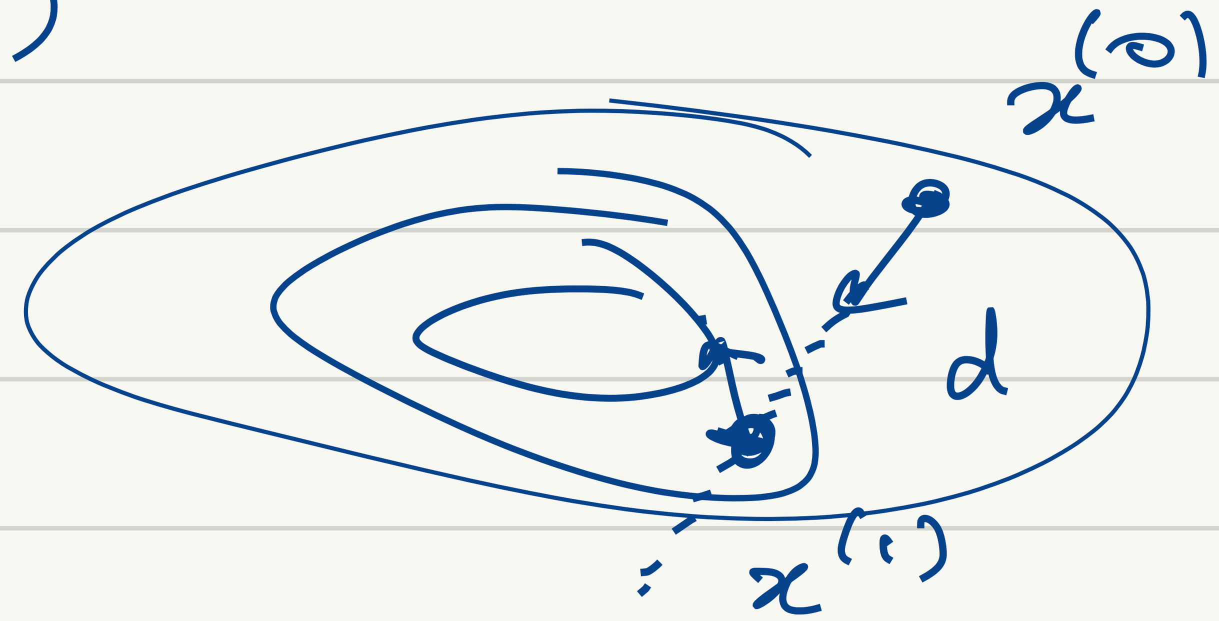
Descent methods

$$f(x^+) \approx f(x) + \underbrace{t\, \nabla f(x)^T \vec{d}} + O(t^2)$$

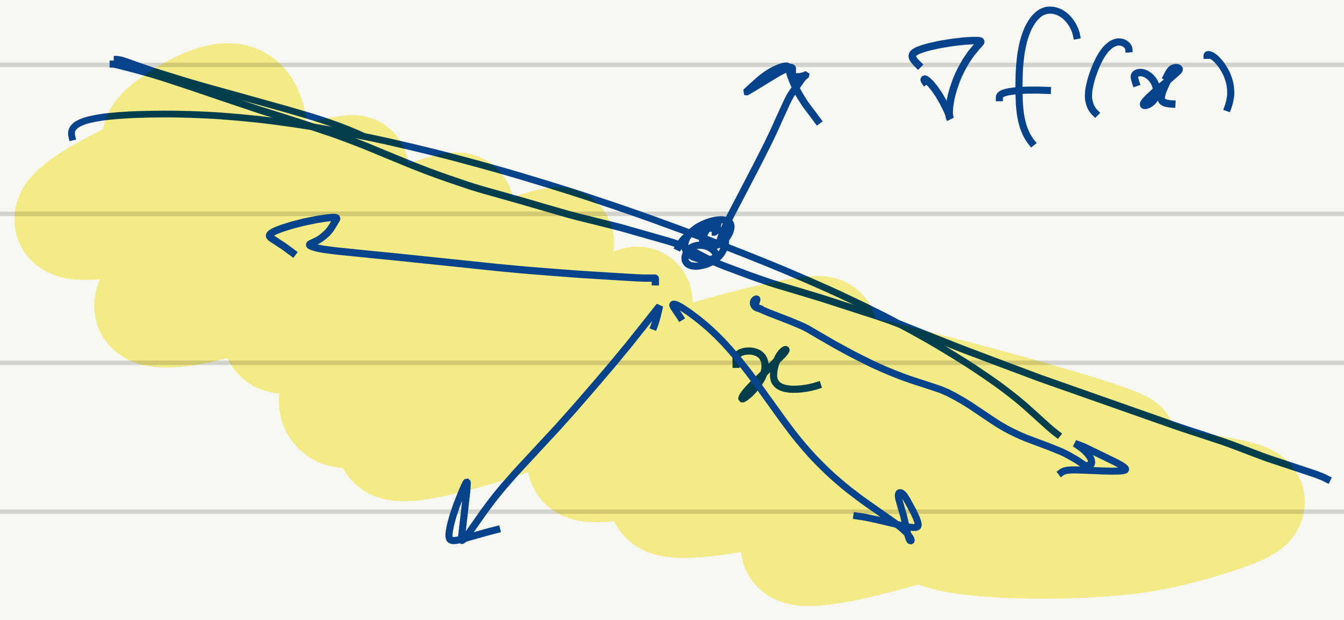


minimize $f(\vec{x})$ over all $\vec{x} \in \mathbb{R}^n$

$\vec{x}^{(0)} \to \vec{x}^{(1)} \to \vec{x}^{(2)} \to \cdots$ $\qquad f(x^{(k+1)}) < f(x^{(k)})$

$$f(x^+) < f(x)$$

Repeat:



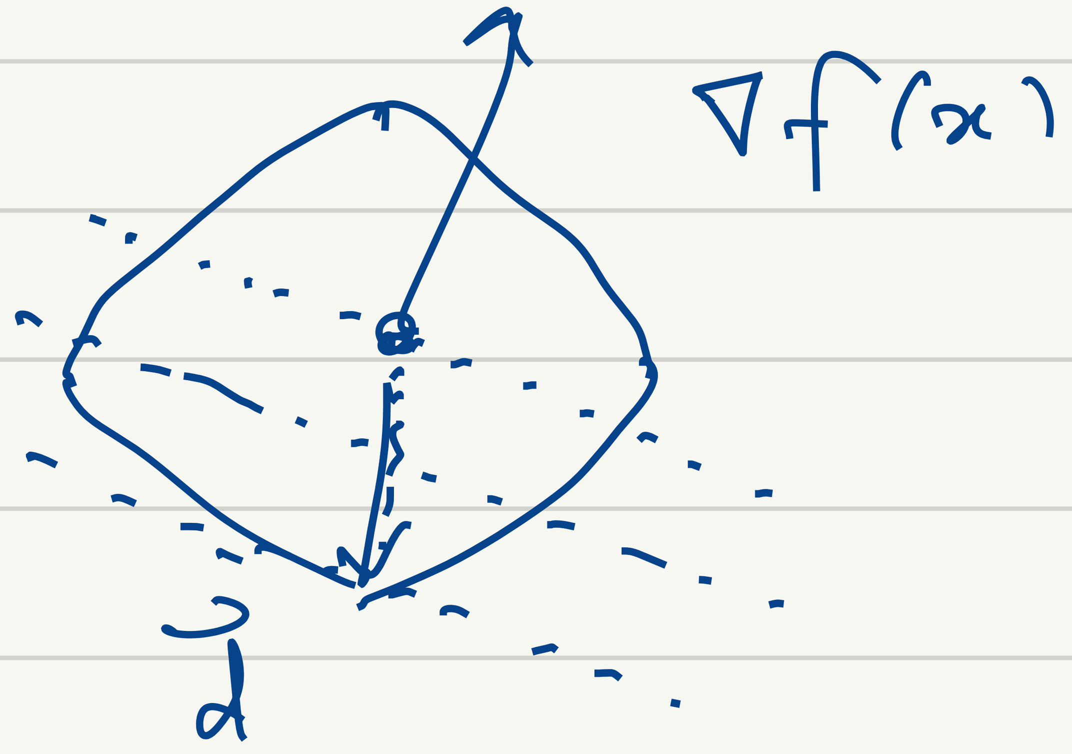
- Pick <u>descent direction</u> $\vec{d}$ s.t. $\nabla f(x)^T \vec{d} < 0$
- Pick step size $t \in \mathbb{R}_+$ s.t. $f(\vec{x} + t\vec{d}) < f(\vec{x}) + \alpha t \underbrace{\nabla f(\vec{x})^T \vec{d}}$
- Set $\vec{x}^+ = \vec{x} + t\vec{d}$

line search

$\boxed{\text{Steepest descent}}$: Choose $\vec{d}$ to decrease f as fast as possible : $\underline{\nabla f(x)^T \vec{d}}$ is as negative as possible

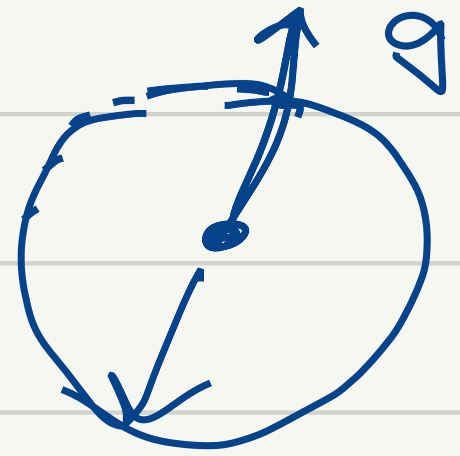
Pat bound: Pick $\vec{d}$ to minimizes $\nabla f(x)^T \vec{d}$ s.t. $\|\vec{d}\| = 1$

(normalized) $\boxed{\text{steepest descent direction}}$



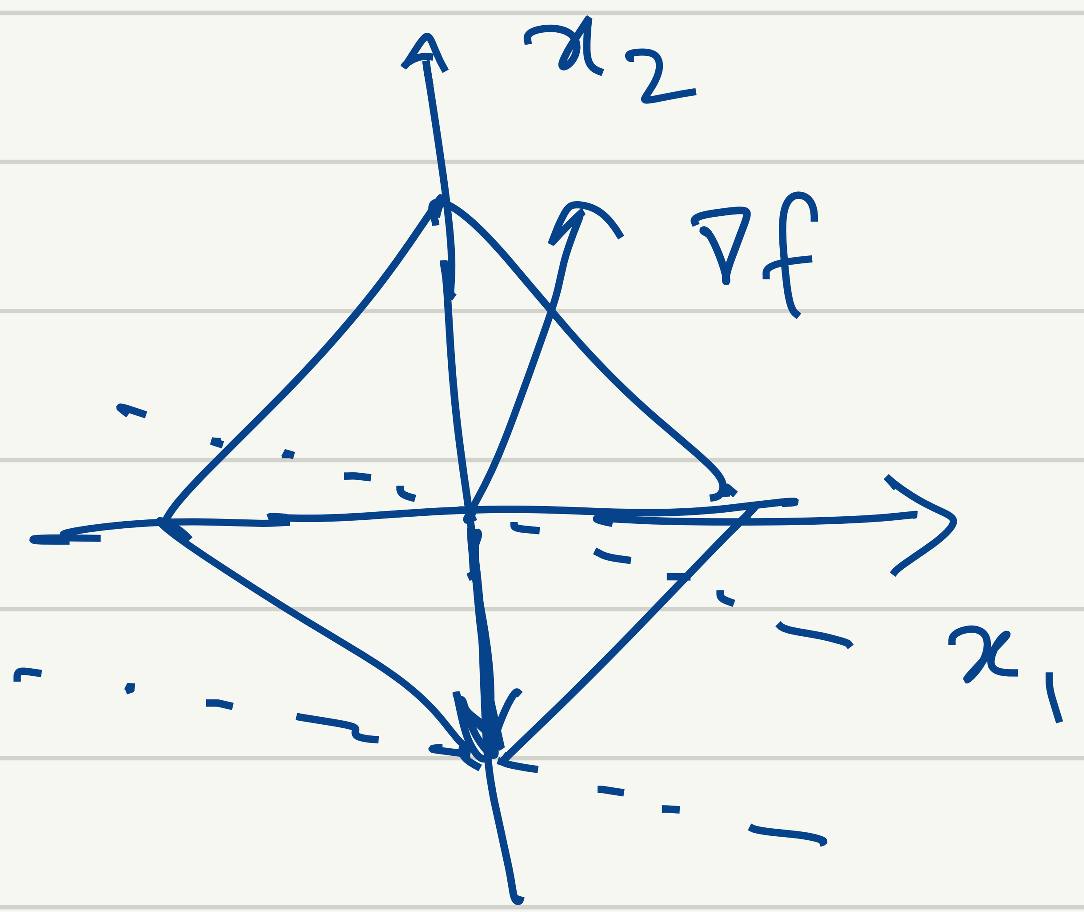
If $\|\cdot\|_2$: $\min \vec{g}^T \vec{d}$ s.t. $\vec{d}^T \vec{d} = 1$

$\Rightarrow \vec{d} = -\vec{g}$



$\vec{d} = -\dfrac{\vec{g}}{\|\vec{g}\|}$ $\boxed{\text{gradient descent}}$

$\|\cdot\|_1$:



If i is the index with $\max \left| \dfrac{\partial f}{\partial x_i} \right|$

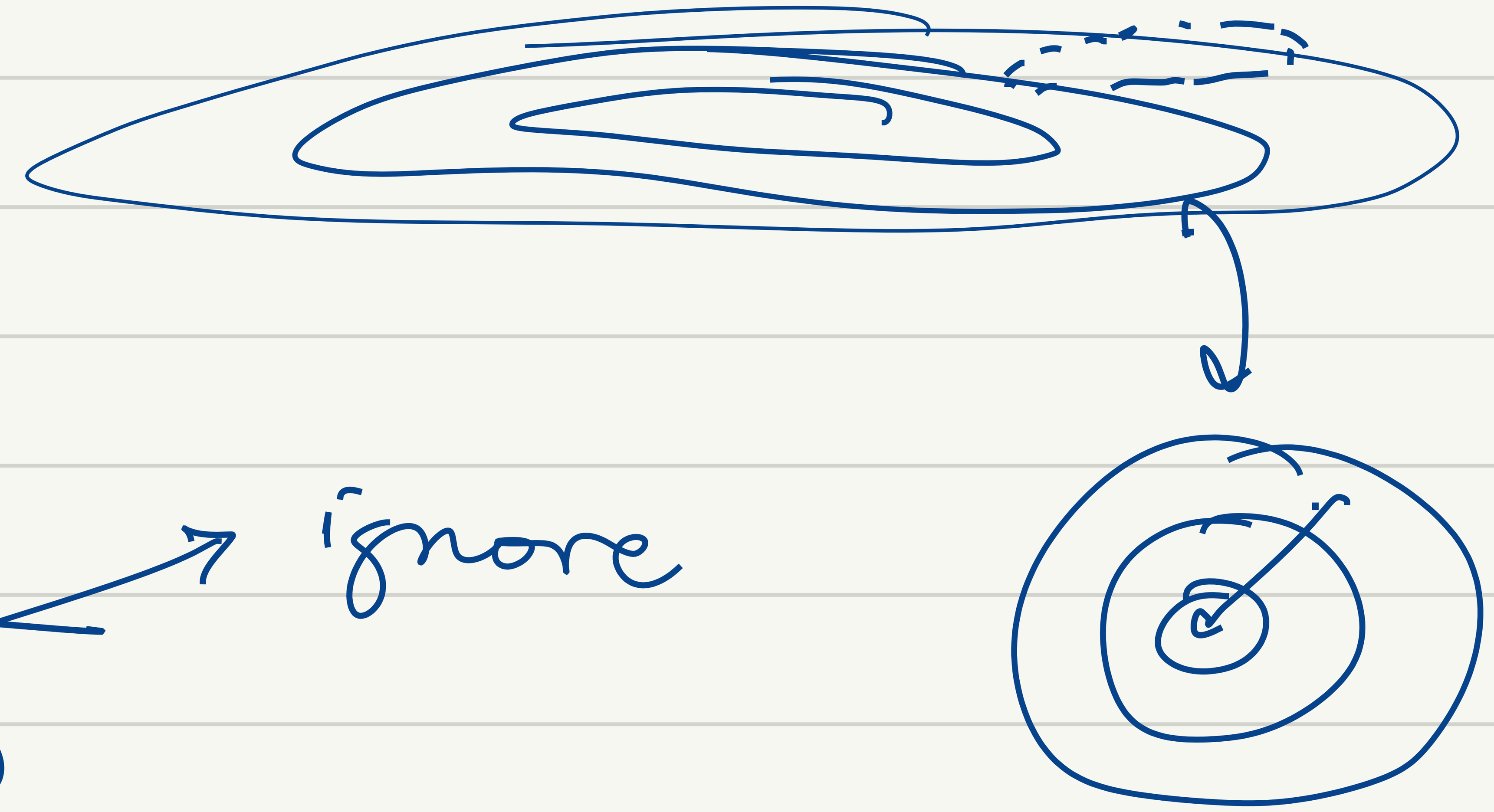
Then $\vec{d} = -\operatorname{sign}\left(\dfrac{\partial f}{\partial x_i}\right) \vec{e}_i$

$\boxed{\text{coordinate descent}}$

$x^+ = x + t\vec{e}_i$

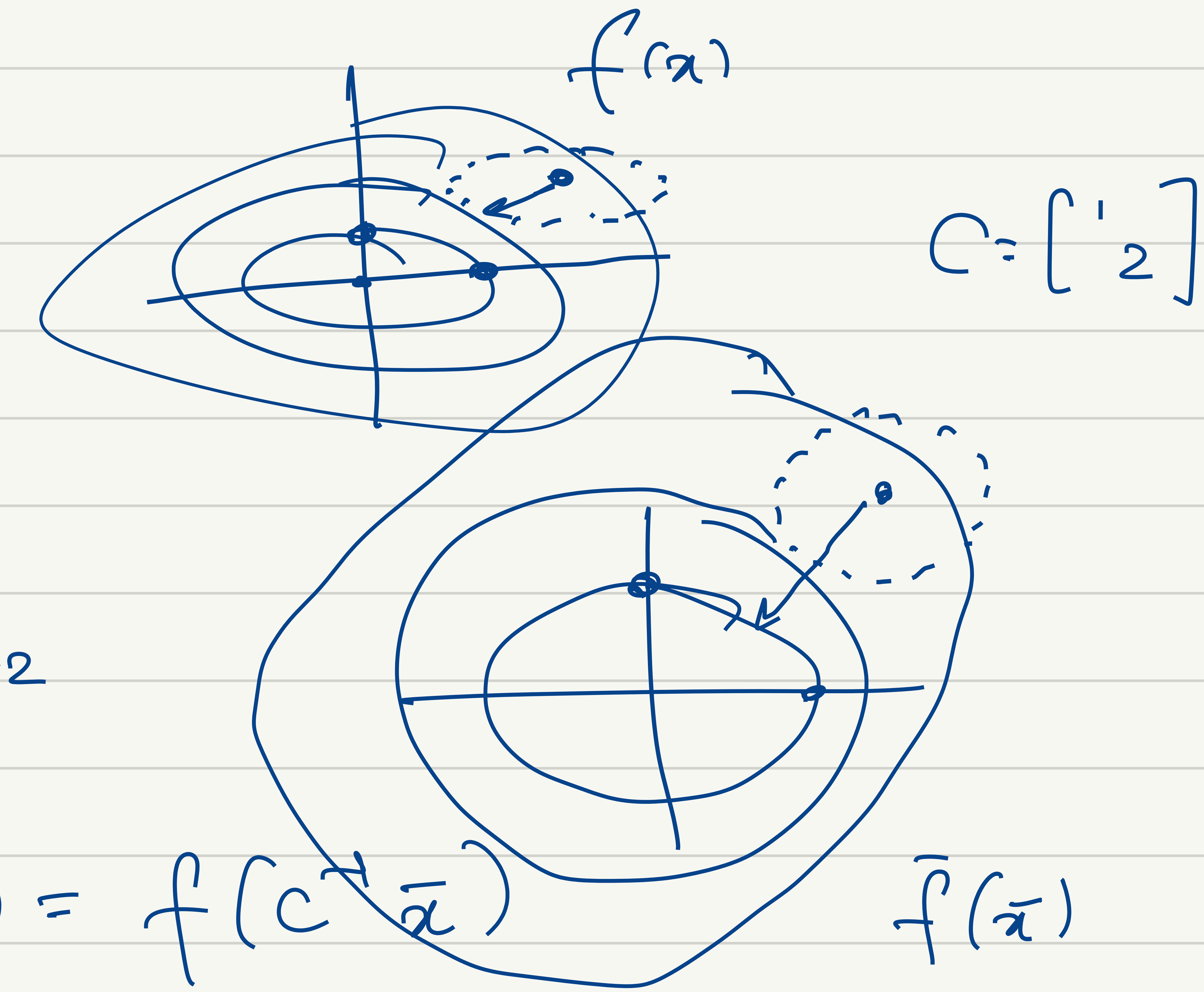
Let $P \succ 0$: symm. pos. def.

$\|\vec{x}\|_P = \sqrt{\vec{x}^T P \vec{x}}$ then $\vec{d} = -\dfrac{P^{-1} \nabla f(\vec{x})}{\cancel{\sqrt{\nabla f(x)^T P \nabla f(x)}}}$ ignore



(quadratic norm)

$$P \vec{d} = -\nabla f(x)$$



Suppose $P = C^T C$

(eg. Cholesky, or $P = Q \Lambda Q^T$, $C = Q \sqrt{\Lambda} Q^T$)

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

min $f(x)$ $\longrightarrow$ $\bar{x} = Cx$, $\bar{f}(\bar{x}) = f(x) = f(C^{-1}\bar{x})$

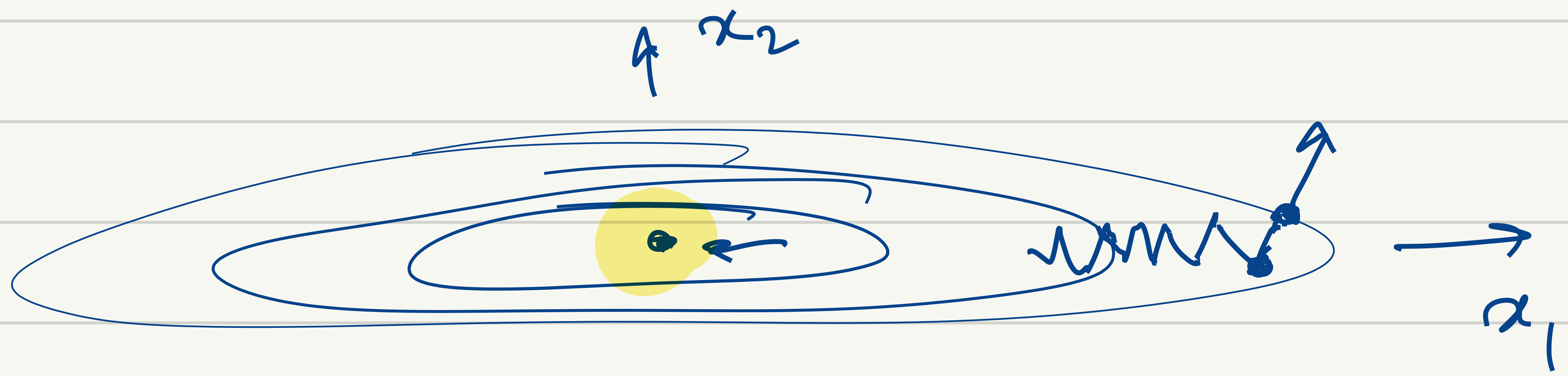
$\equiv$

min $\bar{f}(\vec{x})$

Steepest descent of f on x $\equiv$ grad. desc. of $\bar{f}$ on $\bar{x}$!

Analysis of GD & SD

$\vec{x} \in \mathbb{R}^2 \quad , \quad \gamma \in \mathbb{R}_+$



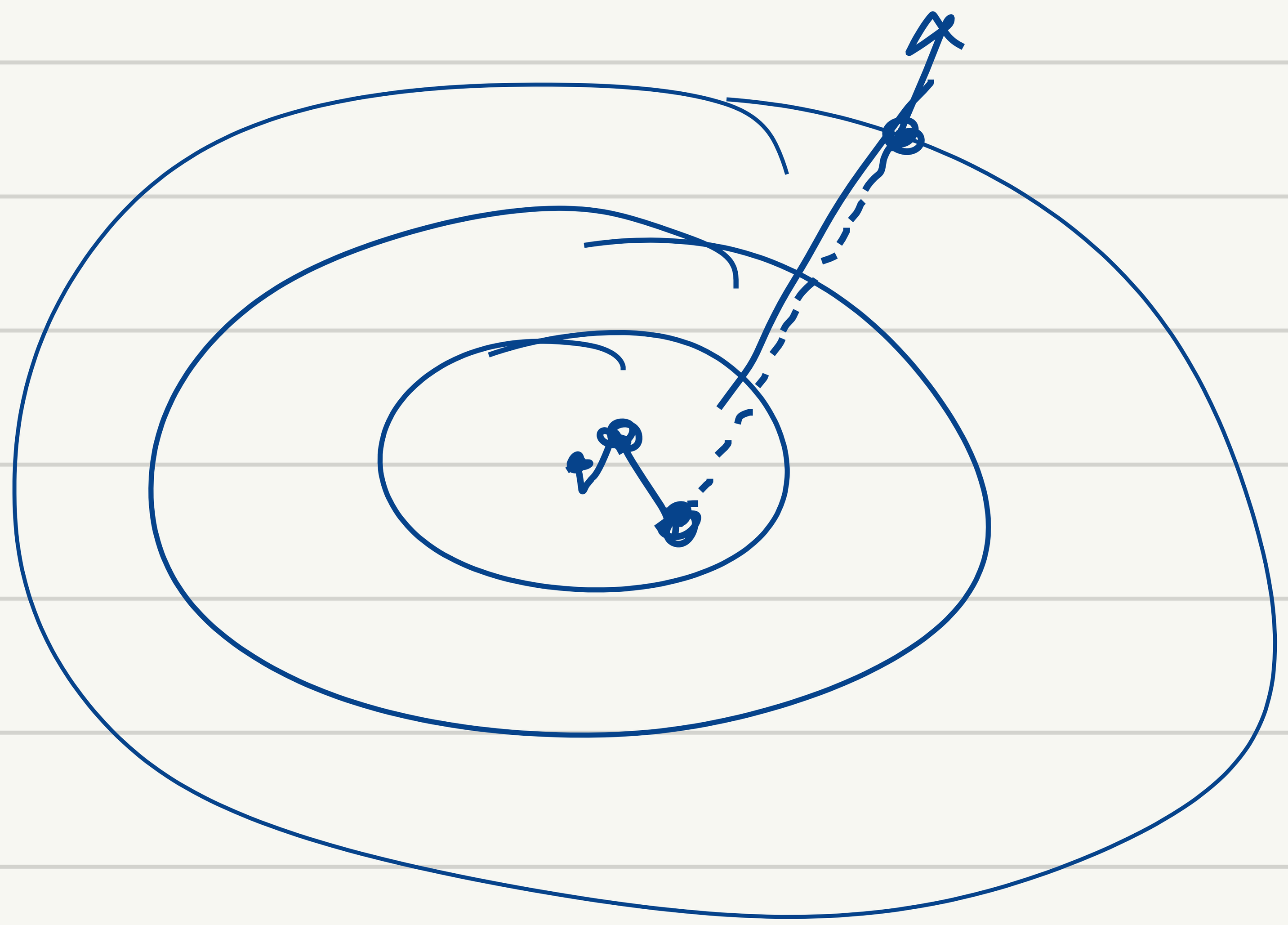
$$f(\vec{x}) = \frac{1}{2}\left(x_1^2 + \gamma x_2^2\right)$$

$$\nabla f(x) = \begin{bmatrix} x_1 \\ \gamma x_2 \end{bmatrix}$$

$\gamma \gg 1$

$$\underbrace{\nabla^2 f = \begin{bmatrix} 1 & \\ & \gamma \end{bmatrix}}$$

$\gamma \approx 1$



Assume : f convex, twice diff. $\nabla^2 f(x)$

Assume $\exists\ m, M \in \mathbb{R}$ s.t. $\boxed{mI \preceq \nabla^2 f(x) \preceq MI}$ for all x

$$\underbrace{\nabla^2 f(x) - mI} \succeq 0 \iff \lambda_i\left(\nabla^2 f(x)\right) \geq m \quad , \text{ similarly } \lambda_i \leq M$$

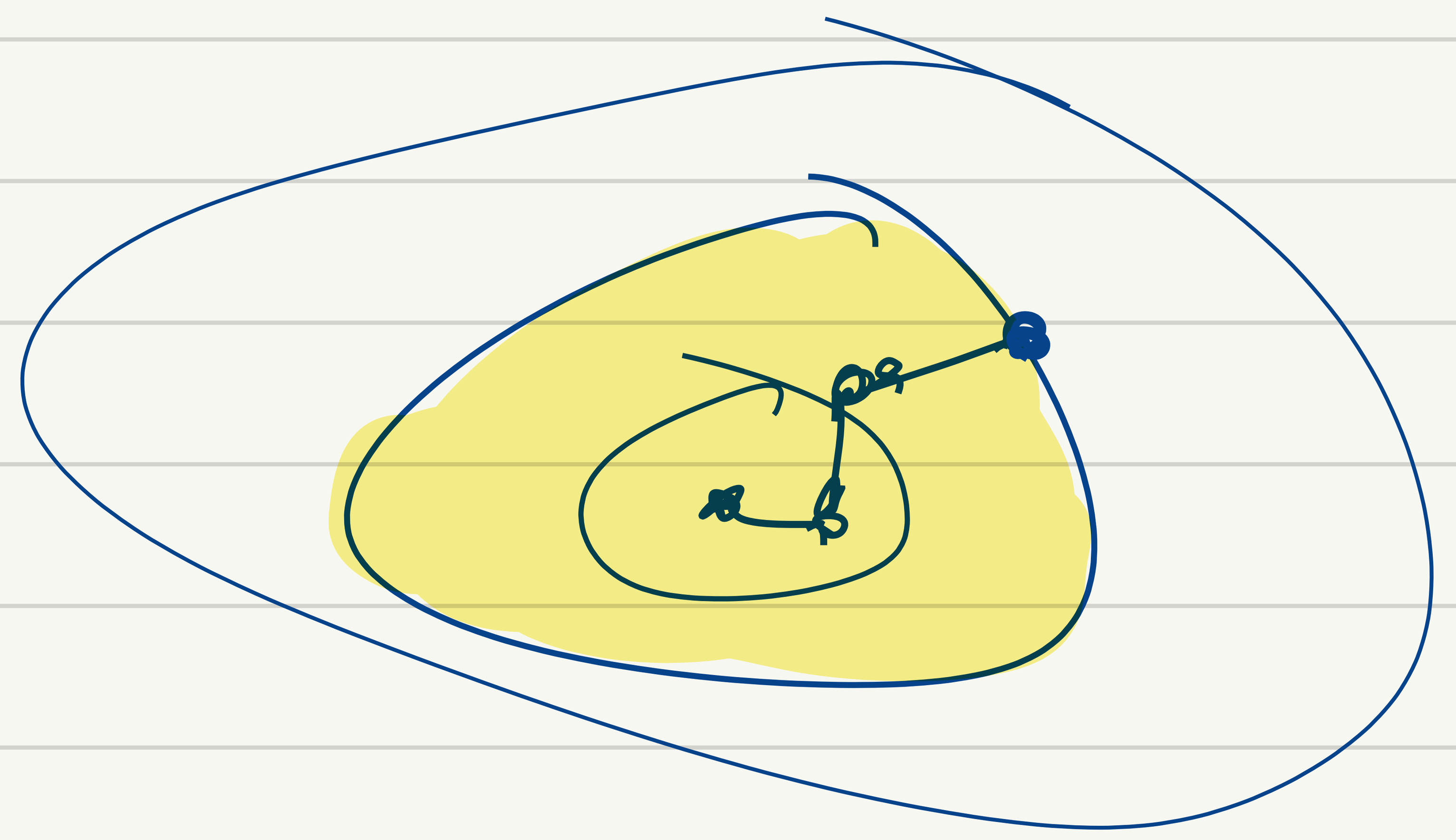
for any vector $\vec{v}$, $\underbrace{m\|v\|^2}_{} \leq v^T \underbrace{\nabla^2 f(x)}_{} v \leq \underbrace{M\|v\|^2}_{}$

$x^{(0)} \rightarrow x^{(1)} \rightarrow x^{(2)} \rightarrow \cdots$

$f(x^{(k)}) < f(x^{(k-1)}) < \cdots < f(x^{(0)}) \Rightarrow x^{(k)} \in S_0 = \{x : f(x) \leq f(x^{(0)})\}$

$\exists\ m, M$ s.t.

$mI \preceq \nabla^2 f(x) \preceq MI \quad \forall\ x \in S_0$



$\Rightarrow \lambda_i \in [m, M]$

$\Rightarrow \kappa(\nabla^2 f(x)) \leq M/m$

Assume $m > 0$: f is $\boxed{\text{strongly convex}}$

Taylor's thm: for any y,

$$f(y) = f(x) + \nabla f(x)^T (y-x) + \frac{1}{2} (y-x)^T \underbrace{\nabla^2 f(z)} (y-x) \quad \text{for some } z = x + t(y-x), \; 0 \le t \le 1$$

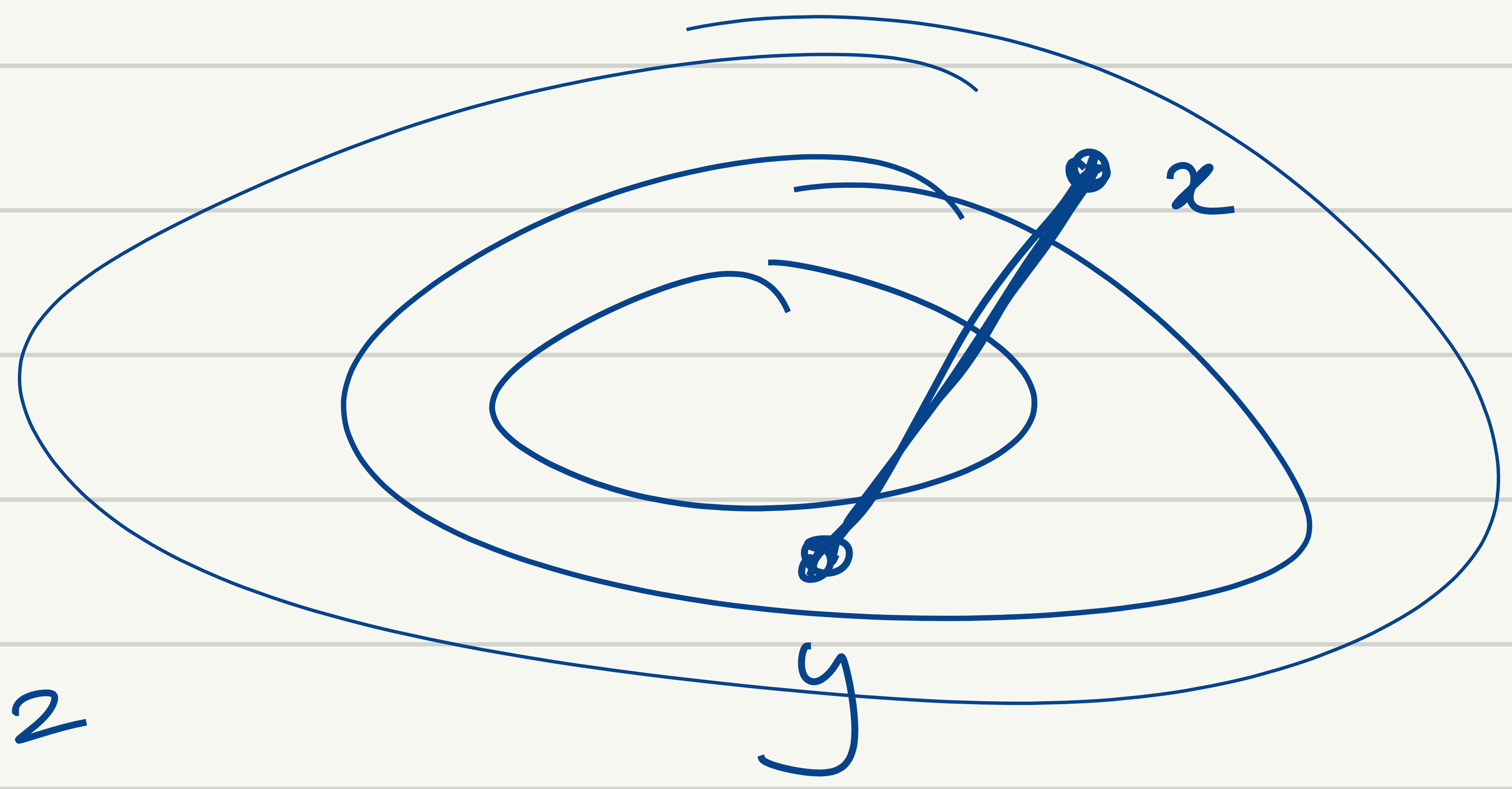
$$\in f(x) + \nabla f(x)^T (y-x) + \frac{1}{2} \underbrace{[m, M]} \cdot \|y-x\|^2$$

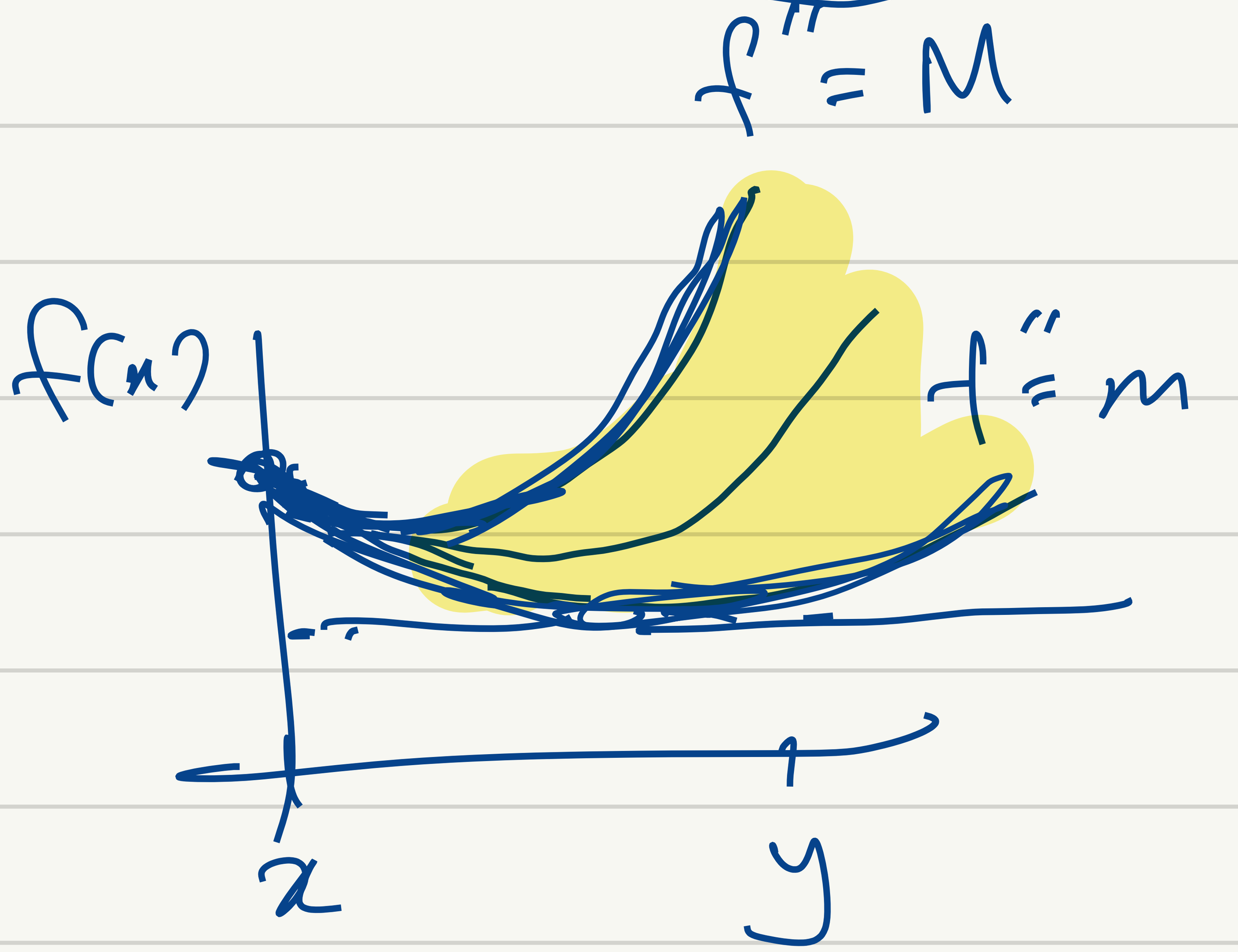
$$\underbrace{f(x) + \nabla f(x)^T (y-x) + \frac{1}{2} m \|y-x\|^2} \le f(y) \le \underbrace{f(x) + \cdots + \frac{1}{2} M \|y-x\|^2}$$

for all y, $f(y) \ge f(x) + \nabla f(x)^T (y-x)$

$$+ \underbrace{\frac{1}{2} m \|y-x\|^2}$$

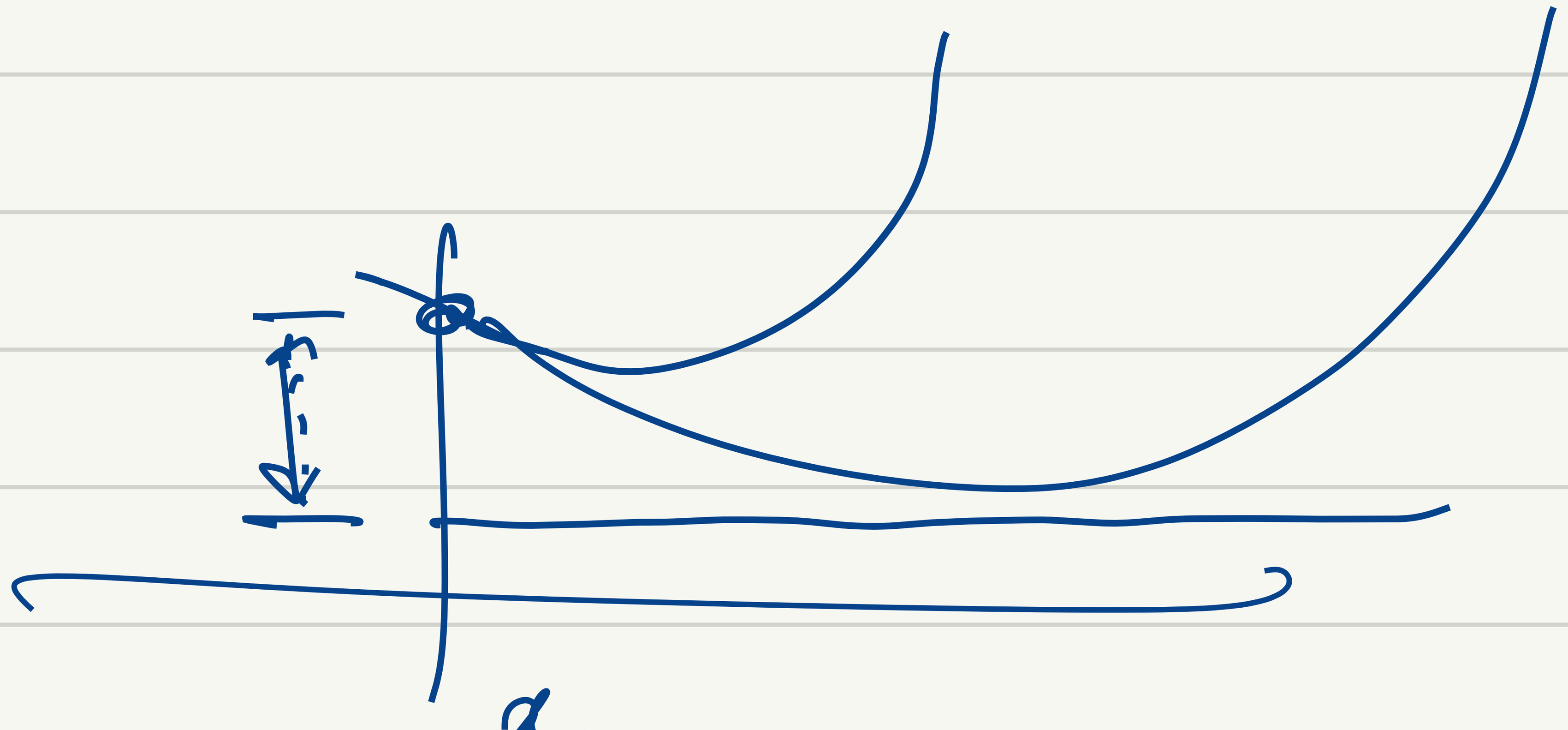
$$\ge \underbrace{f(x) - \frac{1}{2m} \|\nabla f(x)\|^2}$$





optimal value $p^* = \min_y f(y) \geq f(x) - \frac{1}{2m} \|\nabla f(x)\|^2$

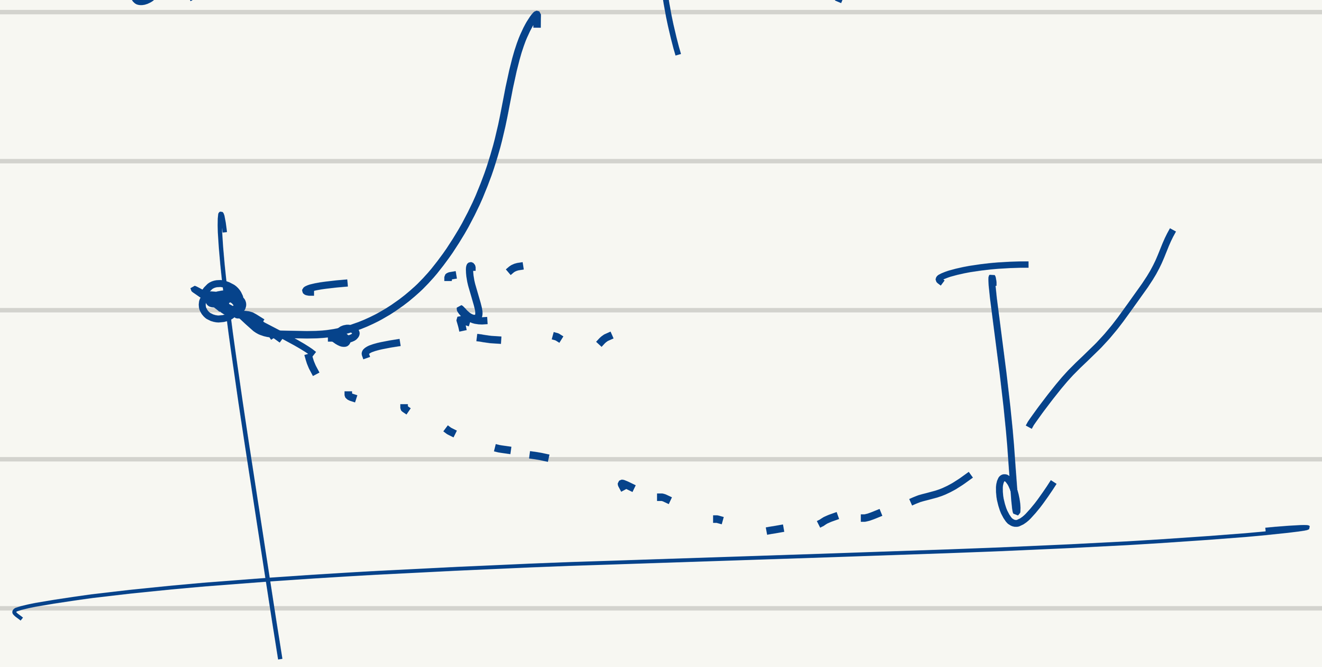
$\boxed{\text{Suboptimality}}$ $f(x) - p^* \leq \frac{1}{2m} \|\nabla f(x)\|^2$



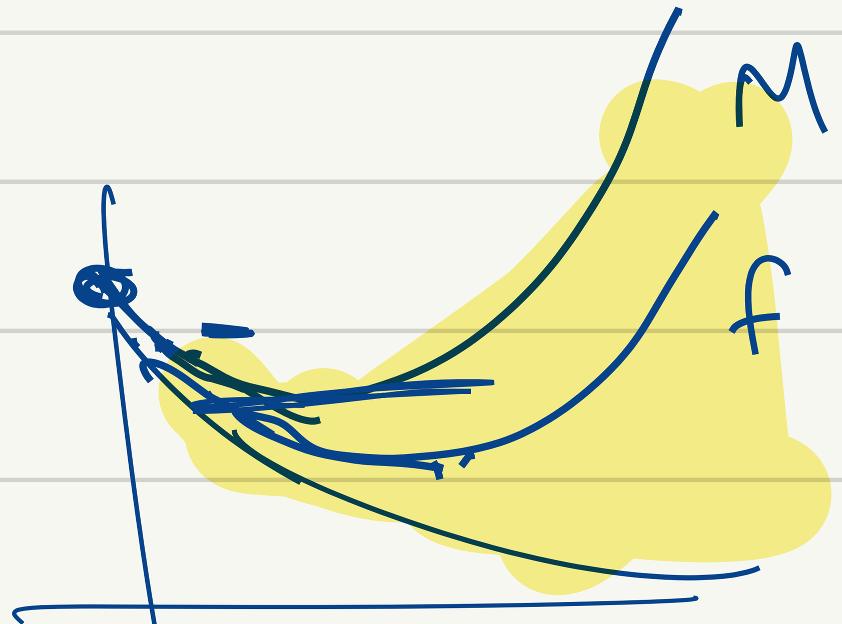
Gradient descent:

$d = -\nabla f(x)$

$$f(y) \leq \overbrace{f(x) + \nabla f(x)^T (y - x) + \frac{M}{2} \|y - x\|^2}$$

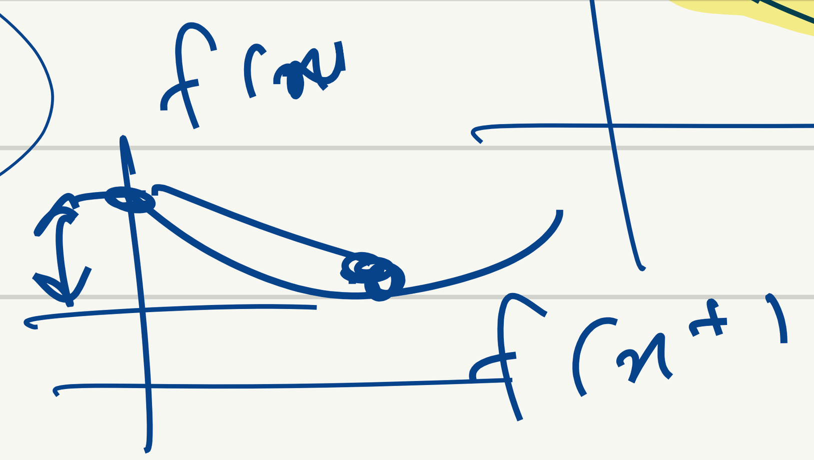


$y = x - t \nabla f(x)$, exact line search:

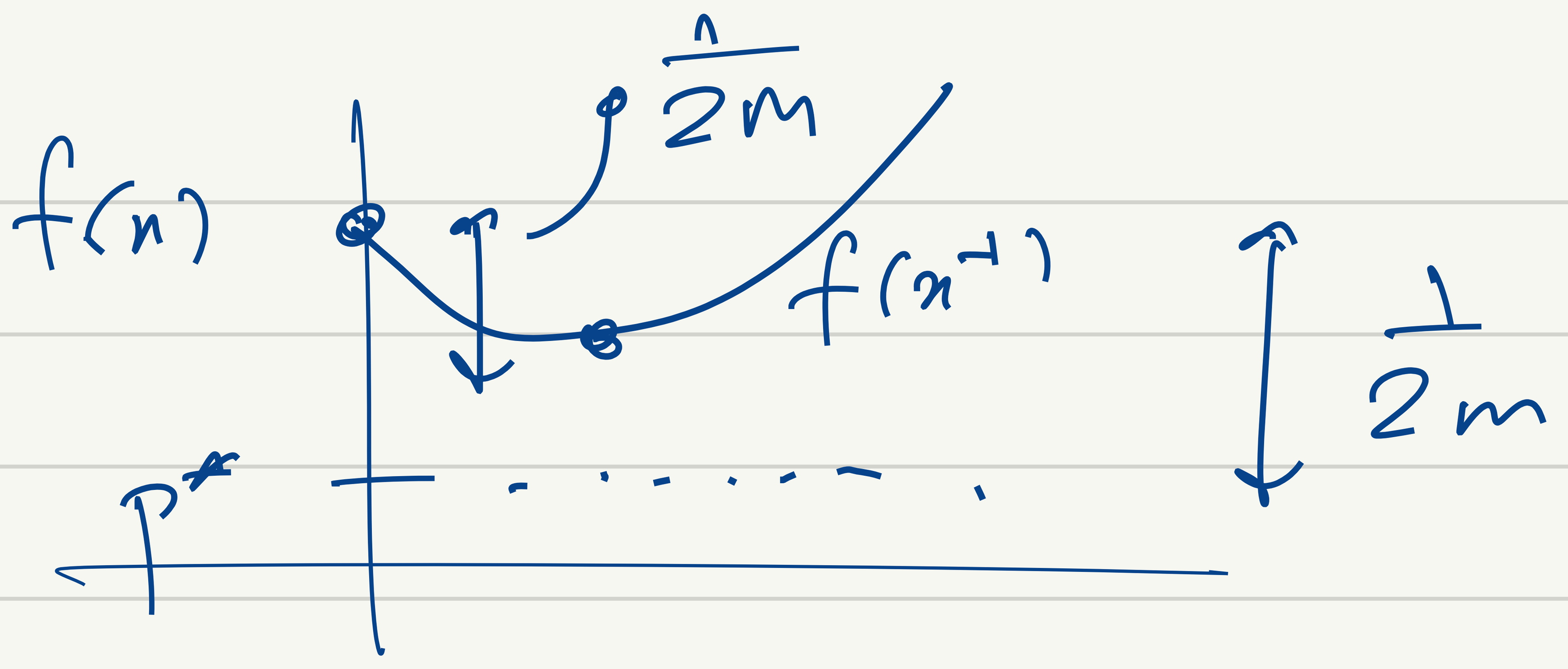


$$f(x^+) \leq f(x) - \frac{1}{2M} \|y - x\|^2$$

$$\Rightarrow \underbrace{f(x) - f(x^+)} \geq \underbrace{\frac{1}{2M} \|y - x\|^2}$$



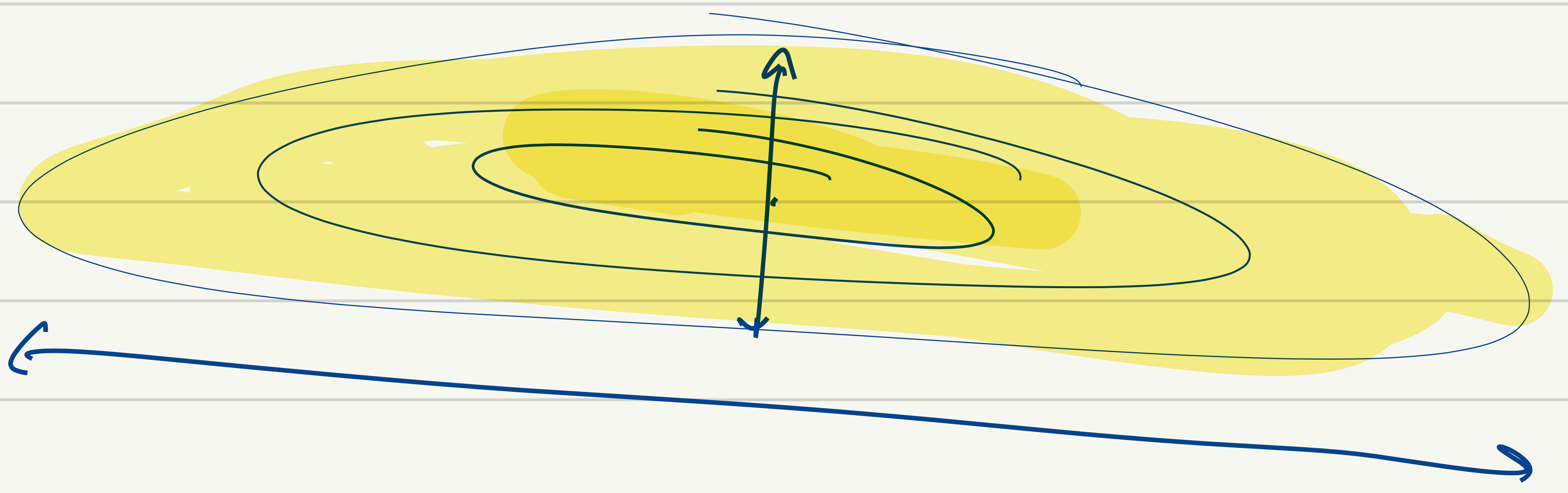
Subopt : $f(x) - p^* \le \frac{1}{2m} \|\nabla f(x)\|^2$



progress : $f(x) - f(x^+) \ge \frac{1}{2M} \|\nabla f(x)\|^2$

Each iter of GD decreases suboptimality by $1 - \frac{m}{M}$

$$f(x^+) - p^* \le \underbrace{\left(1 - \frac{m}{M}\right)}_{\text{on } f} \left(f(x) - p^*\right)$$



Steepest descent with $\|\cdot\|_P$:

$$\equiv \quad \min \bar{f}(\bar{x}) = f(C^{-1}\bar{x}) \quad \text{where} \quad \bar{x} = Cx, \quad C^T C = P$$

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

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

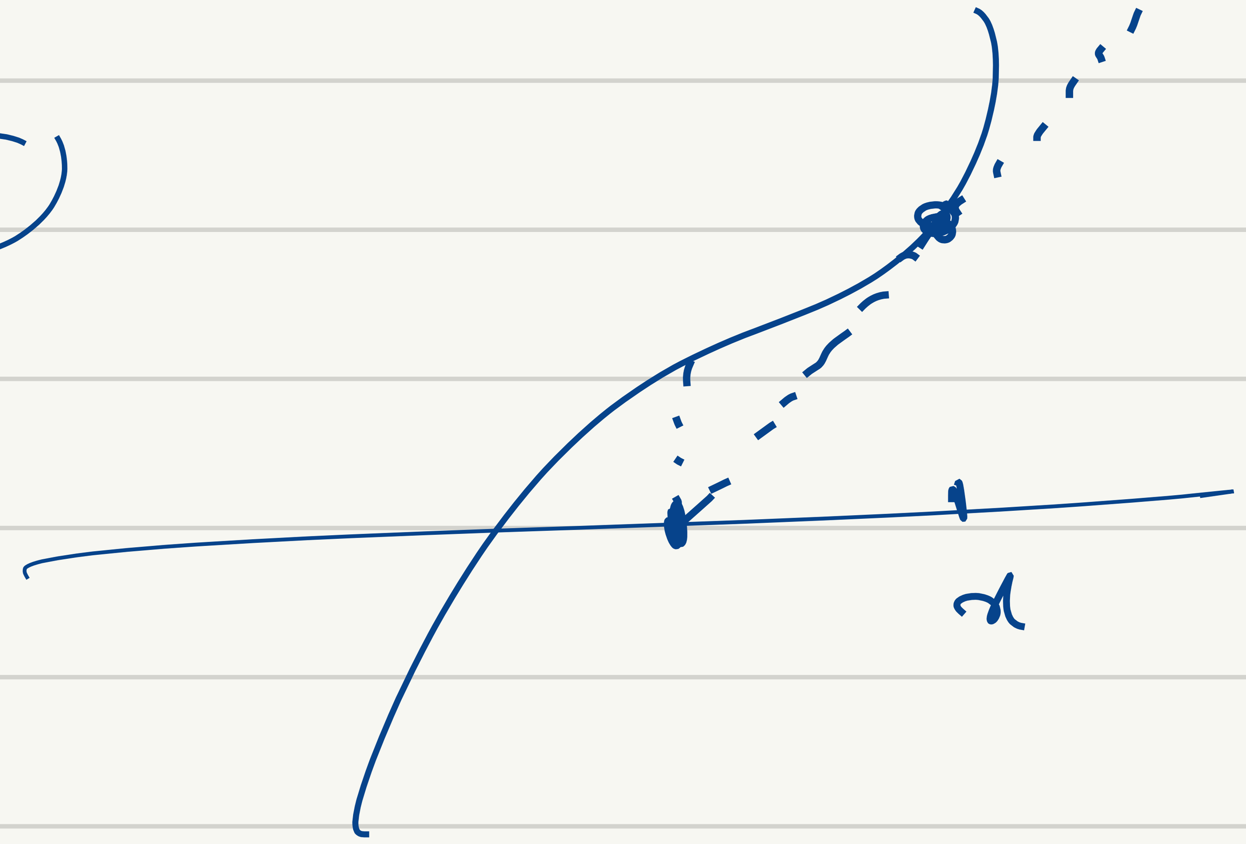
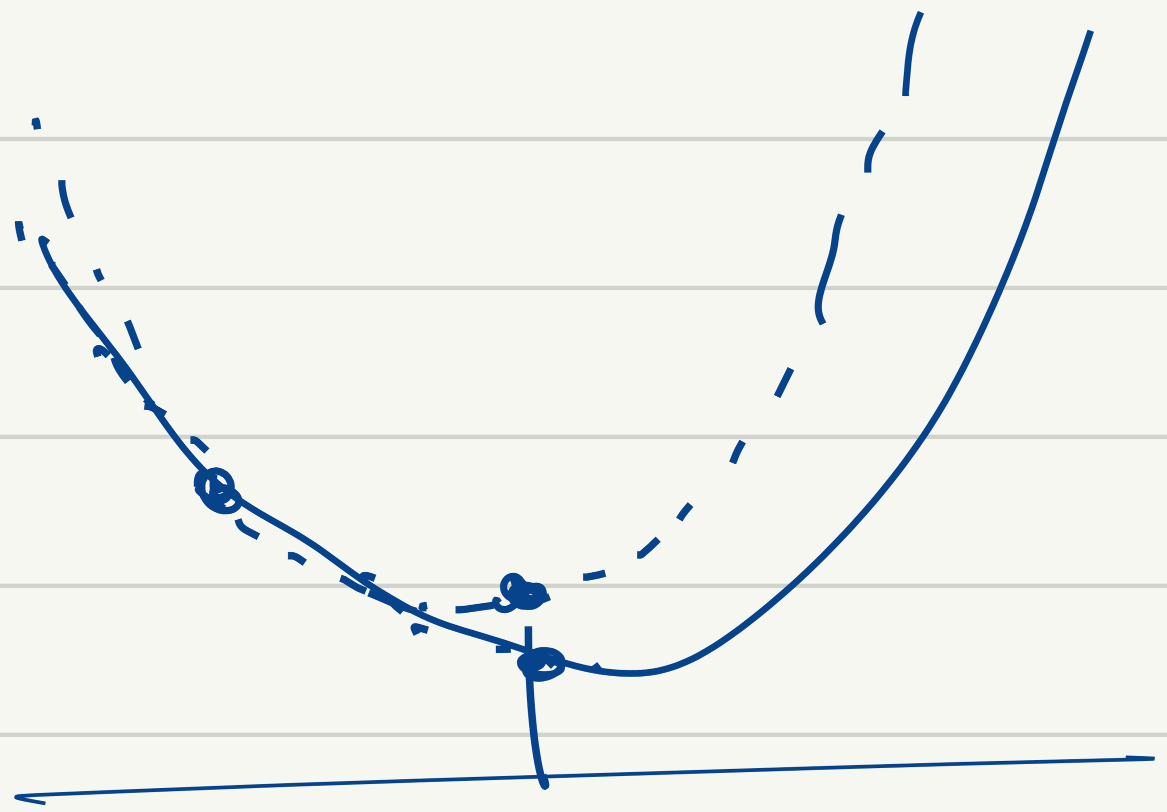
$$\Rightarrow C^{-T} \underbrace{\nabla^2 f(x)} C^{-1} \approx ?$$

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

Choose $\vec{d} = P^{-1} \nabla f(x) = \boxed{(\nabla^2 f(x))^{-1} \nabla f(x)}$: $\boxed{\text{Newton's method}}$ for optimization

for strongly convex fns, $\nabla^2 f(x) \succ 0$,

So well defined, $\vec{d}$ is descent direction

②  

$f(x)$

2nd-order Taylor series:

$$\hat{f}(y) = f(x) + \nabla f(x)^T (y-x) + \frac{1}{2}(y-x)^T \nabla^2 f(x) (y-x)$$

minimize $\hat{f}(y)$

→ solve $\nabla \hat{f}(y) = 0 \Rightarrow \nabla f(x) + \nabla^2 f(x) \cdot (y-x) = 0$

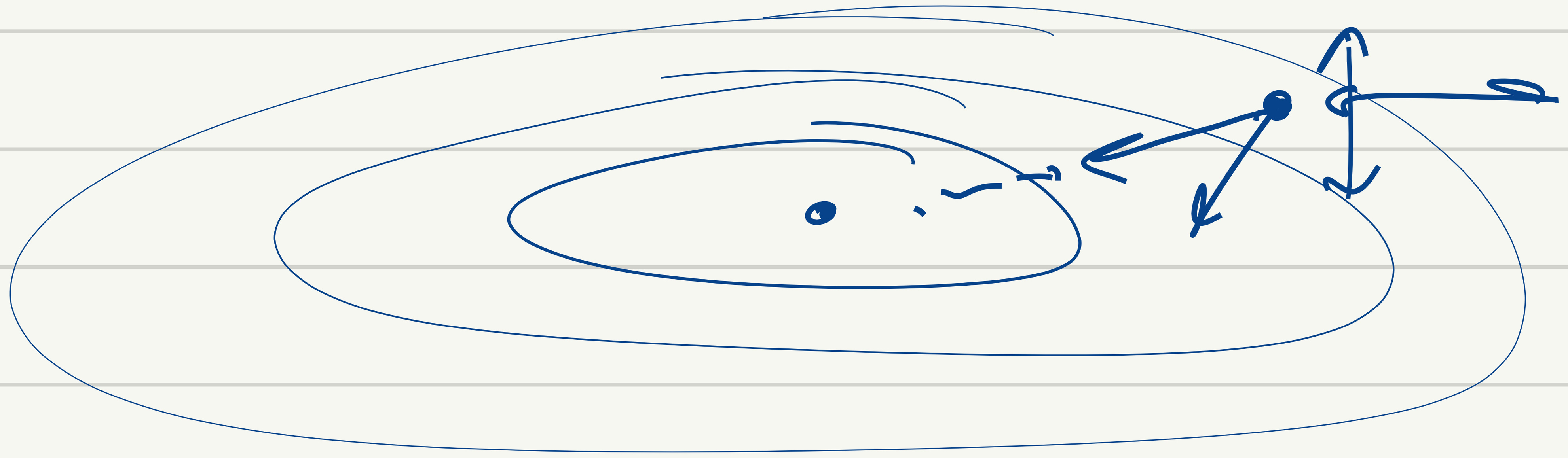
$$\Rightarrow y = x - \underbrace{(\nabla^2 f(x))^{-1} \nabla f(x)}$$

(3) optimum : $\nabla f(x) = 0$ $\qquad$ $g(x) = \nabla f(x)$

$$g(x) = 0$$

$$x^+ = x - J(x)^{-1} g(x) = x - \nabla^2 f(x)^{-1} \nabla f(x)$$

$$J(x) = \left[\partial g_i / \partial x_j \right] = \nabla^2 f(x)$$



$$f(x) = \frac{1}{2} (x_1^2 + \gamma x_2^2)$$

<u>Properties</u>: affine invariant : $\min f(x)$ $\qquad$ $x = Az + b$

$\iff \min g(z)$ $\qquad$ $g(z) = f(Az + b)$