Lecture 16. Proximal mapping; Lagrange multiplier; submanifolds.

Recall gradient descent: $x_{k+1} = \arg\min_y f(x_k) + \nabla f(x_k)^T(y-x) + \frac{1}{2t}\|y-x\|^2$.

Let $f = g + h$. $x_{k+1} = \arg\min_y g(x_k) + \nabla g(x_k)^T(y-x_k) + \frac{1}{2t}\|y-x_k\|^2 + h(y)$.

convex differentiable (g) convex (h)

$$= \arg\min_y \frac{1}{2t}\|y-(x_k - t\nabla g(x_k))\|^2 + h(y).$$

proximal gradient descent: $x_{k+1} = \mathrm{prox}_{h,t}(x_k - t\nabla g(x_k))$.

$\mathrm{prox}_{h,t}(x) = \arg\min_y \frac{1}{2t}\|y-x\|^2 + h(y)$. proximal mapping.

$x_{k+1} = x_k - tG_k$ where $G_k = \frac{1}{t}(x_{k+1} - \mathrm{prox}_{h,t}(x_k - t\nabla g(x_k)))$.

Then we obtain $f(x_{k+1}) \le f(y) - G_t(x_k)^T(y-x_k) - \frac{m}{2}\|x_k-y\|^2 - \frac{t}{2}\|G_t(x_k)\|^2$ $\forall y$.

key ingredients: $w = \mathrm{prox}_{h,t}(x) \Rightarrow \frac{1}{t}(x-w) \in \partial h(w)$.

let $y = x_k \Rightarrow f(x_{k+1}) \le f(x_k) - \frac{t}{2}\|G_t(x_k)\|^2 < f(x_k)$.

note that $G_t(x_k) = 0 \Rightarrow x_k = \mathrm{prox}_{h,t}(x_k - t\nabla g(x_k)) \Rightarrow -\nabla g(x_k) \in \partial h(x_k)$.

$g(y) \ge g(x_k) + \nabla g(x_k)^T(y-x_k)$. $h(y) \ge h(x_k) - \nabla g(x_k)^T(y-x_k) \Rightarrow x_k = x^*$

let $y = x^* \Rightarrow f(x_{k+1}) - f(x^*) \le \frac{1}{2t}((1-mt)\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2)$

if $m=0 \Rightarrow \sum_{k=0}^{T-1} f(x_{k+1}) - f(x^*) \le \frac{1}{2t}\|x_0 - x^*\|^2$.

if $m>0 \Rightarrow \|x_{k+1} - x^*\|^2 \le (1-mt)\|x_k - x^*\|^2$.

Remark: if $h(x) \equiv 0$. it is exactly the same as gradient descent.

Require L-smooth. if L is unknown? Backtracking line search.

choose $0<\alpha, \beta<1$. and initial $\hat{t}>0$. often choose $\alpha = 1/2$.

Recall in analysis we need $g(y)-g(x)-\nabla g(x)^T(y-x) \le \frac{L}{2}\|x-y\|^2$

while $g(x_k - t_k G_{t_k}(x_k)) > g(x_k) - t\nabla g(x_k)^T G_{t_k}(x_k) + \frac{t_k}{2}\|G_{t_k}(x_k)\|^2$ $t_k = \beta t_k$

Convergence analysis for backtracking line search. note $t_k \ge \min\{\hat{t}, \beta/L\} \triangleq t_{min}$

$f(x_{k+1}) - f(x^*) \le \frac{1}{2t_k}\left((1-mt_k)\|x_k - x^*\|^2 - \|x_{k+1}-x^*\|^2\right)$.

$\Rightarrow \sum_{k=0}^{T-1} t_k(f(x_{k+1}) - f(x^*)) \le \frac{1}{2}(\|x_0 - x^*\|^2 - \|x_T - x^*\|^2)$

$\Rightarrow (\Sigma t_k)(f(x_T)-f(x^*)) \le \frac{1}{2}\|x_0 - x^*\|^2 \Rightarrow f(x_T) - f(x^*) \le \frac{\|x_0 - x^*\|^2}{2t_{min}T}$

Again. if m-strongly convex for $m>0$. $\|x_{k+1}-x^*\|^2 \le (1-mt_{min})\|x_k - x^*\|^2$.

Properties of proximal mapping: nonexpansive. (firmly).

Firmly nonexpansive: $(prox_h(x) - prox_h(y))^T(x-y) \ge \|prox_h(x) - prox_h(y)\|^2$

let $u = prox_h(x)$. $v = prox_h(y)$. then $x-u \in \partial h(u)$. $y-v \in \partial h(v)$.

$\Rightarrow h(v) \ge h(u) + (x-u)^T(v-u)$. and $h(u) \ge h(v) + (y-v)^T(u-v)$.

$\Rightarrow (x-u)^T(v-u) + (y-v)^T(u-v) \le 0 \Rightarrow (x-u-y+v)^T(u-v) \ge 0$.

Nonexpansive: $\|prox_h(x) - prox_h(y)\| \le \|x-y\|$ by Cauchy-Schwarz.

Equality constrained optimization

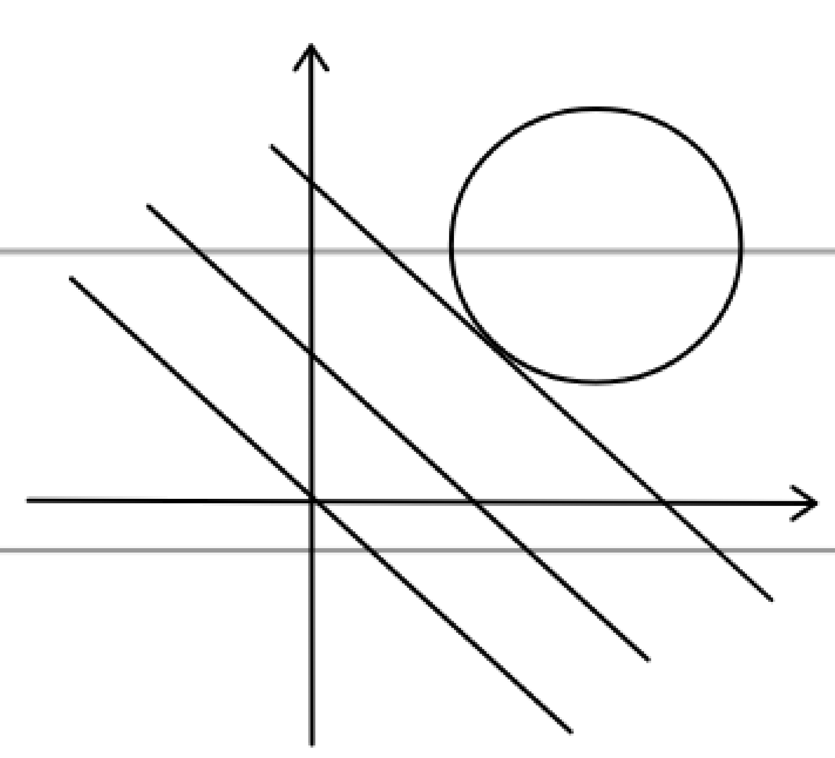
$f: \mathbb{R}^n \to \mathbb{R}$. $\min_x f(x)$ s.t. $g_i(x) = 0$. $g_i: \mathbb{R}^n \to \mathbb{R}$.

let $g: \mathbb{R}^n \to \mathbb{R}^m = (g_1(x), \cdots, g_m(x))$. $\min_x f(x)$ s.t. $g(x) = 0$.

The first question: how to verify optimality?

If unconstrained. $f(x^*)$ optimal $\Rightarrow \nabla f(x^*) = 0$. $\nabla^2 f(x^*) \geq 0$.

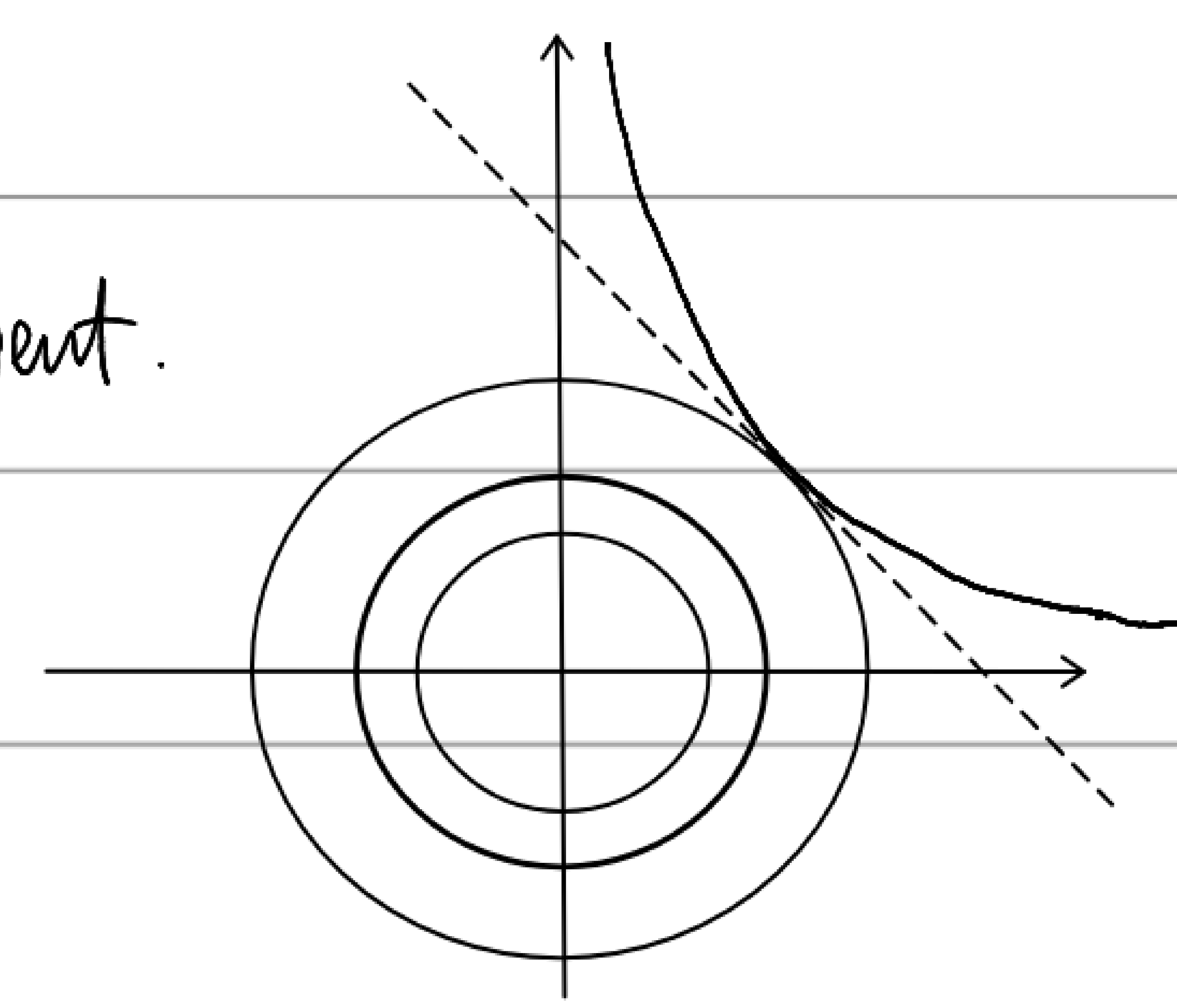
but how about constrained? $\uparrow$ $\Leftarrow \nabla f(x^*) = 0$. $\begin{cases} \nabla^2 f(x^*) > 0 \\ f \text{ convex} \end{cases}$.



$\min\ x+y$. s.t. $(x-2)^2 + (y-2)^2 = 1$

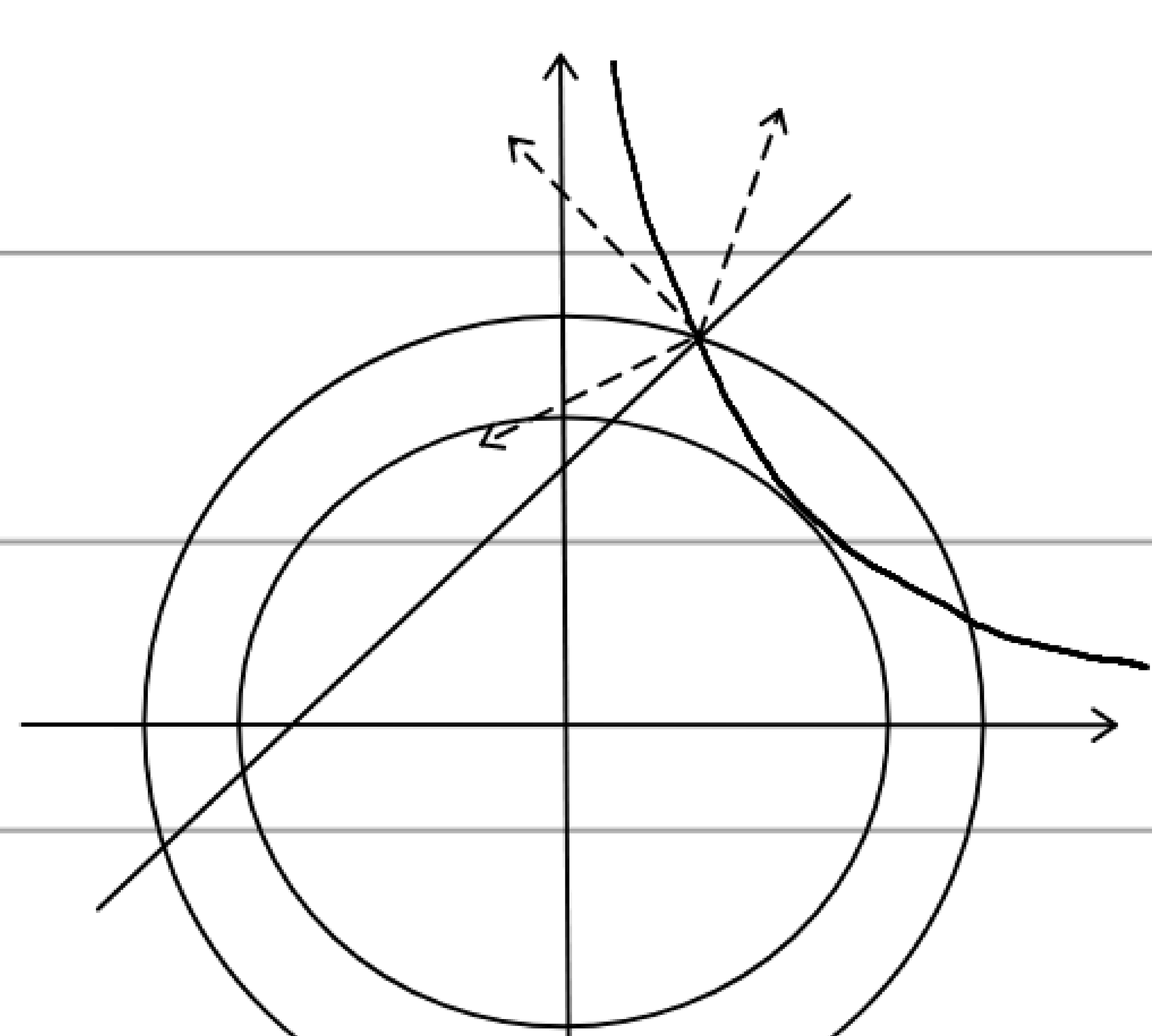
$f(x) = f(x^*)$
$g(x) = 0$

tangent.



$\min\ x^2 + y^2$. s.t. $xy = 1$.

$\cup$ tangent $f(x^*)$ and $g(x^*)$ at $x^* \Rightarrow \nabla f(x^*) = \lambda \nabla g(x^*)$.



if > 1 constraints $\begin{cases} xy = 1 \\ y - x = 2 \end{cases}$

no longer tangent. but $\nabla f(x^*)$ is

linear combination of $\nabla g_1(x^*)$. $\nabla g_2(x^*)$.

$\Rightarrow \nabla f(x^*) - \lambda_1 \nabla g_1(x^*) - \lambda_2 \nabla g_2(x^*) = 0$.

Lagrange multiplier method: let $\lambda_1, \dots, \lambda_m$ be multiplier.

define Lagrange function $L(x,\lambda) = f(x) - \sum_{i=1}^{m} \lambda_i g_i(x)$.

$\min_x f(x) = \min_{x,\lambda} L(x,\lambda)$. (or $\exists \lambda$, $\nabla L(x^*, \lambda) = 0$). ?

The answer is no! consider $\min x$. s.t. $g(x) = \begin{cases} x^2 & x<0 \\ 0 & x\in[0,1] \\ (x-1)^2 & x>1 \end{cases} = 0$.

or. $\min (x+1)^2 + (y+1)^2$. s.t. $g(x,y) = (x^2+y^2)^2 - 2x(x^2+y^2) + 3y^2 = 0$.

$(x^*, y^*) = (0,0)$. $\nabla f \neq \lambda \nabla g$. or $\begin{cases} g_1 = g + x - y \\ g_2 = x - y \end{cases}$ $\nabla f \notin \text{span}\{\nabla g_1, \nabla g_2\}$.

$Dg \neq 0$.

Now we see a good example: g_i is linear function.

$\min_x f(x)$. s.t. $g(x) = Ax + b = 0$. $x \in \mathbb{R}^n$. $g \in \mathbb{R}^{m\times n}$. $m<n$. ind. $\text{rank}(A) = m$.

Then $\{x : g(x) = 0\}$ is an affine set $= \mathbb{R}^{n-m} + x_0 \triangleq G$

suppose x^* is the optimal point. $\Rightarrow \nabla f(x^*)^T (x - x^*) \geq 0$. $\forall x \in G$.

$G - x^* = \ker(A)$. since $A(x - x^*) = 0$, $\forall x \in G$. $\dim \ker(A) = n-m$.

note that $x^* + v \in G \Rightarrow x^* - v \in G$. so $\nabla f(x^*)^T v = 0$.

$\Rightarrow \nabla f(x^*) \perp \ker(A)$. $\Rightarrow \nabla f(x^*) \in \text{span}\{a_1, a_2, \dots, a_m\}$

$= \text{span}\{\nabla g_1, \nabla g_2, \dots, \nabla g_m\}$.

How about general g?

Hope $G = \{x : g(x) = 0\}$ is an affine space. f. g. differentiable on G.

In fact. we do not care about the shape of G. but need local properties.

Manifold: any point has a neighbourhood the "same" as Eculidean space.

manifold: 流形 (江泽涵译). 云行雨施. 品物流形. 易经. 天地有正气. 杂然赋流形. 文天祥.

Trivially $\mathbb{R}^n$ is a manifold. S^2 is also a manifold. 经纬度.

We now give formal definition. (note that we need differentiable manifolds).

Homeomorphism 同胚: $\exists f: \Omega_1 \to \Omega_2$. f invertible. f, f^{-1} continuous.

Diffeomorphism 微分同胚. f, f^{-1} smooth. $\in C^\infty$

smooth submanifold of $\mathbb{R}^n$: parameterize curve $\gamma: (-\varepsilon, \varepsilon) \to \mathbb{R}^n$. $\gamma'(t) \neq 0$.

neighbourhood of $\gamma(0)$ is similar to a line (tagent line).

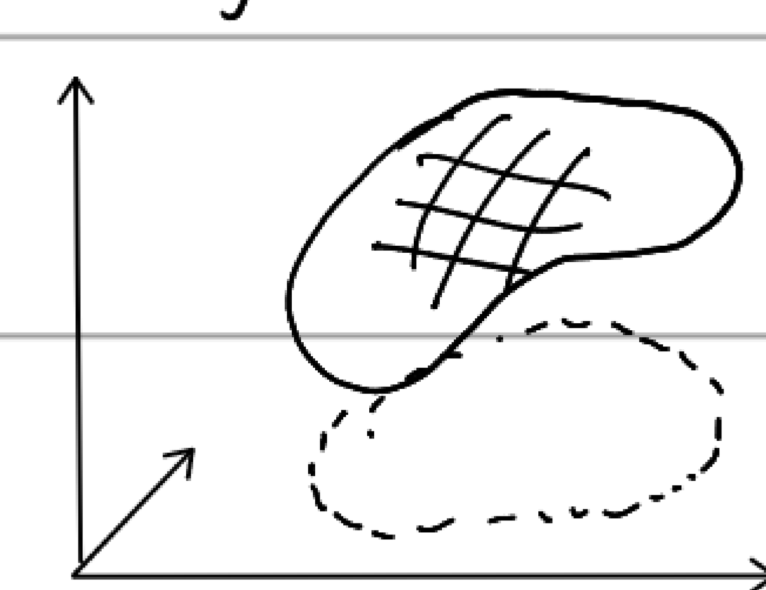
$\Phi: \mathbb{R}^n \to \mathbb{R}^n$. $(x_1, \dots x_n) \to \gamma^{-1}(x_1, \dots, x_n) \cdot \gamma'(0)$. $\Phi^{-1} = \gamma(\frac{x}{\gamma'(0)})$.

Next, consider the graph of smooth function f: $\{(x, f(x)) \in \mathbb{R}^{n+1}\}$.

$\Phi: \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ $(x_1, \dots x_{n+1}) \to (x_1, \dots, x_n, x_{n+1} - f(x_1, \dots, x_n))$.

Now we consider $S^1 = \{(x,y): x^2+y^2=1\}$.

partition it into images of smooth functions.



d-dimensional differentiable manifold: $M \neq \Phi \subseteq \mathbb{R}^n$ (with coordinates $\{x_i\}$)

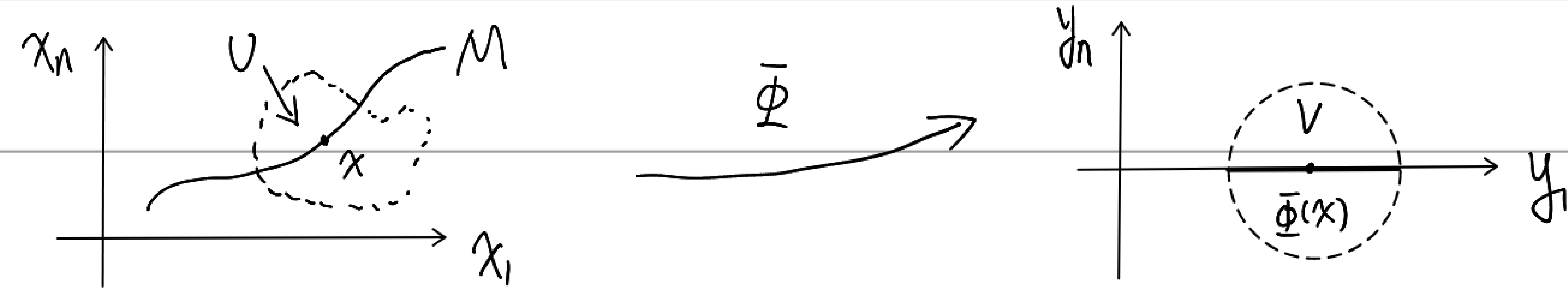
if $\exists d \in \mathbb{N}$. s.t. $\forall x \in M$. $\exists x \in U \subseteq \mathbb{R}^n$. $V \subseteq \mathbb{R}^n$ (with coordinates $\{y_i\}$).

and diffeomorphism $\Phi: U \to V$. s.t. $\Phi(U \cap M) = V \cap (\mathbb{R}^d \times \{0\}^{n-d})$.

Then M is a submanifold with $\dim M = d$. and $\operatorname{codim} M = n-d$.

In fact. $\exists$ $\operatorname{codim} M$ functions s.t. M is the set of common zero points.

Let $f_j : \mathbb{R}^n \to \mathbb{R}$ $(x_1, \dots, x_n) \to y_j(\Phi(x_1, \dots, x_n))$. $j = d+1, \dots, n$.



We believe differentiable manifold is good enough to consider $\nabla f. \nabla g$.

The question is : whether $\{x : g(x) = 0\}$ is a submanifold ?

Note that graph of continuous differentiable functions are submanifolds.

if $g(x,y) = 0 \Rightarrow y = h(x)$. then $\{(x,y) : g(x,y) = 0\}$ is.

so we need the implicit function theorem.

Another question : differential on which linear space ?