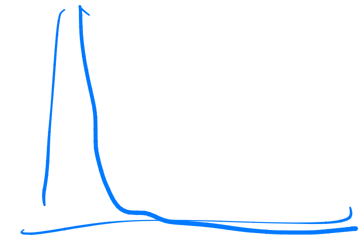
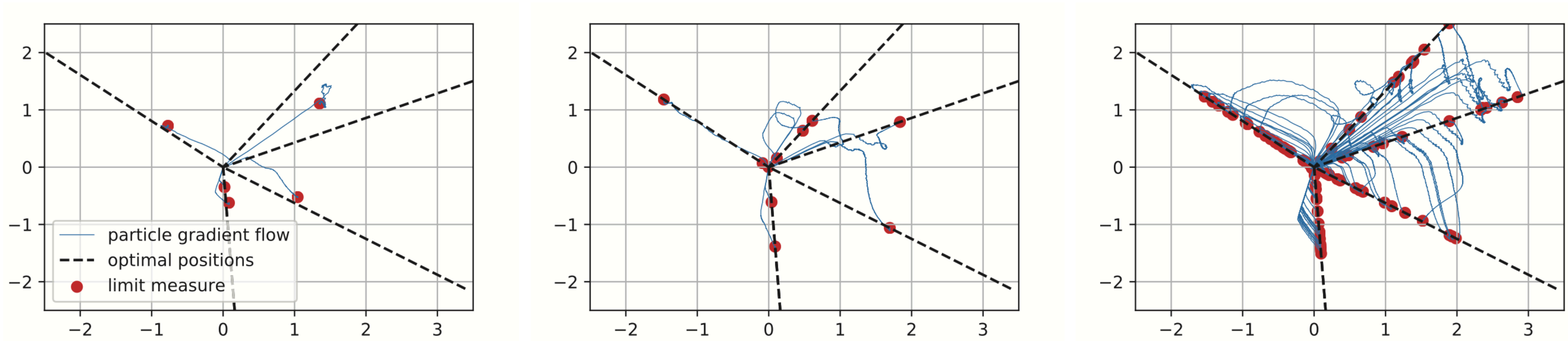




5 neurons 10 neurons 100 neurons

$$\mathcal{X} = \mathbb{R}^2 \qquad f_\sigma = \sum_{j=1}^{5} a_j \, (w_j^T x)_+$$

$$\underline{\eta_j w_j}$$

Updated note!

Linear model: $f_\theta(x) = \langle \theta, \varphi(x) \rangle$
↳ $\varphi : X \to$ Hilbert space

$X = \mathbb{R}^d$, f^* has t derivatives

then excess risk was $O(n^{-t/2s})$ $s > \frac{d}{2}$

regularity of model

adaptivity to smoothness

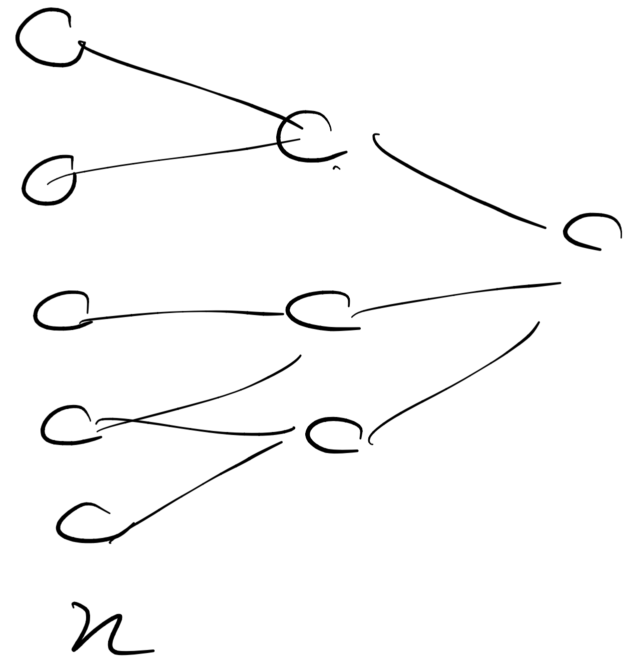
⇒ adaptivity to linear latent variable

for kernel

One hidden layer NN on $\mathbb{R}^d$

$$f_\Theta(x) = \sum_{j=1}^{m} \eta_j \, \sigma(w_j^T x + b_j)$$

output weight   input weights



σ = activation function

$\sigma(u) = (u)_+ = \max(u, 0)$ RELU

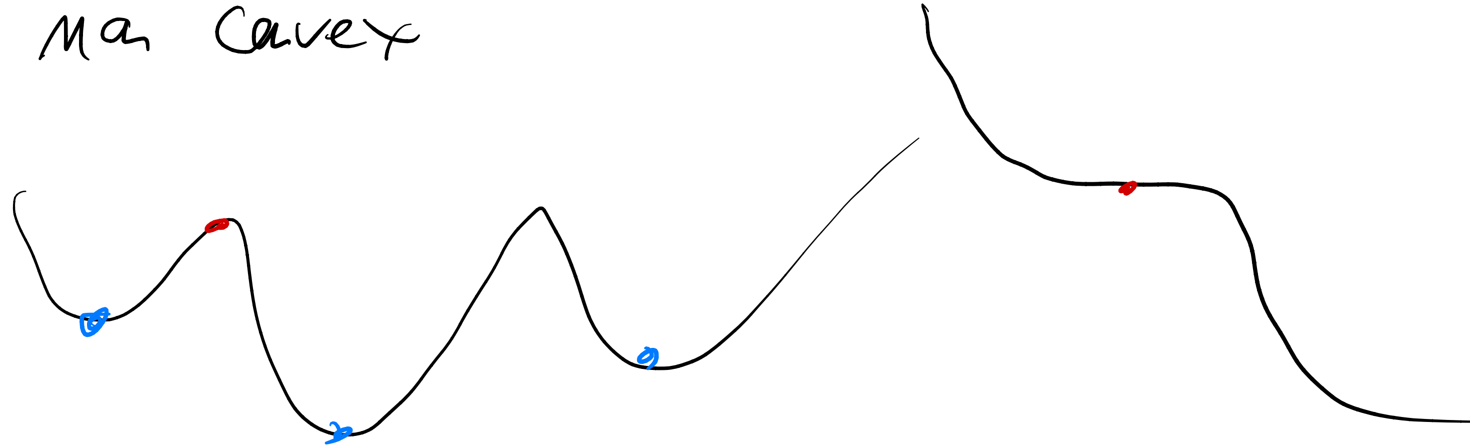
positively homogeneous

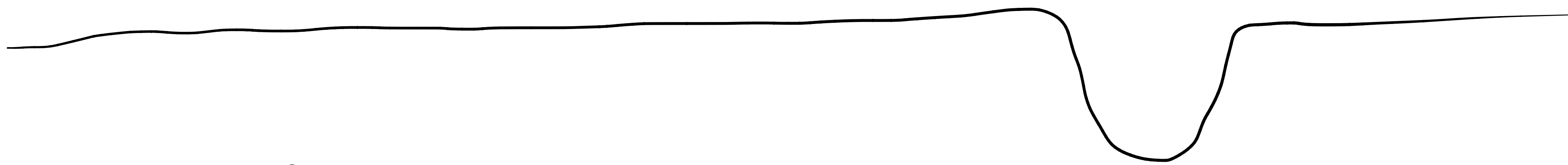
$(\lambda u)_+ = \lambda u_+$ if $\lambda > 0$

$\Theta = \{ \eta_j, w_j, b_j \} \in \mathbb{R}^{m(d+2)}$

f_Θ : Non linear in Θ.

optimisation : non convex





prediction function

$$f(x) = \frac{1}{m}\sum_{j=1}^{m} \underbrace{a_j\,\sigma(w_j^\top x + b_j)}_{\phi(\theta_j)} = \frac{1}{m}\sum_{j=1}^{m} \phi(\theta_j)$$

$\theta_j = (a_j, w_j, b_j)$

$$= \int \phi(\theta)\, d\mu(\theta)$$

proba measure

(goal $\min R(f)$ convex)

(Lénaïc Chizat, FB, 2018)

$R(\theta) = J(\mu) = R\left(\int \phi(\theta)\, d\mu(\theta)\right).$

in practice $d\mu(\theta) = \frac{1}{m} \sum_{j=1}^{m} \delta_{\theta_j}$

GD (backpropagation) $\Rightarrow$ $\dot{\theta} = \nabla R(\theta)$

gradient flow

Result (C & B) : when $m \to +\infty$

GF $\Leftarrow$ flow on measures

"Wasserstein grad. flow"

to minimize $J(\mu)$

satisfied a PDE

$\partial_t \mu = \boxed{}$

(transport)

Estimation error

$$f(x) = \sum_{j=1}^{m} \underset{d_j}{\underline{a_j}} \left(\alpha_j (w_j^T x + b_j) \right)_+$$

penalize $\sum_j a_j^2 + \|w_j\|_2^2 + b_j^2$

"weight decay"

$d_j > c$,

Uniform deviation

Exp $|\hat{R}(G_\theta) - R(G_\theta)|$ penalty $\sum_j \frac{a_j^2}{d_j^2} + d_j^2 (\|w_j\|^2 + b_j^2)$

$\|u\|_2 \le D$ radius of data

$\le 16 \frac{D\, RG}{\sqrt{m}}$ Lipschitz constant of G_θ

opt $d_j \to 2 \sum_j |a_j| \underbrace{\sqrt{\|w_j\|_2^2 + b_j^2}}$

normalize to unit norm

Rademacher averages

$\to \boxed{\text{penalty} = 2\|a\|_1}$

Approximation error $=$ $f_*: \mathbb{R}^d \to \mathbb{R}$

estimated by $f(x) = \sum_{j=1}^{m} a_j (w_j^T x + b_j)_+$

$\|a\|_1$ $\qquad$ (w_j, b_j) bounded

$\|w_j\| = 1$

$|b_j| \le 1$

- trade off between

$$\|f - f_*\|^2_{L^2(\rho)} \quad \text{and} \quad \|a\|_1$$

- take "$m = +\infty$" $\qquad$ $f(x) = \int (w^T x + b)_+ \, dq(w,b)$

signed measure $\neq$ proba measure

$$dq = \sum_{j=1}^{m} a_j \delta_{w_j, b_j}$$

$$\int |dq(w,b)| = \|a\|_1$$

↳ total variation

$$\|f\| = \inf \int |dq(w,b)|$$

variation norm

$$\|f\| = \inf_{\eta} \int |d\eta(w,b)| \quad \text{such that}$$

variation norm

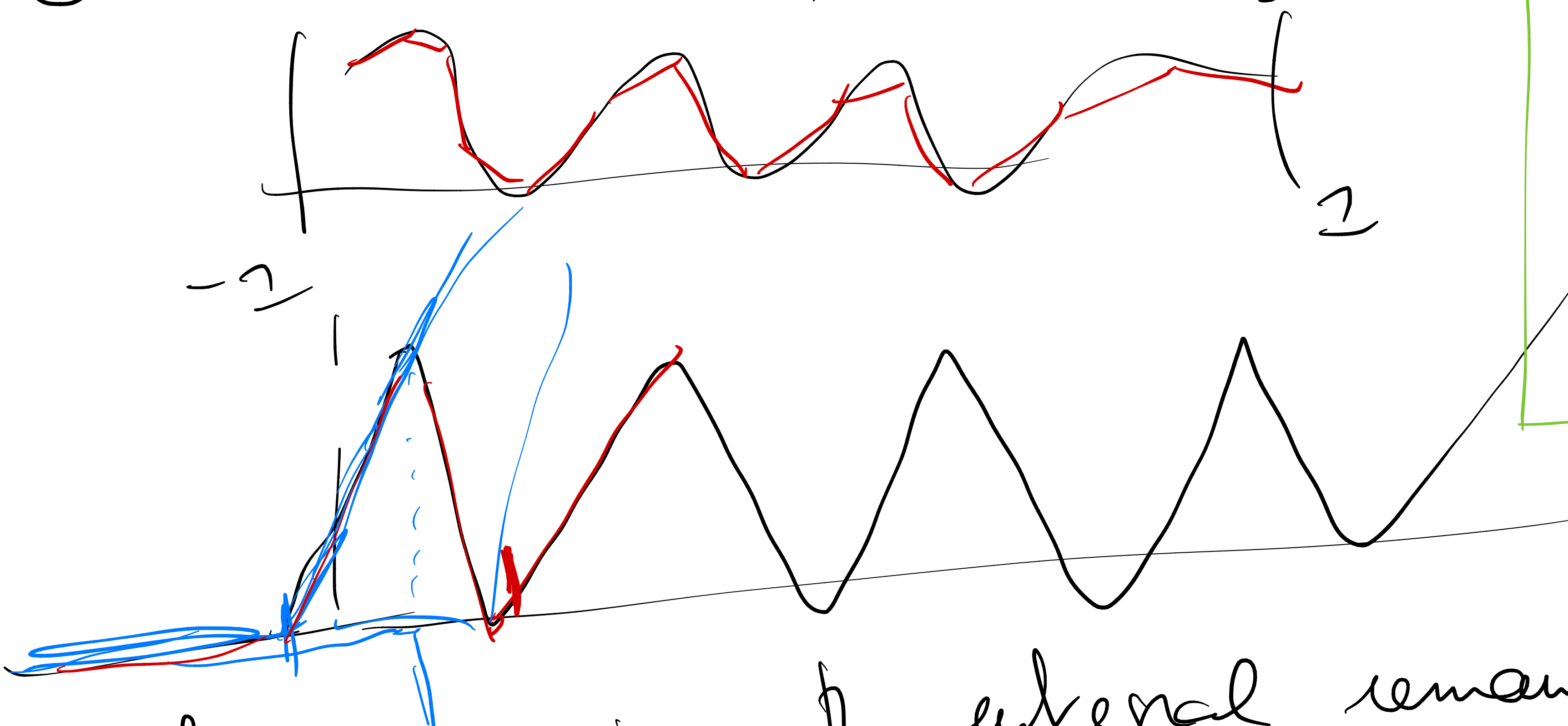
$$f(x) = \int (w^T x + b)_+ \, d\eta(w,b).$$

Key property: ① if $f(x) = g(a^T x)$, $g : \mathbb{R} \to \mathbb{R}$, $a \in \mathbb{R}^d$, $\|a\|_2 = 1$

$$\|f\| \leq \|g\|$$

② if $f(x) = (w^T x + b)_+$ then $\|f\| \leq 1$

1 one dimension = approximation by ReLU



Taylor expansion with integral remainder $f \in C^2$

$$f(x) = f(-1) + f'(-1)(x+1) + \int_{-1}^{x} (x-t)\, f''(t)\, dt$$

$$\underline{\qquad\qquad\qquad\qquad} + \int_{-1}^{1} (x-t)_+ \underbrace{f''(t)\,dt}_{d\mu(1,-t)}$$

$$\Rightarrow \|f\| \sim \int |f''(x)|\, dx + \dots$$

Kernel method: $\int f(x)^2 + \int f'(x)^2$

NN $\int |f(x)| + \int |f^k(x)| \leq$ Sobolev norm

on $(-1,1)$

from the avg dim.

$$f(x) = \frac{1}{(2\pi)^d} \int \hat{f}(\omega) e^{i\omega^T x} \, d\omega$$

$$\|f\| \leq \frac{1}{(2\pi)^d} \int |\hat{f}(\omega)| \underbrace{\| x \mapsto e^{i\omega^T x} \|}_{(1+\|\omega\|^2)}$$

$$\leq \int |\hat{f}(\omega)| (1+\|\omega\|_2^2)$$ Barron norm

$\leq$ Sobolev norm of order $\frac{d+3}{2}$

if $f_*(x) = g(a^T x)$ $\quad g$ t-times differentiable

, then f_* is t-differentiable

linear latent variable

$\Rightarrow$ rate of $O(-n^{t/\frac{d+3}{2}})$

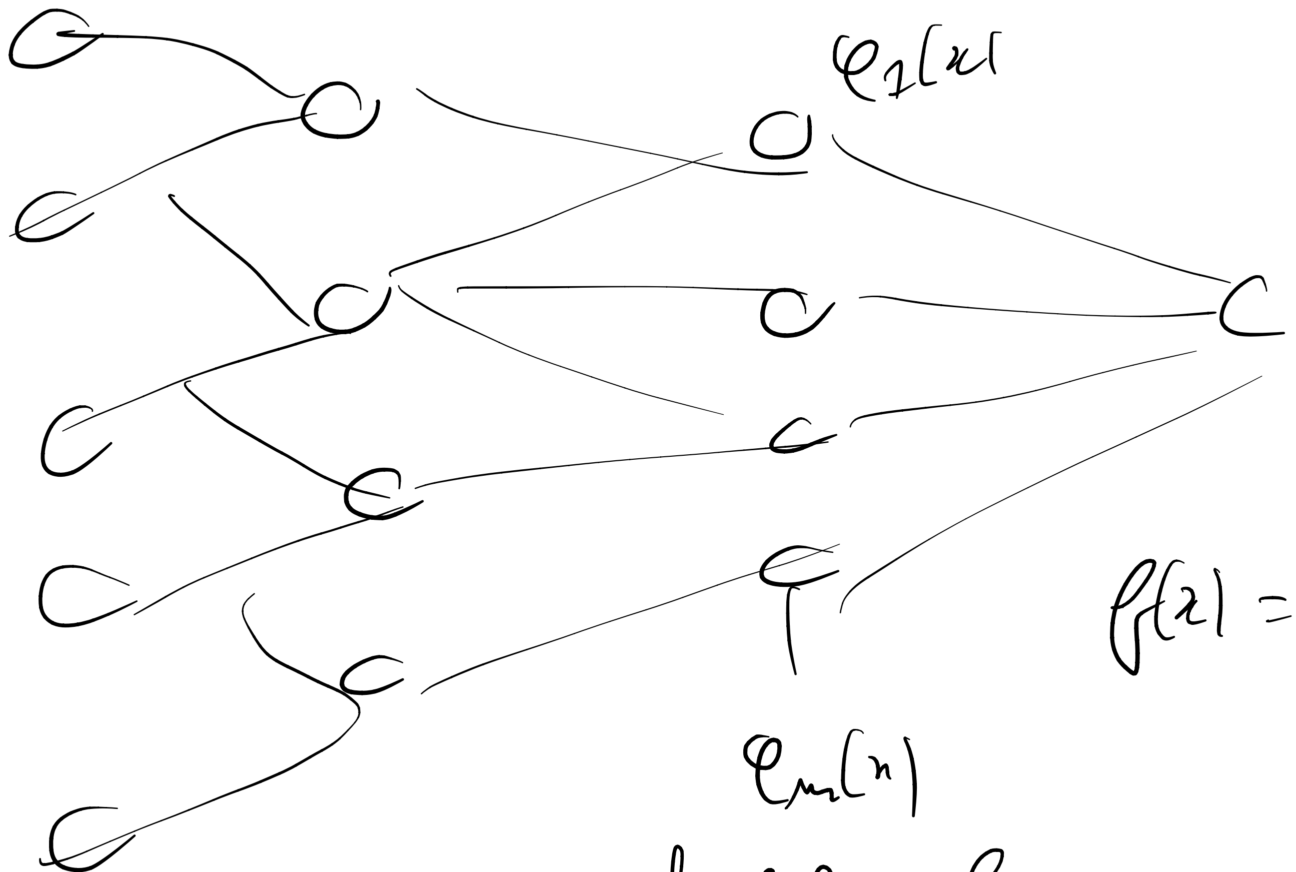
, but $\|f_*\| \le \|g\|$ for variation norm

$$A(\lambda) = \inf_f \|f - f_*\|^2 + \lambda \|f\|^2$$

$\boxed{\begin{array}{c} \|w_j\|_2 \\ \|w_j\|_2 \end{array}}$

$$\le \inf_g \|g - g_*\|^2 + \lambda \|g\|^2 =$$

rate $\Rightarrow$ $O(-n^{t/2})$ instead of $O(n^{-t/\frac{d+3}{2}})$



$$f(x) = \sum_j a_j \varphi_j(x)$$

NN with one hidden layers

optimization
estimation
approx.

local ave.
k-NN
adaptivity
to low-dim inputs

$\longrightarrow$ kernel methods

adaptivity
to
smoothness

$\longrightarrow$ neural nets

adaptivity
to
features