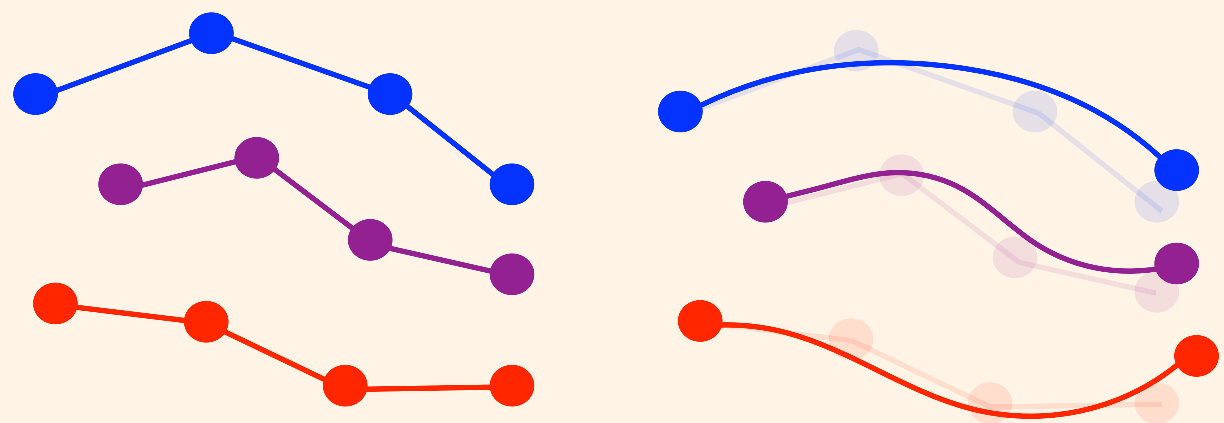
On the Training of Infinitely Deep and Wide ResNets

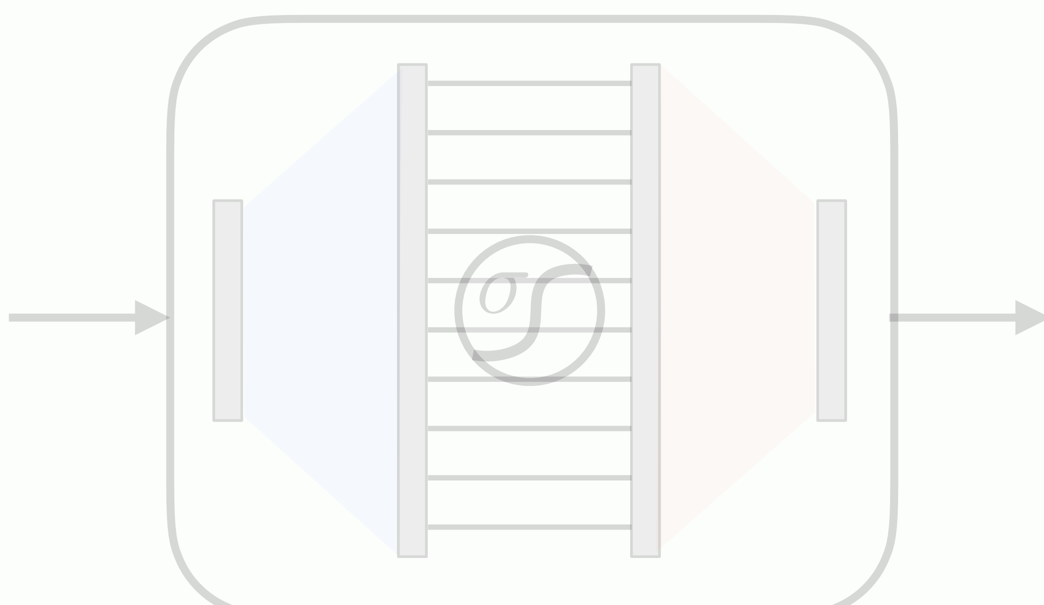
Gabriel Peyré

ResNet and Neural-ODEs

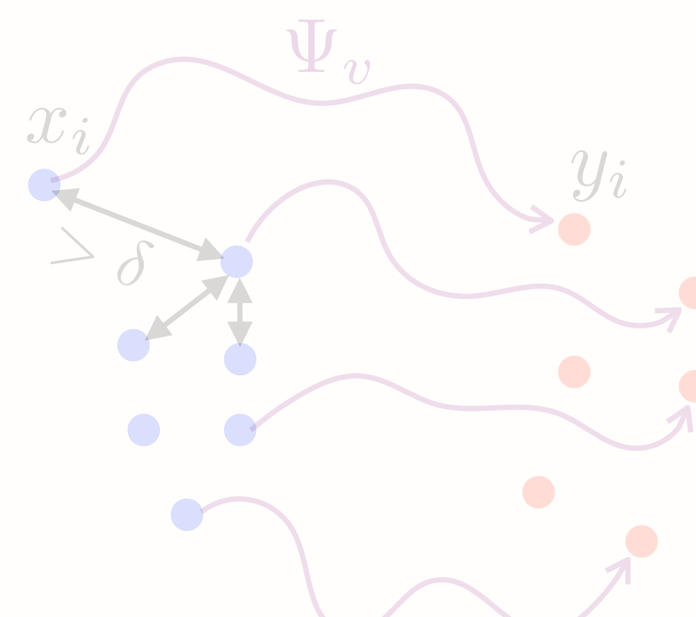


Global and local Polyak-Łojasiewicz conditions

RKHS Neural-ODEs

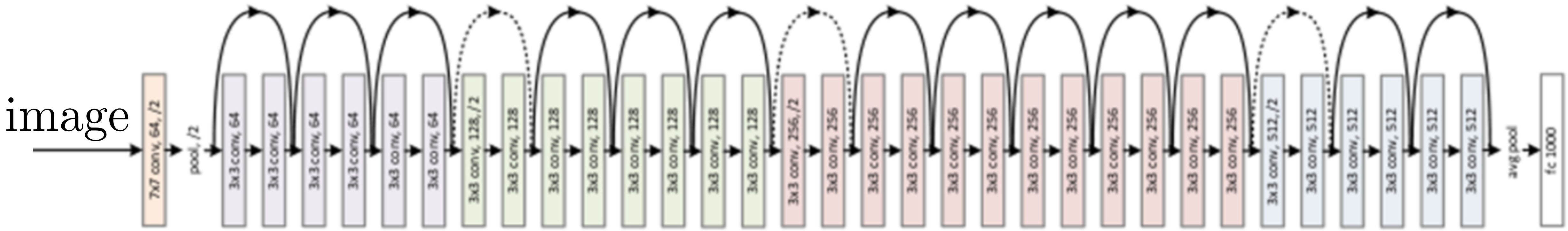


P-Ł condition for Neural-ODEs

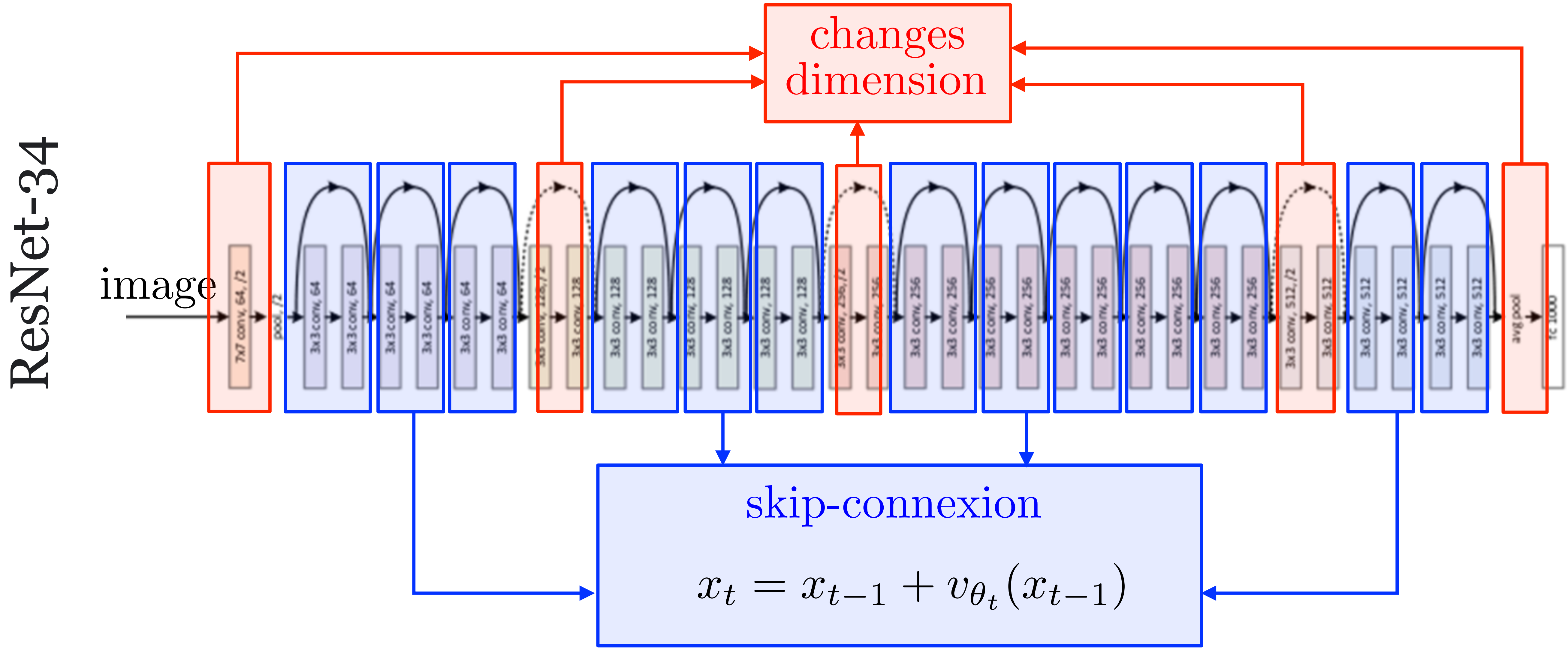


ResNet-type Architectures [He et al’ 16]

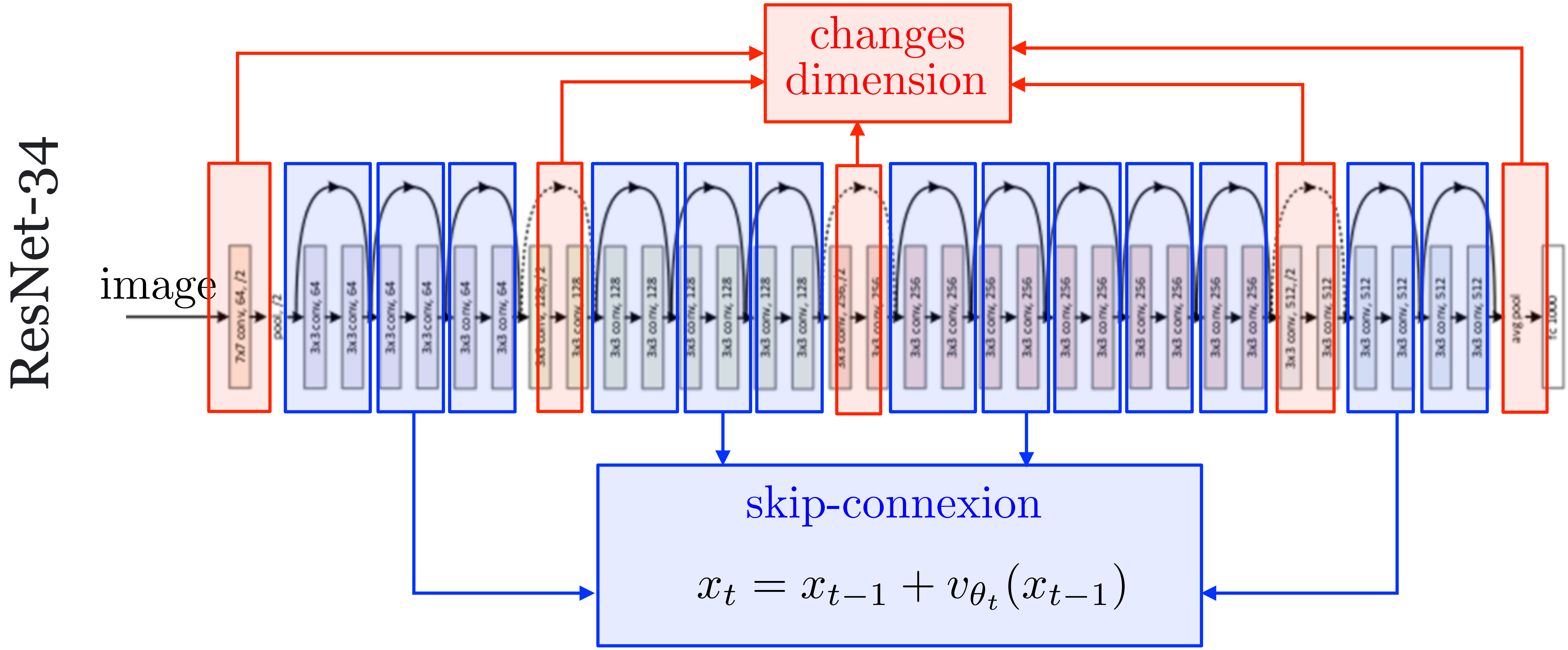
ResNet-34



ResNet-type Architectures [He et al' 16]

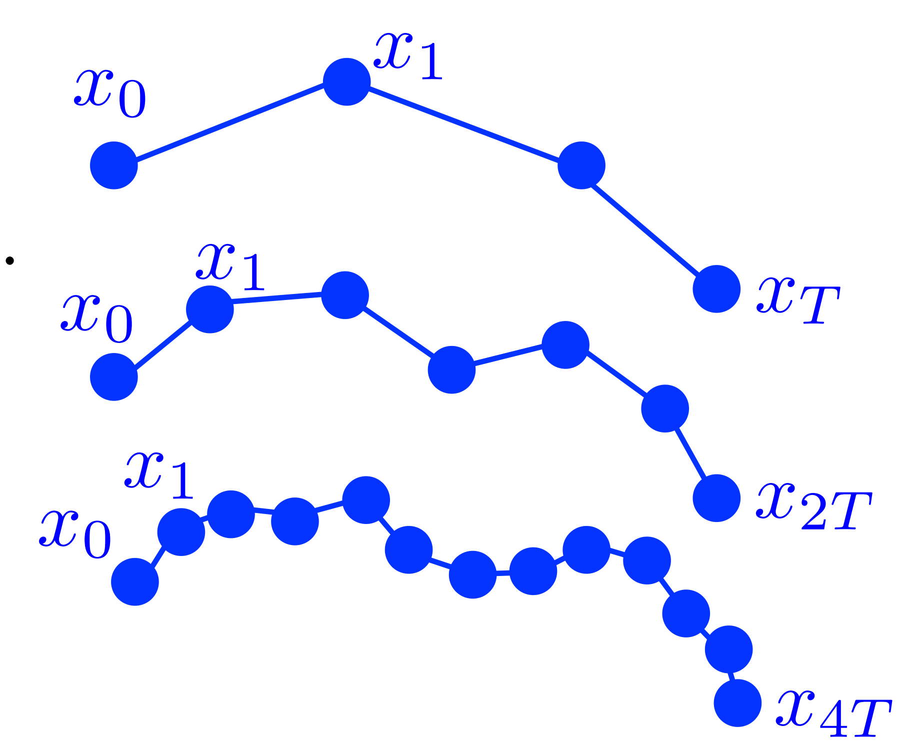


ResNet-type Architectures [He et al' 16]



$\rightarrow$ Makes the "infinite depth" limit non-degenerate.

$\rightarrow$ Enable $v_\theta = 0$ initialization, i.e. identity map.

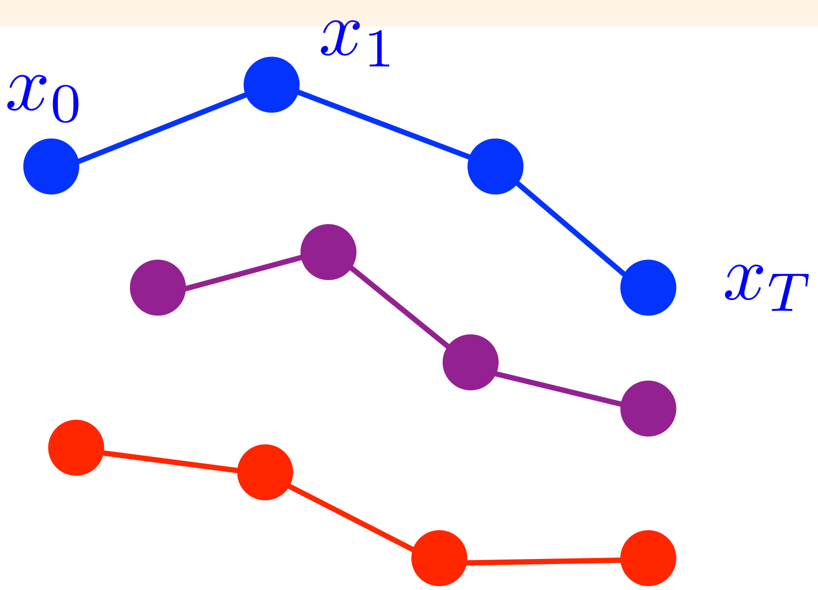


Infinite Depth and Neural-ODEs

ResNet [He et al, 2016]

$$\Phi_\theta(x_0) \triangleq x_T \quad \text{where}$$

$$x_{t+1} = x_t + \frac{1}{T} v_{\theta_t}(x_t)$$



Infinite Depth and Neural-ODEs

ResNet [He et al, 2016]

$\Phi_\theta(x_0) \triangleq x_T$ where

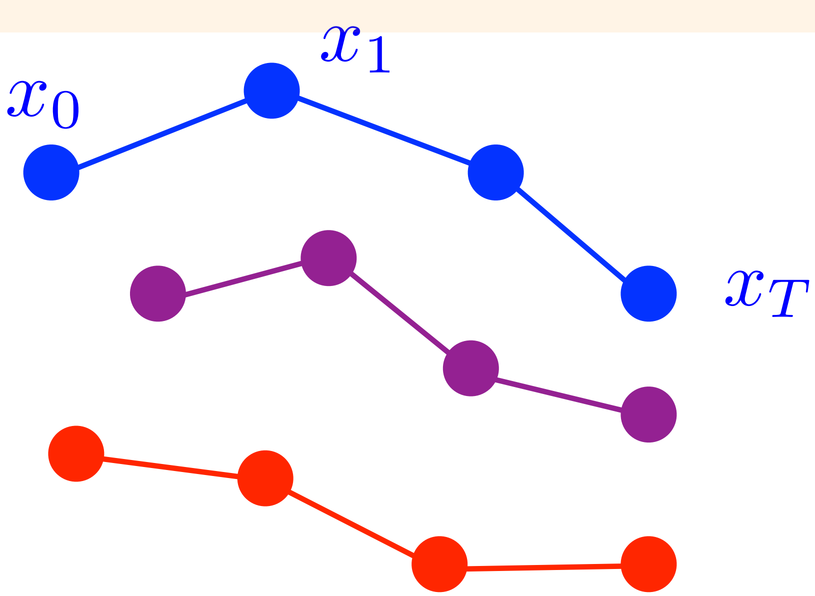
$$x_{t+1} = x_t + \frac{1}{T} v_{\theta_t}(x_t)$$

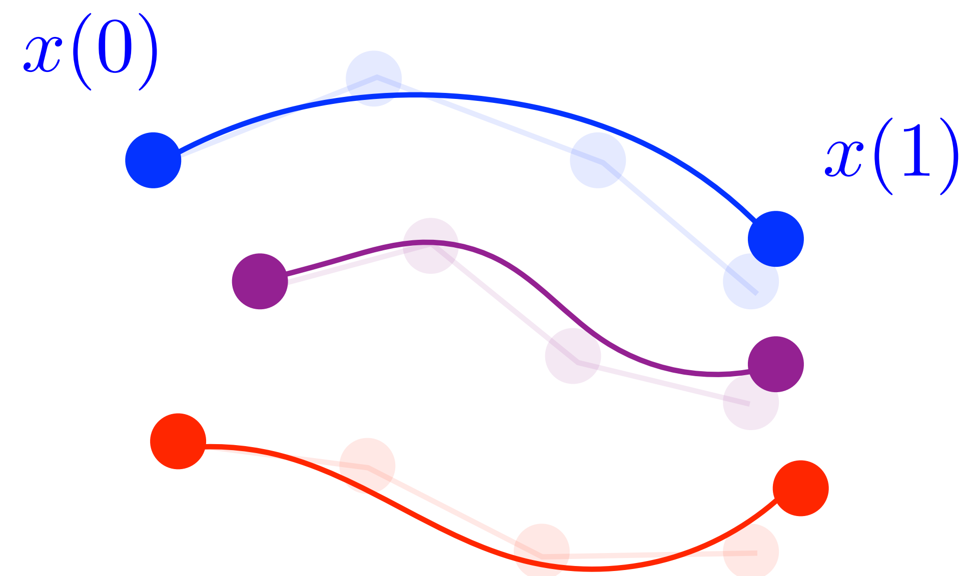
$T \to +\infty$

Neural ODE [Chen et al, 2018]

$\Phi_\theta(x(0)) \triangleq x(1)$ where

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = v_{\theta(t)}(x(t))$$





Infinite Depth and Neural-ODEs

ResNet [He et al, 2016]

$$\Phi_\theta(x_0) \triangleq x_T \quad \text{where}$$

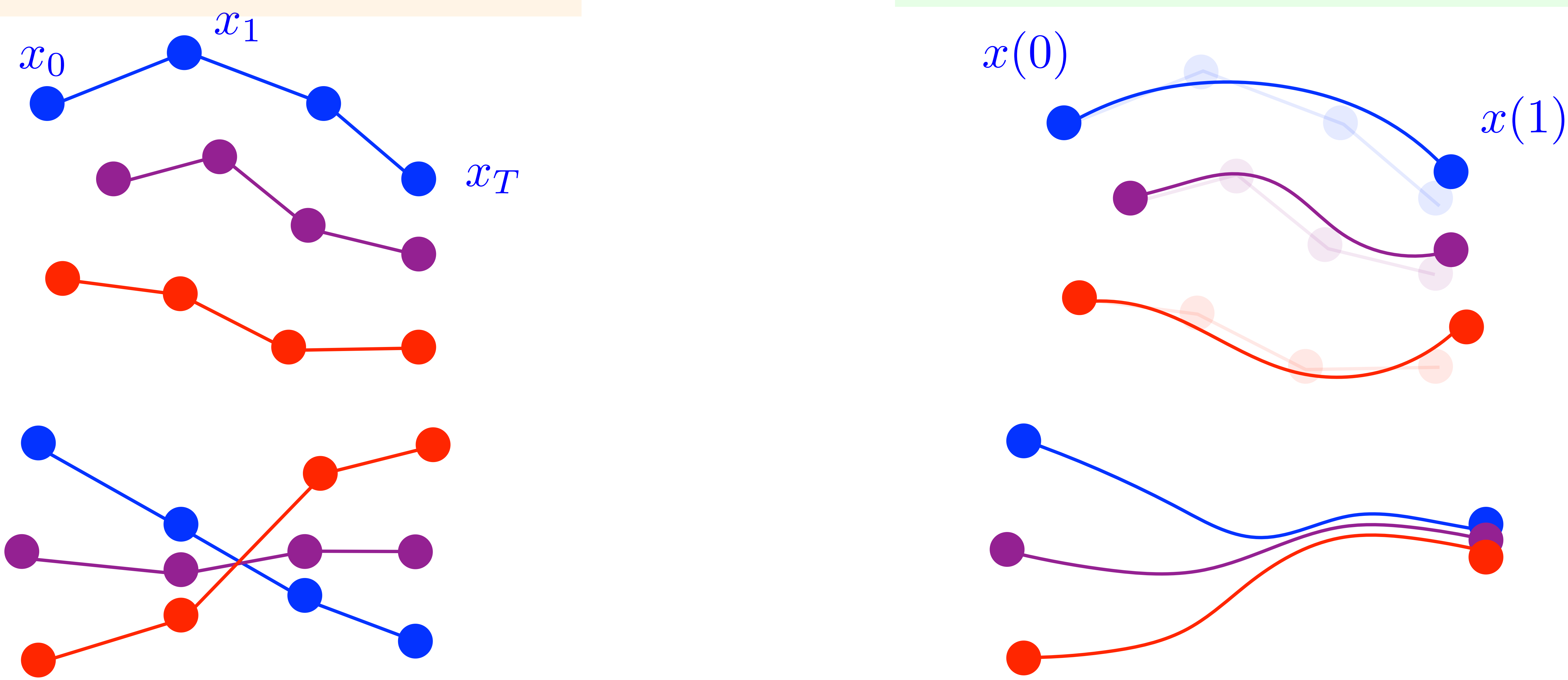
$$x_{t+1} = x_t + \frac{1}{T} v_{\theta_t}(x_t)$$

$T \to +\infty$

Neural ODE [Chen et al, 2018]

$$\Phi_\theta(x(0)) \triangleq x(1) \quad \text{where}$$

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = v_{\theta(t)}(x(t))$$



Trajectories cannot cross: Φ_θ defines a diffeomorphism.

$T \to +\infty$ is a singular limit (θ can "explodes" during training)

On the importance of scale and initialization

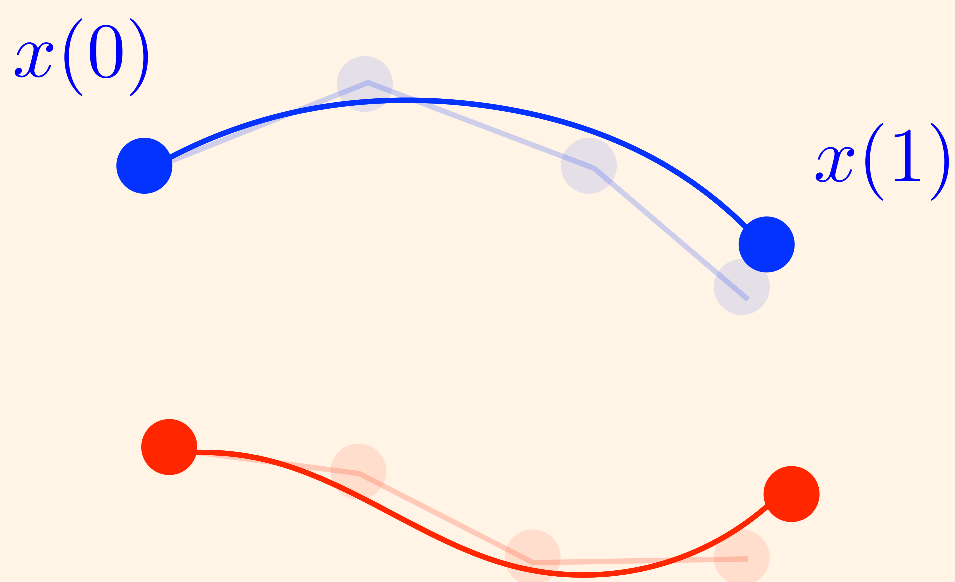
$$x_{t+1} = x_t + \frac{1}{T} v_{\theta_t}(x_t)$$

Zero/smooth initialization of $(\theta_t)_t$

$T \to +\infty$

Deterministic ODE

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = v_{\theta(t)}(x(t))$$



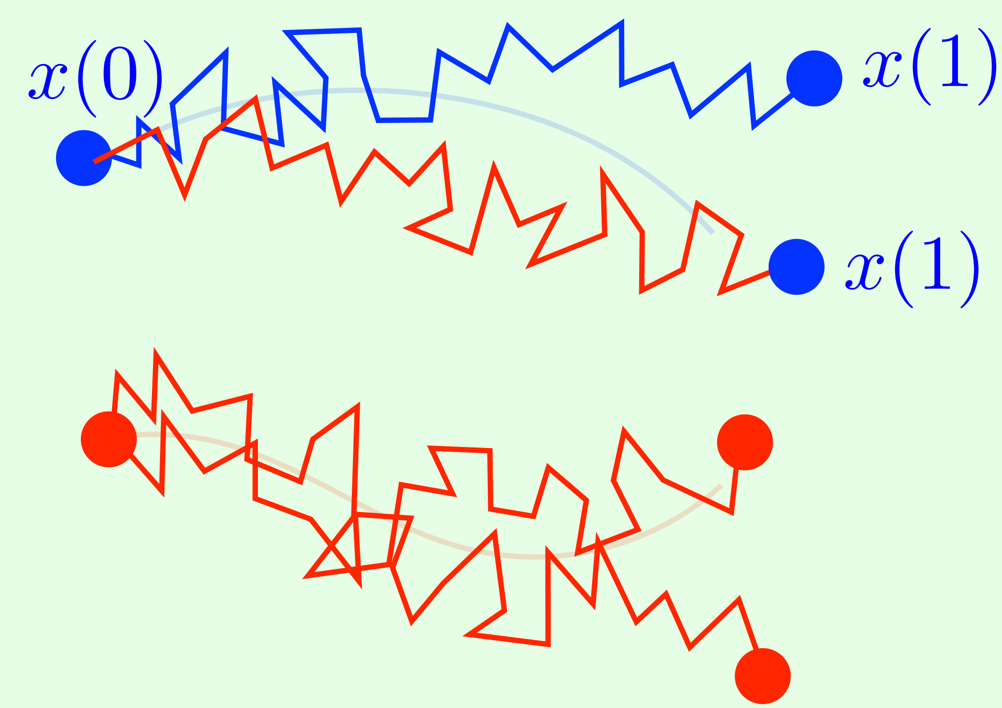
$$x_{t+1} = x_t + \frac{1}{\sqrt{T}} v_{\theta_t}(x_t)$$

Random initialization of $(\theta_t)_t$

$T \to +\infty$

Stochastic ODE

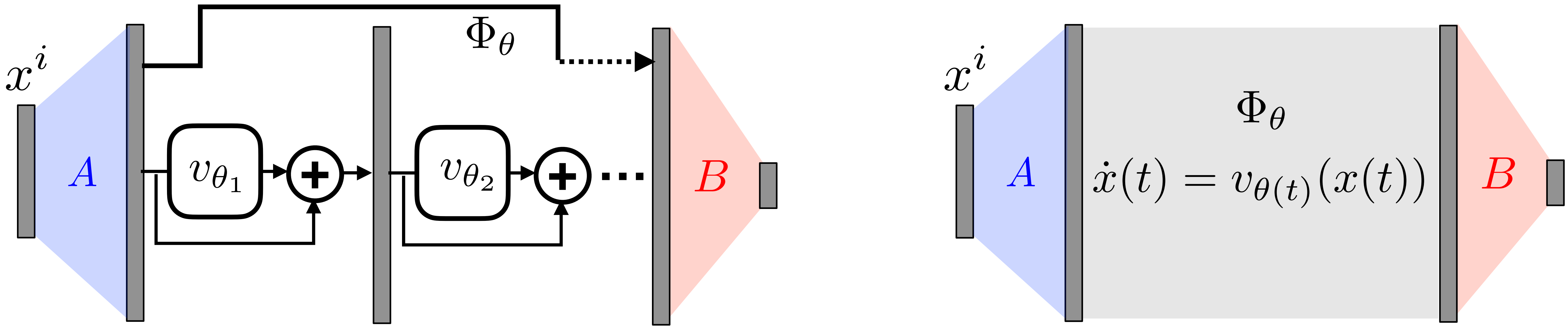
$$\mathrm{d}x(t) = v_{\theta(t)}(x(t))\mathrm{d}t + \mathrm{dW}(t)$$



[R. Cont, A. Rossier, R. Xu, 2022]

[P. Marion, Fermanian, Biau, Vert, 2022]

Training Dynamic

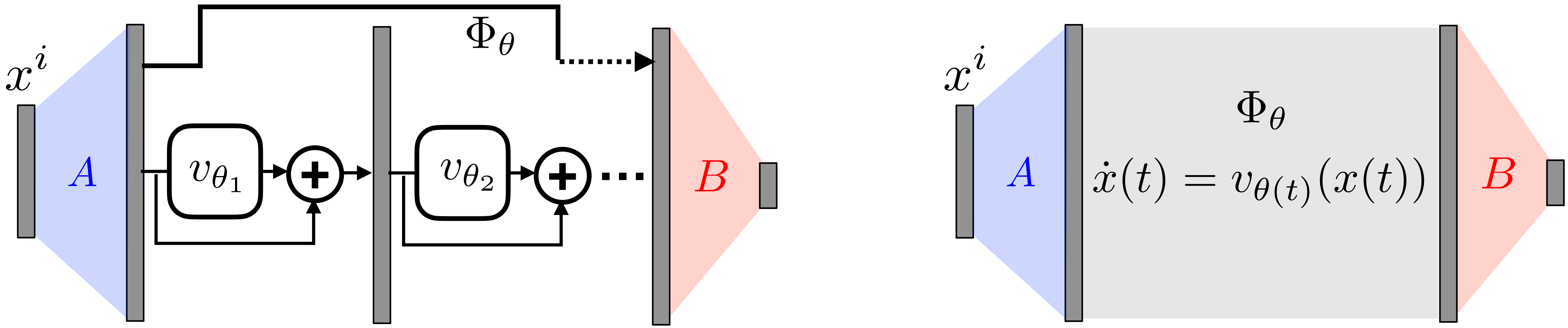


Training: $\min_{\theta} f(\theta) \triangleq \frac{1}{N} \sum_{i=1}^{N} \| B \Phi_\theta(A x^i) - y^i \|^2$

Gradient descent: $\theta^{(k+1)} = \theta^{(k)} - \tau \nabla f(\theta^{(k)})$

$\rightarrow$ **No explicit regularization!**

Training Dynamic



Training: $\min_\theta f(\theta) \triangleq \frac{1}{N} \sum_{i=1}^N \| B \Phi_\theta(A x^i) - y^i \|^2$

Gradient descent: $\theta^{(k+1)} = \theta^{(k)} - \tau \nabla f(\theta^{(k)})$

$\rightarrow$ **No explicit regularization!**

Question: convergence of θ^k toward global minimum?

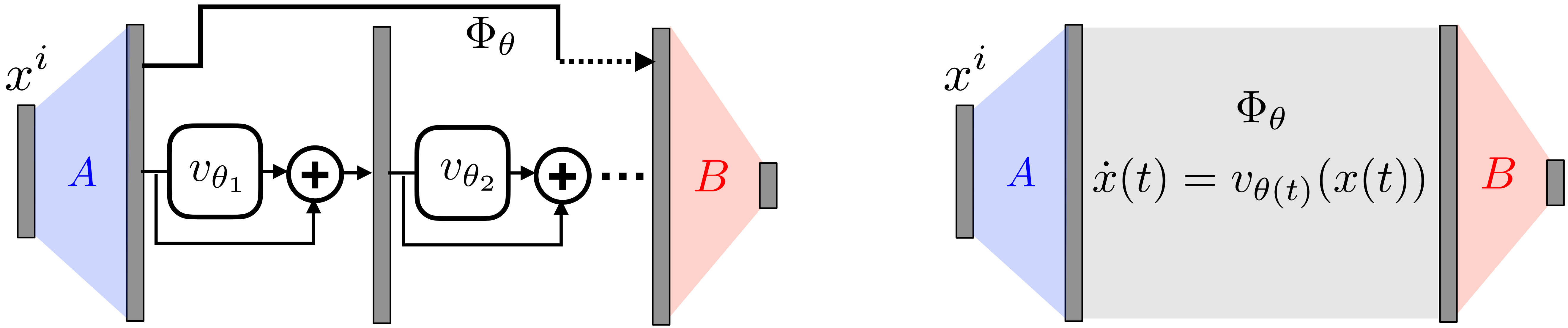
Neural tangent kernel [Jacot et al'18]: local linear expansion.

Polyak-Łojasiewicz inequality [Liu, Zhu, Belkin 2021]:

$\rightarrow$ conditionning might explodes as $T \rightarrow +\infty$.

$\rightarrow$ find a suitable limit model and show "implicit" regularization effect.

Training Dynamic



Training: $\min_\theta f(\theta) \triangleq \frac{1}{N} \sum_{i=1}^N \| B \Phi_\theta (A x^i) - y^i \|^2$

Gradient descent: $\theta^{(k+1)} = \theta^{(k)} - \tau \nabla f(\theta^{(k)})$

$\rightarrow$ **No explicit regularization!**

Question: convergence of θ^k toward global minimum?

Neural tangent kernel [Jacot et al'18]: local linear expansion.

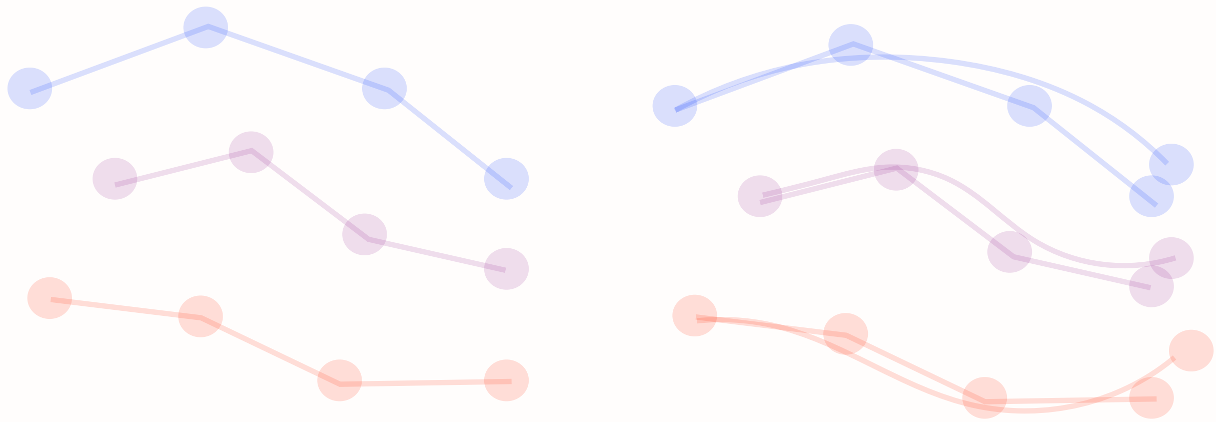
Polyak-Łojasiewicz inequality [Liu, Zhu, Belkin 2021]:

$\rightarrow$ conditionning might explodes as $T \rightarrow +\infty$.

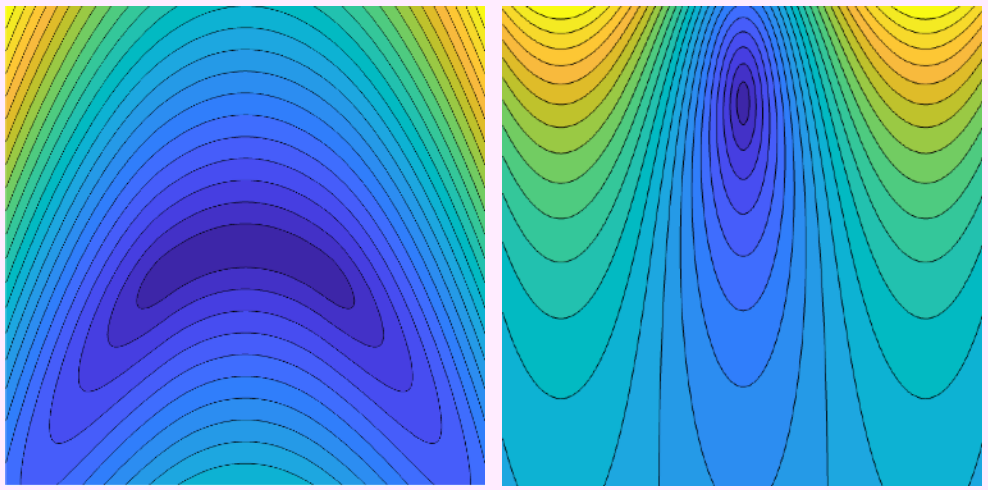
$\rightarrow$ find a suitable limit model and show "implicit" regularization effect.

Mean field for 2 layers perceptron [Chizat-Bach 2018]: **global** convergence.

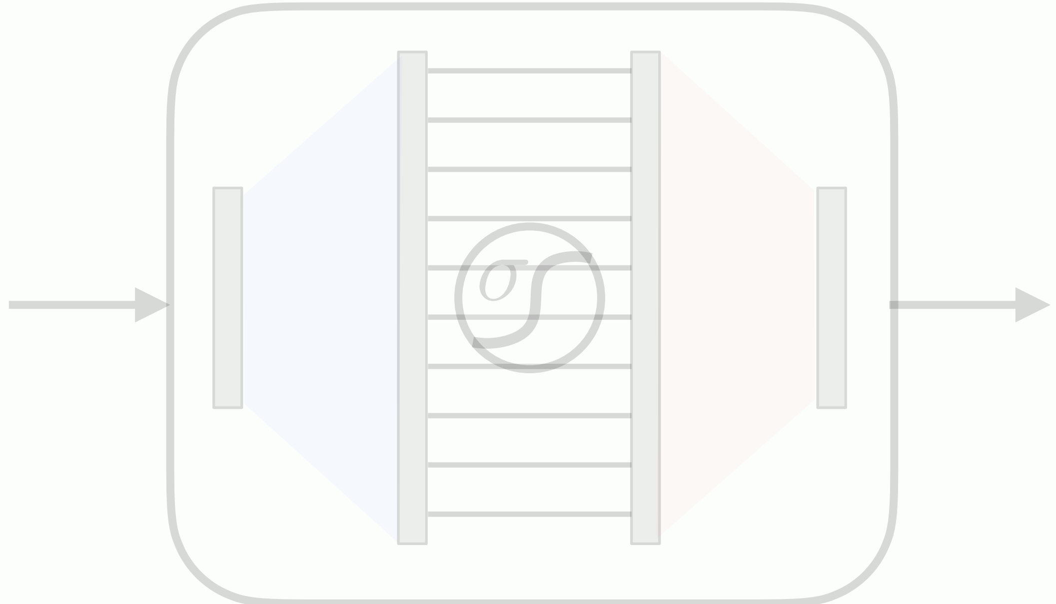
ResNet and Neural-ODEs



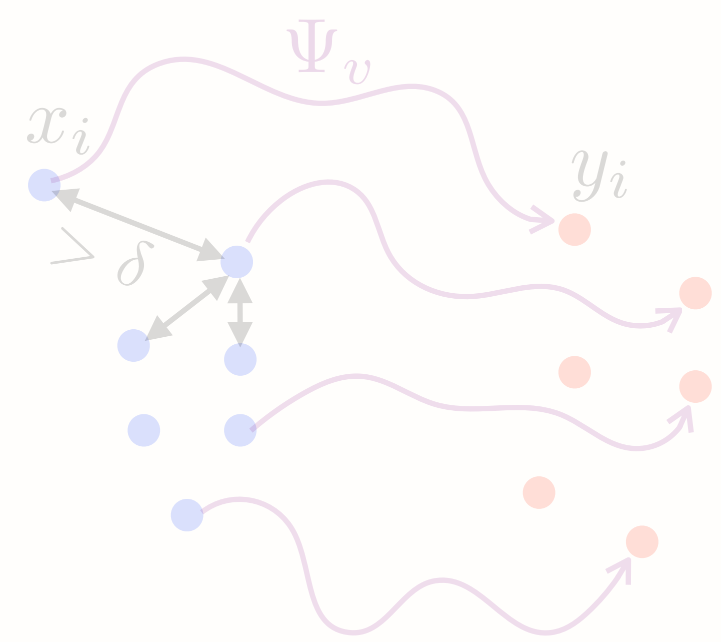
Global and local Polyak-Łojasiewicz conditions



RKHS Neural-ODEs



P-Ł condition for Neural-ODEs

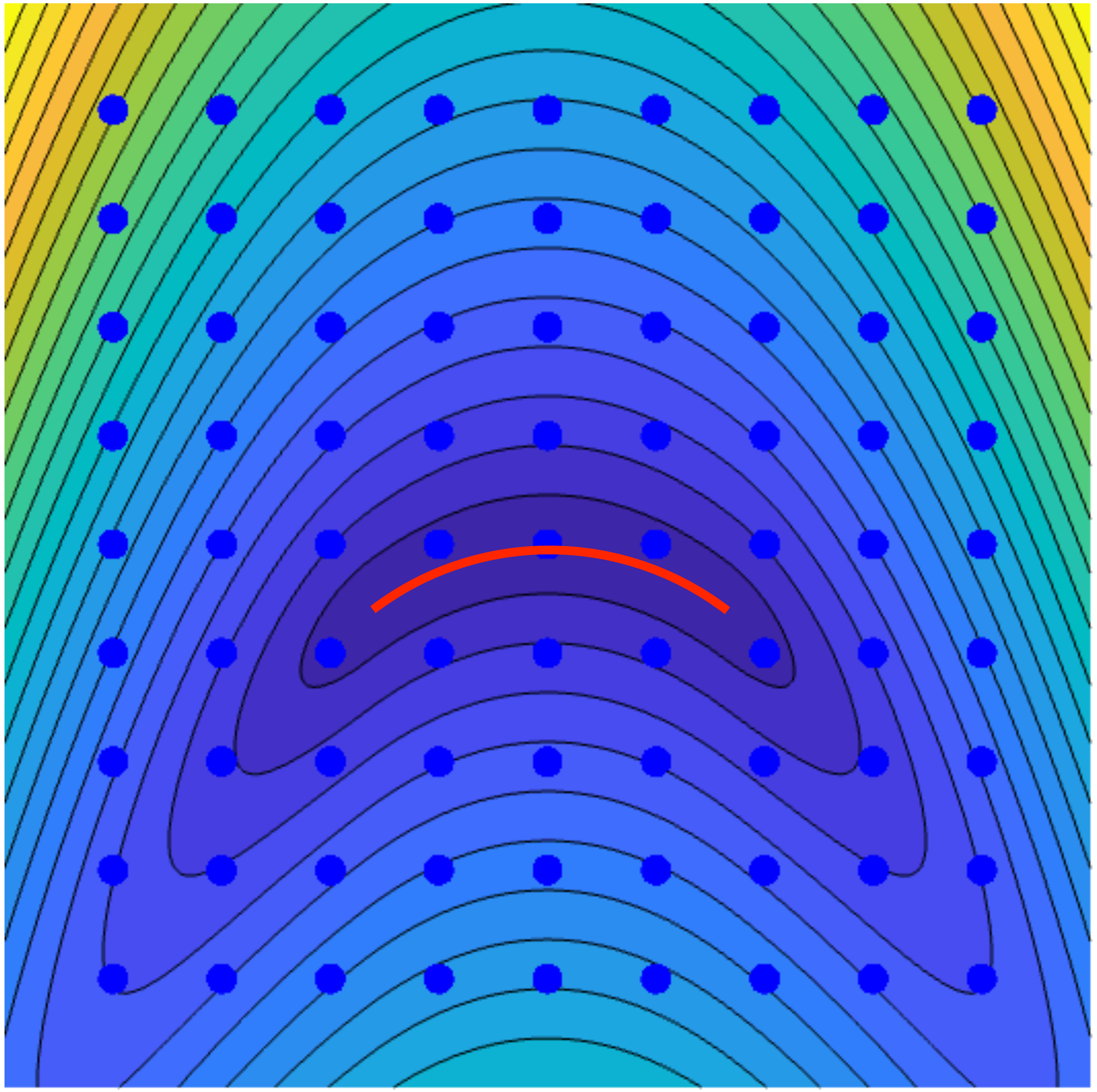


Polyak-Łojasiewicz Condition

Polyak-Łojasiewicz inequality: $0 \leqslant \ mf(\theta) \leqslant \|\nabla f(\theta)\|^2$

$\rightarrow$ no spurious stationary points.

Example: f strongly convex.

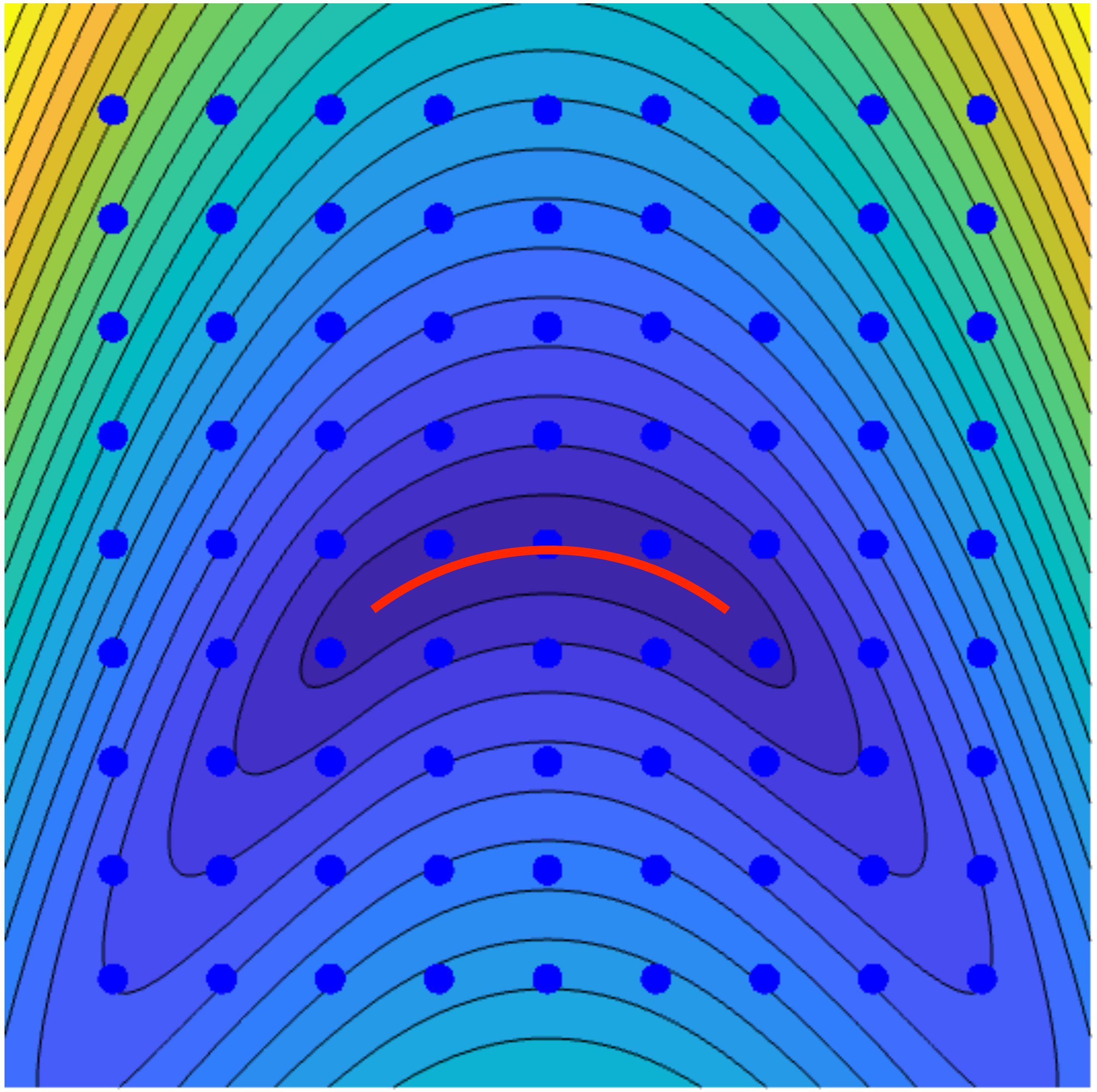


Polyak-Łojasiewicz Condition

Polyak-Łojasiewicz inequality: $0 \leqslant \; mf(\theta) \leqslant \|\nabla f(\theta)\|^2$

$\rightarrow$ no spurious stationary points.

Example: f strongly convex.

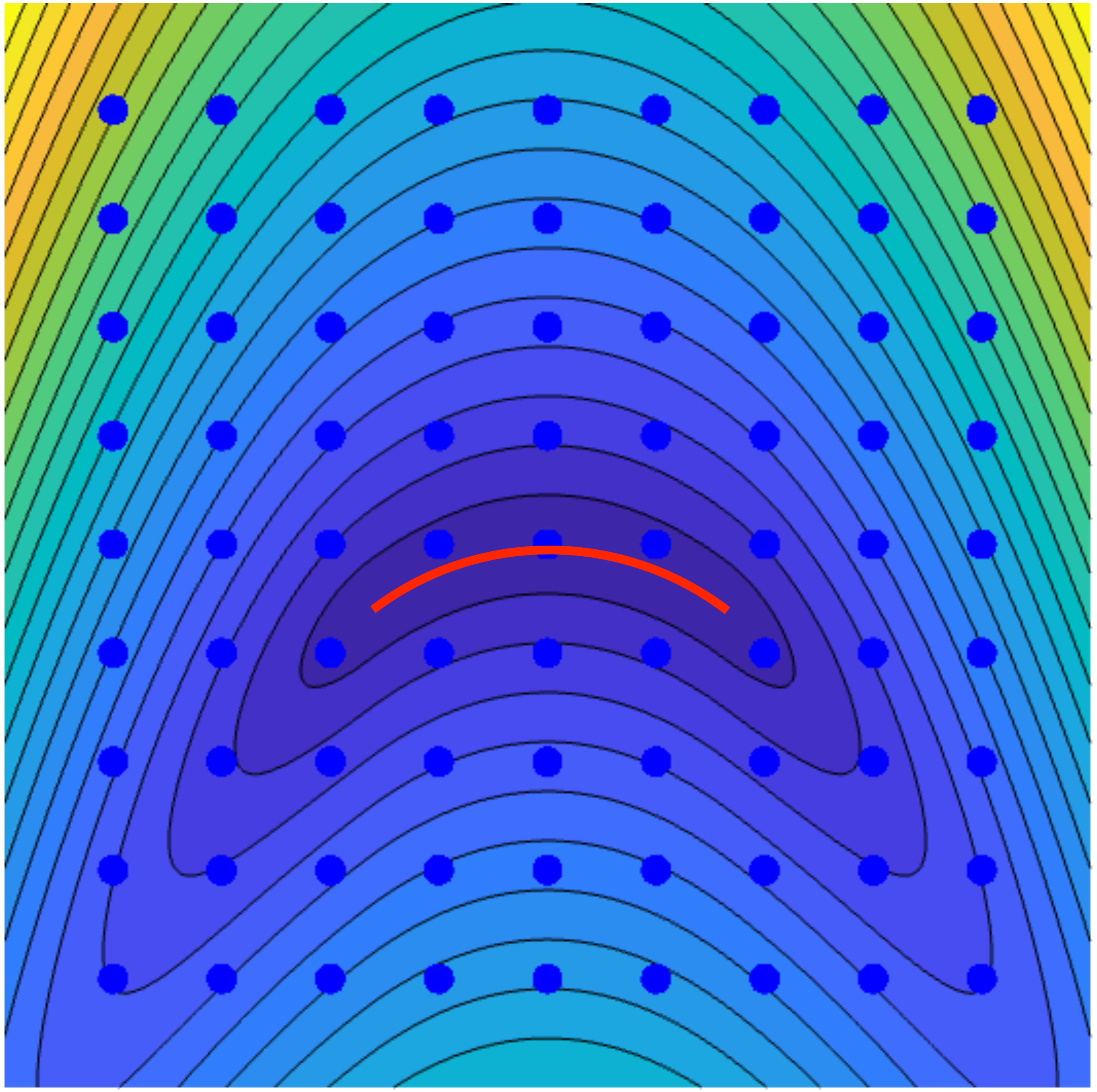


Polyak-Łojasiewicz Condition

Polyak-Łojasiewicz inequality: $0 \leqslant \; mf(\theta) \leqslant \|\nabla f(\theta)\|^2$

$\rightarrow$ no spurious stationary points.

Example: f strongly convex.



Gradient descent:

If ∇f is β-Lipschitz:

$$\theta^{(k+1)} = \theta^{(k)} - \frac{1}{\beta}\nabla f(\theta^{(k)})$$

Theorem: [Polyak 1963]

$$f(\theta^{(k)}) \leqslant \left(1 - \tfrac{m}{2\beta}\right)^k f(\theta^{(0)})$$

Obstruction for P-Ł for Neural ODE

Linear ResNet, time-independant weights: $\dot{x} = \theta x$ $\Phi_\theta(x) = e^\theta x$

For $y^j = -x^i$, i.e. learning $-\mathrm{Id}$:

$$f(\theta) \triangleq \|e^\theta + \mathrm{Id}\|^2$$

$$\theta^{(k+1)} = \theta^{(k)} - \frac{1}{\beta}\nabla f(\theta^{(k)})$$

Obstruction for P-Ł for Neural ODE

Linear ResNet, time-independant weights: $\dot{x} = \theta x \qquad \Phi_\theta(x) = e^\theta x$

For $y^j = -x^i$, i.e. learning $-\mathrm{Id}$:

$$f(\theta) \triangleq \|e^\theta + \mathrm{Id}\|^2$$

$$\theta^{(k+1)} = \theta^{(k)} - \frac{1}{\beta}\nabla f(\theta^{(k)})$$

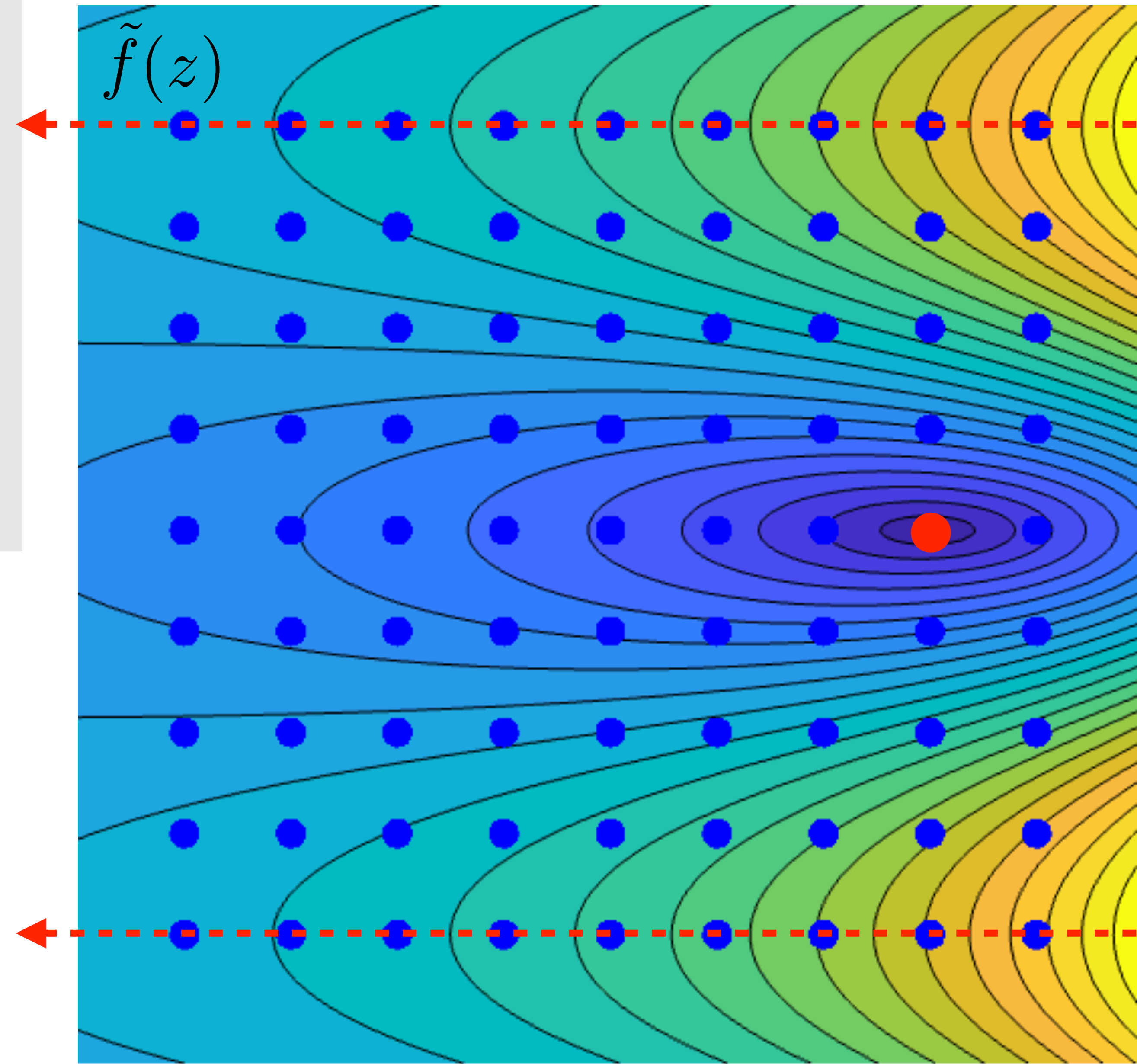
Proposition:

If $\theta^{(0)} = U\,\mathrm{diag}(z_1^{(0)}, \ldots, z_d^{(0)})U^*$,

then $\theta^{(k)} = U\,\mathrm{diag}(z_1^{(k)}, \ldots, z_d^{(k)})U^*$

where $z_i^{(k)} \in \mathbb{C}$ is a gradient descent of

$$\tilde{f}(z) \triangleq |e^z + 1|^2$$



Problem: $\tilde{f}$ does not satisfies P-Ł

For $\mathrm{Im}(z^{(0)}) = 0\ [2\pi]$, $\mathrm{Re}(z^{(k)}) \to -\infty$.

If $\theta^{(0)} = 0$, $e^{\theta^{(k)}} \to 0$ (not invertible)

Local P-Ł Condition

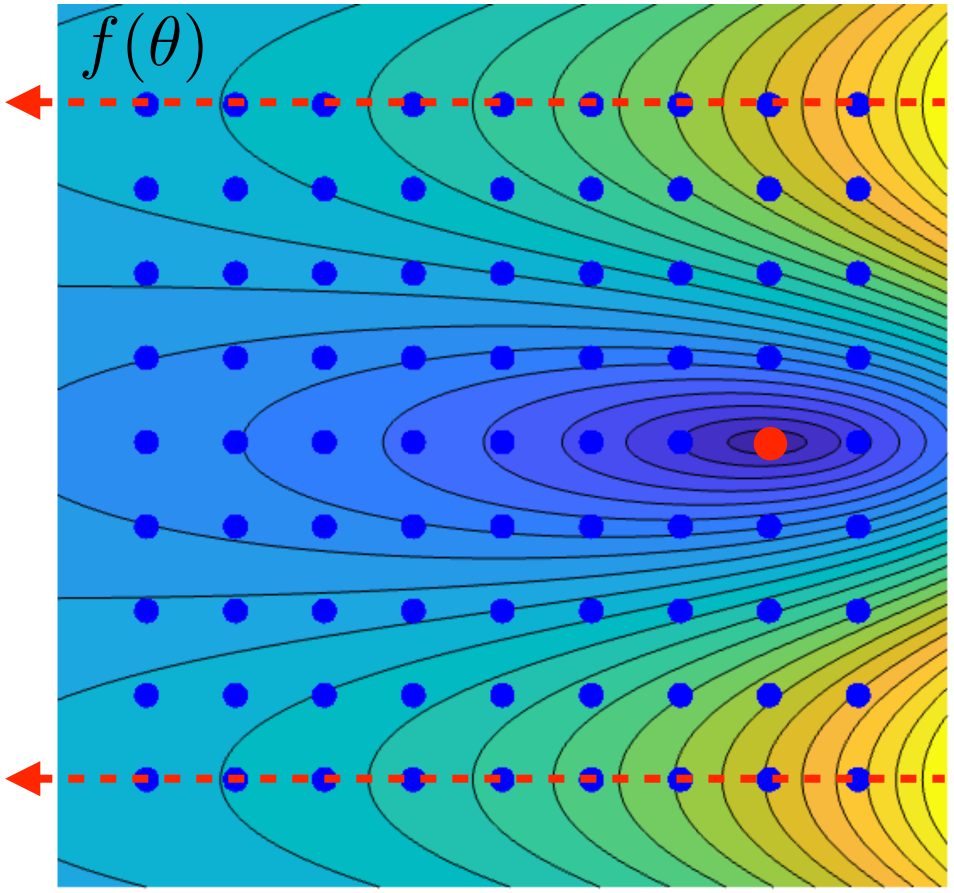
$$0 \leqslant m(\|\theta\|) f(\theta) \leqslant \|\nabla f(\theta)\|^2$$

m degenerates as $\theta \to +\infty$

$\to$ repulses spurious minima at $+\infty$

Example: $f(\theta) = |e^{\theta_1 + \mathrm{i}\theta_2} + 1|^2$

$$m(R) = e^{-2\|\theta\|}$$



Local P-Ł Condition

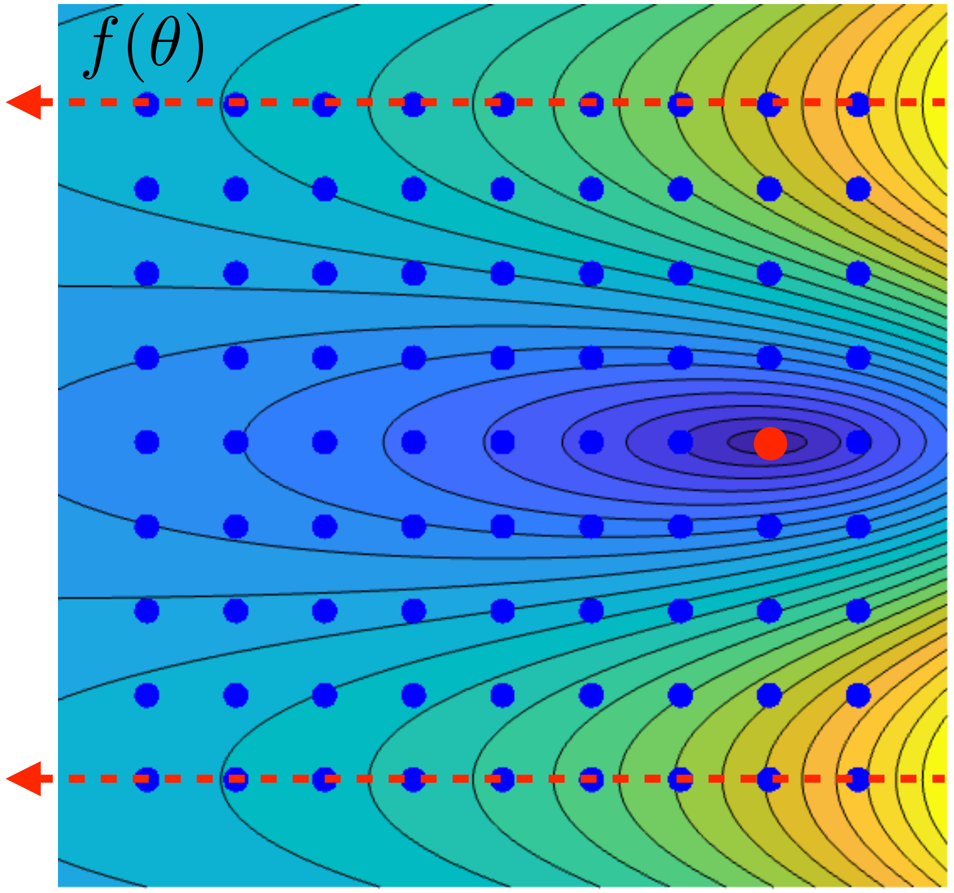
$$0 \leqslant \; m(\|\theta\|) f(\theta) \leqslant \|\nabla f(\theta)\|^2$$

m degenerates as $\theta \to +\infty$

$\to$ repulses spurious minima at $+\infty$

Example: $f(\theta) = |e^{\theta_1 + \mathrm{i}\theta_2} + 1|^2$

$$m(R) = e^{-2\|\theta\|}$$



Local P-Ł Condition

$$0 \leqslant m(\|\theta\|)f(\theta) \leqslant \|\nabla f(\theta)\|^2 \leqslant M(\|\theta\|)f(\theta)$$

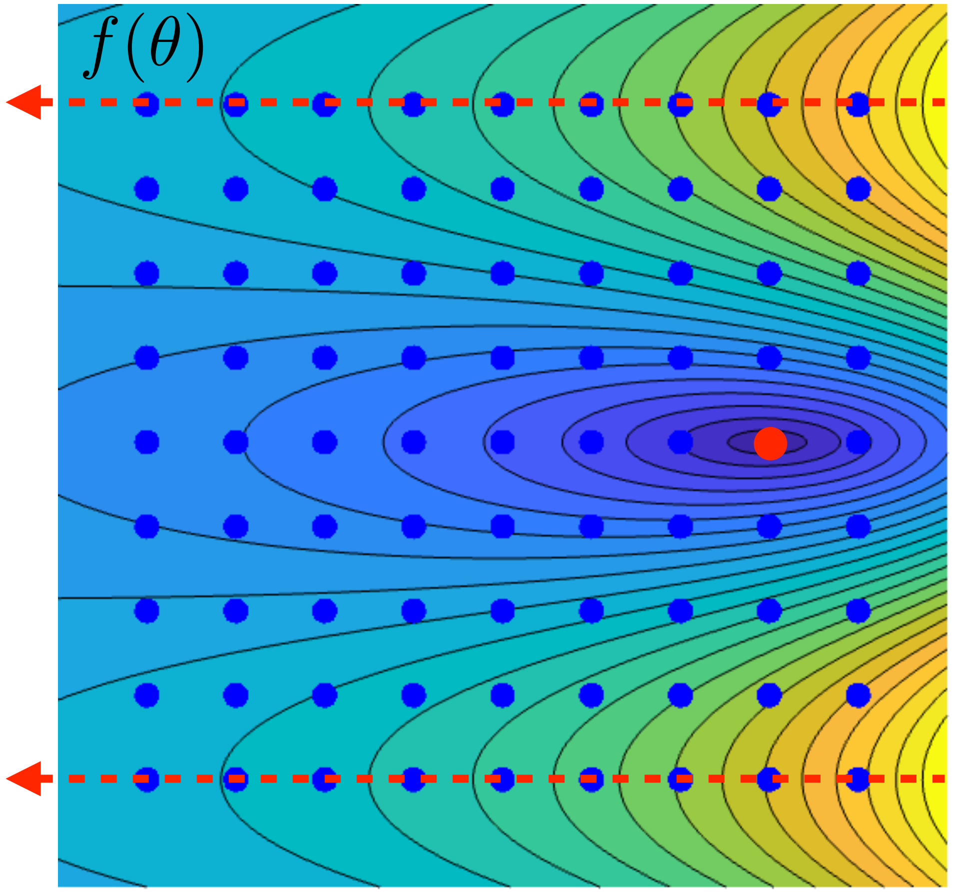
m degenerates as $\theta \to +\infty$

$\to$ repulses spurious minima at $+\infty$

Localize trajectories

Example: $f(\theta) = |e^{\theta_1 + \mathrm{i}\theta_2} + 1|^2$

$m(R) = e^{-2\|\theta\|}$ $M(R) = e^{2\|\theta\|}$



Theorem : [Liu, Zhu, Belkin 2021] If $\theta^{(0)}$ and $R > 0$ satisfies

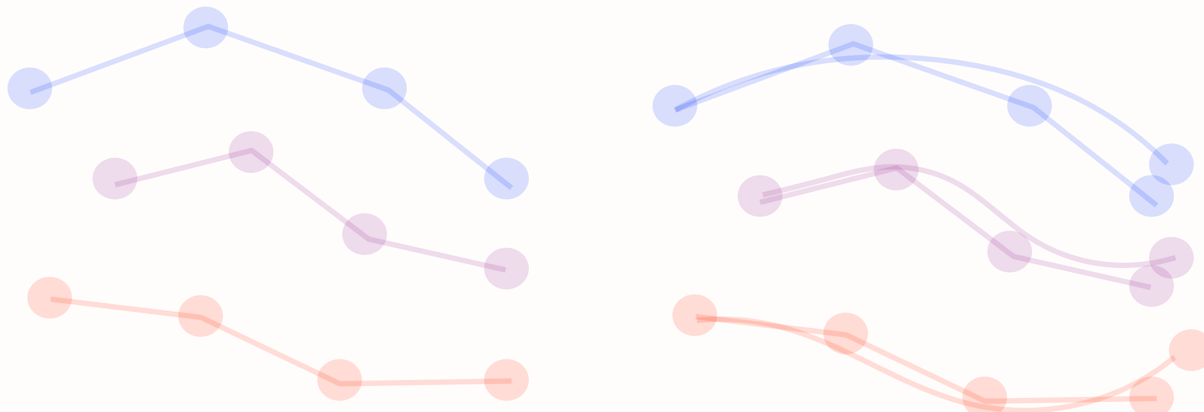
$$f(\theta^{(0)}) \leqslant \frac{m(\|\theta^{(0)}\| + R)^2}{M(\|\theta^{(0)}\| + R)} R^2$$

then

$$f(\theta^{(k)}) \leqslant \left(1 - \frac{m(\|\theta^{(0)} + R\|)}{2\beta}\right)^k f(\theta^{(0)})$$

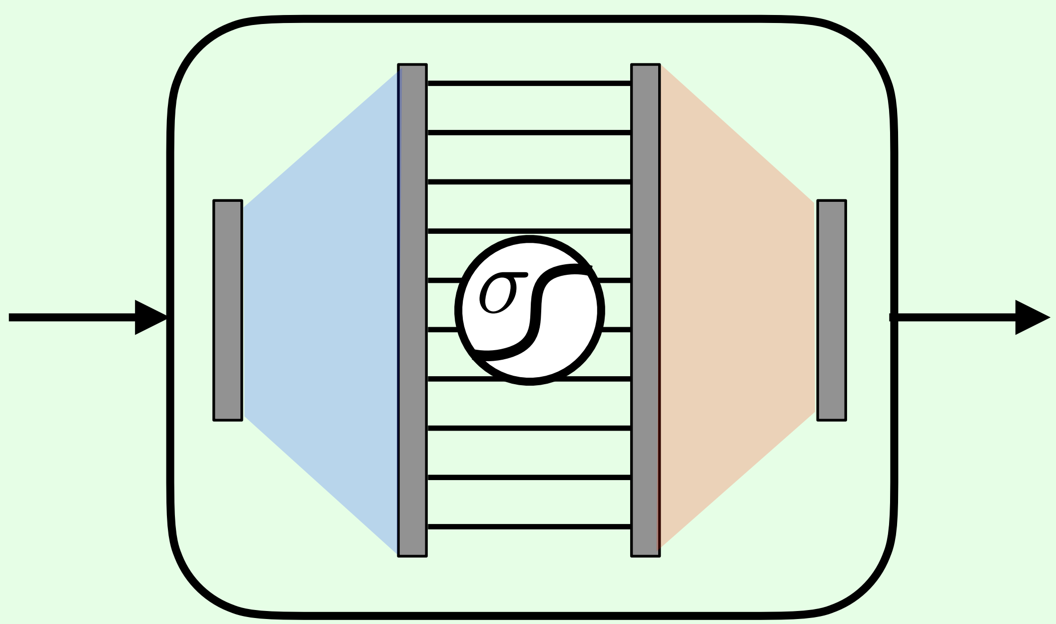
and $\|\theta^{(k)} - \theta^{(0)}\| \leqslant R$.

ResNet and Neural-ODEs

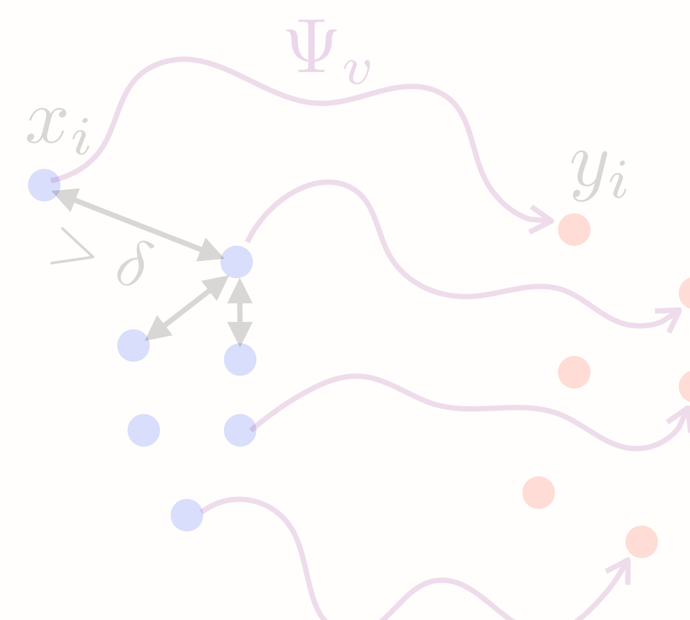


Global and local Polyak-Łojasiewicz conditions

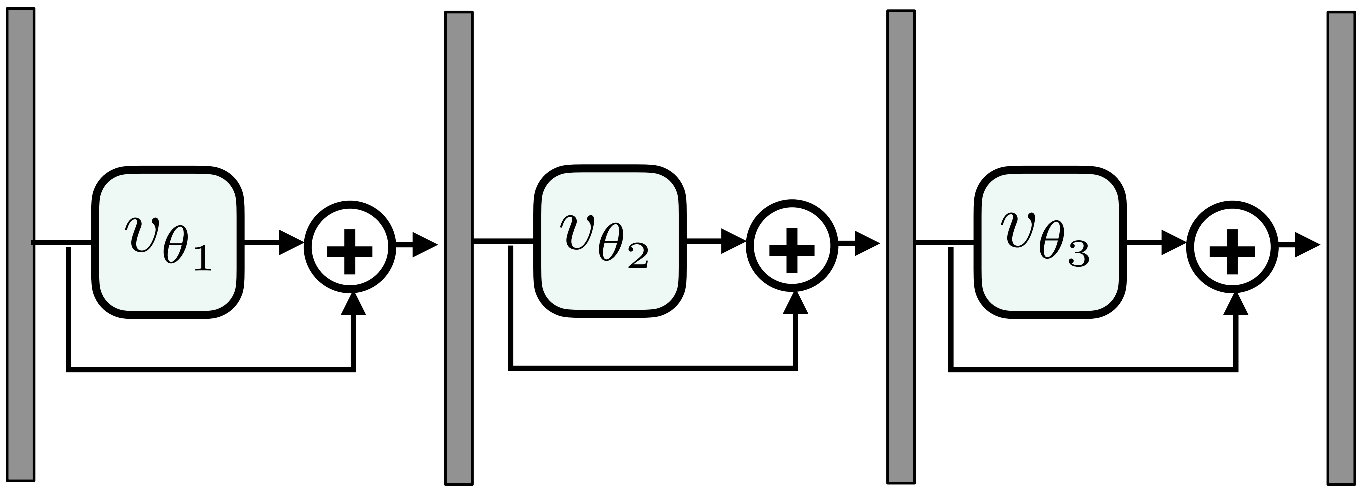
RKHS Neural-ODEs

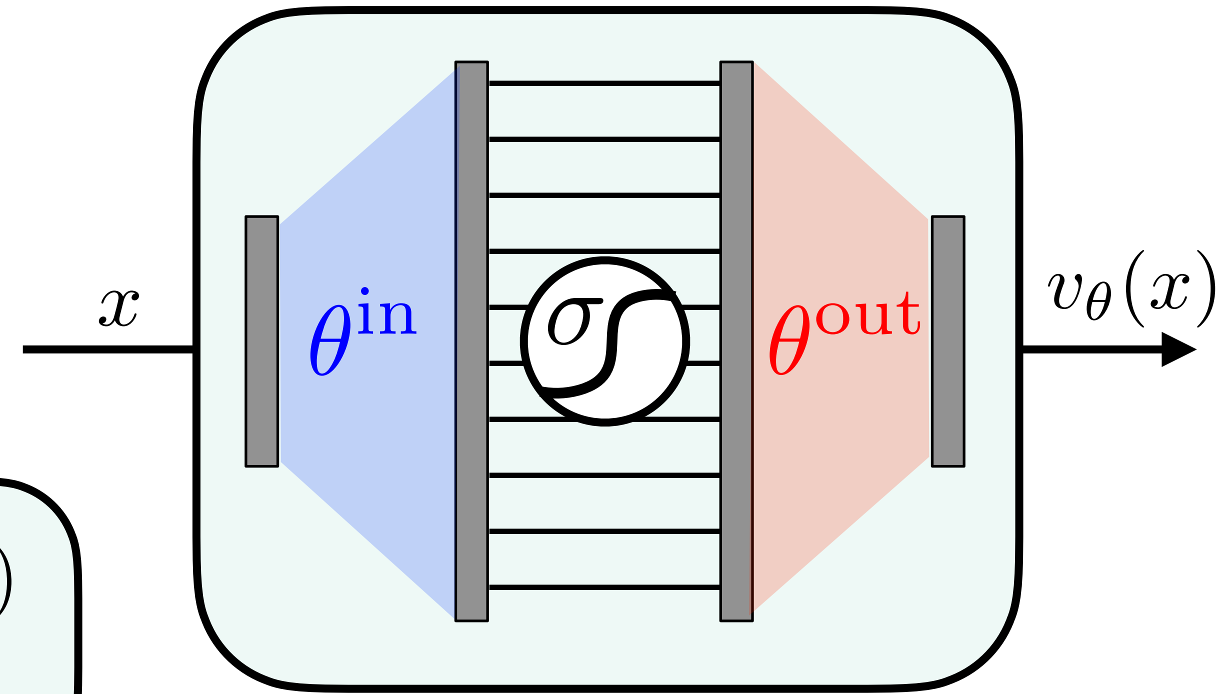


P-Ł condition for Neural-ODEs



Infinite Width and RKHS Parameterization

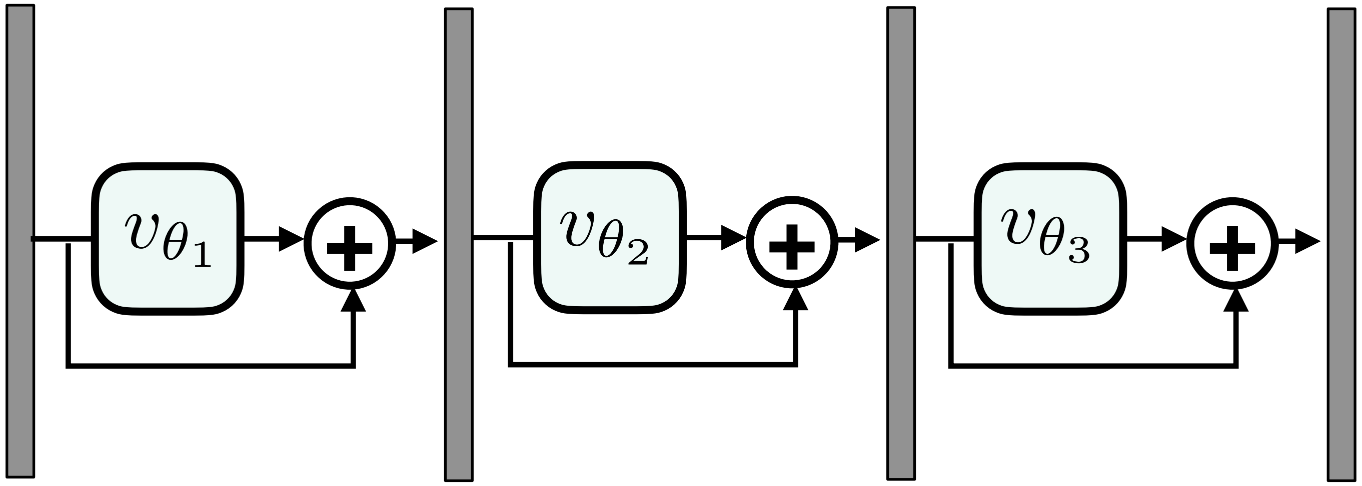


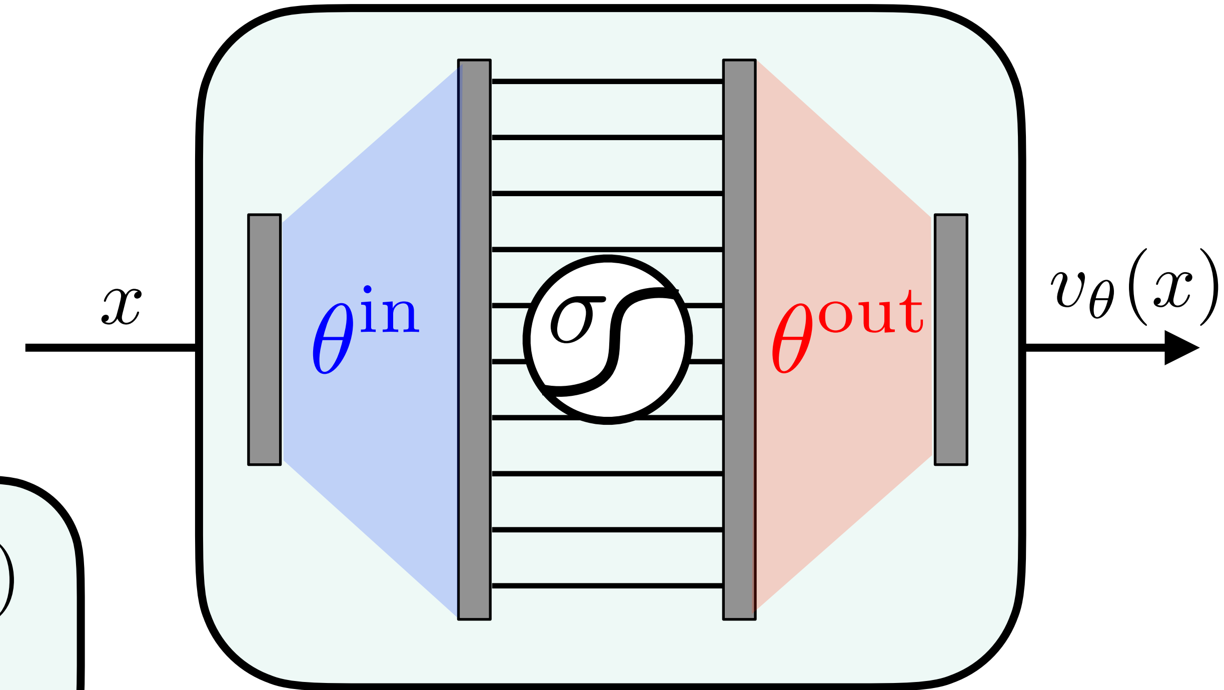


2 layers perceptron: $v_\theta(x) = \theta^{\text{out}} \sigma(\theta^{\text{in}} x)$

$$(\theta^{\text{in}}, \theta^{\text{out}}) \in \mathbb{R}^{d \times q} \times \mathbb{R}^{q \times d}$$

Infinite Width and RKHS Parameterization





2 layers perceptron: $v_\theta(x) = \theta^{\text{out}}\sigma(\theta^{\text{in}}x)$

$(\theta^{\text{in}}, \theta^{\text{out}}) \in \mathbb{R}^{d\times q} \times \mathbb{R}^{q\times d}$

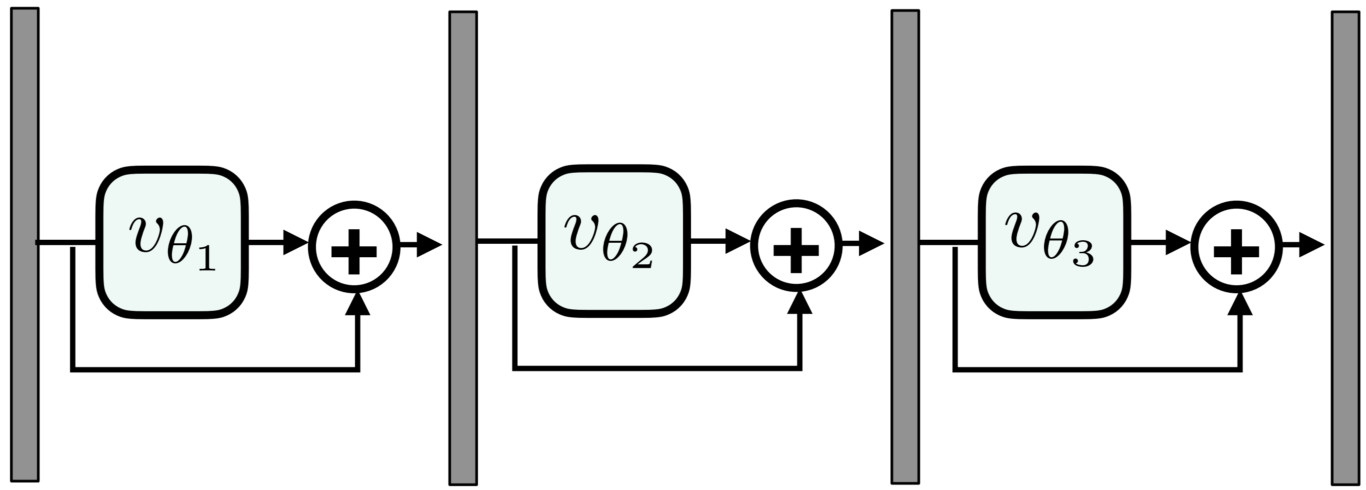
Simplification: only train θ^{out}.

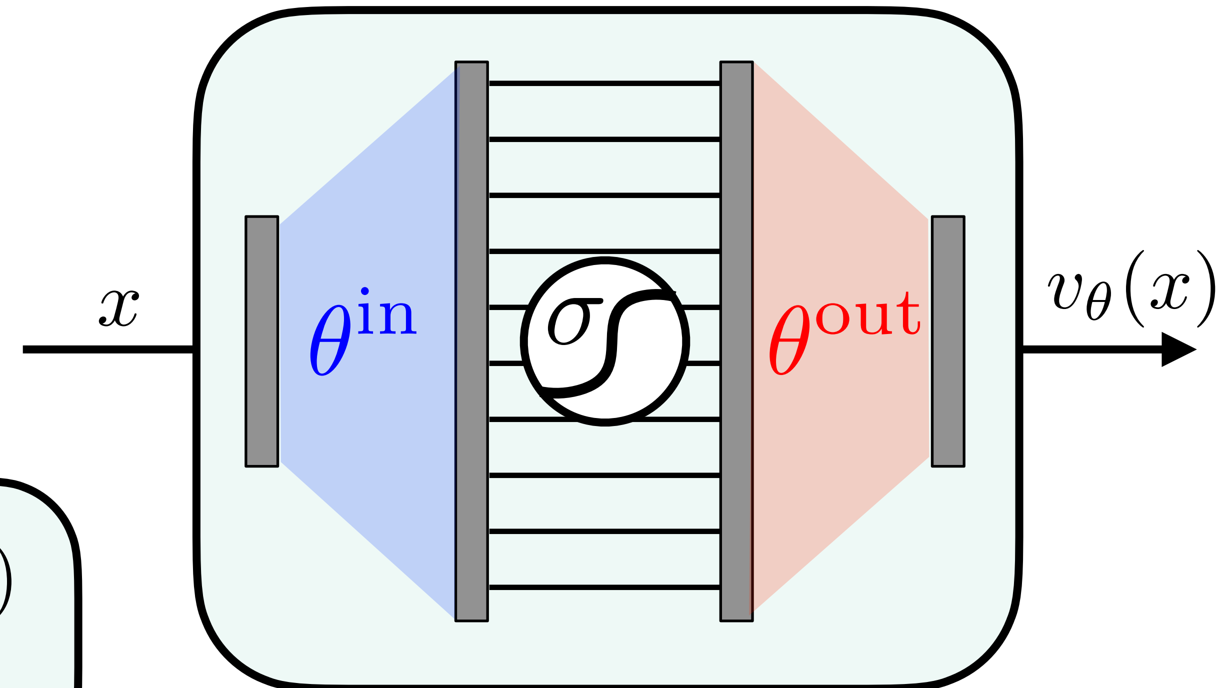
$\rightarrow$ (finite dimensional) Reproducing Kernel Hilbert Space $\mathbb{V}$.

$$\|v\|_{\mathbb{V}} \triangleq \inf_{v=v_\theta} \|\theta^{\text{out}}\|_{\mathbb{R}^{q\times d}}$$

Kernel: $k(x,x') \triangleq \langle \sigma(\theta^{\text{out}}x),\, \sigma(\theta^{\text{out}}x')\rangle_{\mathbb{R}^q}$

Infinite Width and RKHS Parameterization





2 layers perceptron: $v_\theta(x) = \theta^{\text{out}}\sigma(\theta^{\text{in}}x)$

$(\theta^{\text{in}}, \theta^{\text{out}}) \in \mathbb{R}^{d\times q} \times \mathbb{R}^{q\times d}$

Simplification: only train θ^{out}.

$\rightarrow$ (finite dimensional) Reproducing Kernel Hilbert Space $\mathbb{V}$.

$$\|v\|_{\mathbb{V}} \triangleq \inf_{v=v_\theta} \|\theta^{\text{out}}\|_{\mathbb{R}^{q\times d}}$$

Kernel: $k(x,x') \triangleq \langle \sigma(\theta^{\text{out}}x), \sigma(\theta^{\text{out}}x')\rangle_{\mathbb{R}^q}$

Gradient descent on $(\theta^{\text{out}}, \|\cdot\|_{\mathbb{R}^{d\times q}})$ $\Leftrightarrow$ Gradient descent on $(v, \|\cdot\|_{\mathbb{V}})$

Infinite width limit $q \rightarrow +\infty$: infinite dimensional RKVS $v \in \mathbb{V}$.

RKHS Neural ODE

Neural ODE:

$$\Phi_\theta(x(0)) \triangleq x(1) \quad \text{where}$$

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = v_{\theta(t)}(x(t))$$

Replace v_θ by

$v \in L^2([0,1], \mathbb{V})$

RKHS-Neural ODE:

$$\Psi_v(x(0)) \triangleq x(1) \quad \text{where}$$

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = v_t(x(t))$$

$$f(\theta) \triangleq \frac{1}{N}\sum_{i=1}^{N} \|B\Phi_\theta(Ax^i) - y^i\|^2$$

$$\theta^{(k+1)} = \theta^{(k)} - \tau \nabla f(\theta^{(k)})$$

$$\mathcal{F}(v) \triangleq \frac{1}{N}\sum_{i=1}^{N} \|B\Psi_v(Ax^i) - y^i\|^2$$

$$v^{(k+1)} = v^{(k)} - \tau \nabla_{\mathbb{V}} \mathcal{F}(v^{(k)})$$

RKHS Neural ODE

Neural ODE:

$\Phi_\theta(x(0)) \triangleq x(1)$ where

$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = v_{\theta(t)}(x(t))$

Replace v_θ by $v \in L^2([0,1], \mathbb{V})$

RKHS-Neural ODE:

$\Psi_v(x(0)) \triangleq x(1)$ where

$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = v_t(x(t))$

$$f(\theta) \triangleq \frac{1}{N}\sum_{i=1}^{N} \|B\Phi_\theta(Ax^i) - y^i\|^2$$

$$\theta^{(k+1)} = \theta^{(k)} - \tau \nabla f(\theta^{(k)})$$

$$\mathcal{F}(v) \triangleq \frac{1}{N}\sum_{i=1}^{N} \|B\Psi_v(Ax^i) - y^i\|^2$$

$$v^{(k+1)} = v^{(k)} - \tau \nabla_{\mathbb{V}} \mathcal{F}(v^{(k)})$$

Local P-Ł of $f(\theta)$ on $(\mathbb{R}^{q\times d})^T$

$\|\theta\|^2 = \sum_i \|\theta_i^{\mathrm{out}}\|^2$

width $q \to +\infty$

$\not\Longrightarrow$

depth $T \to +\infty$

Local P-Ł of $\mathcal{F}(v)$ on $L^2([0,1], \mathbb{V})$

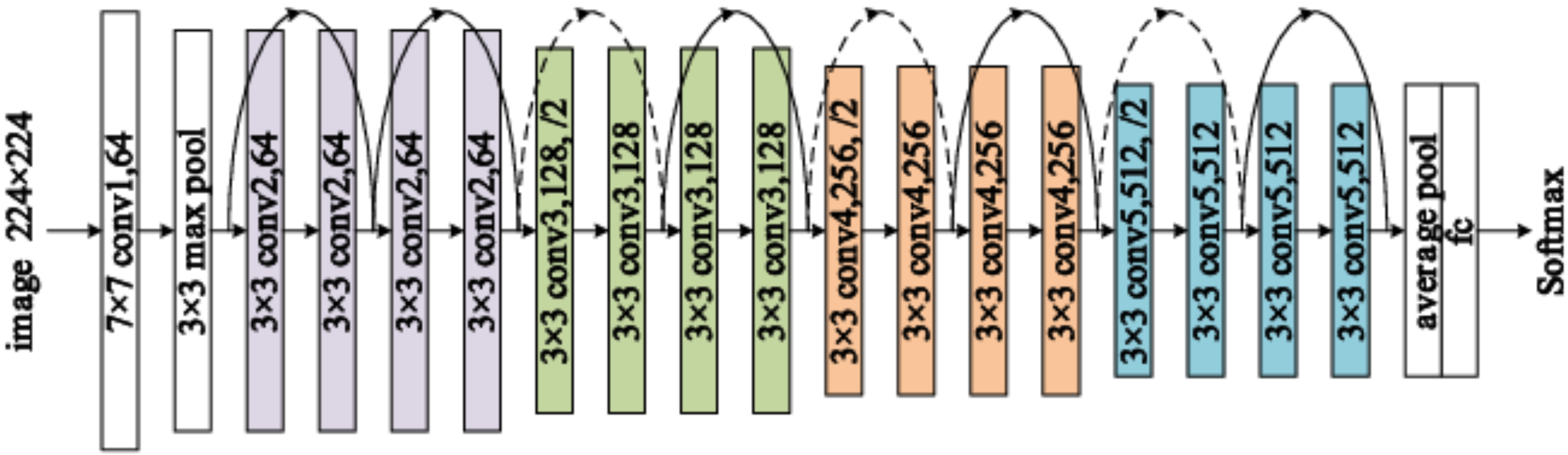
$\|v\|^2 \triangleq \int_0^1 \|v_t\|_{\mathbb{V}}^2 \mathrm{d}t$

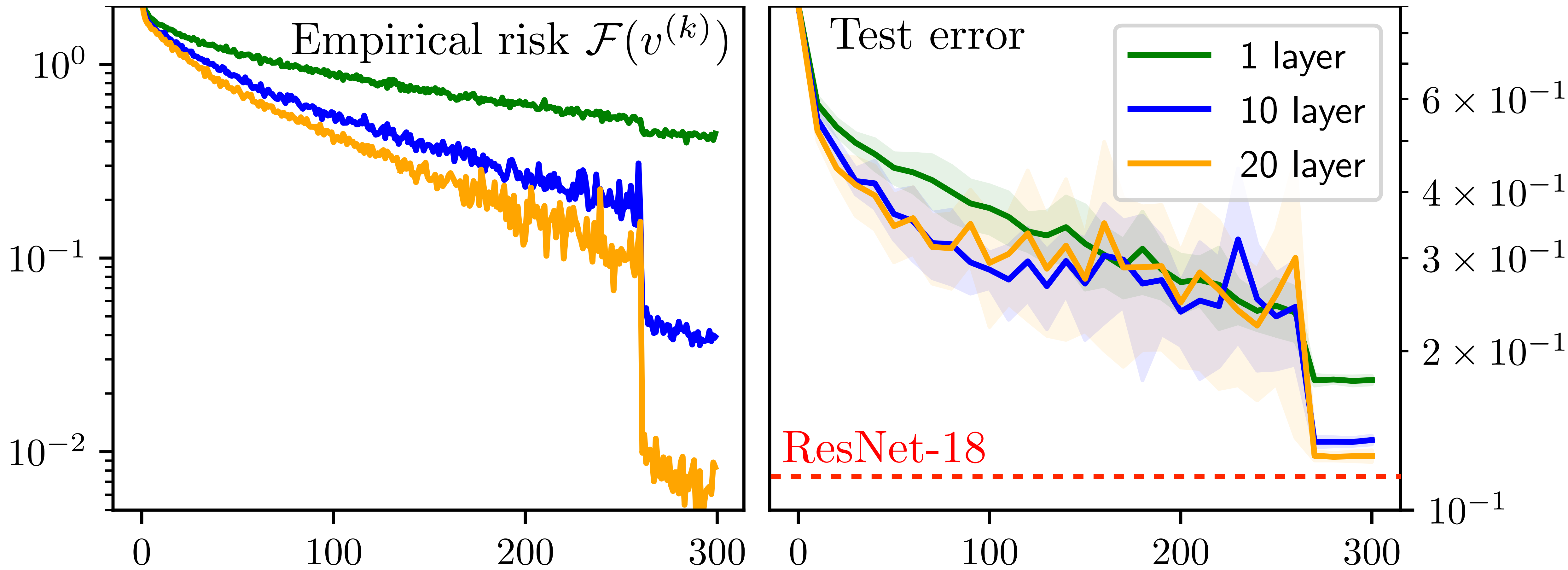
Questions: conditions on data $(x_i)_i$ / kernel $k(x, x')$ to ensure local PŁ of $\mathcal{F}(v)$?

Numerical Example

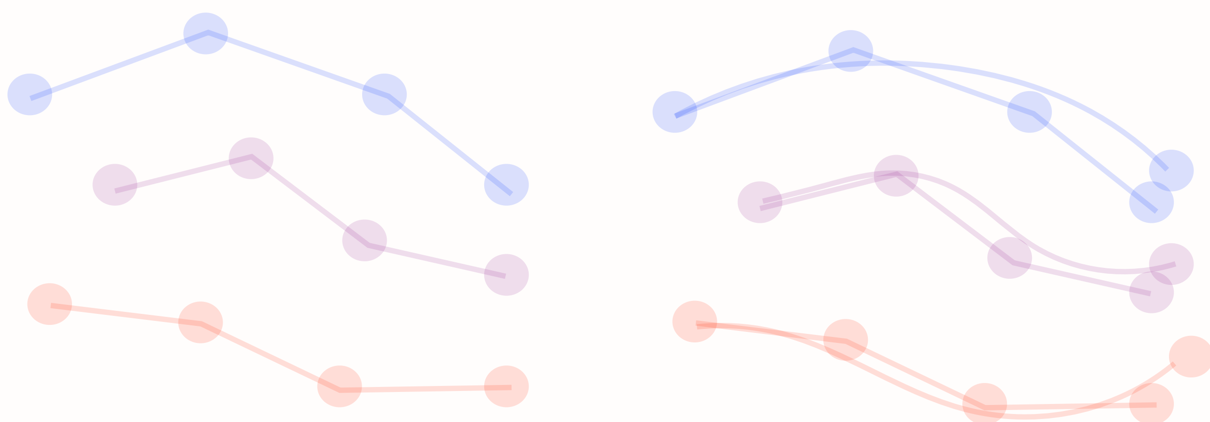
RKHS Neural ODE trained on CIFAR10

Random Fourier features (Sobolev RKHS)



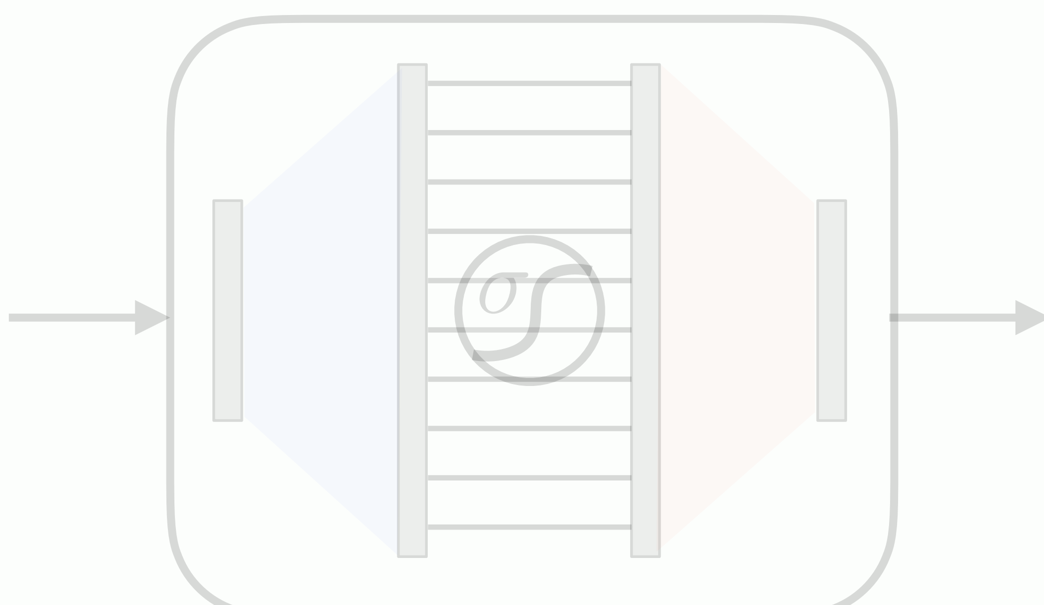


ResNet and Neural-ODEs

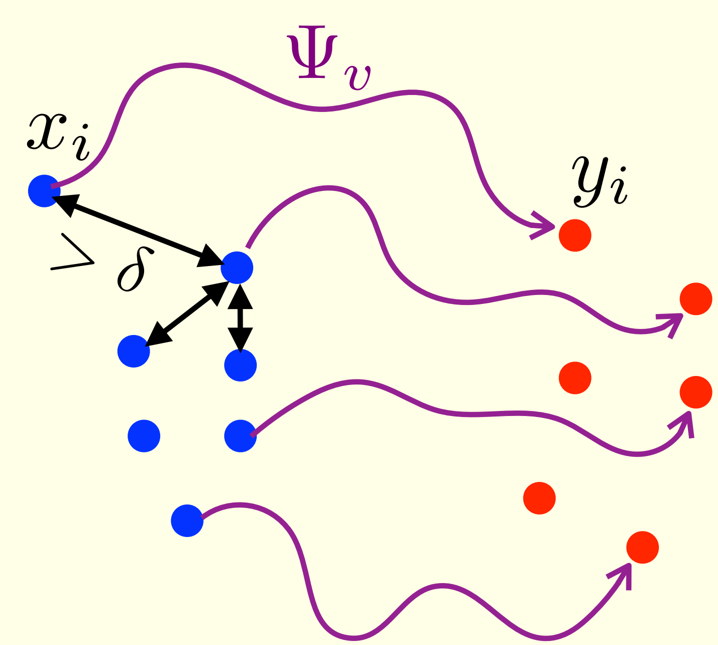


Global and local Polyak-Łojasiewicz conditions

RKHS Neural-ODEs



P-Ł condition for Neural-ODEs



Regularity Condition

Condition 1: regularity. For $v_t \in \mathbb{V}$, $\|v_t\|_\infty + \|Dv_t\|_\infty + \|D^2 v_t\|_\infty \leqslant \kappa \|v_t\|_{\mathbb{V}}$

Needed for ODE (Cauchy-Lipschitz)

Needed for gradient descent

$\rightarrow$ problem with ReLu!

$\rightarrow \sigma = e^{\mathrm{i}\cdot}$: Fourier features $\Rightarrow$ translation invariant kernels.

Regularity Condition

Condition 1: regularity. For $v_t \in \mathbb{V}$, $\|v_t\|_\infty + \|Dv_t\|_\infty + \|D^2 v_t\|_\infty \leqslant \kappa \|v_t\|_{\mathbb{V}}$

Needed for ODE (Cauchy-Lipschitz) — applies to $\|Dv_t\|_\infty$

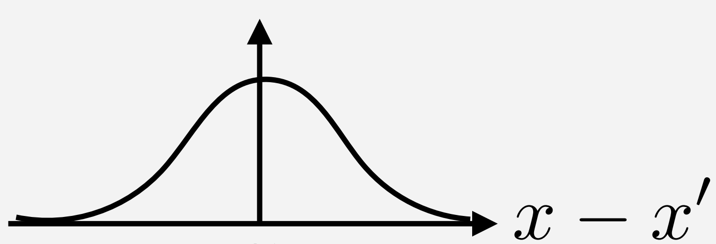
Needed for gradient descent — applies to $\|D^2 v_t\|_\infty$

→ problem with ReLu!

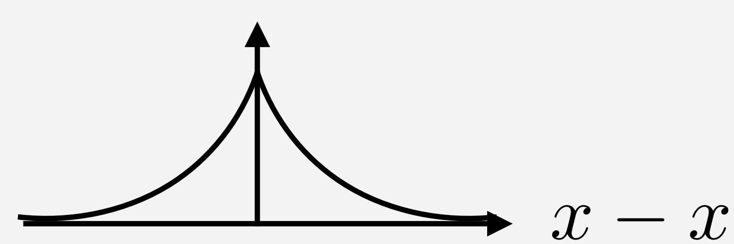
→ $\sigma = e^{\mathrm{i}\cdot}$: Fourier features ⇒ translation invariant kernels.

Example: Mattern kernel $\mathbb{V} = H^\nu$ (Sobolev)

$$\hat{k}(\omega) \propto (1 + \|\omega\|^2/\nu)^{-(d/2+\nu)}$$



$\nu = \infty$: Gaussian.



$\nu = 0$: Laplacian.

Proposition: ($\nu > 2$)

$$\kappa \leqslant 1 + \sqrt{\frac{\nu}{\nu - 1}} + \sqrt{\frac{3\nu^2}{(\nu-1)(\nu-2)}}.$$

Expressivity Condition and P-Ł

Condition 1: regularity. For $v_t \in \mathbb{V}$, $\|v_t\|_\infty + \|Dv_t\|_\infty + \|D^2 v_t\|_\infty \leqslant \kappa \|v_t\|_\mathbb{V}$

Condition 2: quantitative universality. $K_X \triangleq (k(x_i, x_j))_{i,j}$.

$$\lambda(\delta) \triangleq \inf_{\#X \leqslant N} \{\lambda_{\min}(K_X) \; ; \; \forall i \neq j, \|x_i - x_j\| \geqslant \delta\} > 0$$

$\lambda(\delta)$: depends on N, explodes as $\delta \to 0$.

Expressivity Condition and P-Ł

Condition 1: regularity. For $v_t \in \mathbb{V}$, $\|v_t\|_\infty + \|Dv_t\|_\infty + \|D^2 v_t\|_\infty \leqslant \kappa \|v_t\|_{\mathbb{V}}$

Condition 2: quantitative universality. $K_X \triangleq (k(x_i, x_j))_{i,j}$.

$$\lambda(\delta) \triangleq \inf_{\#X \leqslant N} \{\lambda_{\min}(K_X) \; ; \; \forall i \neq j, \|x_i - x_j\| \geqslant \delta\} > 0$$

$\lambda(\delta)$: depends on N, explodes as $\delta \to 0$.

Theorem: If $\forall i \neq j, \|x^i - x^j\| \geqslant \delta$, then

$$m(\|v\|_{\mathcal{H}})\mathcal{F}(v) \leqslant \|\nabla_{\mathcal{H}}\mathcal{F}(v)\|_{\mathcal{H}}^2 \leqslant M(\|v\|_{\mathcal{H}})\mathcal{F}(v)$$

where
$$\begin{cases} M(R) \leqslant \sigma_{\max}(B)^2 e^{2\kappa R} \\ m(R) \geqslant \sigma_{\min}(B)^2 \lambda(\sigma_{\min}(A)\delta e^{-\kappa R}) e^{-2\kappa R} \end{cases}$$

Hilbert space: $\mathcal{H} \triangleq L^2([0,1], \mathbb{V})$, $\|v\|_{\mathcal{H}}^2 \triangleq \int_0^1 \|v_t\|_{\mathbb{V}}^2 \mathrm{d}t$

Expressivity Condition and P-Ł

Condition 1: regularity. For $v_t \in \mathbb{V}$, $\|v_t\|_\infty + \|Dv_t\|_\infty + \|D^2 v_t\|_\infty \leqslant \kappa \|v_t\|_{\mathbb{V}}$

Condition 2: quantitative universality. $K_X \triangleq (k(x_i, x_j))_{i,j}$.

$$\lambda(\delta) \triangleq \inf_{\#X \leqslant N} \{\lambda_{\min}(K_X) \;;\; \forall i \neq j, \|x_i - x_j\| \geqslant \delta\} > 0$$

$\lambda(\delta)$: depends on N, explodes as $\delta \to 0$.

Theorem: If $\forall i \neq j, \|x^i - x^j\| \geqslant \delta$, then

$$m(\|v\|_{\mathcal{H}})\mathcal{F}(v) \leqslant \|\nabla_{\mathcal{H}}\mathcal{F}(v)\|_{\mathcal{H}}^2 \leqslant M(\|v\|_{\mathcal{H}})\mathcal{F}(v)$$

where
$$\begin{cases} M(R) \leqslant \sigma_{\max}(B)^2 e^{2\kappa R} \\ m(R) \geqslant \sigma_{\min}(B)^2 \lambda(\sigma_{\min}(A)\delta e^{-\kappa R}) e^{-2\kappa R} \end{cases}$$

Hilbert space: $\mathcal{H} \triangleq L^2([0,1], \mathbb{V})$, $\|v\|_{\mathcal{H}}^2 \triangleq \int_0^1 \|v_t\|_{\mathbb{V}}^2 \mathrm{d}t$

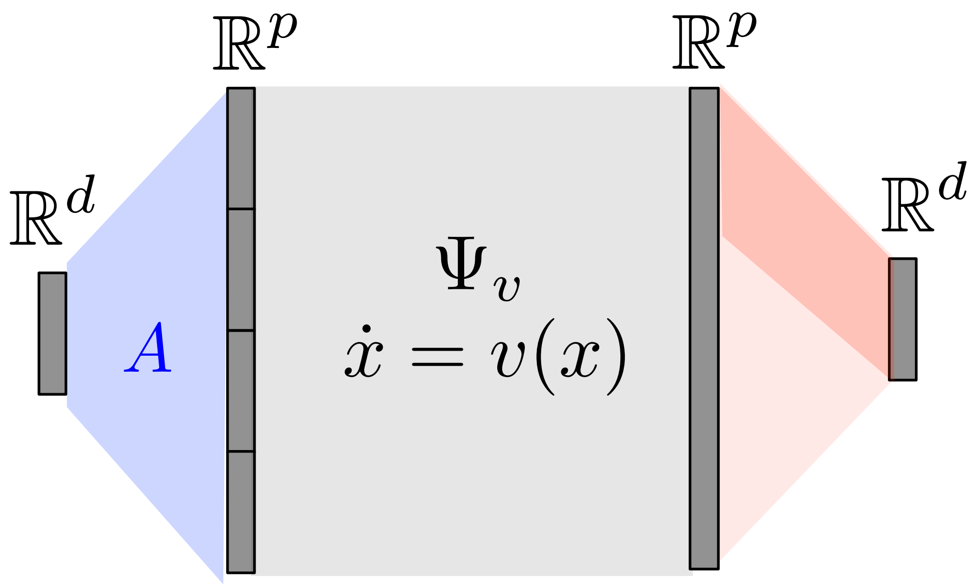
Corrolary: if $\mathcal{F}(v^{(0)})$ small enough, linear convergence of $\mathcal{F}(v^{(k)})$ to 0.

Enforcing Convergence via Lifting

Neural ODE constraint: $\Psi_v(x)$ is a diffeomorphism.

$\rightarrow$ density in continuous transformations require lifting.

$Lifting:$ $$A \propto (\mathrm{Id}_d, \ldots, \mathrm{Id}_d)^\top$$ $$B \propto (\mathrm{Id}_d, 0, \ldots, 0)$$



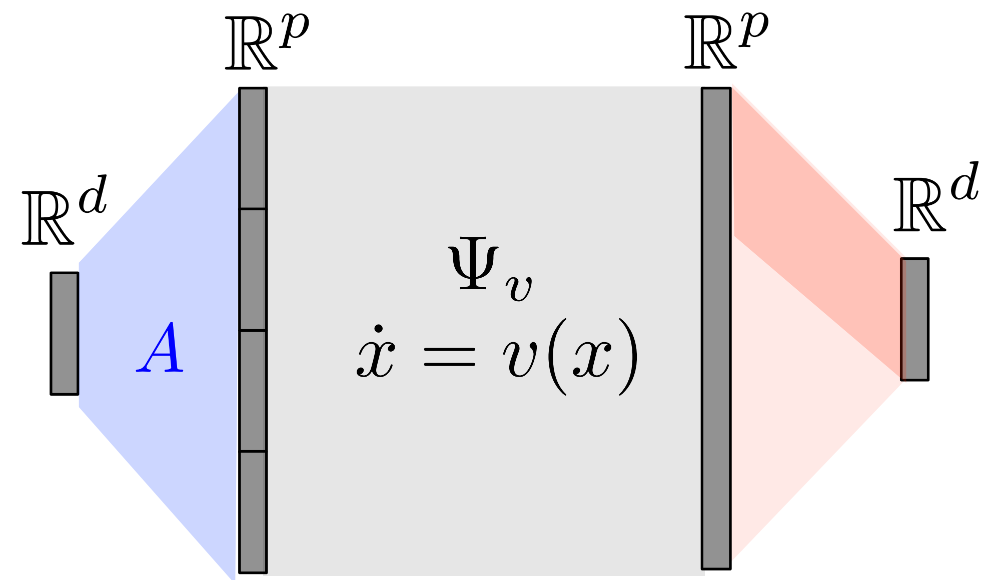
Enforcing Convergence via Lifting

Neural ODE constraint: $\Psi_v(x)$ is a diffeomorphism.

$\rightarrow$ density in continuous transformations require lifting.

Lifting : $A \propto (\mathrm{Id}_d, \ldots, \mathrm{Id}_d)^\top$

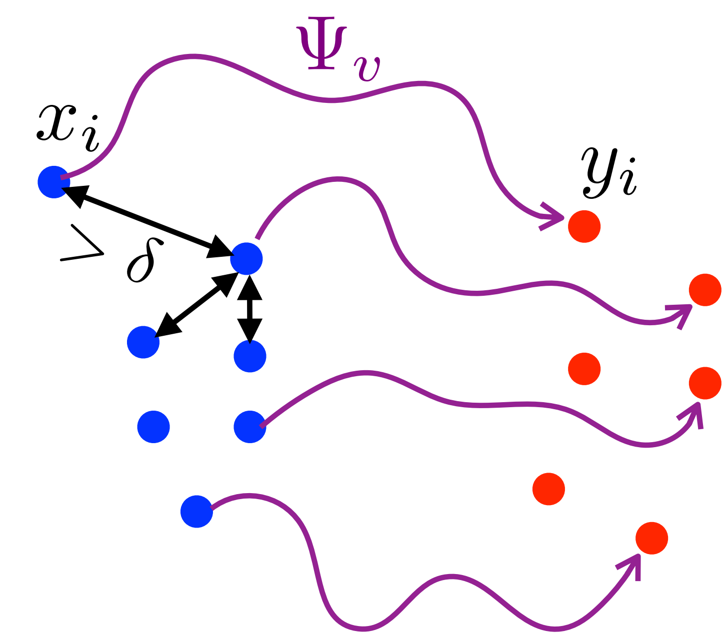
$B \propto (\mathrm{Id}_d, 0, \ldots, 0)$



Proposition: For $\nu > 2$, $\hat{k}(\omega) \propto (1 + \|\omega\|^2/\nu)^{-(d/2+\nu)}$

Given $v^{(0)}$ and $R > 0$, then for $p = O(N^4 + \delta \log(N)^4)$,

$$f(v^{(0)}) \leqslant \frac{m(\|v^{(0)}\| + R)^2}{M(\|v^{(0)}\| + R)} R^2$$

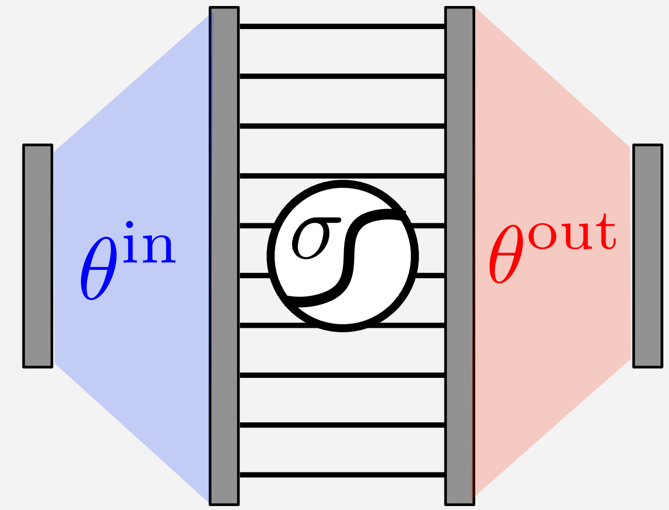


$M(R) \leqslant \sigma_{\max}(B)^2 e^{2\kappa R}$ $\qquad$ $m(R) \geqslant \sigma_{\min}(B)^2 \lambda(\sigma_{\min}(A)\delta e^{-\kappa R}) e^{-2\kappa R}$

Open Problems!

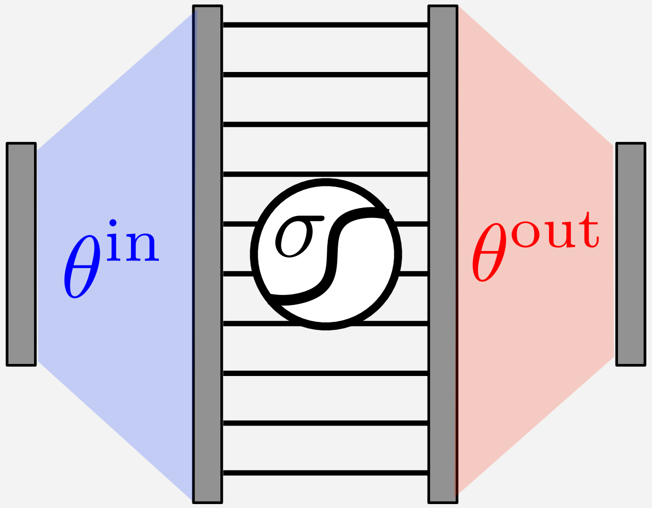
Training inner weights:

RKHS → Banach space (Barron)



Open Problems!

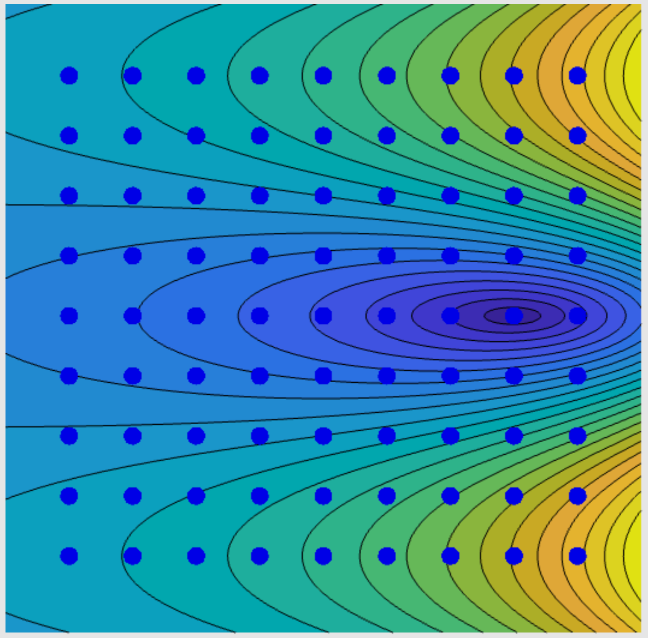
Training inner weights:



RKHS → Banach space (Barron)

Global convergence: (for generic initialization)

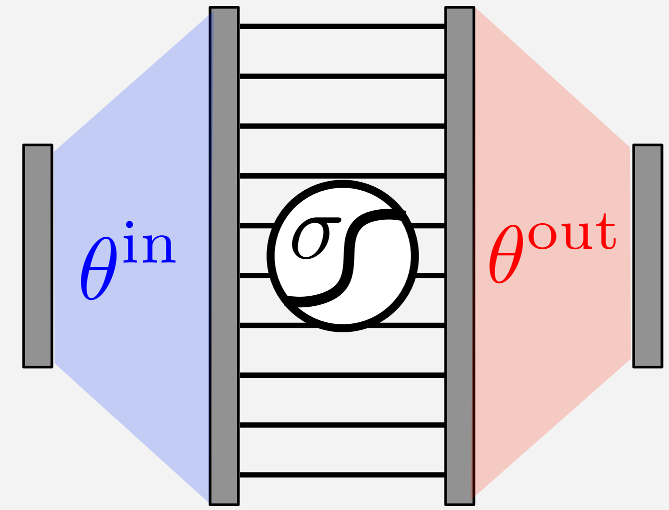
Open problem (even for linear networks)



Open Problems!

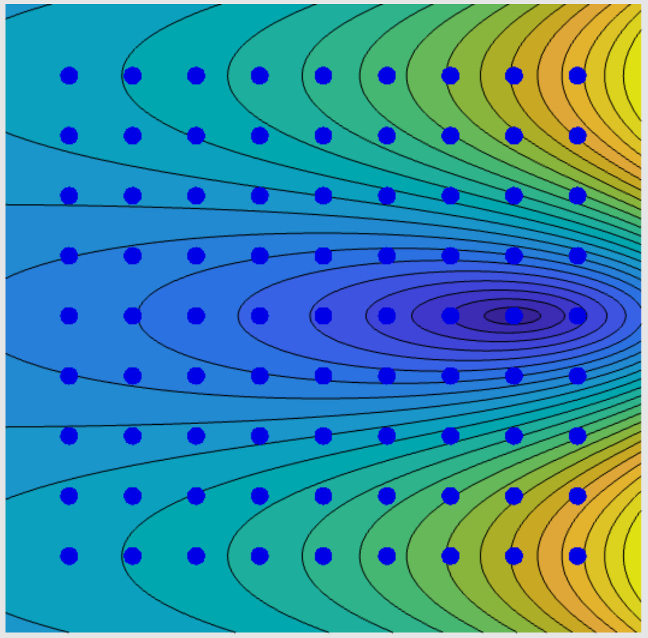
Training inner weights:

RKHS → Banach space (Barron)



Global convergence: (for generic initialization)

Open problem (even for linear networks)



Transformers architecture:

ODEs (single point) → Wasserstein flows (group of points)