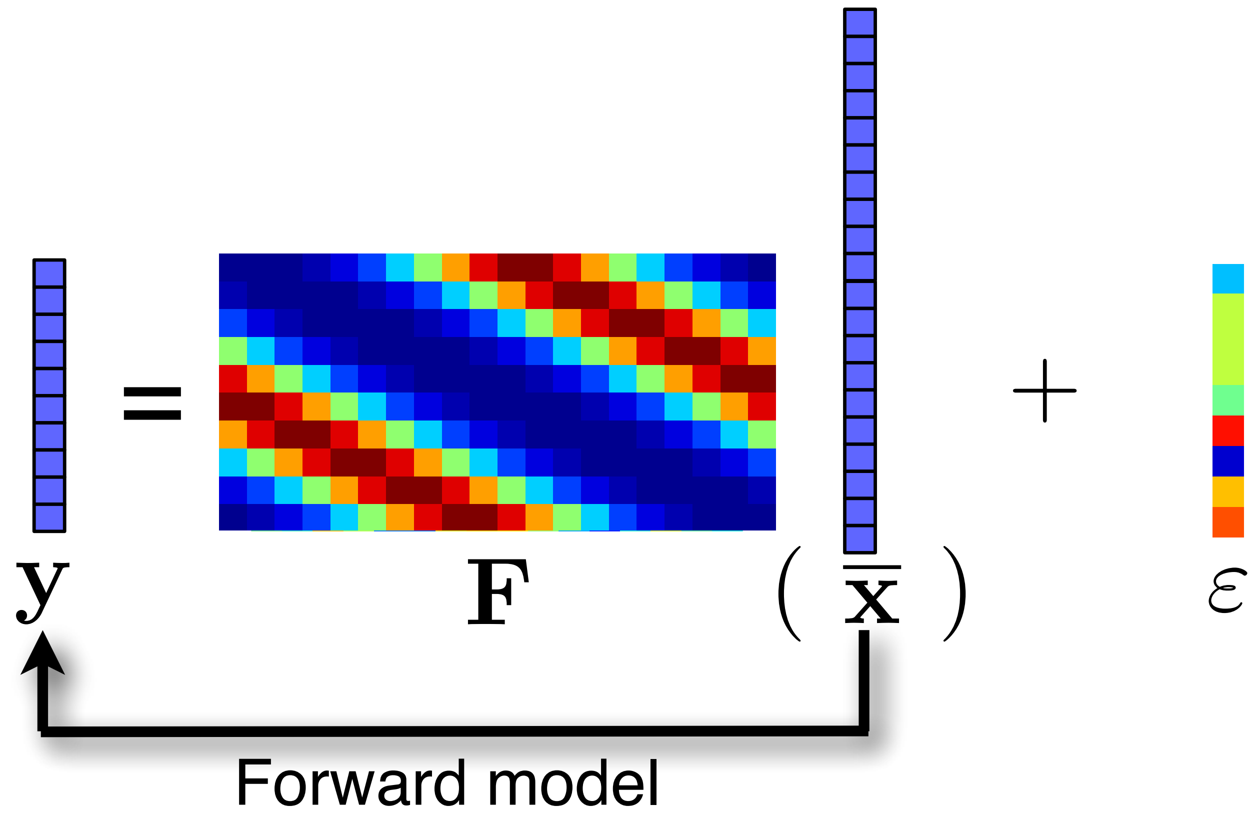
Inertial Algorithms Meet NN-Based Methods for Inverse Problems

Jalal Fadili
Normandie Université-ENSICAEN, CNRS

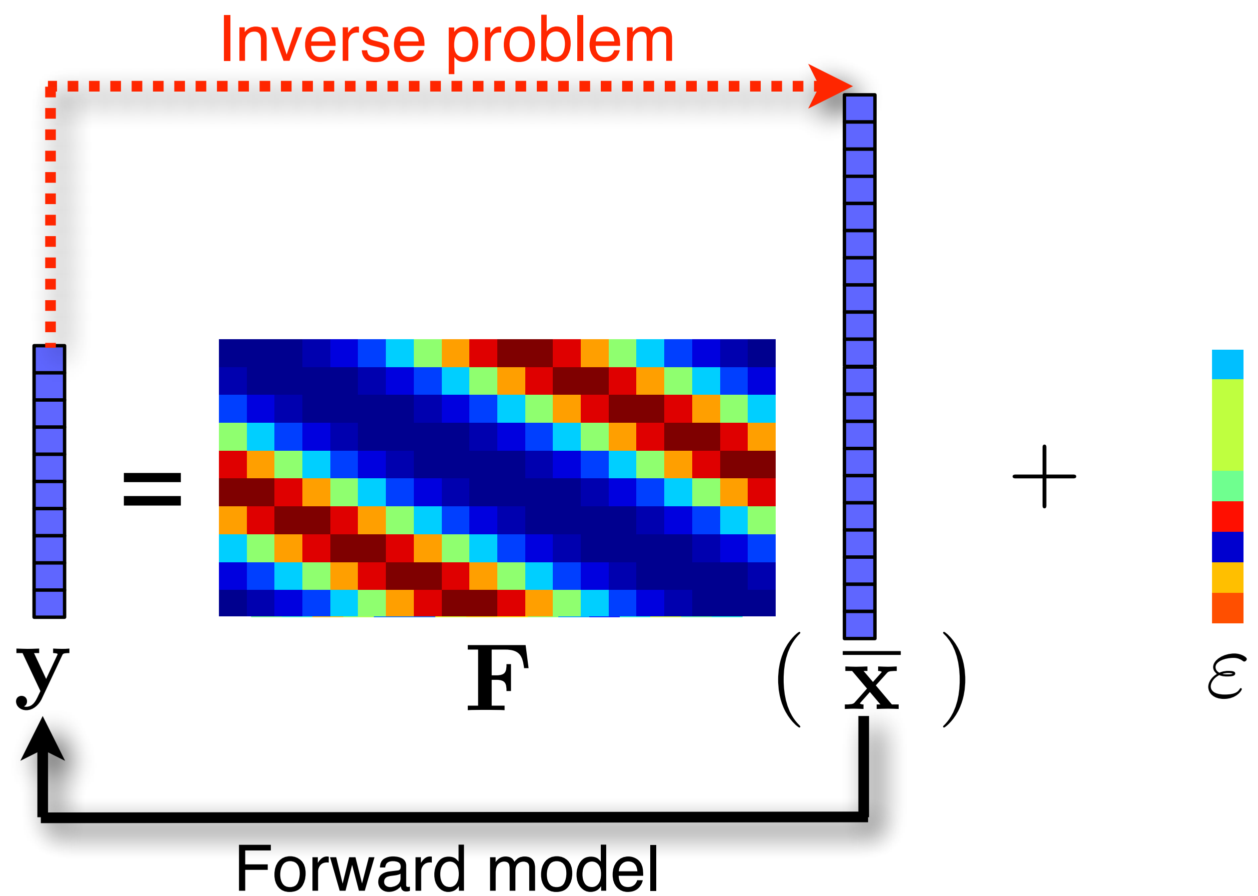
Joint ARGOS-TITAN-TOSCA workshop
7-8 July 2025
Join work with Rodrigo Maulen and Nathan Buskulic

Motivation



- Throughout the talk : finite-dimensional setting.
- $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^m$ is the forward operator (physics of the observation formation model).
- ε : noise.

Motivation

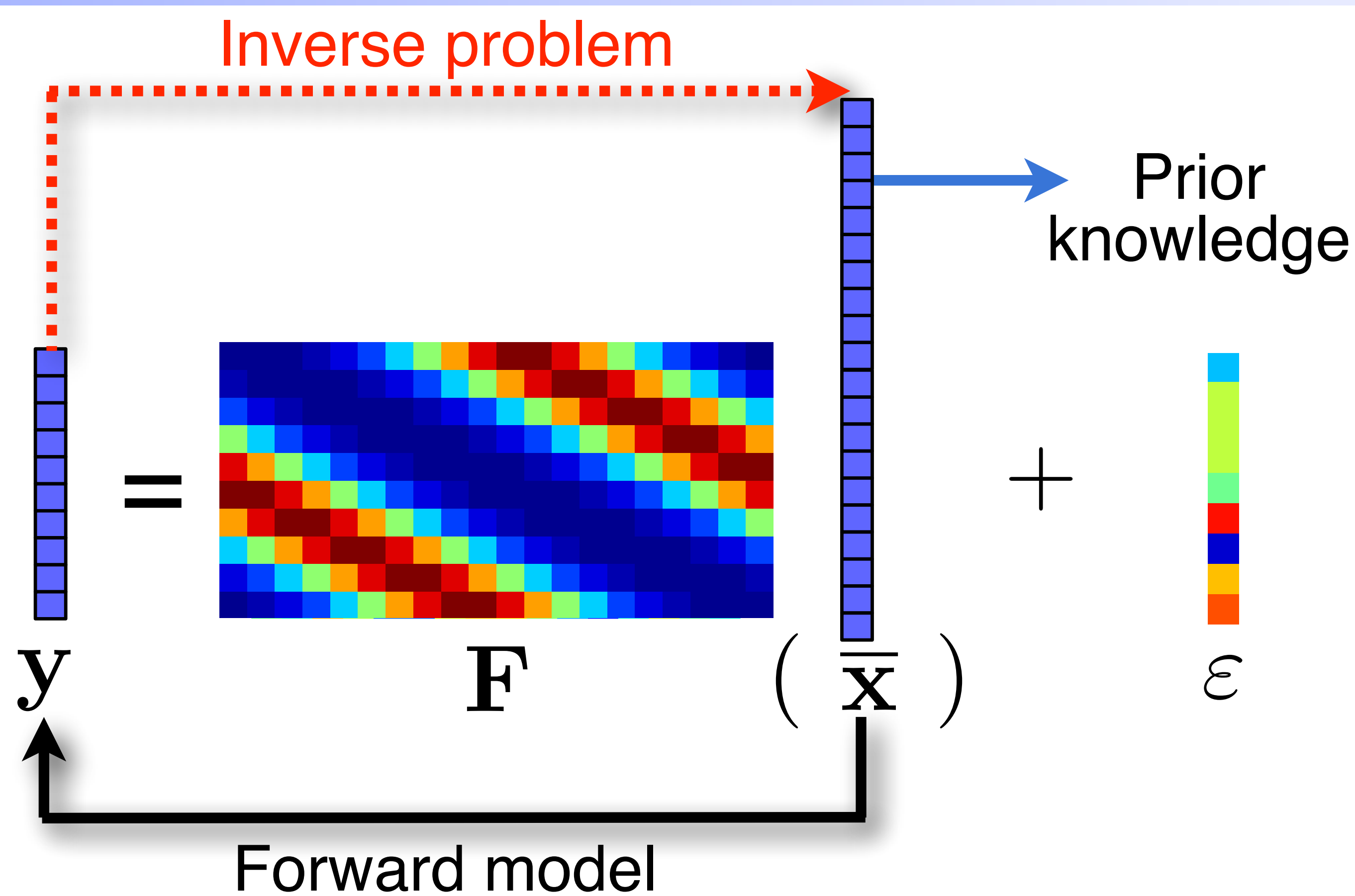


- Throughout the talk : finite-dimensional setting.
- $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^m$ is the forward operator (physics of the observation formation model).
- ε : noise.

Goal

Recover $\overline{\mathbf{x}}$ from $\mathbf{y}$ is generally an ill-posed inverse problem.

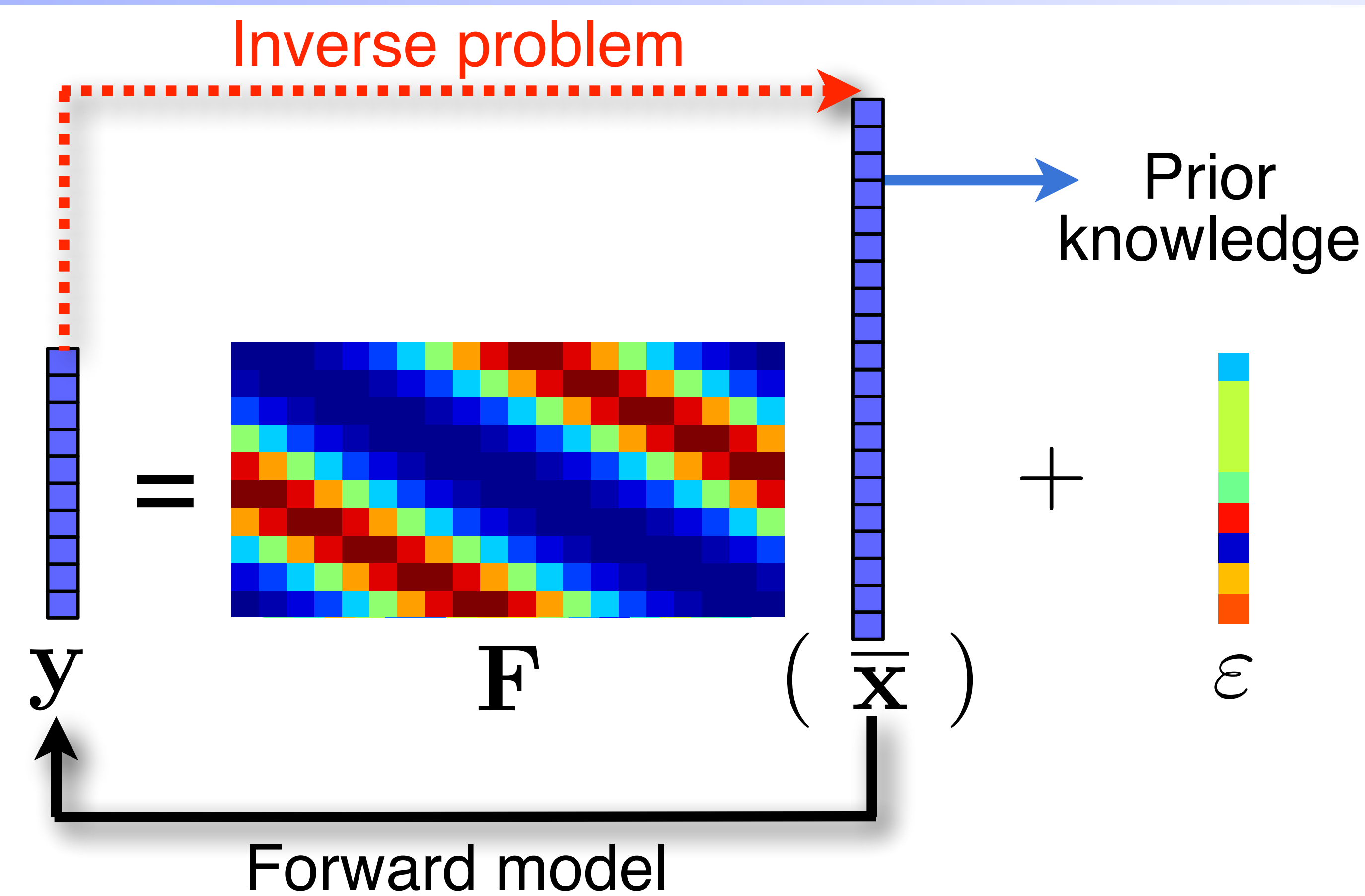
Model-based variational approach



- Solve :

$$\min_{\mathbf{x}\in\mathbb{R}^n} \underbrace{\mathcal{L}_{\mathbf{y}}(\mathbf{F}(\mathbf{x}))}_{\text{Data fidelity}} + \sum_{i=1}^{r} \underbrace{R_i(\mathbf{x})}_{\substack{\text{Model knowledge} \\ \text{Low complexity prior}}}$$

Model-based variational approach



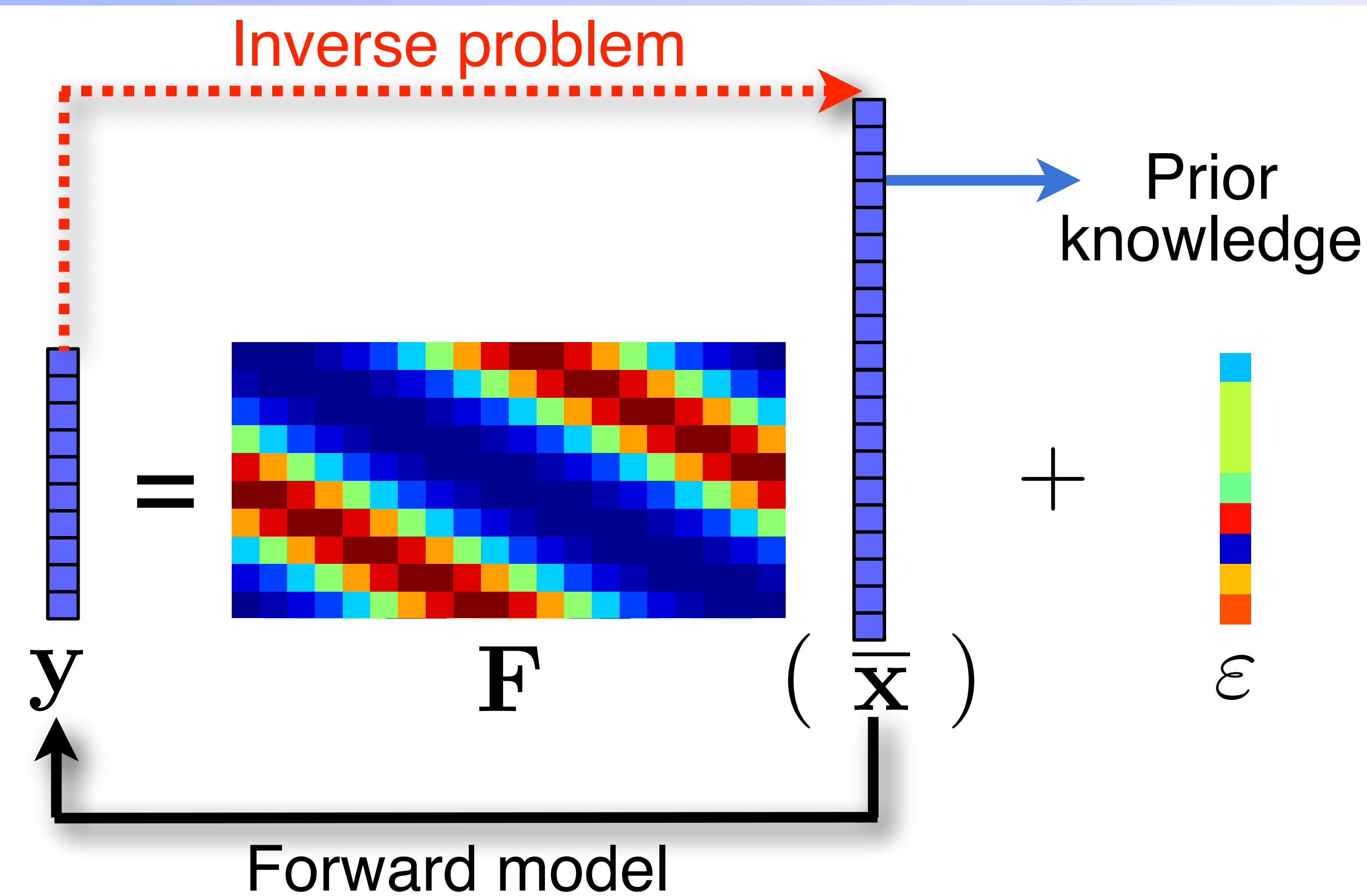
- Solve :

$$\min_{\mathbf{x}\in\mathbb{R}^n} \underbrace{\mathcal{L}_{\mathbf{y}}(\mathbf{F}(\mathbf{x}))}_{\text{Data fidelity}} + \sum_{i=1}^{r} \underbrace{R_i(\mathbf{x})}_{\text{Model knowledge}}$$

Pros

- Well-understood.
- Wealth of theoretical guarantees:
  - recovery: exact, stability.
  - algorithms.
  - explainability/interpretability.
  - etc.

Model-based variational approach



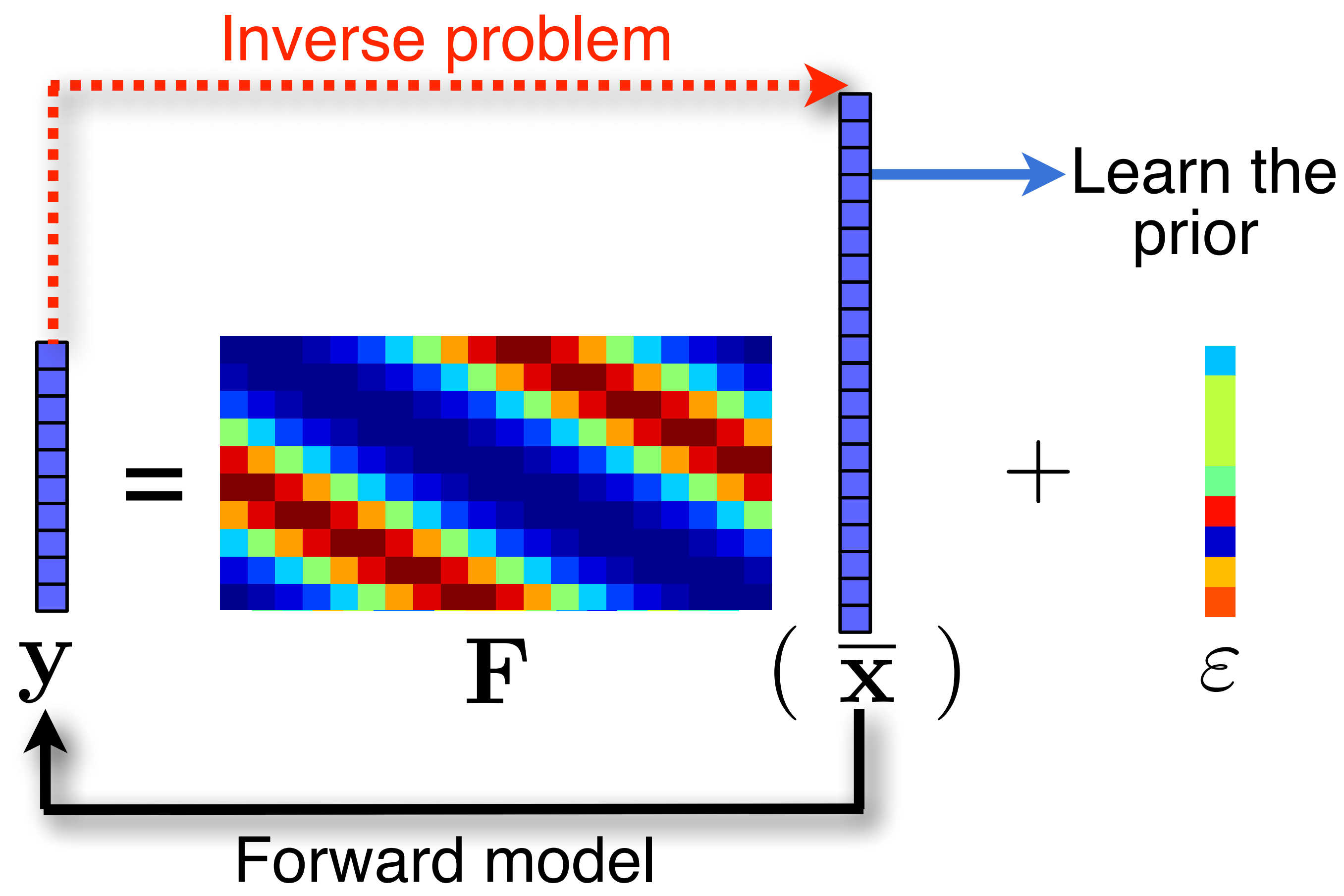
- Solve :

$$\min_{\mathbf{x}\in\mathbb{R}^n} \underbrace{\mathcal{L}_{\mathbf{y}}(\mathbf{F}(\mathbf{x}))}_{\text{Data fidelity}} + \sum_{i=1}^{r} \underbrace{R_i(\mathbf{x})}_{\text{Model knowledge}}$$

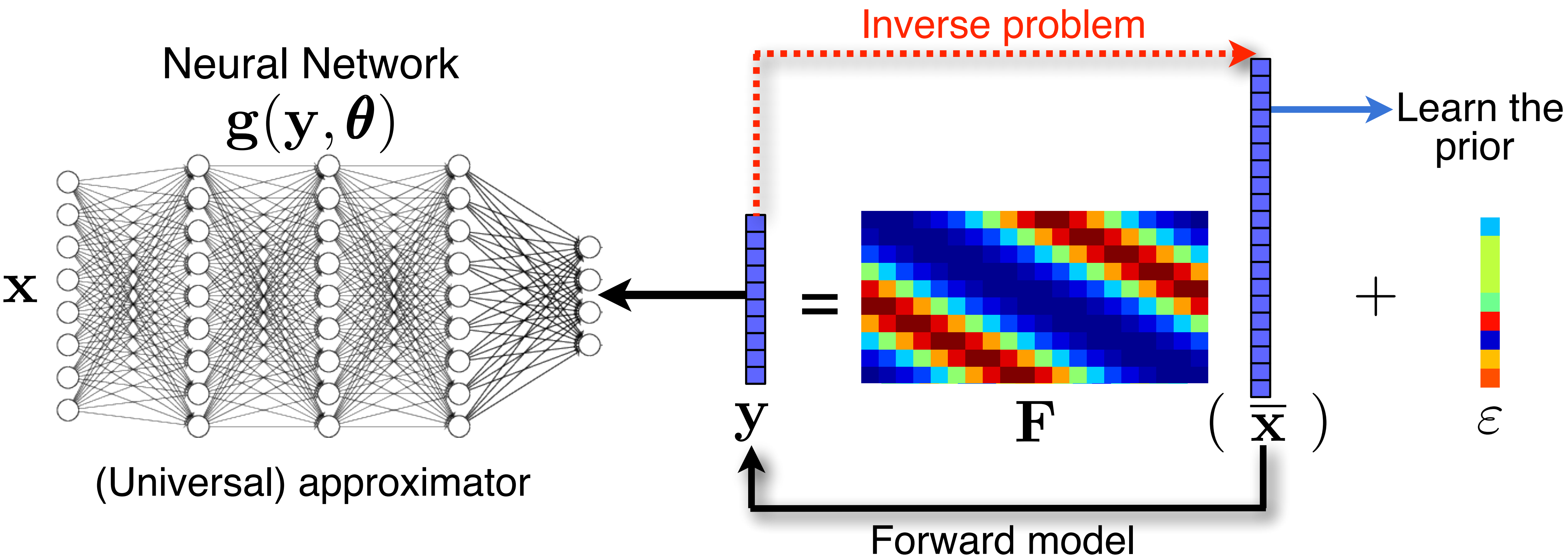
Pros	***Cons***

• Well-understood.
• Wealth of theoretical guarantees:
  ◦ recovery: exact, stability.
  ◦ algorithms.
  ◦ explainability/interpretability.
  ◦ etc.
	• Choice of the prior class not always easy.
	• Diversity and complexity of objects to recover.

Data-based: learning the inverse

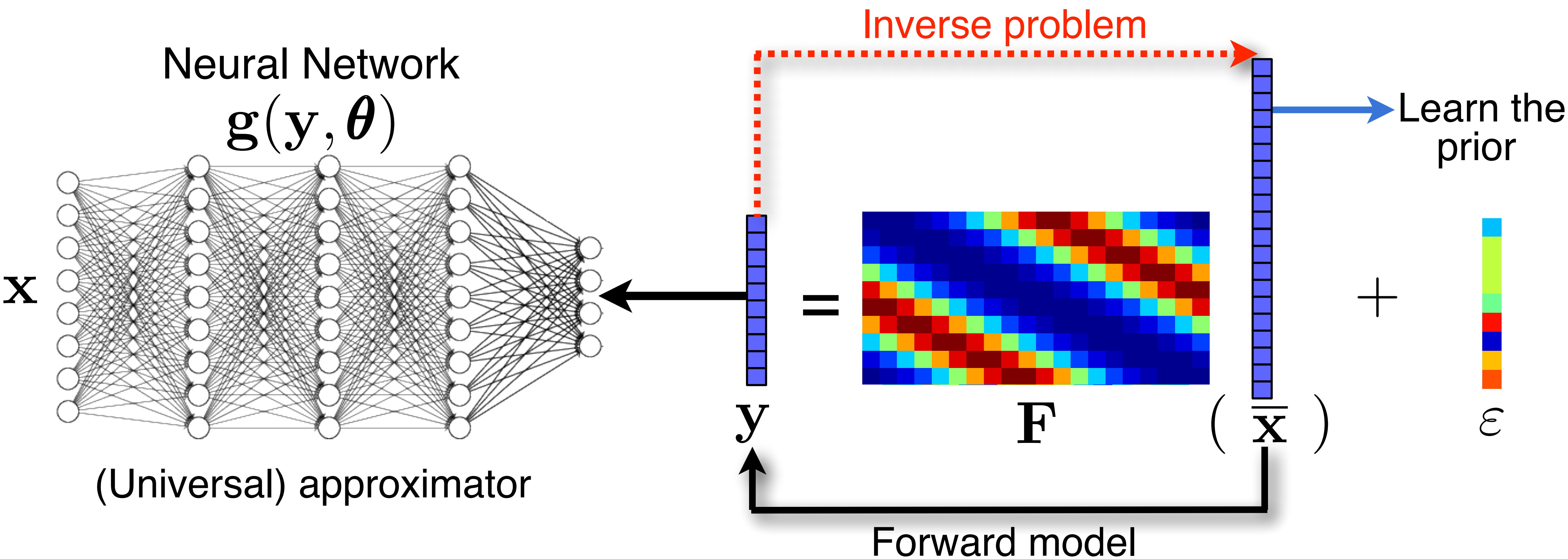


Data-based: learning the inverse



$$\min_{\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p} \frac{1}{N} \sum_{i=1}^{N} \ell(\mathbf{x}_i, \mathbf{g}(\mathbf{y}_i, \boldsymbol{\theta}))$$

Data-based: learning the inverse

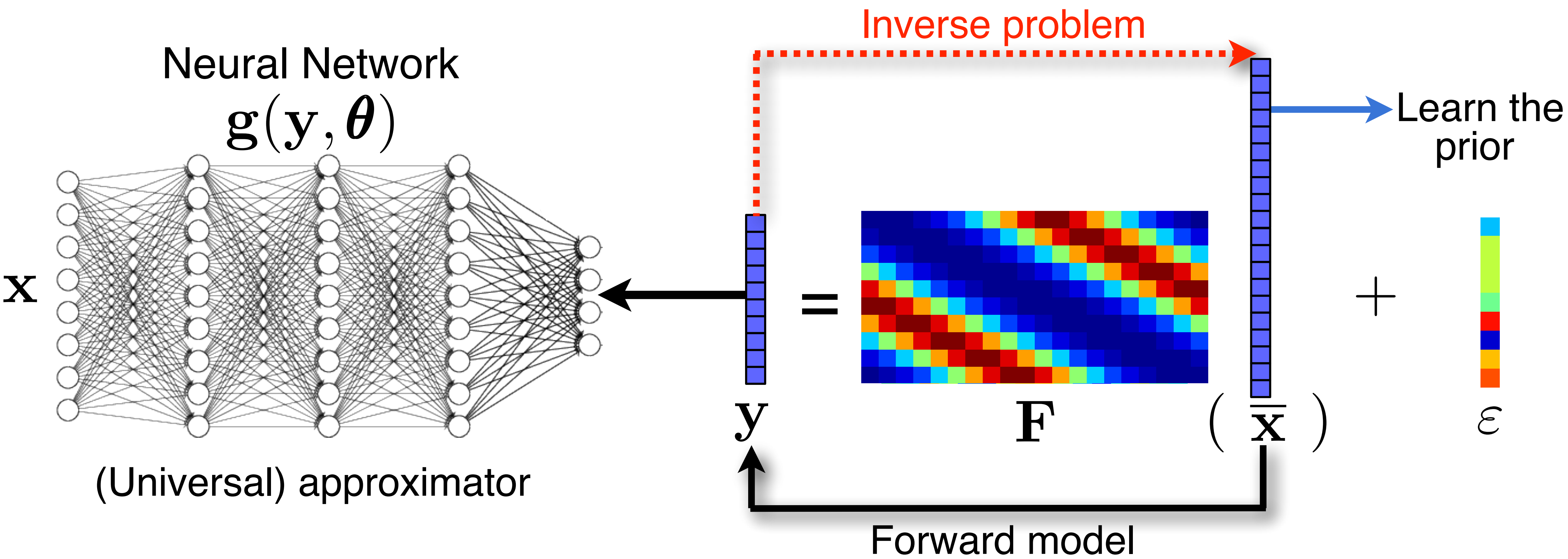


$$\min_{\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p} \frac{1}{N} \sum_{i=1}^{N} \ell(\mathbf{x}_i, \mathbf{g}(\mathbf{y}_i, \boldsymbol{\theta}))$$

Pros

- Off-the-shelf NN learning frameworks.
- No model to think about (… not quite so).
- Training once for all.

Data-based: learning the inverse



$$\min_{\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p} \frac{1}{N} \sum_{i=1}^{N} \ell(\mathbf{x}_i, \mathbf{g}(\mathbf{y}_i, \boldsymbol{\theta}))$$

Pros

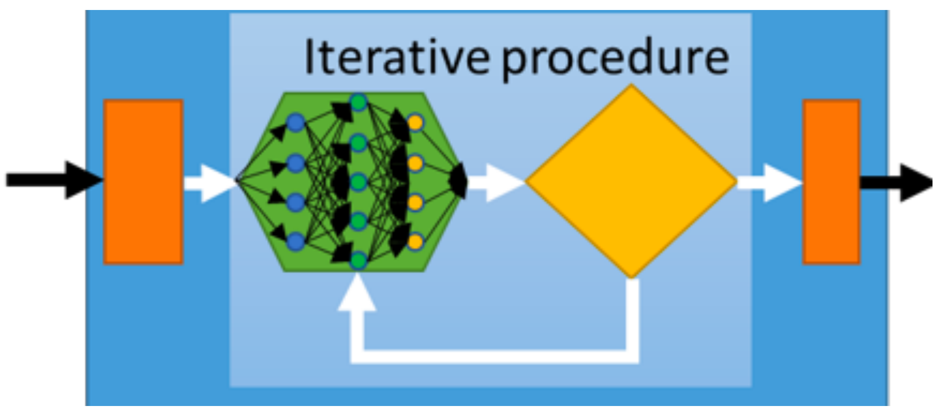
- Off-the-shelf NN learning frameworks.
- No model to think about (… not quite so).
- Training once for all.

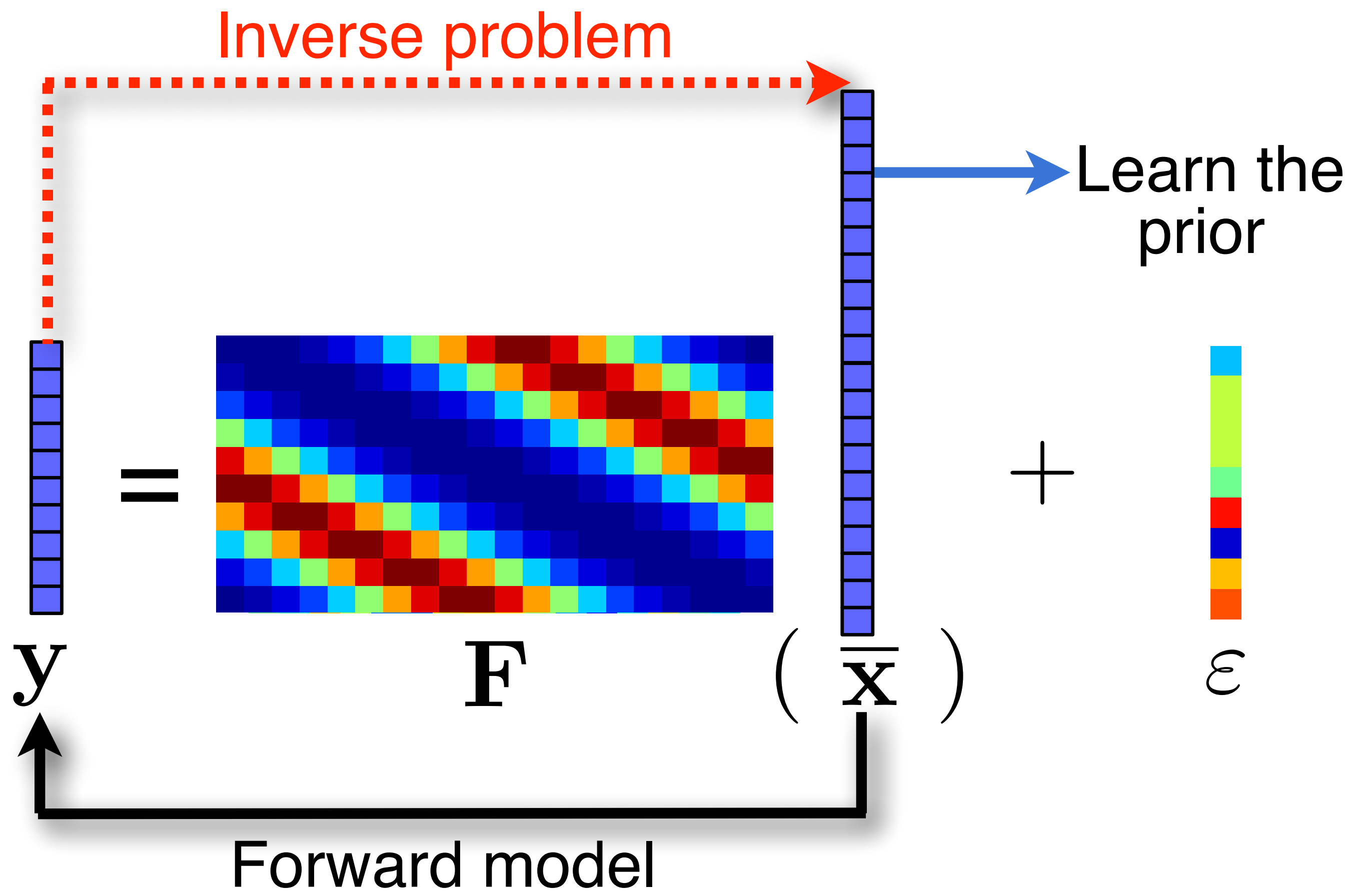
Cons

- Supervised: availability of training data.
- NN design (prior design is traded for NN design).
- No physical/forward model included.
- Guarantees from IP perspective: recovery, stability, explainability, etc.

Hybrid (model-based) learning

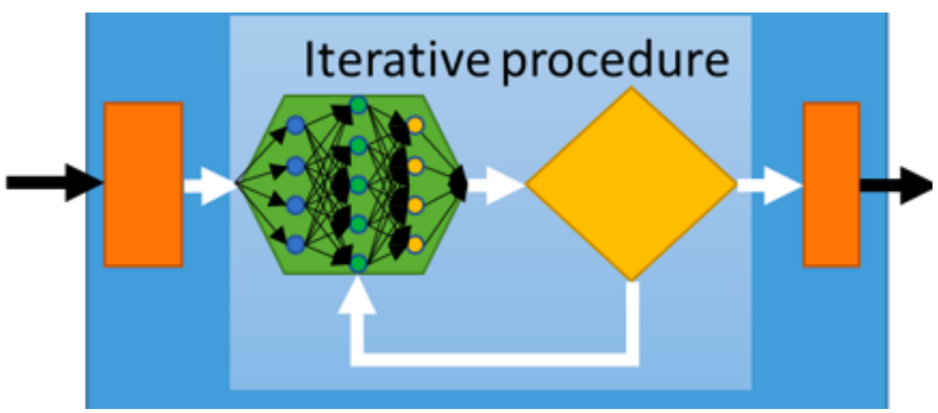
- Mix model- and data-driven methods in various ways: e.g.
  - Learn the regularizer.
  - Plug-and-Play.
  - Unrolling.
  - Deep equilibrium.
  - Generative models.
  - etc.
- An extremely active area, with extensive literature and reviews.

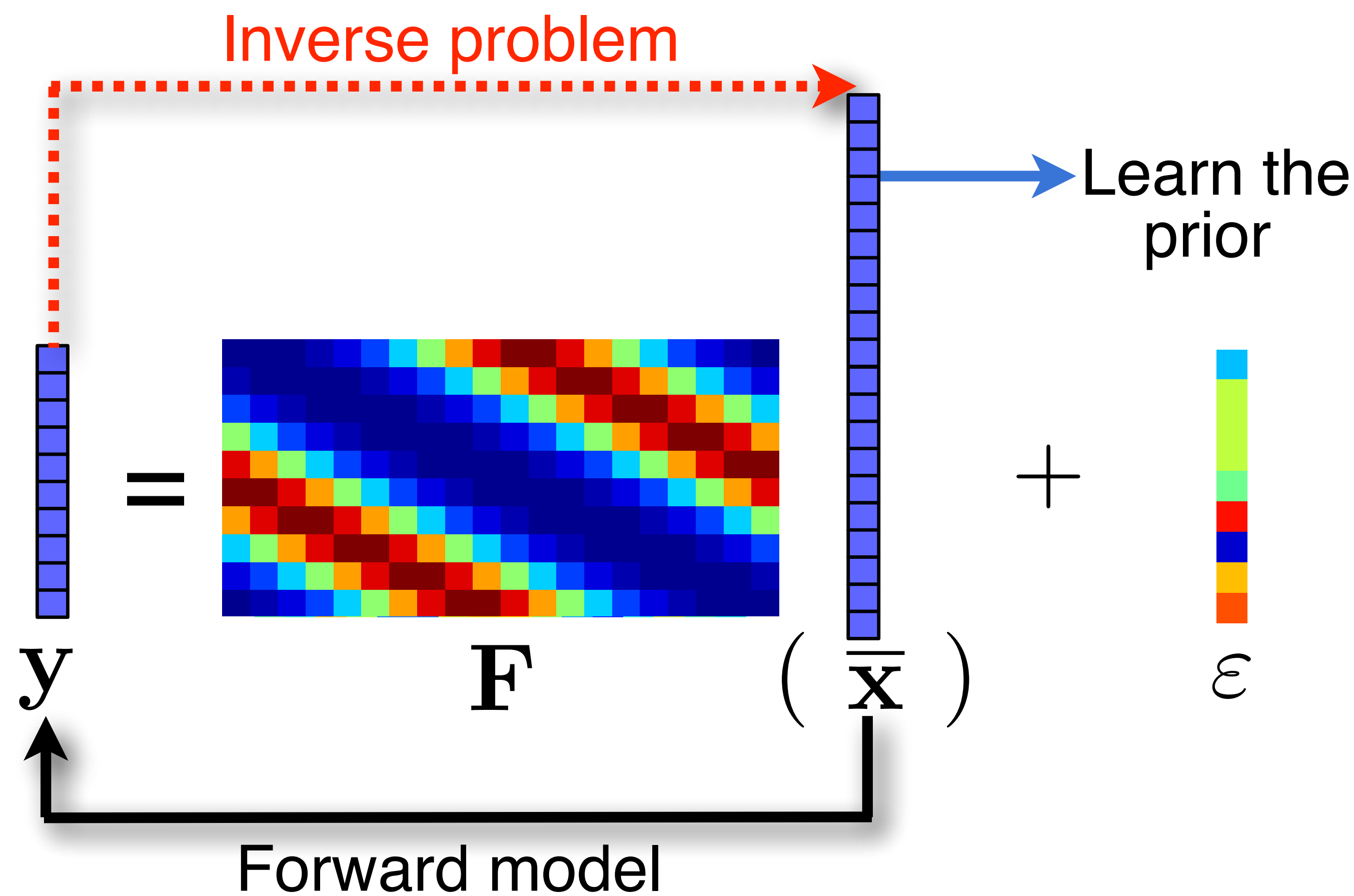




Hybrid (model-based) learning

- Mix model- and data-driven methods in various ways: e.g.
  - Learn the regularizer.
  - Plug-and-Play.
  - Unrolling.
  - Deep equilibrium.
  - Generative models.
  - etc.
- An extremely active area, with extensive literature and reviews.



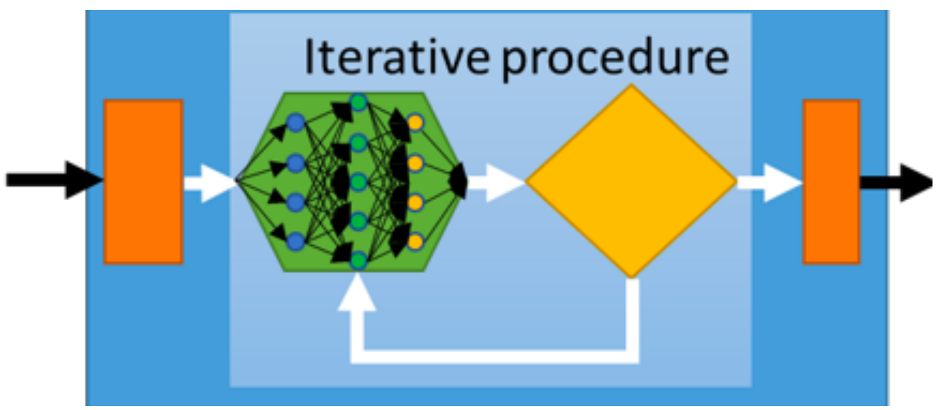


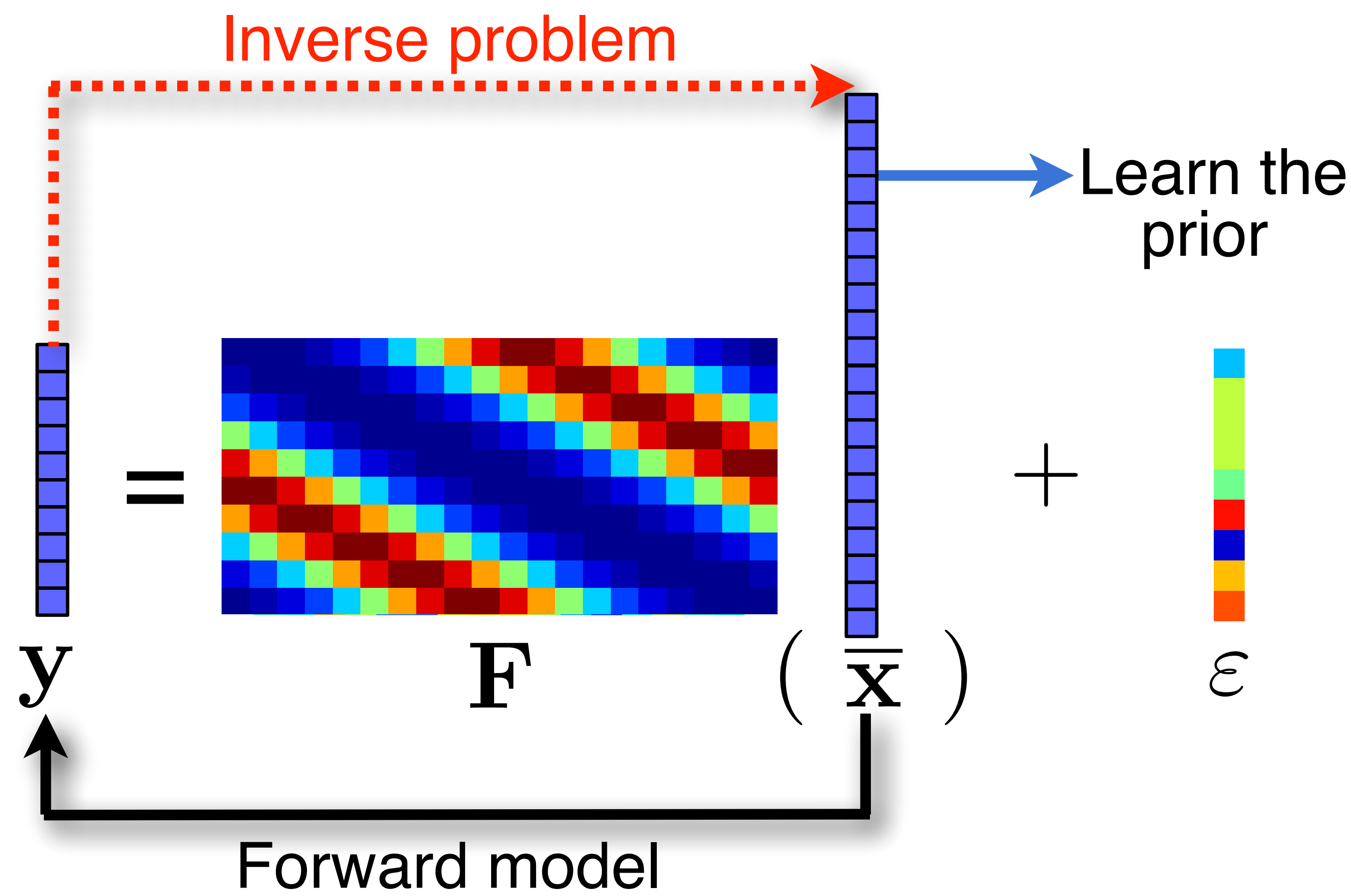
Pros

- Tries to get the best of both worlds.
- Accounts for the forward model.
- Prior learned explicitly/implicitly.
- Training once for all.
- Some guarantees: e.g. non-expansiveness/ Lipschitz constant in unrolling or PnP.

Hybrid (model-based) learning

- Mix model- and data-driven methods in various ways: e.g.
  - Learn the regularizer.
  - Plug-and-Play.
  - Unrolling.
  - Deep equilibrium.
  - Generative models.
  - etc.
- An extremely active area, with extensive literature and reviews.





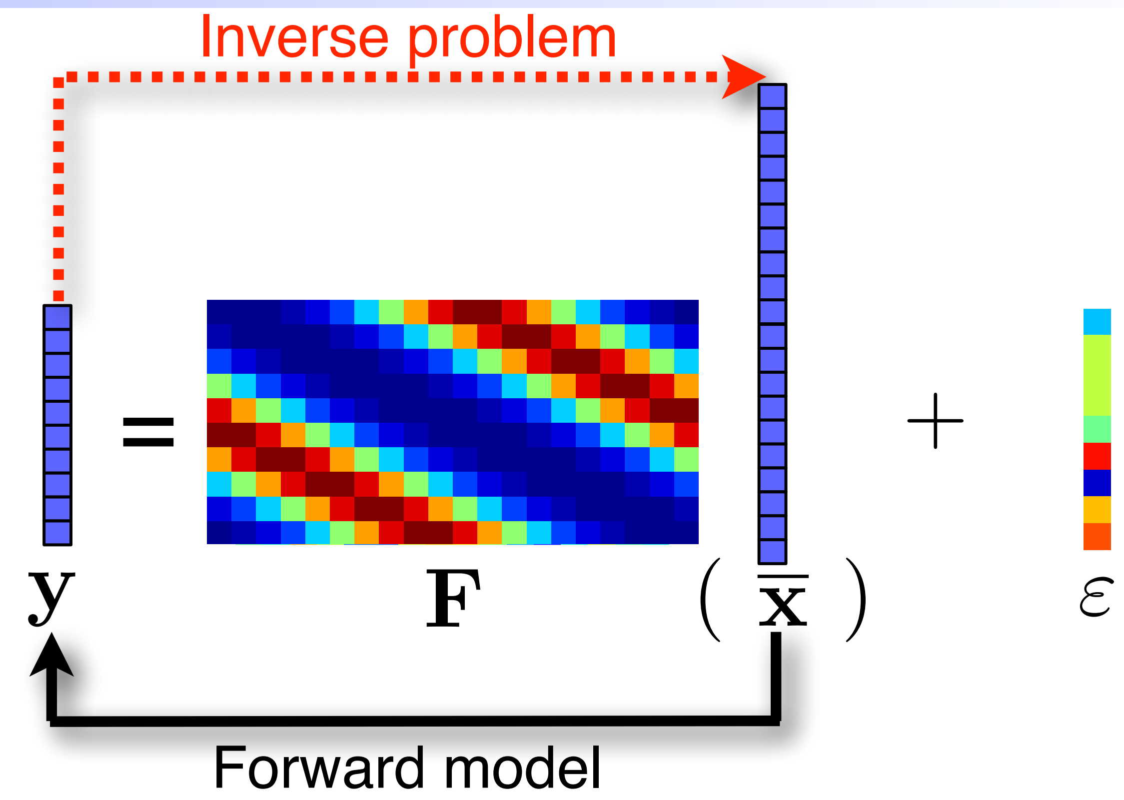
Pros

- Tries to get the best of both worlds.
- Accounts for the forward model.
- Prior learned explicitly/implicitly.
- Training once for all.
- Some guarantees: e.g. non-expansiveness/ Lipschitz constant in unrolling or PnP.

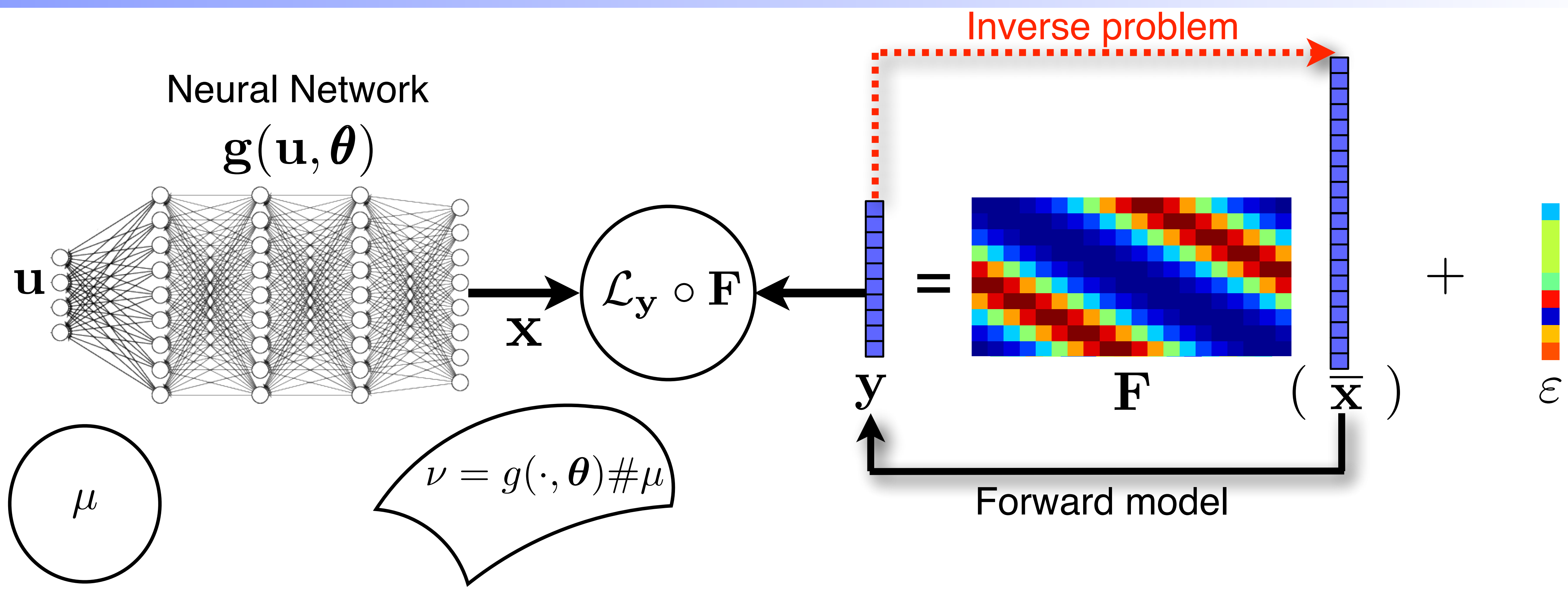
Cons

- Supervised: availability of training data.
- NN design (or even many NNs).
- Lack of guarantees from IP perspective: recovery, stability, explainability, etc.

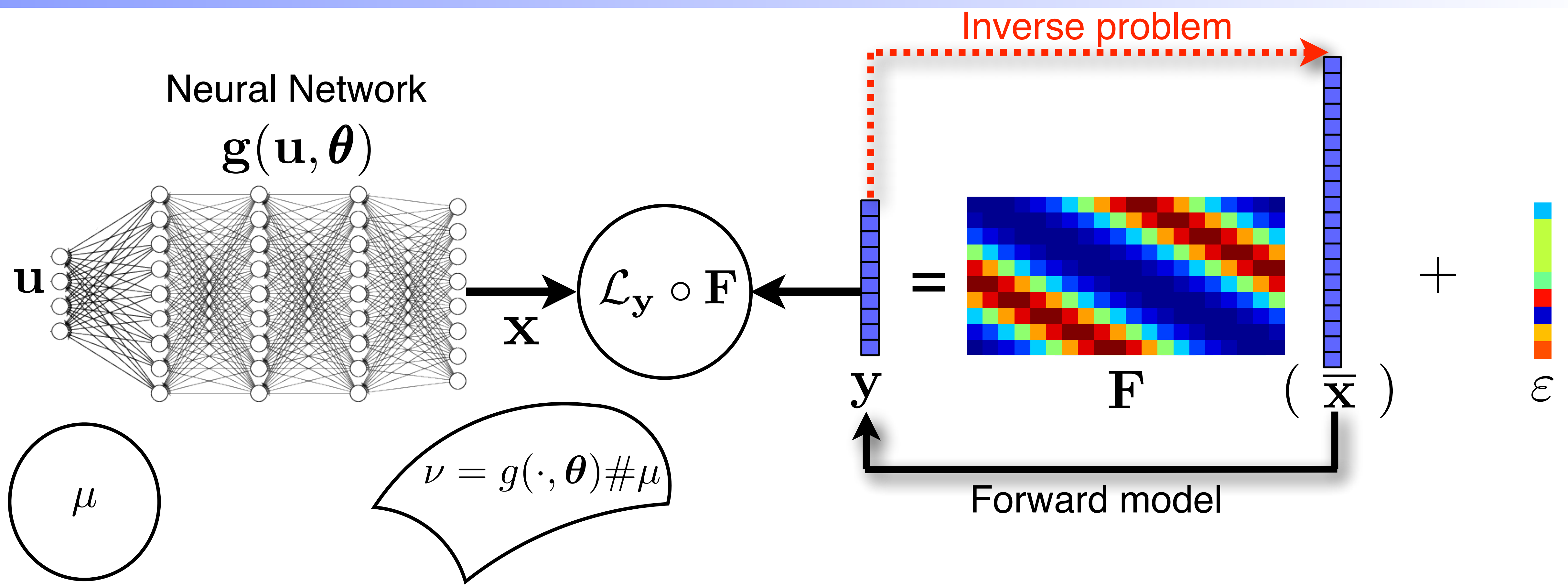
DIP: Deep Inverse/Image Prior



DIP: Deep Inverse/Image Prior



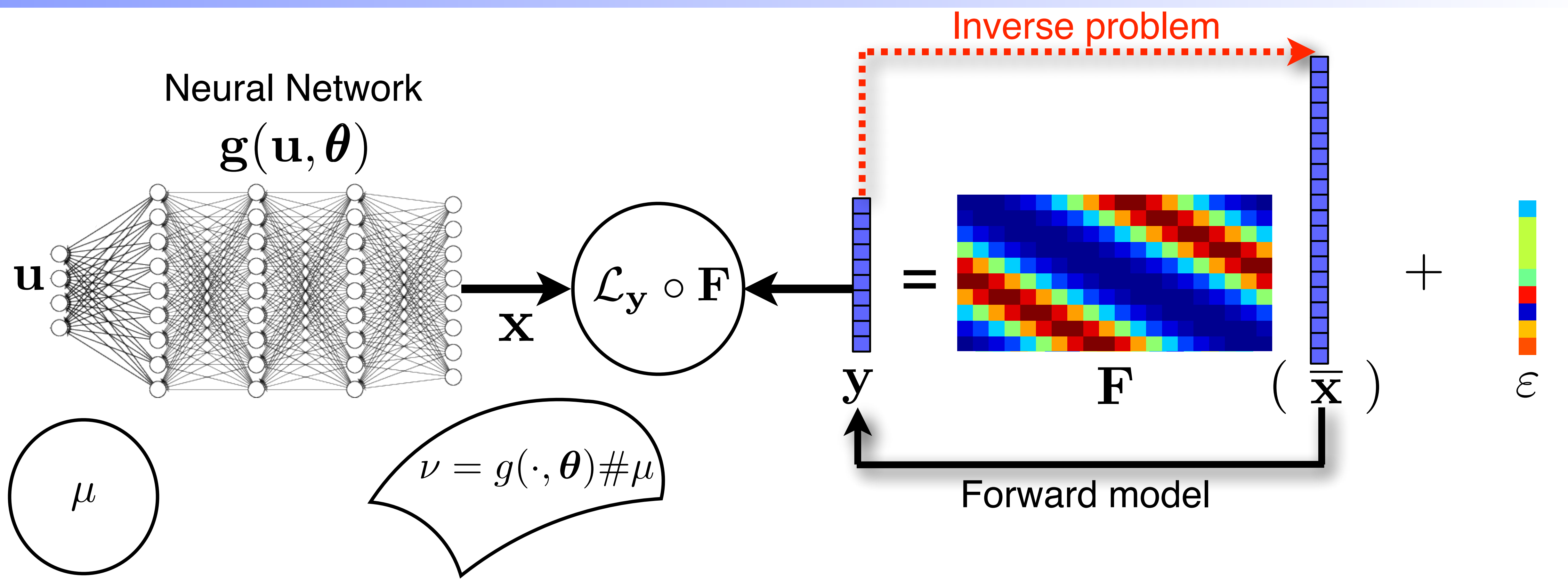
DIP: Deep Inverse/Image Prior



$$\min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}_{\mathbf{y}}(\mathbf{F}(\mathbf{x}_{\boldsymbol{\theta}})) \qquad \text{s.t.} \quad \mathbf{x}_{\boldsymbol{\theta}} \in \Sigma$$

$$\Sigma = \{\mathbf{g}(\mathbf{u},\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$$

DIP: Deep Inverse/Image Prior

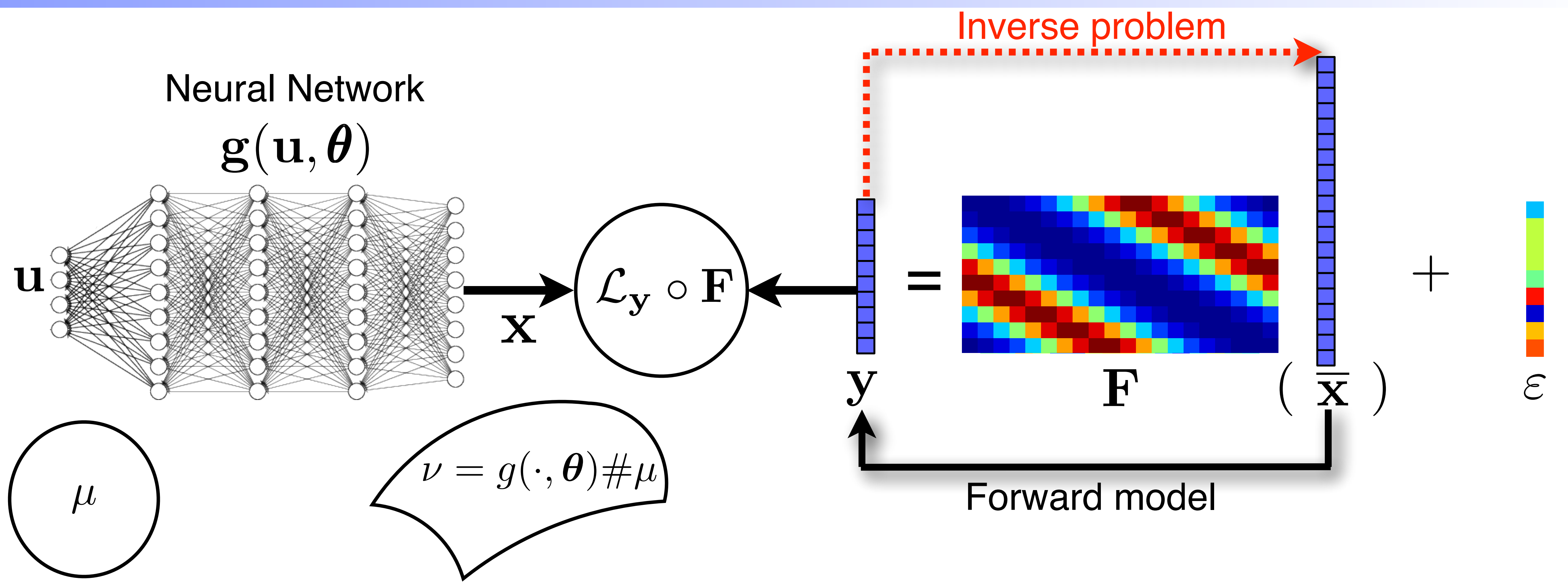


$$\min_{\boldsymbol{\theta} \in \Theta} \mathcal{L}_{\mathbf{y}}(\mathbf{F}(\mathbf{x}_{\boldsymbol{\theta}})) \qquad \text{s.t.} \quad \mathbf{x}_{\boldsymbol{\theta}} \in \Sigma$$

$$\Sigma = \{\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$$

- An unsupervised approach : generator from a latent variable $\mathbf{u} \sim \mu$.
- Hope for NN to induce "implicit regularization" and produce meaningful content before overfitting.
- A early stopping strategy for the NN to generate a vector close to $\overline{\mathbf{x}}$.

DIP: Deep Inverse/Image Prior



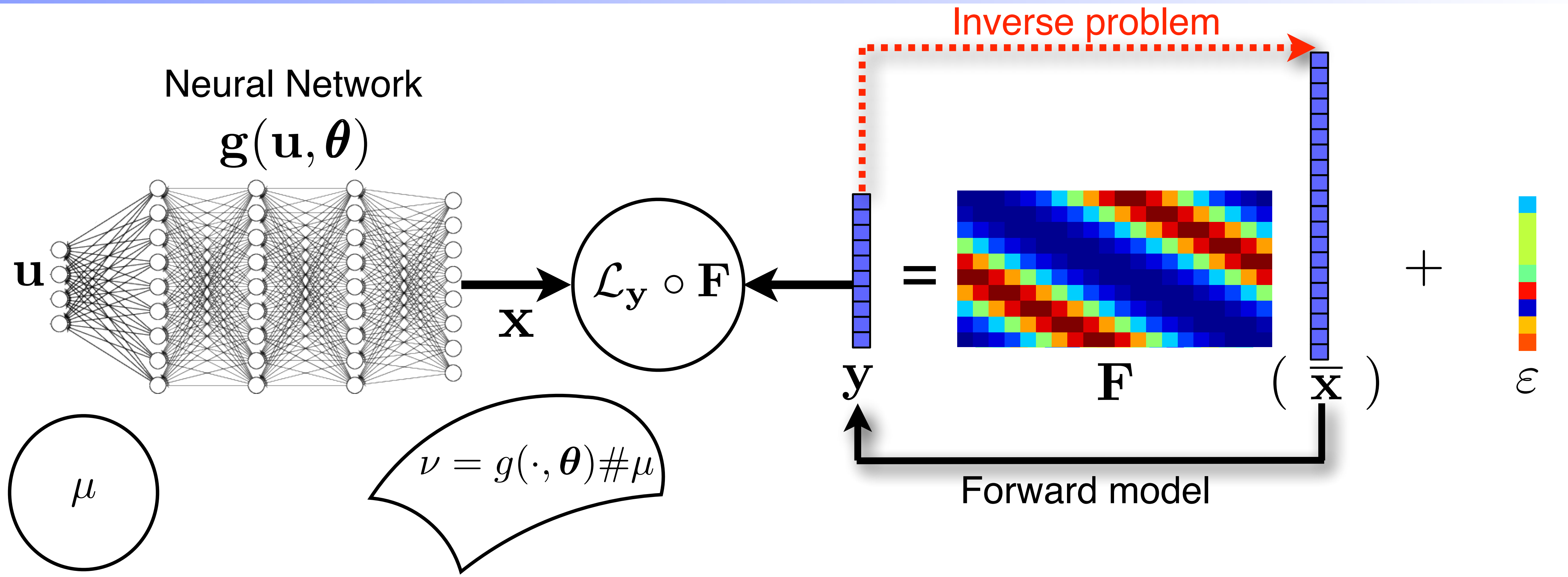
$$\min_{\boldsymbol{\theta} \in \Theta} \mathcal{L}_\mathbf{y}(\mathbf{F}(\mathbf{x}_{\boldsymbol{\theta}})) \qquad \text{s.t.} \quad \mathbf{x}_{\boldsymbol{\theta}} \in \Sigma$$

$$\Sigma = \{\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$$

Pros

- Unsupervised.
- Accounts for the forward model.
- Easy to implement with (very) good empirical success.

DIP: Deep Inverse/Image Prior



$$\min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}_{\mathbf{y}}(\mathbf{F}(\mathbf{x}_{\boldsymbol{\theta}})) \qquad \text{s.t.} \quad \mathbf{x}_{\boldsymbol{\theta}} \in \Sigma$$

$$\Sigma = \{\mathbf{g}(\mathbf{u},\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$$

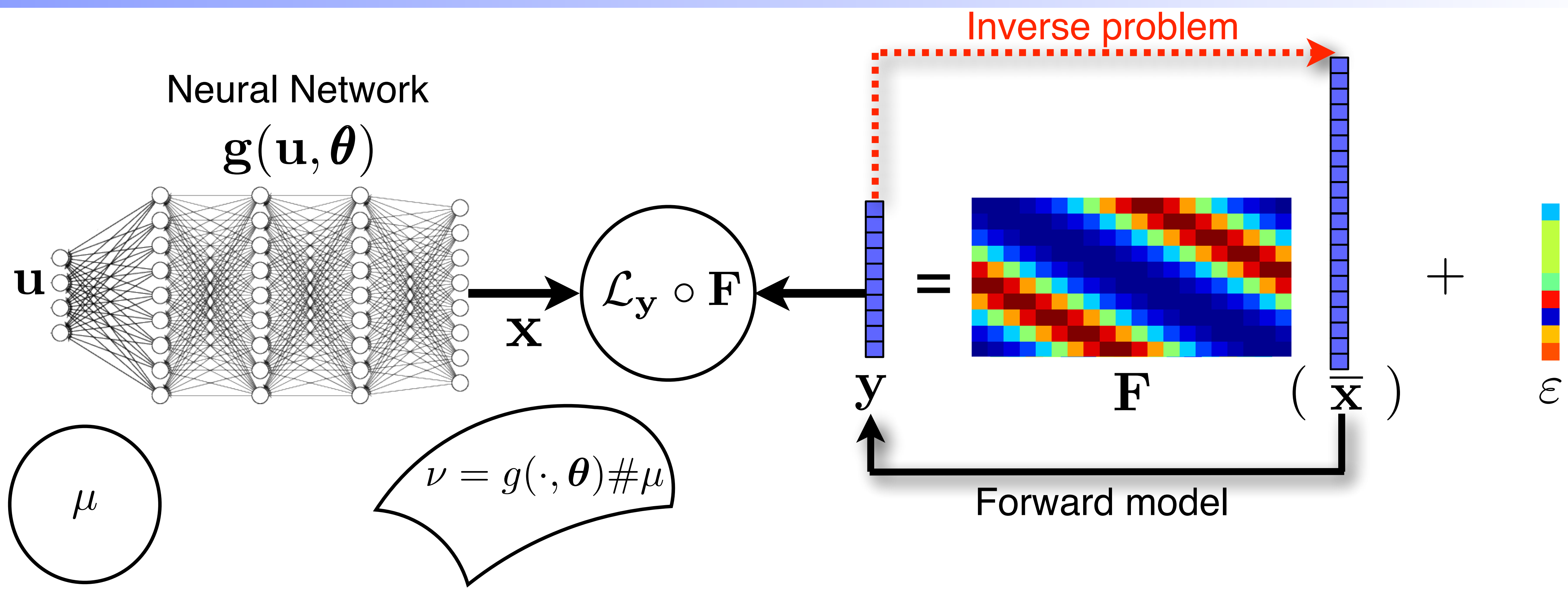
Pros

- Unsupervised.
- Accounts for the forward model.
- Easy to implement with (very) good empirical success.

Cons

- Optimize/train for each signal to recover.
- No theoretical guarantees: recovery, stability, NN design.

DIP: Deep Inverse/Image Prior



$$\min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}_{\mathbf{y}}(\mathbf{F}(\mathbf{x}_{\boldsymbol{\theta}})) \qquad \text{s.t.} \quad \mathbf{x}_{\boldsymbol{\theta}} \in \Sigma$$

$$\Sigma = \{\mathbf{g}(\mathbf{u},\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$$

Pros

- Unsupervised.
- Accounts for the forward model.
- Easy to implement with (very) good empirical success.

Cons

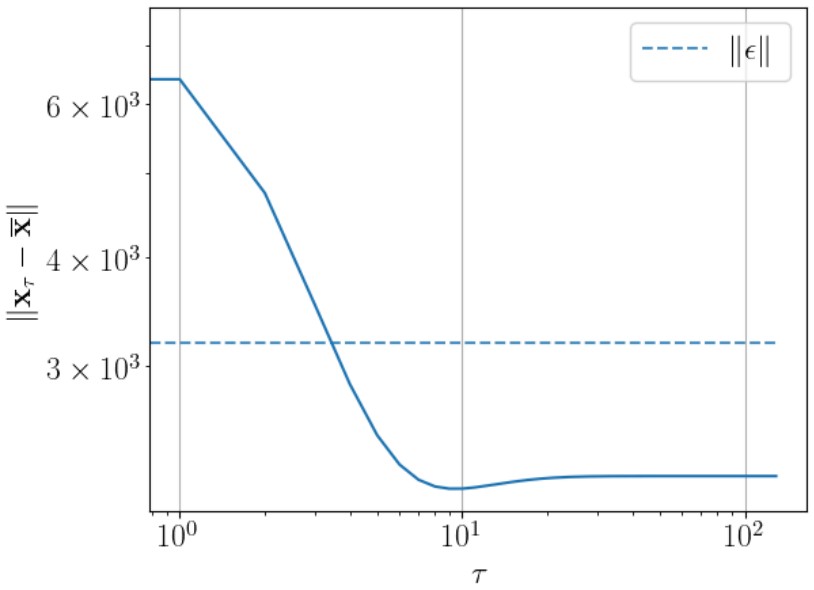
- Optimize/train for each signal to recover.
- No theoretical guarantees: recovery, stability, NN design.

In the rest of the talk, linear forward operator

Example: Image deblurring

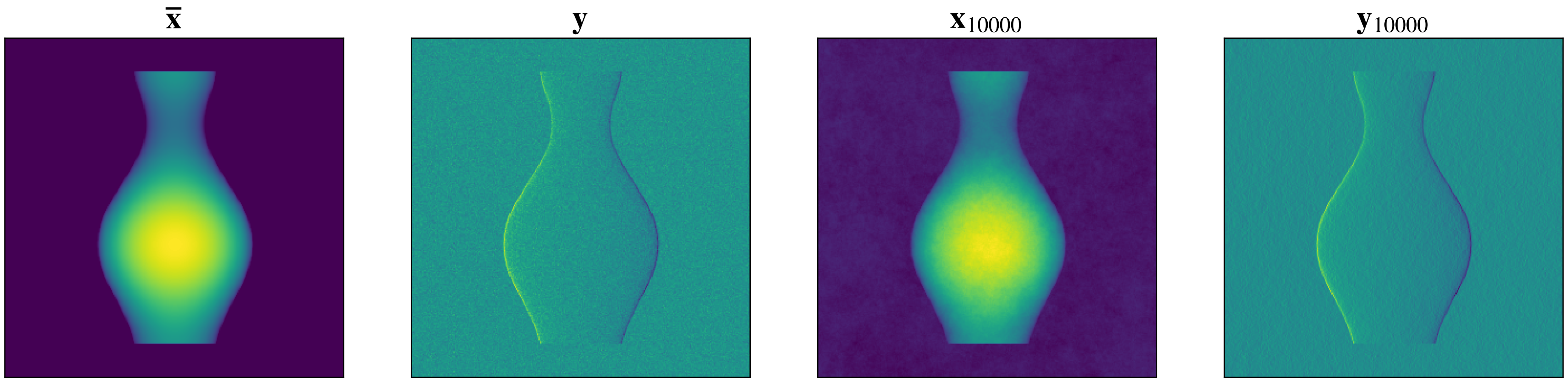
$$\mathbf{y} = \mathbf{A}\overline{\mathbf{x}} + \varepsilon \qquad \varepsilon \sim \mathcal{N}(0, 50^2)$$

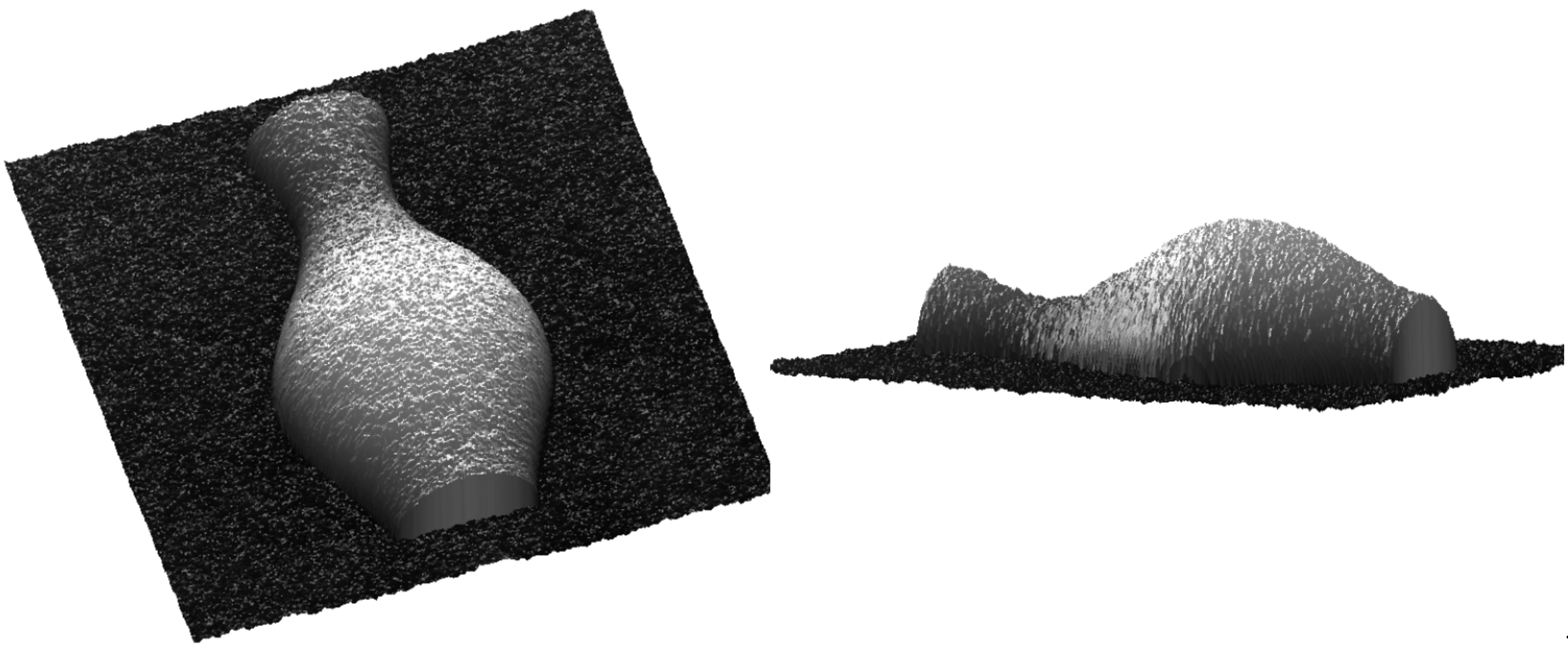
Early stopping



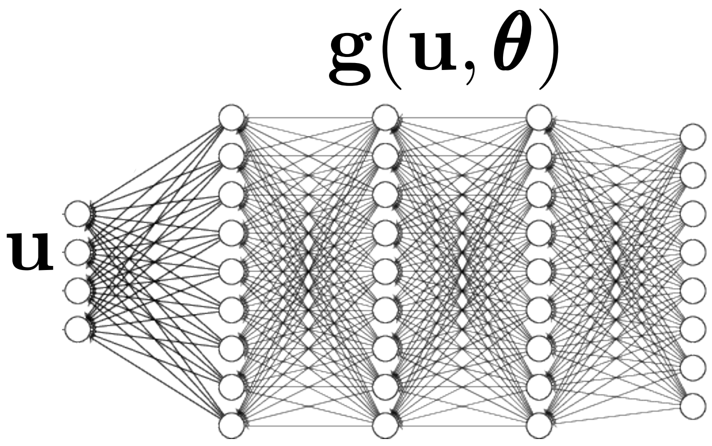
Example: Normal integration

$$\mathbf{y} = \nabla_{\text{diff}} \overline{\mathbf{x}} + \varepsilon \qquad \varepsilon \sim \mathcal{N}(0, 1.5)$$



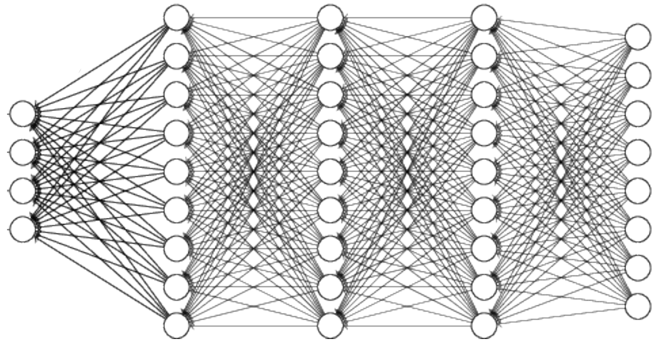


DIP training with inertia



$$\min_{\boldsymbol{\theta} \in \Theta} \mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta})) \qquad \mathbf{A} \in \mathbb{R}^{m \times n}$$

DIP training with inertia

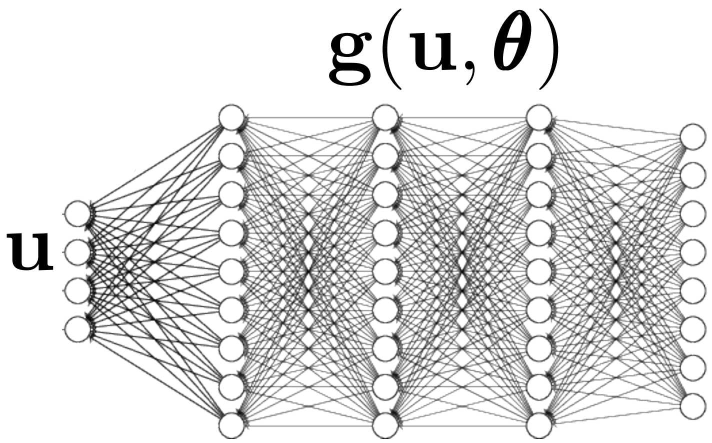


$$\min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta})) \qquad \mathbf{A}\in\mathbb{R}^{m\times n}$$

$$\text{(ISEHD)}\begin{cases}\ddot{\boldsymbol{\theta}}(t)+\alpha\dot{\boldsymbol{\theta}}(t)+\beta\frac{\mathrm{d}}{\mathrm{d}t}\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(t)))+\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(t)))=0\\ \boldsymbol{\theta}(0)=\boldsymbol{\theta}_0,\dot{\boldsymbol{\theta}}(0)=0.\end{cases}$$

$$\text{(IGAHD)}\begin{cases}\boldsymbol{\eta}_\ell &=\boldsymbol{\theta}_\ell+(1-\alpha\sqrt{s_\ell})(\boldsymbol{\theta}_\ell-\boldsymbol{\theta}_{\ell-1})-\beta\sqrt{s_\ell}\left(\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}_\ell))-\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}_{\ell-1}))\right),\\ \boldsymbol{\theta}_{\ell+1} &=\boldsymbol{\eta}_\ell-s_\ell\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}_\ell)).\end{cases}$$

DIP training with inertia



$$\min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta})) \qquad \mathbf{A}\in\mathbb{R}^{m\times n}$$

$$\text{(ISEHD)}\begin{cases}\ddot{\boldsymbol{\theta}}(t)+\alpha\dot{\boldsymbol{\theta}}(t)+\beta\frac{\mathrm{d}}{\mathrm{d}t}\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(t)))+\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(t)))=0\\ \boldsymbol{\theta}(0)=\boldsymbol{\theta}_0,\dot{\boldsymbol{\theta}}(0)=0.\end{cases}$$

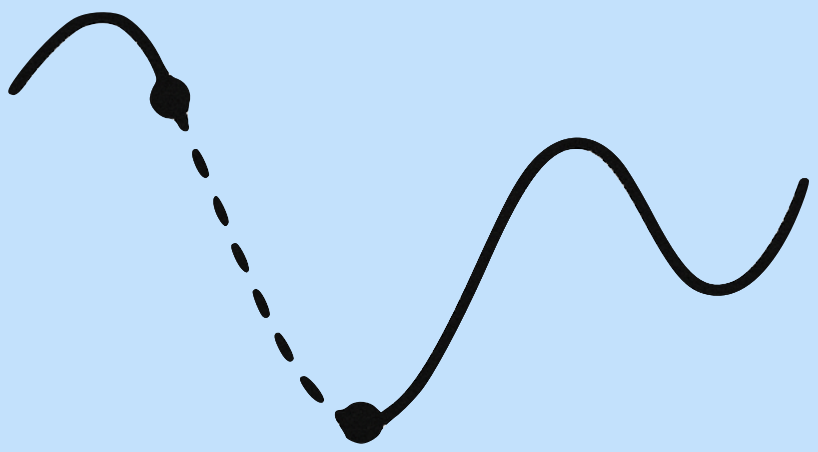
$$\text{(IGAHD)}\begin{cases}\boldsymbol{\eta}_\ell &=\boldsymbol{\theta}_\ell+(1-\alpha\sqrt{s_\ell})(\boldsymbol{\theta}_\ell-\boldsymbol{\theta}_{\ell-1})-\beta\sqrt{s_\ell}\left(\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}_\ell))-\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}_{\ell-1}))\right),\\ \boldsymbol{\theta}_{\ell+1} &=\boldsymbol{\eta}_\ell-s_\ell\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}_\ell)).\end{cases}$$

- Recovery guarantees of DIP when optimized with inertial gradient descent in :
  - Observation space : convergence to zero-loss $\Rightarrow$ implicit regularization.
  - Object space : restricted injectivity of the forward operator on Σ.
- NN architecture : role of overparametrization.

Outline

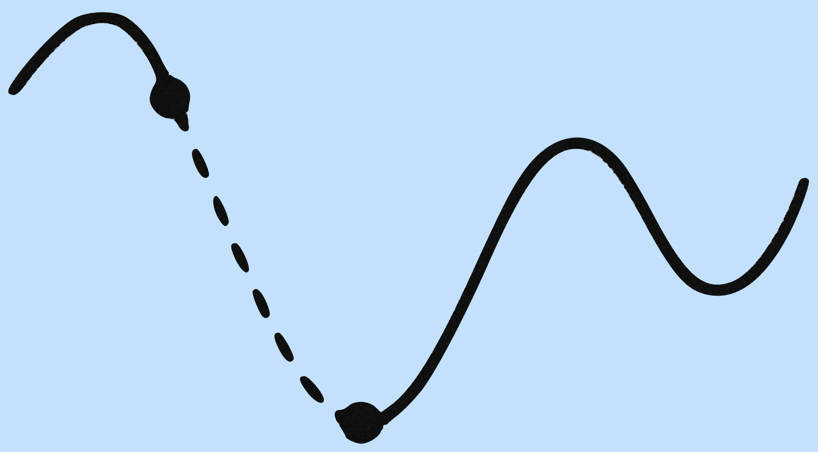
Outline

Convergence

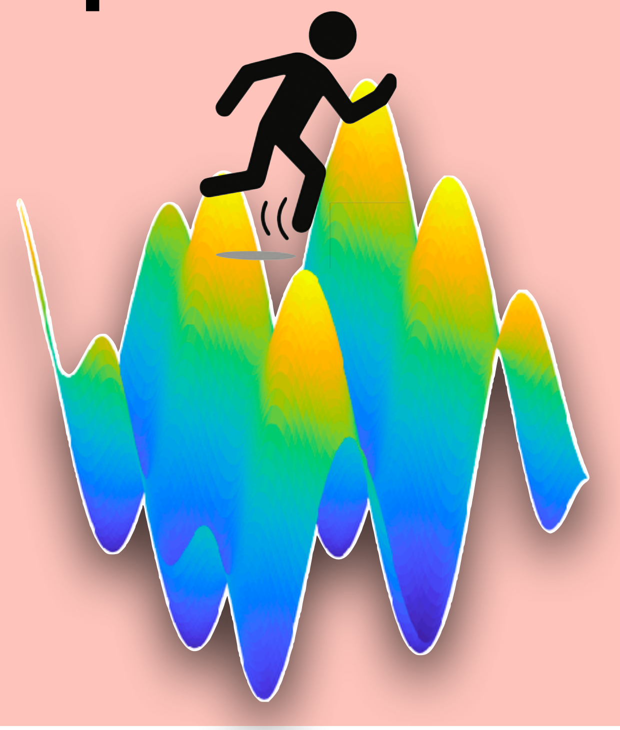


Outline

Convergence

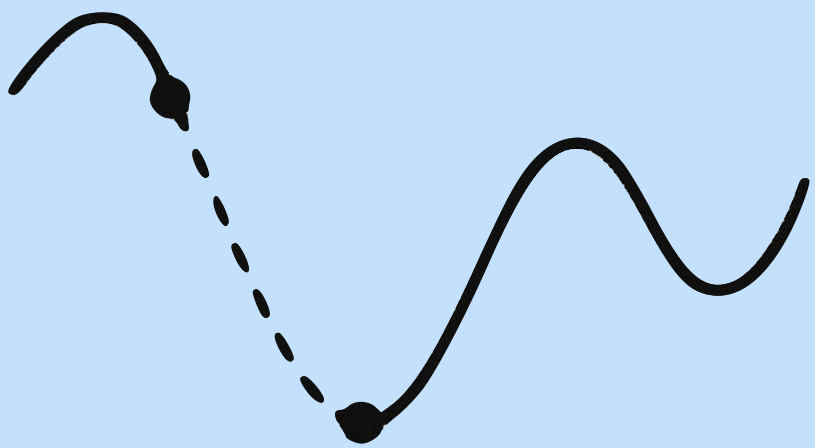


Trap avoidance

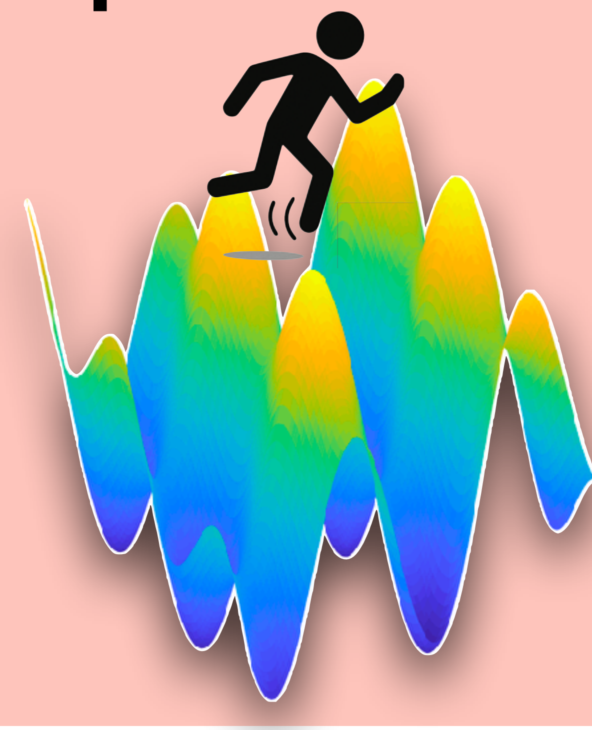


Outline

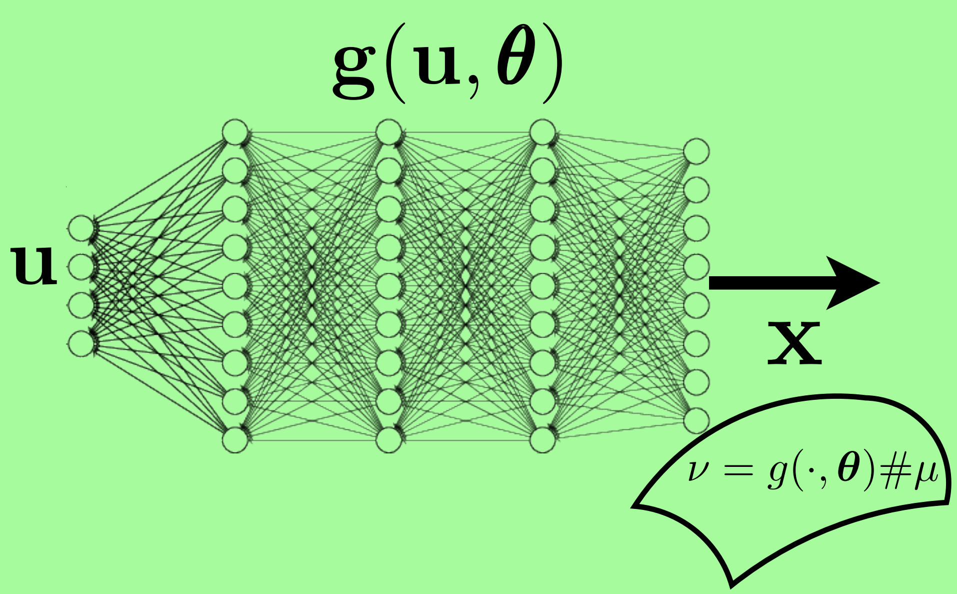
Convergence



Trap avoidance

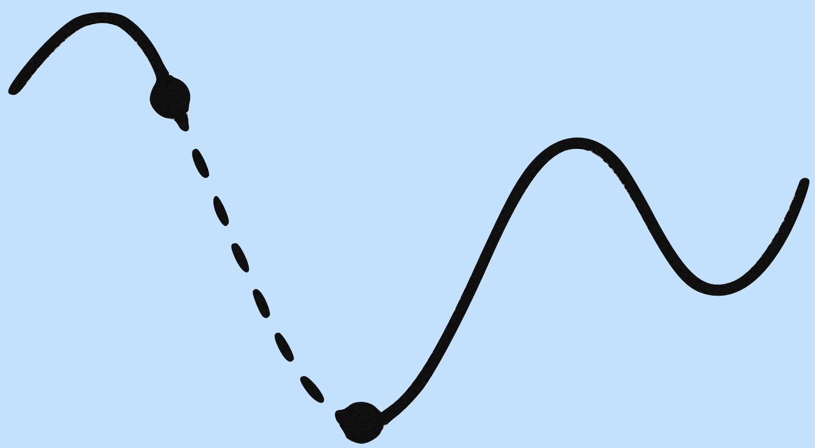


DIP recovery guarantees

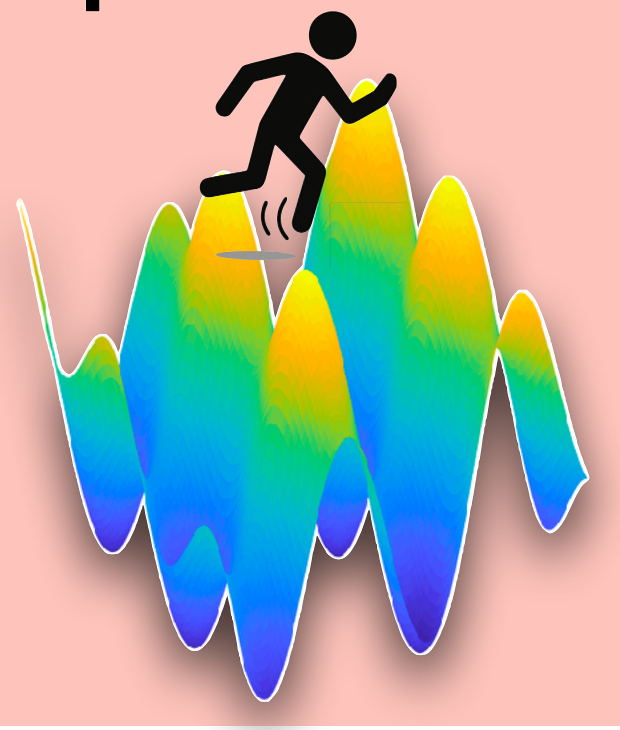


Outline

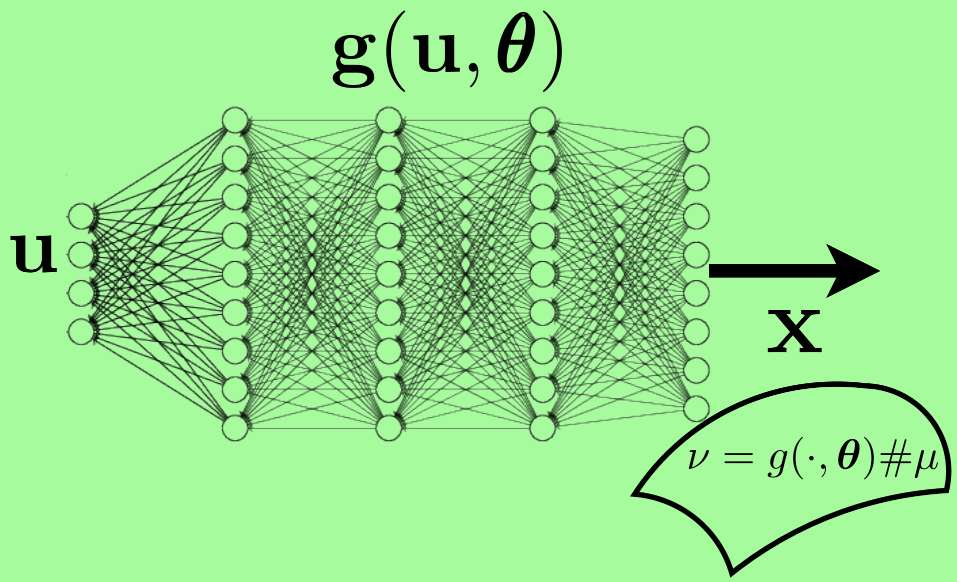
Convergence



Trap avoidance



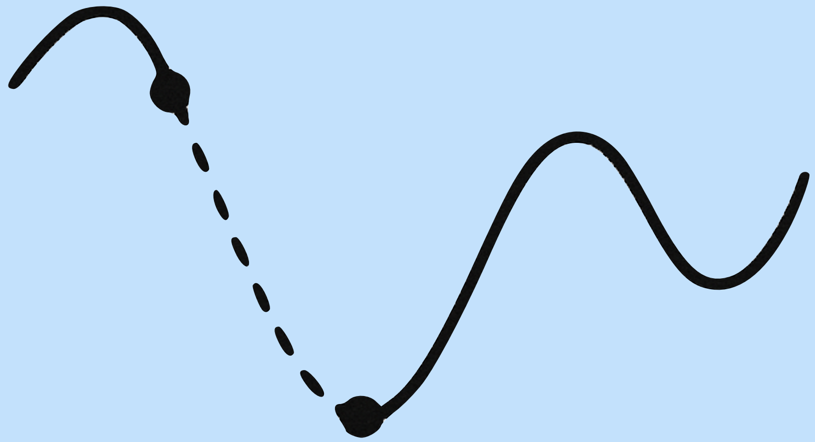
DIP recovery guarantees



Conclusion

Outline

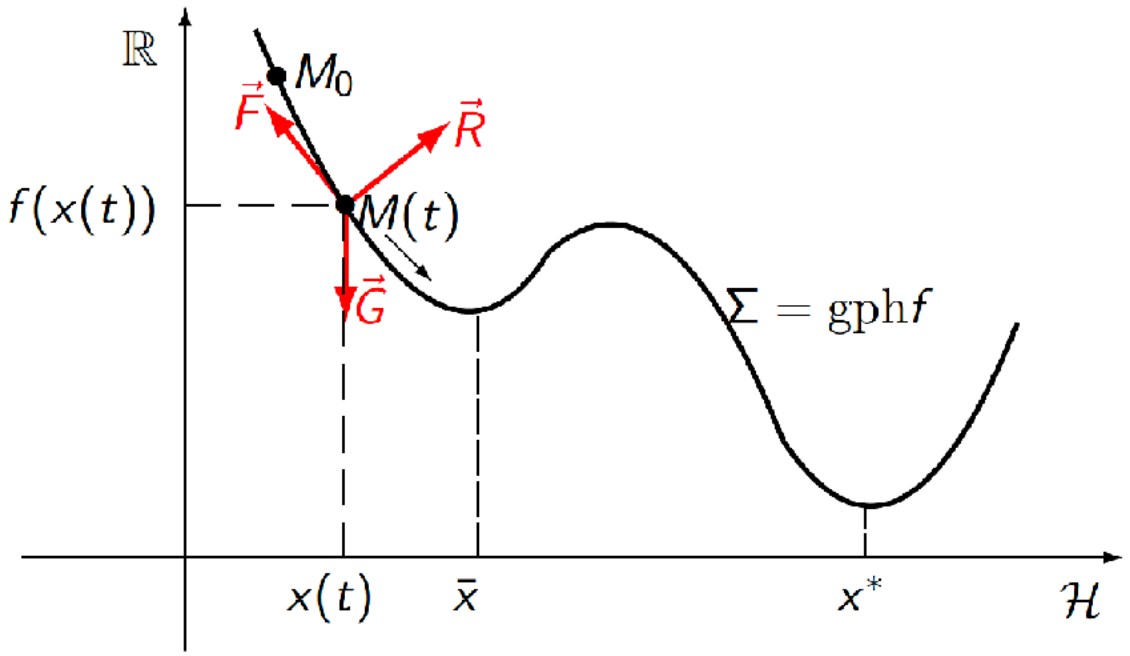
Convergence



Inertial Systems with Hessian Damping

$$\min_{x\in\mathbb{R}^d} f(x), \qquad f \in C^2(\mathbb{R}^d), \inf f > -\infty.$$

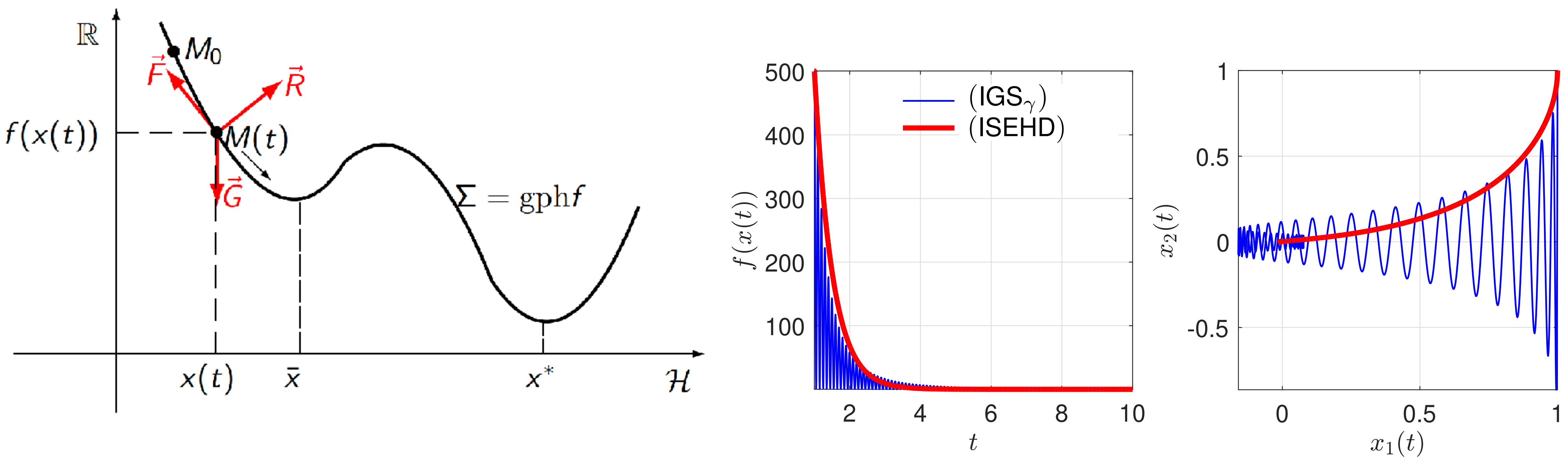
$$\ddot{x}(t) + \gamma(t)\dot{x}(t) + \nabla f(x(t)) = 0, \quad t > t_0, \qquad (\text{IGS}_\gamma)$$



Inertial Systems with Hessian Damping

$$\min_{x \in \mathbb{R}^d} f(x), \qquad f \in C^2(\mathbb{R}^d), \inf f > -\infty.$$

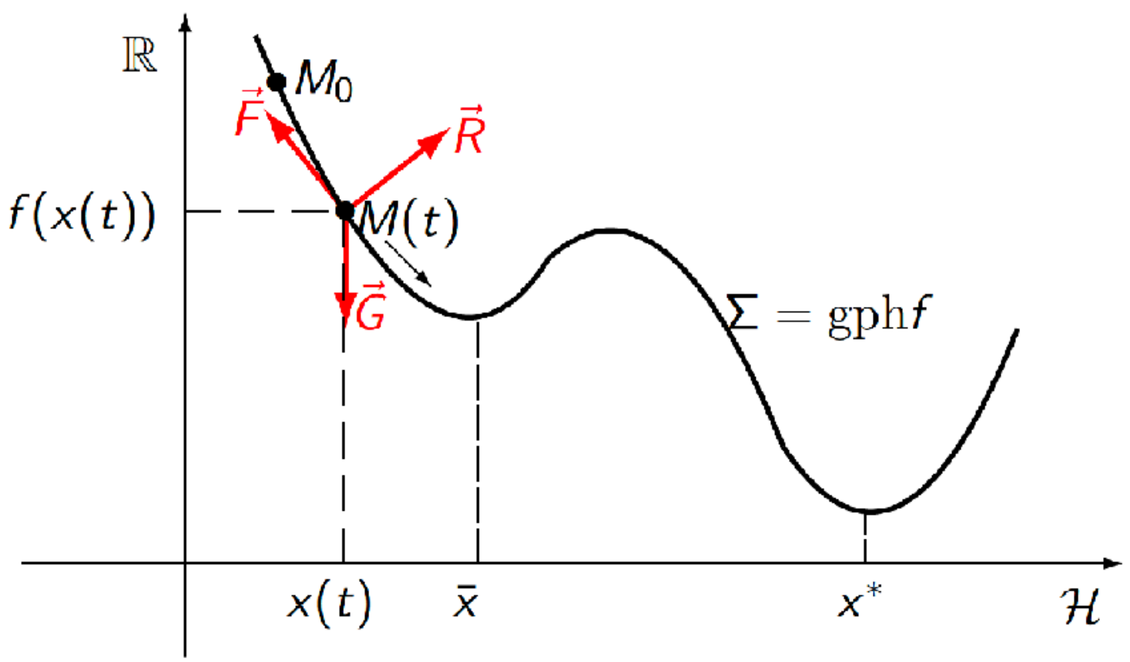
$$\ddot{x}(t) + \gamma(t)\dot{x}(t) + \nabla f(x(t)) = 0, \quad t > t_0, \qquad (\text{IGS}_\gamma)$$



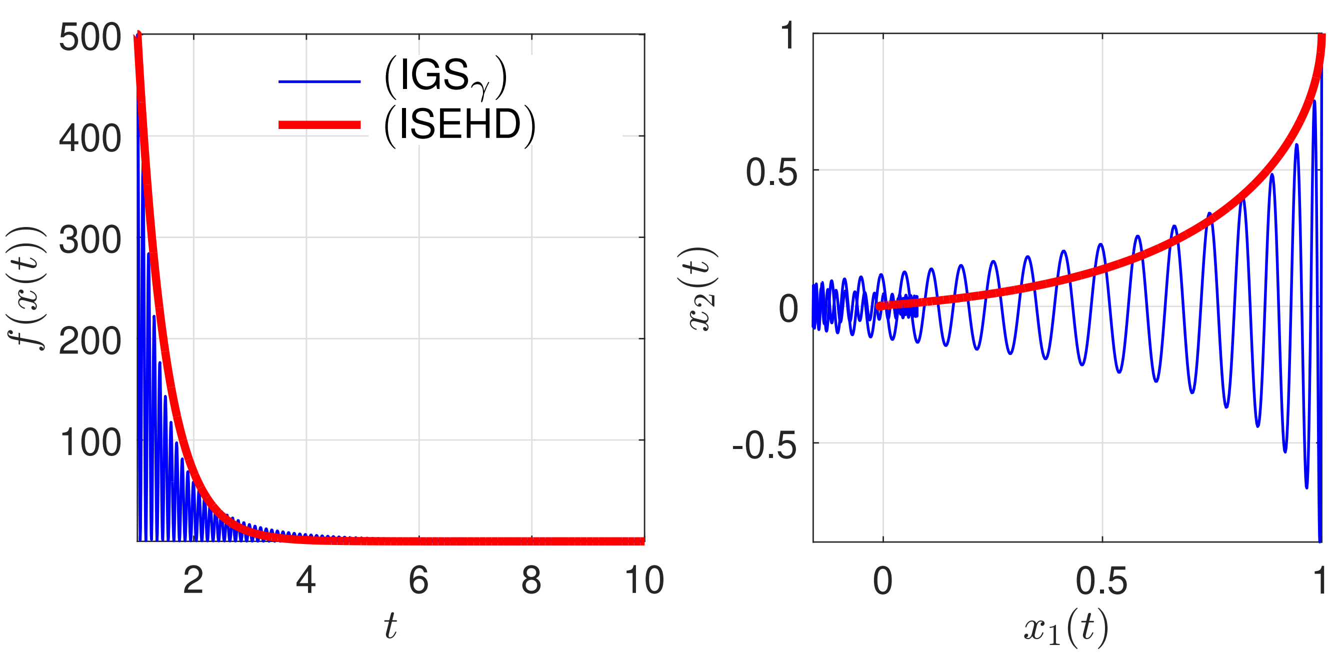
Inertial Systems with Hessian Damping

$$\min_{x\in\mathbb{R}^d} f(x), \qquad f\in C^2(\mathbb{R}^d), \inf f > -\infty.$$

$$\ddot{x}(t) + \gamma(t)\dot{x}(t) + \nabla f(x(t)) = 0, \quad t > t_0, \qquad (\text{IGS}_\gamma)$$



Neutralize oscillations by geometric damping



$$\ddot{x}(t) + \gamma(t)\dot{x}(t) + \beta(t)\nabla^2 f(x(t))\dot{x}(t) + \nabla f(x(t)) = 0 \qquad (\text{ISEHD})$$

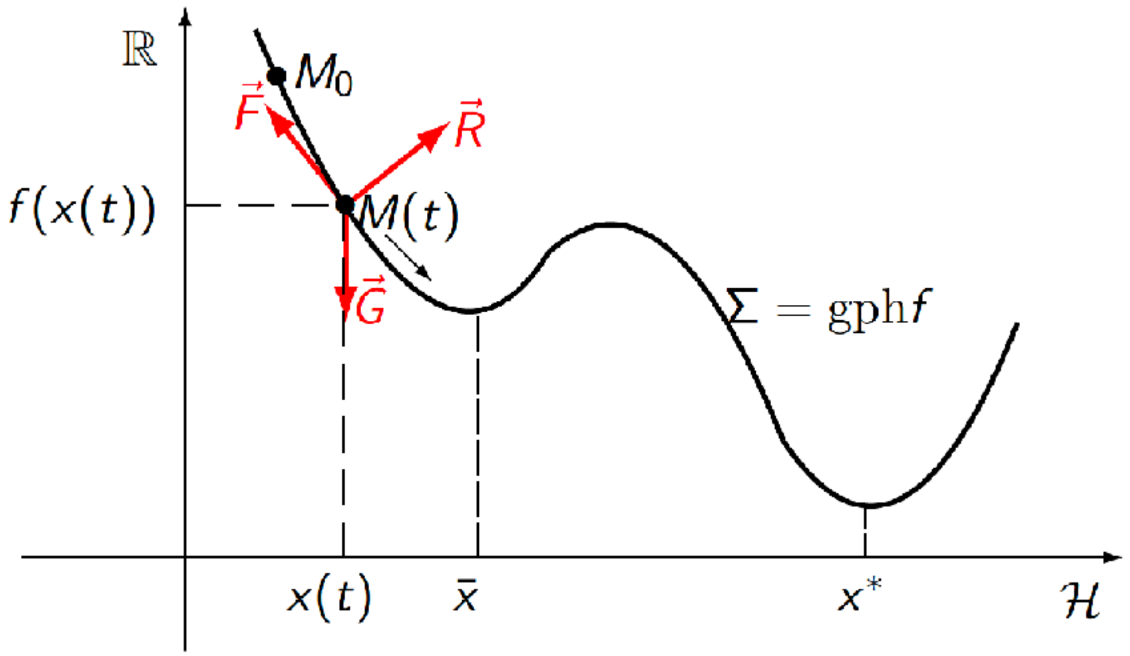
Viscous damping ($\gamma(t)$)

Geometric Hessian-driven damping ($\beta(t)$)

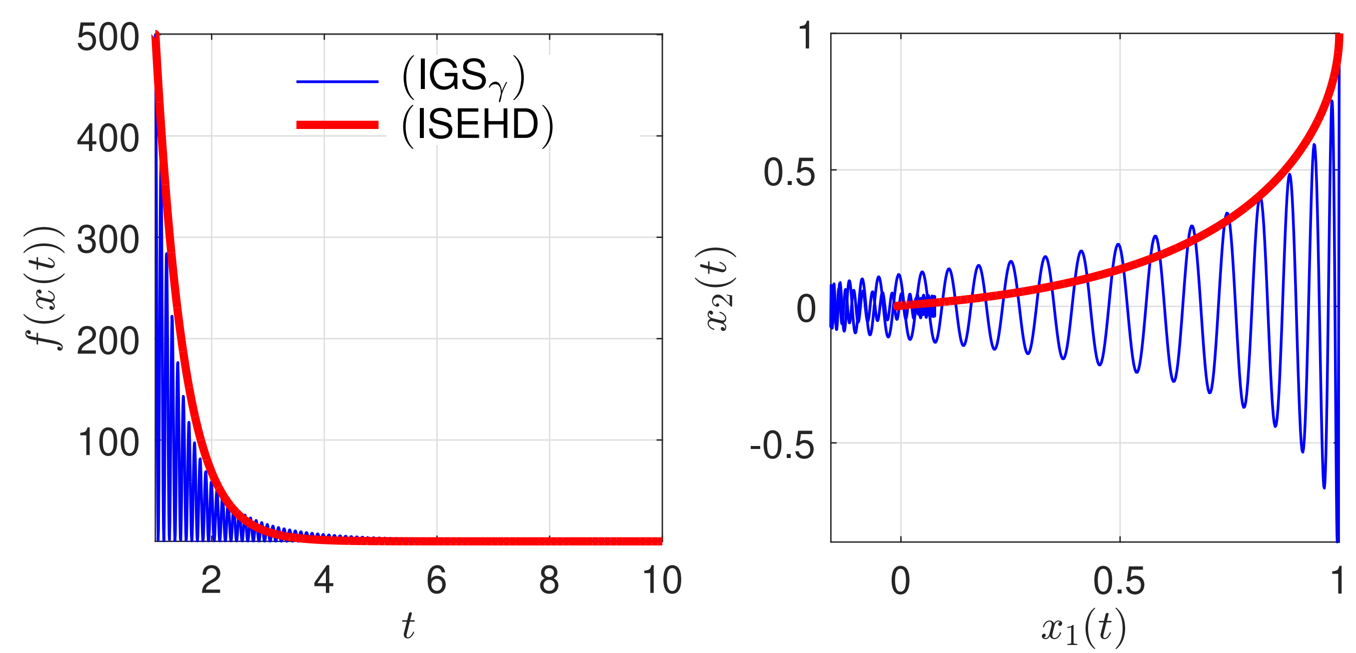
Inertial Systems with Hessian Damping

$$\min_{x \in \mathbb{R}^d} f(x), \qquad f \in C^2(\mathbb{R}^d), \inf f > -\infty.$$

$$\ddot{x}(t) + \gamma(t)\dot{x}(t) + \nabla f(x(t)) = 0, \quad t > t_0, \qquad (\text{IGS}_\gamma)$$



Neutralize oscillations by geometric damping



$$\ddot{x}(t) + \gamma(t)\dot{x}(t) + \beta(t)\nabla^2 f(x(t))\dot{x}(t) + \nabla f(x(t)) = 0 \qquad (\text{ISEHD})$$

Viscous damping

Geometric Hessian-driven damping

$$\ddot{x}(t) + \gamma(t)\dot{x}(t) + \nabla f(x(t) + \beta(t)\dot{x}(t)) = 0 \qquad (\text{ISIHD})$$

Main results

$$\min_{x\in\mathbb{R}^d} f(x), \qquad f \in C^2(\mathbb{R}^d), \inf f > -\infty.$$

$$\ddot{x}(t) + \gamma(t)\dot{x}(t) + \beta(t)\nabla^2 f(x(t))\dot{x}(t) + \nabla f(x(t)) = 0 \qquad \text{(ISEHD)}$$

$$\ddot{x}(t) + \gamma(t)\dot{x}(t) + \nabla f(x(t) + \beta(t)\dot{x}(t)) = 0 \qquad \text{(ISIHD)}$$

- For both systems:
  - Convergence of the **gradient to zero** and convergence of the **values**.
  - **Global convergence and rates** of the trajectories to a critical point for "nice" functions.
  - **Trap avoidance**: generic convergence of the trajectory to a local minimum.
- Same results for several **discrete algorithms**.
- Results transfer to the **DIP training**.

Main results

$$\min_{x\in\mathbb{R}^d} f(x), \qquad f \in C^2(\mathbb{R}^d), \inf f > -\infty.$$

$$\ddot{x}(t) + \gamma(t)\dot{x}(t) + \beta(t)\nabla^2 f(x(t))\dot{x}(t) + \nabla f(x(t)) = 0 \qquad \text{(ISEHD)}$$

$$\ddot{x}(t) + \gamma(t)\dot{x}(t) + \nabla f(x(t) + \beta(t)\dot{x}(t)) = 0 \qquad \text{(ISIHD)}$$

- For both systems:
  - Convergence of the **gradient to zero** and convergence of the **values**.
  - **Global convergence and rates** of the trajectories to a critical point for "nice" functions.
  - **Trap avoidance**: generic convergence of the trajectory to a local minimum.
- Same results for several **discrete algorithms**.
- Results transfer to the **DIP training**.

In the rest of the talk, focus on (ISEHD) and its discrete version (IGAHD)

IGAHD Algorithm

$$\min_{x \in \mathbb{R}^d} f(x), \qquad f \in C^2(\mathbb{R}^d), \inf f > -\infty.$$

$$\ddot{x}(t) + \gamma(t)\dot{x}(t) + \beta(t)\nabla^2 f(x(t))\dot{x}(t) + \nabla f(x(t)) = 0 \qquad \text{(ISEHD)}$$

$$\frac{x_{k+1} - 2x_k + x_{k-1}}{h^2} + \gamma(kh)\frac{x_{k+1} - x_k}{h} + \beta\frac{\nabla f(x_k) - \nabla f(x_{k-1})}{h} + \nabla f(x_k) = 0.$$

$$\begin{cases} y_k & = x_k + \alpha_k(x_k - x_{k-1}) - \beta_k(\nabla f(x_k) - \nabla f(x_{k-1})), \\ x_{k+1} & = y_k - s_k \nabla f(x_k). \end{cases} \qquad \text{(IGAHD)}$$

$$\alpha_k \stackrel{\text{def}}{=} \tfrac{1}{1+\gamma_k h}, \gamma_k \stackrel{\text{def}}{=} \gamma(kh), \beta_k \stackrel{\text{def}}{=} \beta h \alpha_k, s_k \stackrel{\text{def}}{=} h^2 \alpha_k.$$

Convergence and rates of IGAHD

$$\min_{x \in \mathbb{R}^d} f(x), \qquad f \in C^2(\mathbb{R}^d), \inf f > -\infty.$$

$$\begin{cases} y_k & = x_k + \alpha_k(x_k - x_{k-1}) - \beta_k(\nabla f(x_k) - \nabla f(x_{k-1})), \\ x_{k+1} & = y_k - s_k \nabla f(x_k). \end{cases} \qquad \text{(IGAHD)}$$

$$\alpha_k \overset{\text{def}}{=} \frac{1}{1+\gamma_k h}, \gamma_k \overset{\text{def}}{=} \gamma(kh), \beta_k \overset{\text{def}}{=} \beta h \alpha_k, s_k \overset{\text{def}}{=} h^2 \alpha_k.$$

Theorem *Let $f \in C^2(\mathbb{R}^d) \cap C_L^{1,1}(\mathbb{R}^d)$. Assume that $h > 0$, $\beta \geq 0$ and $c \leq \gamma_k \leq C$ for some $c, C > 0$.*

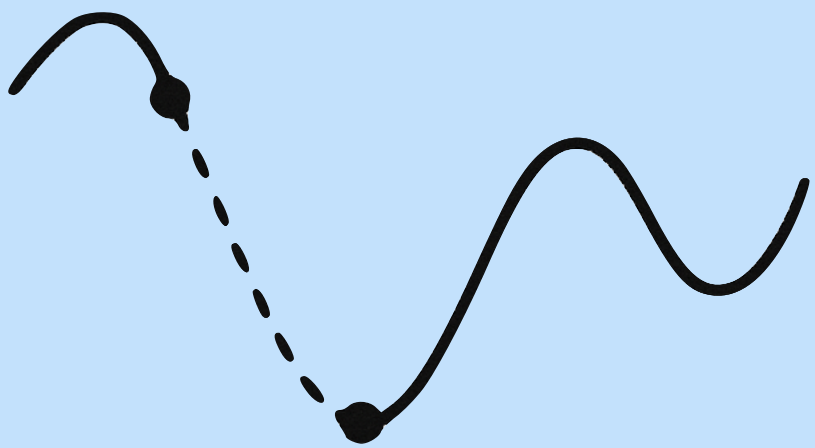
(i) If $\beta + \frac{h}{2} < \frac{c}{L}$, f is definable and $(x_k)_{k \in \mathbb{N}}$ is bounded, then $(\|x_{k+1} - x_k\|)_{k \in \mathbb{N}} \in \ell^1(\mathbb{N})$ and $x_k \to x_\infty \in \mathrm{Crit}(f)$.

(ii) If f is Łojasiewicz with exponent $q \in [0, 1[$, then

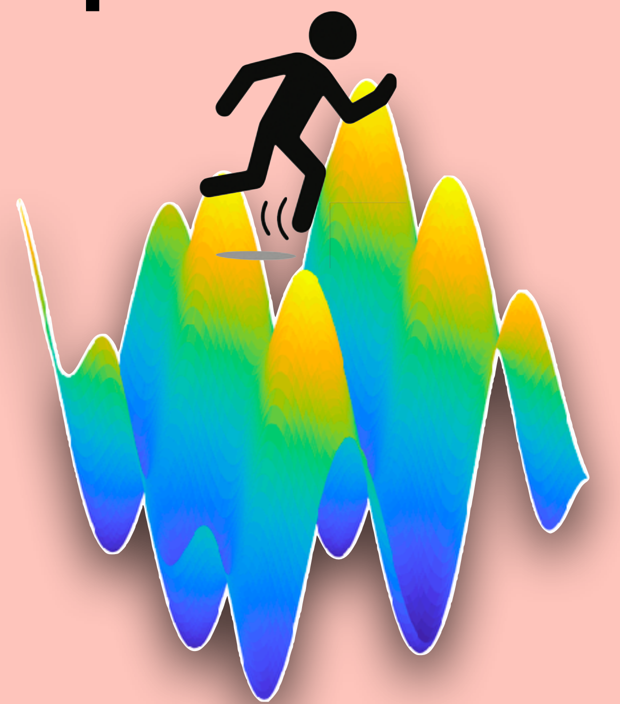
- *if $q \in [0, \frac{1}{2}]$ then there exists $\rho \in]0, 1[$ such that*
$$\|x_k - x_\infty\| = \mathcal{O}(\rho^k).$$
- *If $q \in]\frac{1}{2}, 1[$ then* $$\|x_k - x_\infty\| = \mathcal{O}\left(k^{-\frac{1-q}{2q-1}}\right).$$

Outline

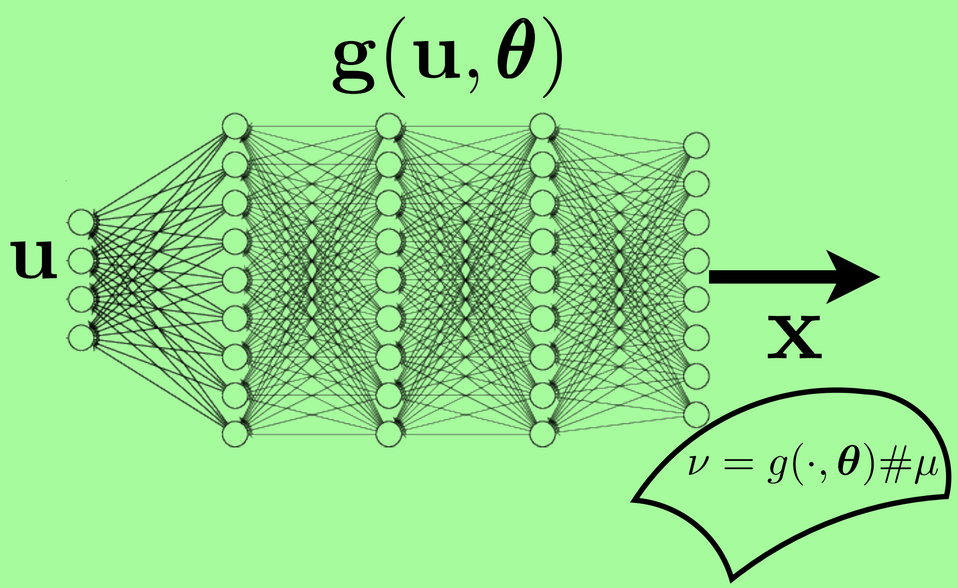
Convergence



Trap avoidance

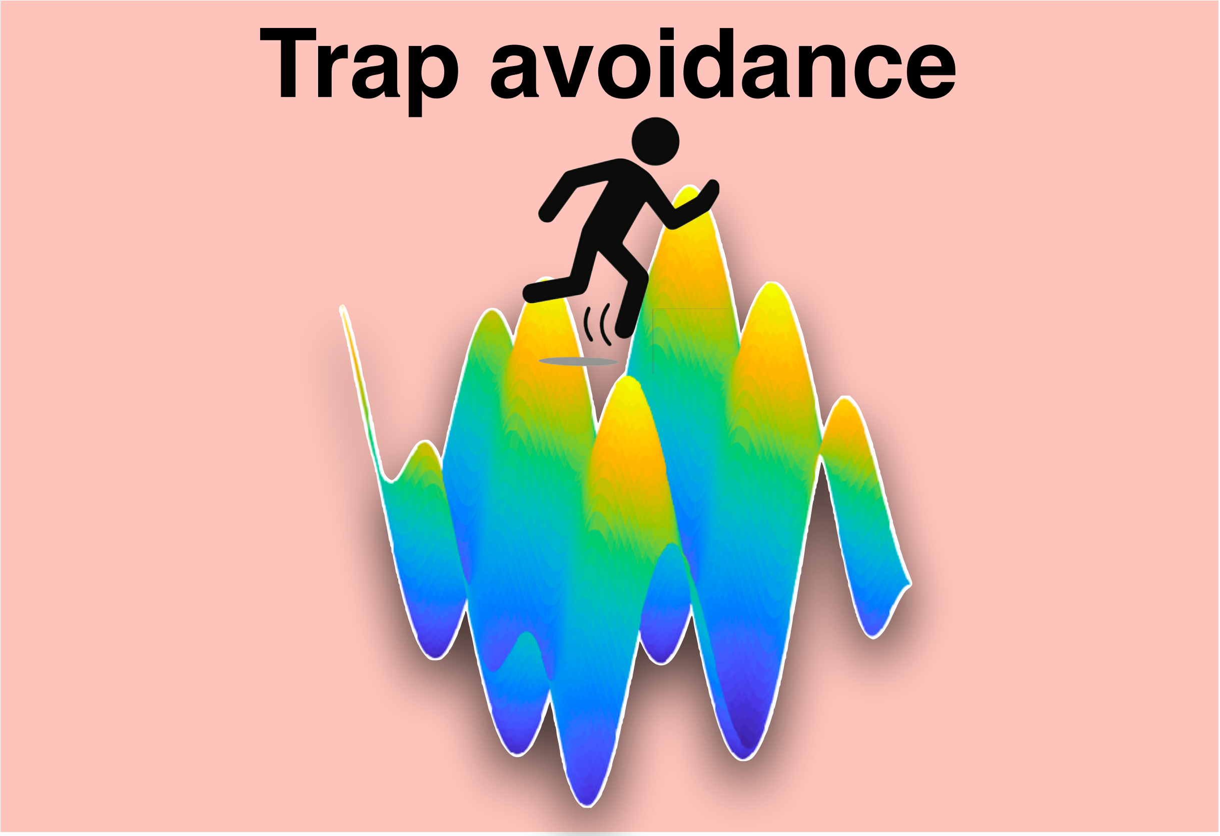


DIP recovery guarantees



Conclusion

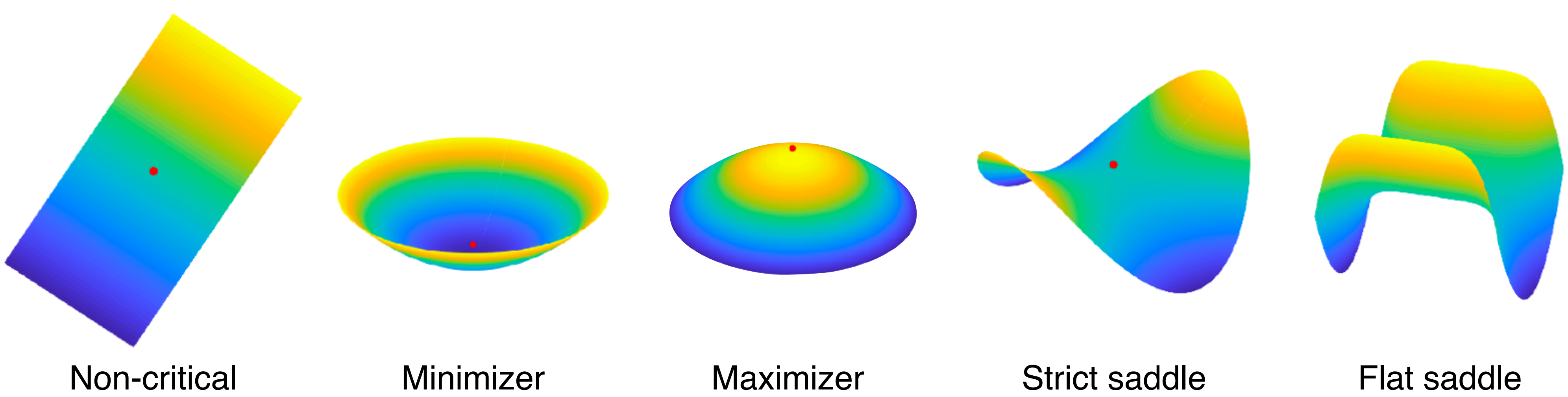
Outline



Trap avoidance: what is it ?

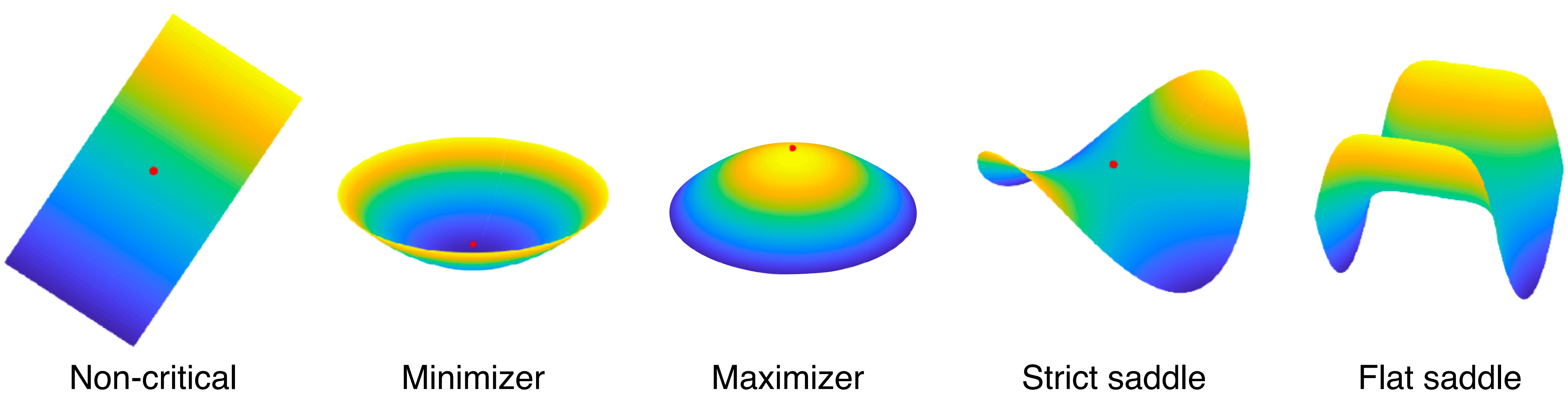
Trap avoidance: what is it ?

- We proved only convergence to critical points.



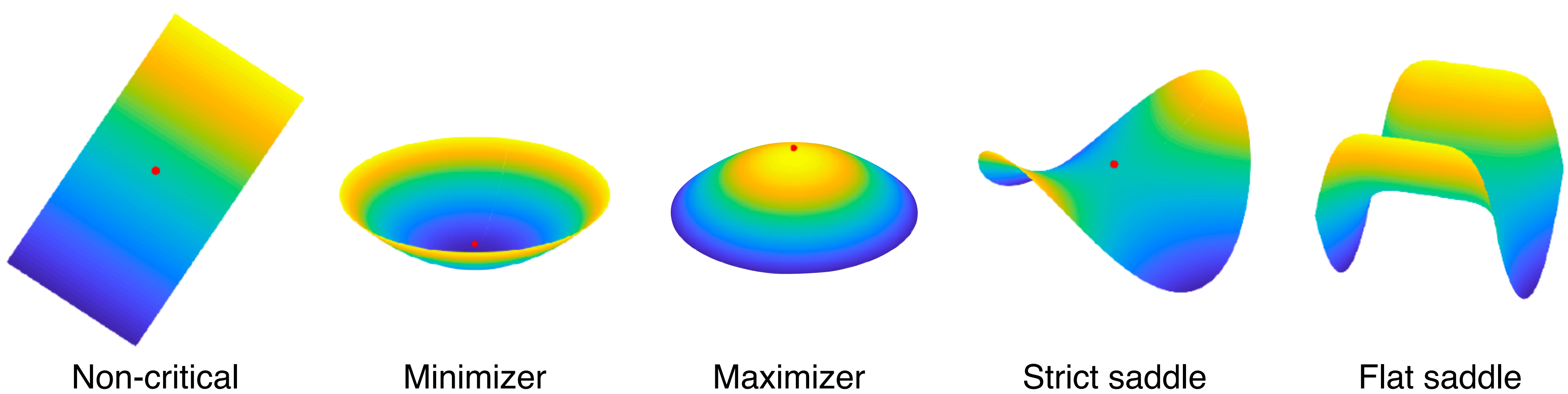
Trap avoidance: what is it ?

- We proved only convergence to critical points.
- Finding global (and even local) minima is (NP-)hard in general.



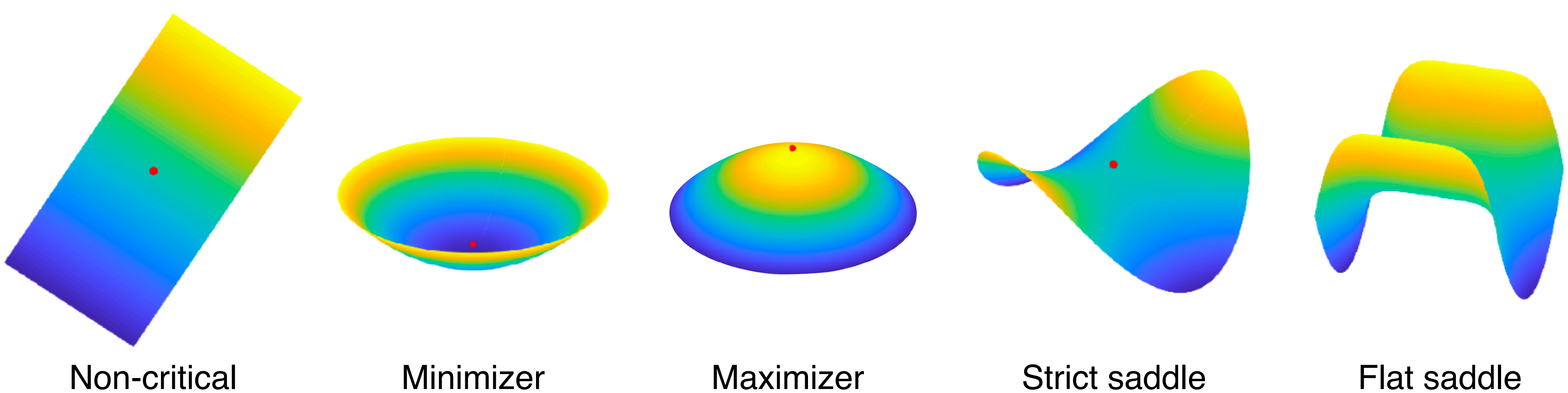
Trap avoidance: what is it ?

- We proved only convergence to critical points.
- Finding global (and even local) minima is (NP-)hard in general.
- Local descent methods can get trapped at saddle points.



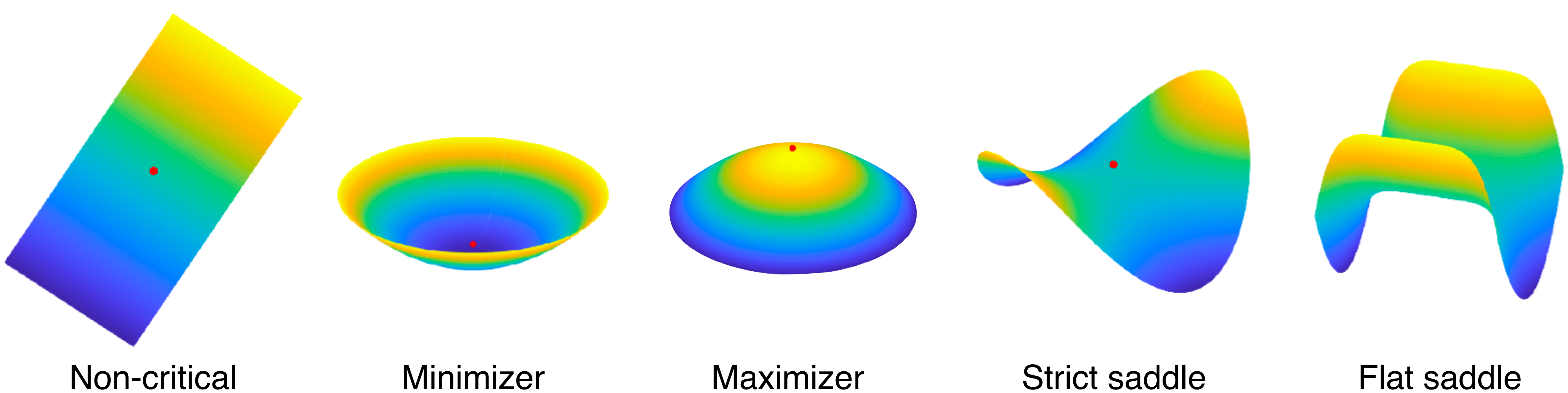
Trap avoidance: what is it ?

- We proved only convergence to critical points.
- Finding global (and even local) minima is (NP-)hard in general.
- Local descent methods can get trapped at saddle points.
- Can this be avoided ?



Trap avoidance: what is it ?

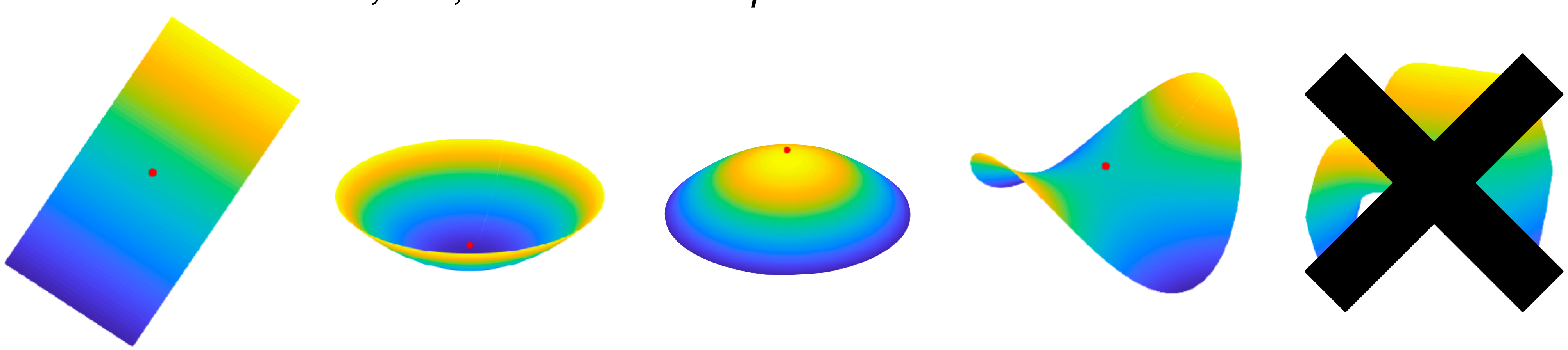
- We proved only convergence to critical points.
- Finding global (and even local) minima is (NP-)hard in general.
- Local descent methods can get trapped at saddle points.
- Can this be avoided ?
- Yes: **center stable manifold theorem**.



Trap avoidance: what is it ?

- We proved only convergence to critical points.
- Finding global (and even local) minima is (NP-)hard in general.
- Local descent methods can get trapped at saddle points.
- Can this be avoided ?
- Yes: **center stable manifold theorem.**

Definition *We will say that* $\hat{x}$ *is a strict saddle point of* $f \in C^2(\mathbb{R}^d)$ *if* $\hat{x} \in \mathrm{Crit}(f)$ *and* $\lambda_{\min}(\nabla^2 f(\hat{x})) < 0$.
$f \in C^2(\mathbb{R}^d)$ *has the strict saddle property if every critical point is either a local minimum or a strict saddle, i.e., no flat saddle points.*

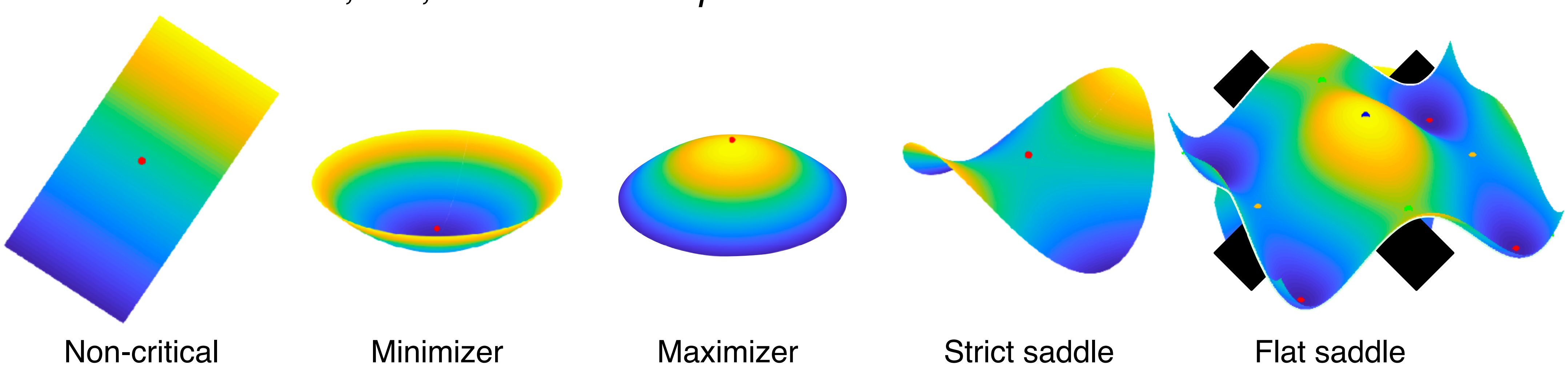


Non-critical   Minimizer   Maximizer   Strict saddle   Flat saddle

Trap avoidance: what is it ?

- We proved only convergence to critical points.
- Finding global (and even local) minima is (NP-)hard in general.
- Local descent methods can get trapped at saddle points.
- Can this be avoided ?
- Yes: **center stable manifold theorem.**

Definition *We will say that* $\hat{x}$ *is a strict saddle point of* $f \in C^2(\mathbb{R}^d)$ *if* $\hat{x} \in \mathrm{Crit}(f)$ *and* $\lambda_{\min}(\nabla^2 f(\hat{x})) < 0$.
$f \in C^2(\mathbb{R}^d)$ *has the strict saddle property if every critical point is either a local minimum or a strict saddle, i.e., no flat saddle points.*



- This property is generic over the space of C^2 (Morse) functions.

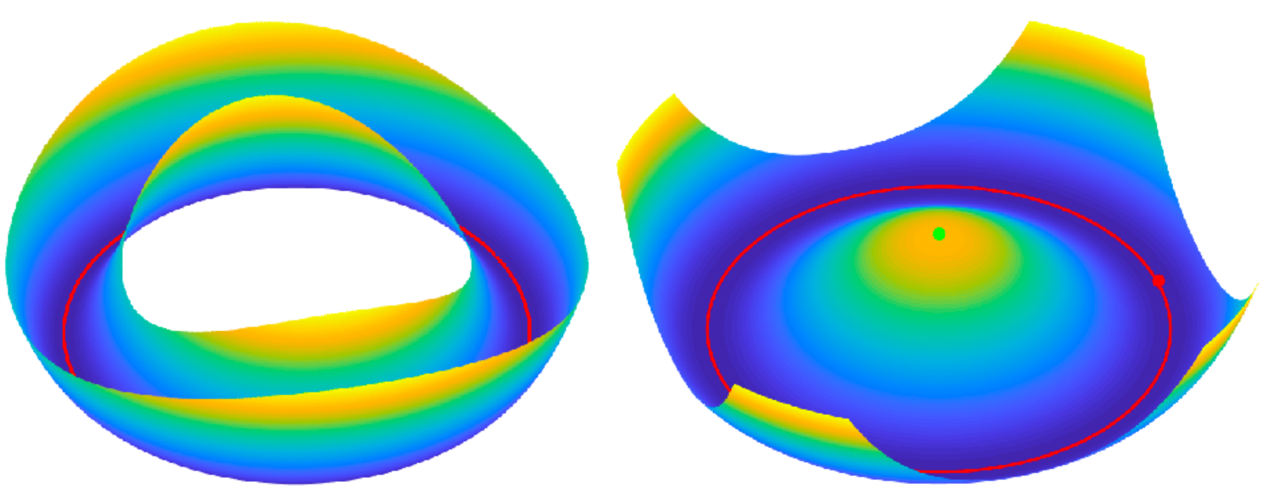
Trap avoidance of IGAHD

$$\min_{x \in \mathbb{R}^d} f(x), \qquad f \in C^2(\mathbb{R}^d), \inf f > -\infty.$$

$$\begin{cases} y_k &= x_k + \alpha_k(x_k - x_{k-1}) - \beta_k(\nabla f(x_k) - \nabla f(x_{k-1})), \\ x_{k+1} &= y_k - s_k \nabla f(x_k). \end{cases}$$

$$\alpha_k \overset{\text{def}}{=} \tfrac{1}{1+\gamma_k h}, \gamma_k \overset{\text{def}}{=} \gamma(kh), \beta_k \overset{\text{def}}{=} \beta h \alpha_k, s_k \overset{\text{def}}{=} h^2 \alpha_k.$$

Theorem *Let $f \in C^2(\mathbb{R}^d) \cap C_L^{1,1}(\mathbb{R}^d)$ be a definable function. Assume that $\gamma_k \equiv c > 0$, $0 < \beta < \frac{c}{L}, \beta \neq \frac{1}{c}$, and $h < \min(2\left(\frac{c}{L} - \beta\right), \frac{1}{L\beta})$, then for almost all $x_0, x_1 \in \mathbb{R}^d$, x_k converges to a critical point of f that is not a strict saddle. Consequently, if f satisfies the strict saddle property then for almost all $x_0, x_1 \in \mathbb{R}^d$, x_k converges to a local minimum of f.*



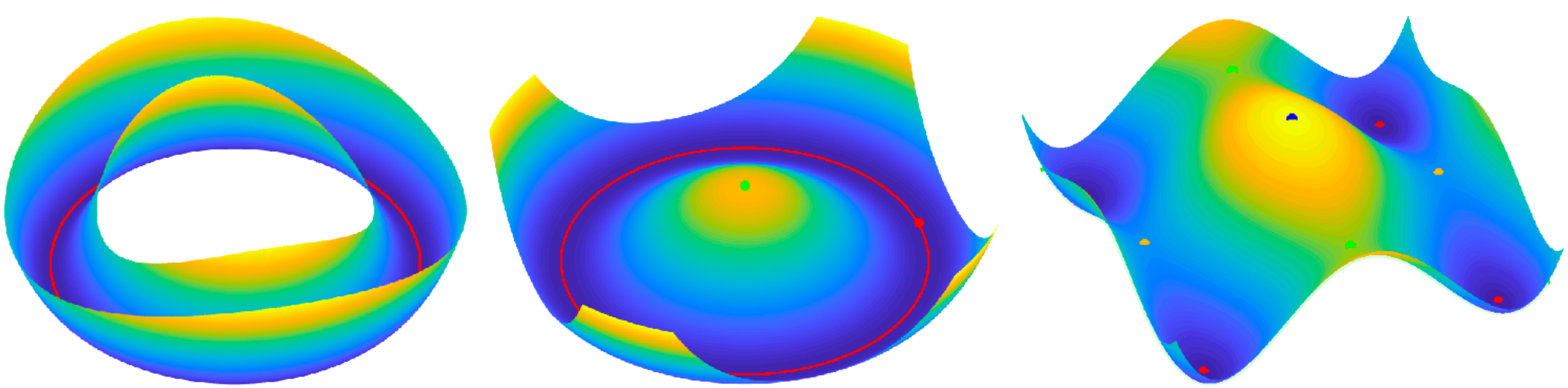
Trap avoidance of IGAHD

$$\min_{x\in\mathbb{R}^d} f(x), \qquad f\in C^2(\mathbb{R}^d), \inf f > -\infty.$$

$$\begin{cases} y_k & = x_k + \alpha_k(x_k - x_{k-1}) - \beta_k(\nabla f(x_k) - \nabla f(x_{k-1})), \\ x_{k+1} & = y_k - s_k \nabla f(x_k). \end{cases}$$

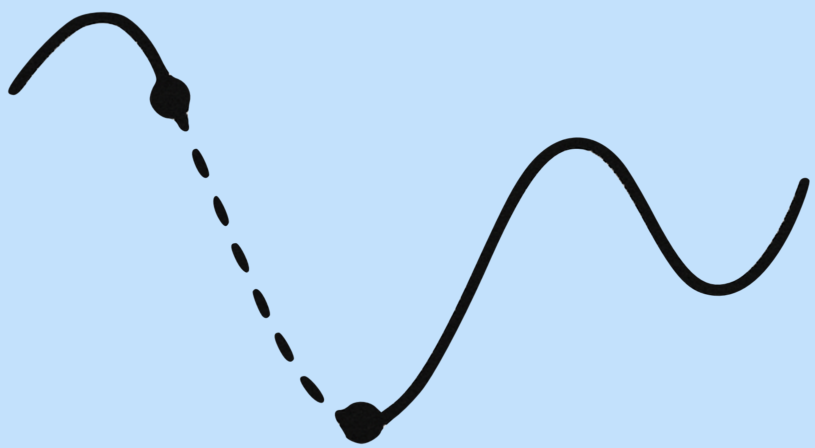
$$\alpha_k \stackrel{\text{def}}{=} \frac{1}{1+\gamma_k h}, \gamma_k \stackrel{\text{def}}{=} \gamma(kh), \beta_k \stackrel{\text{def}}{=} \beta h \alpha_k, s_k \stackrel{\text{def}}{=} h^2\alpha_k.$$

Theorem *Let $f \in C^2(\mathbb{R}^d) \cap C_L^{1,1}(\mathbb{R}^d)$ be a definable function. Assume that $\gamma_k \equiv c > 0$, $0 < \beta < \frac{c}{L}, \beta \neq \frac{1}{c}$, and $h < \min(2\left(\frac{c}{L} - \beta\right), \frac{1}{L\beta})$, then for almost all $x_0, x_1 \in \mathbb{R}^d$, x_k converges to a critical point of f that is not a strict saddle. Consequently, if f satisfies the strict saddle property then for almost all $x_0, x_1 \in \mathbb{R}^d$, x_k converges to a local minimum of f.*

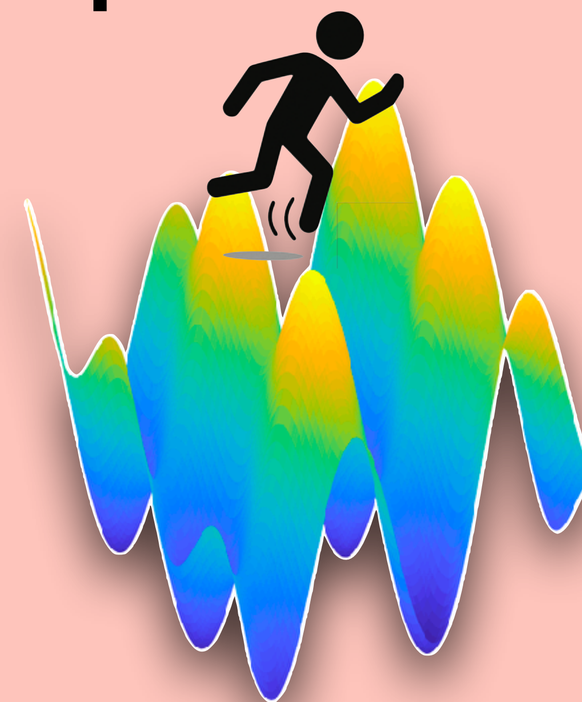


Outline

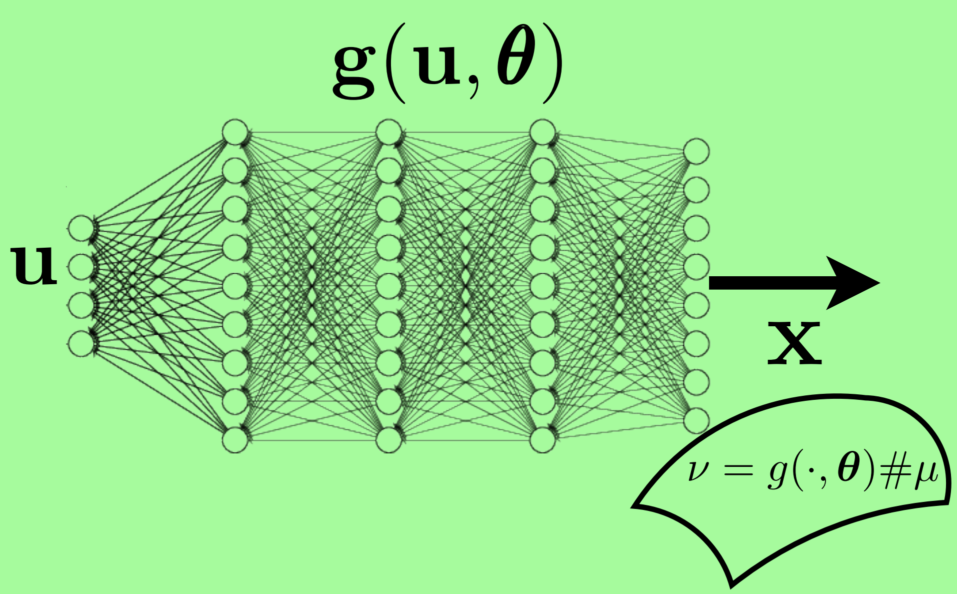
Convergence



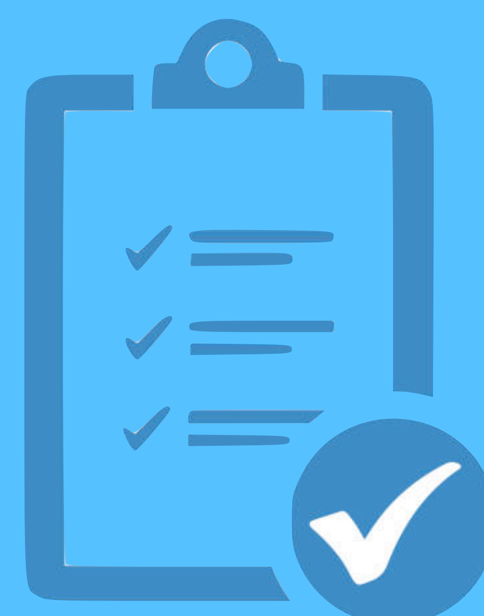
Trap avoidance



DIP recovery guarantees

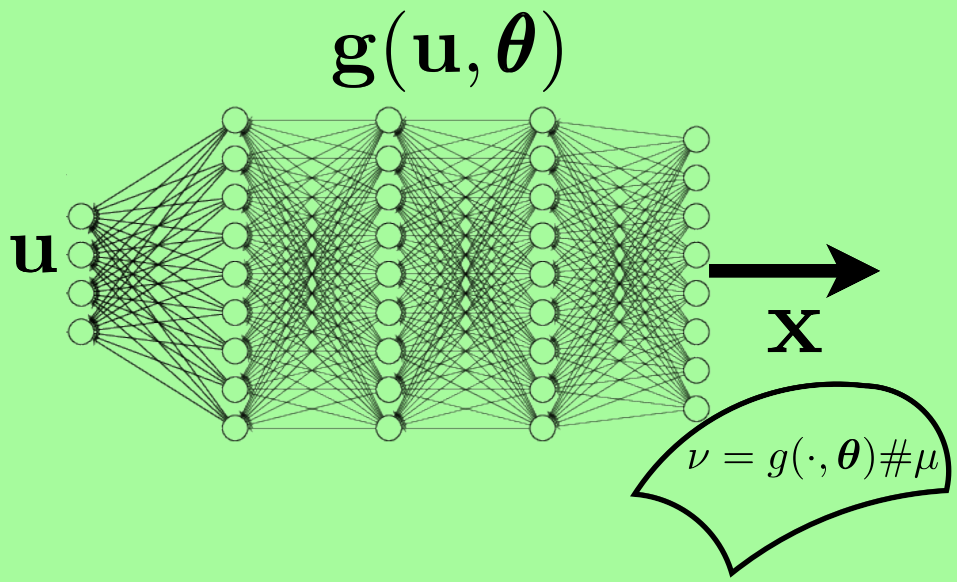


Conclusion



Outline

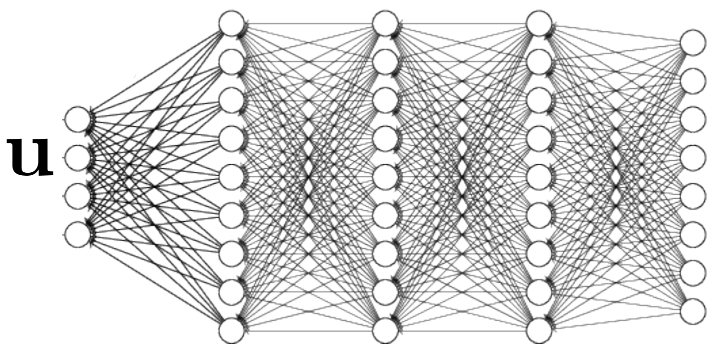
DIP recovery guarantees



DIP training with inertia

$$\mathbf{y} = \mathbf{A}\overline{\mathbf{x}} + \varepsilon \qquad \mathbf{A} \in \mathbb{R}^{m \times n}$$

$\mathbf{g}(\mathbf{u}, \boldsymbol{\theta})$



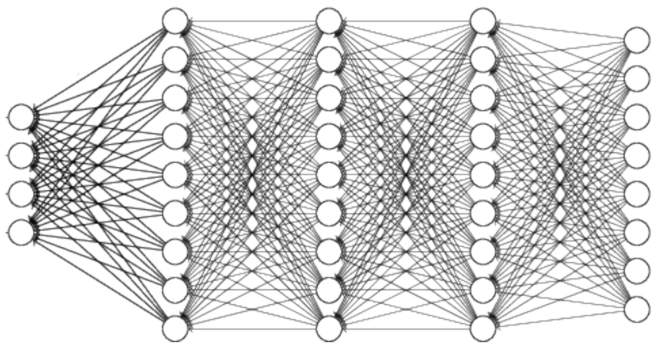
$$\min_{\boldsymbol{\theta} \in \Theta} \mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}))$$

DIP training with inertia

$$\mathbf{y} = \mathbf{A}\overline{\mathbf{x}} + \varepsilon \qquad \mathbf{A} \in \mathbb{R}^{m \times n}$$

$\mathbf{g}(\mathbf{u}, \boldsymbol{\theta})$

$\mathbf{u}$



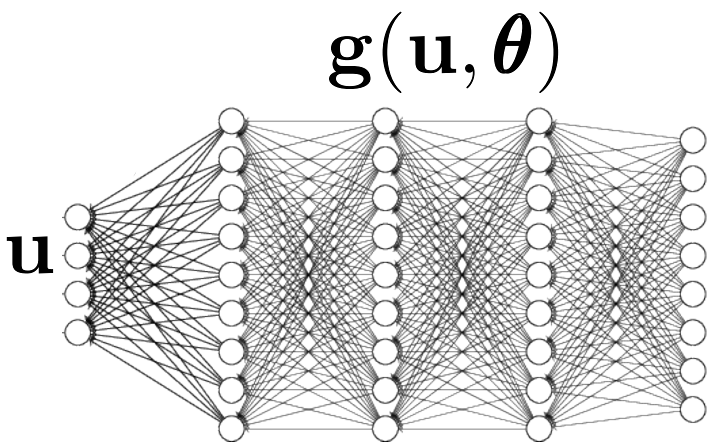
$$\min_{\boldsymbol{\theta} \in \Theta} \mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}))$$

$$\text{(ISEHD)}\begin{cases} \ddot{\boldsymbol{\theta}}(t) + \alpha\dot{\boldsymbol{\theta}}(t) + \beta\frac{\mathrm{d}}{\mathrm{d}t}\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}(t))) + \nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}(t))) = 0 \\ \boldsymbol{\theta}(0) = \boldsymbol{\theta}_0, \dot{\boldsymbol{\theta}}(0) = 0. \end{cases}$$

$$\text{(IGAHD)}\begin{cases} \boldsymbol{\eta}_\ell &= \boldsymbol{\theta}_\ell + \alpha s_\ell(\boldsymbol{\theta}_\ell - \boldsymbol{\theta}_{\ell-1}) - \beta s_\ell^2\left(\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}_\ell)) - \nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}_{\ell-1}))\right), \\ \boldsymbol{\theta}_{\ell+1} &= \boldsymbol{\eta}_\ell - s_\ell\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}_\ell)). \end{cases}$$

DIP training with inertia

$$\mathbf{y} = \mathbf{A}\overline{\mathbf{x}} + \varepsilon \qquad \mathbf{A} \in \mathbb{R}^{m \times n}$$



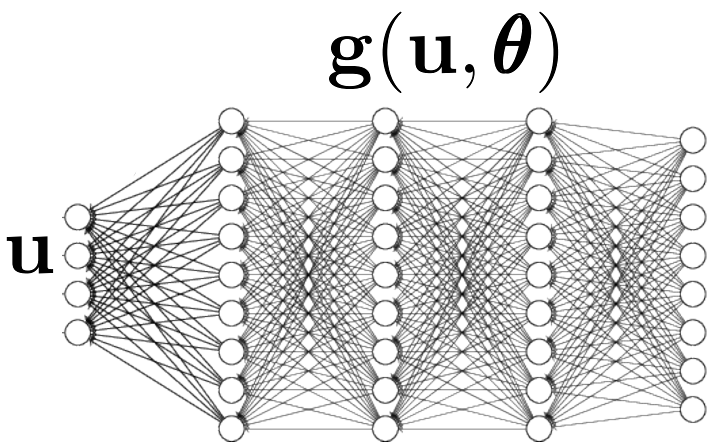
$$\min_{\boldsymbol{\theta} \in \Theta} \mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}))$$

Assumptions

- $\mathcal{L}_{\mathbf{y}}$: quadratic loss.
- $\phi \in \mathcal{C}^1(\mathbb{R})$ and $\exists B > 0$ such that $\sup_{x \in \mathbb{R}} |\phi'(x)| \leq B$ and ϕ' is B-Lipschitz continuous.

DIP training with inertia

$$\mathbf{y} = \mathbf{A}\overline{\mathbf{x}} + \varepsilon \qquad \mathbf{A} \in \mathbb{R}^{m \times n}$$



$$\min_{\boldsymbol{\theta} \in \Theta} \mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}))$$

Assumptions

- $\mathcal{L}_{\mathbf{y}}$: quadratic loss.
- $\phi \in \mathcal{C}^1(\mathbb{R})$ and $\exists B > 0$ such that $\sup_{x \in \mathbb{R}} |\phi'(x)| \leq B$ and ϕ' is B-Lipschitz continuous.

Goal

- Recovery guarantees of DIP when optimized with inertial methods in :
  - Observation ($\mathbf{y}$) space : convergence to zero-loss $\Rightarrow$ implicit regularization.
  - Object ($\mathbf{x}$) space : restricted injectivity of the forward operator on Σ.
- NN architecture : role of overparametrization.

Recovery guarantees: y space

$$\mathbf{y} = \mathbf{A}\overline{\mathbf{x}} + \varepsilon$$

$$\ddot{\boldsymbol{\theta}}(t) + \alpha\dot{\boldsymbol{\theta}}(t) + \beta\frac{\mathrm{d}}{\mathrm{d}t}\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(t))) + \nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(t))) = 0 \qquad \text{(ISEHD)}$$

Recovery guarantees: y space

$$\mathbf{y} = \mathbf{A}\overline{\mathbf{x}} + \varepsilon$$

$$\ddot{\boldsymbol{\theta}}(t) + \alpha\dot{\boldsymbol{\theta}}(t) + \beta\frac{\mathrm{d}}{\mathrm{d}t}\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(t))) + \nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(t))) = 0 \qquad \text{(ISEHD)}$$

$$\sigma_{\mathbf{A}} \stackrel{\text{def}}{=} \inf_{\mathbf{z}\in\mathrm{Ker}(\mathbf{A})^{\perp}} \|\mathbf{A}\mathbf{z}\| \,/\, \|\mathbf{z}\| > 0.$$

Recovery guarantees: y space

$$\mathbf{y} = \mathbf{A}\overline{\mathbf{x}} + \varepsilon$$

$$\ddot{\boldsymbol{\theta}}(t) + \alpha\dot{\boldsymbol{\theta}}(t) + \beta\frac{\mathrm{d}}{\mathrm{d}t}\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(t))) + \nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(t))) = 0 \qquad \text{(ISEHD)}$$

$$\sigma_{\mathbf{A}} \stackrel{\text{def}}{=} \inf_{\mathbf{z}\in\mathrm{Ker}(\mathbf{A})^{\perp}} \|\mathbf{A}\mathbf{z}\| / \|\mathbf{z}\| > 0.$$

Theorem *Let $\boldsymbol{\theta}(\cdot)$ be a solution trajectory of (ISEHD) with $\alpha = \sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}$ and $\beta = \frac{1}{2\alpha}$ where the initialization $\boldsymbol{\theta}_0$ is such that*

$$\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0)) > 0 \quad \textit{and} \quad R' < R,$$

where R' and R obey

$$R' = \eta\sqrt{\xi\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))} \quad \textit{and} \quad R = \frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))}{2\mathrm{Lip}_{\mathbb{B}(\boldsymbol{\theta}_0,R)}(\mathcal{J}_{\mathbf{g}})}$$

with

$$\xi = 1 + \frac{\kappa(\mathcal{J}_{\mathbf{g}}(0))^2\kappa(\mathbf{A})^2}{4} \quad \textit{and} \quad \eta = \frac{4\max\left(\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}, \frac{1+\sqrt{2}}{2}\right)}{\min\left(\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))^2\sigma_{\mathbf{A}}^2, \frac{3}{4}\right)}.$$

Then, the following holds :

(i) the loss converges to 0 at the rate

$$\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)) \leq \xi\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))\exp\left(-\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}}{2}t\right).$$

Moreover, $\boldsymbol{\theta}(t)$ converges to a global minimizer $\boldsymbol{\theta}_\infty$ at the rate

$$\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}_\infty\| \leq \eta\sqrt{\xi\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))}\exp\left(-\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}}{4}t\right).$$

(ii) We have

$$\|\mathbf{y}(t) - \overline{\mathbf{y}}\| \leq 2\|\varepsilon\| \quad \textit{when} \quad t \geq \frac{4}{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}}\ln\left(\frac{\sqrt{2\xi\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))}}{\|\varepsilon\|}\right)$$

Recovery guarantees: y space

$$\mathbf{y} = \mathbf{A}\overline{\mathbf{x}} + \varepsilon$$

$$\ddot{\boldsymbol{\theta}}(t) + \alpha\dot{\boldsymbol{\theta}}(t) + \beta\frac{\mathrm{d}}{\mathrm{d}t}\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(t))) + \nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(t))) = 0 \qquad \text{(ISEHD)}$$

$$\sigma_{\mathbf{A}} \stackrel{\text{def}}{=} \inf_{\mathbf{z}\in\mathrm{Ker}(\mathbf{A})^{\perp}} \|\mathbf{A}\mathbf{z}\| / \|\mathbf{z}\| > 0.$$

Theorem *Let $\boldsymbol{\theta}(\cdot)$ be a solution trajectory of (ISEHD) with $\alpha = \sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}$ and $\beta = \frac{1}{2\alpha}$ where the initialization $\boldsymbol{\theta}_0$ is such that*

$$\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0)) > 0 \quad \textit{and} \quad R' < R,$$

where R' and R obey

$$R' = \eta\sqrt{\xi\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))} \quad \textit{and} \quad R = \frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))}{2\mathrm{Lip}_{\mathbb{B}(\boldsymbol{\theta}_0,R)}(\mathcal{J}_{\mathbf{g}})}$$

Non-degenerate initialization

with

$$\xi = 1 + \frac{\kappa(\mathcal{J}_{\mathbf{g}}(0))^2\kappa(\mathbf{A})^2}{4} \quad \textit{and} \quad \eta = \frac{4\max\left(\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}, \frac{1+\sqrt{2}}{2}\right)}{\min\left(\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))^2\sigma_{\mathbf{A}}^2, \frac{3}{4}\right)}.$$

Then, the following holds :

(i) the loss converges to 0 at the rate

$$\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)) \leq \xi\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))\exp\left(-\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}}{2}t\right).$$

Moreover, $\boldsymbol{\theta}(t)$ converges to a global minimizer $\boldsymbol{\theta}_\infty$ at the rate

$$\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}_\infty\| \leq \eta\sqrt{\xi\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))}\exp\left(-\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}}{4}t\right).$$

(ii) We have

$$\|\mathbf{y}(t) - \overline{\mathbf{y}}\| \leq 2\|\varepsilon\| \quad \textit{when} \quad t \geq \frac{4}{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}}\ln\left(\frac{\sqrt{2\xi\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))}}{\|\varepsilon\|}\right)$$

Recovery guarantees: y space

$$\mathbf{y} = \mathbf{A}\overline{\mathbf{x}} + \varepsilon$$

$$\ddot{\boldsymbol{\theta}}(t) + \alpha\dot{\boldsymbol{\theta}}(t) + \beta \frac{\mathrm{d}}{\mathrm{d}t}\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(t))) + \nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(t))) = 0 \qquad \text{(ISEHD)}$$

$$\sigma_{\mathbf{A}} \stackrel{\text{def}}{=} \inf_{\mathbf{z}\in\mathrm{Ker}(\mathbf{A})^{\perp}} \|\mathbf{A}\mathbf{z}\| / \|\mathbf{z}\| > 0.$$

Theorem *Let $\boldsymbol{\theta}(\cdot)$ be a solution trajectory of (ISEHD) with $\alpha = \sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}$ and $\beta = \frac{1}{2\alpha}$ where the initialization $\boldsymbol{\theta}_0$ is such that*

$$\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0)) > 0 \quad \text{and} \quad R' < R,$$

where R' and R obey

$$R' = \eta\sqrt{\xi\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))} \quad \text{and} \quad R = \frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))}{2\mathrm{Lip}_{\mathbb{B}(\boldsymbol{\theta}_0,R)}(\mathcal{J}_{\mathbf{g}})}$$

Non-degenerate initialization

with

$$\xi = 1 + \frac{\kappa(\mathcal{J}_{\mathbf{g}}(0))^2\kappa(\mathbf{A})^2}{4} \quad \text{and} \quad \eta = \frac{4\max\left(\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}, \frac{1+\sqrt{2}}{2}\right)}{\min\left(\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))^2\sigma_{\mathbf{A}}^2, \frac{3}{4}\right)}.$$

Then, the following holds :

(i) the loss converges to 0 at the rate

$$\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)) \leq \xi\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))\exp\left(-\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}}{2}t\right).$$

Moreover, $\boldsymbol{\theta}(t)$ converges to a global minimizer $\boldsymbol{\theta}_\infty$ at the rate

$$\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}_\infty\| \leq \eta\sqrt{\xi\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))}\exp\left(-\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}}{4}t\right).$$

Trajectory close to initialization

(ii) We have

$$\|\mathbf{y}(t) - \overline{\mathbf{y}}\| \leq 2\|\varepsilon\| \quad \text{when} \quad t \geq \frac{4}{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}}\ln\left(\frac{\sqrt{2\xi\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))}}{\|\varepsilon\|}\right)$$

Recovery guarantees: y space

$$\mathbf{y} = \mathbf{A}\overline{\mathbf{x}} + \varepsilon$$

$$\ddot{\boldsymbol{\theta}}(t) + \alpha\dot{\boldsymbol{\theta}}(t) + \beta\frac{\mathrm{d}}{\mathrm{d}t}\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(t))) + \nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(t))) = 0 \qquad \text{(ISEHD)}$$

$$\sigma_{\mathbf{A}} \stackrel{\text{def}}{=} \inf_{\mathbf{z}\in\mathrm{Ker}(\mathbf{A})^{\perp}} \|\mathbf{A}\mathbf{z}\| \,/\, \|\mathbf{z}\| > 0.$$

Theorem *Let $\boldsymbol{\theta}(\cdot)$ be a solution trajectory of (ISEHD) with $\alpha = \sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}$ and $\beta = \frac{1}{2\alpha}$ where the initialization $\boldsymbol{\theta}_0$ is such that*

$$\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0)) > 0 \quad \textit{and} \quad R' < R,$$

where R' and R obey

$$R' = \eta\sqrt{\xi\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))} \quad \textit{and} \quad R = \frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))}{2\mathrm{Lip}_{\mathbb{B}(\boldsymbol{\theta}_0,R)}(\mathcal{J}_{\mathbf{g}})}$$

Non-degenerate initialization

with

$$\xi = 1 + \frac{\kappa(\mathcal{J}_{\mathbf{g}}(0))^2\kappa(\mathbf{A})^2}{4} \quad \textit{and} \quad \eta = \frac{4\max\left(\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}, \frac{1+\sqrt{2}}{2}\right)}{\min\left(\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))^2\sigma_{\mathbf{A}}^2, \frac{3}{4}\right)}.$$

Then, the following holds :

(i) the loss converges to 0 at the rate

$$\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)) \le \xi\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))\exp\left(-\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}}{2}t\right).$$

Moreover, $\boldsymbol{\theta}(t)$ converges to a global minimizer $\boldsymbol{\theta}_\infty$ at the rate

$$\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}_\infty\| \le \eta\sqrt{\xi\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))}\exp\left(-\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}}{4}t\right).$$

Trajectory close to initialization

(ii) We have

$$\|\mathbf{y}(t) - \overline{\mathbf{y}}\| \le 2\|\varepsilon\| \quad \textit{when} \quad t \ge \frac{4}{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}}\ln\left(\frac{\sqrt{2\xi\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))}}{\|\varepsilon\|}\right)$$

Implicit regularization
Stable recovery by early stopping

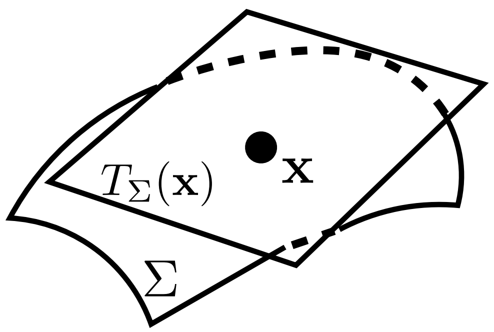
Recovery guarantees: x space

$$\sigma_{\mathbf{A}} = \inf_{\mathbf{z} \in \mathrm{Ker}(\mathbf{A})^{\perp}} \|\mathbf{A}\mathbf{z}\| \,/\, \|\mathbf{z}\| > 0.$$

$$\lambda_{\min}(\mathbf{A}; T_{\Sigma}(\mathbf{x})) = \inf\{\|\mathbf{A}\mathbf{z}\| \,/\, \|\mathbf{z}\| : \mathbf{z} \in T_{\Sigma}(\overline{\mathbf{x}}_{\Sigma})\}.$$

$$\Sigma = \{\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$$

$$T_{\Sigma}(\mathbf{x}) = \overline{\mathrm{conv}}\left(\mathbb{R}_{+}(\Sigma - \mathbf{x})\right)$$



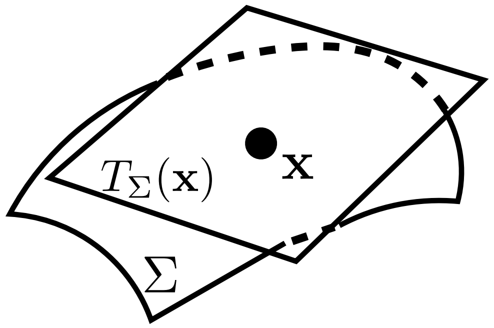
Recovery guarantees: x space

$$\sigma_{\mathbf{A}} = \inf_{\mathbf{z}\in\mathrm{Ker}(\mathbf{A})^\perp} \|\mathbf{A}\mathbf{z}\| \,/\, \|\mathbf{z}\| > 0.$$

$$\Sigma = \{\mathbf{g}(\mathbf{u},\boldsymbol{\theta}) : \; \boldsymbol{\theta} \in \Theta\}$$

$$\lambda_{\min}(\mathbf{A}; T_\Sigma(\mathbf{x})) = \inf\{\|\mathbf{A}\mathbf{z}\| \,/\, \|\mathbf{z}\| : \mathbf{z} \in T_\Sigma(\overline{\mathbf{x}}_\Sigma)\}.$$

$$T_\Sigma(\mathbf{x}) = \overline{\mathrm{conv}}\left(\mathbb{R}_+(\Sigma - \mathbf{x})\right)$$



Theorem *Assume the same assumptions on the parameters and initialization as above. If, moreover,*

$$\ker(\mathbf{A}) \cap T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}) = \{0\} \quad \textit{with} \quad \Sigma' \stackrel{\text{def}}{=} \Sigma_{\mathbb{B}_{R'+\|\boldsymbol{\theta}_0\|}},$$

then

$$\|\mathbf{x}(t) - \overline{\mathbf{x}}\| \leq \frac{\sqrt{2\xi\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))}\exp\left(-\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}}{4}t\right)}{\lambda_{\min}(\mathbf{A}; T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))} + \left(1 + \frac{\|\mathbf{A}\|}{\lambda_{\min}(\mathbf{A}; T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))}\right)\mathrm{dist}(\overline{\mathbf{x}}, \Sigma') + \frac{\|\varepsilon\|}{\lambda_{\min}(\mathbf{A}; T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))}.$$

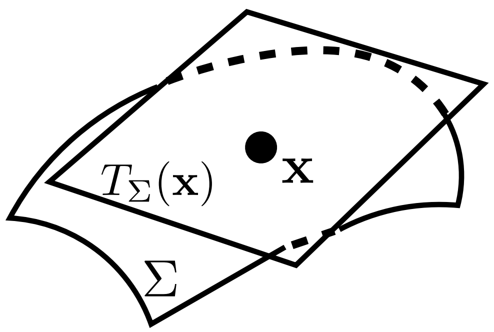
Recovery guarantees: x space

$$\sigma_{\mathbf{A}} = \inf_{\mathbf{z}\in \mathrm{Ker}(\mathbf{A})^{\perp}} \|\mathbf{A}\mathbf{z}\| \,/\, \|\mathbf{z}\| > 0.$$

$$\Sigma = \{\mathbf{g}(\mathbf{u},\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$$

$$\lambda_{\min}(\mathbf{A}; T_{\Sigma}(\mathbf{x})) = \inf\{\|\mathbf{A}\mathbf{z}\| \,/\, \|\mathbf{z}\| : \mathbf{z} \in T_{\Sigma}(\overline{\mathbf{x}}_{\Sigma})\}.$$

$$T_{\Sigma}(\mathbf{x}) = \overline{\mathrm{conv}}\left(\mathbb{R}_+(\Sigma - \mathbf{x})\right)$$



Theorem *Assume the same assumptions on the parameters and initialization as above. If, moreover,*

$$\ker(\mathbf{A}) \cap T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}) = \{0\} \quad \textit{with} \quad \Sigma' \stackrel{\mathrm{def}}{=} \Sigma_{\mathbb{B}_{R'+\|\boldsymbol{\theta}_0\|}},$$

Restricted Injectivity Condition (RIC)

then

$$\|\mathbf{x}(t) - \overline{\mathbf{x}}\| \leq \frac{\sqrt{2\xi\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))}\exp\left(-\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}}{4}t\right)}{\lambda_{\min}(\mathbf{A}; T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))} + \left(1 + \frac{\|\mathbf{A}\|}{\lambda_{\min}(\mathbf{A}; T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))}\right)\mathrm{dist}(\overline{\mathbf{x}}, \Sigma') + \frac{\|\varepsilon\|}{\lambda_{\min}(\mathbf{A}; T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))}.$$

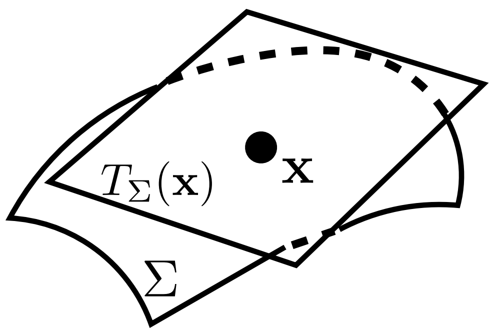
Recovery guarantees: x space

$$\sigma_{\mathbf{A}} = \inf_{\mathbf{z} \in \mathrm{Ker}(\mathbf{A})^\perp} \|\mathbf{A}\mathbf{z}\| / \|\mathbf{z}\| > 0.$$

$$\Sigma = \{\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$$

$$\lambda_{\min}(\mathbf{A}; T_\Sigma(\mathbf{x})) = \inf\{\|\mathbf{A}\mathbf{z}\| / \|\mathbf{z}\| : \mathbf{z} \in T_\Sigma(\overline{\mathbf{x}}_\Sigma)\}.$$

$$T_\Sigma(\mathbf{x}) = \overline{\mathrm{conv}}\left(\mathbb{R}_+(\Sigma - \mathbf{x})\right)$$



Theorem *Assume the same assumptions on the parameters and initialization as above. If, moreover,*

$$\ker(\mathbf{A}) \cap T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}) = \{0\} \quad \textit{with} \quad \Sigma' \stackrel{\text{def}}{=} \Sigma_{\mathbb{B}_{R' + \|\boldsymbol{\theta}_0\|}},$$

Restricted Injectivity Condition (RIC)

then

$$\|\mathbf{x}(t) - \overline{\mathbf{x}}\| \leq \frac{\sqrt{2\xi \mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))} \exp\left(-\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}}{4} t\right)}{\lambda_{\min}(\mathbf{A}; T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))} + \left(1 + \frac{\|\mathbf{A}\|}{\lambda_{\min}(\mathbf{A}; T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))}\right) \mathrm{dist}(\overline{\mathbf{x}}, \Sigma') + \frac{\|\varepsilon\|}{\lambda_{\min}(\mathbf{A}; T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))}.$$

Optimization error

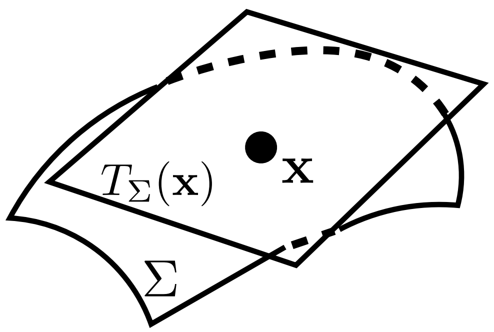
Recovery guarantees: x space

$$\sigma_{\mathbf{A}} = \inf_{\mathbf{z} \in \mathrm{Ker}(\mathbf{A})^{\perp}} \|\mathbf{A}\mathbf{z}\| \,/\, \|\mathbf{z}\| > 0.$$

$$\Sigma = \{\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}) : \ \boldsymbol{\theta} \in \Theta\}$$

$$\lambda_{\min}(\mathbf{A}; T_{\Sigma}(\mathbf{x})) = \inf\{\|\mathbf{A}\mathbf{z}\| \,/\, \|\mathbf{z}\| : \mathbf{z} \in T_{\Sigma}(\overline{\mathbf{x}}_{\Sigma})\}.$$

$$T_{\Sigma}(\mathbf{x}) = \overline{\mathrm{conv}}\left(\mathbb{R}_{+}(\Sigma - \mathbf{x})\right)$$



Theorem *Assume the same assumptions on the parameters and initialization as above. If, moreover,*

$$\ker(\mathbf{A}) \cap T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}) = \{0\} \quad \textit{with} \quad \Sigma' \stackrel{\text{def}}{=} \Sigma_{\mathbb{B}_{R' + \|\boldsymbol{\theta}_0\|}},$$

Restricted Injectivity Condition (RIC)

then

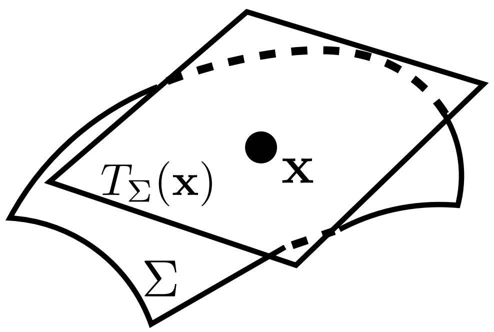
$$\|\mathbf{x}(t) - \overline{\mathbf{x}}\| \le \frac{\sqrt{2\xi \mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))} \exp\left(-\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}}{4} t\right)}{\lambda_{\min}(\mathbf{A}; T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))} + \left(1 + \frac{\|\mathbf{A}\|}{\lambda_{\min}(\mathbf{A}; T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))}\right) \mathrm{dist}(\overline{\mathbf{x}}, \Sigma') + \frac{\|\varepsilon\|}{\lambda_{\min}(\mathbf{A}; T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))}.$$

Optimization error   Approximation error

Recovery guarantees: x space

$$\sigma_{\mathbf{A}} = \inf_{\mathbf{z} \in \mathrm{Ker}(\mathbf{A})^{\perp}} \|\mathbf{A}\mathbf{z}\| \,/\, \|\mathbf{z}\| > 0. \qquad \Sigma = \{\mathbf{g}(\mathbf{u},\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$$

$$\lambda_{\min}(\mathbf{A}; T_{\Sigma}(\mathbf{x})) = \inf\{\|\mathbf{A}\mathbf{z}\| \,/\, \|\mathbf{z}\| : \mathbf{z} \in T_{\Sigma}(\overline{\mathbf{x}}_{\Sigma})\}. \qquad T_{\Sigma}(\mathbf{x}) = \overline{\mathrm{conv}}\left(\mathbb{R}_+(\Sigma - \mathbf{x})\right)$$



Theorem *Assume the same assumptions on the parameters and initialization as above. If, moreover,*

$$\ker(\mathbf{A}) \cap T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}) = \{0\} \quad \textit{with} \quad \Sigma' \stackrel{\mathrm{def}}{=} \Sigma_{\mathbb{B}_{R' + \|\boldsymbol{\theta}_0\|}},$$

Restricted Injectivity Condition (RIC)

then

$$\|\mathbf{x}(t) - \overline{\mathbf{x}}\| \leq \frac{\sqrt{2\xi\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))}\exp\left(-\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}}{4}t\right)}{\lambda_{\min}(\mathbf{A}; T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))} + \left(1 + \frac{\|\mathbf{A}\|}{\lambda_{\min}(\mathbf{A}; T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))}\right)\mathrm{dist}(\overline{\mathbf{x}}, \Sigma') + \frac{\|\varepsilon\|}{\lambda_{\min}(\mathbf{A}; T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))}.$$

Optimization error — Approximation error — Noise error

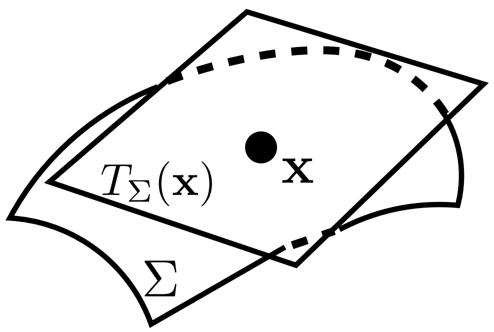
Recovery guarantees: x space

$$\sigma_{\mathbf{A}} = \inf_{\mathbf{z}\in \mathrm{Ker}(\mathbf{A})^\perp} \|\mathbf{A}\mathbf{z}\| \,/\, \|\mathbf{z}\| > 0.$$

$$\Sigma = \{\mathbf{g}(\mathbf{u},\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$$

$$\lambda_{\min}(\mathbf{A}; T_\Sigma(\mathbf{x})) = \inf\{\|\mathbf{A}\mathbf{z}\| \,/\, \|\mathbf{z}\| : \mathbf{z} \in T_\Sigma(\overline{\mathbf{x}}_\Sigma)\}.$$

$$T_\Sigma(\mathbf{x}) = \overline{\mathrm{conv}}\left(\mathbb{R}_+(\Sigma - \mathbf{x})\right)$$



Theorem *Assume the same assumptions on the parameters and initialization as above. If, moreover,*

$$\ker(\mathbf{A}) \cap T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}) = \{0\} \quad \textit{with} \quad \Sigma' \stackrel{\text{def}}{=} \Sigma_{\mathbb{B}_{R'+\|\boldsymbol{\theta}_0\|}},$$

Restricted Injectivity Condition (RIC)

then

$$\|\mathbf{x}(t) - \overline{\mathbf{x}}\| \le \frac{\sqrt{2\xi\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))}\exp\left(-\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}}{4}t\right)}{\lambda_{\min}(\mathbf{A}; T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))} + \left(1 + \frac{\|\mathbf{A}\|}{\lambda_{\min}(\mathbf{A}; T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))}\right)\mathrm{dist}(\overline{\mathbf{x}}, \Sigma') + \frac{\|\varepsilon\|}{\lambda_{\min}(\mathbf{A}; T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))}.$$

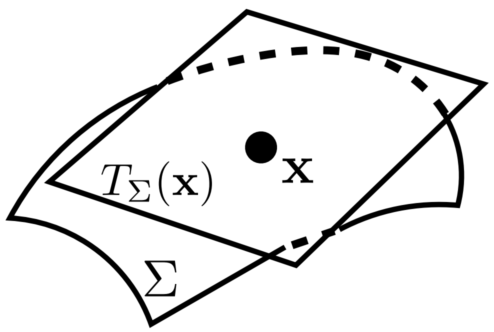
Optimization error — Approximation error — Noise error

- Sample bounds for $\lambda_{\min}$ can be given in a compressed sensing framework via the Gaussian width of the tangent cone.

Recovery guarantees: x space

$$\sigma_{\mathbf{A}} = \inf_{\mathbf{z}\in\mathrm{Ker}(\mathbf{A})^{\perp}} \|\mathbf{A}\mathbf{z}\| \,/\, \|\mathbf{z}\| > 0. \qquad \Sigma = \{\mathbf{g}(\mathbf{u},\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$$

$$\lambda_{\min}(\mathbf{A}; T_{\Sigma}(\mathbf{x})) = \inf\{\|\mathbf{A}\mathbf{z}\| \,/\, \|\mathbf{z}\| : \mathbf{z} \in T_{\Sigma}(\overline{\mathbf{x}}_{\Sigma})\}. \qquad T_{\Sigma}(\mathbf{x}) = \overline{\mathrm{conv}}\,(\mathbb{R}_{+}(\Sigma - \mathbf{x}))$$



Theorem *Assume the same assumptions on the parameters and initialization as above. If, moreover,*

$$\ker(\mathbf{A}) \cap T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}) = \{0\} \quad \textit{with} \quad \Sigma' \stackrel{\mathrm{def}}{=} \Sigma_{\mathbb{B}_{R'+\|\boldsymbol{\theta}_0\|}},$$

Restricted Injectivity Condition (RIC)

then

$$\|\mathbf{x}(t) - \overline{\mathbf{x}}\| \leq \frac{\sqrt{2\xi\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))}\exp\left(-\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}}{4}t\right)}{\lambda_{\min}(\mathbf{A}; T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))} + \left(1 + \frac{\|\mathbf{A}\|}{\lambda_{\min}(\mathbf{A}; T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))}\right)\mathrm{dist}(\overline{\mathbf{x}}, \Sigma') + \frac{\|\varepsilon\|}{\lambda_{\min}(\mathbf{A}; T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))}.$$

Optimization error — Approximation error — Noise error

- Sample bounds for $\lambda_{\min}$ can be given in a compressed sensing framework via the Gaussian width of the tangent cone.
- Trade-off between the expressivity of the model and the RIC.

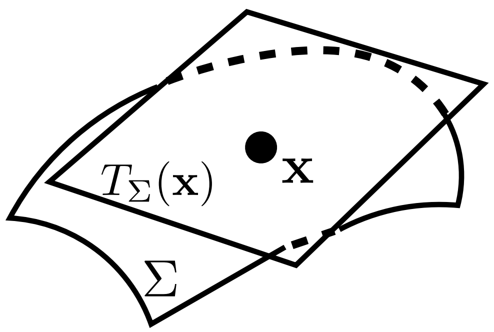
Recovery guarantees: x space

$$\sigma_{\mathbf{A}} = \inf_{\mathbf{z}\in \mathrm{Ker}(\mathbf{A})^{\perp}} \|\mathbf{A}\mathbf{z}\| \,/\, \|\mathbf{z}\| > 0.$$

$$\Sigma = \{\mathbf{g}(\mathbf{u},\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$$

$$\lambda_{\min}(\mathbf{A}; T_{\Sigma}(\mathbf{x})) = \inf\{\|\mathbf{A}\mathbf{z}\| \,/\, \|\mathbf{z}\| : \mathbf{z} \in T_{\Sigma}(\overline{\mathbf{x}}_{\Sigma})\}.$$

$$T_{\Sigma}(\mathbf{x}) = \overline{\mathrm{conv}}\left(\mathbb{R}_{+}(\Sigma - \mathbf{x})\right)$$



Theorem *Assume the same assumptions on the parameters and initialization as above. If, moreover,*

$$\ker(\mathbf{A}) \cap T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}) = \{0\} \quad \textit{with} \quad \Sigma' \stackrel{\text{def}}{=} \Sigma_{\mathbb{B}_{R'+\|\boldsymbol{\theta}_0\|}},$$

Restricted Injectivity Condition (RIC)

then

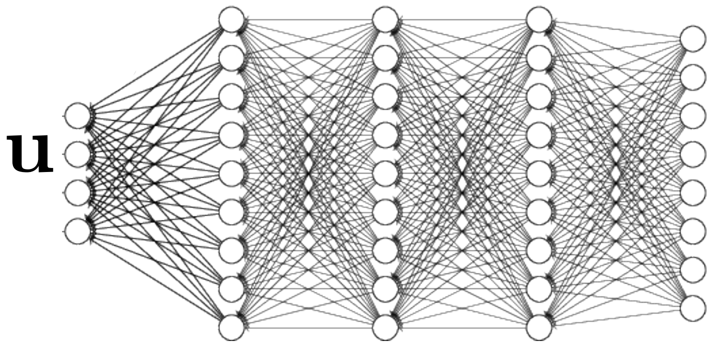
$$\|\mathbf{x}(t) - \overline{\mathbf{x}}\| \leq \frac{\sqrt{2\xi\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))}\exp\left(-\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}}{4}t\right)}{\lambda_{\min}(\mathbf{A}; T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))} + \left(1 + \frac{\|\mathbf{A}\|}{\lambda_{\min}(\mathbf{A}; T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))}\right)\mathrm{dist}(\overline{\mathbf{x}}, \Sigma') + \frac{\|\varepsilon\|}{\lambda_{\min}(\mathbf{A}; T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))}.$$

Optimization error — Approximation error — Noise error

- Sample bounds for $\lambda_{\min}$ can be given in a compressed sensing framework via the Gaussian width of the tangent cone.
- Trade-off between the expressivity of the model and the RIC.
- Optimization error of GF : $O\left(\exp\left(-\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))^2\sigma_{\mathbf{A}}^2}{4}t\right)\right)$.
- Optimization error of ISEHD : $O\left(\exp\left(-\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}}{4}t\right)\right)$.

Non-degenerate initialization

$\mathbf{g}(\mathbf{u}, \boldsymbol{\theta})$



$$\min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta})) \qquad \mathbf{A} \in \mathbb{R}^{m\times n}$$

$$\ddot{\boldsymbol{\theta}}(t) + \alpha\dot{\boldsymbol{\theta}}(t) + \beta\frac{\mathrm{d}}{\mathrm{d}t}\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(t))) + \nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(t))) = 0 \qquad \text{(ISEHD)}$$

Theorem *Let $\boldsymbol{\theta}(\cdot)$ be a solution trajectory of (ISEHD) with $\alpha = \sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}$ and $\beta = \frac{1}{2\alpha}$ where the initialization $\boldsymbol{\theta}_0$ is such that*

$$\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0)) > 0 \quad \textit{and} \quad R' < R,$$

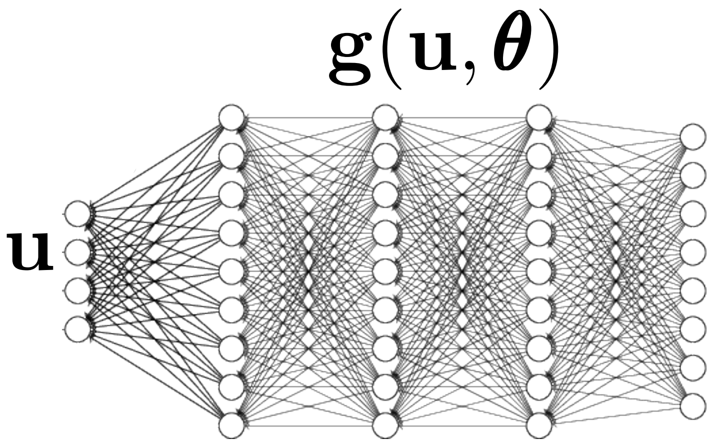
where R' and R obey

$$R' = \eta\sqrt{\xi\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))} \quad \textit{and} \quad R = \frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))}{2\mathrm{Lip}_{\mathbb{B}(\boldsymbol{\theta}_0,R)}(\mathcal{J}_{\mathbf{g}})}$$

Non-degenerate initialization

etc.

Non-degenerate initialization



$$\min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta})) \qquad \mathbf{A} \in \mathbb{R}^{m\times n}$$

$$\ddot{\boldsymbol{\theta}}(t) + \alpha\dot{\boldsymbol{\theta}}(t) + \beta\frac{\mathrm{d}}{\mathrm{d}t}\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(t))) + \nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(t))) = 0 \qquad \text{(ISEHD)}$$

Theorem *Let* $\boldsymbol{\theta}(\cdot)$ *be a solution trajectory of (ISEHD) with* $\alpha = \sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}$ *and* $\beta = \frac{1}{2\alpha}$ *where the initialization* $\boldsymbol{\theta}_0$ *is such that*

$$\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0)) > 0 \quad \textit{and} \quad R' < R,$$

where R' *and* R *obey*

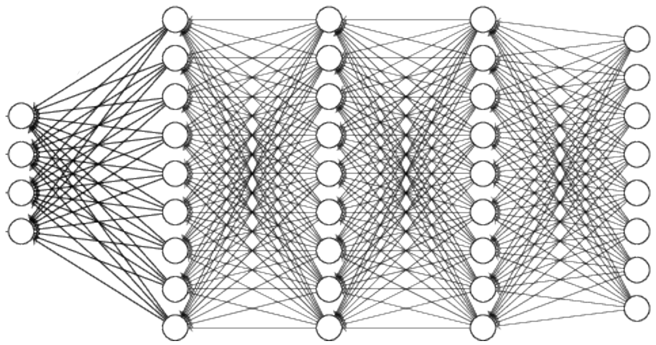
$$R' = \eta\sqrt{\xi\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))} \quad \textit{and} \quad R = \frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))}{2\mathrm{Lip}_{\mathbb{B}(\boldsymbol{\theta}_0,R)}(\mathcal{J}_{\mathbf{g}})}$$

Non-degenerate initialization

etc.

How to ensure this ?

Non-degenerate initialization



$$\min_{\boldsymbol{\theta} \in \Theta} \mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta})) \qquad \mathbf{A} \in \mathbb{R}^{m \times n}$$

$$\ddot{\boldsymbol{\theta}}(t) + \alpha \dot{\boldsymbol{\theta}}(t) + \beta \frac{\mathrm{d}}{\mathrm{d}t} \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}(t))) + \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}(t))) = 0 \qquad \text{(ISEHD)}$$

Theorem *Let $\boldsymbol{\theta}(\cdot)$ be a solution trajectory of (ISEHD) with $\alpha = \sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}$ and $\beta = \frac{1}{2\alpha}$ where the initialization $\boldsymbol{\theta}_0$ is such that*

$$\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0)) > 0 \quad \textit{and} \quad R' < R,$$

where R' and R obey

$$R' = \eta \sqrt{\xi \mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))} \quad \textit{and} \quad R = \frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))}{2\mathrm{Lip}_{\mathbb{B}(\boldsymbol{\theta}_0, R)}(\mathcal{J}_{\mathbf{g}})}$$

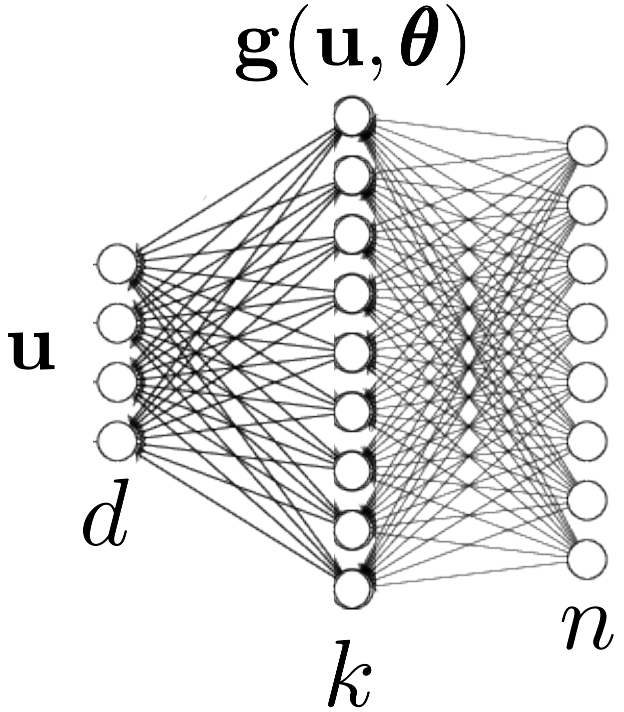
Non-degenerate initialization

etc.

How to ensure this ?

The role of overparametrization

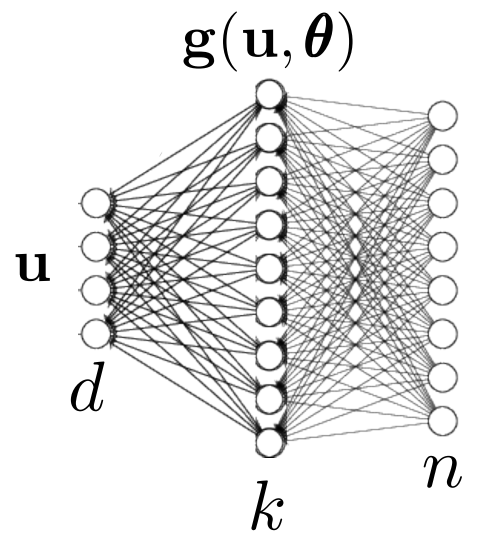
Wide two-layer DIP



$$\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}) = \frac{1}{\sqrt{k}} \mathbf{V} \phi(\mathbf{W}\mathbf{u})$$

- $\mathbf{u}$ uniform vector on $\mathbb{S}^{d-1}$.
- $\mathbf{W}(0)$ has iid $\mathcal{N}(0, 1)$ entries.
- $\mathbf{V}(0)$ independent from $\mathbf{W}(0)$ and $\mathbf{u}$, and its entries are zero-mean independent D-bounded random variables of unit variance.

Overparametrization bound



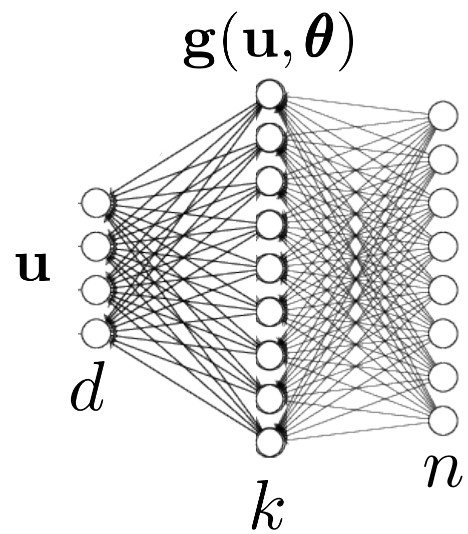
$$\mathbf{g}(\mathbf{u},\boldsymbol{\theta}) = \frac{1}{\sqrt{k}}\mathbf{V}\phi(\mathbf{W}\mathbf{u})$$

Theorem *Consider the one-hidden layer DIP network with the architecture parameters where both layers are trained with the architecture parameters obeying*

$$k \gtrsim (1+\kappa(\mathbf{A})^4)\frac{\max\left(\sigma_{\mathbf{A}}^4, c_1\right)}{\min\left(\sigma_{\mathbf{A}}^8, c_2\right)} n \left(\left\|\mathbf{A}\right\|^4 n^2 + \left(1+\mathrm{SNR}^{-1}\right)^4 m^2\right).$$

Then with probability at least $1-5e^{-(n-1)}-2n^{-1}$, $\boldsymbol{\theta}(0) = (\mathbf{W}(0), \mathbf{V}(0))$ *is a non-degenerate initial point. Here* c_1, c_2 *are absolute constants.*

Overparametrization bound



$$\mathbf{g}(\mathbf{u},\boldsymbol{\theta}) = \frac{1}{\sqrt{k}}\mathbf{V}\phi(\mathbf{W}\mathbf{u})$$

Theorem *Consider the one-hidden layer DIP network with the architecture parameters where both layers are trained with the architecture parameters obeying*

$$k \gtrsim (1+\kappa(\mathbf{A})^4)\frac{\max\left(\sigma_{\mathbf{A}}^4, c_1\right)}{\min\left(\sigma_{\mathbf{A}}^8, c_2\right)} n \left(\|\mathbf{A}\|^4 n^2 + \left(1+\mathrm{SNR}^{-1}\right)^4 m^2\right).$$

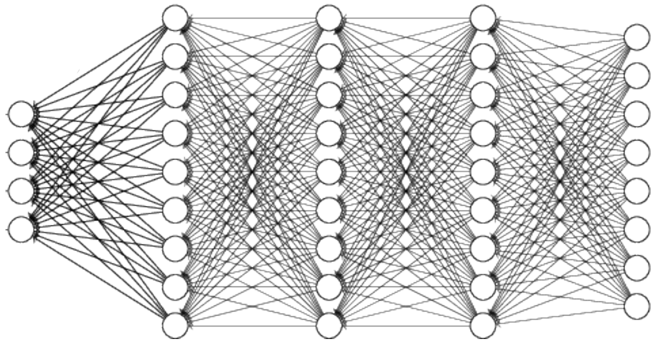
Then with probability at least $1-5e^{-(n-1)}-2n^{-1}$, $\boldsymbol{\theta}(0) = (\mathbf{W}(0), \mathbf{V}(0))$ *is a non-degenerate initial point. Here* c_1, c_2 *are absolute constants.*

- The bound scales as $k \gtrsim n^3 + nm^2$.
- Improved to $k \gtrsim n^2 m$ if $\mathbf{V}$ is fixed and only is $\mathbf{W}$ is optimized.
- (ISEHD) achieves an optimal exponential rate but at the price of a more stringent condition on compared to GF.

What about (IGAHD)

$$\mathbf{y} = \mathbf{A}\overline{\mathbf{x}} + \varepsilon \qquad \mathbf{A} \in \mathbb{R}^{m \times n}$$

$\mathbf{g}(\mathbf{u}, \boldsymbol{\theta})$



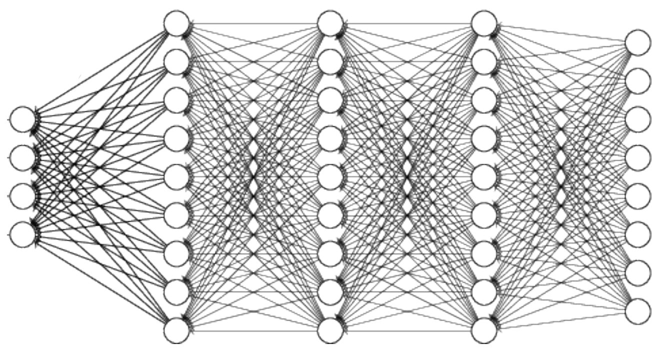
$$\min_{\boldsymbol{\theta} \in \Theta} \mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}))$$

$$\text{(IGAHD)} \begin{cases} \boldsymbol{\eta}_\ell &= \boldsymbol{\theta}_\ell + \alpha s_\ell(\boldsymbol{\theta}_\ell - \boldsymbol{\theta}_{\ell-1}) - \beta s_\ell^2 \left(\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}_\ell)) - \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}_{\ell-1}))\right), \\ \boldsymbol{\theta}_{\ell+1} &= \boldsymbol{\eta}_\ell - s_\ell \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}_\ell)). \end{cases}$$

What about (IGAHD)

$$\mathbf{y} = \mathbf{A}\overline{\mathbf{x}} + \varepsilon \qquad \mathbf{A} \in \mathbb{R}^{m \times n}$$

$\mathbf{g}(\mathbf{u}, \boldsymbol{\theta})$



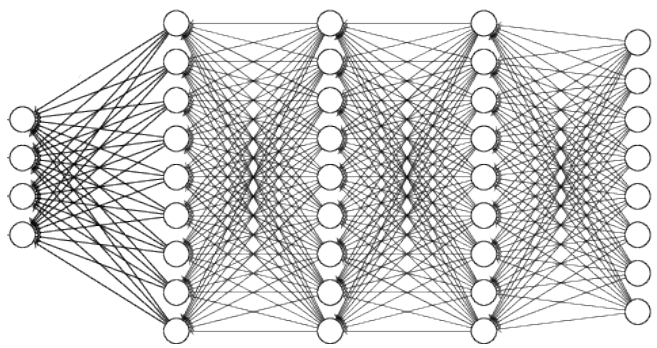
$$\min_{\boldsymbol{\theta} \in \Theta} \mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}))$$

$$\text{(IGAHD)} \begin{cases} \boldsymbol{\eta}_\ell & = \boldsymbol{\theta}_\ell + \alpha s_\ell(\boldsymbol{\theta}_\ell - \boldsymbol{\theta}_{\ell-1}) - \beta s_\ell^2 \left(\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}_\ell)) - \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}_{\ell-1}))\right), \\ \boldsymbol{\theta}_{\ell+1} & = \boldsymbol{\eta}_\ell - s_\ell \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}_\ell)). \end{cases}$$

Beware of local Lipschitz continuity only of $\mathbf{g}(\mathbf{u},.)$.

What about (IGAHD)

$$\mathbf{y} = \mathbf{A}\overline{\mathbf{x}} + \varepsilon \qquad \mathbf{A} \in \mathbb{R}^{m\times n}$$



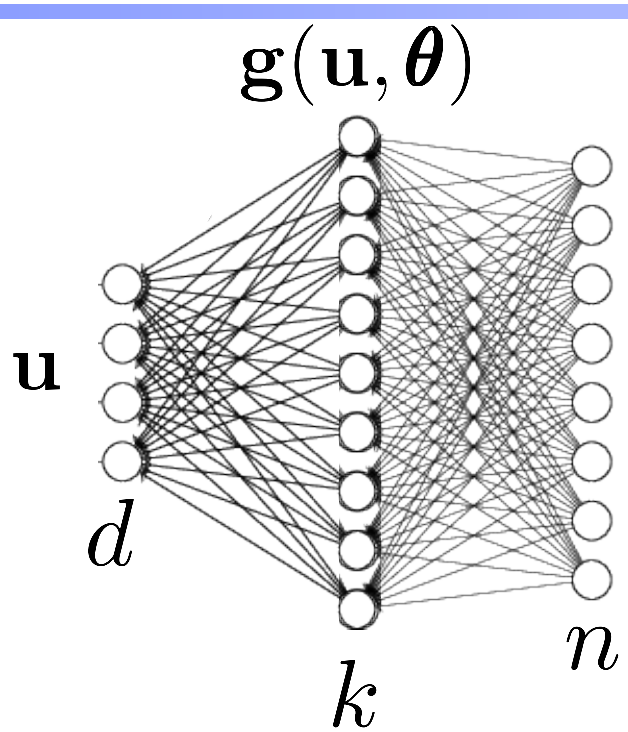
$$\min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}))$$

$$\text{(IGAHD)}\begin{cases} \boldsymbol{\eta}_\ell & = \boldsymbol{\theta}_\ell + \alpha s_\ell(\boldsymbol{\theta}_\ell - \boldsymbol{\theta}_{\ell-1}) - \beta s_\ell^2 \left(\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}_\ell)) - \nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}_{\ell-1}))\right), \\ \boldsymbol{\theta}_{\ell+1} & = \boldsymbol{\eta}_\ell - s_\ell \nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}_\ell)). \end{cases}$$

Beware of local Lipschitz continuity only of $\mathbf{g}(\mathbf{u},.)$.

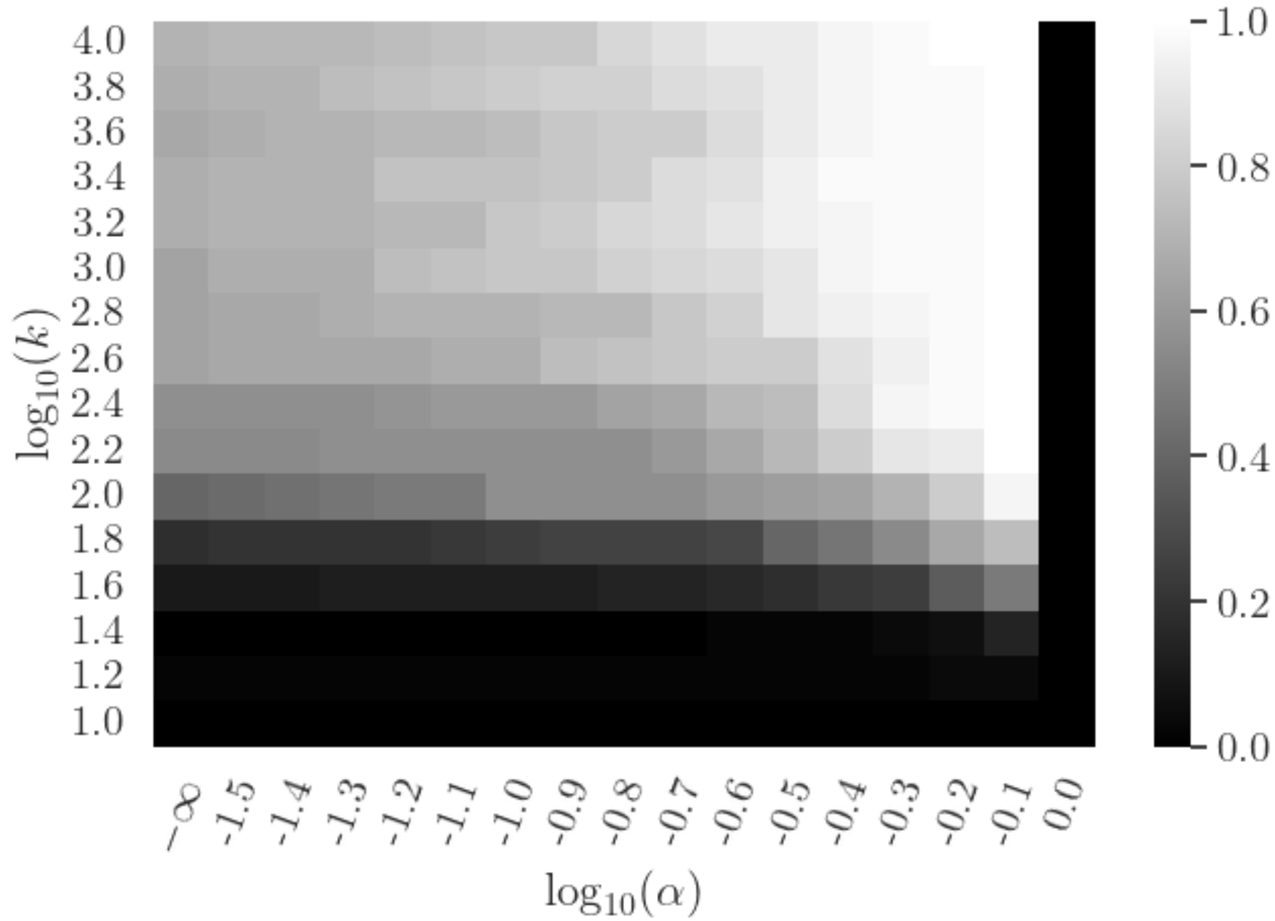
Similar guarantees hold with a backtracking procedure within (IGAHD)

Flexibility of (IGAHD)



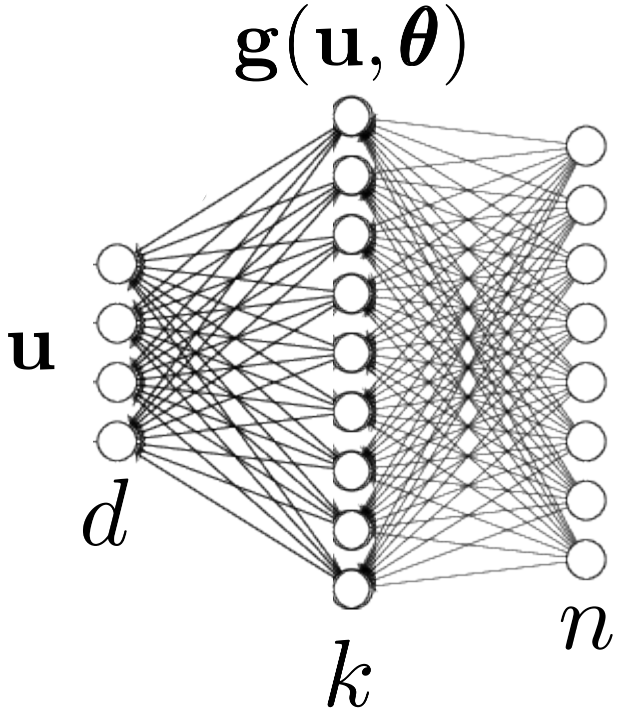
$\mathbf{A}_{ij}$ iid $\mathcal{N}(0, 1/m)$

$$\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}) = \frac{1}{\sqrt{k}} \mathbf{V} \phi(\mathbf{W}\mathbf{u})$$



Empirical probability of (IGAHD) to achieve numerical accuracy over the loss in less than 15000 iterations for varying (k,α). β=0.05.

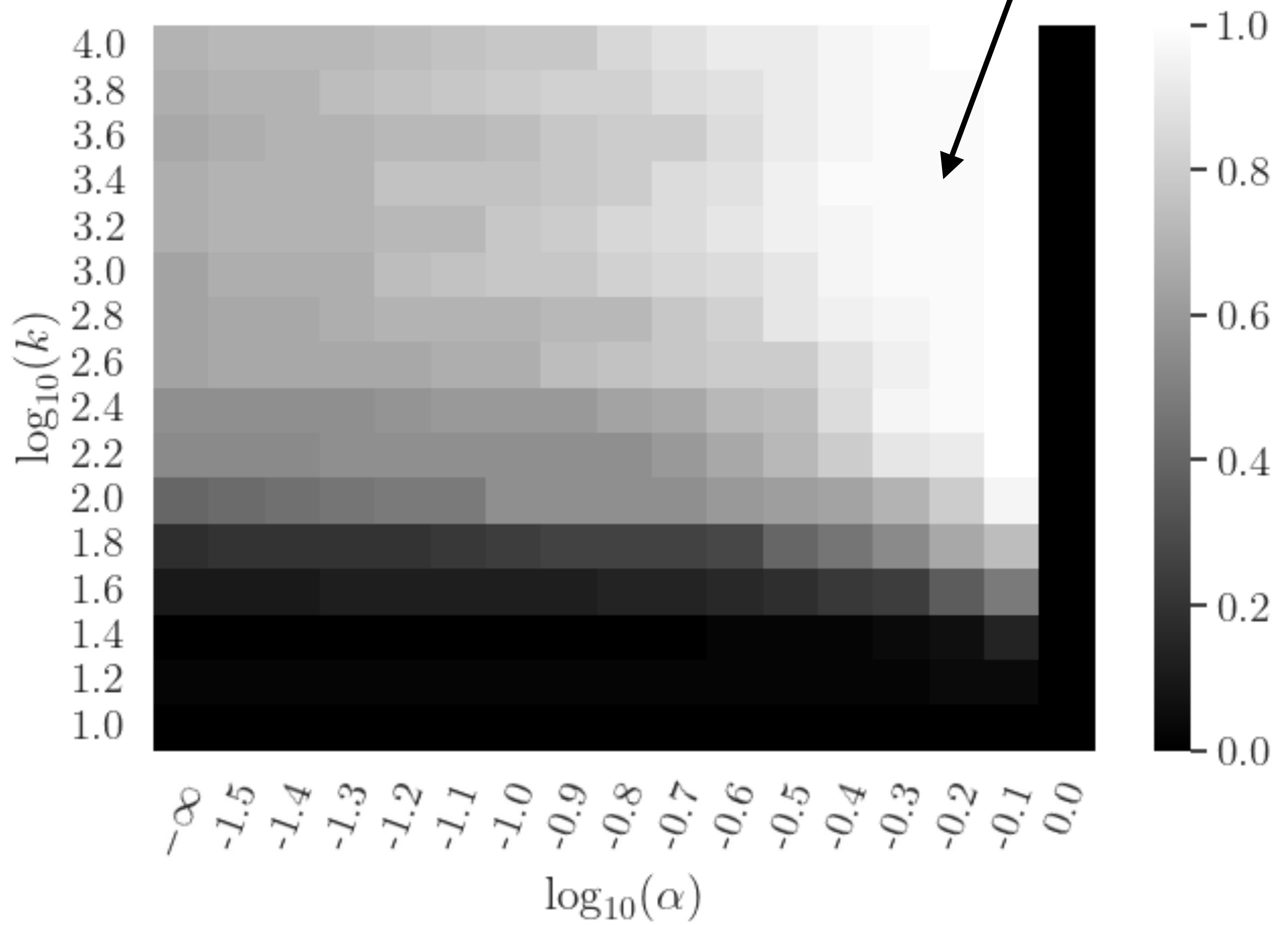
Flexibility of (IGAHD)



$\mathbf{A}_{ij}$ iid $\mathcal{N}(0,1/m)$

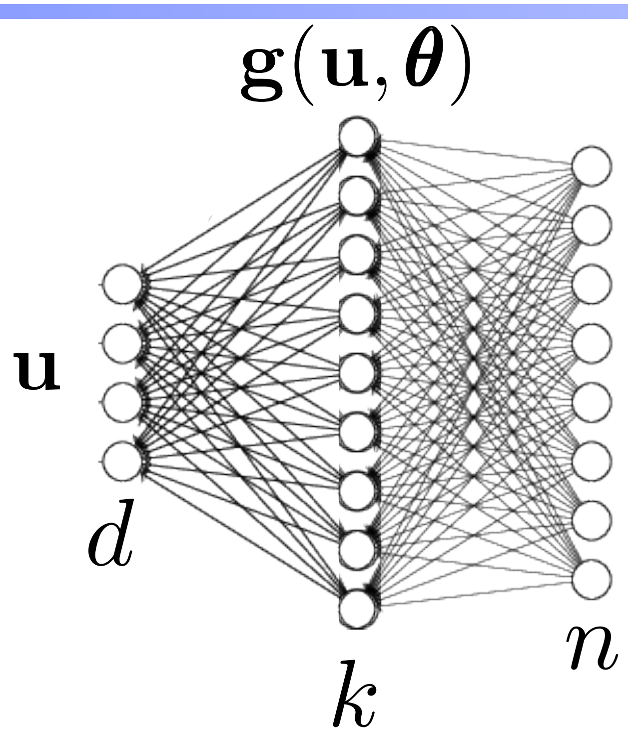
$$\mathbf{g}(\mathbf{u},\boldsymbol{\theta}) = \frac{1}{\sqrt{k}}\mathbf{V}\phi(\mathbf{W}\mathbf{u})$$

Substantial gain in the overparametrized regime



Empirical probability of (IGAHD) to achieve numerical accuracy over the loss in less than 15000 iterations for varying (k,α). β=0.05.

Flexibility of (IGAHD)

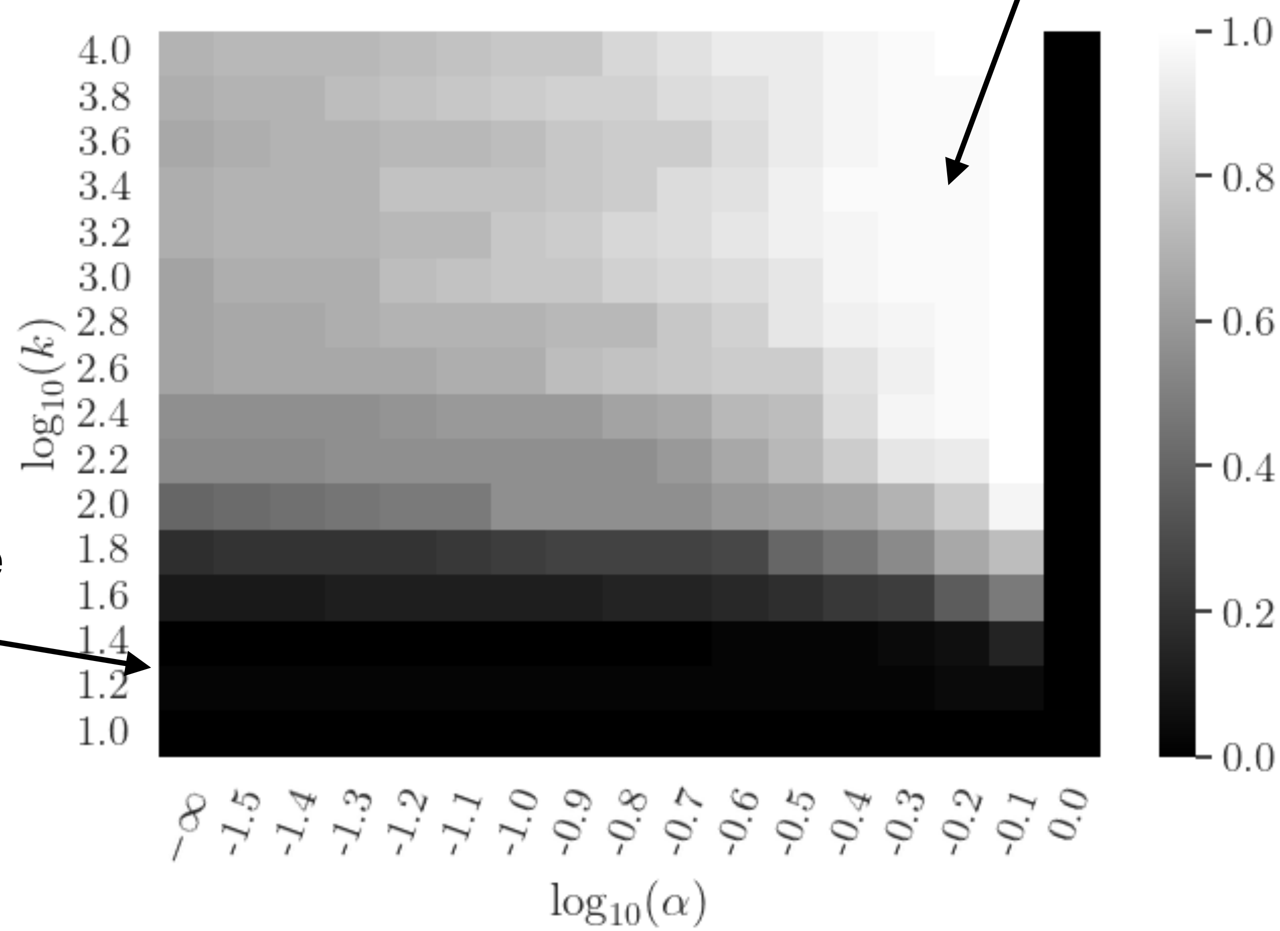


$$\mathbf{A}_{ij} \text{ iid } \mathcal{N}(0, 1/m)$$

$$\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}) = \frac{1}{\sqrt{k}} \mathbf{V} \phi(\mathbf{W}\mathbf{u})$$

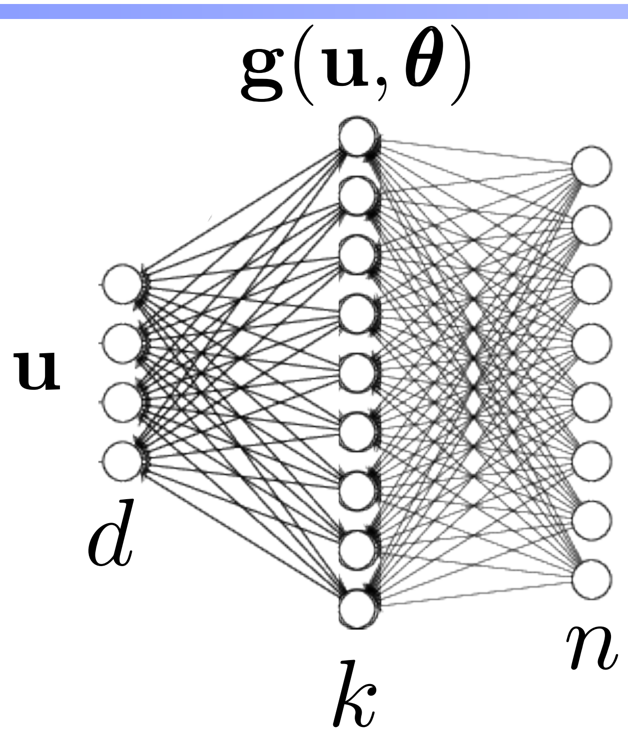
Substantial gain in the overparametrized regime

No gain in the underparametrized regime



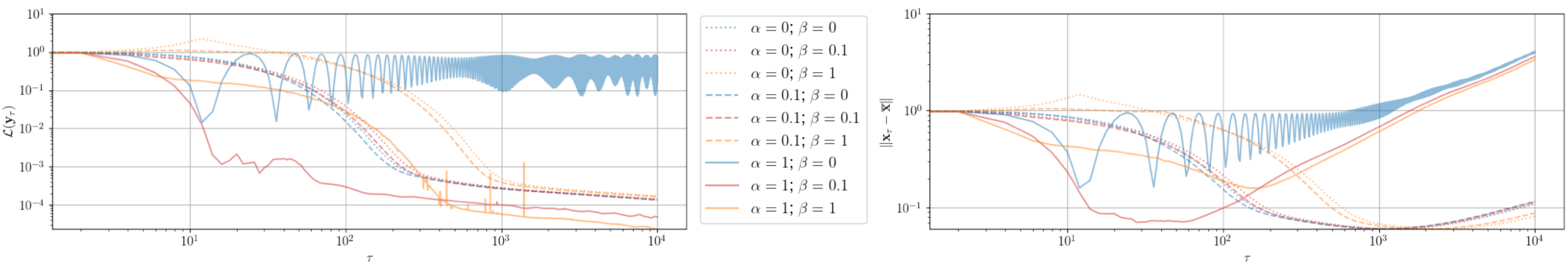
Empirical probability of (IGAHD) to achieve numerical accuracy over the loss in less than 15000 iterations for varying (k,α). β=0.05.

Flexibility of (IGAHD)

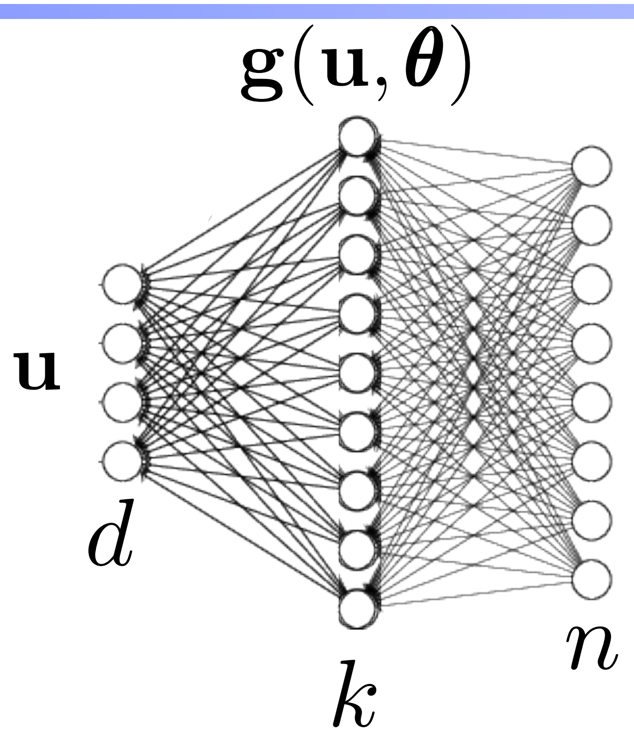


$$\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}) = \frac{1}{\sqrt{k}} \mathbf{V} \phi(\mathbf{W}\mathbf{u})$$

Well-adjusted parameters: acceleration and oscillation reduction.

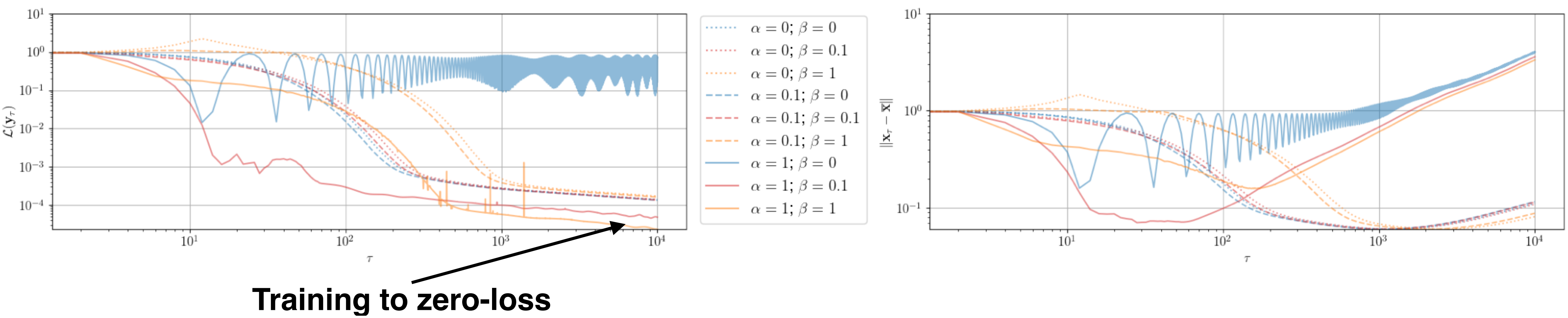


Flexibility of (IGAHD)

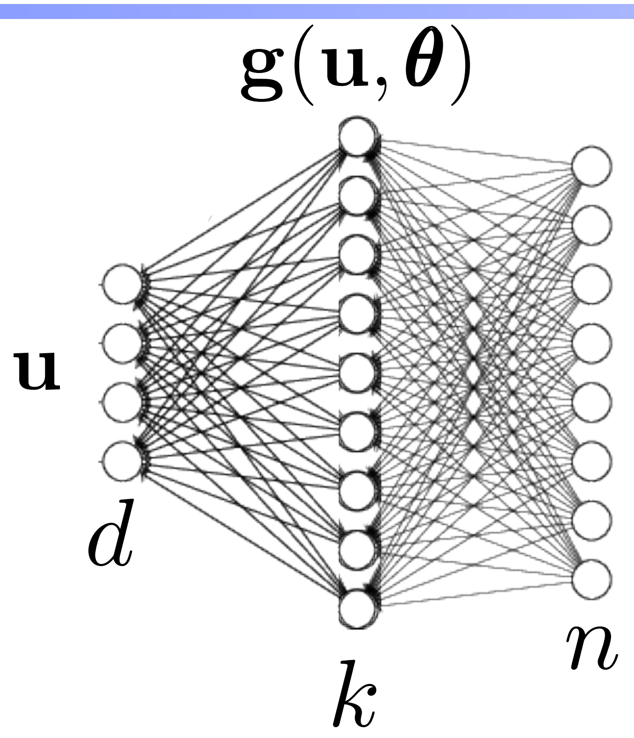


$$\mathbf{g}(\mathbf{u},\boldsymbol{\theta}) = \frac{1}{\sqrt{k}}\mathbf{V}\phi(\mathbf{W}\mathbf{u})$$

Well-adjusted parameters:
acceleration and oscillation reduction.

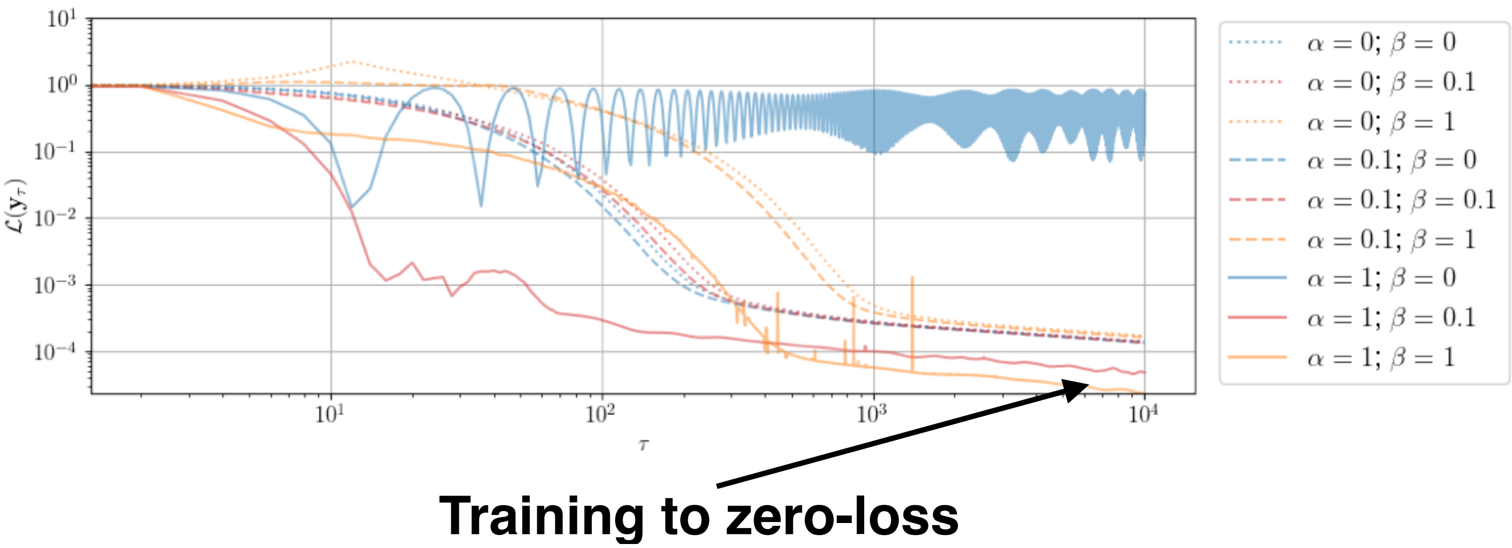


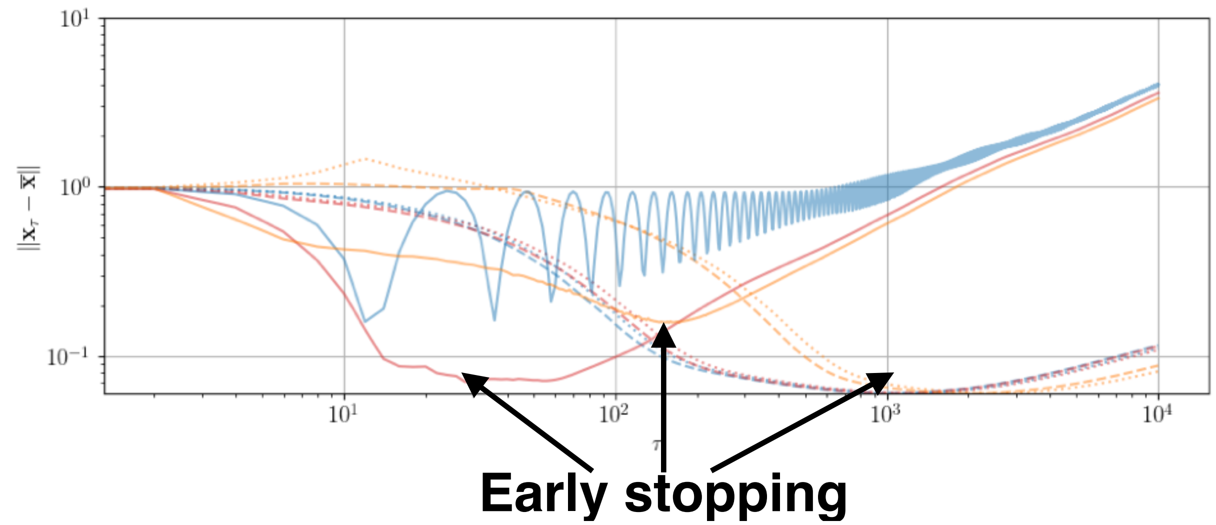
Flexibility of (IGAHD)



$$\mathbf{g}(\mathbf{u},\boldsymbol{\theta}) = \frac{1}{\sqrt{k}}\mathbf{V}\phi(\mathbf{W}\mathbf{u})$$

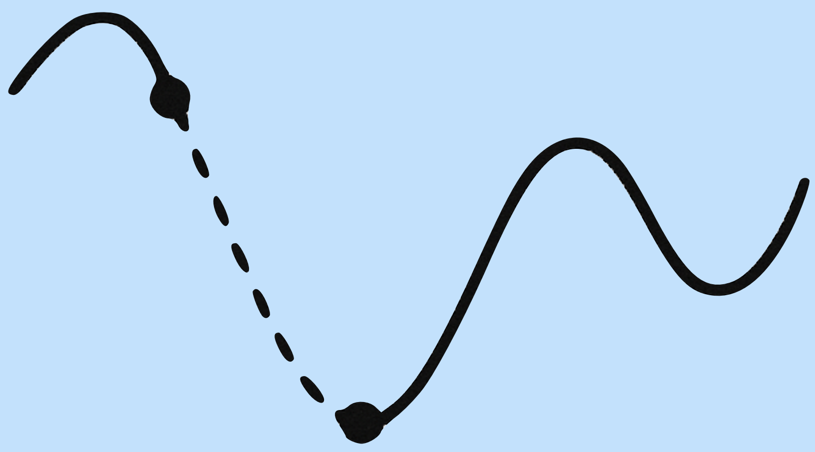
Well-adjusted parameters:
acceleration and oscillation reduction.



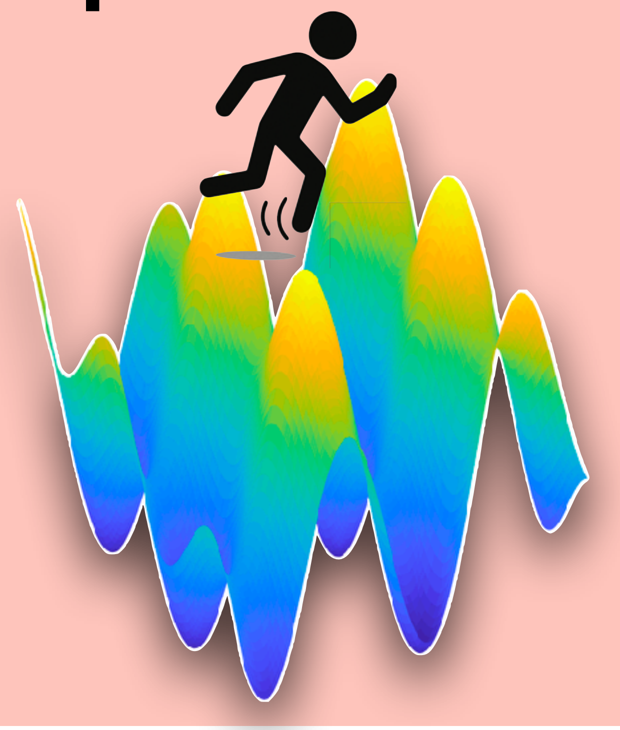


Outline

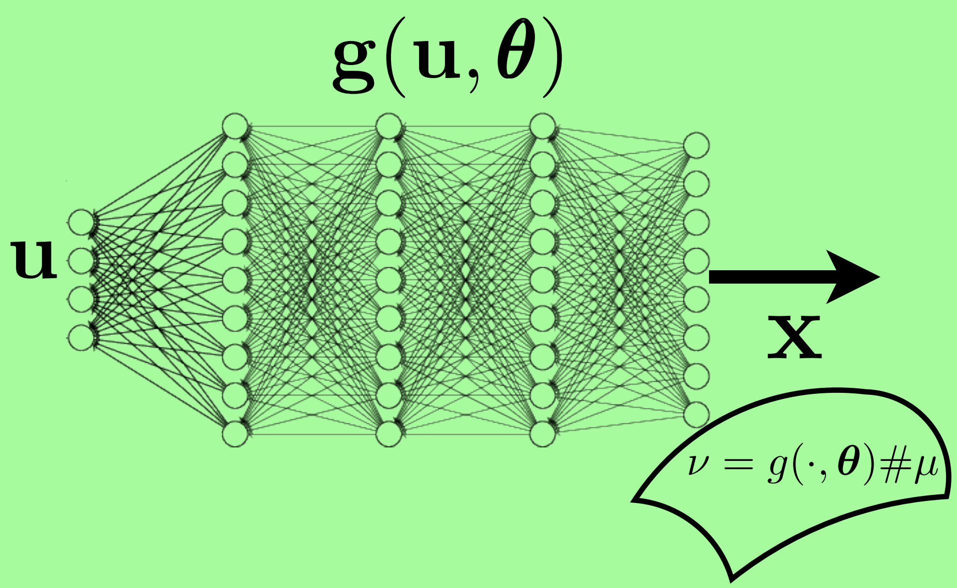
Convergence



Trap avoidance



DIP recovery guarantees



Conclusion

Outline

Conclusion

Take away messages

Take away messages

- **Inertia** (viscous and geometric) is good even for **non-convex** problems if **properly** used.

Take away messages

- **Inertia** (viscous and geometric) is good even for **non-convex** problems if **properly** used.
- **Convergence** and **trap avoidance**.

Take away messages

- **Inertia** (viscous and geometric) is good even for **non-convex** problems if **properly** used.
- **Convergence** and **trap avoidance**.
- Impact on r**ecovery guarantees** of DIP when optimized with inertia.

Take away messages

- **Inertia** (viscous and geometric) is good even for **non-convex** problems if **properly** used.
- **Convergence** and **trap avoidance**.
- Impact on r**ecovery guarantees** of DIP when optimized with inertia.
- NN **design**: need for **overparametrization**.

Take away messages

- **Inertia** (viscous and geometric) is good even for **non-convex** problems if **properly** used.
- **Convergence** and **trap avoidance**.
- Impact on r**ecovery guarantees** of DIP when optimized with inertia.
- NN **design**: need for **overparametrization**.
- **Empirical** results agree with **theoretical** predictions.

Take away messages

- **Inertia** (viscous and geometric) is good even for **non-convex** problems if **properly** used.
- **Convergence** and **trap avoidance**.
- Impact on r**ecovery guarantees** of DIP when optimized with inertia.
- NN **design**: need for **overparametrization**.
- **Empirical** results agree with **theoretical** predictions.

- Stochastic setting.

Take away messages

- **Inertia** (viscous and geometric) is good even for **non-convex** problems if **properly** used.
- **Convergence** and **trap avoidance**.
- Impact on r**ecovery guarantees** of DIP when optimized with inertia.
- NN **design**: need for **overparametrization**.
- **Empirical** results agree with **theoretical** predictions.

- Stochastic setting.
- Non-smooth setting.

Take away messages

- **Inertia** (viscous and geometric) is good even for **non-convex** problems if **properly** used.
- **Convergence** and **trap avoidance**.
- Impact on r**ecovery guarantees** of DIP when optimized with inertia.
- NN **design**: need for **overparametrization**.
- **Empirical** results agree with **theoretical** predictions.

- Stochastic setting.
- Non-smooth setting.
- Long time behaviour (occupation measures).

Take away messages

- **Inertia** (viscous and geometric) is good even for **non-convex** problems if **properly** used.
- **Convergence** and **trap avoidance**.
- Impact on r**ecovery guarantees** of DIP when optimized with inertia.
- NN **design**: need for **overparametrization**.
- **Empirical** results agree with **theoretical** predictions.

- Stochastic setting.
- Non-smooth setting.
- Long time behaviour (occupation measures).
- Other NN-based frameworks: PINNs, supervised setting.

Take away messages

- **Inertia** (viscous and geometric) is good even for **non-convex** problems if **properly** used.
- **Convergence** and **trap avoidance**.
- Impact on r**ecovery guarantees** of DIP when optimized with inertia.
- NN **design**: need for **overparametrization**.
- **Empirical** results agree with **theoretical** predictions.

- Stochastic setting.
- Non-smooth setting.
- Long time behaviour (occupation measures).
- Other NN-based frameworks: PINNs, supervised setting.
- Other overparametrization regimes.

Take away messages

- **Inertia** (viscous and geometric) is good even for **non-convex** problems if **properly** used.
- **Convergence** and **trap avoidance**.
- Impact on r**ecovery guarantees** of DIP when optimized with inertia.
- NN **design**: need for **overparametrization**.
- **Empirical** results agree with **theoretical** predictions.

- Stochastic setting.
- Non-smooth setting.
- Long time behaviour (occupation measures).
- Other NN-based frameworks: PINNs, supervised setting.
- Other overparametrization regimes.
- Other data-driven methods for IP: PnP, unrolling, generative models.

Preprint on arxiv and paper on

https://fadili.users.greyc.fr/

Thanks
Any questions ?