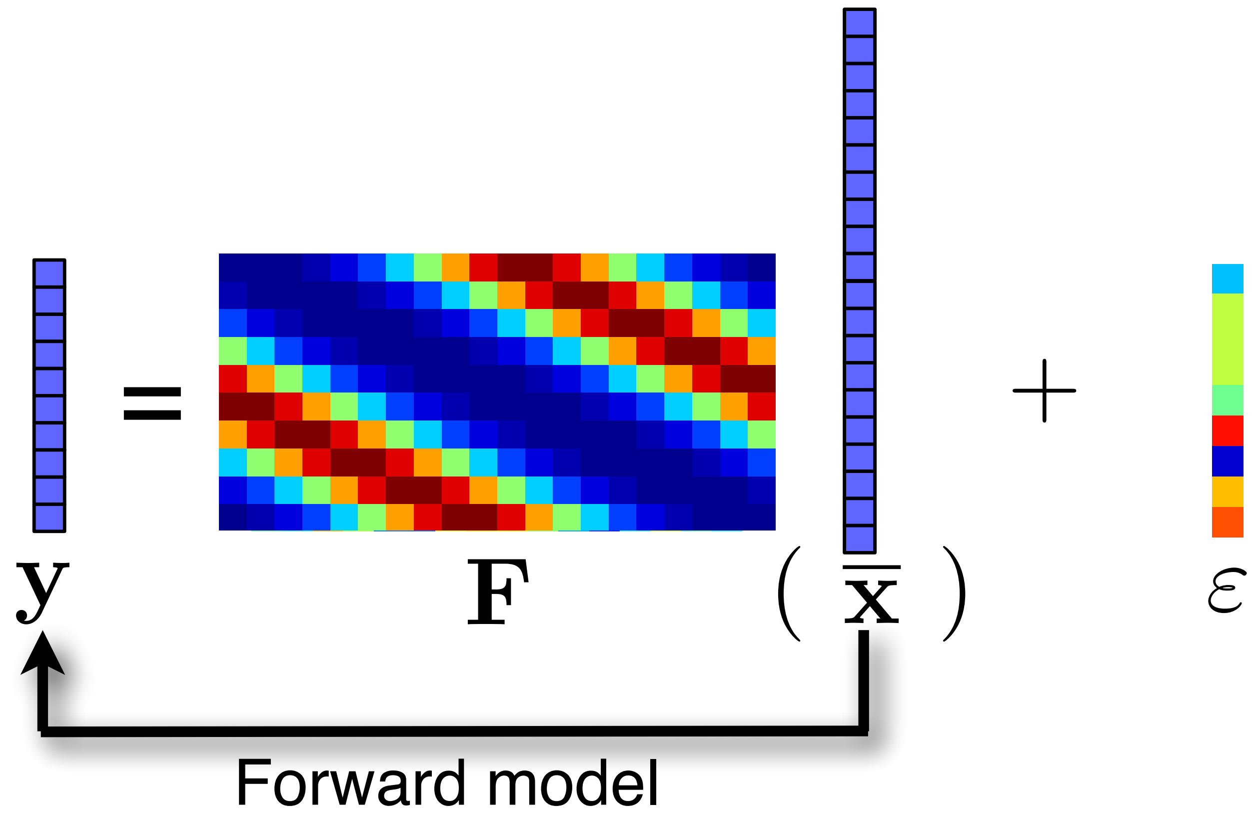
Iterative Regularization of NN-Based Inverse Problems via Gradient Flow

Jalal Fadili
Normandie Université-ENSICAEN, GREYC CNRS UMR 6072

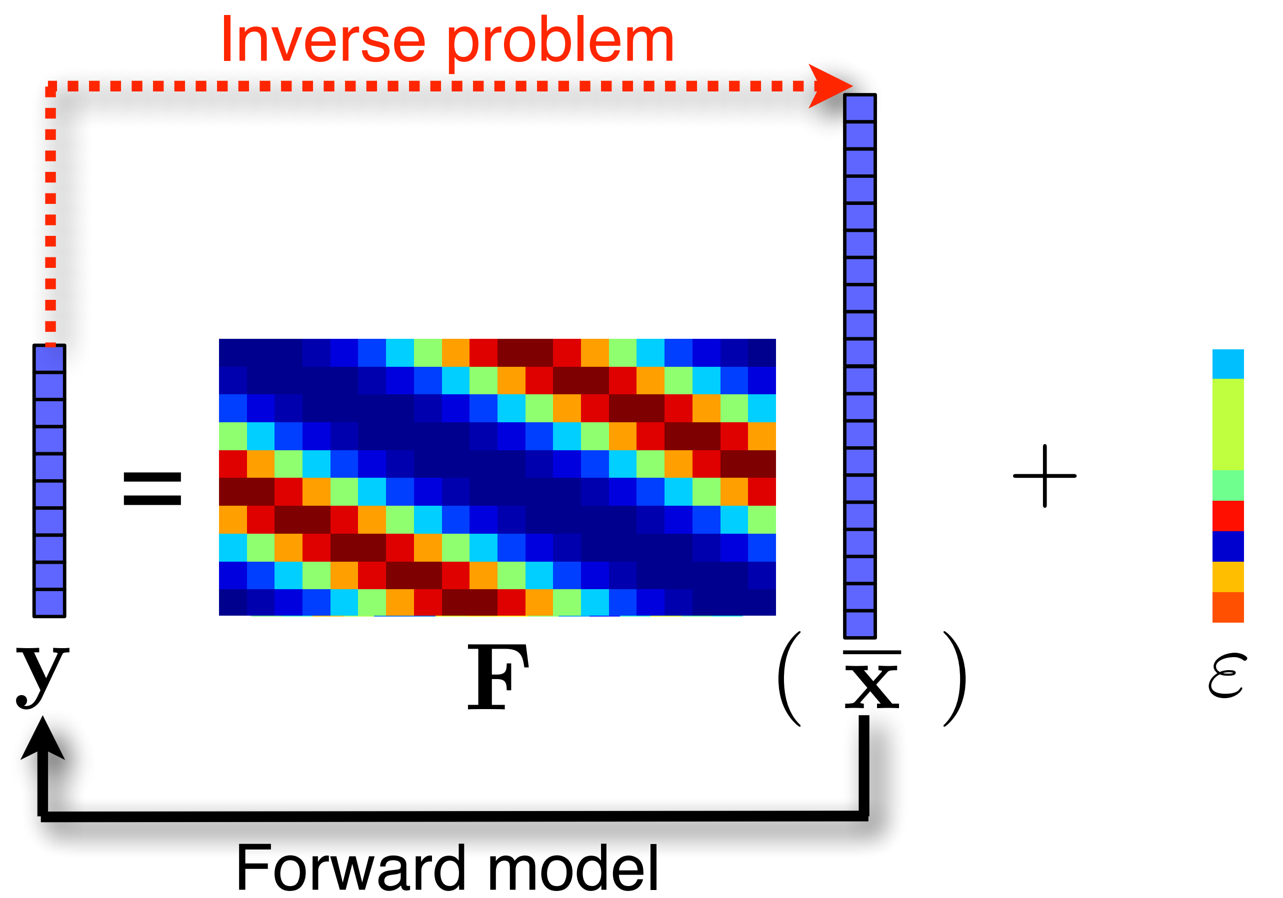
Joint ARGOS-TITAN-TOSCA workshop
6-7 June 2024
Join work with Nathan Buskulic and Yvain Quéau

Inverse problems



- 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.

Inverse problems

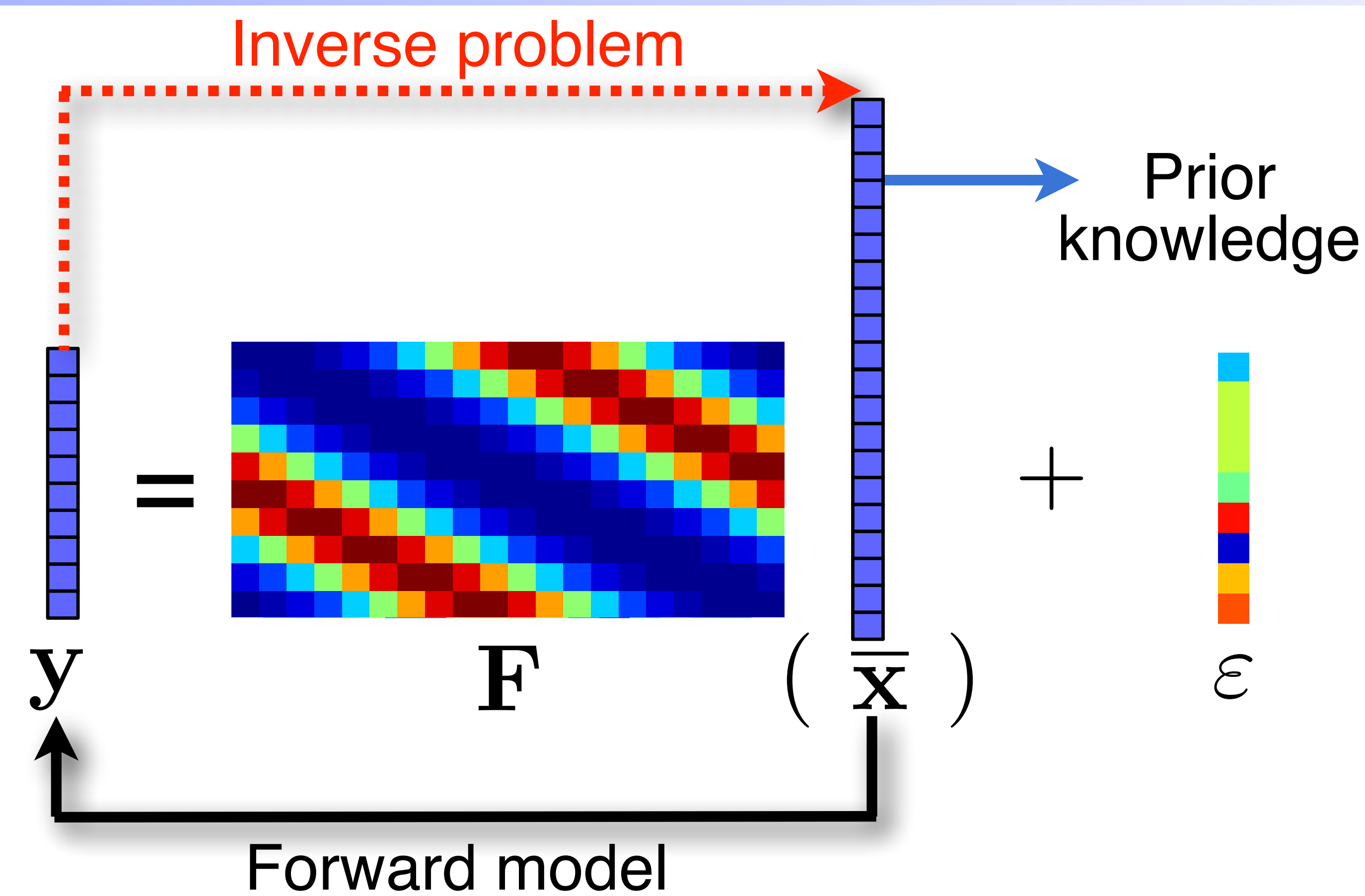


- Throughout the talk : finite-dimensional setting.
- $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^m$ is the forward operator (physics of the observation formation model).
- ε : noise.

Goal

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

Model-based variational approach

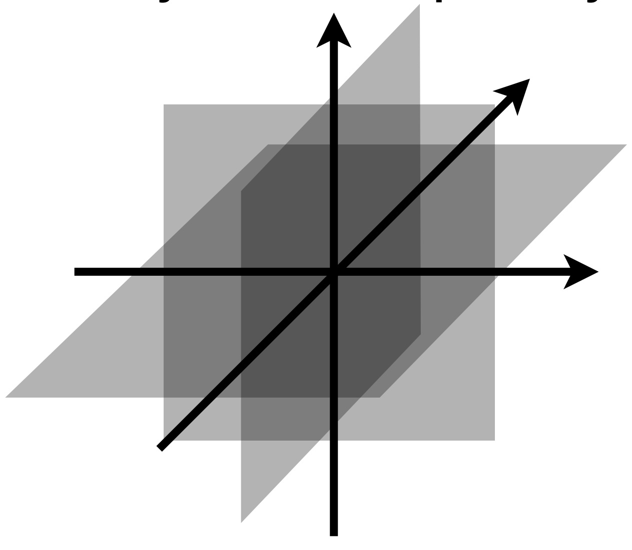


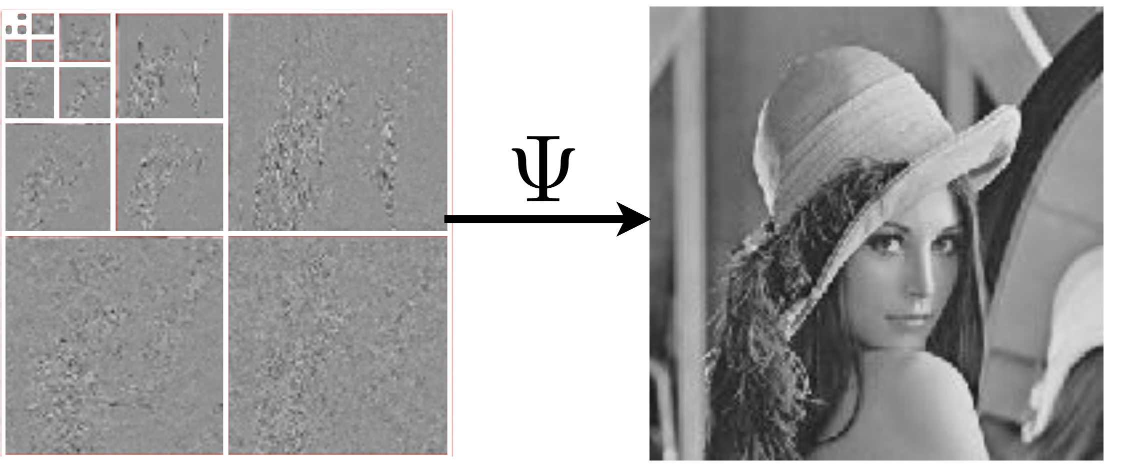
- Solve :

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

Low-complexity models

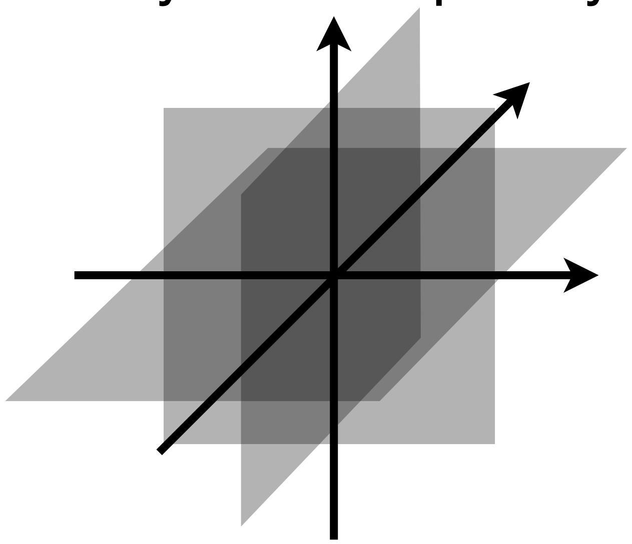
Synthesis sparsity



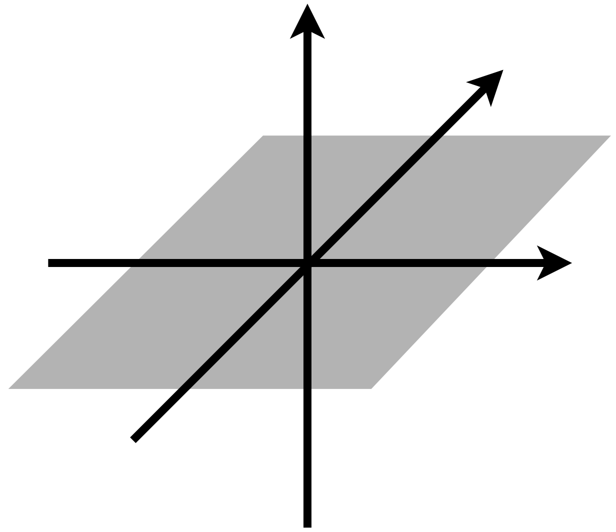


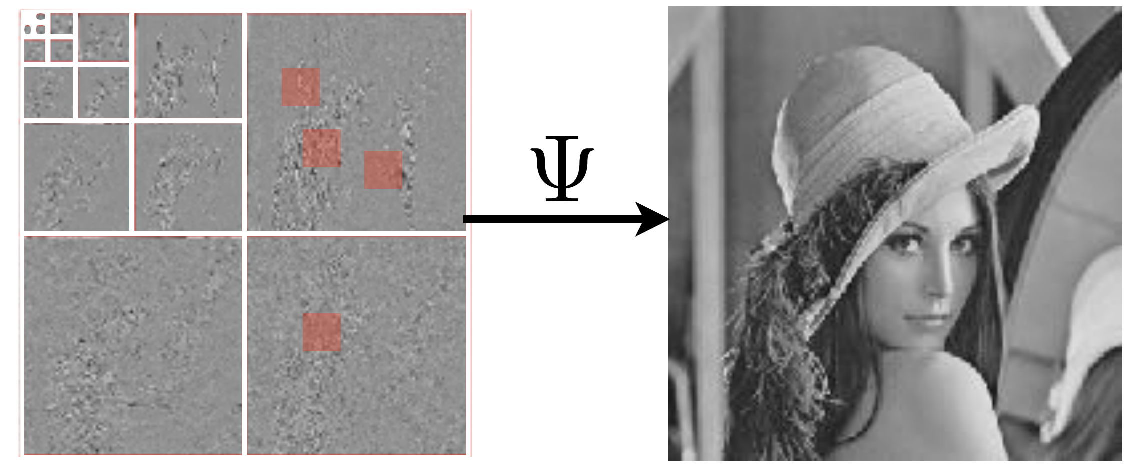
Low-complexity models

Synthesis sparsity



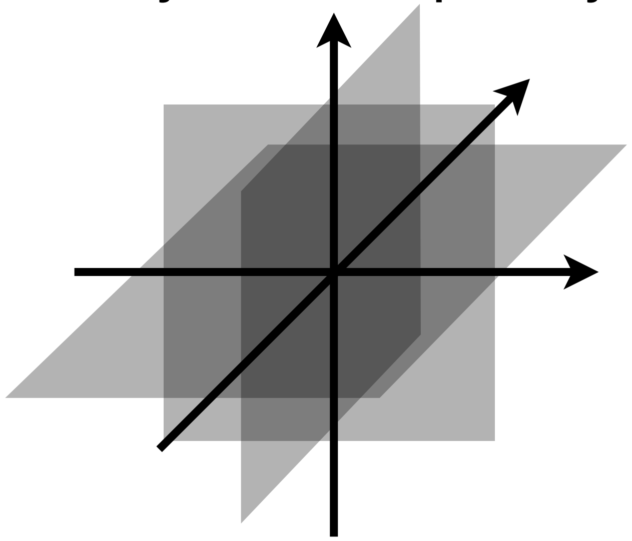
Group sparsity



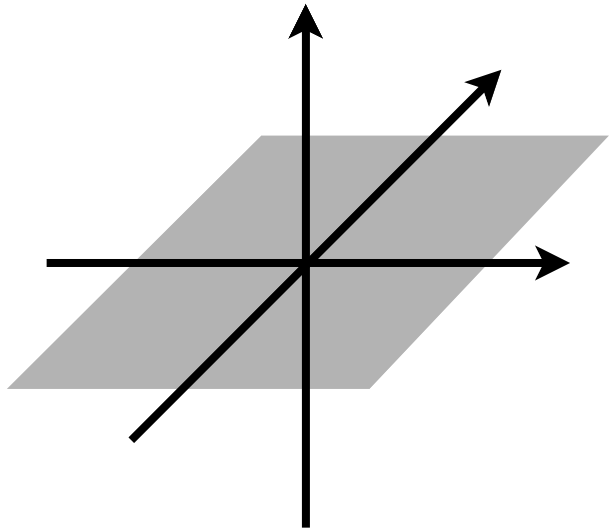


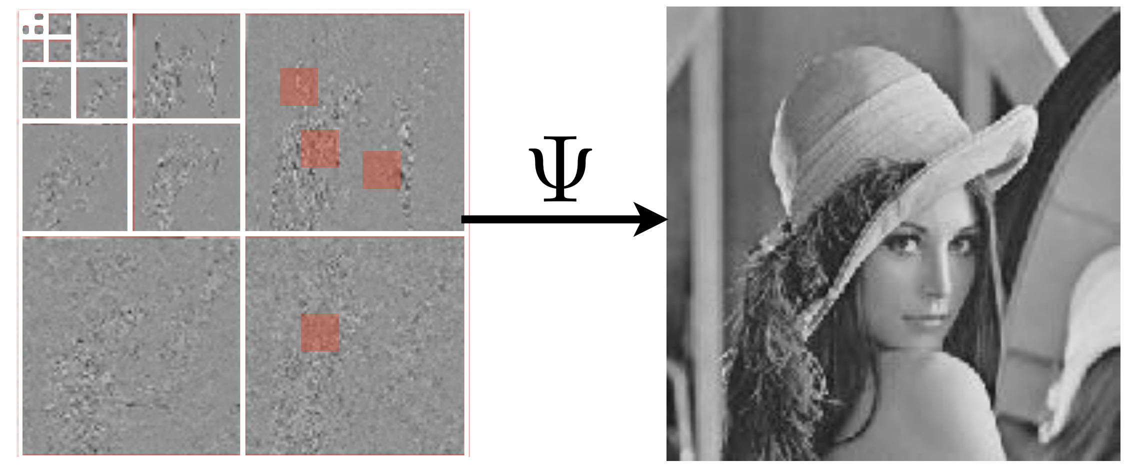
Low-complexity models

Synthesis sparsity

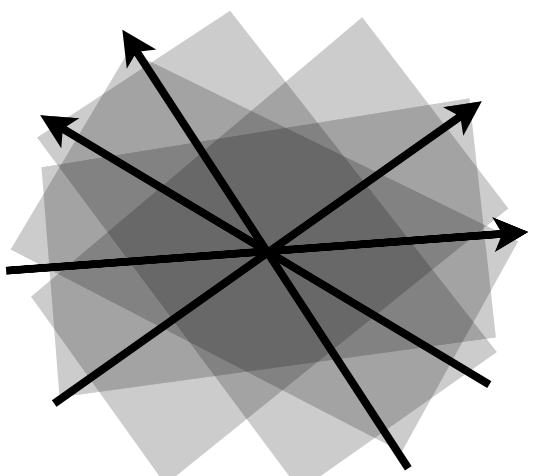


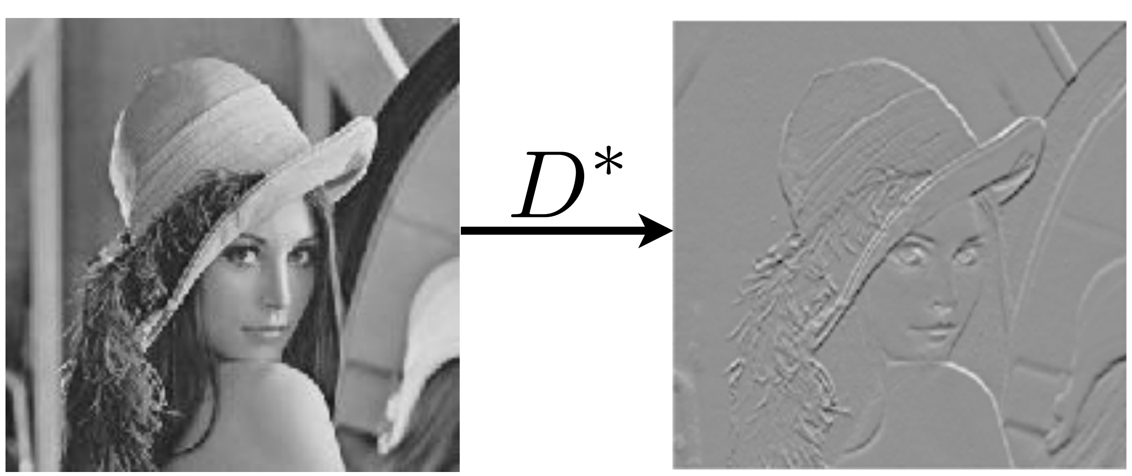
Group sparsity





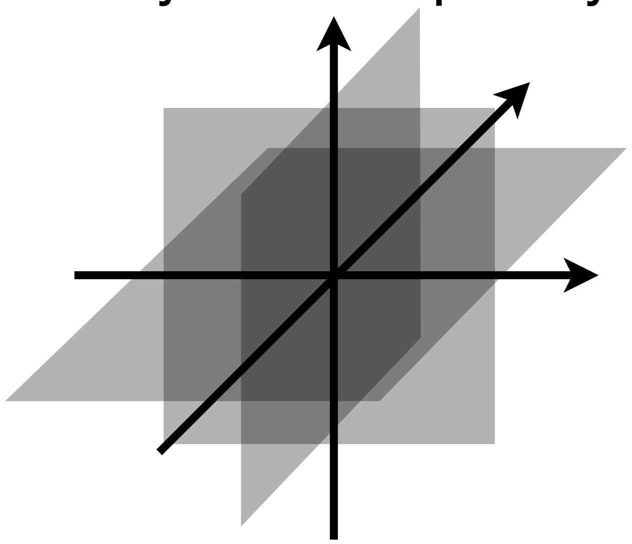
Analysis sparsity



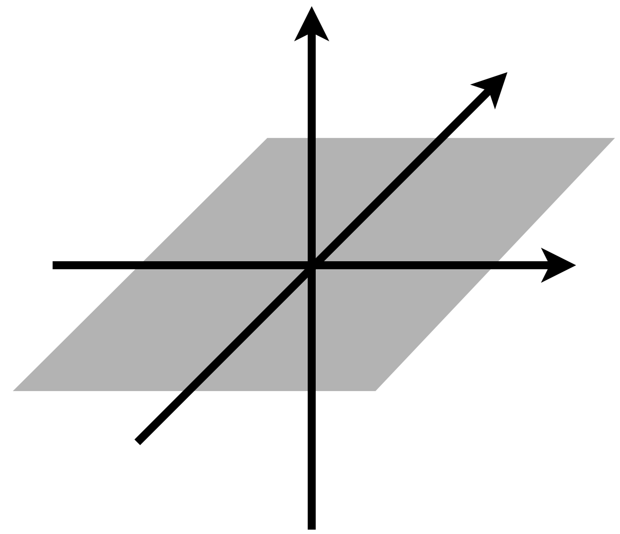


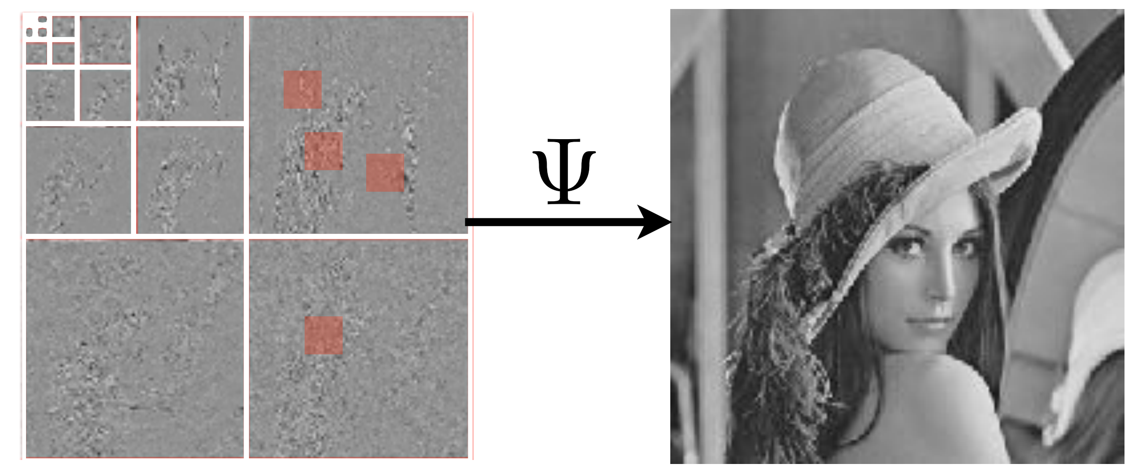
Low-complexity models

Synthesis sparsity

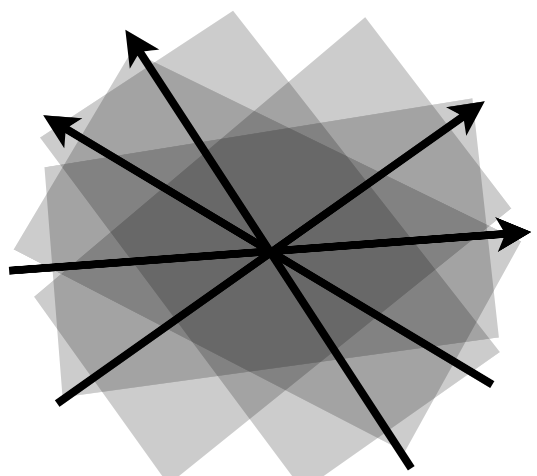


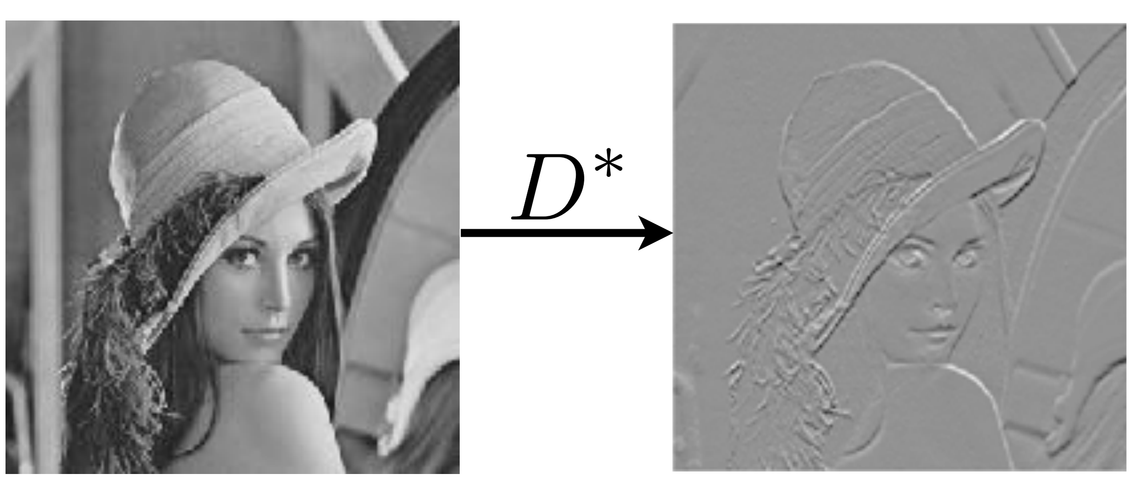
Group sparsity



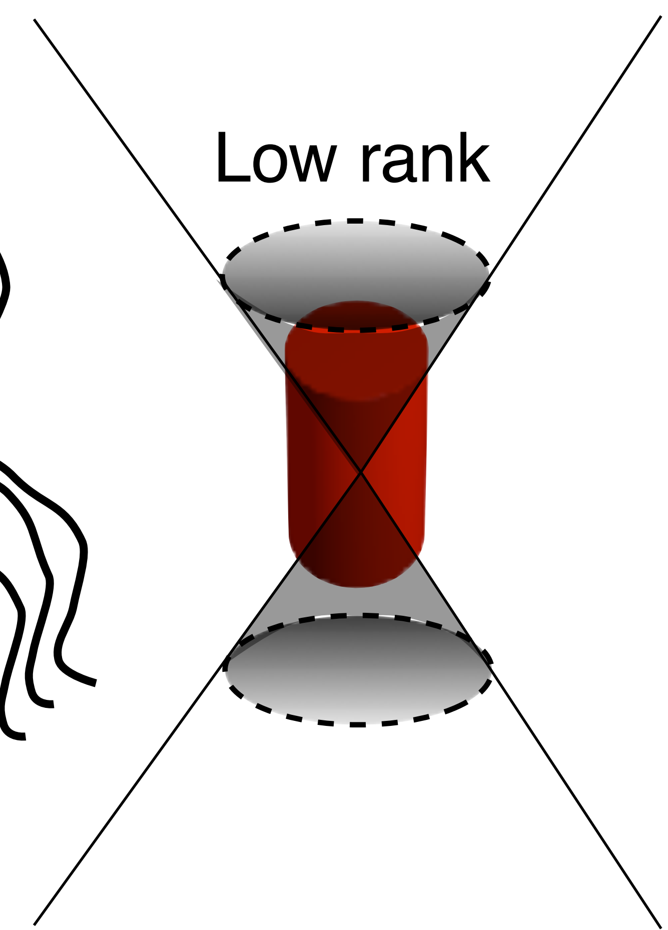


Analysis sparsity

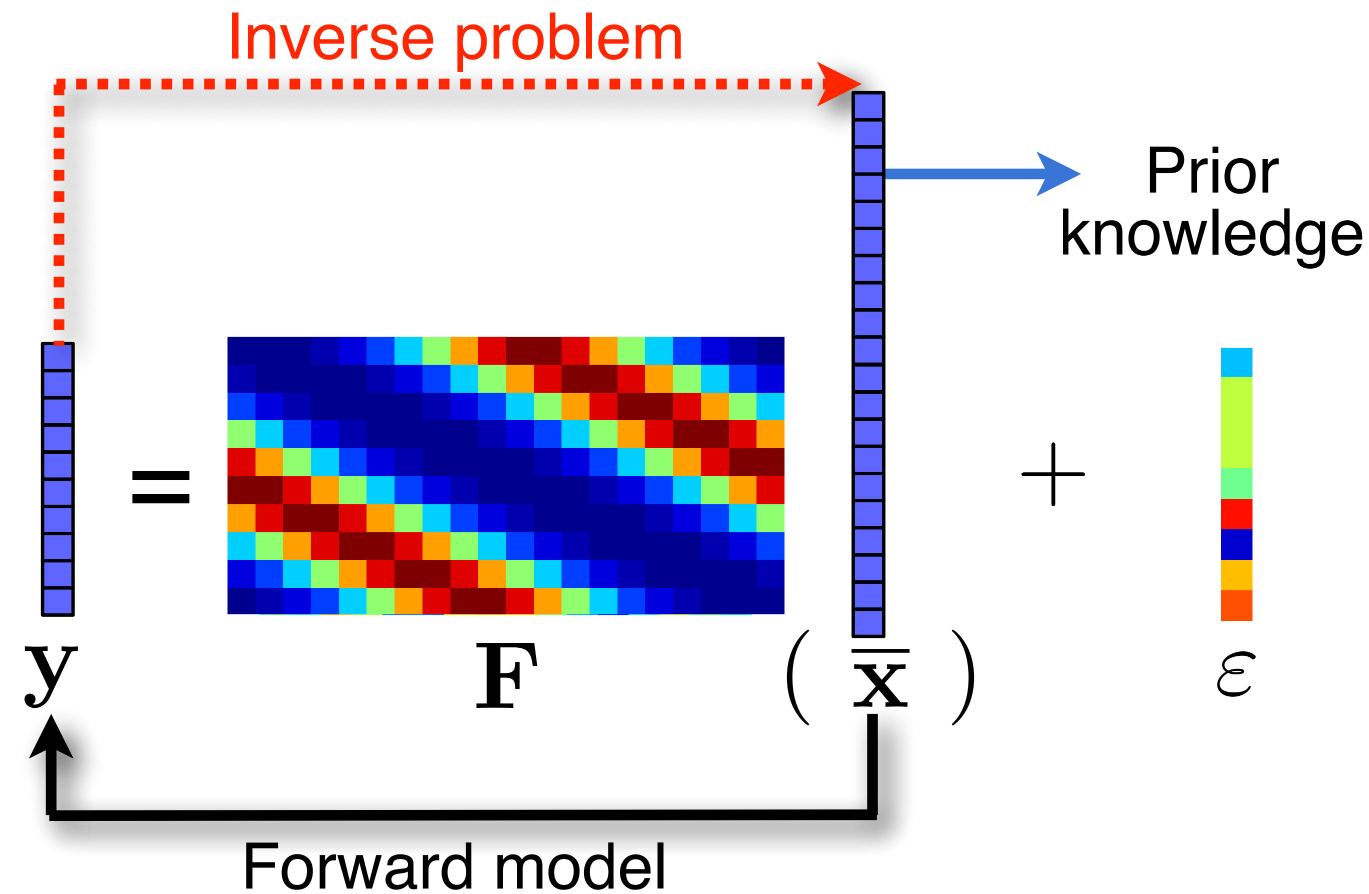




Hyperspectral imaging



Model-based variational approach



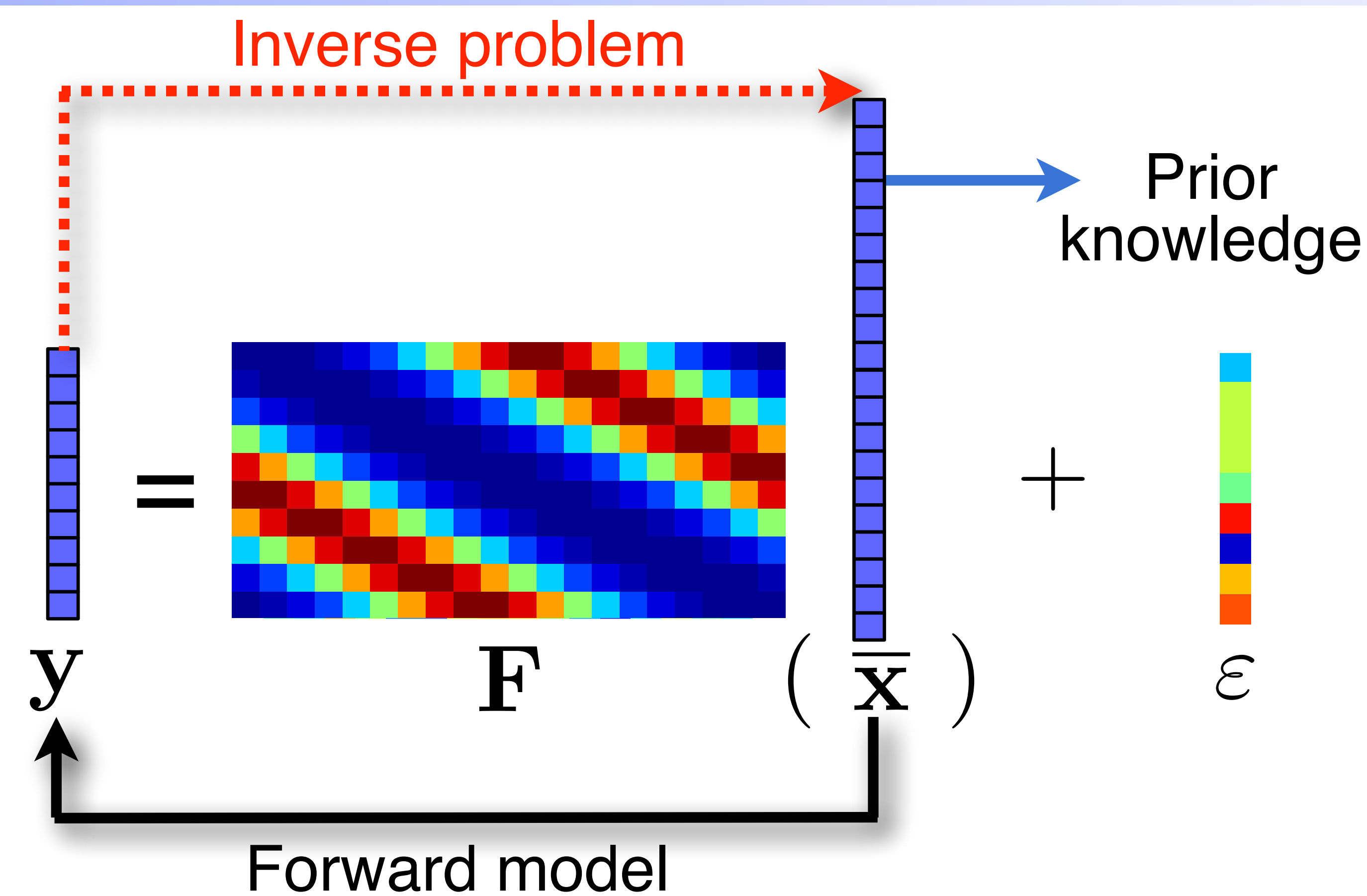
- Solve :

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

Pros

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

Model-based variational approach



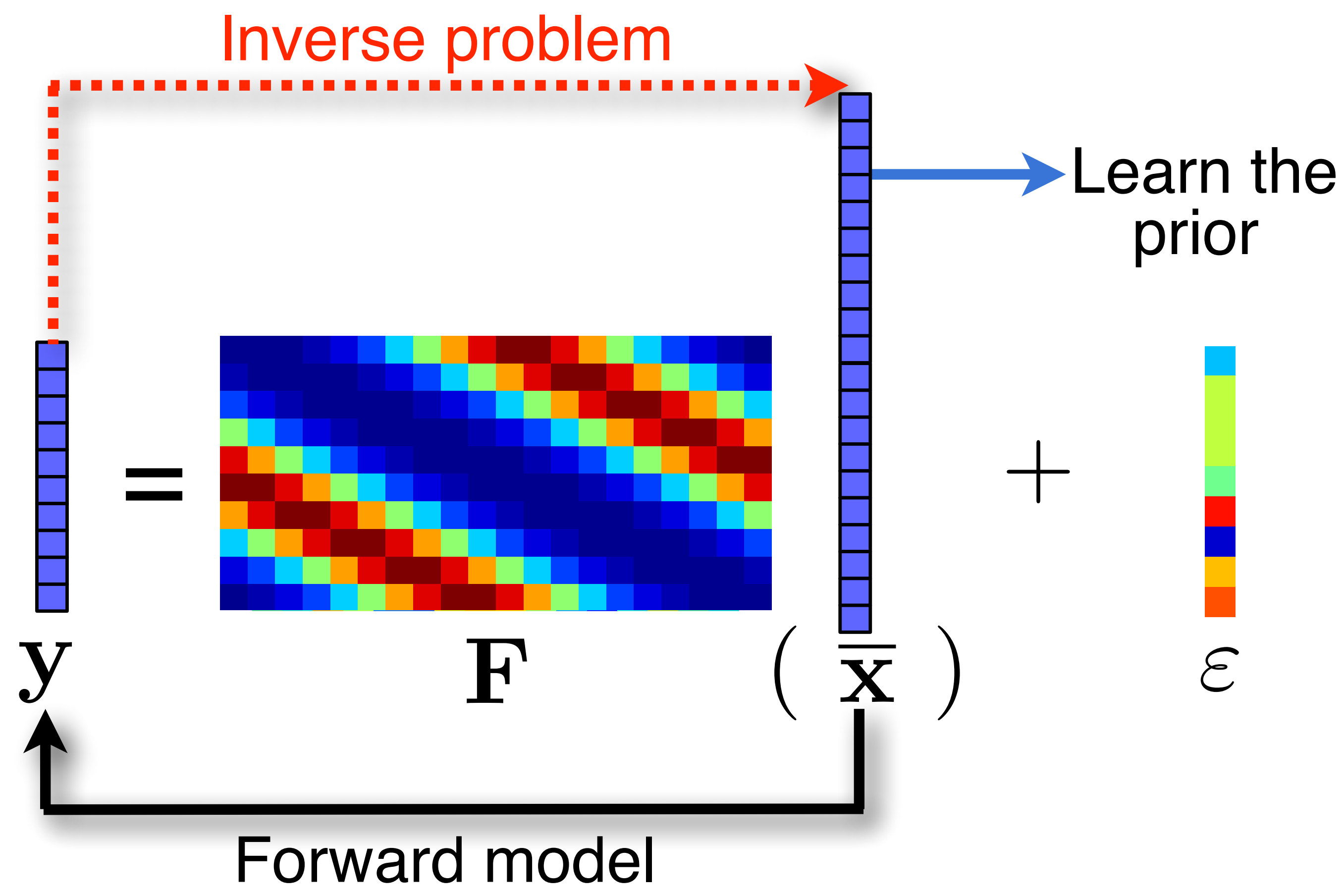
- Solve :

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

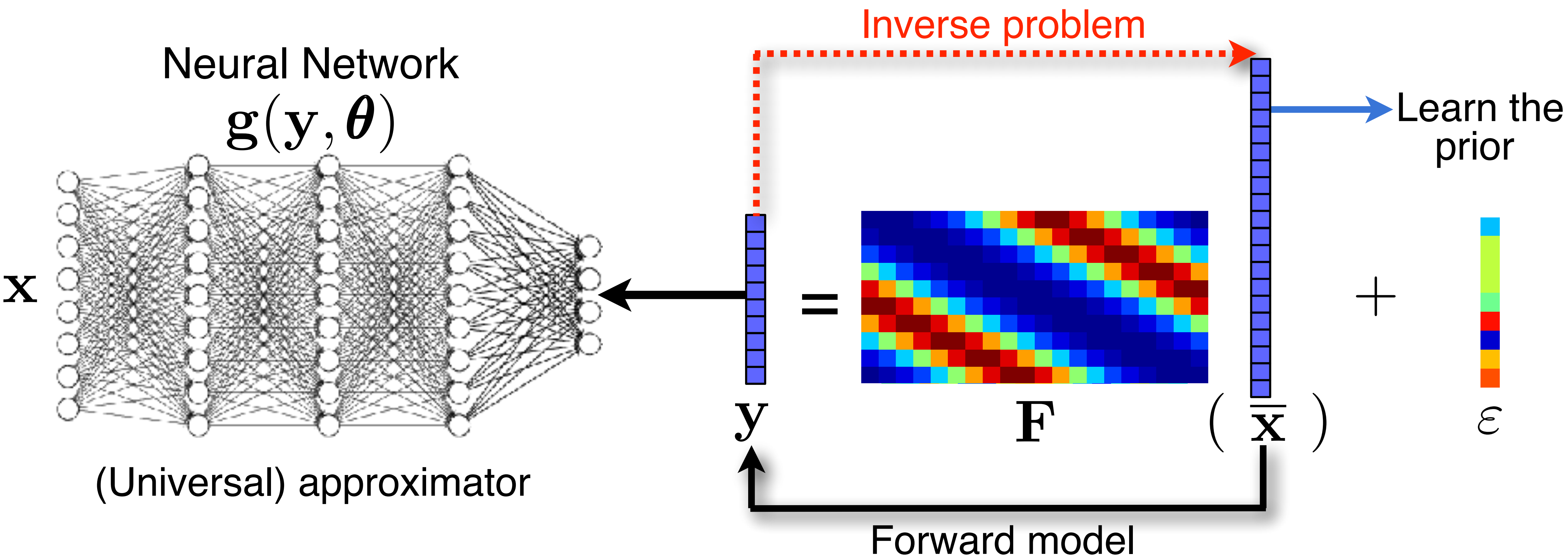
Pros	***Cons***

• 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 approach

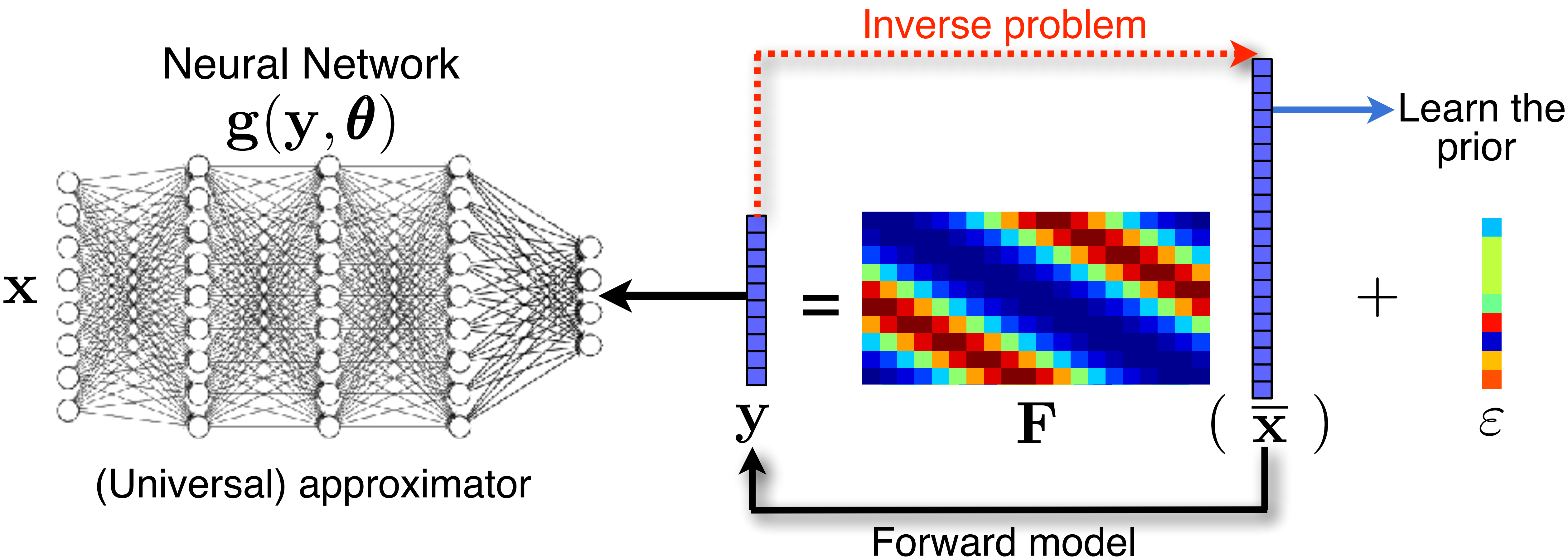


Data-based approach



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

Data-based approach

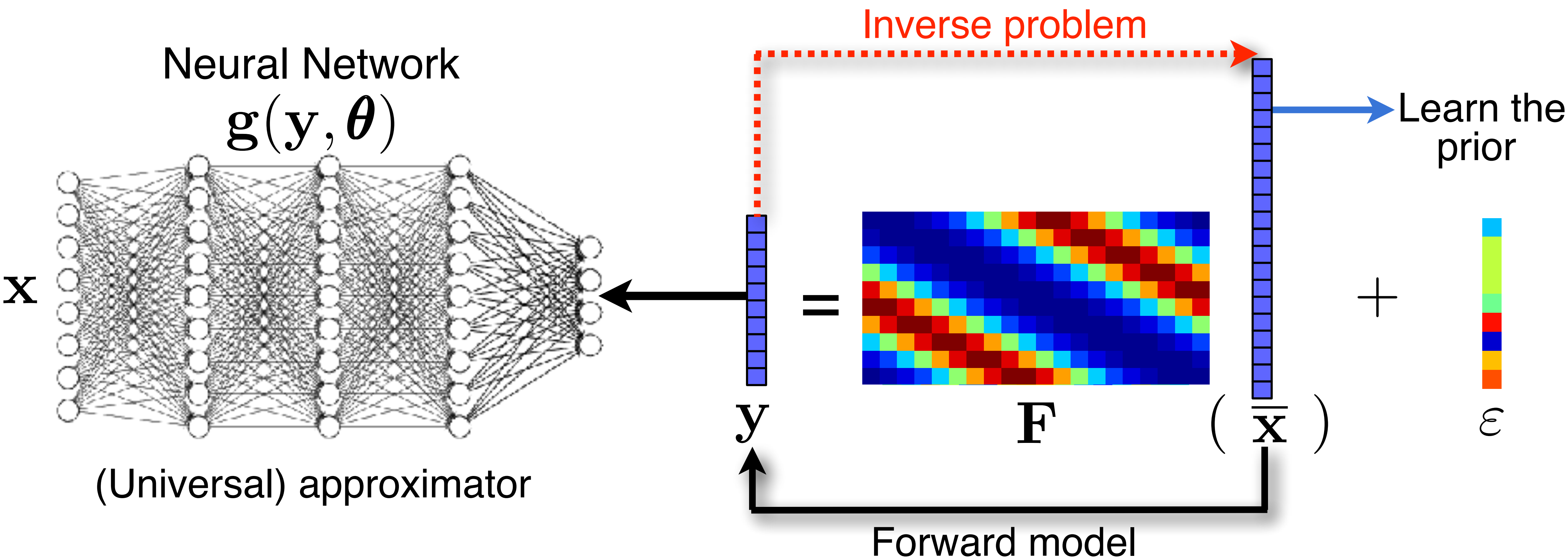


$$\min_{\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p} \frac{1}{N} \sum_{i=1}^{N} \ell(\mathbf{x}_i, \mathbf{g}(\mathbf{y}_i, \boldsymbol{\theta}))$$

Pros

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

Data-based approach



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

Pros

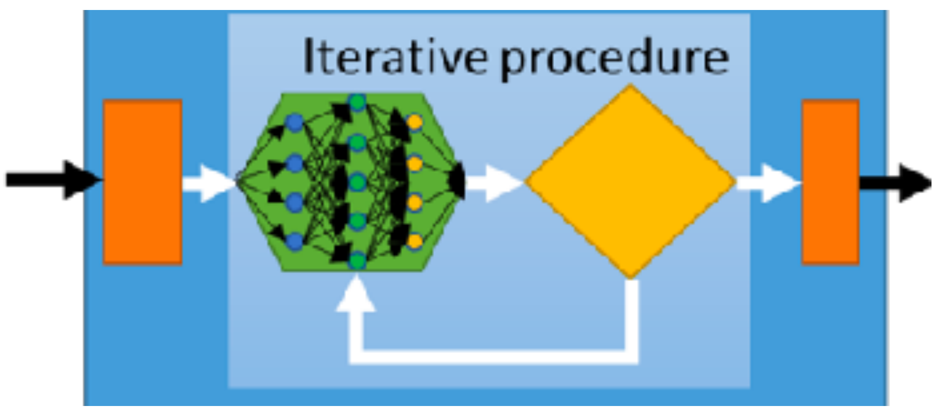
- Off-the-shelf NN learning frameworks.
- No model to think about (… not quite so).
- Training once for all.

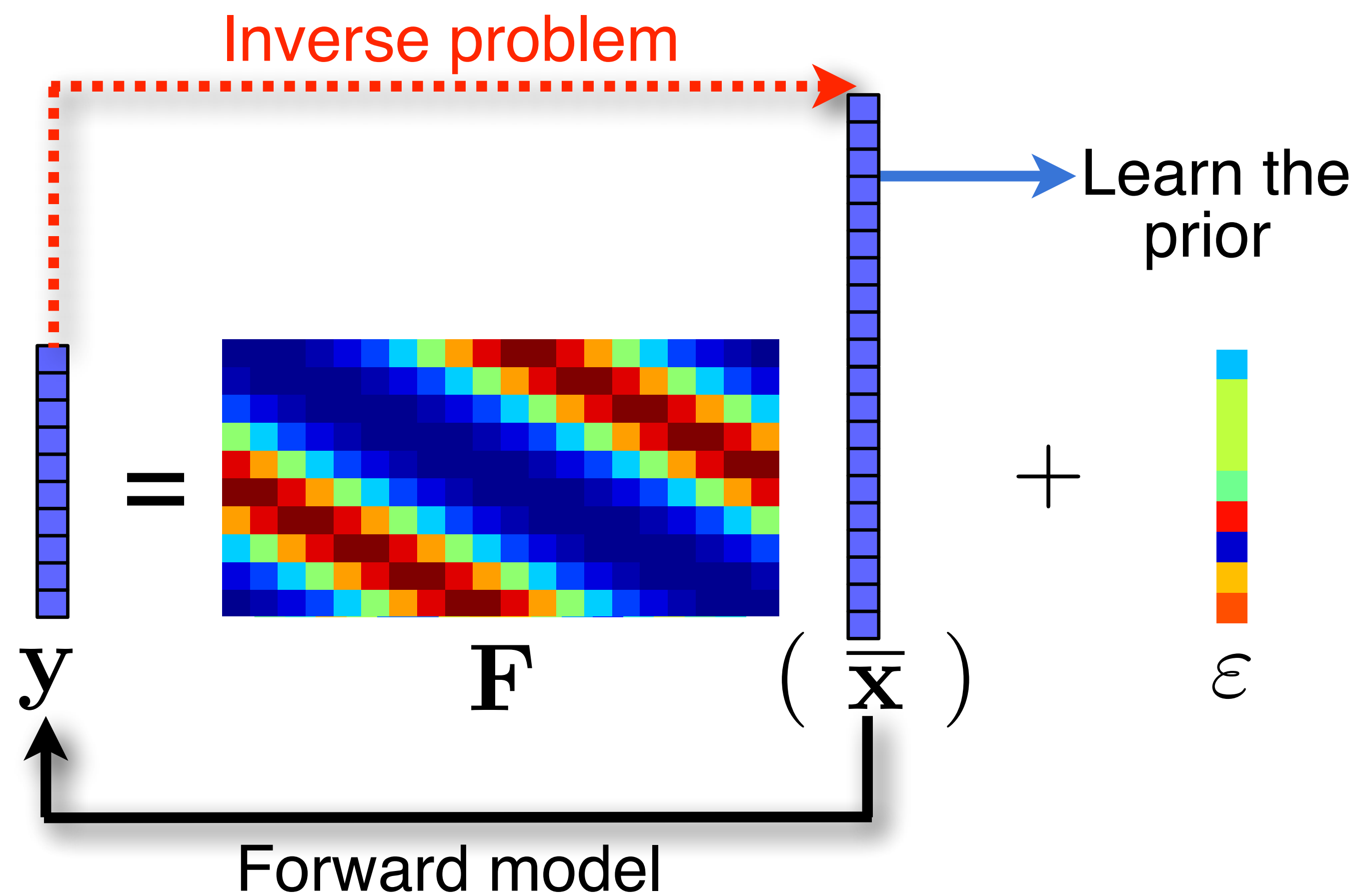
Cons

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

Hybrid (model-based) learning

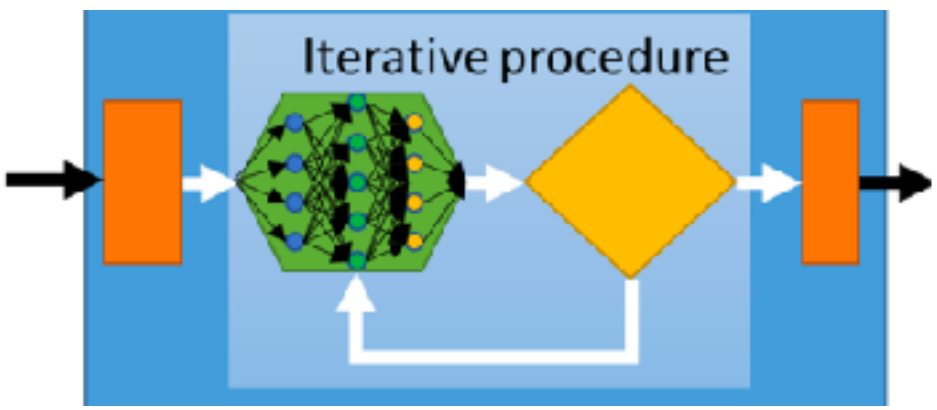
- Mix model- and data-driven methods in various ways: e.g.
  - Learn the regularizer.
  - Plug-and-Play.
  - Unrolling.
  - Deep equilibrium.
  - Learn other inference methods and/or generative priors.
  - etc.
- An extremely active area, with extensive literature and reviews.

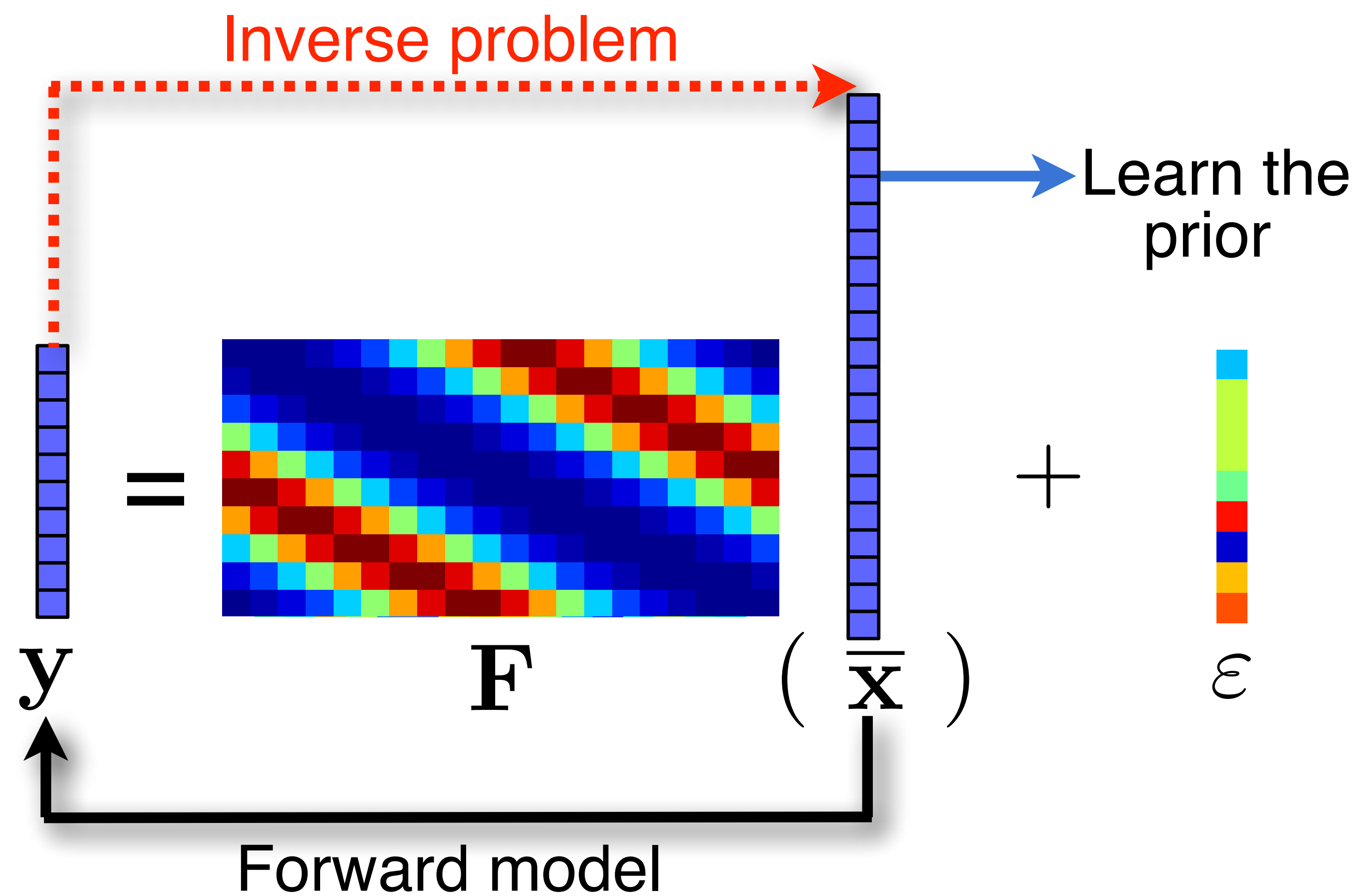




Hybrid (model-based) learning

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



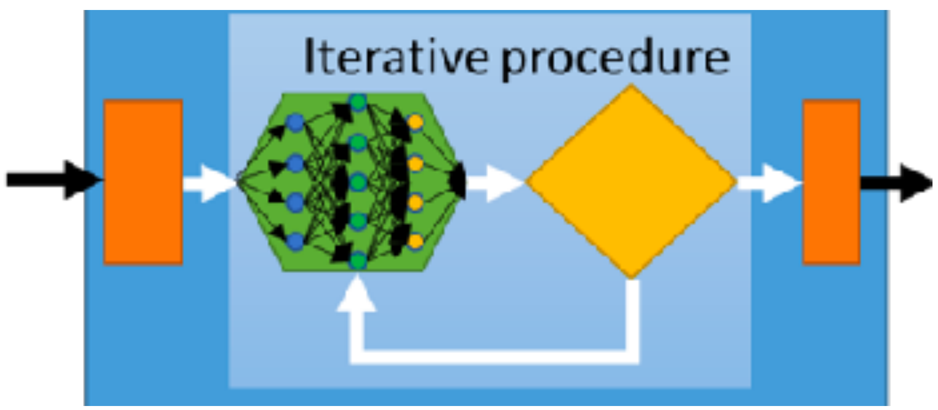


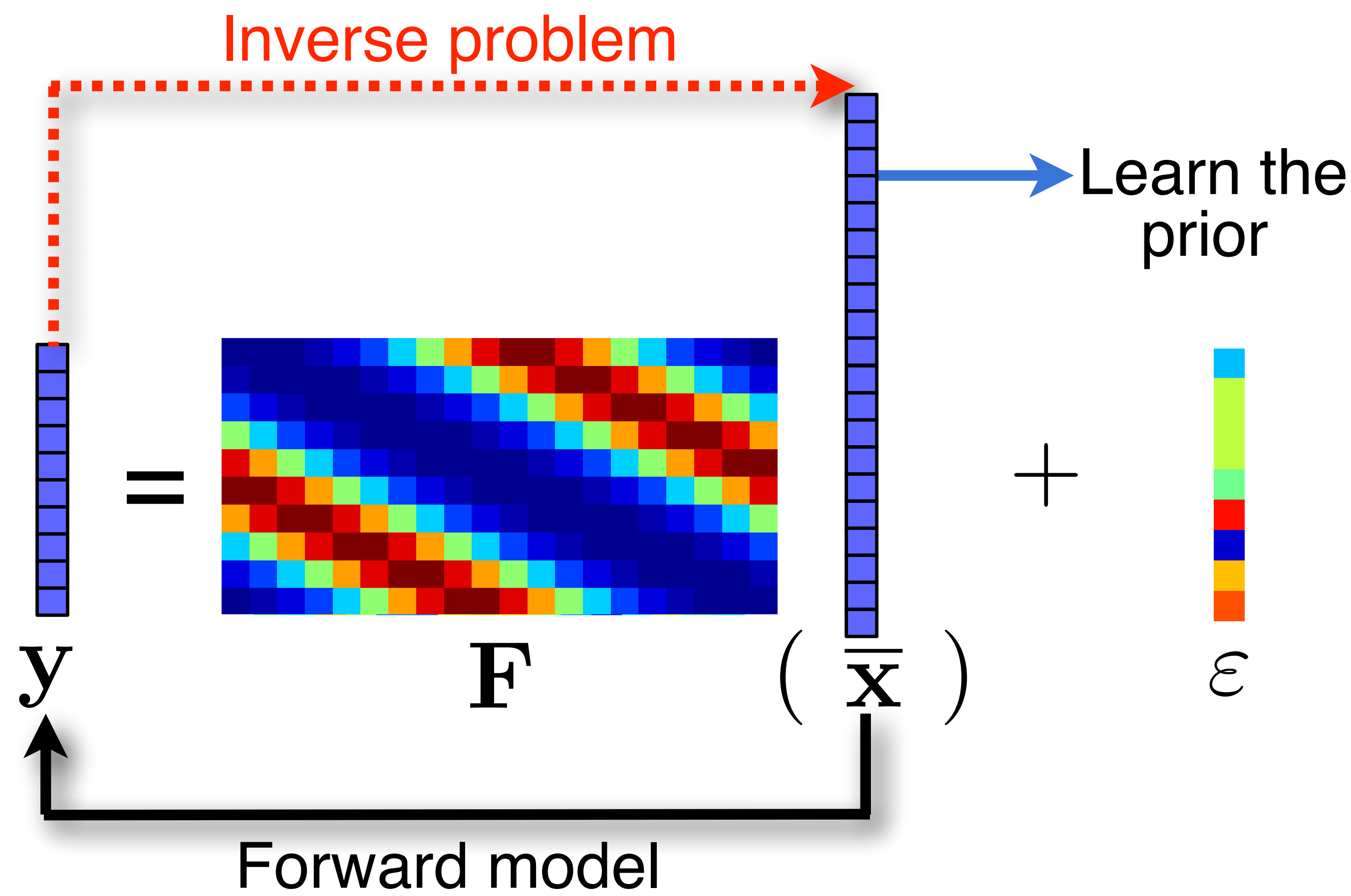
Pros

- 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.
  - Learn other inference methods and/or generative priors.
  - etc.
- An extremely active area, with extensive literature and reviews.





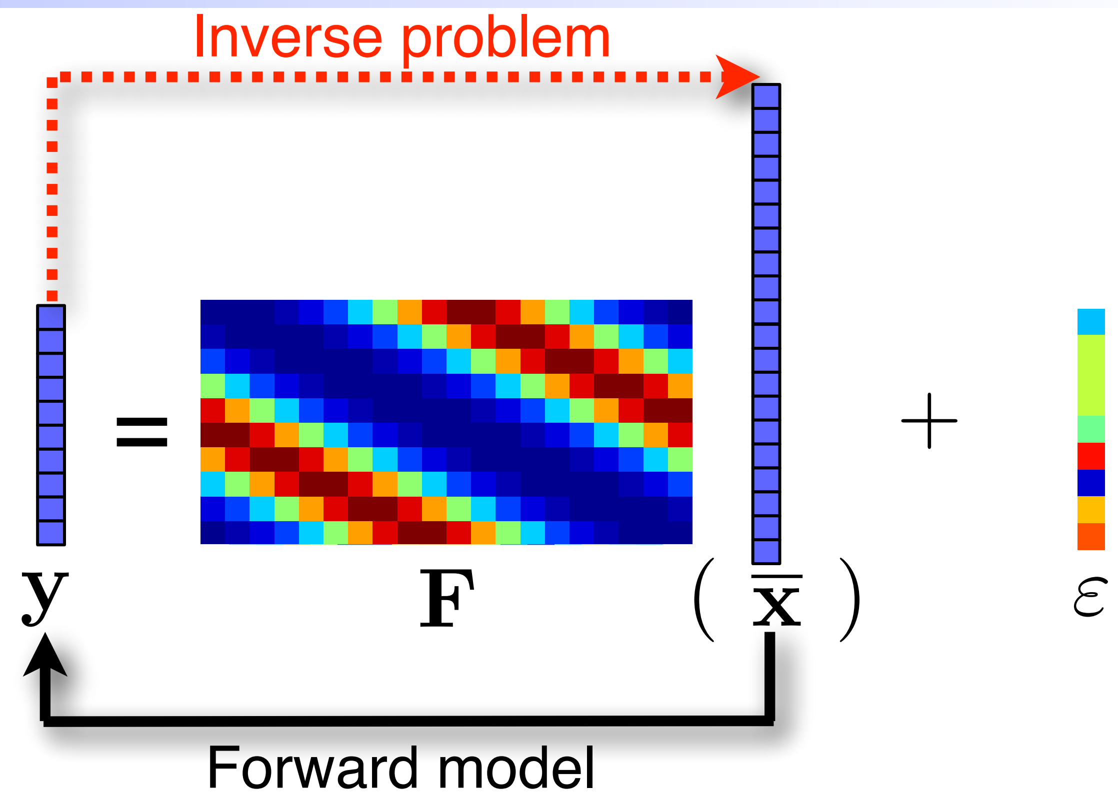
Pros

- 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

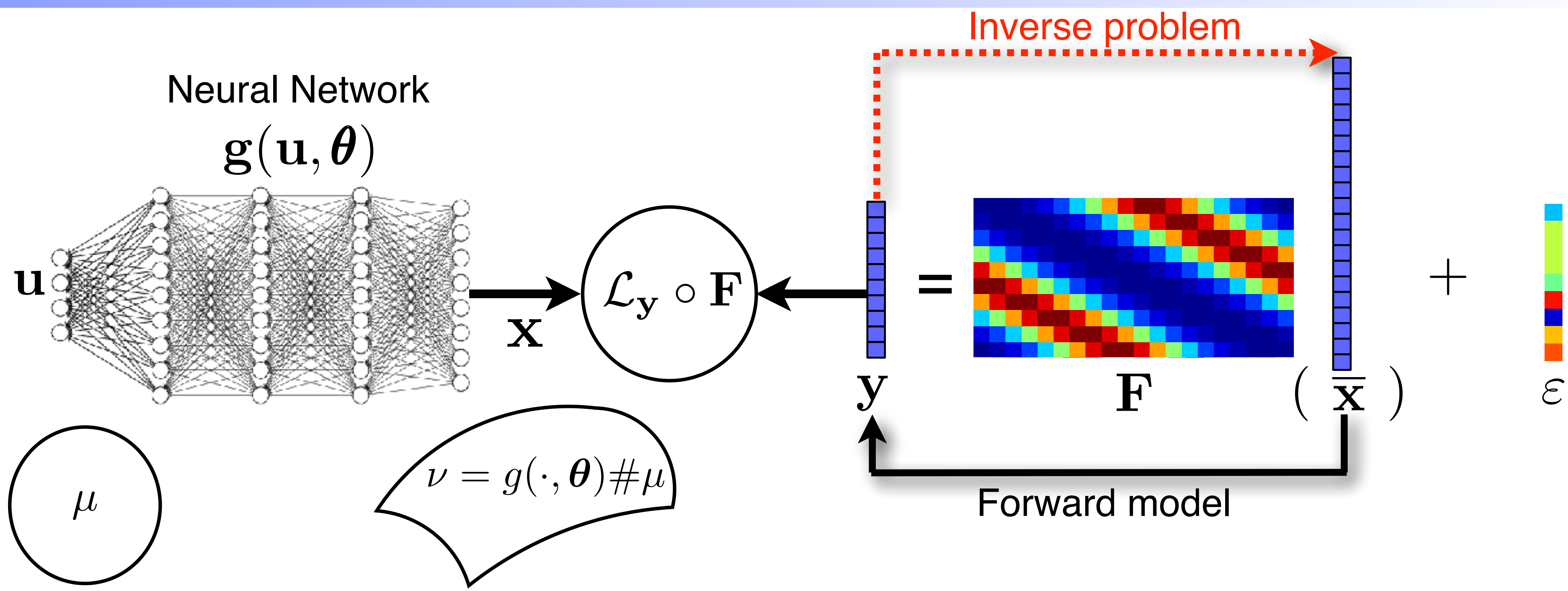
- Supervised: availability of training data.
- NN design (or even many NNs).
- Lack of guarantees from IP perspective: recovery, stability, explainability, etc.

DIP: Deep Inverse/Image Prior

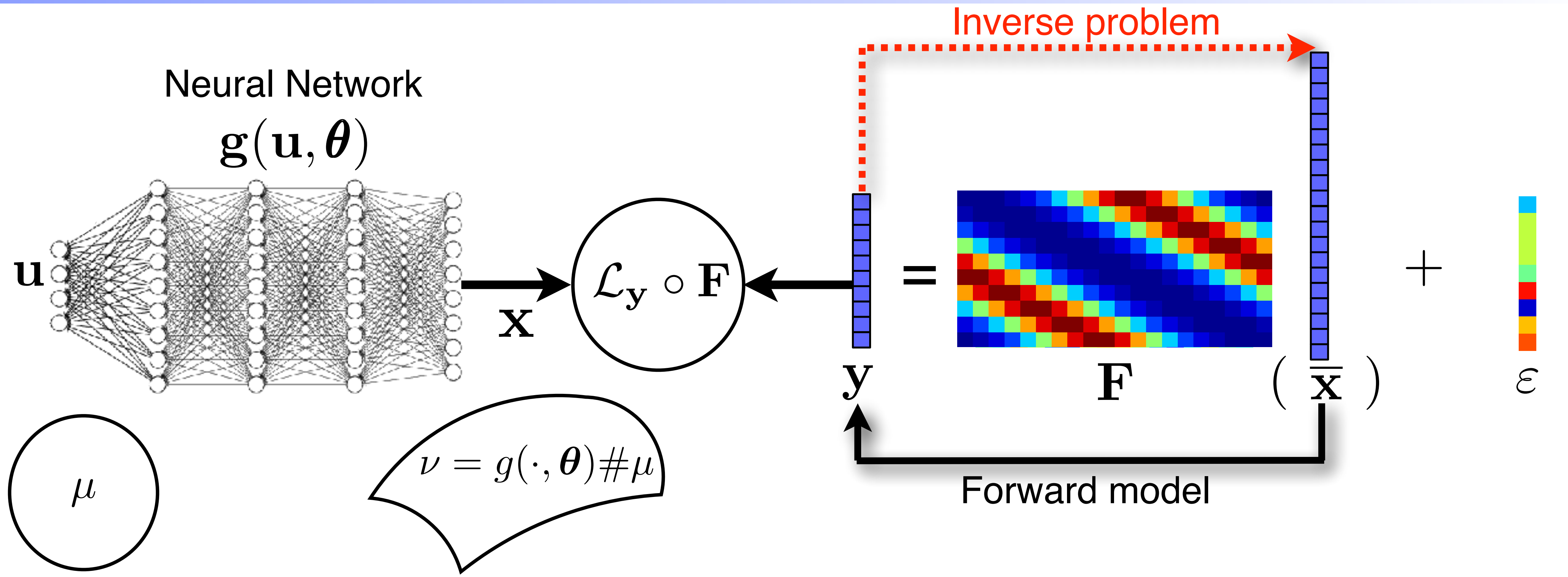


$$\mathbf{y} = \mathbf{F}(\bar{\mathbf{x}}) + \varepsilon$$

DIP: Deep Inverse/Image Prior



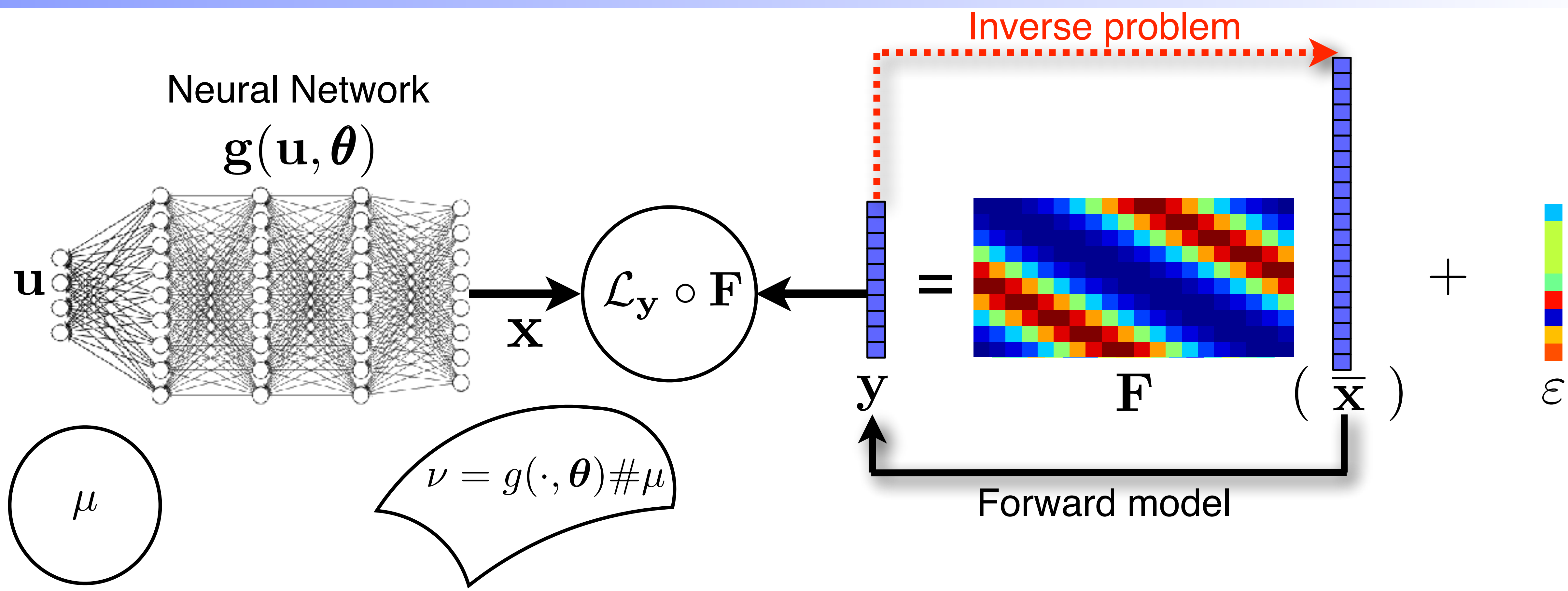
DIP: Deep Inverse/Image Prior



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

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

DIP: Deep Inverse/Image Prior



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

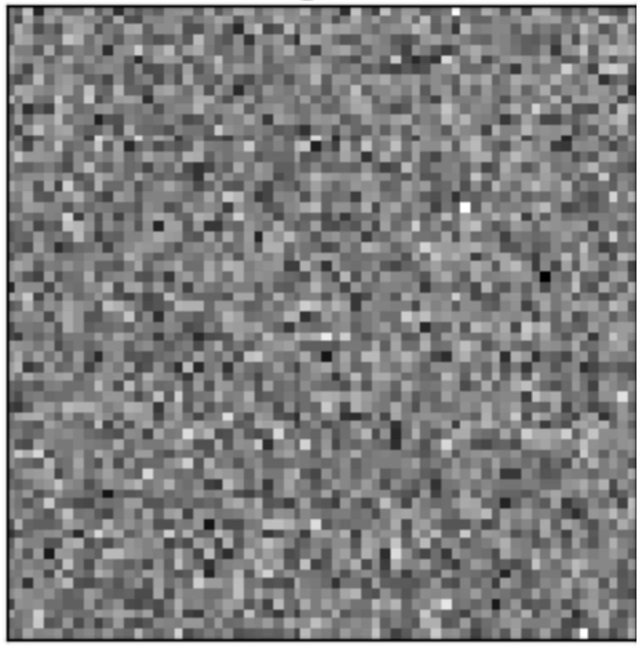
$$\Sigma = \{\mathbf{g}(\mathbf{u},\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$$

- 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 $\overline{\mathbf{x}}$.

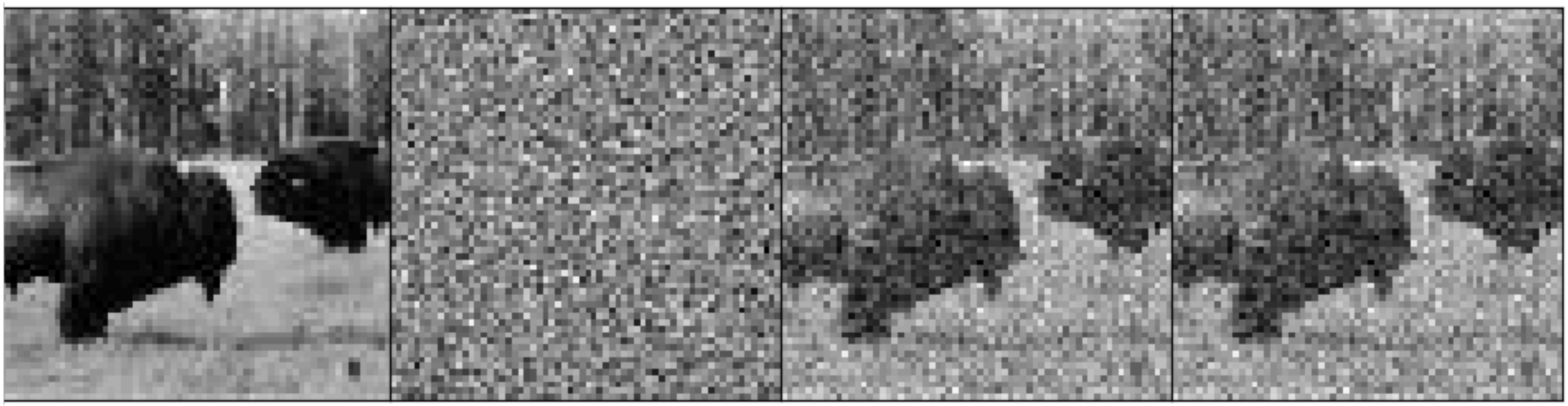
Example: DIP for image deblurring

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

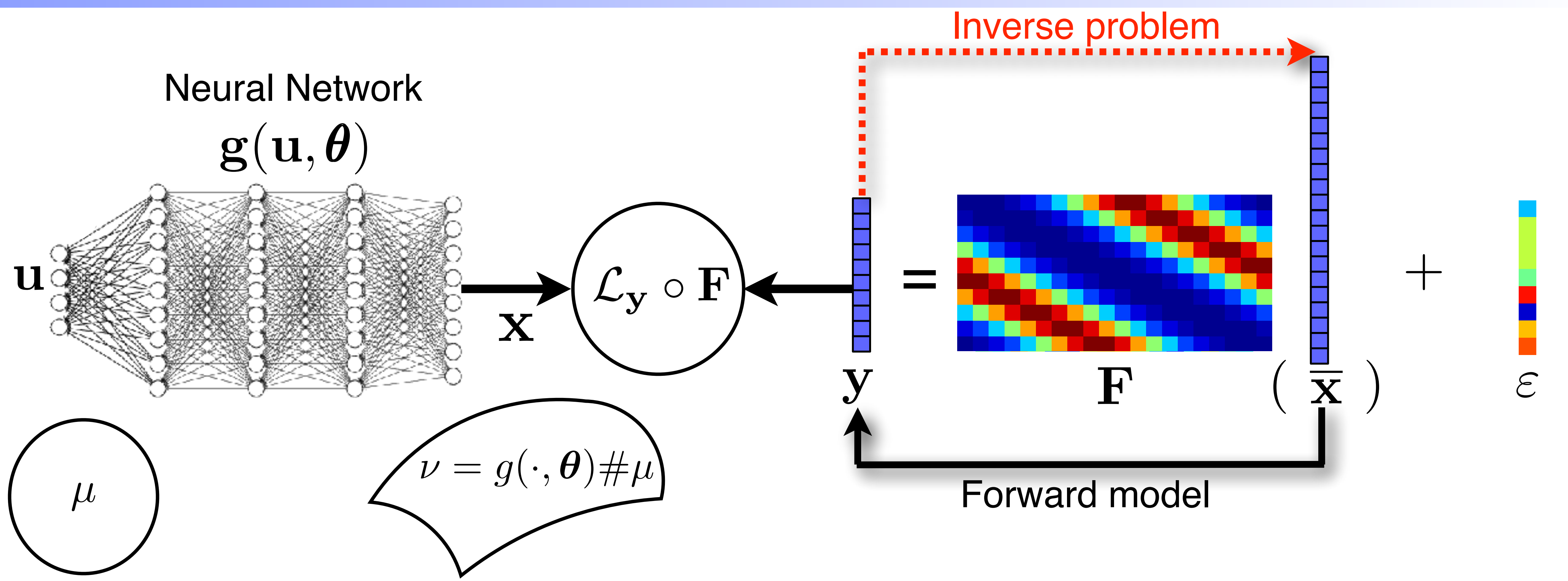
$\varepsilon \sim \mathcal{N}(0, 50^2)$



$\overline{\mathbf{x}}$ $\mathbf{x}_0$ $\mathbf{x}_{80}$ $\mathbf{x}_{1000}$



DIP: Deep Inverse/Image Prior



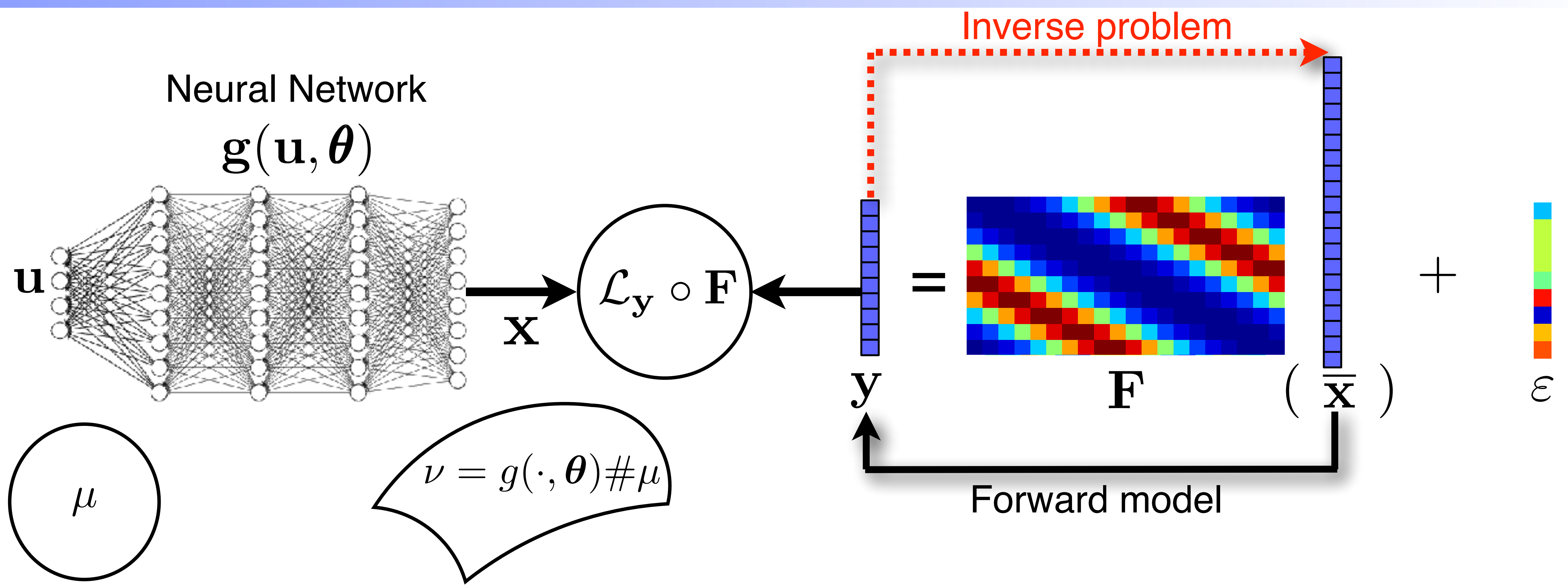
$$\min_{\boldsymbol{\theta} \in \Theta} \mathcal{L}_{\mathbf{y}}(\mathbf{F}(\mathbf{x}_{\boldsymbol{\theta}})) \qquad \text{s.t.} \quad \mathbf{x}_{\boldsymbol{\theta}} \in \Sigma$$

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

Pros

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

DIP: Deep Inverse/Image Prior

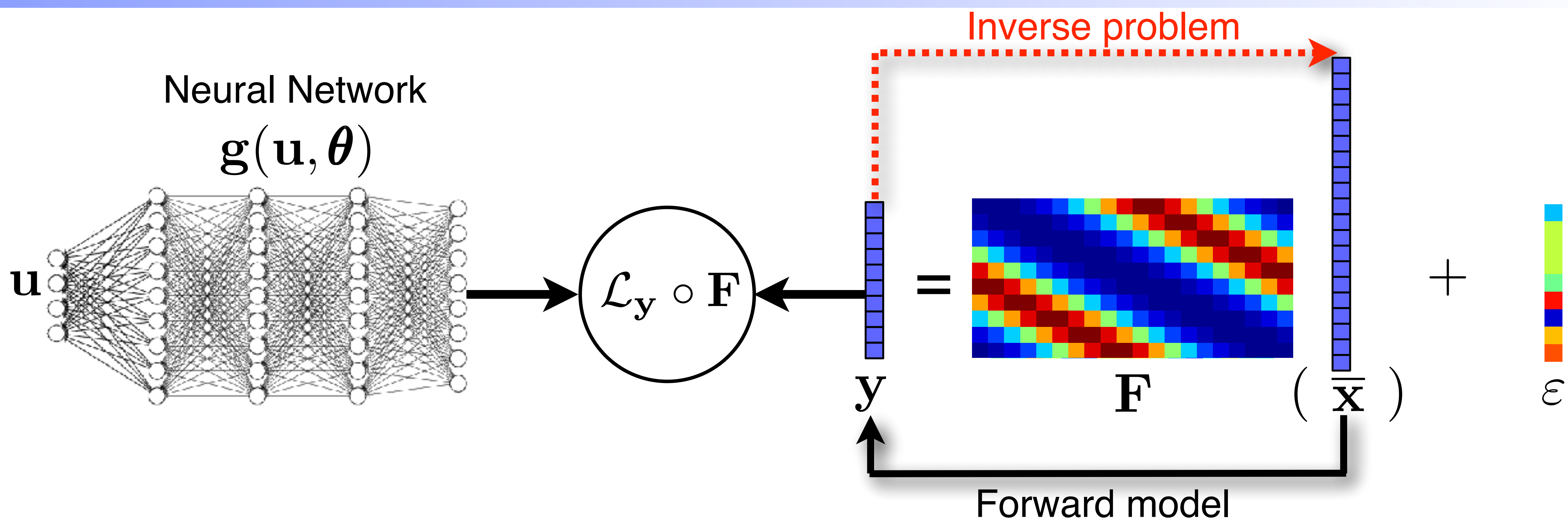


$$\min_{\boldsymbol{\theta} \in \Theta} \mathcal{L}_{\mathbf{y}}(\mathbf{F}(\mathbf{x}_{\boldsymbol{\theta}})) \qquad \text{s.t.} \quad \mathbf{x}_{\boldsymbol{\theta}} \in \Sigma$$

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

Pros	***Cons***
• Unsupervised. • Accounts for the forward model. • Easy to implement with (very) good empirical success.	• Optimize/train for each signal to recover. • No theoretical guarantees: recovery, stability, NN design.

Today's talk: Guarantees of DIP

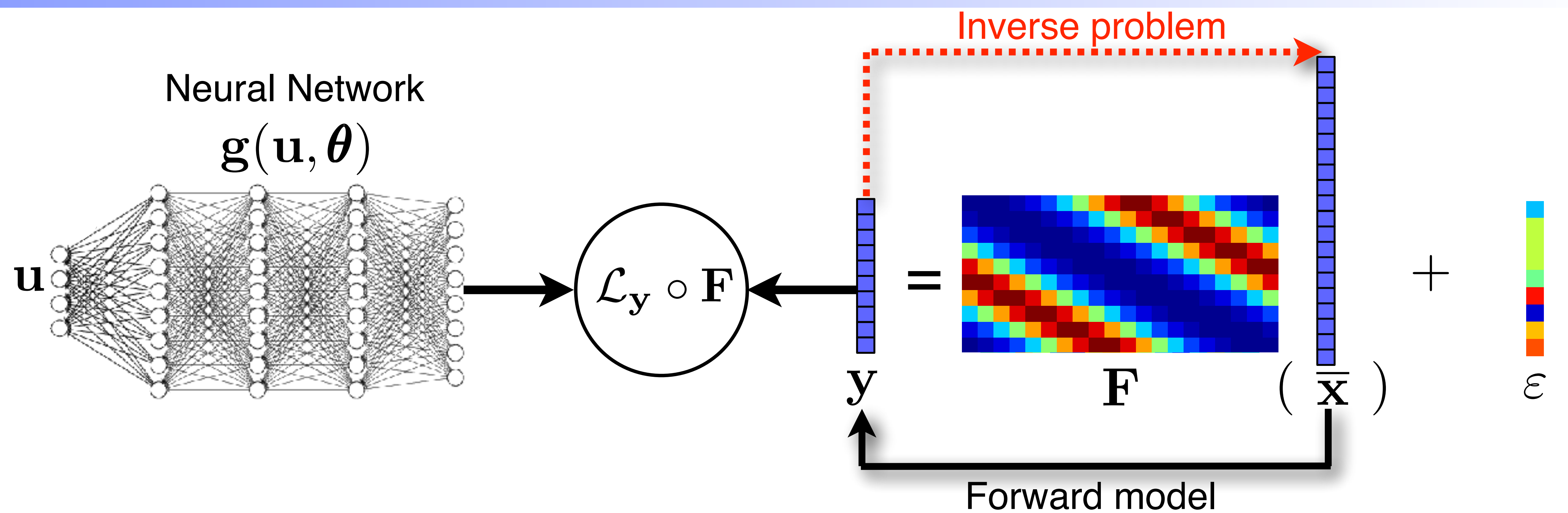


$$\min_{\boldsymbol{\theta} \in \Theta} \mathcal{L}_{\mathbf{y}}(\mathbf{F}(\mathbf{x}_{\boldsymbol{\theta}})) \qquad \text{s.t.} \quad \mathbf{x}_{\boldsymbol{\theta}} \in \Sigma$$

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

- Recovery guarantees of DIP when optimized with gradient descent in :
  - Observation space : convergence to zero-loss $\Rightarrow$ early stopping strategy.
  - Object space : restricted injectivity of the forward operator on Σ.
- General loss functions verifying the Kurdyka-Łojasewicz (KL) property : role of the desingularizing function on the convergence rate.
- NN design : role of overparametrization for the two-layer DIP setting.

Today's talk: Guarantees of DIP



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

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

- Recovery guarantees of DIP when optimized with gradient descent in :
  - Observation space : convergence to zero-loss $\Rightarrow$ early stopping strategy.
  - Object space : restricted injectivity of the forward operator on Σ.
- General loss functions verifying the Kurdyka-Łojasewicz (KL) property : role of the desingularizing function on the convergence rate.
- NN design : role of overparametrization for the two-layer DIP setting.

In the rest of the talk, linear forward operator

Outline

- Our setting.
- Main recovery guarantees.
- Case of the two-layer DIP.
- Numerical results.
- Conclusion.

Outline

- Our setting.
- Main recovery guarantees.
- Case of the two-layer DIP.
- Numerical results.
- Conclusion.

Globalized KL functions

Definition (KL inequality) *A continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the KL inequality if there exists $r_0 > 0$ and a strictly increasing function $\psi \in \mathcal{C}^0([0, r_0[) \cap \mathcal{C}^1(]0, r_0[)$ with $\psi(0) = 0$ such that*

$$\psi'(f(\mathbf{z}) - \min f)\,\|\nabla f(\mathbf{z})\| \geq 1, \quad \forall \mathbf{z} \in [\min f < f < \min f + r_0].$$

We use the shorthand notation $f \in$ KŁ$_\psi(r_0)$ for a function satisfying this inequality.

Globalized KL functions

Definition (KL inequality) *A continuously differentiable function* $f : \mathbb{R}^n \to \mathbb{R}$ *satisfies the KL inequality if there exists* $r_0 > 0$ *and a strictly increasing function* $\psi \in \mathcal{C}^0([0, r_0[) \cap \mathcal{C}^1(]0, r_0[)$ *with* $\psi(0) = 0$ *such that*

$$\psi'(f(\mathbf{z}) - \min f)\, \|\nabla f(\mathbf{z})\| \geq 1, \quad \forall \mathbf{z} \in [\min f < f < \min f + r_0].$$

We use the shorthand notation $f \in$ KŁ$_\psi(r_0)$ *for a function satisfying this inequality.*

- KL is a gradient domination inequality.
- KL expresses the fact that f is sharp under a reparameterization of its values :

$$\|\nabla(\psi \circ (f - \min f))(\mathbf{z})\| \geq 1, \quad \forall \mathbf{z} \in [\min f < f < \min f + r_0],$$

  hence the name ”desingularizing function” for ψ.
- Popular Łojasiewicz inequality : $\psi(s) = cs^\alpha$ with $\alpha \in [0, 1]$.
- KL inequality plays a fundamental role in several fields of applied mathematics among which optimization, neural networks, PDE’s, to cite a few.
- KL closely related to error bounds used to derive complexity bounds of descent-like algorithms.

Globalized KL functions

Definition (KL inequality) *A continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the KL inequality if there exists $r_0 > 0$ and a strictly increasing function $\psi \in \mathcal{C}^0([0, r_0[) \cap \mathcal{C}^1(]0, r_0[)$ with $\psi(0) = 0$ such that*

$$\psi'(f(\mathbf{z}) - \min f) \, \|\nabla f(\mathbf{z})\| \geq 1, \quad \forall \mathbf{z} \in [\min f < f < \min f + r_0].$$

We use the shorthand notation $f \in$ KŁ$_\psi(r_0)$ for a function satisfying this inequality.

- KL is a gradient domination inequality.
- KL expresses the fact that f is sharp under a reparameterization of its values :

$$\|\nabla(\psi \circ (f - \min f))(\mathbf{z})\| \geq 1, \quad \forall \mathbf{z} \in [\min f < f < \min f + r_0],$$

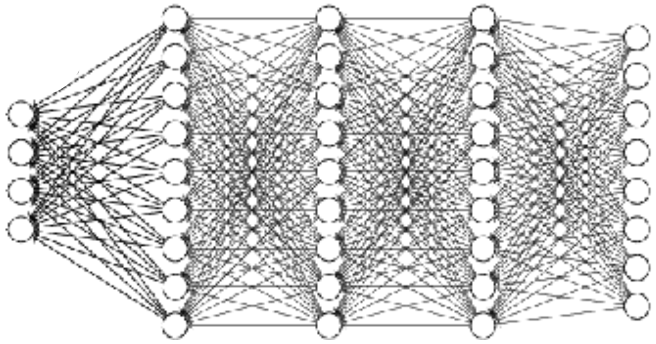
  hence the name ”desingularizing function” for ψ.
- Popular Łojasiewicz inequality : $\psi(s) = cs^\alpha$ with $\alpha \in [0, 1]$.
- KL inequality plays a fundamental role in several fields of applied mathematics among which optimization, neural networks, PDE’s, to cite a few.
- KL closely related to error bounds used to derive complexity bounds of descent-like algorithms.

- Examples :
  - Convex functions with sufficient growth.
  - Uniformly convex functions.
  - Real semi-algebraic functions and more generally, definable functions are KL.
  - Most examples of losses in applications are KL : MSE, ℓ_p-loss, Kullback-Leibler divergence, cross-entropy, etc.

(Inertial) Gradient flow for DIP

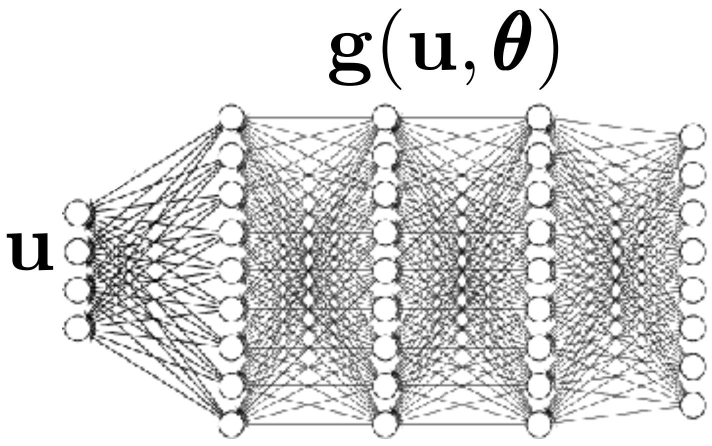
$\mathbf{g}(\mathbf{u}, \boldsymbol{\theta})$

$\mathbf{u}$



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

(Inertial) Gradient flow for DIP

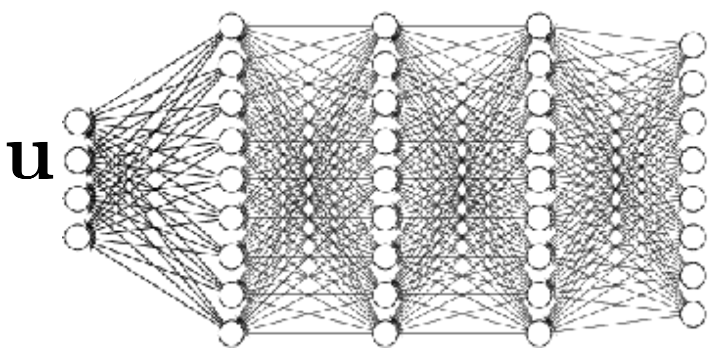


$$\min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta})) \qquad \mathbf{A} \in \mathbb{R}^{m\times n}$$

GF $$\begin{cases} \dot{\boldsymbol{\theta}}(t) = -\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(t))) \\ \boldsymbol{\theta}(0) = \boldsymbol{\theta}_0. \end{cases}$$

GD $$\boldsymbol{\theta}_{\ell+1} = \boldsymbol{\theta}_\ell - s_\ell \nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}_\ell)).$$

(Inertial) Gradient flow for DIP



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

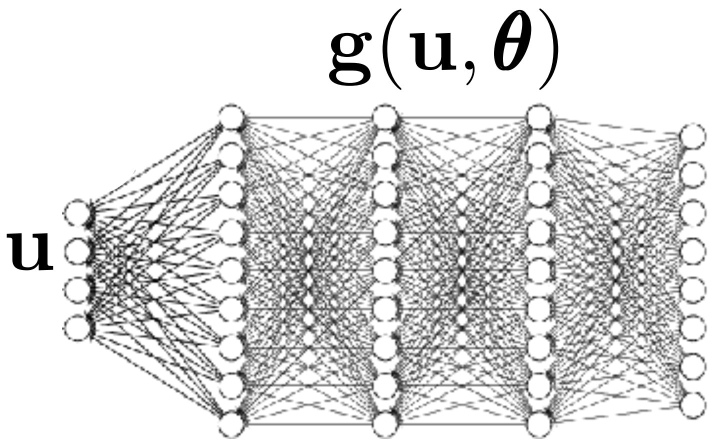
GF
$$\begin{cases}\dot{\boldsymbol{\theta}}(t) = -\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(t)))\\ \boldsymbol{\theta}(0)=\boldsymbol{\theta}_0.\end{cases}$$

GD
$$\boldsymbol{\theta}_{\ell+1} = \boldsymbol{\theta}_\ell - s_\ell\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}_\ell)).$$

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

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

(Inertial) Gradient flow for DIP

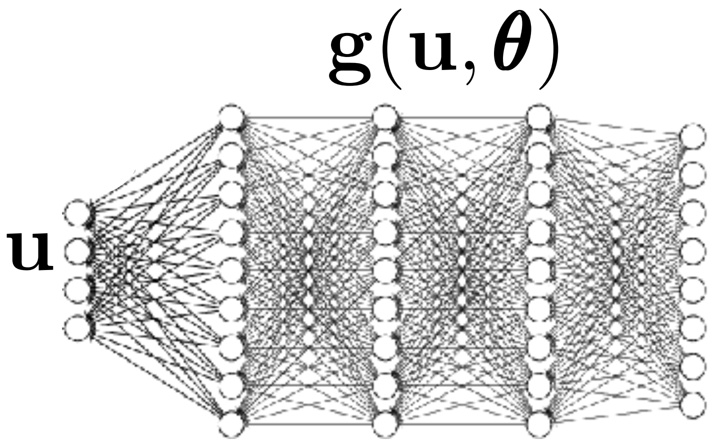


$$\min_{\boldsymbol{\theta} \in \Theta} \mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta})) \qquad \mathbf{A} \in \mathbb{R}^{m \times n}$$

Assumptions on the loss

- WLOG $\min \mathcal{L}_{\mathbf{y}}(\cdot) = 0$.
- $\mathcal{L}_{\mathbf{y}}(\cdot) \in \mathcal{C}^1(\mathbb{R}^m)$ whose gradient is locally Lipschitz continuous.
- $\mathcal{L}_{\mathbf{y}}(\cdot) \in$ KŁ$_{\psi}(\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0)) + \eta)$ for some $\eta > 0$.

(Inertial) Gradient flow for DIP



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

Assumptions on the loss

- WLOG $\min \mathcal{L}_{\mathbf{y}}(\cdot) = 0$.
- $\mathcal{L}_{\mathbf{y}}(\cdot) \in \mathcal{C}^1(\mathbb{R}^m)$ whose gradient is locally Lipschitz continuous.
- $\mathcal{L}_{\mathbf{y}}(\cdot) \in$ KŁ$_\psi(\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0)) + \eta)$ for some $\eta > 0$.

Assumptions on the activation

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

Outline

- Our setting.
- Main recovery guarantees.
- Case of the two-layer DIP.
- Numerical results.
- Conclusion.

Recovery guarantees: observation space

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

Theorem *Suppose that our assumptions hold. Assume that 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

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

Then

(i) the loss converges to 0 at a rate depending solely on ψ, $\sigma_{\mathbf{A}}$ and $\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))$.

(ii) $\boldsymbol{\theta}(t)$ (resp. $\mathbf{x}(t) = \mathbf{g}(\mathbf{u}, \boldsymbol{\theta}(t))$) converges to a global minimizer $\boldsymbol{\theta}_\infty$ of $\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \cdot))$ (resp. $\mathbf{x}_\infty = \mathbf{g}(\mathbf{u}, \boldsymbol{\theta}_\infty)$), at a rate depending solely on the desingularizing function ψ.

(iii) If $\mathrm{Argmin}\,(\mathcal{L}_{\mathbf{y}}(\cdot)) = \{\mathbf{y}\}$, then $\lim_{t \to +\infty} \mathbf{y}(t) = \mathbf{y}$. In addition, if $\mathcal{L}_{\mathbf{y}}$ is convex then

$$\left\| \mathbf{y}(t) - \overline{\mathbf{y}} \right\| \leq 2 \left\| \varepsilon \right\| \quad \textit{when} \quad t \geq \frac{4\Psi(\psi^{-1}(\|\varepsilon\|))}{\sigma_{\mathbf{A}}^2 \sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))^2} - \Psi(\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))),$$

with Ψ a primitive of $-\psi'^2$.

Recovery guarantees: observation space

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

Theorem *Suppose that our assumptions hold. Assume that 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

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

Non degenerate initialization

Then

(i) the loss converges to 0 at a rate depending solely on ψ, $\sigma_{\mathbf{A}}$ and $\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))$.

(ii) $\boldsymbol{\theta}(t)$ (resp. $\mathbf{x}(t) = \mathbf{g}(\mathbf{u}, \boldsymbol{\theta}(t))$) converges to a global minimizer $\boldsymbol{\theta}_\infty$ of $\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \cdot))$ (resp. $\mathbf{x}_\infty = \mathbf{g}(\mathbf{u}, \boldsymbol{\theta}_\infty)$), at a rate depending solely on the desingularizing function ψ.

(iii) If $\mathrm{Argmin}\,(\mathcal{L}_{\mathbf{y}}(\cdot)) = \{\mathbf{y}\}$, then $\lim_{t \to +\infty} \mathbf{y}(t) = \mathbf{y}$. In addition, if $\mathcal{L}_{\mathbf{y}}$ is convex then

$$\|\mathbf{y}(t) - \overline{\mathbf{y}}\| \leq 2\|\varepsilon\| \quad \textit{when} \quad t \geq \frac{4\Psi(\psi^{-1}(\|\varepsilon\|))}{\sigma_{\mathbf{A}}^2 \sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))^2} - \Psi(\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))),$$

with Ψ a primitive of $-\psi'^2$.

Recovery guarantees: observation space

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

Theorem *Suppose that our assumptions hold. Assume that 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*

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

Non degenerate initialization

Then

(i) the loss converges to 0 *at a rate depending solely on* ψ, $\sigma_{\mathbf{A}}$ *and* $\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))$.

(ii) $\boldsymbol{\theta}(t)$ *(resp.* $\mathbf{x}(t) = \mathbf{g}(\mathbf{u}, \boldsymbol{\theta}(t))$*) converges to a global minimizer* $\boldsymbol{\theta}_\infty$ *of* $\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \cdot))$ *(resp.* $\mathbf{x}_\infty = \mathbf{g}(\mathbf{u}, \boldsymbol{\theta}_\infty)$*), at a rate depending solely on the desingularizing function* ψ.

(iii) If $\mathrm{Argmin}\,(\mathcal{L}_{\mathbf{y}}(\cdot)) = \{\mathbf{y}\}$, *then* $\lim_{t\to+\infty} \mathbf{y}(t) = \mathbf{y}$. *In addition, if* $\mathcal{L}_{\mathbf{y}}$ *is convex then*

Trajectory close to initialization

$$\|\mathbf{y}(t) - \overline{\mathbf{y}}\| \leq 2\,\|\varepsilon\| \quad \textit{when} \quad t \geq \frac{4\Psi(\psi^{-1}(\|\varepsilon\|))}{\sigma_{\mathbf{A}}^2\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))^2} - \Psi(\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))),$$

with Ψ *a primitive of* $-\psi'^2$.

Recovery guarantees: observation space

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

Theorem *Suppose that our assumptions hold. Assume that 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

$$R' = \frac{2}{\sigma_{\mathbf{A}} \sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))} \psi(\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))) \quad \textit{and} \quad R = \frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))}{2\text{Lip}_{\mathbb{B}(\boldsymbol{\theta}_0, R)}(\mathcal{J}_{\mathbf{g}})}.$$

Non degenerate initialization

Then

(i) the loss converges to 0 at a rate depending solely on ψ, $\sigma_{\mathbf{A}}$ and $\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))$.

(ii) $\boldsymbol{\theta}(t)$ (resp. $\mathbf{x}(t) = \mathbf{g}(\mathbf{u}, \boldsymbol{\theta}(t))$) converges to a global minimizer $\boldsymbol{\theta}_{\infty}$ of $\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \cdot))$ (resp. $\mathbf{x}_{\infty} = \mathbf{g}(\mathbf{u}, \boldsymbol{\theta}_{\infty})$), at a rate depending solely on the desingularizing function ψ.

(iii) If $\text{Argmin}\,(\mathcal{L}_{\mathbf{y}}(\cdot)) = \{\mathbf{y}\}$, then $\lim_{t \to +\infty} \mathbf{y}(t) = \mathbf{y}$. In addition, if $\mathcal{L}_{\mathbf{y}}$ is convex then

Trajectory close to initialization

$$\|\mathbf{y}(t) - \overline{\mathbf{y}}\| \leq 2\,\|\varepsilon\| \quad \textit{when} \quad t \geq \frac{4\Psi(\psi^{-1}(\|\varepsilon\|))}{\sigma_{\mathbf{A}}^2 \sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))^2} - \Psi(\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))),$$

with Ψ a primitive of $-\psi'^2$.

Stable recovery of $\overline{\mathbf{y}}$ by early stopping

Recovery guarantees: observation space

$$\psi(s) = cs^{\alpha},\ \alpha \in [0,1]$$

$$\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)) \leq \begin{cases} O(t^{-\frac{1}{1-2\alpha}}) & 0 < \alpha < \frac{1}{2} \\ O\left(\exp\left(-\frac{\sigma_{\mathbf{A}}^2 \sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))^2}{c^2} t\right)\right) & \alpha = \frac{1}{2} \\ C(\hat{t} - t)^{-\frac{1}{2\alpha-1}} & \frac{1}{2} < \alpha < 1 \text{ and } t \leq \hat{t} \\ 0 & \frac{1}{2} < \alpha < 1 \text{ and } t > \hat{t}. \end{cases}$$

$$\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}_{\infty}\| \leq \begin{cases} O(t^{-\frac{\alpha}{1-2\alpha}}) & 0 < \alpha < \frac{1}{2} \\ O\left(\exp\left(-\frac{\sigma_{\mathbf{A}}^2 \sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))^2}{c^2} t\right)\right) & \alpha = \frac{1}{2} \\ C(\hat{t} - t)^{-\frac{\alpha}{2\alpha-1}} & \frac{1}{2} < \alpha < 1 \text{ and } t \leq \hat{t} \\ 0 & \frac{1}{2} < \alpha < 1 \text{ and } t > \hat{t}. \end{cases}$$

Recovery guarantees: observation space

$$\psi(s) = cs^{\alpha},\ \alpha \in [0,1]$$

$$\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)) \leq \begin{cases} O(t^{-\frac{1}{1-2\alpha}}) & 0 < \alpha < \frac{1}{2} \\ O\left(\exp\left(-\frac{\sigma_{\mathbf{A}}^2 \sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))^2}{c^2} t\right)\right) & \alpha = \frac{1}{2} \\ C(\hat{t}-t)^{-\frac{1}{2\alpha-1}} & \frac{1}{2} < \alpha < 1 \text{ and } t \leq \hat{t} \\ 0 & \frac{1}{2} < \alpha < 1 \text{ and } t > \hat{t}. \end{cases}$$

Sublinear rate

$$\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}_{\infty}\| \leq \begin{cases} O(t^{-\frac{\alpha}{1-2\alpha}}) & 0 < \alpha < \frac{1}{2} \\ O\left(\exp\left(-\frac{\sigma_{\mathbf{A}}^2 \sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))^2}{c^2} t\right)\right) & \alpha = \frac{1}{2} \\ C(\hat{t}-t)^{-\frac{\alpha}{2\alpha-1}} & \frac{1}{2} < \alpha < 1 \text{ and } t \leq \hat{t} \\ 0 & \frac{1}{2} < \alpha < 1 \text{ and } t > \hat{t}. \end{cases}$$

Sublinear rate

Recovery guarantees: observation space

$$\psi(s) = cs^{\alpha},\ \alpha \in [0,1]$$

$$\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)) \leq \begin{cases} O(t^{-\frac{1}{1-2\alpha}}) & 0 < \alpha < \frac{1}{2} \\ O\left(\exp\left(-\frac{\sigma_{\mathbf{A}}^2 \sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))^2}{c^2} t\right)\right) & \alpha = \frac{1}{2} \\ C(\hat{t}-t)^{-\frac{1}{2\alpha-1}} & \frac{1}{2} < \alpha < 1 \text{ and } t \leq \hat{t} \\ 0 & \frac{1}{2} < \alpha < 1 \text{ and } t > \hat{t}. \end{cases}$$

Sublinear rate

Linear rate

$$\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}_{\infty}\| \leq \begin{cases} O(t^{-\frac{\alpha}{1-2\alpha}}) & 0 < \alpha < \frac{1}{2} \\ O\left(\exp\left(-\frac{\sigma_{\mathbf{A}}^2 \sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))^2}{c^2} t\right)\right) & \alpha = \frac{1}{2} \\ C(\hat{t}-t)^{-\frac{\alpha}{2\alpha-1}} & \frac{1}{2} < \alpha < 1 \text{ and } t \leq \hat{t} \\ 0 & \frac{1}{2} < \alpha < 1 \text{ and } t > \hat{t}. \end{cases}$$

Sublinear rate

Linear rate

Recovery guarantees: observation space

$$\psi(s) = cs^{\alpha},\ \alpha \in [0,1]$$

$$\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)) \leq \begin{cases} O\left(t^{-\frac{1}{1-2\alpha}}\right) & 0 < \alpha < \frac{1}{2} \\ O\left(\exp\left(-\frac{\sigma_{\mathbf{A}}^2 \sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))^2}{c^2} t\right)\right) & \alpha = \frac{1}{2} \\ C(\hat{t} - t)^{-\frac{1}{2\alpha-1}} & \frac{1}{2} < \alpha < 1 \text{ and } t \leq \hat{t} \\ 0 & \frac{1}{2} < \alpha < 1 \text{ and } t > \hat{t}. \end{cases}$$

Sublinear rate

Linear rate

Finite termination

$$\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}_{\infty}\| \leq \begin{cases} O\left(t^{-\frac{\alpha}{1-2\alpha}}\right) & 0 < \alpha < \frac{1}{2} \\ O\left(\exp\left(-\frac{\sigma_{\mathbf{A}}^2 \sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))^2}{c^2} t\right)\right) & \alpha = \frac{1}{2} \\ C(\hat{t} - t)^{-\frac{\alpha}{2\alpha-1}} & \frac{1}{2} < \alpha < 1 \text{ and } t \leq \hat{t} \\ 0 & \frac{1}{2} < \alpha < 1 \text{ and } t > \hat{t}. \end{cases}$$

Sublinear rate

Linear rate

Finite termination

Recovery guarantees: observation space

Recovery guarantees: observation space

- All claims rely on the fact for a good initial point, the whole trajectory remains in a ball around it.

Recovery guarantees: observation space

- All claims rely on the fact for a good initial point, the whole trajectory remains in a ball around it.
- Closely related to the Hartman–Grobman theorem:
  - local behaviour of an autonomous dynamical system in the neighbourhood of a hyperbolic equilibrium point is topologically conjugate to its linearization.

Recovery guarantees: observation space

- All claims rely on the fact for a good initial point, the whole trajectory remains in a ball around it.
- Closely related to the Hartman–Grobman theorem:
  - local behaviour of an autonomous dynamical system in the neighbourhood of a hyperbolic equilibrium point is topologically conjugate to its linearization.
- Relation to conservation laws (and symmetries of variational problems via E. Noether's Theorem) of the gradient flow seen as an isolated evolving physical system.

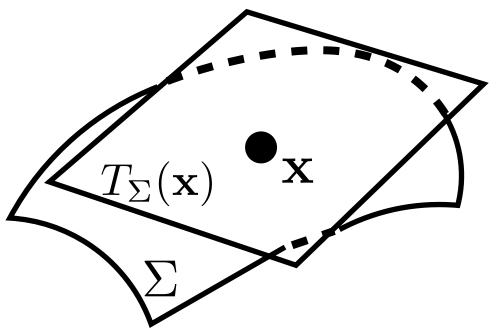
Recovery guarantees: parameter space

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

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

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

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



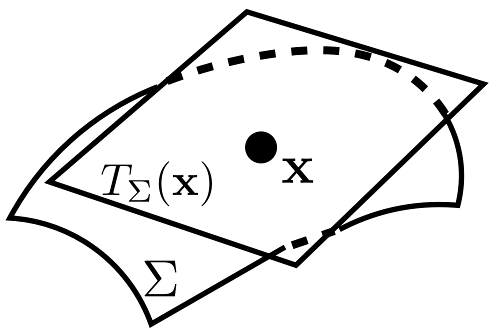
Recovery guarantees: parameter space

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

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

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

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



Theorem *Suppose that our assumptions hold. Assume that the gradient flow is initialized as before. If $\mathcal{L}_{\mathbf{y}}$ is convex,* $\mathrm{Argmin}\,(\mathcal{L}_{\mathbf{y}}(\cdot)) = \{\mathbf{y}\}$*, and*

$$\mathrm{Ker}(\mathbf{A}) \cap T_{\Sigma}(\overline{\mathbf{x}}_{\Sigma}) = \{0\},$$

then

$$\|\mathbf{x}(t) - \overline{\mathbf{x}}\| \leq \frac{2\psi\left(\Psi^{-1}\left(\frac{\sigma_{\mathbf{A}}^2 \sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))^2}{4} t - \hat{t}\right)\right)}{\lambda_{\min}(\mathbf{A}; T_{\Sigma}(\overline{\mathbf{x}}_{\Sigma}))\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}} + \left(1 + \frac{\|\mathbf{A}\|}{\lambda_{\min}(\mathbf{A}; T_{\Sigma}(\overline{\mathbf{x}}_{\Sigma}))}\right) \mathrm{dist}(\overline{\mathbf{x}}, \Sigma) + \frac{\|\varepsilon\|}{\lambda_{\min}(\mathbf{A}; T_{\Sigma}(\overline{\mathbf{x}}_{\Sigma}))}.$$

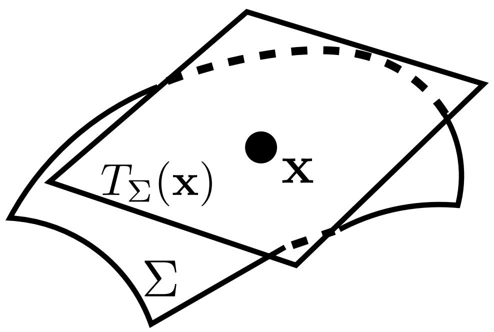
Recovery guarantees: parameter space

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

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

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

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



Theorem *Suppose that our assumptions hold. Assume that the gradient flow is initialized as before. If $\mathcal{L}_{\mathbf{y}}$ is convex,* $\mathrm{Argmin}\,(\mathcal{L}_{\mathbf{y}}(\cdot)) = \{\mathbf{y}\}$*, and*

$$\mathrm{Ker}(\mathbf{A}) \cap T_{\Sigma}(\overline{\mathbf{x}}_{\Sigma}) = \{0\},$$

Restricted Injectivity Condition (RIC)

then

$$\|\mathbf{x}(t) - \overline{\mathbf{x}}\| \leq \frac{2\psi\left(\Psi^{-1}\left(\frac{\sigma_{\mathbf{A}}^2 \sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))^2}{4} t - \hat{t}\right)\right)}{\lambda_{\min}(\mathbf{A}; T_{\Sigma}(\overline{\mathbf{x}}_{\Sigma}))\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}} + \left(1 + \frac{\|\mathbf{A}\|}{\lambda_{\min}(\mathbf{A}; T_{\Sigma}(\overline{\mathbf{x}}_{\Sigma}))}\right)\mathrm{dist}(\overline{\mathbf{x}}, \Sigma) + \frac{\|\varepsilon\|}{\lambda_{\min}(\mathbf{A}; T_{\Sigma}(\overline{\mathbf{x}}_{\Sigma}))}.$$

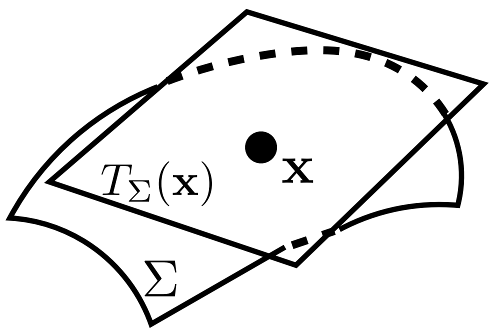
Recovery guarantees: parameter space

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

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

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

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



Theorem *Suppose that our assumptions hold. Assume that the gradient flow is initialized as before. If $\mathcal{L}_{\mathbf{y}}$ is convex,* $\mathrm{Argmin}\,(\mathcal{L}_{\mathbf{y}}(\cdot)) = \{\mathbf{y}\}$*, and*

$$\mathrm{Ker}(\mathbf{A}) \cap T_\Sigma(\overline{\mathbf{x}}_\Sigma) = \{0\},$$

Restricted Injectivity Condition (RIC)

then

$$\|\mathbf{x}(t) - \overline{\mathbf{x}}\| \leq \frac{2\psi\left(\Psi^{-1}\left(\frac{\sigma_{\mathbf{A}}^2 \sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))^2}{4} t - \hat{t}\right)\right)}{\lambda_{\min}(\mathbf{A}; T_\Sigma(\overline{\mathbf{x}}_\Sigma))\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}} + \left(1 + \frac{\|\mathbf{A}\|}{\lambda_{\min}(\mathbf{A}; T_\Sigma(\overline{\mathbf{x}}_\Sigma))}\right)\mathrm{dist}(\overline{\mathbf{x}}, \Sigma) + \frac{\|\varepsilon\|}{\lambda_{\min}(\mathbf{A}; T_\Sigma(\overline{\mathbf{x}}_\Sigma))}.$$

Optimization error

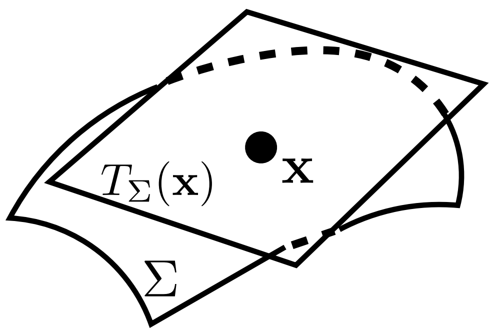
Recovery guarantees: parameter space

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

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

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

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



Theorem *Suppose that our assumptions hold. Assume that the gradient flow is initialized as before. If $\mathcal{L}_{\mathbf{y}}$ is convex,* $\mathrm{Argmin}\,(\mathcal{L}_{\mathbf{y}}(\cdot)) = \{\mathbf{y}\}$, *and*

$$\mathrm{Ker}(\mathbf{A}) \cap T_\Sigma(\overline{\mathbf{x}}_\Sigma) = \{0\},$$

Restricted Injectivity Condition (RIC)

then

$$\|\mathbf{x}(t) - \overline{\mathbf{x}}\| \leq \frac{2\psi\left(\Psi^{-1}\left(\frac{\sigma_{\mathbf{A}}^2 \sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))^2}{4} t - \hat{t}\right)\right)}{\lambda_{\min}(\mathbf{A}; T_\Sigma(\overline{\mathbf{x}}_\Sigma))\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}} + \left(1 + \frac{\|\mathbf{A}\|}{\lambda_{\min}(\mathbf{A}; T_\Sigma(\overline{\mathbf{x}}_\Sigma))}\right)\mathrm{dist}(\overline{\mathbf{x}}, \Sigma) + \frac{\|\varepsilon\|}{\lambda_{\min}(\mathbf{A}; T_\Sigma(\overline{\mathbf{x}}_\Sigma))}.$$

Optimization error — Approximation error

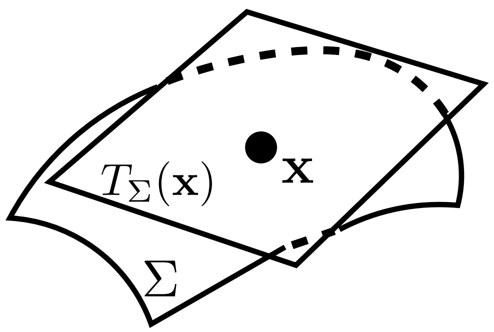
Recovery guarantees: parameter space

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

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

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

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



Theorem *Suppose that our assumptions hold. Assume that the gradient flow is initialized as before. If $\mathcal{L}_{\mathbf{y}}$ is convex,* $\mathrm{Argmin}\,(\mathcal{L}_{\mathbf{y}}(\cdot)) = \{\mathbf{y}\}$*, and*

$$\mathrm{Ker}(\mathbf{A}) \cap T_\Sigma(\overline{\mathbf{x}}_\Sigma) = \{0\},$$

Restricted Injectivity Condition (RIC)

then

$$\|\mathbf{x}(t) - \overline{\mathbf{x}}\| \le \frac{2\psi\left(\Psi^{-1}\left(\frac{\sigma_{\mathbf{A}}^2\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))^2}{4}t - \hat{t}\right)\right)}{\lambda_{\min}(\mathbf{A}; T_\Sigma(\overline{\mathbf{x}}_\Sigma))\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}} + \left(1 + \frac{\|\mathbf{A}\|}{\lambda_{\min}(\mathbf{A}; T_\Sigma(\overline{\mathbf{x}}_\Sigma))}\right)\mathrm{dist}(\overline{\mathbf{x}}, \Sigma) + \frac{\|\varepsilon\|}{\lambda_{\min}(\mathbf{A}; T_\Sigma(\overline{\mathbf{x}}_\Sigma))}.$$

Optimization error — Approximation error — Noise error

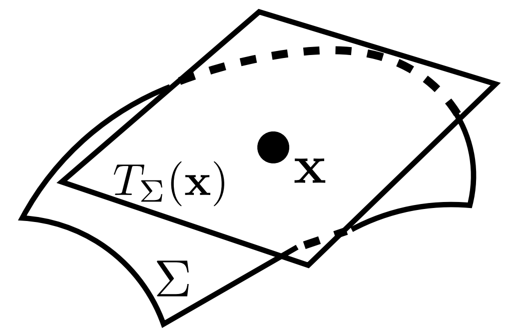
Recovery guarantees: parameter space

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

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

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

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



Theorem *Suppose that our assumptions hold. Assume that the gradient flow is initialized as before. If* $\mathcal{L}_{\mathbf{y}}$ *is convex,* $\mathrm{Argmin}\,(\mathcal{L}_{\mathbf{y}}(\cdot)) = \{\mathbf{y}\}$*, and*

$$\mathrm{Ker}(\mathbf{A}) \cap T_{\Sigma}(\overline{\mathbf{x}}_{\Sigma}) = \{0\},$$

Restricted Injectivity Condition (RIC)

then

$$\|\mathbf{x}(t) - \overline{\mathbf{x}}\| \leq \frac{2\psi\left(\Psi^{-1}\left(\frac{\sigma_{\mathbf{A}}^2 \sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))^2}{4} t - \hat{t}\right)\right)}{\lambda_{\min}(\mathbf{A}; T_{\Sigma}(\overline{\mathbf{x}}_{\Sigma}))\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}} + \left(1 + \frac{\|\mathbf{A}\|}{\lambda_{\min}(\mathbf{A}; T_{\Sigma}(\overline{\mathbf{x}}_{\Sigma}))}\right) \mathrm{dist}(\overline{\mathbf{x}}, \Sigma) + \frac{\|\varepsilon\|}{\lambda_{\min}(\mathbf{A}; T_{\Sigma}(\overline{\mathbf{x}}_{\Sigma}))}.$$

Optimization error — Approximation error — Noise error

- Sample bounds for RIC can be given in a compressed sensing framework via the Gaussian width of the tangent cone.

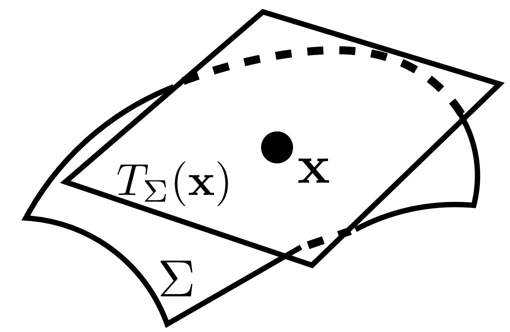
Recovery guarantees: parameter space

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

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

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

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



Theorem *Suppose that our assumptions hold. Assume that the gradient flow is initialized as before. If $\mathcal{L}_{\mathbf{y}}$ is convex, $\mathrm{Argmin}\,(\mathcal{L}_{\mathbf{y}}(\cdot)) = \{\mathbf{y}\}$, and*

$$\mathrm{Ker}(\mathbf{A}) \cap T_{\Sigma}(\overline{\mathbf{x}}_{\Sigma}) = \{0\},$$

Restricted Injectivity Condition (RIC)

then

$$\|\mathbf{x}(t) - \overline{\mathbf{x}}\| \leq \frac{2\psi\left(\Psi^{-1}\left(\frac{\sigma_{\mathbf{A}}^{2}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))^{2}}{4}t - \hat{t}\right)\right)}{\lambda_{\min}(\mathbf{A}; T_{\Sigma}(\overline{\mathbf{x}}_{\Sigma}))\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}} + \left(1 + \frac{\|\mathbf{A}\|}{\lambda_{\min}(\mathbf{A}; T_{\Sigma}(\overline{\mathbf{x}}_{\Sigma}))}\right)\mathrm{dist}(\overline{\mathbf{x}}, \Sigma) + \frac{\|\varepsilon\|}{\lambda_{\min}(\mathbf{A}; T_{\Sigma}(\overline{\mathbf{x}}_{\Sigma}))}.$$

Optimization error — Approximation error — Noise error

- Sample bounds for RIC 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.

Inertial system with Hessian damping

$$\text{ISEHD}\begin{cases}\ddot{\boldsymbol{\theta}}(t)+\alpha\dot{\boldsymbol{\theta}}(t)+\beta\frac{\mathrm{d}}{\mathrm{d}t}\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(t)))+\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(t)))=0\\ \boldsymbol{\theta}(0)=\boldsymbol{\theta}_0,\dot{\boldsymbol{\theta}}(0)=0.\end{cases}\qquad \begin{aligned}\alpha&=\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}\\ \beta&=\frac{1}{2\alpha}\end{aligned}$$

Theorem *Suppose that our assumptions hold. Assume that the inertial gradient flow is initialized merely as before. If* $\mathcal{L}_{\mathbf{y}}$ *is* $\|\cdot\|^2$ *and*

$$\operatorname{Ker}(\mathbf{A})\cap T_{\Sigma}(\overline{\mathbf{x}}_{\Sigma})=\{0\}\,,$$

then

$$\|\mathbf{x}(t)-\overline{\mathbf{x}}\|\leq\frac{C\sqrt{\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))}\exp\left(-\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}}{8}t\right)}{\lambda_{\min}(\mathbf{A};T_{\Sigma}(\overline{\mathbf{x}}_{\Sigma}))}+\frac{\|\varepsilon\|}{\lambda_{\min}(\mathbf{A};T_{\Sigma}(\overline{\mathbf{x}}_{\Sigma}))}+\left(1+\frac{\|\mathbf{A}\|}{\lambda_{\min}(\mathbf{A};T_{\Sigma}(\overline{\mathbf{x}}_{\Sigma}))}\right)\operatorname{dist}(\overline{\mathbf{x}},\Sigma)$$

Inertial system with Hessian damping

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

$$\alpha=\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}$$
$$\beta=\frac{1}{2\alpha}$$

Theorem *Suppose that our assumptions hold. Assume that the inertial gradient flow is initialized merely as before. If $\mathcal{L}_{\mathbf{y}}$ is $\|\cdot\|^2$ and*

$$\mathrm{Ker}(\mathbf{A})\cap T_{\Sigma}(\overline{\mathbf{x}}_{\Sigma})=\{0\},$$ Restricted Injectivity Condition (RIC)

then

$$\|\mathbf{x}(t)-\overline{\mathbf{x}}\|\leq\frac{C\sqrt{\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))}\exp\left(-\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}}{8}t\right)}{\lambda_{\min}(\mathbf{A};T_{\Sigma}(\overline{\mathbf{x}}_{\Sigma}))}+\frac{\|\varepsilon\|}{\lambda_{\min}(\mathbf{A};T_{\Sigma}(\overline{\mathbf{x}}_{\Sigma}))}+\left(1+\frac{\|\mathbf{A}\|}{\lambda_{\min}(\mathbf{A};T_{\Sigma}(\overline{\mathbf{x}}_{\Sigma}))}\right)\mathrm{dist}(\overline{\mathbf{x}},\Sigma)$$

Inertial system with Hessian damping

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

$$\alpha=\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}$$
$$\beta=\frac{1}{2\alpha}$$

Theorem *Suppose that our assumptions hold. Assume that the inertial gradient flow is initialized merely as before. If $\mathcal{L}_{\mathbf{y}}$ is $\|\cdot\|^2$ and*

$$\mathrm{Ker}(\mathbf{A})\cap T_\Sigma(\overline{\mathbf{x}}_\Sigma)=\{0\},$$

Restricted Injectivity Condition (RIC)

then

$$\|\mathbf{x}(t)-\overline{\mathbf{x}}\|\leq\frac{C\sqrt{\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))}\exp\left(-\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}}{8}t\right)}{\lambda_{\min}(\mathbf{A};T_\Sigma(\overline{\mathbf{x}}_\Sigma))}+\frac{\|\varepsilon\|}{\lambda_{\min}(\mathbf{A};T_\Sigma(\overline{\mathbf{x}}_\Sigma))}+\left(1+\frac{\|\mathbf{A}\|}{\lambda_{\min}(\mathbf{A};T_\Sigma(\overline{\mathbf{x}}_\Sigma))}\right)\mathrm{dist}(\overline{\mathbf{x}},\Sigma)$$

Optimization error

Inertial system with Hessian damping

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

$$\alpha=\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}$$

$$\beta=\frac{1}{2\alpha}$$

Theorem *Suppose that our assumptions hold. Assume that the inertial gradient flow is initialized merely as before. If $\mathcal{L}_{\mathbf{y}}$ is $\|\cdot\|^2$ and*

$$\mathrm{Ker}(\mathbf{A})\cap T_\Sigma(\overline{\mathbf{x}}_\Sigma)=\{0\},$$ Restricted Injectivity Condition (RIC)

then

$$\|\mathbf{x}(t)-\overline{\mathbf{x}}\|\leq\frac{C\sqrt{\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))}\exp\left(-\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}}{8}t\right)}{\lambda_{\min}(\mathbf{A};T_\Sigma(\overline{\mathbf{x}}_\Sigma))}+\frac{\|\varepsilon\|}{\lambda_{\min}(\mathbf{A};T_\Sigma(\overline{\mathbf{x}}_\Sigma))}+\left(1+\frac{\|\mathbf{A}\|}{\lambda_{\min}(\mathbf{A};T_\Sigma(\overline{\mathbf{x}}_\Sigma))}\right)\mathrm{dist}(\overline{\mathbf{x}},\Sigma)$$

Optimization error

Approximation error

Inertial system with Hessian damping

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

$$\alpha=\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}$$
$$\beta=\frac{1}{2\alpha}$$

Theorem *Suppose that our assumptions hold. Assume that the inertial gradient flow is initialized merely as before. If $\mathcal{L}_{\mathbf{y}}$ is $\|\cdot\|^2$ and*

$$\mathrm{Ker}(\mathbf{A})\cap T_\Sigma(\overline{\mathbf{x}}_\Sigma)=\{0\},$$ Restricted Injectivity Condition (RIC)

then

$$\|\mathbf{x}(t)-\overline{\mathbf{x}}\|\leq\frac{C\sqrt{\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))}\exp\left(-\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}}{8}t\right)}{\lambda_{\min}(\mathbf{A};T_\Sigma(\overline{\mathbf{x}}_\Sigma))}+\frac{\|\varepsilon\|}{\lambda_{\min}(\mathbf{A};T_\Sigma(\overline{\mathbf{x}}_\Sigma))}+\left(1+\frac{\|\mathbf{A}\|}{\lambda_{\min}(\mathbf{A};T_\Sigma(\overline{\mathbf{x}}_\Sigma))}\right)\mathrm{dist}(\overline{\mathbf{x}},\Sigma)$$

Optimization error | Noise error | Approximation error

Inertial system with Hessian damping

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

$$\alpha=\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}$$
$$\beta=\frac{1}{2\alpha}$$

Theorem *Suppose that our assumptions hold. Assume that the inertial gradient flow is initialized merely as before. If $\mathcal{L}_{\mathbf{y}}$ is $\|\cdot\|^2$ and*

$$\mathrm{Ker}(\mathbf{A})\cap T_{\Sigma}(\overline{\mathbf{x}}_{\Sigma})=\{0\},$$
Restricted Injectivity Condition (RIC)

then

$$\|\mathbf{x}(t)-\overline{\mathbf{x}}\|\leq\frac{C\sqrt{\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))}\exp\left(-\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}}{8}t\right)}{\lambda_{\min}(\mathbf{A};T_{\Sigma}(\overline{\mathbf{x}}_{\Sigma}))}+\frac{\|\varepsilon\|}{\lambda_{\min}(\mathbf{A};T_{\Sigma}(\overline{\mathbf{x}}_{\Sigma}))}+\left(1+\frac{\|\mathbf{A}\|}{\lambda_{\min}(\mathbf{A};T_{\Sigma}(\overline{\mathbf{x}}_{\Sigma}))}\right)\mathrm{dist}(\overline{\mathbf{x}},\Sigma)$$

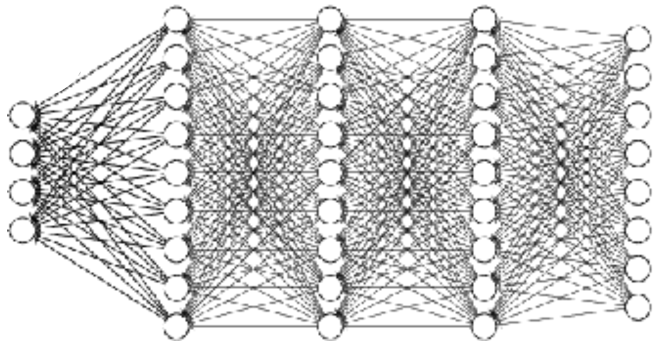
Optimization error — Noise error — Approximation error

- 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 : $O\left(\exp\left(-\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{A}}}{8}t\right)\right)$.

Outline

- Our setting.
- Main recovery guarantees.
- Case of the two-layer DIP.
- Numerical results.
- Conclusion.

Non degenerate initialization



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

$$\begin{cases} \dot{\boldsymbol{\theta}}(t) = -\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(t))) \\ \boldsymbol{\theta}(0) = \boldsymbol{\theta}_0. \end{cases}$$

Theorem *Suppose that our assumptions hold. Assume that 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*

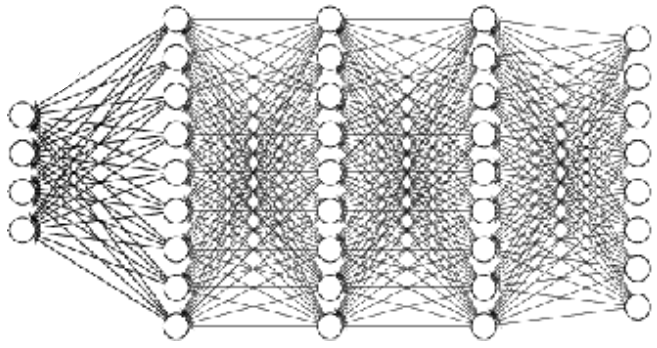
$$R' = \frac{2}{\sigma_{\mathbf{A}}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))}\psi(\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))) \quad \textit{and} \quad R = \frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))}{2\mathrm{Lip}_{\mathbb{B}(\boldsymbol{\theta}_0,R)}(\mathcal{J}_{\mathbf{g}})}.$$

Non degenerate initialization

etc.

Non degenerate initialization

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



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

$$\begin{cases} \dot{\boldsymbol{\theta}}(t) = -\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(t))) \\ \boldsymbol{\theta}(0) = \boldsymbol{\theta}_0. \end{cases}$$

Theorem *Suppose that our assumptions hold. Assume that 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*

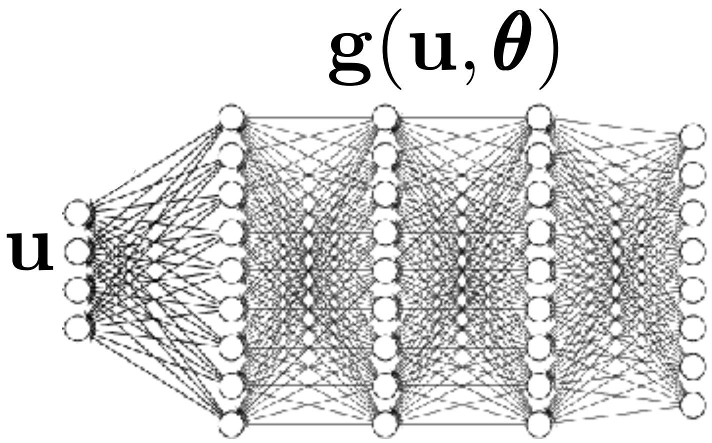
$$R' = \frac{2}{\sigma_{\mathbf{A}}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))}\psi(\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))) \quad \textit{and} \quad R = \frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))}{2\mathrm{Lip}_{\mathbb{B}(\boldsymbol{\theta}_0,R)}(\mathcal{J}_{\mathbf{g}})}.$$

Non degenerate initialization

etc.

How to ensure this ?

Non degenerate initialization



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

$$\begin{cases} \dot{\boldsymbol{\theta}}(t) = -\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(t))) \\ \boldsymbol{\theta}(0) = \boldsymbol{\theta}_0. \end{cases}$$

Theorem *Suppose that our assumptions hold. Assume that 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*

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

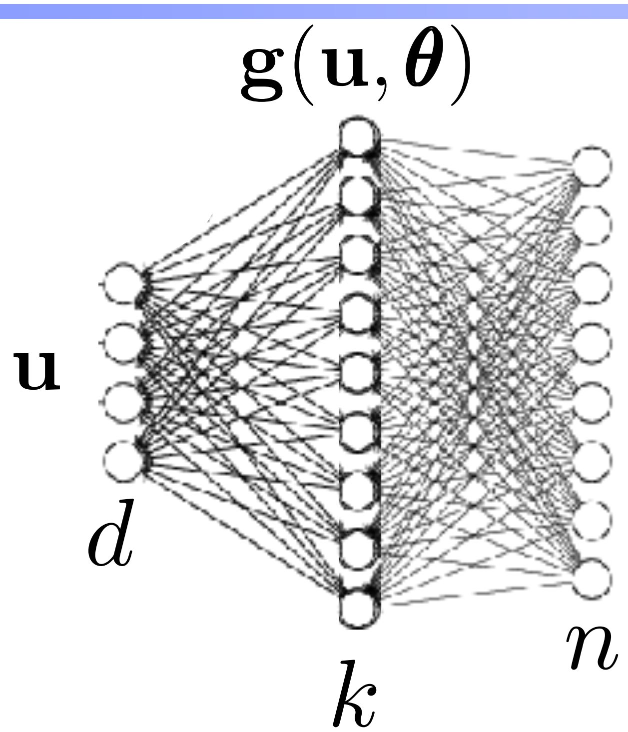
Non degenerate initialization

etc.

How to ensure this ?

The role of overparametrization

Two-layer DIP

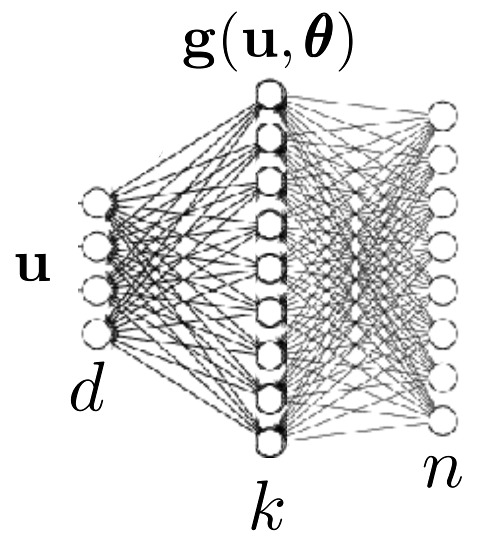


$$\mathbf{g}(\mathbf{u},\boldsymbol{\theta}) = \frac{1}{\sqrt{k}}\mathbf{V}\phi(\mathbf{W}\mathbf{u})$$

$$\begin{cases} \dot{\boldsymbol{\theta}}(t) = -\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(t))) \\ \boldsymbol{\theta}(0) = \boldsymbol{\theta}_0. \end{cases}$$

- $\mathbf{u}$ is a 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 has iid columns with identity covariance and D-bounded centred entries.

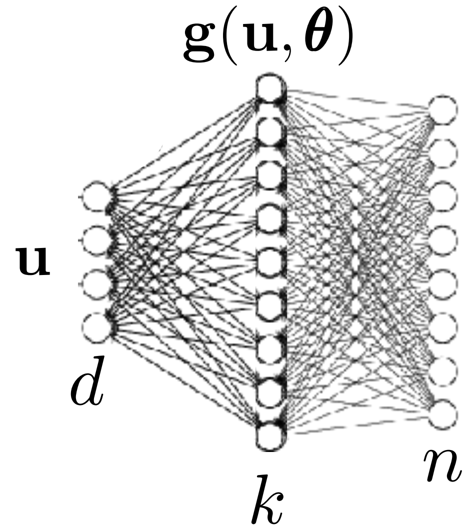
Overparametrization bound



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

$$\begin{cases} \dot{\boldsymbol{\theta}}(t) = -\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(t))) \\ \boldsymbol{\theta}(0) = \boldsymbol{\theta}_0. \end{cases}$$

Overparametrization bound



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

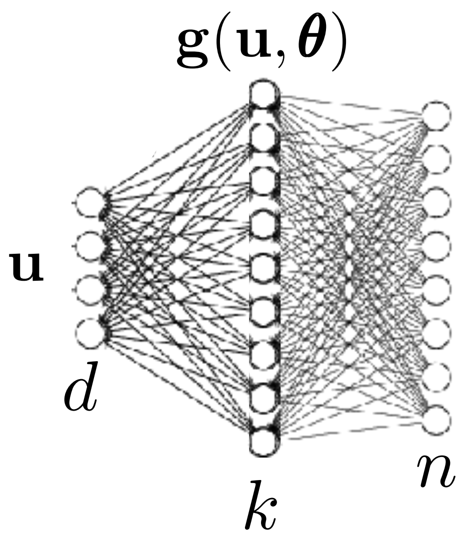
$$\begin{cases} \dot{\boldsymbol{\theta}}(t) = -\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(t))) \\ \boldsymbol{\theta}(0) = \boldsymbol{\theta}_0. \end{cases}$$

Theorem *Consider the one-hidden DIP layer network with the architecture parameters obeying*

$$k \geq C'\sigma_{\mathbf{A}}^{-4} n\psi \left(C\left(\sqrt{n\log(d)} + \sqrt{m}\right)^2\right)^4.$$

Then with probability at least $1 - n^{-1} - d^{-1}$, $\boldsymbol{\theta}(0) = (\mathbf{W}(0), \mathbf{V}(0))$ *is a nondegenerate initial point.*

Overparametrization bound



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

$$\begin{cases} \dot{\boldsymbol{\theta}}(t) = -\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(t))) \\ \boldsymbol{\theta}(0) = \boldsymbol{\theta}_0. \end{cases}$$

Theorem *Consider the one-hidden DIP layer network with the architecture parameters obeying*

$$k \geq C'\sigma_{\mathbf{A}}^{-4} n\psi\left(C\left(\sqrt{n\log(d)} + \sqrt{m}\right)^2\right)^4.$$

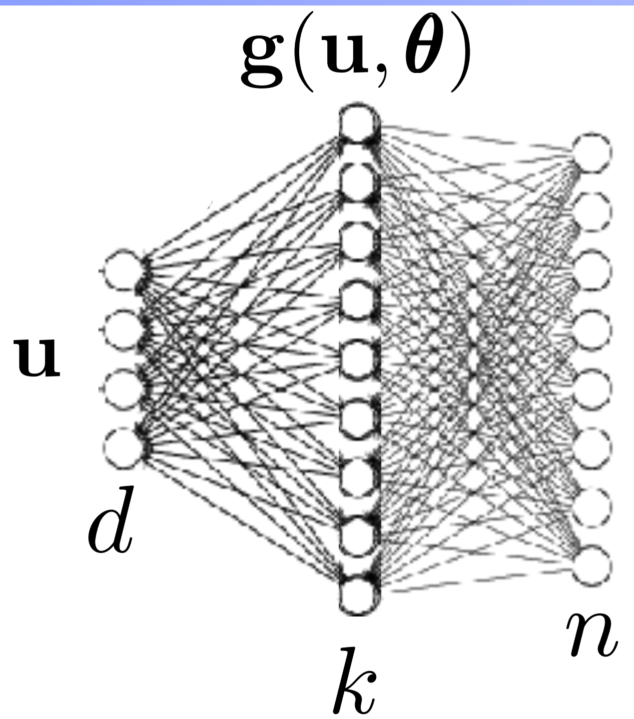
Then with probability at least $1 - n^{-1} - d^{-1}$, $\boldsymbol{\theta}(0) = (\mathbf{W}(0), \mathbf{V}(0))$ *is a nondegenerate initial point.*

- For the MSE loss, the bounds reads : $k \gtrsim n^3 m^2$.
- If $\mathbf{V}$ is fixed and only is $\mathbf{W}$ is optimized for :
  - $k \gtrsim \sigma_{\mathbf{A}}^{-2} n\psi(C(n+m))^2$.
  - MSE : $k \gtrsim n^2 m$.

Outline

- Our setting.
- Main recovery guarantees.
- Case of the two-layer DIP.
- Numerical results.
- Conclusion.

Overparamerization for noiseless MSE



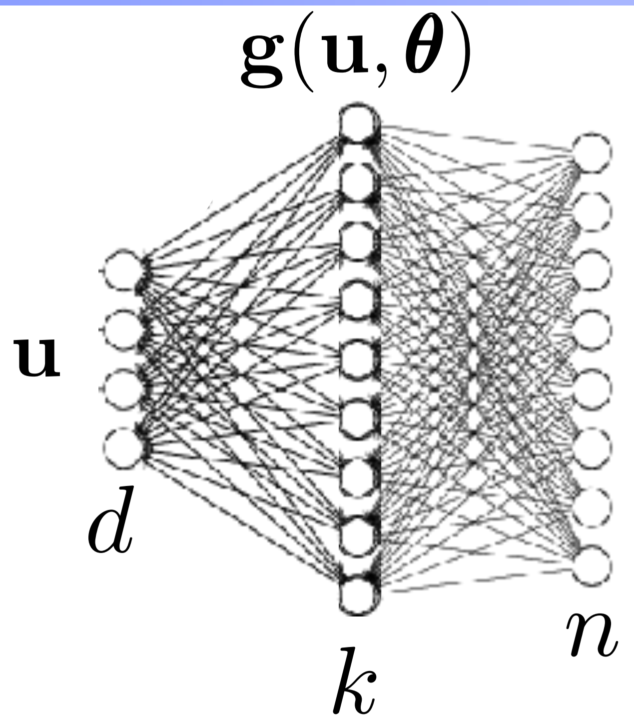
$$\mathbf{g}(\mathbf{u},\boldsymbol{\theta}) = \frac{1}{\sqrt{k}}\mathbf{V}\phi(\mathbf{W}\mathbf{u})$$

$$\begin{cases} \dot{\boldsymbol{\theta}}(t) = -\nabla_{\mathbf{W}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(t))) \\ \boldsymbol{\theta}(0) = \boldsymbol{\theta}_0. \end{cases}$$

$\mathcal{L}_{\mathbf{y}}$: MSE

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

Overparamerization for noiseless MSE

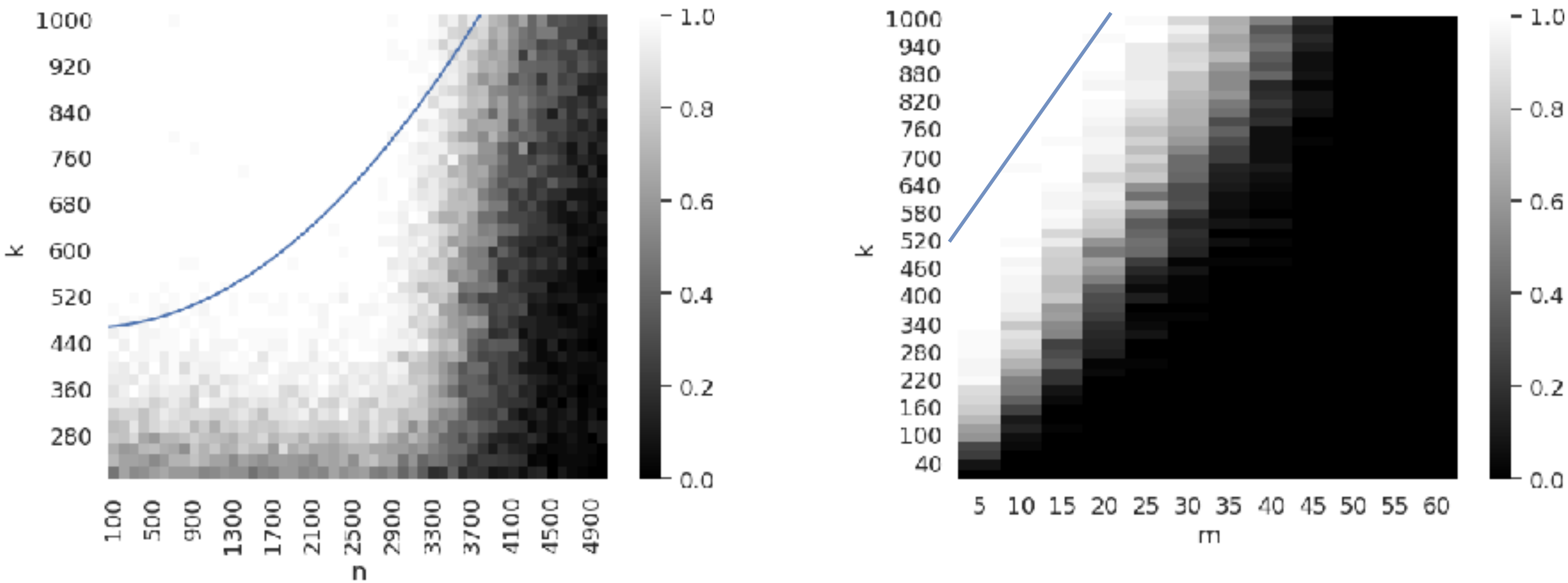


$$\mathbf{g}(\mathbf{u},\boldsymbol{\theta}) = \frac{1}{\sqrt{k}}\mathbf{V}\phi(\mathbf{W}\mathbf{u})$$

$$\begin{cases} \dot{\boldsymbol{\theta}}(t) = -\nabla_{\mathbf{W}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(t))) \\ \boldsymbol{\theta}(0) = \boldsymbol{\theta}_0. \end{cases}$$

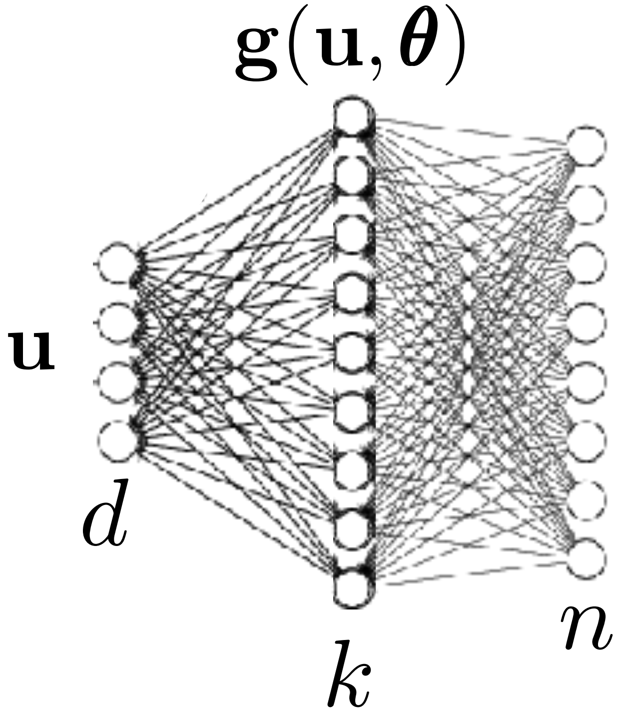
$\mathcal{L}_{\mathbf{y}}$: MSE

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



Probability of converging to zero-loss for networks with different architecture parameters confirming our theoretical predictions $k \gtrsim n^2 m$.

Signal recovery under ill-conditioning



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

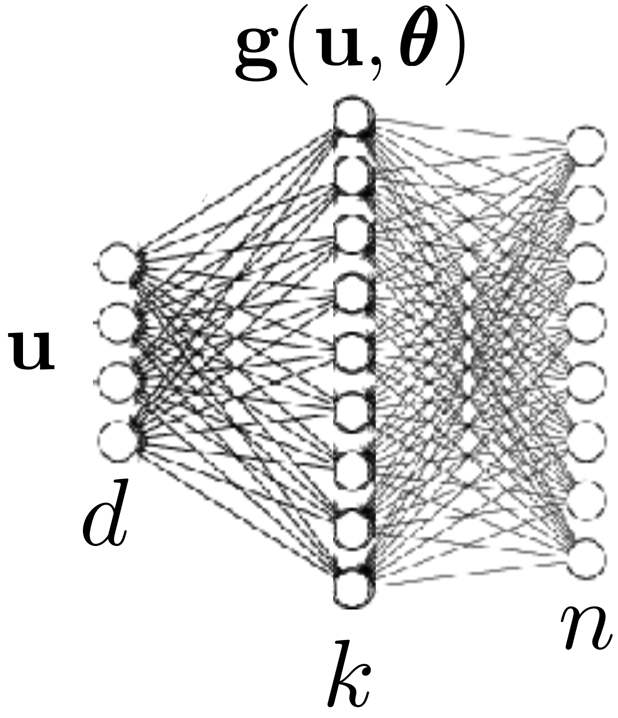
$$\begin{cases} \dot{\boldsymbol{\theta}}(t) = -\nabla_{\mathbf{W}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(t))) \\ \boldsymbol{\theta}(0) = \boldsymbol{\theta}_0. \end{cases}$$

$\eta(s) = s^{p+1}/(2(p+1)),\ p \in [0,1]$

$$\mathcal{L}_{\mathbf{y}} = \eta(\|\cdot - \mathbf{y}\|^2)$$

$$\sigma_i(\mathbf{A}) = \frac{1}{1+i^2}$$

Signal recovery under ill-conditioning



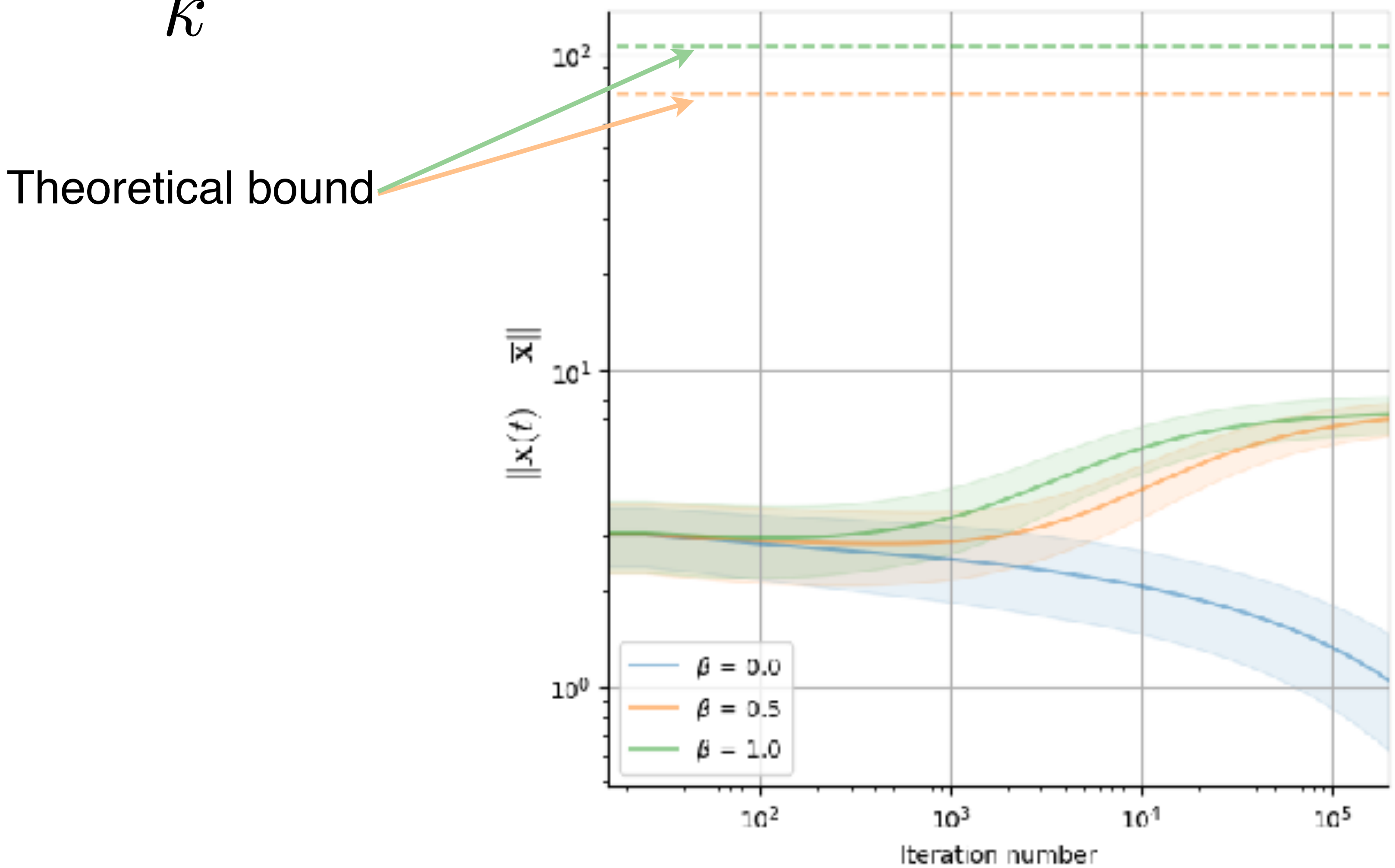
$$\mathbf{g}(\mathbf{u},\boldsymbol{\theta}) = \frac{1}{\sqrt{k}}\mathbf{V}\phi(\mathbf{W}\mathbf{u})$$

$$\begin{cases} \dot{\boldsymbol{\theta}}(t) = -\nabla_{\mathbf{W}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(t))) \\ \boldsymbol{\theta}(0) = \boldsymbol{\theta}_0. \end{cases}$$

$\eta(s) = s^{p+1}/(2(p+1)),\, p \in [0,1]$

$$\mathcal{L}_{\mathbf{y}} = \eta(\|\cdot - \mathbf{y}\|^2)$$

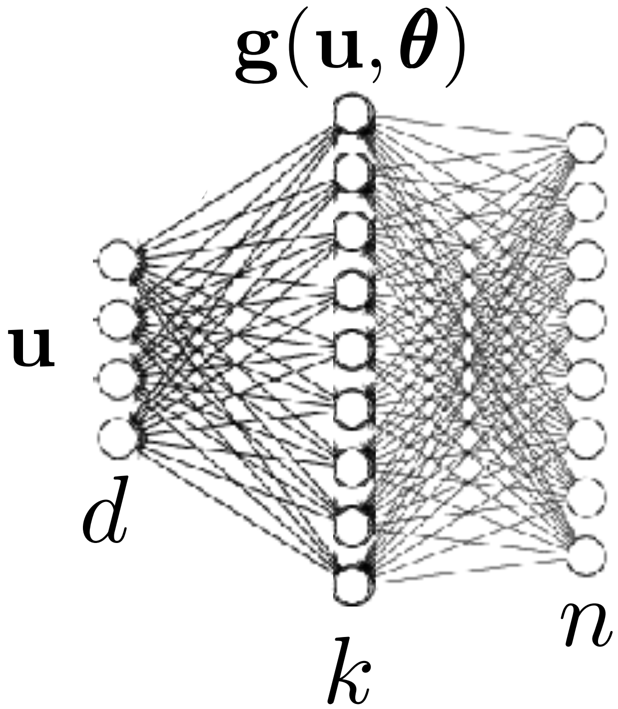
$$\sigma_i(\mathbf{A}) = \frac{1}{1+i^2}$$



$$p = 0.2$$

Convergence to a noise-dominated region for different noise levels.

Impact of the KL property



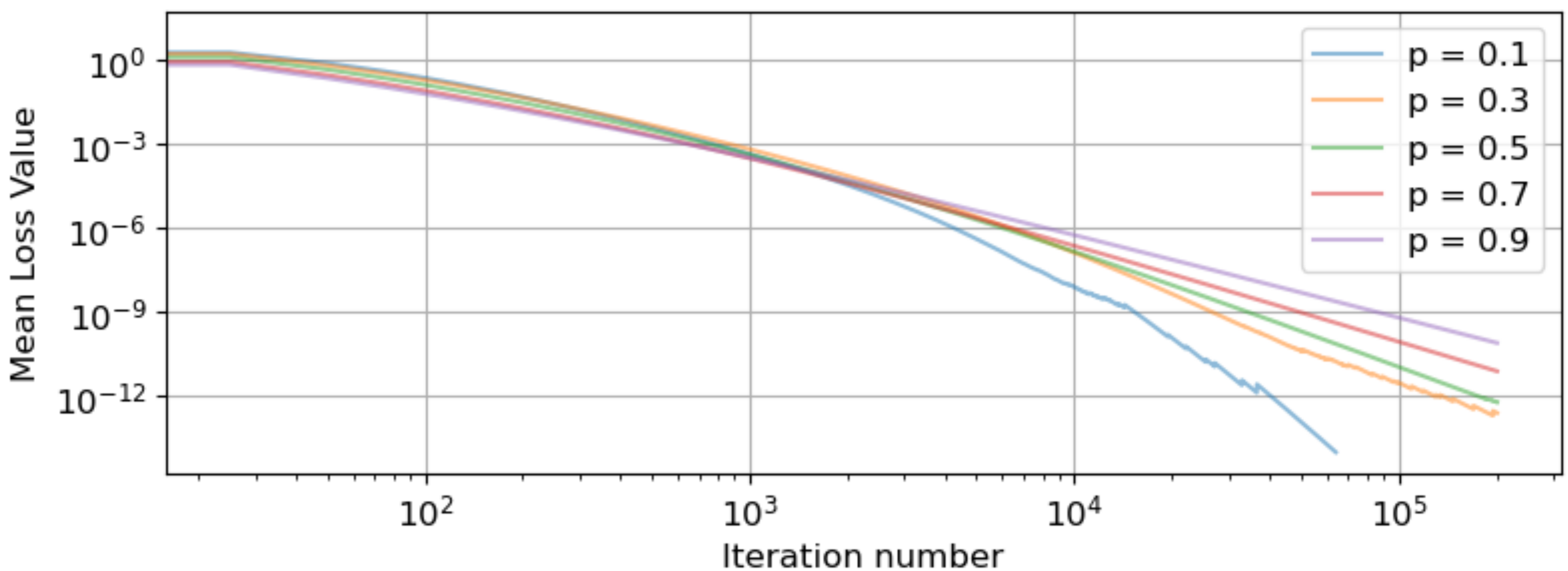
$$\mathbf{g}(\mathbf{u},\boldsymbol{\theta}) = \frac{1}{\sqrt{k}}\mathbf{V}\phi(\mathbf{W}\mathbf{u})$$

$$\begin{cases} \dot{\boldsymbol{\theta}}(t) = -\nabla_{\mathbf{W}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(t))) \\ \boldsymbol{\theta}(0) = \boldsymbol{\theta}_0. \end{cases}$$

$\eta(s) = s^{p+1}/(2(p+1)),\ p \in [0,1]$

$$\mathcal{L}_{\mathbf{y}} = \eta(\|\cdot - \mathbf{y}\|^2)$$

$$\sigma_i(\mathbf{A}) = \frac{1}{1+i^2}$$

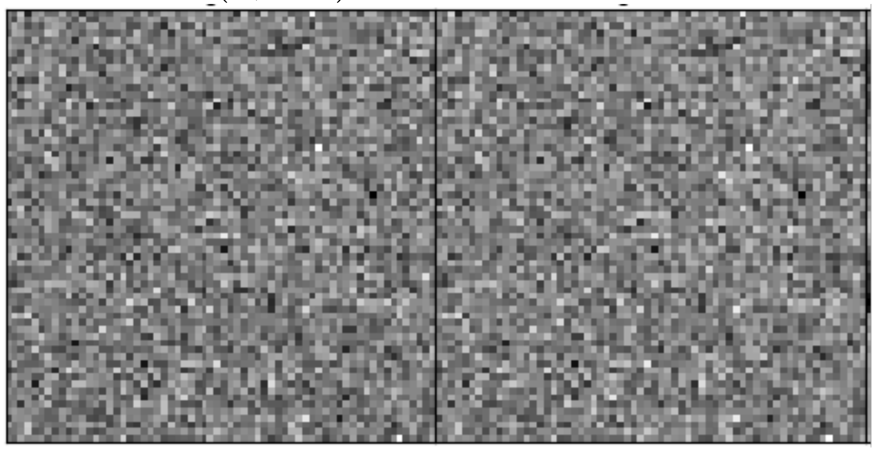


As expected
the smaller p the faster the convergence rate.

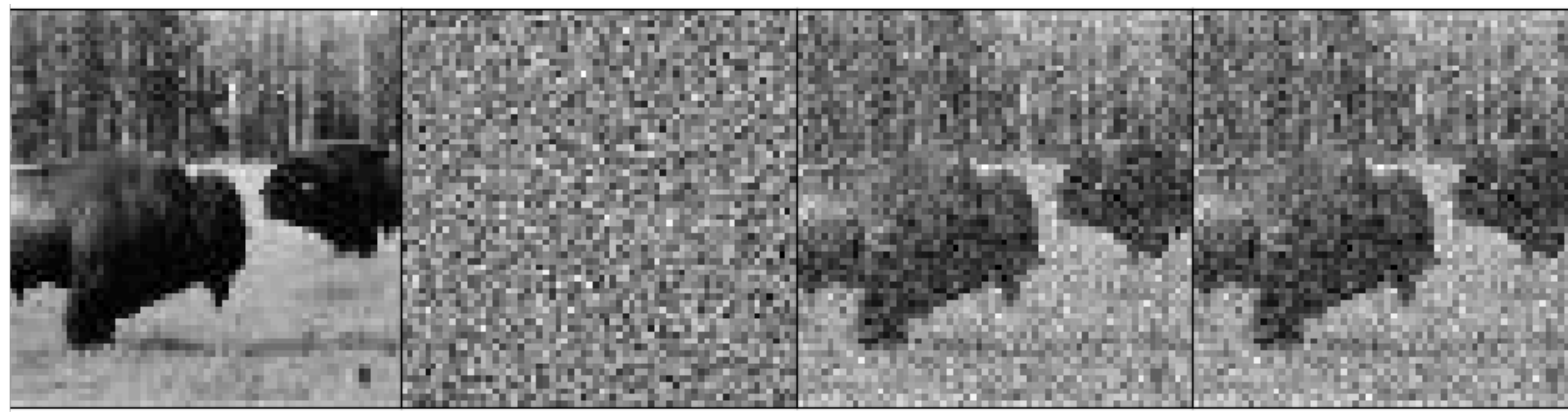
Application to image recovery: deblurring

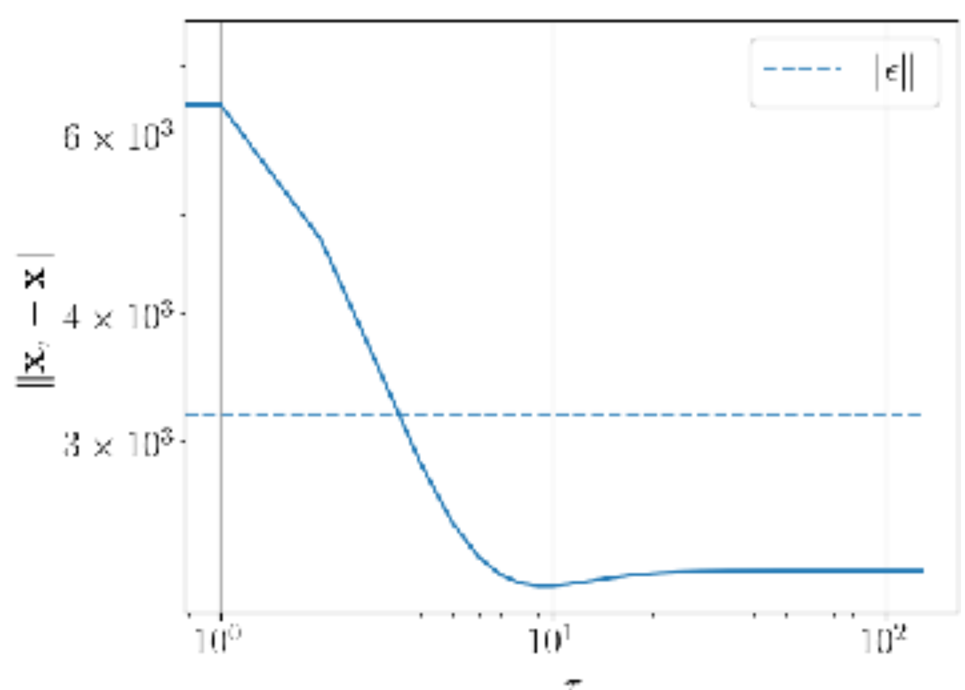
$\mathbf{y} = \mathbf{A}\overline{\mathbf{x}} + \varepsilon$

$\varepsilon \sim \mathcal{N}(0, 50^2)$ $\qquad \mathbf{y}_{1000}$



$\overline{\mathbf{x}}$ $\qquad \mathbf{x}_0$ $\qquad \mathbf{x}_{80}$ $\qquad \mathbf{x}_{1000}$





Take away messages

Take away messages

- **Recovery guarantees** of DIP when optimized with gradient descent in both **observation** and **signal** spaces.

Take away messages

- **Recovery guarantees** of DIP when optimized with gradient descent in both **observation** and **signal** spaces.
- Influence if the **forward operator** and the **loss function** via its **desingularizing** function.

Take away messages

- **Recovery guarantees** of DIP when optimized with gradient descent in both **observation** and **signal** spaces.
- Influence if the **forward operator** and the **loss function** via its **desingularizing** function.
- NN **design**: need for **overparametrization**.

Take away messages

- **Recovery guarantees** of DIP when optimized with gradient descent in both **observation** and **signal** spaces.
- Influence if the **forward operator** and the **loss function** via its **desingularizing** function.
- NN **design**: need for **overparametrization**.
- **Empirical** results agree with **theoretical** predictions.

Take away messages

- **Recovery guarantees** of DIP when optimized with gradient descent in both **observation** and **signal** spaces.
- Influence if the **forward operator** and the **loss function** via its **desingularizing** function.
- NN **design**: need for **overparametrization**.
- **Empirical** results agree with **theoretical** predictions.

- Discrete setting ✓.

Take away messages

- **Recovery guarantees** of DIP when optimized with gradient descent in both **observation** and **signal** spaces.
- Influence if the **forward operator** and the **loss function** via its **desingularizing** function.
- NN **design**: need for **overparametrization**.
- **Empirical** results agree with **theoretical** predictions.

- Discrete setting ✓.
- Stochastic setting.

Take away messages

- **Recovery guarantees** of DIP when optimized with gradient descent in both **observation** and **signal** spaces.
- Influence if the **forward operator** and the **loss function** via its **desingularizing** function.
- NN **design**: need for **overparametrization**.
- **Empirical** results agree with **theoretical** predictions.

- Discrete setting ✓.
- Stochastic setting.
- Non-smooth setting.

Take away messages

- **Recovery guarantees** of DIP when optimized with gradient descent in both **observation** and **signal** spaces.
- Influence if the **forward operator** and the **loss function** via its **desingularizing** function.
- NN **design**: need for **overparametrization**.
- **Empirical** results agree with **theoretical** predictions.

- Discrete setting ✓.
- Stochastic setting.
- Non-smooth setting.
- Other NN-based frameworks: PINNs, supervised setting.

Take away messages

- **Recovery guarantees** of DIP when optimized with gradient descent in both **observation** and **signal** spaces.
- Influence if the **forward operator** and the **loss function** via its **desingularizing** function.
- NN **design**: need for **overparametrization**.
- **Empirical** results agree with **theoretical** predictions.

- Discrete setting ✓.
- Stochastic setting.
- Non-smooth setting.
- Other NN-based frameworks: PINNs, supervised setting.
- Other overparametrization regimes.

Preprint on arxiv and paper on

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

Thanks
Any questions ?