A machine learning approach to optimal regularization: Affine Manifolds

Valeriya Naumova

(joint work with Ernesto De Vito, Massimo Fornasier, Zeljko Kereta)

Simula Research Laboratory AS

Department of Mathematical Sciences,
NTNU,

18 September 2018

Motivation, Goal, and Tools

- **Motivation:** Data from many real-life acquisitions (signals, images, etc.) are affected by noise of various distribution and intensity.
- **Methods:** Regularization-based approaches balance the discrepancy between data and complexity of the solution, measured by some norm (total variation, ℓ_1-norm etc.), depending on so-called regularization parameter(s).
- **Problem:** To find appropriate regularization parameter in a fast way is difficult.
- **Goal:** To learn the nonlinear function defined in high dimension that describes the relation between special features of the data and the optimal regularization parameter from a number of given examples.
- **Tools:** regularization methods, statistical learning theory, data representation, probability theory, sparsity.

Motivation, Goal, and Tools

- **Motivation:** Data from many real-life acquisitions (signals, images, etc.) are affected by noise of various distribution and intensity.
- **Methods:** Regularization-based approaches balance the discrepancy between data and complexity of the solution, measured by some norm (total variation, ℓ_1-norm etc.), depending on so-called regularization parameter(s).
- **Problem:** To find appropriate regularization parameter in a fast way is difficult.
- **Goal:** To learn the nonlinear function defined in high dimension that describes the relation between special features of the data and the optimal regularization parameter from a number of given examples.
- **Tools:** regularization methods, statistical learning theory, data representation, probability theory, sparsity.

Motivation, Goal, and Tools

- **Motivation:** Data from many real-life acquisitions (signals, images, etc.) are affected by noise of various distribution and intensity.
- **Methods:** Regularization-based approaches balance the discrepancy between data and complexity of the solution, measured by some norm (total variation, ℓ_1-norm etc.), depending on so-called regularization parameter(s).
- **Problem:** To find appropriate regularization parameter in a fast way is difficult.
- **Goal:** To learn the nonlinear function defined in high dimension that describes the relation between special features of the data and the optimal regularization parameter from a number of given examples.
- **Tools:** regularization methods, statistical learning theory, data representation, probability theory, sparsity.

Motivation, Goal, and Tools

- **Motivation:** Data from many real-life acquisitions (signals, images, etc.) are affected by noise of various distribution and intensity.
- **Methods:** Regularization-based approaches balance the discrepancy between data and complexity of the solution, measured by some norm (total variation, ℓ_1-norm etc.), depending on so-called regularization parameter(s).
- **Problem:** To find appropriate regularization parameter in a fast way is difficult.
- **Goal:** To learn the nonlinear function defined in high dimension that describes the relation between special features of the data and the optimal regularization parameter from a number of given examples.
- **Tools:** regularization methods, statistical learning theory, data representation, probability theory, sparsity.

Introduction

Problem Statement

Consider a linear inverse problem $Y = AX + \sigma W$, where

- $X \in \mathbb{R}^d$ is the quantity of interest (e.g., ground truth image),
- $A \in \mathbb{R}^{m \times d}$ is a measurement operator (e.g., convolution, mask),
- $W \in \mathbb{R}^m$ is a random variable / noise,
- $Y \in \mathbb{R}^m$ is the observed quantity (e.g., noisy image).

We consider regularization approaches, where the **regularized solution** is given as the result of minimizing functionals of the type

$$Z^\alpha = \operatorname*{argmin}_{z \in \mathbb{R}^d} \|Az - Y\|^2 + \alpha J(z).$$

Introduction

Problem Statement

Consider a linear inverse problem $Y = AX + \sigma W$, where

- $X \in \mathbb{R}^d$ is the quantity of interest (e.g., ground truth image),
- $A \in \mathbb{R}^{m \times d}$ is a measurement operator (e.g., convolution, mask),
- $W \in \mathbb{R}^m$ is a random variable / noise,
- $Y \in \mathbb{R}^m$ is the observed quantity (e.g., noisy image).

We consider regularization approaches, where the **regularized solution** is given as the result of minimizing functionals of the type

$$Z^\alpha = \operatorname*{argmin}_{z \in \mathbb{R}^d} \|Az - Y\|^2 + \alpha J(z).$$

Introduction

Problem Statement

Consider a linear inverse problem $Y = AX + \sigma W$, where

- $X \in \mathbb{R}^d$ is the quantity of interest (e.g., ground truth image),
- $A \in \mathbb{R}^{m \times d}$ is a measurement operator (e.g., convolution, mask),
- $W \in \mathbb{R}^m$ is a random variable / noise,
- $Y \in \mathbb{R}^m$ is the observed quantity (e.g., noisy image).

We consider regularization approaches, where the **regularized solution** is given as the result of minimizing functionals of the type

$$Z^\alpha = \underset{z \in \mathbb{R}^d}{\operatorname{argmin}} \|Az - Y\|^2 + \alpha J(z).$$

- Tikhonov regularization: $J(z) = \|z\|_2^2$,
- Elastic-net regularization: $J(z) = \|z\|_1 + \epsilon\|z\|_2^2$,
- ℓ_1-regularization: $J(z) = \|z\|_1$,
- TV- regularization: $J(z) = \int_\Omega |\nabla z|$,
- ...

Introduction

What about α choice?

- The optimal regularization parameter is given as

$$\begin{cases} \alpha^* = \text{argmin}_{\alpha \in (0,+\infty)} \|Z^\alpha - X\|^2 \\ \text{s.t. } Z^\alpha = \text{argmin}_{z \in \mathbb{R}^d} \|Az - Y\|^2 + \alpha J(z) \end{cases}$$

However, both X and σ are unknown.

- Techniques for regularization parameter choice:
  - *A priori* choice rules based on the noise level and some knowledge about the solution.
  - *A posteriori* choice rules based on the datum Y and the noise level: **Examples: discrepancy principle, L-curve, balancing principle, MSE-based methods, etc.**
  - *Heuristic* choice rules based on the datum Y: **Examples: quasi-balancing principle, quasi-optimality criterion, generalized cross validation, etc.**
- (Unsupervised) data-driven method for parameter selection.

Introduction

What about α choice?

- The optimal regularization parameter is given as

$$\begin{cases} \alpha^* = \text{argmin}_{\alpha \in (0,+\infty)} \|Z^\alpha - X\|^2 \\ \text{s.t. } Z^\alpha = \text{argmin}_{z \in \mathbb{R}^d} \|Az - Y\|^2 + \alpha J(z) \end{cases}$$

However, both X and σ are unknown.

- Techniques for regularization parameter choice:
  - *A priori* choice rules based on the noise level and some knowledge about the solution.
  - *A posteriori* choice rules based on the datum Y and the noise level: **Examples: discrepancy principle, L-curve, balancing principle, MSE-based methods, etc.**
  - *Heuristic* choice rules based on the datum Y: **Examples: quasi-balancing principle, quasi-optimality criterion, generalized cross validation, etc.**
- (Unsupervised) data-driven method for parameter selection.

Introduction

What about α choice?

- The optimal regularization parameter is given as

$$\begin{cases} \alpha^* = \text{argmin}_{\alpha \in (0,+\infty)} \|Z^\alpha - X\|^2 \\ \text{s.t. } Z^\alpha = \text{argmin}_{z \in \mathbb{R}^d} \|Az - Y\|^2 + \alpha J(z) \end{cases}$$

  However, both X and σ are unknown.

- Techniques for regularization parameter choice:
  - *A priori* choice rules based on the noise level and some knowledge about the solution.
  - *A posteriori* choice rules based on the datum Y and the noise level:
    Examples: discrepancy principle, L-curve, balancing principle, MSE-based methods, etc.
  - *Heuristic* choice rules based on the datum Y:
    Examples: quasi-balancing principle, quasi-optimality criterion, generalized cross validation, etc.
- (Unsupervised) data-driven method for parameter selection.

Introduction

What about α choice?

- The optimal regularization parameter is given as

$$\begin{cases} \alpha^* = \text{argmin}_{\alpha \in (0,+\infty)} \|Z^\alpha - X\|^2 \\ \text{s.t. } Z^\alpha = \text{argmin}_{z \in \mathbb{R}^d} \|Az - Y\|^2 + \alpha J(z) \end{cases}$$

  However, both X and σ are unknown.

- Techniques for regularization parameter choice:
  - *A priori* choice rules based on the noise level and some knowledge about the solution.
  - *A posteriori* choice rules based on the datum Y and the noise level:
    Examples: discrepancy principle, L-curve, balancing principle, MSE-based methods, etc.
  - *Heuristic* choice rules based on the datum Y:
    Examples: quasi-balancing principle, quasi-optimality criterion, generalized cross validation, etc.
- (Unsupervised) data-driven method for parameter selection.

Introduction

Parameter learning under supervised machine learning setting

- Assume we are provided with $\{(X_i, Y_i)\}_{i=1}^n$.
- We can compute the optimal parameters

$$\begin{aligned} (X_1, Y_1) &\rightarrow \alpha_1^* = \underset{\alpha \in (0,+\infty)}{\operatorname{argmin}} \|Z^\alpha(Y_1) - X_1\| \\ &\dots \qquad \dots \\ (X_n, Y_n) &\rightarrow \alpha_n^* = \underset{\alpha \in (0,+\infty)}{\operatorname{argmin}} \|Z^\alpha(Y_n) - X_n\| \end{aligned}$$

- We want to compute α for previously unseen data: $(??, Y) \rightarrow \bar{\alpha}$
  $\Longrightarrow$ We want to find the regression function

$$\mathcal{R} : Y \mapsto \bar{\alpha} := \mathcal{R}(Y) = \int_0^\infty \alpha d\mu(\alpha \mid Y),$$

μ is the (unknown) joint distribution of $(Y_1, \alpha_1^*), \dots, (Y_n, \alpha_n^*)$,
$\mu(\cdot \mid Y)$ is its conditional distribution, very much concentrated.

Introduction

Parameter learning under supervised machine learning setting

- Assume we are provided with $\{(X_i, Y_i)\}_{i=1}^n$.
- We can compute the optimal parameters

$$\begin{aligned} (X_1, Y_1) &\rightarrow \alpha_1^* = \underset{\alpha \in (0,+\infty)}{\operatorname{argmin}} \|Z^\alpha(Y_1) - X_1\| \\ &\dots \qquad \dots \\ (X_n, Y_n) &\rightarrow \alpha_n^* = \underset{\alpha \in (0,+\infty)}{\operatorname{argmin}} \|Z^\alpha(Y_n) - X_n\| \end{aligned}$$

- We want to compute α for previously unseen data: $(??, Y) \rightarrow \bar{\alpha}$

$\Longrightarrow$ We want to find the regression function

$$\mathcal{R} : Y \mapsto \bar{\alpha} := \mathcal{R}(Y) = \int_0^\infty \alpha d\mu(\alpha \mid Y),$$

μ is the (unknown) joint distribution of $(Y_1, \alpha_1^*), \dots, (Y_n, \alpha_n^*)$,
$\mu(\cdot \mid Y)$ is its conditional distribution, very much concentrated.

Introduction

Parameter learning under supervised machine learning setting

- Assume we are provided with $\{(X_i, Y_i)\}_{i=1}^n$.
- We can compute the optimal parameters

$$\begin{aligned} (X_1, Y_1) &\rightarrow \alpha_1^* = \underset{\alpha \in (0,+\infty)}{\operatorname{argmin}} \|Z^\alpha(Y_1) - X_1\| \\ &\dots \qquad \dots \\ (X_n, Y_n) &\rightarrow \alpha_n^* = \underset{\alpha \in (0,+\infty)}{\operatorname{argmin}} \|Z^\alpha(Y_n) - X_n\| \end{aligned}$$

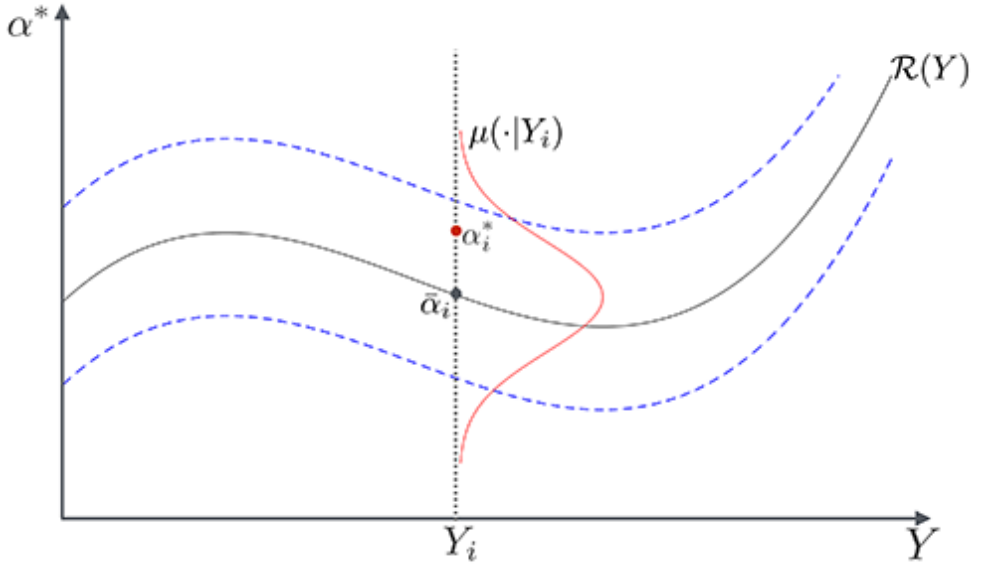
- We want to compute α for previously unseen data: $(??, Y) \rightarrow \bar{\alpha}$
  $\Longrightarrow$ We want to find the regression function

$$\mathcal{R} : Y \mapsto \bar{\alpha} := \mathcal{R}(Y) = \int_0^\infty \alpha d\mu(\alpha \mid Y),$$

μ is the (unknown) joint distribution of $(Y_1, \alpha_1^*), \dots, (Y_n, \alpha_n^*)$,
$\mu(\cdot \mid Y)$ is its conditional distribution, very much concentrated.

Introduction

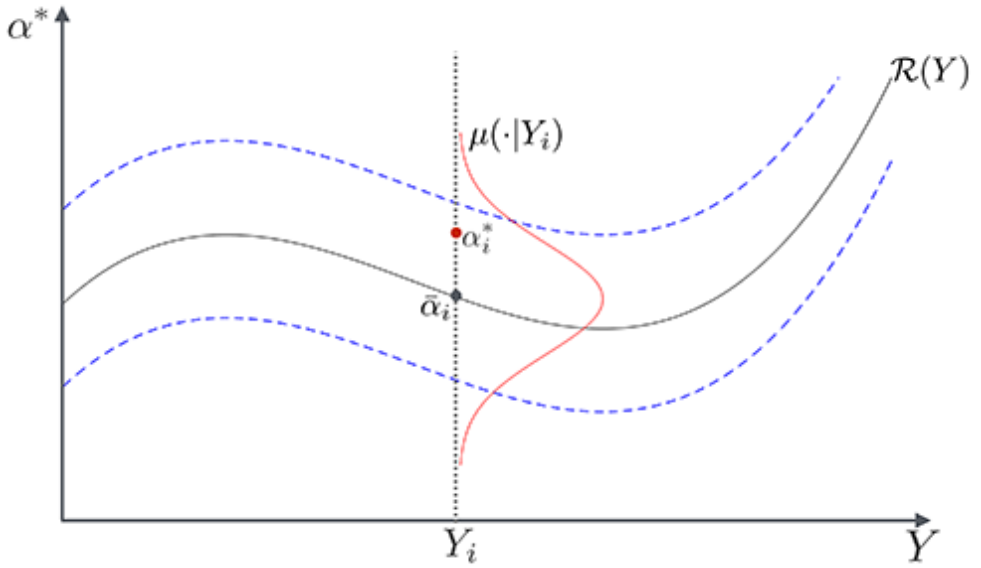
Parameter learning under supervised machine learning setting



- We want to find an approximation $\hat{\mathcal{R}}$ to the regression function $\mathcal{R}$ using only (a small number of) samples n.
- We do not know the conditional distribution $\mu(\cdot \mid Y)$.
- The problem is known to be intractable (Novak & Wozniakowski '09) even for infinitely differentiable functions.
- The number of training points must grow exponentially in m.

Introduction

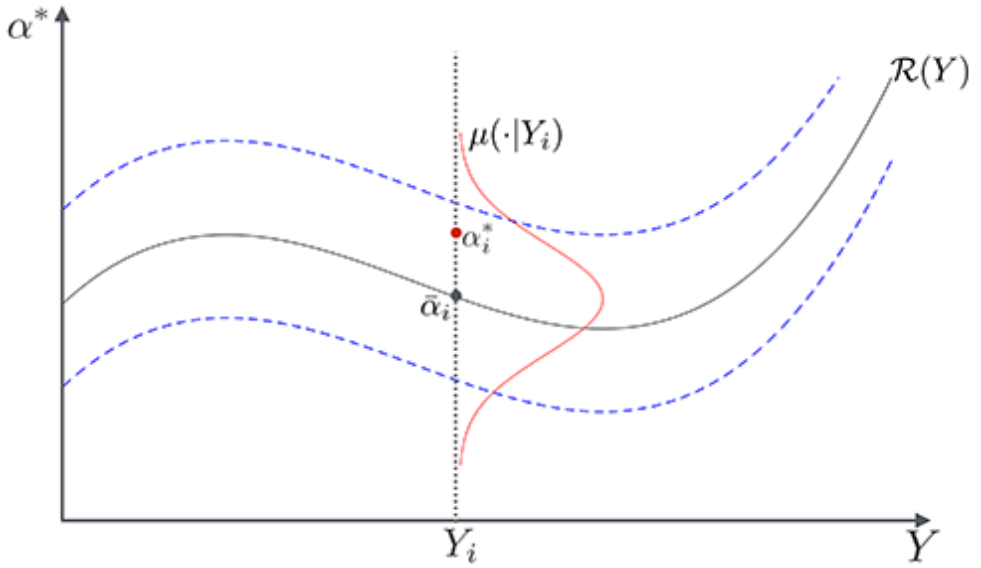
Parameter learning under supervised machine learning setting



- We want to find an approximation $\hat{\mathcal{R}}$ to the regression function $\mathcal{R}$ using only (a small number of) samples n.
- We do not know the conditional distribution $\mu(\cdot \mid Y)$.
- The problem is known to be intractable (Novak & Wozniakowski '09) even for infinitely differentiable functions.
- The number of training points must grow exponentially in m.

Introduction

Parameter learning under supervised machine learning setting



- We want to find an approximation $\hat{\mathcal{R}}$ to the regression function $\mathcal{R}$ using only (a small number of) samples n.
- We do not know the conditional distribution $\mu(\cdot \mid Y)$.
- The problem is known to be intractable (Novak & Wozniakowski '09) even for infinitely differentiable functions.
- The number of training points must grow exponentially in m.

Introduction

Parameter learning under unsupervised machine learning setting

Conclusion: High smoothness does not help!

$\Longrightarrow$ Way out: Solutions concentrate around lower dimensional sets (manifolds), $h \ll d$.

Method idea:

- Using noisy samples $\{Y_i\}_{i=1}^n$, construct an approximation $\widehat{X}$ to X;

  $\Longrightarrow$ does not depend on the regularisation method.

  $\Longrightarrow$ require theoretical analysis for different model types.

- Find the optimal parameter $\hat{\alpha}$ as

$$\hat{\alpha} = \arg\min \widehat{\mathcal{R}}(Y) \text{ and } \widehat{\mathcal{R}}(Y) = \|Z^{\alpha}(Y) - \widehat{X}\|^2.$$

  $\Longrightarrow$ depends on the regularisation method.

  $\Longrightarrow$ require development of efficient numerical methods.

Introduction

Parameter learning under unsupervised machine learning setting

Conclusion: High smoothness does not help!

$\Longrightarrow$ Way out: Solutions concentrate around lower dimensional sets (manifolds), $h \ll d$.

Method idea:

- Using noisy samples $\{Y_i\}_{i=1}^n$, construct an approximation $\widehat{X}$ to X;

  $\Longrightarrow$ does not depend on the regularisation method.

  $\Longrightarrow$ require theoretical analysis for different model types.

- Find the optimal parameter $\hat{\alpha}$ as

$$\hat{\alpha} = \arg\min \widehat{\mathcal{R}}(Y) \text{ and } \widehat{\mathcal{R}}(Y) = \|Z^{\alpha}(Y) - \widehat{X}\|^2.$$

  $\Longrightarrow$ depends on the regularisation method.

  $\Longrightarrow$ require development of efficient numerical methods.

Introduction

Parameter learning under unsupervised machine learning setting

Conclusion: High smoothness does not help!

$\Longrightarrow$ Way out: Solutions concentrate around lower dimensional sets (manifolds), $h \ll d$.

Method idea:

- Using noisy samples $\{Y_i\}_{i=1}^n$, construct an approximation $\widehat{X}$ to X;

  $\Longrightarrow$ does not depend on the regularisation method.

  $\Longrightarrow$ require theoretical analysis for different model types.

- Find the optimal parameter $\hat{\alpha}$ as

$$\hat{\alpha} = \arg\min \widehat{\mathcal{R}}(Y) \text{ and } \widehat{\mathcal{R}}(Y) = \|Z^{\alpha}(Y) - \widehat{X}\|^2.$$

  $\Longrightarrow$ depends on the regularisation method.

  $\Longrightarrow$ require development of efficient numerical methods.

Introduction

Parameter learning under unsupervised machine learning setting

Conclusion: High smoothness does not help!

$\Longrightarrow$ Way out: Solutions concentrate around lower dimensional sets (manifolds), $h \ll d$.

Method idea:

- Using noisy samples $\{Y_i\}_{i=1}^n$, construct an approximation $\widehat{X}$ to X;

  $\Longrightarrow$ does not depend on the regularisation method.

  $\Longrightarrow$ require theoretical analysis for different model types.

- Find the optimal parameter $\hat{\alpha}$ as

$$\hat{\alpha} = \arg\min \widehat{\mathcal{R}}(Y) \text{ and } \widehat{\mathcal{R}}(Y) = \|Z^{\alpha}(Y) - \widehat{X}\|^2.$$

  $\Longrightarrow$ depends on the regularisation method.

  $\Longrightarrow$ require development of efficient numerical methods.

Introduction

Parameter learning under unsupervised machine learning setting

Conclusion: High smoothness does not help!

$\Longrightarrow$ Way out: Solutions concentrate around lower dimensional sets (manifolds), $h \ll d$.

Method idea:

- Using noisy samples $\{Y_i\}_{i=1}^n$, construct an approximation $\widehat{X}$ to X;

  $\Longrightarrow$ does not depend on the regularisation method.

  $\Longrightarrow$ require theoretical analysis for different model types.
- Find the optimal parameter $\hat{\alpha}$ as

$$\hat{\alpha} = \arg\min \widehat{\mathcal{R}}(Y) \text{ and } \widehat{\mathcal{R}}(Y) = \|Z^{\alpha}(Y) - \widehat{X}\|^2.$$

  $\Longrightarrow$ depends on the regularisation method.

  $\Longrightarrow$ require development of efficient numerical methods.

Introduction

Parameter learning under unsupervised machine learning setting

Conclusion: High smoothness does not help!

$\Longrightarrow$ Way out: Solutions concentrate around lower dimensional sets (manifolds), $h \ll d$.

Method idea:

- Using noisy samples $\{Y_i\}_{i=1}^n$, construct an approximation $\widehat{X}$ to X;

  $\Longrightarrow$ does not depend on the regularisation method.

  $\Longrightarrow$ require theoretical analysis for different model types.

- Find the optimal parameter $\hat{\alpha}$ as

$$\hat{\alpha} = \arg\min \widehat{\mathcal{R}}(Y) \text{ and } \widehat{\mathcal{R}}(Y) = \|Z^{\alpha}(Y) - \widehat{X}\|^2.$$

$\Longrightarrow$ depends on the regularisation method.

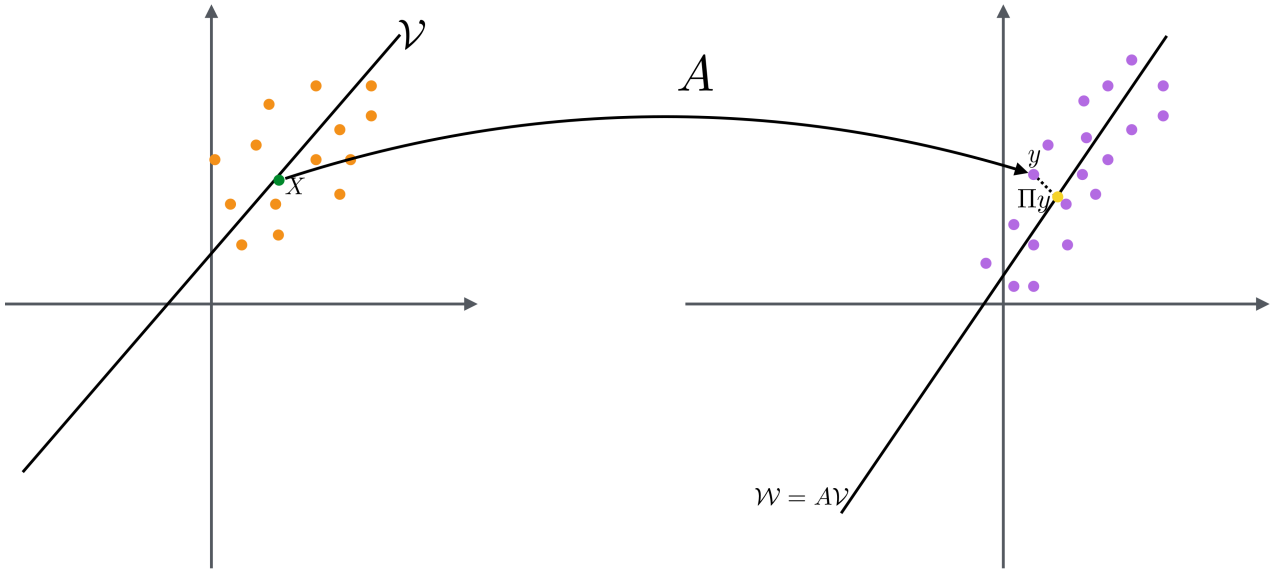
$\Longrightarrow$ require development of efficient numerical methods.

Problem setting

Empirical estimators

Consider $Y = AX + \sigma W$ such that

- $X \in \mathbb{R}^d$ has a sub-Gaussian distribution over a linear subspace $\mathcal{V}$,
- $\dim \mathcal{V} = \text{range}\, \Sigma(X) = h \ll d$,
- W is an independent sub-Gaussian vector with $\Sigma(W) = \mathbb{I}$.



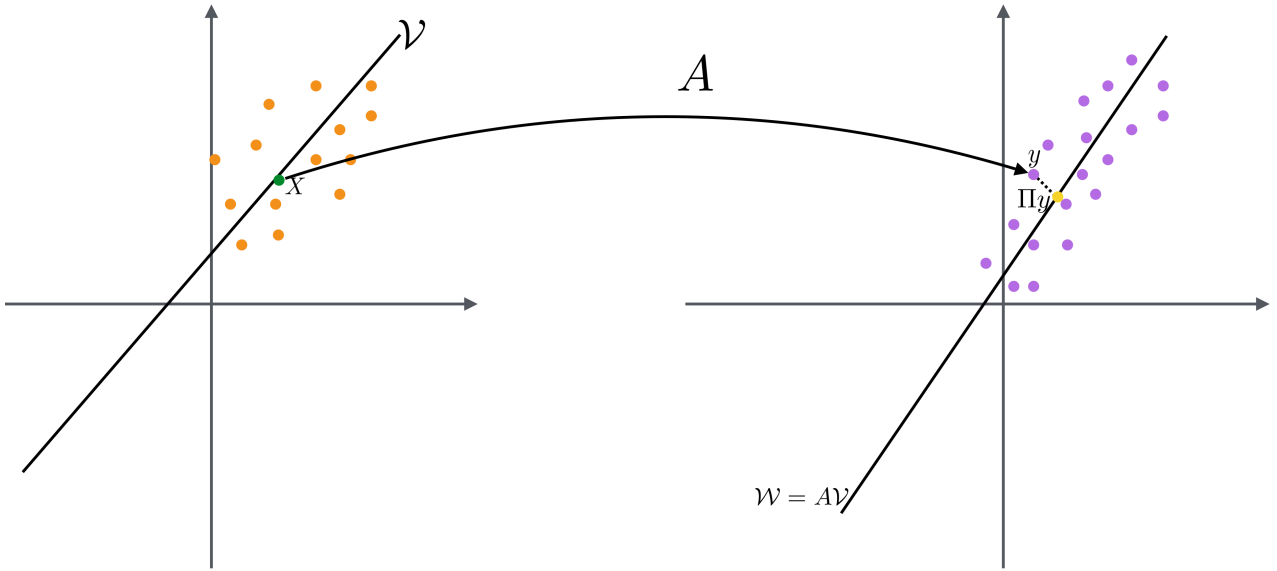
- We define the projection Π onto $\mathcal{W}$, where $\mathcal{W} = A\mathcal{V}$.
- $\dim \mathcal{W} = \dim \mathcal{V} = h$.

Problem setting

Empirical estimators

Consider $Y = AX + \sigma W$ such that

- $X \in \mathbb{R}^d$ has a sub-Gaussian distribution over a linear subspace $\mathcal{V}$,
- $\dim \mathcal{V} = \text{range}\, \Sigma(X) = h \ll d$,
- W is an independent sub-Gaussian vector with $\Sigma(W) = \mathbb{I}$.

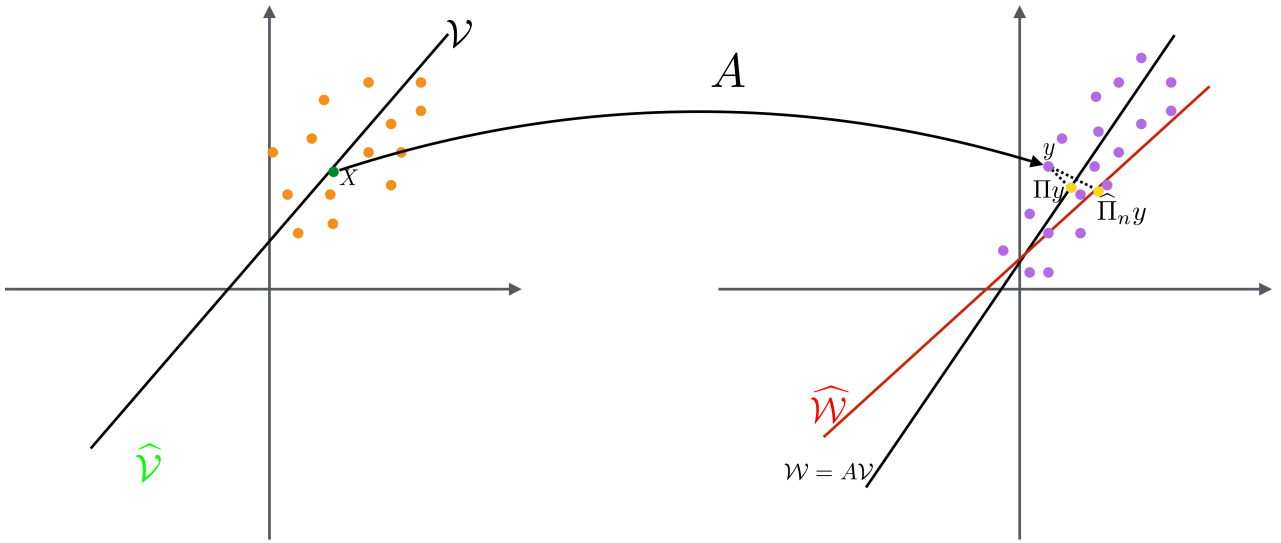


- We define the projection Π onto $\mathcal{W}$, where $\mathcal{W} = A\mathcal{V}$.
- $\dim \mathcal{W} = \dim \mathcal{V} = h$.

Empirical estimators

Consider $Y = AX + \sigma W$ such that

- $X \in \mathbb{R}^d$ has a sub-Gaussian distribution over a linear subspace $\mathcal{V}$,
- $\dim \mathcal{V} = \text{range}\, \Sigma(X) = h \ll d$,
- W is an independent sub-Gaussian vector with $\Sigma(W) = \mathbb{I}$.



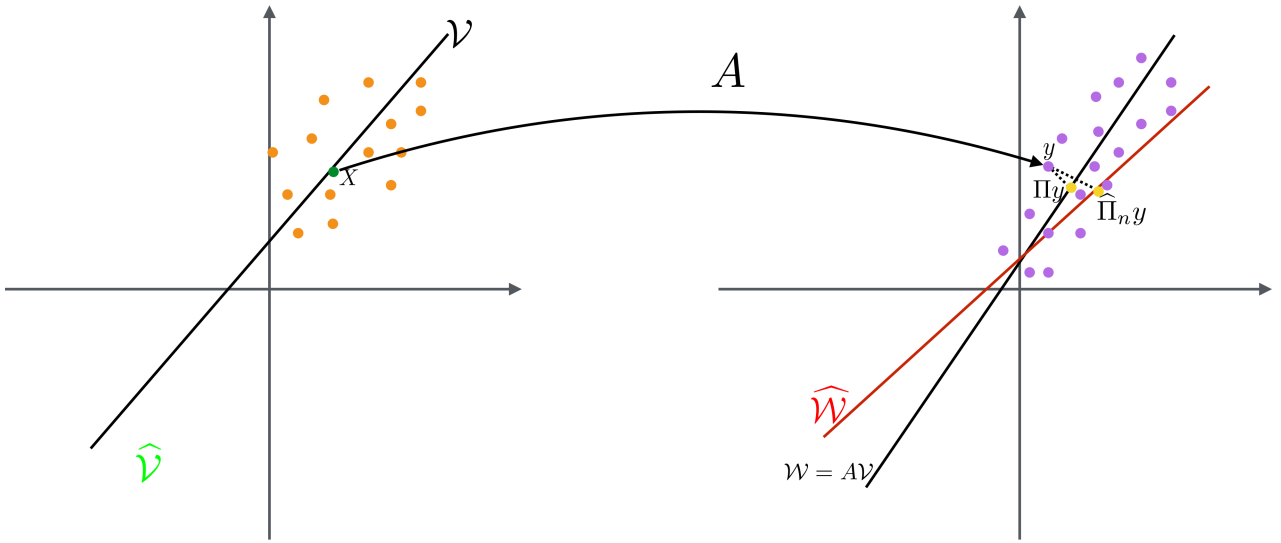
- We define the empirical projection $\widehat{\Pi}_n$ onto the space spanned by the first h eigenvectors of $\widehat{\Sigma}(Y) = \frac{1}{n}\sum_{i=1}^{n} Y_i \otimes Y_i$.

$\Longrightarrow$ with high probability $\widehat{\Pi}_n \sim \Pi$ and is unique for $n = O(m)$ and small σ.

Empirical estimators

Consider $Y = AX + \sigma W$ such that

- $X \in \mathbb{R}^d$ has a sub-Gaussian distribution over a linear subspace $\mathcal{V}$,
- $\dim \mathcal{V} = \text{range}\, \Sigma(X) = h \ll d$,
- W is an independent sub-Gaussian vector with $\Sigma(W) = \mathbb{I}$.



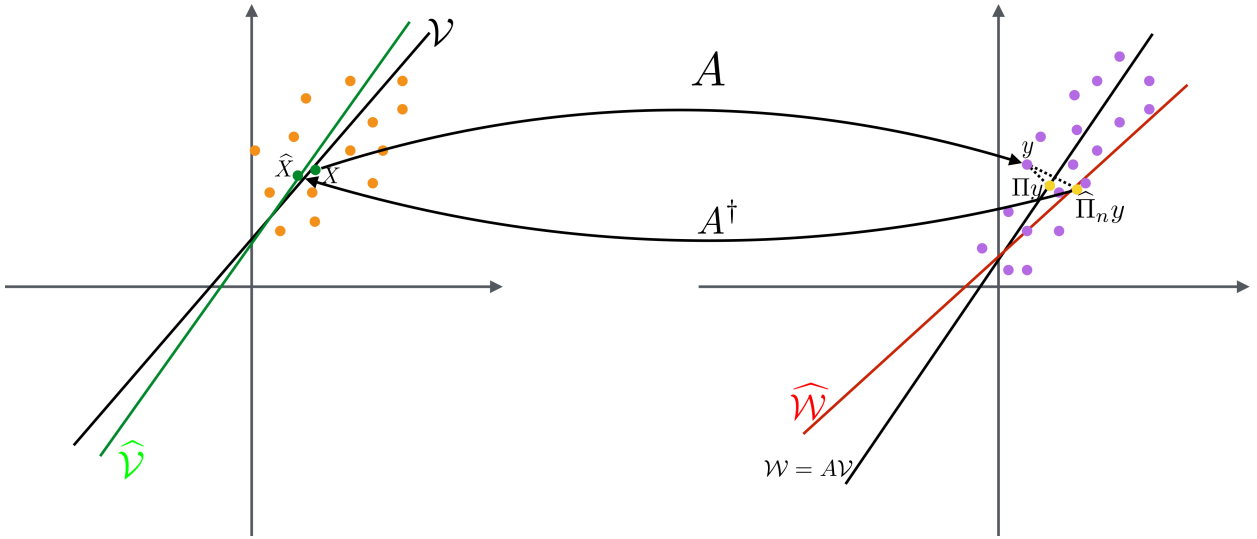
- We define the empirical projection $\widehat{\Pi}_n$ onto the space spanned by the first h eigenvectors of $\widehat{\Sigma}(Y) = \frac{1}{n}\sum_{i=1}^{n} Y_i \otimes Y_i$.

  $\Longrightarrow$ with high probability $\widehat{\Pi}_n \sim \Pi$ and is unique for $n = O(m)$ and small σ.

Empirical estimators

Consider $Y = AX + \sigma W$ such that

- $X \in \mathbb{R}^d$ has a sub-Gaussian distribution over a linear subspace $\mathcal{V}$,
- $\dim \mathcal{V} = \text{range}\, \Sigma(X) = h \ll d$,
- W is an independent sub-Gaussian vector with $\Sigma(W) = \mathbb{I}$.



- We define the empirical estimators of X and W as

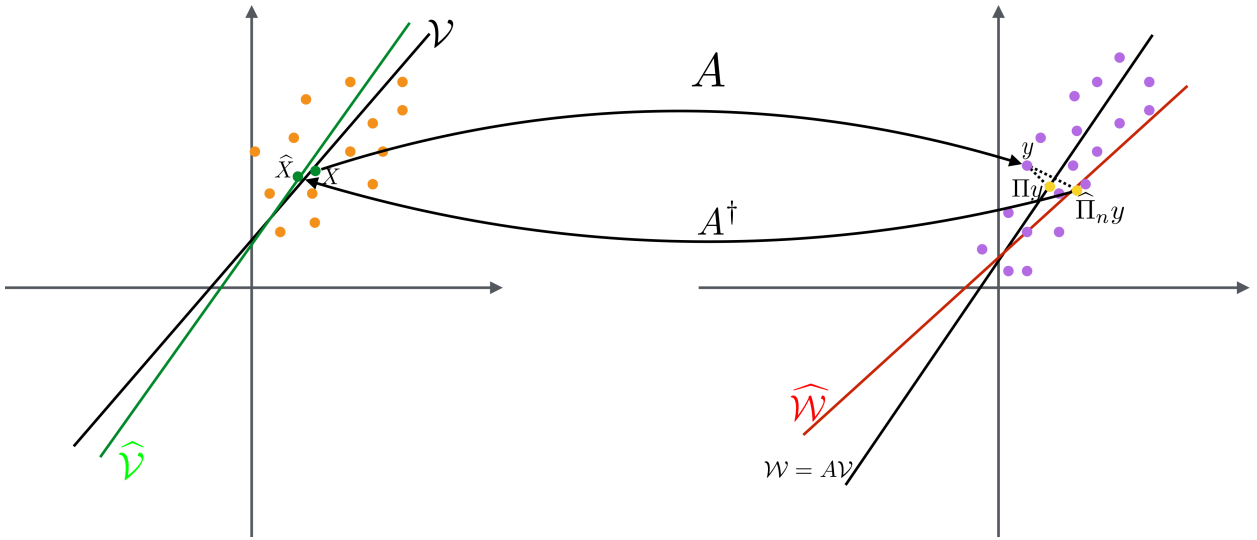
$$\widehat{X} = A^\dagger \widehat{\Pi}_n Y \qquad \text{and} \qquad \widehat{W} = (Y - \widehat{\Pi}_n Y),$$

which satisfies empirical inverse problem $A\widehat{X} + Q\widehat{W} = QY$ and $Q = AA^\dagger$

Empirical estimators

Consider $Y = AX + \sigma W$ such that

- $X \in \mathbb{R}^d$ has a sub-Gaussian distribution over a linear subspace $\mathcal{V}$,
- $\dim \mathcal{V} = \operatorname{range} \Sigma(X) = h \ll d$,
- W is an independent sub-Gaussian vector with $\Sigma(W) = \mathbb{I}$.



- Let $\widehat{Z}^\alpha(Y) = \arg\min \|Az - Qy\| + \alpha J(z)$. Then $\widehat{Z}^\alpha(Y) = Z^\alpha(Y)$.
- We consider

$$\hat{\alpha} = \min \|Z^\alpha - \widehat{X}\|^2.$$

Introduction

Parameter learning under unsupervised machine learning setting

Conclusion: High smoothness does not help!

$\Longrightarrow$ Way out: Solutions concentrate around lower dimensional sets (manifolds), $h \ll d$.

Method idea:

- Using noisy samples $\{Y_i\}_{i=1}^n$, construct an approximation $\widehat{X}$ to X;

  $\Longrightarrow$ does not depend on the regularisation method.

  $\Longrightarrow$ require theoretical analysis for different model types.

- Find the optimal parameter $\hat{\alpha}$ as

$$\hat{\alpha} = \arg\min \widehat{\mathcal{R}}(Y) \text{ and } \widehat{\mathcal{R}}(Y) = \|Z^{\alpha}(Y) - \widehat{X}\|^2.$$

$\Longrightarrow$ depends on the regularisation method.

$\Longrightarrow$ require development of efficient numerical methods.

Parameter learning for different regularisation

$$\begin{cases} \hat{\alpha}^* = \arg\min \|Z^\alpha - \widehat{X}\|^2 \\ \text{s.t. } Z^\alpha = \operatorname{argmin}_{z \in \mathbb{R}^d} \|Az - Y\|^2 + \alpha J(z) \end{cases}$$

- **Tikhonov:** $J(z) = \|z\|_2^2$, (Theoretical results)
- **Elastic-net:** $J(z) = \|z\|_1 + \epsilon\|z\|_2^2$, (Theoretical results)
- **ℓ_1:** $J(z) = \|z\|_1$, (Encouraging numerical results)
- **TV:** $J(z) = \int_\Omega |\nabla z|$. (Encouraging numerical results)

$$Z^t = \operatorname*{argmin}_{z \in \mathbb{R}^d} \left(t \|Az - Y\|^2 + (1-t) J(z) \right), \quad t \in [0, 1]$$

$$\Longleftrightarrow$$

$$Z^\alpha = \operatorname*{argmin}_{z \in \mathbb{R}^d} \|Az - Y\|^2 + \alpha J(z), \quad \alpha = (1-t)/t \in [0, +\infty]$$

Parameter learning for different regularisation

$$\begin{cases} \hat{\alpha}^* = \arg\min \|Z^\alpha - \widehat{X}\|^2 \\ \text{s.t. } Z^\alpha = \operatorname{argmin}_{z\in\mathbb{R}^d} \|Az - Y\|^2 + \alpha J(z) \end{cases}$$

- **Tikhonov:** $J(z) = \|z\|_2^2$, (Theoretical results)
- **Elastic-net:** $J(z) = \|z\|_1 + \epsilon\|z\|_2^2$, (Theoretical results)
- ℓ_1**:** $J(z) = \|z\|_1$, (Encouraging numerical results)
- **TV:** $J(z) = \int_\Omega |\nabla z|$. (Encouraging numerical results)

$$Z^t = \operatorname*{argmin}_{z\in\mathbb{R}^d} \left(t\|Az - Y\|^2 + (1-t)J(z)\right), \quad t \in [0,1]$$

$$\Longleftrightarrow$$

$$Z^\alpha = \operatorname*{argmin}_{z\in\mathbb{R}^d} \|Az - Y\|^2 + \alpha J(z), \quad \alpha = (1-t)/t \in [0,+\infty]$$

Parameter learning for different regularisation

$$\begin{cases} \hat{\alpha}^* = \arg\min \|Z^\alpha - \widehat{X}\|^2 \\ \text{s.t. } Z^\alpha = \operatorname{argmin}_{z\in\mathbb{R}^d} \|Az - Y\|^2 + \alpha J(z) \end{cases}$$

- **Tikhonov:** $J(z) = \|z\|_2^2$, (Theoretical results)
- **Elastic-net:** $J(z) = \|z\|_1 + \epsilon\|z\|_2^2$, (Theoretical results)
- ℓ_1**:** $J(z) = \|z\|_1$, (Encouraging numerical results)
- **TV:** $J(z) = \int_\Omega |\nabla z|$. (Encouraging numerical results)

$$Z^t = \operatorname*{argmin}_{z\in\mathbb{R}^d} \left(t\|Az - Y\|^2 + (1-t)J(z) \right), \quad t \in [0,1]$$

$$\Longleftrightarrow$$

$$Z^\alpha = \operatorname*{argmin}_{z\in\mathbb{R}^d} \|Az - Y\|^2 + \alpha J(z), \quad \alpha = (1-t)/t \in [0,+\infty]$$

Parameter learning for different regularisation

$$\begin{cases} \hat{\alpha}^* = \arg\min \|Z^\alpha - \widehat{X}\|^2 \\ \text{s.t. } Z^\alpha = \operatorname{argmin}_{z\in\mathbb{R}^d} \|Az - Y\|^2 + \alpha J(z) \end{cases}$$

- **Tikhonov:** $J(z) = \|z\|_2^2$, (Theoretical results)
- **Elastic-net:** $J(z) = \|z\|_1 + \epsilon\|z\|_2^2$, (Theoretical results)
- ℓ_1**:** $J(z) = \|z\|_1$, (Encouraging numerical results)
- **TV:** $J(z) = \int_\Omega |\nabla z|$. (Encouraging numerical results)

$$Z^t = \operatorname*{argmin}_{z\in\mathbb{R}^d} \left(t\,\|Az - Y\|^2 + (1-t)J(z)\right), \quad t \in [0,1]$$

$$\Longleftrightarrow$$

$$Z^\alpha = \operatorname*{argmin}_{z\in\mathbb{R}^d} \|Az - Y\|^2 + \alpha J(z), \quad \alpha = (1-t)/t \in [0,+\infty]$$

Optimal Parameter Choice

Minimizers

- Tikhonov minimizer

$$Z^t_{Tik} = \underset{z \in \mathbb{R}^d}{\operatorname{argmin}} \left(t \|Az - Y\|^2 + (1-t)\|z\|^2 \right)$$

$\Longrightarrow$ close-form solution exists

- Elastic-net minimizer

$$Z^t_{EN} = \underset{z \in \mathbb{R}^d}{\operatorname{argmin}} \left(t \|Az - Y\|^2 + (1-t) \left[\|z\|_1 + \epsilon \|z\|^2 \right] \right)$$

$\Longrightarrow$ close form solution exists only when $A^T A = \mathbb{I}$,

$\Longrightarrow$otherwise, solution is given via soft-thresholding

Optimal Parameter Choice

Quadratic loss

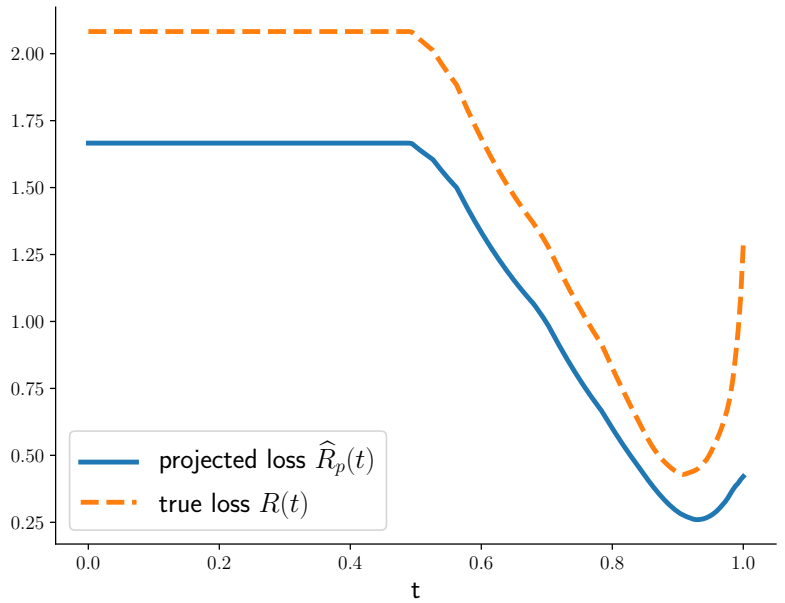
We study the behaviour of the

- **True quadratic loss**

$$R(t) = \|Z^t - X\|^2$$

- **Empirical quadratic loss**

$$\widehat{R}(t) = \|Z^t - \widehat{X}\|^2$$



When A is injective $\hat{t}^* \sim t^*$

Optimal Parameter Choice

Quadratic loss

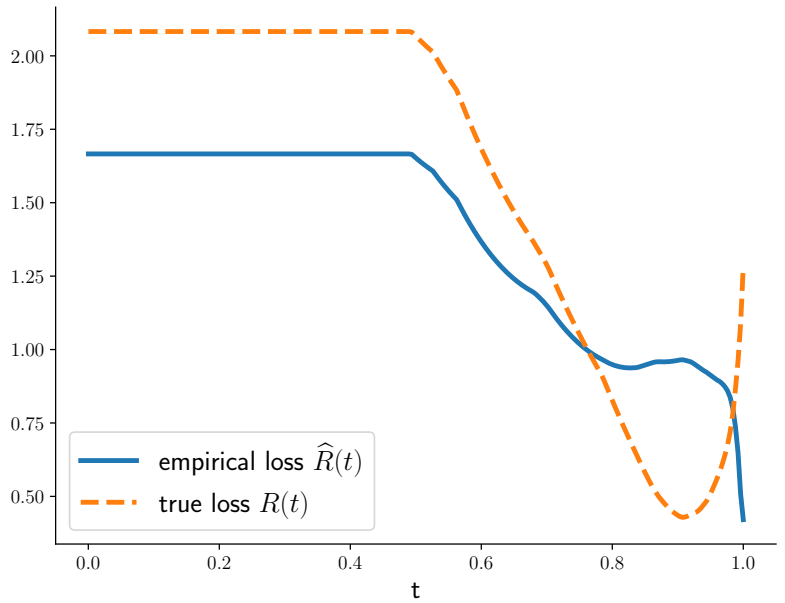
We study the behaviour of the

- **True quadratic loss**

$$R(t) = \|Z^t - X\|^2$$

- **Empirical quadratic loss**

$$\widehat{R}(t) = \|Z^t - \widehat{X}\|^2$$



When A is non injective $\hat{t}^* \nsim t^*$

Optimal Parameter Choice

Quadratic loss

We study the behaviour of the

- **True quadratic loss**

$$R(t) = \|Z^t - X\|^2$$

- **Empirical quadratic loss**

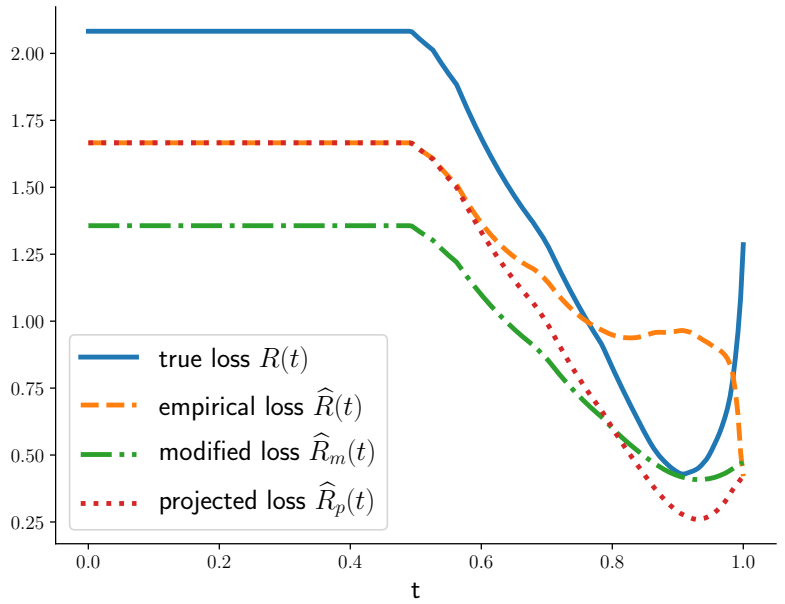
$$\widehat{R}(t) = \|Z^t - \widehat{X}\|^2$$

- **Projected empirical loss**

$$\widehat{R}_p(t) = \|PZ^t - \widehat{X}\|^2, \quad P = A^\dagger A$$

- **Modified projected loss**

$$\widehat{R}_m(t) = \|AZ^t - \widehat{\Pi} y\|^2$$



Optimal Parameter Choice

Tikhonov regularisation with $A = \mathbb{I}$

Theorem (De Vito, Fornasier, Naumova)

For $\tau \geqslant 1$ *with probability greater than* $1 - 6\exp^{-\tau^2}$

$$|\hat{t}^* - t^*| \leqslant \frac{1}{\lambda_h}\left(\sqrt{\frac{d}{n}} + \frac{\tau}{\sqrt{n}} + \sigma^2\right) + \frac{\tau}{d}(\sqrt{h} + \tau)$$

for $n = \mathcal{O}(d)$ *and* $\lambda_h > 0$ *is the smallest non-zero eigenvalue of* $\Sigma(Ax)$.

- Explicit formula for calculating $\hat{t}^*$.

Optimal Parameter Choice

Tikhonov regularisation with $A = \mathbb{I}$

Theorem (De Vito, Fornasier, Naumova)

For $\tau \geqslant 1$ with probability greater than $1 - 6\exp^{-\tau^2}$

$$|\hat{t}^* - t^*| \leqslant \frac{1}{\lambda_h}\left(\sqrt{\frac{d}{n}} + \frac{\tau}{\sqrt{n}} + \sigma^2\right) + \frac{\tau}{d}(\sqrt{h} + \tau)$$

for $n = \mathcal{O}(d)$ and $\lambda_h > 0$ is the smallest non-zero eigenvalue of $\Sigma(Ax)$.

- Explicit formula for calculating $\hat{t}^*$.

Optimal Parameter Choice

Elastic-net with $A = \mathbb{I}$ and Bernoulli noise

Theorem (De Vito, Kereta, Naumova)

For $\tau > 0$ with probability of at least $1 - 2\exp^{-\tau}$

$$|\hat{t}^* - t^*| \leqslant \frac{\lambda_1}{\lambda_h}\left(\sqrt{\frac{h+\tau+\sigma^2 m}{n}} + \frac{h+\tau+\sigma^2 m}{n}\right) + \sigma\sqrt{\frac{h}{m}}$$

for $n = \mathcal{O}(h+m)$ and $\lambda_1 > \lambda_h > 0$ is the largest and the smallest non-zero eigenvalue of $\Sigma(Ax)$.

- Existence and uniqueness results for bounded noise.
- **OptEN Algorithm** for finding $\hat{t}^*$ using a line search method.

Optimal Parameter Choice

Elastic-net with $A = \mathbb{I}$ and Bernoulli noise

Theorem (De Vito, Kereta, Naumova)

For $\tau > 0$ with probability of at least $1 - 2\exp^{-\tau}$

$$|\hat{t}^* - t^*| \leqslant \frac{\lambda_1}{\lambda_h}\left(\sqrt{\frac{h+\tau+\sigma^2 m}{n}} + \frac{h+\tau+\sigma^2 m}{n}\right) + \sigma\sqrt{\frac{h}{m}}$$

for $n = \mathcal{O}(h+m)$ and $\lambda_1 > \lambda_h > 0$ is the largest and the smallest non-zero eigenvalue of $\Sigma(Ax)$.

- Existence and uniqueness results for bounded noise.
- **OptEN Algorithm** for finding $\hat{t}^*$ using a line search method.

Numerical Examples

Parameter learning for Tikhonov regularization

Data:

- $\{(X_i, Y_i)\}_{i=1}^{n}$, $n = 50$.
- $X_i \in \mathbb{R}^{1000}$, $Y_i \in \mathbb{R}^{60}$, and $\sigma = 0.06$.
- $X \in \mathcal{V}$ for $\mathcal{V} = span\{e_1, e_2, \ldots, e_5\}$.

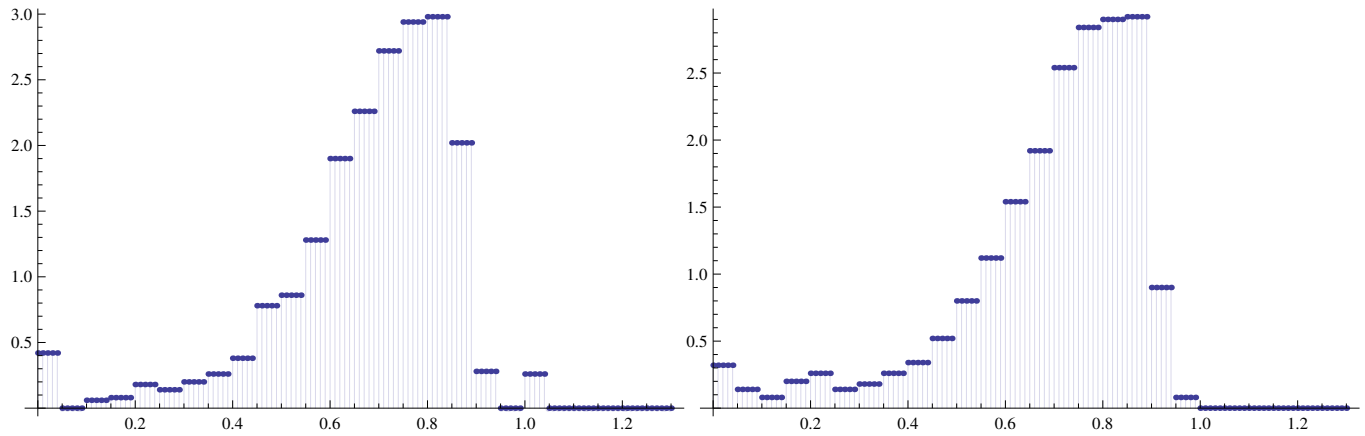


Figure: Empirical distribution of the optimal parameters t^* (left) and the learned parameter $\widehat{t^*}$ (right), $n = 1000$.

Numerical Examples

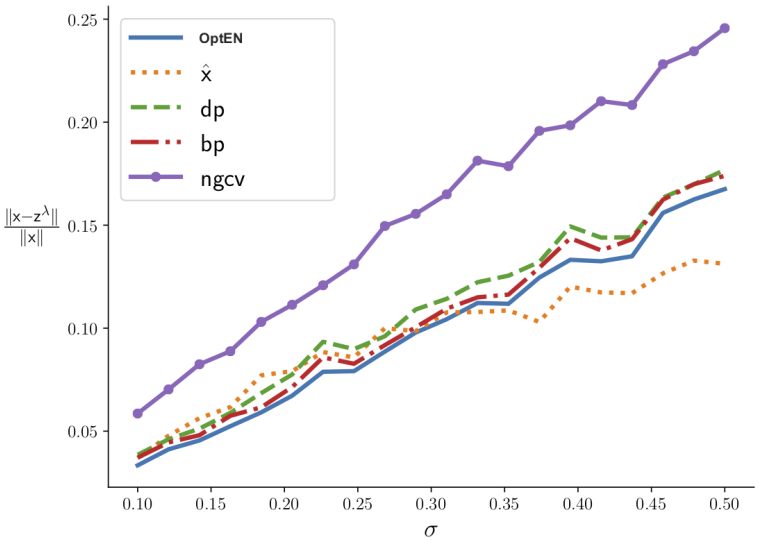
Parameter learning for Elastic-net

Data:

- $\{(X_i, Y_i)\}_{i=1}^n$, $n = 50$.
- $X_i \in \mathbb{R}^{100}$, $Y_i \in \mathbb{R}^{500}$, and $\sigma \in [0.1, 0.5]$.
- $X \in \mathcal{V}$ for $\mathcal{V} = span\{e_1, e_2, \ldots, e_{10}\}$.

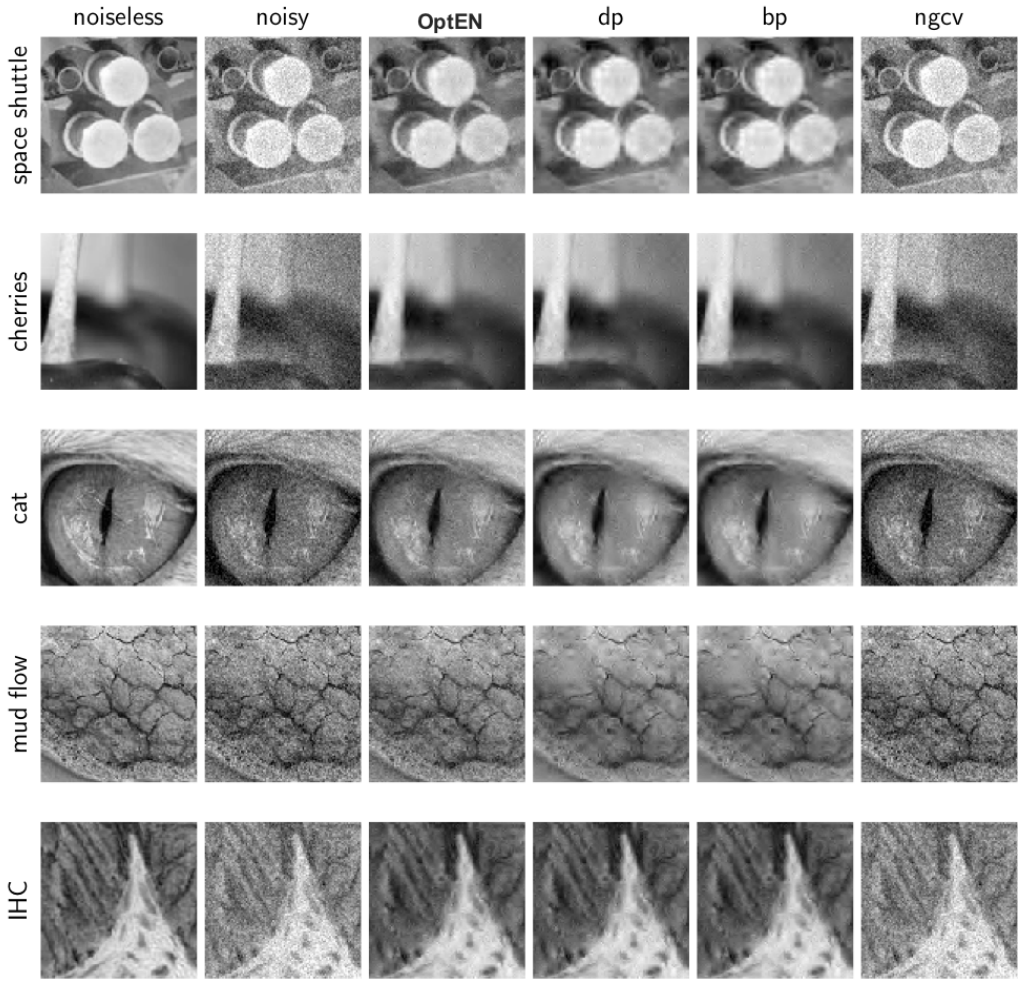
Compare with state-of-the-art parameter choice methods

- discrepancy principle (dp)
- balancing principle (bp)
- non-linear generalised cross-validation (ngcv)
- ....



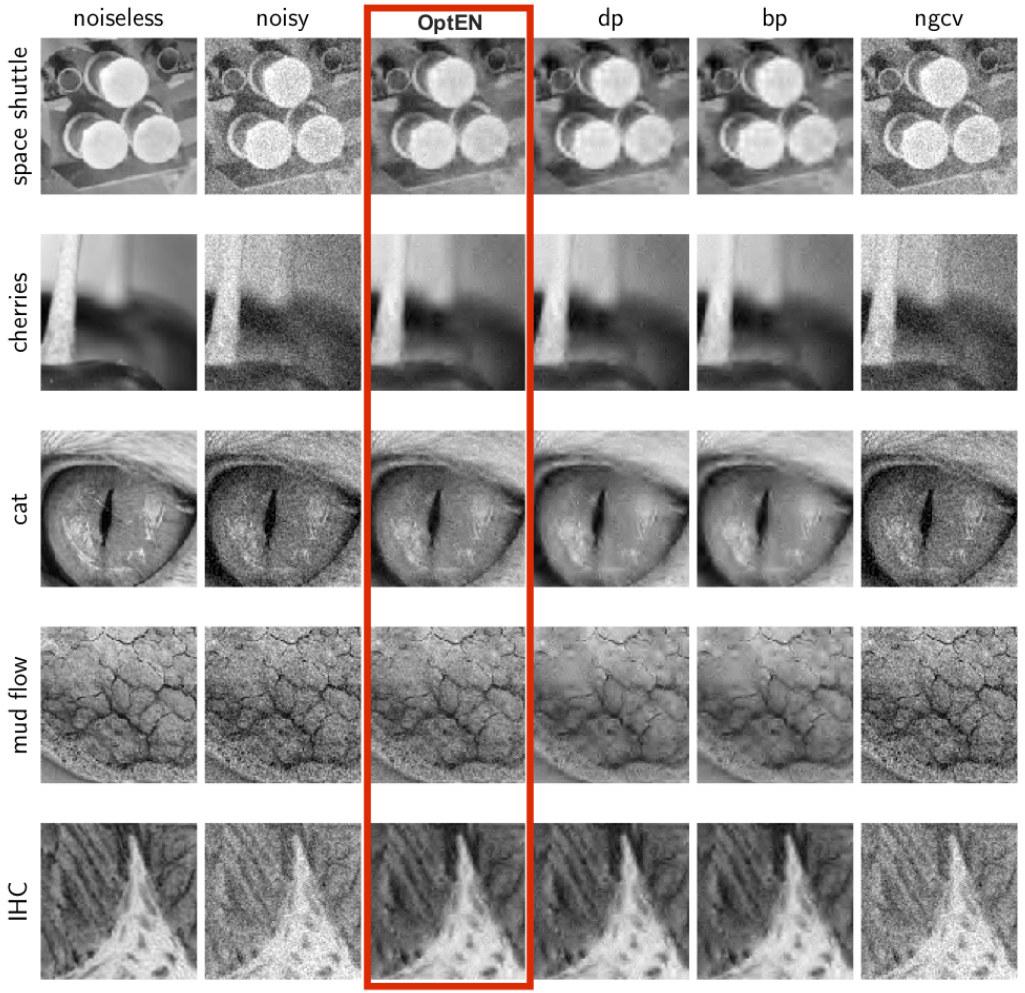
Numerical Examples

Image denoising using Elastic-net



Numerical Examples

Image denoising using Elastic-net



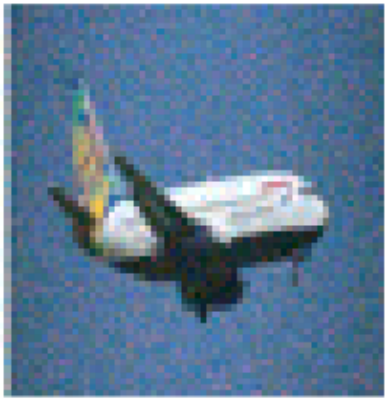
OptEN delivers the best PSNR and SSIM for all images for various noise levels

Numerical Examples

Image denoising using TV

Original

Noisy



Estimated parameter

Optimal parameter

Numerical Examples

Image denoising using TV

Original

Noisy

Estimated parameter

Optimal parameter

Conclusion and Future Directions

Conclusion:

- Unsupervised machine learning approach for optimal regularization:
  - Theoretical results for data-driven parameter learning in Tikhonov and Elastic-net;
  - Practical implementation of the method;
  - Promising numerical results for TV-regularization.
- The approach determines the parameter that allows for achievement of the same quality of reconstruction in terms of PSNR and visual quality as the optimal parameter.

Future direction:

- Theoretical results for $\mathcal{V}$ being a lower-dimensional nonlinear manifold;
- Theoretical results when X belongs to unions of linear subspaces;
- Consider different noise models (results of J. C. Reyes and C. Schönlieb '13, '16);
- Applicability of the method for practical problems.

Conclusion and Future Directions

Conclusion:

- Unsupervised machine learning approach for optimal regularization:
  - Theoretical results for data-driven parameter learning in Tikhonov and Elastic-net;
  - Practical implementation of the method;
  - Promising numerical results for TV-regularization.
- The approach determines the parameter that allows for achievement of the same quality of reconstruction in terms of PSNR and visual quality as the optimal parameter.

Future direction:

- Theoretical results for $\mathcal{V}$ being a lower-dimensional nonlinear manifold;
- Theoretical results when X belongs to unions of linear subspaces;
- Consider different noise models (results of J. C. Reyes and C. Schönlieb '13, '16);
- Applicability of the method for practical problems.