Inverse Problems meets Statistical Learning

Gabriel Peyré

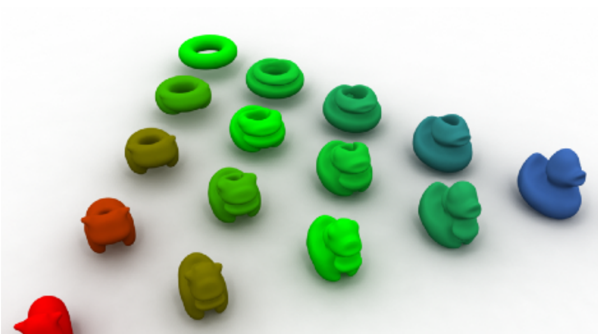
www.numerical-tours.com

Mathematical Coffees

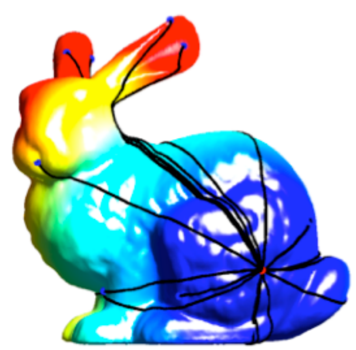
Huawei-FSMP joint seminars

https://mathematical-coffees.github.io

Organized by: Mérouane Debbah & Gabriel Peyré

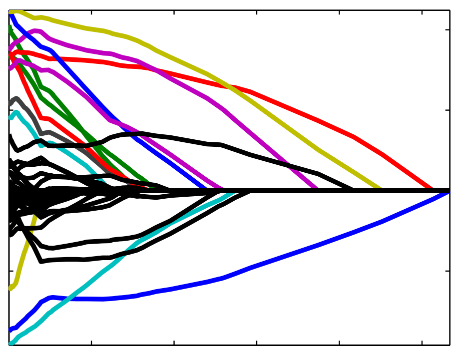


Optimal Transport

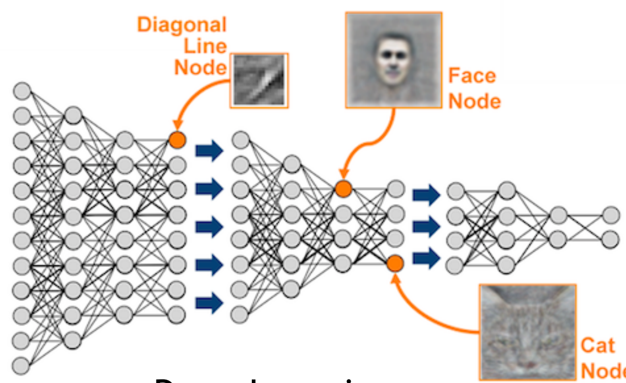


Geodesics

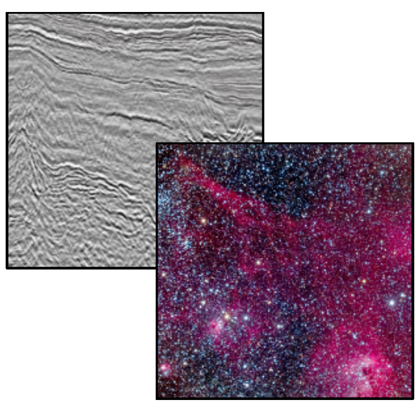
Meshes



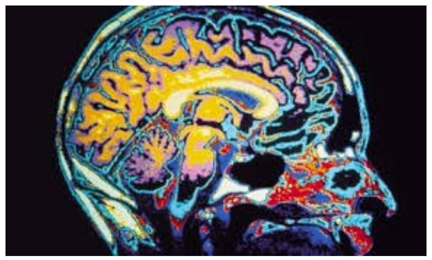
Optimization



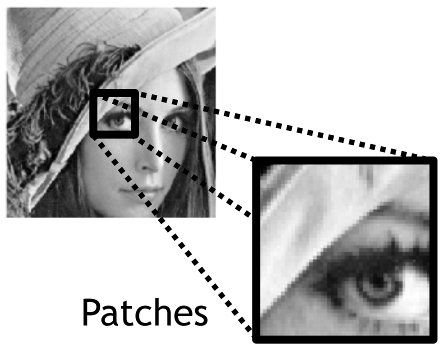
Deep Learning



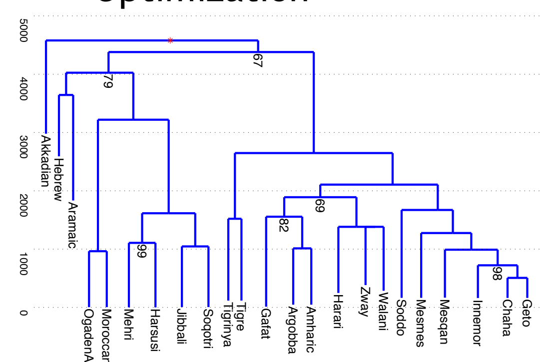
Sparsity



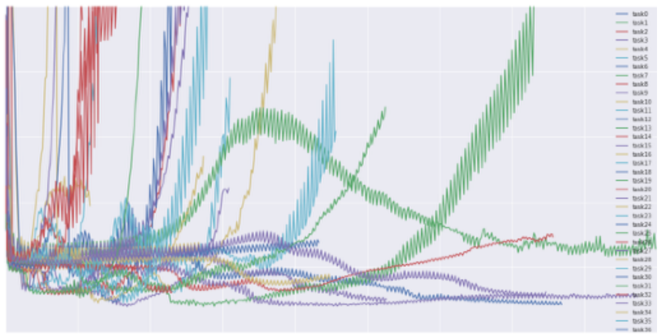
Neuro-imaging



Patches



Bayesian



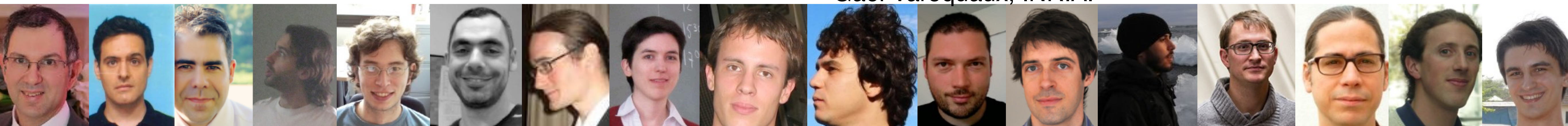
Parallel/Stochastic

Alexandre Allauzen, Paris-Sud.
Pierre Alliez, INRIA.
Guillaume Charpiat, INRIA.
Emilie Chouzenoux, Paris-Est.

Nicolas Courty, IRISA.
Laurent Cohen, CNRS Dauphine.
Marco Cuturi, ENSAE.
Julie Delon, Paris 5.

Fabian Pedregosa, INRIA.
Guillaume Lecué, CNRS ENSAE
Julien Tierny, CNRS and P6.
Robin Ryder, Paris-Dauphine.
Gael Varoquaux, INRIA.

Jalal Fadili, ENSICaen.
Alexandre Gramfort, INRIA.
Matthieu Kowalski, Supelec.
Jean-Marie Mirebeau, CNRS,P-Sud.



Inverse Problems

Forward model: $y = A\, f + w \in \mathbb{R}^m$

Observations

Operator
$A \in \mathbb{R}^{m \times p} : \mathbb{R}^p \to \mathbb{R}^m$

(Unknown)
Input

Noise

Inverse Problems

Forward model: $y = A\,f + w \in \mathbb{R}^m$

Observations

Operator

$A \in \mathbb{R}^{m \times p} : \mathbb{R}^p \to \mathbb{R}^m$

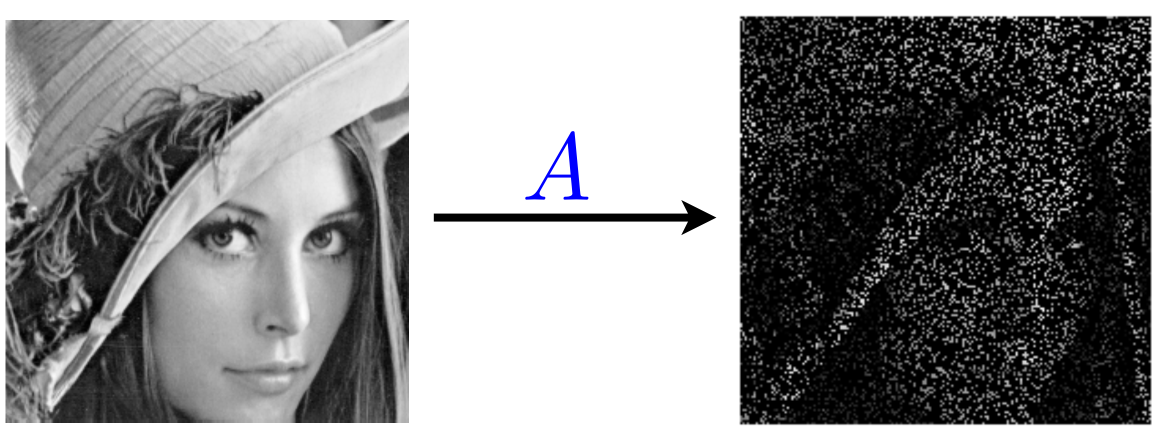
(Unknown) Input

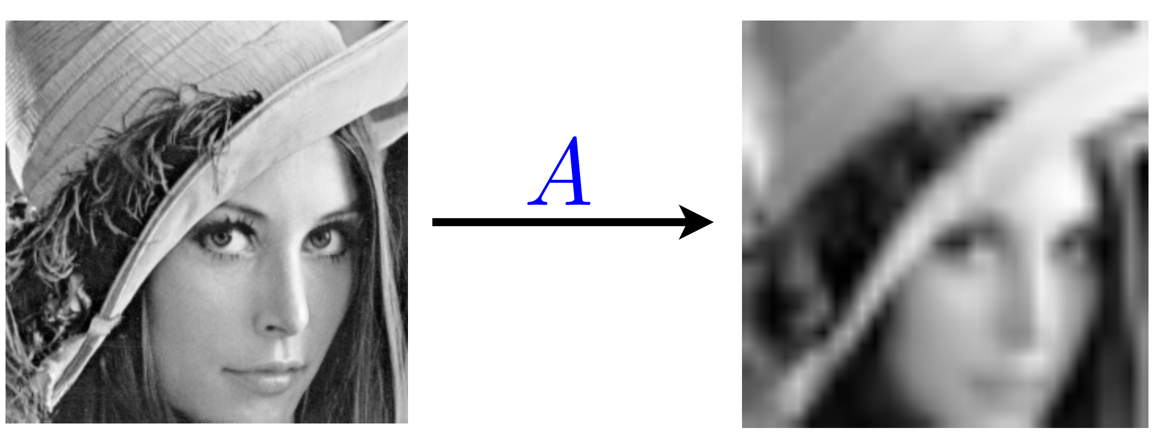
Noise

Denoising: $A = \mathrm{Id}_p$, $m = p$

Inpainting: set Ω of available pixels, $m = |\Omega|$, $\quad Af = (f_i)_{i \in \Omega}$

Super-resolution: $Af = (f \star k) \downarrow_\tau$, $m = p/\tau$.





Inverse Problems

Forward model: $y = A\, f + w \in \mathbb{R}^m$

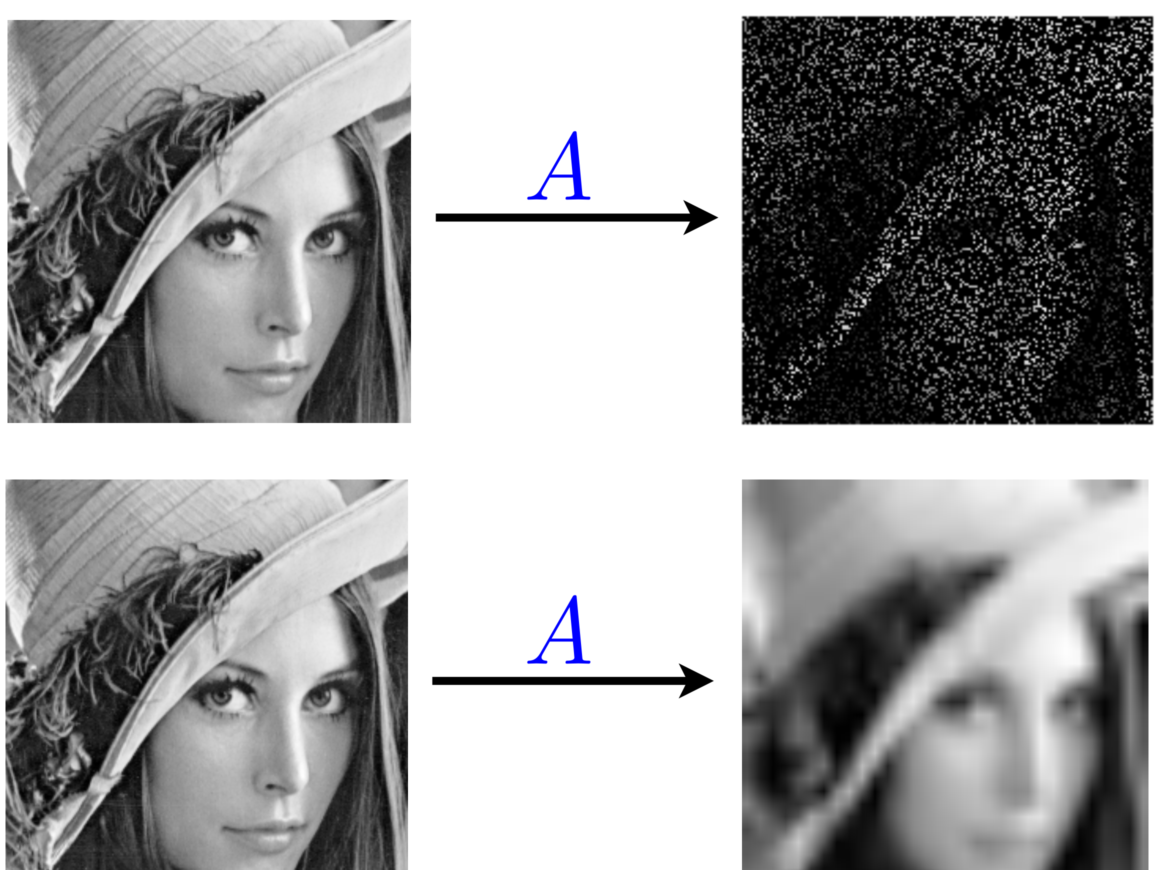
Observations, Operator, (Unknown) Input, Noise

$A \in \mathbb{R}^{m \times p} : \mathbb{R}^p \to \mathbb{R}^m$

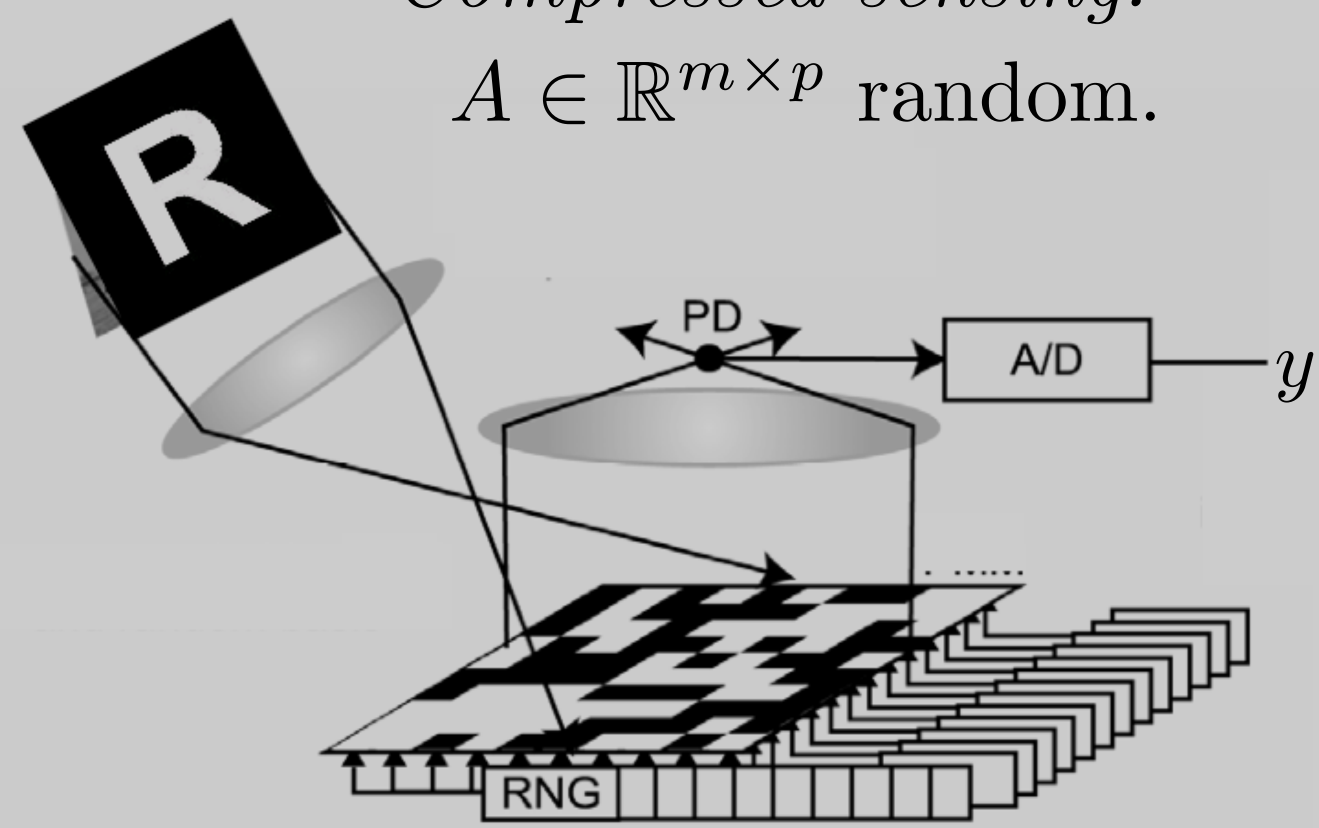
Denoising: $A = \mathrm{Id}_p$, $m = p$

Inpainting: set Ω of available pixels, $m = |\Omega|$, $\quad Af = (f_i)_{i \in \Omega}$

Super-resolution: $Af = (f \star k) \downarrow_\tau$, $m = p/\tau$.

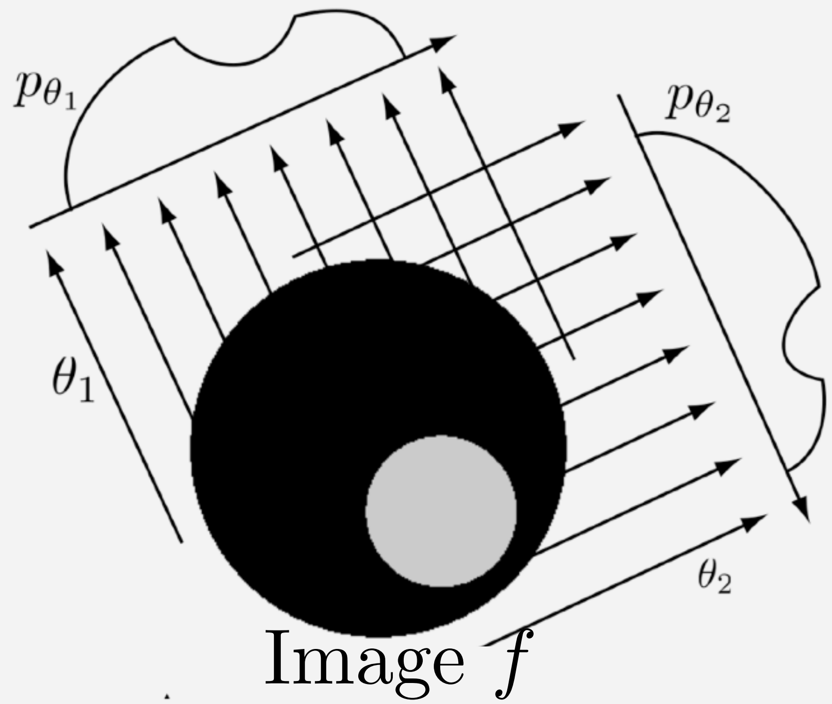


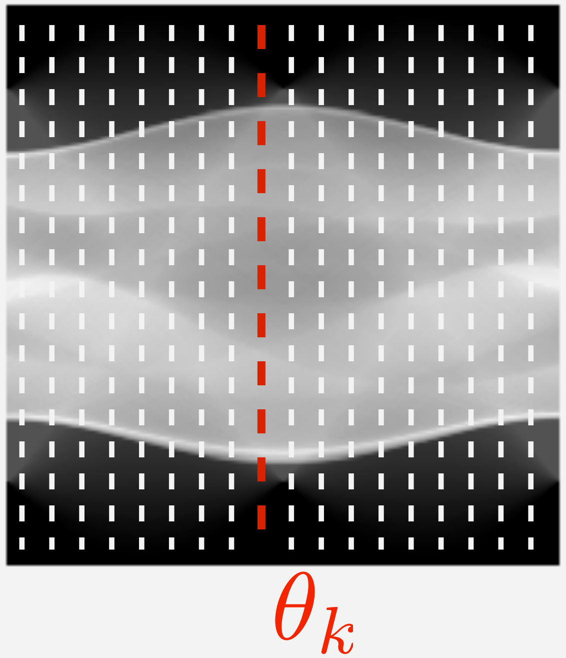
Compressed sensing:
$A \in \mathbb{R}^{m \times p}$ random.



Inverse Problem in Medical Imaging

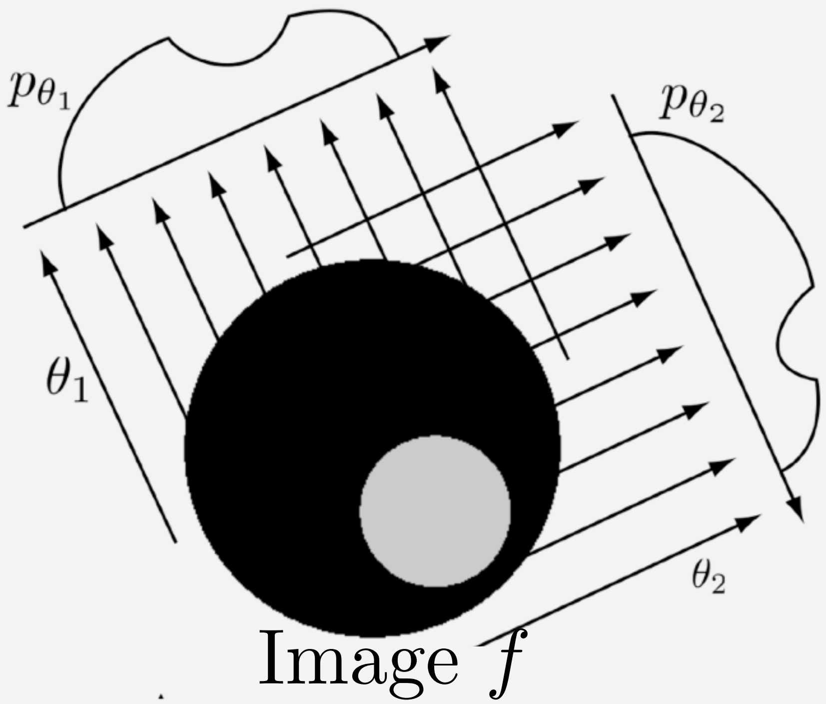
Tomography projection: $Af = (p_{\theta_k})_{k=1}^{K}$

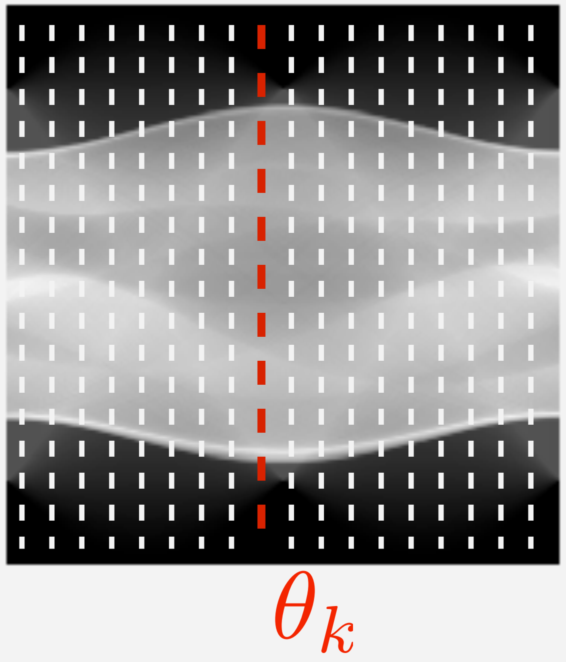




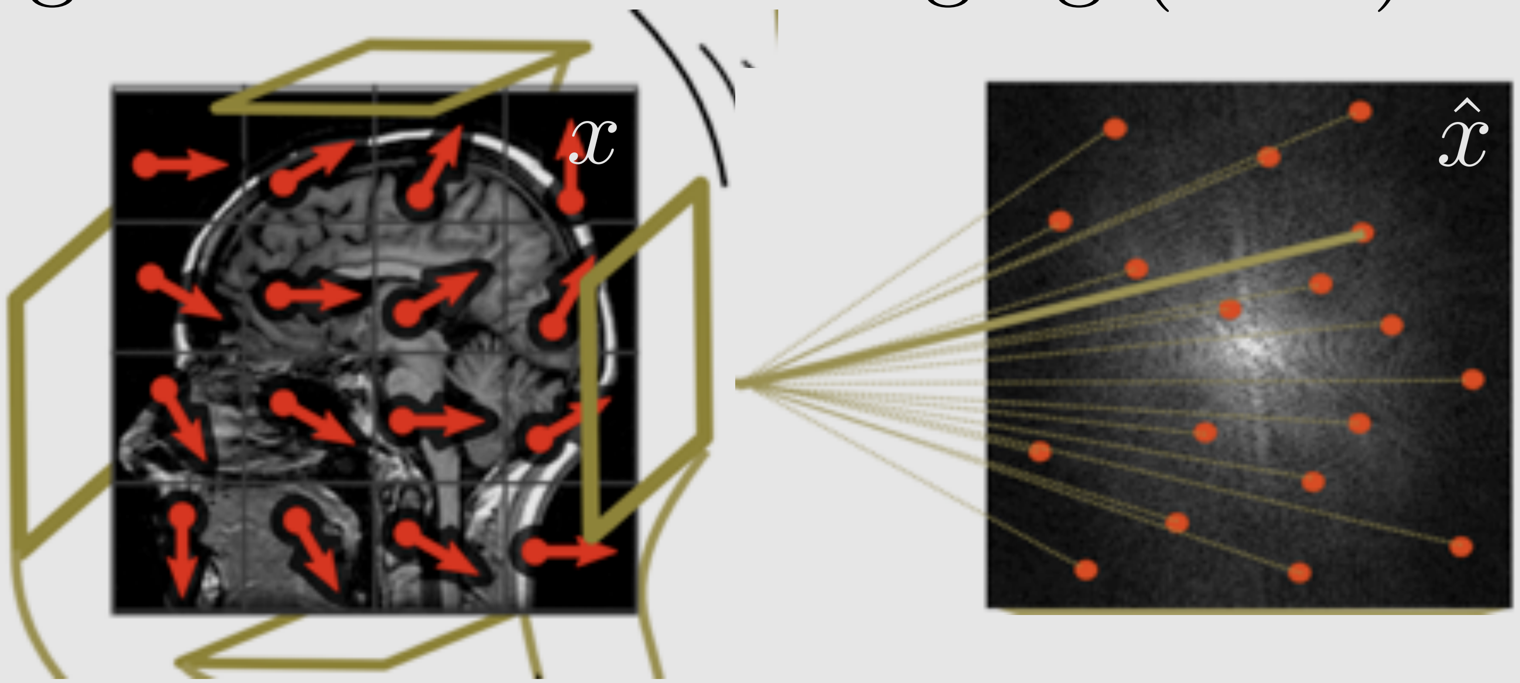
Inverse Problem in Medical Imaging

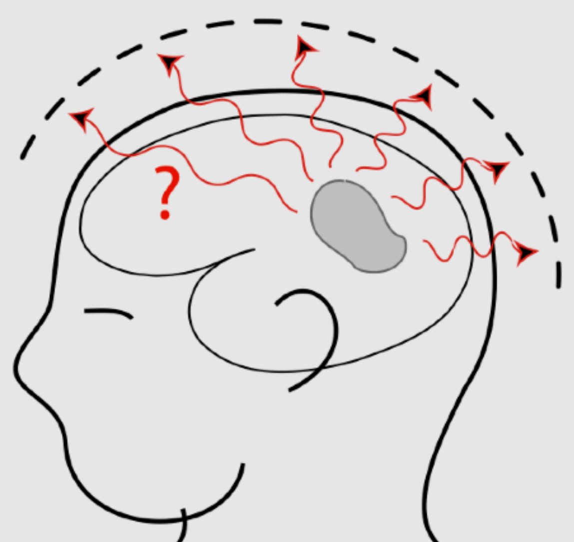
Tomography projection: $Af = (p_{\theta_k})_{k=1}^K$





Magnetic resonance imaging (MRI): $Af = (\hat{f}(\omega))_{\omega \in \Omega}$

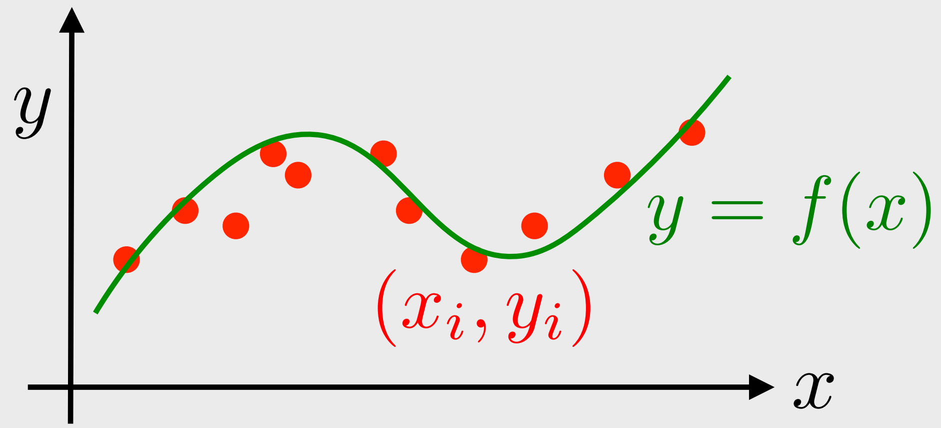




Other examples: MEG, EEG, . . .

Regression in Statistical Learning

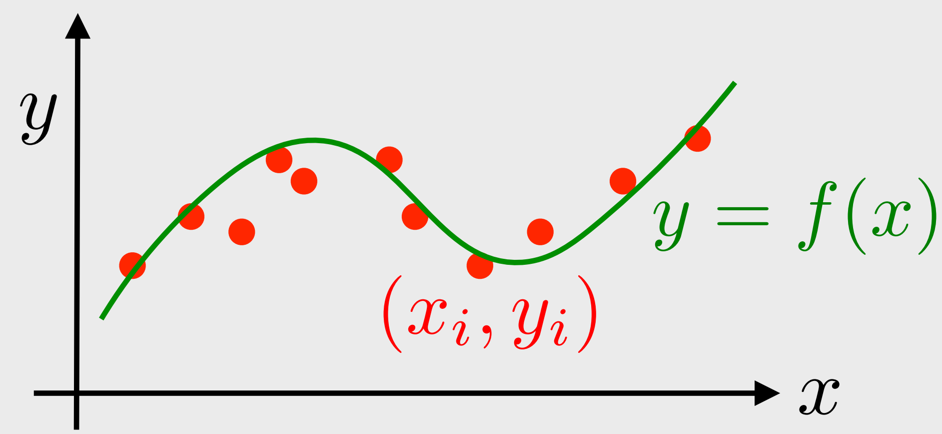
(Noisy) observations (x_i, y_j), try to infer $y = f(x)$.



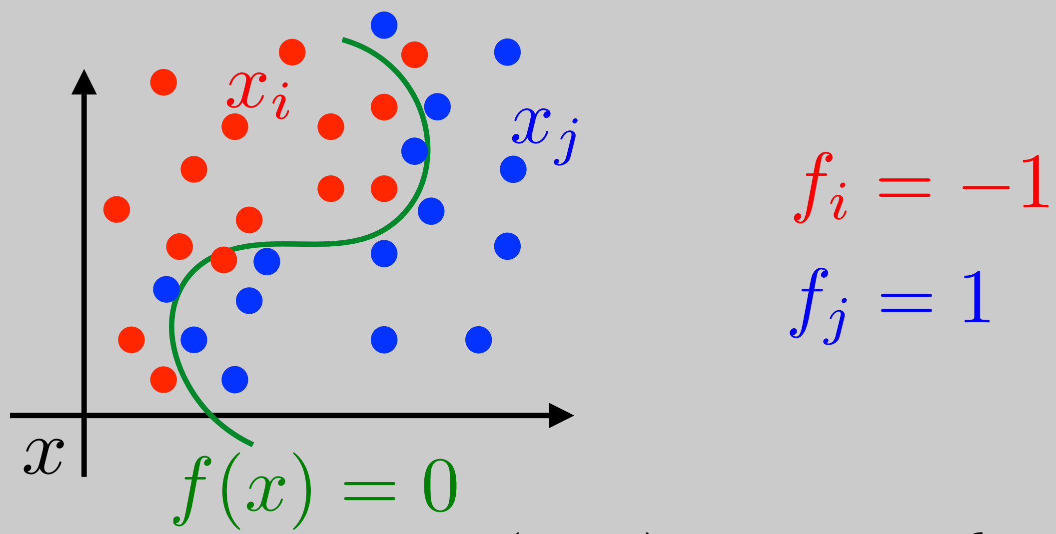
Regression $\quad (x, y) \in \mathbb{R}^p \times \mathbb{R}$

Regression in Statistical Learning

(Noisy) observations (x_i, y_j), try to infer $y = f(x)$.



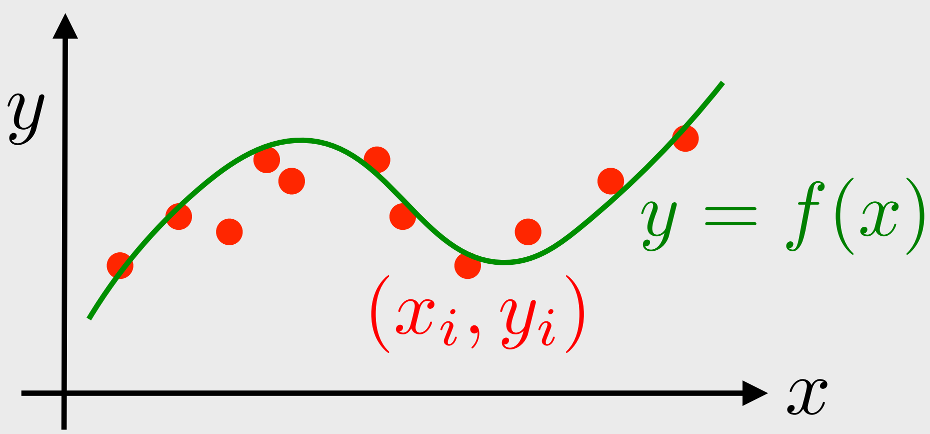
Regression $(x, y) \in \mathbb{R}^p \times \mathbb{R}$



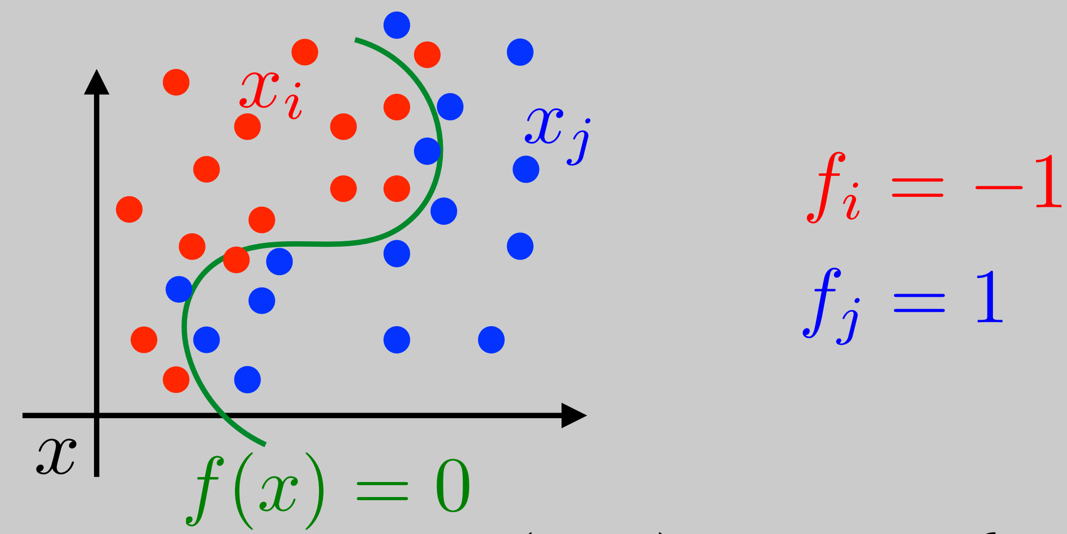
Classification $(x, y) \in \mathbb{R}^p \times \{-1, 1\}$

Regression in Statistical Learning

(Noisy) observations (x_i, y_j), try to infer $y = f(x)$.

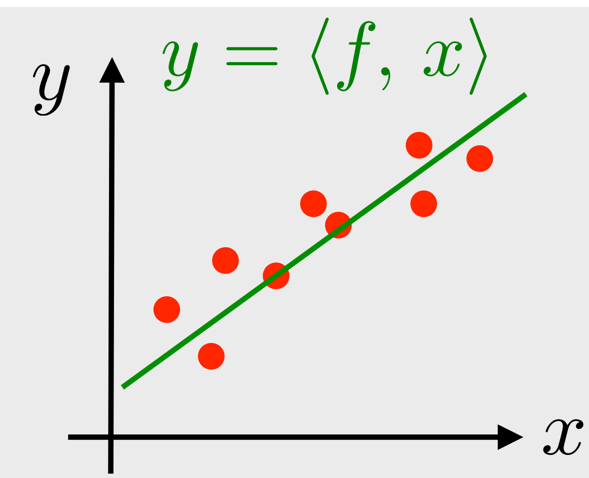


Regression $(x, y) \in \mathbb{R}^p \times \mathbb{R}$



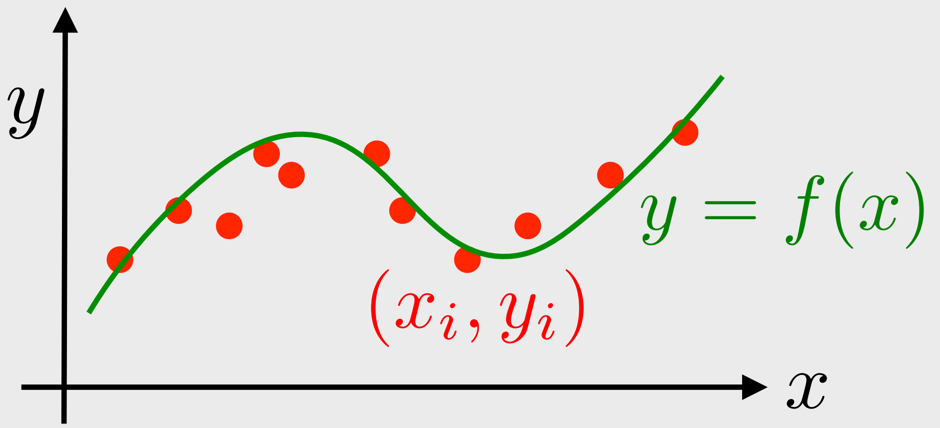
Classification $(x, y) \in \mathbb{R}^p \times \{-1, 1\}$

Linear models: $\forall\, i = 1, \ldots, n, \quad y_i = \langle x_i,\, f \rangle + \underbrace{\overbrace{\varepsilon_i}^{\text{noise}}}_{\text{model error}}$

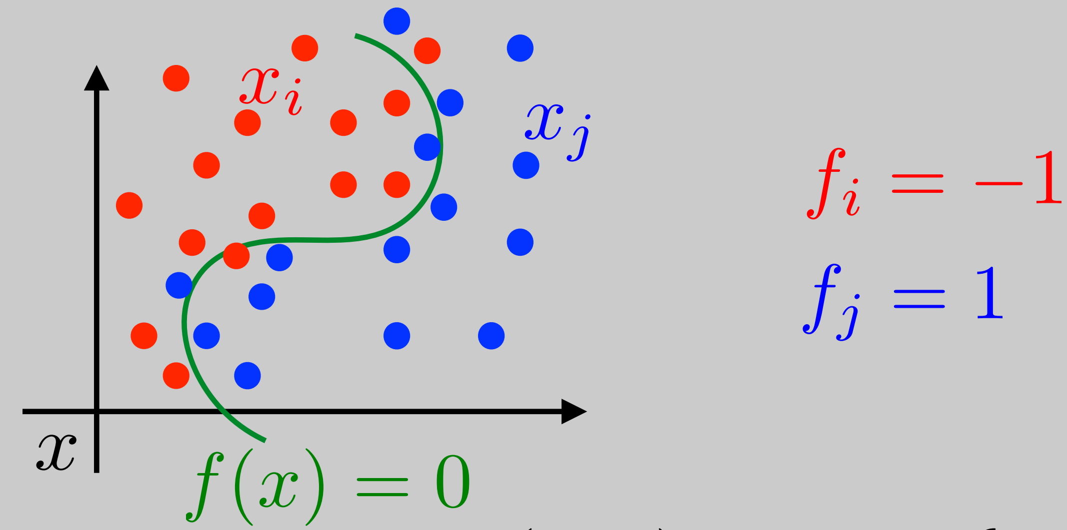


Regression in Statistical Learning

(Noisy) observations (x_i, y_j), try to infer $y = f(x)$.

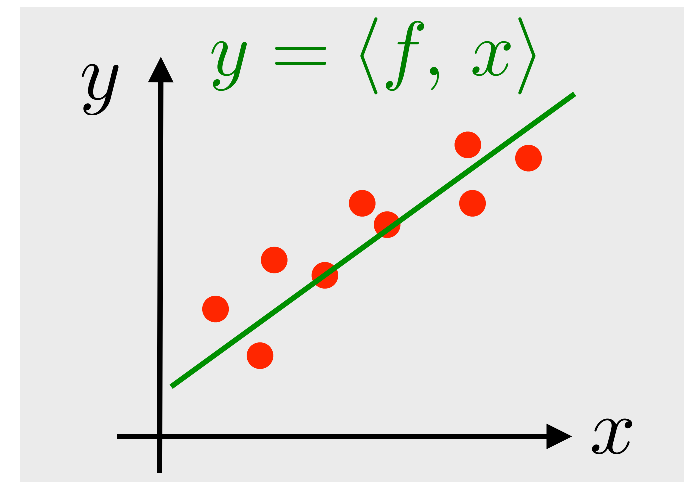


Regression $\quad (x, y) \in \mathbb{R}^p \times \mathbb{R}$

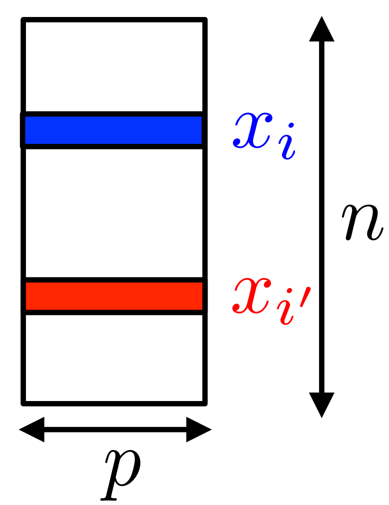


Classification $\quad (x, y) \in \mathbb{R}^p \times \{-1, 1\}$

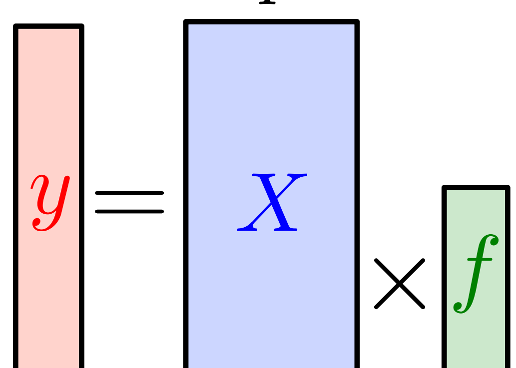
Linear models: $\quad \forall\, i = 1, \ldots, n, \quad y_i = \langle x_i,\, f\rangle + \overset{\text{noise}}{\underset{\text{model error}}{\varepsilon_i}}$



Empirical design matrix: $\quad X =$



Model: $\; y = Xf + \varepsilon \in \mathbb{R}^n$



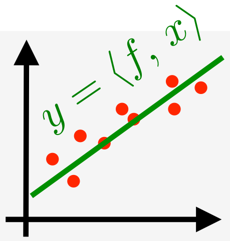
Inverse Problems vs. Statistical Learning

Inverse Problems

$$y = Af + w$$

Statistical Learning

$$y = Xf + \varepsilon$$



Inverse Problems vs. Statistical Learning

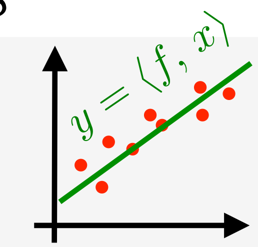
Inverse Problems

$$y = Af + w$$

$$\underbrace{A^\top y}_{\stackrel{\text{def.}}{=}\, u} = \underbrace{(A^\top A)}_{\stackrel{\text{def.}}{=}\, C} f + \underbrace{A^\top w}_{\stackrel{\text{def.}}{=}\, r}$$

Statistical Learning

$$y = Xf + \varepsilon$$



$$\underbrace{\frac{1}{n} X^\top y}_{\stackrel{\text{def.}}{=}\, u_n} = \underbrace{\frac{1}{n}(X^\top X)}_{\stackrel{\text{def.}}{=}\, C_n} f + \underbrace{\frac{1}{n} X^\top \varepsilon}_{\stackrel{\text{def.}}{=}\, r_n}$$

Inverse Problems vs. Statistical Learning

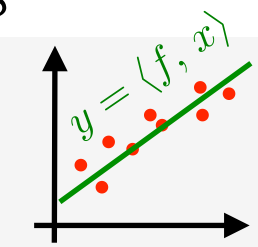
Inverse Problems

$$y = Af + w$$

$$\underbrace{A^\top y}_{\stackrel{\text{def.}}{=}\, u} = \underbrace{(A^\top A)}_{\stackrel{\text{def.}}{=}\, C} f + \underbrace{A^\top w}_{\stackrel{\text{def.}}{=}\, r}$$

Statistical Learning

$$y = Xf + \varepsilon$$



$$\underbrace{\frac{1}{n} X^\top y}_{\stackrel{\text{def.}}{=}\, u_n} = \underbrace{\frac{1}{n}(X^\top X)}_{\stackrel{\text{def.}}{=}\, C_n} f + \underbrace{\frac{1}{n} X^\top \varepsilon}_{\stackrel{\text{def.}}{=}\, r_n}$$

$n \to +\infty$, $(x_i, y_i)_i$ i.i.d.

$$u_n \downarrow \quad u = \mathbb{E}(yx) \qquad C_n \downarrow \quad C = \mathbb{E}(xx^\top)$$

Inverse Problems vs. Statistical Learning

Inverse Problems

$$y = Af + w$$

$$\underbrace{A^\top y}_{\stackrel{\text{def.}}{=}\, u} = \underbrace{(A^\top A)}_{\stackrel{\text{def.}}{=}\, C} f + \underbrace{A^\top w}_{\stackrel{\text{def.}}{=}\, r}$$

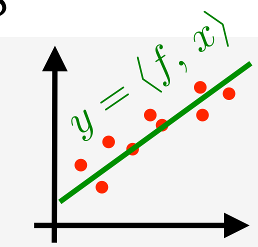
Regularized inversion:

$$\min_f \frac{1}{2}\|Af - y\|^2 + \lambda \|f\|^2$$

$$f_\lambda = (C + \lambda \mathrm{Id}_p)^{-1} u$$

Statistical Learning

$$y = Xf + \varepsilon$$



$$\underbrace{\frac{1}{n} X^\top y}_{\stackrel{\text{def.}}{=}\, u_n} = \underbrace{\frac{1}{n}(X^\top X)}_{\stackrel{\text{def.}}{=}\, C_n} f + \underbrace{\frac{1}{n} X^\top \varepsilon}_{\stackrel{\text{def.}}{=}\, r_n}$$

$n \to +\infty$ $\quad$ $(x_i, y_i)_i$ i.i.d.

$$u = \mathbb{E}(yx) \qquad C = \mathbb{E}(xx^\top)$$

Empirical risk minimization:

$$\min_f \frac{1}{2n}\|Xf - y\|^2 + \lambda \|f\|^2$$

$$f_{\lambda,n} = (C_n + \lambda \mathrm{Id}_p)^{-1} u_n$$

Inverse Problems vs. Statistical Learning

Inverse Problems

$$y = Af + w$$

$$\underbrace{A^\top y}_{\overset{\text{def.}}{=}\, u} = \underbrace{(A^\top A)}_{\overset{\text{def.}}{=}\, C} f + \underbrace{A^\top w}_{\overset{\text{def.}}{=}\, r}$$

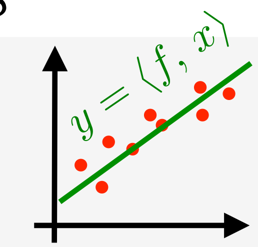
Regularized inversion:

$$\min_f \frac{1}{2}\|Af - y\|^2 + \lambda \|f\|^2$$

$$f_\lambda = (C + \lambda \mathrm{Id}_p)^{-1} u$$

Statistical Learning

$$y = Xf + \varepsilon$$



$$\underbrace{\frac{1}{n} X^\top y}_{\overset{\text{def.}}{=}\, u_n} = \underbrace{\frac{1}{n}(X^\top X)}_{\overset{\text{def.}}{=}\, C_n} f + \underbrace{\frac{1}{n} X^\top \varepsilon}_{\overset{\text{def.}}{=}\, r_n}$$

$n \to +\infty$, $(x_i, y_i)_i$ i.i.d.

$$u_n \to u = \mathbb{E}(yx) \qquad C_n \to C = \mathbb{E}(xx^\top)$$

Empirical risk minimization:

$$\min_f \frac{1}{2n}\|Xf - y\|^2 + \lambda \|f\|^2$$

$$f_{\lambda,n} = (C_n + \lambda \mathrm{Id}_p)^{-1} u_n$$

Inverse Problems		Statistical Learning
Exact covariance C	$\longleftrightarrow$	Noisy covariance C_n
Deterministic bounded noise r	$\longleftrightarrow$	Random noise r_n
Noise level $\|r\|$	$\longleftrightarrow$	Noise level $\|r_n\| \sim n^{-\frac{1}{2}}$

Theory: Convergence Rates

Inverse Problems

$$y = Af_0 + w$$

Statistical Learning

$$y_i = \langle x_i,\, f \rangle + \varepsilon_i \quad \text{i.i.d.}$$

$$y = Xf_0 + \varepsilon$$

Theory: Convergence Rates

Inverse Problems

$$y = Af_0 + w$$

Statistical Learning

$$y_i = \langle x_i,\, f\rangle + \varepsilon_i \quad \text{i.i.d.}$$

$$y = Xf_0 + \varepsilon$$

Source condition: $\exists z, \quad f_0 = \Phi^* z$

$\longrightarrow$ smoothness constraint.

$\longrightarrow$ $f_0 \perp \ker(\Phi)$

"no free lunch"

Theory: Convergence Rates

Inverse Problems

$$y = Af_0 + w$$

Statistical Learning

$$y_i = \langle x_i,\, f\rangle + \varepsilon_i \quad \text{i.i.d.}$$

$$y = Xf_0 + \varepsilon$$

Source condition: $\exists z, \quad f_0 = \Phi^* z$

$\longrightarrow$ smoothness constraint.

$\longrightarrow f_0 \perp \ker(\Phi)$

“no free lunch”

Theorem: setting $\lambda \sim \|w\|$,

$$\|f_\lambda - f_0\| \sim \sqrt{\|w\|}$$

$$\|Af_\lambda - Af_0\| \sim \|w\|$$

Theory: Convergence Rates

Inverse Problems

$$y = Af_0 + w$$

Statistical Learning

$$y_i = \langle x_i,\, f\rangle + \varepsilon_i \quad \text{i.i.d.}$$

$$y = Xf_0 + \varepsilon$$

Source condition: $\exists z, \quad f_0 = \Phi^* z$

$\longrightarrow$ smoothness constraint.

$\longrightarrow$ $f_0 \perp \ker(\Phi)$

"no free lunch"

Theorem: setting $\lambda \sim \|w\|$,

$$\|f_\lambda - f_0\| \sim \sqrt{\|w\|}$$

$$\|Af_\lambda - Af_0\| \sim \|w\|$$

Theorem: setting $\lambda \sim n^{-\frac{1}{2}}$,

$$\mathbb{E}(\|f_{\lambda,n} - f_0\|) \sim n^{-\frac{1}{4}}$$

$$\mathbb{E}(|\langle f - f_0,\, x\rangle|) \sim n^{-\frac{1}{2}}$$

Theory: Convergence Rates

Inverse Problems

$$y = Af_0 + w$$

Statistical Learning

$$y_i = \langle x_i,\, f\rangle + \varepsilon_i \quad \text{i.i.d.}$$
$$y = Xf_0 + \varepsilon$$

Source condition: $\exists z, \quad f_0 = \Phi^* z$

$\longrightarrow$ smoothness constraint.

$\longrightarrow$ $f_0 \perp \ker(\Phi)$

"no free lunch"

Theorem: setting $\lambda \sim \|w\|$,

$$\|f_\lambda - f_0\| \sim \sqrt{\|w\|}$$
$$\|Af_\lambda - Af_0\| \sim \|w\|$$

Theorem: setting $\lambda \sim n^{-\frac{1}{2}}$,

$$\mathbb{E}(\|f_{\lambda,n} - f_0\|) \sim n^{-\frac{1}{4}}$$
$$\mathbb{E}(|\langle f - f_0,\, x\rangle|) \sim n^{-\frac{1}{2}}$$

Faster $O(\|w\|, n^{-\frac{1}{2}})$ estimation rates

Super-resolution effect

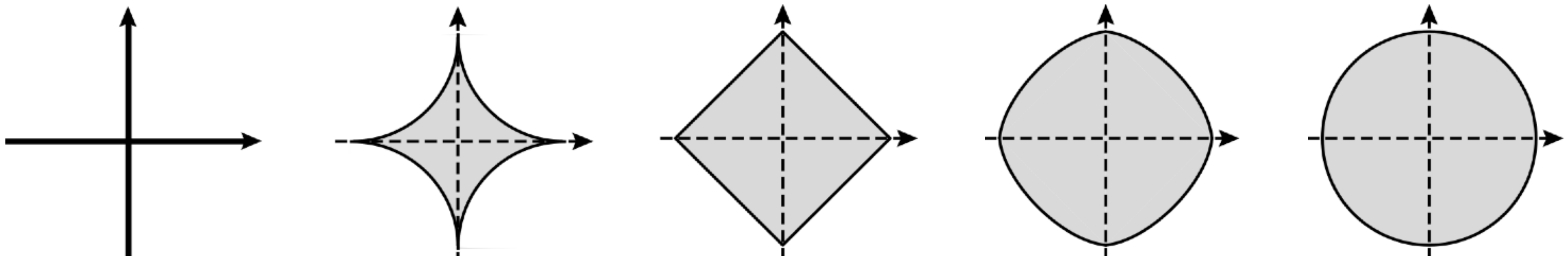
(recover information in $\ker(\Phi)$)

$\longrightarrow$

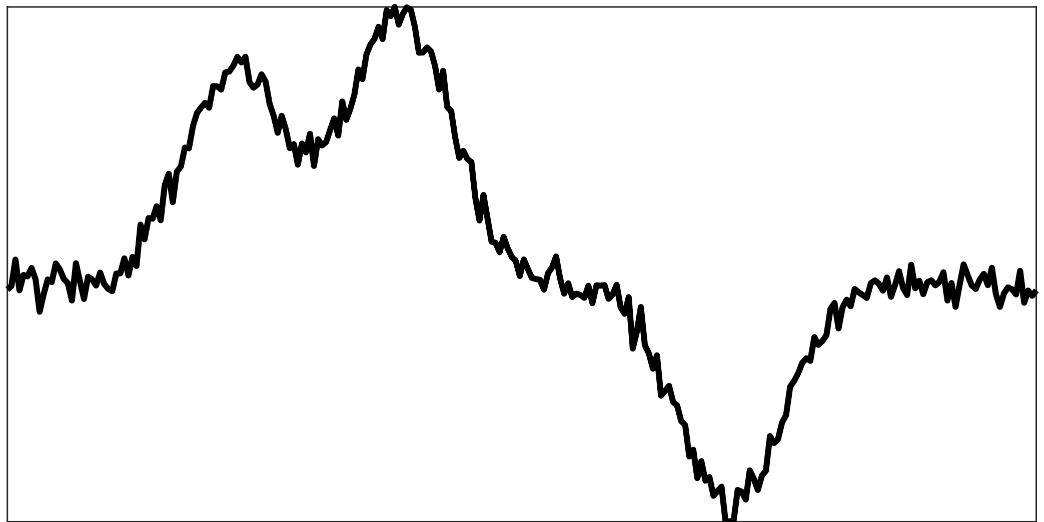
Needs non-quadratic &

non-smooth regularization

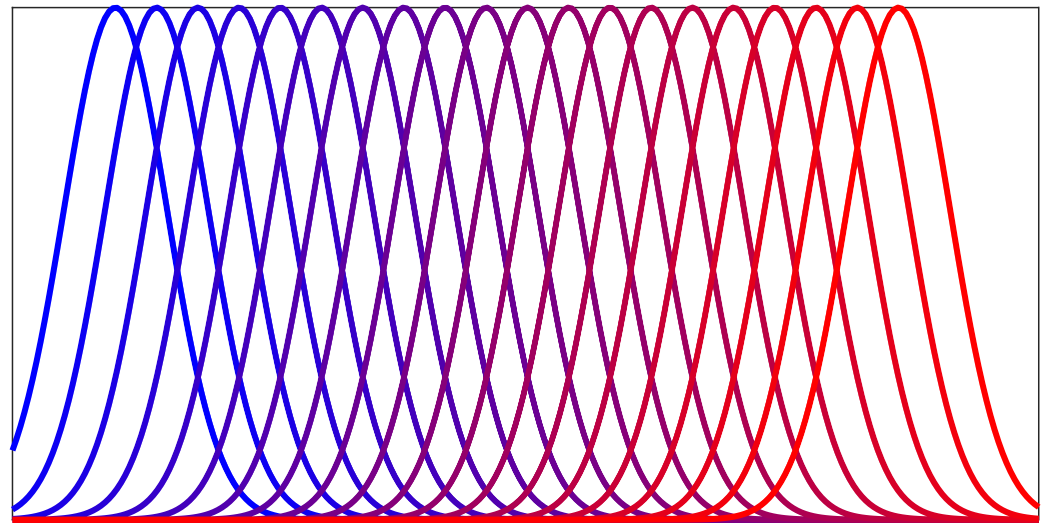
(ℓ_1, TV, trace norm, $\dots$)



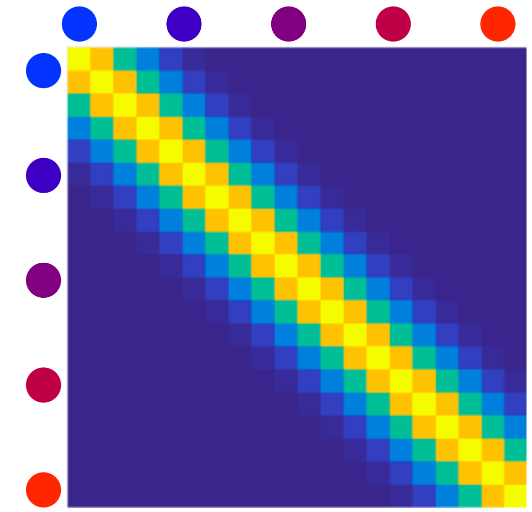
L2 vs. L1 Regularization



Observations y

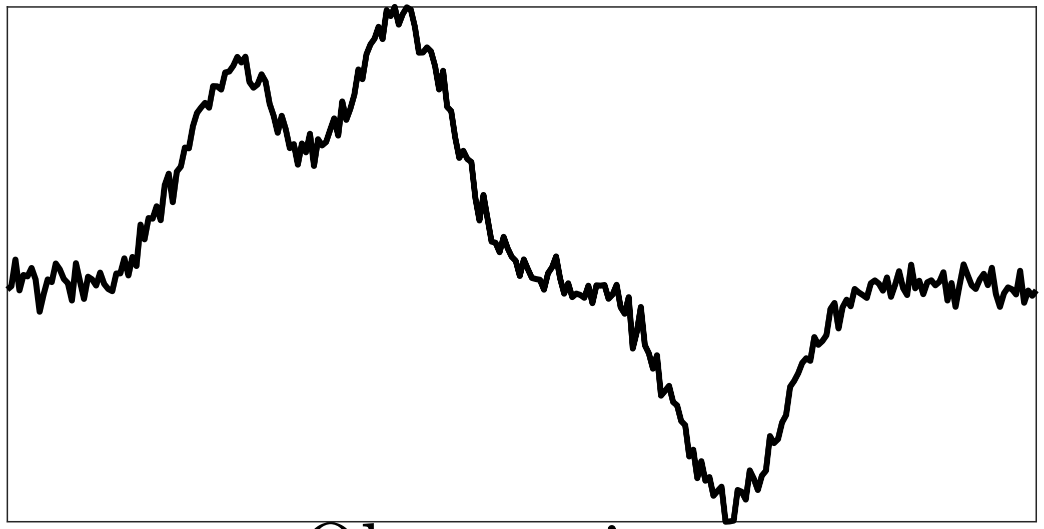


Columns of A

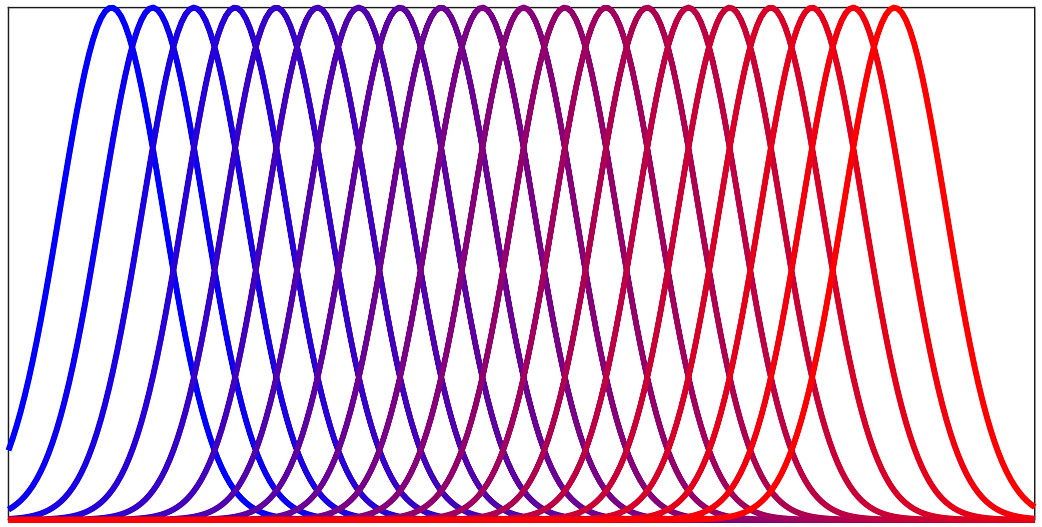


$C = A^\top A$

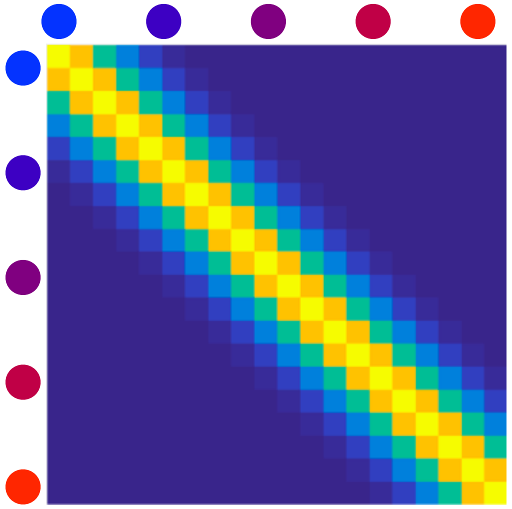
L2 vs. L1 Regularization



Observations y

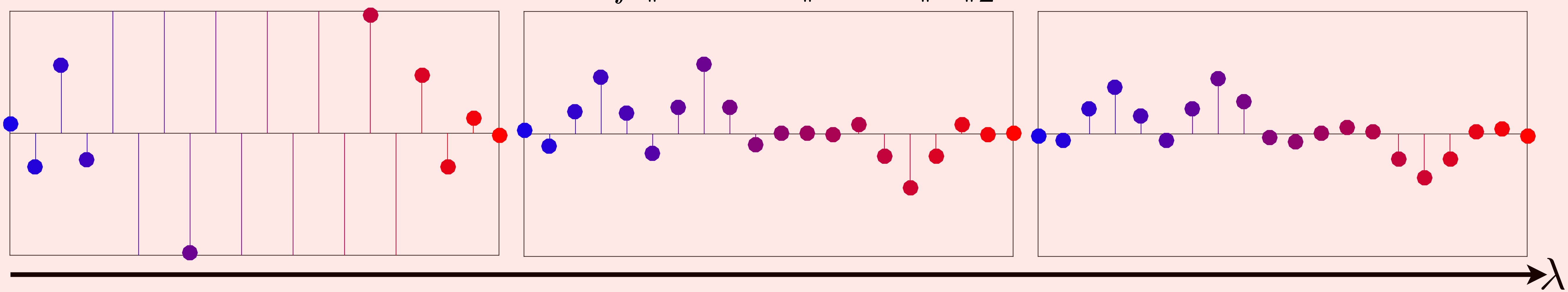


Columns of A



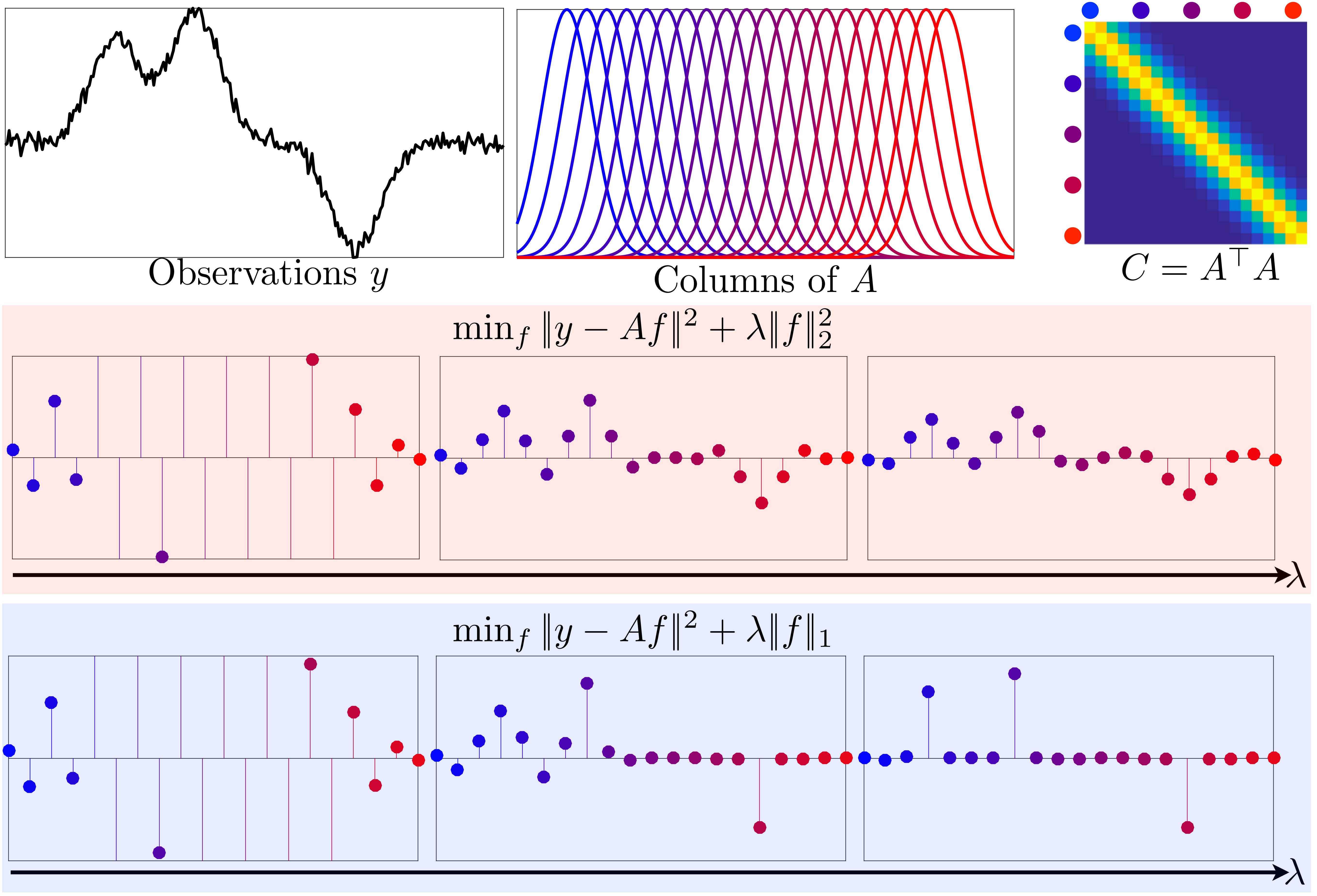
$C = A^\top A$

$$\min_f \|y - Af\|^2 + \lambda \|f\|_2^2$$



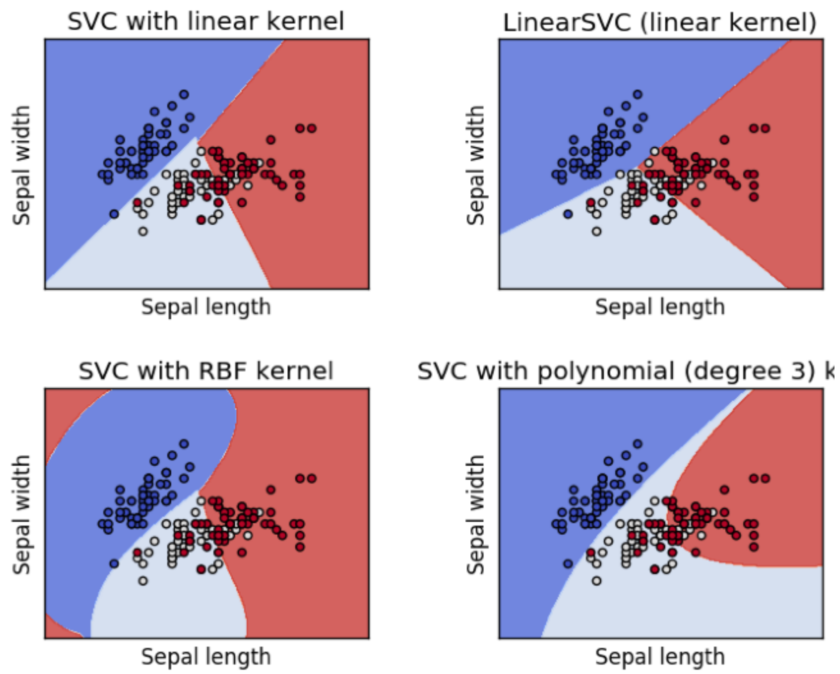
λ

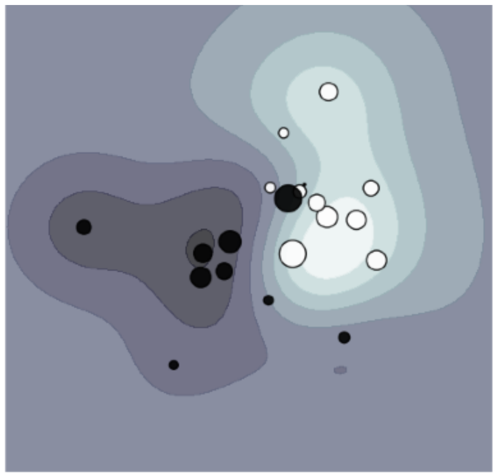
L2 vs. L1 Regularization

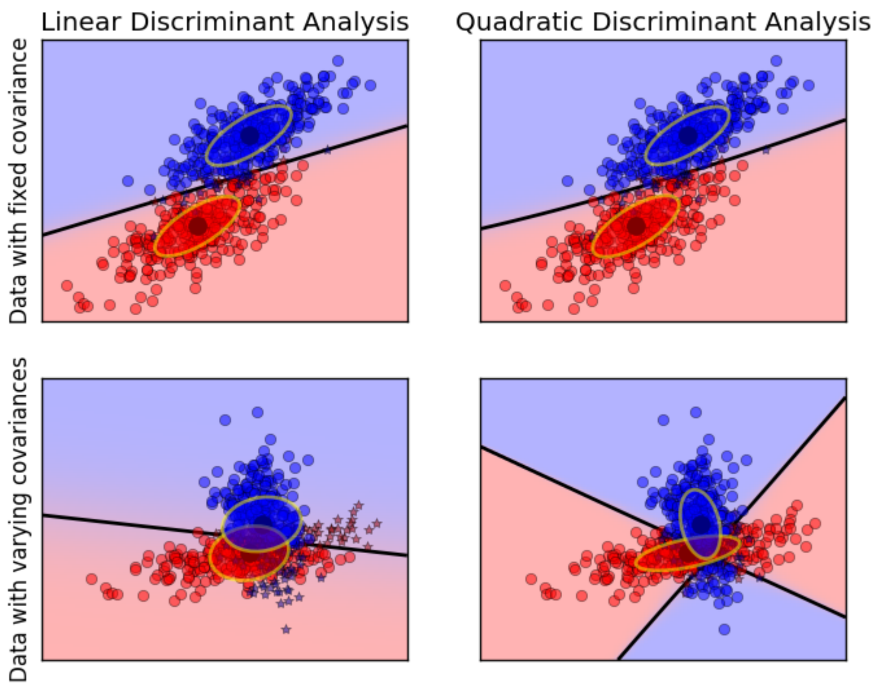


What’s Next

Alexandre Gramfort: ML for classification.







Gael Varoquaux: