Learning Filter Functions in Regularisers by Minimising Quotients

SIAM annual meeting, Pittsburgh
13 July 2017

Martin Benning[1], Guy Gilboa[2], Joana Sarah Grah[1,2], **Carola-Bibiane Schönlieb**[1]

[1]**Department of Applied Mathematics and Theoretical Physics, University of Cambridge, United Kingdom**

[2]**Electrical Engineering Department, Technion - Israel Institute of Technology, Haifa, Israel**

UNIVERSITY OF CAMBRIDGE

Technion
Israel Institute of Technology

There would be no talk if it had not been for...

Motivation

Martin Benning, Guy Gilboa, and CBS. "Learning Filter Functions in Regularisers by Minimising Quotients." *PAMM* 16.1 (2016): 933-936.

Motivation

Martin Benning, Guy Gilboa, and CBS. "Learning Filter Functions in Regularisers by Minimising Quotients." *PAMM* 16.1 (2016): 933-936.

Idea: Learning parametrised regularisation functions by quotient minimisation, where both wanted and unwanted outcomes are incorporated in the model by integrating the former in the numerator and the latter in the denominator

Motivation

Martin Benning, Guy Gilboa, and CBS. "Learning Filter Functions in Regularisers by Minimising Quotients." *PAMM* 16.1 (2016): 933-936.

Idea: Learning parametrised regularisation functions by quotient minimisation, where both wanted and unwanted outcomes are incorporated in the model by integrating the former in the numerator and the latter in the denominator

$$\hat{h} \in \operatorname*{arg\,min}_{\|h\|_2=1} \frac{J(u^+; h)}{J(u^-; h)}, \qquad J(u; h) := \|u * h\|_1,$$

Motivation

Martin Benning, Guy Gilboa, and CBS. "Learning Filter Functions in Regularisers by Minimising Quotients." *PAMM* 16.1 (2016): 933-936.

Idea: Learning parametrised regularisation functions by quotient minimisation, where both wanted and unwanted outcomes are incorporated in the model by integrating the former in the numerator and the latter in the denominator

$$\hat{h} \in \operatorname*{arg\,min}_{\|h\|_2=1} \frac{J(u^+; h)}{J(u^-; h)}, \qquad J(u; h) := \|u * h\|_1,$$

Aim: Learn convolution kernel h that makes u^+ sparse! Hope that u^+ can be represented by just a few building components, cf. works on dictionary learning, e.g.

Bruckstein, A.M., D.L. Donoho, and M. Elad. "From sparse solutions of systems of equations to sparse modeling of signals and images." *SIAM review* 51.1 (2009): 34-81.

Algorithm: Generalised Inverse Power Method

Algorithm 1 Inverse Power Method (Hein and Bühler)

Initialisation: Randomly generalised f^0 with $\|f^0\|_2 = 1$

while $\frac{|\lambda^{k+1} - \lambda^k|}{\lambda^k} < \varepsilon$ **do**

$$f^{k+1} = \underset{\|u\|_2 \leq 1}{\arg\min} \left\{ R(u) - \lambda^k \langle u, s(f^k) \rangle \right\}, \quad s(f^k) \in \partial S(f^k)$$

$$\lambda^{k+1} = \frac{R(f^{k+1})}{S(f^{k+1})}$$

end while

Output: Eigenvalue λ^{k+1}, Eigenfunction f^{k+1}

Matthias Hein and Thomas Bühler. "An inverse power method for nonlinear eigenproblems with applications in 1-spectral clustering and sparse PCA." *Advances in Neural Information Processing Systems.* 2010.

Eigenproblems

Linear Eigenproblem

$$Af - \lambda f = 0$$

$$F(f) = \frac{\langle f, Af \rangle}{\|f\|_2^2}$$

Non-linear Eigenproblem

$$\partial R(f) - \lambda \partial S(f) \ni 0$$

assuming $R, S \colon \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ convex, Lipschitz continuous, positively one-homogeneous and $S(f) = 0 \Leftrightarrow f = 0$

$$F(f) = \frac{R(f)}{S(f)}$$

Methods to Solve Eigenproblems

Linear case:
Inverse Power Method

$$Af^{k+1} = f^k$$
$$\Leftrightarrow Af^{k+1} - f^k = 0$$

$$f^{k+1} = \arg\min_u \frac{1}{2}\langle u, Au\rangle - \langle u, f^k\rangle$$

Non-linear case:
Generalised Inverse Power Method

$$r(f^{k+1}) - s(f^k) = 0$$

$$r(f^{k+1}) \in \partial R(f^{k+1}),\ s(f^k) \in \partial S(f^k)$$

Optimisation problem:

$$f^{k+1} = \arg\min_u R(u) - \langle u, s(f^k)\rangle$$

Algorithm: Generalised Inverse Power Method

Algorithm 1 Inverse Power Method (Hein and Bühler)

Initialisation: Randomly generalised f^0 with $\|f^0\|_2 = 1$

while $\frac{|\lambda^{k+1} - \lambda^k|}{\lambda^k} < \varepsilon$ **do**

$$f^{k+1} = \underset{\|u\|_2 \leq 1}{\arg\min} \left\{ R(u) - \lambda^k \langle u, s(f^k) \rangle \right\}, \quad s(f^k) \in \partial S(f^k)$$

$$\lambda^{k+1} = \frac{R(f^{k+1})}{S(f^{k+1})}$$

end while

Output: Eigenvalue λ^{k+1}, Eigenfunction f^{k+1}

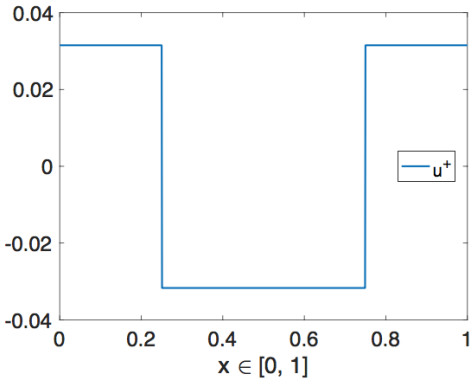
Matthias Hein and Thomas Bühler. "An inverse power method for nonlinear eigenproblems with applications in 1-spectral clustering and sparse PCA." *Advances in Neural Information Processing Systems.* 2010.

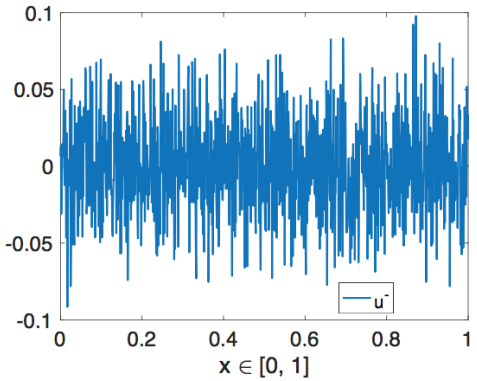
Algorithm: Regularisation Learning Model

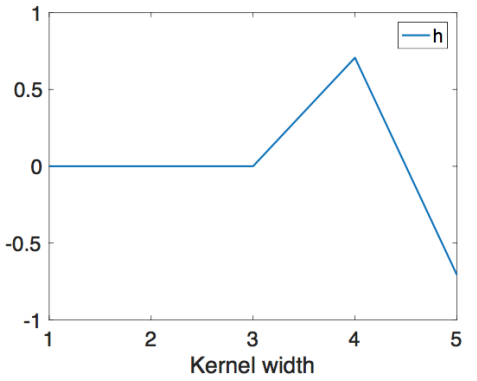
$$\hat{h} \in \operatorname*{arg\,min}_{\|h\|_2=1} \frac{J(u^+; h)}{J(u^-; h)}$$

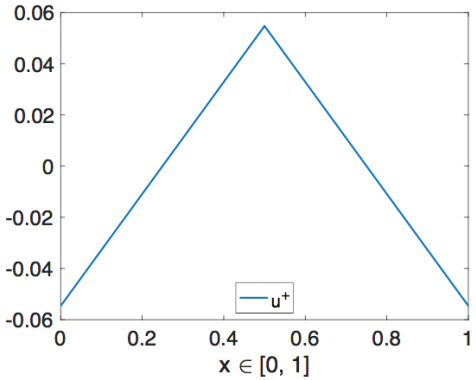
$$\begin{cases} h^{k+\frac{1}{2}} & = \operatorname*{arg\,min}_{\sum_{j=1}^n h_j=0} \left\{ J(u^+; h) - \mu^k \langle h, p^k \rangle \right\} \\ h^{k+1} & = \dfrac{h^{k+\frac{1}{2}}}{\left\| h^{k+\frac{1}{2}} \right\|_2} \\ p^{k+1} & \in \partial J(u^-; h^{k+1}) \\ \mu^{k+1} & = \dfrac{J(u^+; h^{k+1})}{J(u^-; h^{k+1})} \end{cases}$$

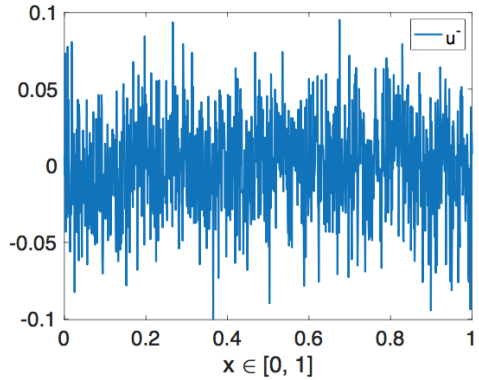
1D Examples: First- and second-order filters

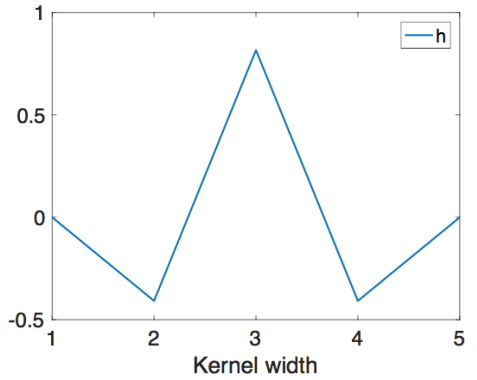




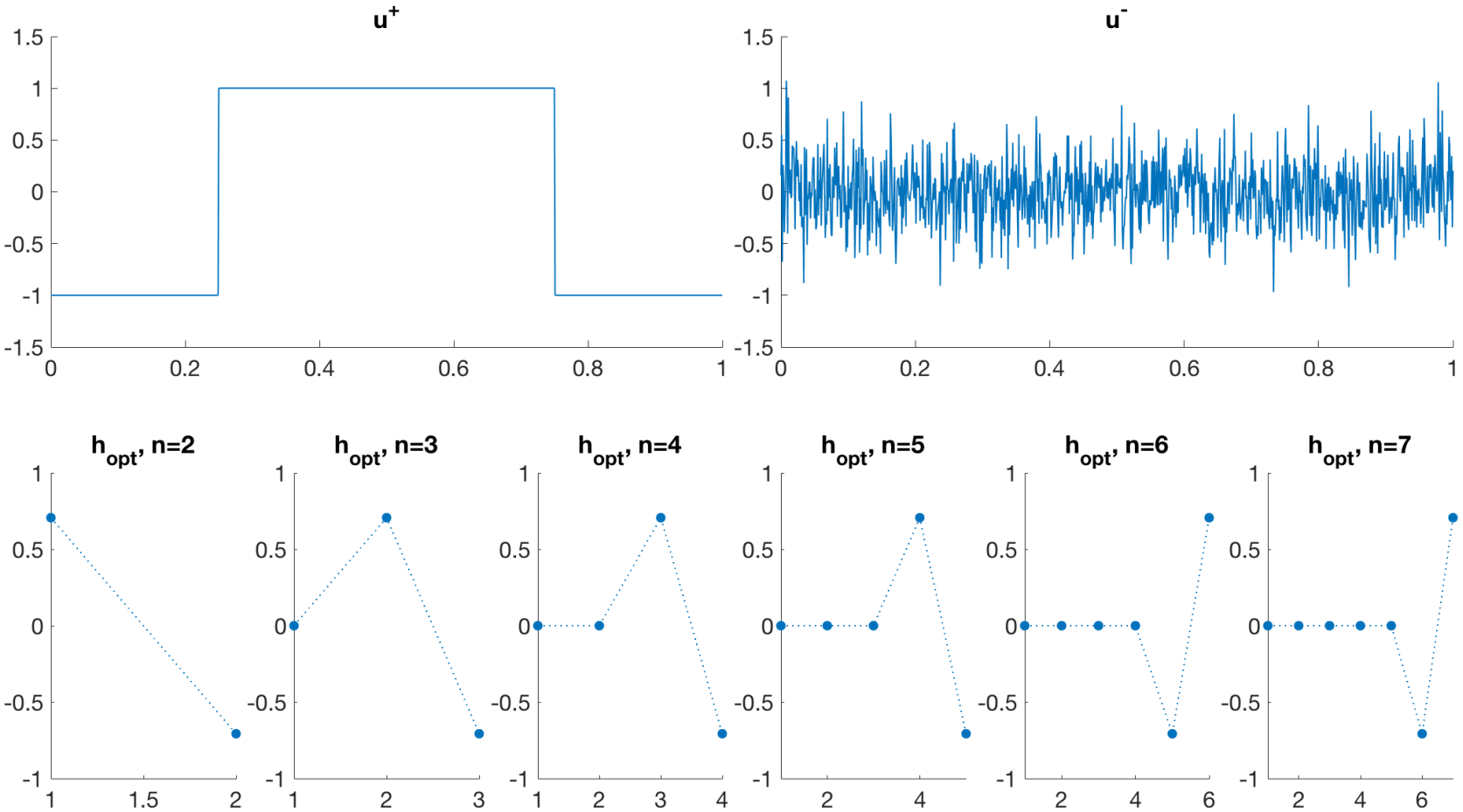








1D Example: Increasing filter size



1D Examples: Reconstruction

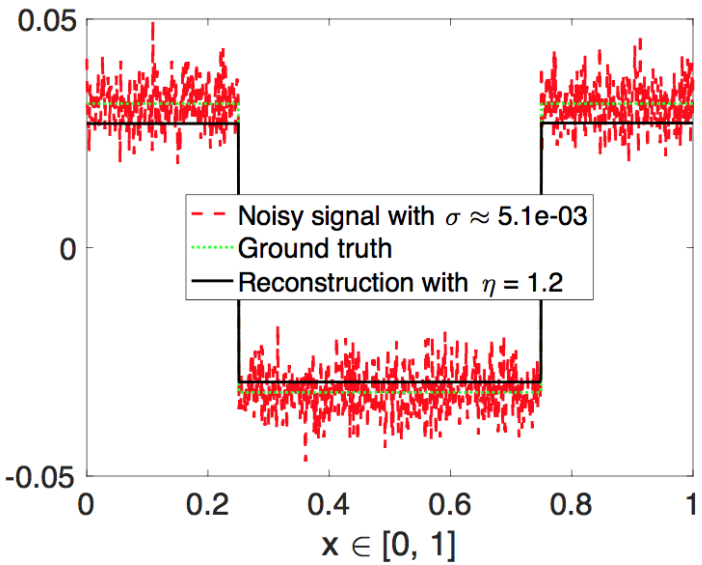
Reconstruction of a noisy signal in order to test the behaviour of the optimal filter is obtained by solving the following constrained optimisation problem:

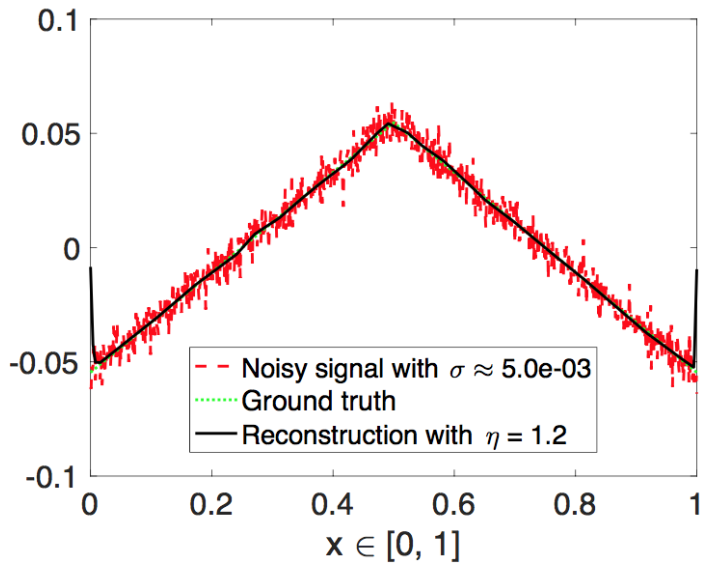
$$\hat{u} = \underset{u \in \mathbb{R}^m}{\arg\min}\, J(u; \hat{h}) \qquad \text{subject to} \quad \|u - f\|_2 \leq \eta\sigma\sqrt{m}\,.$$

1D Examples: Reconstruction

Reconstruction of a noisy signal in order to test the behaviour of the optimal filter is obtained by solving the following constrained optimisation problem:

$$\hat{u} = \operatorname*{arg\,min}_{u \in \mathbb{R}^m} J(u; \hat{h}) \qquad \text{subject to} \quad \|u - f\|_2 \leq \eta\sigma\sqrt{m}\,.$$





Remark

- In this setting there is indeed an even more efficient way of finding suitable filter functions h
- Simplifying the model to a variant without the need of having a negative input function u^- yields the same results, which is a clear indicator that in the above-mentioned framework the numerator plays a dominant role
- In fact, varying the model to

$$\hat{h} \in \arg\min_h \frac{\|u^+ * h\|_1}{\|h\|_2}$$

returns exactly the same solutions as the basic model

Extension of the Model

$$\hat{h} \in \underset{\substack{\|h\|_2=1 \\ \text{mean}(h)=0}}{\arg\min} \frac{\frac{1}{M}\sum_{i=1}^{M}\sum_{k=1}^{K} J(u_i^+; h_k)}{\frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{K} J(u_j^-; h_k)}, \qquad J(u; h) = \|u * h\|_1$$

- h is a K-dimensional parametrisation of a regularisation functional J; here, each h_k corresponds to a convolution kernel
- The signals u^+ and u^- are desired and undesired input signals, respectively

Martin Benning, Guy Gilboa, Joana Sarah Grah, and CBS. "Learning Filter Functions in Regularisers by Minimising Quotients." *Scale Space and Variational Methods in Computer Vision*, 2017, accepted.

Algorithm: Extended Model

$$\hat{h} \in \arg\min_{h} \left\{ \frac{F(h)}{G(h)} \quad \text{subject to} \quad \|h\|_2 = 1 \quad \text{and} \quad \text{mean}(h) = 0 \right\}$$

$$\begin{cases} h^{k+\frac{1}{2}} & = \arg\min\limits_{\text{mean}(h)=0} \left\{ F(h) - \mu^k \langle h - h^k, s^k \rangle + \left\| h - h^k \right\|_2^2 \right\} \\ \mu^{k+1} & = \dfrac{F(h^{k+\frac{1}{2}})}{G(h^{k+\frac{1}{2}})} \\ s^{k+1} & \in \partial G(h^{k+\frac{1}{2}}) \\ h^{k+1} & = \dfrac{h^{k+\frac{1}{2}}}{\left\| h^{k+\frac{1}{2}} \right\|_2} \end{cases}$$

Bresson, X., Laurent, T., Uminsky, D., Brecht, J.V.. "Convergence and energy landscape for Cheeger cut clustering." *Advances in Neural Information Processing Systems* (2012).

Brief Convergence Analysis

Following Section 3.2 of the below paper, we show two results that are essential for proving global convergence of our algorithm: a descent lemma and a bound of the subgradient by the iterates gap.

Jérôme Bolte, Shoham Sabach, and Marc Teboulle. "Proximal alternating linearized minimization for nonconvex and nonsmooth problems." *Mathematical Programming* 146.1-2 (2014): 459-494.

Brief Convergence Analysis

Lemma

Let F and G be proper, lower semi-continuous and convex functions. Then the iterates of our algorithm satisfy

$$\mu^{k+\frac{1}{2}} + \frac{1}{G(h^{k+\frac{1}{2}})}\|h^{k+\frac{1}{2}} - h^k\|^2 \leq \mu^k \,,$$

if we further assume $G(h^{k+\frac{1}{2}}) \neq 0$ for all $k \in \mathbb{N}$.

Brief Convergence Analysis

Lemma

Let F and G be proper, lower semi-continuous and convex functions. Then the iterates of our algorithm satisfy

$$\mu^{k+\frac{1}{2}} + \frac{1}{G(h^{k+\frac{1}{2}})}\|h^{k+\frac{1}{2}} - h^k\|^2 \leq \mu^k \,,$$

if we further assume $G(h^{k+\frac{1}{2}}) \neq 0$ for all $k \in \mathbb{N}$.

Main ingredients for proof:

- Write down first line of the algorithm (minimisation problem) for minimiser $h^{k+\frac{1}{2}}$, estimate upper bound inserting h^k
- Use convexity of G

Brief Convergence Analysis

Lemma

Let F and G be proper, lower semi-continuous and convex functions, and let G be differentiable with L-Lipschitz-continuous gradient, i.e. $\|\nabla G(h_1) - \nabla G(h_2)\|_2 \leq L\|h_1 - h_2\|_2$ for all h_1 and h_2 and a fixed constant L.

Then the iterates of our algorithm satisfy

$$\|r^{k+\frac{1}{2}} - \mu^{k+\frac{1}{2}} \nabla G(x^{k+\frac{1}{2}})\|_2 \leq (2 + C^{k+\frac{1}{2}} L)\|h^{k+\frac{1}{2}} - h^k\|_2 \,,$$

for some constant $C^{k+\frac{1}{2}}$, $r^{k+\frac{1}{2}} \in \partial F(h^{k+\frac{1}{2}})$ and $\mu^{k+\frac{1}{2}} := \mu^{k+1} = F(h^{k+\frac{1}{2}})/G(h^{k+\frac{1}{2}})$.

Brief Convergence Analysis

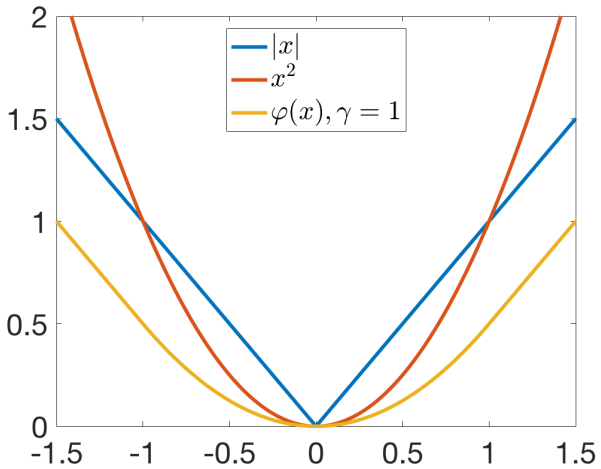
Note: Now we assume that G is smooth and has a Lipschitz-continuous gradient ∇G. Neither of those assumptions is fulfilled for the one-norm.

Brief Convergence Analysis

Note: Now we assume that G is smooth and has a Lipschitz-continuous gradient ∇G. Neither of those assumptions is fulfilled for the one-norm.

$\rightarrow$ Replace one-norm by Huber one-norm, i.e. replace modulus in one-norm by Huber function

$$\phi_\gamma(x) = \begin{cases} \frac{x^2}{2}, & |x| \leq \gamma \\ \gamma\left(|x| - \frac{\gamma}{2}\right), & |x| > \gamma \end{cases}.$$

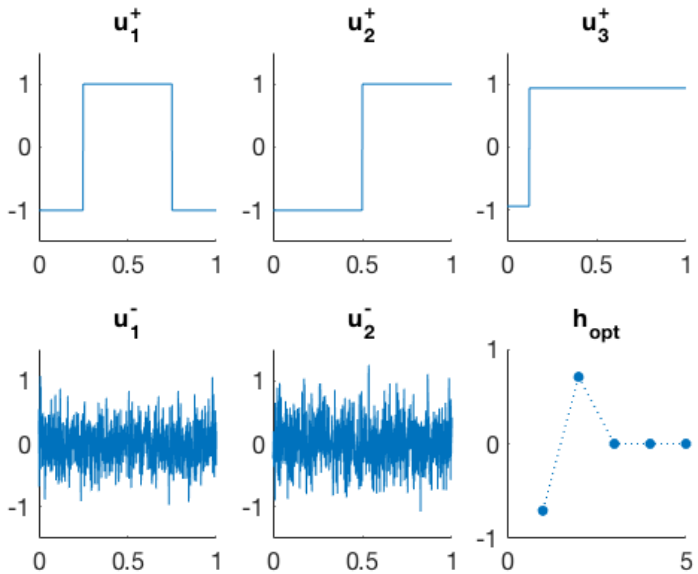


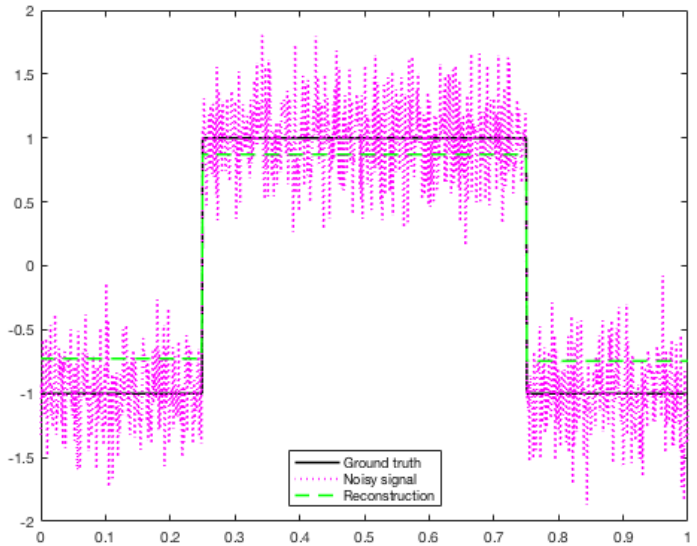
Brief Convergence Analysis

Main ingredients for proof:

- Write down optimality condition for inner minimisation problem in the first line of the algorithm
- Use differentiability of G
- Add zero
- Triangle inequality
- Exploit Lipschitz-continuity of G

1D Example: Multiple u^+ and u^- as Input

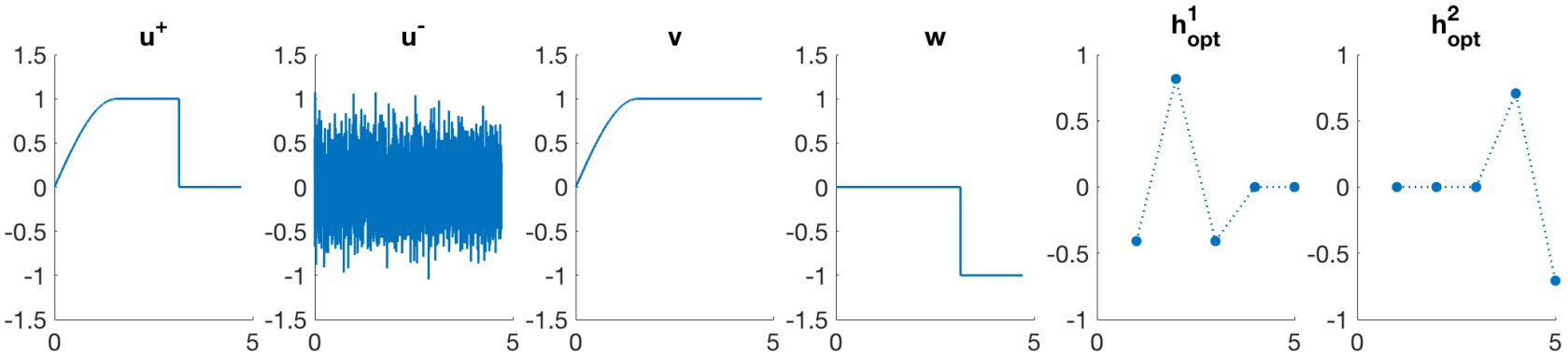




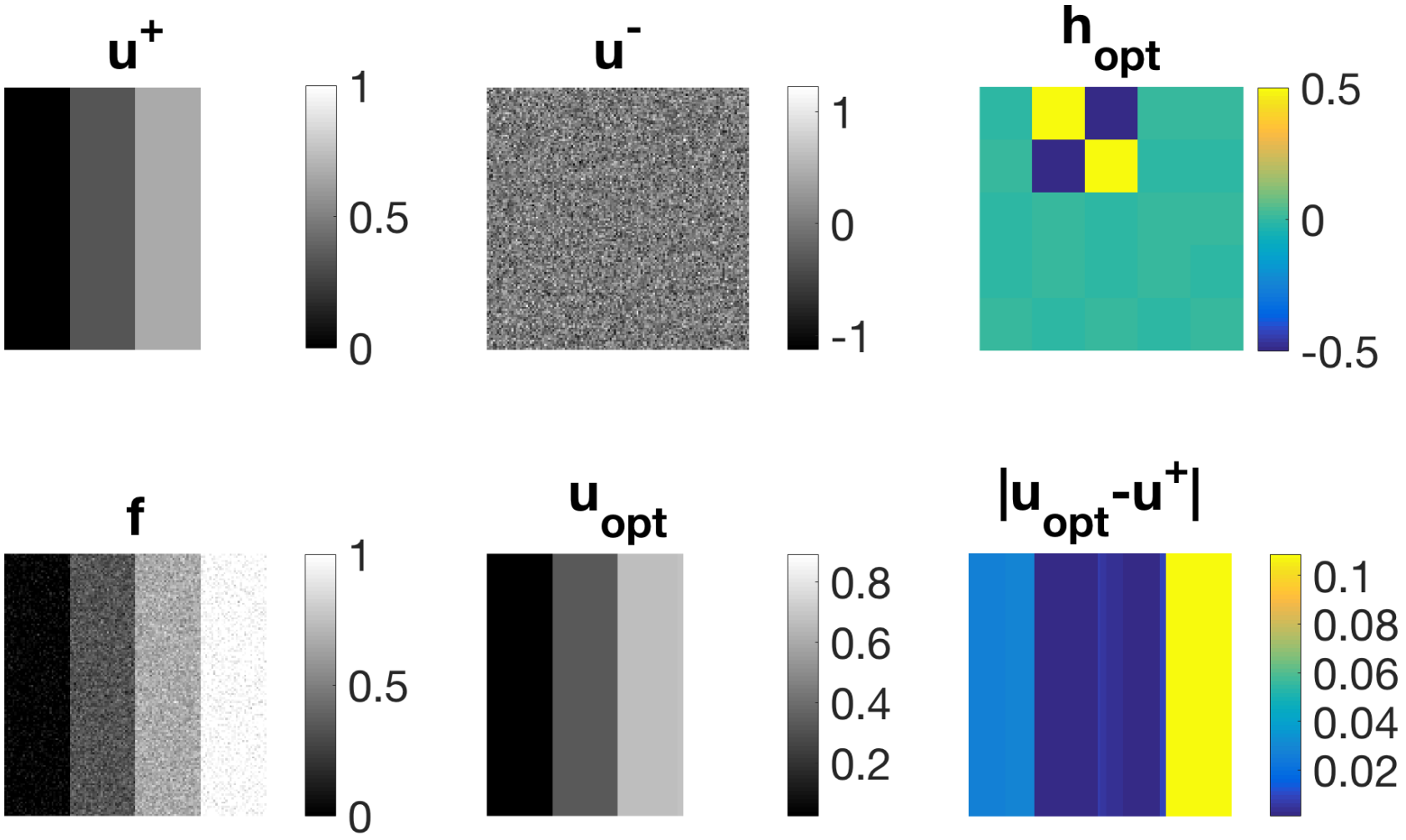
1D Example: TV-TV^2 Infimal Convolution

Given a known decomposition $u^+ = v + w$, i.e. u^+ consists of a smooth part v and a piecewise constant part w, we minimise with respect to h:

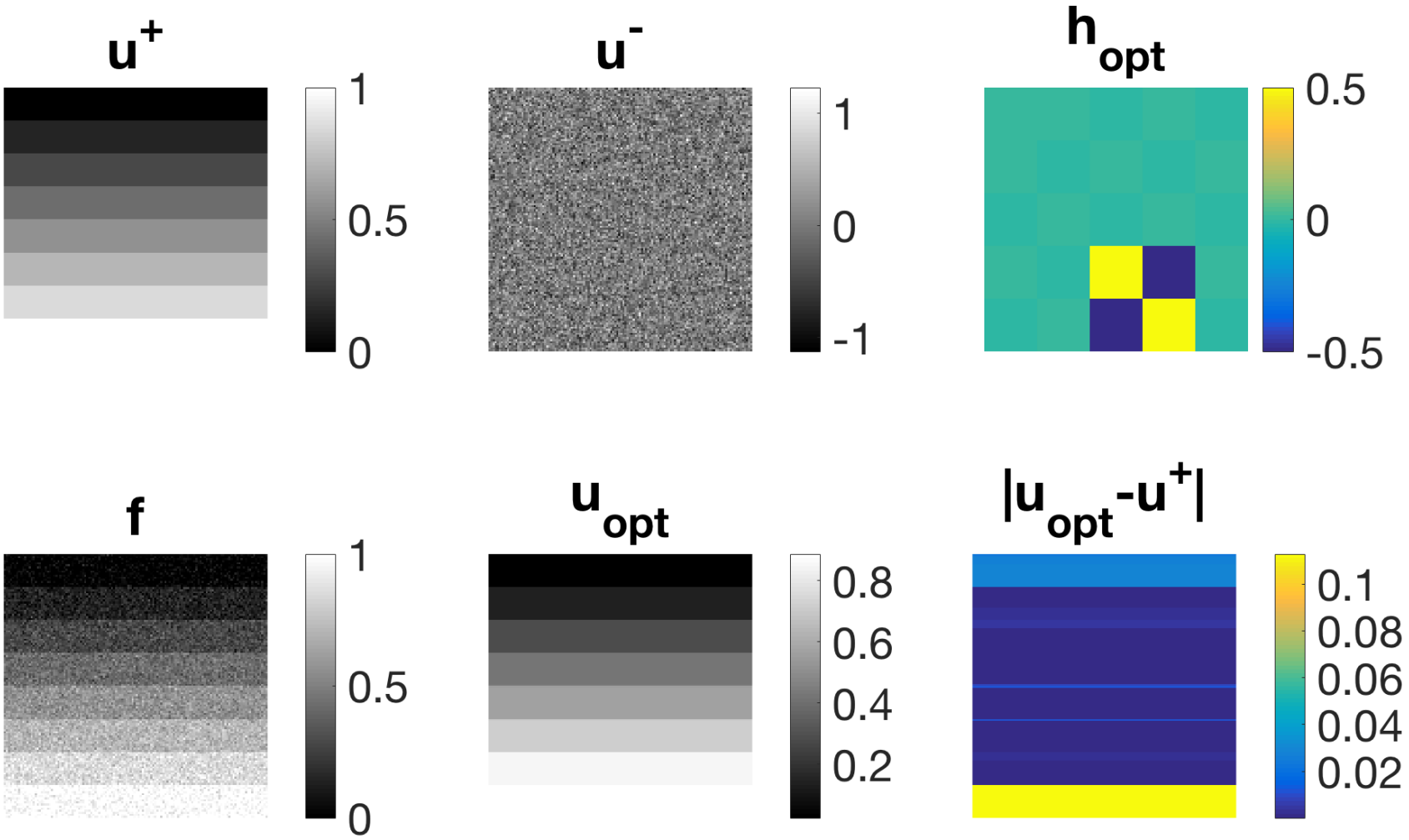
$$\frac{\|h_1 * v\|_1 + \|h_2 * (u^+ - v)\|_1}{\|h_1 * u^-\|_1 + \|h_2 * u^-\|_1}$$



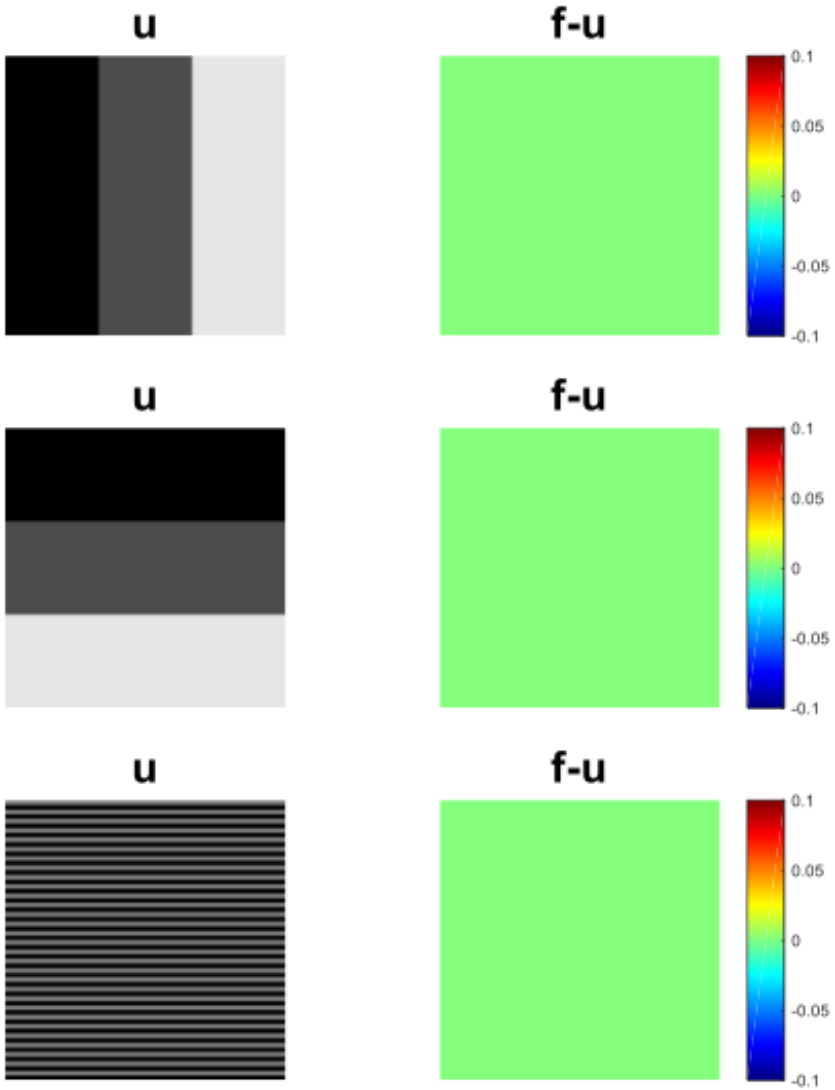
2D Example: Piecewise-constant images



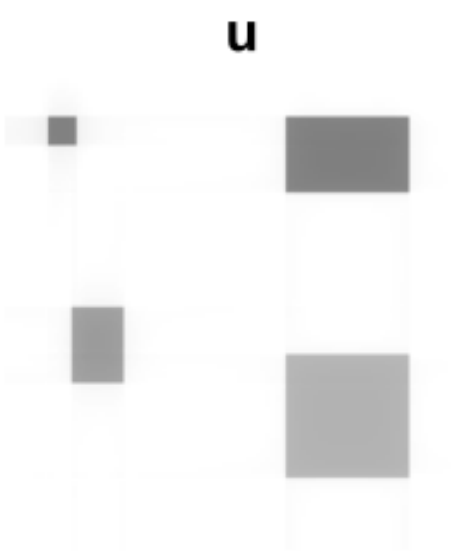
2D Example: Piecewise-constant images

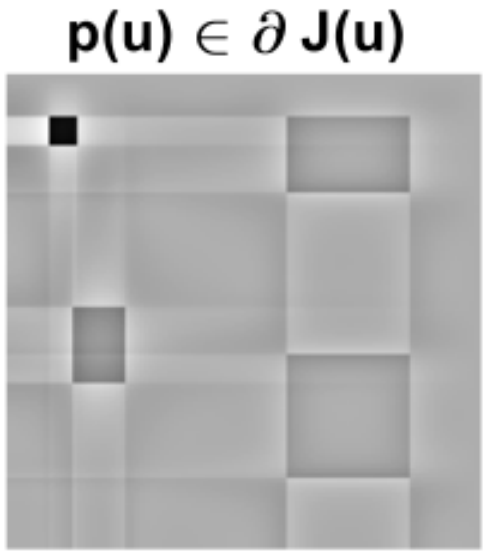


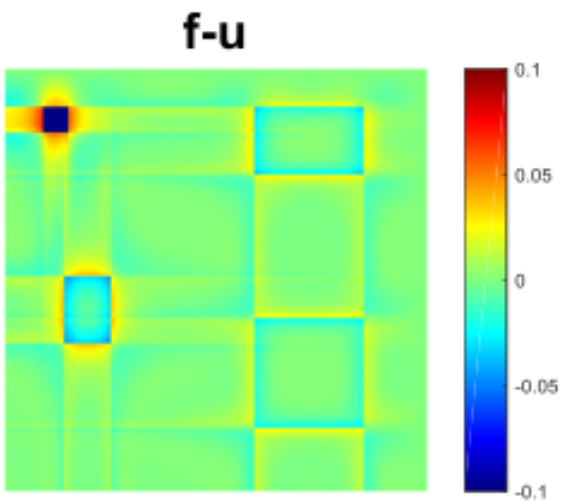
2D Example: Nullspace Property

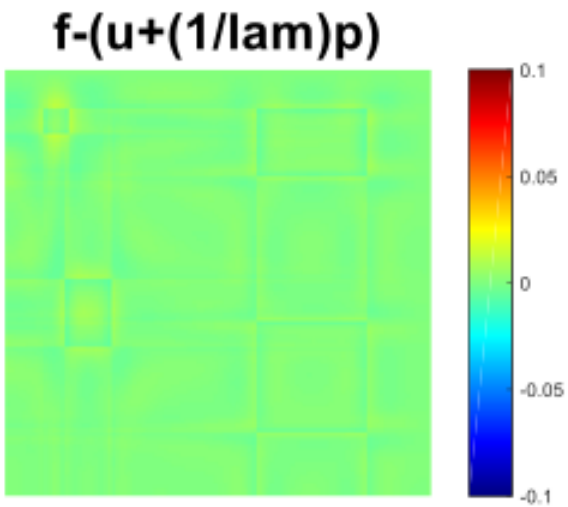


2D Example: Rectangle reconstruction

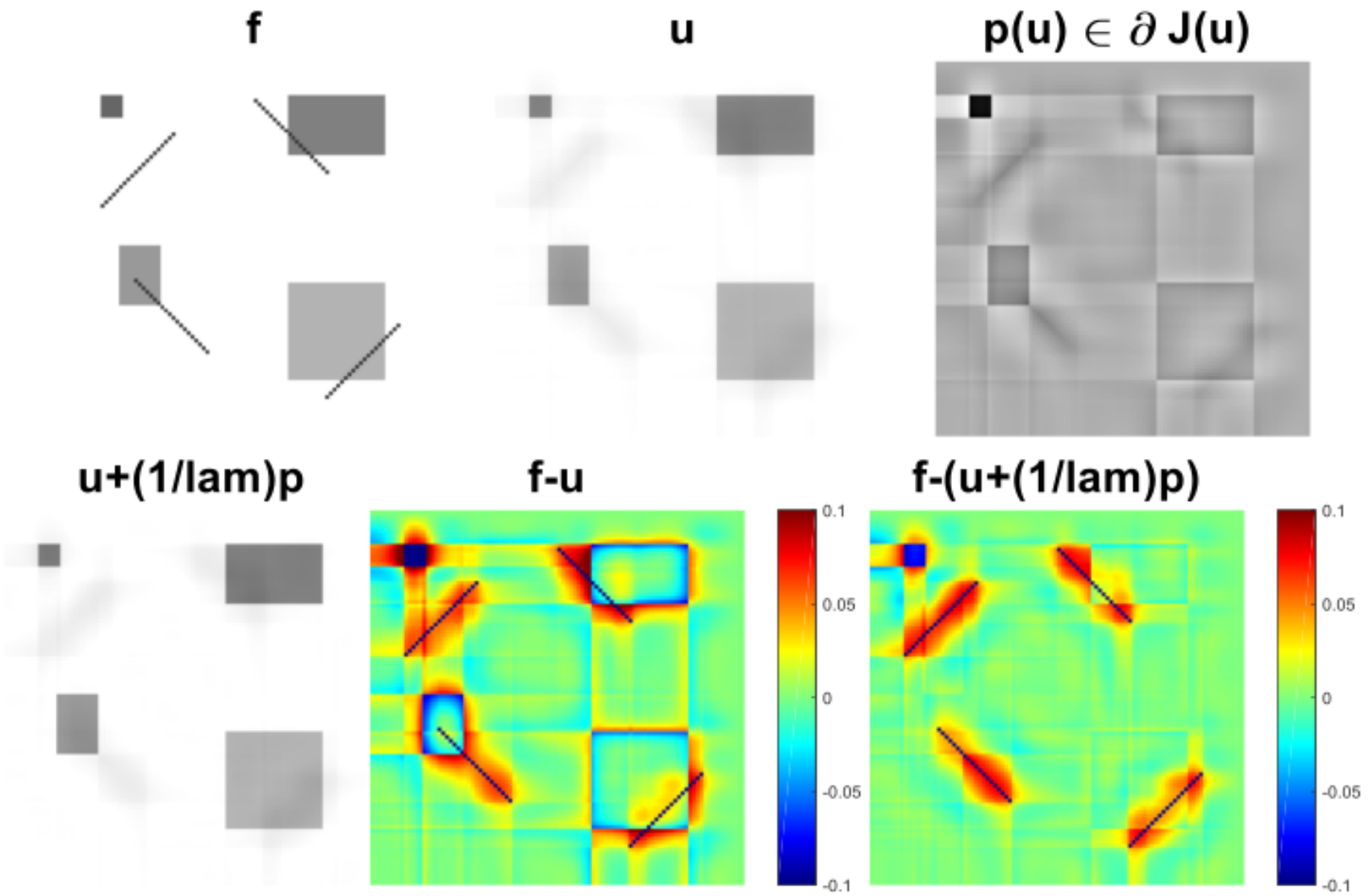




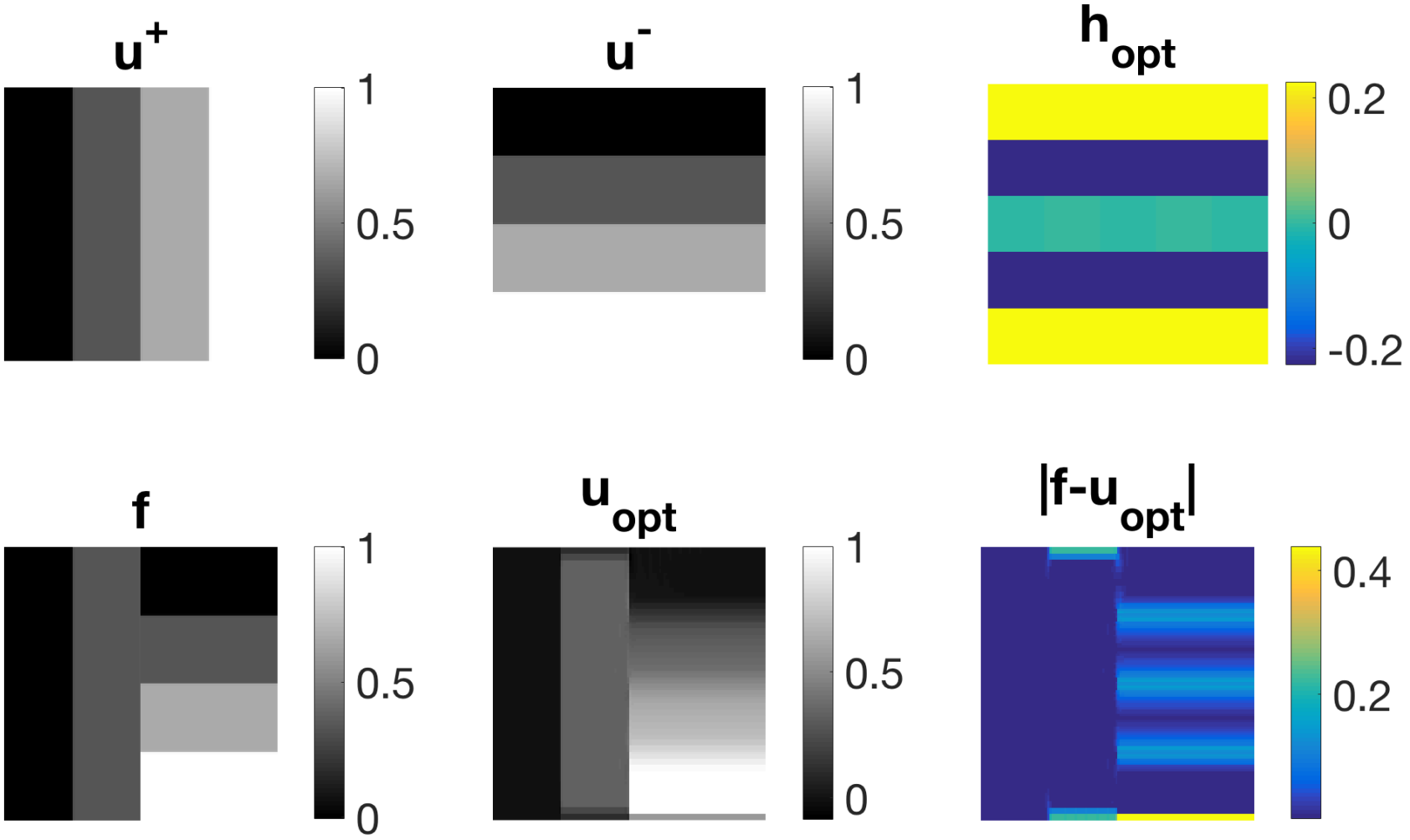




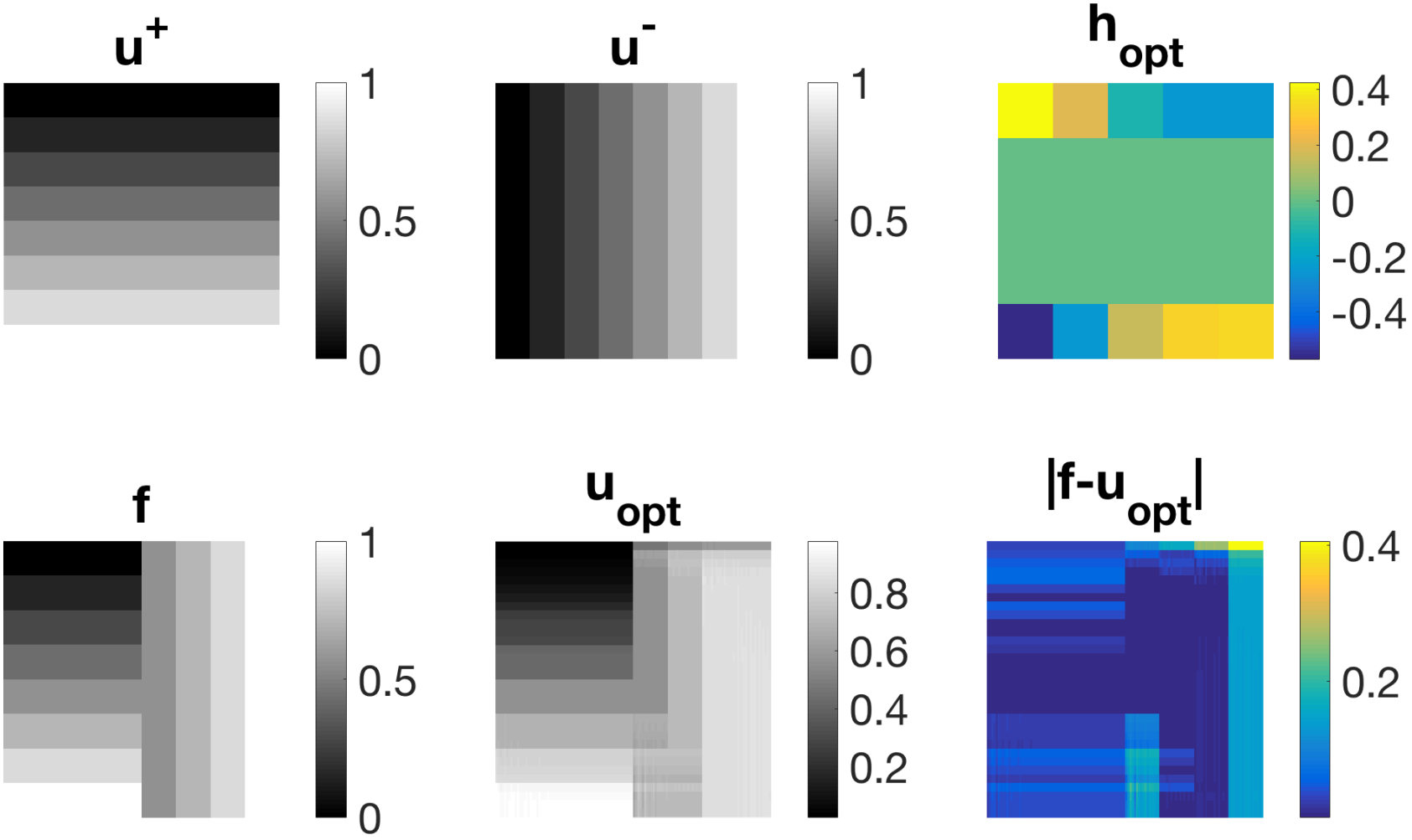
2D Example: Removing diagonals



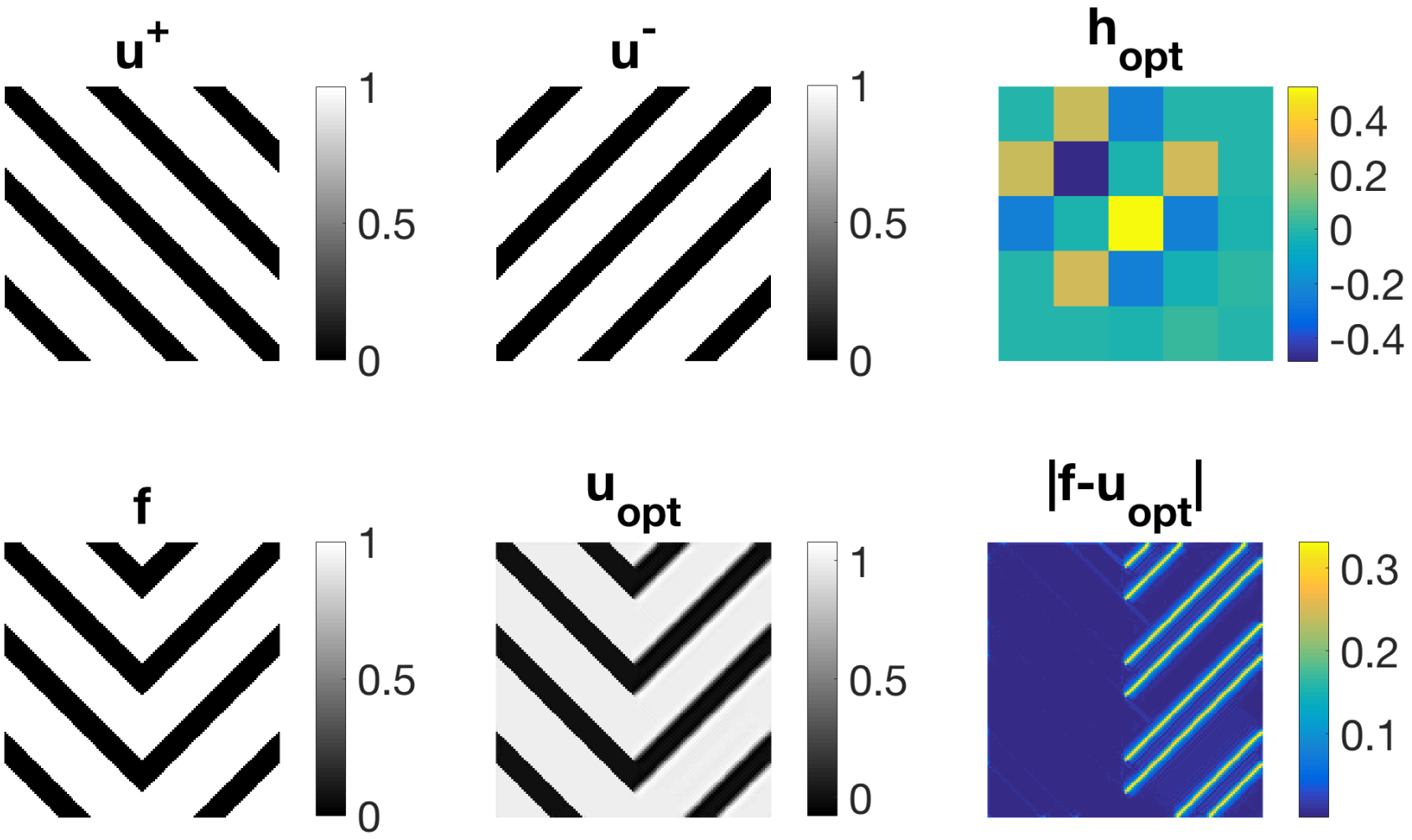
2D Examples: Distinguishing between shapes, angles, scales



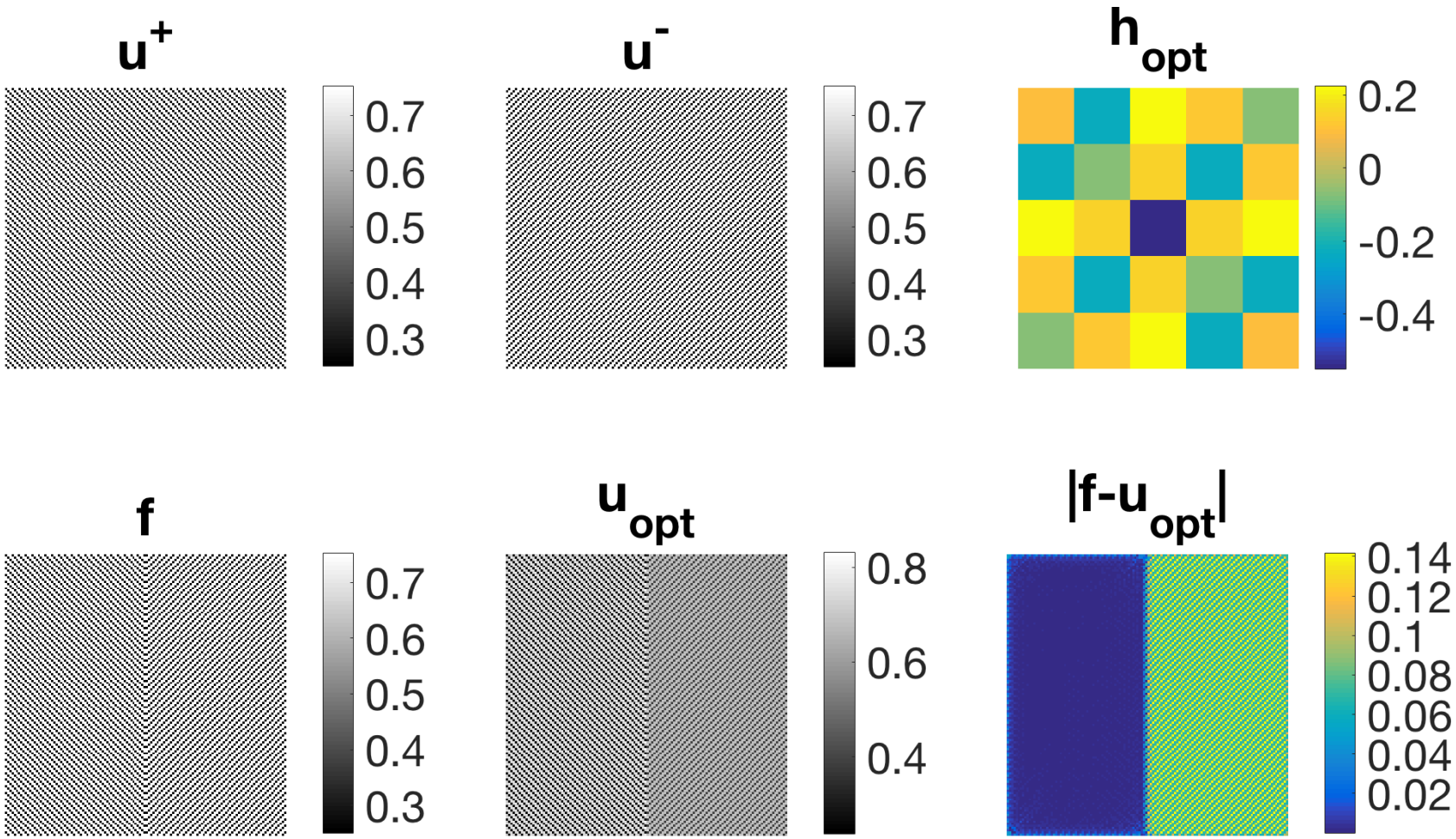
2D Examples: Distinguishing between shapes, angles, scales



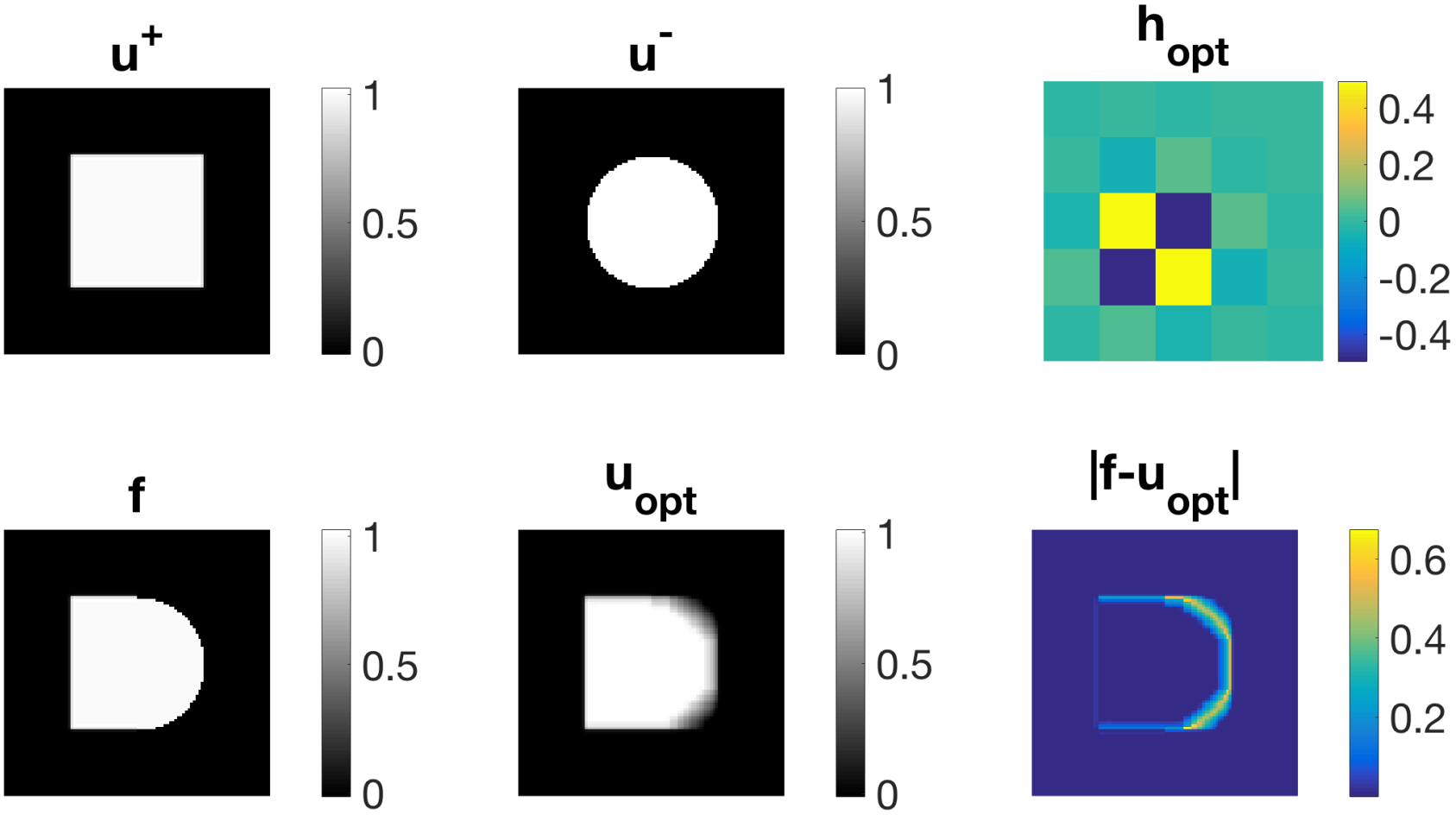
2D Examples: Distinguishing between shapes, angles, scales



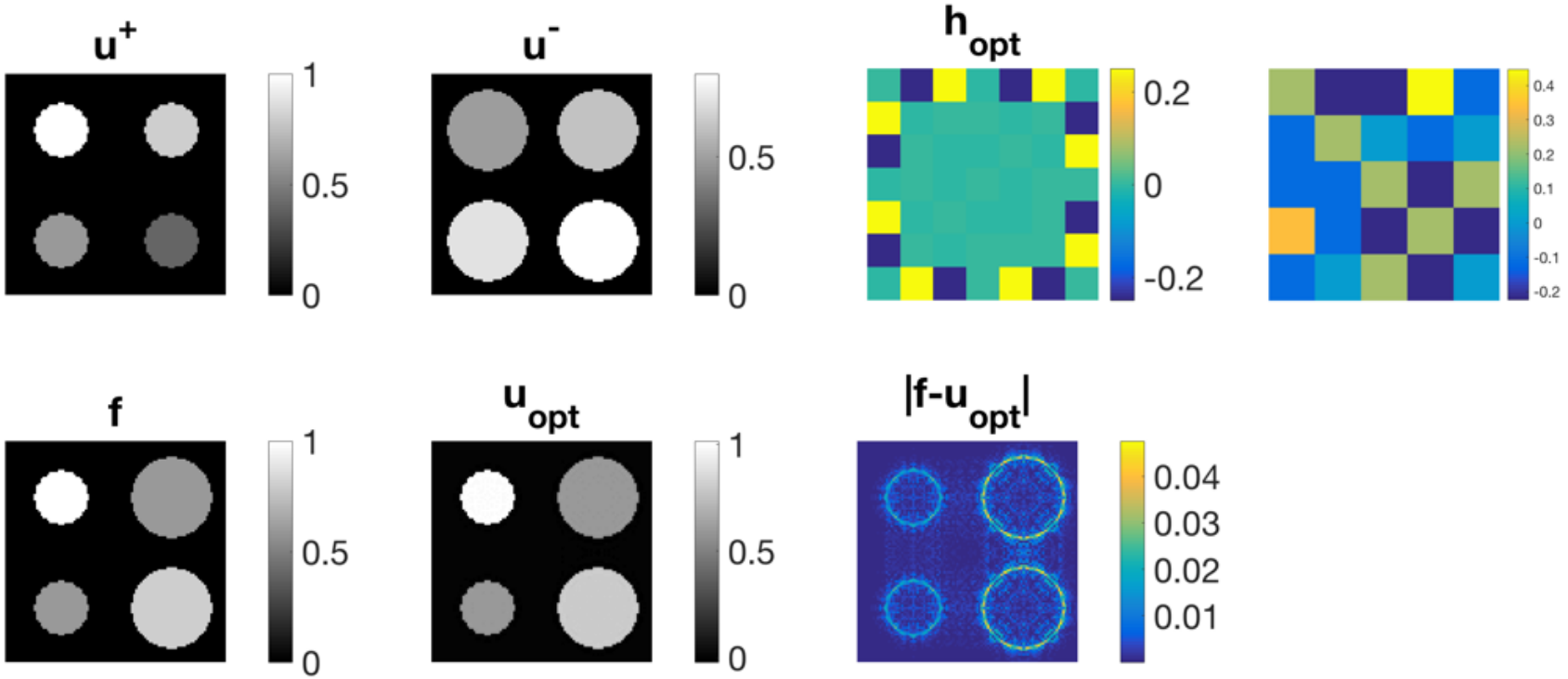
2D Examples: Distinguishing between shapes, angles, scales



2D Examples: Distinguishing between shapes, angles, scales

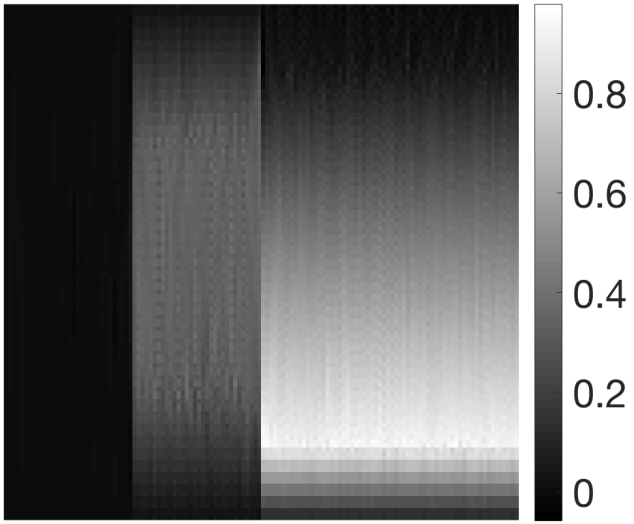


2D Examples: Distinguishing between shapes, angles, scales

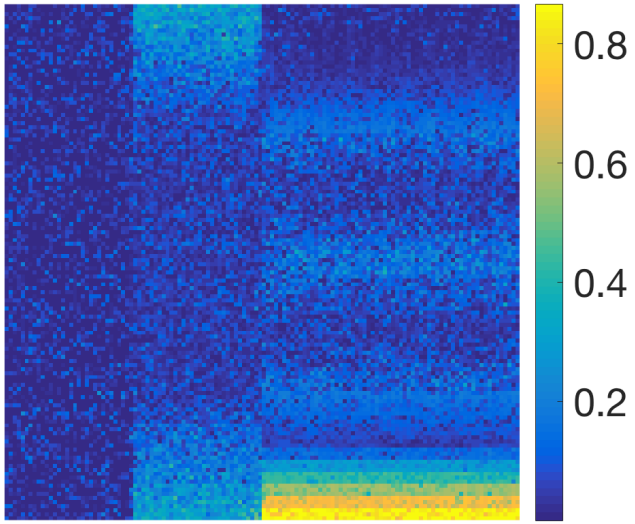


2D Examples: Denoising performance of filters

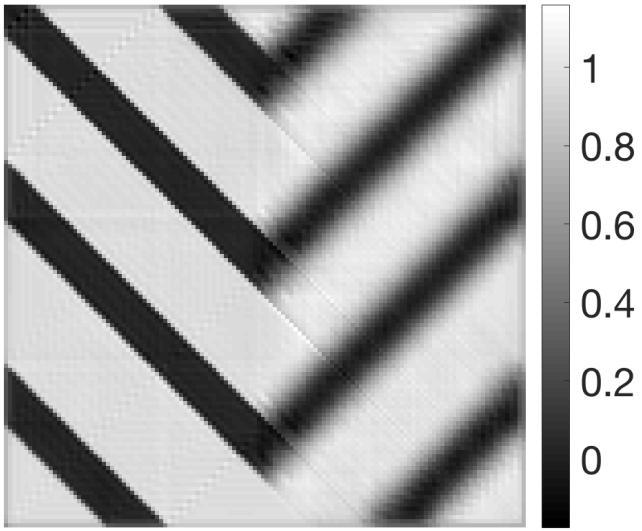
u_{opt}, $\eta = 2.5$, $\sigma^2 = 0.005$



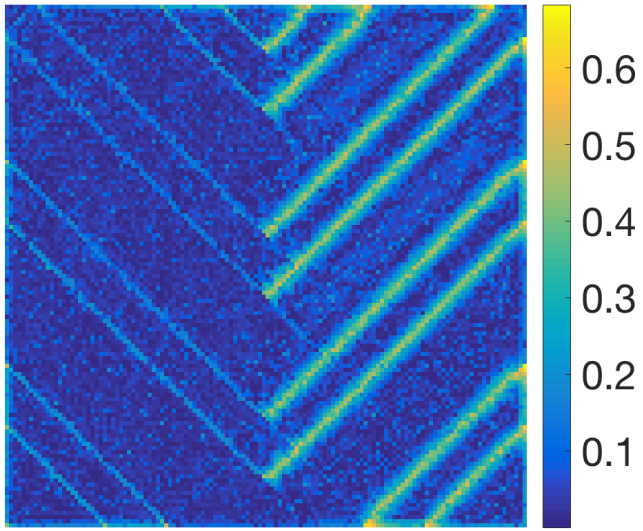
$|f-u_{opt}|$



u_{opt}, $\eta = 2$, $\sigma^2 = 0.005$



$|f-u_{opt}|$



Infimal Convolution Model

Learning

$$\min_{h_i} \frac{\|h_i * u_i\|_1}{\frac{1}{n-1}\sum_{j \neq i} \|h_i * u_j\|_1}, \quad i = 1, \ldots, n$$

Infimal Convolution Model

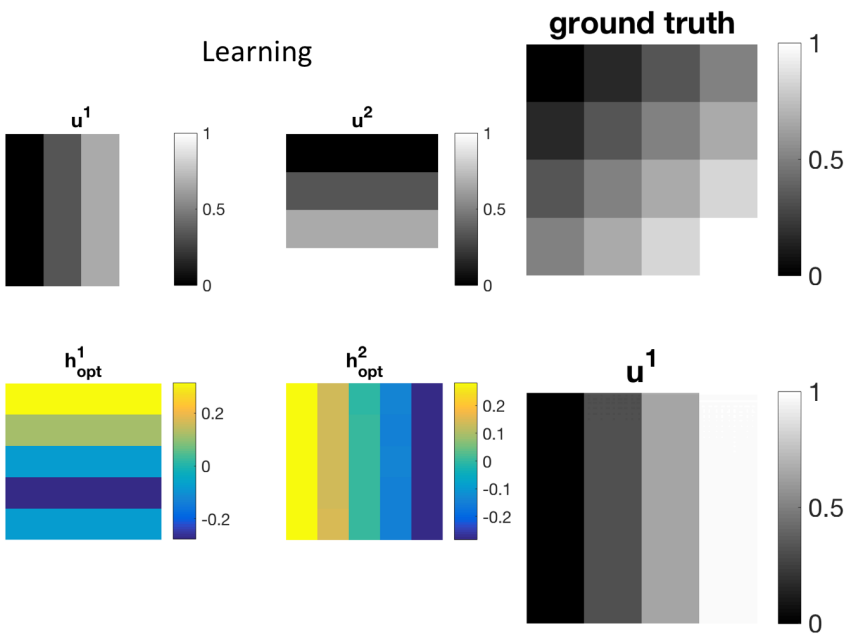
Learning

$$\min_{h_i} \frac{\|h_i * u_i\|_1}{\frac{1}{n-1}\sum_{j\neq i}\|h_i * u_j\|_1}, \quad i = 1, \dots, n$$

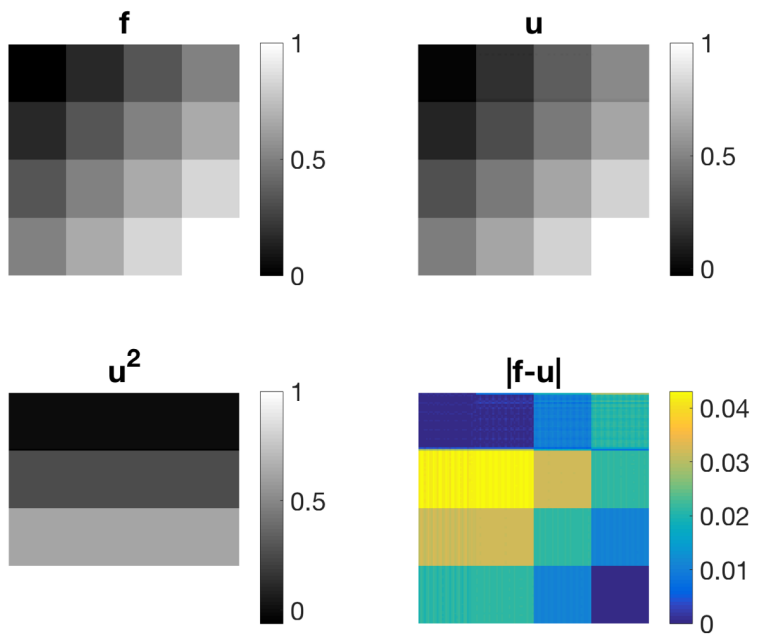
Reconstruction

$$\inf_{u=u_1+\cdots+u_n} \sum_{j=1}^{n} \|h_j * u_j\|_1$$

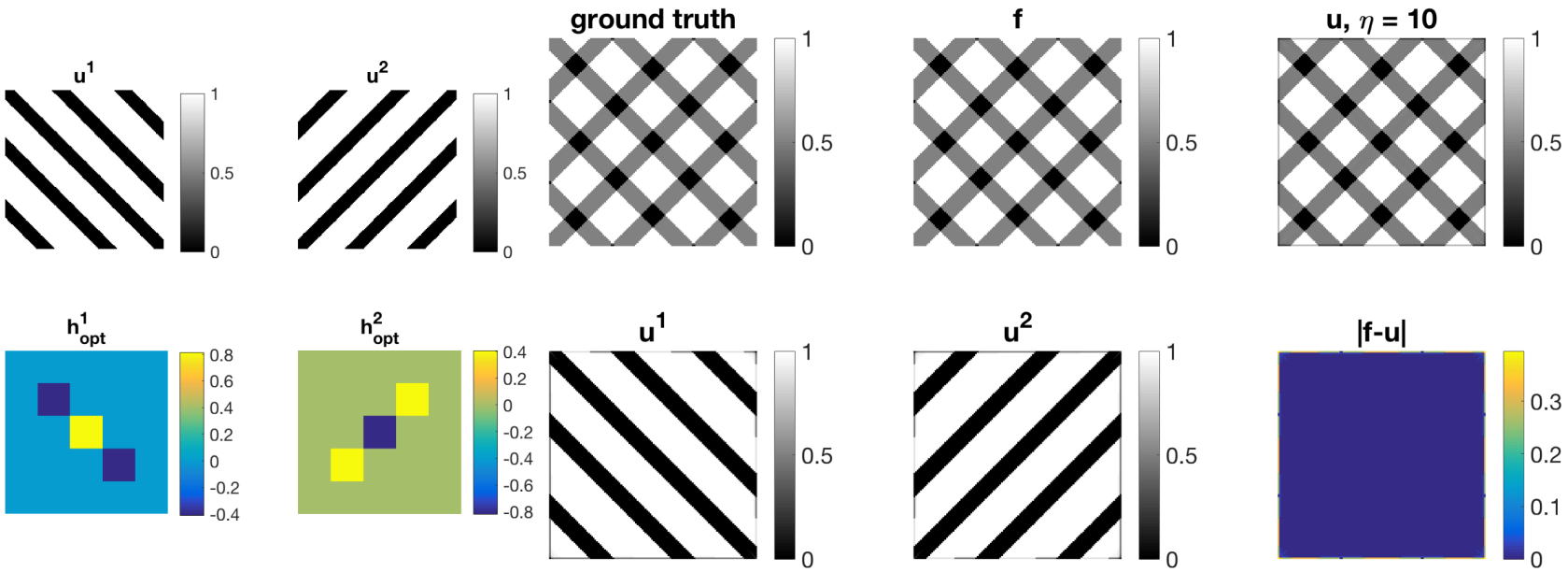
Additive test image



Reconstruction – without noise

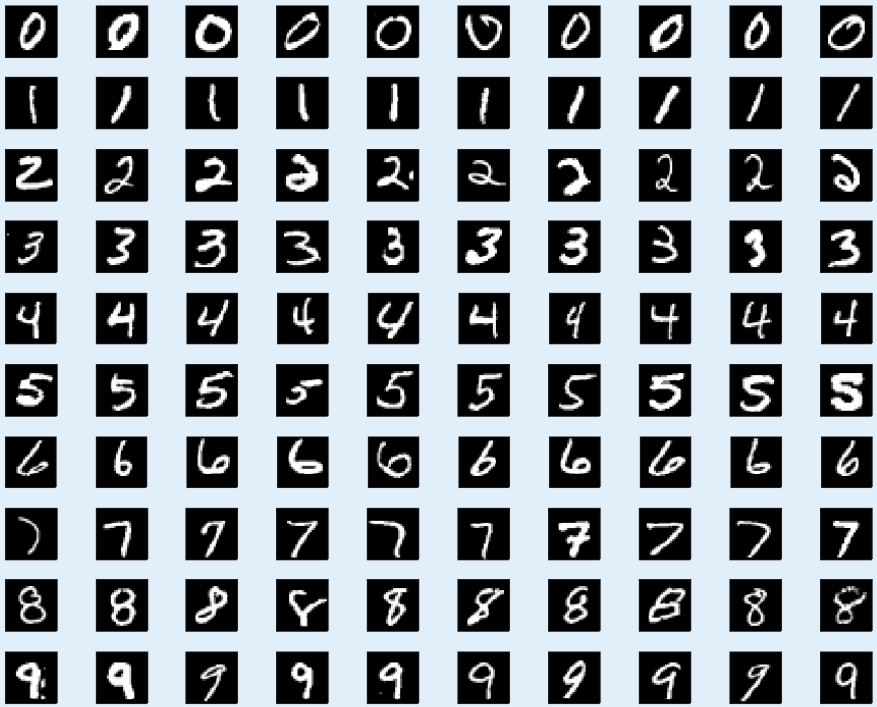


Additive test image: Diagonal stripes (noise-free)



MNIST classification

To conclude this talk we want to discuss the incorporation of misfit data in the context of a basic digit classification task



UNIVERSITY OF CAMBRIDGE

MNIST classification

Setup: we pick two digit classes, lets say 0 and 3, and learn the four filter functions

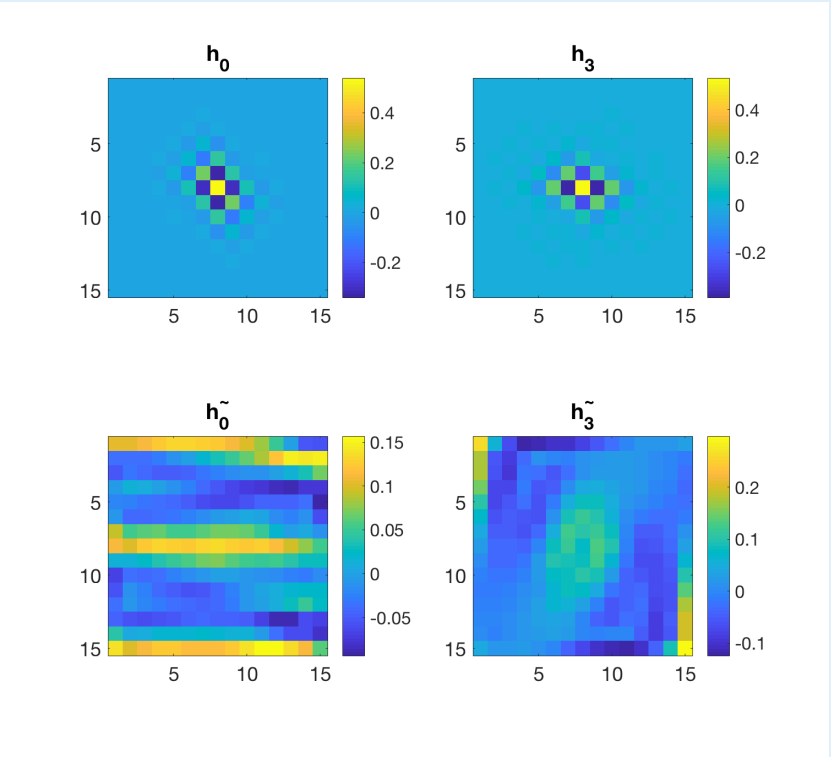
$$h_0 = \operatorname*{arg\,min}_{\|h\|_2=1} \|h * U_0\|_1 \qquad \text{and} \qquad h_3 = \operatorname*{arg\,min}_{\|h\|_2=1} \|h * U_3\|_1$$

as well as

$$\tilde{h}_0 = \operatorname*{arg\,min}_{\|h\|_2=1} \frac{\|h * U_0\|_1}{\|h * U_3\|_1} \qquad \text{and} \qquad \tilde{h}_3 = \operatorname*{arg\,min}_{\|h\|_2=1} \frac{\|h * U_3\|_1}{\|h * U_0\|_1}\,.$$

Here $U_0 \in \mathbb{R}^{784\times 5923}$ and $U_3 \in \mathbb{R}^{784\times 6131}$ are training-data matrices that contain images of hand-written zeros, respectively threes, only.

MNIST classification



Classification results:

$$\min\left(\|h_0 * u_{\text{test}}\|_1, \|h_3 * u_{\text{test}}\|_1\right)$$
$\rightarrow$ 77.7387 % success rate

$$\min\left(\|\tilde{h}_0 * u_{\text{test}}\|_1, \|\tilde{h}_3 * u_{\text{test}}\|_1\right)$$
$\rightarrow$ 95.1759 % success rate

Conclusions & Outlook

Conclusions

- Incorporation of misfit information into simple learning process
- Quotient minimisation yields generalised Eigenvalue problem
- Potential applications in signal decomposition and classification

Outlook

- Quotients of more sophisticated models
- Multiple classes of fit- and misfit-training-data
- More sophisticated applications

Thank you very much for your attention!

Are there any questions?

http://www.damtp.cam.ac.uk/research/cia/

cbs31@cam.ac.uk