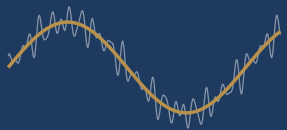
CNI Seminar Series

Plug-and-Play Denoising for Inverse Problems



From Fixed Points to Convergent Regularization

Subhadip Mukherjee

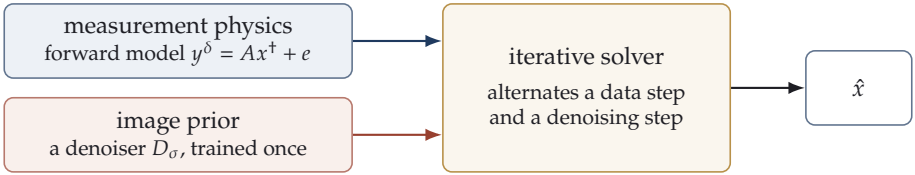
Department of Electronics & Electrical Communication Engineering, IIT Kharagpur

Centre for Networked Intelligence, Indian Institute of Science, Bengaluru · 2 November 2026

Joint work with H. Y. Tan, J. Tang, C.-B. Schönlieb and with A. Hauptmann, F. Sherry, C.-B. Schönlieb

Plug-and-play in one slide

Plug-and-play (PnP): measurement physics inside an iterative solver, with a powerful denoiser in place of a hand-crafted proximal prior



The mathematical challenge: what do the resulting iterations actually compute?

1. **Convergence**
   Do PnP iterations converge?
2. **An explicit energy**
   Is the limit a minimizer of an explicit energy?
3. **Speed**
   How fast can we get there?
4. **Regularization**
   Is it a convergent regularization as the noise vanishes?

The inverse problem

Noisy, indirect data:

$$y^\delta = Ax^\dagger + e, \qquad \|e\| \le \delta$$

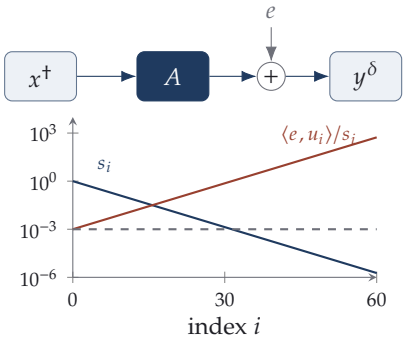
- A: blur, Radon transform, subsampled Fourier transform
- naive inversion is unstable; with the SVD $A = \sum_i s_i\, u_i v_i^\top$,

$$A^\dagger y^\delta - A^\dagger y^0 = \sum_i \frac{\langle e, u_i\rangle}{s_i}\, v_i$$

- small $s_i \Rightarrow$ noise amplified without bound

Variational regularization

$$\hat{x} \in \arg\min_x \underbrace{\tfrac{1}{2}\|Ax - y^\delta\|^2}_{\text{data consistency}} + \lambda \underbrace{R(x)}_{\text{prior}}$$



Smoothing operator: fast-decaying s_i (navy), noise level (dashed), exploding noise components $\langle e, u_i\rangle / s_i$ (rust)

The proximal viewpoint

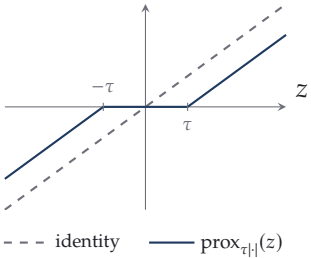
Proximal gradient (forward–backward splitting); f smooth, R possibly nonsmooth:

$$x^{k+1} = \mathrm{prox}_{\gamma\lambda R}\big(x^k - \gamma\nabla f(x^k)\big),$$

$$\mathrm{prox}_{\gamma\lambda R}(z) = \arg\min_x \ \tfrac{1}{2}\|x - z\|^2 + \gamma\lambda R(x)$$

physics step $z^k = x^k - \gamma\nabla f(x^k)$ → prior step $x^{k+1} = \mathrm{prox}_{\gamma\lambda R}(z^k)$

- prior step = a Gaussian denoiser
- MAP estimate of x from $z = x + n$, $n \sim \mathcal{N}(0, \gamma\mathrm{I})$, prior $p(x) \propto e^{-\lambda R(x)}$



$R = \|\cdot\|_1$: soft thresholding at $\tau = \gamma\lambda$; shrinks every coefficient, zeroes the small ones

Every proximal algorithm already contains a Gaussian denoiser: the entry point for PnP

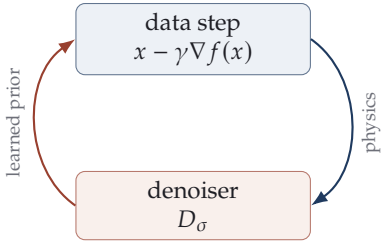
Plug-and-play: the one-line idea

Replace the proximal map by a denoiser:

$$x^{k+1} = D_\sigma\left(x^k - \gamma \nabla f(x^k)\right)$$

[Venkatakrishnan, Bouman, Wohlberg 2013]

- D_σ: BM3D, non-local means (NLM), CNNs such as DnCNN or DRUNet, proximal networks, diffusion denoisers
- one denoiser, reused across many forward models A
- often remarkably good, even when D_σ is not a proximal map



Physics only through ∇f, prior only through D_σ: one Gaussian denoiser, trained once, for every forward operator

> ***Core tension:*** *empirical power from strong black-box denoisers vs. guarantees that need structure*

The same move in other splittings

PnP-ADMM

$$x^{k+1} = \arg\min_x \; f(x) + \tfrac{\rho}{2}\|x - z^k + u^k\|^2$$
$$z^{k+1} = D_\sigma(x^{k+1} + u^k)$$
$$u^{k+1} = u^k + x^{k+1} - z^{k+1}$$

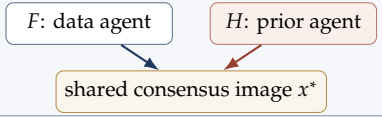
- D_σ in place of $\mathrm{prox}_{\lambda R/\rho}$
- data step: exploits the structure of A
- denoising step: any mature denoiser

[Sreehari et al. 2016; Chan, Wang, Elgendy 2017]

Consensus equilibrium

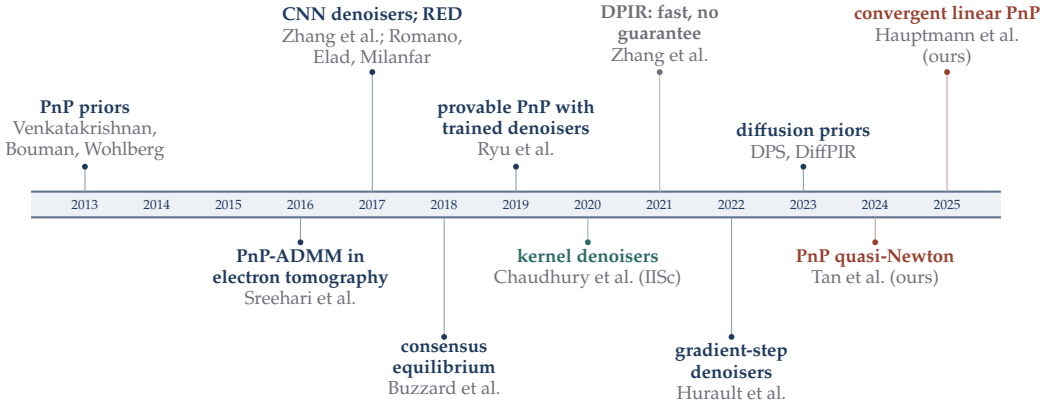
- data agent $F(z) = \arg\min_x f(x) + \frac{\rho}{2}\|x - z\|^2$
- prior agent $H = D_\sigma$
- equilibrium (x^*, u^*) [Buzzard et al. 2018]:

$$x^* = F(x^* + u^*) = H(x^* - u^*)$$



Same substitution in half-quadratic splitting and Douglas–Rachford; equilibria need no proximal D_σ, but are not yet a regularization method

A decade of plug-and-play



Fast-moving practice; theory catching up, one question at a time

Why PnP is attractive, and why it is subtle

For practitioners

- physics and learning decoupled
- one denoiser, many imaging modalities
- better denoiser $\Rightarrow$ better old algorithms
- no paired training data per inverse problem

For theorists

- a denoiser defines an implicit prior: which one?
- fixed-point map: a nonstandard optimization algorithm
- stability and regularization: subtle
- diffusion models: denoising $\leftrightarrow$ score-based Bayesian inference

> *From a brilliantly simple engineering trick to the search for the right mathematical object behind it*

Four questions about plug-and-play

1	**Do PnP iterations converge?**	$x^{k+1} = T(x^k) \;\longrightarrow\; x^\star \in \operatorname{Fix} T$	*an algorithmic statement*
2	**Is the limit a minimizer of an explicit energy?**	$x^\star \in \arg\min \Phi_y$, or at least a critical point of Φ_y	*a variational statement*
3	**How fast can we get there?**	few iterations, without giving up questions 1 and 2	*a computational statement* **our work: PnP quasi-Newton**
4	**Is it a convergent regularization?**	$x^\delta_{\alpha(\delta)} \to x^\dagger$ with $Ax^\dagger = y^0$, as the noise $\delta \to 0$	*an inverse-problem statement* **our work: linear PnP**

Converged iterate ≠ minimizer ≠ convergent regularization: each question needs its own structure on the denoiser and its own proof technique

What does each assumption on the denoiser buy?

Assumption on D_σ	What it buys	Price paid	Question
nonexpansive	convergence of the relaxed (Krasnosel'skii–Mann) iteration	may restrict expressivity	1
averaged	convergence of the plain PnP iteration	hard to impose on arbitrary CNNs	1
contractive	a unique fixed point and a linear rate	strong bias towards shrinkage	1, 3, 4
$D_\sigma = \mathrm{prox}_\phi$ with ϕ convex	an objective, with global minimizers	needs explicit proximal structure	2
$D_\sigma = \mathrm{prox}_\phi$ with ϕ weakly convex	critical points of a nonconvex energy	careful step sizes, a merit function	2, 3
linear and symmetric	spectral analysis and an explicit quadratic prior	a limited class of denoisers	4

Each theorem: not only a proof, but a lens on what a denoiser may be

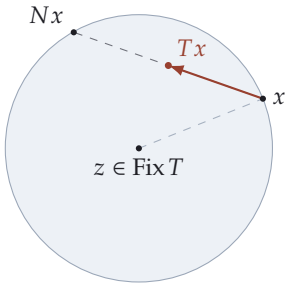
Averaged operators

Definition

N *nonexpansive*: $\|Nx - Ny\| \le \|x - y\|$ for all x, y
T *θ-averaged*, $\theta \in (0,1)$: $T = (1-\theta)\mathrm{I} + \theta N$, N nonexpansive

- nonexpansive is not enough: the reflection $N = -\mathrm{I}$ gives $x^{k+1} = -x^k$, oscillating forever
- averaging fixes it: $(1-\theta)\mathrm{I} + \theta N = (1-2\theta)\mathrm{I}$, so $x^k \to 0$
- key inequality [Bauschke, Combettes 2017]: for T θ-averaged and every $z \in \operatorname{Fix} T$,

$$\|Tx - z\|^2 \le \|x - z\|^2 - \tfrac{1-\theta}{\theta}\,\|x - Tx\|^2$$



Nx: anywhere in the ball around z, so no guaranteed progress. Tx ($\theta = \frac{1}{2}$): strictly inside, with a gain that grows with $\|x - Tx\|$

Averagedness: guaranteed progress towards every fixed point, at every step

A clean fixed-point theorem

Theorem (cf. Sun, Wohlberg, Kamilov 2019)

Assume f convex with L-Lipschitz gradient, $0 < \gamma < 2/L$, D_σ averaged, and a fixed point of $T = D_\sigma \circ (\mathrm{I} - \gamma \nabla f)$. Then the PnP iterates $x^{k+1} = Tx^k$ converge to a fixed point of T.

The proof in three steps

Step 1. gradient step $\mathrm{I} - \gamma \nabla f$ is averaged

cocoercivity of ∇f

→

Step 2. composition $T = D_\sigma \circ (\mathrm{I} - \gamma \nabla f)$ is averaged

Combettes–Yamada

→

Step 3. averaged + fixed point ⇒ Fejér monotone ⇒ convergent

Krasnosel'skii–Mann

No explicit regularizer needed (Step 1 uses only f, Step 2 only the averagedness of D_σ): its elegance, and its limitation

Step 1: the gradient step is averaged

(1) Cocoercivity [Baillon, Haddad 1977]

f convex, ∇f L-Lipschitz, $\Delta := \nabla f(x) - \nabla f(y) \Rightarrow$

$$\langle \Delta, x - y \rangle \geq \tfrac{1}{L}\|\Delta\|^2$$

(2) Expand the square for $G_\gamma = \mathrm{I} - \gamma \nabla f$:

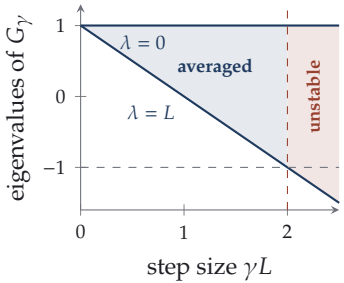
$G_\gamma x - G_\gamma y = (x - y) - \gamma \Delta$

$$\begin{aligned} \|G_\gamma x - G_\gamma y\|^2 &= \|x - y\|^2 - 2\gamma\langle \Delta, x - y\rangle + \gamma^2\|\Delta\|^2 \\ &\leq \|x - y\|^2 - \tfrac{2\gamma}{L}\|\Delta\|^2 + \gamma^2\|\Delta\|^2 \qquad \text{by (1)} \\ &= \|x - y\|^2 - \gamma\left(\tfrac{2}{L} - \gamma\right)\|\Delta\|^2 \end{aligned}$$

(3) Read off

- $\gamma = \frac{2}{L}$: last term vanishes $\Rightarrow N := \mathrm{I} - \frac{2}{L}\nabla f$ nonexpansive
- $\alpha := \frac{\gamma L}{2}$: $G_\gamma = (1 - \alpha)\mathrm{I} + \alpha N$, so α-averaged for $\gamma \in (0, \frac{2}{L})$

Step 1: only convexity and smoothness of the data term f



Quadratic f: eigenvalues $1 - \gamma\lambda$, $\lambda \in [0, L]$, fill the wedge; all in $(-1, 1]$ iff $\gamma L < 2$

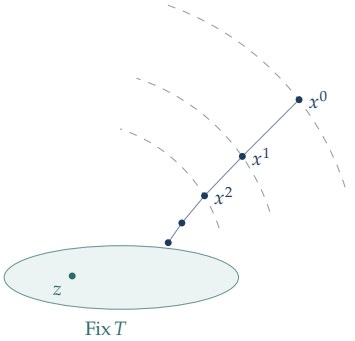
Steps 2 and 3: composition, then Fejér monotonicity

- averaged ∘ averaged is averaged [Combettes, Yamada 2015], so $T = D_\sigma \circ G_\gamma$ is θ-averaged
- key inequality, for every $z \in \operatorname{Fix} T$:

$$\|x^{k+1} - z\|^2 \le \|x^k - z\|^2 - \tfrac{1-\theta}{\theta}\,\|x^k - x^{k+1}\|^2$$

Consequences:

1. $\|x^k - z\|$ non-increasing $\Rightarrow$ bounded iterates
2. $\sum_k \|x^k - x^{k+1}\|^2 < \infty \Rightarrow$ residual $x^k - Tx^k \to 0$
3. cluster point $\bar{x} \in \operatorname{Fix} T$; take $z = \bar{x} \Rightarrow$ whole sequence $\to \bar{x}$



Fejér monotonicity: each iterate at least as close as the previous one to every fixed point z (dashed arcs)

Convergence to an equilibrium, not necessarily to a minimizer of any objective

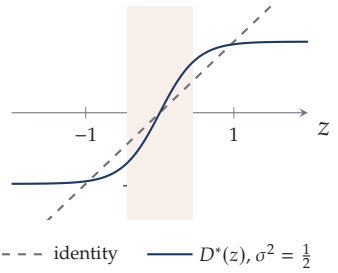
Which denoisers qualify, and at what cost?

Fixed-point assumptions, satisfied by design:

- averaged networks with constrained Lipschitz constants [Hertrich et al. 2021; Pesquet et al. 2021]
- small Lipschitz constant of $D_\sigma - \mathrm{I}$, strongly convex f [Ryu et al. 2019]
- kernel denoisers, e.g. NLM with frozen weights [Gavaskar, Chaudhury 2020]

The cost: the minimum mean-squared-error (MMSE) denoiser need not be nonexpansive

- $x = \pm 1$ equally likely, $z = x + \sigma\xi$: $D^*(z) = \mathbb{E}[x \mid z] = \tanh(z/\sigma^2)$
- slope $1/\sigma^2 > 1$ at $z = 0$ whenever $\sigma < 1$



Shaded: slope above 1. Ambiguous inputs pushed to the nearest mode, distances expanded

Constraints cost expressivity; the limit still has no variational meaning

When does PnP have an objective?

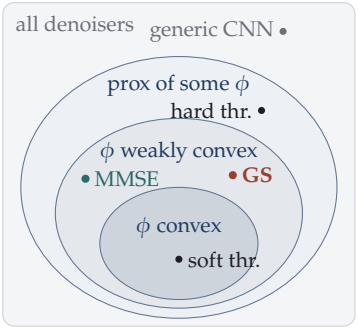
$D_\sigma = \text{prox}_{\gamma g} \Rightarrow$ PnP is proximal gradient for $\Phi = f + g$, with fixed points = critical points:

$$x = D_\sigma\big(x - \gamma \nabla f(x)\big) \iff 0 \in \nabla f(x) + \partial g(x)$$

Theorem (Moreau 1965; Gribonval, Nikolova 2020)

$D = \text{prox}_\phi$ for some ϕ $\iff$ $D(x) \in \partial\psi(x)$ for a convex ψ
ϕ convex possible $\iff$ in addition, D nonexpansive

- smooth proximal denoiser $\Rightarrow$ symmetric Jacobian
- generic CNN denoisers: not symmetric [Reehorst, Schniter 2019]



Soft / hard thresholding: proximal maps of ℓ^1 / ℓ^0. GS: gradient-step denoiser (next slide)

Can expressive networks be exact proximal maps?

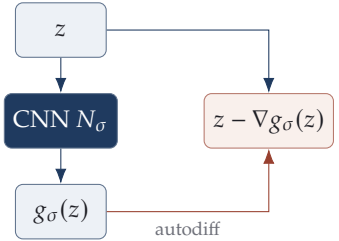
Gradient-step denoisers

Tweedie's formula ($z = x + \sigma\xi$, $\xi \sim \mathcal{N}(0, \mathrm{I})$, p_σ the density of z):

$$D_\sigma^*(z) = \mathbb{E}[x \mid z] = z + \sigma^2 \nabla \log p_\sigma(z)$$

- optimal denoiser = identity minus a gradient
- so parametrize exactly that, and train it as a denoiser [Hurault, Leclaire, Papadakis, ICLR 2022]:

$$D_\sigma = \mathrm{I} - \nabla g_\sigma, \quad g_\sigma(x) = \tfrac{1}{2}\|x - N_\sigma(x)\|^2, \quad N_\sigma\text{: a CNN}$$



Gradient structure built into the architecture, not hoped for

Proposition (Hurault, Leclaire, Papadakis, ICML 2022)

$g_\sigma \in C^2$, ∇g_σ L-Lipschitz with $L < 1 \Rightarrow D_\sigma = \mathrm{prox}_{\phi_\sigma}$, with $\phi_\sigma(x) = g_\sigma(D_\sigma^{-1}x) - \frac{1}{2}\|D_\sigma^{-1}x - x\|^2$ on the image of D_σ, and ϕ_σ is $\frac{L}{L+1}$-weakly convex

Prox-PnP: an energy that decreases

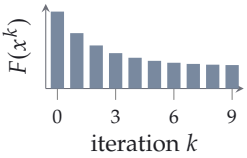
Theorem (Hurault, Leclaire, Papadakis, ICML 2022)

Assume ∇f L_f*-Lipschitz,* $\lambda L_f < 1$*, and* $D_\sigma = \operatorname{prox}_{\phi_\sigma}$ *with* ϕ_σ M*-weakly convex,* $M < 1$*. Then* $x^{k+1} = D_\sigma\big(x^k - \lambda \nabla f(x^k)\big)$ *decreases* $F = \lambda f + \phi_\sigma$*, with* $\min_{i \le k} \|x^{i+1} - x^i\|^2 = O(1/k)$*; bounded iterates converge to a critical point of* F *if* F *has the Kurdyka–Łojasiewicz (KŁ) property.*

Proof: sufficient decrease. prox subproblem $(1 - M)$-strongly convex: test it at x^k, add the descent lemma

$$F(x^{k+1}) \le F(x^k) - \frac{2 - M - \lambda L_f}{2} \|x^{k+1} - x^k\|^2$$

summing ⇒ rate; KŁ [Attouch, Bolte, Svaiter 2013], as a black box ⇒ the whole sequence converges □

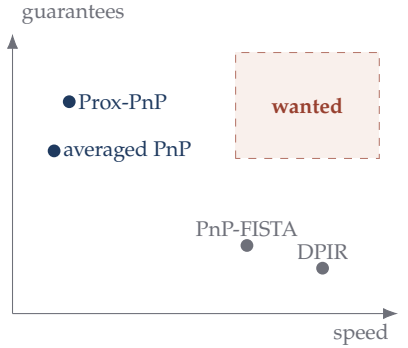


$F(x^k)$ *drops at every step*

Question 2 answered by sufficient decrease; the same pattern returns in our proof

Provable PnP is slow; fast PnP is unprovable

- **Provable PnP** (averaged denoisers, Prox-PnP): questions 1 and 2 answered, but typically hundreds to thousands of iterations
- **Fast PnP**, e.g. DPIR [Zhang et al. 2021]: state-of-the-art quality in a handful of iterations (decreasing σ schedule), but no guarantee; run it longer and the quality drops
- classical cure for slow first-order methods: curvature, as in **quasi-Newton** methods; combinable with a denoiser, without losing the guarantees?



Schematic: provable methods (navy) slow, fast methods (grey) unprovable; our first paper targets the corner

Goal: Newton-type speed with the energy and the guarantees of Prox-PnP

Our work: provably convergent PnP quasi-Newton

OUR WORK

H. Y. Tan, S. Mukherjee, J. Tang, C.-B. Schönlieb. *Provably convergent plug-and-play quasi-Newton methods.* SIAM Journal on Imaging Sciences, 2024.

Question: can quasi-Newton methods (curvature learned from the iterates) accelerate PnP without losing its guarantees?

Denoiser
gradient-step denoiser $D_\sigma = \mathrm{prox}_{\phi_\sigma}$, with ϕ_σ weakly convex

→

Merit function
forward–backward envelope φ_γ: smooth, computable from D_σ

→

Algorithm
quasi-Newton step on φ_γ with a line search, then one PnP step

Answer: yes

- convergence to critical points of the weakly convex energy $f + \frac{1}{\gamma}\phi_\sigma$
- 2–8× faster convergence than other provable PnP methods (deblurring, super-resolution)

Why Newton does not simply plug in

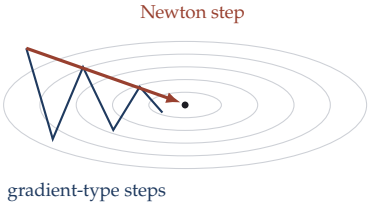
PnP fixed points = zeros of the residual

$$R_\gamma(x) = \tfrac{1}{\gamma}\big(x - D_\sigma(x - \gamma\nabla f(x))\big)$$

- R_γ plays the role of a gradient; Newton would precondition it with curvature
- but proximal Newton [Lee, Sun, Saunders 2014] needs the *scaled* proximal map

$$\operatorname{prox}_g^B(x) = \arg\min_u \; g(u) + \tfrac{1}{2}\|u - x\|_B^2$$

- metric $B \approx \nabla^2 f(x^k)$: a new one at every iteration
- a trained denoiser knows only the Euclidean metric; retraining for every B is out of the question



Ill-conditioned problem: gradient-type steps zigzag across the valley; a curvature-aware step goes straight to the minimizer

Needed: a Newton-type method that touches the prior only through the denoiser D_σ

The forward–backward envelope

Definition (Patrinos, Bemporad 2013)

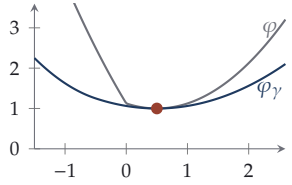
For $\varphi = f + g$ and a step $\gamma > 0$:

$$\varphi_\gamma(x) = f(x) - \tfrac{\gamma}{2}\|\nabla f(x)\|^2 + g^\gamma\big(x - \gamma\nabla f(x)\big),$$

$$g^\gamma(y) = \min_u \; g(u) + \tfrac{1}{2\gamma}\|u - y\|^2 \quad \text{(Moreau envelope)}$$

For small γ [Stella, Themelis, Patrinos 2017]:

- $\varphi_\gamma \le \varphi$, equal at critical points, same minimizers
- φ_γ differentiable, $\nabla\varphi_\gamma = (\mathrm{I} - \gamma\nabla^2 f)\, R_\gamma$
- R_γ: the PnP residual (previous slide), with $D_\sigma = \mathrm{prox}_{\gamma g}$



$\varphi(x) = \frac{1}{2}(x - 1.5)^2 + |x|$ *(grey) and its envelope for* $\gamma = \frac{1}{2}$ *(navy): smooth, below* φ*, equal at the minimizer*

A smooth surrogate, stationary exactly at the PnP fixed points: quasi-Newton can run on it

For a gradient-step denoiser, the envelope is free

OUR WORK

Key observation (Tan, M., Tang, Schönlieb 2024)

$D_\sigma = \mathrm{I} - \nabla g_\sigma = \mathrm{prox}_{\phi_\sigma}$, $g = \frac{1}{\gamma}\phi_\sigma \Rightarrow \varphi = f + \frac{1}{\gamma}\phi_\sigma$, T_γ = exactly the PnP step; Moreau envelope of g = the network potential itself:

$$g^\gamma = \tfrac{1}{\gamma}\, g_\sigma \qquad \Longrightarrow \qquad \varphi_\gamma(x) = f(x) - \tfrac{\gamma}{2}\|\nabla f(x)\|^2 + \tfrac{1}{\gamma}\, g_\sigma\big(x - \gamma \nabla f(x)\big)$$

Why. proximal point of y: $D_\sigma y$; the formula for ϕ_σ gives $\phi_\sigma(D_\sigma y) = g_\sigma(y) - \frac{1}{2}\|y - D_\sigma y\|^2$; inserted into $g^\gamma(y) = \frac{1}{\gamma}\left[\phi_\sigma(D_\sigma y) + \frac{1}{2}\|D_\sigma y - y\|^2\right]$, the quadratic terms cancel □

What this buys

- one forward and one backward pass through N_σ: both $\varphi_\gamma(x)$ and $\nabla\varphi_\gamma(x)$
- no inverse and no Jacobian of D_σ
- no scaled proximal map

Merit function and gradient for the cost of one PnP step

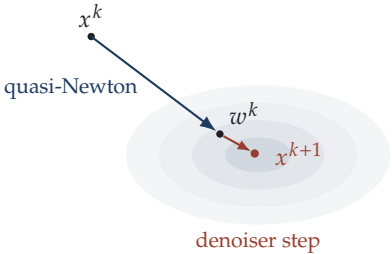
The algorithm: PnP-LBFGS

OUR WORK

PnP-LBFGS (Tan, M., Tang, Schönlieb 2024)

For $k = 0, 1, 2, \ldots$

1. *Direction:* $d^k = -H_k \nabla \varphi_\gamma(x^k)$, with H_k the limited-memory BFGS inverse-Hessian estimate
2. *Line search:* $w^k = x^k + \tau_k d^k$, $\tau_k \in [0, 1]$ (try $\tau_k = 1$ first), with $\varphi_\gamma(w^k) \leq \varphi_\gamma(x^k)$
3. *PnP step:* $x^{k+1} = D_\sigma\big(w^k - \gamma \nabla f(w^k)\big)$



Level sets of φ_γ: the quasi-Newton step (navy) moves far, the denoiser step (rust) keeps the theory intact

- a PnP version of MINFBE [Stella et al. 2017], for weakly convex g
- late iterations: typically $\tau_k = 1$

Any envelope-decreasing direction allowed; but why does φ itself still decrease?

Proof idea: a sandwich

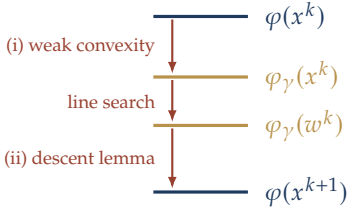
OUR WORK

Proposition (Tan et al. 2024, Prop. 2.6)

∇f L_f-Lipschitz, g M-weakly convex $\Rightarrow$ for all x and $\gamma > 0$:

(i) $\varphi_\gamma(x) \le \varphi(x) - \frac{\gamma - M\gamma^2}{2}\|R_\gamma(x)\|^2$

(ii) $\varphi(T_\gamma x) \le \varphi_\gamma(x) - \frac{\gamma}{2}(1 - \gamma L_f)\|R_\gamma(x)\|^2$



One iteration, top to bottom: one inequality per arrow

Chain (i), the line search and (ii) along $x^k \to w^k \to x^{k+1}$; telescope:

$$\min_{i \le k} \|R_\gamma(x^i)\|^2 \le \frac{2\big(\varphi(x^0) - \inf \varphi\big)}{(k+1)(\gamma - M\gamma^2)}$$

Sufficient decrease again, as for Prox-PnP; with KŁ, the whole sequence converges

Main result I: PnP quasi-Newton converges

OUR WORK

Theorem (Tan, M., Tang, Schönlieb, SIIMS 2024, Cor. 3.2)

Assume $f \in C^1$ *KŁ with* L_f*-Lipschitz gradient;* $g_\sigma \in C^2$ *semialgebraic with* L_g*-Lipschitz gradient,* $L_g < 1$ *(so* ϕ_σ *is* M*-weakly convex,* $M = \frac{L_g}{L_g+1}$*);* $\gamma < \frac{1}{2M}$ *as in the paper,* $\tau_k \le \bar{\tau}$*,* $\|d^k\| \le c\,\|R_\gamma(x^k)\|$*. Then, for* $\varphi = f + \frac{1}{\gamma}\phi_\sigma$*:*

1. $\varphi(x^k)$ *decreases monotonically*
2. $\min_{i \le k}\|R_\gamma(x^i)\| = O(1/\sqrt{k})$
3. *bounded iterates converge to a critical point of* φ*, at which* $\varphi = \varphi_\gamma$
4. *under the Dennis–Moré condition: superlinear convergence*

- the energy and guarantees of Prox-PnP, plus a Newton-type mechanism
- experiments: 2–8× faster convergence than other provable PnP methods

Acceleration and guarantees can coexist

Why fixed-point convergence is not enough

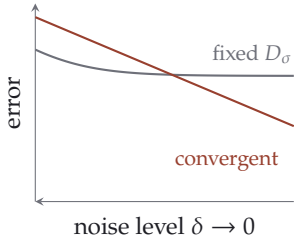
A *family* of reconstruction maps $\mathcal{R}_\alpha : y^\delta \mapsto x_\alpha^\delta$, $\alpha > 0$

Definition

$(\mathcal{R}_\alpha)_{\alpha>0}$ is a *convergent regularization* if each $\mathcal{R}_\alpha$ is continuous and, for a parameter rule $\alpha = \alpha(\delta)$, whenever $\|y^\delta - y^0\| \le \delta \to 0$:

$$\mathcal{R}_{\alpha(\delta)}(y^\delta) \;\to\; x^\dagger, \quad \text{a solution of } Ax = y^0$$

- questions 1–3: one y; question 4: a family $y^\delta \to y^0$
- fixed denoiser: good at one noise level, possibly not convergent [Ebner, Haltmeier, IEEE TIP 2024]



Schematic: with a fixed denoiser the error stalls at a bias floor; a convergent method drives it to zero

Question 4 needs a knob that weakens the denoiser as $\delta \to 0$

The classical template: Tikhonov-type regularization

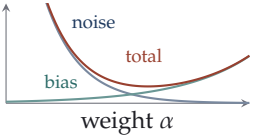
Theorem (e.g. Scherzer et al. 2009)

Assume $R \geq 0$ lower semicontinuous with compact sublevel sets, $x_\alpha^\delta \in \arg\min_x \frac{1}{2}\|Ax - y^\delta\|^2 + \alpha R(x)$, and $\alpha(\delta) \to 0$ with $\delta^2/\alpha(\delta) \to 0$. Then $x_{\alpha(\delta)}^\delta$ converges, along subsequences, to R-minimizing solutions of $Ax = y^0$.

Proof idea. compare with an R-minimizing $x^\dagger$:

$$\tfrac{1}{2}\|Ax_\alpha^\delta - y^\delta\|^2 + \alpha R(x_\alpha^\delta) \leq \tfrac{\delta^2}{2} + \alpha R(x^\dagger)$$

$\Rightarrow$ residual $\to 0$, $R(x_\alpha^\delta)$ bounded; compactness $\Rightarrow$ convergent subsequence, R-minimizing limit $\square$



Total error smallest at an intermediate α

Knob = the weight α; where is the knob in a learned denoiser?

Convergent data-driven methods before our work

Learned regularizers

- null-space networks [Schwab et al. 2019]
- NETT [Li et al. 2020]
- learned convex regularizers [Mukherjee et al. 2020]
- weakly convex regularizers [Shumaylov et al. 2024]
- a survey [Mukherjee et al. 2023]

Plug-and-play

- suitable family of contractive denoisers + parameter rule: a convergent regularization [Ebner, Haltmeier, IEEE TIP 2024]
- no recipe for *how* a given trained denoiser should be weakened as $\delta \to 0$

- learned regularizers: explicit regularizer, and its weight is the knob
- PnP: implicit regularizer, so the knob has to be built into the denoiser

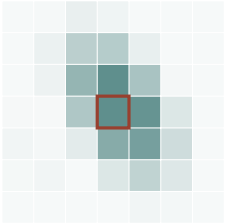
> *Natural starting point: linear denoisers, whose fixed-noise theory is already complete*

Kernel denoisers: a complete fixed-noise theory

Kernel denoisers (NLM, bilateral filter) with frozen weights are linear:

$$W = D^{-1}K, \quad K \text{ symmetric, nonnegative}, \quad D = \operatorname{diag}(K\mathbf{1})$$

- W similar to a symmetric matrix: a proximal map in the D-weighted inner product, so PnP solves a convex problem [Gavaskar, Athalye, Chaudhury 2021; Nair, Gavaskar, Chaudhury 2021]
- mild conditions on A: contractive PnP operator, linear convergence [Athalye, Chaudhury, Kumar 2023; Sinha et al. 2025]
- induced problem strongly convex [Sinha, Chaudhury 2024]



NLM weights K_{ij} around one pixel (outlined): large where the surrounding patches look alike; frozen weights make the denoiser linear

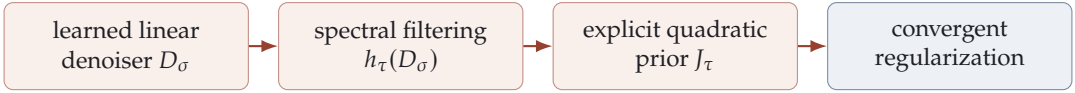
Questions 1–3 settled (much of it at IISc); open: how should the denoiser fade as $\delta \to 0$?

Our work: convergent linear PnP

OUR WORK

A. Hauptmann, S. Mukherjee, C.-B. Schönlieb, F. Sherry. *Convergent regularization in inverse problems and linear plug-and-play denoisers.* Foundations of Computational Mathematics 25, 2025.

- **Denoiser class:** learned linear denoisers
- **Key bridge:** implicit prior of a linear denoiser = an explicit quadratic functional
- **Technical move:** spectral filtering of the denoiser controls the regularization strength
- **Conclusion:** with a parameter rule $\tau \sim \delta$, linear PnP is a convergent regularization method



From "does the iteration converge?" to "does the family of methods converge to a meaningful exact-data solution as the noise vanishes?"

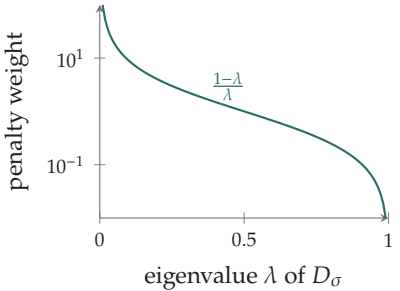
A linear denoiser reveals its regularizer

D_σ linear, self-adjoint, positive definite,
spec $D_\sigma \subset (0,1)$, eigenpairs (λ_i, v_i):

- $D_\sigma = \mathrm{prox}_J = (\mathrm{I} + \partial J)^{-1} \Rightarrow \partial J = D_\sigma^{-1} - \mathrm{I}$ is linear
- hence, up to a constant,

$$J(x) = \tfrac{1}{2}\langle x, (D_\sigma^{-1} - \mathrm{I})x\rangle = \tfrac{1}{2}\sum_i \frac{1-\lambda_i}{\lambda_i}\langle x, v_i\rangle^2$$

- modes the denoiser keeps ($\lambda_i \approx 1$): almost free
- modes it removes ($\lambda_i \approx 0$): heavily penalized
- example: a bias-free, symmetric linear U-net, trained for denoising (our experiments)



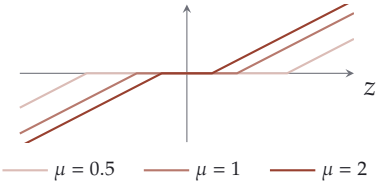
Weight $(1-\lambda)/\lambda$ that J puts on a mode with denoiser eigenvalue λ

An explicit prior, to be weakened on purpose: but how?

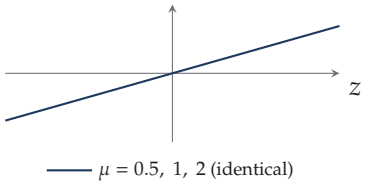
Scaling the denoiser does nothing

Denoiser scaling [Xu et al. 2020]: $D_{\sigma,\mu}(x) = \frac{1}{\mu} D_\sigma(\mu x)$

- $D_\sigma = \mathrm{prox}_J$ with J positively 1-homogeneous (ℓ^1, total variation): $D_{\sigma,\mu} = \mathrm{prox}_{J/\mu}$, so μ acts as a regularization parameter
- linear denoiser = proximal map of a quadratic (2-homogeneous): $\frac{1}{\mu} D_\sigma(\mu x) = D_\sigma x$ for every μ



Soft thresholding (1-homogeneous prior): changes with μ



Linear denoiser (2-homogeneous prior): the same map for every μ

Scaling is invisible to linear denoisers: a different knob is needed

Spectral filtering of the denoiser

OUR WORK

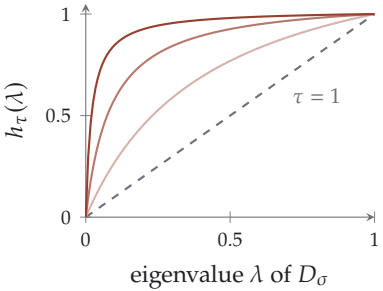
Construction (Hauptmann et al. 2025)

Regularizer $J \mapsto \tau J \Rightarrow$ a *filtered* denoiser

$$\mathrm{prox}_{\tau J} = h_\tau(D_\sigma), \qquad h_\tau(\lambda) = \frac{\lambda}{\tau(1-\lambda)+\lambda}$$

with h_τ acting on the eigenvalues of D_σ

- any filter family $h_\tau \Rightarrow$ a quadratic prior
  $J_\tau(x) = \frac{1}{2}\langle x, (h_\tau(D_\sigma)^{-1} - \mathrm{I})x\rangle = \frac{\tau}{2}\langle x, r_\tau(D_\sigma)x\rangle$
- $r_\tau = (1-h_\tau)/(\tau h_\tau)$; PnP runs with $h_\tau(D_\sigma)$ in place of D_σ



h_τ for $\tau = 1$ (dashed) and $\tau = 0.3, 0.08, 0.02$ (darker is smaller); as $\tau \to 0$ the denoiser fades into the identity

τ: the missing knob, built into the denoiser rather than into an energy

Main result II: convergent linear PnP

OUR WORK

Theorem (Hauptmann, M., Schönlieb, Sherry, FoCM 2025, Thm. 5)

Assume D_σ linear and self-adjoint, $\operatorname{spec} D_\sigma \subset (0,1)$, $\|A\| = 1$, *and continuous filters h_τ with*

(A1) $h_\tau(\operatorname{spec} D_\sigma) \subset (0,1)$ *for every* $\tau > 0$;
(A2) $r_\tau \to r^*$ *uniformly on* $\operatorname{spec} D_\sigma$, *with* $\|r_\tau - r^*\|_\infty = o(\tau)$;
(A3) $0 < \underline{c} \le r_\tau \le \overline{c} < \infty$ *on* $\operatorname{spec} D_\sigma$, *uniformly in* τ.

Then PnP-FBS, $x^{k+1} = h_\tau(D_\sigma)\big(x^k - A^*(Ax^k - y)\big)$, *converges to the minimizer* $\hat{x}(y,\tau)$ *of* $\frac{1}{2}\|Ax - y\|^2 + J_\tau(x)$, *and for* $\tau \sim \delta$

$$\hat{x}(y^\delta, \tau) \;\to\; x^\dagger = \arg\min\Big\{\tfrac{1}{2}\langle x, r^*(D_\sigma)x\rangle \;:\; A^*(Ax - y^0) = 0\Big\}$$

- the limit is selected by the denoiser's own prior
- scaling filter h_τ: $r_\tau = r^* = (1-\lambda)/\lambda$, so (A1)–(A3) hold automatically

Trained linear denoiser + spectral filtering = a convergent regularization method

Proof: reduce to Tikhonov, then perturb

OUR WORK

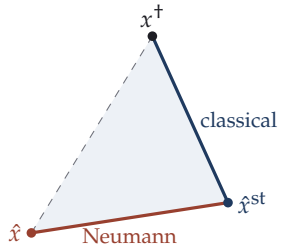
Proof sketch. *static* problem (r^* in place of r_τ) = classical Tikhonov, so $\hat{x}^{\mathrm{st}} \to x^\dagger$ for $\tau \sim \delta$; both solutions linear in the data:

$$\hat{x} = M_\tau^{-1} A^* y^\delta, \qquad M_\tau = A^*A + \tau r_\tau(D_\sigma),$$
$$\hat{x}^{\mathrm{st}} = (M_\tau^{\mathrm{st}})^{-1} A^* y^\delta, \qquad M_\tau^{\mathrm{st}} = A^*A + \tau r^*(D_\sigma)$$

(A3): $\|(M_\tau^{\mathrm{st}})^{-1}\| \le 1/(\underline{c}\,\tau)$; Neumann series:

$$\|M_\tau^{-1} - (M_\tau^{\mathrm{st}})^{-1}\| \le \frac{\|r_\tau - r^*\|_\infty}{\tau\left(\underline{c}^2 - \underline{c}\,\|r_\tau - r^*\|_\infty\right)} \xrightarrow{(A2)} 0$$

□

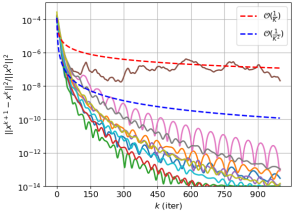


Triangle inequality: a perturbation estimate (rust) plus the classical theorem (navy)

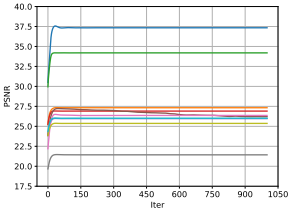
Classical Tikhonov theory does the heavy lifting; (A2) makes the perturbation vanish

Fast but unprovable PnP drifts

OUR WORK



(a) Residual PnP-FISTA



(b) PSNR PnP-FISTA



(c) Failure example

PnP-FISTA for super-resolution, $\sigma = 7.65$, one curve per CBSD10 image [Tan et al., SIIMS 2024]

What to look for

- some residuals stall or oscillate (a)
- PSNR can peak early, then degrade (b)
- some runs fail, leaving artifacts (c)

> *Without guarantees, speed comes with unpredictable failures*

PnP quasi-Newton: same quality, far fewer iterations

OUR WORK

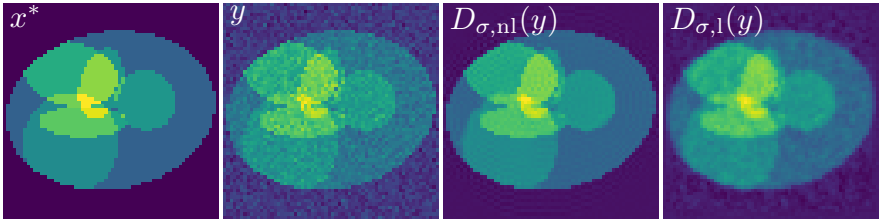
Table: Averaged PSNR (dB) on the CBSD68 dataset for super-resolution.

Scale	$s = 2$				$s = 3$			
σ	2.55	7.65	12.75	Time (s)	2.55	7.65	12.75	Time (s)
PnP-LBFGS[1]	27.89	26.62	25.80	3.19	26.12	25.32	24.68	4.80
PnP-LBFGS[2]	27.89	26.62	25.80	9.81	26.12	25.30	24.68	13.15
PnP-PGD	27.44	26.57	25.82	25.99	25.60	25.20	24.63	37.33
PnP-DRSdiff	27.44	26.58	25.82	18.24	25.60	25.19	24.63	32.83
PnP-DRS	27.93	26.61	25.79	15.74	26.13	25.29	24.67	27.00
PnP-$\hat{a}$PGD	27.94	26.62	25.72	4.24	26.11	25.32	24.69	8.78
PnP-FISTA	26.38	26.44	25.79	24.61	24.96	25.15	24.63	33.13
DPIR (iter 10^3)	18.58	26.36	25.74	19.58	17.53	24.96	24.55	19.67
DPIR (iter 24)	27.82	26.60	25.85	0.98	26.06	25.29	24.67	0.97

PSNR on par with other provable PnP; 2–8× faster convergence, 4–5× shorter run times

A learned linear denoiser

OUR WORK



ground truth x^* noisy y: 24.3 dB nonlinear U-net: 34.0 dB linear U-net: 27.8 dB

Nonlinear vs. symmetric linear U-net, same architecture: the linear one struggles with sharp edges [Hauptmann et al., FoCM 2025]

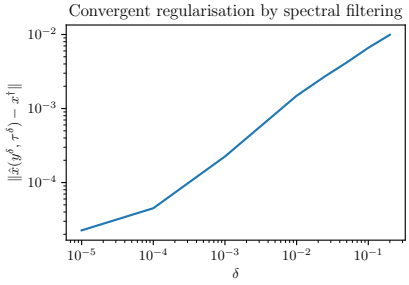
Setting

- computed tomography: 64×64 ellipse phantoms, parallel-beam geometry, 150 views
- denoiser: symmetric, bias-free linear U-net, trained on noisy phantoms

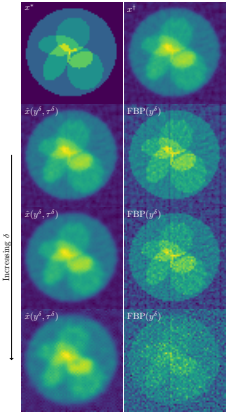
Linear denoisers: weaker, but with a complete theory

Convergent regularization, observed

OUR WORK



Error to $x^\dagger$ versus noise level, with spectral filtering and $\tau \propto \delta$ [Hauptmann et al., FoCM 2025]



Snapshots for the plot on the left, δ increasing downward [Hauptmann et al., FoCM 2025]

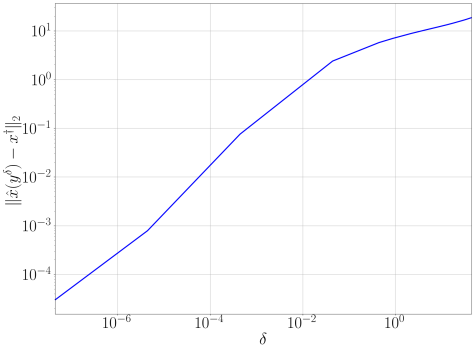
What to look for

- spectral filtering (left): clean even at large δ, tends to $x^\dagger$ as $\delta \to 0$
- filtered backprojection (right): noise grows with δ
- top: $x^\dagger$ lacks the sharpness of x^*, a limitation of linear denoisers

The theorem, visible in practice: the limit is $x^\dagger$, not the ground truth (A has a non-trivial kernel)

A hint beyond linear denoisers

OUR WORK



Gradient-step DRUNet PnP for deblurring, $\lambda = c\,\sigma_0 + \epsilon$ ($c = 0.04$, $\epsilon = 10^{-4}$): distance to the g_σ-minimizing solution $x^\dagger$ versus noise level [Hauptmann et al., FoCM 2025]

Observation

- error decreases with the noise level, as for a convergent regularization
- no theorem yet for this nonlinear case

Nonlinear PnP also seems to converge as $\delta \to 0$: why?

Nonlinear PnP as the noise vanishes

OUR WORK



$\sigma_0 = 10\%$



$\sigma_0 = 8\%$



$\sigma_0 = 5\%$



$\sigma_0 = 2\%$



$\sigma_0 = 1\%$



$\sigma_0 = 0.5\%$



$\sigma_0 = 0.1\%$



$\sigma_0 = 0$: $x^\dagger$



ground truth

Same experiment: reconstructions $\hat{x}(y^{\sigma_0})$ as the noise level σ_0 (in % of the peak value 255) decreases [Hauptmann et al., FoCM 2025]

> *Visually too: the reconstructions settle on $x^\dagger$ as $\sigma_0 \to 0$*

From denoisers to scores: Tweedie's bridge

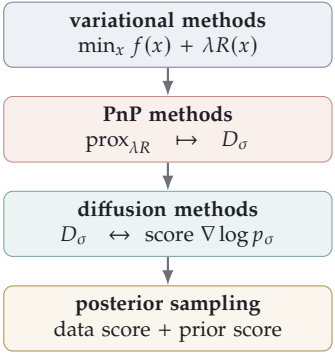
Tweedie ($z = x + \sigma\xi$):
$D_\sigma(z) = \mathbb{E}[x \mid z] = z + \sigma^2 \nabla \log p_\sigma(z)$, so

$$\frac{D_\sigma(z) - z}{\sigma^2} = \nabla \log p_\sigma(z)$$

- an MMSE denoiser *is* a score model
- posterior $p(x \mid y) \propto \exp\left(-\|Ax - y\|^2/2\sigma_n^2\right) p(x)$, with score

$$\nabla \log p(x \mid y) = \tfrac{1}{\sigma_n^2} A^\top (y - Ax) + \nabla \log p(x)$$

- diffusion samplers for inverse problems: denoising across noise levels σ_t, alternated or blended with data consistency [DDRM, Kawar et al. 2022; DPS, Chung et al. 2023; DiffPIR, Zhu et al. 2023]



PnP as the bridge between variational reconstruction and score-based posterior inference

Diffusion posterior sampling: a stochastic, multiscale PnP with a noise-dependent denoiser

What changes in the diffusion setting?

Classical PnP	Diffusion and posterior sampling	New difficulty
one denoiser D_σ	a family D_{σ_t}, or a score $s_\theta(\cdot, t)$	multiscale analysis
a fixed-point iteration	a stochastic reverse process	convergence in distribution
a point estimate	posterior samples and uncertainty	calibration
a static data step	time-dependent likelihood guidance	stability versus bias
operator assumptions on D_σ	assumptions on a learned score field	score errors accumulate

Open mathematical question

A theory as clean as fixed-point PnP, but strong enough for stochastic, diffusion-based posterior sampling?

Main messages

What we learned

1. PnP is powerful because it decouples the physics from the image prior
2. its mathematical meaning depends on the structure of the denoiser
3. convergence, an explicit energy, speed and convergent regularization: four questions, answered by averaged operators, sufficient decrease, and comparison with classical regularization

Our contributions

4. PnP quasi-Newton: acceleration and guarantees can coexist
5. linear PnP: a denoiser defines an honest convergent regularization once spectral filtering controls its strength

Next chapter: diffusion-based sampling, a stochastic and multiscale PnP with uncertainty

Open questions

Nonlinear denoisers

The knob
What replaces spectral filtering for a nonlinear denoiser: σ, λ, or a filter on the Jacobian?

Convergent regularization
Which structure makes nonlinear PnP converge as $\delta \to 0$? Is compatibility with $\ker A$ enough?

Certified structure
Weak convexity and similar constraints hold only approximately after training: how to certify them?

Diffusion-based samplers

A theory for samplers
What does convergent regularization mean for a stochastic posterior sampler?

Faster samplers
Can quasi-Newton or variable-metric ideas accelerate diffusion and PnP samplers?

Errors across scales
How do score errors at many noise levels accumulate in the samples, and can we control them?

> *PnP: from a substitution to a framework for how learned image models, physics, optimization and Bayesian inference should interact*