MAP Estimation with Denoisers: Convergence Rates and Guarantees

Julien Mairal

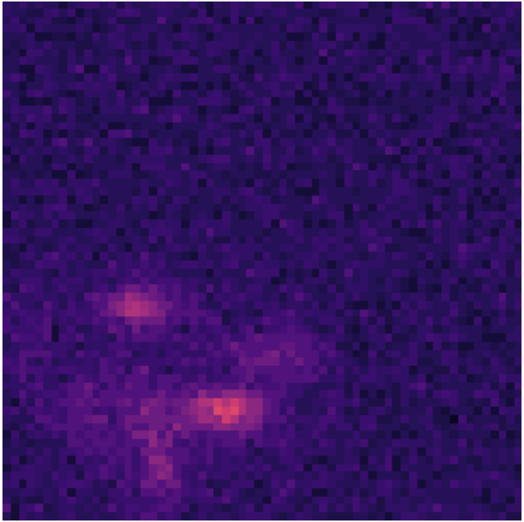
Univ. Grenoble-Alpes, Inria

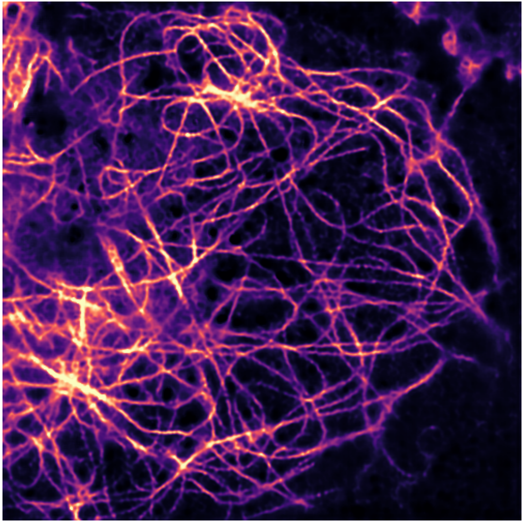
Collaborators and Publications

- S. Pesme, G. Meanti, M. Arbel and J. Mairal. MAP Estimation with Denoisers: Convergence Rates and Guarantees. *preprint arXiv:2507.15307*. 2025.

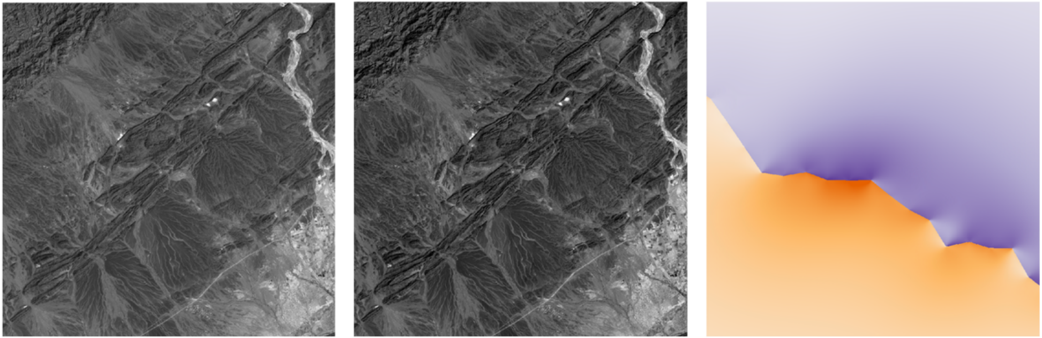
   

Inverse imaging problems: molecular microscopy





Inverse imaging problems: ground displacement estimation



Estimating ground displacement fields from satellite imagery. Picture from [Montagnon, Hollingsworth, Pathier, Marchandon, Dalla Mura, Giffard-Roisin, 2022].

Inverse imaging problems: super-resolution from raw image bursts

Inverse imaging problems: super-resolution from raw image bursts

Classical approaches with (fake) MAP estimation

Find a reasonable model of degradation

For instance: $$\underbrace{z}_{\text{observations}} = A \underbrace{x}_{\text{true signal}} + \underbrace{\varepsilon}_{\text{noise}}.$$

Estimate the true signal by optimizing a reasonable cost function

For instance: $$\min_x \underbrace{\|z - Ax\|^2}_{\text{data fitting term}} + \underbrace{\lambda\phi(x)}_{\text{prior information}}.$$

Some classical priors

- Smoothness $\|\mathcal{L}x\|^2$.
- Total variation $\|\nabla x\|_1$.
- Sparsity $\|x\|_1$.

Classical approaches with (fake) MAP estimation

Probabilistic interpretation

$$\min_x \underbrace{-\log p(z|x)}_{\text{data fitting term}} - \underbrace{\log p(x)}_{\text{log prior information}} .$$

Classical issues

- All the classical priors are part of a **model** and have nothing to do with the real $p(x)$ (assuming it exists).
- We also have to trust the degradation and the noise models.

Classical approaches with (fake) MAP estimation

Probabilistic interpretation

$$\min_x \underbrace{-\log p(z|x)}_{\text{data fitting term}} - \underbrace{\log p(x)}_{\text{log prior information}} .$$

Classical issues

- All the classical priors are part of a **model** and have nothing to do with the real $p(x)$ (assuming it exists).
- We also have to trust the degradation and the noise models.

These approaches are very useful to encode a desired property in the solution.

Supervised (deep) learning

- **Engineer a realistic dataset**: Produce enough pairs (x_i, z_i) of clean/degraded images (semi-synthetic setting).
- **Choose a class of parametrized models** $\{f_\theta : \theta \in \Theta\}$.
- **Learn the parameters**:

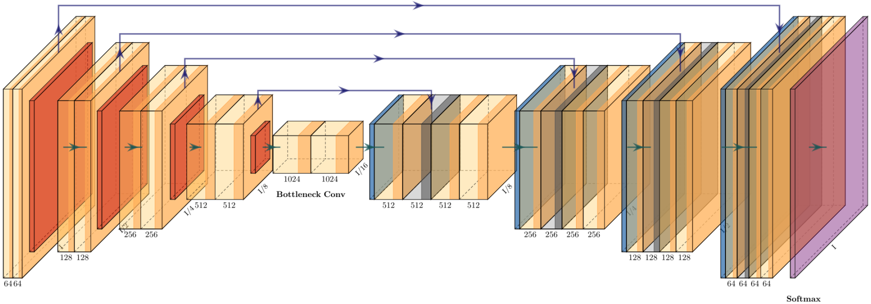
$$\min_\theta \frac{1}{n} \sum_{i=1}^{n} \|f_\theta(z_i) - x_i\|^2.$$

Supervised (deep) learning

- **Engineer a realistic dataset**: Produce enough pairs (x_i, z_i) of clean/degraded images (semi-synthetic setting).
- **Choose a class of parametrized models** $\{f_\theta : \theta \in \Theta\}$.
- **Learn the parameters**:

$$\min_\theta \frac{1}{n} \sum_{i=1}^{n} \|f_\theta(z_i) - x_i\|^2.$$

Example: U-Net



Supervised (deep) learning

- **Engineer a realistic dataset**: Produce enough pairs (x_i, z_i) of clean/degraded images (semi-synthetic setting).
- **Choose a class of parametrized models** $\{f_\theta : \theta \in \Theta\}$.
- **Learn the parameters**:

$$\min_\theta \frac{1}{n} \sum_{i=1}^{n} \|f_\theta(z_i) - x_i\|^2.$$

Approximation of the Bayes (MMSE) estimator

$$f_{\theta^\star}(z) \approx \mathbb{E}[X|Z = z].$$

- large capacity is required.
- large amounts of data are typically easy to produce in semi-synthetic settings.
- **do we have the right learning theory for image processing?**

(true) MAP vs. MMSE

We are considering two types well-motivated estimators? Do we care?

MMSE

$$\mathsf{MMSE}(z) = \mathbb{E}[X|Z = z].$$

Interpretation: "an average of plausible solutions".

true MAP

$$\mathsf{MAP}(z) = \arg\min_x \{-\log p(z|x) - \log p(x)\}\,.$$

Interpretation: "a good solution".

In both cases, we can only obtain an approximation. true MAP requires modeling $p(x)$, which suggests using generative models. **Do we need sampling or optimization?**

(true) MAP vs. MMSE: motivation on a simple denoising case

Approximation of the MMSE (left) and of the true MAP (right)

Approximation of the MMSE (left) and of the true MAP (right)

Are these details true? approximation of the MAP

Are these details true? ground truth

(true) MAP vs. MMSE

- approx true MAP is visually more pleasant but it hallucinates some details.
- approx MMSE has fewer hallucinations but tend to be less sharp (slightly blurry).

Related work

- Recently, inverse problems have been seen as a **sampling** task [Delbracio and Milanfar, 2023, Chung et al., 2023, Boys et al., 2024], using diffusion tools.
- Heuristic algorithms have been proposed that "may" approximate the MAP such as Indi, Cold Diffusion [Bansal et al., 2023], ...

(true) MAP vs. MMSE

- approx true MAP is visually more pleasant but it hallucinates some details.
- approx MMSE has fewer hallucinations but tend to be less sharp (slightly blurry).

Related work

- Recently, inverse problems have been seen as a **sampling** task [Delbracio and Milanfar, 2023, Chung et al., 2023, Boys et al., 2024], using diffusion tools.
- Heuristic algorithms have been proposed that "may" approximate the MAP such as Indi, Cold Diffusion [Bansal et al., 2023], ...

How to address (true) MAP estimation with convergence guarantees assuming we have access to an optimal denoiser (MMSE)?

a few attempts: [Laumont et al., 2023, Zhang et al., 2024]

Our problem: computing the proximal operator of $-\tau \log p$

$$\mathsf{Prox}_{-\tau \log p}[z] := \arg\min_x \left\{ F(x) := \frac{1}{2}\|z - x\|^2 - \tau \log p(x) \right\},$$

assuming of course the argmin is unique...

Our problem: computing the proximal operator of $-\tau \log p$

$$\mathsf{Prox}_{-\tau \log p}[z] := \arg\min_x \left\{ F(x) := \frac{1}{2}\|z - x\|^2 - \tau \log p(x) \right\},$$

assuming of course the argmin is unique...

Why the prox? (i) it can serve as a swiss-army knife in almost any inverse problem; (ii) it will yield well-grounded plug-and-play algorithms with convergence guarantees [Venkatakrishnan et al., 2013, Hurault et al., 2021].

Our problem: computing the proximal operator of $-\tau \log p$

$$\mathsf{Prox}_{-\tau \log p}[z] := \arg\min_x \left\{ F(x) := \frac{1}{2}\|z - x\|^2 - \tau \log p(x) \right\},$$

assuming of course the argmin is unique...

What is specific here?

- We **can neither evaluate** $p(x)$, nor $\nabla \log p(x)$.
- We do not have access to classical quantities such as Lipschitz constants. The algorithm needs to be **parameter-free**!
- We assume we can sample from p and we have access to an **optimal MMSE denoiser**:

$$\mathsf{MMSE}_\sigma(z) = \mathbb{E}[X | X + \sigma\varepsilon = z] \quad \text{with} \quad \varepsilon \sim \mathcal{N}(0, I) \quad \text{and} \quad X \sim p.$$

The algorithm

Consider an input $x_0 = z$, and sequences $\alpha_k = \frac{1}{k+2}$ and $\sigma_k^2 = \frac{\tau}{k+1}$.

$$x_{k+1} = (1 - \alpha_k)\,\mathrm{MMSE}_{\sigma_k}(x_k) + \alpha_k z. \qquad \text{(MMSE Averaging)}$$

- Same structure as cold diffusion [Bansal et al., 2023], related to Indi [Delbracio and Milanfar, 2023].
- Related to flow matching [Liu et al., 2022] and score matching [Hyvärinen, 2005].

We are talking about deterministic variants.

The algorithm

Consider an input $x_0 = z$, and sequences $\alpha_k = \frac{1}{k+2}$ and $\sigma_k^2 = \frac{\tau}{k+1}$.

$$x_{k+1} = (1 - \alpha_k)\,\mathrm{MMSE}_{\sigma_k}(x_k) + \alpha_k z. \qquad \text{(MMSE Averaging)}$$

Where does it come from? Remember Tweedie's formula:

$$\mathsf{MMSE}_\sigma(z) = z + \sigma^2 \nabla \log p_\sigma(z),$$

where $p_\sigma = p \star \mathcal{N}(0, \sigma^2 I)$ is the density of $x + n$ with $x \sim p$ and $n \sim \mathcal{N}(0, \sigma^2 I)$.
It turns out that the Lipschitz constant of $\nabla \log p_\sigma(z)$ is upper-bounded by $1/\sigma^2$.
Then, the algorithm is equivalent to

$$x_{k+1} = x_k - \alpha_k \nabla F_{\sigma_k}(x_k), \quad \text{with} \quad F_{\sigma_k}(x) := \frac{1}{2}\|z - x\|^2 - \tau \log p_{\sigma_k}(x),$$

with the step-size $\alpha_k = 1/L_{F_{\sigma_k}}$ (classical step-size).

An optimization point of view

Original problem:

$$\min_x \left\{ F(x) := \frac{1}{2}\|z - x\|^2 - \tau \log p(x) \right\}.$$

Algorithm:

$$x_{k+1} = x_k - \alpha_k \nabla F_{\sigma_k}(x_k), \quad \text{with} \quad F_{\sigma_k}(x) := \frac{1}{2}\|z - x\|^2 - \tau \log p_{\sigma_k}(x),$$

- gradient descent on a smoothed objective with **vanishing step-sizes and smoothing parameters**.
- the algorithm does not depend on the smoothness properties of F!

Our result: assumptions

This is a nonconvex problem with no known convergence result. It seems it can be approached via randomized smoothing techniques (but the $\log$ changes everything).

Our result: assumptions

This is a nonconvex problem with no known convergence result. It seems it can be approached via randomized smoothing techniques (but the $\log$ changes everything).

Assumption

The density p is log-concave, and strictly positive on $\mathbb{R}^d$ **:-(**

Our result: assumptions

This is a nonconvex problem with no known convergence result. It seems it can be approached via randomized smoothing techniques (but the $\log$ changes everything).

Assumption

The density p is log-concave, and strictly positive on $\mathbb{R}^d$ **:-(**

Assumption

p is three times differentiable and the third derivative of $\log p$ is bounded:

$$\sup_{x \in \mathbb{R}^d} \left\| \nabla^3 \log p(x) \right\|_F = M.$$

Note that

- For a Gaussian density, $M = 0$.
- $\log p$ could be ill-conditioned (no bound on second derivative).

Our result: theorem

Theorem (Convergence to the Proximal operator)

Under the previous assumptions, the MMSE Averaging iterates with parameters $\alpha_k = 1/(k+2)$, $\sigma_k^2 = \tau/(k+1)$ *and initialised at* $x_0 = z$ *satisfy:*

$$\|x_k - \mathit{Prox}_{-\tau \log p}(z)\| \leq \frac{(\log k) + 7}{k+1} \left[\|z - \mathit{Prox}_{-\tau \log p}(z)\| + \tau^2 M \sqrt{d} \right].$$

Recipe of the proof

- analysis of gradient descent with inexact gradients.
- Controlling $\sigma \mapsto x_\sigma^\star$ (the interesting part).

Case of the Gaussian density

Consider a Gaussian density with covariance Σ.

A few remarks

- the algorithm MMSE Averaging with $\alpha_k = 1$ converges in **one step**.
- with the defaults parameters, it converges in $\tilde{O}(1/\varepsilon)$.
- if naive GD was authorized (requires access to $\nabla \log p(x)$), it would converge in $O(L \log(1/\varepsilon))$ with $L = \|\Sigma\|$ assuming L is known.

Sketch of proof (1/2): improved conditioning

Key property: A "second-order Tweedie" identity leads to:

$$-\nabla^2 \ln p_\sigma(z) \preceq \tfrac{1}{\sigma^2} I_d,$$

which implies that the conditioning of F_σ satisfies $\kappa_\sigma = 1 + \frac{\tau}{\sigma^2}$.

Now let $x^\star_{\sigma_k} := \arg\min F_{\sigma_k}$. A single gradient step on F_{σ_k} yields:

$$\begin{aligned}\|x_{k+1} - x^\star_{\sigma_k}\| &\le \left(\tfrac{\kappa_{\sigma_k}-1}{\kappa_{\sigma_k}+1}\right)^{1/2} \|x_k - x^\star_{\sigma_k}\| \\ &\le \left(\tfrac{k+1}{k+3}\right)^{1/2} \Big(\|x_k - x^\star_{\sigma_{k-1}}\| + \overbrace{\|x^\star_{\sigma_{k-1}} - x^\star_{\sigma_k}\|}^{\text{shift of the minimisers}} \Big)\end{aligned}$$

The main challenge is to control $\sigma \mapsto x^\star_\sigma$. Assuming we can:

$$\|x_{k+1} - x^\star_{\sigma_k}\| \le \left(\tfrac{k+1}{k+3}\right)^{1/2} \Big(\|x_k - x^\star_{\sigma_{k-1}}\| + C\,(\sigma^2_{k-1} - \sigma^2_k) \Big).$$

Unrolling this recursion yields the desired convergence bound.

Sketch of proof (2/2): Lipschitz continuity of $\sigma^2 \mapsto x_\sigma^\star$

PDE satisfied by $x_\sigma^\star$. Using the optimality conditions and the heat equation $\partial_{\sigma^2} p_\sigma = \Delta p_\sigma$, one obtains

$$\frac{\mathrm{d}x_\sigma^\star}{\mathrm{d}\sigma^2} = \tfrac{1}{2}\left[-\nabla^2 \ln p_\sigma(x_\sigma^\star) + \tfrac{1}{\tau} I_d\right]^{-1}\Big(2\nabla^2 \ln p_\sigma(x_\sigma^\star)\,\nabla \ln p_\sigma(x_\sigma^\star) + \overbrace{\nabla\Delta \ln p_\sigma(x_\sigma^\star)}^{\text{annoying term}}\Big).$$

Which immediately yields

$$\left\|\frac{\mathrm{d}x_\sigma^\star}{\mathrm{d}\sigma^2}\right\| \le \|\nabla \ln p_\sigma(x_\sigma^\star)\| + \tfrac{\tau}{2}\,\|\nabla\Delta \ln p_\sigma(x_\sigma^\star)\|.$$

The first term is readily bounded via the optimality condition. For the second term, we prove that (technical crux of the paper):

$$\sup_{x\in\mathbb{R}^d} \|\nabla\Delta \ln p_\sigma(x)\| \le \sqrt{d} \sup_{x\in\mathbb{R}^d} \|\nabla^3 \ln p(x)\|_F.$$

And we obtain the uniform bound $\left\|\frac{\mathrm{d}x_\sigma^\star}{\mathrm{d}\sigma^2}\right\| \le C = \frac{1}{\tau}\|y - \mathrm{prox}_{-\tau \ln p}(y)\| + \tau\sqrt{d}\,\sup_{x\in\mathbb{R}^d}\|\nabla^3 \ln p(x)\|_F$, which proves that $\sigma^2 \mapsto x_\sigma^\star$ is Lipschitz.

Questions and perspectives

Questions

- do we have the right rate of convergence?
- What is the right quantity the analysis should depend on?
- what is the right dependence on the dimension? Should it depend on a local dimension?
- how to deal with inexact MMSE estimators (on-going).
- how to deal with non-convexity? Would stochastic variants be useful?
- non-smooth distributions: uniform densities on compact sets?

Perspective

- a PnP algorithm with convergence guarantees, which optimizes a natural objective function. **Finally!**

References I

Arpit Bansal, Eitan Borgnia, Hong-Min Chu, Jie Li, Hamid Kazemi, Furong Huang, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Cold diffusion: Inverting arbitrary image transforms without noise. *Advances in Neural Information Processing Systems*, 36: 41259–41282, 2023.

Benjamin Boys, Mark Girolami, Jakiw Pidstrigach, Sebastian Reich, Alan Mosca, and O. Deniz Akyildiz. Tweedie moment projected diffusions for inverse problems, 2024.

Hyungjin Chung, Jeongsol Kim, Sehui Kim, and Jong Chul Ye. Parallel diffusion models of operator and image for blind inverse problems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6059–6069, 2023.

Mauricio Delbracio and Peyman Milanfar. Inversion by direct iteration: An alternative to denoising diffusion for image restoration. *arXiv preprint arXiv:2303.11435*, 2023.

Samuel Hurault, Arthur Leclaire, and Nicolas Papadakis. Gradient step denoiser for convergent plug-and-play. In *International Conference on Learning Representations*, 2021.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005.

References II

Rémi Laumont, Valentin De Bortoli, Andrés Almansa, Julie Delon, Alain Durmus, and Marcelo Pereyra. On maximum a posteriori estimation with plug & play priors and stochastic gradient descent. *Journal of Mathematical Imaging and Vision*, 65(1):140–163, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Singanallur V Venkatakrishnan, Charles A Bouman, and Brendt Wohlberg. Plug-and-play priors for model based reconstruction. In *2013 IEEE global conference on signal and information processing*, pages 945–948. IEEE, 2013.

Yasi Zhang, Peiyu Yu, Yaxuan Zhu, Yingshan Chang, Feng Gao, Ying Nian Wu, and Oscar Leong. Flow priors for linear inverse problems via iterative corrupted trajectory matching. *arXiv preprint arXiv:2405.18816*, 2024.