Inverse Problems and Imaging
Volume 1, No. 2, 2007, 399–422

Web site: http://www.aimSciences.org

MODEL DISTORTIONS IN BAYESIAN MAP RECONSTRUCTION

Mila Nikolova

CMLA, ENS Cachan, CNRS,
PRES UniverSud
61 Av. President Wilson,
F-94230 Cachan, FRANCE

(Communicated by Jianhong (Jackie) Shen)

Abstract. The Bayesian approach and especially the *maximum a posteriori* (MAP) estimator is most widely used to solve various problems in signal and image processing, such as denoising and deblurring, zooming, and reconstruction. The reason is that it provides a coherent statistical framework to combine observed (noisy) data with prior information on the unknown signal or image which is optimal in a precise statistical sense. This paper presents an objective critical analysis of the MAP approach. It shows that the MAP solutions substantially deviate from both the data-acquisition model and the prior model that underly the MAP estimator. This is explained theoretically using several analytical properties of the MAP solutions and is illustrated using examples and experiments. It follows that the MAP approach is not relevant in the applications where the data-observation and the prior models are accurate. The construction of solutions (estimators) that respect simultaneously two such models remains an open question.

1. MAP estimators to combine noisy data and priors

We address a wide variety of inverse problems where an estimate $\hat{x} \in \mathbb{R}^p$ of an original (unknown) $X = x$ (e.g. an image, a signal or some parameters) is recovered from a realization[1] of noisy data $Y = y \in \mathbb{R}^q$ using a statistical model for their production as well as a prior model for the original X. Typical applications are signal and image restoration, segmentation, motion estimation, sequence processing, color reproduction, optical imaging, tomography, seismic and nuclear imaging, and many others. The likelihood function $f_{Y|X}(y|x)$—the distribution for the observed data $Y = y$ given an original $X = x$—is governed by physical considerations concerned with the data-acquisition device. By far the most common models are of the form

$$Y = AX + N, \tag{1}$$

where $A : \mathbb{R}^p \to \mathbb{R}^q$ is a linear operator (e.g. a blurring kernel, a Fourier or a Radon transform, or a subsampling operator) and N is additive noise which is independent of X. If the noise samples $N_i, 1 \le i \le q$ are independent and identically distributed

2000 *Mathematics Subject Classification:* Primary: 62H99, 68U99; Secondary: 49K99.
Key words and phrases: MAP estimation, restoration, regularization, modeling.

[1]When necessary, random variables are identified by uppercase letters and their values by lowercase letters.

(i.i.d.) with marginal distribution f_N, then

$$f_{Y|X}(y|x) = \prod_{i=1}^{q} f_N\big(a_i^T x - y_i\big), \tag{2}$$

where a_i^T, $1 \le i \le q$ are the rows of A. Frequently f_N is a zero-mean Gaussian density on $\mathbb{R}$.

A meaningful solution $\hat{x}$ can seldom be recovered based on the data-acquisition model $f_{Y|X}$ solely, without the help of prior information[2] on the unknown X [61, 59, 15, 62]. Priors on X can take various forms [26, 59, 15, 60, 31, 62]. We will focus on two approaches to model statistical priors f_X on X. Markov random models [8, 49, 23, 35, 9, 20, 30, 3] concentrate on the local characteristics of X, namely the distribution for each pixel x_i conditionally to its neighbors x_j, $j \in \mathcal{N}_i$:

$$f_X(x_i \Big| x_j, j \neq i) = f_X(x_i \Big| x_j, j \in \mathcal{N}_i), \ \forall i \in \{1, \ldots, p\}.$$

For a 2D image, $\mathcal{N}_i$ is often the set of the 4, or the 8 pixels adjacent to i. Assume that the prior has the usual Gibbsian form

$$f_X(x) \propto \exp\{-\lambda \Phi(x)\}, \tag{3}$$

where Φ is a prior energy function and $\lambda > 0$ is a parameter. The Hammersley-Clifford theorem [8] is a powerful tool to conceive Markovian priors by factorizing f_X according to the set of the cliques involved in (3). A very useful class of priors obtained in this way correspond to[3] [9, 7, 34, 11]

$$\Phi(x) = \frac{1}{2} \sum_i \sum_{j \in \mathcal{N}_i} \varphi(x_i - x_j), \tag{4}$$

where $\varphi : \mathbb{R} \to \mathbb{R}_+$ is a suitable symmetric function, increasing on $\mathbb{R}_+$, as those given in Table 1. The fit of φ to the empirical distribution of the differences $\{x_i - x_j : j \in \mathcal{N}_i, 1 \le i \le p\}$ in real-world images is considered e.g. in [52].

Another approach is to use wavelets or more generally frame expansions from the outset. Let $\{w_i : 1 \le i \le r\}$ be a family of wavelet functions on $\mathbb{R}^p$. In numerous papers [36, 64, 56, 38, 33, 55, 37, 6, 60, 1, 10] the coefficients

$$u_i = \langle w_i, x \rangle, \ 1 \le i \le r, \tag{5}$$

are assumed to be i.i.d. and their statistical distribution f_U is described using priors of the form

$$f_U(u) = \frac{1}{Z} \prod_{i=1}^{r} \exp\big(-\lambda_i \varphi(t)\big), \tag{6}$$

where φ is a function as those given in Table 1 and $\lambda_i > 0$ for all i. In (6) and in what follows, Z denotes a normalization constant.

[2] If A is the identity, the maximum-likelihood estimate—the maximizer of $f_{Y|X}(y|.)$—is useless since it returns back the noisy data $\hat{x} = y$. When A is ill-conditioned, it is well known that any maximizer $\hat{x}$ of $f_{Y|X}(y|.)$ is unstable with respect to the noise and the numerical errors.

[3] For Φ of the form (4), $\Phi(x) = 0$ if $x_i = x_j$ for all i, j, so $x \to e^{-\Phi(x)}$ is non-integrable on $\mathbb{R}^p$ and f_X is an improper prior [9, 7, 19]. This impropriety can be easily removed either by restricting x to belong to a bounded domain or by preventing x to shift up and down. As noticed by many authors, this is hardly worthwhile since under quite general assumptions, the posterior distribution $f_{X|Y}(x|y)$ is proper.

Notice that the set of the cliques involved in (4) reads $\{(i,j) : j \in \mathcal{N}_i, 1 \le i \le p\}$.

Convex functions φ			
Smooth at zero functions φ		Nonsmooth at zero functions φ	
(f1)	$\varphi(t) = \|t\|^\alpha,\ 1<\alpha\leq 2$	(f8)	$\varphi(t) = \|t\|$
(f2)	$\varphi(t) = \sqrt{\alpha + t^2}$		
(f3)	$\varphi(t) = \begin{cases} t^2/2 & \text{if } \|t\| \leq \alpha \\ \alpha t - \alpha^2/2 & \text{else} \end{cases}$		
Nonconvex functions φ			
Smooth at zero functions φ		Nonsmooth at zero functions φ	
(f4)	$\varphi(t) = \min\{\alpha t^2, 1\}$	(f9)	$\varphi(t) = \|t\|^\alpha,\ 0 < \alpha < 1$
(f5)	$\varphi(t) = \dfrac{\alpha t^2}{1+\alpha t^2}$	(f10)	$\varphi(t) = \dfrac{\alpha\|t\|}{1+\alpha\|t\|}$
(f6)	$\varphi(t) = \log(\alpha t^2 + 1)$	(f11)	$\varphi(t) = \log(\alpha\|t\| + 1)$
(f7)	$\varphi(t) = 1 - \exp(-\alpha t^2)$	(f12)	$\varphi(0)=0,\ \varphi(t)=1 \text{ if } t\neq 0$

TABLE 1. Commonly used functions φ where $\alpha > 0$ is a parameter. Some references for these functions are [40, 12, 24, 9, 27, 32, 48, 58, 21, 51, 13, 63, 52].

The posterior distribution $f_{X|Y}(x|y)$, given by the Bayesian chain rule,

$$f_{X|Y}(x|y) = f_{Y|X}(y|x) f_X(x) \frac{1}{Z},\ Z = f_Y(y),$$

combines the information brought by the likelihood $f_{Y|X}(y|.)$ with the prior f_X. Bayesian estimators are based on $x \to f_{X|Y}(x|y)$ and they realize a *compromise* between $f_{Y|X}(y|.)$ and f_X which is optimal with respect to a loss function[4]. Our focus is on the most popular Bayesian estimator—the *Maximum a Posteriori (MAP)*—which selects $\hat{x}$ as *the most likely solution given the observed data $Y = y$:*

$$\begin{aligned} \hat{x} &= \arg\max_x f_{X|Y}(x|y) \\ &= \arg\min_x \big(-\ln f_{Y|X}(y|x) - \ln f_X(x) \big). \end{aligned}$$

If $\Psi(x,y) \propto -\ln f_{Y|X}(y|x)$ and (3) holds, $\hat{x}$ equivalently minimizes a posterior energy $\mathcal{E}_y$ of the form

$$\mathcal{E}_y(x) = \Psi(x,y) + \beta\Phi(x) \tag{7}$$

where β ensures that $\mathcal{E}_y \propto -\ln f_{X|Y}(.|y) + const$. Under the classical assumption that the noise in (1) is i.i.d. and has a zero-mean Gaussian density with variance σ^2—which situation will be indicated by $N \sim \text{Normal}(0, \sigma^2 I)$ where I is the $q \times q$ identity matrix—the MAP estimate $\hat{x}$ minimizes

$$\mathcal{E}_y(x) = \|Ax - y\|^2 + \beta\Phi(x) \text{ where } \beta = 2\sigma^2\lambda. \tag{8}$$

[4]For every y, a Bayesian estimate $\hat{x}$ minimizes $\int \mathcal{L}(\hat{x}, x) f_{X|Y}(x|y)dx$ where $\mathcal{L}$ is a loss function. The MAP estimator corresponds to

$$\mathcal{L}(\hat{x}, x) = \begin{cases} 0 & \text{if } \hat{x} = x \\ 1 & \text{otherwise} \end{cases}$$

Other Bayesian estimators are for instance the *Posterior Mean* which is defined by $\hat{x} = \int xp(x|y)dx$ and the *Marginal Posterior Mean* where $\hat{x}_i = \int x_i p(x_i|y)dx$ for $1 \leq i \leq p$. They usually require cumbersome numerical integration on $\mathbb{R}^p$ which considerably restricts their practical interest in signal and image applications. By far, the MAP is the most widely used Bayesian estimator [53].

Since [16], denoising of signals and images is efficiently dealt by restoring the noisy wavelet coefficients $\langle w_i, y\rangle$, $1 \le i \le r$, with the aid of priors of the form (6). Such methods were considered by many authors, e.g. [36, 64, 56, 38, 33, 55, 37, 6, 60, 1, 10], and they amount to calculating $\hat{u} = \arg\min_u \mathcal{E}_y(u)$ for

$$\mathcal{E}_y(u) = \sum_i \Big((u_i - \langle w_i, y\rangle)^2 + \lambda_i \varphi(|u_i|) \Big). \tag{9}$$

The sought-after solution is then $\hat{x} = W^\dagger \hat{u}$ where $W^\dagger$ is a left-inverse of $\{w_i, 1 \le i \le r\}$.

Realistic statistical modeling of the physical phenomena in data-acquisition devices on the one hand, and modeling of the priors for real-world images and signals on the other hand, focuses more and more efforts in research and applications, and the references are abundant [3, 7, 31, 52, 39, 57, 54, 25]. This is naturally done with *the expectation to obtain solutions $\hat{x}$ that are coherent with all the two models $f_{Y|X}$ and f_X*. The adequacy of the most popular MAP estimator has essentially been considered in asymptotical conditions when $\beta \searrow 0$, $q \to \infty$, or $\beta \nearrow \infty$, i.e. when either f_X or $f_{Y|X}$ vanishes. Stronger results were derived in [18], namely the distribution of the MAP estimator $\hat{X}$ conditionally to an original image x, for priors of the form (4) and φ as (f1) or (f2) in Table 1, and these results confirm what is presented in the following. In regular (non-asymptotic) conditions, we exhibit important contradictions in the MAP approach since *the MAP solutions substantially deviate from the data-acquisition model $f_{Y|X}$ on the one hand, and from the prior model f_X on the other hand*. This gap between modeling and solution is first illustrated in section 2 using tractable examples that consider the ideal situation when both the data-acquisition and the prior models are known exactly. The remaining of the paper explains rigorously the reasons for this gap between models and solutions for the main families of posterior distributions used in signal and image processing. More precisely, we consider posterior energies involving non-smooth priors (section 3), or non-smooth likelihood functions (section 4), or non-convex prior energies (section 5). For clarity, we skip some technical details but reference where they can be found. The theoretical arguments come from several analytical properties characterizing the MAP solutions as a function of the shape of $f_{X|Y}$, or equivalently as a function of the shape of $\mathcal{E}_y$, see e.g. [42, 44, 47]. Relevant numerical experiments are used to illustrate the gap between the goals of the modeling and the resultant solutions. Obviously the MAP estimator deforms the information contained in both the data-acquisition and the prior models. Even though MAP is *optimal* in a precise statistical sense, the distortions it imposes on the data-acquisition and on the prior models are embarrassing in many applications where these models are accurate. It can hence not be recommended in such cases. We can conjecture that similar problems arise along with other Bayesian estimators as far as the posterior $f_{X|Y}$ mix these models. The conception of solutions that are coherent with two such models remains an open problem.

2. Gaps between models and estimate

Let us consider a measurement model of the form (1)-(2) where the sought-after X and the noise N have distributions f_X and f_N, respectively. In full rigor, an estimator $\hat{X}$ for X, based on data Y, can be said to be coherent with the underlying models if $\hat{X} \sim f_X$ (i.e. $\hat{X}$ has the same distribution as the prior) and if the resultant noise estimator $\hat{N} = Y - A\hat{X}$ satisfies $\hat{N} \sim f_N$ (i.e. it has the same distribution as the noise N). Usually, neither of the distributions $f_{\hat{X}}$ nor $f_{\hat{N}}$ can be calculated. Below we calculate $f_{\hat{X}}$ and $f_{\hat{N}}$ for scalar variables with tractable distributions and the result is illuminating.

2.1. Analytical example on $\mathbb{R}$. Let us assume a scalar additive measurement model

$$Y = X + N,$$

where the distribution of X is

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0, \\ 0 & \text{else,} \end{cases} \tag{10}$$

and $N \sim \text{Normal}(0, \sigma^2)$ (i.e. $N \sim f_N$ for $f_N(n) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{n^2}{2\sigma^2}}$). For every $y \in \mathbb{R}$, the MAP solution $\hat{x}$ is the minimizer on $[0, +\infty)$ of $\mathcal{E}_y$ where

$$\mathcal{E}_y(x) = (x - y)^2 + \beta x \text{ for } \beta = 2\sigma^2\lambda.$$

It is easy to find that

$$\hat{x} = \begin{cases} 0 & \text{if} \quad y < \frac{\beta}{2}, \\ y - \frac{\beta}{2} > 0 & \text{if} \quad y \geq \frac{\beta}{2}. \end{cases}$$

After some calculations[5], the (unconditional) distribution $f_{\hat{X}}$ of $\hat{X}$ reads

$$f_{\hat{X}}(\hat{x}) = f_X(\hat{x})\ \xi(\hat{x}) + c\delta(\hat{x}),$$

where δ stands for Dirac distribution and

$$\begin{aligned} \xi(\hat{x}) &= e^{\frac{\lambda}{2}(\lambda\sigma^2 - \beta)} \int_0^\infty f_N(x - \hat{x} - \frac{\beta}{2} + \lambda\sigma^2) dx, \\ c &= \int_0^\infty f_X(x) \int_{-\infty}^{\frac{\beta}{2} - x} f_N(n) dn dx > 0. \end{aligned}$$

The distribution of $\hat{X}$ is fundamentally dissimilar to the prior f_X since $f_{\hat{X}}$ involves a Dirac delta at zero while for every $\hat{x} > 0$ it is weighted by $\xi(\hat{x})$. Furthermore, the noise estimate $\hat{n} = y - \hat{x}$ reads

$$y - \hat{x} = \begin{cases} y & \text{if} \quad y < \frac{\beta}{2}, \\ \frac{\beta}{2} & \text{if} \quad y \geq \frac{\beta}{2}. \end{cases}$$

Its distribution is given by

$$\begin{aligned} f_{\hat{N}}(\hat{n}) &= f_N(\hat{n})\ \mathbb{1}(\hat{n} < \frac{\beta}{2})\ \zeta(\hat{n}) + (1 - c)\ \delta(\hat{n} - \frac{\beta}{2}), \\ \zeta(\hat{n}) &= \int_0^\infty f_X(x) e^{-\frac{x^2 - 2\hat{n}x}{2\sigma^2}} dx. \end{aligned}$$

Unlike N, the distribution of $\hat{N}$ is upper bounded by $\frac{\beta}{2}$, presents a Dirac delta at $\frac{\beta}{2}$ and is deformed by $\zeta(\hat{n})$ on $(-\infty, \frac{\beta}{2})$. It follows that the MAP estimator does not match the underlying model.

[5]First we compute $f(\hat{x}|x)$ and then $f(\hat{x}) = \int f(\hat{x}|x) f_X(x) dx$. Similarly, $f_{\hat{N}}(\hat{n}) = \int f(\hat{n}|x) f_X(x) dx$.

2.2. DISTRIBUTION OF THE MAP FOR GENERALIZED GAUSSIAN PRIORS. To see the practical importance of the example below, one can think of the MAP restoration of noisy wavelet coefficients that was sketched in (5)-(6) and (9). The statistical distribution of the (noise-free) wavelet coefficients (5) in real-world image data has been shown to be fairly described using generalized Gaussian (GG) distribution laws [36, 37, 6]

$$f_X(x) = \frac{1}{Z}e^{-\lambda|x|^\alpha},\ x \in \mathbb{R}, \tag{11}$$

for appropriate choices of the parameters $\lambda > 0$ and $\alpha > 0$. Under the usual assumption for i.i.d. Gaussian noise, the MAP estimate $\hat{u}_i$ of each noisy coefficient $\langle w_i, y\rangle$ in (9) is done independently, by minimizing a scalar function $\mathcal{E}_y : \mathbb{R} \to \mathbb{R}_+$ of the form

$$\mathcal{E}_y(x) = (x-y)^2 + \beta|x|^\alpha \text{ for } \beta = 2\sigma^2\lambda, \tag{12}$$

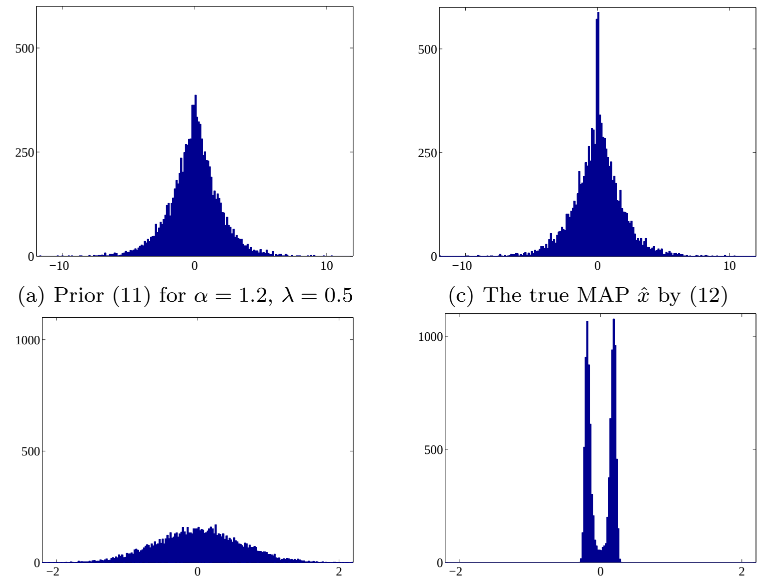
where we identify x with u_i, y with $\langle w_i, y\rangle$ and λ with λ_i. This is a situation where both the prior and the data-acquisition models are pertinent. It is then crucial to know how accurately the MAP estimate $\hat{x}$ fits these models. We address this question with the aid of numerical experiments.

For (α, λ) and σ fixed, we realize 10 000 independent trials. In each trial, an original $x \in \mathbb{R}$ is sampled from f_X and then $y = x + n$ for n sampled from f_N =Normal$(0, \sigma^2)$. After this, the true MAP solution $\hat{x}$ is calculated using (12). According to the value of α, the posterior distribution $f_{X|Y}(., y)$ has one or two modes.

(a) *Case* $\alpha \geq 1$. The results in Fig. 1 correspond to $\alpha = 1.2$, $\lambda = 0.5$ and $\sigma = 0.6$ which yields an SNR of 10 dB. The histograms of the samples x drawn from f_X and the samples n drawn from f_N are shown in Figs. 1 (a) and (b), respectively. For every $y \in \mathbb{R}$, the function $\mathcal{E}_y$ is strictly convex and has a unique minimizer $\hat{x}$. Unlike the prior f_X, the histogram of the MAP estimates $\hat{x}$ in all trials, plotted in Fig. 1 (c), is very concentrated in the vicinity of zero (even though $|\hat{x}| < 10^{-3}$ for only 2.35% of the trials). The histogram of the resultant noise estimates $\hat{n} = y - \hat{x}$ seen in Fig. 1 (d) is far from approximating f_N: it is clearly bounded while its value in the vicinity of zero is very small which means that almost all MAP solutions $\hat{x}$ are biased.

(b) *Case* $0 < \alpha < 1$. Fig. 2 corresponds to $\alpha = 0.5$, $\lambda = 2$ and $\sigma = 0.8$ in which case the SNR is 10.3 dB. The samples drawn from f_X and f_N are represented in Figs. 2(a) and (b), respectively. For every $y \neq 0$ the function $\mathcal{E}_y$ has two local minimizers[6], $\hat{x}_1 = 0$ and $\hat{x}_2$ such that $|\hat{x}_2| > \theta$ for $\theta = \left(\frac{2}{\alpha(1-\alpha)\beta}\right)^{\frac{1}{\alpha-2}} \approx 0.47$, and the global one is found by comparing $\mathcal{E}_y(\hat{x}_1)$ and $\mathcal{E}_y(\hat{x}_2)$. We deduce that the distribution $f_{\hat{X}}$ of the true MAP solution $\hat{x}$ contains a Dirac-delta at zero and is null on a subset containing $\big(-\theta, 0\big) \bigcup \big(0, \theta\big)$. The empirical histogram of all MAP estimates $\hat{x}$, shown in 2(c), reflects these two special features: we have $\hat{x} = 0$ in 77% of the trials while the smallest non-zero $|\hat{x}|$ is $0.77 > \theta$. The shape of this histogram is essentially different from the prior f_X. The resultant estimate of the noise $\hat{n}$ shown in (d) is clearly bounded on $\mathbb{R}$. The

[6] Since $\alpha \in (0, 1)$, we have $\varphi'(0^+) = +\infty$, hence for any $y \neq 0$, the subdifferential of $\mathcal{E}_y$ at zero contains the point $\hat{x}_1 = 0$. One can show that $\hat{x}_2$ has the same sign as $y \neq 0$ and that it satisfies $\frac{\partial}{\partial x}\mathcal{E}_y(x)|_{\hat{x}_2} = 0$. Using that $\mathcal{E}''_y(x) < 0$ if $|x| \in (0, \theta)$, no local minimizer can belong to $(-\theta, 0) \cup (0, \theta)$, hence the inequality on $\hat{x}_2$.



(a) Prior (11) for $\alpha = 1.2$, $\lambda = 0.5$ (c) The true MAP $\hat{x}$ by (12)

(b) Noise: Normal$(0, \sigma^2)$ for $\sigma = 0.6$ (d) The noise estimate $\hat{n} = y - \hat{x}$

Figure 1. Histograms for 10 000 independent trials, case $\alpha \geq 1$. Left column: realizations of the original models. Right column: the relevant true MAP solutions.

samples near the origin are better than in the case of Fig. 1: this suggests there are MAP solutions $\hat{x} \neq 0$ that are close to the relevant y.

Obviously, the MAP estimate does not fit neither the GG prior model nor the additive Gaussian noise model. This is especially unfortunate in a case when these models are accurate. We will see that the gap between the prior and the data-acquisition models on the one hand, and the *effective* distributions realized by the true MAP estimate on the other hand, is a permanent contradiction in Bayesian MAP estimation.

3. Non-smooth at zero priors

This section is devoted to MAP solutions corresponding to Gibbsian priors (3) where Φ is nonsmooth. We start with an example that gives a flavor of the kind of contradictions entailed by such priors.

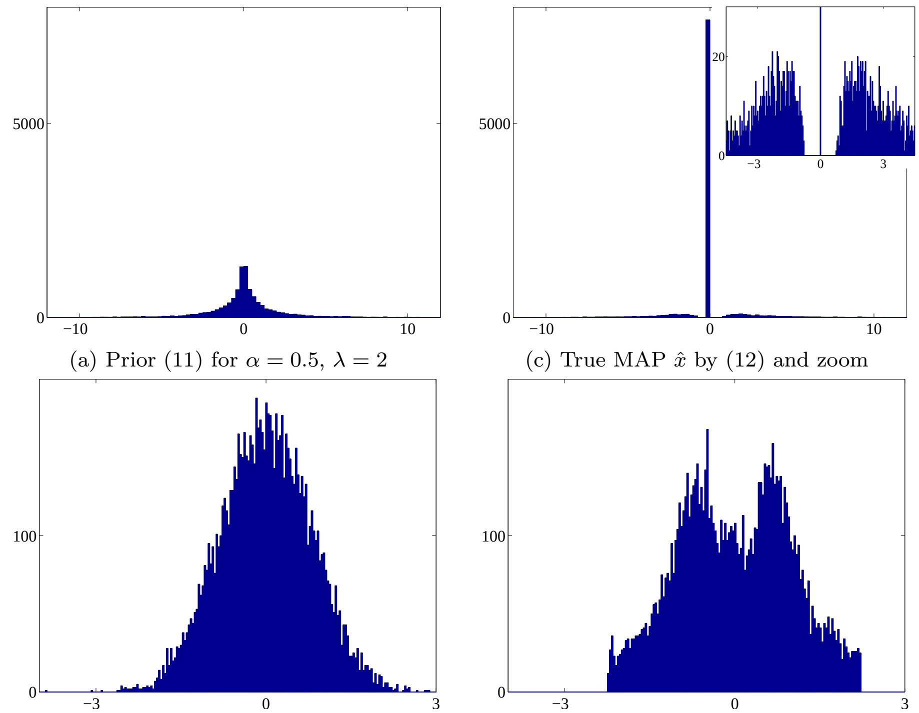
3.1. A Laplacian Markov chain corrupted with Gaussian noise.

The model for the true signal is a Markov chain with a Gibbsian distribution (3) where

$$\Phi(x) = \lambda \sum_{i=1}^{p-1} |x_i - x_{i+1}|,\ \lambda > 0. \tag{13}$$

Then the differences $X_i - X_{i+1}$, $1 \leq i \leq p-1$ are i.i.d. with the same Laplacian density:

$$f_{\Delta X}(t) = \frac{\lambda}{2} e^{-\lambda|t|},\ t \in \mathbb{R}. \tag{14}$$



(a) Prior (11) for $\alpha = 0.5$, $\lambda = 2$

(c) True MAP $\hat{x}$ by (12) and zoom

(b) Noise: Normal$(0, \sigma^2)$, $\sigma = 0.8$

(d) Noise estimate $\hat{n} = y - \hat{x}$

FIGURE 2. Histograms for 10 000 independent trials, case $0 < \alpha < 1$. Left column: realizations of the original models. Right side: the relevant true MAP solutions.

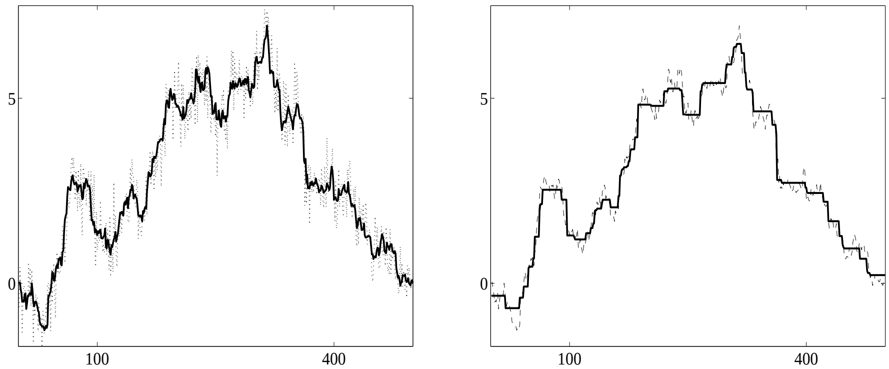
Since the prior defined by (13) is improper, any realization $X = x$ involves an arbitrary shifting constant which plays no role in what follows. The observed data read $Y = X + N$ where $f_N =$Normal$(0, \sigma^2 I)$. The posterior distribution $f_{X|Y}(x|y)$ is proper and reads

$$\begin{aligned} f_{X|Y}(x|y) &= \exp\left(-\frac{1}{2\sigma^2}\mathcal{E}_y(x)\right)\frac{1}{Z}, \\ \mathcal{E}_y(x) &= \|x-y\|^2 + \beta\sum_{i=1}^{p-1}\Big|x_i - x_{i+1}\Big|,\ \beta = 2\sigma^2\lambda. \end{aligned} \tag{15}$$

Coherence with the prior modeling done in (3) and (13) requires[7] that the solution $\hat{x}$, when $p = q$ is large enough, is such that the normalized empirical distribution of its differences, $\hat{x}_i - \hat{x}_{i+1}$, $1 \le i \le p-1$, approximates $f_{\Delta X}$ in (14). Similarly, the empirical distribution of the noise estimate $\hat{n}_i = y_i - \hat{x}_i$, $1 \le i \le q$, must approximate f_N.

The experiments presented next concern 500-length Laplacian Markov chains corresponding to $\lambda = 8$ and $\sigma = 0.5$, in which case $\beta = 4$ in (15). Realizations of an original $X = x$ and data $Y = y$ are shown in Fig. 3 (a). The solution $\hat{x}$—corresponding to the true parameters (λ, σ) used to generate x and y—is displayed

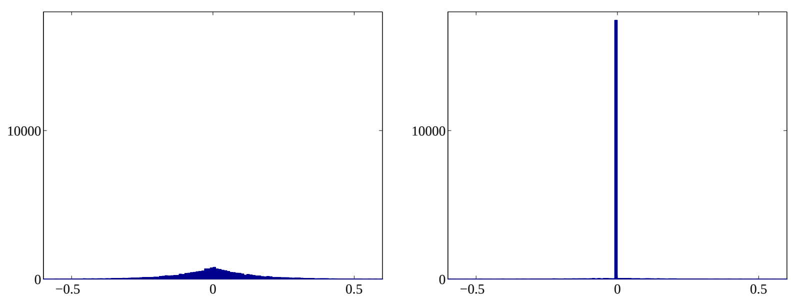
[7] We can notice that this requirement involves an ergodicity assumption which is easily admitted in this kind of modeling.



(a) Original x (—) for $\lambda = 8$ in (14) and data $y = x + n$ ($\cdots$) for $\sigma = 0.5$.

(b) The true MAP $\hat{x}$ (—) versus the original x (- - -).

Figure 3. The true MAP restoration of an original Laplacian Markov chain corrupted with white Gaussian noise.



(a) 40×499 differences $x_i - x_{i+1}$ sampled from (14) for $\lambda = 8$.

(b) The differences $\hat{x}_i - \hat{x}_{i+1}$ of the true MAP solutions.

Figure 4. Histograms for 40 trials with 500-length signals: the differences sampled from the prior $f_{\Delta X}$ and the differences of the true MAP solutions.

in Fig. 3 (b). Even though the original x has some slightly homogeneous regions, the restored $\hat{x}$ has a very different aspect since it is constant on many regions—almost 92% of its differences are null, $|\hat{x}_i - \hat{x}_{i+1}| < 10^{-30}$. *Visually, $\hat{x}$ is far from fitting the prior model.* Next we repeat the same experiment 40 times. The histogram of all original differences $x_i - x_{i+1}$, for all trials, is shown in Fig. 4 (a). The histogram of all restored differences $\hat{x}_i - \hat{x}_{i+1}$ in all trials is shown in Fig. 4 (b). It is quite dissimilar to the prior model since it contains a huge spike at zero—87% of all restored differences are null (their magnitude is less than 10^{-30}). Obviously, the MAP solution is far from fitting the prior.

The observed incoherence between the models f_X and $f_{Y|X}$ on the one hand, and the estimator $\hat{X}$ on the other hand, is inherent since it originates from the analytical properties of the MAP solution corresponding to nonsmooth prior energies combined with smooth data-acquisition models [42, 46]. This will be explained below.

3.2. Analytical results on the MAP and their statistical meaning. Let us more generally consider Gibbsian priors (3) where

$$\Phi(x) = \lambda \sum_{i=1}^{r} \varphi(\|G_i x\|), \tag{16}$$

G_i, $1 \le i \le r$, are linear operators (e.g. they can yield the finite differences or discrete derivatives of x) and $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ is an increasing $\mathcal{C}^m$-function such that

$$\varphi'(0) > 0.$$

By the latter condition, Φ is nonsmooth at any x such that $G_i x = 0$ for some $i \in \{1, \ldots, r\}$. Examples of such functions φ are (f8)-(f12) in Table 1. The prior model corresponding to (16) reads

$$f_X(x) \propto \prod_{i=1}^{r} e^{-\lambda\varphi(\|G_i x\|)}.$$

Suppose that the observed data Y correspond to a likelihood $f_{Y|X}(y|x) \propto e^{-\Psi(x,y)}$ where Ψ is a $\mathcal{C}^m$-function, $m \ge 2$, such that the posterior distribution $f_{X|Y}(.|y)$ is proper. The likelihood $f_{Y|X}$ can for instance be of the form (2) for a $\mathcal{C}^m$ function f_N. The MAP estimator $\hat{X}$ then minimizes

$$\mathcal{E}_y(x) = \Psi(x, y) + \lambda\Phi(x).$$

The result below comes from [42] and [46] (Theorems 6.1 and 2, respectively). In the following, $\delta\mathcal{E}_y(x)(u)$ will denote the one-sided derivative[8] of $\mathcal{E}_y$ at x in the direction of $u \ne 0$.

Theorem 3.1. *Given $y \in \mathbb{R}^q$, let $\hat{x} \in \mathbb{R}^p$ be such that if we put*

$$\begin{aligned} J &= \Big\{ i \in \{1, \ldots, r\} : G_i \hat{x} = 0 \Big\}, \\ K_J &= \Big\{ u \in \mathbb{R}^p : G_i u = 0,\ \forall i \in J \Big\}, \end{aligned} \tag{17}$$

we have

(a) *$\delta\mathcal{E}_y(\hat{x})(u) > 0$ for every $u \in K_J^{\perp} \setminus \{0\}$;*
(b) *$D\mathcal{E}_y|_{K_J}(\hat{x})u = 0$ and $D^2\mathcal{E}_y|_{K_J}(\hat{x})(u, u) > 0$, for every $u \in K_J \setminus \{0\}$.*

Then $\mathcal{E}_y$ has a strict (local) minimum at $\hat{x}$. Moreover, there are a neighborhood O_J of y and a continuous function $\mathcal{X} : O_J \to \mathbb{R}^p$ such that $\mathcal{X}(y) = \hat{x}$ and that for every $y' \in O_J$, $\mathcal{E}_{y'}$ has a (local) minimum at $\hat{x}' = \mathcal{X}(y')$ satisfying

$$G_i \hat{x}' = 0,\ \forall i \in J,$$

or equivalently, that $\hat{x}' \in K_J$ for every $y' \in O_J$.

Conditions (a) and (b) ensure that $\mathcal{E}_y$ has a strict local minimum at $\hat{x}$. They are quite general, as confirmed by the following result considering a linear Gaussian measurement model (see Proposition 2 and 3 in [17]).

Proposition 1. *Let $\Psi(x, y) = \frac{1}{2\sigma^2}\|Ax - y\|^2$ with $A^T A$ invertible. Define $\Omega \subset \mathbb{R}^q$ to be such that if $y \in \Omega$ then every (local) minimizer $\hat{x}$ of $\mathcal{E}_y$ is strict, and that (a) and (b) in Theorem 3.1 hold. Then*

[8] The one-sided derivative $\delta\mathcal{E}_y(x)(u)$ exists if the following limit (possibly infinite) exists: $\delta\mathcal{E}_y(x)(u) = \lim_{t \downarrow 0} \frac{\mathcal{E}_y(x+tu) - \mathcal{E}_y(x)}{t}$. This holds under mild conditions [28, 50] that are satisfied by the energies used in practice. If $\mathcal{E}_y$ is differentiable at x, then $\delta\mathcal{E}_y(x)(u) = D\mathcal{E}_y(x)u$ for any u.

(i) Ω^c *(the complement of* Ω *in* $\mathbb{R}^q$*) is of Lebesgue measure zero;*
(ii) *if in addition* $\lim_{t\to\infty}\varphi'(t)/t = 0$*, then the closure of* Ω^c *is of Lebesgue measure zero as well.*

The requirement on φ' in (ii) is standard for edge-preserving signal and image restoration methods, see for instance [14, 2]. It holds for all functions φ in Table 1, except for (f1) with $\alpha = 2$ which is not edge-preserving.

It is essential to notice that O_J contains an open subset of $\mathbb{R}^q$ and that the theorem addresses many nonempty J. By Theorem 3.1,

$$y \in O_J \text{ and } \hat{x} = \arg\max_{x\in\mathbb{R}^p} f_{X|Y}(x|y) \;\Rightarrow\; G_i\hat{x} = 0,\ \forall i \in J, \tag{18}$$

or equivalently $\hat{x} \in K_J$. Then the probability to have $\hat{X} \in K_J$ satisfies

$$\Pr(\hat{X} \in K_J) \geq \Pr(Y \in O_J) = \int_{O_J} f_Y(y)dy > 0.$$

The strict positivity of the integral above comes from the facts that O_J contains an open subset of $\mathbb{R}^q$ and that

$$f_Y(y) = \int f_{Y|X}(y|x)f_X(x)dx = \frac{1}{Z}\int e^{-\mathcal{E}_y(x)}dx > 0,\ \forall y \in \mathbb{R}^q. \tag{19}$$

The model on the unknown X which is *effectively* realized by the MAP estimator $\hat{X}$ hence corresponds to images and signals such that $G_i\hat{X} = 0$ for a certain number of indexes i. If $\{G_i\}$ are first-order differences or discrete gradients, then we have an *effective prior model for locally constant images and signals*. This is in contradiction with the prior model f_X involved in $\mathcal{E}_y$. The function f_X being continuous, for any nonempty $J \subset \{1,\dots,r\}$ the probability that $X \in K_J$ is null:

$$\Pr(X \in K_J) = \int_{K_J} f_X(x)dx = 0,$$

since $K_J \subset \mathbb{R}^p$ is a subspace of $\mathbb{R}^p$ of dimension strictly smaller than p.

Laplacian Markov chain. In the case of a linear Gaussian measurement model with A invertible and a Laplacian Markov chain prior as in (13), we have $f_{X|Y}(x|y) \propto \exp(-\mathcal{E}_y(x)) + const$ for

$$\mathcal{E}_y(x) = \|Ax - y\|^2 + \beta\sum_{i=1}^{p-1}|x_i - x_{i+1}|,\ \beta = 2\sigma^2\lambda. \tag{20}$$

The following striking phenomena then occur (see [46] for details):

(a) for every $\hat{x} \in \mathbb{R}^p$, there is a polyhedron $Q_{\hat{x}} \subset \mathbb{R}^q$ of dimension $\#J$ for $J = \{i : G_i\hat{x} = 0\}$, such that for every $y \in Q_{\hat{x}}$, the same point $\hat{x}$ is the unique minimizer of $\mathcal{E}(., y)$;
(b) for every $J \subset \{1,\dots,p-1\}$, there is a subset $\tilde{O}_J \subset \mathbb{R}^q$, composed of $2^{n-\#J-1}$ unbounded polyhedra of $\mathbb{R}^q$, such that for every $y \in \tilde{O}_J$, the minimizer $\hat{x}$ of $\mathcal{E}_y$ satisfies $\hat{x}_i = \hat{x}_{i+1}$ for all $i \in J$ and $\hat{x}_i \neq \hat{x}_{i+1}$ for all $i \in J^c$. Moreover, their closure forms a covering of $\mathbb{R}^q$.

As a consequence, for every $J \subset \{1,\dots,p-1\}$ we have

$$\Pr\big(\hat{X}_i = \hat{X}_{i+1}, \forall i \in J\big) \geq \Pr\big(Y \in \tilde{O}_J\big) > 0.$$

These are solutions composed of constant pieces. However, the prior model involved in (20) yields $\Pr\big(X_i = X_{i+1}\big) = 0$ for every $i \in \{1,\dots,p-1\}$.

4. Non-smooth at zero noise models

Consider a linear measurement model (1) corrupted with i.i.d. additive noise N as in (2) where

$$f_N(t) = \frac{1}{Z} e^{-\sigma\psi(t)},\ t \in \mathbb{R}, \tag{21}$$

$\sigma > 0$ is a parameter and $\psi : \mathbb{R} \to \mathbb{R}$ is $\mathcal{C}^m$, $m \geq 2$, on $\mathbb{R} \setminus \{0\}$ and nonsmooth at zero, such that

$$0 < \psi'(0^+) = -\psi'(0^-) < \infty. \tag{22}$$

By (2), the likelihood function is $f_{Y|X}(y|x) \propto \exp(-\sigma\Psi(x,y))$ where

$$\Psi(x,y) = \sum_{i=1}^{q} \psi(a_i^T x - y_i). \tag{23}$$

Then Ψ is nonsmooth at any (x,y) such that $a_i^T x = y_i$ for some $i \in \{1,\ldots,q\}$. If N is Laplacian i.i.d. noise, $\psi(t) = |t|$ which leads to an ℓ_1 data-fidelity term $\Psi(x,y) = \|Ax - y\|_1$. We can notice that even though ψ is non-smooth, f_N is a continuous function, hence $\Pr\big(N_i = 0\big) = 0$ for every $i \in \{1,\ldots,q\}$.

Furthermore, let X correspond to a Gibbsian prior (3) where $\Phi : \mathbb{R}^p \to \mathbb{R}$ is a $\mathcal{C}^m$-function. For instance, Φ can be of the form (4) or (16) where φ is any $\mathcal{C}^m$ function in Table 1. Given $y \in \mathbb{R}^q$, the MAP solution $\hat{x}$ minimizes $\mathcal{E}_y$ given below

$$\mathcal{E}_y(x) = \Psi(x,y) + \beta\Phi(x),\ \beta = \frac{\lambda}{\sigma}. \tag{24}$$

We will start with a numerical example.

4.1. Generalized Gaussian Markov chain under Laplace noise. Let X be a 100-length Markov chain whose differences $X_i - X_{i+1} \sim f_{\Delta X}$, $1 \leq i \leq p-1$ are i.i.d. and $f_{\Delta X}$ is a GG density

$$f_{\Delta X}(t) = \frac{1}{Z} e^{-\lambda|t|^\alpha},\ t \in \mathbb{R}. \tag{25}$$

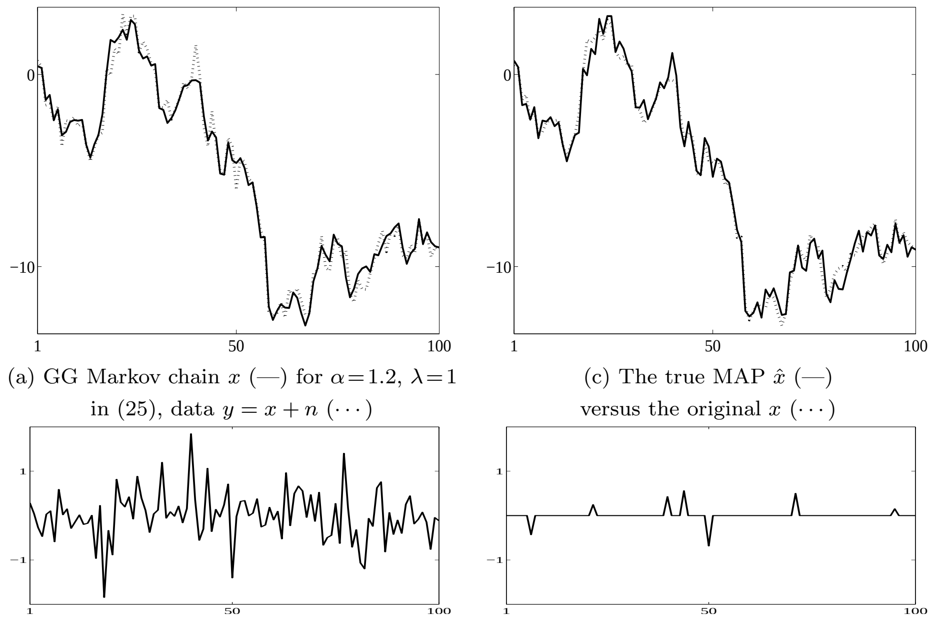
Suppose we have data $Y = X + N$ where N_i, $1 \leq i \leq p$ are i.i.d. with marginal density

$$f_N(t) = \frac{\sigma}{2} e^{-\sigma|t|},\ t \in \mathbb{R}. \tag{26}$$

The posterior distribution $f_{X|Y}$ is proper and reads

$$\begin{aligned} f_{X|Y}(x|y) &= \exp\Big(-\sigma\mathcal{E}_y(x)\Big)\frac{1}{Z}, \\ \mathcal{E}_y(x) &= \sum_{i=1}^{p} \big|x_i - y_i\big| + \beta\sum_{i=1}^{p-1} |x_i - x_{i+1}|^\alpha \text{ where } \beta = \frac{\lambda}{\sigma}. \end{aligned}$$

The experiments presented next correspond to $\alpha = 1.2$, $\lambda = 1$ and $\sigma = 2.5$ in which case $\beta = 0.4$ and $f_{X|Y}(.,y)$ has a unique mode. Realizations of an original $X = x$ and data $Y = y$ are shown in Fig. 5 (a) while the noise $N = n$ contained in the data is plotted in Fig. 5 (b). Notice that $x_i \neq y_i$ for all i (more precisely, $|x_i - y_i| > 0.04$ for all $i \in \{1,\ldots,100\}$ in this experiment). The MAP solution $\hat{x}$ obtained for the true value of β, shown in Fig. 5 (c), contains 93% samples satisfying $\hat{x}_i = y_i$. Obviously, $\hat{x}$ does not have the aspect of a GG Markov chain. Correspondingly, the



(a) GG Markov chain x (—) for $\alpha = 1.2$, $\lambda = 1$ in (25), data $y = x + n$ $(\cdots)$

(c) The true MAP $\hat{x}$ (—) versus the original x $(\cdots)$

(b) Laplacian i.i.d. noise n for $\sigma = 2.5$ in (26)

(d) The noise estimate $\hat{n} = y - \hat{x}$.

Figure 5. The true MAP restoration of an original GG Markov chain x from data $y = x + n$ corrupted with white Laplacian noise.

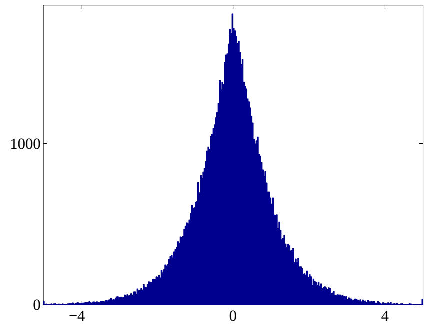
estimate of the noise $\hat{n} = y - \hat{x}$, shown in Fig. 5 (d), is far from approximating a Laplacian noise since it involves only 7% non-zero samples.

Next, we repeat the same experiment 1000 times. Fig. 6 (a) shows the histogram of all 99×10^3 differences $x_i - x_{i+1}$ sampled from $f_{\Delta X}$ in (25) for $\alpha = 1.2$ and $\lambda = 1$ in order to form 1000 original GG Markov chains x. Below in Fig. 6 (b) one can see the histogram of all the 10^4 Laplacian noise samples n_i generated by f_N in (26) for $\sigma = 2.5$, used to form 1000 noise vectors n. The true MAP solution $\hat{x}$ is then computed for each data set $y = x + n$. The histogram of all differences $\hat{x}_i - \hat{x}_{i+1}$ of the MAP solutions in all trials is shown in Fig. 6 (c), while the histogram of the samples $\hat{n}_i = y_i - \hat{n}_i$ of the resultant noise estimates is seen in Fig. 6 (d). For 87% of the samples in all trials, $\hat{x}_i = y_i$, hence the huge spike at zero in Fig. 6 (d). This means that most of the samples $\hat{x}_i$ of the MAP solution keep the noise intact. Correspondingly, many differences of the MAP solutions read $\hat{x}_i - \hat{x}_{i+1} = x_i - x_{i+1} + n_i - n_{i+1}$: even if originally $x_i - x_{i+1} \approx 0$, for the MAP $\hat{x}_i - \hat{x}_{i+1}$ is no longer close to zero. Hence the flattening of the histogram in Fig. 6 (c) near the origin.

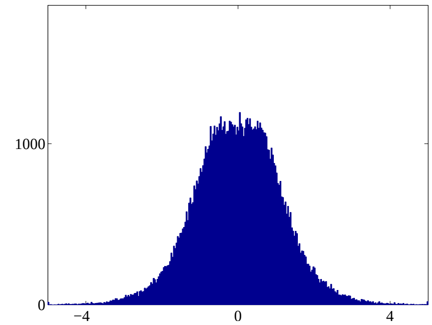
4.2. Main analytical result and statistical interpretation. We consider posterior energies of the form (24) and posit the assumptions made in the introduction of section 4. The result stated next, established in [44], is the key to explain the behavior observed in Figs. 5 and 6.

Theorem 4.1. *Given $y \in \mathbb{R}^q$, suppose that $\hat{x} \in \mathbb{R}^p$ is such that if we put*

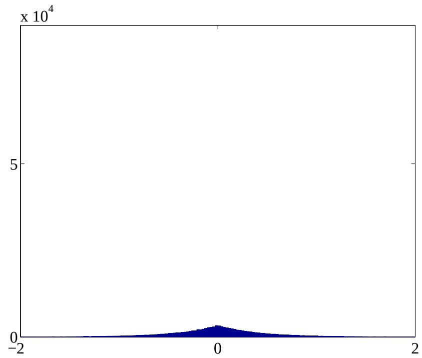
$$\begin{aligned} J &= \big\{ i \in \{1, \ldots, q\} : a_i^T \hat{x} = y_i \big\}, \\ K_J &= \{ u \in \mathbb{R}^p : \ a_i^T u = 0, \ \forall i \in J \}, \end{aligned}$$



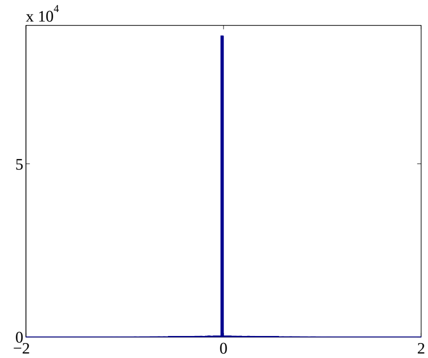
(a) All original differences $x_i - x_{i+1}$ sampled from $f_{\Delta X}$ in (25) for $\alpha = 1.2$ and $\lambda = 1$



(c) The differences $\hat{x}_i - \hat{x}_{i+1}$ of the true MAP solutions $\hat{x}$.



(b) Laplacian i.i.d. noise by (26) for $\sigma = 2.5$.



(d) All the residuals $y - \hat{x}$.

FIGURE 6. Histograms for 1000 independent trials with 100-length signals. Left column: samples from the original models $f_{\Delta X}$ and f_N. Right column: the models *effectively* realized by the MAP—$\hat{x}$ and the resultant noise estimate $\hat{n} = y - \hat{x}$.

we have:

(a) *the set* $\{a_i : i \in J\}$ *is linearly independent;*
(b) $D\mathcal{E}_y|_{\hat{x}+K_J}(\hat{x})u = 0$ *and* $D^2\mathcal{E}_y|_{\hat{x}+K_J}(\hat{x})(u,u) > 0$*, for every* $u \in K_J \setminus \{0\}$*;*
(c) $\delta\mathcal{E}_y(\hat{x})(u) > 0$*, for every* $u \in K_J^{\perp} \setminus \{0\}$*.*

Then $\mathcal{E}_y$ *has a strict (local) minimum at* $\hat{x}$*. Moreover, there are a neighborhood* $O_J \subset \mathbb{R}^q$ *containing* y *and a* $\mathcal{C}^{m-1}$ *function* $\mathcal{X} : O_J \to \mathbb{R}^p$ *such that for every* $y' \in O_J$*, the function* $\mathcal{E}_{y'}$ *has a (local) minimum at* $\hat{x}' = \mathcal{X}(y')$ *and*

$$\begin{aligned} a_i^T \hat{x}' &= y_i' \quad \text{if} \quad i \in J, \\ a_i^T \hat{x}' &\neq y_i' \quad \text{if} \quad i \in J^c. \end{aligned} \tag{27}$$

Hence $\mathcal{X}(y') \in \hat{x} + K_J$ *for every* $y' \in O_J$*.*

It is shown in [44] that for any $A \in \mathbb{R}^{q \times p}$ the assumption (a) holds for all $y \in \mathbb{R}^q$ except those included in a subspace of dimension strictly smaller than q. Hence the probability that (a) fails is null. Noticing that $\mathcal{E}_y|_{\hat{x}+K_J}$ is $\mathcal{C}^m$ near $(\hat{x}, y)$, (b) is the classical sufficient condition for a strict local minimum of a smooth function. Next,

(c) is a weak condition ensuring that $\mathcal{E}_y$ has a strict local minimum at $\hat{x}$ along all non-zero directions in $K_J^\perp$.

A crucial consequence is that O_J contains an open subset of $\mathbb{R}^q$ and that we have many nonempty J when y ranges on $\mathbb{R}^q$. By Theorem 4.1, the distribution of the MAP estimator $\hat{X}$ is such that

$$\Pr\left(a_i^T\hat{X} - Y_i = 0\right) \geq \Pr\left(Y \in O_J\right) = \int_{O_J} f_Y(y)dy > 0,\ \forall i \in J,$$

where the last inequality comes from (19). For all $i \in J$, the prior has no influence on the solution and the noise remains intact. This result contradicts the model for the noise assumed in (21) since by the continuity of ψ we have

$$\Pr\left(a_i^T X - Y_i = 0\right) = \Pr\left(N_i = 0\right) = 0,\ \forall i \in \{1, \ldots, q\}.$$

For simplicity, consider now that A is invertible and that Φ is of the form (16). Define $O_\infty \subset \mathbb{R}^p$ by

$$O_\infty = \left\{y \in \mathbb{R}^p : \|D\Phi(A^{-1}y)\| < \frac{\psi'(0^+)}{\beta} \min_{\|u\|=1} \sum_{i=1}^p |a_i^T u|\right\}.$$

The set O_∞ is clearly non-empty and contains an open subset of $\mathbb{R}^q$. Consequently,

$$\Pr(A\hat{X} = Y) \geq \Pr(Y \in O_\infty) > 0.$$

It is amazing to see that[9]

$$y \in O_\infty \ \Rightarrow\ a_i^T\hat{x} = y_i,\ \forall i \in \{1 \ldots, n\},$$

i.e. that $\hat{x} = A^{-1}y$ which means that the prior has no influence on the solution. This property violates the prior model.

4.3. A LAPLACE NOISE MODEL TO REMOVE IMPULSE NOISE. We confine our attention to an important class of posterior energies for denoising (then $p = q$). For any $y \in \mathbb{R}^p$, let us consider the minimization of $\mathcal{E}_y$ below:

$$\mathcal{E}_y(x) = \sum_{i=1}^p |x_i - y_i| + \frac{\beta}{2}\sum_i \sum_{j \in \mathcal{N}_i} \varphi(x_i - x_j), \tag{28}$$

where φ is a symmetric $\mathcal{C}^1$ strictly convex edge-preserving function, e.g. (f1)-(f3) in Table 1. From a Bayesian standpoint, the ℓ_1 data-fidelity in (28) corresponds to data $Y = X + N$ where N is Laplacian white noise, $f_{Y|X}(x|y) = \prod_{i=1}^p f_N(y_i - x_i)$ for f_N as in (26), while the prior distribution is of the form (3) for $\beta = \lambda/\sigma$. However, Theorem 4.1 and the example in § 4.1 have shown that the MAP solution $\hat{x}$ cannot efficiently clean Laplacian noise since all $\hat{x}_i$ such that $\hat{x}_i = y_i$ keep the noise intact, $\hat{x}_i = x_i + n_i$ while $n_i \neq 0$ almost surely. Instead, we will describe the noise model which is *effectively* realized by the MAP estimator defined by (28).

Our reasoning is based upon the conditions for a minimum of $\mathcal{E}_y$. More precisely, $\mathcal{E}_y$ reaches its minimum[10] at a point $\hat{x} \in \mathbb{R}^p$ if, and only if, for $J = \Big\{i \in \{1, \ldots, p\}$:

[9] Assumption (a) of Theorem 4.1 holds, (b) is void since $K_J^\perp = \mathbb{R}^p$ for $J = \{1, \ldots, q\}$ and (c) is satisfied since $\forall y \in O_\infty$, $\hat{x} = A^{-1}y$ and $\|u\| = 1$, hence $\delta\mathcal{E}_y(\hat{x})(u) = \psi'(0^+)\sum_{i=1}^q |a_i^T u| + \beta D\Phi(\hat{x})u \geq \psi'(0^+)\min_{\|u\|=1}\sum_{i=1}^q |a_i^T u| + \beta D\Phi(\hat{x})u > \beta\|D\Phi(\hat{x})\|\,\|u\| + \beta D\Phi(\hat{x})u \geq 0$.

$\left.\hat{x}_i = y_i\right\}$ we have

$$i \in J \quad \Rightarrow \quad \Big|\sum_{j\in\mathcal{N}_i} \varphi'(y_i - \hat{x}_j)\Big| \leq \frac{1}{\beta}, \tag{29}$$

$$i \in J^c \quad \Rightarrow \quad \sum_{j\in\mathcal{N}_i} \varphi'(\hat{x}_i - \hat{x}_j) = \frac{\sigma_i}{\beta}, \tag{30}$$

$$\sigma_i = \text{sign}\Big(\sum_{j\in\mathcal{N}_i} \varphi'(y_i - \hat{x}_j)\Big) \in \{-1, 1\}. \tag{31}$$

The details can be found in [45] (see Theorem 1 and Corollary 1 there). These conditions underly the next Proposition 2 which reinforces the result of Theorem 4.1 in the context of (28). Its proof is outlined in the Appendix.

Proposition 2. *Let $\beta > 1$ and $\varphi''(t) > 0$ for all $t \in \mathbb{R}$. Choose a nonempty $J \subset \{1, \ldots, p\}$ as well as $\sigma_i \in \{-1, 1\}$ for every $i \in J^c$. Then there are $y \in \mathbb{R}^p$ and $\rho > 0$ such that if O_J reads*

$$O_J = \left\{ y' \in \mathbb{R}^p : \left| \begin{array}{ll} |y'_i - y_i| \leq \rho & \forall i \in J \\ \sigma_i y'_i \geq \sigma_i y_i - \rho & \forall i \in J^c \end{array} \right. \right\},$$

then for every $y' \in O_J$ the function $\mathcal{E}_{y'}$ reaches its minimum at an $\hat{x}' \in \mathbb{R}^p$ such that

$$\hat{x}'_i = y'_i, \quad \forall i \in J, \tag{32}$$

$$\hat{x}'_i = \mathcal{X}_i(\{y'_i : i \in J\}), \quad \forall i \in J^c, \tag{33}$$

where $\mathcal{X}_i$, $i \in J^c$ are continuous functions that depend only on $\{y'_i : i \in J\}$.

For every $J \in \{1, \ldots, p\}$ the set O_J contains an open subset of $\mathbb{R}^p$, hence $\Pr(Y \in O_J) > 0$. Using that the sets O_J are disjoint, we can write that

$$\Pr(\hat{X}_i - Y_i = 0) \geq \sum_{\{J : i \in J\}} \Pr(Y \in O_J) > 0, \ \forall i \in \{1, \ldots, p\}.$$

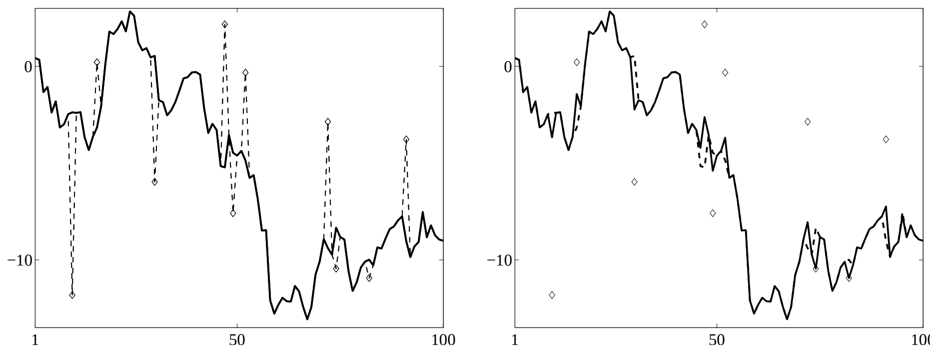
This result contradicts the Laplacian noise model involved in (28) since the latter implies

$$\Pr(X_i - Y_i = 0) = 0, \ \forall i \in \{1, \ldots, p\}.$$

By (32), the data samples y'_i, $i \in J$ are fitted exactly, hence they must be free of noise in the effective noise model realized by the MAP solution. These data samples y'_i, $i \in J$ satisfy (29). Since φ' is increasing and $\varphi'(0) = 0$, and since O_J is open, (29) shows that a noise-free sample y'_i for $i \in J$ can be dissimilar with respect to its neighbors only up to some degree that depends on β. Otherwise, if y'_i is too dissimilar with respect to its neighbors, then $i \in J^c$ and according to (30) its value is replaced by the estimate $\hat{x}'_i = \mathcal{X}_i(\{y'_i : i \in J\})$ which depends only on the noise-free data samples. The samples y'_i for $i \in J^c$ are hence *outliers* that can take any value on the half-line contained in O_J. In conclusion, *the MAP estimator defined by (28) corresponds in fact to an impulse noise model on the data.*

By way of illustration, let us consider again the GG Markov chain x plotted in Fig. 5 in § 4.1 along with data y containing 10% random-valued impulse noise in the range $[\min_i x_i, \max_i x_i]$. Both x and y are plotted in Fig. 7 (a), the former with a solid line and the latter with a dashed line. The minimizer $\hat{x}$ of $\mathcal{E}_y$ in (28) for $\beta = 0.4$

[10] One can show that the minimizer $\hat{x}$ is unique if J is nonempty and that it is reached on a bounded subset if $J = \emptyset$.



(a) Original x (—), data y (- - -) with 10% random valued impulse noise.

(b) The minimizer $\hat{x}$ of $\mathcal{E}_y$ for $\beta = 0.4$ (—), the original x (- - -), and $y_i \neq x_i$ ($\diamond$) .

Figure 7. Restoration of a GG Markov chain corrupted with impulse noise by minimizing an energy $\mathcal{E}_y$ involving an ℓ_1 data-fidelity term (a Laplacian noise model).

is presented in Fig. 7 (b) with a solid line. Almost all noisy samples are restored well while $\hat{x}_i = y_i$ for 89 among all the 90 noise-free samples (i.e. 99%). Indeed, the restoration of images corrupted with impulse noise using energy functions with an ℓ_1 data-fidelity term was considered in [45, 4, 5].

5. Priors with non-convex energies

Let us now consider a linear model for the data (1) with $N \sim \text{Normal}(0, \sigma^2 I)$ and a Gibbsian prior (3) with a nonconvex prior energy Φ

$$\Phi(x) = \sum_{i=1}^{r} \varphi(g_i^T x), \tag{34}$$

where $g_i \in \mathbb{R}^p$, $1 \leq i \leq r$, are difference operators and $\varphi : \mathbb{R} \to \mathbb{R}_+$ is *nonconvex*, see Table 1 for examples. More precisely, we will assume that φ is symmetric, $\mathcal{C}^2$ and increasing on $(0, +\infty)$ with a strict minimum at zero, and that there is $\theta > 0$ such that $\varphi''(\theta) < 0$ and $\lim_{t\to\infty} \varphi''(t) = 0$. Given $y \in \mathbb{R}^q$, the MAP solution $\hat{x}$ is the (global) minimizer $\hat{x}$ of a posterior energy $\mathcal{E}_y$ of the form

$$\mathcal{E}_y(x) = \|Ax - y\|^2 + \beta\Phi(x), \text{ where } \beta = 2\sigma^2\lambda. \tag{35}$$

Since the inaugural work of Geman and Geman [23], nonconvex functions φ are used to produce solutions $\hat{x}$ comprising well smoothed regions and sharp edges. Various nonconvex prior energies have been considered in the literature, e.g. [24, 41, 9, 48, 21, 22, 34, 11, 2].

5.1. Piecewise Gaussian Markov chain in Gaussian noise.

A famous prior model, that was the object of a huge amount of studies during the last 20 years, is the piecewise Gaussian Markov chain [23], known also as the discrete one-dimensional Mumford-Shah model [40], or the weak-string model [12]. According to this model, X is such that its differences $X_{i+1} - X_i$, $1 \leq i \leq p-1$, are i.i.d. with distribution

$f_{\Delta X}(t) \propto e^{-\lambda\varphi(t)}$ where

$$\varphi(t) = \begin{cases} \alpha t^2 & \text{if } |t| < \sqrt{\frac{1}{\alpha}}, \\ 1 & \text{else.} \end{cases} \tag{36}$$

In a closed form, $\varphi(t) = \min\{\alpha t^2, 1\}$. This φ does not satisfy the general assumptions stated in the beginning of this section. By (36), $f_{\Delta X}$ describes a Gaussian distribution truncated on $\left(-\sqrt{\frac{1}{\alpha}}, \sqrt{\frac{1}{\alpha}}\right)$ and a uniform distribution on $\left(-\gamma, -\sqrt{\frac{1}{\alpha}}\right) \cup \left(\sqrt{\frac{1}{\alpha}}, \gamma\right)$ for an arbitrarily fixed $\gamma > \sqrt{\frac{1}{\alpha}}$. The histogram of 200×299 random variables sampled from $f_{\Delta X}$ for $\alpha = 1$, $\lambda = 5$ and $\gamma = 15$ is seen in Fig. 8. The posterior energy $\mathcal{E}_y$ is of the form (35) where Φ reads

$$\Phi(x) = \sum_{i=1}^{p-1} \varphi(x_i - x_{i+1}).$$

Our reasoning is based on the following result (see [43] for the details).

Theorem 5.1. *For every $i \in \{1, \ldots, p-1\}$, define $u_i \in \mathbb{R}^p$ by $u_i[j] = 0$ if $1 \le j \le i$ and $u_i[j] = 1$ if $i+1 \le j \le m$. Let $P = I - \frac{A\mathbb{1}\mathbb{1}^T A^T}{\|A\mathbb{1}\|^2}$ denote the orthogonal projector onto $\{\mathbb{1}\}^\perp$.*
If $\mathcal{E}_y$ reaches its global minimum at $\hat{x}$, then for every $i \in \{1, \ldots, p-1\}$ we have:

$$\text{either } \; |\hat{x}_i - \hat{x}_{i+1}| \le \frac{1}{\sqrt{\alpha}}\,\Gamma_i \;\text{ or }\; |\hat{x}_i - \hat{x}_{i+1}| \ge \frac{1}{\sqrt{\alpha}\,\Gamma_i}, \tag{37}$$

where

$$\Gamma_i = \sqrt{\frac{\|PAu_i\|^2}{\|PAu_i\|^2 + \alpha\beta}} < 1.$$

In particular, $\hat{x}_i - \hat{x}_{i+1} = 0$ if $PAu_i = 0$.

By the theorem, for any realization $Y = y$, no difference $\hat{X}_i - \hat{X}_{i+1}$ of the MAP solution can have its magnitude in the interval $(\frac{\Gamma_i}{\sqrt{\alpha}}, \frac{1}{\sqrt{\alpha}\Gamma_i})$, hence

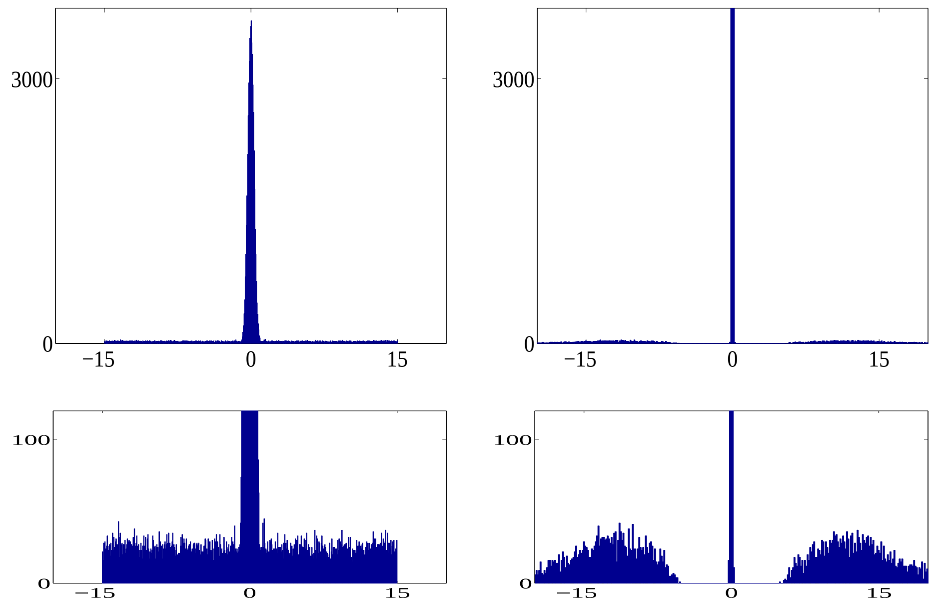
$$\Pr\left(\frac{\Gamma_i}{\sqrt{\alpha}} < |\hat{X}_i - \hat{X}_{i+1}| < \frac{1}{\sqrt{\alpha}\Gamma_i}\right) = 0, \ \forall i \in \{1, \ldots, p-1\}.$$

The sample space of each $\hat{X}_i - \hat{X}_{i+1}$ is disconnected since it is included in $\mathbb{R} \setminus \left\{(-\frac{1}{\sqrt{\alpha}\Gamma_i}, -\frac{\Gamma_i}{\sqrt{\alpha}}) \cup (\frac{\Gamma_i}{\sqrt{\alpha}}, \frac{1}{\sqrt{\alpha}\Gamma_i})\right\}$. The distribution of the MAP estimator $\hat{X}$ is definitely dissimilar to the prior model f_X since for the latter[11]

$$\Pr\left(\frac{\Gamma_i}{\sqrt{\alpha}} < |X_i - X_{i+1}| < \frac{1}{\sqrt{\alpha}\Gamma_i}\right) > 0, \ \forall i \in \{1, \ldots, p-1\}. \tag{38}$$

To illustrate the theorem, we repeat 200 times the following experiment. We generate an original $X = x$ of length $p = 300$ whose differences $x_i - x_{i+1}$, $1 \le i \le p-1$ are sampled from $f_{\Delta X}$ for $\alpha = 1$, $\lambda = 5$ and $\gamma = 15$. The histogram

[11] More precisely, (36) yields $\Pr\left(\frac{\Gamma_i}{\sqrt{\alpha}} < |X_i - X_{i+1}| \le \frac{1}{\sqrt{\alpha}}\right) = \frac{2}{Z}\int_{\frac{\Gamma_i}{\sqrt{\alpha}}}^{\frac{1}{\sqrt{\alpha}}} e^{-\lambda\alpha t^2} dt > 0$ and $\Pr\left(\frac{1}{\sqrt{\alpha}} < |X_i - X_{i+1}| < \frac{1}{\sqrt{\alpha}\Gamma_i}\right) = \frac{2e^{-\lambda}}{Z\sqrt{\alpha}}\left(\frac{1}{\Gamma_i} - 1\right) > 0$, where $Z = 2\left(e^{-\lambda}\left(\gamma - \frac{1}{\sqrt{\alpha}}\right) + \int_0^{\frac{1}{\sqrt{\alpha}}} e^{-\alpha\lambda t^2} dt\right)$. Hence the inequality in (38).



(a) Histogram of all original differences $x_i - x_{i+1}$ (up) and zoom (bottom).

(b) Histogram of the differences for all the true MAP solutions $\hat{x}$ (up) and zoom (bottom).

Figure 8. First row: Distribution of the differences $x_i - x_{i+1}$ of the original signals (left) and of the global minimizers $\hat{x}_i - \hat{x}_{i+1}$ (right). Second row: Zoom of the above distributions.

of all original differences $x_i - x_{i+1}$ in all trials is shown in Fig. 8(a). For each original x we generate $y = x + n$ where n is sampled from Normal$(0, \sigma^2 I)$ with $\sigma = 4$. Then the global minimizer $\hat{x}$ of $\mathcal{E}_y$ is computed for the true value of the parameter $\beta = 2\sigma^2\lambda = 160$. As predicted by Theorem 5.1, no difference of any MAP solution $\hat{x}$ has its magnitude $|\hat{x}_i - \hat{x}_{i+1}|$ in $\Big(\Gamma_i/\sqrt{\alpha}, (\sqrt{\alpha}\Gamma_i)^{-1}\Big)$. The histogram of the differences $\hat{x}_i - \hat{x}_{i+1}$ of all MAP solutions in all trials, shown in Fig. 8(b), is very different from the prior. If we define

$$\theta_0 = \frac{\Gamma_{\max}}{\sqrt{\alpha}} \text{ and } \theta_1 = \frac{1}{\sqrt{\alpha}\Gamma_{\max}} \text{ where } \Gamma_{\max} = \max_{1 \le i \le p-1} \Gamma_i < 1,$$

(37) implies that for every $i \in \{1, \ldots, r\}$ we have either $|\hat{x}_i - \hat{x}_{i+1}| \le \theta_0$ or $|\hat{x}_i - \hat{x}_{i+1}| \ge \theta_1$. For our choice of the parameters, $\theta_0 = 0.56$ and $\theta_1 = 1.77$ and the histogram shows that latter inequalities are strongly satisfied. One can also observe that most of the differences satisfy $|\hat{x}_i - \hat{x}_{i+1}| \le \theta_0$ since they belong to homogeneous zones in $\hat{x}$.

5.2. MAP for smooth at zero functions φ. We posit the assumptions given in the introduction of this section 5. We suppose in addition that φ is $\mathcal{C}^2$ and that there are $\tau > 0$ and $\mathcal{T} \in (\tau, \infty)$ such that $\varphi''(t) \ge 0$ if $t \in [0, \tau]$ and $\varphi''(t) \le 0$ if $t \ge \tau$, where φ'' is decreasing on $(\tau, \mathcal{T})$ and increasing on $(\mathcal{T}, \infty)$ (in fact $\mathcal{T}$ is the point where φ'' reaches its minimum on $\mathbb{R}_+$ and φ'' is never positive for $t > \mathcal{T}$.) These assumptions are satisfied by all smooth non-convex functions φ used in practice, such as (f4)-(f7) in Table 1. The MAP energy $\mathcal{E}_y$ is as defined by (34) and (35). In the following, G will denote the $r \times p$ matrix whose rows are g_i^T, $1 \leq i \leq r$. Notice that $\operatorname{rank} G = r \leq m$ means that g_i, $1 \leq i \leq r$ are linearly independent. We will write e_i for the ith vector of the canonical basis of $\mathbb{R}^p$. The result given below exhibits an important feature of the minimizers $\hat{x}$ of $\mathcal{E}_y$ in this context (details can be found in [47]).

Theorem 5.2. *Assume that* $\operatorname{rank} G = r$ *and that* $\beta > \frac{2\mu^2 \|A^T A\|}{|\varphi''(\mathcal{T})|}$ *where* $\mu = \max_{1 \leq i \leq r} \|G^T(GG^T)^{-1}e_i\|$. *Then there are* $\theta_0 \in (\tau, \mathcal{T})$ *and* $\theta_1 \in (\mathcal{T}, \infty)$ *such that for every* $y \in \mathbb{R}^q$, *every minimizer* $\hat{x}$ *of* $\mathcal{E}_y$ *satisfies*

$$\text{either } |g_i^T \hat{x}| \leq \theta_0 \text{ or } |g_i^T \hat{x}| \geq \theta_1, \ \forall i \in \{1, \ldots, r\}. \tag{39}$$

This result is qualitatively similar to (37), but it holds for a wide range of functions φ and concerns any local minimizer $\hat{x}$ of $\mathcal{E}_y$. In particular, for any realization $Y = y$, if $\hat{x}$ is the MAP solution, then $|g_i^T \hat{x}| \notin (\theta_0, \theta_1)$ for every $i \in \{1, \ldots, r\}$. It follows that the distribution of the MAP estimator $\hat{X}$ is such that

$$\Pr\Big(\theta_0 < |g_i^T \hat{X}| < \theta_1\Big) = 0, \ \forall i \in \{1, \ldots, r\}.$$

The prior model *effectively* realized by the MAP estimator corresponds to images and signals whose differences are either smaller than θ_0 or larger than θ_1. Nothing similar holds for the prior model f_X involved in $\mathcal{E}_y$ since for the latter, $\Pr\Big(\theta_0 < |g_i^T X| < \theta_1\Big) > 0$, $\forall i \in \{1, \ldots, r\}$.

5.3. MAP FOR NON-SMOOTH AT ZERO FUNCTIONS φ. Beyond the assumptions made in the introduction of section 5, we assume also that $\varphi'(0^+) > 0$ and that φ'' is increasing on $(0, \infty)$ with $\varphi''(t) \leq 0$, for all $t > 0$. This additional assumption is general enough and is satisfied by all nonsmooth at zero nonconvex functions φ in Table 1. The MAP energy $\mathcal{E}_y$ has the form defined by (34) and (35). Notice that the theorem below does not involve any assumption on $\{g_i : 1 \leq i \leq r\}$.

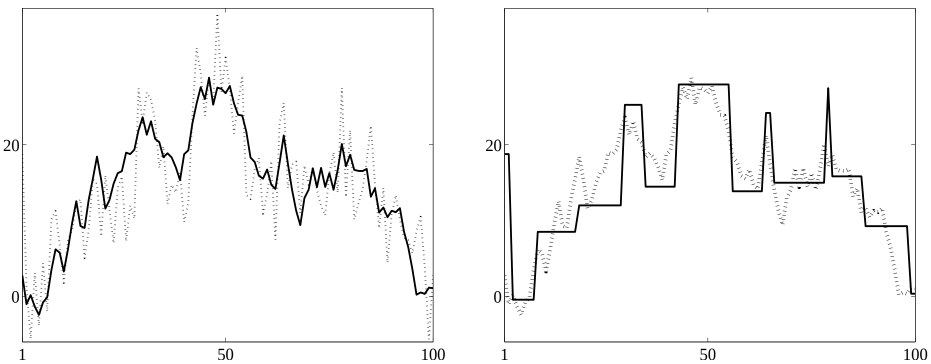
Theorem 5.3. *There is a constant* $\mu > 0$ *such that if* $\beta > \dfrac{2\mu^2 \|A^T A\|}{|\varphi''(0^+)|}$, *then there exists* $\theta_1 > 0$ *such that for every* $y \in \mathbb{R}^q$, *every minimizer* $\hat{x}$ *of* $\mathcal{E}_y$ *satisfies*

$$\text{either } |g_i^T \hat{x}| = 0 \text{ or } |g_i^T \hat{x}| \geq \theta_1, \ \forall i \in \{1, \ldots, r\}. \tag{40}$$

The constant μ is described in [47]. If $|\varphi''(0^+)| = \infty$—as with (f9) in Table 1—the condition on β in the theorem is simplified to $\beta > 0$. Observe that the alternative (40) holds for any realization $Y = y$. It follows that the distribution of the MAP estimator $\hat{X}$ is such that for every $i \in \{1, \ldots, r\}$ we have

$$\begin{aligned} \Pr\Big(|g_i^T \hat{X}| = 0\Big) &> 0, \\ \Pr\Big(0 < |g_i^T \hat{X}| < \theta_1\Big) &= 0. \end{aligned}$$

Hence the sample space of $\hat{X}$ is disconnected and semi-discrete. If $\{g_i, 1 \leq i \leq r\}$ correspond to the first-order differences between neighboring samples, (40) shows that every minimizer $\hat{x}$ of $\mathcal{E}_y$ is composed out of constant patches separated by edges higher than θ_1. This is the *effective* prior model on X realized by the MAP estimator $\hat{X}$. This result is in clear disagreement with the prior model f_X for which $\Pr\Big(|g_i^T X| = 0\Big) = 0$ and $\Pr\Big(0 < |g_i^T X| < \theta_1\Big) > 0$.



(a) Original x (—) by (41) for $\alpha = 10$, $\lambda = 1$ data $y = x + n$ $(\cdots)$, $N \sim \text{Normal}(0, \sigma^2 I)$, $\sigma = 5$.

(b) The true MAP $\hat{x}$ (—) versus the original x $(\cdots)$.

Figure 9. The true MAP restoration of a Markov chain with a nonsmooth nonconvex prior energy from data $y = x + n$ corrupted with white Gaussian noise.

The generalized Gaussian model for $0 < \alpha < 1$, considered in item (b) in § 2.2, provides a first illustration of Theorem 5.3. Other practical consequences are illustrated next. The original x, plotted in Fig. 9(a) with a solid line, is a realization of a 100-length Markov chain whose differences $X_i - X_{i+1}, 1 \le i \le p-1$, are i.i.d. on $[-\gamma, \gamma]$ with density

$$f_{\Delta X}(t) \propto e^{-\lambda \varphi(t)}, \ \varphi(t) = \frac{\alpha |t|}{1 + \alpha |t|}, \tag{41}$$

for $\alpha = 10$, $\lambda = 1$ and $\gamma = 4$. The model for the data is $Y = X + N$ where $N \sim \text{Normal}(0, \sigma^2 I)$. A realization $Y = y$ for $\sigma = 5$ is plotted in the same Fig. 9(a) with a dotted line, and the resultant SNR is 10.65 dB. The MAP solution $\hat{x}$ corresponds to the minimum of $\mathcal{E}_y(x) = \|x - y\|^2 + \beta \sum_{i=1}^{p-1} \varphi(x_i - x_{i+1})$ for $\beta = 2\sigma^2 \lambda$. It is plotted in Fig. 9(b) with a solid line. As predicted by Theorem 5.3, $\hat{x}$ is constant on many pieces which are separated by large edges. Its visual aspect is fundamentally different from the original x since the latter does not involve constant zones and its differences take any value on $[-\gamma, \gamma]$.

6. Conclusion

We have shown both experimentally and theoretically that MAP estimators do not match the underlying models for the production of the data and for the prior. Even though MAP is *optimal* in a precise statistical sense, the distortions of the data-acquisition and the prior models it introduces are embarrassing in many applications where these models are accurate. Instead, based on some analytical properties of the MAP solutions, we partially characterize the models that are *effectively realized by the MAP solutions*. The latter are qualitatively different from the models that underly the posterior energy $\mathcal{E}_y$. Conversely, the obtained results suggest that our analytical approach can be at the basis of a rigorous way to define solutions that realize a priori expected features.

7. APPENDIX

PROOF OF PROPOSITION 2. For any $y \in \mathbb{R}^q$, put $\hat{x}_i = y_i$ for all $i \in \hat{J}$. All equations in (30) thus depend only on $\{y_i, i \in \hat{J}\}$ and are independent of $\{y_i^c, i \in \hat{J}\}$. According to Lemma 4 in [45], there are continuous functions $\mathcal{X}_i : \mathbb{R}^q \to \mathbb{R}$, for $i \in \hat{J}^c$, such that for every $y \in \mathbb{R}^q$, (30) is solved by $\hat{x}_i = \mathcal{X}_i(y)$, $i \in \hat{J}^c$ and $\hat{x}_i = y_i$, $i \in \hat{J}$. Since $\varphi'(0) = 0$ and φ' is continuous, the continuity of $\mathcal{X}_i$, $i \in \hat{J}^c$ shows that there are $y_i : i \in \hat{J}$ such that all the inequalities in (29) are strict. Last, we can easily find y_i, $i \in \hat{J}^c$, such that (31) holds as well. The ultimate results comes from Theorem 3 in [45].

REFERENCES

[1] A. Antoniadis and J. Fan, *Regularization of wavelet approximations*, Journal of Acoustical Society America,**96** (2001), 939–967.
[2] G. Aubert and P. Kornprobst, "Mathematical Problems in Images Processing," Springer-Verlag, Berlin, 2002.
[3] R. G. Aykroyd and P. J. Green, *Global and local priors, and the location of lesions using gamma-camera imagery*, Phil. Trans. R. Soc. Lond. A, **337** (1991).
[4] L. Bar, N. Sochen, and N. Kiryati, *Image deblurring in the presence of salt-and-pepper noise*, in Proceeding of 5th International Conference on Scale Space and PDE methods in Computer Vision, ser. LNCS, **3439** (2005), 107–118.
[5] L. Bar, N. Sochen, and N. Kiryati, *Semi-blind image restoration via mumfordshah regularization*, IEEE Transactions on Image Processing, **15** (2006), 483–493.
[6] M. Belge, M. Kilmer, and E. Miller, *Wavelet domain image restoration with adaptive edge-preserving regularization*, IEEE Transactions on Image Processing, **9** (2000), 597–608.
[7] J. Besag, P. Green, D. Higdon, and K. Mengersen, *Bayesian computation and stochastic systems*, Statistical Science, **10** (1995), 3–66.
[8] J. E. Besag, *Spatial interaction and the statistical analysis of lattice systems (with discussion)*, Journal of the Royal Statistical Society B, **36** (1974), 192–236.
[9] ———, *Digital image processing : Towards Bayesian image analysis*, Journal of Applied Statistics, **16** (1989), 395–407.
[10] J. M. Bioucas-Dias, *Bayesian wavelet-based image deconvolution: A gem algorithm exploiting a class of heavy-tailed priors*, IEEE Transactions on Image Processing, **15** (2006), 937–951.
[11] M. Black and A. Rangarajan, *On the unification of line processes, outlier rejection, and robust statistics with applications to early vision*, International Journal of Computer Vision, **19** (1996), 57–91.
[12] A. Blake and A. Zisserman, *Visual reconstruction*, The MIT Press, Cambridge, 1987.
[13] C. Bouman and K. Sauer, *A generalized Gaussian image model for edge-preserving map estimation*, IEEE Transactions on Image Processing, **2** (1993), 296–310.
[14] P. Charbonnier, L. Blanc-Féraud, G. Aubert, and M. Barlaud, *Deterministic edge-preserving regularization in computed imaging*, IEEE Transactions on Image Processing, **6** (1997), 298–311.
[15] G. Demoment, *Image reconstruction and restoration : Overview of common estimation structure and problems*, IEEE Transactions on Acoustics Speech and Signal Processing, ASSP-**37** (1989), 2024–2036.
[16] D. L. Donoho and I. M. Johnstone, *Ideal spatial adaptation by wavelet shrinkage*, Biometrika, **81** (1994), 425–455.
[17] S. Durand and M. Nikolova, *Stability of minimizers of regularized least squares objective functions I: study of the local behaviour*, Applied Mathematics and Optimization (Springer-Verlag New York), **53** (2006), 185–208.
[18] F. Fessler, *Mean and variance of implicitly defined biased estimators (such as penalized maximum likelihood): Applications to tomography*, IEEE Transactions on Image Processing, **5** (1996), 493–506.
[19] A. Gelman, J. Carlin, H. Stern, and D. Rubin, "Bayesian Data Analysis," CRC Press, 2 ed., 2003.

[20] D. Geman, *Random fields and inverse problems in imaging*, vol. 1427, École d'Été de Probabilités de Saint-Flour XVIII - 1988, Springer-Verlag, lecture notes in mathematics ed., 1990, 117–193.

[21] D. Geman and G. Reynolds, *Constrained restoration and recovery of discontinuities*, IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-**14** (1992), 367–383.

[22] D. Geman and C. Yang, *Nonlinear image recovery with half-quadratic regularization*, IEEE Transactions on Image Processing, IP-**4** (1995), 932–946.

[23] S. Geman and D. Geman, *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images*, IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-**6** (1984), 721–741.

[24] S. Geman and D. E. McClure, *Statistical methods for tomographic image reconstruction*, in Proc. of the 46-th Session of the ISI, Bulletin of the ISI, **52** (1987), 22–26.

[25] J. Gutiérrez, F. Ferri, and J. Malo, *Regularization operators for natural images based on nonlinear perception models*, IEEE Transactions on Image Processing, **15** (2006), 189–200.

[26] R. M. Haralick, *Decision making in context*, IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-**5** (1983), 417–428.

[27] T. Hebert and R. Leahy, *A generalized EM algorithm for 3-D Bayesian reconstruction from Poisson data using Gibbs priors*, IEEE Transactions on Medical Imaging, **8** (1989), 194–202.

[28] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex analysis and Minimization Algorithms, vol. I and II*, Springer-Verlag, Berlin, 1996.

[29] P. J. Huber, "Robust Statistics," Wiley, New York, 1981.

[30] F. Jeng and J. Woods, *Compound Gauss-Markov random fields for image estimation*, IEEE Transactions on Signal Processing, SP-**39** (1991), 683–697.

[31] R. Kass and L. Wasserman, *The selection of prior distributions by formal rules*, Journal of Acoustical Society America, **91** (1996), 1343–1370.

[32] Y. Leclerc, *Constructing simple stable description for image partitioning*, International Journal of Computer Vision, **3** (1989), 73–102.

[33] D. Leporini, J.-C. Pesquet, and H. Krim, "Best Basis Representations with Prior Statistical Models"(Eds: P. Muller and B. Vidakovic), springer verlag: berlin ed., 1999.

[34] S. Li, "Markov Random Field Modeling in Computer Vision," Springer-Verlag, New York, 1 ed., 1995.

[35] J. Marroquin, S. Mitter, and T. Poggio, *Probabilistic solution of ill-posed problems in computational vision*, Journal of Acoustical Society America, **82** (1987), 76–89.

[36] P. Mathieu, M. Antonini, M. Barlaud, and I. Daubechies, *Image coding using wavelet transform*, IEEE Transactions on Image Processing, **1** (1992), p. 205220.

[37] P. Moulin and J. Liu, *Analysis of multiresolution image denoising schemes using generalized gaussian and complexity priors*, IEEE Transactions on Image Processing, 45 (1999), 909–919.

[38] P. Muller, B. Vidakovic, and Eds., "Bayesian Inference in Wavelet-Based Models," Springer-Verlag, New York, 1999.

[39] D. Mumford, "The Dawning of the Age of Stochasticity," AMS, ed. by arnold, atiyah, lax and mazur ed., 2000.

[40] D. Mumford and J. Shah, *Boundary detection by minimizing functionals*, in Proceedings of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 1985, 22–26.

[41] ———, *Optimal approximations by piecewise smooth functions and associated variational problems*, Communications on Pure and Applied Mathematics, (1989), 577–684.

[42] M. Nikolova, *Local strong homogeneity of a regularized estimator*, SIAM Journal on Applied Mathematics, **61** (2000), 633–658.

[43] ———, *Thresholding implied by truncated quadratic regularization*, IEEE Transactions on Signal Processing, **48** (2000), 3437–3450.

[44] ———, *Minimizers of cost-functions involving nonsmooth data-fidelity terms. Application to the processing of outliers*, SIAM Journal on Numerical Analysis, **40** (2002), 965–994.

[45] ———, *A variational approach to remove outliers and impulse noise*, Journal of Mathematical Imaging and Vision, **20** (2004).

[46] ———, *Weakly constrained minimization. Application to the estimation of images and signals involving constant regions*, Journal of Mathematical Imaging and Vision, **21** (2004), 155–175.

[47] ———, *Analysis of the recovery of edges in images and signals by minimizing nonconvex regularized least-squares*, SIAM Journal on Multiscale Modeling and Simulation, **4** (2005), 960–991.

[48] P. Perona and J. Malik, *Scale-space and edge detection using anisotropic diffusion*, IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-**12** (1990), 629–639.
[49] B. D. Ripley and F. P. Kelly, *Markov point processes.*, Journal of the London Mathematical Society, **15** (1977), 188192.
[50] R. T. Rockafellar and J. B. Wets, "Variational Analysis," Springer-Verlag, New York, 1997.
[51] L. Rudin, S. Osher, and C. Fatemi, *Nonlinear total variation based noise removal algorithm*, Physica, **60** D (1992), 259–268.
[52] S. Saquib, C. Bouman, and K. Sauer, *ML parameter estimation for Markov random fields, with appications to Bayesian tomography*, IEEE Transactions on Image Processing, **7** (1998), 1029–1044.
[53] R. R. Schultz and R. L. Stevenson, *A bayesian approach to image expansion for improved definition*, IEEE Transactions on Image Processing, **3** (1994), 233–242.
[54] H. Sidenbladh and M. Black, *Learning the statistics of people in images and video*, Journal of Mathematical Imaging and Vision, **54** (2003), 183–209.
[55] E. P. Simoncelli, *Bayesian denoising of visual images in the wavelet domain*, Lecture Notes in Statistics, Vol. **41**, Springer Verlag: Berlin, 1999.
[56] E. P. Simoncelli and E. H. Adelson, *Noise removal via Bayesian wavelet coding*, in Proceedings of the IEEE International Conference on Image Processing, Lausanne, Switzerland, Sep. 1996, 379–382.
[57] A. Srivastava, A. B. Lee, E. P. Simoncelli, and S.-C. Zhu, *On advances in statistical modeling of natural images*, Journal of Mathematical Imaging and Vision, **18** (2003).
[58] R. Stevenson and E. Delp, *Fitting curves with discontinuities*, in "Proc. of the 1st Int. Workshop on Robust Comput. Vision," Seattle, WA, 1990, 127–136.
[59] A. Tarantola, "Inverse Problem Theory : Methods for Data Fitting and Model Parameter Estimation," Elsevier Science Publishers, Amsterdam, 1987.
[60] L. Tenorio, *Statistical regularization of inverse problems*, SIAM Review, **43** (2001), 347–366.
[61] A. Tikhonov and V. Arsenin, "Solutions of Ill-Posed Problems," Winston, Washington DC, 1977.
[62] C. Vogel, "Computational Methods for Inverse Problems," Frontiers in Applied Mathematics Series, Number 23, SIAM, 2002.
[63] C. R. Vogel and M. E. Oman, *Iterative method for total variation denoising*, SIAM Journal on Scientific Computing, **17** (1996), 227–238.
[64] G. Wang, J. Zhang, and G.-W. Pan, *Solution of inverse problems in image processing by wavelet expansion*, IEEE Transactions on Image Processing, **4** (1995), 579–593.

Received for publication December 2006.

E-mail address: nikolova@cmla.ens-cachan.fr