Contre-examples for Bayesian MAP restoration

Mila Nikolova

CMLA—ENS de Cachan, 61 av. du Président Wilson, 94235 Cachan cedex
(nikolova@cmla.ens-cachan.fr)

Obergurgl, September 2006
1. MAP estimators to combine noisy data and priors

   Combining observed data $y$ for the unknown $x$ with priors on $x$

   $$\hat{x} = \arg\min \left\{ \Psi(x, y) + \beta \Phi(x) \right\}$$

2. Examples of gaps between models and estimate

   *MAP solutions (substantially) deviate from the data model and from the prior*

   *Instead — effective prior (based on properties of minimizers)*

3. Non-smooth at zero priors

4. Non-smooth at zero noise models

5. Priors with non-convex energies

6. Concluding remarks
1. MAP estimators to combine noisy data and priors

- Forward model $= f_{Y|X}(y|x)$ likelihood - physical considerations on data-acquisition

  E.g. $Y = AX + N$

  $A$ — blur, Fourier, Radon, subsampling... and $N$ — noise

  $\{N_i\}$ i.i.d. $\sim f_N \Rightarrow f_{Y|X}(y|x) = \prod_i f_N(a_i^T x - y_i)$

  If $f_N =$Normal$(0, \sigma^2) \Rightarrow f_{Y|X} = \frac{1}{Z} e^{-\frac{\|Ax-y\|^2}{2\sigma^2}}$

- Prior $= f_X(x)$

  - Markov models —local characteristics— $f_X(x_i|x_j, j \neq i) = f_X(x_i|x_j, j \in N_i)$

    Gibbsian form $f_X(x) \propto \exp\{ -\lambda \Phi(x) \}$

    The Hammersley-Clifford theorem $\Rightarrow \Phi(x) = \frac{1}{2} \sum_i \sum_{j \in N_i} \varphi(x_i - x_j)$

  - Wavelet expansions — coefficients $u_i = \langle w_i, x \rangle$ are i.i.d. $\sim f_{U_i}(t) = e^{(-\lambda_i \varphi(t))} \frac{1}{Z}$
Customary functions $\varphi$

$\varphi(t) = t^\alpha$, $0 < \alpha \leq 2$
$\varphi(t) = \sqrt{\alpha + t^2}$
$\varphi(t) = \log(\cosh(t/\alpha))$
$\varphi(t) = 1 - \exp(-\alpha t^2)$
$\varphi(t) = \alpha t^2/(1 + \alpha t^2)$
$\varphi(t) = \alpha |t|/(1 + \alpha |t|)$
$\varphi(t) = \min\{\alpha t^2, 1\}$
$\varphi(t) = \log(\alpha |t| + 1)$
and many others...

- The posterior (Bayesian rule) $f_{X|Y}(x|y) = f_{Y|X}(y|x)f_X(x)\frac{1}{Z}$ $Z = f_Y(y)$

MAP $\hat{x} = \text{the most likely solution given the recorded data } Y = y$:

$\hat{x} = \arg \max_x f_{X|Y}(x|y) = \arg \min_x ( - \ln f_{Y|X}(y|x) - \ln f_X(x) )$

$= \arg \min_x ( \Psi(x, y) + \beta \Phi(x) )$

Examples:

$E_y(x) = \|Ax - y\|^2 + \beta \Phi(x), \quad \beta = 2\sigma^2 \lambda$
$E_y(u) = \sum_i \left( (u_i - \langle w_i, y \rangle)^2 + \lambda_i \varphi(|u_i|) \right), \quad \hat{x} = W^\dagger \hat{u}$

More and more realist models for data-acquisition $f_{Y|X}$ and prior $f_X$

... natural expectation that $\hat{x}$ is coherent with $f_{Y|X}$ and $f_X$

(If $X \sim f_X$ and $AX - Y \sim f_N$ then $\hat{X} \sim f_X$ and $A\hat{X} - Y \sim f_N$)

Contradiction: the MAP solution substantially deviates from the models!
2. Gap between models and estimate

Analytical example on \( \mathbb{R} \)

\[
Y = X + N \quad f_X(x) = \begin{cases} 
\lambda e^{-\lambda x} & \text{if } x \geq 0 \\
0 & \text{else}
\end{cases}
\]

\( N \sim \text{Normal}(0, \sigma^2) \)

The MAP \( \hat{x} \) is the minimizer on \([0, +\infty)\) of \( E_y(x) = (x - y)^2 + \beta x \) for \( \beta = 2\sigma^2\lambda \)

\[
\hat{x} = \begin{cases} 
0 & \text{if } y < \frac{\beta}{2} \\
\frac{y}{2} > 0 & \text{if } y \geq \frac{\beta}{2}
\end{cases}
\]

\[ f_{\hat{X}}(\hat{x}) = f_X(\hat{x}) \xi(\hat{x}) + c \text{ Dirac}(\hat{x}) \quad \text{where} \quad \begin{cases} 
\xi(\hat{x}) = \frac{1}{2}(\lambda\sigma^2 - \beta) \int_{0}^{\infty} f_N(x - \hat{x} - \frac{\beta}{2} + \lambda\sigma^2)dx \\
c = \int_{0}^{\infty} f_X(x) \int_{-\infty}^{\frac{\beta}{2} - x} f_N(n)dndx \in (0, 1).
\end{cases}\]

\[ \Rightarrow f_{\hat{X}} \text{ is fundamentally dissimilar to } f_X \]

The noise estimate \( \hat{n} = y - \hat{x} = \begin{cases} 
y & \text{if } y < \frac{\beta}{2} \\
\frac{\beta}{2} & \text{if } y \geq \frac{\beta}{2}
\end{cases} \]

\[ f_{\hat{N}}(\hat{n}) = f_N(\hat{n}) \mathbb{1}(\hat{n} < \frac{\beta}{2}) \zeta(\hat{n}) + (1 - c) \text{ Dirac}(\hat{n} - \frac{\beta}{2}) \quad \text{for} \quad \zeta(\hat{n}) = \int_{0}^{\infty} f_X(x)e^{-\frac{x^2 - 2\hat{n}x}{2\sigma^2}}dx \]

\[ \Rightarrow f_{\hat{N}} \text{ is upper bounded by } \frac{\beta}{2}, \text{ dissimilar to } f_N \]
In general $f_X$ and $f_N$ cannot be calculated

**Distribution of the MAP for generalized Gaussian priors**

MAP restoration of noisy wavelet coefficients with Gaussian noise

Noise-free wavelet coefficients are i.i.d. and follow GG

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

MAP $\hat{u}_i$ of each noisy coefficient $\langle w_i, y \rangle$ minimizes

$$E_y(x) = (x - y)^2 + \beta |x|^\alpha \quad \text{for} \quad \beta = 2\sigma^2 \lambda$$

For $(\alpha, \lambda)$ and $\sigma$ fixed, we realize 10 000 independent trials:

- sample $x \in \mathbb{R}$ from $f_X$
- $y = x + n$ for $n \sim \text{Normal}(0, \sigma^2)$
- compute the true MAP solution $\hat{x}$
$f_{X|Y}(., y)$ has one mode if $\alpha \geq 1$

GG prior for $\alpha = 1.2$, $\lambda = 0.5$

Noise Normal$(0, \sigma^2)$ for $\sigma = 0.6$

The true MAP $\hat{x}$

The noise estimate $\hat{n} = y - \hat{x}$
If $0 < \alpha < 1$, $f_{X|Y}(., y)$ has two modes, $\hat{x}_1 = 0$ and $\hat{x}_2$ with $|\hat{x}_2| > \theta$ for

$$\theta = \left( \frac{2}{\alpha (1-\alpha) \beta} \right)^{\frac{1}{\alpha-2}} \approx 0.47$$

$\Rightarrow f_\hat{X}$ has a Dirac at zero and is null on $(-\theta, 0) \cup (0, \theta)$

Prior $f_X$ for $\alpha = 0.5$, $\lambda = 2$

True MAP $\hat{x}$

Zoom of the histogram of $\hat{x}$

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

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

$\hat{x} = 0$ in 77% of the trials and $\min\{|\hat{x}_i| : x_i \neq 0\} = 0.77 > \theta$
3. Non-smooth at zero priors

A Laplacian Markov chain corrupted with Gaussian noise

Markov chain with a Gibbsian distribution \( f_X \propto e^{-\lambda \Phi(x)} \)

\[
\Phi(x) = \lambda \sum_{i=1}^{p-1} |x_i - x_{i+1}|, \quad \lambda > 0
\]

\( X_i - X_{i+1}, 1 \leq i \leq p - 1 \) are Laplacian and i.i.d.

\[
f_{\Delta X}(t) = \frac{\lambda}{2} e^{-\lambda |t|}
\]

\( Y = X + N, \quad N \sim \text{Normal}(0, \sigma^2 I) \)

\[
f_{X|Y}(x|y) = \exp \left( -\frac{1}{2\sigma^2} E_y(x) \right) \frac{1}{Z}
\]

\[
E_y(x) = \|x - y\|^2 + \beta \sum_{i=1}^{p-1} |x_i - x_{i+1}|, \quad \beta = 2\sigma^2 \lambda
\]
Coherence with the models: for $p \to \infty$ \begin{align*}
\text{Hist}(\hat{x}_i - \hat{x}_{i+1}) & \approx f_{\Delta X} \\
\text{Hist}(y_i - \hat{x}_i) & \approx f_N
\end{align*}
The same experiment (500-length signals) 40 times:

\[ 40 \times 499 \text{ differences } x_i - x_{i+1} \]

sampled from \( f_{\Delta X} \) for \( \lambda = 8 \).

87\% of all restored differences are null

The MAP solution is far from representing the prior

The observed incoherence is inherent — it originates from the analytical properties of the MAP solution.
Analytical results on the MAP and their statistical meaning

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

$G_i, 1 \leq i \leq r$ linear operators (e.g. finite differences or discrete derivatives)

$\varphi : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ is increasing, $C^m$ and

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

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

$$f_{Y|X}(y|x) \propto e^{-\Psi(x,y)} \text{ where } \Psi \sim C^m, m \geq 2$$

The MAP estimator $\hat{X}$ minimizes

$$E_y(x) = \Psi(x, y) + \lambda \Phi(x)$$
Theorem [Nikolova 2000, 2004] Given $y \in \mathbb{R}^q$, let $\hat{x} \in \mathbb{R}^p$ be such that for $J = \left\{ i \in \{1, \ldots, r\} : G_i \hat{x} = 0 \right\}$ and $K_J = \left\{ u \in \mathbb{R}^p : G_i u = 0, \forall i \in J \right\}$, we have

(a) $\delta E_y(\hat{x})(u) > 0$ for every $u \in K_J^\perp \setminus \{0\}$;

(b) $D E_y|_{K_J}(\hat{x})u = 0$ and $D^2 E_y|_{K_J}(\hat{x})(u, u) > 0$, for every $u \in K_J \setminus \{0\}$.

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

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

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

(a) and (b) ensure that $E_y$ has a strict local minimum at $\hat{x}$ they are quite general:

Proposition [Durand & Nikolova 2006] 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 $E_y$ is strict, and that (a) and (b) hold. Then

(i) $\Omega^c$ (the complement of $\Omega$ 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 $\Omega^c$ is of Lebesgue measure zero as well.
\( O_J \) contains an open subset of \( \mathbb{R}^q \)

\[
y \in O_J \quad \text{and} \quad \hat{x} = \arg\max_{x \in \mathbb{R}^p} f_{X|Y}(x|y) \quad \Rightarrow \quad G_i\hat{x} = 0 \quad \forall i \in J
\]

or equivalently \( \hat{x} \in K_J \)

\[
\Rightarrow \quad \Pr(\hat{X} \in K_J) \geq \Pr(Y \in O_J) = \int_{O_J} f_Y(y)dy > 0
\]

since \( f_Y(y) = \int f_{Y|X}(y|x)f_X(x)dx = \frac{1}{Z} \int e^{-E_y(x)}dx > 0, \quad \forall y \)

The “prior” model on the unknown \( X \) which is effectively realized by the MAP estimator \( \hat{X} \)
corresponds to images and signals such that \( G_i\hat{X} = 0 \) for a certain number of indexes \( i \).

If \( \{G_i\} = \text{first-order} \), then effective prior model for locally constant images and signals.

According to the prior, for any nonempty \( J \subset \{1, \ldots, r\} \)

\[
\Pr(X \in K_J) = \int_{K_J} f_X(x)dx = 0
\]

since \( \dim K_J \subset \mathbb{R}^p < p \) and \( x \in \mathbb{R}^p \)
Linear Gaussian data model with $A$ invertible and a Laplacian Markov chain prior

$$f_{X|Y}(x|y) \propto \exp(-E_y(x)) + \text{const}$$

$$E_y(x) = \|Ax - y\|^2 + \beta \sum_{i=1}^{p-1} |x_i - x_{i+1}|, \quad \beta = 2\sigma^2\lambda$$

Striking phenomena:

(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 $E(. , y)$;

(b) for every $J \subset \{1, \ldots, p-1\}$, 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 $E_y$ 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 $\mathbb{R}^q$.

$$\Rightarrow \forall J \subset \{1, \ldots, p-1\} \quad \Pr(\hat{X}_i = \hat{X}_{i+1}, \forall i \in J) \geq \Pr(Y \in \tilde{O}_J) > 0.$$ 

$$\Rightarrow \quad \hat{x} \text{ are composed of constant pieces.}$$

However, the prior model yields $\Pr(X_i = X_{i+1}) = 0$ for every $i \in \{1, \ldots, p-1\}$. 

15
4. Non-smooth at zero noise models

\[ Y = AX + N \] where \( N_i \sim f_N \) are i.i.d.

\[ f_N(t) = \frac{1}{Z} e^{-\sigma \psi(t)} \]

\( \psi : \mathbb{R} \to \mathbb{R} \) is \( C^m, m \geq 2 \), on \( \mathbb{R} \setminus \{0\} \) and

\[ 0 < \psi'(0^+) = -\psi'(0^-) < \infty \]

\[ f_{Y \mid X}(y \mid x) \propto \exp(-\sigma \Psi(x, y)) \]

\[ \Psi(x, y) = \sum_{i=1}^{q} \psi(a_i^T x - y_i) \]

If \( N \sim \text{Laplacian i.i.d. noise} \Rightarrow \Psi(x, y) = \|Ax - y\|_1 \]

Notice \( \Pr(N_i = 0) = 0 \) for every \( i \in \{1, \ldots, q\} \)

Let \( X \sim \text{Gibbsian} \) where \( \Phi : \mathbb{R}^p \to \mathbb{R} \) is \( C^m \)

The MAP \( \hat{x} \) minimizes

\[ E_y(x) = \Psi(x, y) + \beta \Phi(x), \quad \beta = \frac{\lambda}{\sigma} \]
Generalized Gaussian Markov chain under Laplace noise

\( X \) — Markov chain, \( X_i - X_{i+1} \sim f_{\Delta X} \) are i.i.d.

\[
f_{\Delta X}(t) = \frac{1}{Z} e^{-\lambda |t|^\alpha}
\]

\( Y = X + N \) where \( N_i, 1 \leq i \leq p \) are i.i.d. with \( f_N(t) = \frac{\sigma}{2} e^{-\sigma |t|} \)

\[
f_{X|Y}(x|y) = \exp \left( -\sigma E_y(x) \right) \frac{1}{Z}
\]

\[
E_y(x) = \sum_{i=1}^{p} |x_i - y_i| + \beta \sum_{i=1}^{p-1} |x_i - x_{i+1}|^{\alpha} \quad \text{where} \quad \beta = \frac{\lambda}{\sigma}.
\]
GG Markov chain $x$ (—) for $\alpha = 1.2$, $\lambda = 1$

data $y = x + n$ (⋯)

Laplacian i.i.d. noise $n$ for $\sigma = 2.5$

The true MAP $\hat{x}$ (—)

versus the original $x$ (⋯)

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

Notice $x_i \neq y_i$ for all $i$

The MAP $\hat{x}$ contains 93% samples satisfying $\hat{x}_i = y_i$. 
The same experiment 1000 times

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

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

Laplacian i.i.d. noise by for $\sigma = 2.5$.

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

$\hat{x}_i = y_i$ for 87% of the samples in all trials $\Rightarrow$ most of the samples $\hat{x}_i$ keep the noise intact
Main analytical result and statistical interpretation

**Theorem [Nikolova2001]** Given $y \in \mathbb{R}^q$, suppose that $\hat{x} \in \mathbb{R}^p$ is such that for $J = \{i \in \{1, \ldots, q\} : a_i^T \hat{x} = y_i\}$ and $K_J = \{u \in \mathbb{R}^p : a_i^T u = 0 \ \forall i \in J\}$ we have:

(a) the set $\{a_i : i \in J\}$ is linearly independent;

(b) $D E_y|_{\hat{x} + K_J}(\hat{x})u = 0$ and $D^2 E_y|_{\hat{x} + K_J}(\hat{x})(u, u) > 0$, for every $u \in K_J \setminus \{0\}$;

(c) $\delta E_y(\hat{x})(u) > 0$, for every $u \in K_{\perp J} \setminus \{0\}$.

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

$$a_i^T \hat{x}' = y_i' \quad \text{if} \quad i \in J,$$

$$a_i^T \hat{x}' \neq y_i' \quad \text{if} \quad i \in J^c.$$

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

Weak assumptions: Pr that (a) fails =0, (b)-(c) sufficient conditions for a strict local minimum.
Crucial: $O_J$ contains an open subset of $\mathbb{R}^q$

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

*For all $i \in J$, the prior has no influence on the solution and the noise remains intact*

This contradicts the noise model since

$$\Pr \left( a_i^T X - Y_i = 0 \right) = \Pr \left( N_i = 0 \right) = 0, \quad \forall i$$

Let $A$ invertible and $\Phi$ Gibbsian

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

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

*Amazing: on $O_{\infty}$ the prior has no influence on the solution*

$$y \in O_{\infty} \quad \Rightarrow \quad a_i^T \hat{x} = y_i, \quad \forall i$$
A Laplace noise model to remove impulse noise

\[ E_y(x) = \sum_{i=1}^{p} |x_i - y_i| + \frac{\beta}{2} \sum_{i} \sum_{j \in \mathcal{N}_i} \varphi(x_i - x_j) \]

\( \varphi \) symmetric \( C^1 \) strictly convex edge-preserving

Bayesian standpoint: \( Y = X + N \) with \( N \) Laplacian white noise

Previous results: the MAP cannot efficiently clean Laplacian noise (all \( \hat{x}_i \) such that \( \hat{x}_i = y_i = x_i + n_i \) keep the noise intact while \( n_i \neq 0 \) almost surely)

What is the noise model which is effectively realized by the MAP?

\( E_y \) reaches its minimum at a point \( \hat{x} \in \mathbb{R}^p \), for which we define
\( J = \{ i \in \{1, \ldots, p\} : \hat{x}_i = y_i \} \), if, and only if,

\[
  i \in J \quad \Rightarrow \quad \left| \sum_{j \in \mathcal{N}_i} \varphi'(y_i - \hat{x}_j) \right| \leq \frac{1}{\beta},
\]

\[
  i \in J^c \quad \Rightarrow \quad \sum_{j \in \mathcal{N}_i} \varphi'(\hat{x}_i - \hat{x}_j) = \frac{\sigma_i}{\beta}, \quad \sigma_i = \text{sign} \left( \sum_{j \in \mathcal{N}_i} \varphi'(y_i - \hat{x}_j) \right) \in \{-1, 1\}.
\]
Proposition  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

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

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

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

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

where $X_i$, $i \in J^c$ are continuous functions that depend only on $y'_i$ for $i \in J$.

- $\Pr(Y \in O_J) > 0$ since $O_J$ contains an open of $\mathbb{R}^p$
- $O_J$ are disjoint, hence
  $$\Pr(\hat{X}_i - Y_i = 0) \geq \sum_{J : i \in J} \Pr(Y \in O_J) > 0, \quad \forall i$$
- Contradicts the Laplacian noise model involved in $E_y$: $\Pr(X_i - Y_i = 0) = 0, \quad \forall i \in \{1, \ldots, p\}$
- The data samples $y'_i$, $i \in J$ are fitted exactly, hence they must be free of noise. Otherwise $i \in J^c$ and $y'_i$ is replaced by the estimate $\hat{x}'_i = X_i(\{y'_i : i \in J\})$
  Hence $y'_i$, $i \in J^c$ is outlier and can take any value on the half-line contained in $O_J$.
- The MAP estimator defined by $E_y$ corresponds to an impulse noise model on the data
Original $x$ (---), data $y$ (---)
with 10% random valued impulse noise.

The minimizer $\hat{x}$ of $E_y$ for $\beta = 0.4$ (---),
the original $x$ (---), and $y_i \neq x_i$ (◇)
$\hat{x}_i = y_i$ for 89/90 of the noise-free samples.
5. Priors with non-convex energies

\[ Y = AX + N \] with \( N \sim \text{Normal}(0, \sigma^2 I) \) and a Gibbsian prior with a nonconvex \( \Phi \)

\[
\Phi(x) = \sum_{i=1}^{r} \varphi(g_i^T x)
\]

\( g_i \) difference operators

\( \varphi \) \{ symmetric, \( C^2 \) and increasing on \((0, +\infty)\) with a strict minimum at zero

and \( \exists \theta > 0 \) such that \( \varphi''(\theta) < 0 \) and \( \lim_{t \to \infty} \varphi''(t) = 0 \) (nonconvex)

The MAP \( \hat{x} \) yields the (global) minimum of

\[
E_y(x) = \|Ax - y\|^2 + \beta \Phi(x), \quad \beta = 2\sigma^2 \lambda
\]

Since [Geman 1984] various nonconvex \( \varphi \) to produce \( \hat{x} \) with smooth regions and sharp edges.
Piecewise Gaussian Markov chain in Gaussian noise

The piecewise GM chain = discrete 1D Mumford-Shah model = the weak-string model

\[ X \text{ such that } X_{i+1} - X_i \text{ are i.i.d. } \sim f_{\Delta X}(t) \propto e^{-\lambda \varphi(t)} \]

\[ \varphi(t) = \begin{cases} \alpha t^2 & \text{if } |t| < \sqrt{\frac{1}{\alpha}} \\ 1 & \text{else} \end{cases} = \min\{\alpha t^2, 1\} \]

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

**Theorem [Nikolova 2000]** Define \( u_i \in \mathbb{R}^p \) by \( u_i[j] = 0 \) if \( 1 \leq j \leq i \) and \( u_i[j] = 1 \) if \( j \geq i + 1 \) (step), and \( P = I - \frac{A_1 A_1^T A_1^T}{\|A_1\|^2} \) (projection). If \( E_y \) has a global minimum at \( \hat{x} \), then \( \forall i \in \{1, \ldots, p - 1\} \)

\[ \text{either } |\hat{x}_i - \hat{x}_{i+1}| \leq \frac{1}{\sqrt{\alpha}} \Gamma_i \text{ or } |\hat{x}_i - \hat{x}_{i+1}| \geq \frac{1}{\sqrt{\alpha}} \Gamma_i \]

\[ \Gamma_i = \sqrt{\frac{\|P A u_i\|^2}{\|P A u_i\|^2 + \alpha \beta}} < 1. \text{ In particular, } \hat{x}_i - \hat{x}_{i+1} = 0 \text{ if } P A u_i = 0. \]

\[ \implies \Pr\left( \frac{\Gamma_i}{\sqrt{\alpha}} < |\hat{X}_i - \hat{X}_{i+1}| < \frac{1}{\sqrt{\alpha} \Gamma_i} \right) = 0 \]

whereas a priori \( \Pr\left( \frac{\Gamma_i}{\sqrt{\alpha}} < |X_i - X_{i+1}| < \frac{1}{\sqrt{\alpha} \Gamma_i} \right) > 0 \)
We repeat 200 times the following experiment:

- generate $X = x$ of length $p = 300$ where $x_i - x_{i+1}$ are sampled from $f_{\Delta X}$ for $\alpha = 1$, $\lambda = 5$ and $\gamma = 15$

- $y = x + n$ where $n \sim \text{Normal}(0, \sigma^2 I)$, $\sigma = 4$

- compute $\hat{x} = \arg \min E_y$ for the true parameter $\beta = 2\sigma^2 \lambda = 160$.

![Histogram of all original differences $x_i - x_{i+1}$ (up) and zoom (bottom).](image1)

![Histogram of the differences for all the true MAP solutions $\hat{x}$ (up) and zoom (bottom).](image2)
Additional assumption: $\varphi$ is $C^2$ and $\exists \tau > 0$, $T \in (\tau, \infty)$ such that $\varphi''(t) \geq 0$ if $t \in [0, \tau]$ and $\varphi''(t) \leq 0$ if $t \geq \tau$, where $\varphi''$ is decreasing on $(\tau, T)$ and increasing on $(T, \infty)$

$G \in \mathbb{R}^{r \times p}$, row $i = g_i^T$

$e_i$ — the $i$th vector of the canonical basis of $\mathbb{R}^p$

**Theorem [Nikolova05]** Let $\text{rank } G = r$ and $\beta > \frac{2\|A^T A\|}{|\varphi''(T)|} \max_i \|G^T (GG^T)^{-1} e_i\|^2$. Then $\exists \theta_0 \in (\tau, T)$ and $\exists \theta_1 \in (T, \infty)$ such that $\forall y$, every minimizer $\hat{x}$ of $E_y$ satisfies

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

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

The prior model effectively realized by the MAP estimator corresponds to images and signals whose differences are either smaller than $\theta_0$ or larger than $\theta_1$.

Different from the prior since $\text{Pr} \left( \theta_0 < |g_i^T X| < \theta_1 \right) > 0, \forall i \in \{1, \ldots, r\}$. 

28
Additional assumption: $\varphi'(0^+) > 0$ and that $\varphi''$ is increasing on $(0, \infty)$ with $\varphi''(t) \leq 0$, $\forall t > 0$

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

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

If $|\varphi''(0^+)| = \infty$ the condition is $\beta > 0$.

The alternative holds for any realization $Y = y$. Hence

$$\Pr(|g_i^T \hat{X}| = 0) > 0,$$

$$\Pr(0 < |g_i^T \hat{X}| < \theta_1) = 0.$$

(The sample space of $\hat{X}$ is disconnected and semi-discrete)

If $\{g_i, 1 \leq i \leq r\}$ — first-order differences between neighbors, every minimizer $\hat{x}$ of $E_y$ is composed out of constant patches separated by edges higher than $\theta_1 \equiv$ the effective prior model realized by the MAP

Disagreement with the prior $f_X$ for which $\Pr(|g_i^T X| = 0) = 0$ and $\Pr(0 < |g_i^T X| < \theta_1) > 0$
Original $x$ with differences $X_i - X_{i+1}$ i.i.d. on $[-\gamma, \gamma]$ with density

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

Original $x$ (—) by $f_{\Delta X}$ for $\alpha = 10$, $\lambda = 1$, $\gamma = 4$

data $y = x + n$ ($\cdots$), $N \sim \text{Normal}(0, \sigma^2 I)$, $\sigma = 5$.

- $\hat{x}$ is constant on many pieces which are separated by large edges.
  Its visual aspect is fundamentally different from the original $x$

- $x$ does not involve constant zones and its differences take any value on $[-\gamma, \gamma]$.  

The true MAP $\hat{x}$ (—), $\beta = 2\sigma^2 \lambda$ versus the original $x$ ($\cdots$).
6. Conclusion

- MAP estimators do not match the underlying models for the production of the data and for the prior
  
  Experimental demonstration and theoretical explanation
  
  Embarrassing... the problem of $\beta$ never solved

- Based on some analytical properties of the MAP solutions, we partially characterize the models that are effectively realized by the MAP solutions.

- Conjecture: similar problems generally arise with other Bayesian estimators too.

- Combining models is an open problem

- Papers available at http://www.cmla.ens-cachan.fr/ nikolova/