# AXDA : efficient sampling through variable splitting inspired bayesian hierarchical models

**P. Chainais**

with Maxime Vono & Nicolas Dobigeon

March 12th 2019

# Flight schedule

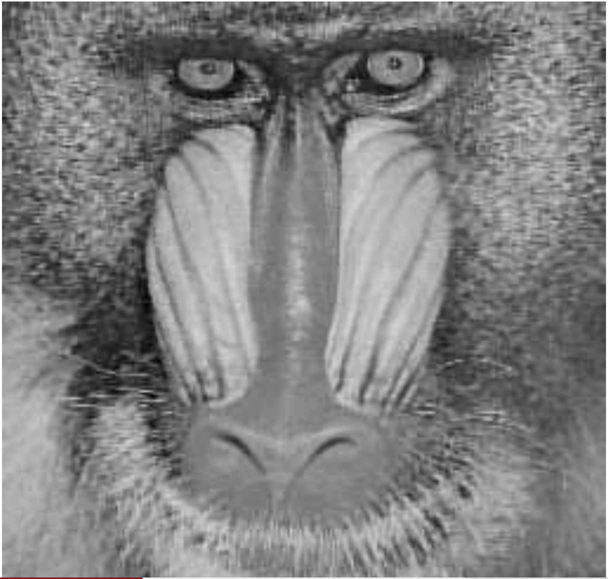
# Outline

# Motivations: applications in image processing

Image deblurring

# Motivations: applications in image processing

Image deblurring



# Motivations: applications in image processing

Image inpainting

# Motivations: applications in image processing

Image inpainting

# Motivations: applications in image processing

## Confidence intervals

# Motivations

I have a dream...

- **solve** complex ill-posed inverse problems
- **big** data in **large** dimensions
- **excellent** performances
- **fast** inference algorithms
- **credibility intervals**

with maybe some additional options such as:

- parallel **distributed** computing
- **privacy** preserving

# Motivations

I have a dream...

- **solve** complex ill-posed inverse problems
- **big** data in **large** dimensions
- **excellent** performances
- **fast** inference algorithms
- **credibility intervals**

with maybe some additional options such as:

- parallel **distributed** computing
- **privacy** preserving

$\Longrightarrow$ **Bayesian approach + MCMC method!**

Inverse problems & **optimization**

= define a **cost function** : $f(\mathbf{x}) = f_1(\mathbf{x}|\mathbf{y}) + f_2(\mathbf{x})$

where $f_2$ is typically

- convex (or not)
- not differentiable $\Rightarrow$ proximal operators
- a sum of various penalties

**Solution: proximal operators**

# Motivations

## The optimization-based approach

Inverse problems & **optimization**

= define a **cost function** : $f(\mathbf{x}) = f_1(\mathbf{x}|\mathbf{y}) + f_2(\mathbf{x})$

where $f_2$ is typically

- convex (or not)
- not differentiable $\Rightarrow$ proximal operators
- a sum of various penalties

**Solution:** **proximal operators** and **splitting** techniques

$$\arg\min_{\mathbf{x}} f_1(\mathbf{x}) + f_2(\mathbf{z}) \text{ such that } \mathbf{x} = \mathbf{z}$$

maybe relaxed to (simplified version of **ADMM**)

$$\arg\min_{\mathbf{x}} f_1(\mathbf{x}) + f_2(\mathbf{z}) + \frac{\alpha}{2}\|\mathbf{x} - \mathbf{z}\|_2^2 + \mathbf{u}^T(\mathbf{x} - \mathbf{z})$$

# Motivations

## The Bayesian approach

Inverse problems & **Bayes** posterior $\propto$ likelihood$(f1)$ $\times$ prior$(f2)$

$$= \text{define a } \textbf{posterior distribution } p(\mathbf{x}|\mathbf{y}) = p_1(\mathbf{x}|\mathbf{y}) \cdot p_2(\mathbf{x})$$

where $p_2$ is typically

- log-concave (or not) $\leftrightarrow$ $f_2$ convex
- conjugate $\Rightarrow$ easy sampling/inference
- a combination of various prior

**Solution: MCMC methods** and **Gibbs** sampling

$$\boxed{x_i \sim p(x_i|x_{\setminus i}) \quad \forall 1 \leq i \leq d}$$

# Motivations

## The Bayesian approach

Inverse problems & **Bayes** posterior $\propto$ likelihood$(f1)$ $\times$ prior$(f2)$

= define a **posterior distribution** $p(\mathbf{x}|\mathbf{y}) = p_1(\mathbf{x}|\mathbf{y}) \cdot p_2(\mathbf{x})$

where $p_2$ is typically

- log-concave (or not) $\leftrightarrow$ $f_2$ convex
- conjugate $\Rightarrow$ easy sampling/inference
- a combination of various prior

**Solution:** **MCMC methods** and **Gibbs** sampling

$$\boxed{x_i \sim p(x_i|x_{\setminus i}) \quad \forall 1 \leq i \leq d}$$

**Can we adapt *splitting* and *augmentation* from optimization ?**

$$\pi_\rho(\mathbf{x}, \mathbf{z}, \mathbf{u}) \propto \exp\left[-f_1(\mathbf{x}) - f_2(\mathbf{z}) - \frac{1}{2\rho^2}\|\mathbf{u} - \mathbf{x} + \mathbf{z}\|_2^2 - \frac{1}{2\alpha^2}\|\mathbf{u}\|^2\right]$$

Inverse problems & **Bayes** posterior $\propto$ likelihood$(f1)$ $\times$ prior$(f2)$

= define a **posterior distribution** $p(\mathbf{x}|\mathbf{y}) = p_1(\mathbf{x}|\mathbf{y}) \cdot p_2(\mathbf{x})$

**Computational motivations: difficult sampling**

- non-conjugate priors [conj. priors $\Rightarrow$ easy inference]
- rich models: complicated prior distributions
- big datasets: expensive likelihood computation

**Strategy:** **DIVIDE-TO-CONQUER**

$\Longrightarrow$ ***splitting (SP)*** and ***augmentation (SPA)***

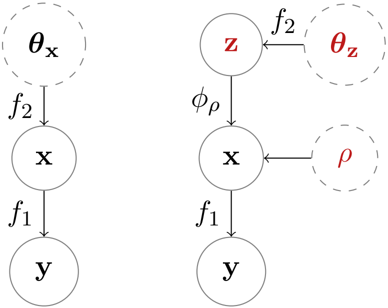
# Outline

# Splitted Gibbs sampling (SP)

$$\pi(\mathbf{x}) \propto \exp\left[-f_1(\mathbf{x}) - f_2(\mathbf{x})\right]$$

$$\Downarrow$$

$$\pi_\rho(\mathbf{x}, \mathbf{z}) \propto \exp\left[-f_1(\mathbf{x}) - f_2(\mathbf{z}) - \frac{1}{2\rho^2}\|\mathbf{x} - \mathbf{z}\|_2^2\right]$$

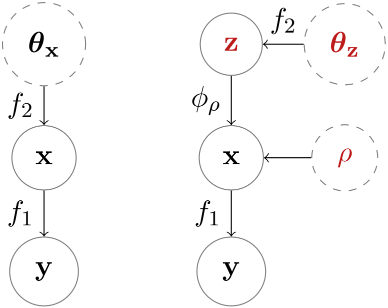


# Splitted Gibbs sampling (SP)

$$\pi(\mathbf{x}) \propto \exp\left[-f_1(\mathbf{x}) - f_2(\mathbf{x})\right]$$

$$\Downarrow$$

$$\pi_\rho(\mathbf{x}, \mathbf{z}) \propto \exp\left[-f_1(\mathbf{x}) - f_2(\mathbf{z}) - \phi_\rho(\mathbf{x}, \mathbf{z})\right]$$



# Splitted Gibbs sampling (SP): Theorem 1

Consider the marginal of $\mathbf{x}$ under $\pi_\rho$:

$$p_\rho(\mathbf{x}) = \int_{\mathbb{R}^d} \pi_\rho(\mathbf{x}, \mathbf{z}) \mathrm{d}\mathbf{z} \propto \int_{\mathbb{R}^d} \exp\left[-f_1(\mathbf{x}) - f_2(\mathbf{z}) - \phi_\rho(\mathbf{x}, \mathbf{z})\right] \mathrm{d}\mathbf{z} \ .$$

**Theorem**

*Assume that in the limiting case $\rho \to 0$, $\phi_\rho$ is such that*

$$\frac{\exp\left(-\phi_\rho(\mathbf{x}, \mathbf{z})\right)}{\int_{\mathbb{R}^d} \exp\left(-\phi_\rho(\mathbf{x}, \mathbf{z})\right) \mathrm{d}\mathbf{x}} \xrightarrow[\rho \to 0]{} \delta_{\mathbf{x}}(\mathbf{z})$$

*Then $p_\rho$ coincides with $\pi$ when $\rho \to 0$, that is*

$$\|p_\rho - \pi\|_{\mathrm{TV}} \xrightarrow[\rho \to 0]{} 0$$

# Splitted Gibbs sampling (SP): marginal distributions

Full conditional distributions under the split distribution $\pi_\rho$:

$$\pi_\rho(\mathbf{x}|\mathbf{z}) \propto \exp\left(-f_1(\mathbf{x}) - \phi_\rho(\mathbf{x},\mathbf{z})\right)$$

$$\pi_\rho(\mathbf{z}|\mathbf{x}) \propto \exp\left(-f_2(\mathbf{z}) - \phi_\rho(\mathbf{x},\mathbf{z})\right).$$

**Note that** $f_1$ and $f_2$ are now separated **in 2 distinct distributions**

# Splitted Gibbs sampling (SP): marginal distributions

Full conditional distributions under the split distribution $\pi_\rho$:

$$\pi_\rho(\mathbf{x}|\mathbf{z}) \propto \exp\left(-f_1(\mathbf{x}) - \frac{1}{2\rho^2}\|\mathbf{x}-\mathbf{z}\|_2^2\right)$$

$$\pi_\rho(\mathbf{z}|\mathbf{x}) \propto \exp\left(-f_2(\mathbf{z}) - \frac{1}{2\rho^2}\|\mathbf{x}-\mathbf{z}\|_2^2\right).$$

**Note that** $f_1$ and $f_2$ are now separated **in 2 distinct distributions**

**State of the art sampling methods:**

- P-MYULA = proximal MCMC, (Pereyra 2016; Durmus et al. 2018)
- Fourier or Aux-V1 or E-PO for Gaussian variables
- ...

$$\mathbf{y} = \mathbf{P}\mathbf{x} + \mathbf{n},$$

where $\mathbf{P}$ = damaging operator and $\mathbf{n} \sim \mathcal{N}\left(\mathbf{0}_d, \sigma^2 \mathbf{I}_d\right)$ = noise.

$$\begin{cases} f_1(\mathbf{x}) & = \dfrac{1}{2\sigma^2} \left\| \mathbf{y} - \mathbf{P}\mathbf{x} \right\|_2^2 \quad \forall \mathbf{x} \in \mathbb{R}^d, \\ f_2(\mathbf{x}) & = \tau \psi(\mathbf{x}), \quad \tau > 0. \end{cases}$$

Then the SP conditional distributions are:

$$\pi_\rho(\mathbf{x}|\mathbf{z}) = \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{x}}, \mathbf{Q}_{\mathbf{x}}{}^{-1}\right)$$
$$\pi_\rho(\mathbf{z}|\mathbf{x}) \propto \exp\left(-\tau \psi(\mathbf{z}) - \frac{1}{2\rho^2} \left\| \mathbf{z} - \mathbf{x} \right\|_2^2\right),$$

# Splitted Gibbs sampling (SP): efficient sampling

## Linear Gaussian inverse problems

$$\pi_\rho(\mathbf{x}|\mathbf{z}) = \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{x}}, \mathbf{Q}_{\mathbf{x}}{}^{-1}\right)$$

$$\pi_\rho(\mathbf{z}|\mathbf{x}) \propto \exp\left(-\tau\psi(\mathbf{z}) - \frac{1}{2\rho^2}\|\mathbf{z}-\mathbf{x}\|_2^2\right),$$

**Examples:**

- Convex non-smooth

  $\psi(\mathbf{x}) = $ **TV**, $\ell_{\mathbf{1}}$ sparsity... $\Rightarrow$ **proximal MCMC**

- Tikhonov regularization

  $\psi(\mathbf{z}) = \|\mathbf{Qz}\|_2^2 \Rightarrow$ **Gaussian variables**

  (e.g. **P** or **Q** diagonalizable in Fourier$\rightarrow$ E-PO)

# Splitted Gibbs sampling (SP): TV deblurring

Posterior distribution

$$p(\mathbf{x}|\mathbf{y}) \propto \exp\left[-\frac{1}{2\sigma^2}\|\mathbf{P}\mathbf{x}-\mathbf{y}\|_2^2 - \beta\mathrm{TV}(\mathbf{x})\right]$$

where $\mathbf{P}$ = damaging operator (blur, binary mask...) and

$$\mathrm{TV}(\mathbf{x}) = \sum_{1\leq i,j\leq N} \left\|(\nabla\mathbf{x})_{i,j}\right\|_2$$

**Direct sampling is challenging**

1. generally high dimension of the image,
2. non-conjugacy of the TV-based prior,
3. non-differentiability of $g$ ($\neq$ Hamiltonian Monte Carlo algorithms)

# Splitted Gibbs sampling (SP): TV deblurring

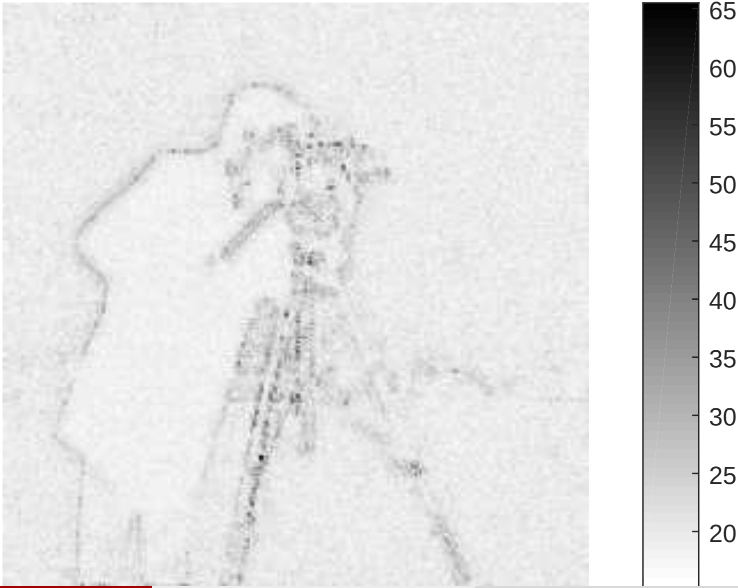
## Linear Gaussian inverse problems

# Splitted Gibbs sampling (SP): TV deblurring

## Linear Gaussian inverse problems

# Splitted Gibbs sampling (SP): TV deblurring

## Linear Gaussian inverse problems



# Splitted Gibbs sampling (SP): TV deblurring

## Linear Gaussian inverse problems

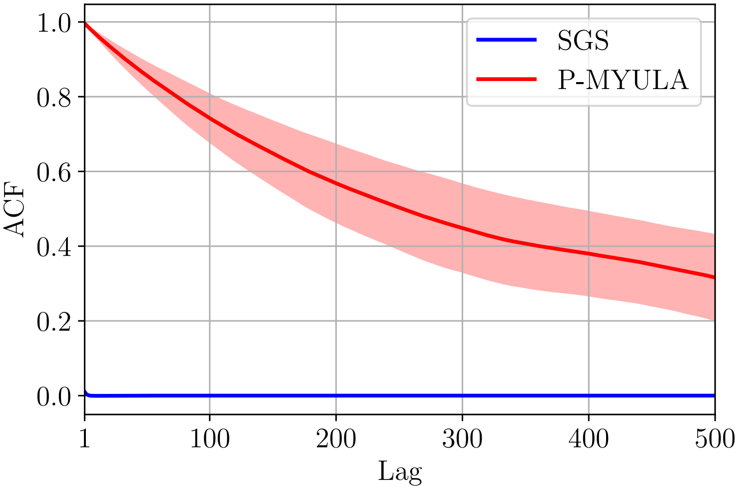
| | **SALSA** | **FISTA** | **SGS** | **P-MYULA** |
|---|---|---|---|---|
| time (s) | 1 | 10 | 470 | 3600 |
| time ($\times$ var. split.) | 1 | 10 | 1 | **7.7** |
| nb. iterations | 22 | 214 | $\sim 10^4$ | $10^5$ |
| SNR (dB) | 17.87 | 17.86 | **18.36** | 17.97 |

$$\text{Rk} : \rho^2 = 9$$

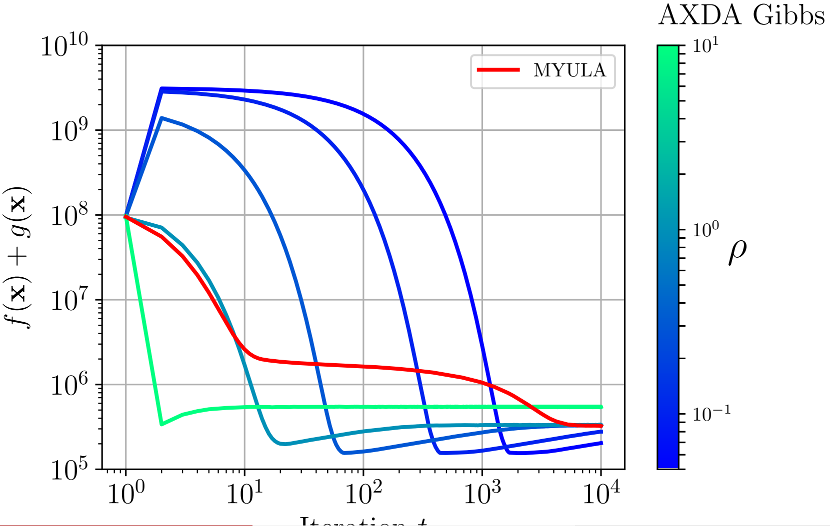
# Splitted Gibbs sampling (SP): TV deblurring

## Linear Gaussian inverse problems

Short auto-correlation of the Markov chain



$\rho$ = comput. time compromise/quality



# Outline

# Splitted & Augmented Gibbs sampling (SPA)

**Motivation for *augmentation*:**

**better mixing properties of the Markov chain**

$$\begin{aligned}\pi_{\rho,\alpha} &\triangleq p(\mathbf{x}, \mathbf{z}, \mathbf{u}; \rho, \alpha) \\ &\propto \exp\left[-f(\mathbf{x}) - g(\mathbf{z})\right] \\ &\times \exp\left[-\phi_1(\mathbf{x}, \mathbf{z} - \mathbf{u}; \rho) - \phi_2(\mathbf{u}; \alpha)\right]\end{aligned}$$

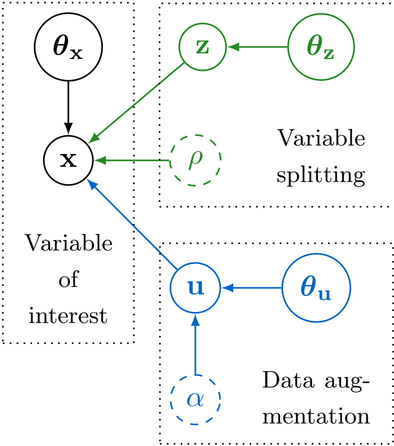
**Assumption 2**

$\phi_2$ and $\phi_1$ are such that $\forall \mathbf{x}, \mathbf{z} \in \mathbb{R}^d$,

$$\begin{aligned}\int_{\mathbb{R}^d} \exp\left[-\phi_1(\mathbf{x}, \mathbf{z} - \mathbf{u}; \rho) - \phi_2(\mathbf{u}; \alpha)\right] \mathrm{d}\mathbf{u} \\ \propto \exp\left[-\phi_1(\mathbf{x}, \mathbf{z}; \eta(\rho, \alpha))\right].\end{aligned} \tag{1}$$

# Splitted & Augmented Gibbs sampling (SPA)

$$\begin{aligned}\pi_{\rho,\alpha} &\triangleq p(\mathbf{x}, \mathbf{z}, \mathbf{u}; \rho, \alpha) \\ &\propto \exp\left[-f(\mathbf{x}) - g(\mathbf{z})\right] \\ &\times \exp\left[-\phi_1(\mathbf{x}, \mathbf{z} - \mathbf{u}; \rho) - \phi_2(\mathbf{u}; \alpha)\right]\end{aligned}$$



The conditional split-augmented distributions are:

$$p(\mathbf{x}|\mathbf{z},\mathbf{u};\rho) \propto \exp\left[-f(\mathbf{x}) - \phi_1(\mathbf{x},\mathbf{z}-\mathbf{u};\rho)\right]$$

$$p(\mathbf{z}|\mathbf{x},\mathbf{u};\rho) \propto \exp\left[-g(\mathbf{z}) - \phi_1(\mathbf{x},\mathbf{z}-\mathbf{u};\rho)\right]$$

$$p(\mathbf{u}|\mathbf{x},\mathbf{z};\rho,\alpha) \propto \exp\left[-\phi_2(\mathbf{u};\alpha)\right] \times \exp\left[-\phi_1(\mathbf{x},\mathbf{z}-\mathbf{u};\rho)\right].$$

The conditional split-augmented distributions are:

$$p(\mathbf{x}|\mathbf{z},\mathbf{u};\rho) \propto \exp\left[-f(\mathbf{x}) - \frac{1}{2\rho^2}\|\mathbf{x}-\mathbf{z}+\mathbf{u}\|_2^2\right]$$

$$p(\mathbf{z}|\mathbf{x},\mathbf{u};\rho) \propto \exp\left[-g(\mathbf{z}) - \frac{1}{2\rho^2}\|\mathbf{x}-\mathbf{z}+\mathbf{u}\|_2^2\right]$$

$$p(\mathbf{u}|\mathbf{x},\mathbf{z};\rho,\alpha) \propto \exp\left[-\frac{\|\mathbf{u}\|_2^2}{2\alpha^2} - \frac{1}{2\rho^2}\|\mathbf{x}-\mathbf{z}+\mathbf{u}\|_2^2\right].$$

By replacing each Gibbs sampling step by optimizations, ADMM appears:

**Algorithm 1:** ADMM (scaled version)

**Input**: Functions $f$, $g$, penalty $\rho^2$, init. $t \leftarrow 0$, $\mathbf{z}^{(0)}$,$\mathbf{u}^{(0)}$

**while** *stopping criterion not satisfied* **do**

$\quad \mathbf{x}^{(t)} \in \arg\min_{\mathbf{x}} -\log p\left(\mathbf{x}|\mathbf{z}^{(t-1)}, \mathbf{u}^{(t-1)}; \rho\right);$

$\quad \mathbf{z}^{(t)} \in \arg\min_{\mathbf{z}} -\log p\left(\mathbf{z}|\mathbf{x}^{(t)}, \mathbf{u}^{(t-1)}; \rho\right);$

$\quad \mathbf{u}^{(t)} = \mathbf{u}^{(t-1)} + \mathbf{x}^{(t)} - \mathbf{z}^{(t)}$ ;

$\quad t \leftarrow t + 1$ ;

**end**

**Output**: Approximate solution of the optimization problem $\hat{\mathbf{x}}$.

# Splitted Gibbs sampling (SPA): TV restoration

## Linear Gaussian inverse problems

The conditional distributions associated to SPA are:

$$p(\mathbf{x}|\mathbf{z},\mathbf{u}) \propto \exp\left[-\frac{1}{2\sigma^2}\left\|\mathbf{P}\mathbf{x}-\mathbf{y}\right\|_2^2\right] \times \exp\left[-\frac{1}{2\rho^2}\left\|\mathbf{x}-(\mathbf{z}-\mathbf{u})\right\|_2^2\right]$$

$$p(\mathbf{z}|\mathbf{x},\mathbf{u}) \propto \exp\left[-\beta \mathrm{TV}(\mathbf{z}) - \frac{1}{2\rho^2}\left\|\mathbf{z}-(\mathbf{x}+\mathbf{u})\right\|_2^2\right]$$

$$p(\mathbf{u}|\mathbf{x},\mathbf{z}) \propto \exp\left[-\frac{1}{2\alpha^2}\left\|\mathbf{u}\right\|_2^2 - \frac{1}{2\rho^2}\left\|\mathbf{u}-(\mathbf{z}-\mathbf{x})\right\|_2^2\right]$$

Rk: sampling from $p(\mathbf{z}|\mathbf{x},\mathbf{u}) \Rightarrow$ P-MYULA + Chambolle's algorithm

# Splitted Gibbs sampling (SPA): TV inpainting
## Linear Gaussian inverse problems

# Splitted Gibbs sampling (SPA): TV inpainting

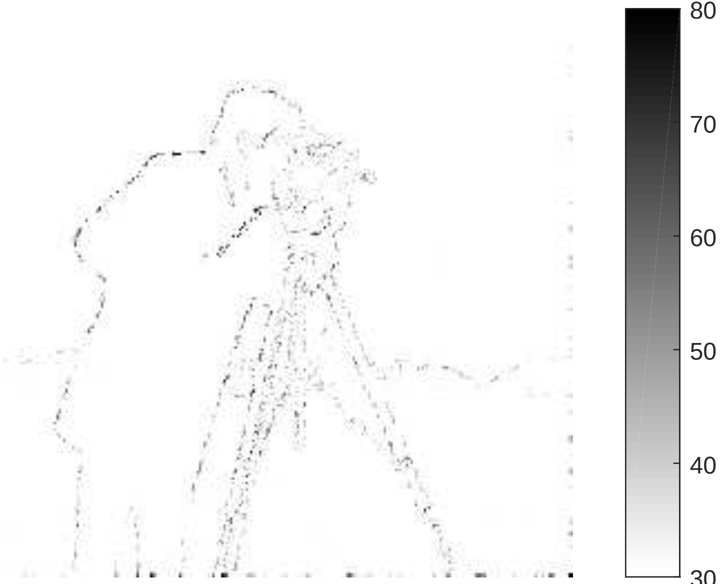
## Linear Gaussian inverse problems

# Splitted Gibbs sampling (SPA): TV inpainting

## Linear Gaussian inverse problems

# Splitted Gibbs sampling (SPA): **confidence intervals**

## Linear Gaussian inverse problems



# Splitted Gibbs sampling (SPA): TV inpainting

## Linear Gaussian inverse problems

| | **SALSA** | **P-MYULA** | **SP** | **SPA** |
|---|---|---|---|---|
| Balloons | 26.18 | 23.00 | 26.19 | 26.18 |
| Baboon | 14.37 | 13.35 | 14.60 | 14.59 |
| Elaine | 23.61 | 21.21 | 23.86 | 23.84 |
| Clock | 25.72 | 24.50 | 25.45 | 25.42 |
| Donna | 24.71 | 21.69 | 23.87 | 23.82 |
| House | 20.21 | 19.59 | 20.43 | 20.43 |
| Peppers | 20.35 | 19.20 | 20.22 | 20.20 |
| Cameraman | 19.48 | 18.76 | 19.34 | 19.34 |
| Boat | 20.81 | 19.80 | 20.74 | 20.71 |

**Many problems can be considered using SPA:**

- Laplacian + $\ell_2$ regularizer for deconvolution

  M. Vono et al., "Split-and-augmented Gibbs sampler - Application to large-scale inference problems," in *IEEE Trans. Signal Processing*, 2019

- Poisson noise + blur + non-negativity + ...

  M. Vono et al., "Bayesian image restoration under Poisson noise and log-concave prior," in *Proc. ICASSP 2019*

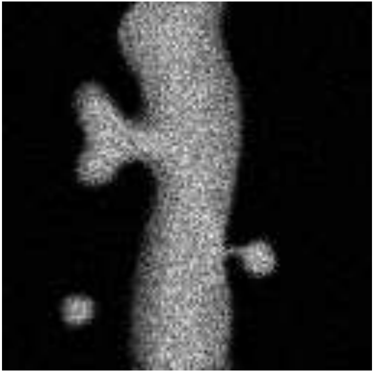
- Machine learning: logistic regression,...

  M. Vono et al. (2018), "Sparse Bayesian binary logistic regression using the split-and-augmented Gibbs sampler," in *Proc. IEEE MLSP 2018*

# Splitted & Augmented Gibbs sampling (SPA)

## Poisson denoising + deblurring: sparse wavelet transform + non-negativity

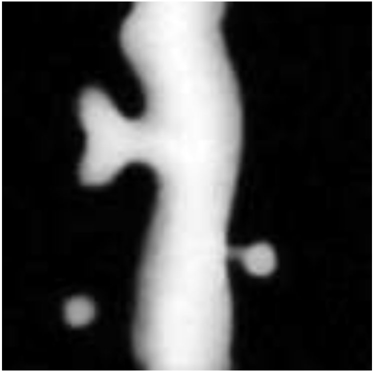
- 6 times faster than P-MYULA,
- no approximation (e.g., Anscombe in P-MYULA)
- ... but using SPA + P-MYULA !
- performances similar to PIDAL (Figueiredo & Bioucas-Dias 2010)
- **+ confidence intervals**



# Splitted & Augmented Gibbs sampling (SPA)

## Poisson denoising + deblurring: sparse wavelet transform + non-negativity

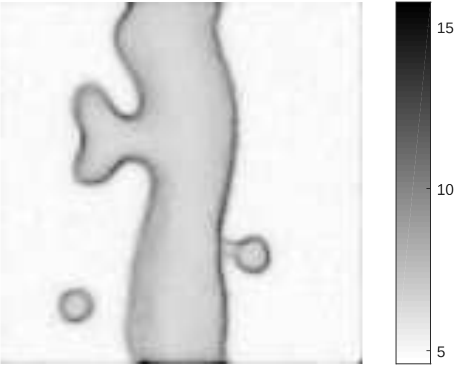
- 6 times faster than P-MYULA,
- no approximation (e.g., Anscombe in P-MYULA)
- ... but using SPA + P-MYULA !
- performances similar to PIDAL (Figueiredo & Bioucas-Dias 2010)
- **+ confidence intervals**



# Splitted & Augmented Gibbs sampling (SPA)

## Poisson denoising + deblurring: sparse wavelet transform + non-negativity

- 6 times faster than P-MYULA,
- no approximation (e.g., Anscombe in P-MYULA)
- ... but using SPA + P-MYULA !
- performances similar to PIDAL (Figueiredo & Bioucas-Dias 2010)
- **+ confidence intervals**



# Outline

Let $\pi \in L^1$ a target **probability distribution** with density with respect to (w.r.t.) the Lebesgue measure

$$\boxed{\pi(\mathbf{x}) \propto \exp(-f(\mathbf{x}))}$$

where $f : \mathcal{X} \subseteq \mathbb{R}^d \to (-\infty, +\infty]$ stands for a **potential** function.

With a slight abuse of notations, $\pi$ shall refer to

- a prior $\pi(\mathbf{x})$,
- a likelihood $\pi(\mathbf{x}) \triangleq \pi(\mathbf{y}|\mathbf{x})$,
- a posterior $\pi(\mathbf{x}) \triangleq \pi(\mathbf{x}|\mathbf{y})$,

where $\mathbf{y}$ are observations.

Let $\pi \in L^1$ a target **probability distribution** with density with respect to (w.r.t.) the Lebesgue measure

$$\boxed{\pi(\mathbf{x}) \propto \exp(-f(\mathbf{x}))}$$

where $f : \mathcal{X} \subseteq \mathbb{R}^d \to (-\infty, +\infty]$ stands for a **potential** function.

**Assumption 1**

Inference from $\pi$ is difficult and possibly inefficient.

**Examples:**

- non-trivial maximum likelihood estimation
- difficult posterior sampling with poor mixing chains

# Data augmentation (DA)

One surrogate is to introduce auxiliary variables $\mathbf{z} \in \mathcal{Z} \subseteq \mathbb{R}^k$ such that

$$\int_{\mathcal{Z}} \pi(\mathbf{x}, \mathbf{z}) \mathrm{d}\mathbf{z} = \pi(\mathbf{x}).$$

Numerous well-known **advantages**:

- augmented likelihood $\pi(\mathbf{x}, \mathbf{z}) \triangleq \pi(\mathbf{y}, \mathbf{z} | \mathbf{x})$ **easier** to work with
- joint posterior $\pi(\mathbf{x}, \mathbf{z}) \triangleq \pi(\mathbf{x}, \mathbf{z} | \mathbf{y})$ with **simpler** full conditionals
- **improved** inference (multimodal problems, mixing properties)

# The art of exact data augmentation: XDA

Unfortunately, satisfying

$$\int_{\mathcal{Z}} \pi(\mathbf{x}, \mathbf{z}) \mathrm{d}\mathbf{z} = \pi(\mathbf{x}) \quad \text{(XDA)}$$

is a matter of **art** (van Dyk and Meng 2001).

**Difficulties**:

- **finding** $\pi(\mathbf{x}, \mathbf{z})$ (Geman and Yang 1995)
- **scaling** in high-dimensional/big data settings (Neal 2003; Polson et al. 2013).

**Goal: relax (XDA) while keeping XDA's advantages**

# Asymptotically exact data augmentation (AXDA)

Let consider an augmented density $p_\rho(\mathbf{x}, \mathbf{z})$ and define

$$\pi_\rho(\mathbf{x}) = \int_{\mathcal{Z}} p_\rho(\mathbf{x}, \mathbf{z}) \mathrm{d}\mathbf{z},$$

where $\rho > 0$.

**Assumption 2**

For all $\mathbf{x} \in \mathcal{X}$, $\lim_{\rho \to 0} \pi_\rho(\mathbf{x}) = \pi(\mathbf{x})$.

**Theorem 1** (Scheffé 1947)

Under Assumption 2,

$$\|\pi_\rho - \pi\|_{\mathrm{TV}} \xrightarrow[\rho \to 0]{} 0.$$

# Choice of the augmented density

Take inspiration from variable splitting in optimization (Boyd et al. 2011)...

This motivates the choice (Vono et al. 2019)

$$\boxed{p_\rho(\mathbf{x}, \mathbf{z}) \propto \exp(-f(\mathbf{z}) - \phi_\rho(\mathbf{x}, \mathbf{z}))}$$

- **simplify** the inference (splitting complicated potentials) (Vono et al. 2019)
- **distribute** the inference (Rendell et al. 2018)
- **accelerate** the inference (Vono et al. 2019).

# Properties based on convolution

**(H1)** $\pi \in L^1$ is log-concave.
**(H2)** $\phi_\rho(\mathbf{x},\mathbf{z}) = \tilde{\phi}_\rho(\mathbf{x}-\mathbf{z})$, such that

$$\pi_\rho(\mathbf{x}) = \int_{\mathcal{Z}} p_\rho(\mathbf{x},\mathbf{z})\mathrm{d}\mathbf{z}$$

$K_\rho \propto \exp(-\tilde{\phi}_\rho)$ is $\mathcal{C}^\infty$ log-concave,
$\forall k \geq 0$, $\partial^k K_\rho$ is bounded
$\lim_{\rho\to 0} K_\rho(\mathbf{u}) = \delta(\mathbf{u})$ with $\mathbb{E}_{K_\rho}(U) = 0$.

Then,

i) $\pi_\rho \xrightarrow[\rho\to 0]{} \pi$

ii) $\pi_\rho$ is log-concave

iii) $\pi_\rho$ is infinitely differentiable on $\mathcal{X}$

iv) $\pi(\mathbf{x}) \Longrightarrow \mathbb{E}_{\pi_\rho}(X) = \mathbb{E}_\pi(X)$

$$\mathrm{var}_{\pi_\rho}(X) = \mathrm{var}_\pi(X) + \mathrm{var}_{K_\rho}(X).$$

# Non-asymptotic bound on the TV distance

**(H3)** $f$ is $L_f$-Lipschitz,

**(H4)** $\phi_\rho(\mathbf{x}, \mathbf{z}) = \frac{1}{2\rho^2} \|\mathbf{x} - \mathbf{z}\|_2^2$.

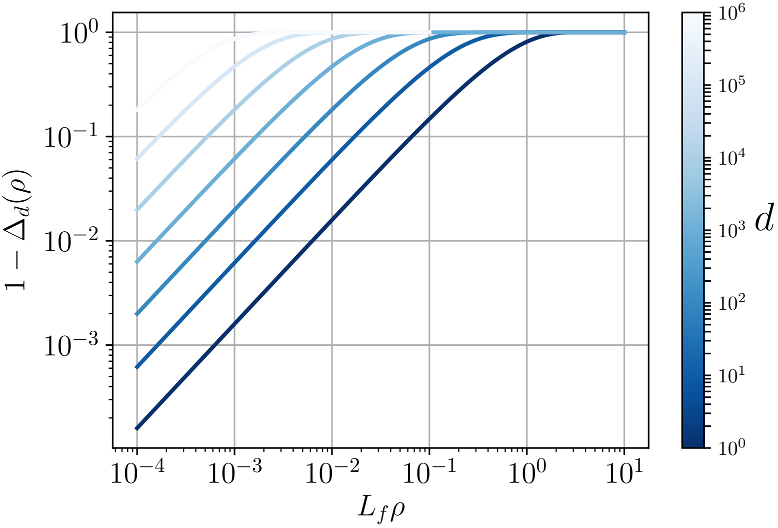
Let $d = \dim(\mathcal{X})$. For all $\rho > 0$,

$$\|\pi_\rho - \pi\|_{\mathrm{TV}} \leq 1 - \Delta_d(\rho) = 1 - \frac{D_{-d}(L_f \rho)}{D_{-d}(-L_f \rho)}$$

$$1 - \Delta_d(\rho) \underset{\rho \to 0}{\sim} \frac{2\sqrt{2}\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} L_f \rho$$

The function $D_{-d}$ is the parabolic cylinder special function.

# Behavior when $\rho \to 0$ & illustration



# Bounds on potentials

$$f_\rho(\mathbf{x}) = \frac{d}{2}\log(2\pi\rho^2) - \log\int_{\mathcal{X}} \exp\left(-f(\mathbf{z}) - \frac{1}{2\rho^2}\|\mathbf{z}-\mathbf{x}\|_2^2\right)\mathrm{d}\mathbf{z}$$

For all $\rho > 0$ and $\mathbf{x} \in \mathcal{X}$,

$$\boxed{L_\rho \;\leq\; f_\rho(\mathbf{x}) - f(\mathbf{x}) \;\leq\; U_\rho}$$

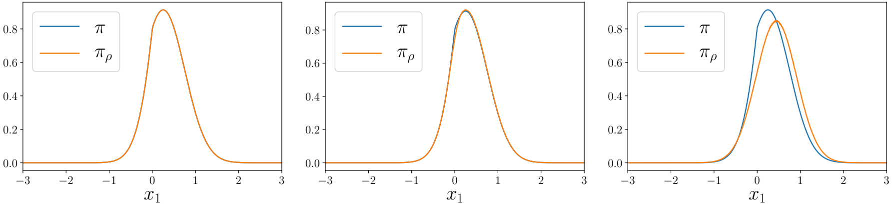
with

$$\begin{aligned}
L_\rho &= \log M_\rho - \log D_{-d}(-L_f\rho)\\
U_\rho &= \log M_\rho - \log D_{-d}(L_f\rho)\\
M_\rho &= \frac{2^{d/2-1}\Gamma\left(d/2\right)}{\Gamma(d)\exp\left(L_f^2\rho^2/4\right)}.
\end{aligned}$$

# Bounds on credibility intervals

## Illustration

$$(1-\alpha)\frac{M_\rho}{D_{-d}(-L_f\rho)} \;\leq\; \int_{\mathcal{C}_\alpha^\rho} \pi(\mathbf{x})\mathrm{d}\mathbf{x} \;\leq\; \min\left(1,(1-\alpha)\frac{M_\rho}{D_{-d}(L_f\rho)}\right)$$



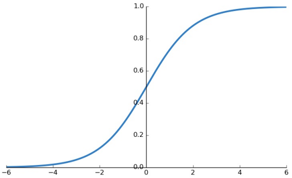

| $\rho$ | $\mathcal{C}_\alpha$ | $\mathcal{C}_\alpha^\rho$ | $\int_{\mathcal{C}_\alpha^\rho}\pi(x_1)\mathrm{d}x_1$ | $\mathcal{I}_\alpha^\rho$ |
|---|---|---|---|---|
| $10^{-2}$ | [-0.47,1.24] | [-0.47,1.24] | 0.95 | [0.948,0.952] |
| $10^{-1}$ | idem | [-0.47,1.24] | 0.95 | [0.88,1] |
| 1 | idem | [-0.47,1.37] | 0.96 | [0.34,1] |

# Distributed sampling and data privacy

## Regularized logistic regression

$$\forall\ i \in [\![1, n]\!], \quad y_i \sim \text{Bernoulli}\left(\sigma(\mathbf{a}_i^T \mathbf{x})\right)$$

$$\pi(\mathbf{x}|\mathbf{y}) \propto \exp\left(-f(\mathbf{x}) - \sum_{j=1}^{b} g^{(j)}(\mathbf{x})\right)$$




where

- $g^{(j)}(\mathbf{x}) = \sum_{i \in \mathcal{D}_j} \log\left(1 + \exp\left(-y_i \mathbf{a}_i^T \mathbf{x}\right)\right)$,
- $\mathcal{D}_j$ indices associated to the $j$th block of data,
- $f$ = prior on the regressor $\mathbf{x}$

**Issues**:

- the full data set is distributed over $b$ nodes, $b \in [\![1, n]\!]$
- data privacy.

# Distributed sampling and data privacy

## Applying AXDA $b$ times

$$p_\rho(\mathbf{x}, \mathbf{z}_{1:b}) \propto \exp\left(-f(\mathbf{x}) - \sum_{j=1}^{b}\left[\frac{1}{2\rho^2}\left\|\mathbf{x} - \mathbf{z}_j\right\|^2 + \sum_{i \in \mathcal{D}_j} \log\left(1 + \exp\left(-y_i \mathbf{a}_i^T \mathbf{z}_j\right)\right)\right]\right)$$

**Benefits of AXDA**:

- inference via a Gibbs sampler distributed on $b$ nodes
- the master node never *sees* the data set: **privacy**
- theoretical guarantees on the approximation

# Conclusion

- **SP & SPA split-and-augment strategy**
  - Bayesian inference for complex models
  - large scale problems (big & tall)
  - **confidence intervals**
- **Efficient algorithms** for inference
  - **acceleration** of state-of-the-art sampling algorithms
  - **distributed** inference (simulation, optimization, variational approx.)
- **AXDA**: **unifying** statistical framework
  - asymptotically exact: control parameter $\rho$
  - **non-asymptotic theoretical guarantees** on the approximation under mild assumptions

# Interested in AXDA for your statistical problems?

## Theory and methods

- M. Vono et al. (2019), "Asymptotically exact data augmentation: models, properties and algorithms". Technical report. https://arxiv.org/abs/1902.05754/
- M. Vono et al. (2019), "Split-and-augmented Gibbs sampler - Application to large-scale inference problems," *IEEE Transactions on Signal Processing*.
- L. J. Rendell et al. (2018), "Global consensus Monte Carlo". Technical report. https://arxiv.org/abs/1807.09288/

## Applications

- M. Vono et al. (2019), "Bayesian image restoration under Poisson noise and log-concave prior," in *Proc. ICASSP*.
- M. Vono et al. (2018), "Sparse Bayesian binary logistic regression using the split-and-augmented Gibbs sampler," in *Proc. MLSP*.

## Code

- https://github.com/mvono

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002), "Approximate Bayesian Computation in Population Genetics," *Genetics*, 162, 2025–2035.

Besag, J. and Green, P. J. (1993), "Spatial Statistics and Bayesian Computation," *Journal of the Royal Statistical Society, Series B*, 55, 25–37.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011), "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *Foundations and Trends in Machine Learning*, 3, 1–122.

Damien, P., Wakefield, J., and Walker, S. (1999), "Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables," *Journal of the Royal Statistical Society, Series B*, 61, 331–344.

Del Moral, P., Doucet, A., and Jasra, A. (2012), "An adaptive sequential Monte Carlo method for approximate Bayesian computation," *Statistics and Computing*, 22, 1009–1020.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, Series B*, 39, 1–38.

Doucet, A., Godsill, S. J., and Robert, C. P. (2002), "Marginal maximum a posteriori estimation using Markov chain Monte Carlo," *Statistics and Computing*, 12, 77–84.

Durmus, A., Moulines, E., and Pereyra, M. (2018), "Efficient Bayesian Computation by Proximal Markov chain Monte Carlo: When Langevin Meets Moreau," *SIAM Journal on Imaging Sciences*, 11, 473–506.

Filstroff, L., Lumbreras, A., and Févotte, C. (2018), "Closed-form Marginal Likelihood in Gamma-Poisson Matrix Factorization," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, vol. 80, pp. 1506–1514.

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

Ising, E. (1925), "Beitrag zur Theorie des Ferromagnetismus," *Zeitschrift für Physik*, 31, 253–258.

Neal, R. M. (2003), "Slice sampling," *The Annals of Statistics*, 31, 705–767.

Peel, D. and McLachlan, G. J. (2000), "Robust mixture modelling using the t distribution," *Statistics and Computing*, 10, 339–348.

Pereyra, M. (2016), "Proximal Markov chain Monte Carlo algorithms," *Statistics and Computing*, 26, 745–760.

Polson, N. G., Scott, J. G., and Windle, J. (2013), "Bayesian Inference for Logistic Models Using Polya-Gamma Latent Variables," *Journal of the American Statistical Association*, 108, 1339–1349.

Potts, R. B. (1952), "Some generalized order-disorder transformations," *Mathematical Proceedings of the Cambridge Philosophical Society*, 48, 106–109.

Ratmann, O., Andrieu, C., Wiuf, C., and Richardson, S. (2009), "Model criticism based on likelihood-free inference, with an application to protein network evolution," *Proceedings of the National Academy of Sciences*, 106, 10576–10581.

Rendell, L. J., Johansen, A. M., Lee, A., and Whiteley, N. (2018), "Global consensus Monte Carlo," [online]. Technical report. Available at `https://arxiv.org/abs/1807.09288/`.

Scheffé, H. (1947), "A useful convergence theorem for probability distributions," *The Annals of Mathematical Statistics*, 18, 434–438.

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I., and McCulloch, R. E. (2016), "Bayes and Big Data: The Consensus Monte Carlo Algorithm," *International Journal of Management Science and Engineering Management*, 11, 78–88.

Sisson, S., Fan, Y., and Beaumont, M. (eds.) (2018), *Handbook of Approximate Bayesian Computation*, Chapman and Hall/CRC Press, 1st ed.

Swendsen, R. H. and Wang, J.-S. (1987), "Nonuniversal critical dynamics in Monte Carlo simulations," *Physical Review Letters*, 58, 86–88.

van Dyk, D. A. and Meng, X.-L. (2001), "The Art of Data Augmentation," *Journal of Computational and Graphical Statistics*, 10, 1–50.

Vono, M., Dobigeon, N., and Chainais, P. (2018), "Sparse Bayesian binary logistic regression using the split-and-augmented Gibbs sampler," in *IEEE International Workshop on Machine Learning for Signal Processing*, Aalborg, Denmark.

— (2019), "Split-and-augmented Gibbs sampler - Application to large-scale inference problems," *IEEE Transactions on Signal Processing*, 67, 1648–1661.

Wang, C. and Blei, D. M. (2018), "A General Method for Robust Bayesian Modeling," *Bayesian Analysis*, 13, 1163–1191.