Majorize-Minimize adapted Metropolis-Hastings algorithm. Application to multichannel image recovery
Yosra Marnissi, Amel Benazza-Benyahia, Emilie Chouzenoux,
Jean-Christophe Pesquet

To cite this version:
Yosra Marnissi, Amel Benazza-Benyahia, Emilie Chouzenoux, Jean-Christophe Pesquet. Majorize-Minimize adapted Metropolis-Hastings algorithm. Application to multichannel image recovery. 22th European Signal Processing Conference (EUSIPCO 2014), Sep 2014, Lisbon, Portugal. hal-01077273

HAL Id: hal-01077273
https://hal.archives-ouvertes.fr/hal-01077273
Submitted on 24 Oct 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
MAJORIZE-MINIMIZE ADAPTED METROPOLIS–HASTINGS ALGORITHM.
APPLICATION TO MULTICHANNEL IMAGE RECOVERY

Y. Marnissi1, A. Benazza-Benyahia2, E. Chouzenoux1, and J.-C. Pesquet1

1 Université Paris-Est, LIGM, UMR CNRS 8049, Champs sur Marne, France
2 COSIM Lab., SUP’COM, Carthage Univ., Cité Technologique des Communications, Tunisia

ABSTRACT
One challenging task in MCMC methods is the choice of the proposal density. It should ideally provide an accurate approximation of the target density with a low computational cost. In this paper, we are interested in Langevin diffusion where the proposal accounts for a directional component. We propose a novel method for tuning the related drift term. This term is preconditioned by an adaptive matrix based on a Majorize-Minimize strategy. This new procedure is shown to exhibit a good performance in a multispectral image restoration example.

Index Terms— MCMC methods, Langevin diffusion, Majorize-Minimize, MMSE, multichannel image restoration.

1. INTRODUCTION
Recovering the signal of interest from degraded observations embedded in an additive noise is a key issue for many applications such as remote sensing imaging. In this respect, a Bayesian framework can be adopted to compute the Minimum Mean Squared Estimator (MMSE). However, it is not always possible to derive a closed expression of the posterior distribution involved in the MMSE. To alleviate this problem, Markov Chain Monte Carlo (MCMC) approaches have been developed. They consist of constructing an irreducible Markov chain whose stationary distribution is the unknown posterior distribution. Building the Markov chain corresponds to a specific way of exploring the state space. For this purpose, Metropolis–Hastings (MH) algorithms have been intensively used [1]. The key issue however is the choice of a proposal density. Recent methods such as those based on Langevin diffusion have incorporated a directional component for the proposal [2]. More precisely, two parameters (a stepsize and a scale matrix) are introduced to guide the directional component. The problem of setting the scale matrix must be carefully addressed, especially for high dimensional problems. Several solutions have been considered [2–4]. In this work, we propose a novel approach for choosing the scale matrix based on a Majorize-Minimize (MM) strategy. The paper is organized as follows: In Section 2, we formulate the problem and we give a brief overview of the Langevin diffusion process. In Section 3, we describe the new MM adapted MH algorithm. Section 4 is devoted to experimental results for multicomponent image restoration. Finally, some concluding remarks are drawn in Section 5.

2. RELATED WORKS

2.1. Problem statement
In this paper, we address a wide array of problems where the vector $x$ in $\mathbb{R}^Q$ of samples of an unknown signal $X$ is estimated from the observation vector $z$ in $\mathbb{R}^N$ given the following observation model:

$$z = Hx + w$$

where the matrix $H \in \mathbb{R}^{N \times Q}$ corresponds to a linear degradation operator, and $w$ in $\mathbb{R}^N$ is the vector of a Gaussian additive noise samples. In this work, we adopt a Bayesian framework and we aim at computing the MMSE $\hat{x}_{\text{MMSE}} = \mathbb{E}_{x \mid z} x$ where the posterior distribution $\pi_X$ is related to the prior distribution $p_X$ of the unknown vector $X$:

$$\pi_X(x) \propto p_X(x)p(z|x).$$

Generally, the computation of $\hat{x}_{\text{MMSE}}$ involves integrals that are both analytically and numerically intractable. The Monte Carlo approach is a classical alternative solution which consists of simulating a sufficient number of i.i.d. random variables from the posterior distribution $\pi_X$ and approximating the MMSE estimator by the empirical average over all these samples. However, the target posterior is often complex and does not present a closed form, so that direct sampling is not always possible. To alleviate this difficulty, MCMC methods have been developed. They consist of building an irreducible Markov chain whose stationary distribution is $\pi_X$. The asymptotic state of the chain is then considered as a sample of the target distribution. From an initial state $x^0$, the problem reduces to exploring the state space according to the transition probabilities that characterize the Markov chain. Different ways of moving from a state to another have been reported. Among them, much attention was paid to MH algorithms that generate a random walk according to a pro-
propositional density and implement a method for rejecting proposed moves [1].

### 2.2. Langevin diffusion

The choice of the proposal distribution is crucial as it impacts the statistical properties of the resulting Markov chain especially for complex and high-dimensional target distributions. Advanced MH methods introduce a directional component for the proposal. In this respect, Langevin diffusion strategies adjust the state transition by accounting for the gradient direction and implement a method for rejecting proposed values than a standard random walk. Indeed, most of the samples should lie and hence, enables to achieve a chain to be guided toward regions of higher probability, where two scale parameters play an important role: the length of proposed jumps whereas 

determines the determinant of this matrix. It can be proved that the Langevin process has as its stationary distribution and is more likely to accept proposed values than a standard random walk. Indeed, the gradient information of the target distribution allows the chain to be guided toward regions of higher probability, where most of the samples should lie and hence, enables to achieve high acceptance rates. To this end, it is worth noting that the two scale parameters play an important role: determines the length of proposed jumps whereas controls the direction.

Three classes of algorithms have been developed from this diffusion depending on the choice of .

### 2.3. Choice of the scale matrix

The standard Metropolis adjusted Langevin algorithm (MALA) is the simplest form of this diffusion when equals , the identity matrix of , the smallest sample. Let be a differentiable function with an Lipschitzian gradient and .

Recent algorithms [6–10] propose adaptive procedures where is tuned according to the past behavior of the Markov chain resorting to some deterministic optimization tools. For example, when setting to the inverse of the Hessian matrix of 

then, a new sample of the Newton-based MCMC is more likely drawn from a highly probable region and then more likely accepted, which can speed up the convergence of the sampling process [6–8]. However, in practice, this method has a high computational load since it requires the computation of the Hessian matrix and its inverse at each iteration. This is particularly critical for large scale problems and/or when the Hessian matrix is not definite positive. One appealing solution is to replace the Hessian by a scale matrix that can efficiently accelerate the algorithm with a lower computational cost. In particular, many methods have proposed the Fisher information matrix as a preconditioning matrix in the Langevin diffusion [9, 10] which can be interpreted as the discretization of the MALA algorithm directly on a natural Riemannian manifold where the parameters live. In this work, we propose a new approach where the scale matrix of the Langevin diffusion is chosen according to a Majorize-Minimize strategy.

### 3. MAJORIZE-MINIMIZE ADAPTED METROPOLIS–HASTINGS

We focus on the case when the minus-log of the target density function can be expressed as:

\[
\forall x \in \mathbb{R}^Q \quad \Theta(x) = \Phi(Hx - z) + \Psi(x),
\]

where \( z \in \mathbb{R}^N \), \( H \in \mathbb{R}^{N \times Q} \), \( \Phi \) is a continuous coercive differentiable function with an -Lipschitzian gradient and \( \Psi(x) = \sum_{s=1}^{S} \psi_s(\|V_s x - c_s\|) \),

where \( \psi_s \in \{1, ..., S\} \) \( V_s \in \mathbb{R}^{P_s \times Q} \), \( c_s \in \mathbb{R}^{P_s} \) and \( \{\psi_s\}_{1 \leq s \leq S} \) is a set of positive continuous functions satisfying the following properties:

- \( \forall s \in \{1, ..., S\} \) \( \psi_s \) is a differentiable function,
- \( \forall s \in \{1, ..., S\} \) \( \psi_s'(\sqrt{u}) \) is concave over \( \mathbb{R}^+ \),
- \( \forall s \in \{1, ..., S\} (3 \omega_s > 0) \) such that \( \forall u > 0 \) \( 0 \leq \psi_s(u) \leq \omega_s u \) and \( \lim_{t \to 0} \psi_s(u)/u < \infty \).

The minimization of (6) using the MM approach consists of performing successive minimizations of its tangent majorant functions [11]. Let \( x' \in \mathbb{R}^Q \). A function \( f \) is said to be a tangent majorant function of \( \Theta \) at \( x' \) provided that

\[
\left\{ \begin{array}{l}
  f(x', x') = \Theta(x'), \\
  f(x, x') \geq \Theta(x) \quad (\forall x \in \mathbb{R}^Q).
\end{array} \right.
\]
Let us assume the existence, for every $x' \in \mathbb{R}^Q$, of a positive definite matrix $Q(x') \in \mathbb{R}^{Q \times Q}$ such that the following quadratic function, defined for every $x \in \mathbb{R}^Q$,

$$f(x, x') = \Theta(x') + (x - x')^\top \nabla \Theta(x') + \frac{1}{2} (x - x')^\top Q(x')(x - x')$$

(9)

is a tangent majorant of (6) at $x'$. Then, the MM optimization algorithm reduces to building a sequence $(x^t)_{t \in \mathbb{N}}$ through the following scheme:

$$(\forall t \in \mathbb{N}) \quad x^{t+1} = x^t + \frac{\varepsilon^2}{2} Q^{-1}(x^t) \nabla \log \pi_X(x^t),$$

(10)

with $\varepsilon \in (0, \sqrt{2}]$. According to the majorization properties (8), the MM update rule (10) will produce a monotonically decreasing sequence $(\Theta(x^t))_{t \in \mathbb{N}}$ that converges to a local minimum of $\Theta$. We take up this idea to speed up the Langevin diffusion by using the inverse of the curvature matrix $Q(x')$ as a scale matrix in (4). Similarly to Newton-based MCMC algorithms, the drift term, assuming zero curvature changes, proposes, from a current state $x'$, a state with a higher value of $\log \pi_X$, resulting from an iteration of MM algorithm on $- \log \pi_X$. Then, the obtained proposal reduces to a noisy version of a MM iteration for minimizing $- \log \pi_X$. Since the deterministic MM optimization approach can suffer from convergence to a local minimum in the nonconvex case, the addition of the noise component can solve this issue. The resulting 3MH sampler is described by Algorithm 1.

**Algorithm 1:** Majorize–Minimize adapted Metropolis–Hastings algorithm

1. **Initialize** $x^0, t = 0, \varepsilon \in (0, \sqrt{2}]$.
2. **Compute** $A(x^t) = Q^{-1}(x^t)$ and $g(x^t) = \nabla \log \pi_X(x^t)$
3. **Generate** $x^t \sim q(x^t, \cdot)$, where $q(x^t, \cdot) = N'(x^t + \frac{\varepsilon^2}{2} A(x^t) g(x^t), \varepsilon^2 A(x^t))$
4. **Accept with probability**

$$\alpha(x^t, x^* ) = \min \left(1, \frac{\pi_X(x^*) q(x^t, x^*)}{\pi_X(x^t) q(x^*, x^t)} \right)$$

5. **Set** $t \leftarrow t + 1$ and go to 1 until convergence.

Moreover, $\zeta \geq 0$ is a constant that ensures the invertibility of $Q(x)$ for every $x \in \mathbb{R}^Q$. In the context of large scale problems, the inversion of the curvature matrix (11) at each iteration may become intractable. We thus also propose to resort to the following alternative choice described in [13], which can be understood as a diagonal approximation of (11):

$$\forall x \in \mathbb{R}^Q \quad Q(x) = (\mu \|H\|^2 + \zeta) I_Q + \text{Diag}(P^\top \omega(x)),$$

(13)

where $I_Q$ is the unit vector of $\mathbb{R}^Q$ and $P \in \mathbb{R}^{P \times Q}$, with $P = \sum P_s$, is the matrix whose elements are given by

$$(\forall i \in \{1, \ldots, P\})(\forall j \in \{1, \ldots, Q\}) P_{i,j} = |V_{i,j}|$$

4. **Experimental results**

**Objective**

The experiments we carried out deal with the recovery of a multicomponent image with $B$ components degraded by a blur modelled by a linear operator $D$ and an additive Gaussian noise $w$ with covariance matrix $A$. The restoration is performed in the wavelet transform domain: our objective is to compute the MMSE of $x \in \mathbb{R}^Q$ through the linear model defined in (1) with $H = DF^\top$ where $F^\top$ denotes a linear synthesis wavelet operator. Note that the wavelet coefficients are grouped into $M$ subbands of size $Q_m, m \in \{1, \ldots, M\}$ and, for each subband $m$, we can extract the set of vectors $(x_{m,q})_{q=1}^{Q_m}$ containing the wavelet coefficients located at the same spatial position $q$ through all the $B$ channels using a $B \times Q$ permutation matrix $P_{m,q}$ such as $x_{m,q} = P_{m,q} x$.

**Prior distribution**

Similarly to [14], we assume that the vectors $(x_{m,q})_{q=1}^{Q_m}$ are realizations of a random vector with a Generalized Multivariate Exponential Power (GMEP) distribution whose multivariate probability density function $p_{\text{GMEP}}$ is defined, for every $u \in \mathbb{R}^B$, by

$$p_{\text{GMEP}}(u; \theta_m) = C_m |\Sigma_m|^{1/2} y \left(u^\top \Sigma_m^{-1} u; \beta_m, \delta_m\right),$$

(14)

where $\theta_m = \{\beta_m > 0, \delta_m > 0, \Sigma_m\}$, for every $t \in \mathbb{R}^+$, $y(t; \beta_m, \delta_m) = \exp \left(- \frac{1}{2} (t + \delta_m)^{\beta_m}\right)$, and $C_m$ is the associated normalization constant. $\Sigma_m$ is related to the covariance matrix $\Gamma_m$ through $\Sigma_m = K_{\beta_m, \delta_m}^2 \Gamma_m$ with

$$K_{\beta_m, \delta_m}^2 = \frac{\int_0^\infty t^{\beta_m - 1} e^{-\frac{1}{2} \left(t + \delta_m\right)^{\beta_m}} dt}{\int_0^\infty \frac{1}{2} e^{-\frac{1}{2} (t + \delta_m)^{\beta_m}} dt}.$$  

(15)

The GMEP is an elliptical distribution that reflects the sparsity of the coefficients and, following [15] it can be proved that it is a scale mixture of Gaussian distributions when $\beta_m < 1$. Its mixing density is expressed as follows:

$$h_{\beta_m, \delta_m}(\nu) = \frac{2^{\frac{1}{2}} \Gamma\left(\frac{1}{2}\right) \nu^{\beta_m - 3} \delta_m^{3/2} \left(\frac{1}{2} \nu^{-2} - \frac{\nu^{-1}}{\delta_m}\right)}{\int_0^\infty \frac{1}{2} e^{-\frac{1}{2} (t + \delta_m)^{\beta_m}} dt}.$$  

(16)
where, for every $\alpha_1 \in (0, 1)$, $\alpha_2 > 0$, $\sigma > 0$, $S_{\alpha_1,\alpha_2}(\cdot, \sigma) = e^{-\frac{\sigma^2}{2} S_{\alpha_1}(\cdot, \sigma)}$, and $S_{\alpha_1}(\cdot, \sigma)$ is the alpha-stable density whose characteristic function is $\exp(-\sigma|\cdot|(|\alpha_1| e^{\frac{\sigma^2}{2} \cdot \text{sign}(|\cdot|)})$ [16].

**Proposed priors for the hyperparameters** In the following, we suppose that $(\delta_m)_{1 \leq m \leq M}$ is known. We also assume that, for every $m \in \{1, \ldots, M\}$, parameters $\beta_m$ and $\Gamma_m$ are independent and we denote by $p_{\beta_m}$ and $p_{\Gamma_m}$ their respective prior density functions. Since wavelet coefficient images have inverse Wishart prior for $\Omega$, we suppose that $\Omega$ has a $\text{inv-Wishart}(\Gamma, M)$ prior. Hence, the posterior distribution of $\Omega$ reduces to a product of densities of stable distributions [17].

**Posterior distributions** The posterior distributions of the GMEP hyperparameters have complicated form and there is no practical way for designing algorithms to simulate samples from them especially for the scale matrix. Since for every $m \in \{1, \ldots, M\}$, $\beta_m \in (0, 1)$, we propose to exploit the fact that GMEP is a scale mixture of normal distributions. Then, there exists a vector $v_{m,q} = (v_{m,q})_{q=1}^{Q_m}$ of random variables $v_{m,q}$ such that $K_{\beta_m,\delta_m}v_{m,q} \sim h_{\beta_m,\delta_m}$ and, for all $q \in \{1, \ldots, Q_m\}$, $x_{m,q}$ is drawn independently from a zero mean Gaussian distribution with covariance matrix $\Sigma_{m,q}^{-1}$. Hence, the posterior distributions of $\Gamma_m$ reduces to an inverse Wishart distribution and the posterior distribution of $\beta_m$ is related to a product of densities of stable distributions [15].

Let $\Omega = H^T \Lambda^{-1} H + \sum_{m=1}^{M} \sum_{q=1}^{Q_m} v_{m,q}^2 P_{m,q}^{-1} \Gamma_m^{-1} P_{m,q}$ then the posterior distribution of $x$ reduces to a normal distribution with mean $\mu = \Omega^{-1} H^T \Lambda^{-1} z$ and covariance matrix $\Sigma^{-1}$. Note that sampling from high-dimensional Gaussian distributions is often very difficult since matrix factorization (Cholesky, QR, square root) is not always possible because of its high computation cost and/or memory requirements. Some solutions have been proposed for some special structures of the covariance matrix (circular, sparse,...) [18, 19]. In this work, we propose to use a step of the 3MH algorithm with the diagonal curvature matrix. The interest of this alternative solution is that it does not require any assumption on the structure of the covariance matrix.

Note that we follow the method proposed in [15] for the sampling of $v_{m,q}$. Moreover, MH steps are used to generate samples from the posterior distribution $\beta_m$.

**Results** The test image is a remote sensing multispectral SPOT image of size $128 \times 128$ with three components ($B = 3$) and, corresponding to a scene depicting the city of Tunis. This image is artificially blurred with a cosine blur FTM and corrupted with a zero-mean white Gaussian noise with a variance adjusted so as to correspond to an initial averaged Blurred Signal-to-Noise Ratio (BSNR) equal to 21.64 dB. We apply a 2-resolution wavelet orthonormal decomposition using a Symlet wavelet transform of order 8. We run the hybrid Gibbs sampler for 8,000 iterations, reject the 6,000 first ones as a burn-in, and take the last 2,000 results as samples for the target data. Fig. 1 provides the evolution of improved SNR (ISNR) with respect to the computational time using different algorithms to generate $\pi_X$, when performing tests on an Intel Core i7 CPU, @ 3.00 GHz and using a Matlab 7.12 implementation. It can be observed that the 3MH algorithm reaches stability faster than MALA. In fact, MALA algorithm requires less time per iteration but our algorithm converges in a significantly smaller number of iterations. The obtained samples of the wavelet coefficients are then used to compute the empirical MMSE estimator for the original image. We have used the hyperparameters estimated by the Gibbs sampler to run the method described in [14] which computes the MAP estimate with a GMEP prior using the MM Memory Gradient algorithm. Table 1 reports the results obtained for the different components in terms of SNR, BSNR and Structural Similarity Index (SSIM). It can be observed that the MMSE estimator shows better performance than the MAP.

This can also be observed on Fig. 2 showing the visual improvement for the first component of the image.

![Fig. 1. Convergence speed of 3MH, MALA and MH.](image-url)
Fig. 2. (a) Original image, (b) degraded version of the first component (SNR = 21.72 dB, SSIM = 0.729), (c) restored version using the MAP estimator (SNR = 24.95 dB, SSIM = 0.860), (d) restored version using the MMSE estimator (SNR = 25.20 dB, SSIM = 0.872).

5. CONCLUSION

In this paper, we have proposed a new MCMC algorithm that can be considered as a scaled MALA where the scale matrix is adapted at each iteration with a MM strategy. We have applied this algorithm to compute the MMSE estimator of a multicomponent image from its blurred version. Experimental results indicate the good performance of this new MCMC method. Note that the proposed approach can be applied to general models where the posterior distribution is non-Gaussian.

REFERENCES


