Fast Scalable Image Restoration using Total Variation Priors and Expectation Propagation

Dan Yao, Stephen McLaughlin Fellow, IEEE, and Yoann Altmann, Member, IEEE

Abstract

This paper presents a scalable approximate Bayesian method for image restoration using total variation (TV) priors. In contrast to most optimization methods based on maximum a posteriori estimation, we use the expectation propagation (EP) framework to approximate minimum mean squared error (MMSE) estimators and marginal (pixel-wise) variances, without resorting to Monte Carlo sampling. For the classical anisotropic TV-based prior, we also propose an iterative scheme to automatically adjust the regularization parameter via expectation-maximization (EM). Using Gaussian approximating densities with diagonal covariance matrices, the resulting method allows highly parallelizable steps and can scale to large images for denoising, deconvolution and compressive sensing (CS) problems. The simulation results illustrate that such EP methods can provide a posteriori estimates on par with those obtained via sampling methods but at a fraction of the computational cost. Moreover, EP does not exhibit strong underestimation of posterior variances, in contrast to variational Bayes alternatives.

Index Terms

Variational inference, image restoration, Expectation Propagation (EP), Expectation Maximization (EM), hyperparameter estimation.

I. INTRODUCTION

IMAGE restoration, the recovery of an unknown true image from its degraded measurement, is a fundamental problem in modern image processing [1], [2]. It has found numerous applications in remote sensing [3], medical imaging [4], astronomical imaging [5], defense and security [6], to name a few. To cope with the ill-posed nature of image restoration, a large number of image processing algorithms have been proposed using the Bayesian formalism or penalty-based formulations [7]–[9].

Bayesian image restoration often relies on different prior models for the unknown image of interest. Among the broad variety of existing image prior models, this work focuses on priors promoting small image gradients, centered around the classical Total Variation (TV) prior introduced in [10]. Although more advanced priors, such as priors built using convolution neural network models, can provide better image estimates, the TV prior model still offers practical advantages. Being a local (Markovian) model, the prior can be evaluated efficiently, i.e., without large matrix multiplication as when using wavelet/dictionary-based priors. When defined using the $\ell_1$ or $\ell_2$ norm of image gradients, the TV prior is log-concave. It also requires only a reduced number of hyperparameters to be tuned and does not require to be trained using external images. This makes TV-based image restoration methods appealing for applications where fast and flexible restoration methods and reliable estimates are preferred over high-quality image estimates.

While most image restoration methods aim at maximizing a penalised likelihood or a posterior distribution, they often only provide point estimates and limited tools to assess the uncertainties associated with the estimated images [14]–[16]. The classical approach to uncertainty quantification (UQ) a posteriori when exact computation is not possible remains Monte Carlo sampling. For high-dimensional images, efficient Markov chain Monte Carlo (MCMC) methods have been proposed over the last few years [17]–[20], in particular for log-concave but non-smooth posteriors [19], [20], allowing shorter chains to be used, with lower per iteration cost, e.g., using variable splitting [21]. However, there is still a widely held perception that scaling MCMC to modern high-dimensional problems is not (yet) feasible for fast inference [22]. The most popular alternative to sampling is variational approaches [23], [24], which aim to approximate the posterior distribution of interest by a more tractable distribution whose moments are easier to compute. Variational Bayes (VB) methods [25] are a classical family of tools used in such cases, yet they can be difficult to implement when likelihoods and priors are not conjugate. While VB methods can efficiently approximate (marginal) posterior means, they tend to underestimate marginal posterior variances and these estimated quantities should be handled carefully within any subsequent decision-making process.

Expectation propagation (EP) [26], [27] is another variational alternative to sampling, which has become a popular approximation method for inference in large scale statistical models. This family of algorithms provides efficient solutions to perform approximate Bayesian inference, and has recently been applied to solve high-dimensional imaging problems [28]–[31]. One of the key differences between EP and VB is the form of the divergences to be minimized, which leads to different approximate...
probability distributions. By choosing approximations within the exponential family [32], EP can be applied to wider sets of problems compared to VB. In contrast to VB, EP tends to overestimate marginal variances and can thus be used more reliably for uncertainty upper bounds in decision-making [33, 34].

EP has recently been used in combination with gradient-based priors for tomographic image reconstruction [31]. However, the resulting EP algorithm has several limitations. Firstly, for an $N$ dimensional image vector, it requires $O(N^4)$ arithmetic operations per iteration to update the variance of the approximating factor and it is limited to a sequential update scheme, which prevents scalable inference for high-dimensional imaging problems. In this work, a set of fast and scalable EP algorithms, which only capture marginal (pixel-wise) variances a posteriori, are proposed to solve high-dimensional image restoration problems. It allows the use of a reduced number of sequential EP updates, where furthermore most steps can be implemented in a parallel fashion. This paper first considers the $\ell_1$-TV prior, which allows the comparison with existing MCMC and VB methods. Furthermore, as in [31], we also show that our method can be used with more aggressive (non-convex) gradient-based priors, e.g. spike-and-slab prior, which tends to perform better than $\ell_1$-TV prior in homogeneous regions denoising while preserving sharp edges. Finally, for $\ell_1$-TV prior we illustrate how the proposed EP method can be embedded within larger inference schemes. More precisely, we introduce auxiliary variables, without increasing significantly the computational footprint of EP, to allow for the hyperparameter estimation via an Expectation Maximization (EM)-like procedure.

The main contributions of this paper are twofold:

- A set of new scalable EP algorithms with convex and non-convex gradient-based priors are proposed to solve high-dimensional image restoration problems, including denoising, non-blind deconvolution, and compressive sensing (CS) reconstruction. These algorithms benefit from closed-form expressions for most updates.
- We illustrate how the EP method can be used efficiently within larger inference schemes, e.g., to estimate the hyperparameter of $\ell_1$-TV prior, despite only capturing marginal variances a posteriori.

The reminder of this paper is organized as follows. Section II presents an exact Bayesian model with three gradient-based priors considered for image restoration. Section III proposes the EP algorithms to perform approximate Bayesian inference. Hyperparameter estimation is discussed in Section IV. Section V evaluates the performance of the proposed EP algorithms on high-dimensional image restoration problems. Conclusions and further work are finally reported in Section VI.

II. BAYESIAN MODEL FOR IMAGE RESTORATION

The image restoration problem investigated in this work consists of recovering, from a set of observations $y \in \mathbb{R}^M$, an unknown image $x \in \mathbb{R}^N$ which has been linearly transformed by a known degradation operator $H \in \mathbb{R}^{M \times N}$ ($M \leq N$) and corrupted by additive noise. More precisely, using the Bayesian formalism, we aim at providing a point estimate for $x$ (e.g., posterior mean) and pixel-wise a posteriori uncertainty measures.

A. Likelihood

The observations in $y = [y_1, \ldots, y_M]^T$ are assumed to be corrupted by independently and identically distributed (i.i.d.) zero-mean Gaussian noise with variance $\xi$ and the mean of $y$ is $Hx$. The resulting likelihood $f_y(y|Hx)$ can be expressed as

$$f_y(y|Hx) = \prod_{m=1}^{M} \mathcal{N}(y_m; h_m x, \xi),$$

where $\{h_m\}_{m=1,\ldots,M}$ are the rows of $H$. Different structures of matrix $H$ are considered in this work: (i) $H = I_N$ is the identity matrix (denoising problem), (ii) $H \in \mathbb{R}^{N \times N}$ is a convolution matrix (image deconvolution problem), (iii) $H \in \mathbb{R}^{M \times N}$ ($M \ll N$) is a sensing matrix in CS, such as a random matrix with Gaussian i.i.d. entries or a subsampled Paley-ordered 2D Hadamard matrix [35].

B. Image gradient-based priors

In this work, we consider, as in [31], a set of prior distributions based on the expected properties of the discrete gradients of natural images. The gradients of a 2D gray scale image $x$ are defined as the difference between adjacent pixels along the horizontal and vertical directions. To keep the inference process scalable, we consider priors which can be expressed, using the Hammersley-Clifford theorem [36], as

$$f_x(x|\theta) \propto \prod_{(i,j) \in \mathcal{V}} \phi(x_i - x_j; \theta),$$

where $\mathcal{V}$ denotes the set of pairs of pixels that are direct neighbours (with a 4-neighbourhood structure) and $\phi(\cdot; \theta)$ is a positive function parameterised by $\theta$. This factorization allows each factor in (2) to only depend on two pixels, which in turn enables efficient EP updates. Moreover, the set of cliques in $\mathcal{V}$ can be partitioned into a four (disjoined) sets of cliques, denoted by
In Fig. 1, the prior in (2) is factorized as

\[
f_x(x|\theta) = \frac{1}{C(\theta)} \prod_{k=1}^{4} \phi_{\nu_k}(x|\theta),
\]

where \( \phi_{\nu_k}(x|\theta) \propto \prod_{(i,j) \in \nu_k} \phi(x_i-x_j; \theta). \)

We consider three parametric functions for \( \phi(\cdot; \theta) \), leading to three gradient-based priors.

\( \ell_1 \)-TV prior: the first type of \( \phi(\cdot; \theta) \) is given by

\[
\phi(x_i-x_j; \theta) = \exp(-\lambda|x_i-x_j|),
\]

where \( \theta := \lambda \geq 0 \) is scalar hyperparameter, resulting in \( f_x(x|\theta) \) in (3) being the classical anisotropic (\( \ell_1 \)-norm) TV prior. A convenient property of this gradient-based prior is that it is log-concave (and thus unimodal), which, when combined with the likelihood (1), makes the posterior distribution of \( x \) log-concave. Note that (2) cannot be used directly to model the isotropic TV-based prior.

In natural images, the gradients can present different (local) distributions in homogeneous regions and at the boundaries of objects. It can be difficult using \( \ell_1 \)-TV prior to recover simultaneously sharp boundaries and large, textured regions. For such tasks, it is preferable to use more aggressive and flexible distributions. In the following we consider Gaussian mixtures, which leads to closed-form EP updates and good performance in practice.

MoG2-TV prior: the second type of \( \phi(\cdot; \theta) \) is constructed using a mixture of two Gaussian (MoG2) profiles, i.e.,

\[
\phi(x_i-x_j; \theta) = \omega \mathcal{N}(x_i-x_j; 0, s_1^2) + (1-\omega) \mathcal{N}(x_i-x_j; 0, s_2^2),
\]

where \( \theta := (\omega, s_1^2, s_2^2) \) includes three hyperparameters \( \omega, s_1, \) and \( s_2. \) Without loss of generality, we assume \( s_1 > s_2 > 0. \) The first Gaussian term encodes the distribution of the large image gradients expected at object boundaries, while the second term represents the distribution of image gradients within homogeneous regions (where image gradients are expected to be smaller). The parameter \( \omega \in (0, 1) \) controls the prior fraction of small/large image gradients. The mixture of more than two Gaussian distributions could also be used, however it would introduce additional hyperparameters whose setting would remain challenging.

BG-TV prior: the last type of \( \phi(\cdot; \theta) \) considered is obtained by letting \( s_2^2 \) above tend to 0. In that case, \( \phi(\cdot; \theta) \) reduces to a Bernoulli-Gaussian (BG) mixture, whereby image gradients are a priori either exactly zero, or Gaussian distributed, i.e.,

\[
\phi(x_i-x_j; \theta) = \omega \mathcal{N}(x_i-x_j; 0, s^2) + (1-\omega) \delta(x_i-x_j),
\]

where \( \theta := (\omega, s^2). \) Although this prior seems very informative/restrictive, it only depends on two hyperparameters, which can be easier to tune than those involved in (5) and can be appropriate when the scene of interest presents piece-wise constant intensity profiles [31].

C. Exact posterior distribution

Irrespective of the form of \( \phi(\cdot; \theta) \) in (4) - (6), using the factorization in (3) and the Bayes rule, the posterior distribution of \( x \) conditioned on \( \theta \) is given by

\[
f(x|y, \theta) = \frac{f_y(y|Hz)f_x(x|\theta)}{\int f_y(y|Hz)f_x(x|\theta)dx},
\]
Bayesian inference based on \(f(x|y, \theta)\) is usually intractable as the denominator in (2) is typically intractable. Although sampling from the posterior is possible, in particular for the \(\ell_1\)-TV prior where the posterior is log-concave \([19], [37]\), it remains challenging in high-dimensional settings, in particular with the multimodal priors induced by \([5]\) and \([6]\).

### D. Extended Bayesian model and EP posterior approximation

Although the proposed EP method can be applied directly to approximate the posterior distribution in (7), we introduce an extended model including a set of auxiliary variables gathered in a set denoted by \(u = \{u_i\}_{i=1}^{2N}\). \(u\) contains all the vertical and horizontal gradients of \(x\), i.e., \(u = \{x_i - x_j\}_{(i,j) \in V}\). Using the partitioning described in Section II-B, \(u\) can be partitioned as \(u = \{u_1, u_2, u_3, u_4\}\), such that \(u_k = \{x_i - x_j\}_{(i,j) \in V_k}, \forall k\). One of the key ingredients of the proposed EP algorithms is the one-to-one mapping between \(u_k\) and the pixel pairs in \(V_k\), i.e., \(u_n = x_i - x_j\). The introduction of \(u\) (i) does not change the approximation of the posterior of \(x\), (ii) generalizes the description of the proposed EP methods as the same unifying equations and updates can be used irrespective of \(\phi(\cdot; \theta)\), and (iii) allows to efficiently estimate \(\lambda\) (if unknown) when using the \(\ell_1\)-TV prior (see Section IV).

In the extended Bayesian model, the prior of \(u\) conditioned on \(x\), is defined as \(f_u(u|x) = \prod_{k=1}^{4} f_u(u_k|x)\), where
\[
f_u(u_k|x) = \delta(u_k - D_k x), \quad \forall k \in \{1, 2, 3, 4\}. \tag{8}
\]
The \(N/2 \times N\) matrix \(D_k\) allows the computation of the gradients of \(x\) associated with the edges in \(V_k\), and \(\delta(\cdot)\) denotes a product of Dirac delta functions, applied element-wise to its (multivariate) input. Eq. (8) ensures that the elements of \(u_k\) correspond to the gradients of \(x\) in \(V_k\). Combining \(f_x(x|\theta)\) in (5) and \(f_u(u|x)\), the resulting extended posterior distribution becomes
\[
f(x, u|y, \theta) \propto f_y(y|Hx) \prod_{k=1}^{4} \phi_{V_k}(x|\theta) f_u(u_k|x). \tag{9}
\]
We then approximate the extended posterior in (9) using a Gaussian distribution \(Q(x, u)\)\(^\dagger\) such that
\[
Q(x, u) \approx f(x, u|y, \theta). \tag{10}
\]

To ensure the proposed EP algorithm remains tractable, a classical mean-field approximation framework \([23], [38]\) is used to factorize \(Q(x, u)\) such that
\[
Q(x, u) = Q(x)Q(u), \tag{11}
\]
with \(Q(x) = \mathcal{N}(x; \mu_x, \Sigma_x)\) and \(Q(u) = \mathcal{N}(u; \mu_u, \Sigma_u)\). This approximation basically decouples the approximate distribution of the image \(x\) from that of its gradients (encoded in \(u\)). Note that the marginal distribution obtained by integrating \(f(x, u|y, \theta)\) over \(u\) is the original posterior distribution in (7). Thus, using the marginal posterior approximation \(\int Q(x, u)du\) is equivalent to approximating the original posterior in (7). Moreover, to keep Bayesian inference scheme scalable and numerically stable, the covariance matrices \(\Sigma_x\) and \(\Sigma_u\) are enforced to be diagonal.

### III. PROPOSED EP ALGORITHM WITH TV PRIORS

This section proposes new EP algorithm to find the two Gaussian densities \(Q(x)\) and \(Q(u)\).

#### A. EP factorization

EP leverages the factorization of the exact posterior in (9), which reduces to five factors depicted in Fig. 2. More precisely, each exact factor is associated with an approximating factor such that
\[
\begin{align*}
q_{x,0}(x) &\approx K_0 f_y(y|Hx), \\
q_{x,k}(x|u_k) &\approx K_k \phi_{V_k}(x|\theta) f_u(u_k|x), \quad \forall k \in \{1, \ldots, 4\},
\end{align*}
\]
where \(\{K_k\}_{k=0}^{4}\) are constants and the approximating factors are Gaussian densities whose moments are denoted as follows
\[
\begin{align*}
q_{x,k}(x) &= \mathcal{N}(x; \mu_{x,k}, \Sigma_{x,k}), \quad \forall k \in \{0, \ldots, 4\}, \\
q(u_k) &= \mathcal{N}(u_k; \mu_{u,k}, \Sigma_{u,k}), \quad \forall k \in \{1, \ldots, 4\}.
\end{align*}
\]
The global EP marginal approximations \(Q(x)\) and \(Q(u)\) in (11) can then be obtained by
\[
Q(x) \propto q_{x,0}(x) \prod_{k=1}^{4} q_{x,k}(x), \quad Q(u) = \prod_{k=1}^{4} Q(u_k). \tag{12}
\]
To ensure that \(\Sigma_x\) and \(\Sigma_u\) are diagonal, \(\{\Sigma_{x,k}\}_{k=0}^{4}\) and \(\{\Sigma_{u,k}\}_{k=1}^{4}\) are forced to be diagonal (and positive-definite) during the EP updates, as will be discussed in subsection III-C.

\(^\dagger\)To simplify the notation, \(Q(\cdot)\) is used to represent multivariate Gaussian distributions with respect to (w.r.t.) \(x, (x, u)\), or \(u\) in the remainder of this paper.
B. KL divergence minimization

EP based on Gaussian approximations can be seen as a sequential message passing algorithm which, at each iteration, updates the means and covariance matrices of the five EP approximating factors in Fig. 2. More precisely, each iteration consists of solving sequentially the following KL divergence minimization problems

$$\min_{q_{x,0}(x)} KL(f_y(y|hx)Q(0)(x)||Q(x)),$$

and

$$\min_{q_{x,k}(x|u_k)} KL(\phi_{V_k}(x|\theta)f_u(u_k|\theta)Q^k(x)||Q(x)Q(u_k)),$$

for $k \geq 1$, where $\{Q^k(x)\}_{k=0}^\infty$ denote the so-called cavity distributions obtained by removing a Gaussian approximating factor from the global posterior approximation $Q(x)$, i.e., $Q^k(x) \propto Q(x)/q_{x,k}(x)$ is the ratio of two Gaussian densities (see Appendix 3). The cavity distributions are also Gaussian distributions, with means and covariance matrices defined by $\Sigma^k(x) = N(x; \mu^k, \Sigma^k_x), \forall k \geq 0$. Moreover, note that the covariance matrices $\{\Sigma^k_{x|k}\}_{k}$ are also diagonal by construction.

The first arguments of the KL divergences in (13) are called tilted distributions, and are formed by the product of an exact factor to be approximated and the corresponding cavity distributions. To ease notations, they are denoted by $P_k(x, u), \forall k \geq 0$ (although $P_0(x, u)$ does not depend on $u$). The second arguments of the KL divergences are the global EP approximations or one of its marginals. Since $Q(x)$ and $Q(u)$ are Gaussian densities with diagonal covariance matrices, solving the problems in (13) reduces to matching the marginal moments (means and variances) of $P_k(x, u)$ and $Q(x, u)$ (see [23, Chap. 10]). For instance, to update the factors $q_{x,k}(x)$ and $Q(u_k)$ (for $k \geq 1$), one can compute the marginal moments of $P_k(x, u)$ w.r.t. $x$ and $u_k$. This provides directly the marginal moments of $Q(u_k)$ and the updated moments of $Q(x)$. The updated parameters of $q_{x,k}(x)$ are then obtained using $q_{x,k}(x) \propto Q(x)/Q^k(x)$. Similarly, updating $q_{x,0}(x)$ simply requires the computation of the marginal moments (w.r.t. $x$) of $P_0(x, u)$. The next section details the update of the different approximating factors, including the computation of the marginal moments of the different tilted distributions.

C. Updating the approximating factors

Update of $q_{x,0}(x)$: $q_{x,0}(x)$ is the minimizer of (13a), which is used to approximate $f_y(y|hx)$. Although the tilted distribution $P_0(x) = f_y(y|hx)Q(0)(x)$ is a multivariate Gaussian distribution whose mean and covariance matrix can be obtained in closed-form as follows

$$\begin{align*}
\text{Cov}_{P_0}(x) &= \left[\frac{1}{\xi}H^TH + (\Sigma^0_x)^{-1}\right]^{-1}, \\
\text{E}_{P_0}[x] &= \text{Cov}_{P}(x)\left[\frac{1}{\xi}H^Ty + (\Sigma^0_x)^{-1}\mu^0_x\right],
\end{align*}$$

(14)

the computation of $\text{Cov}_{P_0}(x)$ can be computationally expensive (due to the matrix inversion) depending on the structure of $H^TH$. This costly matrix inversion with complexity $O(N^3)$ over multiple iterations remains the computational bottleneck of EP algorithms in general [39]. Computing $\text{E}_{P_0}[x]$ in the second line of (14) is less challenging as it can be achieved efficiently using conjugate gradient methods [40], given that left-multiplying by $\text{Cov}_{P_0}(x)$ is often simple. Since $Q(x)$ has a diagonal covariance matrix, only the diagonal elements of $\text{Cov}_{P_0}(x)$ are actually needed. These marginal variances are gathered in a diagonal matrix denoted $\text{Var}_{P_0}(x)$, such that diag($\text{Var}_{P_0}(x)$) = diag($\text{Cov}_{P_0}(x)$). Given that $\Sigma^0_x$ is diagonal, if $H^TH$ is diagonal, computing $\text{Cov}_{P_0}(x)$ is trivial (inversion of a diagonal matrix) and if $H^TH$ is low-rank, the Woodbury matrix identity can be used to compute $\text{Cov}_{P_0}(x)$. For more general $H$, diag($\text{Var}_{P_0}(x)$) can be approximated via Monte Carlo sampling, e.g., Rao-Blackwellized Monte Carlo (RBMC) method proposed in [47]. Note that if the noise is not i.i.d., its covariance matrix can be easily integrated in (14), which makes the denoising problem with non i.i.d. noise trivial using our EP methods.
Once \( \text{Var}_{P_0}(x) \) and \( \mathbb{E}_{P_0}[x] \) are computed, the parameters of \( q_{x,0}(x) \) can be updated via

\[
\Sigma_{x,0} = \left[ \left( \text{Var}_{P_0}(x) \right)^{-1} - \left( \Sigma_{x}^0 \right)^{-1} \right]^{-1},
\]

\[
\mu_{x,0} = \Sigma_{x,0} \left[ \left( \Sigma_{x,0}^{-1} + \left( \Sigma_{x}^0 \right)^{-1} \right) \mathbb{E}_{P_0}[x] - \left( \Sigma_{x}^0 \right)^{-1} \mu_{x}^0 \right].
\]

(15)

**Update of** \( q_{x,k}(x) \) **and** \( Q(u_k) \): Update of these approximating factors requires the computation of the marginal moments of tilted distribution \( P_k(x, u_k) = \phi_{V_k}(x; \theta) f_u(u_k | \theta) Q^{k}(x) \) in (13b), \( \forall k \geq 1 \). It can be shown that marginalizing over \( u_k \) leads to

\[
P_k(x) = \int P_k(x, u_k) du_k = \prod_{(i,j) \in V_k} P_k(x_i, x_j)
\]

\[
= \prod_{(i,j) \in V_k} \phi(x_i - x_j; \theta) Q_x^{k}(x_i) Q_x^{k}(x_j),
\]

(16)

where \( Q_x^{k}(x_i) \) and \( Q_x^{k}(x_j) \), whose mean and variance are \( (m_i, c_i) \) and \( (m_j, c_j) \), denote the marginals of \( x_i \) and \( x_j \) associated with \( Q^{k}(x) \). Moreover, we define \( Q_x^{k}(x_i, x_j) = Q_x^{k}(x_i) Q_x^{k}(x_j) \), where \( x_i \) and \( x_j \) are pixel pairs appearing in a single pair/edge in \( V_k \). Thus, \( P_k(x) \) is the product of \( N/2 \) two-dimensional unnormalized densities \( P_k(x_i, x_j) \) associated with the \( N/2 \) pixel pairs in \( V_k \).

Similarly, it can be shown that marginalizing over \( x \) yields

\[
P_k(u_k) = \int P_k(x, u_k) dx = \prod_{n \in V_k} P_k(u_n).
\]

(17)

The update of \( Q(u_k) \) requires computing the moments of \( P_k(u_k) \), where

\[
P_k(u_n) \propto \mathcal{N}(u_n; m_i - m_j, c_i + c_j) \phi(u_n; \theta).
\]

(18)

For the three priors considered, \( P_k(u_n) \) reduces to simple mixtures of two distributions involving (truncated) Gaussian distributions. Thus, its mean \( \bar{m}_n \) and variance \( \bar{S}_n \) can be computed easily. Prior-dependent expressions are omitted here for brevity. If more exotic functions were used for \( \phi(\cdot; \theta) \), the moments of \( P_k(u_n) \) could still be computed, e.g., via numerical integration.

As for the update of \( q_{x,k}(x) \), it requires computing the moments of \( P_k(x) \) in (16). Using the projection scheme in (42), the marginal moments of \( P_k(x_i, x_j) \) can be obtained via

\[
\mathbb{E}_{P_k}[(x_i, x_j)] = (m_i, m_j) + \frac{\bar{m}_n - (m_i - m_j)}{c_i + c_j} (c_i, c_j),
\]

\[
\text{Var}_{P_k}(x_i, x_j) = (c_i, c_j) + \frac{\bar{S}_n - (c_i + c_j)}{(c_i + c_j)^2} (c_i^2, c_j^2).
\]

It can be seen from (17) that the moments of the \( N/2 \) marginals w.r.t. \( u_k \) can all be computed independently (i.e., in parallel), after which all the edges/pairs \( x_i, x_j \) in (16) can be processed independently as well. Once the marginal moments of \( \text{Var}_{P_k}(x) \) and \( \mathbb{E}_{P_k}[x] \) for \( x_i, x_j \in V_k \) are computed, the parameters of \( q_{x,k}(x) \) can be obtained via

\[
\Sigma_{x,k} = \left[ \text{Var}_{P_k}(x)^{-1} - \left( \Sigma_{x}^k \right)^{-1} \right]^{-1},
\]

\[
\mu_{x,k} = \Sigma_{x,k} \left[ \left( \Sigma_{x,k}^{-1} + \left( \Sigma_{x}^k \right)^{-1} \right) \mathbb{E}_{P_k}[x] - \left( \Sigma_{x}^k \right)^{-1} \mu_{x,k}^k \right].
\]

(19)

Therefore, the update of \( Q(u_k) \) and \( q_{x,k}(x) \) is very efficient in practice.

As mentioned earlier, EP iterates by updating sequentially \( q_{x,0}(x) \) and \( \{q_{x,k}(x), Q(u_k)\}_{k=1}^{4} \) until convergence. The final EP approximation \( Q(x) \) are obtained by

\[
\Sigma_x = (\Sigma_{x}^0 + \sum_{k=1}^{4} \Sigma_{x,k})^{-1},
\]

\[
\mu_x = \Sigma_x (\Sigma_{x,0}^{-1} \mu_{x,0} + \sum_{k=1}^{4} \Sigma_{x,k}^{-1} \mu_{x,k}).
\]

(20)

The first lines of (15) and (19) do not ensure that the updated covariance matrices \( \Sigma_{x,0}, \Sigma_{x,k} \) are strictly positive definite. If negative variances are obtained, these values are usually replaced by large positive values (e.g., \( 10^8 \)) before computing \( \mu_{x,0} \) and \( \mu_{x,k} \). Algorithm 1 summarizes the pseudo-code for the proposed algorithm. In general, there is no guarantee for the convergence of EP algorithms. A standard damping strategy is used here to prevent potential oscillations.
between successive iterations. In practice, if \( \theta \) was set reasonably to reflect the expected distribution (scale) of the gradients of \( x \), we did not experienced convergence issues. Such issues may arise if \( \theta \) is set such that the resulting posterior distribution \( f(x | y, \theta) \) is extremely ill-conditioned.

Thanks to the auxiliary variables and the priors considered, all EP updates admit closed-form solutions and enable efficient parallel computation. Moreover, the mean and marginal variance of auxiliary variable \( u \) can be further used to estimate the hyperparameter \( \lambda \) of \( \ell_1 \)-TV prior under an EP-EM framework, as will be discussed in next section.

IV. UNSUPERVISED EP ALGORITHM WITH \( \ell_1 \)-TV PRIOR

In this section, we propose to embed the proposed EP algorithm using the \( \ell_1 \)-TV prior within a larger inference problem where the hyperparameter \( \lambda \) is unknown and needs to be adjusted for the image of interest. We consider only this prior as it has been shown in \[44\] that the normalizing constant \( C(\lambda) \) in this case has a closed-form expression, leading to

\[
 f(x | \lambda) = \frac{1}{D\lambda^{-N}} \exp \left[ -\lambda \sum_{(i,j) \in V} |x_i - x_j| \right],
\]  

where \( D \) is a constant independent of \( \lambda \). Although this prior is not proper, the posterior distribution often is in practice. We do not consider MoG2-TV and BG-TV priors in this section as the associated constants \( C(\theta) \) in those cases are intractable.

It has been shown in \[45\] that when the prior is parametrized by a reduced number of hyperparameters (only a scalar-valued \( \lambda \) here), estimating these hyperparameters by maximum marginal likelihood estimation (MMLE) or marginal maximum a posteriori (MMAP) estimation tends to lead to better image estimates than estimating \( x \) and the hyperparameters by joint MMSE estimation. Here, we present a MMLE approach which aims at maximizing

\[
 f(y | \lambda) = \int f_y(y | \lambda) f_x(x | \lambda) dx,
\]  

i.e., the normalizing constant of the exact posterior in \(7\). However, a similar approach could be used to maximize \( f(\lambda | y) \), in particular if \( \lambda \) is assigned a conjugate prior \[44\].

Maximizing \( f(y | \lambda) \) can be achieved using an iterative procedure based on EM, where the standard update rule at iteration \( t \) is

\[
 \lambda^{(t)} = \arg\max_{\lambda} \mathbb{E}_{f(x | y, \lambda^{(t-1)})} \left[ \log f(x | y, \lambda) \right],
\]  

yielding the cost function \( F^{(t)}(\lambda) \) to be maximized as

\[
 F^{(t)}(\lambda) = N \log \lambda - \lambda \mathbb{E}_{f(x | y, \lambda^{(t-1)})} \left[ \sum_{(i,j) \in V} |x_i - x_j| \right].
\]  

By zeroing its derivative w.r.t. \( \lambda \), the estimate of \( \lambda \) at iteration \( t \) is obtained by \( \lambda^{(t)} = N/a_0 \), with

\[
 a_0 = \mathbb{E}_{f(x | y, \lambda^{(t-1)})} \left[ \sum_{(i,j) \in V} |x_i - x_j| \right].
\]

When exact computation of \( a_0 \) using expectation w.r.t. the exact posterior \( f(x | y, \lambda^{(t-1)}) \) is intractable, classical options turn to stochastic approximations (such as stochastic EM variants) or variational approximation (such as variational EM) \[46\].
Algorithm 2: Proposed EP-EM algorithm—unknown $\lambda$

**Input**: observation $y$, degradation matrix $H$, $\xi$, $\lambda^{(0)}$

**Output**: $Q(x) \propto N(x; \mu_x, \Sigma_x), \lambda$

1. for $t=1$ to StopRule do
   2. Obtain $Q^{(t)}(x|y, \lambda^{(t-1)})$ and $\{P_k^{(t)}(u_k|\lambda^{(t-1)})\}_{k=1}^4$ using Algorithm 1
   3. Compute $\bar{a}_u = \sum_{k=1}^4 \mathbb{E}_{P_k^{(t)}(u_k|\lambda^{(t-1)})}[\|u_k\|_1]$
   4. Set $\lambda^{(t)} = N/\bar{a}_u$

end for

---

**Fig. 3:** Impact of the strategy used to approximate $a_0$ for a denoising problem (256×256-pixel *Cameraman* image). The MSE (see definition in Section V) is presented as a function of $\lambda$ and $\xi$. $\lambda_1^*(\text{Oracle})$ is the oracle value of $\lambda$ that minimizes the MSE and the other markers represent the final EP-EM estimates for Options 1-3.

The latter option is adopted here, and the use of the EP posterior approximation leads to an EP-PM algorithm. Three options to approximate $a_0$ in (25) are considered by replacing directly the expectation w.r.t. $f(x|y, \lambda^{(t-1)})$ to another approximate distribution as follows:

- **Option 1**: $a_x = \mathbb{E}_{Q(x|\lambda^{(t-1)})} \left[ \sum_{(i,j) \in V} |x_i - x_j| \right]$, (26)
- **Option 2**: $a_u = \mathbb{E}_{Q(u|\lambda^{(t-1)})} \left[ \|u\|_1 \right]$, (27)
- **Option 3**: $\bar{a}_u = \sum_{k=1}^4 \mathbb{E}_{P_k(u_k|\lambda^{(t-1)})}[\|u_k\|_1]$. (28)

All the expectations in (26)–(28) can be computed analytically. In practice, (28) performs better for hyperparameter estimation and in turn for unsupervised image restoration than (26) and (27), as illustrated via the image denoising examples in Fig. 3. This is due to $Q(x|\lambda^{(t-1)})$ in (26), which has a diagonal covariance matrix that fails to capture the correlation between adjacent pixels. Using $\sum_{(i,j) \in V} |x_i - x_j| = \|u\|_1$, (27) is a better alternative since $Q(u|\lambda^{(t-1)})$ better captures the posterior variances of the image gradients. Eq. (28) computes expectations w.r.t. tilted distributions, which are expected to be closer to the exact marginal distributions of the gradients than their Gaussian approximations. Thus (28) is adopted in our EP-EM algorithm whose pseudo-code is presented in Algorithm 2. The algorithm can be stopped after a fixed number of iterations or when $\lambda$ stabilises (within 10 to 20 iterations in our experiments, depending on the problem considered). In practice, we observed that a faster version of EP-EM can provide similar results at a lower cost. Instead of running Algorithm 1 until convergence (line 2 of Algorithm 2), it can be run only for a few iterations (in Section V we used a single iteration). This generally allows us to reduce the overall number of EP iterations and seems more robust if the scale of $\lambda^{(0)}$ is set inappropriately.

V. Experimental Results

This section evaluates the performance of the proposed algorithms for image denoising, non-blind deconvolution, and CS reconstruction. The EP posterior mean $\mu_x$ is used as image estimate $\hat{x}$, and the marginal variances in $\text{diag}(\Sigma_x)$ quantify the pixel-wise posterior uncertainty. The quality of the restored images is assessed using the mean squared error (MSE) and peak signal-to-noise ratio (PSNR) computed between any estimated image $\hat{x}$ and the corresponding ground truth $x$, i.e.,

$$\text{MSE} = \frac{1}{N}\|x - \hat{x}\|^2_2, \quad \text{PSNR} = 10 \times \log_{10}\left(\frac{\max_x^2}{\text{MSE}}\right).$$
We will first consider the three gradient-based priors discussed in Section III for image denoising to show the effectiveness of the priors. In this context, the hyperparameters $\theta$ are set using grid search to minimize the MSE, i.e.,

$$\theta_{\text{argmin}} = \arg\min_{\theta} \left\{ \| x - \hat{x}(y, \theta) \|_2^2 \right\},$$

where $\hat{x}(y, \theta)$ is the EP-based image estimate. For completeness, we will also consider (in Fig. 5) a second oracle estimator of $\theta$, where $\hat{x}(y, \theta)$ will be the MAP estimator of $x$ based on the exact posterior distribution.

The damping parameter of EP is set to be 0.9 and it has been observed in our experiments that EP converges within 20 iterations. For EP-EM, unless otherwise stated, the results presented are obtained with the fast alternative with 20 EM iterations and a single EP loop per EP-EM iteration. The pixel-wise uncertainty is then represented as a 2D image/map. The chains are used for EB as it is simply used for hyperparameter estimation. While exact MAP estimation using the multimodal distribution is known. Our EP-EM algorithm is also compared to the empirical Bayes (EB) method from [45]. The SK-ROCK and EB results are considered as reference and for these MCMC algorithms, the burn-in period and chain length (after burn-in) are set to (2000; 10^5) samples and (300; 1500) samples, respectively. Short chains are used for EB as it is simply used for hyperparameter estimation. While exact MAP estimation using the multimodal priors is not easily achievable, for completeness we will use SALSA [48] for MAP estimation with the $\ell_1$-TV prior.

### A. Image denoising

We first discuss the use of the three priors within EP (Algorithm 1) for image denoising problems, i.e., with $H = I_N$. We also include EP-EM (Algorithm 2) to assess its hyperparameter estimation performance. Experiments are conducted using three widely used grayscale images scaled in $[0, 255]$ ($I_1$: Cameraman $256 \times 256$ pixels, $I_2$: Flinstones $512 \times 512$ pixels and $I_3$: Mandrill $512 \times 512$ pixels).

The PSNRs obtained after denoising with hyperparameters set to be $\theta_{\text{(Oracle)}}$ are reported in 3rd-5th columns of Table I for three noise variances. Note that for the three prior models and images, $\theta_{\text{(Oracle)}}$ changes depending on the noise variance, which highlights the need for automatic hyperparameter setting. The two non-convex priors generally present comparable performance (MoG2-TV is marginally better in the high PSNR regime) and provide better PSNRs than the $\ell_1$-TV prior. However, as mentioned before, automatic tuning of the non-convex prior hyperparameters is more challenging than when using the $\ell_1$-TV prior. The PSNRs obtained using EP-EM with the $\ell_1$-TV prior are depicted in the last column of Table I. Although slightly worse compared to $\theta_{\text{(Oracle)}}$, the results in the two last columns are comparable, illustrating the benefit of the proposed EP-EM strategy.

Fig. 4 shows examples of denoising results for the Flinstones image. The denoised images obtained by EP with the MoG2-TV and BG-TV priors present better visual quality and lower uncertainties in homogeneous regions than using the $\ell_1$-TV prior. The UQ maps obtained with the non-convex priors also present higher contrast between homogeneous regions and object boundaries, mainly because these priors penalise more strongly intermediate gradients. In order to verify the accuracy of UQ maps obtained with the $\ell_1$-TV prior, EP-EM is compared with SK-ROCK and VB for a cropped Flinstones image (only a $128 \times 128$ pixels portion to keep the processing time of SK-ROCK relatively short). For fair comparisons, SK-ROCK and VB are run with the final hyperparameter estimated via EP-EM, and denoted $\lambda_{\text{EP-EM}}$. The images denoised by EP-EM and VB are run with the final hyperparameter estimated via EP-EM, and denoted $\lambda_{\text{EP-EM}}$.

![Fig. 4: Denoising results of Flinstones image ($\xi = 30^2$). 2nd-5th columns: results obtained by EP with MoG2-TV, BG-TV, $\ell_1$-TV priors, and EP-EM with $\ell_1$-TV prior. Top row: denoised images. Bottom row: UQ maps. 6th-8th columns: a cropped results by EP-EM, SK-ROCK, and VB with $\ell_1$-TV prior using $\lambda_{\text{EP-EM}}$.](image-url)
VB methods are easier.

**table 1:** Image denoising: PSNR (dB) after denoising using EP. The highest (resp. second highest) PSNR values in each row are bold (resp. underlined). The values in brackets are the values of the corresponding hyperparameter(s).

| ξ | image | EP MoG2-TV | EP BG-TV | EP ℓ₁-TV | EP-EM ℓ₁-TV |
|---|---|---|---|---|---|
| 10² | I₁ | 33.09 | 32.81 | 31.05 | 30.77 |
| | | (0.20, 11, 3400) | (0.85, 2800) | (0.0320) | (0.0377) |
| | I₂ | 35.58 | 31.39 | 30.62 | 30.43 |
| | | (0.25, 11, 1700) | (0.80, 1300) | (0.0360) | (0.0278) |
| | I₃ | 29.07 | 29.02 | 28.90 | 28.85 |
| | | (0.25, 71, 3300) | (0.85, 2700) | (0.0310) | (0.0263) |
| 30² | I₁ | 27.09 | 26.94 | 25.55 | 25.27 |
| | | (0.25, 11, 4000) | (0.75, 3100) | (0.0330) | (0.0254) |
| | I₂ | 25.64 | 25.44 | 24.57 | 24.11 |
| | | (0.25, 11, 4000) | (0.75, 4000) | (0.0270) | (0.0195) |
| | I₃ | 22.91 | 22.93 | 22.66 | 22.66 |
| | | (0.20, 1, 4000) | (0.80, 4700) | (0.0240) | (0.0233) |
| 50² | I₁ | 24.24 | 24.62 | 23.05 | 22.87 |
| | | (0.25, 1, 3400) | (0.75, 8000) | (0.0260) | (0.0213) |
| | I₂ | 22.35 | 22.75 | 21.94 | 21.65 |
| | | (0.25, 11, 4000) | (0.75, 8200) | (0.0210) | (0.0162) |
| | I₃ | 20.86 | 20.88 | 20.77 | 20.75 |
| | | (0.20, 1, 4000) | (0.80, 4900) | (0.0230) | (0.0213) |

VB are visually similar and present a slightly higher PSNR than the approximate MMSE estimate obtained via SK-ROCK. Although SK-ROCK can include a bias such that the image reported is not the true MMSE estimate, it is more likely that this difference is mostly due to the biases of EP and VB, which are beneficial here and yield better PSNRs. When comparing the UQ maps obtained by EP, VB and SK-ROCK (using SK-ROCK as reference), we observe that EP tends to overestimate the uncertainties in homogeneous regions and extend the high-uncertainty regions at object boundaries. VB tends to underestimate the marginal variances, although they remain close to the SK-ROCK results.

These denoising results illustrate that the MoG2-TV and BG-TV priors can provide better image estimates (when the hyperparameters are correctly set) and we expect similar trends for deconvolution and CS problems. However, hyperparameter tuning by grid search can be computationally intensive and automatic tuning is not straightforward for the two priors. For ℓ₁-tv prior, for which comparisons with existing MCMC and VB methods are easier.

B. Non-blind image deconvolution

This subsection illustrates the performance of EP-EM algorithm with ℓ₁-Tv prior for image deconvolution. It is also used to assess its ability to estimate the regularization parameter. Sub-images (128 × 128, 164 × 164, 165 × 165 pixels) of the three test images I₁–I₃ are used as reference to keep the computational cost of the competing methods relatively low. The matrix H corresponds to a 9 × 9 pixels uniform blurring kernel. The observed images are generated with blurred signal-to-noise-ratio (BSNR) of 15 dB, 25 dB and 35 dB. The alternative hyperparameter estimation methods include EB [45] and the hierarchical Bayes (HB) method from [44], whose original implementations have been modified to include the ℓ₁-Tv prior. We also consider SUGAR [49] and the Morozov’s discrepancy principle (DP) method [50], as in [45].

Fig. 5 first presents image MSEs obtained via MAP estimation (blue curves) and EP (approximate MMSE estimation, red curves), as a function of λ. We did not include the MSE curves associated with ‘exact’ MMSE estimation, which could be approximated via SK-ROCK as it would require running too many Markov chains (the SK-ROCK parameters would also need to be tuned over the range of λ considered). For a given value of λ, the MSE of the MAP estimator is generally lower than that of EP and it is interesting to observe that each of the blue and red curves does not reach its minimum at the same λ. EB and EP-EM aim at maximizing the same marginal likelihood and λEP–EM and λEB are thus expected to be close. This is confirmed in Fig. 5 where they are generally close, although λEP–EM can be smaller than λEB. Moreover, λEP–EM and λEB are also close to the Oracle values that minimize the MSE using MAP or approximate MMSE estimation. The main benefit of EP-EM over EB is the computational cost since EP-EM does not require high-dimensional Monte Carlo sampling. This makes EP-EM particularly attractive for fast hyperparameter setting, e.g., for subsequent use with MAP-based algorithm.

We now investigate the quality of the approximate MMSE estimator and associated posterior marginal variances obtained via EP-EM. Fig. 6 shows the deconvolved images and their UQ maps by EP-EM, SK-ROCK and VB, for BSNR = 35 dB. Here again, SK-ROCK and VB are run with λEP–EM. Compared to the denoising experiment in Fig. 4 the three methods provide closer PSNRs, which seem to indicate smaller biases affecting the approximate MMSE estimators of EP and VB. Using SK-ROCK as reference, the EP bias is smaller than that of VB, whose mean tends to shift toward the mode of the posterior. The benefit of EP over VB in terms of UQ is more significant here, where EP slightly overestimates the marginal variances of SKROCK, while VB drastically underestimates the scale of the marginal variances.
Oracle by MAP estimator
Oracle by approximate MMSE estimator
MAP estimator (SALSA)
approximate MMSE estimator (EP)

BSNR=35 dB BSNR=25 dB BSNR=15 dB

EB
HB
SUGAR
DP
EP-EM
MSE
10^{-3} 10^{-2} 10^{-1}
20
30
40
50

EB
HB
SUGAR
DP
EP-EM
MSE
10^{-3} 10^{-2} 10^{-1}
20
30
40
50

EB
HB
SUGAR
DP
EP-EM
MSE
10^{-3} 10^{-2} 10^{-1}
20
30
40
50

EB
HB
SUGAR
DP
EP-EM
MSE
10^{-3} 10^{-2} 10^{-1}
20
30
40
50

EB
HB
SUGAR
DP
EP-EM
MSE
10^{-3} 10^{-2} 10^{-1}
20
30
40
50

Mandrill
Flinstones
Cameraman

Fig. 5: MSE for different $\lambda$ of $\ell_1$-TV prior estimated by different methods. The values on x-y axis are presented on a log scale.

C. CS reconstruction

We now evaluate the performance of EP-EM for CS reconstruction. Experiments are conducted on synthetic observations generated using $H$, where $H \in \mathbb{R}^{M \times N}$ is a (a) Gaussian i.i.d. matrix with mean zero and variance $1/M$, and (b) 2D Hadamard matrix with $M$ randomly and uniformly selected patterns, $\xi = 10^{-4}$. The reference $x$ is the Shepp-Logan phantom image of size $128 \times 128$ pixels with pixel intensity in $[0,1]$.

Fig. 7 depicts reconstruction results using EP-EM, SK-ROCK and VB using the same $\lambda_{EP EM}$ and $M/N = 0.3$. In both cases, EP-EM provides an estimate $\lambda_{EP EM}$ which effectively regularizes the reconstruction problem. With $\lambda_{EP EM}$, EP and provides PSNRs closer to SK-ROCK than VB. While in (a) the UQ maps of EP and SK-ROCK are almost identical, EP underestimates, on average, the marginal uncertainties in (b), using SK-ROCK as reference. We believe this is due to the long-range pixel dependencies induced by the Hadamard patterns, and this observation will be further discussed in the conclusion of the paper. The marginal variances estimated by VB are significantly larger in both cases.

D. Computational time

All the experiments in this paper were carried out using MATLAB R2018b on an Intel(R) Core(TM) i7-8700K CPU @ 3.70GHz workstation. Table II reports the computational time to obtain some of the results presented in Figs. 4, 6 and 7. The top row shows that the complexity of EP does not change significantly when changing the prior ($\ell_1$-TV slightly more expensive due to the truncated Gaussian distributions). While EP-EM would take approximately $T$ times longer than EP for $T$ EM iterations (with $\ell_1$-TV), its cheaper implementation, using a single update of the approximating factor, took only 5.5 seconds (when EP takes 5.4 seconds). The three bottom rows compare computational time of EP, SK-ROCK and VB, used to estimate posterior means and marginal variances (using the same $\lambda_{EP EM}$). EP is slightly faster than VB because it does not use large matrix multiplications during the approximation of the prior and the two variational methods are significantly faster than the sampling method, which requires many samples for accurate variance estimation. Note that SK-ROCK is less expensive in the deconvolution case mostly because the sub-routine used to compute proximity operators converges more quickly. The fast implementation of EP-EM in these cases has approximately the same cost as EP. If the estimation of $\lambda$ is the main objective, the cost of EP-EM is generally higher than competing methods (unless $H^TH$ is diagonal), as EB only
Observation $\mathbf{y}$

| Method | PSNR: 17.00 dB | PSNR: 17.64 dB | PSNR: 18.48 dB |
|--------|----------------|----------------|----------------|
| EP-EM  | PSNR: 23.95 dB | PSNR: 25.49 dB | PSNR: 25.84 dB |
| SK-ROCK| PSNR: 23.93 dB | PSNR: 25.48 dB | PSNR: 25.48 dB |
| VB     | PSNR: 24.34 dB | PSNR: 25.49 dB | PSNR: 25.48 dB |

PSNR: 17.00 dB
PSNR: 23.95 dB
PSNR: 23.93 dB
PSNR: 24.34 dB
PSNR: 17.64 dB
PSNR: 25.49 dB
PSNR: 25.48 dB
PSNR: 25.84 dB

**Fig. 6:** Comparison of deconvolution results obtained by EP-EM, SK-ROCK and VB, with $\ell_1$-TV prior using $\lambda_{\text{EP-EM}}$. The uncertainty estimates in the UQ maps are presented on a logarithmic scale.

**Fig. 7:** Comparison of CS reconstruction results obtained by EP-EM, SK-ROCK and VB, with $\ell_1$-TV prior using $\lambda_{\text{EP-EM}}$. (a) $\mathbf{H}$ is a Gaussian i.i.d. matrix with mean zero and variance $1/M$. (b) $\mathbf{H}$ is a subsampled 2D Hadamard matrix. The uncertainty estimates in the UQ maps are presented on a logarithmic scale.

requires a few iterations to estimate $\lambda$. However EP-EM remains attractive as it allows joint estimation of the image and $\lambda$, using a single fast algorithm, requiring minimum parameter tuning.

| Method     | PSNR: 36.59 dB | PSNR: 37.85 dB | PSNR: 21.47 dB |
|------------|----------------|----------------|----------------|
| EP-EM      | $\ell_1$-TV (512 × 512) | 4 seconds | 4 seconds | 5.4 seconds |
| SK-ROCK    | EP MoG2-TV (512 × 512) | 4 seconds | 4 seconds | 5.4 seconds |
| VB         | EP BG-TV (512 × 512) | 4 seconds | 4 seconds | 5.4 seconds |

| Method     | PSNR: 20.73 dB | PSNR: 23.47 dB | PSNR: 17.99 dB |
|------------|----------------|----------------|----------------|
| EP-EM      | $\ell_1$-TV (128 × 128) | 0.3 seconds | 14.3 hours | 25 seconds |
| SK-ROCK    | EP $\ell_1$-TV (512 × 512) | 0.3 seconds | 14.3 hours | 25 seconds |
| VB         | VB $\ell_1$-TV (512 × 512) | 0.3 seconds | 14.3 hours | 25 seconds |

**TABLE II:** CPU computational time.

**VI. CONCLUSIONS AND FUTURE WORK**

In this paper, we proposed a series of new EP algorithms with convex or non-convex gradient-based priors for scalable image restoration. We also discussed how EP can be embedded within more complex problems where additional problem parameters are unknown and can be estimated with EM-like procedures. We showed that the results are generally more accurate than using the VB alternative, and close to the MCMC-based alternative (which is taken as reference), at a fraction of the computational cost. In a denoising context, we showed that our fast EP methods can be used for rapid approximate MMSE estimation,
together with marginal variance estimation. The method can be very easily adapted to Gaussian noise models with non-isotropic covariance matrices. Thus, it makes it particularly attractive for use within Plug-and-Play methods, beyond MAP-like denoisers.

The scalability of the proposed algorithms relies strongly on the diagonal structure of the covariance matrix of the global EP approximation. These constraints are well suited for most problems considered in this paper, where the exact posterior covariance matrix is close to diagonal. Indeed, the gradient-based priors do not induce strong long-range dependencies, nor does the matrix $H$ if the noise level is sufficiently high. In such cases, the EP approximation is particularly accurate. However, the method is expected to fail if the posterior distribution exhibits strong correlation structures. In such cases, less restrictive covariance constraints should be considered, but how to keep the resulting variational method scalable would require further investigation.

REFERENCES

[1] M. Banham and A. Katsaggelos, “Digital image restoration,” IEEE Signal Processing Magazine, vol. 14, no. 2, pp. 24–41, 1997.
[2] B. K. Gunturk and X. Li, Image restoration: fundamentals and advances. CRC Press, 2012.
[3] J. M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. Nasrabadi, and J. Chanussot, “Hyperspectral remote sensing data analysis and future challenges,” IEEE Geoscience and remote sensing magazine, vol. 1, no. 2, pp. 6–36, 2013.
[4] S. Webb, A. Long, R. Ott, M. Leach, and M. Flower, “Constrained deconvolution of spelt liver tomograms by direct digital image restoration,” Medical Physics, vol. 12, no. 1, pp. 53–58, 1985.
[5] R. Molina, J. Núñez, F. J. Cortijo, and J. Mateos, “Image restoration in astronomy: a bayesian perspective,” IEEE Signal Processing Magazine, vol. 18, no. 2, pp. 11–29, 2001.
[6] T. Bourlai, A. Ross, and A. K. Jain, “Restoring degraded face images: A case study in matching faxed, printed, and scanned photos,” IEEE Transactions on Information Forensics and Security, vol. 6, no. 2, pp. 371–384, 2011.
[7] K.-H. H. H., “On the regularization of ill-posed problems,” Proc. Int. Symp. Figure and Dynamics of the Earth, Moon, and Planets, Prague, vol. -1, p. 365, 01 1987.
[8] A. N. Tikhonov, “On the solution of ill-posed problems and the method of regularization,” in Doklady Akademii Nauk, vol. 151, no. 3. Russian Academy of Sciences, 1963, pp. 501–504.
[9] J. M. Bioucas-Dias, M. A. Figueiredo, and J. P. Oliveira, “Adaptive total variation image deconvolution: A majorization-minimization approach,” in 2006 14th European Signal Processing Conference. IEEE, 2006, pp. 1–4.
[10] L. I. Rudin, S. Osher, and E. Fatemi, “Nonlinear total variation based noise removal algorithms,” Physica D: nonlinear phenomena, vol. 60, no. 1-4, pp. 259–268, 1992.
[11] D. Ulyanov, A. Vedaldi, and V. Lempitsky, “Deep image prior,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 9446–9454.
[12] G. Mataev, P. Milanfar, and M. Elad, “Deepred: Deep image prior powered by red,” in Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019, pp. 0–0.
[13] G. Vaksman, M. Elad, and P. Milanfar, “Lidia: Lightweight learned image denoising with instance adaptation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 524–525.
[14] C. Robert, The Bayesian choice: from decision-theoretic foundations to computational implementation. Springer Science & Business Media, 2007.
[15] A. Repetti, M. Pereyra, and Y. Wiaux, “Scalable bayesian uncertainty quantification in imaging inverse problems via convex optimization,” SIAM Journal on Imaging Sciences, vol. 12, no. 1, pp. 87–118, 2019.
[16] M. Pereyra, “Maximum-a-posteriori estimation with bayesian confidence regions,” SIAM Journal on Imaging Sciences, vol. 10, no. 1, pp. 285–302, 2017.
[17] A. Pakman and L. Paninski, “Exact hamiltonian monte carlo for truncated multivariate gaussians,” Journal of Computational and Graphical Statistics, vol. 23, no. 2, pp. 518–542, 2014.
[18] S. Bubeck, R. Eldan, and J. Lehec, “Finite-time analysis of projected langevin monte carlo,” in Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1, ser. NIPS’15. Cambridge, MA, USA: MIT Press, 2015, p. 1243–1251.
[19] A. Durmus, E. Moulines, and M. Pereyra, “Efficient bayesian computation by proximal markov chain monte carlo: when langevin meets moreau,” SIAM Journal on Imaging Sciences, vol. 11, no. 1, pp. 473–506, 2018.
[20] J. Lehec, “The langevin monte carlo algorithm in the non-smooth log-concave case,” arXiv preprint arXiv:2101.10695, 2021.
[21] M. Vono, “Asymptotically exact data augmentation–models and Monte Carlo sampling with applications to bayesian inference,” Ph.D. dissertation, Institut National Polytechnique de Toulouse, 2020.
[22] B. Rajaratnam and D. Sparks, “Mcmc-based inference in the era of big data: A fundamental analysis of the convergence complexity of high-dimensional chains,” arXiv preprint arXiv:1508.00947, 2015.
[23] C. M. Bishop, Pattern recognition and machine learning. springer, 2006.
[24] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, “Variational inference: A review for statisticians,” Journal of the American Statistical Association, vol. 112, no. 518, p. 859–877, Apr 2017. [Online]. Available: http://dx.doi.org/10.1080/01621459.2017.1285775.
[25] H. Attias, “A variational Bayesian framework for graphical models,” Advances in neural information processing systems, vol. 12, no. 1-2, pp. 209–215, 2000.
[26] T. P. Minka, “Expectation propagation for approximate bayesian inference,” in Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence, ser. UAI’01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 362–369.
[27] M. J. Wainwright and M. I. Jordan, Graphical models, exponential families, and variational inference. Now Publishers Inc, 2008.
[28] Y.-J. Ko and M. W. Seeger, “Expectation propagation for rectified linear poisson regression,” in Asian Conference on Machine Learning. PMLR, 2016, pp. 253–268.
[29] D. Yao, Y. Altman, S. McLaughlin, and M. E. Davies, “Joint robust linear regression and anomaly detection in poisson noise using expectation-propagation,” in 2020 28th European Signal Processing Conference (EUSIPCO). IEEE, 2021, pp. 2463–2467.
[30] A. Braunstein, A. P. Muntoni, A. Pagnani, and M. Pieropan, “Compressed sensing reconstruction using expectation propagation,” Journal of Physics A: Mathematical and Theoretical, vol. 53, no. 18, p. 184001, 2020.
[31] A. P. Muntoni, R. D. H. Rojas, A. Braunstein, A. Pagnani, and I. P. Castillo, “Nonconvex image reconstruction via expectation propagation,” Physical Review E, vol. 100, no. 3, p. 032134, 2019.
[32] M. Seeger, “Expectation propagation for exponential families,” 2005. [Online]. Available: http://infoscience.epfl.ch/record/161464.
[33] P. Jylänki, J. Vanhatalo, and A. Vehtari, “Robust gaussian process regression with a student-t likelihood,” Journal of Machine Learning Research, vol. 12, no. 11, 2011.
[34] B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and scalable predictive uncertainty estimation using deep ensembles,” p. 6405–6416, 2017.
[35] A. Moshtaghpour, J. M. Bioucas-Dias, and L. Jacques, “Close encounters of the binary kind: Signal reconstruction guarantees for compressive hadamard sampling with haar wavelet basis,” IEEE Transactions on Information Theory, vol. 66, no. 11, pp. 7253–7273, 2020.

[36] P. Clifford, “Markov random fields in statistics,” Disorder in physical systems: A volume in honour of John M. Hammersley, pp. 19–32, 1990.

[37] L. Vargas, M. Pereyra, and K. C. Zygalakis, “Accelerating proximal markov chain monte carlo by using an explicit stabilised method,” SIAM Journal on Imaging Sciences, vol. 13, no. 2, pp. 905–935, 2019.

[38] G. Parisi, Statistical field theory. Addison-Wesley, 1988.

[39] J. M. Hernández-Lobato, D. Hernández-Lobato, and A. Suárez, “Expectation propagation in linear regression models with spike-and-slab priors,” Machine Learning, vol. 99, no. 3, pp. 437–487, 2015.

[40] M. R. Hestenes, E. Stiefel et al., Methods of conjugate gradients for solving linear systems. NBS Washington, DC, 1952, vol. 49, no. 1.

[41] P. Sidén, F. Lindgren, D. Bolin, and M. Villani, “Efficient covariance approximations for large sparse precision matrices,” Journal of Computational and Graphical Statistics, vol. 27, no. 4, pp. 898–909, 2018.

[42] M. Gelman, A. Vehtari, P. Jylänki, C. Robert, N. Chopin, and J. P. Cunningham, “Expectation propagation as a way of life,” Journal of Machine Learning Research, vol. 21, pp. 1–53, 2020.

[43] M. Pereyra, J. M. Bioucas-Dias, and M. A. Figueiredo, “Maximum-a-posteriori estimation with unknown regularisation parameters,” in 2015 23rd European Signal Processing Conference (EUSIPCO). IEEE, 2015, pp. 230–234.

[44] A. F. Vidal, V. De Bortoli, M. Pereyra, and A. Durmus, “Maximum likelihood estimation of regularization parameters in high-dimensional inverse problems: An empirical bayesian approach part i: Methodology and experiments,” SIAM Journal on Imaging Sciences, vol. 13, no. 4, pp. 1945–1989, 2020.

[45] G. Celeux, F. Forbes, and N. Peyrard, “EM procedures using mean field-like approximations for Markov model-based image segmentation,” Pattern recognition, vol. 36, no. 1, pp. 131–144, 2003.

[46] P. Ruiz, X. Zhou, J. Mateos, R. Molina, and A. K. Katsaggelos, “Variational bayesian blind image deconvolution: A review,” Digital Signal Processing, vol. 47, pp. 116–127, 2015.

[47] M. V. Afonso, J. M. Bioucas-Dias, and M. A. Figueiredo, “Fast image recovery using variable splitting and constrained optimization,” IEEE transactions on image processing, vol. 19, no. 9, pp. 2345–2356, 2010.

[48] C.-A. Deledalle, S. Vaiter, J. Fadili, and G. Peyré, “Stein unbiased gradient estimator of the risk (sugar) for multiple parameter selection,” SIAM Journal on Imaging Sciences, vol. 7, no. 4, pp. 2448–2487, 2014.

[49] V. A. Morozov, Methods for solving incorrectly posed problems. Springer Science & Business Media, 2012.