Journal of Machine Learning Research 15 (2014) 3775-3814 Submitted 5/13; Revised 5/14; Published 11/14

Revisiting Bayesian Blind Deconvolution

David Wipf DAVIDWIPF@GMAIL.COM
Visual Computing Group
Microsoft Research
Building 2, No. 5 Danling Street
Beijing, P.R. China, 100080

Haichao Zhang HCZHANG1@GMAIL.COM
School of Computer Science
Northwestern Polytechnical University
127 West Youyi Road
Xi'an, P.R. China, 710072

Editor: Lawrence Carin

Abstract

Blind deconvolution involves the estimation of a sharp signal or image given only a blurry observation. Because this problem is fundamentally ill-posed, strong priors on both the sharp image and blur kernel are required to regularize the solution space. While this naturally leads to a standard MAP estimation framework, performance is compromised by unknown trade-off parameter settings, optimization heuristics, and convergence issues stemming from non-convexity and/or poor prior selections. To mitigate some of these problems, a number of authors have recently proposed substituting a variational Bayesian (VB) strategy that marginalizes over the high-dimensional image space leading to better estimates of the blur kernel. However, the underlying cost function now involves both integrals with no closed-form solution and complex, function-valued arguments, thus losing the transparency of MAP. Beyond standard Bayesian-inspired intuitions, it thus remains unclear by exactly what mechanism these methods are able to operate, rendering understanding, improvements and extensions more difficult. To elucidate these issues, we demonstrate that the VB methodology can be recast as an unconventional MAP problem with a very particular penalty/prior that conjoins the image, blur kernel, and noise level in a principled way. This unique penalty has a number of useful characteristics pertaining to relative concavity, local minima avoidance, normalization, and scale-invariance that allow us to rigorously explain the success of VB including its existing implementational heuristics and approximations. It also provides strict criteria for learning the noise level and choosing the optimal image prior that, perhaps counter-intuitively, need not reflect the statistics of natural scenes. In so doing we challenge the prevailing notion of why VB is successful for blind deconvolution while providing a transparent platform for introducing enhancements and extensions. Moreover, the underlying insights carry over to a wide variety of other bilinear models common in the machine learning literature such as independent component analysis, dictionary learning/sparse coding, and non-negative matrix factorization.

Keywords: blind deconvolution, blind image deblurring, variational Bayes, sparse priors, sparse estimation

1. Introduction

Blind deconvolution problems involve the estimation of some latent sharp signal of interest given only an observation that has been compromised by an unknown filtering process. Although relevant algorithms and analysis apply in a general setting, this paper will focus on the particular case of blind image deblurring, where an unknown convolution or blur operator, as well as additive noise, corrupt the image capture of an underlying natural scene. Such blurring is an undesirable consequence that often accompanies the image formation process and may arise, for example, because of camera-shake during acquisition. Blind image deconvolution or deblurring strategies aim to recover a sharp image from only a blurry, compromised observation, a long-standing problem (Richardson, 1972; Lucy, 1974; Kundur and Hatzinakos, 1996) that remains an active research topic (Fergus et al., 2006; Shan et al., 2008; Levin et al., 2009; Cho and Lee, 2009; Krishnan et al., 2011). Moreover, applications extend widely beyond standard photography, with astronomical, bio-imaging, and other signal processing data eliciting particular interest (Zhu and Milanfar, 2013; Kenig et al., 2010).

Assuming a convolutional blur model with additive noise (Fergus et al., 2006; Shan et al., 2008), the low quality image observation process is commonly modeled as

$$\mathbf{y} = \mathbf{k} * \mathbf{x} + \mathbf{n}, \tag{1}$$

where $\mathbf{k}$ is the point spread function (PSF) or blur kernel, $*$ denotes the 2D convolution operator, and $\mathbf{n}$ is a zero-mean Gaussian noise term (although as we shall see, these assumptions about the noise distribution can easily be relaxed via the framework described herein). The task of blind deconvolution is to estimate both the sharp image $\mathbf{x}$ and blur kernel $\mathbf{k}$ given only the blurry observation $\mathbf{y}$, where we will mostly be assuming that $\mathbf{x}$ and $\mathbf{y}$ represent filtered (e.g., gradient domain) versions of the original pixel-domain images. Because $\mathbf{k}$ is non-invertible, some (typically) high frequency information is lost during the observation process, and thus even if $\mathbf{k}$ were known, the non-blind estimation of $\mathbf{x}$ is ill-posed. However, in the blind case where $\mathbf{k}$ is also unknown, the difficulty is exacerbated considerably, with many possible image/kernel pairs explaining the observed data equally well. This is analogous to the estimation challenges associated with a wide variety of bilinear models, where the observation model (1) is generalized to

$$\mathbf{y} = \mathbf{H}(\mathbf{k})\mathbf{x} + \mathbf{n}. \tag{2}$$

Here $\mathbf{y}$, $\mathbf{x}$, and $\mathbf{k}$ can be arbitrary matrices or vectors, and $\mathbf{k}$ represents unknown parameters embedded in the linear operator $\mathbf{H}(\mathbf{k})$. Note that (1) represents a special case of (2) when $\mathbf{y}$ and $\mathbf{x}$ are vectorized images and $\mathbf{H}(\mathbf{k})$ is the Toeplitz convolution matrix associated with $\mathbf{k}$. Other important instances prevalent in the machine learning and signal processing literature include independent component analysis (ICA) (Hyvarinen and Oja, 2000), dictionary learning for sparse coding (Mairal et al., 2010), and non-negative matrix factorization (Lee and Seung, 2001).

To alleviate the intrinsic indeterminacy, prior assumptions must be adopted to constrain the space of candidate solutions, which naturally suggests a Bayesian framework. In Section 2, we briefly review the two most common classes of Bayesian algorithms for blind deconvolution used in the literature, (i) Maximum a Posteriori (MAP) estimation and (ii)

Variational Bayes (VB), and then later detail their fundamental limitations, which include heuristic implementational requirements and complex cost functions that are difficult to disentangle. Section 3 uses ideas from convex analysis to reformulate these Bayesian methods promoting greater understanding and suggesting useful enhancements, such as rigorous criteria for choosing appropriate image priors. Section 4 then situates our theoretical analysis within the context of existing analytic studies of blind deconvolution, notably the seminal work from Levin et al. (2009, 2011b), and discusses the relevance of natural image statistics. Learning noise variances is later addressed in Section 5, while experiments are carried out in Section 6 to provide corroborating empirical evidence for some of our theoretical claims. Finally, concluding remarks are contained in Section 7. While nominally directed at the challenges of blind deconvolution, we envision that the underlying principles analyses developed herein will nonetheless contribute to better understanding of generalized bilinear models in broad application domains.

2. MAP versus VB

As mentioned above, to compensate for the ill-posedness of the blind deconvolution problem, a strong prior is required for both the sharp image and kernel to regularize the solution space. Recently, natural image statistics over image gradients have been invoked to design prior (regularization) models (Roth and Black, 2009; Levin et al., 2007; Krishnan and Fergus, 2009; Cho et al., 2012), and MAP estimation using these priors has been proposed for blind deconvolution (Shan et al., 2008; Krishnan et al., 2011). While some specifications may differ, the basic idea is to find the mode (maximum) of

$$p(\mathbf{x},\mathbf{k}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x},\mathbf{k})p(\mathbf{x})p(\mathbf{k})}{p(\mathbf{y})} \propto p(\mathbf{y}|\mathbf{x},\mathbf{k})p(\mathbf{x})p(\mathbf{k}),$$

where $\mathbf{x}$ and $\mathbf{y}$ are now assumed to be vectorized gradient domain sharp and blurry images respectively, and $\mathbf{k}$ is the corresponding vectorized kernel.[1] $p(\mathbf{y}|\mathbf{x},\mathbf{k})$ is a Gaussian likelihood function with mean $\mathbf{k}*\mathbf{x}$ and covariance $\lambda\mathbf{I}$, and $p(\mathbf{x})$ and $p(\mathbf{k})$ are priors, with the former often assumed to be sparsity-promoting consistent with estimates of natural image statistics (Buccigrossi and Simoncelli, 1999; Levin et al., 2011b). After a $-2\log$ transformation, and ignoring constant factors, this is equivalent to computing

$$\min_{\mathbf{x},\mathbf{k}} -2\log p(\mathbf{x},\mathbf{k}|\mathbf{y}) \equiv \min_{\mathbf{x},\mathbf{k}} \frac{1}{\lambda}\|\mathbf{k}*\mathbf{x}-\mathbf{y}\|_2^2 + g_{\mathbf{x}}(\mathbf{x}) + g_{\mathbf{k}}(\mathbf{k}), \tag{3}$$

where $g_{\mathbf{x}}(\mathbf{x})$ is a penalty term over the desired image, typically of the form $g_{\mathbf{x}}(\mathbf{x}) = \sum_i g_x(x_i)$, while $g_{\mathbf{k}}(\mathbf{k})$ regularizes the blur kernel. Both penalties generally have embedded parameters that must be balanced along with the unknown λ. It is also typical to assume that $\sum_i k_i = 1$, with $k_i \geq 0$ and we will adopt this assumption throughout; however, Sections 3.5 and 3.7 will discuss a type of scale invariance such that this assumption becomes irrelevant in important cases.

Although straightforward, there are many problems with existing MAP approaches including ineffective global minima, e.g., poor priors may lead to degenerate global solutions

1. Even in vectorized form, we will still use $\mathbf{k}*\mathbf{x}$ to denote the standard 2D convolution operator, where the result is then subsequently vectorized.

like the delta kernel (frequently called the no-blur solution), or many suboptimal local minima and subsequent convergence issues. Therefore, the generation of useful solutions requires a delicate balancing of various factors such as dynamic noise levels, trade-off parameter values, and other heuristic regularizers such as salient structure selection (Shan et al., 2008; Cho and Lee, 2009; Krishnan et al., 2011) (we will discuss these issues more in Section 3).

To mitigate some of these shortcomings of MAP, the influential work by Levin et al. (2009) and others proposes to instead solve

$$\max_{\mathbf{k}} p(\mathbf{k}|\mathbf{y}) \equiv \min_{\mathbf{k}} -2\log p(\mathbf{y}|\mathbf{k})p(\mathbf{k}), \tag{4}$$

where $p(\mathbf{y}|\mathbf{k}) = \int p(\mathbf{x},\mathbf{y}|\mathbf{k})d\mathbf{x}$. This technique is sometimes referred to as Type II estimation in the statistics literature.[2] Once $\mathbf{k}$ is estimated in this way, $\mathbf{x}$ can then be obtained via conventional non-blind deconvolution techniques. One motivation for the Type II strategy is based on the inherent asymmetry in the dimensionality of the image relative to the kernel (Levin et al., 2009). By integrating out (or averaging over) the high-dimensional image, the estimation process can then more accurately focus on the few remaining low-dimensional parameters in $\mathbf{k}$.

The challenge with (4) is that evaluation of $p(\mathbf{y}|\mathbf{k})$ requires a marginalization over $\mathbf{x}$, which is a computationally intractable integral given realistic image priors. Consequently a variational Bayesian (VB) strategy is used to approximate the troublesome marginalization (Levin et al., 2011a). A similar idea has also been explored by a number of other authors (Miskin and MacKay, 2000; Fergus et al., 2006; Babacan et al., 2012). In brief, VB provides a convenient way of computing a rigorous upper bound on $-\log p(\mathbf{y}|\mathbf{k})$, which can then be substituted into (4) for optimization purposes leading to an approximate Type II estimator.

The VB methodology can be easily applied whenever the image prior $p(\mathbf{x})$ is expressible as a Gaussian scale mixture (GSM) (Palmer et al., 2006), meaning

$$p(\mathbf{x}) = \exp\left[-\frac{1}{2}g_{\mathbf{x}}(\mathbf{x})\right] = \prod_i \exp\left[-\frac{1}{2}g_x(x_i)\right] = \prod_i \int \mathcal{N}(x_i;0,\gamma_i)p(\gamma_i)d\gamma_i, \tag{5}$$

where each $\mathcal{N}(x_i;0,\gamma_i)$ represents a zero mean Gaussian with variance γ_i and prior distribution $p(\gamma_i)$. The role of this decomposition will become apparent below. Also, with some abuse of notation, $p(\gamma_i)$ may characterize a discrete distribution, in which case the integral in (5) can be reduced to a summation. Note that all prior distributions expressible via (5) will be supergaussian (Palmer et al., 2006), and therefore will to varying degrees favor a sparse $\mathbf{x}$ (we will return to this issue in Sections 3 and 4).

Given this $p(\mathbf{x})$, the negative log of $p(\mathbf{y}|\mathbf{k})$ can be upper bounded via

$$-\log p(\mathbf{y}|\mathbf{k}) \leq \underbrace{-\iint q(\mathbf{x},\boldsymbol{\gamma})\log\frac{p(\mathbf{x},\boldsymbol{\gamma},\mathbf{y}|\mathbf{k})}{q(\mathbf{x},\boldsymbol{\gamma})}d\mathbf{x}d\boldsymbol{\gamma}}_{F[q(\mathbf{x},\boldsymbol{\gamma}),\mathbf{k}]},$$

2. To be more specific, Type II estimation refers to the case where we optimize over one set of unknown variables after marginalizing out another set, in our case the image $\mathbf{x}$. In this context, standard MAP over both $\mathbf{x}$ and $\mathbf{k}$ via (3) can be viewed as Type I.

where $F[q(\mathbf{x},\boldsymbol{\gamma}),\mathbf{k}]$ is called the *free energy*, $q(\mathbf{x},\boldsymbol{\gamma})$ is an arbitrary distribution over $\mathbf{x}$, and $\boldsymbol{\gamma} = [\gamma_1,\gamma_2,\ldots]^T$, the vector of all the variances from (5). Equality is obtained when $q(\mathbf{x},\boldsymbol{\gamma}) = p(\mathbf{x},\boldsymbol{\gamma}|\mathbf{y},\mathbf{k})$. In fact, if we were able to iteratively minimize this F over $q(\mathbf{x},\boldsymbol{\gamma})$ and $\mathbf{k}$ (i.e., a form of coordinate descent), this would be exactly equivalent to the standard expectation-maximization (EM) algorithm for minimizing $-\log p(\mathbf{y}|\mathbf{k})$ with respect to $\mathbf{k}$, treating $\boldsymbol{\gamma}$ and $\mathbf{x}$ as hidden data and assuming $p(\mathbf{k})$ is flat within the specified constraint set mentioned previously (see Bishop 2006, Ch.9.4 for a detailed examination of this fact). However, optimizing over $q(\mathbf{x},\boldsymbol{\gamma})$ is intractable since $p(\mathbf{x},\boldsymbol{\gamma}|\mathbf{y},\mathbf{k})$ is generally not available in closed-form. Likewise, there is no closed-form update for $\mathbf{k}$, and hence no exact EM solution is possible.

The contribution of VB theory is to show that if we restrict the form of $q(\mathbf{x},\boldsymbol{\gamma})$ via structural assumptions, the updates can now actually be computed, albeit approximately. For this purpose the most common constraint is that $q(\mathbf{x},\boldsymbol{\gamma})$ must be factorized, namely, $q(\mathbf{x},\boldsymbol{\gamma}) = q(\mathbf{x})q(\boldsymbol{\gamma})$, sometimes called a mean-field approximation (Bishop, 2006, Ch.10.1). With this approximation we are effectively utilizing the revised (and looser) upper bound

$$-\log p(\mathbf{y}|\mathbf{k}) \leq -\iint q(\mathbf{x})q(\boldsymbol{\gamma})\log\frac{p(\mathbf{x},\boldsymbol{\gamma},\mathbf{y}|\mathbf{k})}{q(\mathbf{x})q(\boldsymbol{\gamma})}d\mathbf{x}d\boldsymbol{\gamma}, \tag{6}$$

which may be iteratively minimized over $q(\mathbf{x})$, $q(\boldsymbol{\gamma})$, and $\mathbf{k}$ independently while holding the other two fixed. In each case, closed-form updates are now possible, although because of the factorial approximation, we are of course no longer guaranteed to minimize $-\log p(\mathbf{y}|\mathbf{k})$.

Compared to the original Type II problem from (4), minimizing the bound from (6) is considerably simplified because the problematic marginalization over $\mathbf{x}$ has been effectively decoupled from $\boldsymbol{\gamma}$. However, when solving for $q(\mathbf{x})$ at each iteration, it can be shown that a full covariance matrix of $\mathbf{x}$ conditioned on $\boldsymbol{\gamma}$, denoted as $\mathbf{C}$, must be computed. While this is possible in closed form, it requires $O(m^3)$ operations, where m is the number of pixels in the image. Because this is computationally impractical for reasonably-sized images, a diagonal approximation to $\mathbf{C}$ must be adopted (Levin et al., 2011a). This assumption is equivalent to incorporating an additional factorization into the VB process such that now we are enforcing the constraint $q(\mathbf{x},\boldsymbol{\gamma}) = \prod_i q(x_i)q(\gamma_i)$. This leads to the considerably looser upper bound

$$-\log p(\mathbf{y}|\mathbf{k}) \leq -\iint \prod_i q(x_i)q(\gamma_i)\log\frac{p(\mathbf{x},\boldsymbol{\gamma},\mathbf{y}|\mathbf{k})}{\prod_i q(x_i)q(\gamma_i)}d\mathbf{x}d\boldsymbol{\gamma}.$$

In summary then, the full Type II approach can be approximated by minimizing the VB upper bound via the optimization problem

$$\min_{q(\mathbf{x},\boldsymbol{\gamma}),\mathbf{k}} \; F\left[q(\mathbf{x},\boldsymbol{\gamma}),\mathbf{k}\right], \quad \text{s.t. } q(\mathbf{x},\boldsymbol{\gamma}) = \prod_i q(x_i)q(\gamma_i). \tag{7}$$

The requisite update rules are shown in Algorithm 1.[3] Numerous methods fall within this category with some implementational differences, and the estimation steps are equivalent

3. For simplicity we have ignored image boundary effects when presenting the computation for c_j in Algorithm 1. In fact, the complete expression for c_j is described in Appendix A in the proof of Theorem 1. Additionally, Algorithm 1 in its present form includes a modest differentiability assumption on g_x for

Algorithm 1 VB Blind Deblurring (Levin et al., 2011a; Palmer et al., 2006; Babacan et al., 2012)

1: **Input:** blurry gradient domain image $\mathbf{y}$, noise level reduction factor β ($\beta > 1$), minimum noise level λ_0, image prior $p(\mathbf{x}) = \exp[-\frac{1}{2}g_{\mathbf{x}}(\mathbf{x})] = \prod_i \exp[-\frac{1}{2}g_x(x_i)]$

2: **Initialize:** blur kernel $\mathbf{k}$, noise level λ

3: **While** stopping criteria is not satisfied, repeat

- **Update sufficient statistics for $q(\boldsymbol{\gamma}) = \prod_i q(\gamma_i)$:**

$$\omega_i \triangleq \mathrm{E}_{q(\gamma_i)}[\gamma_i^{-1}] \leftarrow \frac{{g_x}'(\sigma_i)}{2\sigma_i},$$

  with $\sigma_i^2 \triangleq \mathrm{E}_{q(x_i)}[x_i^2] = \mu_i^2 + C_{ii}$.

- **Update sufficient statistics for $q(\mathbf{x}) = \prod_i q(x_i)$:**

$$\boldsymbol{\mu} \triangleq \mathrm{E}_{q(\mathbf{x})}[\mathbf{x}] \leftarrow \mathbf{A}^{-1}\mathbf{b}, \quad C_{ii} \triangleq \mathrm{Var}_{q(x_i)}[x_i] \quad \leftarrow A_{ii}^{-1},$$

  where $\mathbf{A} = \frac{\mathbf{H}^T\mathbf{H}}{\lambda} + \mathrm{diag}[\boldsymbol{\omega}]$, $\mathbf{b} = \frac{\mathbf{H}^T\mathbf{y}}{\lambda}$, $\mathbf{H}$ is the convolution matrix of $\mathbf{k}$.

- **Update k:**

$$\mathbf{k} \leftarrow \arg\min_{\mathbf{k}\geq 0} \|\mathbf{y} - \mathbf{W}\mathbf{k}\|_2^2 + \sum_j c_j k_j^2$$

  where $c_j = \sum_i C_{i+j,i+j}$ and $\mathbf{W}$ is the convolution matrix of $\boldsymbol{\mu}$.

- **Noise level reduction:** If $\lambda > \lambda_0$, then $\lambda \leftarrow \lambda/\beta$.

4: **Final Non-Blind Step:** In original image domain, estimate sharp image using fixed $\mathbf{k}$ from above

to simply inserting the kernel update rule and noise reduction heuristic from Levin et al. (2011a) into the general VB sparse estimation framework from Palmer et al. (2006). Results using this strategy for blind deblurring with different priors can be found in Babacan et al. (2012). Note that the full distributions for each $q(x_i)$ and $q(\gamma_i)$ are generally not needed; only certain sufficient statistics are required (certain means and variances, see Algorithm 1), analogous to standard EM. These can be efficiently computed using techniques from Palmer et al. (2006) for any $p(\mathbf{x})$ produced by (5). In the VB algorithm from Levin et al. (2011a), the sufficient statistic for $\boldsymbol{\gamma}$ is computed using an alternative methodology which applies only to finite Gaussian scale mixtures. However, the resulting updates are nonetheless equivalent to Algorithm 1 as shown in the proof of Theorem 1 presented later.

updating the sufficient statistics of $q(\gamma_i)$). Finally, while it is trivial to include multiple image filters in this pipeline (Levin et al., 2011a; Babacan et al., 2012), we avoid including such additional notation to simplify the exposition. Here we are already assuming that $\mathbf{y}$ and $\mathbf{x}$ represent blurry and sharp gradient domain images, obtained by simple horizontal and vertical first-order difference filters (see Section 3.1).

While possibly well-motivated in principle, the Type II approach relies on rather severe factorial assumptions which may compromise the original high-level justifications. In fact, at any minimizing solution denoted $q^*(x_i), q^*(\gamma_i), \forall i, \mathbf{k}^*$, it is easily shown that the gap between F and $-\log p(\mathbf{y}|\mathbf{k}^*)$ is given explicitly by

$$KL\left(\prod_i q^*(x_i)q^*(\gamma_i)||p(\mathbf{x}, \boldsymbol{\gamma}|\mathbf{y}, \mathbf{k}^*)\right), \tag{8}$$

where $KL(p_1||p_2)$ denotes the standard KL divergence between the distributions p_1 and p_2. Because the posterior $p(\mathbf{x}, \boldsymbol{\gamma}, |\mathbf{y}, \mathbf{k})$ is generally highly coupled (non-factorial), this divergence will typically be quite high, indicating that the associated approximation can be poor. We therefore have no reason to believe that this $\mathbf{k}^*$ is anywhere near the maximizer of $p(\mathbf{y}|\mathbf{k})$, which was the ultimate goal and motivation of Type II to begin with.

Other outstanding issues persist as well. For example, the free energy cost function, which involves both integration and function-valued arguments, is not nearly as transparent as the standard MAP estimation from (3). Moreover for practical use, VB depends on an appropriate schedule for reducing the noise variance λ during each iteration (see Algorithm 1), which implements a form of coarse-to-fine, multiresolution estimation scheme (Levin et al., 2011b) while potentially improving the convergence rate (Levin et al., 2011a).

It therefore becomes difficult to rigorously explain exactly why VB has often been empirically more successful than MAP in practice (see Babacan et al. 2012; Levin et al. 2011b,a for such comparisons), nor how to decide which image priors operate best in the VB framework.[4] While Levin *et al.* have suggested that at a high level, marginalization over the latent sharp image using natural-image-statistic-based priors is a good idea to overcome some of the problems faced by MAP estimation (Levin et al., 2009, 2011b), this argument only directly motivates substituting (4) for (3) rather than providing explicit rationalization for (7). Thus, we intend to more meticulously investigate the exact mechanism by which VB operates, explicitly accounting for all of the approximations and assumptions involved by drawing on convex analysis and sparse estimation concepts from Palmer et al. (2006); Wipf et al. (2011) (Section 4 will discuss direct comparisons with Levin *et al.* in detail). This endeavor then naturally motivates extensions to the VB framework and a simple prescription for choosing an appropriate image prior $p(\mathbf{x})$. Overall, we hope that we can further demystify VB providing an entry point for broader improvements such as robust non-uniform blur and noise estimation.

Several surprising, possibly counterintuitive conclusions emerge from this investigation which challenge some of the prevailing wisdom regarding why and how Bayesian algorithms can be advantageous for blind deconvolution. These include:

- The optimal image prior for blind deconvolution purposes using VB or MAP is likely not the one which most closely reflects natural images statistics. Rather, we argue that it is the distribution that most significantly discriminates between blurry and sharp images, meaning a prior such that, for some good sharp image estimate $\hat{\mathbf{x}}$, we

4. Note that, as discussed later, certain MAP algorithms can perform reasonably well when carefully balanced with additional penalty factors and tuning parameters added to (3). However, in direct comparisons using the same basic probabilistic model, VB can perform substantially better, even achieving state-of-the-art performance without additional tuning.

have $p(\hat{\mathbf{x}}) \gg p(\mathbf{k} * \hat{\mathbf{x}})$. Natural image statistics typically fail in this regard for explicit reasons, which apply to both MAP and VB, as discussed in Sections 3 and 4.

- The advantage of VB over MAP is not directly related to the dimensionality differences between $\mathbf{k}$ and $\mathbf{x}$ and the conventional benefits of marginalization over the latter. In fact, we prove in Section 3.2 that the underlying cost functions are formally equivalent in ideal noiseless environments given the factorial assumptions required by practical VB algorithms, and the same basic line of reasoning holds equally well in the noisy case. Instead, there is an intrinsic mechanism built into VB that allows bad locally minimizing solutions to be largely avoided even when using the highly non-convex, discriminative priors needed to distinguish between blurry and sharp images. This represents a new perspective on the relative advantages of VB.
- The VB algorithm can be reformulated in such a way that non-Gaussian noise models, non-uniform blur operators, and other extensions are easily incorporated, circumventing one important perceived advantage of MAP. In fact, we have already obtained practical success in more complex non-uniform and multi-image models using similar principles (Zhang et al., 2013; Zhang and Wipf, 2013).

3. Analysis of Variational Bayes

Drawing on ideas from Palmer et al. (2006); Wipf et al. (2011), in this section we will reformulate the VB methodology to elucidate its behavior. Simply put, we will demonstrate that VB is actually equivalent to using an unconventional MAP estimation-like cost function but with a particular penalty that links the image, blur kernel, and noise in a well-motivated fashion. This procedure removes the ambiguity introduced by the VB approximation, the subsequent diagonal covariance approximation, and the λ reduction heuristic that all contribute still somewhat mysteriously to the effectiveness of VB. It will then allow us to pinpoint the exact reasons why VB represents an improvement over conventional MAP estimations in the form of (3), and it provides us with a specific criteria for choosing the image prior $p(\mathbf{x})$.

3.1 Notation and Definitions

As mentioned above, and following many previous works (Fergus et al., 2006; Levin et al., 2011a), we will henceforth work entirely in the derivative domain of images, with the exception of an implicit final non-blind deconvolution step once the kernel $\mathbf{k}$ has been estimated. From a practical standpoint, these derivatives are computed by convolving the raw image with standard first-order horizontal and vertical difference filters $[1, -1]$ and $[1, -1]^T$. Given that convolution is a commutative operator, the blur kernel is unaltered. For latent sharp image derivatives of size $M \times N$ and blur kernel of size $P \times Q$, we denote the lexicographically ordered vector of the sharp image derivatives, blurry image derivatives, and blur kernel as $\mathbf{x} \in \mathbb{R}^m$, $\mathbf{y} \in \mathbb{R}^n$ and $\mathbf{k} \in \mathbb{R}^l$ respectively, with $m \triangleq MN$, $n \triangleq (M - P + 1)(N - Q + 1)$, and $l \triangleq PQ$. This assumes a single derivative filter. The extension to multiple filters, for example one for each image dimension as described above, follows naturally. For simplicity of notation however, we omit explicit referencing of multiple filters throughout this paper, although all related analysis directly follow through.

The likelihood model (1) can be rewritten as

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n} = \mathbf{W}\mathbf{k} + \mathbf{n}, \tag{9}$$

where $\mathbf{H} \in \mathbb{R}^{n\times m}$ and $\mathbf{W} \in \mathbb{R}^{n\times l}$ are the Toeplitz convolution matrices constructed from the blur kernel and sharp image respectively. We introduce a matrix $\bar{\mathbf{I}} \in \mathbb{R}^{l\times m}$, where the j-th row of $\bar{\mathbf{I}}$ is a binary vector with 1 indicating that the j-th element of $\mathbf{k}$ (i.e., k_j) appears in the corresponding column of $\mathbf{H}$ and 0 otherwise. We define $\|\bar{\mathbf{k}}\|_2 \triangleq \sqrt{\sum_j k_j^2 \bar{I}_{ji}}$, which is equivalent to the norm of the i-th column of $\mathbf{H}$. It can also be viewed as the effective norm of $\mathbf{k}$ accounting for the boundary effects.[5] The element-wise magnitude of $\mathbf{x}$ is given by $|\mathbf{x}| \triangleq [|x_1|, |x_2|, \ldots]^T$.

Finally we introduce the definition of *relative concavity* (Palmer, 2003) which will serve subsequent analyses:

Definition 1 *Let u be a strictly increasing function on $[a, b]$. The function ν is **concave relative** to u on the interval $[a, b]$ if and only if*

$$\nu(y) \leq \nu(x) + \frac{\nu'(x)}{u'(x)}\left[u(y) - u(x)\right] \tag{10}$$

holds $\forall x, y \in [a, b]$.

In the following, we will use $\nu \prec u$ to denote that ν is concave relative to u on $[0, \infty)$. This can be understood as a natural generalization of the traditional notion of a concavity, in that a concave function is equivalently *concave relative to a linear function* per Definition 1. In general, if $\nu \prec u$, then when ν and u are set to have the same functional value and the same slope at any given point (i.e., by an affine transformation of u), then ν lies completely under u.

It is well-known that functions concave in $|\mathbf{x}|$ favor sparsity (meaning a strong preference for zero and relatively little distinction between nonzero values) (Rao et al., 2003; Wipf et al., 2011). The notion of relative concavity induces an ordering for many of the common sparsity promoting functions. Intuitively, a non-decreasing function ν of $|x_i|$ is more aggressive in promoting sparsity than some u if it is concave relative to u. For example, consider the class of ℓ_p functionals $\sum_i |x_i|^p$ that are concave in $|x_i|$ whenever $0 < p \leq 1$. The smaller p, the more a sparse $\mathbf{x}$ will be favored, with the extreme case $p \to 0$ producing the ℓ_0 norm (a count of the number of nonzero elements in $\mathbf{x}$), which is the most aggressive penalty for promoting sparsity. Meanwhile, using Definition 1 it is easy to verify that, as a function of $|\mathbf{x}|$, $\ell_{p_1} \prec \ell_{p_2}$ whenever $p_1 < p_2$.

5. Technically $\|\bar{\mathbf{k}}\|_2$ depends on i, the index of image pixels, but it only makes a difference near the image boundaries. We prefer to avoid an explicit notational dependency on i to keep the presentation concise, although the proofs in Appendix A do consider i-dependency when it is relevant. The subsequent analysis will also omit this dependency keeping in mind that all of the results nonetheless carry through in the general case. The same is true for the other quantities that depend on $\|\bar{\mathbf{k}}\|_2$, e.g., the ρ parameter defined later in (12).

3.2 Connecting VB with MAP

As mentioned previously, the VB algorithm of Levin et al. (2011a) can be efficiently implemented using any image prior expressible in the form of (5). However, for our purposes we require an alternative representation with roots in convex analysis. Based on Palmer et al. (2006), it can be shown that any prior given by (5) can also be represented as a maximization over scaled Gaussians with different variances leading to the alternative representation

$$p(x_i) = \exp\left[-\frac{1}{2}g_x(x_i)\right] = \max_{\gamma_i \geq 0} \mathcal{N}(x_i; 0, \gamma_i) \exp\left[-\frac{1}{2}f(\gamma_i)\right], \tag{11}$$

where $f(\gamma_i)$ is some non-negative energy function; the associated exponentiated factor is sometimes treated as a hyperprior, although it will not generally integrate to one. This f, which determines the form of g_x in (5), will ultimately play a central role in how VB penalizes images $\mathbf{x}$ as will be explored via the results of this section.

Theorem 1 *Consider the objective function*

$$\mathcal{L}(\mathbf{x}, \mathbf{k}) \triangleq \frac{1}{\lambda}\|\mathbf{y} - \mathbf{k} * \mathbf{x}\|_2^2 + \sum_i g_{\text{VB}}(x_i, \rho) + m \log \|\bar{\mathbf{k}}\|_2^2, \tag{12}$$

where

$$g_{\text{VB}}(x_i, \rho) \triangleq \min_{\gamma_i \geq 0}\left[\frac{x_i^2}{\gamma_i} + \log(\rho + \gamma_i) + f(\gamma_i)\right], \text{ and } \rho \triangleq \frac{\lambda}{\|\bar{\mathbf{k}}\|_2^2}. \tag{13}$$

Algorithm 1 minimizing (7) is equivalent to coordinate descent minimization of (12) over $\mathbf{x}$*,* $\mathbf{k}$*, and the latent variables* $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_m]^T$*.*

Proofs will be deferred to the Appendix A. This reformulation of VB closely resembles (3), with a quadratic data fidelity term combined with additive image and kernel penalties. The penalty on $\mathbf{k}$ in (12) is not unlike those incorporated into standard MAP schemes, meaning $g_{\mathbf{k}}(\mathbf{k})$ from (3). However, quite unlike MAP, for $\lambda > 0$ the penalty g_{VB} on the image pixels x_i is dependent on both the noise level λ and the kernel $\mathbf{k}$ through the parameter ρ, the ratio of the noise level to the squared kernel norm. The remainder of Section 3 will explore the consequences of this crucial, yet previously unexamined distinction from typical MAP formulations.

In contrast, with $\lambda = 0$, both MAP and VB possess a formally equivalent penalty on each x_i via the following corollary:

Corollary 1 *If* $\lambda = 0$*, then* $g_{\text{VB}}(x_i, 0) = g_x(x_i) \equiv -2 \log p(x_i)$*.*

Therefore the underlying VB cost function *is effectively no different than regular MAP from (3) in the noiseless setting*, a surprising conclusion that seems to counter much of the prevailing understanding of VB deconvolution algorithms. So then what exactly is the true advantage, if any, of VB over MAP? And is the advantage limited to cases when the noise level is significant? The remainder of Section 3 will address these questions, demonstrating that VB maintains a significant advantage over MAP, and that this advantage persists, perhaps paradoxically, even when the noise level is small or zero. These conclusions ultimately hinge on detailed properties of the image penalty g_{VB} in the context of practical deblurring pipelines. We will also examine the related issue of choosing the optimal image prior $p(\mathbf{x})$, which is equivalent to choosing the optimal f in (11).

3.3 Evaluating the VB Image Penalty g_{VB}

While in a few special cases g_{VB} can be computed in closed-form for general $\rho \neq 0$ leading to greater transparency, as we shall see below the VB algorithm and certain attendant analyses can nevertheless be carried through even when closed-form solutions for g_{VB} are not possible. Importantly, we can assess properties that may potentially affect the sparsity and quality of resulting solutions as λ and $\|\bar{\mathbf{k}}\|_2^2$ are varied.

A highly sparse prior, and therefore penalty function, is generally more effective in differentiating sharp images with fine structures from blurry ones (details in Section 4). Recall that concavity with respect to coefficient magnitudes is a signature property of such sparse penalties (Rao et al., 2003; Wipf et al., 2011). A potential advantage of MAP is that it is very straightforward to characterize the associated image penalty; namely, if g_x from (3) is a highly concave, nondecreasing function of each $|x_i|$, then we may expect that sparse image gradients will be heavily favored. And for two candidate image penalties ${g_x}^{(1)}$ and ${g_x}^{(2)}$, if ${g_x}^{(1)} \prec {g_x}^{(2)}$, then we may expect the former to promote an even sparser solution than the latter (provided we are not trapped at a bad local solution). Section 4 will argue that ${g_x}^{(1)}$ will then lead to a better estimate of $\mathbf{x}$ and $\mathbf{k}$.

In contrast, with VB it is completely unclear to what degree g_{VB} favors sparse solutions, except in the special case from the previous section when g_{VB} is equal to the MAP penalty g_x. We now explicitly describe sufficient and necessary conditions for g_{VB} to be a concave, nondecreasing function of $|x_i|$, which turn out to be much stricter than the conditions required for MAP.

Theorem 2 *The VB penalty g_{VB} will be a concave, non-decreasing function of $|x_i|$ for any ρ if and only if f from (11) is a concave, non-decreasing function on $[0, \infty)$. Moreover, at least $m-n$ elements of $\mathbf{x}$ will equal zero at any locally minimizing solution to (12) (however typically many more will equal zero in practice).*

Theorem 2 explicitly quantifies what class of image priors leads to a strong, sparsity-promoting $\mathbf{x}$ penalty when fully propagated through the VB framework. Yet while this attribute may anchor VB as a legitimate sparse estimator in the image (filter) domain given an appropriate f, it does not explain precisely why VB often produces superior results to MAP. In fact, the associated MAP penalty g_x (when generated from the same f) will actually promote sparse solutions under much weaker conditions as follows:

Corollary 2 *The MAP penalty g_x will be a concave, non-decreasing function of $|x_i|$ if and only if $\vartheta(z) \triangleq \log(z) + f(z)$ is a concave, non-decreasing function on $[0, \infty)$.*

The extra log factor implies that f itself need not be concave to ensure that g_x is concave. For example, the selection $f(z) = z - \log(z)$ it not concave and yet the associated g_x still will be since now $\vartheta(z) = z$, which is concave and non-decreasing as required by Corollary 2. The stronger proclivity for producing sparsity of MAP over VB is further quantified by the following:

Corollary 3 *Let f be a differentiable, non-decreasing function which induces the penalty functions g_x and g_{VB} associated with MAP and VB respectively. As $z \to \infty$, then $g_{\text{VB}}(z) - g_x(z) \to 0$, and therefore g_{VB} and g_x penalize large magnitudes of $\mathbf{x}$ equally. In contrast,*

for any z,$z' \geq 0$, if $z < z'$ then $g_{\text{VB}}(z) - g_x(z) \geq g_{\text{VB}}(z') - g_x(z')$. Therefore, as $z \to 0$, $g_{\text{VB}}(z) - g_x(z)$ is maximized, implying that the MAP penalty g_x favors zero-valued coefficients (sparsity) more heavily than g_{VB}.

This result implies that for a broad class of image priors, VB actually leads to a *weaker* enforcement of sparsity than the corresponding MAP estimator. This occurs because large magnitudes of any x_i are penalized nearly equivalently with VB and MAP, while small magnitudes are penalized much more aggressively with MAP. Taken together then, Corollaries 2 and 3 superficially suggest that perhaps MAP should be preferred over VB to the extent that we believe sparsity is important for distinguishing sharp and blurry images. However, a closer investigation will reveal why this conclusion is premature.

For this purpose we will consider closely the simplest choice for f which satisfies the conditions of Theorem 2, and a choice that has been advocated in the sparse estimation literature in different contexts: namely, a constant value, $f(\gamma) = b$. This in turn implies that the resulting prior $p(x_i)$ is a Jeffreys non-informative distribution on the coefficient magnitudes $|x_i|$ after solving the maximization from (11), and is attractive in part because there are no embedded hyperparameters (the constant b is irrelevant).[6] This selection for f leads to a particularly interesting closed-form penalty g_{VB} as follows:

Theorem 3 *In the special case where $f(\gamma_i) = b$, then*

$$g_{\text{VB}}(x_i, \rho) \equiv \frac{2|x_i|}{|x_i| + \sqrt{x_i^2 + 4\rho}} + \log\left(2\rho + x_i^2 + |x_i|\sqrt{x_i^2 + 4\rho}\right). \tag{14}$$

Figures 1 (a) and (b) display 1D and 2D plots of this penalty function. It is worth spending some time here to examine this particular selection for f (and therefore g_{VB}) in detail since it elucidates many of the mechanisms whereby VB, with all of its attendant approximations and heuristics, can affect improvement over MAP regardless of the true noise level.

In the limit as $\rho \to 0$, the first term in (14) converges to the indicator function $I[x_i \neq 0]$, and thus when we sum over i we obtain the ℓ_0 norm of $\mathbf{x}$, which represents a canonical measure of sparsity, or a count of the nonzero elements in a vector.[7] The second term in (14), when we again sum over i, converges to $\sum_i \log|x_i|$, ignoring a constant factor. Sometimes referred to as Gaussian entropy, this term can also be connected to the ℓ_0 norm via the relations $\|\mathbf{x}\|_0 \equiv \lim_{p\to 0} \sum_i |x_i|^p$ and $\lim_{p\to 0} \frac{1}{p}\sum_i (|x_i|^p - 1) = \sum_i \log|x_i|$ (Wipf et al., 2011). Thus the cumulative effect when ρ becomes small is an image prior that closely mimics the highly non-convex (in $|x_i|$) ℓ_0 norm which favors maximally sparse solutions. In contrast, when ρ becomes large, it can be shown that both terms in (14), when combined for all i, approach scaled versions of the convex ℓ_1 norm. Additionally, this scaling turns out to be principled in the sense described in Wipf and Wu (2012); Zhang and Wipf (2013).

For intermediate values of ρ between these two extremes, we obtain a g_{VB} that becomes *less* concave with respect to each $|x_i|$ as ρ increases in the formal sense of relative concavity discussed in Section 3.1. In particular, we have the following:

6. The Jeffreys prior is of the form $p(x) \propto 1/|x|$, which represents an improper distribution that does not integrate to one.
7. Although with $\rho = 0$, this term reduces to a constant, and therefore has no impact.

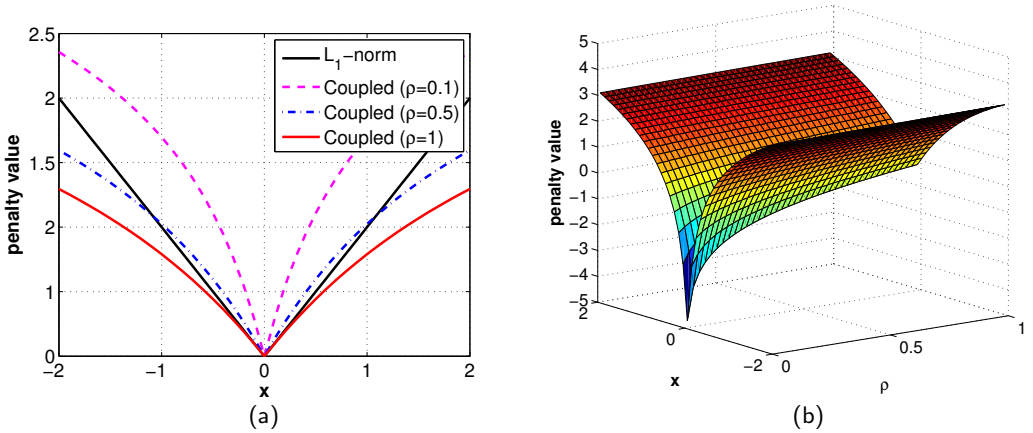

Figure 1: (a) A 1D example of the coupled penalty $g_{\text{VB}}(x, \rho)$ (normalized) with different ρ values assuming f is a constant. The ℓ_1 norm is included for comparison. (b) A 2D example surface plot of the coupled penalty function $g_{\text{VB}}(x, \rho)$; f is a constant.

Corollary 4 *If* $f(\gamma_i) = b$, *then* $g_{\text{VB}}^{\rho_1} \prec g_{\text{VB}}^{\rho_2}$ *for* $\rho_1 < \rho_2$.

Thus, as the noise level λ is increased, ρ increases and we have a penalty that behaves more like a convex (less sparse) function, and so becomes less prone to local minima. In contrast, as $\|\bar{\mathbf{k}}\|_2^2$ is increased, meaning that ρ is now reduced, the penalty actually becomes *more* concave with respect to $|x_i|$. This phenomena is in some ways similar to certain homotopy continuation sparse estimation schemes (e.g., Chartrand and Yin 2008), where heuristic hyperparameters are introduced to gradually introduce greater non-convexity into canonical compressive sensing problems, but without any dependence on the noise or other factors. The key difference here with VB however is that penalty shape modulation is explicitly dictated by both the noise level λ and the kernel $\mathbf{k}$ in an integrated fashion. [8]

To summarize then, the ratio ρ can be viewed as modulating a smooth transition of the penalty function shape from something akin to the non-convex ℓ_0 norm to a properly-scaled ℓ_1 norm. In contrast, conventional MAP-based penalties on $\mathbf{x}$ are independent from $\mathbf{k}$ or λ, and thus retain a fixed shape. The crucial ramifications of this coupling and ρ-controlled shape modification/augmentation exclusive to the VB framework will be addressed in the following two subsections. Other choices for f, which exhibit a partially muted form of this coupling, will be considered in Section 3.7, which will also address a desirable form of invariance that only exists when f is a constant.

3.4 Noise Dependency Analysis

The success of practical VB blind deconvolution algorithms is heavily dependent on some form of stagewise coarse-to-fine approach, whereby the kernel is repeatedly re-estimated at

8. After our original submission, a new MAP-related algorithm was published that applies a continuation strategy, not unlike Chartrand and Yin (2008)), to blind deblurring and achieves very promising results (Xu et al., 2013). However, unlike VB this algorithm requires two tuning parameters balancing kernel and image penalties and a fixed, pre-defined schedule for modulating the image penalty shape.

successively higher resolutions. At each stage, a lower resolution version is used to initialize the estimate at the next higher resolution. One way to implement this approach is to initially use large values of λ (regardless of the true noise level) such that only dominant, primarily low-frequency image structures dictate the optimization (Levin et al., 2009). During subsequent iterations as the blur kernel begins to reflect the correct coarse shape, λ can be gradually reduced to allow the recovery of more detailed, fine structures.

A highly sparse (concave) prior can ultimately be more effective in differentiating sharp images and fine structures than a convex one. Detailed supported evidence for this claim can be found in Fergus et al. (2006); Levin et al. (2007); Krishnan and Fergus (2009); Cho et al. (2012), as well as in Section 4 below. However, if such a prior is applied at the initial stages of estimation, the iterations are likely to become trapped at suboptimal local minima, of which there will always be a combinatorial number. Moreover, in the early stages, the effective noise level is actually high due to errors contained in the estimated blur kernel, and exceedingly sparse image penalties are likely to produce unstable solutions.

Given the reformulation outlined above, we can now argue that VB implicitly avoids these problems by beginning with a large λ (and therefore a large ρ), such that the penalty function is initially nearly convex in $|x_i|$ (see Figure 1). In this situation, the data fidelity term $\frac{1}{\lambda}\|\mathbf{y}-\mathbf{k}*\mathbf{x}\|_2^2$ from (12) is de-emphasized because of the inverse dependency on λ, and small edges and structures will be ignored. Hence an approximately convex penalty is generally sufficient for resolving the remaining strong edges. As the iterations proceed and fine structures need to be resolved, the penalty function is made less convex (more concave) as λ is reduced, but the risk of local minima and instability is ameliorated by the fact that we are likely to be already in the neighborhood of a desirable basin of attraction. Additionally, the implicit noise level (or modeling error) is now substantially lower.

This kind of automatic 'resolution' adaptive penalty shaping is arguably superior to conventional MAP approaches based on (3), where the concavity/shape of the induced separable penalty function is kept fixed regardless of the variation in the noise level or scale, i.e., at different resolutions across the coarse-to-fine hierarchy. In general, it would seem very unreasonable that the same penalty shape would be optimal across vastly different noise scales. This advantage over MAP can be easily illustrated by simple head-to-head comparisons where the underlying prior distributions are identical; Section 3.6 below contains one such example. Additionally, this phenomena can be significantly enhanced by automatically learning λ as discussed in Section 5.

3.5 Blur Dependency Analysis

There are many different viewpoints for understanding how the blur dependency of the VB penalty g_{VB} contributes to successful deblurring results. Here we consider potentially one of the most transparent.

Because the blur kernel appears judiciously in all three terms in (12), with a slight reparameterization and subsequent algebraic manipulation, we may consolidate terms leading to the equivalent revised formulation of (12) given by

$$\mathcal{L}(\tilde{\mathbf{x}},\tilde{\mathbf{k}}) \triangleq \frac{1}{\lambda}\left\|\mathbf{y}-\tilde{\mathbf{k}}*\tilde{\mathbf{x}}\right\|_2^2 + \sum_i g_{\text{VB}}(\tilde{x}_i,\lambda), \quad \text{s.t. } \|\tilde{\tilde{\mathbf{k}}}\|_2 = 1, \tag{15}$$

where $\tilde{x}_i \triangleq x_i \|\bar{\mathbf{k}}\|_2$ for all i and, with slight abuse of notation, $\tilde{\mathbf{k}}$ denotes a normalized kernel such that the resulting convolution matrix $\mathbf{H}$ has columns of unit norm. In other words, if $\mathbf{x}^*$ and $\mathbf{k}^*$ represent the optimal solution to (12), then $\tilde{\mathbf{x}}^* = \mathbf{x}^* \|\bar{\mathbf{k}}^*\|_2$ and $\tilde{\mathbf{k}}^* = \mathbf{k}^* / \|\bar{\mathbf{k}}^*\|_2$ are the optimal solution to (15), at least when g_{VB} is given by (14). Hence we see that, once the kernel operator is constrained to have unit ℓ_2 norm, *no additional kernel penalization is included whatsoever*. Consequently then, to the extent VB is successful, we observe that an additional kernel penalty, and any associated tuning parameter, is completely unnecessary. Additionally, with the kernel fixed, solving for $\tilde{\mathbf{x}}$ now represents a standardized sparse estimation problem, with a quadratic data-fit term now characterized by a design matrix $\mathbf{H}$ with consistent ℓ_2 normalized columns.

Note that essentially all previous blind deblurring algorithms assume what amounts to an ℓ_1 normalized kernel satisfying $\sum_i k_i = 1$ (since each element of the sum must be positive, the corresponding convolution matrix $\mathbf{H}$ will have ℓ_1 normalized columns except at the boundary). But in the context of a quadratic data fit term as used by deblurring algorithms, this is unlikely to be optimal as it will apply some implicit pressure on the estimated kernel towards the no-blur solution ($\mathbf{k} = \delta$), potentially counteracting, at least in part, the push for sparse image estimates. This occurs because kernels near the delta solution will increase the ℓ_2 column norms of $\mathbf{H}$ when the ℓ_1 norm is fixed, which then allows for relatively smaller image coefficients $\mathbf{x}$ by virtue of the quadratic data term. These smaller coefficients then reduce any non-decreasing penalty on the magnitudes of $\mathbf{x}$, reducing the overall cost function. Additionally, in much more complex non-uniform deblurring environments, this undesirable effect is considerably more pronounced (Zhang and Wipf, 2013).

In the context of VB however, this ℓ_1 normalization is implicitly switched to ℓ_2 normalization via the mechanism outlined above leading to (15), and hence it is entirely inconsequential to VB. In contrast, MAP has no such agency, and therefore it is not surprising that the majority of MAP algorithms explicitly include an additional ℓ_2 kernel penalty which helps to counteract movement towards no-blur solutions. The disadvantage of course is that an additional image-dependent tuning parameter is required as well to balance the resulting contribution. We could, however, alternatively consider replacing the ℓ_1 norm constraint in MAP with ℓ_2-norm constraints as in (15), although this complicates the optimization process considerably, whereas VB handles this automatically.[9]

3.6 Illustrative Example using 1D Signals

Here we will briefly illustrate some of the distinctions between MAP and VB discussed thus far where other confounding factors have been conveniently removed. For this purpose we consider a simplified noiseless situation where the optimal λ value is zero, and we consider the image prior produced when $f(\gamma) = b$ as introduced in Section 3.3 (later Section 3.7 will

9. One potential way around these complications for MAP would be to replace the non-convex constraint $\|\tilde{\bar{\mathbf{k}}}\|_2 = 1$ with the convex quadratic one $\|\tilde{\bar{\mathbf{k}}}\|_2 \leq 1$, ignoring boundary effects which would complicate things dramatically by requiring m additional constraints, one for each element of $\mathbf{x}$. While in uniform blur models this could be effective since the optimal solution should satisfy $\|\tilde{\bar{\mathbf{k}}}\|_2 = 1$ anyway, in non-uniform models this is unlikely the case, and such a substitution could still be difficult to implement and could undermine performance. But the central point remains that VB handles all of these situations naturally and seamlessly.

argue that this selection is in some sense optimal). For MAP we also include the kernel penalty $m \log \|\bar{\mathbf{k}}\|_2^2$ from (12) which will facilitate more direct comparisons with VB below. Given these assumptions, the associated MAP problem from (3) is easily shown to be

$$\min_{\mathbf{x},\mathbf{k}} \frac{1}{\lambda} \|\mathbf{y} - \mathbf{k} * \mathbf{x}\|_2^2 + \sum_i 2 \log |x_i| + 2m \log \|\bar{\mathbf{k}}\|_2, \tag{16}$$

where the image penalty is obtained by applying a $-2 \log$ transformation to (11) giving

$$-2 \log \left[\max_{\gamma_i \geq 0} \mathcal{N}(x_i; 0, \gamma_i) \right] \equiv 2 \log |x_i| \,. \tag{17}$$

Irrelevant additive constants have been excluded. Conveniently, since

$$\sum_i 2 \log |x_i| + 2m \log \|\bar{\mathbf{k}}\|_2 = \sum_i 2 \log \left(|x_i| \, \|\bar{\mathbf{k}}\|_2 \right), \tag{18}$$

we can reparameterize (16) using $\tilde{\mathbf{x}}$ such that the kernel penalty is removed and the constraint $\|\tilde{\bar{\mathbf{k}}}\|_2 = 1$ is enforced just as with VB. Consequently, in the limit as $\lambda \to 0$, based on the equivalency derived from Corollary 1, both VB and MAP are then effectively solving

$$\min_{\tilde{\mathbf{x}},\tilde{\mathbf{k}}} \sum_i \log |\tilde{x}_i| \,, \quad \text{s.t. } \mathbf{y} = \tilde{\mathbf{k}} * \tilde{\mathbf{x}}, \; \|\tilde{\bar{\mathbf{k}}}\|_2 = 1. \tag{19}$$

Moreover, given the arguments made in Section 3.3, the penalty on $\tilde{\mathbf{x}}$ is more or less equivalent to the ℓ_0 norm up to inconsequential scaling and translations. Thus, (19) effectively reduces to

$$\min_{\tilde{\mathbf{x}},\tilde{\mathbf{k}}} \|\tilde{\mathbf{x}}\|_0, \quad \text{s.t. } \mathbf{y} = \tilde{\mathbf{k}} * \tilde{\mathbf{x}}, \; \|\tilde{\bar{\mathbf{k}}}\|_2 = 1. \tag{20}$$

Therefore at this simplified, stripped-down level both VB and MAP are merely minimizing the ℓ_0 norm of $\mathbf{x}$ subject to the linear convolutional constraint. Of course we do not attempt to solve (20) directly, which is a difficult combinatorial problem in nature. Instead for both VB and MAP we begin with a large λ and gradually reduce it towards zero as part of a multi-resolution approach designed to avoid bad local minima as described in Section 3.4. For this reduction schedule of λ we use $\beta = 1.15$ in Algorithm 1 (this value is taken from Levin et al. 2011a).[10] While equivalent when $\lambda \to 0$, before λ becomes small the VB and MAP cost functions will behave very differently, leading to a radically different optimization trajectory terminating at different locally minimizing solutions to (20).

The superiority of the VB convergence path will now be demonstrated with a synthetic 1D signal composed of multiple spikes. This signal is convolved with two different blur kernels, one uniform and one random, creating two different blurry observations. Refer to Figure 2 (first row) for the ground-truth spike signal and associated blur kernels. We then apply the MAP and VB blind deconvolution algorithms, with the same prior (f equals a constant) and λ reduction schedule, to the blurry test signals and compare the quality of the

10. The corresponding MAP algorithm can be implemented by simply setting $\mathbf{C}$ to zero before the $q(\gamma_i)$ update in Algorithm 1, with guaranteed convergence to some local minima. For both MAP and VB, the γ sufficient statistic update is simply $\omega_i = \sigma_i^{-2}$ whenever f is a constant.

reconstructed blur kernels and signals. The recovery results are shown in Figure 2 (second and third rows), where it is readily apparent that VB produces superior estimation quality of both kernel and image. Additionally, the signal recovered by VB is considerably sparser than MAP, indicating that it has done a better job of optimizing (20), consistent with our previous analysis. This is not to say that MAP cannot potentially be effective with careful tuning and initialization (perhaps coupled with additional regularization factors or clever optimization schemes), only that VB is much more robust in its present form.

Note that this demonstrable advantage of VB is entirely based on an improved convergence path, since VB and MAP possess an identical constellation of local minima once $\lambda = 0$. Moreover, it is unrelated to any putative advantage of solving (4) over (3). We will revisit this latter point in Section 4.

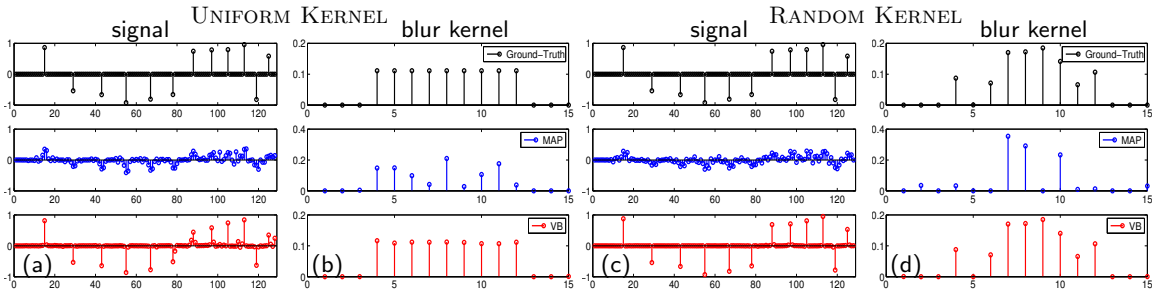

Figure 2: 1D deblurring example using MAP and VB approaches assuming the same underlying image prior $p(\mathbf{x})$. (a)-(b) results with a uniform blur kernel; (c)-(d) results with a random blur kernel.

3.7 Other Choices for f

Because essentially any sparse prior on $\mathbf{x}$ can be expressed using the alternative variational form from (11), choosing such a prior is tantamount to choosing f which then determines g_{VB}. Theorem 2 suggests that a concave, non-decreasing f is useful for favoring sparsity (assumed to be in the gradient domain). Moreover, Theorem 3 and subsequent analyses suggest that the simplifying choice where $f(\gamma) = b$ possesses several attractive properties regarding the relative concavity of the resulting g_{VB}. But what about other selections for f and therefore g_{VB}?

While directly working with g_{VB} can sometimes be limiting (except in certain special cases like $f(\gamma) = b$ from before), the variational form of (13) allows us to closely examine the relative concavity of a useful proxy. Let

$$\psi(\gamma_i, \rho) \triangleq \log(\rho + \gamma_i) + f(\gamma_i). \tag{21}$$

Then for fixed λ and $\mathbf{k}$ the VB estimation problem can equivalently be viewed as solving

$$\min_{\mathbf{x}, \boldsymbol{\gamma} \geq 0} \frac{1}{\lambda} \|\mathbf{y} - \mathbf{k} * \mathbf{x}\|_2^2 + \sum_i \left[\frac{x_i^2}{\gamma_i} + \psi(\gamma_i, \rho) \right]. \tag{22}$$

It now becomes clear that the sparsity of $\mathbf{x}$ and $\boldsymbol{\gamma}$ are intimated related. More concretely, assuming f is concave and non-decreasing (as motivated by Theorems 2 and 3), then there is actually a one-to-one correspondence in that whenever $x_i = 0$, the optimal γ_i equals zero as well, and vice versa.[11] Therefore we may instead examine the relative concavity of ψ for different ρ values, which will directly determine the sparsity of $\boldsymbol{\gamma}$ and in turn, the sparsity of $\mathbf{x}$. This then motivates the following result:

Theorem 4 *Let $\rho_1 < \rho_2$ and assume that f satisfies the conditions of Theorem 2. Then $\psi^{\rho_1} \prec \psi^{\rho_2}$ if and only if $f(\gamma) = a\gamma + b$, with $a \geq 0$.*

Thus, although we have not been able to formally establish a relative concavity result for all general g_{VB} directly, Theorem 4 provides a nearly identical analog allowing us to draw similar conclusions to those detailed in Sections 3.4 and 3.5 whenever a general affine f is adopted. Perhaps more importantly, it also suggests that as f deviates from an affine function, we may begin to lose some of the desirable effects regarding the described penalty shape modulation.

While previously we closely scrutinized the special affine case where $f(\gamma) = b$, it still remains to examine the more general affine form $f(\gamma) = a\gamma + b$, $a > 0$. In fact, it is not difficult to show that as a is increased, the resulting penalty on $\mathbf{x}$ increasingly resembles an ℓ_1 norm with lesser dependency on ρ, thus severely muting the effect of the shape modulation that appears to be so effective (see arguments above and empirical results section below). So there currently does not seem to be any advantage to choosing some $a > 0$ and we are left, out of the multitude of potential image priors, with the conveniently simple choice of $f(\gamma) = b$, where the value of b is inconsequential. Experimental results support this conclusion: namely, as a is increased from zero performance gradually degrades (results not shown for space considerations).

As a final justification for simply choosing $f(\gamma) = b$, there is a desirable form of invariance that uniquely accompanies this selection.

Theorem 5 *If $\mathbf{x}^*$ and $\mathbf{k}^*$ represent the optimal solution to (12) under the constraint $\sum_i k_i = 1$, then $\alpha^{-1}\mathbf{x}^*$ and $\alpha\mathbf{k}^*$ will always represent the optimal solution under the modified constraint $\sum_i k_i = \alpha$ if and only if $f(\gamma) = b$. Additionally, minimizing (12) is equivalent to minimizing (15) if and only if $f(\gamma) = b$.*

This is unlike the myriad of MAP estimation techniques or VB with other choices of f, where the exact calibration of the constraint can fundamentally alter the form of the optimal solution beyond a mere rescaling. Moreover, if such a constraint on $\mathbf{k}$ is omitted altogether, these other methods must then carefully tune associated trade-off parameters, so in one way or another this lack of invariance will require additional tuning.

Interestingly, Babacan et al. (2012) experiments with a variety of VB algorithms using different underlying image priors, and empirically find that f as a constant works best;

11. To see this first consider $x_i = 0$. The x_i^2/γ_i term can be ignored and so the optimal γ_i need only minimize $\log(\rho + \gamma_i) + f(\gamma_i)$, which is concave and non-decreasing whenever f is. Therefore the optimal γ_i is trivially zero. Conversely if $\gamma_i = 0$, then there is effectively an infinite penalty on x_i, and so the optimal x_i must also be zero.

however, no rigorous explanation is given for why this should be the case.[12] Thus, our results provide a powerful theoretical confirmation of this selection, along with a number of useful attendant intuitions.

3.8 Analysis Summary

To summarize this section, we have shown that the shape of the effective VB image penalty is explicitly controlled by the ratio of the noise variance to the squared kernel norm, and that in many circumstances this leads to a desired mechanism for controlling relative concavity and balancing sparsity, largely mitigating issues such as local minima that compromise the convergence of more traditional MAP estimators. We have then demonstrated a unique choice for the image prior (i.e., when f is constant) such that this mechanism is in some sense optimal and scale-invariant. Of course we readily concede that different choices for the image prior could still be useful when other factors are taken in to account. We also emphasize that none of this is meant to suggest that real imaging data follows a Jeffreys prior distribution (which is produced when f is constant). We will return to this topic in Section 4 below. Overall, this perspective provides a much clearer picture of how VB is able to operate effectively and how we might expect to optimize performance.

While space precludes a detailed treatment, many natural extensions to VB are suggested by these developments. For example, in the original formation of VB given by (7) it is not clear the best way to incorporate alternative noise models because the required integrations are no longer tractable. However, when viewed alternatively using (12) it becomes obvious that different data-fidelity terms can easily be substituted in place of the quadratic likelihood factor. Likewise, given additional prior knowledge about the blur kernel, there is no difficulty in substituting for the ℓ_2-norm on $\mathbf{k}$ or the uniform convolutional observation model to reflect additional domain knowledge. Thus, the proposed reformulation allows VB to inherit most of the transparent extensibility previously reserved for MAP.

We may also consider these ideas in the context of existing MAP algorithms, which adopt various structure selection heuristics, implicitly or explicitly, to achieve satisfactory performance (Shan et al., 2008; Cho and Lee, 2009; Xu and Jia, 2010). This can be viewed as adding additional image penalty terms and trade-off parameters to (3). For example, Shan et al. (2008) incorporates an extra local penalty on the latent image, such that the gradients of small-scale structures in the recovered image are close to those in the blurry image. Thus they will actually contribute less to the subsequent kernel estimation step, allowing larger structures to be captured first. Similarly, a bilateral filtering step is used for pruning out small scale structures in Cho and Lee (2009). Finally, Xu and Jia (2010) develop an empirical structure selection metric designed such that small scale structures can be pruned away by thresholding the corresponding response map, allowing subsequent kernel estimation to be dominated by only large-scale structures.

12. Based on a strong simplifying assumption that the covariance $\mathbf{C}$ from Algorithm 1 is a constant, Babacan et al. (2012) provides some preliminary discussion regarding possibly why VB may be advantageous over MAP. However, when $\mathbf{C}$ is constant, the analysis easily reduces to a standard penalized regression problem, and hence this material can already be found in the sparse estimation literature (e.g., see Palmer et al. 2006; Wipf et al. 2011 and related references). Our key contribution is to explicitly account for the actual dynamic nature of $\mathbf{C}$ and expose the true behavior of VB blind deblurring.

Generally speaking, existing MAP strategies face a trade-off: either they must adopt a highly sparse image prior needed for properly resolving fine structures (see Section 4) and then deal with the attendant constellation of problematic local minima,[13] or rely on a more smooth image prior augmented with compensatory structure-selection measures such as those described above to avoid bad global solutions. In contrast, we may interpret the coupled penalty function intrinsic to VB as a principled alternative with a transparent, integrated functionality for estimation at different resolutions without any additional penalty factors, trade-off parameters, or complexity.

4. Maximal Sparsity vs. Natural Image Statistics

Levin et al. (2009, 2011a,b), which represents the initial inspiration for our work, presents a compelling and highly influential case that joint MAP estimation over $\mathbf{x}$ and $\mathbf{k}$ generally favors a degenerate, no-blur solution, meaning that $\mathbf{k}$ will be a delta function, even when the assumed image prior $p(\mathbf{x})$ reflects the true underlying distribution of $\mathbf{x}$, meaning $p(\mathbf{x}) = p_{\text{true}}(\mathbf{x})$, and $p(\mathbf{k})$ is assumed flat in the feasible region.[14] In turn, this is presented as a primary argument for why MAP is inferior to VB. As this line of reasoning is considerably different from that given in Section 3, here we will take a closer look at these orthogonal perspectives in the hopes of providing a clarifying resolution.

To begin, it helps to revisit the formal analysis of MAP failure from Levin et al. (2011b), where the following specialized scenario is presented. Assume that a blurry image $\mathbf{y}$ is generated by $\mathbf{y} = \mathbf{k}^* * \mathbf{x}^*$, where $\|\mathbf{k}^*\|_2 \ll 1$ and each true sharp image gradient x_i^* is drawn iid from the *generalized Gaussian distribution* $p_{\text{true}}(x_i^*) \sim \exp(-|x_i^*|^p)$, $0 < p \leq 1$. Now consider the minimization problem

$$\min_{\mathbf{x},\mathbf{k}} \sum_i |x_i|^p \quad \text{s.t. } \mathbf{y} = \mathbf{k} * \mathbf{x}. \tag{23}$$

Solving (23) is equivalent to MAP estimation over $\mathbf{x}$ and $\mathbf{k}$ under the true image prior $p_{\text{true}}(\mathbf{x})$ and an implicitly assumed flat prior on $\mathbf{k}$ within the previously specified kernel constraint set. In the limit as the image grows arbitrarily large, (Levin et al., 2011b, Claim 2) proves that the no-blur delta solution $\{\mathbf{x} = \mathbf{y}, \mathbf{k} = \delta\}$ will be favored over the true solution $\{\mathbf{x} = \mathbf{x}^*, \mathbf{k} = \mathbf{k}^*\}$. Intuitively, this occurs because the blurring operator $\mathbf{k}$ contributes two opposing effects:

1. It reduces a measure of the image *sparsity*, which increases $\sum_i |y_i|^p$, and
2. It broadly reduces the overall image *variance*, which reduces $\sum_i |y_i|^p$.

Depending on the relative contributions, we may have the situation where the second effect dominates such that $\sum_i |y_i|^p$ may be less than $\sum_i |x_i^*|^p$, meaning the cost function value

13. Appropriate use of continuation methods such as the algorithm from Chartrand and Yin (2008) may help in this regard.

14. Note that Levin *et al.* frequently use $\text{MAP}_{x,k}$ to refer to joint MAP estimation over both $\mathbf{k}$ and $\mathbf{x}$ (Type I) while using MAP_k for MAP estimation of $\mathbf{k}$ alone after $\mathbf{x}$ has been marginalized out (Type II). In this terminology, MAP_k then represents the inference ideal that VB purports to approximate, equivalent to (4) herein.

at the blurred image is actually lower than at the true, sharp image. Consequently, MAP estimation may not be reliable.

Our conclusions then suggest a sort of paradox: in Section 3 we have argued that VB is actually equivalent to an unconventional form of MAP estimation over $\mathbf{x}$, but with an intrinsic mechanism for avoiding bad local minima, increasing the chances that a good global or near-global minima can be found. Moreover, at least in the noiseless case ($\lambda \to 0$), any such minima will be exactly equivalent to the standard MAP solution by virtue of Corollary 1. However, based on the noiseless analysis from Levin *et al.* above, any global MAP solution is unlikely to involve the true sharp image when the true image statistics are used for $p(\mathbf{x})$, meaning that VB performance should be poor as well even at a global solution. Thus how can we reconcile the positive performance of VB actually observed in practice, and avoidance of degenerate no-blur solutions, with Levin *et al.*'s characterization of the MAP cost function?

First, when analyzing MAP, Levin *et al.* consider only a flat prior on $\mathbf{k}$ within the constraint set $\sum_i k_i = 1$ and $k_i \geq 0$. However, MAP estimation may still avoid no-blur solutions when equipped with an appropriate non-flat kernel prior and associated trade-off parameter. Likewise under certain conditions described in Section 3.5, VB naturally substitutes in a quadratic normalization constraint for $\mathbf{k}$ that we have argued disfavors no-blur solutions automatically. Moreover, VB introduces this normalization in a convenient form devoid of additional tuning parameters.

Secondly, the argument in Levin *et al.* breaks down when the true sharp image $\mathbf{x}^*$ is actually sparse in the canonical sense, meaning the distribution of each element includes a delta function at zero, i.e.,

$$p_{\text{true}}(x_i^*) = \alpha\delta(x_i^*) + (1-\alpha)\rho(\mathbf{x}_i), \tag{24}$$

where ρ is an arbitrary distribution and $\alpha \in [0,1]$ is a constant. Clearly samples from (24) will include some elements exactly equal to zero with probability at least α.

Lemma 1 *Let $\mathbf{x}^*$ be distributed iid with elements drawn from (24) and let $\mathbf{y} = \mathbf{k}^* * \mathbf{x}^*$ for some non-negative kernel $\mathbf{k}^*$. Then with probability approaching one as the image size grows large*

$$\mathbf{k}^*, \mathbf{x}^* = \arg \min_{\mathbf{k},\mathbf{x}:\mathbf{y}=\mathbf{k}*\mathbf{x}} \log p_{\text{true}}(x_i^*) = \arg \min_{\mathbf{k},\mathbf{x}:\mathbf{y}=\mathbf{k}*\mathbf{x}} \|\mathbf{x}\|_0. \tag{25}$$

The proof is straightforward and we do not reproduce it here. Regardless, this result demonstrates that exactly sparse images can in fact be recovered using MAP or equivalently the ℓ_0 norm, the latter of which is actually blind to the distribution of non-zero coefficients $\rho(\mathbf{x}_i)$. Intuitively, this occurs because these measures are entirely immune to changes in variance and only sensitive to sparsity; hence any blurring operation will only increase either penalty function in the feasible region. So immediately we may conclude that, assuming we have some way of solving (25), we should not discount MAP or ℓ_0 minimization as a viable means for recovering sparse images. And importantly, the exact distribution of nonzero coefficients is irrelevant as long as some degree of sparsity exists.

Of course most practical images of interest are not exactly sparse. Rather, a commonly-reported estimate of true image gradient statistics is the generalized Gaussian distribution with $p \approx [0.5, 0.8]$, samples from which will have no exactly zero-valued elements (Buccigrossi and Simoncelli, 1999). In this regime then the original claim from Levin *et al.* will

hold and MAP estimation seems to have been discredited. However, we would argue that MAP can still be salvaged if we are willing to intentionally allow mismatch between the true image prior and the image prior which forms the basis of our MAP estimator. More specifically, we suggest replacing the true image statistics p_{true} with the ℓ_0 norm. However, solving (25) directly will obviously not be effective when there are no exactly zero-valued coefficients.

Fortunately there is a simple way around this. In the regime where $n \approx m$, meaning the sharp and blurry images $\mathbf{y}$ and $\mathbf{x}$ are large relative to the size of $\mathbf{k}$ and therefore of nearly equal dimension, the generalized Gaussian distribution with $p \approx [0.5, 0.8]$ is a *compressible distribution* in the sense described in Cevher (2009). In words, this means that the sorted magnitudes of samples from this distribution exhibit a power-law decay and hence can be well-approximated by sparse signals. Consequently, there will exist some sparse $\hat{\mathbf{x}}$ with $\|\hat{\mathbf{x}}\|_0 \ll m$ such that $\|\mathbf{y} - \mathbf{k}^* * \hat{\mathbf{x}}\|_2^2 < \epsilon$ for some small ϵ. In contrast, each element of the blurry image $\mathbf{y}$ is a summation of many elements of $\mathbf{x}^*$ via the blur operation, and therefore, by central limit theorem arguments each element, while not exactly iid, will approach samples from a Gaussian distribution (exactly so for large enough blur kernels), which is not a compressible distribution. Therefore, if we solve a relaxed version of (25) given by

$$\min_{\mathbf{x},\mathbf{k}} \|\mathbf{x}\|_0, \quad \text{s.t. } \|\mathbf{y} - \mathbf{k} * \mathbf{x}\|_2^2 < \epsilon, \tag{26}$$

with an appropriate choice for ϵ, then we are very likely to obtain the true blur kernel $\mathbf{k}^*$, and a close sparse approximation to $\mathbf{x}^*$. Conversely it is very *unlikely* that the solution will be $\mathbf{x} = \mathbf{y}$ and $\mathbf{k} = \delta$. Therefore, just because $\mathbf{x}^*$ may not be exactly sparse, we may nonetheless locate a sparse approximation $\hat{\mathbf{x}}$ that is sufficiently reasonable such that the unknown $\mathbf{k}^*$ can still be estimated accurately, facilitating a later non-blind, image domain estimation step.

Overall then, the success of the ℓ_0 norm penalty in the context of MAP estimation speaks to the following point: it is more important that *the assumed image prior* $p(\mathbf{x}) = \exp[-\frac{1}{2} g_{\mathbf{x}}(\mathbf{x})]$ *be maximally discriminative with respect to blurred and sharp images, as opposed to accurately reflecting the statistics of real images.* Mathematically, this implies that it is much more important that we have $p(\mathbf{k} * \mathbf{x}^*) \ll p(\hat{\mathbf{x}})$ for some $\hat{\mathbf{x}}$ such that $\mathbf{k} * \mathbf{x}^* \approx \mathbf{k} * \hat{\mathbf{x}}$, than we enforce $p(\mathbf{x}) = p_{\text{true}}(\mathbf{x})$, even if $p_{\text{true}}(\mathbf{x})$ were known exactly. This is because the sparsity/variance trade-off described above implies that it may often be the case that $p_{\text{true}}(\mathbf{k} * \mathbf{x}^*) > p_{\text{true}}(\hat{\mathbf{x}})$ leading to the no-blur solution.

Obviously from a practical standpoint solving (26) represents a difficult, combinatorial optimization problem with numerous local minima. However, to the extent that the VB image penalty g_{VB} approximates the ℓ_0 norm, the VB cost (12) can be viewed as an approximate Lagrangian form of (26), but augmented with an adaptive shape modulation that helps to circumvent these local minima. Thus we can briefly summarize largely why VB can be superior to MAP: *VB allows us to use a near-optimal image penalty, one that is maximally discriminative between blurry and sharp images, but with a reduced risk of getting stuck in bad local minima during the optimization process.* Overall, these conclusions provide a more complete picture of the essential differences between MAP and VB.

Before proceeding to the next section, we emphasize that none of the arguments presented herein discredit the use of natural image statistics when directly solving (4). In fact

Levin et al. (2011b) prove that when $p(\mathbf{x}) = p_{\text{true}}(\mathbf{x})$, then in the limit as the image grows large the MAP estimate for $\mathbf{k}$, after marginalizing over $\mathbf{x}$ (Type II), will equal the true $\mathbf{k}^*$. But there is no inherent contradiction with our results, since it should now be readily apparent that VB is fundamentally different than solving $\min_{\mathbf{k}} p(\mathbf{k}|\mathbf{y})$, and therefore justification for the latter cannot be directly transferred to justification for the former. This highlights the importance of properly differentiating various forms of Bayesian inference, both in the context of blind image deblurring and beyond to widespread application domains.

Natural image statistics are ideal in cases where $\mathbf{y}$ and $\mathbf{x}$ grow large and we are able to integrate out the unknown $\mathbf{x}$, benefiting from central limit arguments when estimating $\mathbf{k}$ alone. However, when we jointly compute MAP estimates of both $\mathbf{x}$ and $\mathbf{k}$ (Type I) as in (3), we enjoy no such asymptotic welfare since the number of unknowns increases proportionally with the sample size. One of the insights of our paper is to show that, at least in this regard, VB is on an exactly equal footing with Type I MAP, and thus we must look for theoretical VB justification elsewhere, leading to the analysis of relative concavity, local minima, invariance, maximal sparsity, etc. presented herein.

5. Learning λ

While existing VB blind deconvolution algorithms typically utilize some pre-assigned decreasing sequence for λ as described in Section 3.4 and noted in Algorithm 1, it is preferable to have λ learned automatically from the data itself as is common in other applications of VB. In the case of blind deblurring, we expect that such a learned λ, with an image-dependent trajectory, may better modulate the penalty curvature discussed in Section 3.4. In contrast, a fixed, pre-defined decreasing sequence is likely to be miscalibrated as it will not reflect the current quality of image and kernel estimates during each iteration. Additionally, the alternative strategy of learning λ has the conceptual appeal of an integrated cost function that is universally reduced even as λ is updated, unlike Algorithm 1 where the λ reduction step may in fact increase the overall cost.

However, current VB deblurring papers either do not mention such a seemingly obvious alternative (perhaps suggesting that the authors unsuccessfully tried such an approach) or explicitly mention that learning λ is problematic but without concrete details. For example, Levin et al. (2011b) observed that the noise level learning used in Fergus et al. (2006) represents a source of problems as the optimization diverges when the estimated noise level decreases too much. But there is no mention of why λ might decrease too much, and further details or analyses are absent.

Interestingly, the perspective presented herein provides some direct insights into how λ may be effectively learned. Consider minimization of the revised VB cost function (12) over $\mathbf{x}$, $\mathbf{k}$, and now λ as well. Because $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^n$ with $m > n$, for a fixed $\mathbf{k}$ there are an infinite number of candidate solutions such that $\mathbf{y} = \mathbf{k} * \mathbf{x}$ since the corresponding null-space of the convolution matrix is of dimension $m - n$. Therefore there exist an infinite set images $\mathbf{x}$ such that the term $\frac{1}{\lambda} \|\mathbf{y} - \mathbf{k} * \mathbf{x}\|_2^2$ in the VB cost function (12) can be minimized to exactly zero even in the limit as $\lambda \to 0$. Included in this set are basic feasible solutions (in the linear programming sense), each of which have at least $m - n$ elements of $\mathbf{x}$ equal to zero.

A problem then arises because it can be shown that $g_{\text{VB}}(0,\rho) \rightarrow -\infty$ as $\rho \rightarrow 0$ for all non-decreasing f.[15] Consequently, we can always trivially drive the VB cost function (12) to $-\infty$ using any basic feasible solution combined with $\lambda \rightarrow 0$, regardless of the quality of the solution. Now because of the disparity in dimensionality mentioned above, there will always be feasible solutions to $\mathbf{y} = \mathbf{k} * \mathbf{x}$ with at least $m - n$ or more elements of $\mathbf{x}$ equal to zero. Thus, at any one of these solutions the VB cost function (12) can then be driven to $-\infty$ with $\lambda \rightarrow 0$. Unless the true $\mathbf{x}$ actually has many exactly zero-valued elements, this will represent a globally degenerate minimizing solution for a broad class of f. And even for other choices for f, a slightly more subdued form of this same degeneracy will still exist since the VB-specific regularization fundamentally favors λ being small: essentially the $\log(\gamma_i + \rho)$ factor in (13) will always favor ρ, and therefore λ being small. The $1/\lambda$ weighting of $\|\mathbf{y} - \mathbf{k} * \mathbf{x}\|_2^2$ is not sufficient for counteracting this effect given the multitude of feasible solutions such that $\mathbf{y} = \mathbf{k} * \mathbf{x}$.

And even for other choices for f, a slightly more subdued form of this same degeneracy will still persist since the existing VB-specific regularization fundamentally favors λ being small: essentially the $\log(\gamma_i + \rho)$ factor in (13) will always favor ρ, and therefore λ being small. The $1/\lambda$ weighting of $\|\mathbf{y} - \mathbf{k} * \mathbf{x}\|_2^2$ is not sufficient for counteracting this effect given the multitude of feasible solutions such that $\mathbf{y} = \mathbf{k} * \mathbf{x}$.

However, these degeneracies can be circumvented with an additional penalty factor on λ that is naturally motivated by this framework. Specifically, we propose to append the penalty function

$$v(\lambda) = (n - m)\log\lambda + \frac{d}{\lambda} \tag{27}$$

to (12), where d is assumed to be a small positive constant. The first term in (27) directly counteracts the degeneracy of basic feasible solutions by providing an equal and opposite barrier to arbitrary solutions with $\lambda \rightarrow 0$ and $\|\mathbf{x}\|_0 = m - n$. Additionally, when f is a constant as we have argued previously represents a well-motivated selection, then it can be shown that this additional penalty represents a very principled approximation to what the true λ penalty should be if the original VB formulation from (6) were not factorized as in (7). Additionally, for other choices of f, (12) can be similarly modified to provide a consistent estimator for λ in the sense described in Wipf and Wu (2012).

As justification for the second term in (27), note that this added factor is proportional to $1/\lambda\,\|\mathbf{y} - \mathbf{k} * \mathbf{x}\|_2^2$, but acts as an interpretable barrier preventing λ from ever going below d/n, which remains a possibility even with the $(n - m)\log\lambda$ term in place. In fact it is easily shown (see Appendix B) that any λ minimizing the cost function (12) augmented with the penalty $v(\lambda)$ must satisfy $\lambda \geq d/n$, which can be viewed as a lower-bound on what $1/n\,\|\mathbf{y} - \mathbf{k} * \mathbf{x}\|_2^2$ should be.[16]

In practice, we have found the fixed value $d = n \times 10^{-4}$ to be highly effective across a wide range of images and testing scenarios, including all reported results in Section 6 and

15. Based on (13), it is clear that the optimizing γ_i value for computing $g_{\text{VB}}(0,\rho)$ will be $\gamma_i = 0$. When $\rho \rightarrow 0$, we then have $\log(\gamma_i + \rho) \rightarrow -\infty$, and therefore $g_{\text{VB}}(0,\rho) \rightarrow -\infty$. Graphically, Figure 1 (b) also reveals this effect, showing that if we were to jointly minimize over both x and ρ, the $\{0,0\}$ solution is heavily favored.

16. While it could be argued that setting d to a larger value could obviate the need for the $(n - m)\log\lambda$ penalty altogether, we would lose considerable interpretability, connection with the original VB problem, and from a practical standpoint, we would likely be saddled with a more sensitive tuning heuristic for d.

Algorithm 2 VB Blind Deblurring with Jeffreys Prior and Learned λ (VB+).

1: **Input:** blurry image $\mathbf{y}$, noise level estimation hyper-parameter $d = n \times 10^{-4}$
2: **Initialize:** blur kernel $\mathbf{k}$, noise level λ
3: **While** stopping criteria is not satisfied, modify ω_i and λ updates from Algorithm 1 with $\omega_i \leftarrow \sigma_i^{-2}, \forall i$ and $\lambda \leftarrow \frac{\|\mathbf{y}-\boldsymbol{\mu}*\mathbf{k}\|_2^2+\sum_i(\|\bar{\mathbf{k}}\|_2^2 \cdot C_{ii})+d}{n}$
4: **End**

all of the real-world, more complex non-uniform deblurring experiments from Zhang and Wipf (2013). Regardless, use of a single, fixed value is likely to be less burdensome than producing an entire λ reduction schedule, which also requires a user-specified minimal λ value anyway. Moreover, the VB update rules only require a slight modification to account for this additional term while retaining existing convergence properties (see Appendix B for the derivation). By estimating the noise level together with the image and kernel, we not only make the deblurring algorithm more noise-aware and mostly parameter-free. But more importantly, by initializing with a large value and allowing the iterations to learn the optimal reduction schedule, it offers a natural coarse-to-fine process for blind deblurring, which has been found as one of the crucial factors for blind deblurring algorithms as discussed above. The experimental results from Section 6 support this conclusion.

6. Experimental Results

We emphasize that the primary purpose of this paper is the formal analysis of existing VB blind deconvolution methodology, not the development and validation of an entirely practical system per se. Some empirical support for recent VB algorithms, complementary to our theoretical presentation, already exist (Babacan et al., 2012; Levin et al., 2011a; Zhang et al., 2013; Zhang and Wipf, 2013). Nonetheless, motivated by our results herein, we will briefly evaluate two simple refinements of Algorithm 1 that help corroborate some of our analytical findings while demonstrating that an extremely simplified version of VB, albeit with theoretically sound underpinnings, can outperform recent published state-of-the-art MAP and VB algorithms with considerably more complexity and/or manual parameters. In doing so, we hope to motivate the optimal usage of VB for more sophisticated and realistic blind deblurring problems.

To this end we will (i) use an image prior obtained when f is flat (Jeffreys prior) as motivated in Section 3 instead of a prior based on natural image statistics, and (ii) we will learn the λ parameter automatically per the discussion in Section 5. The revised estimation steps are summarized in Algorithm 2, which is obtained by adopting the basic procedure from Algorithm 1 under the special case $f(\gamma) = b$ and with the λ updates derived in Appendix B. We will refer to this algorithm as *VB+*. Note that estimation is performed in the gradient domain using the filters described in Section 3.1; however, the recovered kernel is applied to a non-blind deconvolution step to obtain the final latent image estimate. This filtering and final non-blind step, taken from Levin et al. (2011a), is standardized across all algorithms compared in this section, with the exception of approach from Babacan et al. (2012), which uses its own specially-tailored non-blind step.

Given this variant of VB, we reproduce the experiments from Levin et al. (2011a) using the useful benchmark test data from Levin et al. (2009).[17] This consists of 4 base images of size 255×255 and 8 different blurring effects, leading to a total of 32 blurry images. Ground truth blur kernels were estimated by recording the trace of focal reference points on the boundaries of the sharp images (see Levin et al. 2011b, Figure 7 and related text for details of the experimental setup and data collection). The kernel sizes range from 13×13 to 27×27. All evaluations are based on the SSD (Sum of Squared Difference) metric defined in Levin et al. (2009), which quantifies the error between estimated and the ground-truth images. To normalize for the fact that harder kernels give a larger image reconstruction error even when the true kernel is known (because the corresponding non-blind deconvolution problem is also harder), the SSD ratio between the image deconvolved with the estimated kernel and the image deconvolved with the ground-truth kernel is used as the final evaluation measure.

We first compare VB+ as described in Algorithm 2 with the related variational Bayesian methods from Fergus et al. (2006), Levin et al. (2011a), and Babacan et al. (2012), labeled VB-Fergus, VB-Levin, and VB-Babacan respectively. For VB-Fergus and VB-Levin, results directly accompany the Levin *et al.* data set; for VB-Babacan the results are produced using a script provided directly from the first author's website explicitly designed for producing competitive results with the Levin *et al.* data (note that this code contains an additional kernel penalty with added trade-off parameters set by the authors for working with this data set). While all three existing VB algorithms can be effective in practice, they have not been optimized with respect to the considerations provided herein. With VB-Fergus and VB-Levin, the adopted image priors are loosely based on the statistics of natural scenes and, as we have argued in Sections 3 and 4, may not be optimal. While VB-Babacan also involves a prior with f flat like VB+, it adopts the fixed λ reduction schedule originally from VB-Levin, and hence cannot exploit the full potential of VB.

The cumulative histogram of the SSD error ratios is shown in Figure 3(a) for all VB methods. The height of the bar indicates the percentage of images having error ratio below that level. High bars indicate better performance. As mentioned by Levin *et al.*, the results with error ratios above 2 may already have some visually implausible regions (Levin et al., 2009). VB+ can achieve close to 90% success with error ratio below 2, significantly higher than the others.

Regardless, all of the VB algorithms still exhibit reasonable performance, especially given that they do not benefit from any additional prior information or regularization heuristics that facilitate blur-adaptive structure selection (meaning the additional regularization based on domain knowledge added to Equation 3 that boost typical MAP algorithms as discussed previously). However, one curious phenomenon is that both VB-Fergus and VB-Levin experience a relatively large drop-off in performance when the error ratio reduces from 1.5 to 1.1. While it is difficult to be absolutely certain, one very plausible explanation for this decline relates to the prior selection employed by these algorithms. In both cases, the prior is based on a finite mixture of zero mean Gaussians with different variances roughly matched to natural image statistics. While such a prior does heavily favor approximately sparse signals, it will never produce any exactly sparse estimates at any resolution of the course-to-fine hierarchy, and hence, especially at high resolutions the penalty shape modu-

17. This data is available online at `http://www.wisdom.weizmann.ac.il/~levina/papers/LevinEtalCVPR09Data.rar`

lation effect of VB will be highly muted, as will be the beneficial sparsity/variance trade-off that accompanies more strongly sparse priors. Thus these algorithms may not be optimal for resolving extremely fine details, which is required for reliably producing image estimates with low error ratios. In contrast, to achieve high error ratios only lower resolution features need be resolved, and in this regime VB-Levin, which is also based directly on Algorithm 1, performs nearly as well as VB+. Also note that VB-Babacan performs relatively poorly at the higher error ratios, which is likely attributable to the fact that local minima and more catastrophic errors are a serious problem when using the highly non-convex Jeffreys distribution without a proper schedule for reducing λ that is tuned to the image prior.

We next compare VB+ with several state-of-the-art MAP algorithms from Shan et al. (2008), Xu and Jia (2010), and Cho and Lee (2009). Shan *et al.* (denoted MAP-Shan) adopts an additional local smoothness prior designed to reduce ringing artifacts. Xu *et al.* (MAP-Xu) includes two phases for kernel estimation and incorporates an explicit scheme for edge structure selection. Finally, Cho *et al.* (MAP-Cho) is also a carefully-engineered MAP approach coupled with structure selection and sharp edge prediction schemes, which help the algorithm to avoid the degenerate delta solution. Recall that previously we have argued that standard MAP algorithms may suffer from one of two problems: either the pixel-wise image prior is highly sparse and convergence to sub-optimal local solutions becomes a problem, or the prior is less sparse and global solutions do not sufficiently distinguish blurry from sharp images. All of the MAP algorithms tested here can be viewed as addressing this conundrum by including additional regularization schemes (priors) such that global or near global minima favor sharp images even when the basic pixel-wise image prior is convex (i.e., minimally sparse). This is a very different strategy than VB, which adopts a simpler underlying model with no additional regularizers beyond the canonical pixel-wise sparse prior. Figure 3(b) reveals that the simple VB strategy, when properly implemented, can still outperform specially tuned MAP estimates. Note that the results of MAP-Cho are from the data set accompanying Levin et al. (2011a) directly, while the results of MAP-Shan and MAP-Xu are produced using the software provided by the authors, for which we adjust the parameters carefully. For all algorithms we run every test image with the same parameters, similar to Levin et al. (2009, 2011b). Overall, VB+ obtains the highest reported result of any existing VB or MAP algorithm on this important benchmark.

7. Conclusion

This paper presents an insightful reformulation and subsequent analysis of MAP and VB blind deconvolution algorithms revealing why practical success is possible and suggesting valuable improvements for the latter. We summarize the contributions of this perspective as follows:

- Levin et al. (2009, 2011b) have provided an interesting analysis of VB and related MAP algorithms. We push the limits of understanding further, demonstrating that rigorous evaluation of VB and its associated priors cannot be separated from implementation heuristics, and we have meticulously examined the interplay of the relevant underlying algorithmic details employed by practical VB systems. Consequently, what may initially appear to be a plausible rationale for achieving high performance may

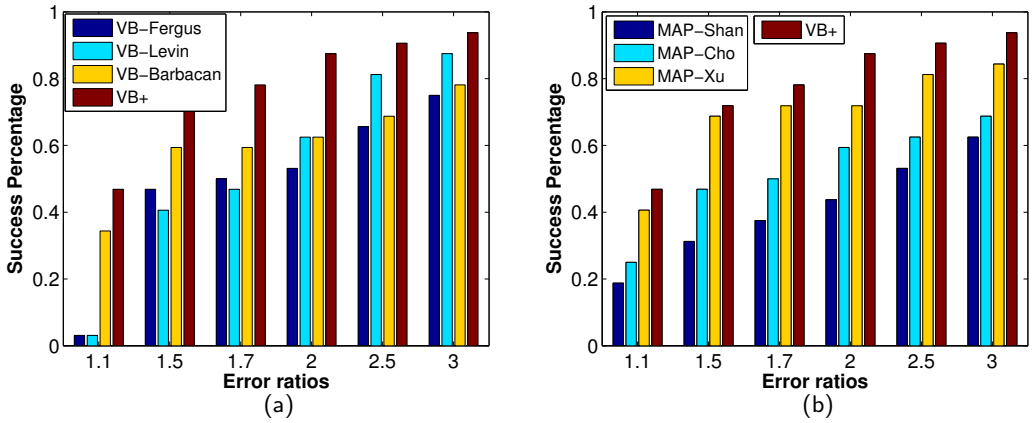

Figure 3: Evaluation of the restoration results: Cumulative histogram of the deconvolution error ratio across 32 test examples. The height of the bar indicates the percentage of images having error ratio below that level. High bars indicate better performance. (a) comparison with several other VB algorithms. (b) comparison with several state-of-the-art MAP algorithms.

have limited applicability given the assumptions required to implement scalable versions of VB.

- We have proven that in an ideal, noiseless setting, VB and MAP have an identical underlying cost function once the requisite approximations are accounted for (and they are intimately related even when noise is present). This is in direct contrast to conventional assumptions explaining the presumed performance advantages of VB.

- We carefully examine the underlying VB objective function in a transparent form, leading to principled criteria for choosing the optimal image prior. It is crucial to emphasize that this image prior need not, and generally should not, reflect the most accurate statistics of real imaging data. Instead, the preferred distribution is one that is most likely to guide VB iterations to high quality global solutions by strongly differentiating between blurry and sharp images. In this context, we have motivated a unique selection, out of the infinite set of possible sparse image priors, that simultaneously allows for maximal discrimination between $\mathbf{k}*\mathbf{x}$ and $\mathbf{x}$, displays a desirable form of scale invariance, and leads to an intrinsic coupling between the blur kernel, noise level, and image penalty such that bad local minima can largely be avoided. To the best of our knowledge, this represents a completely new viewpoint for understanding VB algorithms.

- The cause of failure when using standard MAP algorithms depends on the choice of image prior. If $-2\log p(\mathbf{x})$ is only marginally concave in $|\mathbf{x}|$, or is tuned to natural image statistics, then the problem is often that global or near-global solutions do not properly differentiate blurry from sharp images. In contrast, if $p(\mathbf{x})$ is highly sparse, while global solutions may be optimally selective for sharp image gradients,

convergence to bad local solutions is more-or-less inevitable. It is with the latter that VB offers a compelling advantage.

- We have derived and analyzed a simple yet powerful extension of VB deconvolution algorithms for learning the noise level.
- By reframing VB as a nearly parameter-free sparse regression problem in standard form, we demonstrate that it is no longer difficult to enhance performance and generality by inheriting additional penalty functions (such as those from Shan et al. 2008) or noise models (e.g., Laplacian, Poisson, etc.) commonly reserved for MAP. Moreover, we anticipate that these contributions will lead to a wider range of principled VB applications, such as non-uniform deconvolution (Whyte et al., 2012; Zhu and Milanfar, 2013) and multi-frame and video deblurring (Sroubek and Milanfar, 2012; Takeda and Milanfar, 2011). Preliminary results show tremendous promise (Zhang et al., 2013; Zhang and Wipf, 2013). Additionally, the analysis we conducted for blind deconvolution may well be relevant to other related bilinear models like robust dictionary learning in the presence of noise.

Overall, we hope that these observations will ensure that VB is not under-utilized in blind deconvolution and related tasks. We conclude by mentioning that, given the new perspective on VB provided herein, it may be possible to derive new blind deblurring algorithms and penalty functions that deviate from the VB script but nonetheless adopt some of its attractive properties. This is a direction of ongoing research.

8. Acknowledgements

We would especially like to thank Yi Ma and Aaron Hertzmann for providing some insightful comments on an earlier version of this manuscript, and to the reviewers for several useful suggestions.

Appendix A. Proofs

This section provides proofs of various technical results described in this paper.

A.1 Proof of Theorem 1

We begin with the cost function

$$\mathcal{L}(\mathbf{x}, \mathbf{k}, \boldsymbol{\gamma}) \triangleq \frac{1}{\lambda}\|\mathbf{y} - \mathbf{k} * \mathbf{x}\|_2^2 + \sum_i \left[\frac{x_i^2}{\gamma_i} + \log(\lambda + \|\bar{\mathbf{k}}\|_2^2\gamma_i) + f(\gamma_i)\right], \tag{28}$$

which is obtained starting with (12) and then simply removing the minimization over $\boldsymbol{\gamma}$ from the definition of g_{VB} in (13), plugging in the value of ρ, and simplifying. The basic strategy here will be to use a majorization-minimization approach (Hunter and Lange, 2004) akin to the concave-convex procedure (Yuille and Rangarajan, 2001) to derive coordinate-wise updates that are guaranteed to reduce or leave unchanged $\mathcal{L}(\mathbf{x}, \mathbf{k}, \boldsymbol{\gamma})$, and then show that these are in fact the same updates as Algorithm 1. In doing so we show that (12) is an equally valid explanatory cost function with which to interpret VB.

As an initial proposition, we may attempt to directly minimize $\mathcal{L}(\mathbf{x}, \mathbf{k}, \boldsymbol{\gamma})$ over $\mathbf{x}$, $\mathbf{k}$, and $\boldsymbol{\gamma}$ independently, in each case while holding the other two variables fixed. Beginning with $\mathbf{x}$, we collect relevant terms and find that we must solve

$$\min_{\mathbf{x}} \frac{1}{\lambda}\|\mathbf{y} - \mathbf{k} * \mathbf{x}\|_2^2 + \sum_i \frac{x_i^2}{\gamma_i}, \tag{29}$$

which has a convenient closed-form solution $\mathbf{x}^{\text{opt}}$ given by

$$\mathbf{x}^{\text{opt}} = \left[\frac{1}{\lambda}\mathbf{H}^T\mathbf{H} + \boldsymbol{\Gamma}^{-1}\right]^{-1} \frac{1}{\lambda}\mathbf{H}^T\mathbf{y}, \tag{30}$$

where $\boldsymbol{\Gamma} \triangleq \text{diag}[\boldsymbol{\gamma}]$ and $\mathbf{H}$ is the convolution matrix of the blur kernel defined in Section 3.1.

Next we consider updating $\boldsymbol{\gamma}$, where the associated cost function conveniently decouples so we may solve for each γ_i independently. For this purpose, we use the fact that

$$\lambda + \|\bar{\mathbf{k}}\|_2^2\gamma_i = \lambda\gamma_i\left(\frac{1}{\gamma_i} + \frac{\|\bar{\mathbf{k}}\|_2^2}{\lambda}\right) \tag{31}$$

to obtain the following minimization problem for each γ_i:

$$\min_{\gamma_i \geq 0} \frac{x_i^2}{\gamma_i} + \log\gamma_i + \log\left[\frac{\|\bar{\mathbf{k}}\|_2^2}{\lambda} + \gamma_i^{-1}\right] + f(\gamma_i), \tag{32}$$

where γ_i-independent terms are omitted. Because no closed-form solution is available, we instead use basic principles from convex analysis to form a strict upper bound that will facilitate subsequent optimization. In particular, we use

$$\frac{z_i}{\gamma_i} - \phi^*(z_i) \geq \log\left[\frac{\|\bar{\mathbf{k}}\|_2^2}{\lambda} + \gamma_i^{-1}\right], \tag{33}$$

which holds for all $z_i \geq 0$, where ϕ^* is the concave conjugate (Boyd and Vandenberghe, 2004) of the concave function $\phi(\alpha) \triangleq \log\left[\frac{\|\bar{\mathbf{k}}\|_2^2}{\lambda} + \alpha\right]$. It can be shown that equality in (33) is obtained using

$$z_i^{\text{opt}} = \frac{\partial\phi}{\partial\alpha}\bigg|_{\alpha=\gamma_i^{-1}} = \frac{1}{\frac{\sum_j k_j^2 \bar{I}_{ji}}{\lambda} + \gamma_i^{-1}}, \forall i, \tag{34}$$

where we have used the fact that $\|\bar{\mathbf{k}}\|_2^2 \triangleq \sum_j k_j^2 \bar{I}_{ji}$ is the squared norm of $\mathbf{k}$ reincorporating the i-dependent image boundary conditions (see Section 3.1), which will become somewhat relevant for a more comprehensive version of the proof. Plugging (33) into (32) we obtain the revised problem

$$\min_{\gamma_i \geq 0} \frac{x_i^2 + z_i}{\gamma_i} + \log\gamma_i + f(\gamma_i). \tag{35}$$

This sub-problem can be handled in multiple ways. First, if the underlying g_x associated with f (obtained from (11)) is differentiable, then (35) has a convenient closed-form solution obtained as follows. After a $\exp[-1/2(\cdot)]$ transformation (35) assumes the same variational

form as the sparse prior given by (11) evaluated at the point $\sqrt{x_i^2 + z_i}$, ignoring irrelevant constants. Consequently, based on Palmer et al. (2006) we know that the optimizing γ_i is given by

$$\gamma_i^{\text{opt}} = \left.\frac{2\sigma}{g_x{}'(\sigma)}\right|_{\sigma=\sqrt{x_i^2+z_i}}, \forall i. \tag{36}$$

This covers the vast majority of practical sparse priors (and all of those amenable to Algorithm 1). Secondly, if for some reason g_x is not differentiable at some point(s), then (35) may still be solved numerically as a 1D optimization problem, or perhaps analytically leveraging the structure of f. For example, if f is a non-decreasing function (as motivated in Section 3.3), then g_x will not be differentiable at zero. However, since $\gamma_i^{\text{opt}} = 0$ whenever $x_i^2 + z_i = 0$, so this does not pose a problem.

We now examine optimization over $\mathbf{k}$. Isolating terms, this requires that we solve

$$\min_{\mathbf{k}\geq 0} \frac{1}{\lambda}\|\mathbf{y} - \mathbf{k} * \mathbf{x}\|_2^2 + \sum_i \log\left[\frac{\|\bar{\mathbf{k}}\|_2^2}{\lambda} + \gamma_i^{-1}\right]. \tag{37}$$

There is no closed-form solution; however, as before we may use strict upper bounds derived from convex analysis for optimization purposes. Accounting again for the fact that $\|\bar{\mathbf{k}}\|_2^2 \triangleq \sum_j k_j^2 \bar{I}_{ji}$ actually depends on i, we choose

$$\Big(\sum_j k_j^2 \bar{I}_{ji}\Big) v_i - \varphi_i^*(v_i) \geq \log\Big[\frac{1}{\lambda}\Big(\sum_j k_j^2 \bar{I}_{ji}\Big) + \gamma_i^{-1}\Big], \tag{38}$$

which holds for all $v_i \geq 0$, where φ^* is the concave conjugate of the concave function $\varphi_i(\alpha) \triangleq \log\left[\frac{\alpha}{\lambda} + \gamma_i^{-1}\right]$. Similar to the $\boldsymbol{\gamma}$ updates from above, it can be shown that equality in (38) is obtained with the minimizing v_i given by

$$v_i^{\text{opt}} = \left.\frac{\partial \varphi_i}{\partial \alpha}\right|_{\alpha=\sum_j k_j^2 \bar{I}_{ji}} = \frac{z_i}{\lambda}, \forall i. \tag{39}$$

Plugging (38) and (39) into (37) leads to the quadratic optimization problem

$$\mathbf{k}^{\text{opt}} = \arg\min_{\mathbf{k}\geq 0} \frac{1}{\lambda}\|\mathbf{y}-\mathbf{W}\mathbf{k}\|_2^2 + \sum_i \frac{z_i}{\lambda}\left(\sum_j k_j^2 \bar{I}_{ji}\right) = \arg\min_{\mathbf{k}\geq 0} \|\mathbf{y}-\mathbf{W}\mathbf{k}\|_2^2 + \sum_j k_j^2 \left(\sum_i z_i \bar{I}_{ji}\right), \tag{40}$$

where $\mathbf{W}$ is the convolution matrix constructed from the image $\mathbf{x}$ (see Section 3.1). As a simple convex program, there exist many high-performance algorithms for solving (40).

To review, we would originally like to minimize $\mathcal{L}(\mathbf{x}, \mathbf{k}, \boldsymbol{\gamma})$ over $\mathbf{x}$, $\mathbf{k}$, and the latent variables $\boldsymbol{\gamma}$. To simplify the optimization we introduce additional latent variables $\mathbf{z} \triangleq [z_1, \ldots, z_m]^T$ and $\mathbf{v} \triangleq [v_1, \ldots, v_m]^T$, such that, after combining terms from above we are now equivalently minimizing

$$\begin{aligned}\mathcal{L}(\mathbf{x}, \mathbf{k}, \boldsymbol{\gamma}, \mathbf{z}, \mathbf{v}) \triangleq{}& \frac{1}{\lambda}\|\mathbf{y} - \mathbf{k} * \mathbf{x}\|_2^2 \\ &+ \sum_i \left[\frac{x_i^2 + z_i}{\gamma_i} + \log\gamma_i + f(\gamma_i) - \phi^*(z_i)\right] + \sum_i \left[\sum_j \left(k_j^2 \bar{I}_{ji}\right) v_i - \varphi_i^*(v_i)\right]\end{aligned} \tag{41}$$

over $\mathbf{x}$, $\mathbf{k}$, and the latent variables $\boldsymbol{\gamma}$, $\mathbf{z}$, and $\mathbf{v}$. The associated coordinate descent updates rules, meaning the cyclic iteration of (30), (34), (36), (39), and (40), are guaranteed to reduce or leave unchanged $\mathcal{L}(\mathbf{x}, \mathbf{k}, \boldsymbol{\gamma})$ by standard properties of majorization-minimization algorithms. And importantly, at least for our purposes, these updates are in one-to-one correspondence with those from Algorithm 1, albeit with some inconsequential differences in notation and statistical interpretation. Specifically, the $\boldsymbol{\gamma}$ update from (36) is equivalent to the $\boldsymbol{\omega}$ update in Algorithm 1, the $\mathbf{x}$ update from (30) is equivalent to the $\boldsymbol{\mu}$ update, the $\mathbf{z}$ update becomes equivalent to computing the diagonal of $\mathbf{C}$, and finally the $\mathbf{k}$ update from (40) is the same as that in Algorithm 1 but with the requisite boundary conditions explicitly incorporated via $\bar{\mathbf{I}}$.

Note that the $\boldsymbol{\omega}$ update from Algorithm 1 appears somewhat different from that originally presented in Levin et al. (2011a), which only considers the special case where the assumed image prior is a finite Gaussian scale mixture given by

$$p(x_i) = \sum_j \frac{\pi_j}{\sqrt{2\pi\bar{\gamma}_j}} \exp\left[-\frac{1}{2}\frac{x_i^2}{\bar{\gamma}_j}\right], \tag{42}$$

where $\pi_j \geq 0$ and $\sum_j \pi_j = 1$. However, using Palmer et al. (2006) it is easily shown that

$$\frac{2\sigma}{g_x{}'(\sigma)} = \left(\mathrm{E}_{p(\gamma|x_i=\sigma)}[\gamma^{-1}]\right)^{-1} = \frac{\sum_j \frac{\pi_j}{\sqrt{2\pi\bar{\gamma}_j}} \exp\left[-\frac{1}{2}\frac{\sigma^2}{\bar{\gamma}_j}\right]}{\sum_j \frac{\pi_j}{\sqrt{2\pi\bar{\gamma}_j}} \exp\left[-\frac{1}{2}\frac{\sigma^2}{\bar{\gamma}_j}\right]\frac{1}{\bar{\gamma}_j}} \tag{43}$$

such that formal equivalence with Levin et al. (2011a) is maintained.

In closing, we emphasize that the upper bounds utilized here were specifically chosen so as to establish a connection with Algorithm 1. However, once we have motivated that $\mathcal{L}(\mathbf{x}, \mathbf{k}, \boldsymbol{\gamma})$ is an equally valid cost function, other bounds can be used to potentially improve the convergence rate or other properties of the algorithm. This is a direction of future research. ■

A.2 Proof of Corollary 1

Here we omit the pixel-wise subscript i for simplicity. Likewise for later proofs where appropriate. From the definition of g_{VB} we know that $g_{\text{VB}}(x, 0) = \min_{\gamma \geq 0} \frac{x^2}{\gamma} + \log(\gamma) + f(\gamma)$. After a $-2\log$ transformation of (11), and ignoring constant terms, we have $g_x(x) = -2\log p(x) = \min_{\gamma \geq 0} \frac{x^2}{\gamma} + \log(\gamma) + f(\gamma)$, and so it follows that $g_{\text{VB}}(x, 0) = g_x(x)$. ■

A.3 Proof of Theorem 2

We first assume that f is a concave, non-decreasing function and express $g_{\text{VB}}(x, \rho)$ as

$$g_{\text{VB}}(x, \rho) \triangleq \min_{\gamma \geq 0} \frac{x^2}{\gamma} + \psi(\gamma), \tag{44}$$

where $\psi(\gamma) \triangleq \log(\rho+\gamma) + f(\gamma)$ is also a concave, non-decreasing function of γ (because $\log(\rho+\gamma)$ is). Thus we can always express $\psi(\gamma)$ as

$$\psi(\gamma) = \min_{z \geq 0} z\gamma - \psi^*(z), \tag{45}$$

where $\psi^*(z)$ is the concave conjugate (Boyd and Vandenberghe, 2004) of $\psi(\gamma)$. Therefore, it follows that

$$g_{\text{VB}}(x,\rho) = \min_{\gamma,z \geq 0} \frac{x^2}{\gamma} + z\gamma - \psi^*(z). \tag{46}$$

Optimizing over γ for fixed x and z, the optimal solution is

$$\gamma^{\text{opt}} = z^{-1/2}|x|. \tag{47}$$

Plugging this result into (46) gives

$$g_{\text{VB}}(x,\rho) = \min_{z \geq 0} \frac{x^2}{z^{-1/2}|x|} + zz^{-1/2}|x| - \psi^*(z) = \min_{z \geq 0} 2z^{1/2}|x| - \psi^*(z). \tag{48}$$

This implies that $g_{\text{VB}}(x,\rho)$ can be expressed as a minimum over upper-bounding hyperplanes in $|x|$, with different z implying different slopes. Any function expressible in this form is necessarily concave, and also non-decreasing since $z \geq 0$ (Boyd and Vandenberghe, 2004).

Now in the other direction, assume that $g_{\text{VB}}(x,\rho)$ is a concave, non-decreasing function of $|x|$. It then follows that

$$g_{\text{VB}}(x,\rho) = \min_{z \geq 0} 2z|x| + h(z) \tag{49}$$

for some function h. Using the fact that

$$2|x| = \min_{\alpha \geq 0} \frac{x^2}{\alpha} + \alpha \tag{50}$$

and defining $\gamma \triangleq \alpha z^{-1}$, we can re-express $g_{\text{VB}}(x,\rho)$ as

$$\begin{aligned} g_{\text{VB}}(x,\rho) &= \min_{\alpha,z \geq 0} z\left[\frac{x^2}{\alpha} + \alpha\right] + h(z) = \min_{\alpha,z \geq 0} \frac{x^2}{\alpha z^{-1}} + z\alpha + h(z) \\ &= \min_{\gamma,z \geq 0} \frac{x^2}{\gamma} + z^2\gamma + h(z) = \min_{\gamma \geq 0} \frac{x^2}{\gamma} + \varphi(\gamma), \end{aligned} \tag{51}$$

where $\varphi(\gamma) \triangleq \min_{z \geq 0} z^2\gamma + h(z)$ is necessarily a concave, non-decreasing function of γ by construction and arguments made previously. This implies that $\psi(\gamma)$ from (44) must be a concave, non-decreasing function of γ for all ρ. Of course as $\rho \to \infty$, $\log(z+\rho)$ becomes arbitrarily flat, with derivative approaching zero for all γ. Consequently, the only way to ensure that $\psi(\gamma)$ is concave and non-decreasing for any ρ is to require that f is a concave, non-decreasing function.

Finally, any locally minimizing solution $\mathbf{x}^{\text{opt}}$ to (12) must necessarily be a local minimum to

$$\min_{\mathbf{x}} \frac{1}{\lambda}\|\mathbf{y} - \mathbf{H}\mathbf{x}\|_2^2 + \sum_i g_{\text{VB}}(x_i,\rho). \tag{52}$$

If f is concave and non-decreasing, then so is $g_{\text{VB}}(x_i, \rho)$ based on the arguments presented above, and so (52) is a canonical sparse estimation problem with a separable concave in $|\mathbf{x}|$ regularizer. Based on (Rao et al., 2003, Theorem 1), we may then conclude that $m - n$ elements of $\mathbf{x}^{\text{opt}}$ will be zero at any local minimizer. ■

A.4 Proof of Corollaries 2 and 3

For Corollary 2, the proof in both directions follows from similar arguments to those used for proving Theorem 2. For Corollary 3, the proof follows from several modifications of the proof of Theorem 2 from Zhang et al. (2013). We omit details for the sake of brevity. ■

A.5 Proof of Theorem 3

For $f(\gamma) = b$, we have

$$g_{\text{VB}}(x, \rho) \equiv \min_{\gamma \geq 0} \underbrace{\frac{x^2}{\gamma} + \log(\rho + \gamma)}_{\varphi} \tag{53}$$

since constant terms are irrelevant. We first calculate the optimal γ by differentiating φ and equating terms to zero. Since

$$\frac{\partial \varphi}{\partial \gamma} = -\frac{x^2}{\gamma^2} + \frac{1}{\rho + \gamma}, \tag{54}$$

it follows after some algebra that

$$\gamma^{\text{opt}} = \frac{x^2 + |x|\sqrt{x^2 + 4\rho}}{2}. \tag{55}$$

Based on the unimodality of φ it follows that γ^{opt} represents the unique minimizer. Substituting (55) into (53) and omitting irrelevant constant factors, we have

$$g_{\text{VB}}(x, \rho) \equiv \frac{2|x|}{|x| + \sqrt{x^2 + 4\rho}} + \log\left(2\rho + x^2 + |x|\sqrt{x^2 + 4\rho}\right).$$

■

A.6 Proof of Corollary 4

Assuming $f(\gamma) = b$ and $\rho_1 < \rho_2$, we want to show that $g_{\text{VB}}^{\rho_1} \prec g_{\text{VB}}^{\rho_2}$. For this purpose it is sufficient to show that $\frac{\partial^2 g_{\text{VB}}^{\rho}(x)}{\partial x^2} / \frac{\partial g_{\text{VB}}^{\rho}(x)}{\partial x}$ is an increasing function of ρ, which represents an equivalent condition for relatively concavity to one given by Definition 1, assuming the requisite derivatives exist (Palmer, 2003).

Defining $\eta \triangleq \gamma^{-1}$, we have

$$g_{\text{VB}}^{\rho}(x) = \min_{\eta \geq 0} \eta x^2 + \log(\rho + \eta^{-1}) \tag{56}$$

where the optimal η^{opt} is given by the gradient of $g^{\rho}_{\text{VB}}(x)$ with respect to x^2, which follows from basic concave duality theory. Let $h^{\rho}(z) \triangleq g^{\rho}_{\text{VB}}(\sqrt{z})$. Then $\eta^{\text{opt}} = \frac{\partial h^{\rho}(z)}{\partial z}$. With $z \triangleq x^2$, we can readily compute the expression for $\frac{\partial g^{\rho}_{\text{VB}}}{\partial x}(x)$ via

$$\frac{\partial g^{\rho}_{\text{VB}}(x)}{\partial x} = \frac{\partial h^{\rho}(z)}{\partial z}\frac{dz}{dx} = 2x\frac{\partial h^{\rho}(z)}{\partial z} = \frac{x}{\rho}\left(\sqrt{1+\frac{4\rho}{x^2}}-1\right). \tag{57}$$

Using (57) it is also straightforward to derive $\frac{\partial^2 g^{\rho}_{\text{VB}}(x)}{\partial x^2}$ as

$$\frac{\partial^2 g^{\rho}_{\text{VB}}(x)}{\partial x^2} = 2\frac{\partial h^{\rho}(z)}{\partial z} - \frac{4}{x^2\sqrt{1+\frac{4\rho}{x^2}}}. \tag{58}$$

We must then show that

$$\frac{\partial^2 g^{\rho}_{\text{VB}}(x)/\partial x^2}{\partial g^{\rho}_{\text{VB}}(x)/\partial x} = \frac{1}{x} - \frac{\frac{4}{x^2\sqrt{1+\frac{4\rho}{x^2}}}}{\frac{x}{\rho}\left(\sqrt{1+\frac{4\rho}{x^2}}-1\right)} \tag{59}$$

is an increasing function of ρ. By neglecting irrelevant additive and multiplicative factors (and recall that $x \geq 0$ from the definition of g^{ρ}_{VB}), this is equivalent to showing that

$$\xi(\rho) = \frac{1}{\rho}\left(\sqrt{1+\frac{4\rho}{x^2}}-1\right) \tag{60}$$

is a decreasing function of ρ. It is easy to check that

$$\xi'(\rho) = \frac{\sqrt{1+\frac{4\rho}{x^2}}-1-\frac{2\rho}{x^2}}{\sqrt{1+\frac{4\rho}{x^2}}} < 0. \tag{61}$$

Therefore, $\xi(\rho)$ is a decreasing function of ρ, implying that $\frac{\partial^2 g^{\rho}_{\text{VB}}(x)}{\partial x^2}/\frac{\partial g^{\rho}_{\text{VB}}(x)}{\partial x}$ is an increasing function of ρ, completing the proof. ■

A.7 Proof of Theorem 4

For simplicity assume that f is twice differentiable. From the definition of relative concavity, $\psi^{\rho_1} \prec \psi^{\rho_2}$ if and only if $\frac{\partial^2 \psi^{\rho}(\gamma)}{\partial \gamma^2}/\frac{\partial \psi^{\rho}(\gamma)}{\partial \gamma}$ is an increasing function of ρ (Palmer, 2003). It is easy to show that

$$\xi(\rho) \triangleq \frac{\partial^2 \psi^{\rho}(\gamma)}{\partial \gamma^2}/\frac{\partial \psi^{\rho}(\gamma)}{\partial \gamma} = \frac{-\frac{1}{(\gamma+\rho)^2}+f''(\gamma)}{\frac{1}{\gamma+\rho}+f'(\gamma)}. \tag{62}$$

To avoid notation clutter, we let $\omega \triangleq \gamma+\rho$, so that the objective is then to prove that

$$\xi(\rho) = \frac{-\frac{1}{\omega^2}+f''(\gamma)}{\frac{1}{\omega}+f'(\gamma)} \tag{63}$$

is an increasing function of ρ, for all $\gamma, \rho \geq 0$ if and only if $f''(\gamma) = 0$ and $f'(\gamma) \geq 0$, or equivalently that f is affine with positive slope. For this purpose it suffices to examine conditions whereby

$$\xi'(\rho) = \frac{f''(\gamma)\omega^2 + 2f'(\gamma)\omega + 1}{\left(f'(\gamma)\omega^2 + \omega\right)^2} \geq 0, \forall \rho, \gamma \geq 0. \tag{64}$$

First, assume $f''(\gamma) = 0$. We also have that $f'(\gamma) \geq 0$ by virtue of the Theorem statement. Clearly (64) will always be true and so $\xi(\rho)$ must be an increasing function of ρ. In the other direction, assume that (64) is true for all ρ and γ. Because f is a concave function, $f''(\gamma) \leq 0$. Now consider the case where $f''(\gamma) < 0$. The denominator of (64) is always non-negative and can be ignored. For the numerator, allow ρ to become arbitrarily large while keeping γ fixed. The quadratic term will then dominate such that $\xi'(\gamma) < 0$, violating our assumption that $\xi'(\rho) \geq 0$. Therefore it must be that $f''(\gamma) = 0$.

To conclude, $\psi^{\rho_1} \prec \psi^{\rho_2}$ if and only if $f''(\gamma) = 0$ and $f'(\gamma) \geq 0$, which is equivalent to the requirement that $f(\gamma) = a\gamma + b$ with $a \geq 0$. ■

A.8 Proof of Theorem 5

Consider the VB cost function (12) with g_{VB} defined via (13). Given an optimal solution pair $\{\mathbf{x}^*, \mathbf{k}^*\}$, we equivalently want to prove that $\{\alpha^{-1}\mathbf{x}^*, \alpha\mathbf{k}^*\}$ is also always an optimal solution pair if and only if $f(\gamma_i) = b$.

First we assume that f is a constant. It is easy to see that the value of the data fidelity term in (12) is unchanged since

$$\frac{1}{\lambda}\left\|\mathbf{y} - \mathbf{k}^* * \mathbf{x}^*\right\|_2^2 \equiv \frac{1}{\lambda}\left\|\mathbf{y} - \alpha\mathbf{k}^* * \frac{\mathbf{x}^*}{\alpha}\right\|_2^2. \tag{65}$$

For the penalty terms, after defining $\bar{\gamma}_i \triangleq \alpha^2\gamma_i$ for each i and $\rho^* \triangleq \lambda/\|\bar{\mathbf{k}}^*\|_2^2$, we have

$$\begin{aligned}
g_{\text{VB}}\left(\frac{x_i^*}{\alpha}, \frac{\rho^*}{\alpha^2}\right) + \log\left(\alpha^2\left\|\bar{\mathbf{k}}^*\right\|_2^2\right) &= \min_{\gamma_i \geq 0} \frac{x_i^{*2}}{\alpha^2\gamma_i} + \log\left(\frac{\rho^*}{\alpha^2} + \gamma_i\right) + \log\left(\alpha^2\left\|\bar{\mathbf{k}}^*\right\|_2^2\right) \\
&= \min_{\bar{\gamma}_i \geq 0} \frac{x_i^{*2}}{\bar{\gamma}_i} + \log\left(\frac{\rho^*}{\alpha^2} + \frac{\bar{\gamma}_i}{\alpha^2}\right) + \log\alpha^2 + \log\left\|\bar{\mathbf{k}}^*\right\|_2^2 \\
&= \min_{\bar{\gamma}_i \geq 0} \frac{x_i^{*2}}{\bar{\gamma}_i} + \log(\rho^* + \bar{\gamma}_i) + \log\|\bar{\mathbf{k}}^*\|_2^2 \\
&\equiv g_{\text{VB}}(x_i^*, \rho^*) + \log\left(\|\bar{\mathbf{k}}^*\|_2^2\right),
\end{aligned} \tag{66}$$

Therefore, the rescaled solution pair $\{\alpha^{-1}\mathbf{x}^*, \alpha\mathbf{k}^*\}$ does not change the cost function value, and must therefore also represent an optimal solution.

On the other hand, assume that $\{\alpha^{-1}\mathbf{x}^*, \alpha\mathbf{k}^*\}$ is an optimal solution for any $\alpha > 0$, from which it must follow that

$$g_{\text{VB}}\left(\frac{x_i^*}{\alpha}, \frac{\rho}{\alpha^2}\right) + \log(\alpha^2\|\bar{\mathbf{k}}^*\|_2^2) = g_{\text{VB}}(x_i^*, \rho) + \log(\|\bar{\mathbf{k}}^*\|_2^2)$$

and therefore

$$\min_{\bar{\gamma}_i \geq 0} \frac{x_i^{*2}}{\bar{\gamma}_i} + \log(\rho + \bar{\gamma}_i) + \log \|\bar{\mathbf{k}}\|_2^2 + f\left(\frac{\bar{\gamma}_i}{\alpha^2}\right) = \min_{\gamma_i \geq 0} \frac{x_i^{*2}}{\gamma_i} + \log(\rho + \gamma_i) + \log \|\bar{\mathbf{k}}\|_2^2 + f(\gamma_i). \quad (67)$$

To satisfy the above equivalence for all possible $\mathbf{x}^*$, $\mathbf{k}^*$, and λ, f must be a constant (with the exception of an irrelevant, zero-measure discontinuity at zero). The second part of the theorem is likewise straightforward to show; however, we omit additional details. ■

Appendix B. Noise Level Estimation

As introduced in Section 5, we would like to minimize the VB cost function (12) after the inclusion of an additional λ-dependent penalty. This is tantamount to solving

$$\min_{\lambda \geq 0} \frac{1}{\lambda}\left[d + \|\mathbf{y} - \mathbf{k} * \mathbf{x}\|_2^2\right] + n \log \lambda + \sum_i \log\left(\frac{\|\bar{\mathbf{k}}\|_2^2}{\lambda} + \gamma_i^{-1}\right),$$

where VB factors irrelevant to λ estimation have been omitted. We set $d = n \times 10^{-4}$ for all simulations which leads to good performance. While there is no closed-form, minimizing solution for λ, similar to the $\boldsymbol{\gamma}$ updates described in the proof of Theorem 1, we may utilize a convenient upper bound for optimization purposes. Here we use

$$\frac{\theta}{\lambda} - \phi^*(\theta) \geq \sum_i \log\left(\frac{\|\bar{\mathbf{k}}\|_2^2}{\lambda} + \gamma_i^{-1}\right) \quad (68)$$

where ϕ^* is the concave conjugate of $\phi(\theta) \triangleq \sum_i \log\left(\theta \|\bar{\mathbf{k}}\|_2^2 + \gamma_i^{-1}\right)$. Equality is obtained with

$$\theta^{\text{opt}} = \left.\frac{\partial \phi}{\partial \theta}\right|_{\theta = \lambda^{-1}} = \sum_i \frac{\|\bar{\mathbf{k}}\|_2^2}{\frac{\|\bar{\mathbf{k}}\|_2^2}{\lambda} + \gamma_i^{-1}}. \quad (69)$$

To optimize over λ, we may iteratively solve

$$\min_{\lambda, \theta \geq 0} \frac{1}{\lambda}\left(\|\mathbf{y} - \mathbf{k} * \mathbf{x}\|_2^2 + d\right) + n \log \lambda + \frac{1}{\lambda}\theta - \phi^*(\theta). \quad (70)$$

For fixed θ, the minimizing λ is easily computed as

$$\lambda^{\text{opt}} = \frac{\|\mathbf{y} - \mathbf{k} * \mathbf{x}\|_2^2 + \theta + d}{n}, \quad (71)$$

where λ^{opt} has a lower bound of d/n. Thus we may set d so as to reflect some expectation regarding the minimal about of noise or modeling error. In practice, these updates can be merged into Algorithm 1 without disrupting the convergence properties (see Algorithm 2).

References

S.D. Babacan, R. Molina, M.N. Do, and A.K. Katsaggelos. Bayesian blind deconvolution with general sparse image priors. In *ECCV*, 2012.

C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2004.

R.W. Buccigrossi and E.P. Simoncelli. Image compression via joint statistical characterization in the wavelet domain. *IEEE Trans. Image Processing*, pages 1688–1701, 1999.

V. Cevher. Learning with compressible priors. In *NIPS*, 2009.

R. Chartrand and W. Yin. Iteratively reweighted algorithms for compressive sensing. In *ICASSP*, 2008.

S. Cho and S. Lee. Fast motion deblurring. In *ACM SIGGRAPH ASIA*, 2009.

T.S. Cho, C.L. Zitnick, N. Joshi, S.B. Kang, R. Szeliski, and W.T. Freeman. Image restoration by matching gradient distributions. *IEEE Trans. PAMI*, 34(4):683–694, 2012.

R. Fergus, B. Singh, A. Hertzmann, S.T. Roweis, and W.T. Freeman. Removing camera shake from a single photograph. In *ACM SIGGRAPH*, 2006.

D.R. Hunter and K. Lange. A tutorial on MM algorithms. *Amer. Statist*, 58(1):30–37, 2004.

A. Hyvarinen and E. Oja. Independent component analysis: Algorithms and application. *Neural Networks*, (4-5):411–430, 2000.

T. Kenig, Z. Kam, and A. Feuer. Blind image deconvolution using machine learning for three-dimensional microscopy. *IEEE Trans. PAMI*, 32(12):2191–2204, 2010.

D. Krishnan and R. Fergus. Fast image deconvolution using hyper-Laplacian priors. In *NIPS*, 2009.

D. Krishnan, T. Tay, and R. Fergus. Blind deconvolution using a normalized sparsity measure. In *CVPR*, 2011.

D. Kundur and D. Hatzinakos. Blind image deconvolution. *IEEE Signal Process. Mag.*, 13 (3):43–64, 1996.

D.D. Lee and H.S. Seung. Algorithms for non-negative matrix factorization. In *NIPS*, 2001.

A. Levin, R. Fergus, F. Durand, and W. T. Freeman. Deconvolution using natural image priors. Technical report, MIT, 2007.

A. Levin, Y. Weiss, F. Durand, and W.T. Freeman. Understanding and evaluating blind deconvolution algorithms. In *CVPR*, 2009.

A. Levin, Y. Weiss, F. Durand, and W.T. Freeman. Efficient marginal likelihood optimization in blind deconvolution. In *CVPR*, 2011a.

A. Levin, Y. Weiss, F. Durand, and W.T. Freeman. Understanding blind deconvolution algorithms. *IEEE Trans. PAMI*, 33(12):2354–2367, 2011b.

L.B. Lucy. An iterative technique for the rectification of observed distributions. *The Astronomical Journal*, 79:745–754, 1974.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *J. Machine Learning Research*, pages 19–60, 2010.

J.W. Miskin and D.J.C. MacKay. Ensemble learning for blind image separation and deconvolution. In *Advances in Independent Component Analysis*, 2000.

J.A. Palmer. Relative convexity. Technical report, UCSD, 2003.

J.A. Palmer, D.P. Wipf, K. Kreutz-Delgado, and B.D. Rao. Variational EM algorithms for non-Gaussian latent variable models. In *NIPS*, 2006.

B.D. Rao, K. Engan, S.F. Cotter, J.A. Palmer, and K. Kreutz-Delgado. Subset selection in noise based on diversity measure minimization. *IEEE Trans. Signal Processing*, 51(3): 760–770, 2003.

H.W. Richardson. Bayesian-Based Iterative Method of Image Restoration. *J. Optical Society of America*, 62(1):55–59, 1972.

S. Roth and M.J. Black. Fields of experts. *Int. J. Computer Vision*, 82(2):205–229, 2009.

Q. Shan, J. Jia, and A. Agarwala. High-quality motion deblurring from a single image. In *ACM SIGGRAPH*, 2008.

F. Sroubek and P. Milanfar. Robust multichannel blind deconvolution via fast alternating minimization. *IEEE Trans. Image Processing*, 21(4):1687–1700, 2012.

H. Takeda and P. Milanfar. Removing motion blur with space-time processing. *IEEE Trans. Image Processing*, 20(10):2990–3000, 2011.

O. Whyte, J. Sivic, A. Zisserman, and J. Ponce. Non-uniform deblurring for shaken images. *Int. J. Computer Vision*, 98(2):168–186, 2012.

D.P. Wipf and Y. Wu. Dual-space analysis of the sparse linear model. In *NIPS*, 2012.

D.P. Wipf, B.D. Rao, and S.S. Nagarajan. Latent variable Bayesian models for promoting sparsity. *IEEE Trans. Info Theory*, 57(9):6236–6255, 2011.

L. Xu and J. Jia. Two-phase kernel estimation for robust motion deblurring. In *ECCV*, 2010.

L. Xu, S. Zheng, and J. Jia. Unnatural L0 sparse representation for natural image deblurring. In *CVPR*, 2013.

A.L. Yuille and A. Rangarajan. The concave-convex procedure (CCCP). In *NIPS*, 2001.

H. Zhang and D.P. Wipf. Non-uniform camera shake removal using a spatially-adaptive sparse penalty. In *NIPS*, 2013.

H. Zhang, D.P. Wipf, and Y. Zhang. Multi-image blind deblurring using a coupled adaptive sparse prior. In *CVPR*, 2013.

X. Zhu and P. Milanfar. Removing atmospheric turbulence via space-invariant deconvolution. *IEEE Trans. PAMI*, 35(1):157–170, 2013.