Abstract—We propose a data-dependent denoising procedure to restore noisy images. Different from existing denoising algorithms which search for patches from either the noisy image or a generic database, the new algorithm finds patches from a database that contains only relevant patches. We formulate the denoising problem as an optimal filter design problem and make two contributions. First, we determine the basis function of the denoising filter by solving a group sparsity minimization problem. The optimization formulation generalizes existing denoising algorithms and offers systematic analysis of the performance. Improvement methods are proposed to enhance the patch search process. Second, we determine the spectral coefficients of the denoising filter by considering a localized Bayesian prior. The new algorithm over existing methods.

Index Terms—Patch-based filtering, image denoising, external database, optimal filter, non-local means, BM3D, group sparsity, Bayesian estimation.

I. INTRODUCTION

A. Patch-based Denoising

Image denoising is a classical signal recovery problem where the goal is to restore a clean image from its observations. Although image denoising has been studied for decades, the problem remains a fundamental one as it is the test bed for a variety of image processing tasks.

Among the numerous contributions in image denoising, the most highly-regarded class of methods, to date, is the class of patch-based image denoising algorithms [1–9]. The idea of a patch-based denoising algorithm is simple: Given a $\sqrt{d} \times \sqrt{d}$ patch $q \in \mathbb{R}^d$ from the noisy image, the algorithm finds a set of reference patches $p_1, \ldots, p_k \in \mathbb{R}^d$ and applies some linear (or non-linear) function $\Phi$ to obtain an estimate $\hat{p}$ of the unknown clean patch $p$ as

$$\hat{p} = \Phi(q; p_1, \ldots, p_k).$$

For example, in non-local means (NLM) [1], $\Phi$ is a weighted average of the reference patches, whereas in BM3D [3], $\Phi$ is a transform-shrinkage operation.

B. Internal vs External Denoising

For any patch-based denoising algorithm, the denoising performance is intimately related to the reference patches $p_1, \ldots, p_k$. Typically, there are two sources of these patches: the noisy image itself and an external database of patches. The former is known as internal denoising [10], whereas the latter is known as external denoising [11–12].

Internal denoising is practically more popular than external denoising because it is computationally less expensive. Moreover, internal denoising does not require a training stage, hence making it free of training bias. Furthermore, Glasner et al. [13] showed that patches tend to recur within an image, e.g., at a different location, orientation, or scale. Thus searching for patches in the noisy image is often a plausible approach. However, on the downside, internal denoising often fails for rare patches — patches that seldom recur in an image. This phenomenon is known as “rare patch effect”, and is widely regarded as a bottleneck of internal denoising [14–15]. Therefore, while there is still room for patch-based algorithms to improve, continuing to focus on internal denoising is a challenging task.

External denoising [6–16] is an alternative solution to internal denoising. Levin et al. [15–20] showed that in the limit, the theoretical minimum mean squared error of denoising is achievable using an infinitely large external database. Recently, Chan et al. [18–19] developed a computationally efficient sampling scheme to reduce the complexity and demonstrated practical usage of large databases. However, these existing external denoising algorithms only consider generic databases [6, 16–17]. The databases, although large in volume, do not necessarily contain useful information to denoise the noisy image of interest. For example, a database of natural images is not helpful to denoise a noisy portrait image.

C. Adaptive Image Denoising

In this paper, we propose an adaptive image denoising algorithm using a targeted external database instead of a generic database. Here, a targeted database refers to a database that contains images relevant to the noisy image only. As will be illustrated in later parts of this paper, targeted external databases could be obtained in many practical scenarios, such as text images (e.g., newspapers and documents), human faces (under certain conditions), and images captured by multiview camera systems. Other possible scenarios include images of license plates, medical CT and MRI images, and images of landmarks.

The concept of using targeted external databases has been proposed in various occasions, e.g., [21–26]. However, none of these methods are tailored for image denoising problems.
The objective of this paper is to bridge the gap by addressing the following question:

Suppose we are given a targeted external database, how should we design a denoising algorithm which can maximally utilize the database?

Here, we assume that the reference patches \( p_1, \ldots, p_k \) are given. We emphasize that this assumption is application specific — for the examples we mentioned earlier (e.g., text, multiview, face, etc), the assumption is typically true because these images have relatively less variety in content.

When the reference patches are given, the above question perhaps becomes a simple one: We can extend existing internal multi-image denoising algorithms, e.g., [1, 3, 5, 27], so that the patches are searched from a database instead of the noisy image. Likewise, one can also treat an external database as a “video” and feed the data to multi-image denoising algorithms, e.g., [28, 31]. However, the problem of these approaches is that the brute force modifications are heuristic. There is no theoretical guarantee of performance.

An alternative response to the above question is to train a statistical prior of the targeted database, e.g., [6, 16, 17, 32, 54]. The merit of this approach is that the performance often has theoretical guarantee because the denoising problem can now be formulated as a maximum a posteriori (MAP) estimation. However, the drawback is that many of these methods require a large number of training samples which is not always available in practice.

D. Contributions and Organization

In view of the above seemingly easy yet challenging question, we introduced a new denoising algorithm using targeted external databases in [35]. Compared to existing methods, the method proposed in [35] achieves better performance and only requires a small number of external images. In this paper, we extend [35] by offering the following new contributions:

1) Generalization of Existing Methods. We propose a generalized framework which encapsulates a number of denoising algorithms. In particular, we show (in Section III-B) that the proposed group sparsity minimization generalizes both fixed basis and PCA methods. We also show (in Section IV-B) that the proposed local Bayesian MSE solution is a generalization of many spectral operations in existing methods.

2) Improvement Strategies. We propose two improvement strategies for the generalized denoising framework. In Section III-C, we present a patch selection optimization to improve the patch search process. In Section IV-C, we present a soft-thresholding and a hard-thresholding method to improve the spectral coefficients learned by the algorithm.

3) Detailed Proofs. Proofs of the results in this paper and [35] are presented in the Appendix.

The rest of the paper is organized as follows. After outlining the design framework in Section II, we present the above contributions in Section III – IV. Experimental results are discussed in Section V, and concluding remarks are given in Section VI.

II. OPTIMAL LINEAR DENOISING FILTER

The foundation of our proposed method is the classical optimal linear denoising filter design problem [36]. In this section, we give a brief review of the design framework and highlight its limitations.

A. Optimal Filter

The design of an optimal denoising filter can be posed as follows: Given a noisy patch \( q \in \mathbb{R}^d \), and assuming that the noise is i.i.d. Gaussian with zero mean and variance \( \sigma^2 \), we want to find a linear operator \( A \in \mathbb{R}^{d \times d} \) such that the estimate \( \hat{p} = Aq \) has the minimum mean squared error (MSE) compared to the ground truth \( p \in \mathbb{R}^d \). That is, we want to solve the optimization

\[
A = \arg\min_A \mathbb{E} \left[ \|Aq - p\|_2^2 \right].
\]

Here, we assume that \( A \) is symmetric, or otherwise the Sinkhorn-Knopp iteration [37] can be used to symmetrize \( A \). Because \( A \) is symmetric, one can apply the eigen-decomposition, \( A = U\Lambda U^T \), where \( U = [u_1, \ldots, u_d] \in \mathbb{R}^{d \times d} \) is the basis matrix and \( \Lambda = \text{diag} \{\lambda_1, \ldots, \lambda_d\} \in \mathbb{R}^{d \times d} \) is the diagonal matrix containing the spectral coefficients. With \( U \) and \( \Lambda \), the optimization problem in (2) becomes

\[
(U, \Lambda) = \arg\min_{U, \Lambda} \mathbb{E} \left[ \|U\Lambda U^T q - p\|_2^2 \right],
\]

subject to the constraint that \( U \) is an orthonormal matrix.

The joint optimization (3) can be solved by noting the following Lemma.

Lemma 1: Let \( u_i \) be the \( i \)-th column of the matrix \( U \), and \( \lambda_i \) be the \( (i, i) \)-th entry of the diagonal matrix \( \Lambda \). If \( q = p + \eta \), where \( \eta \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2 I) \), then

\[
\mathbb{E} \left[ \|U\Lambda U^T q - p\|_2^2 \right] = \sum_{i=1}^{d} \left( 1 - \lambda_i \right)^2 (u_i^T p)^2 + \sigma^2 \lambda_i^2.
\]

The proof of Lemma 1 is given in [38]. With Lemma 1, the denoised patch as a consequence of (3) is as follows.

Lemma 2: The denoised patch \( \hat{p} \) using the optimal \( U \) and \( \Lambda \) of (3) is

\[
\hat{p} = U \left( \text{diag} \left\{ \frac{\|p\|_2^2}{\|p\|_2^2 + \sigma^2}, 0, \ldots, 0 \right\} \right) U^T q,
\]

where \( U \) is any orthonormal matrix with the first column \( u_1 = p/\|p\|_2 \), and \( \text{diag} \{\cdot\} \) denotes the diagonalization operator.

Proof: See Appendix A

Lemma 2 states that if hypothetically we are given the ground truth \( p \), the optimal denoising process is to first project the noisy observation \( q \) onto the subspace spanned by \( p \), perform a Wiener shrinkage \( \|p\|_2^2/(\|p\|_2^2 + \sigma^2) \), and then re-project the shrinkage coefficients to obtain the denoised estimate. However, since in reality we never have access to the ground truth \( p \), this optimal result is not achievable.
B. Problem Statement

Since the oracle optimal filter is not achievable in practice, the question becomes whether it is possible to find a surrogate solution that does not require the ground truth \( p \).

To answer this question, it is helpful to separate the joint optimization \( (3) \) by first fixing \( U \) and minimize the MSE with respect to \( \Lambda \). In this case, one can show that \( (4) \) achieves the minimum when

\[
\lambda_i = \frac{(u_i^T p)^2}{(u_i^T p)^2 + \sigma^2},
\]

in which the minimum MSE estimator is given by

\[
\hat{p} = U \left( \text{diag} \left\{ \frac{(u_1^T p)^2}{(u_1^T p)^2 + \sigma^2}, \ldots, \frac{(u_d^T p)^2}{(u_d^T p)^2 + \sigma^2} \right\} \right) U^T q,
\]

where \( \{u_1, \ldots, u_d\} \) are the columns of \( U \).

Inspecting \( (6) \), we identify two parts of the problem:

1) Determine \( U \). The choice of \( U \) plays a critical role in the denoising performance. In literature, \( U \) are typically chosen as the FFT or the DCT bases \([3, 4]\). In \([5, 7, 8]\), the PCA bases of various data matrices are proposed. However, the optimality of these bases is not fully understood.

2) Determine \( \Lambda \). Even if \( U \) is fixed, the optimal \( \Lambda \) in \( (5) \) still depends on the unknown ground truth \( p \). In \([3]\), \( \Lambda \) is determined by hard-thresholding a stack of DCT coefficients or applying an empirical Wiener filter constructed from a first-pass estimate. In \([4]\), \( \Lambda \) is formed by the PCA coefficients of a set of relevant noisy patches. Again, it is unclear which of these is optimal.

Motivated by the problems about \( U \) and \( \Lambda \), in the following two sections we present our proposed method for each of these problems. We discuss its relationship to prior works, and present ways to further improve it.

III. DETERMINE \( U \)

In this section we present our proposed method to determine the basis matrix \( U \) and show that it is a generalization of a number of existing denoising algorithms. We also discuss ways to improve \( U \).

A. Database Reduction by \( k \) Nearest Neighbors

Our first task to determine the basis matrix \( U \) is to reduce the database. This step is necessary because while all patches in the database come from images that contain similar content to the noisy image, a big portion of the patches are still not useful to denoise a particular noisy patch.

The database reduction is performed by selecting the \( k \) most similar patches to the noisy patch (i.e., the one to be denoised). The similarity is measured based on the \( \ell_2 \) distance between the noisy patch \( q \) and the database \( \{p_j\}_{j=1}^n \), where \( n > k \), as

\[
d(q, p_j) = \|q - p_j\|_2, \quad \text{for } j = 1, \ldots, n.
\]

Effectively, this amounts to searching \( k \) nearest neighbors (kNN) from a set of \( n \) data points.

The kNN procedure is effective to select a subset of the database. However, it has the drawback that under the \( \ell_2 \) distance, some of the \( k \) selected patches could be irrelevant. We will address this issue more thoroughly in Section III-D by discussing methods to improve the robustness of the kNN.

B. Group Sparsity

Given \( \{p_j\}_{j=1}^k \) from the kNN search, we postulate that a good projection matrix \( U \) should be the one that makes the projected vectors \( \{U^T p_j\}_{j=1}^k \) similar in both magnitude and location. This hypothesis follows from the observation that since \( \{p_j\}_{j=1}^k \) have small \( \ell_2 \) distances from \( q \), it must hold that any \( p_i \) and \( p_j \) (hence \( U^T p_i \) and \( U^T p_j \)) in the set should also be similar.

In addition to being self-similar, we require each projected vector \( U^T p_j \) to contain as few non-zero entries as possible, i.e., sparse. The reason is related to the shrinkage step to be discussed in Section IV because a vector of few non-zero coefficients has higher energy concentration and hence is more effective for denoising.

In order to satisfy these two requirements, we propose to consider the idea of group sparsity\(^1\), which is characterized by the matrix \( \ell_{1,2} \) norm, defined as

\[
\|X\|_{1,2} \overset{\text{def}}{=} \sum_{i=1}^d \|x_i\|_2,
\]

for any matrix \( X \in \mathbb{R}^{d \times k} \), where \( x_i \) is the \( i \)th row of a matrix \( X \). In words, a small \( \|X\|_{1,2} \) makes sure that \( X \) has few non-zero entries, and the non-zero entries are located similarly in each column. A pictorial illustration is shown in Figure 1.

![Fig. 1: Comparison between sparsity (where columns are sparse, but do not coordinate) and group sparsity (where all columns are sparse with similar locations).](image)

Going back to our problem, we propose to minimize the \( \ell_{1,2} \)-norm of the matrix \( U^T P \):

\[
\begin{align*}
\text{minimize} & \quad \|U^T P\|_{1,2} \\
\text{subject to} & \quad U^T U = I,
\end{align*}
\]

where \( P \overset{\text{def}}{=} [p_1, \ldots, p_k] \). Here, the equality constraint ensures that \( U \) is orthonormal. Thus, the solution of \( (8) \) is the projection matrix that generates the most group sparse vector.

An interesting fact of this problem is that the solution is identical to the classical principal component analysis (PCA) result, which is given in the following lemma.

\(^1\)Group sparsity was first proposed by Cotter et al. for group sparse reconstruction \([59]\) and later used by Mairal et al. for denoising \([60]\), but towards a different end from the method presented in this paper.
Lemma 3: The solution to (8) is that
\[
[U, S] = \text{eig}(PP^T),
\]
where \(S\) is the corresponding eigenvalue matrix.

Proof: See Appendix [3]

Remark 1: In practice, we note that the \(k\) reference patches might have deviations in terms of similarity with \(q\). Thus, we improve the data matrix \(P\) by introducing a diagonal weight matrix
\[
W = \frac{1}{Z} \text{diag} \left\{ e^{-\|q-p_1\|^2/h^2}, \ldots, e^{-\|q-p_k\|^2/h^2} \right\},
\]
for some user tunable parameter \(h\) and normalization constant \(Z \equiv 1^T W 1\). In this case, (8) becomes
\[
\begin{align*}
\text{minimize} & \quad \|U^T PW^{1/2}\|_{1,2} \\
\text{subject to} & \quad U^T U = I,
\end{align*}
\]
of which the solution is given by
\[
[U, S] = \text{eig}(PW^T P^T).
\] (12)

C. Relationship to Prior Works

The fact that (12) is the solution to a group sparsity minimization problem allows us to understand the performance of a number of existing denoising algorithms to some extent.

1) BM3D [3]: It is perhaps a misconception that the underlying principle of BM3D is to enforce sparsity of the 3-dimensional data volume (which we shall call it a 3-way tensor). However, what BM3D enforces is the group sparsity of the slices of the tensor, not the sparsity of the tensor.

To see this, we note that the 3-dimensional transforms in BM3D are separable (e.g., DCT2 + Haar in its default setting). Thus, unless the reference patches \(p_1, \ldots, p_k\) are highly dissimilar, the DCT2 coefficients will be similar in both magnitude and location. That means if we fix a location and trace the DCT2 coefficients along the third axis, the signal we observe is almost flat. Hence, applying the Haar transform returns a sparse vector. Clearly, such sparsity is based on the stationarity of the DCT2 coefficients along the third axis. In essence, this is equivalent to being group sparse.

2) HOSVD [9]: The true tensor sparsity can only be utilized by the high order singular value decomposition (HOSVD), which is recently studied in [9]. Let \(P \in \mathbb{R}^{d_k \times d_k \times k}\) be the tensor by stacking the patches \(p_1, \ldots, p_k\) into a 3-dimensional array. HOSVD seeks three orthonormal matrices \(U^{(1)} \in \mathbb{R}^{d_k \times d_k \times d_k}, U^{(2)} \in \mathbb{R}^{d_k \times d_k \times d_k}, U^{(3)} \in \mathbb{R}^{k \times k}\) and an array \(S \in \mathbb{R}^{d_k \times d_k \times k}\), such that
\[
S = P \times_1 U^{(1)^T} \times_2 U^{(2)^T} \times_3 U^{(3)^T},
\]
where \(\times_k\) denotes a tensor mode-\(k\) multiplication [40].

HOSVD ignores the fact that image patches tend to be group sparse instead of being tensor sparse. Consequently, its performance is worse than BM3D, as we observe in [9].

3) Shape-adaptive BM3D [11]: As a variation of BM3D, SA-BM3D groups similar patches according to a mask defined by \(q\). The mask modifies the standard \(\ell_2\) distance between patches to a weighted \(\ell_2\) distance by masking out irrelevant sub-regions.

Shape-adaptive BM3D can be easily generalized in our proposed framework by defining an additional weight matrix \(W_s \in \mathbb{R}^{d_k \times d_k}\) (where the subscript \(s\) denotes a spatial weight) and consider the weighted data
\[
\mathbf{P} = W_s^{1/2} PW_s^{1/2},
\]
where \(W \in \mathbb{R}^{k \times k}\) is defined in (10). Here the matrix \(W_s\) is used to control the relative emphasis of each pixel in the spatial coordinate. For the rest of the paper, we let \(W_s = I\) to improve computational efficiency.

4) BM3D-PCA [5] and LPG-PCA [7]: The idea of both BM3D-PCA and LPG-PCA is that given \(p_1, \ldots, p_k\), \(U\) is determined as the principal components of \(P = [p_1, \ldots, p_k]\). Incidentally, such approaches arrive at the same result as (12), i.e., the principal components are indeed the solution of a group sparse minimization. However, the key of using the group sparsity is not noticed in [5] and [7]. This provides additional theoretical justifications for both methods.

D. Improvement: Patch Selection Refinement

The optimization problem (11) suggests that the \(U\) computed from (12) is the optimal basis with respect to the reference patches \([p_j]_{j=1}^{k}\). However, one issue that remains is how to improve the selection of the \(k\) patches.

1) Patch Selection as Linear Programming: To facilitate the discussion of our proposed scheme, it is useful to revisit the kNN search from an optimization perspective.

It is not difficult to see that the kNN search can be formulated as the following optimization problem
\[
\begin{align*}
\text{minimize} & \quad \sum_{j=1}^{n} x_j \|q - p_j\|_2 \\
\text{subject to} & \quad \sum_{j=1}^{n} x_j = k.
\end{align*}
\]
(13)

Here, the optimization variables \(x_j \in \{0, 1\}\) form a vector of indicators. If \(x_j = 1\), then the corresponding \(p_j\) should be selected. Thus, by minimizing \(\sum_{j=1}^{n} x_j \|q - p_j\|_2\) we obtain the \(k\) nearest neighbors of \(q\).

Problem (13) is a combinatorial problem as it seeks for one out of the \(\binom{n}{k}\) configurations that minimizes the objective. However, a close inspection reveals that such a combinatorial search is unnecessary. In fact, (13) is equivalent to a relaxed convex optimization (more specifically, a linear programming problem)
\[
\begin{align*}
\text{minimize} & \quad c^T x \\
\text{subject to} & \quad x^T 1 = k, \quad 0 \leq x \leq 1,
\end{align*}
\]
(14)

where we define \(c = [c_1, \ldots, c_k]^T\) with \(c_j \equiv \|q - p_j\|_2\). To see the equivalence between (13) and (14), we consider a simple case where \(n = 2\) and \(k = 1\). In this case, the constraints \(x^T 1 = k\) and \(0 \leq x \leq 1\) form a closed line segment in the positive quadrant. Since the objective function \(c^T x\) is linear, the optimal point must be at one of the vertices...
Thus, for computational efficiency we choose similar to sum using (which could be problematic due to the noise present in patches should not be determined merely from regularized optimization.

The difference is that the patch in the set. In practice, we find that patch matching results: (a) ground truth, (b) 10 best reference patches using \( \varphi(x) = 1^T B x \) (where \( \tau = 1/(2n) \)), (c) 10 best reference patches using \( \varphi(x) = 1^T B x \) (where \( \tau = 1 \)).

Remark 2: Because the optimal solution must be a vertex of the polytope defined by the constraints \( x^T 1 = k \) and \( 0 \leq x \leq 1 \). \( (14) \) can be solved efficiently by locating the smallest entries in \( e \), eliminating the need of an iterative linear programming solver.

2) Regularization by Cross Similarity: Our proposed patch selection scheme is to modify \( (14) \) by adding an appropriate penalty term to the objective function:

\[
\begin{align*}
\text{minimize} & \quad c^T x + \tau \varphi(x) \\
\text{subject to} & \quad x^T 1 = k, \quad 0 \leq x \leq 1, \quad (15)
\end{align*}
\]

where \( \varphi(x) \) is the penalty function and \( \tau > 0 \) is a parameter. In this paper we present two possible choices of \( \varphi(x) \).

The first choice of \( \varphi(x) \) is to consider \( \varphi(x) = x^T B x \), where \( B \in \mathbb{R}^{n \times n} \) is a symmetric matrix with \( B_{ij} \equiv \| p_i - p_j \|_2 \). Writing \( (15) \) explicitly, we see that the optimization problem \( (13) \) becomes

\[
\begin{align*}
\text{minimize} & \quad \sum_{i,j \leq 1, i \neq j} x_j \| q - p_j \|_2 + \tau \sum_{i,j \leq 1, i \neq j} x_i x_j \| p_i - p_j \|_2 \cdot (16)
\end{align*}
\]

The penalized problem \( (16) \) suggests that the optimal \( k \) reference patches should not be determined merely from \( \| q - p_j \|_2 \) (which could be problematic due to the noise present in \( q \)). Instead, a good reference patch should also be similar to all other patches that are selected. The cross similarity term \( x_i x_j \| p_i - p_j \|_2 \) provides a way for such measure. This shares some similarities to the patch ordering concept proposed by Cohen and Elad. \( (27) \). The difference is that the patch ordering proposed in \( (27) \) is a shortest path problem that tries to organize the noisy patches, whereas ours is to solve a regularized optimization.

Problem \( (16) \) is in general not convex because the matrix \( B \) is not positive semidefinite. One way to relax the formulation is to consider \( \varphi(x) = 1^T B x \). Geometrically, the solution of using \( \varphi(x) = 1^T B x \) tends to identify patches that are close to \( \sum \) of all other patches in the set. In many cases, this is similar to \( \varphi(x) = x^T B x \) which finds patches that are similar to every individual patch in the set. In practice, we find that the difference between \( \varphi(x) = x^T B x \) and \( \varphi(x) = 1^T B x \) is not significant. Thus, for computational efficiency we choose \( \varphi(x) = 1^T B x \).
In this section we present our proposed method to determine \( \Lambda \) for a fixed \( U \). Our proposed method is based on the concept of a Bayesian MSE estimator.

**A. Bayesian MSE Estimator**

Recall that the noisy patch is related to the latent clean patch as \( q = p + \eta \), where \( \eta \sim N(0, \sigma^2 I) \) denotes the noise. Therefore, the conditional distribution of \( q \) given \( p \) is

\[
f(q | p) = N(p, \sigma^2 I).
\]

Assuming that the prior distribution \( f(p) \) is known, it is natural to consider the Bayesian mean squared error (BMSE) between the estimate \( \hat{p} \equiv U \Lambda U^T q \) and the ground truth \( p \):

\[
\text{BMSE} \equiv \mathbb{E}_p \left[ \mathbb{E}_{q|p} \left[ \| \hat{p} - p \|_2^2 \mid p \right] \right]. \tag{19}
\]

Here, the subscripts remark the distributions under which the expectations are taken.

The BMSE defined in (19) suggests that the optimal \( \Lambda \) should be the minimizer of the optimization problem

\[
\Lambda = \arg \min_{\Lambda} \mathbb{E}_p \left[ \mathbb{E}_{q|p} \left[ \| U \Lambda U^T q - p \|_2^2 \mid p \right] \right]. \tag{20}
\]

In the next subsection we discuss how to solve (20).

**B. Localized Prior from the Targeted Database**

Minimizing BMSE over \( \Lambda \) involves knowing the prior distribution \( f(p) \). However, in general, the exact form of \( f(p) \) is never known. This leads to many popular models in the literature, e.g., Gaussian mixture model \( [34] \), the field of expert model \( [33, 41] \), and the expected patch log-likelihood model (EPLL) \( [17, 42] \).

One common issue of all these models is that the prior \( f(p) \) is built from a generic database of patches. In other words, the \( f(p) \) models all patches in the database. As a result, \( f(p) \) is often a high dimensional distribution with complicated shapes.

In our targeted database setting, the difficult prior modeling becomes a much simpler task. The reason is that while the shape of the distribution \( f(p) \) is still unknown, the subsampled reference patches (which are few but highly representative) could be well approximated as samples drawn from a single Gaussian centered around some mean \( \mu \) and covariance \( \Sigma \). Therefore, by appropriately estimating \( \mu \) and \( \Sigma \) of this localized prior, we can derive the optimal \( \Lambda \) as given by the following Lemma:

**Lemma 4**: Let \( f(q | p) = N(p, \sigma^2 I) \), and let \( f(p) = N(\mu, \Sigma) \) for any vector \( \mu \) and matrix \( \Sigma \), then the optimal \( \Lambda \) that minimizes (19) is

\[
\Lambda = \frac{\text{diag} \left\{ U^T \Sigma U + U^T \mu \mu^T U \right\}}{\text{diag} \left\{ U^T \Sigma U + U^T \mu \mu^T U + \sigma^2 I \right\}}, \tag{21}
\]

where the division operation is element-wise.

**Proof**: See Appendix C.

To specify \( \mu \) and \( \Sigma \), we let

\[
\mu = \sum_{j=1}^k w_j p_j, \quad \Sigma = \sum_{j=1}^k w_j (p_j - \mu)(p_j - \mu)^T, \tag{22}
\]

where \( w_j \) is the \( j \)-th diagonal entry of \( W \) defined in (10). Intuitively, an interpretation of (22) is that \( \mu \) is the non-local mean of the reference patches. However, the more important part of (22) is \( \Sigma \), which measures the uncertainty of the reference patches with respect to \( \mu \). This uncertainty measure makes some fundamental improvements to existing methods which will be discussed in Section IV-C.

We note that Lemma 4 holds even if \( f(p) \) is not Gaussian. In fact, for any distribution \( f(p) \) with the first cumulant \( \mu \) and the second cumulant \( \Sigma \), the optimal solution in (20) still holds. This links our work to the classical linear minimum MSE (LMMSE) estimation \( [43] \).

From a computational perspective, \( \mu \) and \( \Sigma \) defined in (22) lead to a very efficient implementation as illustrated by the following lemma.
**Lemma 5:** Using $\mu$ and $\Sigma$ defined in (22), the optimal $\Lambda$ is given by

$$\Lambda = \frac{S}{S + \sigma^2 I}. \quad (23)$$

where $[U, S] = \text{eig}(PW^{1/2})$ is the eigen-decomposition of the weighted matrix $PW^{1/2}$.

**Proof:** See Appendix [1].

Combining Lemma 5 with Lemma 1, we observe that for any set of reference patches $\{p_i\}_{i=1}^k$, $U$ and $\Lambda$ can be determined simultaneously through the eigen-decomposition of $PW^{1/2}$. Therefore, we arrive at the overall algorithm shown in Algorithm 1.

---

**Algorithm 1 Proposed Algorithm**

Input: Noisy patch $q$, noise variance $\sigma^2$, and clean reference patches $p_1, \ldots, p_k$.

Output: Estimate $\hat{p}$

Learn $U$ and $\Lambda$:

- Form data matrix $P$ and weight matrix $W$.
- Compute eigen-decomposition $[U, S] = \text{eig}(PW^{1/2})$.
- Compute $\Lambda = \frac{S}{S + \sigma^2 I}$. (division is element-wise)

Denoise: $\hat{p} = U\Lambda U^T q$.

---

**C. Relationship to Prior Works**

It is interesting to note that many existing patch-based denoising algorithms assume some notions of prior, either explicitly or implicitly. In this subsection, we mention a few of the important ones. For notational simplicity, we will focus on the $i$th diagonal entry of $\Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_d\}$.

1. **BM3D** [3], **Shape-Adaptive BM3D** [4] and **BM3D-PCA** [5]: BM3D and its variants have two denoising steps. In the first step, the algorithm applies a basis matrix $U$ (either a pre-defined basis such as DCT, or a basis learned from PCA). Then, it applies a hard-thresholding to the projected coefficients to obtain a filtered image $\overline{p}$. In the second step, the filtered image $\overline{p}$ is used as a pilot estimate to the desired spectral component

$$\lambda_i = \frac{(u_i^T \overline{p})^2}{(u_i^T p)^2 + \sigma^2}. \quad (24)$$

Following our proposed Bayesian framework, we observe that the role of using $\overline{p}$ in (24) is equivalent to assuming a dirac delta prior

$$f(p) = \delta(p - \overline{p}). \quad (25)$$

In other words, the prior that BM3D assumes is concentrated at one point, $\overline{p}$, and there is no measure of uncertainty. As a result, the algorithm becomes highly sensitive to the first-pass estimate. In contrast, (22) suggests that the first-pass estimate can be defined as a non-local mean solution. Additionally, we incorporate a covariance matrix $\Sigma$ to measure the uncertainty of observing $\mu$. These provide a more robust estimate to the denoising algorithm which is absent from BM3D and its variants.

2. **LPG-PCA** [7]: In LPG-PCA, the $i$th spectral component $\lambda_i$ is defined as

$$\lambda_i = \frac{(u_i^T q)^2 - \sigma^2}{(u_i^T q)^2}. \quad (26)$$

where $q$ is the noisy patch. The (implicit) assumption in (27) is that $(u_i^T q)^2 \approx (u_i^T p)^2 + \sigma^2$, and so by substituting $(u_i^T p)^2 \approx (u_i^T q)^2 - \sigma^2$ into (5), yields (26). However, the assumption implies the existence of a perturbation $\Delta p$ such that $(u_i^T q)^2 = (u_i^T (p + \Delta p))^2 + \sigma^2$. Letting $\overline{p} = p + \Delta p$, we see that LPG-PCA implicitly assumes a dirac prior as in (24) and (25). The denoising result depends on the magnitude of $\Delta p$.

3. **Generic Global Prior** [6] [16] [17] [32]: As a comparison to methods using generic databases such as [6] [16] [17] [32], we note that the key difference lies in the usage of a **global** prior versus a **local** prior. Figure 5 illustrates the concept pictorially. The generic (global) prior $f(p)$ covers the entire space, whereas the targeted (local) prior is concentrated at its mean. The advantage of the local prior is that it allows one to denoise an image with few reference patches. It saves us from the intractable computation of learning the global prior, which is a high-dimensional non-parametric function.

---

**D. Improving $\Lambda$**

The Bayesian framework proposed above can be generalized to further improve the denoising performance. Referring to (20), we observe that the BMSE optimization can be reformulated to incorporate a penalty term in $\Lambda$. Here, we consider the following $\ell_1$ penalized BMSE:

$$\text{BMSE}_\alpha \overset{\text{def}}{=} \mathbb{E}_p \left[ \mathbb{E}_{q|p} \left[ \left\| U\Lambda U^T q - p \right\|_2^2 \right] \right] + \gamma \| \Lambda \|_\alpha, \quad (27)$$

where $\gamma > 0$ is the penalty parameter, and $\alpha \in \{0, 1\}$ controls which norm to be used. The solution to the minimization of (27) is given by the following lemma.
For a refined database of BMSE solution.

A. Comparison Methods

To evaluate the performance of the proposed algorithm against several existing methods, we compare our method against BM3D [3], BM3D-PCA [5], LPG-PCA [7], NLM [1] and EPLL [17].

Except for EPLL, all other four methods are internal denoising methods. We re-implement and modify the internal methods so that patch search is performed over the targeted external databases. These methods are iterated for two times where the solution of the first step is used as a basic estimate for the second step. The specific settings of each algorithm are as follows:

1) BM3D-PCA [5] and LPG-PCA [7]: $U$ is learned from the best $k$ external patches, which is the same as in our proposed method. $A$ is computed following (24) for BM3D-PCA and (26) for LPG-PCA. In BM3D-PCA’s first step, the threshold is set to 2.7σ.

2) NLM [1]: The weights in NLM are computed according to a Gaussian function of the $\ell_2$ distance of two patches [44, 45]. However, instead of using all reference patches in the database, we use the best $k$ patches following [2].

3) BM3D [3]: As a benchmark, we run the original BM3D code of Dabov et al. to show the performance of internal image denoising. For a fair comparison, we adjust the patch size, step size and search window size so that they are consistent with other methods. We use the default settings of BM3D for other parameters.

4) EPLL [17]: EPLL is an external denoising method, but the default patch prior is learned from a generic database. For a fair comparison, we use a targeted database that is used by the proposed method. We train the prior distribution using the EM algorithm mentioned in [17].

To emphasize the difference between the original algorithms (which are single-image denoising algorithms) and the corresponding new implementations for external databases, we denote the original, (single-image) denoising algorithms with “i” (internal), and the corresponding new implementations for external databases with “e” (external).

We add zero-mean Gaussian noise with standard deviations from $\sigma = 20$ to $\sigma = 80$ to the test images. The patch size is set as $8 \times 8$ (i.e., $d = 64$), and the sliding step size is 6 in the first step and 4 in the second step. Two quality metrics, namely Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM) are used to evaluate the objective quality of the denoised images.

B. Denoising Text and Documents

Our first experiment considers denoising a text image. The purpose is to simulate the case where we want to denoise a noisy document with the help of other similar but non-identical texts. This idea can be easily generalized to other scenarios such as handwritten signatures, bar codes and license plates.

To prepare this scenario, we capture a region ($127 \times 104$) of a document and add noise. We then build the targeted external database by cropping 9 arbitrary portions from a different document but with the same font sizes.

1) Denoising Performance: Figure 7 shows the denoising results when we add excessive noise ($\sigma = 100$) to the query image. Among all the methods, the proposed method yields the

Fig. 6: Comparisons of the $\ell_1$ and $\ell_0$ adaptive solutions over the original solution with $\gamma = 0$. The PSNR value for each noise level is averaged over 100 independent trials to reduce the bias due to a particular noise realization.

Lemma 6: Let $s_i$ be the $i$th diagonal entry in $S$, where $S$ is the eigenvalue matrix of $PW^TP$, then the optimal $A$ that minimizes BMSE$_{\alpha}$ is $\text{diag}\{\lambda_1, \cdots, \lambda_d\}$, where

$$\lambda_i = \max \left( \frac{s_i - \gamma/2}{s_i + \sigma^2}, 0 \right), \quad \text{for } \alpha = 1,$$

and

$$\lambda_i = \frac{s_i}{s_i + \sigma^2} \left( \frac{s_i^2}{s_i + \sigma^2} > \gamma \right), \quad \text{for } \alpha = 0. \quad (29)$$

Proof: See Appendix 2

The motivation of introducing an $\ell_\alpha$-norm penalty in (27) is related to the group sparsity used in defining $U$. Recall from Section III that since $U$ is the optimal solution to a group sparsity optimization, only few of the entries in the ideal projection $U^T p$ should be non-zero. Consequently, it is desired to require $A$ to be also sparse so that the reconstruction $UAU^T q$ has similar spectral components as that of $p$.

To demonstrate the effectiveness of the proposed $\ell_\alpha$ formulation, we consider the example patch shown in Figure 3. For a refined database of $k = 40$ patches, we consider the original minimum BMSE solution ($\gamma = 0$), the $\ell_0$ solution with $\gamma = 0.02$, and the $\ell_1$ solution with $\gamma = 0.02$. The results in Figure 6 show that with the proposed penalty term, the new BMSE$_{\alpha}$ solution performs consistently better than the original BMSE solution.

V. EXPERIMENTAL RESULTS

In this section, we present a set of denoising experiments to evaluate the performance of the proposed algorithm against several existing methods.

A. Comparison Methods

The methods we choose for comparison are BM3D [3], BM3D-PCA [5], LPG-PCA [7], NLM [1] and EPLL [17].
procedure determine the experimental result
(a) clean (b) noisy \( \sigma = 100 \)
(c) iBM3D 16.68 dB (0.7100)
(d) EPLL(generic) 16.93 dB (0.7341)
(e) EPLL(target) 18.65 dB (0.8234)
(f) eNLM 20.72 dB (0.8422)
(g) eBM3D 20.33 dB (0.8228)
(h) eBM3D-PCA 21.39 dB (0.8435)
(i) eLPG-PCA 20.37 dB (0.7299)
(j) ours 22.20 dB (0.9069)

Fig. 7: Denoising text images: Visual comparison and objective comparison (PSNR and SSIM in the parenthesis). Prefix “i” stands for internal denoising (i.e., single-image denoising), and prefix “e” stands for external denoising (i.e., using external databases).

highest PSNR and SSIM values. The PSNR is 5 dB better than the benchmark BM3D (internal) denoising algorithm. Some existing training-based methods, such as EPLL, do not perform well due to the insufficient training samples from the targeted database. Compared to other external denoising methods, the proposed method shows a better utilization of the targeted database.

We plot and compare the PSNR values over a range of noise levels in Figure 8. Our proposed method outperforms other competitors, especially at high noise levels. For example, for \( \sigma = 60 \), our restored result is 0.82 dB better than the second best result by eBM3D-PCA.

![Fig. 8](image_url)

Fig. 8: Text image denoising: PSNR vs noise levels. In this plot, each PSNR value is averaged over 8 independent Monte-Carlo trials to reduce the bias due to a particular noise realization.

2) Database Quality: We show how the database quality affects our denoising performance. Given a database, we compute its average distance from the clean image of interest. Specifically, for each patch \( p_i \in \mathbb{R}^d \) in a clean image containing \( m \) patches and a database \( \mathcal{P} \) of \( n \) patches, we compute its minimum distance

\[
d(p_i, \mathcal{P}) \overset{\text{def}}{=} \min_{p_j \in \mathcal{P}} \|p_i - p_j\|_2 / \sqrt{d}.
\]

The average patch-database distance is then defined as \( \overline{d}(\mathcal{P}) \overset{\text{def}}{=} (1/m) \sum_{i=1}^{m} d(p_i, \mathcal{P}) \). Therefore, a smaller \( \overline{d}(\mathcal{P}) \) indicates that the database is more relevant to the ground truth (clean) image.

Figure 9 shows the results. For all noise levels, when the average distance between an image and a database decreases,
the proposed method benefits from the database quality improvement and thus enhances the denoising performance.

C. Denoising Multiview Images

Our second experiment considers the scenario of capturing images using a multiview camera system. The multiview images are captured at different viewing positions. Suppose that one or more cameras are not functioning properly so that some images are corrupted with noise. Our goal is to demonstrate that with the help of the other clean views, the noisy view could be restored.

To simulate the experiment, we download 2 multiview datasets from Middlebury Computer Vision Page

\[\text{http://vision.middlebury.edu/stereo/}\]. Each set of images consists of 5 views. We add i.i.d. Gaussian noise to one view and then use the rest 4 views to assist in denoising.

In Figure 10, we visually show the denoising results of the “Barn” and “Cone” multiview datasets. In comparison to the competing methods, our proposed method has the highest PSNR values. The magnified areas indicate that our proposed method removes the noise significantly and better reconstructs some fine details. In Figure 11, we plot and compare the PSNR values over a range of noise levels. The proposed method is consistently better than its competitors. For example, for \(\sigma = 50\), our proposed method is 0.80 dB better than eBM3D-PCA and 1.94 dB better than iBM3D.

D. Denoising Human Faces

Our third experiment considers denoising human face images. In low light conditions, images captured are typically corrupted by noise. To facilitate other high-level vision tasks such as recognition and tracking, denoising is a necessary preprocessing step. This experiment demonstrates the ability of denoising face images.

\[\text{http://vision.middlebury.edu/stereo/}\]

In this experiment, we use the Gore face database from [46], of which some examples are shown in the top row of Figure 12 (each image is 60 × 80). We simulate the denoising task by adding noise to a randomly chosen image and then use the other images (19 other face images in our experiment) in the database to assist in denoising.

In the bottom row of Figure 12 we show the noisy face and denoising results. We observe that while the facial expressions are different and there are misalignments between images, the proposed method still generates robust results. In Figure 13 we plot the PSNR curves, where we see consistent gain compared to other methods.
Fig. 12: Face denoising of Gore dataset [46]. [Top] Database images; [Bottom] Denoising results.

Fig. 13: Face denoising results: PSNR vs noise levels. In this plot, each PSNR value is averaged over 8 independent Monte-Carlo trials to reduce the bias due to a particular noise realization.

VI. CONCLUSION

Classical image denoising methods based on a single noisy input or generic databases are approaching their performance limits. We proposed an adaptive image denoising algorithm using targeted databases. The proposed method applies a group sparsity minimization and a localized prior to learn the basis matrix and the spectral coefficients of the optimal denoising filter, respectively. We show that the new method generalizes a number of existing patch-based denoising algorithms such as BM3D, BM3D-PCA, Shape-adaptive BM3D, LPG-PCA, and EPLL. Based on the new framework, we proposed improvement schemes, namely an improved patch selection procedure for determining the basis matrix and a penalized minimization for determining the spectral coefficients. For a variety of scenarios including text, multiview images and faces, we demonstrated empirically that the proposed method has superior performance over existing methods. With the increasing amount of image data available online, we anticipate that the proposed method is an important first step towards a data-dependent generation of denoising algorithms.

APPENDIX

A. Proof of Lemma 2

Proof: From (4), the optimization to be solved is

\[
\begin{align*}
\min_{u_1, \ldots, u_d, \lambda_1, \ldots, \lambda_d} & \quad \sum_{i=1}^{d} [(1 - \lambda_i)(u_i^T p)^2 + \sigma^2 \lambda_i^2] \\
\text{subject to} & \quad u_i^T u_i = 1, \quad u_i^T u_j = 0.
\end{align*}
\]

Since each term in the sum of the objective function is non-negative, we can consider the minimization over each individual term separately. This gives

\[
\begin{align*}
\min_{u_i, \lambda_i} & \quad (1 - \lambda_i)(u_i^T p)^2 + \sigma^2 \lambda_i^2 \\
\text{subject to} & \quad u_i^T u_i = 1.
\end{align*}
\]

The Lagrangian function of the above equality-constrained problem is

\[
\mathcal{L}(u_i, \lambda_i, \beta) = (1 - \lambda_i)(u_i^T p)^2 + \sigma^2 \lambda_i^2 + \beta (1 - u_i^T u_i),
\]

where \(\beta\) is the Lagrange multiplier. Differentiating \(\mathcal{L}\) with respect to \(u_i, \lambda_i\) and \(\beta\) yields

\[
\begin{align*}
\frac{\partial \mathcal{L}}{\partial \lambda_i} & = -2(1 - \lambda_i)(u_i^T p)^2 + 2 \lambda_i \sigma^2 \\
\frac{\partial \mathcal{L}}{\partial u_i} & = 2(1 - \lambda_i)(u_i^T p)p - 2 \beta u_i \\
\frac{\partial \mathcal{L}}{\partial \beta} & = 1 - u_i^T u_i.
\end{align*}
\]

Setting \(\partial \mathcal{L}/\partial \lambda_i = 0\) yields \(\lambda_i = (u_i^T p)^2/(u_i^T p)^2 + \sigma^2\). Substituting this \(\lambda_i\) into (31) and setting \(\partial \mathcal{L}/\partial u_i = 0\) yields

\[
\frac{2 \sigma^4 (u_i^T p) p}{((u_i^T p)^2 + \sigma^2)^2} - 2 \beta u_i = 0. \quad (33)
\]

There are two solutions. The first one is the trivial one: \(u_i = \text{any unit vector orthogonal to } p\) and \(\beta = 0\). In this case, \(u_i^T p = 0\) and \(\beta u_i = 0\) so that the left hand side of (33) is 0. The non-trivial solution is

\[
u_i = p/\|p\|, \quad \text{and} \quad \beta = \frac{\sigma^4 \|p\|^2}{((\|p\|^2 + \sigma^2)^2}, \quad (34)
\]

which can be proved easily by substituting (34) into (33). Therefore, the denoising result is

\[
\hat{p} = U \left( \text{diag} \left\{ \frac{\|p\|^2}{\|p\|^2 + \sigma^2}, 0, \ldots, 0 \right\} \right) U^T q.
\]

B. Proof of Lemma 3

Proof: Let \(u_i\) be the \(i\)th column of \(U\). Then, (8) becomes

\[
\begin{align*}
\min_{u_1, \ldots, u_d} & \quad \sum_{i=1}^{d} \|u_i^T P\|_2 \\
\text{subject to} & \quad u_i^T u_i = 1, \quad u_i^T u_j = 0.
\end{align*}
\]

Since each term in the sum of (35) is non-negative, we can consider each individual term

\[
\begin{align*}
\min_{u_i} & \quad \|u_i^T P\|_2 \\
\text{subject to} & \quad u_i^T u_i = 1,
\end{align*}
\]
which is equivalent to

$$\text{minimize } \|u_i^T P\|^2_2$$

subject to

$$u_i^T u_i = 1.$$ (36)

The constrained problem (36) can be solved by considering the Lagrange function,

$$L(u_i, \beta) = \|u_i^T P\|^2_2 + \beta(1 - u_i^T u_i).$$ (37)

Taking derivatives $\frac{dL}{du_i} = 0$ and $\frac{dL}{d\beta} = 0$ yield

$$PP^T u_i = \beta u_i, \quad \text{and} \quad u_i^T u_i = 1.$$ Therefore, $u_i$ is the eigenvector of $PP^T$, and $\beta$ is the corresponding eigenvalue.

C. Proof of Lemma 4

Proof: First, by plugging $q = p + \eta$ into BMSE we get

$$\text{BMSE} = \mathbb{E}_p \left[ \mathbb{E}_q \left[ \left\| UAU^T (p + \eta) - p \right\|^2_2 \right] \right]$$

$$= \mathbb{E}_p \left[ p^T U (I - \Lambda)^2 U^T p + \sigma^2 \text{Tr} (\Lambda^2) \right] = \text{Tr} \left( (I - \Lambda)^2 U^T U \mu^T U + (I - \Lambda)^2 U^T \Sigma U \right) + \sigma^2 \text{Tr} (\Lambda^2) = \sum_{i=1}^d \left[ (1 - \lambda_i)^2 s_i + \sigma^2 \lambda_i^2 \right]$$

$$= \sum_{i=1}^d \left[ \left( \lambda_i - \frac{s_i}{s_i + \sigma^2} \right)^2 \frac{s_i^2 + \sigma^2}{s_i + \sigma^2} \right].$$

Adding the penalty term $\gamma \|\Lambda\|_1$, the minimization problem with respect to $\lambda_i$ becomes

$$\text{minimize } \lambda_i \sum_{i=1}^d \left[ \left( \lambda_i - \frac{s_i}{s_i + \sigma^2} \right)^2 \right] + \gamma \|\Lambda\|_1,$$ (42)

where $\gamma \|\Lambda\|_1 = \gamma \sum_{i=1}^d |\lambda_i|$ or $\gamma \sum_{i=1}^d 1 (\lambda_i \neq 0)$ for $\alpha = 1$ or 0. We note that when $\alpha = 1$ or 0, (42) is the standard shrinkage problem [47], in which a closed form solution exists. Following from [48], the solutions are given by

$$\lambda_i = \max \left( \frac{s_i - \gamma/2}{s_i + \sigma^2}, 0 \right), \quad \text{for } \alpha = 1,$$

and

$$\lambda_i = \frac{s_i}{s_i + \sigma^2}, \quad \text{for } \alpha = 0.$$

D. Proof of Lemma 5

Proof: First, we write $\Sigma$ in (22) in a matrix form

$$\Sigma = (P - \mu 1^T) W (P - \mu 1^T)^T = PW P^T - \mu 1^T W P^T - PW 1 \mu^T + \mu^T W 1 \mu^T.$$

It is not difficult to see that $1^T W P^T = \mu^T$, $PW 1 = \mu$ and $1^T W 1 = 1$. Therefore,

$$\Sigma = PW P^T - \mu 1^T - \mu 1^T + \mu^T = PW P^T - \mu 1^T,$$

which gives

$$\Sigma + \mu 1^T = PW P^T.$$ (40)

Substituting (40) into (39), we have

$$\Lambda = \frac{\text{diag} \left\{ U^T (\Sigma + \mu 1^T) U \right\}}{\text{diag} \left\{ U^T (\Sigma + \mu 1^T) U \right\} + \sigma^2 I}$$

$$= \frac{\text{diag} \left\{ U^T P W P^T U \right\}}{\text{diag} \left\{ U^T P W P^T U \right\} + \sigma^2 I} = \frac{1}{S + \sigma^2 I},$$ (41)

where the divisions are element-wise.

E. Proof of Lemma 6

Proof: To prove Lemma 6 we first apply the results in (38) and (41)

$$\frac{\partial}{\partial \Lambda} \text{BMSE} = \frac{\partial}{\partial \Lambda} \mathbb{E}_q \left[ \left\| U A U^T q - p \right\|^2_2 \right] = \text{Tr} \left( (I - \Lambda)^2 U^T \mu^T U + (I - \Lambda)^2 U^T \Sigma U \right) + \sigma^2 \text{Tr} (\Lambda^2)$$

$$= \text{Tr} (I - \Lambda)^2 S + \sigma^2 \text{Tr} (\Lambda^2) = \sum_{i=1}^d \left[ (1 - \lambda_i)^2 s_i + \sigma^2 \lambda_i^2 \right].$$

Adding the penalty term $\gamma \|\Lambda\|_1$, the minimization problem with respect to $\lambda_i$ becomes

$$\text{minimize } \lambda_i \sum_{i=1}^d \left[ \left( \lambda_i - \frac{s_i}{s_i + \sigma^2} \right)^2 \right] + \gamma \|\Lambda\|_1,$$ (42)

where $\gamma \|\Lambda\|_1 = \gamma \sum_{i=1}^d |\lambda_i|$ or $\gamma \sum_{i=1}^d 1 (\lambda_i \neq 0)$ for $\alpha = 1$ or 0. We note that when $\alpha = 1$ or 0, (42) is the standard shrinkage problem [47], in which a closed form solution exists. Following from [48], the solutions are given by

$$\lambda_i = \max \left( \frac{s_i - \gamma/2}{s_i + \sigma^2}, 0 \right), \quad \text{for } \alpha = 1,$$

and

$$\lambda_i = \frac{s_i}{s_i + \sigma^2}, \quad \text{for } \alpha = 0.$$

REFERENCES

[1] A. Buades, B. Coll, and J. Morel, "A review of image denoising algorithms, with a new one," SIAM Multiscale Model and Simulation, vol. 4, no. 2, pp. 490–530, 2005.
[2] C. Kervrann and J. Boulanger, "Local adaptivity to variable smoothness for exemplar-based image regularization and representation," International Journal of Computer Vision, vol. 79, no. 1, pp. 45–69, 2008.
[3] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3D transform-domain collaborative filtering," IEEE Trans. Image Process., vol. 16, no. 8, pp. 2080–2095, Aug. 2007.
[4] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "A nonlocal and shape-adaptive transform-domain collaborative filtering," in Proc. Intl. Workshop on Local and Non-Local Approx. in Image Process. (LNLA’08), pp. 1–8, Aug. 2008.
[5] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, “BM3D image denoising with shape-adaptive principal component analysis,” in Signal Process. with Adaptive Sparse Structured Representations (SPARS’09), pp. 1–6, Apr. 2009.
[6] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, “Non-local sparse models for image restoration,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR’09), pp. 2272–2279, Sep. 2009.
[7] L. Zhang, W. Dong, D. Zhang, and G. Shi, “Two-stage image denoising by principal component analysis with local pixel grouping,” Pattern Recognition, vol. 43, pp. 1531–1549, Apr. 2010.
[8] W. Dong, L. Zhang, G. Shi, and X. Li, “Nonlocally centralized sparse representation for image restoration,” IEEE Trans. Image Process., vol. 22, no. 4, pp. 1620 – 1630, Apr. 2013.
[9] A. Rajwade, A. Rangarajan, and A. Banerjee, “Image denoising using the higher order singular value decomposition,” IEEE Trans. Pattern Anal. and Mach. Intell., vol. 35, no. 4, pp. 849 – 862, Apr. 2013.
[10] M. Zontak and M. Irani, “Internal statistics of a single natural image,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR’11), pp. 977–984, Jun. 2011.
[11] I. Mosseri, M. Zontak, and M. Irani, “Combining the power of internal and external denoising,” in Proc. Intl. Conf. Computational Photography (ICCP’12), pp. 1–9, Apr. 2013.
[12] H.C. Burger, C.J. Schuler, and S. Harmeling, “Learning how to combine internal and external denoising methods,” Pattern Recognition, pp. 121–130, 2013.
[13] D. Glasner, S. Bagon, and M. Irani, “Super-resolution from a single image,” in Proc. Intl. Conf. Computer Vision (ICCV’09), pp. 349–356, Sep. 2009.
[14] P. Chatterjee and P. Milanfar, “Is denoising dead?,” IEEE Trans. Image Process., vol. 19, no. 4, pp. 895–911, Apr. 2010.
[15] A. Levin and B. Nadler, “Natural image denoising: Optimality and inherent bounds,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR’11), pp. 2833–2840, Jun. 2011.
[16] M. Elad and M. Aharon, “Image denoising via sparse and redundant representations over learned dictionaries,” IEEE Trans. Image Process., vol. 15, no. 12, pp. 3565–3584, Dec. 2006.
[17] D. Zoran and Y. Weiss, “From learning models of natural image patches to whole image restoration,” in Proc. IEEE Intl. Conf. Computer Vision (ICCV’11), pp. 479–486, Nov. 2011.
[18] S.H. Chan, T. Zickler, and Y. M. Lu, “Fast non-local filtering by random iterative denoising,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR’09), pp. 1542–1549, Jun. 2009.
[19] T. Buades, Y. Lou, J. Morel, and Z. Tang, “A note on multi-image denoising,” in Proc. IEEE Intl. Workshop on Local and Non-Local Approx. in Image Process. (LNLA’09), pp. 1–15, Aug. 2009.
[20] S. Roth and M. J. Black, “Fields of experts,” Intl. J. Computer Vision, vol. 82, no. 2, pp. 205–229, 2009.
[21] G. Yu, G. Sapiro, and S. Mallat, “Solving inverse problems with piecewise linear estimators: From gaussian mixture models to structured sparsity,” IEEE Trans. Image Process., vol. 21, no. 5, pp. 2481–2499, May 2012.
[22] E. Luo, S.H. Chan, and T.Q. Nguyen, “Image denoising by targeted external databases,” in Proc. IEEE Intl. Conf. Acoustics, Speech and Signal Process. (ICASSP’14), pp. 2469–2473, May 2014.
[23] P. Milanfar, “A tour of modern image filtering,” IEEE Signal Process. Magazine, vol. 30, no. 106–128, Jan. 2013.
[24] P. Milanfar, “Symmetrizing smoothing filters,” SIAM J. Imaging Sci., vol. 6, no. 1, pp. 263–284, 2013.
[25] H. Talebi and P. Milanfar, “Global image denoising,” IEEE Trans. Image Process., vol. 23, no. 2, pp. 755–768, Feb. 2014.
[26] S. Cotter, B. Rao, K. Engan, and K. Kreutz-Delgado, “Sparse solutions to linear inverse problems with multiple measurement vectors,” IEEE Trans. Signal Process., vol. 53, no. 7, pp. 2477–2488, Jul. 2005.
[27] T. Kolda and B. Bader, “Tensor decompositions and applications,” SIAM Review, vol. 51, no. 3, pp. 455–500, 2009.
[28] S. Roth and M. Black, “Fields of experts: A framework for learning image priors,” in Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR’05), 2005, vol. 2, pp. 860–867 vol. 2, Jun. 2005.
[29] D. Zoran and Y. Weiss, “Natural images, gaussian mixtures and dead leaves,” Advances in Neural Information Process. Systems (NIPS’12), vol. 25, pp. 1745–1753, 2012.
[30] S. M. Kay, “Fundamentals of statistical signal processing: Detection theory,” 1998.
[31] A. Buades, B. Coll, and J. M. Morel, “Non-local means denoising,” [on line] http://www.ipol.im/pub/art/2011/bcm alm/. 2011.
[32] E. Luo, S. Pan, and T. Nguyen, “Generalized non-local means for iterative denoising,” in Proc. 20th Euro. Signal Process. Conf. (EU- SIPCO’12), pp. 260–264, Aug. 2012.
[33] Y. Peng, A. Ganesh, J. Wright, W. Xue, and Y. Ma, “RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images,” IEEE Trans. Pattern Anal. and Mach. Intell., vol. 34, no. 11, pp. 2233–2246, Nov. 2012.
[34] C. Li, “An efficient algorithm for total variation regularization with applications to the single pixel camera and compressive sensing,” PhD. thesis, Rice Univ., 2009, available at http://www.caam.rice.edu/~optimization/L1/TVAl3/tval3_thesis.pdf.
[35] S.H. Chan, K. Khoshabeh, K.B. Gibson, P.E. Gill, and T.Q. Nguyen, “An augmented Lagrangian method for total variation video restoration,” IEEE Trans. Image Process., vol. 20, no. 11, pp. 3097–3111, Nov. 2011.