Int J Comput Vis (2008) 80: 406–420
DOI 10.1007/s11263-008-0148-2

# Limits of Learning-Based Superresolution Algorithms

**Zhouchen Lin · Junfeng He · Xiaoou Tang · Chi-Keung Tang**

Received: 2 October 2007 / Accepted: 10 June 2008 / Published online: 20 August 2008


**Abstract** Learning-based superresolution (SR) is a popular SR technique that uses application dependent priors to infer the missing details in low resolution images (LRIs). However, their performance still deteriorates quickly when the magnification factor is only moderately large. This leads us to an important problem: "Do limits of learning-based SR algorithms exist?" This paper is the first attempt to shed some light on this problem when the SR algorithms are designed for general natural images. We first define an expected risk for the SR algorithms that is based on the root mean squared error between the superresolved images and the ground truth images. Then utilizing the statistics of general natural images, we derive a closed form estimate of the lower bound of the expected risk. The lower bound only involves the covariance matrix and the mean vector of the high resolution images (HRIs) and hence can be computed by sampling real images. We also investigate the sufficient number of samples to guarantee an accurate estimate of the lower bound. By computing the curve of the lower bound w.r.t. the magnification factor, we could estimate the limits of learning-based SR algorithms, at which the lower bound of the expected risk exceeds a relatively large threshold. We perform experiments to validate our theory. And based on our observations we conjecture that the limits may be independent of the size of either the LRIs or the HRIs.

**Keywords** Superresolution · Learning-based · Limits · Resolution

## 1 Introduction

Superresolution (SR) is a technique that produces an image or video with a resolution that is higher than those of any of the input images or frames. Roughly speaking, SR algorithms can be categorized into four classes (Borman and Stevenson 1998; Farsiu et al. 2004; Park et al. 2003). Interpolation-based algorithms (e.g., Komatsu et al. 1993; Nguyen and Milanfar 2000; Shah and Zakhor 1999) register low resolution images (LRIs) with the high resolution image (HRI), then apply nonuniform interpolation to produce an improved resolution image which is further deblurred. Frequency-based algorithms (e.g., Kim and Su 1993; Rhee and Kang 1999; Tsai and Huang 1984) make use of the aliasing that exists in each LRI and try to dealias the LRIs by utilizing the phase difference among the LRIs. Reconstruction-based algorithms (e.g., Elad and Feuer 1997; Hardie et al. 1997; Lin and Shum 2004; Patti et al. 1997) rely on the registration relationship between the LRIs and the HRI to build a linear system that relates the LRIs and the HRI, and then assume various kinds of priors on the HRI in order to regularize this ill-posed inverse problem. Recently, many

A preliminary version of this paper was published in International Conference on Computer Vision 2007 (Lin et al. 2007).

Z. Lin (✉) · X. Tang
Microsoft Research Asia, Sigma Building, Zhichun Road #49, Haidian District, Beijing 100190, People's Republic of China
e-mail: zhoulin@microsoft.com

X. Tang
e-mail: xitang@microsoft.com

J. He · C.-K. Tang
The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, People's Republic of China

J. He
e-mail: hejf@cse.ust.hk

C.-K. Tang
e-mail: cktang@cse.ust.hk

learning-based algorithms (LBAs) have attracted much attention. Learning-based SR algorithms are new SR techniques that may start from the seminal papers by Freeman and Pasztor (1999) and Baker and Kanade (2002). Compared to traditional methods, which basically process images at the signal level, learning-based SR algorithms incorporate application dependent priors to infer the unknown HRI. For instance, a probabilistic model is often learned from a training set of image pairs, i.e., corresponding LRI and HRI images, and then the output HRI is inferred from the input LRIs based on that probabilistic model. Often, they can achieve better results if the training set agrees well with the statistics of the input LRIs.

Existing LBAs vary from generic to specific, incorporating different levels of prior knowledge. Among them, the algorithms in Freeman and Pasztor (1999), Bégin and Ferrie (2004), Bishop et al. (2003), Candocia and Principe (1999), Kursun and Favorov (2003), Miravet and Rodríguez (2003), Pickup et al. (2003), Sun et al. (2003), Zhang and Pan (2002) can be applied to general images or videos (Bishop et al. 2003). As the image/video sizes are not fixed, all of them can only use patch-based approaches. In contrast, the algorithms in Baker and Kanade (2002), Capel and Zisserman (2001), Dedeoğlu et al. (2004), Li and Lin (2004a, 2004b), Liu et al. (2001, 2005a, 2005b, 2005c) are devoted to face hallucination only. They all utilize the strong structural information of faces by aligning the face images. Hence most of them use eigen-faces. Despite some drawbacks such as a fixed magnification factor and dependence of performance on how well the input LRI matches the training low resolution (LR) samples, LBAs have several advantages, which make them very popular nowadays. For example, they require fewer LRIs yet still achieve higher magnification factors than traditional SR algorithms, because they already have more specific prior knowledge of the image/video. Most of the algorithms can even work on a *single* image, which is impossible for traditional algorithms. Moreover, it is possible to design fast LBAs, e.g., eigen-face based face hallucination or neural network based SR algorithms, while traditional SR algorithms (except nonuniform interpolation Park et al. 2003) are usually iterative and slow, making real-time SR difficult. Finally, if we change the prior for LBAs, the HRIs may exhibit an artistic style (Freeman and Pasztor 1999; Pickup et al. 2003). This may enable LBAs to perform style transfer. In contrast, traditional SR algorithms do not have such capability.

Although more specific prior knowledge, e.g., the strong structure of faces, has been incorporated in LBAs, people have found that the SR results are still unsatisfactory even though the magnification factors are still not very large. This poses an important question: "Do limits exist for learning-based superresolution?", i.e., "Does there exist an upper bound for magnification factors such that *no* SR algorithm can produce satisfactory results?" In this paper, we provide some theoretical analysis on the limits of LBAs for general natural images, which is the first work on this problem according to the best of our knowledge. In comparison, the counterpart analysis for reconstruction-based SR algorithms has been presented in Lin and Shum (2004) and Baker and Kanade (2002). This paper has two major contributions:

1. A closed form lower bound of the expected error between the superresolved and the ground truth images is proved. This formula only involves the covariance matrix and the mean of the prior distribution of HRIs. This lower bound is used to estimate the limits of LBAs.
2. A formula on the sufficient number of HRI samples is provided to ensure the accuracy of the sample-based computation of the lower bound.

Moreover, from our experiments, we have also observed that the limits may be independent of the sizes of both LRIs and HRIs.

Currently, we limit our analysis to general natural images, i.e., *the set of all natural images of given size*, because the statistics of general natural images have been studied for a long time (Srivastava et al. 2003) and there have been some pertinent observations on their characteristics that are useful for our analysis. In particular, we will use the following two properties:

1. The distribution of HRIs is not concentrated around several HRIs and the distribution of LRIs is not concentrated around several LRIs either. Noticing that general natural images cannot be classified into a small number of categories will justify this property.
2. Smoother LRIs have a higher probability than non-smooth ones. This property is actually called the "smoothness prior" that is widely used for regularization, for instance, when performing reconstruction-based SR.

In contrast, for specific class of images, e.g., face or text images, there is no similar work on their statistics to the best of our knowledge. So currently we have to focus on the SR of general natural images.

The rest of this paper is organized as follows. We first review the existing LBAs in Sect. 2. Then we formulate our problem and prove the lower bound of the expected risks of LBAs in Sect. 3. In Sect. 4, we investigate the problem of the sufficient number of samples needed for estimating the lower bound. Next, we present the experimental results in Sect. 5 and discuss the case of involving the noise in Sect. 6. Finally, we give the conclusions and future work in Sect. 7.

## 2 Overview of Learning-Based SR Algorithms

Based on our own understandings, the existing LBAs can roughly be categorized into indirect maximum a posteriori

(MAP) inference and direct MAP inference, where the former can be further classified as global and local.

## 2.1 Indirect MAP inference LBAs

Indirect MAP inference algorithms model the SR problems as:

$$\mathbf{H} = \arg\max_{\mathbf{H}} P(\{\dot{\mathbf{L}}_i\}_{i=1}^N|\dot{\mathbf{H}})P(\dot{\mathbf{H}}),$$

where $\mathbf{H}$ is the HRI, $\dot{\mathbf{H}}$ is some feature or quantity related to the HRI ($\dot{\mathbf{H}}$ could be identical to $\mathbf{H}$), and $\dot{\mathbf{L}}_i$ are some features or quantities related to the LRIs ( $\dot{\mathbf{L}}_i$ could be identical to the LRI $\mathbf{L}_i$ too). Different algorithms differ in the definitions of $P(\{\dot{\mathbf{L}}_i\}_{i=1}^N|\dot{\mathbf{H}})$ and $P(\dot{\mathbf{H}})$.

### *2.1.1 Locally Indirect MAP inference LBAs*

Locally indirect MAP LBAs infer an HRI patch by patch. The final HRI is obtained by resolving the mismatch among neighboring patches.

Freeman and Pasztor (1999) model the SR problem as an inference problem of the high frequency:

$$\mathbf{H} = \bar{\mathbf{L}}_1 + \hat{\mathbf{H}},$$

where $\bar{\mathbf{L}}_1$ is the image by interpolating the LRI $\mathbf{L}_1$ to the size of $\mathbf{H}$, and $\hat{\mathbf{H}}$ is the missing high frequency such that:

$$\hat{\mathbf{H}} = \arg\max_{\hat{\mathbf{H}}} P(\tilde{\mathbf{L}}_1|\hat{\mathbf{H}})P(\hat{\mathbf{H}}),$$

where $\tilde{\mathbf{L}}_1$ is the mid-frequency of $\bar{\mathbf{L}}_1$. In their framework, $P(\tilde{\mathbf{L}}_1|\hat{\mathbf{H}})$ and $P(\hat{\mathbf{H}})$ are defined via local patches in the HRI. They use a Markov network to model the relationship between the high resolution (HR) patches and LR patches: nearby HR patches are connected to each other and each HR patch is also connected to its corresponding LR patch. The weights in this Markov network are learnt from sample images and are approximated by mixture of Gaussians. Standard Belief Propagation algorithm is adopted to find the optimal $\hat{\mathbf{H}}$ iteratively.

Bégin and Ferrie's algorithm (Bégin and Ferrie 2004) follows the framework of Freeman and Pasztor (1999). However, they argue that to have good SR performance the LR sample patches matching the input LRI should have similar point spread functions to that of the input LRI. Bishop et al. (2003) and Dedeoğlu et al. (2004) further extend Freeman and Pasztor's work (Freeman and Pasztor 1999) to videos, but the former simplifies the algorithm a lot and the latter only targets on face hallucination.

Baker and Kanade's hallucination algorithm (Baker and Kanade 2002) models the likelihood as

$$P(\{\mathbf{L}_i\}_{i=1}^N|\mathbf{H}) \sim \exp\{-\lambda\|\mathbf{H} - \mathbf{PL}\|^2\}, \quad (1)$$

where $\mathbf{L}$ is the concatenated vector of LR pixels in $\{\mathbf{L}_i\}_{i=1}^N$ and $\mathbf{P}$ is the coefficient matrix relating the HR pixels and the LR pixels. The prior probability $P(\mathbf{H})$ of HRIs is designed according to the pixels in the HRIs that best match the pixels in LRIs. In Pickup et al. (2003), the likelihood is the same as (1), but the prior probability of HRIs is learnt from texture samples.

### *2.1.2 Globally Indirect MAP Inference LBAs*

Globally indirect MAP LBAs do not infer the HRI patch by patch. Instead, the existing algorithms assume that the HRI can be decomposed into "bases". Therefore, what the algorithms compute is actually the combination coefficients among the bases. Usually, for a global algorithm to work effectively, the HRIs should have a consistent structure. Therefore, all existing global LBAs are proposed for face hallucination.

Gunturk et al. (2003) represent both the HR and the LR face images as linear combinations of corresponding eigenfaces. The SR problem becomes:

$$\mathbf{h} = \arg\max_{\mathbf{h}} P(\mathbf{h})P(\{\mathbf{l}_i\}_{i=1}^N|\mathbf{h}),$$

where $\mathbf{h}$ is the combination vector for the HR face image and $\mathbf{l}_i$ are the combination vectors for the LR face images. $P(\mathbf{h})$ and $P(\{\mathbf{l}_i\}_{i=1}^N|\mathbf{h})$ are defined via Gaussians. Following Gunturk et al. (2003), Li and Lin (2004b) further consider the pose variation in the input faces. In contrast, all the rest of the face hallucinating algorithms deal with frontal faces only.

Liu et al.'s face hallucination algorithm (Liu et al. 2001) is actually a hybrid of global and local approaches. The inference of the global face image $\mathbf{H}^g$ in Liu et al. (2001) follows a methodology that is similar to that of Gunturk et al. (2003). As for the local feature image $\mathbf{H}^l$, like Freeman and Pasztor (1999), Liu et al. also adopt an MRF model that partitions the LRI and the HRI into patches and use the global feature image $\mathbf{H}^g$ to infer the local feature image $\mathbf{H}^l$:

$$\mathbf{H}^l = \arg\max_{\mathbf{H}^l} P(\mathbf{H}^l|\mathbf{H}^g).$$

Simulated annealing algorithm is applied to find the optimal $\mathbf{H}^l$. Capel and Zisserman's framework (Capel and Zisserman 2001) resembles that of the inference of the global face image in Gunturk et al. (2003), Liu et al. (2001). However, the faces are partitioned into five regions: left eye, right eye, nose, left cheek, right cheek, and mouth. The SR is then done in each region separately. As the alignment among individual parts of faces is better than aligning the whole face, the SR results exhibit more details than those in Liu et al. (2001) and Gunturk et al. (2003). The side effect is that the boundaries between the regions may be discontinuous. Also

inspired by Liu et al. (2001), Li and Lin (2004a) propose a computationally efficient approach to compute the global face and the residue. When computing the global face, both the HRI and the LRI are assumed to be a linear combination of HR and LR principal components, respectively. As they adopt the smoothness prior, the resultant HRI is inevitably blurred.

### 2.2 Direct MAP Inference LBAs

The existing direct MAP inference LBAs are all patch based. So there is no need to classify them as either global or local. This class of methods model the SR problem as:

$$\mathbf{H} = \arg\max_{\mathbf{H}} P(\dot{\mathbf{H}} | \{\dot{\mathbf{L}}_i\}_{i=1}^{N})$$

where the notations follow those in Sect. 2.1.

Sun et al. (2003) use the primal sketch prior to infer the high frequency:

$$\mathbf{H} = \bar{\mathbf{L}} + \mathbf{H}^p,$$

where $\bar{\mathbf{L}}$ is the image by interpolating the LRI $\mathbf{L}$ to the size of HRI and

$$\mathbf{H}^p = \arg\max_{\mathbf{H}^p} P(\mathbf{H}^p | \bar{\mathbf{L}}),$$

is the missing high frequency. The algorithm first prepares many examples of HR and corresponding LR patches extracted along edges in training images using Gabor filters. To compute $P(\mathbf{H}^p | \bar{\mathbf{L}})$, the authors approximate it as:

$$P(\mathbf{H}^p | \bar{\mathbf{L}}) \approx \prod_k P(C_k | \bar{\mathbf{L}}),$$

where $C_k$ are the contours in $\bar{\mathbf{L}}$ and have already been approximated by candidates of HR patches. To resolve the inconsistency among the overlapping patches, each contour is modelled as a first order Markov chain. Their methodology is very similar to that in Freeman and Pasztor (1999) and Belief Propagation is also adopted to find the optimal HR patches.

Hertzmann et al. (2001) propose a very simple algorithm that transfers styles between images. Given a pair of images $A$ and $A'$ and another image $B$, the goal is to synthesize a new image $B'$ such that the transform from $B$ to $B'$ is similar to that from $A$ to $A'$. According to the algorithm, $B'$ is synthesized pixel by pixel in a scan-line order, based on features extracted from the Gaussian pyramids of $A$, $A'$ and $B$. When $A$ is the LRI and $A'$ is the corresponding HRI, the algorithm outputs an HRI $B'$ for image $B$.

By assuming that the manifolds of LR patches and HR patches have similar local geometry, i.e., if an input LR patch is a linear combination of its $k$-nearest neighbors in the LR training patches, then its HR patch should roughly be the linear combination of corresponding HR training patches using the same combination coefficients, Liu et al. (2005a, 2005b, 2005c) develop algorithms for face hallucination, where the intermediate HR patches and the HR residue are inferred successively. And Chang et al. (2004, 2006) and Fan and Yeung (2007) apply similar methodology to generic images.

There are also several papers that utilize neural networks (NNs) for SR. For example, Zhang and Pan (2002) train an adaptive linear NN that maps LR residual errors to HR residual errors. Candocia and Principe (1999) use an NN to cluster LR patches as well as infer HR patches from LR patches. Miravet and Rodríguez (2003) use a hybrid multilayer perception (MLP) and probabilistic neural network (PNN) for SR via nonuniform interpolation. Kursun and Favorov (2003) build a biologically inspired NN that mimics the SINBAD (Set of INteracting BAckpropagating Dendrites) cells in the visual cortex, and claim that the SINBAD NN can identify high-order regularities in natural images. Therefore, the SINBAD NN can have excellent SR performance when interpolating the missing pixels.

## 3 Limits of Learning-Based SR Algorithms

Figure 1 outlines our analysis on the limits of learning-based SR algorithms. We first define an expected risk of a learning-based SR algorithm. The risk is minimized by an optimal SR function. Using the statistics of general natural images, we derive a closed form formula for the lower bound of the risk, which only involves the covariance matrix and the mean

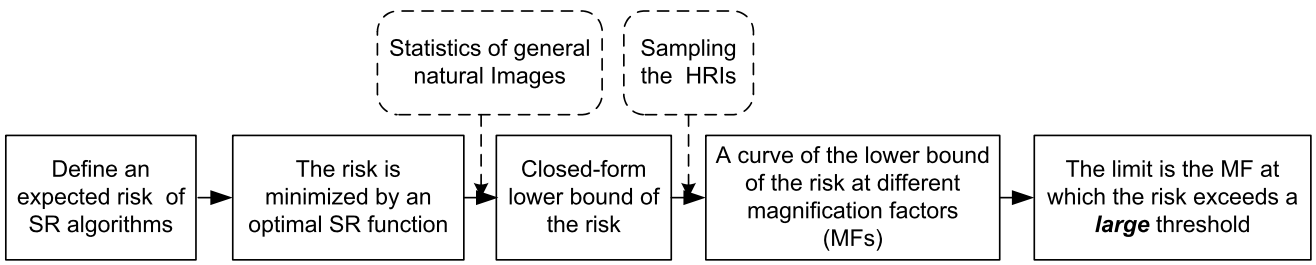


**Fig. 1** Our methodology of estimating the limits of learning-based SR algorithms. Please refer to Sect. 3 for more details

of the distribution of the HRIs. By sampling the real-world HRIs, we can obtain a curve of the lower bound of the risk w.r.t. the magnification factor. Finally, by choosing a relatively large threshold for the lower bound of the risk, we can roughly estimate the limit of learning-based SR algorithms.

### 3.1 What are the Limits of Learning-Based SR Algorithms?

To investigate the limits of LBAs quantitatively, we have to consider this problem in an abstract level in order to build a mathematical model. Although the existing LBAs differ in implementation details, they are simply different functions that map LRIs to corresponding HRIs, i.e., mappings $s$ from a low-dimensional Euclidean space to a high-dimensional one. If we define a function $d$ that downsamples HRIs to LRIs, then the composite function $f = s \circ d$ is a mapping between HRIs, and different LBAs correspond to different $f$'s. Let $A$ be the set of admissible $f$'s: $A = \{f(\cdot, \alpha)|\alpha \in I\}$, where $I$ is an index set for differentiating different $f$'s (Vapnik 1998). Then designing an LBA can be translated into determining a function in $A$. Hence each LBA is associated with an index $\alpha$. Given an HRI, the performance of an LBA $f(\cdot, \alpha)$ should be evaluated by how good this HRI is recovered. In a language of statistical learning theory (Vapnik 1998), this can be depicted by a risk function $r$. As an LBA performs differently for different HRIs (e.g., it is usually more difficult to recover HRIs with more details), we should look at its *average* performance, which corresponds to its expected risk (Vapnik 1998):

$$R(\alpha) = \int r(h, f(h, \alpha))\, p(h) \mathrm{d}h, \tag{2}$$

where $p(h)$ is the probability density function of the HRIs $h$. It is possible that an LBA performs extremely well on a particular LRI. However, if the SR results on many other LRIs are poor, we still do not consider it a good SR algorithm. That is why the average performance is a preferred criterion to evaluate an LBA. More generally, in statistical learning theory the expected risk is often used to evaluate the performance of a learning function.

Now we have to define an appropriate risk function for LBAs. As suggested in Lin and Shum (2004), a good SR algorithm should produce HRIs that are close to the ground truth. Otherwise, the produced HRI will not be what we desire, no matter how high its resolution is (e.g., an HR car image will not be considered as the HRI of an LR face image no matter how many details it presents). Therefore, we may define the risk function as the closeness between an HRI and its superresolved version. As the root mean squared error (RMSE) is a widely used measure of image similarity in the image processing community (e.g., the peak signal to noise ratio in image compression) and also in various kinds of error analysis, we may define the risk function using the RMSE between an HRI and its superresolved version.

Although small RMSEs do not necessarily guarantee good recovery of the HRIs, large RMSEs should nonetheless imply that the recovery is poor. Therefore, we may convert the problem to a tractable one: find the upper bound of the magnification factors such that the expected risk is below a relatively *large* threshold. Such an upper bound can be considered the limit of learning-based SR algorithms.

### 3.2 Problem Formulation

For simplicity, we present the arguments for the 1D case only. Those for the 2D case are similar but the expressions are much more complex.

As argued in Sect. 3.1, we use the RMSE between the recovered HRI and the ground truth HRI to evaluate the performance of a learning-based SR algorithm. This motivates us to define the following expected risk of the SR algorithm:[1]

$$\begin{aligned} g(N, m) &= \left(\frac{1}{mN}\tilde{g}(N, m)\right)^{\frac{1}{2}}, \quad \text{where} \\ \tilde{g}(N, m) &= \int_{\mathbf{h}} \|\mathbf{h} - s(\mathbf{D}\mathbf{h})\|^2 \, p_h(\mathbf{h}) \mathrm{d}\mathbf{h}, \end{aligned} \tag{3}$$

in which $s$ is the learnt SR function that maps $N$-dimensional images to $mN$-dimensional ones, $m > 1$ is the magnification factor and always makes $mN$ an integer, $p_h$ is the probability density function of the HRIs and $\mathbf{D}$ is the downsampling matrix that downsamples $mN$-dimensional signals to $N$-dimensional ones. The downsampling matrix is introduced here to simulate the image formation process. One may argue that the downsampling matrix may not be identical for different images, or the downsampling process may even be nonlinear. However, all these kinds of discrepancy can be viewed as part of the noise, which would be discussed in Sect. 6.

Equation (3) defines the expected risk of a particular SR algorithm $s$, which should be evaluated by running the algorithm on a large number of HRIs. This is very time consuming. Moreover, for a particular SR algorithm, its magnification factor is often fixed. Therefore, estimating the expected risk of a *particular* SR function does not help to find the limits of *all* learning-based SR algorithms. Consequently, we have to study the lower bound of (3).

Before going on, we first introduce the corresponding upsampling matrix $\mathbf{U}$ which upsamples $N$-dimensional signals to $mN$-dimensional ones. We expect that images are unchanged if they are upsampled and then downsampled. This

[1] Throughout our paper, vectors or matrices are written in boldface, while scalars are in normal fonts. Moreover, all the vectors without the transpose are column vectors.

implies that $\mathbf{DU} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. This upsampling matrix is purely a mathematical tool to facilitate the derivation and the representation of our results.

### 3.3 Theorem on the Lower Bound of the Expected Risk

The central theorem of our paper is the following:

**Theorem 3.1** (Lower Bound of the Expected Risk) *When $p_h(\mathbf{h})$ is the distribution of general natural images, namely the set of* all *natural images, $\tilde{g}(N,m)$ is effectively lower bounded by $\tilde{b}(N,m)$, where*

$$\tilde{b}(N,m) = \frac{1}{4}\operatorname{tr}\left[(\mathbf{I}-\mathbf{UD})\Sigma(\mathbf{I}-\mathbf{UD})^t\right] + \frac{1}{4}\left\|(\mathbf{I}-\mathbf{UD})\bar{\mathbf{h}}\right\|^2, \tag{4}$$

*in which* $\operatorname{tr}(\cdot)$ *is the trace operator, the superscript $^t$ represents the matrix or vector transpose, and $\Sigma$ and $\bar{\mathbf{h}}$ denote the covariance matrix and the mean of the HRIs* $\mathbf{h}$*, respectively. Hence $g(N,m)$ is lower bounded by*

$$b(N,m) = \left(\frac{1}{mN}\tilde{b}(N,m)\right)^{\frac{1}{2}}. \tag{5}$$

For an HRI $\mathbf{h}$, $(\mathbf{I}-\mathbf{UD})\mathbf{h} = \mathbf{h} - \mathbf{U}(\mathbf{Dh})$ is its high frequency. So (4) is essentially related to the richness of the high frequency component in the HRIs. Hence Theorem 3.1 implies that the richer the high frequency component in the HRIs is, the more difficult the SR is. This agrees with our intuition.

Note that Theorem 3.1 holds for all possible SR functions $s$ as it gives the lower bound of the risk. By investigating the lower bound, we are freed from taking care of the details of different learning-based SR algorithms.

### 3.4 Sketch of the Proof

In this subsection, we present the idea of proving Theorem 3.1. Now that different HRIs can result in the same LRI ($\mathbf{Dh}$ can be identical for different $\mathbf{h}$), it may be easier to analyze (3) by fixing $\mathbf{Dh}$. This can be achieved by performing a variable transform in (3). To do so, we find a complementary matrix (not unique) $\mathbf{Q}$ such that $\binom{\mathbf{D}}{\mathbf{Q}}$ is a non-singular square matrix and $\mathbf{QU} = \mathbf{0}$. Such a $\mathbf{Q}$ exists. The proof can be found in Appendix. Denote $\mathbf{M} = (\mathbf{R}\ \mathbf{V}) = \binom{\mathbf{D}}{\mathbf{Q}}^{-1}$. From $\binom{\mathbf{D}}{\mathbf{Q}}(\mathbf{R}\ \mathbf{V}) = \mathbf{I}$, we know that $\mathbf{R} = \mathbf{U}$.

Now we perform a variable transform $\mathbf{h} = \mathbf{M}\binom{\mathbf{x}}{\mathbf{y}}$. Then (3) becomes

$$\begin{aligned}\tilde{g}(N,m) &= \int_{\mathbf{x},\mathbf{y}} \left\|(\mathbf{U}\quad \mathbf{V})\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} - s(\mathbf{x})\right\|^2 p_{x,y}\left(\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}\right)\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}\\ &= \int_{\mathbf{x}} p_x(\mathbf{x}) V(\mathbf{x})\mathrm{d}\mathbf{x},\end{aligned} \tag{6}$$

where

$$p_{x,y}\left(\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}\right) = |\mathbf{M}| p_h\left(\mathbf{M}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}\right),$$

$$V(\mathbf{x}) = \int_{\mathbf{y}} \|\mathbf{Vy} - \phi(\mathbf{x})\|^2 \tilde{p}_y(\mathbf{y}|\mathbf{x})\,\mathrm{d}\mathbf{y}.$$

$p_x(\mathbf{x})$ is the marginal distribution of $\mathbf{x}$, $\tilde{p}_y(\mathbf{y}|\mathbf{x})$ is the conditional distribution of $\mathbf{y}$ given $\mathbf{x}$ and $\phi(\mathbf{x}) = s(\mathbf{x}) - \mathbf{Ux}$ is the recovered high frequency component of the HRI given the LRI $\mathbf{x}$. For this reason, we call $\phi(\mathbf{x})$ the high frequency function. Note that $\mathbf{x} = \mathbf{Dh}$, and $\mathbf{Vy} = \mathbf{h} - \mathbf{Ux}$. So $\mathbf{x}$ is the LRI downsampled from $\mathbf{h}$, and $\mathbf{Vy}$ is the high frequency of $\mathbf{h}$.

One can see that there is an optimal high frequency function such that $V(\mathbf{x})$ (hence $g(N,m)$) is minimized:

$$\phi_{opt}(\mathbf{x};\tilde{p}_y) = \mathbf{V}\int_{\mathbf{y}} \mathbf{y}\tilde{p}_y(\mathbf{y}|\mathbf{x})\,\mathrm{d}\mathbf{y}, \tag{7}$$

where $\phi_{opt}(\mathbf{x};p)$ denotes the optimal $\phi$ w.r.t. the distribution $p$. This means that the optimal high frequency component should be the expectation of all possible high frequencies associated to the LRI $\mathbf{x}$. The introduction of the optimal high frequency function (or equivalently, the optimal SR function, since $s_{opt}(\mathbf{x}) = \phi_{opt}(\mathbf{x}) + \mathbf{Ux}$) frees us from dealing with the details of different learning-based SR algorithms, because $s_{opt}$ attains the minimum of the expected risk.

Then one can easily verify that

$$V(\mathbf{x}) = \int_{\mathbf{y}} \|\mathbf{Vy}\|^2 \tilde{p}_y(\mathbf{y}|\mathbf{x})\,\mathrm{d}\mathbf{y} - \left\|\phi_{opt}(\mathbf{x};\tilde{p}_y)\right\|^2. \tag{8}$$

In Appendix, we show that for general natural images,

$$\left\|\phi_{opt}(\mathbf{x};\tilde{p}_y)\right\|^2 \le \frac{3}{4}\frac{\int_{\mathbf{y}} \|\mathbf{Vy}\|^2 p_{x,y}\left(\binom{\mathbf{x}}{\mathbf{y}}\right)\mathrm{d}\mathbf{y}}{p_x(\mathbf{x})}. \tag{9}$$

Therefore, from (6), (8) and (9) we have that

$$\begin{aligned}\tilde{g}(N,m) &= \int_{\mathbf{x}} p_x(\mathbf{x})\left(\int_{\mathbf{y}} \|\mathbf{Vy}\|^2 \tilde{p}_y(\mathbf{y}|\mathbf{x})\,\mathrm{d}\mathbf{y}\right.\\ &\qquad \left. - \left\|\phi_{opt}(\mathbf{x};\tilde{p}_y)\right\|^2\right)\mathrm{d}\mathbf{x}\\ &\ge \frac{1}{4}\int_{\mathbf{x}} p_x(\mathbf{x})\int_{\mathbf{y}} \|\mathbf{Vy}\|^2 \tilde{p}_y(\mathbf{y}|\mathbf{x})\,\mathrm{d}\mathbf{y}\mathrm{d}\mathbf{x}\\ &= \frac{1}{4}\int_{\mathbf{x},\mathbf{y}} \|\mathbf{Vy}\|^2 p_{x,y}\left(\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}\right)\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}\\ &= \frac{1}{4}\int_{\mathbf{h}} \|\mathbf{VQh}\|^2 p_h(\mathbf{h})\mathrm{d}\mathbf{h}\\ &= \frac{1}{4}\operatorname{tr}\left((\mathbf{I}-\mathbf{UD})\Sigma(\mathbf{I}-\mathbf{UD})^t\right) + \frac{1}{4}\left\|(\mathbf{I}-\mathbf{UD})\bar{\mathbf{h}}\right\|^2,\end{aligned}$$

where we have used $\mathbf{VQ} = \mathbf{I} - \mathbf{UD}$, which comes from $(\mathbf{U}\ \mathbf{V})\binom{\mathbf{D}}{\mathbf{Q}} = \mathbf{I}$. This proves Theorem 3.1.

We see that the variance and the mean of the HRIs plays a key role in lower bounding $g(N,m)$. Although it is intuitive that $p_h(\mathbf{h})$ is critical for the limits of learning-based SR algorithms, Theorem 3.1 quantitatively depicts how $p_h(\mathbf{h})$ influences the SR performance.

### 3.5 Limits of Learning-Based SR Algorithms

If at a particular magnification factor, $b(N,m)$ (see (5)) is larger than a threshold $T$, i.e., the expected RMSE between $\mathbf{h}$ and $s_{opt}(\mathbf{Dh})$ is larger than $T$, then for *any* SR function $s$, the RMSE between $\mathbf{h}$ and $s(\mathbf{Dh})$ is also expected to be larger than $T$ because $b(N,m)$ is the lower bound of the expected risk. This implies that at this magnification factor no SR function can effectively recover the original HRI.

Therefore, if we have full knowledge of the variance and the mean of the prior distributions $p_h(\mathbf{h})$ at different magnification factors, we can define a curve of $b(N,m)$ as a function of $m$. Then the limit of learning-based SR algorithms is upper bounded by $b^{-1}(T)$.

Note that the lower bound $b(N,m)$ is proved in a conservative way. Consequently, the estimate on the limits of learning-based SR algorithms using (5) is also conservative. And also note that $p_h(\mathbf{h})$ being the distribution of the set of all natural images is important for us to arrive at (4). Otherwise, we will not come up with the coefficient $1/4$ therein and $\tilde{g}(N,m)$ may be arbitrarily close to 0. For example, if there is only one HRI, we can always recover the HRI no matter of how low resolution the input LRI is.

## 4 The Sufficient Number of HRI Samples

To compute $b(N,m)$ using (4), we have to know the covariance matrix and the mean of HRIs $\mathbf{h}$ for general natural images. There has been a long history of natural image statistics (Srivastava et al. 2003). Unfortunately, all the existing models only solve the problem partially: the natural images fit some models, but not all images that are sampled from these models are natural images. On the other hand, we do not need the full knowledge of $p_h(\mathbf{h})$: its covariance matrix and mean already suffice.

This motivates us to sample HRIs from real data, because accurately estimating the mean and the variance of $p_h(\mathbf{h})$ by sampling is relatively easy. To make sure that sufficient images have been sampled to achieve an accurate estimate of $\tilde{b}(N,m)$, we provide a theorem below.

### 4.1 Theorem on the Sufficient Number of HRI Samples

**Theorem 4.1** (Sufficient Number of Samples) *If we sample $M(p,\varepsilon)$ HRIs independently, then with probability of at least $1-p$, $|\hat{\tilde{b}}(N,m)-\tilde{b}(N,m)|<\varepsilon$, where $\hat{\tilde{b}}(N,m)$ is the value of $\tilde{b}(N,m)$ estimated from real samples,*[2]

$$M(p,\varepsilon)=\frac{(C_1+2C_2)^2}{16p\varepsilon^2}, \tag{10}$$

$C_1=\sqrt{E(\|(\mathbf{I}-\mathbf{UD})(\mathbf{h}-\bar{\mathbf{h}})\|^4)-\mathrm{tr}^2[(\mathbf{I}-\mathbf{UD})\Sigma(\mathbf{I}-\mathbf{UD})^t]}$, *and* $C_2=\sqrt{\bar{\mathbf{b}}^t\Sigma\bar{\mathbf{b}}}$, *in which $E(\cdot)$ is the expectation operator and* $\bar{\mathbf{b}}=(\mathbf{I}-\mathbf{UD})^t(\mathbf{I}-\mathbf{UD})\bar{\mathbf{h}}$.

Note that both $C_1$ and $C_2$ are related to the variance of the high frequency component of the HRIs. So Theorem 4.1 implies that the larger the variance is, the more samples are required.

### 4.2 Sketch of the Proof

We first denote $\hat{\bar{\Sigma}}_M=\frac{1}{M}\sum_{k=1}^{M}(\hat{\mathbf{h}}_k-\bar{\mathbf{h}})(\hat{\mathbf{h}}_k-\bar{\mathbf{h}})^t$ and the estimated covariance matrix $\hat{\Sigma}_M=\frac{1}{M}\sum_{k=1}^{M}(\hat{\mathbf{h}}_k-\hat{\bar{\mathbf{h}}})(\hat{\mathbf{h}}_k-\hat{\bar{\mathbf{h}}})^t$,[3] where $\hat{\mathbf{h}}_k$'s are i.i.d. samples and $\hat{\bar{\mathbf{h}}}=\frac{1}{M}\sum_{k=1}^{M}\hat{\mathbf{h}}_k$ is the estimated mean. Then one may check that

$$\hat{\Sigma}_M=\hat{\bar{\Sigma}}_M-(\hat{\bar{\mathbf{h}}}_M-\bar{\mathbf{h}})(\hat{\bar{\mathbf{h}}}_M-\bar{\mathbf{h}})^t.$$

In the following, we denote $\mathbf{B}=(\mathbf{I}-\mathbf{UD})^t(\mathbf{I}-\mathbf{UD})$ and $\bar{\mathbf{b}}=\mathbf{B}\bar{\mathbf{h}}$ for brevity, and denote the $i$-th entry of a vector $\mathbf{a}$ as $a_i$ and the $(i,j)$-th entry of a matrix $\mathbf{A}$ as $A_{ij}$.

With some calculation we have

$$|\hat{\tilde{b}}(N,m)-\tilde{b}(N,m)|\le\frac{1}{4}\left|\sum_{i,j=1}^{mN}B_{ij}\left(\hat{\bar{\Sigma}}_{M;ij}-\Sigma_{ij}\right)\right|+\frac{1}{2}\left|\sum_{i=1}^{mN}\bar{b}_i(\hat{\bar{h}}_{M;i}-\bar{h}_i)\right|. \tag{11}$$

The details can be found in Appendix. So we have to estimate the convergence rates of both terms.

Therefore, we define $\xi=\sum_{i,j=1}^{mN}B_{ij}\hat{\bar{\Sigma}}_{M;ij}=\mathrm{tr}(\mathbf{B}\hat{\bar{\Sigma}})$ and $\eta=\sum_{i=1}^{mN}\bar{b}_i\hat{\bar{h}}_{M;i}=\bar{\mathbf{b}}^t\hat{\bar{\mathbf{h}}}_M$. Then their expectations are

$$E(\xi)=\mathrm{tr}(\mathbf{B}\Sigma),\quad\text{and}\quad E(\eta)=\bar{\mathbf{b}}^t\bar{\mathbf{h}},$$

respectively. And their variances can be found to be

$$\mathrm{var}(\xi)=\frac{C_1^2}{M}, \tag{12}$$

[2] Throughout our paper, we use the embellishment $\wedge$ above a value to represent the sampled or estimated quantities.

[3] For unbiased estimation of the covariance matrix, the coefficient before the summation should be $1/(M-1)$. However, we are more interested in the error of $\tilde{b}(N,m)$, rather than the covariance matrix itself. If $1/(M-1)$ is used instead of $1/M$, there will be an additional $O(1/M)$ term at the right hand side of (15), indicating that the convergence might be slightly slower. Nonetheless, when $M$ is large, the difference is negligible.

and

$$\mathrm{var}(\eta) = \frac{C_2^2}{M}, \tag{13}$$

respectively. The proofs can be found in Appendix.

Then by Chebyshev's inequality (Shiryaev 1995),

$$P\left(\left|\sum_{i,j=1}^{mN} B_{ij}\left(\hat{\hat{\Sigma}}_{M;ij} - \Sigma_{ij}\right)\right| \geq \delta\right)$$
$$= P\left(|\xi - E(\xi)| \geq \delta\right) \leq \frac{\mathrm{var}(\xi)}{\delta^2} = \frac{C_1^2}{M\delta^2},$$
$$P\left(\left|\sum_{i=1}^{mN} \bar{b}_i\left(\hat{\bar{h}}_{M;i} - \bar{h}_i\right)\right| \geq \delta\right)$$
$$= P\left(|\eta - E(\eta)| \geq \delta\right) \leq \frac{\mathrm{var}(\eta)}{\delta^2} = \frac{C_2^2}{M\delta^2}.$$

Therefore, at least at a probability of $1 - p$,

$$\left|\sum_{i,j=1}^{mN} B_{ij}\left(\hat{\hat{\Sigma}}_{M;ij} - \Sigma_{ij}\right)\right| \leq \frac{C_1}{\sqrt{Mp}},$$
$$\left|\sum_{i=1}^{mN} \bar{b}_i\left(\hat{\bar{h}}_{M;i} - \bar{h}_i\right)\right| \leq \frac{C_2}{\sqrt{Mp}}. \tag{14}$$

Then by (11) and (14), with probability at least $1 - p$, we have

$$\left|\hat{\tilde{b}}(N,m) - \tilde{b}(N,m)\right| \leq \frac{C_1}{4\sqrt{Mp}} + \frac{C_2}{2\sqrt{Mp}}. \tag{15}$$

Now one can check that Theorem 4.1 is true.

Note that

$$\left|\hat{b}(N,m) - b(N,m)\right|$$
$$= \left|\sqrt{\frac{1}{mN}\hat{\tilde{b}}(N,m)} - \sqrt{\frac{1}{mN}\tilde{b}(N,m)}\right|$$
$$= \frac{\frac{1}{mN}\left|\hat{\tilde{b}}(N,m) - \tilde{b}(N,m)\right|}{\sqrt{\frac{1}{mN}\hat{\tilde{b}}(N,m)} + \sqrt{\frac{1}{mN}\tilde{b}(N,m)}}$$
$$\approx \frac{\frac{1}{mN}\left|\hat{\tilde{b}}(N,m) - \tilde{b}(N,m)\right|}{2\sqrt{\frac{1}{mN}\tilde{b}(N,m)}}$$
$$= \frac{\left|\hat{\tilde{b}}(N,m) - \tilde{b}(N,m)\right|}{2mNb(N,m)}.$$

So in practice, we may choose $p = 0.01$ and $\varepsilon = \frac{1}{2}mN \times b(N,m)$ in (10) in order to make $|\hat{b}(N,m) - b(N,m)| \leq 0.25$ at above 99% certainty. Here we choose 0.25 as the threshold because it is roughly the mean of the graylevel quantization error.

## 5 Experiments

*Collecting Samples* We crawled images from the web and collected 100,000+ images. They are of various kinds of scenes: cityscape, landscape, sports, portraits, etc. Therefore, our image library could be viewed as an i.i.d. sampling of general natural images. To sample $mN \times mN$ sized HRIs, we first convert each color image into a graylevel one, then decompose it into non-overlapping patches of size $mN \times mN$ with at least one pixel apart from each of them in order to ensure the mutual independence. So each patch can be regarded as a sample of HRIs of size $mN \times mN$. Then we blindly run our program to estimate the covariance and the mean of the HRIs, where $mN$ varies from 8 to 48 at a step size of 4. The number of samples is of the order of $10^6$ to $10^8$. While one may argue that this number may not be sufficient to estimate $p_h(\mathbf{h})$, recall that our goal is not on estimating $p_h(\mathbf{h})$; we are interested in the values of $b(N,m)$ only.

*Characteristics of* $\mathbf{b}(\mathbf{N},\mathbf{m})$ Next, we have to specify a downsampling matrix in order to compute the lower bound $b(N,m)$ by (5). (The upsampling matrix $\mathbf{U}$ is determined by $\mathbf{D}$. See Sect. 3.2.) We simply choose a downsampling matrix that corresponds to the bicubic B-spline filter.[4] Then the curves of $b(N,m)$ w.r.t. $m$ are shown in Fig. 2(a), where for each individual curve $N$ is fixed.

We can see that for fixed $N$, $b_N^{(1)}(m) = b(N,m)$ increases with $m$. A remarkable observation is that for different $N$'s, the curves in Fig. 2(a) coincide well with each other. This suggests that for general natural images $b(N,m)$ may be *independent* of $N$. Another interesting observation on Fig. 2(a) is that $b_N^{(1)}(m)$ seems to grow at the rate of $(m-1)^{1/2}$. The important implication from these observations is: we may estimate the limits of learning-based SR by trying relatively small sized images and small magnification factors, rather than trying large sized images and large magnification factors, which saves computation and memory without compromising the estimation accuracy.

[4] In the 1D case, a cubic filter can be written as:

$$k(x) = \begin{cases} (a+2)|x|^3 - (a+3)|x|^2 + 1, & \text{if } 0 \leq |x| \leq 1, \\ a|x|^3 - 5a|x|^2 + 8a|x| - 4a, & \text{if } 1 \leq |x| \leq 2, \\ 0, & \text{if } |x| > 2. \end{cases} \tag{16}$$

When $a = -1$, it is the cubic B-spline filter. The downsampling matrix for 2D images is the Kronecker product of the 1D downsampling matrices.

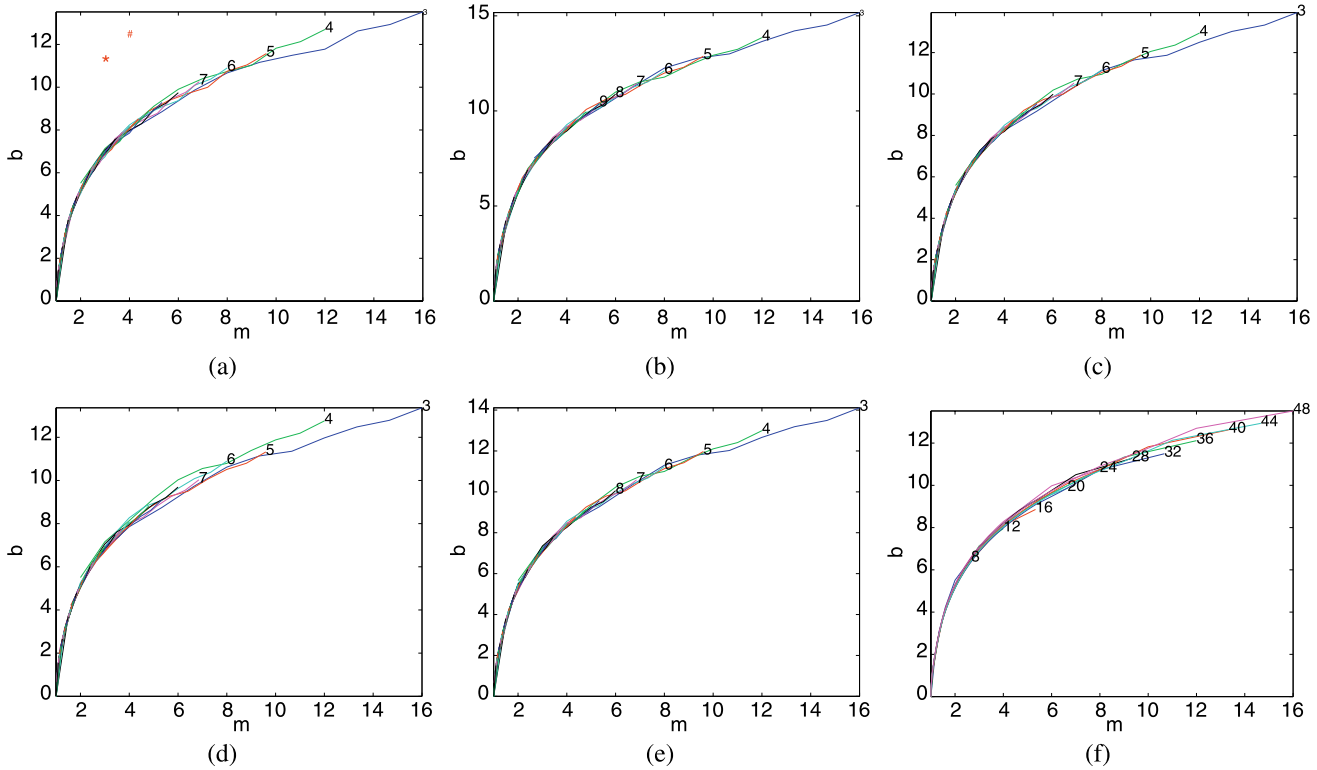


**Fig. 2** (**a**)–(**e**) are curves of $b(N, m)$ using different $\mathbf{D}$'s, drawn with $N$ fixed for each individual curve. The corresponding $N$'s are labelled at the tails of the curves (in order not to make the graph crowded, large $N$'s for short curves are not shown). (**a**) uses a bicubic filter with $a = -1$. The marks '*' at (3, 11.1) and '#' at (4, 12.6) represent the expected risks of Sun et al. (2003) and Freeman and Pasztor (1999) SR algorithms, respectively. (**b**) uses a bicubic filter with $a = 0.5$. (**c**) uses a Gaussian filter with $\sigma = 0.5$. (**d**) uses a Gaussian filter with $\sigma = 1.5$. (**e**) uses the bilinear filter. (**f**) are the curves of $b(N, m)$ with $mN$ fixed. The corresponding $mN$'s are labelled at the tails of the curves. The filter used is the same as that in (**a**)

However, one should be cautious that strictly speaking the $\mathbf{D}$ in (3) should be estimated from real cameras. Fortunately, we have found that our lower bound does not seem to be very sensitive to the choice of $\mathbf{D}$. We have tried the bilinear filter, Gaussian filters (with the variance varying from $0.5^2$ to $1.5^2$) and bicubic filters (with the parameter $a$ varying from $-1$ to 0.5, see (16)), and have found that the lower bounds are fairly close to each other. The curves in Figs. 2(a)–(e) testify to this observation. Moreover, what we have observed in the last paragraph is still true.

When training learning-based SR algorithms, one usually collects HRIs and downsamples them to LRIs. So it is also helpful to draw the curves by fixing $mN$ instead. The same phenomenon mentioned above can also be observed (Fig. 2(f)). And the curves of $b^{(2)}_{mN}(m) = b(N, m)$ by fixing $mN$ also coincide well with those of $b^{(1)}_{N}(m)$ (please compare Figs. 2(a) and (f)), implying that $b(N, m)$ is also independent of the size of HRIs. This can be easily proved: if $b(N, m) = c(m)$ for some function $c$, then $b^{(2)}_{mN}(m) = b(N, m) = b^{(1)}_{N}(m) = c(m)$.

*Testing Theorem 3.1* We run the SR algorithm by Sun et al. (2003) on over 50,000 $16 \times 16$ LRIs that are downsampled from $48 \times 48$ HRIs and that by Freeman and Pasztor (1999) on over 40,000 $12 \times 12$ LRIs that are also downsampled from $48 \times 48$ HRIs. Both algorithms are designed for general images and they work at magnification factors of 3.0 and 4.0, respectively. A few sample results are shown in Fig. 3. The expected risks of Sun et al.'s algorithm and Freeman and Pasztor's are about 11.1 and 12.6, respectively, which are both above our curves (Fig. 2(a)). Therefore, these results are consistent with Theorem 3.1.

*Estimating the Limits* With the curves of $b(N, m)$, we can find the limits of learning-based algorithms by choosing an appropriate threshold $T$ (see Sect. 3.5). Unfortunately, there does not seem to exist a benchmark threshold. So a practitioner may choose a threshold that he/she deems appropriate and then estimate the limits on his/her own. For example, from the SR results of Sun et al.'s algorithm (Sun et al. 2003) (Fig. 3), we see that the fine details are already missing. Therefore, we deem that the estimated risk 11.1 of their algorithm is a large enough threshold. Using $T = 11.1$ we can expect that the limit of learning-based SR algorithms for general natural images is roughly 10 (Fig. 2(a)). This limit

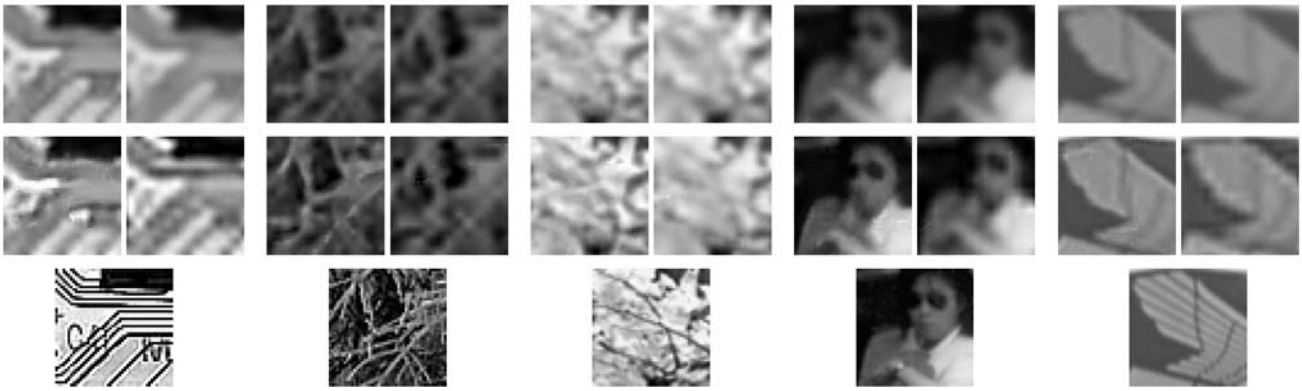

**Fig. 3** Part of the SR results using Sun et al.'s algorithm (Sun et al. 2003) (the magnification factor is 3.0) and Freeman and Pasztor's algorithm (Freeman and Pasztor 1999) (the magnification factor is 4.0). In each group of images, the *top left* one is the LRI of $16 \times 16$, interpolated to $48 \times 48$ using bicubic interpolation. The *middle left* one is the SR result by Sun et al.'s algorithm. The *top right* one is the LRI of $12 \times 12$, interpolated to $48 \times 48$ using bicubic interpolation. The *middle right* one is the SR result by Freeman and Pasztor's algorithm. At the *bottom* is the ground truth HRI

is a bit loose but it can be enhanced when the noise in LRIs (see Sect. 6) is considered.

*Testing Theorem 4.1* Finally, we present an experiment to test Theorem 4.1. We sample over 1.5 million $8 \times 8$ images and set $m = 2$ (hence $N = 4$). Figure 4 shows the curve of predicted sufficient number of samples using the most updated variance and mean of HRIs, where $p$ and $\varepsilon$ are chosen as described at the end of Sect. 4.2. We see that the estimated $b(4, 2)$ already becomes stable even the number of samples is still smaller than the predicted number. There is still small fluctuation in $b(4, 2)$ when $M > \hat{M}(p, \varepsilon)$ because we allow the deviation from the true value to be within 0.25 at above 99% certainty. Therefore, this result is consistent with Theorem 4.1.

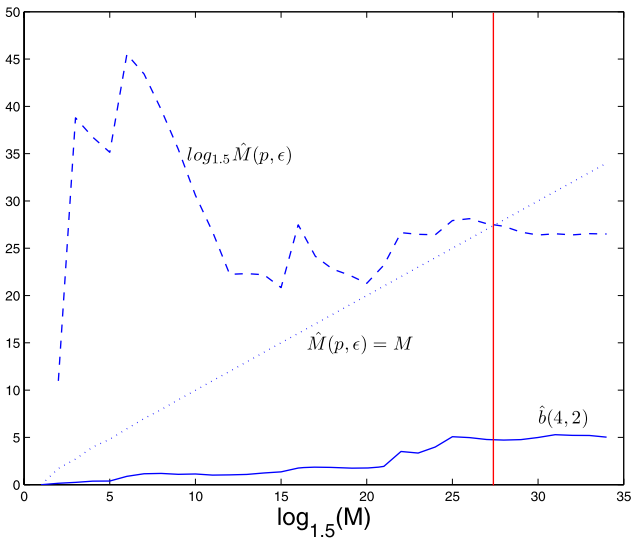


**Fig. 4** The evolution of $\hat{b}(4, 2)$ w.r.t. the number $M$ of HRI samples. The *dashed curve* is the log of the estimated sufficient samples using the currently available covariance matrix and mean. The *dotted line* is used to identify when the estimated number of samples is enough. The *solid curve* at the bottom is the estimated $b(4, 2)$ using $M$ samples. The *horizontal axis* is in log scale

## 6 Discussion on the Noise

In the above analysis, noise is neglected. To take noise into account, $\tilde{g}(m)$ should be changed to

$$\tilde{g}'(m) = \int_{\mathbf{h},\mathbf{n}} \|\mathbf{h} - s\,(\mathbf{D}\mathbf{h} + \mathbf{n})\|^2\, p_{h,n}\left(\begin{pmatrix}\mathbf{h}\\ \mathbf{n}\end{pmatrix}\right) \mathrm{d}\mathbf{h}\mathrm{d}\mathbf{n},$$

where $p_{h,n}$ is the joint probability density functions of the HRIs and the noise. Note that here the noise comes from two sources. One is the "real" noise in the LRIs. The other is the "virtual" noise that results from the inaccuracy of modeling the relationship between the LRIs and the HRIs with a single identical downsampling matrix $\mathbf{D}$ (see Sect. 3.2).

When noise is considered, by assuming the independency between the HRIs and the noise, (4) is changed to

$$\begin{aligned}\tilde{b}'(N, m) = {} & \frac{1}{4}\,\mathrm{tr}\left[(\mathbf{I} - \mathbf{U}\mathbf{D})\,\Sigma\,(\mathbf{I} - \mathbf{U}\mathbf{D})^t\right] \\ & + \frac{1}{4}\,\mathrm{tr}\left(\mathbf{U}\Sigma_n\mathbf{U}^t\right) \\ & + \frac{1}{4}\left\|(\mathbf{I} - \mathbf{U}\mathbf{D})\bar{\mathbf{h}} - \mathbf{U}\bar{\mathbf{n}}\right\|^2, \end{aligned} \tag{17}$$

where $\Sigma_n$ and $\bar{\mathbf{n}}$ are the variance matrix and the mean of the noise, respectively. We omit the details as the proof is analogous. As $\tilde{b}'(N, m) > \tilde{b}(N, m)$, a better estimate on the limit of LBAs is expected.

Unfortunately, according to our knowledge currently there is no good noise model for natural images for our purpose. And estimating the statistical property of the noise by sampling is a very difficult problem: different images can have different types and levels of noise. Things are even more complicated when the virtual noise is also considered. So currently we are still unable to performance experiments where noise is involved.

## 7 Conclusions and Future Work

This paper presents the first attempt to analyze the limits of learning-based SR algorithms. We have proven a closed form lower bound of the expected risk of SR algorithms. We also sample real images to estimate the lower bound. Finally, we prove the formula that gives the sufficient number of HRIs to be sampled in order to ensure the accuracy of the estimate.

We have also observed from experiments that the lower bound $b(N, m)$ may be dependent on $m$ only and the growth rate of $b(N, m)$ may be $(m-1)^{1/2}$. These are important observations, implying that one may more conveniently compute with small sized images and at small magnification factors and then *predict* the limits. This would save much computation and memory. We hope to prove this conjecture in the future.

As no authoritative threshold $T$ is currently available, our estimated limit (roughly 10 times) of learning-based SR algorithms for general natural images is not convincing enough. We are investigating how to propose an objective threshold and how to effectively sample the statistics of the noise in (17) to produce a tighter limit.

Also, we will investigate the limits of learning-based SR algorithms under more specific scenarios, e.g., for face hallucination and text SR. We expect that more specific prior knowledge of the HRI distribution will be required.

## Appendix

**Proposition 8.1** $\mathbf{Q}$ *exists*.

*Proof* Suppose the SVD of $\mathbf{U}$ is: $\mathbf{U} = \mathbf{O}_1 \binom{\Lambda}{\mathbf{0}} \mathbf{O}_2^t$, where $\Lambda$ is a non-degenerate square matrix. Then all the solutions to $\mathbf{XU} = \mathbf{0}$ can be written as: $\mathbf{X} = \mathbf{O}_2(\mathbf{0}\ \mathbf{Y})\mathbf{O}_1^t$, where $\mathbf{Y}$ is any matrix of proper size. On the other hand, from $\mathbf{DU} = \mathbf{I}$ we know that there exists some $\mathbf{Y}_0$ such that $\mathbf{D} = \mathbf{O}_2(\Lambda^{-1}\ \mathbf{Y}_0)\mathbf{O}_1^t$. Therefore, $\binom{\mathbf{D}}{\mathbf{Q}} = \mathbf{O}_2 \begin{pmatrix} \Lambda^{-1} & \mathbf{Y}_0 \\ \mathbf{0} & \mathbf{Y} \end{pmatrix} \mathbf{O}_1^t$. When $\mathbf{Y}$ is of full-rank, $\binom{\mathbf{D}}{\mathbf{Q}}$ is a non-degenerate square matrix. □

**Proposition 8.2** *Equation* (9) *is true*.

*Proof* The optimal high frequency function given in (7) is inconvenient for estimating a lower bound for $\tilde{g}(N, m)$, because we do not know $\mathbf{V}$ and $\tilde{p}_y(\mathbf{y}|\mathbf{x})$ therein. To overcome this, we assume that the probability density of HRIs is provided by the mixture of Gaussians (MoGs):

$$p_h(\mathbf{h}) = \sum_{k=1}^{K} \alpha_k G_{h;k}(\mathbf{h}),$$

where $\alpha_k > 0$, $\sum_{k=1}^{K} \alpha_k = 1$, and $G_{h;k}(\mathbf{h}) = G(\mathbf{h}; \mathbf{h}_k, \Sigma_k)$ is the Gaussian with mean $\mathbf{h}_k$ and variance $\Sigma_k$. Note that the above MoGs approximation may not give an exact $p_h(\mathbf{h})$. However, as every $L_2$ function can be approximated by MoGs at an arbitrary accuracy (in the sense of $L_2$ norm) (Wilson 2000), and $\mathbf{h} - s(\mathbf{Dh})$ must be bounded (e.g., every dimension is between $-255$ and 255), when the MoGs approximation is sufficiently accurate, we will give a sufficiently accurate estimate of $\tilde{g}(N, m)$. Therefore, in order not to introduce new notations, we simply write $p_h(\mathbf{h})$ as MoGs in our proof. More importantly, as we will see, MoGs actually serve as a bridge to pave our proving process. Our final results do *not* involve any parameters from MoGs, as shown in Theorem 3.1. If one wishes, a limit can be carried on throughout the proof to make everything rigorous.

Writing in MoGs, we have

$$p_{x,y}\left(\binom{\mathbf{x}}{\mathbf{y}}\right) = \sum_{k=1}^{K} \alpha_k G_{x,y;k}\left(\binom{\mathbf{x}}{\mathbf{y}}\right),$$

$$p_x(\mathbf{x}) = \sum_{k=1}^{K} \alpha_k G_{x;k}(\mathbf{x}),$$

$$\tilde{p}_y(\mathbf{y}|\mathbf{x}) = \frac{\sum_{k=1}^{K} \alpha_k G_{x;k}(\mathbf{x}) \tilde{G}_{y;k}(\mathbf{y}|\mathbf{x})}{\sum_{k=1}^{K} \alpha_k G_{x;k}(\mathbf{x})},$$

where $G_{x,y;k}\left(\binom{\mathbf{x}}{\mathbf{y}}\right)$ is the Gaussian corresponding to $G_{h;k}(\mathbf{h})$ after the variable transform, $G_{x;k}(\mathbf{x})$ is the marginal distribution of $G_{x,y;k}\left(\binom{\mathbf{x}}{\mathbf{y}}\right)$ and

$$\tilde{G}_{y;k}(\mathbf{y}|\mathbf{x}) = \frac{G_{x,y;k}\left(\binom{\mathbf{x}}{\mathbf{y}}\right)}{G_{x;k}(\mathbf{x})}$$

is the conditional distribution. As we will not use the exact formulation of $G_{x,y;k}\left(\binom{\mathbf{x}}{\mathbf{y}}\right)$, $G_{x;k}(\mathbf{x})$ and $\tilde{G}_{y;k}(\mathbf{y}|\mathbf{x})$, we omit their details.

Now $\phi_{opt}(\mathbf{x}; \tilde{p}_y)$ can be written as

$$\begin{aligned} \phi_{opt}(\mathbf{x}; \tilde{p}_y) &= \frac{\sum_{k=1}^{K} \alpha_k G_{x;k}(\mathbf{x}) \mathbf{V} \int_{\mathbf{y}} \mathbf{y} \tilde{G}_{y;k}(\mathbf{y}|\mathbf{x})\, \mathrm{d}\mathbf{y}}{\sum_{k=1}^{K} \alpha_k G_{x;k}(\mathbf{x})} \\ &= \frac{\sum_{k=1}^{K} \alpha_k G_{x;k}(\mathbf{x}) \phi_{opt}(\mathbf{x}; \tilde{G}_{y;k})}{\sum_{k=1}^{K} \alpha_k G_{x;k}(\mathbf{x})}, \end{aligned} \tag{18}$$

where

$$\phi_{opt}(\mathbf{x};\tilde{G}_{y;k}) = \mathbf{V}\int_{\mathbf{y}} \mathbf{y}\tilde{G}_{y;k}(\mathbf{y}|\mathbf{x})\,\mathrm{d}\mathbf{y}.$$

Next, we highlight two properties of general natural images, which will be used in our argument:

1. The prior distribution $p_h(\mathbf{h})$ is not concentrated around several HRIs and the marginal distribution $p_x(\mathbf{x})$ is not concentrated around several LRIs either. Noticing that general natural images cannot be classified into a small number of categories will testify to this. This property implies that the number $K$ of Gaussians to approximate $p_h(\mathbf{h})$ is not too small, and for every $\mathbf{x}$, $\phi_{opt}(\mathbf{x};\tilde{G}_{y;k})$, $k = 1,\ldots,K$, are most likely quite different from each other.
2. Smoother LRIs have higher probability. This property is actually called the "smoothness prior" that is widely used for regularization, e.g., when doing reconstruction-based SR. An ideal mathematical formulation of this property is: $p_x(\mathbf{x}) \sim \exp(-\frac{1}{2}\beta\|\nabla\mathbf{x}\|^2)$ for some $\beta > 0$ (Srivastava et al. 2003).

Now we utilize the above two properties to argue for (9). We first estimate a reasonable coefficient $\mu$, such that most likely the following inequality holds:

$$\left\|\phi_{opt}\left(\mathbf{x};\tilde{p}_y\right)\right\|^2 \le \mu\cdot\frac{\sum_{k=1}^K \alpha_k G_{x;k}(\mathbf{x})\|\phi_{opt}(\mathbf{x};\tilde{G}_{y;k})\|^2}{\sum_{k=1}^K \alpha_k G_{x;k}(\mathbf{x})},$$
$$\forall\mathbf{x}\text{ such that }\sum_{k=1}^K \alpha_k G_{x;k}(\mathbf{x}) \neq 0. \tag{19}$$

Equation (18) shows that $\phi_{opt}(\mathbf{x};\tilde{p}_y)$ is a convex combination of $\phi_{opt}(\mathbf{x};\tilde{G}_{y;k})$, $k = 1,\ldots,K$. Due to the convexity of the squared vector norm, by Jensen's inequality (Shiryaev 1995), we have that $\mu \le 1$ is always true, where $\mu = 1$ holds only when $\phi_{opt}(\mathbf{x};\tilde{G}_{y;k})$, $k = 1,\ldots,K$, are identical. This will not happen due to the first property of general natural images. Another extreme case is $\mu = 0$. This happens only when $\phi_{opt}(\mathbf{x};\tilde{p}_y) = \mathbf{0}$. This will not happen either as this implies that the simple interpolation $s_{opt}(\mathbf{x}) = \mathbf{U}\mathbf{x}$ produces the optimal HRI.

Therefore, for general natural images $\mu$ can be close to neither 0 nor 1. We also notice that the strong convexity of the squared norm (thinking in 1D, there is large vertical gap between the curve $y = x^2$ and the line segment linking $(x_1, x_1^2)$ and $(x_2, x_2^2)$ when $x_1$ and $x_2$ is not close to each other) implies that the scattering of $\phi_{opt}(\mathbf{x};\tilde{G}_{y;k})$, $k = 1,\ldots,K$, will make $\|\phi_{opt}(\mathbf{x};\tilde{p}_y)\|^2$ *far below* the weighted squared norms of $\phi_{opt}(\mathbf{x};\tilde{G}_{y;k})$, $k = 1,\ldots,K$. This implies that although $\mu$ could be a random number between 0 and 1, it should nevertheless *strongly* bias towards 0, i.e., the probability of $0 < \mu \le 0.5$ should be *much larger* than that of $0.5 < \mu < 1$. For those $\mathbf{x}$ whose $\mu$ is closer to 1, their corresponding $\phi_{opt}(\mathbf{x};\tilde{G}_{y;k})$, $k = 1,\ldots,K$, should be quite cluttered, implying that there is not much choice of adding different high frequency to recover different HRIs. This more likely happens when $\mathbf{x}$ itself is highly textured so that the high frequency is already constrained by the context of the image. Then by the second property of general natural images, such LRIs $\mathbf{x}$ have smaller probability than those requiring smaller $\mu$.

Summing up the bias of $\mu$ and $p_x(\mathbf{x}) = \sum_{k=1}^K \alpha_k G_{x;k}(\mathbf{x})$, we deem that the value 3/4 is sufficient for $\mu$.[5] To further safeguard the upper bound for $\|\phi_{opt}(\mathbf{x};\tilde{p}_y)\|^2$ and obtain a concise mathematical formulation in Theorem 3.1, we add an extra nonnegative term to the right hand side of (19), i.e.,

$$\left\|\phi_{opt}\left(\mathbf{x};\tilde{p}_y\right)\right\|^2 \le \frac{3}{4}\frac{\sum_{k=1}^K \alpha_k G_{x;k}(\mathbf{x})(\|\phi_{opt}(\mathbf{x};\tilde{G}_{y;k})\|^2 + \int_{\mathbf{y}}\|\mathbf{V}\mathbf{y}-\phi_{opt}(\mathbf{x};\tilde{G}_{y;k})\|^2\tilde{G}_{y;k}(\mathbf{y}|\mathbf{x})\mathrm{d}\mathbf{y})}{\sum_{k=1}^K \alpha_k G_{x;k}(\mathbf{x})}$$
$$= \frac{3}{4}\frac{\sum_{k=1}^K \alpha_k G_{x;k}(\mathbf{x})\int_{\mathbf{y}}\|\mathbf{V}\mathbf{y}\|^2\tilde{G}_{y;k}(\mathbf{y}|\mathbf{x})\mathrm{d}\mathbf{y}}{\sum_{k=1}^K \alpha_k G_{x;k}(\mathbf{x})} = \frac{3}{4}\frac{\int_{\mathbf{y}}\|\mathbf{V}\mathbf{y}\|^2 p_{x,y}\left(\binom{\mathbf{x}}{\mathbf{y}}\right)\mathrm{d}\mathbf{y}}{p_x(\mathbf{x})}.$$

This proves (9). □

**Proposition 8.3** *Equation* (11) *is true.*

*Proof*

$$|\hat{\tilde{b}}(N,m) - \tilde{b}(N,m)|$$
$$= \frac{1}{4}\left|\mathrm{tr}\left[(\mathbf{I}-\mathbf{U}\mathbf{D})\left(\hat{\Sigma}_M - \Sigma\right)(\mathbf{I}-\mathbf{U}\mathbf{D})^t\right]\right.$$
$$\left. + \left\|(\mathbf{I}-\mathbf{U}\mathbf{D})\hat{\bar{\mathbf{h}}}_M\right\|^2 - \left\|(\mathbf{I}-\mathbf{U}\mathbf{D})\bar{\mathbf{h}}\right\|^2\right|$$

[5] Actually, we believe that 1/2 already suffices due to the strong bias resulting from the convexity of the squared norm. Here we choose a larger value of 3/4 for additional guarantee.

$$
\begin{aligned}
&= \frac{1}{4}\Big|\mathrm{tr}\Big[\mathbf{B}\Big(\hat{\tilde{\Sigma}}_M - \Sigma\Big)\Big] + \Big\|(\mathbf{I}-\mathbf{UD})[(\hat{\tilde{\mathbf{h}}}_M - \bar{\mathbf{h}}) + \bar{\mathbf{h}}]\Big\|^2 \\
&\quad - \Big\|(\mathbf{I}-\mathbf{UD})\bar{\mathbf{h}}\Big\|^2\Big| \\
&= \frac{1}{4}\Big|\mathrm{tr}\Big[\mathbf{B}\Big(\hat{\tilde{\Sigma}}_M - \Sigma\Big)\Big] + \mathrm{tr}\Big[\mathbf{B}\Big(\hat{\Sigma}_M - \hat{\tilde{\Sigma}}_M\Big)\Big] \\
&\quad + \Big\|(\mathbf{I}-\mathbf{UD})(\hat{\tilde{\mathbf{h}}}_M - \bar{\mathbf{h}})\Big\|^2 \\
&\quad + 2[(\mathbf{I}-\mathbf{UD})\bar{\mathbf{h}}]^t(\mathbf{I}-\mathbf{UD})\ (\hat{\tilde{\mathbf{h}}}_M - \bar{\mathbf{h}})\Big| \\
&= \frac{1}{4}\Bigg|\sum_{i,j=1}^{mN} B_{ij}\Big(\hat{\tilde{\Sigma}}_{M;ij} - \Sigma_{ij}\Big) \\
&\quad - \mathrm{tr}\Big[\mathbf{B}(\hat{\tilde{\mathbf{h}}}_M - \bar{\mathbf{h}})(\hat{\tilde{\mathbf{h}}}_M - \bar{\mathbf{h}})^t\Big] \\
&\quad + \Big\|(\mathbf{I}-\mathbf{UD})(\hat{\tilde{\mathbf{h}}}_M - \bar{\mathbf{h}})\Big\|^2 \\
&\quad + 2\Big[(\mathbf{I}-\mathbf{UD})^t(\mathbf{I}-\mathbf{UD})\bar{\mathbf{h}}\Big]^t(\hat{\tilde{\mathbf{h}}}_M - \bar{\mathbf{h}})\Bigg| \\
&= \frac{1}{4}\Bigg|\sum_{i,j=1}^{mN} B_{ij}\Big(\hat{\tilde{\Sigma}}_{M;ij} - \Sigma_{ij}\Big) + 2\bar{\mathbf{b}}^t(\hat{\tilde{\mathbf{h}}}_M - \bar{\mathbf{h}})\Bigg| \\
&= \frac{1}{4}\Bigg|\sum_{i,j=1}^{mN} B_{ij}\Big(\hat{\tilde{\Sigma}}_{M;ij} - \Sigma_{ij}\Big) + 2\sum_{i=1}^{mN}\bar{b}_i(\hat{\tilde{h}}_{M;i} - \bar{h}_i)\Bigg| \\
&\le \frac{1}{4}\Bigg|\sum_{i,j=1}^{mN} B_{ij}\Big(\hat{\tilde{\Sigma}}_{M;ij} - \Sigma_{ij}\Big)\Bigg| + \frac{1}{2}\Bigg|\sum_{i=1}^{mN}\bar{b}_i(\hat{\tilde{h}}_{M;i} - \bar{h}_i)\Bigg|. \qquad \square
\end{aligned}
$$

**Proposition 8.4** *Equation* (12) *is true.*

*Proof*

$$
\begin{aligned}
&E(\xi^2) \\
&= E\left(\sum_{i,j,i',j'=1}^{mN} B_{ij}B_{i'j'}\hat{\tilde{\Sigma}}_{M;ij}\hat{\tilde{\Sigma}}_{M;i'j'}\right) \\
&= \frac{1}{M^2}\sum_{i,j,i',j'=1}^{mN} B_{ij}B_{i'j'}E\left(\left[\sum_{k=1}^{M}(\hat{h}_{k;i} - \bar{h}_i)(\hat{h}_{k;j} - \bar{h}_j)\right]\right. \\
&\quad \left.\times\left[\sum_{r=1}^{M}(\hat{h}_{r;i'} - \bar{h}_{i'})(\hat{h}_{r;j'} - \bar{h}_{j'})\right]\right) \\
&= \frac{1}{M^2}\sum_{i,j,i',j'=1}^{mN} B_{ij}B_{i'j'}\left\{\sum_{k=1}^{M}E\big[(\hat{h}_{k;i} - \bar{h}_i)\right. \\
&\quad \left.\times(\hat{h}_{k;j} - \bar{h}_j)(\hat{h}_{k;i'} - \bar{h}_{i'})(\hat{h}_{k;j'} - \bar{h}_{j'})\big]\right. \\
&\quad + 2E\Bigg[\sum_{1\le k<r\le M}(\hat{h}_{k;i} - \bar{h}_i)(\hat{h}_{k;j} - \bar{h}_j) \\
&\quad \left.\times(\hat{h}_{r;i'} - \bar{h}_{i'})(\hat{h}_{r;j'} - \bar{h}_{j'})\Bigg]\right\} \\
&= \frac{1}{M^2}\sum_{i,j,i',j'=1}^{mN} B_{ij}B_{i'j'}\big\{ME\big[(h_i - \bar{h}_i)(h_j - \bar{h}_j) \\
&\quad \times(h_{i'} - \bar{h}_{i'})(h_{j'} - \bar{h}_{j'})\big] \\
&\quad + M(M-1)E\big[(h_i - \bar{h}_i)(h_j - \bar{h}_j)\big] \\
&\quad \times E\big[(h_{i'} - \bar{h}_{i'})(h_{j'} - \bar{h}_{j'})\big]\big\} \\
&= \frac{1}{M}\Bigg\{E\Bigg[\sum_{i,j,i',j'=1}^{mN} B_{ij}B_{i'j'}(h_i - \bar{h}_i) \\
&\quad \times(h_j - \bar{h}_j)(h_{i'} - \bar{h}_{i'})(h_{j'} - \bar{h}_{j'})\Bigg] \\
&\quad + (M-1)\sum_{i,j,i',j'=1}^{mN} B_{ij}B_{i'j'}\Sigma_{ij}\Sigma_{i'j'}\Bigg\} \\
&= \frac{1}{M}\Big\{E\Big[\mathrm{tr}^2\big(\mathbf{B}(\mathbf{h}-\bar{\mathbf{h}})(\mathbf{h}-\bar{\mathbf{h}})^t\big)\Big] + (M-1)[\mathrm{tr}(\mathbf{B}\Sigma)]^2\Big\}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\mathrm{var}(\xi) &= E(\xi^2) - [E(\xi)]^2 \\
&= \frac{1}{M}\big\{E\big[\|(\mathbf{I}-\mathbf{UD})(\mathbf{h}-\bar{\mathbf{h}})\|^4\big] - [\mathrm{tr}(\mathbf{B}\Sigma)]^2\big\}. \qquad \square
\end{aligned}
$$

**Proposition 8.5** *Equation* (13) *is true.*

*Proof*

$$
\begin{aligned}
&\mathrm{var}(\eta) \\
&= E\left(\left(\sum_{i=1}^{mN}\bar{b}_i(\hat{\tilde{h}}_{M;i} - \bar{h}_i)\right)^2\right) \\
&= E\left(\sum_{i,j=1}^{mN}\bar{b}_i\bar{b}_j(\hat{\tilde{h}}_{M;i} - \bar{h}_i)(\hat{\tilde{h}}_{M;j} - \bar{h}_j)\right) \\
&= \frac{1}{M^2}\sum_{i,j=1}^{mN}\bar{b}_i\bar{b}_jE\left(\sum_{k=1}^{M}(\hat{h}_{k;i} - \bar{h}_i)\sum_{r=1}^{M}(\hat{h}_{r;j} - \bar{h}_j)\right)
\end{aligned}
$$

$$= \frac{1}{M^2} \sum_{i,j=1}^{mN} \bar{b}_i \bar{b}_j \left[ \sum_{k=1}^{M} E\left( (\hat{h}_{k;i} - \bar{h}_i)(\hat{h}_{k;j} - \bar{h}_j) \right) + \sum_{1 \le k < r \le M} E\left( (\hat{h}_{k;i} - \bar{h}_i)(\hat{h}_{r;j} - \bar{h}_j) \right) \right]$$

$$= \frac{1}{M^2} \sum_{i,j=1}^{mN} \bar{b}_i \bar{b}_j (M \Sigma_{ij})$$

$$= \frac{1}{M} \bar{\mathbf{b}}^t \Sigma \bar{\mathbf{b}}. \qquad \square$$

## References


Baker, S., & Kanade, T. (2002). Limits on super-resolution and how to break them. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(9), 1167–1183.

Bégin, I., & Ferrie, F. R. (2004). Blind super-resolution using a learning-based approach. In *Proc. int. conf. pattern recognition*, August 2004 (Vol. 1, pp. 549–552).

Bishop, C. M., Blake, A., & Marthi, B. (2003). Super-resolution enhancement of video. In C.M. Bishop, & B. Frey (Eds.), *Proc. artificial intelligence and statistics*. Society for Artificial Intelligence and Statistics.

Borman, S., & Stevenson, R. L. (1998). *Spatial resolution enhancement of low-resolution image sequences: a comprehensive review with directions for future research* (Technical report). University of Notre Dame.

Candocia, F. M., & Principe, J. C. (1999). Super-resolution of images based on local correlations. *IEEE Transaction on Neural Network*, *10*(2), 372–380.

Capel, D., & Zisserman, A. (2001). Super-resolution from multiple views using learnt image models. In *Proc. computer vision and pattern recognition* (pp. II 627–634).

Chang, T.-L., Liu, T.-L., & Chuang, J.-H. (2006). Direct energy minimization for super-resolution on nonlinear manifolds. In *Proc. European conf. computer vision* (pp. IV 281–294).

Chang, H., Yeung, D., & Xiong, Y. (2004). Super-resolution through neighbor embedding. In *Proc. computer vision and pattern recognition* (Vol. 1, pp. 275–282).

Dedeoğlu, G., Kanade, T., & August, J. (2004). High-zoom video hallucination by exploiting spatio-temporal regularities. In *Proc. computer vision and pattern recognition* (pp. II 151–158).

Elad, M., & Feuer, A. (1997). Restoration of single super-resolution image from several blurred, noisy and down-sampled measured images. *IEEE Transactions on Image Processing*, *6*(12), 1646–1658.

Fan, W., & Yeung, D.-Y. (2007). Image hallucination using neighbor embedding over visual primitive manifolds. In *Proc. computer vision and pattern recognition*.

Farsiu, S., Robinson, D., Elad, M., & Milanfar, P. (2004). Advances and challenges in super-resolution. *International Journal of Imaging Systems and Technology*, *14*(2), 47–57.

Freeman, W. T., & Pasztor, E. C. (1999). Learning low-level vision. In *Proc. 7th int. conf. computer vision*, Corfu, Greece (pp. 1182–1189).

Gunturk, B. K., Batur, A. U., Altunbasak, Y., Hayes III, M. H., & Mersereau, R. M. (2003). Eigenface-domain super-resolution for face recognition. *IEEE Transactions on Image Processing*, *12*(5), 597–606.

Hardie, R. C., Barnard, K. J., & Amstrong, E. E. (1997). Joint map registration and high-resolution image estimation using a sequence of undersampled images. *IEEE Transactions on Image Processing*, *6*(12), 1621–1633.

Hertzmann, A., Jacobs, C. E., Oliver, N., Curless, B., & Salesin, D. H. (2001). Image analogies. In *SIGGRAPH*.

Kim, S. P., & Su, W. Y. (1993). Recursive high-resolution reconstruction of blurred multiframe images. *IEEE Transactions on Image Processing*, *2*, 534–539.

Komatsu, T., Aizawa, K., Igarashi, T., & Saito, T. (1993). A signal-processing based method for acquiring very high resolution image with multiple cameras and its theoretical analysis. *IEE Proceedings on Communications, Speech and Vision*, *140*(1), 19–25.

Kursun, O., & Favorov, O. (2003). Single-frame super-resolution by inference from learned features. *Istanbul University Journal of Electrical & Electronics Engineering*, *3*(1), 673–681.

Li, Y., & Lin, X. (2004a). An improved two-step approach to hallucinating faces. In *Proc. 3rd int. conf. image and graphics* (pp. 298–301).

Li, Y., & Lin, X. (2004b). Face hallucination with pose variation. In *Proc. 6th IEEE int. conf. automatic face and gesture recognition* (pp. 723–728).

Lin, Z., & Shum, H.-Y. (2004). Fundamental limits of reconstruction-based superresolution algorithms under local translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(1), 83–97.

Lin, Z., He, J., Tang, X., & Tang, C.-K. (2007). Limits of learning-based superresolution algorithms. In *Int. conf. computer vision*.

Liu, C., Shum, H.-Y., & Zhang, C. S. (2001). A two-step approach to hallucinating faces: global parametric model and local nonparametric model. In *Proc. computer vision and pattern recognition* (pp. 192–198).

Liu, W., Lin, D., & Tang, X. (2005a). Hallucinating faces: TensorPatch super-resolution and coupled residue compensation. In *Proc. computer vision and pattern recognition* (pp. II 478–484).

Liu, W., Lin, D., & Tang, X. (2005b). Neighbor combination and transformation for hallucinating faces. In *IEEE int. conf. multimedia and expo (ICME)*, Amsterdam, Netherlands.

Liu, W., Lin, D., & Tang, X. (2005c). Face hallucination through dual associative learning. In *IEEE int. conf. image processing (ICIP)*, Genova, Italy.

Miravet, C., & Rodríguez, F.B. (2003). A hybrid MLP-PNN architecture for fast image superresolution. In *Int. conf. artificial neural network* (pp. 417–424).

Nguyen, N., & Milanfar, P. (2000). An efficient wavelet-based algorithm for image superresolution. *IEEE International Conference Image Processing (ICIP)*, *2*, 351–354.

Park, S. C., Park, M. K., & Kang, M. G. (2003). Super-resolution image reconstruction: a technical overview. *IEEE Signal Processing Magazine*, *20*(3), 21–36.

Patti, A. J., Sezan, M. I., & Tekalp, A. M. (1997). Superresolution video reconstruction with arbitrary sampling lattices and nonzero aperture time. *IEEE Transaction on Image Processing*, *6*(8), 1064–1076.

Pickup, L. C., Roberts, S. J., & Zisserman, A. (2003). A sampled texture prior for image super-resolution. *Advances in Neural Information Processing Systems* (pp. 1587–1594).

Rhee, S. H., & Kang, M. G. (1999). Discrete cosine transform based regularized high-resolution image reconstruction algorithm. *Optical Engineering*, *38*(8), 1348–1356.

Shah, N. R., & Zakhor, A. (1999). Resolution enhancement of color video sequences. *IEEE Transactions on Image Processing*, *8*, 879–885.

Shiryaev, A. N. (1995). *Probability*. Berlin: Springer.

Srivastava, A., Lee, A. B., Simoncelli, E. P., & Zhu, S.-C. (2003). On advances in statistical modeling of natural images. *Journal of Mathematical Imaging and Vision*, *18*, 17–33.

Sun, J., Tao, H., & Shum, H.-Y. (2003). Image hallucination with primal sketch priors. In *Proc. computer vision and pattern recognition* (pp. II 729–736).

Tsai, R. Y., & Huang, T. S. (1984). Multipleframe image restoration and registration. In *Advances in computer vision and image processing* (pp. 317–339). Greenwich: JAI Press.

Vapnik, V. N. (1998). *Statistical learning theory*. New York: Wiley.

Wilson, R. (2000). MGMM: multiresolution Gaussian mixture models for computer vision. In *Proc. int. conf. pattern recognition* (pp. I 212–215).

Zhang, L., & Pan, F. (2002). A new method of images super-resolution restoration by neural networks. In *Proc. 9th int. conf. neural information processing* (Vol. 5, pp. 2414–2418), 18–22 November.