Power-efficient Cameras Using Natural Image Statistics

Roni Feldman
Hebrew University
Jerusalem, Israel
roni.feldman@mail.huji.ac.il

Yair Weiss
Hebrew University
Jerusalem, Israel
yweiss@cs.huji.ac.il

Yonina C. Eldar
Technion
Haifa, Israel
yonina@ee.technion.ac.il

Abstract

Motivated by recent results on compressed sensing cameras we consider cameras that perform an analog linear transformation $\mathbf{\Phi}$ on the signal, followed by scalar quantization. Specifically we ask: is it better to use compressed sensing ($\mathbf{\Phi}$ is an under-sampling random matrix) or direct sensing ($\mathbf{\Phi}$ is the sparsifying basis)? We compare the two approaches using their energy-distortion tradeoffs: assuming most of the energy consumed by such systems is in the ADC and the energy of the quantizer doubles with each bit, which system will give lower distortion for the same energy consumption? We present analytic expressions for the energy-distortion curves for three signal models: signals residing in a known subspace, sparse signals and power-law signals. For all of these models, our analysis shows that direct sensing results in lower distortion for a given energy consumption. We also present simulation results for natural images showing that direct sensing of Haar wavelet coefficients is preferable for these signals. Given the assumptions of our model, direct sensing of Haar wavelets can achieve high quality imaging (PSNR of 40 dB) with 6% the power consumption of standard cameras using 8 bits per channel.

1. Introduction

In the third quarter of 2012, the worldwide smartphone population surpassed 1 billion [1]. The smartphone has become a popular and frequently used platform for both consumption and creation of media. Mobile phones have become increasingly performant, able to do real-time 3D rendering and recording 1080p HD video. However, these devices remain very power-constrained. Reducing power consumption of various phone components has thus become an important goal for manufacturers and researchers.

In [4] and [5], the authors showed that the camera subsystem incurs a high power consumption. In fact, their results show that the camera is among the highest energy consumers among a smartphone's subsystems. In [20] the authors show similar results for Google Glass.

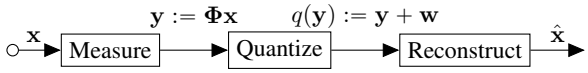

Figure 1. Compression and reconstruction setup. A signal $\mathbf{x}$ with continuous entries is measured using a linear transformation $\mathbf{\Phi}$, quantized using a scalar quantizer $q(\mathbf{y}) = (q_1(y_1), ..., q_m(y_m))$ with $\mathbf{w}$ representing quantization noise, and then reconstructed.

The typical electronic components of a mobile imaging system are the CMOS image sensor, which outputs, per pixel, an analog signal corresponding to the light intensity, an analog-to-digital converter (ADC) which quantizes the signal, and a processing unit which performs post-acquisition tasks such as white balancing and error correction [13]. It is well known that ADCs are often major power consumers in such imaging systems [19]. In view of these observations, optimizing the trade-off between energy consumption and distortion may be of more interest to system designers than the rate-distortion trade-off.

In recent years, compressed sensing (CS) [3, 2, 8, 26, 9] has been suggested as a method to reduce power consumption in imaging systems [17, 25, 21]. In [24], Oike and El Gamal construct a CMOS image sensor with $\Sigma\Delta$ ADC which implements a CS system for natural image acquisition. The authors show a monotonic relation between the energy consumption and the CS compression ratio. Motivated by these results, we ask whether such CS systems have an efficient energy-distortion trade-off, compared to simpler acquisition and reconstruction methods.

Formally, consider a coding scheme for signals $\mathbf{x} \in \mathbb{R}^n$, that consists of a measurement step which is a linear transformation, and a quantization step which converts the continuous measurement into bits. That is, we have a measurement matrix $\mathbf{\Phi} \in \mathbb{R}^{m \times n}$, and for a signal $\mathbf{x} \in \mathbb{R}^n$ we have a measured $\mathbf{y} := \mathbf{\Phi x}$. The measured signal then passes through a scalar quantizer $q(\mathbf{y}) = (q_1(y_1), \cdots, q_m(y_m))$. Fig. 1 shows this setup, including reconstruction. In this work we are interested in the MSE distortion $\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2$. Note that the system described in [24] is a special case of this

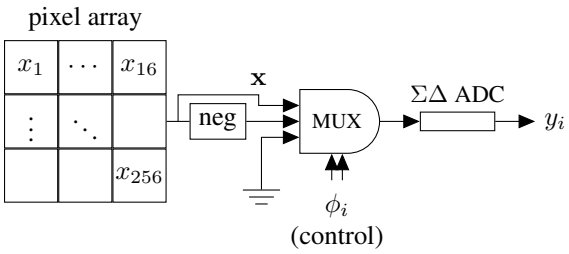

Figure 2. A proposed variation on the design in [24], which allows the system to multiply the pixel array with a $-1, 0, 1$ vector at the quantization stage. A multiplexer is controlled by the projection vector ϕ_i, choosing between the values of $\mathbf{x}$, $-\mathbf{x}$ and 0. The output is passed to a $\Sigma\Delta$ ADC. For details see [24].

coding scheme, where $\mathbf{\Phi}$ is sampled from the ensemble of random binary matrices and quantization is performed per measurement using a $\Sigma\Delta$ ADC.

The main question we address in this work is whether, in terms of energy-distortion trade-off, one should use an under-sampling ($m \leqslant n$) matrix $\mathbf{\Phi} \in \mathbb{R}^{m \times n}$ with i.i.d. entries sampled from a Gaussian, or should $\mathbf{\Phi}$ be the sparsifying basis of the signal. We focus our analysis on the context of natural images.

We will study this question in three settings. We start with a simple result for signals residing in a known subspace. In Section 4 we examine the standard k-sparse model common in the CS literature, and in Section 5 we consider signals obeying a power law decay, which is a good approximation for natural images under common transformations such as DCT and DWT. We show analytically that using direct measurements with efficient rate allocation achieves lower distortion than using random measurements. In Section 6 we corroborate our analysis of the coding schemes with experiments on natural images.

Proofs were omitted due to length constraints, and will appear in the supplementary materials.

1.1. Linear transformations in the analog domain

This work was motivated by the work of Oike and El-Gamal [24] who showed that it is possible to perform a linear transformation on the image before quantization and that the transformation is negligible in terms of power consumption. This was achieved be performing summation and quantization simultaneously in $\Sigma\Delta$ ADCs. The input to each ADC is pixel block values multiplexed with ground, where the multiplexer is controlled by the binary measurement matrix Φ. Their system allows computing the dot product of a pixel block with any binary vector with negligible power consumption. In analyzing the power consumption of a prototype camera that they built, they found that ADCs are main power consumers in the imaging system, and that energy consumption decreases linearly with the measurement ratio of the system.

We propose a small modification to the scheme which allows for the elements of Φ to contain $-1, 0, 1$ values, shown in Fig. 2. This is done by adding an inverting amplifier and using a 3-to-1 multiplexer instead of the 2-to-1 used in [24]. We expect that in our modification as well, the power consumption of the system will be dominated by ADC operations.

2. Modeling assumptions and problem formulation

2.1. Modeling assumptions

To study the energy-distortion behavior of ADCs, we model them as uniform scalar quantizers [28]. We make two strong assumptions in our model:

- The distortion function of fixed rate uniform quantizers has the following form for the input signals we work with:
$$D(R) = c\sigma^2 2^{-2\alpha R}, \tag{1}$$
where σ^2 is the variance of the signal and the constants c, α depend on the quantizer.

- The power consumption of an ADC doubles with every bit
$$E(R) \propto 2^R. \tag{2}$$

In the following we explain these assumptions and formulate our question in the context of our model.

Goyal *et al.* [14] use the following generic distortion function for uniform scalar quantization of signals with bounded support and entropy coded high-resolution uniform scalar quantization of general signals:

$$D(R) = c\sigma^2 2^{-2R}. \tag{3}$$

This distortion model is less accurate for uniform quantization of signals with infinite support. In [15] the authors describe an asymptotically accurate distortion model of uniform quantization for generalized Gaussian signals. This is done by choosing the optimal support of the uniform quantizer given the number of quantization cells and the parameters of the distribution, and results in a fairly complicated expression.

Our distortion model,

$$D(R) = c\sigma^2 2^{-2\alpha R}, \tag{4}$$

was chosen as a generalization of (3) that is still easy to analyze and can be used to approximate the results in [15]. An additional parameter $\alpha > 0$ allows us to configure the model according to the quantization scheme and signal distribution. When $\alpha = 1$ the model reduces to (3), describing the distortion of a uniform quantizer for a signal with

bounded support, and in the supplementary materials we show empirically that for $\alpha \approx 0.85$, this function approximates optimal (according to [15]) uniform quantization of Gaussian and Laplacian signals in the bit rate range $0-15$, which covers most ADC implementations [23]. The multiplicative constant c does not affect our analysis and we ignore it for brevity.

The energy consumption of an ADC depends on its underlying architecture, therefore modeling it accurately is impractical. Instead, we use survey results [23] for ADCs with bit rates in our region of interest as a guideline to define the energy consumption of a converter that outputs a bit rate R as

$$E(R) = c2^R, \tag{5}$$

where we ignore the multiplicative constant for brevity.

In the general case we have m sources (e.g., sensors corresponding to pixels, or wavelet coefficients) which are quantized independently using scalar quantization. Denoting the rates of the codes $\{R_i\}_{i=1}^m$, the total energy consumption of the system is

$$E(R) = \sum_{i=1}^{m} 2^{R_i}. \tag{6}$$

The minimal possible energy consumption according to this model is m, corresponding to an allocation of 0 bits to every element of the quantizer. While a model with 0 minimal energy consumption would be more accurate, the simplicity of (6) allows for easier analysis. We will assume that $E \geqslant m$ throughout the paper.

For a scalar source, the bit rate corresponding to an energy budget E is $R = \log_2 E$, from (6). For a source in $\mathbb{R}^m$, a scalar quantizer $q(\mathbf{y}) = (q_1(y_1), \cdots, q_m(y_m))$ and an energy budget E, one can achieve an energy consumption equal to E with different bit allocations, resulting in different distortions. We define the distortion $D(E)$ of $q(\cdot)$ using the allocation which achieves minimal distortion:

$$\begin{aligned} \min_{R_i} \quad & \sum_{i=1}^{m} D_i(R_i) \\ \text{subject to} \quad & E(R) \leqslant E \\ & R_i \geqslant 0, \forall i, \end{aligned} \tag{7}$$

where $D_i(\cdot)$ is the distortion-rate function of $q_i(\cdot)$ for the specific source distribution, and $E(R)$ is defined in (6). This is similar to rate-distortion formulations for parallel sources constrained by total rate, and is solved using the common "reverse water-filling" algorithm [7]. We note the following observation which we will use in the sequel.

Lemma 1. *Let $E \geqslant m$. If $D_i(\cdot)$ are defined as in (4) and all σ_i are equal, then all the rate allocations $\{R_i\}_{i=1}^m$ of the solution to (7) are equal to $\log_2(E/m)$.*

Equipped with these definitions, our analysis will consist of calculating the measurement distortion $D(E) = \mathbb{E}\|\mathbf{y} - q(\mathbf{y})\|^2$ according to (7) for the measurement schemes described in the following section. From the measurement distortion we infer the reconstruction error $\mathbb{E}\|\mathbf{x} - \hat{\mathbf{x}}\|^2$.

2.2. Quantization systems and problem formulation

Let $\mathbf{x} \in \mathbb{R}^n$ be an input signal with some known distribution. In this work we examine the energy-distortion tradeoff of three quantization schemes of the form portrayed in Fig. 1:

1. Random under-sampling measurements. $m \leqslant n$, $\mathbf{\Phi}_{ij} \sim \mathcal{N}(0, 1/m)$ and a fixed-rate uniform quantizer.

2. Direct measurements. $\mathbf{\Phi} = \mathbf{\Psi}^\dagger$ with a fixed-rate uniform quantizer, where $\mathbf{\Psi}$ is the basis in which $\mathbf{x}$ is sparse, and $\mathbf{\Psi}^\dagger = (\mathbf{\Psi}^T\mathbf{\Psi})^{-1}\mathbf{\Psi}^T$ is the Moore-Penrose pseudoinverse. Since the basis size is smaller than n, $m \leqslant n$ in this case as well.

3. Direct measurements and improved quantization. Again, $\mathbf{\Phi} = \mathbf{\Psi}^\dagger$, followed by a simple variable-rate quantizer, which we call threshold-vr coding.

Given an energy budget E, the signal $\mathbf{x}$ is measured with each of the schemes. It is then quantized and then reconstructed as $\hat{\mathbf{x}}$. The quantizer bit allocations are given by the solution to (7). We ask which scheme achieves minimal MSE distortion $\mathbb{E}\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2$.

3. Distortion-energy of signals in a known subspace

We begin with the simple case in which $\mathbf{x} \in \mathbb{R}^n$ is given by $\mathbf{x} = \mathbf{\Psi}\boldsymbol{\theta}$, where $\mathbf{\Psi} \in \mathbb{R}^{n\times k}$ is a matrix of orthonormal columns, and $\boldsymbol{\theta} \sim \mathcal{N}(0, \mathbf{I}_k)$. We will assume that the sparsifying basis $\mathbf{\Psi}$ is known to both the quantizer and the reconstruction method. Similarly to [14], we show that direct quantization with fixed-rate codes is better than using random measurements. In this case the variable-rate approach threshold-vr is not needed.

3.1. Direct measurements

In the direct approach we have $m = k$, and we directly measure the vector $\boldsymbol{\theta} = \mathbf{\Psi}^\dagger\mathbf{x}$. The elements of $\boldsymbol{\theta}$ are i.i.d.. Therefore, given a total energy budget $E \geqslant m$, the bit rate allocation of each of the k quantizers is $\log_2(E/k)$, from Lemma 1. Thus, the distortion for this approach is:

$$D_{\text{direct}}(E) = k\left(\frac{E}{k}\right)^{-2\alpha}. \tag{8}$$

3.2. Random measurements

To analyze the random measurement distortion, we first note that random matrices $\mathbf{\Phi} \in \mathbb{R}^{m\times n}$ with zero mean Gaussian i.i.d. entries $\mathbf{\Phi}_{ij}$ are orthogonally invariant [10, Section 4.2]. Therefore if $\bar{\mathbf{\Psi}} \in \mathbb{R}^{n\times n}$ is some basis completion of $\mathbf{\Psi}$, then $\mathbf{\Phi}$ and $\mathbf{\Phi}\bar{\mathbf{\Psi}}^{-1}$ have the same distribution. Thus, we will assume in the rest of the paper that the sparsifying basis is the identity $\mathbf{\Psi} = \mathbf{I}_n$ and that $\mathbf{x} = (\boldsymbol{\theta}, 0, \ldots, 0)^T \in \mathbb{R}^n$.

Note that the variances of the elements of $\mathbf{y} := \mathbf{\Phi}\mathbf{x}$ are equal to $\sigma_y^2 = k/m$. Thus, given an energy budget $E \geqslant m$ the distortion of any element of $\mathbf{y}$ is given by

$$D_y(E/m) = \sigma_y^2 \left(\frac{E}{m}\right)^{-2\alpha} = \frac{k}{m}\left(\frac{E}{m}\right)^{-2\alpha}, \qquad (9)$$

according to Lemma 1. We now use the rate-distortion results from [6] slightly modified by replacing distortion-rate functions with distortion-energy functions. This oracle-assisted approach, in which the support of the signal is known to both coder and decoder, provides the following lower bound for reconstruction of randomly measured signals.

Theorem 1. *[6]. Assume that the support of $\mathbf{x} \in \mathbb{R}^n$ is known at the decoder, and that $m > k + 3$. Assume the measurement matrix is $\mathbf{\Phi}_{ij} \sim \mathcal{N}(0, 1/m)$. Let E be the total energy budget of the code, and let $D_y(E)$ be the distortion function of each element of $\mathbf{y}$. Then the distortion $\mathbb{E}\|\mathbf{x} - \hat{\mathbf{x}}\|^2$ of the reconstruction has the following lower bound*

$$D_{\mathsf{cs-oracle}}(E) = \frac{km}{m-k-1} D_y(E/m). \qquad (10)$$

In our case, the support is indeed known, and so this result gives the MSE of using random measurements. Of course, a real CS system would not know the support of the source signal at the decoder, but simulations show that this result acts as a good lower bound on the reconstruction error of CS algorithms [6].

Note that in both the direct and random approaches, the quantized elements are Gaussian. Thus, the α parameter in (8) and (9) has the same value. Using (8), (9) and Theorem 1 we obtain the following simple result.

Corollary 1. *Let $\mathbf{x}$ be a signal drawn from a known subspace as described above, and let $\alpha > 0.5$. If $m > k + 3$ then:*

$$D_{\mathsf{direct}}(E) < D_{\mathsf{cs-oracle}}(E). \qquad (11)$$

The result is intuitive – there is no reason to mix a signal with i.i.d. entries when one knows its support exactly.

4. Distortion-energy of sparse signals

Next, let $\mathbf{x} \in \mathbb{R}^n$ be a k-sparse signal, where the k indices are chosen uniformly from the $\binom{n}{k}$ possible sets, and the non-zero entries are independently sampled from the standard normal distribution.

In this section we show analytically that under reasonable assumptions, the threshold-vr method achieves lower distortion than any reconstruction algorithm which uses random Gaussian measurements.

4.1. Direct measurements

We start with the fixed-rate direct method. In this method the quantization consists of n scalar quantizers, to which we allocate bit rates according to (7). It is easy to see that the variance of any x_i is k/n so that according to Lemma 1 the rate of each quantizer is $R = \log_2(E/n)$. Assuming a distortion model parameter α_0 for this signal, the reconstruction error is:

$$D_{\mathsf{direct}}(E) = n\left(\frac{k}{n}\right) 2^{-2\alpha_0 R} = k\left(\frac{E}{n}\right)^{-2\alpha_0}. \qquad (12)$$

In terms of rate-distortion, it is easy to see that the direct approach is not optimal for high rates since most of the bit-rate is wasted on elements equal to zero. In [14], the authors address this with a simple adaptive approach in which coding consists of spending $R_0 = \log\binom{n}{k}$ bits to code the support indices, and the rest for coding their values, resulting in a distortion of $k2^{-2\alpha(R-R_0)/k}$, where $R \gg R_0$ is the total amount of bits.

Our energy model applies only to quantization, so modeling the energy consumption of coding the support is out of our scope. Instead we assume an oracle model where the support is known to the encoder and decoder. The resulting distortion

$$D_{\mathsf{adaptive}}(E) = k\left(\frac{E}{k}\right)^{-2\alpha}, \qquad (13)$$

is equivalent to (8) and serves as a lower bound for the following approach, which we can model entirely within our framework.

A simple solution for the wasted bits in the direct method is a coding scheme we call "variable rate threshold code" (threshold-vr) which sends the bit '0' when the source is under a negligibly small threshold and otherwise uses a fixed rate code prefixed by the bit '1'. Since the thresholding can be implemented using a simple comparison and separately from the actual quantization, we get that each quantizer uses energy equal to 2^1 for thresholding, and the remaining energy $(E - 2n)$ is available for the k active quantizers.

Given an energy budget of E, the rate of each quantizer is $R = \log_2((E - 2n)/k)$, from Lemma 1. Assume the thresholding correctly identifies the support. From (4) we

get

$$D_{\mathsf{threshold-vr}}(E) = k2^{-2\alpha R} = k\left(\frac{E-2n}{k}\right)^{-2\alpha}. \quad (14)$$

4.2. Random measurements

We now consider the random measurement scheme, where $\mathbf{\Phi}_{ij} \sim \mathcal{N}(0, 1/m)$, and denote the measurement $\mathbf{y} := \mathbf{\Phi x}$. From the central limit theorem, for large enough n the elements of $\mathbf{y}$ are identically distributed as $\mathcal{N}(0, k/m)$, and again from Lemma 1 the energy allocated to each quantizer is E/m.

As in the previous section, we have

$$D_y(E/m) = \sigma_y^2\left(\frac{E}{m}\right)^{-2\alpha} = \frac{k}{m}\left(\frac{E}{m}\right)^{-2\alpha}. \quad (15)$$

Plugging into Theorem 1 provides us a lower bound on the distortion in the case of random measurements

$$D_{\mathsf{cs-oracle}}(E) = \frac{k^2}{m-k-1}\left(\frac{E}{m}\right)^{-2\alpha}. \quad (16)$$

Note that the quantized signal is Gaussian in the threshold-vr and random approaches. Therefore, α has the same value in (14) and (16). Comparing both methods we obtain the following result.

Corollary 2. *Let* $\mathbf{x}$ *be a* k*-sparse signal, and let* $E > 10n$ *and* $\alpha \geqslant 0.85$*. If* $m > 2k$ *then*

$$D_{\mathsf{threshold-vr}}(E) < D_{\mathsf{cs-oracle}}(E). \quad (17)$$

Note that standard CS theory requires that $m > Ck\log(n/k)$ with a large constant $C > 0$ when $\log(n/k)$ is small [12, 11], and so the conditions of Corollary 2 apply to most practical instances.

In Fig. 3 we show the gap between the distortion functions of the oracle-assisted CS (16), the threshold-vr method (14), and the direct method (12). We plot the SNR $= 10\log_{10}(\mathbb{E}\|\mathbf{x}\|^2/D_{\mathsf{method}}(E))$ corresponding to their distortion functions for the quantization model parameter $\alpha = 0.85$, dimension $n = 100$ and sparsity $k = 5$. The CS oracle curve was chosen by searching for the m between $k+4$ and n minimizing (16), and was found to be $m = 15$. We can see that the threshold-vr method achieves SNR comparable to the adaptive bound (13) in most of the energy range.

We conclude that under our assumptions, it is preferable to use direct measurements with variable rate threshold quantization over random measurements.

5. Distortion-energy of signals obeying a power law

While sparse signals have been thoroughly studied in the CS literature, it is interesting to consider signals with statistical properties similar to those of natural images. One

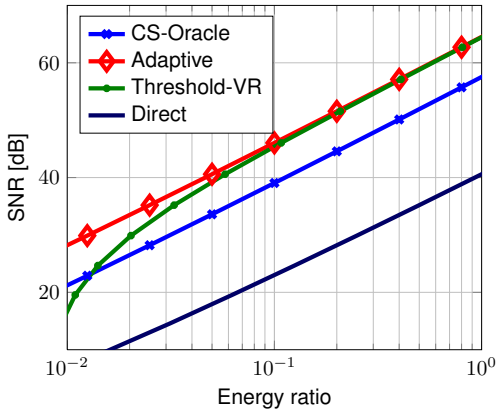

Figure 3. The energy-distortion SNR curves of the distortion functions of oracle-assisted CS (16), threshold variable rate (14) and direct (12) approaches, compared with the adaptive (13) baseline. The horizontal axis is the ratio between the energy consumptions of the methods and the energy consumed when each of the n elements is quantized using 8 bits. The dimension is $n = 100$, and the sparsity is $k = 5$. The quantization parameter is $\alpha = 0.85$, and the best measurement ratio $\beta = m/n$ for the CS bound was chosen using search and was achieved for $m = 15$.

of the fundamental results of frequency-based analysis of natural images is that their power spectrum decay is inversely proportional to the square of the frequency [16]. Signals with decaying magnitudes of their sorted coefficients are generally referred to as *compressible* in the CS literature, and it has been shown that for certain decay rates they are well approximated by sparse signals and so can be tractably decoded using CS techniques, with proven worst case bounds on the error [3, 11].

Our interest lies in the mean error, and as we have done with the previous signal models, we would like to compare the error obtained by direct quantization with optimal allocations with the error of methods that use random measurements.

We remain with the model described in Fig. 1. Instead of a sparse signal, let $\mathbf{x} \in \mathbb{R}^n$ be a random Gaussian with zero mean and independent elements with a diagonal covariance matrix $\mathbf{C}_x$ whose diagonal entries obey the power law $(\mathbf{C}_x)_{ii} = 1/i^2$.

For this signal we cannot use Theorem 1 to calculate the distortion when $\mathbf{\Phi}$ is random, since the signal is not strictly k-sparse. Instead, our analysis will assume a linear measurement model and apply simple results from random matrix theory on the minimum mean square error (MMSE) estimator of the measured signal.

In this setting, $\mathbf{x} \sim \mathcal{N}(0, \mathbf{C_x})$ is measured with

$$\mathbf{y} = \mathbf{\Phi x} + \mathbf{w}, \quad (18)$$

where $\mathbf{\Phi} \in \mathbb{R}^{m\times n}$ (with $m \leqslant n$). The vector $\mathbf{w} \sim \mathcal{N}(0, \mathbf{C_w})$ represents the quantization noise of the system,

and its covariance $\mathbf{C_w}$ is a diagonal matrix with variances equal to the distortion of the corresponding elements, and so dependent on the measurement method and the quantization energy budget.

Let $\hat{\mathbf{x}}$ be the MMSE estimator of the model above, and let the estimation error be $\boldsymbol{\epsilon} = \mathbf{x} - \hat{\mathbf{x}}$. From the Gauss-Markov theorem [18], the covariance of $\boldsymbol{\epsilon}$ is

$$\mathbf{C}_{\boldsymbol{\epsilon}} = \left(\mathbf{C}_{\mathbf{x}}^{-1} + \boldsymbol{\Phi}^T \mathbf{C}_{\mathbf{w}}^{-1} \boldsymbol{\Phi}\right)^{-1} \tag{19}$$

and so the MSE is $\mathrm{tr}(\mathbf{C}_\epsilon)$. In the following we will show that the MMSE of direct measurements is smaller than the MMSE of random measurements.

5.1. Direct measurements

For the direct measurement method ($\boldsymbol{\Phi} = \mathbf{I}_n$) with fixed-rate, the distortion for this signal model is given directly by (7). The following result is a direct consequence of the "reverse water-filling" method applied to this optimization problem.

Lemma 2. *Let $\alpha > 0.5$. If $E > \frac{\alpha+0.5}{\alpha-0.5}(n+1)$ then the distortion of the i-th element is*

$$D_i(E) = \sigma_i^{\frac{1}{\alpha+0.5}} \left(\sum_{j=1}^{n} \sigma_j^{\frac{1}{\alpha+0.5}} \right)^{2\alpha} E^{-2\alpha}, \tag{20}$$

and the total distortion of the quantizer is

$$D_{\mathsf{direct}}(E) = \left(\sum_{i=1}^{n} \sigma_i^{\frac{1}{\alpha+0.5}} \right)^{2\alpha+1} E^{-2\alpha}. \tag{21}$$

Now, assume that $(\mathbf{C_w})_{ii} = D_i(E)$. Then

$$\begin{aligned} \mathsf{mmse}_{\mathsf{direct}}(E) &= \mathrm{tr}\left(\mathbf{C}_{\mathbf{x}}^{-1} + \mathbf{C}_{\mathbf{w}}^{-1}\right)^{-1} && (22) \\ &< \mathrm{tr}(\mathbf{C_w}) = D_{\mathsf{direct}}(E). && (23) \end{aligned}$$

Therefore, it is enough to show that $D_{\mathsf{direct}}(E) < \mathsf{mmse}_{\mathsf{random}}(E)$.

5.2. Random measurements

Now consider the case where $\boldsymbol{\Phi} \in \mathbb{R}^{m\times n}$ is random, with $\boldsymbol{\Phi}_{ij} \sim \mathcal{N}(0, 1/m)$. We assume a large scale regime where $m, n \to \infty$ and $\beta = m/n$ is constant. Our goal is to calculate $\mathsf{mmse}_{\mathsf{random}} = \mathbb{E}\left[\mathrm{tr}(\mathbf{C}_{\boldsymbol{\epsilon}})|\boldsymbol{\Phi}\right]$. In words, we want to find the MSE of the MMSE estimator which receives the measurement matrix $\boldsymbol{\Phi}$ as known input. We will find a simple lower bound for it, beginning with the fact that

$$\begin{aligned} \mathrm{tr}\mathbf{C}_{\boldsymbol{\epsilon}} &= \mathrm{tr}\left(\mathbf{C}_{\mathbf{x}}^{-1} + \boldsymbol{\Phi}^T \mathbf{C}_{\mathbf{w}}^{-1} \boldsymbol{\Phi}\right)^{-1} && (24) \\ &> \mathrm{tr}\left(n^2 \mathbf{I}_n + \boldsymbol{\Phi}^T \mathbf{C}_{\mathbf{w}}^{-1} \boldsymbol{\Phi}\right)^{-1} && (25) \\ &= \frac{1}{n^2} \mathrm{tr}\left(\mathbf{I}_n + \frac{1}{n^2} \boldsymbol{\Phi}^T \mathbf{C}_{\mathbf{w}}^{-1} \boldsymbol{\Phi}\right)^{-1}. && (26) \end{aligned}$$

Similarly to the previous sections, we use Lemma 1 to note that the noise covariance is a scalar matrix $\mathbf{C_w} = \sigma_{\mathbf{w}}^2 \mathbf{I}_m$ with

$$\sigma_{\mathbf{w}}^2 = \frac{1}{m} \sum_{i=1}^{n} \sigma_i^2 \left(\frac{E}{m}\right)^{-2\alpha}, \tag{27}$$

and so it remains to calculate

$$\frac{1}{n^2} \mathrm{tr}\left(\mathbf{I}_n + \frac{1}{n^2} \frac{1}{\sigma_{\mathbf{w}}^2} \boldsymbol{\Phi}^T \boldsymbol{\Phi}\right)^{-1}. \tag{28}$$

Using the Marčenko-Pastur law [27] we get that the above converges a.s. to

$$\mathsf{mmse}_{\mathsf{random}}(E) = \frac{1}{n}\left(1 - \frac{F\left((n\sigma_{\mathbf{w}})^{-2}, \beta^{-1}\right)}{4(n\sigma_{\mathbf{w}})^{-2}\beta^{-1}}\right), \tag{29}$$

where $F(x, z)$ is an elementary function defined in [27]. Note that since the quantized signal is Gaussian in the direct and random approaches, the value of α in (27) and (23) is equal. Therefore, combining (29) with (23) we obtain the following result.

Corollary 3. *Let $0.85 \leqslant \alpha \leqslant 1$, $n > 10$ and $\beta = m/n \leqslant 1$. If*

$$E > cn \tag{30}$$

for a known constant c which depends only on α and β, then

$$\mathsf{mmse}_{\mathsf{direct}}(E) < \mathsf{mmse}_{\mathsf{random}}(E). \tag{31}$$

In fact, c is small enough that for any practical problem dimension n the direct method is essentially always better than using random measurements. See the supplementary materials for specifics. In Fig. 4 we plot the SNR of of both methods for the quantization model parameter $\alpha = 0.85$. The direct curve is given by (21) and the random curve is obtained by evaluating (29) with different measurement ratios β.

6. Experiments with natural images

To test the our numerical results we ran the complete process of quantization and recovery on natural image patches, using direct quantization and CS. Given an image patch $\mathbf{x} \in \mathbb{R}^n$, we used the 2D Haar transform $\mathbf{H} \in \mathbb{R}^{n\times n}$ to get a wavelet representation of the patch $\mathbf{Hx}$ on which we ran our quantization schemes.

We first tried performing optimal uniform quantization (according to [15]) for both direct measurements and random measurements of the Haar coefficients. Since the Haar coefficient distributions are not uniform and have bounded support, uniform quantization turned out to be only slightly better than uniform quantization of pixel values (for the

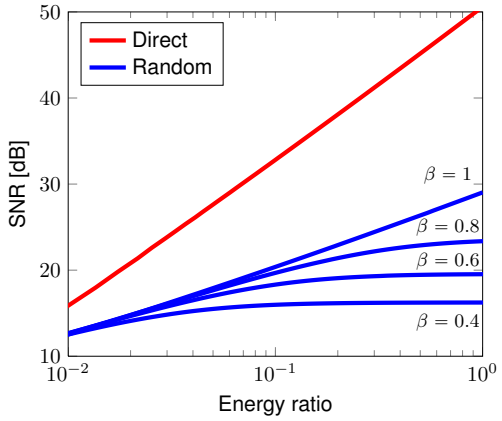

Figure 4. SNR of the lower bound on optimal reconstruction with random measurements (29) versus direct quantization with energy constrained bit allocation (21), for Gaussian signals in dimension $n = 100$ obeying a power law $\sigma_i^2 = 1/i^2$. The horizontal axis is the ratio between the energy consumptions of the methods and the energy consumed when each of the n elements is quantized using 8 bits. We show SNR for several values of $\beta = m/n$.

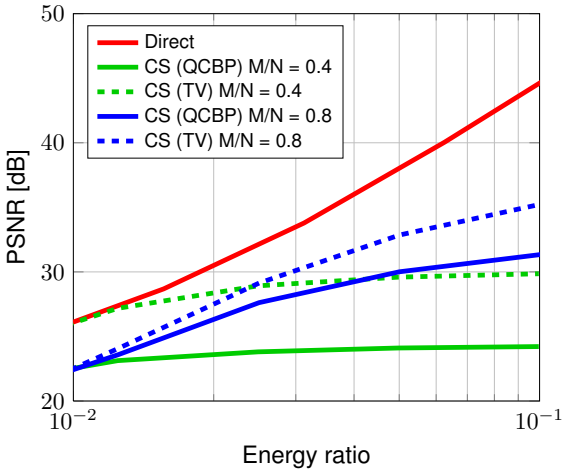

Figure 5. Performance of quantized CS with QCBP reconstruction (32) and direct quantization with quantizer bit-rates allocated according to (7), on 2D Haar coefficients of natural image patches from BSDS300. The horizontal axis is the ratio between the energy consumptions of the methods and the energy consumed when each of the n elements is quantized using 8 bits. The patch size is $n = 16 \times 16$.

same energy budget). To obtain better PSNR we used non-uniform quantization in which the cells were learned with k-means. For the random measurements, using non-uniform quantization didn't result in significant change in PSNR. We note that because of the generality of our distortion model, the theoretical results of the previous sections are relevant even for non-uniform quantization.

Since the Haar wavelet representation does not have equal variance, we allocated a different number of bits to each coefficient. This was done by solving the optimization problem in (7). Specifically, we used reverse water-filling [7] to obtain real-valued rates, and then rounded these rates to integers. Since the random measurements had equal variance in their elements, bit allocation for the random approach was uniform across all elements, as a result of Lemma 1.

CS reconstruction was performed using two methods. The first Quadratically Constrained Basis Pursuit (QCBP), also known as Basis Pursuit DeNoising (BPDN),

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{H}\mathbf{x}\|_1 \quad \text{s.t.} \quad \|\mathbf{\Phi}\mathbf{H}\mathbf{x} - q(\mathbf{y})\|_2 < \epsilon, \quad (32)$$

with $q(\mathbf{y}) = \mathbf{y} + \mathbf{w}$ being the quantized measurements, $\epsilon = c\sigma_w$, and $c > 0$ chosen using simple parameter search. The second is total variation (TV) minimization [3],

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \mathrm{TV}(\mathbf{x}) \quad \text{s.t.} \quad \mathbf{\Phi}\mathbf{H}\mathbf{x} = q(\mathbf{y}), \quad (33)$$

where

$$\mathrm{TV}(\mathbf{x}) = \sum_{i,j} \sqrt{(x_{i+1,j} - x_{i,j})^2 + (x_{i,j+1} - x_{i,j})^2} \quad (34)$$

is the isotropic total variation norm. Note that other constraints could be used in these formulations, such as $-\epsilon \prec \Phi\mathbf{H}\mathbf{x} - \mathbf{y} \prec \epsilon$, where the elements of ϵ are the cells sizes of the quantizers. We have tried several constraint variations for each problem and found that the above reach the lowest error, correspondingly.

As expected, TV minimization performs better than QCBP for natural images. Both approaches have lower PSNR than direct sensing of Haar coefficients at the same energy consumption levels, as can be seen in Fig. 5. Moreover, we see that the behavior of the direct and CS methods is consistent with the analysis in Section 5, and Fig. 4. The direct methods can be further improved using better reconstruction algorithms.

Fig. 6 shows qualitative results of both methods for zoomed-in image patches, and Table 1 shows quantitative results for ten complete images. We also show a subsampling baseline in which the overall amount of sampled pixels is decreased by a factor of 4, corresponding to 25% of the energy consumption of a system sampling all pixels.

We note that these results do not contradict the low power consumption obtained in [24]. Indeed, the amount of quantizations decreases linearly with the amount of CS measurements, and so energy decreases linearly as well. Nevertheless, our experiments show that one can obtain higher quality images with the same power budget by using direct quantization of Haar coefficients.

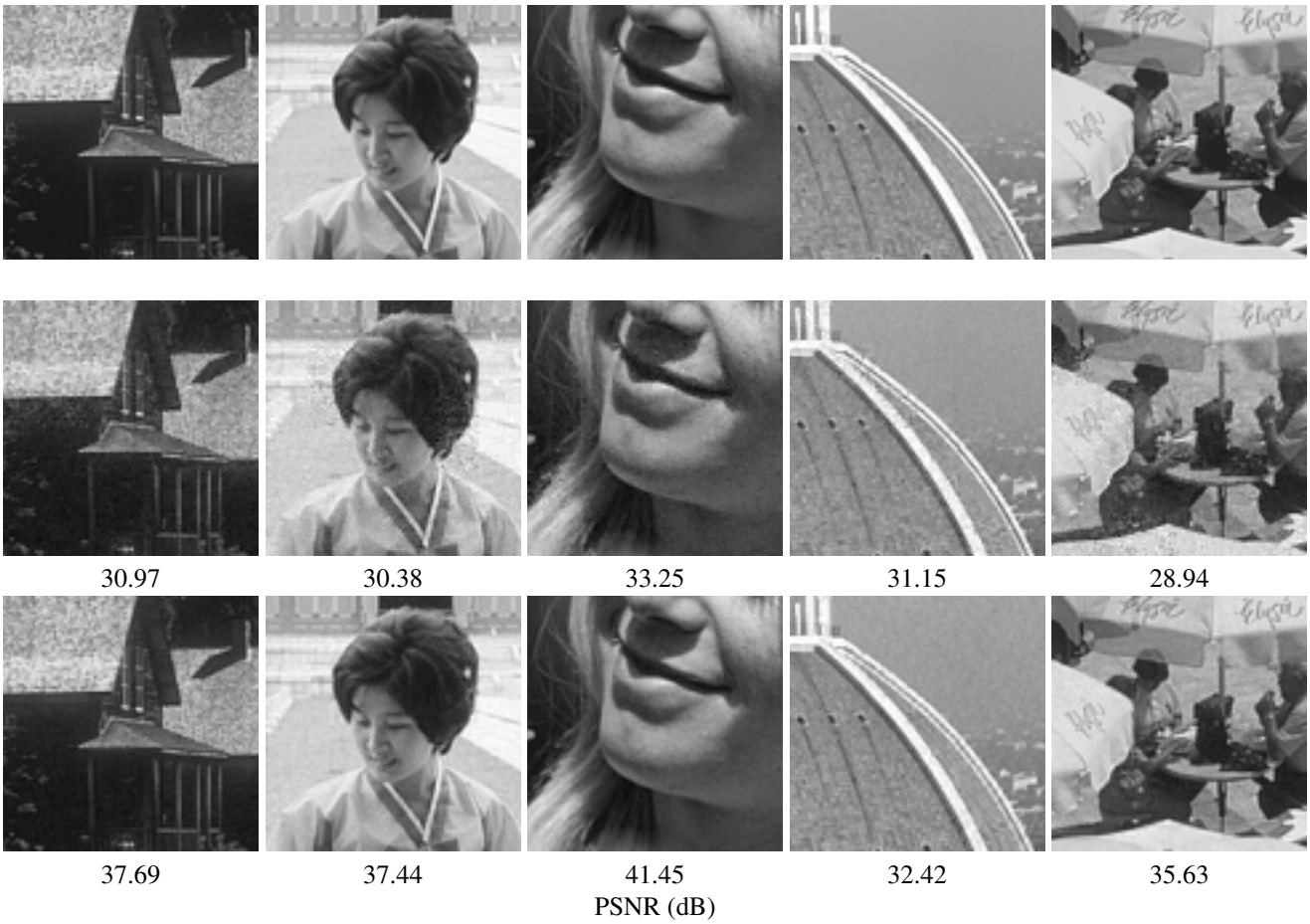

Figure 6. Qualitative comparison of direct quantization and CS ($\beta = 0.8$) of Haar coefficients on several zoomed-in images from BSDS300 [22]. For both methods energy usage is $5\% - 6\%$ of a system which encodes each pixel with 8 bits. The patch size is $n = 16 \times 16$. Note that the examples shown are comprised of several patches. Top: original, middle: CS-TV, bottom: direct.

	PSNR (dB)			
Image Id	CS-QCBP	CS-TV	Subsampling	Direct
232038	29.65	30.97	24.69	**37.69**
236017	31.17	34.08	28.65	**41.65**
238011	41.43	43.37	37.95	**48.30**
239007	28.46	30.38	27.28	**37.44**
239096	30.86	33.25	30.64	**41.45**
24004	28.12	31.15	24.54	**32.42**
24063	41.10	40.91	35.32	**46.59**
242078	26.89	28.94	24.05	**35.63**
245051	28.76	31.15	25.19	**36.68**
246016	32.84	35.02	30.32	**38.44**

Table 1. Reconstruction PSNR of quantized CS ($\beta = 0.8$) and direct quantization of Haar coefficients for 10 images from BSDS300. For the CS and direct methods energy usage is $5\% - 6\%$ of a system which encodes each pixel with 8 bits. The factor 4 subsampling results in an energy usage of 25%, while its PSNR is lower than the other methods. The patch size is $n = 16 \times 16$.

7. Discussion and future work

We presented a model for distortion and energy consumption of signal acquisition using linear measurements followed by quantization. In this model we described a quantizer bit rate allocation which is also constrained by the total energy budget of the system.

We have shown that using direct measurements where quantizer bit rates are allocated with this method is more effective than using random measurements for signals in a known subspace, sparse signals, and signals obeying a power law decay in their variances. Specifically, the results for the power law signals suggest that using compressed sensing for natural image capture may be an inefficient approach for low energy consumption. We supported the theoretical results with simulations on natural images.

In the future, it would be interesting to analyze more accurate signal models, test our results on real hardware, and improve the reconstruction of the direct method using stronger priors.

Acknowledgements

Support from the ISF and Intel ICRI-CI is gratefully acknowledged.

References

[1] S. Bicheno. Global smartphone installed base forecast by operating system for 88 countries: 2007 to 2017. *Strategy Analytics*, 2012. 1

[2] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *Information Theory, IEEE Transactions on*, 52(2):489–509, 2006. 1

[3] E. J. Candes, J. K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006. 1, 5, 7

[4] A. Carroll and G. Heiser. The systems hacker's guide to the galaxy energy usage in a modern smartphone. In *Proceedings of the 4th Asia-Pacific Workshop on Systems*, APSys '13, pages 5:1–5:7, New York, NY, USA, 2013. ACM. 1

[5] X. Chen, Y. Chen, Z. Ma, and F. C. A. Fernandes. How is energy consumed in smartphone display applications? In *Proceedings of the 14th Workshop on Mobile Computing Systems and Applications*, HotMobile '13, pages 3:1–3:6, New York, NY, USA, 2013. ACM. 1

[6] G. Coluccia, A. Roumy, and E. Magli. Operational rate-distortion performance of single-source and distributed compressed sensing. *Communications, IEEE Transactions on*, 62(6):2022–2033, 2014. 4

[7] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012. 3, 7

[8] D. L. Donoho. Compressed sensing. *Information Theory, IEEE Transactions on*, 52(4):1289–1306, 2006. 1

[9] M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, T. Sun, K. E. Kelly, R. G. Baraniuk, et al. Single-pixel imaging via compressive sampling. 1

[10] A. Edelman and N. R. Rao. Random matrix theory. *Acta Numerica*, 14:233–297, 2005. 4

[11] Y. C. Eldar. *Sampling Theory: Beyond Bandlimited Systems*. Cambridge University Press, 2015. 5

[12] S. Foucart and H. Rauhut. *A mathematical introduction to compressive sensing*. Springer, 2013. 5

[13] A. E. Gamal and H. Eltoukhy. Cmos image sensors. *Circuits and Devices Magazine, IEEE*, 21(3):6–20, 2005. 1

[14] V. K. Goyal, A. K. Fletcher, and S. Rangan. Compressive sampling and lossy compression. *Signal Processing Magazine, IEEE*, 25(2):48–56, 2008. 2, 3, 4

[15] D. Hui and D. L. Neuhoff. Asymptotic analysis of optimal fixed-rate uniform scalar quantization. *Information Theory, IEEE Transactions on*, 47(3):957–977, 2001. 2, 3, 6

[16] A. Hyvärinen, J. Hurri, and P. O. Hoyer. *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision.*, volume 39. Springer Science & Business Media, 2009. 5

[17] L. Jacques, P. Vandergheynst, A. Bibet, V. Majidzadeh, A. Schmid, and Y. Leblebici. Cmos compressed imaging by random convolution. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 1113–1116. IEEE, 2009. 1

[18] S. M. Kay. Fundamentals of statistical signal processing: estimation theory. 1993. 6

[19] R. LiKamWa, B. Priyantha, M. Philipose, L. Zhong, and P. Bahl. Energy characterization and optimization of image sensing toward continuous mobile vision. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*, pages 69–82. ACM, 2013. 1

[20] R. LiKamWa, Z. Wang, A. Carroll, F. X. Lin, and L. Zhong. Draining our glass: An energy and heat characterization of google glass. In *Proceedings of 5th Asia-Pacific Workshop on Systems*, page 10. ACM, 2014. 1

[21] V. Majidzadeh, L. Jacques, A. Schmid, P. Vandergheynst, and Y. Leblebici. A (256× 256) pixel 76.7 mw cmos imager/compressor based on real-time in-pixel compressive sensing. In *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, pages 2956–2959. IEEE, 2010. 1

[22] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. 8th Int'l Conf. Computer Vision*, volume 2, pages 416–423, July 2001. 8

[23] B. Murmann. Adc performance survey 1997-2015. *http://www.stanford.edu/˜murmann/adcsurvey.html*, 2015. 3

[24] Y. Oike and A. El Gamal. Cmos image sensor with per-column $\sigma\delta$ adc and programmable compressed sensing. *Solid-State Circuits, IEEE Journal of*, 48(1):318–328, 2013. 1, 2, 7

[25] R. Robucci, J. D. Gray, L. K. Chiu, J. Romberg, and P. Hasler. Compressive sensing on a cmos separable-transform image sensor. *Proceedings of the IEEE*, 98(6):1089–1101, 2010. 1

[26] J. Romberg. Imaging via compressive sampling [introduction to compressive sampling and recovery via convex programming]. *IEEE Signal Processing Magazine*, 25(2):14–20, 2008. 1

[27] A. M. Tulino and S. Verdú. *Random matrix theory and wireless communications*, volume 1. Now Publishers Inc, 2004. 6

[28] T. Wiegand and H. Schwarz. *Source coding: Part I of fundamentals of source and video coding*, volume 10. Now Publishers Inc, 2010. 2