Empirical Bayes and Full Bayes for Signal Estimation

Yanting Ma, Jin Tan, Nikhil Krishnan, and Dror Baron
Department of Electrical and Computer Engineering
North Carolina State University
Raleigh, NC 27695, USA
Email: {yma7, jtan, nkrishn, barondror}@ncsu.edu

***Abstract*—We consider signals that follow a parametric distribution where the parameter values are unknown. To estimate such signals from noisy measurements in scalar channels, we study the empirical performance of an empirical Bayes (EB) approach and a full Bayes (FB) approach. We then apply EB and FB to solve compressed sensing (CS) signal estimation problems by successively denoising a scalar Gaussian channel within an approximate message passing (AMP) framework. Our numerical results show that FB achieves better performance than EB in scalar channel denoising problems when the signal dimension is small. In the CS setting, the signal dimension must be large enough for AMP to work well; for large signal dimensions, AMP has similar performance with FB and EB.**

I. Introduction

A. Motivation

Consider the estimation of an input signal $\mathbf{x} \in \mathbb{R}^N$ from noise-corrupted measurements $\mathbf{y} = \mathbf{x} + \mathbf{z}$, where $\mathbf{z} \in \mathbb{R}^N$ represents the additive noise. Ideally, if the statistical characteristics of both the input $\mathbf{x}$ and the noise $\mathbf{z}$ are known, then the optimal signal estimator in the mean square error sense can be obtained by the Bayesian method of conditional expectation [1]. In many applications, however, the prior distribution of the input signal may not be available, and thus the conditional expectation cannot be computed.

One of the approaches to resolve unknown priors is based on the *minimum description length* (MDL) principle [2, 3]. The main idea of MDL is that the signal of interest $\mathbf{x}$ is usually meaningful and compressible, yet the noise $\mathbf{z}$ is random and incompressible [4]. Therefore, the signal $\mathbf{x}$ and noise $\mathbf{z}$ can be separated by finding the most compressible description of $\mathbf{x}$ subject to constraints on the noisy measurements $\mathbf{y}$ and the noise distribution. However, MDL does not always achieve the *minimum mean square error* (MMSE) [4, 5].

Besides the settings where the prior distribution is completely unavailable, there are applications where we may know the prior distribution class while the parameters that characterize the distribution are unknown. For a parametric source that follows an unknown parameter θ, one can first estimate θ based on observed noisy measurements, and then plug the estimated parameter into an appropriate Bayesian estimator. This approach can be categorized as *empirical Bayes* (EB) [6]. However, EB considers only one possible estimate $\widehat{\theta}$ and discards all other possible estimates that might also be informative. To take into account the uncertainty in θ, a *full Bayes* (FB) approach can be applied. In FB, a prior is assigned to θ, thus the posterior of θ, $f(\theta|\mathbf{y})$, can be computed. The FB estimator is then defined as the weighted sum of the Bayesian estimators with respect to each possible θ, where the weights are the corresponding $f(\theta|\mathbf{y})$.

We notice that the FB approach that we discuss in this paper is a mixture over the parameter space. It is worth mentioning that a closely related estimator that mixes over the signal space [7] was proposed as *universal conditional expectation* (UCE) [8], in which the expectation is computed with respect to a universal prior [2, 9]. It has been proved [7] that UCE achieves the MMSE when the input signal $\mathbf{x}$ is Bernoulli and the noisy measurements $\mathbf{y}$ are observed from a *binary symmetric channel* (BSC) [10].

B. Problem setting

Input distributions: The input $\mathbf{x}$ of dimension N is generated by an independent and identically distributed (i.i.d.) source, and we consider two parametric distributions. Our first distribution is Bernoulli,

$$x_i \sim \text{Bernoulli}(\theta). \tag{1}$$

That is, $\mathbb{P}(x_i = 1) = \theta = 1 - \mathbb{P}(x_i = 0)$, where the subscript $(\cdot)_i$ denotes the i-th component of a vector. Our second distribution is *Bernoulli-Gaussian* (BG),

$$x_i \sim \theta \cdot \mathcal{N}(\mu, \sigma_x^2) + (1-\theta) \cdot \delta(x_i), \tag{2}$$

where $\theta = \mathbb{P}(x_i \neq 0)$, $\mathcal{N}(\mu, \sigma_x^2)$ denotes a Gaussian distribution with mean μ and variance σ_x^2, and $\delta(\cdot)$ is the delta function [11]. The BG model is often used in sparse signal processing [12–14].

Scalar channels: In scalar channels,

$$\mathbf{y} = \mathbf{x} + \mathbf{z}, \tag{3}$$

where $\mathbf{x}, \mathbf{z} \in \mathbb{R}^N$ are the input signal and the additive noise, respectively. The noise $\mathbf{z}$ is i.i.d. Gaussian, $z_i \sim \mathcal{N}(0, \sigma_z^2)$. Given the noisy measurements $\mathbf{y}$, our goal is to find an estimate $\widehat{\mathbf{x}}$ such that $\mathbb{E}[\|\widehat{\mathbf{x}} - \mathbf{x}\|_2^2|\mathbf{y}]$ is minimized.

This work was supported in part by the National Science Foundation under Grant CCF-1217749 and in part by the U.S. Army Research Office under Grant W911NF-04-D-0003. This work was presented at the Information Theory and Application workshop (ITA), San Diego, CA, Feb. 2014.

Matrix channels: In matrix channels,

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{z}, \tag{4}$$

where $\mathbf{A} \in \mathbb{R}^{M \times N}$ is the measurement matrix, $\mathbf{z}$ represents the additive Gaussian noise, and $z_i \sim \mathcal{N}(0, \sigma_z^2)$. We assume that $\mathbf{A}$ is known while the parameters θ, μ, σ_x^2, and σ_z^2 are unknown. Given the measurements $\mathbf{y} \in \mathbb{R}^M$, our goal is to find an estimate $\widehat{\mathbf{x}}$ such that $\mathbb{E}[\|\widehat{\mathbf{x}} - \mathbf{x}\|_2^2 | \mathbf{y}]$ is minimized. This channel model (4) covers applications such as compressed sensing (CS) [15, 16].

II. Signal estimation in scalar channels

Let us denote the parameters of the input distributions by a vector $\mathbf{\Theta}$, i.e., for Bernoulli signals $\mathbf{\Theta} = (\theta)$ and for BG signals $\mathbf{\Theta} = (\theta, \mu, \sigma_x^2)$. If all the parameters in $\mathbf{\Theta}$ are known, then the *Bayesian estimator*, defined as the conditional expectation

$$\widehat{\mathbf{x}}^{\text{Bayes}} = \mathbb{E}[\mathbf{x} | \mathbf{y}, \mathbf{\Theta}], \tag{5}$$

achieves the MMSE. In the remainder of the paper, however, we consider settings where the parameters in $\mathbf{\Theta}$ are unknown.

To perform signal estimation in the EB framework, we could first perform ML estimation of the parameters $\mathbf{\Theta}$, and then plug the estimates into the Bayesian estimator (5), we call this estimator the Plug-in:

$$\widehat{\mathbf{x}}^{\text{Plug-in}} = \mathbb{E}[\mathbf{x} | \mathbf{y}, \widehat{\Theta}_{\text{ML}}]. \tag{6}$$

Instead of taking the estimated $\widehat{\Theta}_{\text{ML}}$ as the true parameters, the FB approach assigns a prior $f(\Theta)$ to Θ, thus the posterior $f(\Theta|\mathbf{y})$ can be computed. The model uncertainty is then incorporated by mixing over all possible Θ, we call this mixed estimator a *mixture denoiser* (MixD):

$$\widehat{\mathbf{x}}^{\text{MixD}} = \int \mathbb{E}[\mathbf{x} | \mathbf{y}, \Theta] f(\Theta | \mathbf{y}) d\Theta, \tag{7}$$

where $f(\mathbf{\Theta}|\mathbf{y}) \propto f(\mathbf{y}|\mathbf{\Theta}) f(\mathbf{\Theta})$. Note that the conditional expectation $E[\mathbf{x}|\mathbf{y}, \mathbf{\Theta}]$ in (7) is identical to the Bayesian estimator (5) when the prior distribution $f(\mathbf{x}|\mathbf{\Theta})$ is available and the true parameters are $\mathbf{\Theta}$. Meir and Zhang [17] proposed a similar Bayesian mixture framework in machine learning problems, while we apply this mixture approach to signal estimation. If we have some side information about Θ, then an informative prior that captures the side information can be applied to improve the estimation stability. MixD considers the settings when there is no side information about Θ, hence a proper noninformative prior needs to be applied.

We expect that when the prior $f(\Theta)$ is properly chosen, $\widehat{\mathbf{x}}^{\text{MixD}}$ can approach the MMSE in scalar channels (3) as the signal dimension N grows. The noninformative prior $f(\mathbf{\Theta})$ is unbiased to any particular $\mathbf{y}$, and hence does not strongly influence the posterior distribution $f(\mathbf{\Theta}|\mathbf{y})$. The uniform distribution is an intuitive but ad hoc choice for a noninformative prior. In contrast, the reference prior, introduced by Bernardo [18], maximizes the mutual information between the posterior and the prior distribution. For single parameter distributions, the reference prior has been shown to be equivalent to Jeffreys' prior [19], which is invariant to reparametrization, and hence is a desirable prior distribution.

It is well-known that the ML estimator asymptotically converges to the true parameter almost surely [1, 20], and thus the Plug-in signal estimator asymptotically converges to the Bayesian conditional expectation (5). It has also been verified [20] that a parameter estimated with a reference prior asymptotically converges to the true parameter asymptotically, and thus it can be conjectured that MixD converges to the Bayesian MMSE. Recent work by Verdú [21] has shown that the excess *mean square error* (MSE) caused by a mismatch between the true distribution f and the estimated distribution $\widehat{f}$ is twice the divergence $\mathbb{D}(f\|\widehat{f})$ [10] between f and $\widehat{f}$. In other words, if the divergence $\mathbb{D}(f\|\widehat{f})$ between the true distribution f and the estimated distribution $\widehat{f}$ converges to 0, then the excess MSE vanishes.

Note that MixD (7) is closely related to another estimator [7] that computes UCE with respect to a universal prior for the input signal $\mathbf{x}$. Consider the estimation for the first entry of $\mathbf{x}$,

$$\mathbb{E}[x_1 | \mathbf{y}] = \frac{P(x_1 = 1, \mathbf{y})}{P(x_1 = 1, \mathbf{y}) + P(x_1 = 0, \mathbf{y})}, \tag{8}$$

where

$$P(x_1, \mathbf{y}) = \sum_{\mathbf{x}_2^N \in \{0,1\}^{N-1}} P(x_1 \& \mathbf{x}_2^N) P(\mathbf{y}|\mathbf{x}),$$

$\mathbf{x}_2^N$ denotes the vector $(x_2, x_3, \ldots, x_N)$, and $\&$ denotes concatenation. We utilized *normalized maximum likelihood* (NML) [20] as the universal prior of $\mathbf{x}$ to compute UCE via (8). When the noisy measurements $\mathbf{y}$ are corrupted by a BSC, we have verified rigorously [7] for Bernoulli inputs that UCE computed via (8) asymptotically achieves the Bayesian MMSE. The main idea in our proof is to show that UCE asymptotically converges to the Plug-in estimator (6), which in turn asymptotically converges to the Bayesian estimator and achieves the MMSE. It can be shown that UCE via (8) is closely related to MixD (7). Keeping the rigorous result for the BSC in mind, we believe that UCE with other input and noise distributions asymptotically converges to the Bayesian estimator.

III. Signal estimation in matrix channels

We study the matrix channel signal estimation problem in the approximate message passing (AMP) [22] framework. AMP can be regarded as an iterative signal estimation algorithm in matrix channels that performs scalar denoising in each iteration. AMP with EB approaches, such as *expectation maximization* (EM) and ML, have been studied [23–25]. In this section, we first briefly review the AMP algorithm, and then apply MixD as the denoiser within AMP iterations.

A. Review of AMP

Consider a matrix channel model (4) where the signal distribution follows $x_i \sim f_X$ and the noise distribution follows $z_i \sim f_Z$. In the specific model (4) described in Section I-B, f_X is Bernoulli (1) or BG (2), and f_Z is $\mathcal{N}(0, \sigma_z^2)$. The measurement matrix $\mathbf{A}$ has i.i.d. Gaussian entries with unit norm columns on average, meaning that the matrix entries are $\mathcal{N}(0, \frac{1}{M})$ distributed and thus the expected value

of the column norm is 1. The AMP algorithm [22] proceeds iteratively according to

$$\mathbf{x}^{t+1} = \eta_t(\mathbf{A}^T\mathbf{r}^t + \mathbf{x}^t), \tag{9}$$

$$\mathbf{r}^t = \mathbf{y} - \mathbf{A}\mathbf{x}^t + \frac{1}{\delta}\mathbf{r}^{t-1}\langle\eta'_{t-1}(\mathbf{A}^T\mathbf{r}^{t-1} + \mathbf{x}^{t-1})\rangle, \tag{10}$$

where $\mathbf{A}^T$ is the transpose of $\mathbf{A}$, $\delta = M/N$ represents the measurement rate, $\eta_t(\cdot)$ is a denoising function, and $\langle\mathbf{u}\rangle = \frac{1}{N}\sum_{i=1}^{N} u_i$ for some vector $\mathbf{u} = (u_1, u_2, \ldots, u_N)$. In the t-th iteration, we obtain the vectors $\mathbf{x}^t \in \mathbb{R}^N$ and $\mathbf{r}^t \in \mathbb{R}^M$. The denoising function $\eta_t(\cdot)$ is separable, meaning that it is applied component-wise to the noisy measurements, and $\eta'_t(\mathbf{s}) = \frac{\partial}{\partial \mathbf{s}}\eta_t(\mathbf{s})$. We highlight that the vector $\mathbf{A}^T\mathbf{r}^t + \mathbf{x}^t \in \mathbb{R}^N$ in (9) can be regarded as noisy measurements of $\mathbf{x}$ in the t-th iteration with noise variance σ_t^2. The asymptotic performance of AMP can be characterized by a *state evolution* (SE) formalism:

$$\sigma_{t+1}^2 = \sigma_z^2 + \frac{1}{\delta}\mathbb{E}\left[\left(\eta_t\left(X + \sigma_t^2 W\right) - X\right)^2\right], \tag{11}$$

where the random variables $W \sim \mathcal{N}(0,1)$ and $X \sim f_X$. Formal statements for SE appear in [26]. SE (11) implies that in each iteration of AMP, the denoiser estimates $\mathbf{x}$ from a scalar channel.

A soft-thresholding denoiser is applied in the original derivation of AMP [22], where the optimal threshold of the denoiser in each AMP iteration can be estimated without knowing the input distribution [27, 28]. Donoho et al. [29] generalized the denoising operator to be various minimax denoisers. With a minimax denoiser, the resulting AMP algorithm is robust to different signal distributions, but may not achieve the Bayesian MMSE.

B. AMP with MixD

We have discussed in Section II that MixD is MMSE optimal for the scalar channel (3). If we apply MixD as the denoiser $\eta_t(\cdot)$ in each AMP step (9), then (11) is MMSE optimal in each iteration. Therefore, it can be expected that using MixD as the denoiser in AMP may achieve the MMSE for the matrix channel (4) (we consider the regions where AMP can achieve the MMSE when the exact distribution of the input is known; cases where the MMSE is not achieved are discussed by Krzakala et al. [23] and Zhu and Baron [30]). In order to make MixD work inside AMP, we need to estimate the effective Gaussian noise in each AMP iteration. The estimated noise variance $\widehat{\sigma}_t^2$ can be calculated as [31]:

$$\widehat{\sigma}_t^2 = \frac{1}{M}\sum_{i=1}^{M}(r_i^t)^2, \tag{12}$$

where $\mathbf{r}^t$ is defined in (10). In each iteration of AMP, we replace the denoiser $\eta_t(\cdot)$ in (9) by $\eta_t^{\text{MixD}}(\mathbf{A}^T\mathbf{r}^t + \mathbf{x}^t, \widehat{\sigma}_t^2)$, which is computed using (7) with $\mathbf{y} = \mathbf{A}^T\mathbf{r}^t + \mathbf{x}^t$ and $\sigma_z^2 = \widehat{\sigma}_t^2$.

IV. Numerical Results

A. MixD in scalar channels

In this subsection, we compare the MSE of MixD and Plug-in to the Bayesian MMSE in scalar channels (3).

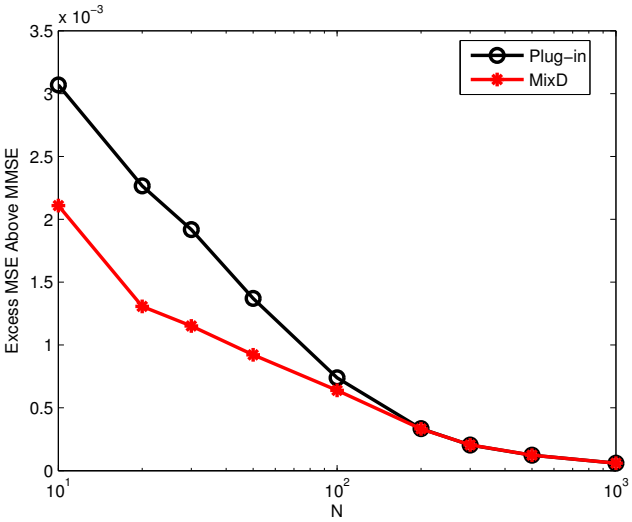

Figure 1: **Bernoulli input measured through scalar channel:** We plot the excess MSE above the MMSE achieved by MixD and the Plug-in as functions of N. MixD has better performance for small N, and both MixD and Plug-in achieve the MMSE asymptotically. (Signal dimension $N = 10{-}1,000$, Bernoulli parameter $\theta = 0.05$, and noise variance $\sigma_z^2 = 0.1$.)

Settings: In the Bernoulli case, the Bernoulli parameter is 0.05, and the Gaussian noise is $\mathcal{N}(0, 0.1)$. The input signal dimension N is evaluated from 10 up to 1,000. We use Jeffreys' prior for the Bernoulli parameter θ,

$$f_{\text{Jeffreys}}(\theta) = \frac{1}{\sqrt{\pi\theta(1-\theta)}}. \tag{13}$$

In the BG case, the Bernoulli parameter is 0.1, the Gaussian part of the signal is $\mathcal{N}(0,1)$, and the noise is $\mathcal{N}(0, 0.1)$. We use Jeffreys' prior (13) for the parameter θ, a uniform prior for μ,

$$f(\mu) = \begin{cases} \frac{1}{4} & \text{if } \mu \in [-2, 2] \\ 0 & \text{else} \end{cases},$$

and a uniform prior for σ_x,

$$f(\sigma_x) = \begin{cases} \frac{1}{2} & \text{if } \sigma_x \in [0, 2] \\ 0 & \text{else} \end{cases}.$$

Note that we use a uniform prior over the standard deviation σ_x and not the variance σ_x^2. Our current implementation limits the ranges of μ and σ_x to $[-2, 2]$ and $[0, 2]$, respectively, and the extension to arbitrary ranges is ongoing work.

Results: Denote the MSE of MixD and Plug-in by e_M and e_P, respectively. We compare the performance of MixD and Plug-in by plotting the excess MSE ($e_M - \text{MMSE}$) and ($e_P - \text{MMSE}$) as functions of N. Figures 1 and 2 illustrate the results for Bernoulli and BG inputs, respectively. It can be seen from Figures 1 and 2 that MixD achieves lower MSE than the Plug-in when N is comparatively small; MixD performs especially well for BG signals. As N increases, ($e_M - \text{MMSE}$) and ($e_P - \text{MMSE}$) tend to zero, suggesting that both MixD and Plug-in asymptotically achieve the MMSE.

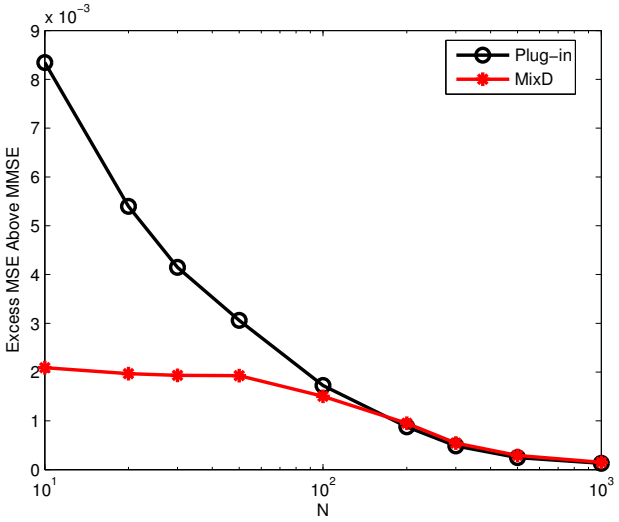

Figure 2: **BG input measured through scalar channel:** We plot the excess the MSE above the MMSE achieved by MixD and the Plug-in as functions of N. MixD has better performance for small N, and both MixD and Plug-in achieve the MMSE asymptotically. (Signal dimension $N = 10-1,000$, Bernoulli parameter $\theta = 0.1$, Gaussian mean $\mu = 0$, variance $\sigma_x^2 = 1$, and noise variance $\sigma_z^2 = 0.1$.)

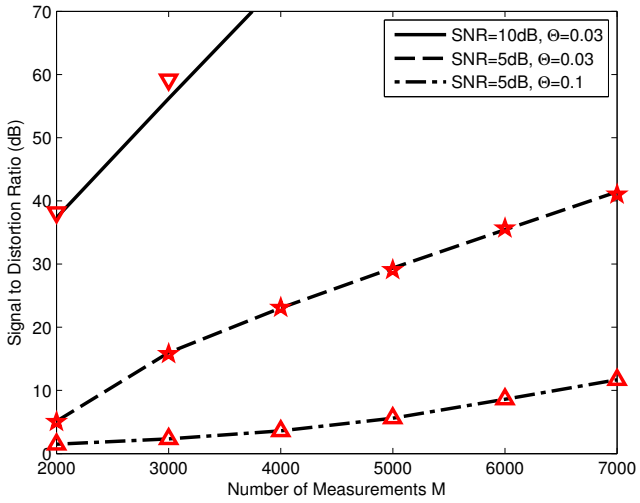

Figure 3: **Bernoulli input measured through matrix channel:** We plot the signal to distortion ratio as a function of the number of measurements M for AMP-MixD. The curves correspond to the MMSE and the markers (triangles and stars) represent the MSE performance of AMP-MixD, which coincides nicely with the theoretically optimal MMSE. (Signal dimension $N = 10,000$.)

B. AMP-MixD in matrix channels

In this subsection, we evaluate the performance of AMP-MixD for both Bernoulli and BG inputs.

Settings: The measurement matrix $\mathbf{A}$ has i.i.d. Gaussian entries distributed as $\mathcal{N}(0, \frac{1}{M})$. Under this setting, the signal to noise ratio (SNR) of the matrix channel (4) is defined as $\frac{N \cdot \text{Var}(\mathbf{x})}{M \cdot \text{Var}(\mathbf{z})}$, where $\text{Var}(\cdot)$ denotes variance.

For Bernoulli inputs, the signal dimension $N = 10,000$, and the number of measurements M varies from 2,000 to 7,000. The Bernoulli parameter $\theta = 0.03$ or 0.1, and the SNR is 5 dB or 10 dB.

For BG inputs, the signal dimension $N = 5,000$, and the number of measurements M varies from 1,000 to 2,500. The Bernoulli parameter $\theta = 0.1$, the Gaussian mean $\mu = 0$, the variance $\sigma_x^2 = 1$, and the SNR is 10dB or 25 dB.

Results: Figures 3 and 4 demonstrate the performance of AMP-MixD for the Bernoulli case and the zero-mean BG case, respectively. The horizontal axis represents the number of measurements M, and the vertical axis represents the signal to distortion ratio, which is defined as the ratio between the signal variance and the MSE of AMP-MixD. The curves correspond to the theoretically optimal MMSE performance, and the markers (triangles and stars) represent the performance of AMP-MixD, which coincides nicely with the theoretically optimal performance.

Finally, we also simulated the state-of-art algorithm EM-BG [14] and AMP-MixD for the nonzero-mean BG input case, and found that the performance of AMP-MixD and EM-BG are comparable. For brevity, results are not included.

V. Discussion

In this paper, we have shown that the MixD approaches the MMSE in solving signal estimation problems while adapting to unknown parametric distributions. Although the results of this paper focus on the Bernoulli and BG parametric distributions, the concepts can be extended to other distributions, in particular non-i.i.d. signals. While the Plug-in also approaches the MMSE when the signal dimension N increases, readers may notice from Figures 1 and 2 that MixD approaches the MMSE faster. For example, in Figure 1 the MSE of MixD is $1.5 \cdot 10^{-3}$ above the MMSE when $N \approx 15$, while the Plug-in achieves the same excess MSE when $N \approx 40$. In addition to the precision of the signal estimation procedure, another criterion for comparing algorithms is their speed. We have noticed that our implementation of the Plug-in runs faster than MixD, which indicates that the Plug-in could be advantageous in some applications where computational speed is of paramount importance. We leave the study of trade-offs between estimation quality and computational requirements for future work.

Acknowledgments

We thank Arian Maleki for detailed explanations about his recent work on parameterless AMP [28]; Phil Schniter for enlightening conversations that significantly influenced this work; Liyi Dai for inspiring conversations; and Junan Zhu for commenting on the manuscript. Special thanks to Neri Merhav for providing the original idea for our mixture denoiser.

References

[1] B. Levy, *Principles of signal detection and parameter estimation*. New York, NY, USA: Springer Verlag, 2008.

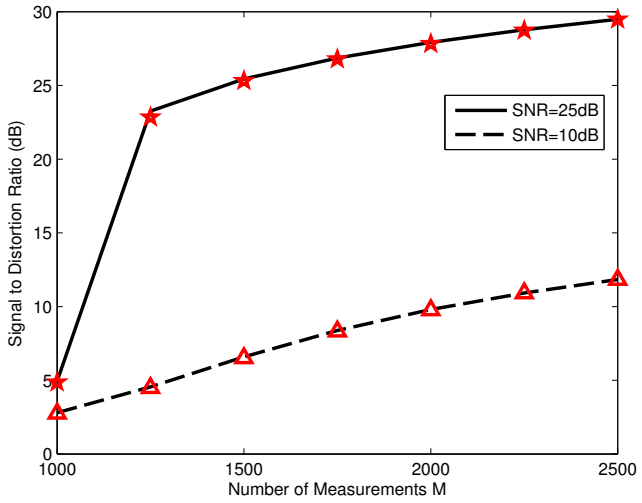

Figure 4: **BG input measured through matrix channel:** We plot the signal to distortion ratio as a function of the number of measurements M for AMP-MixD. The curves correspond to the MMSE and the markers (triangles and stars) represent the MSE performance of AMP-MixD, which coincides nicely with the theoretically optimal MMSE. (Signal dimension $N = 5,000$, Bernoulli parameter $\theta = 0.1$, Gaussian mean $\mu = 0$, variance $\sigma_x^2 = 1$.)

[2] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inf. Theory*, vol. 30, no. 4, pp. 629–636, July 1984.

[3] P. D. Grunwald, *The minimum description length principle (Adaptive computation and machine learning series)*. Prentice-Hall, 2007.

[4] D. L. Donoho, "The Kolmogorov sampler," Stanford University, Stanford, CA, Department of Statistics Technical Report 2002-4, Jan. 2002.

[5] D. Baron and M. F. Duarte, "Universal MAP estimation in compressed sensing," in *Proc. 49th Annual Allerton Conf. Comm., Control, Computing*, Sept. 2011, pp. 768–775.

[6] H. Robbins, *The empirical Bayes approach to statistical decision problems*. Springer, 1985.

[7] J. Tan, N. Krishnan, and D. Baron, "Universal estimation for Bernoulli signals via universal conditional expectation," *in preparation*, May 2014.

[8] D. Baron, "Information complexity and estimation," in *Fourth Workshop Inf. Theoretic Methods Science Eng. (WITMSE 2011)*, Aug. 2011.

[9] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Trans. Inf. Theory*, vol. 23, no. 3, pp. 337–343, May 1977.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley-Interscience, 2006.

[11] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. McGraw Hill Book Co., 1991.

[12] D. Guo and C. C. Wang, "Random sparse linear systems observed via arbitrary channels: A decoupling principle," in *Proc. IEEE Int. Symp. Inf. Theory*, June 2007, pp. 946–950.

[13] J. Starck, F. Murtagh, and J. Fadili, *Sparse image and signal processing: Wavelets, curvelets, morphological diversity*. Cambridge Univ. Press, 2010.

[14] J. Vila and P. Schniter, "Expectation-maximization Bernoulli-Gaussian approximate message passing," in *Proc. IEEE 45th Asilomar Conf. Signals, Syst., and Comput.*, Nov. 2011, pp. 799–803.

[15] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[16] D. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[17] R. Meir and T. Zhang, "Generalization error bounds for Bayesian mixture algorithms," *J. Mach. Learn. Res.*, vol. 4, pp. 839–860, Dec. 2003.

[18] J. M. Bernardo, "Reference posterior distributions for Bayesian inference," *J. Roy. Stat. Soc. Series B (Methodological)*, vol. 41, no. 2, pp. 113–147, Jan. 1979.

[19] B. S. Clarke and A. R. Barron, "Jeffreys' prior is asymptotically least favorable under entropy risk," *J. Stat. Planning Inference*, vol. 41, no. 1, pp. 37–60, Aug. 1994.

[20] J. Rissanen, *Optimal Estimation of Parameters*. Cambridge University Press, 2012.

[21] S. Verdú, "Mismatched estimation and relative entropy," *IEEE Trans. Inf. Theory*, vol. 56, no. 8, pp. 3712–3720, Aug. 2010.

[22] D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing," *Proc. Nat. Acad. Sci.*, vol. 106, no. 45, pp. 18 914–18 919, Nov. 2009.

[23] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová, "Probabilistic reconstruction in compressed sensing: Algorithms, phase diagrams, and threshold achieving matrices," *J. Stat. Mech. - Theory E.*, vol. 2012, no. 08, p. P08009, 2012.

[24] J. Vila and P. Schniter, "Expectation-maximization Gaussian-mixture approximate message passing," *IEEE Trans. Signal Process.*, vol. 61, no. 19, pp. 4658–4672, Oct. 2013.

[25] U. Kamilov, S. Rangan, A. K. Fletcher, and M. Unser, "Approximate message passing with consistent parameter estimation and applications to sparse learning," in *Workshop Neural Info. Proc. Sys. (NIPS)*, Dec. 2012, pp. 2447–2455.

[26] A. Montanari, "Graphical models concepts in compressed sensing," *Compressed Sensing: Theory and Applications*, pp. 394–438, 2012.

[27] D. Donoho and G. Reeves, "Achieving Bayes MMSE performance in the sparse signal + Gaussian white noise model when the noise level is unknown," in *Proc. Int. Symp. Inf. Theory (ISIT2013)*, July 2013, pp. 101–105.

[28] A. Mousavi, A. Maleki, and R. Baraniuk, "Parameterless optimal approximate message passing," *Arxiv preprint arXiv:1311.0035*, Oct. 2013.

[29] D. Donoho, I. Johnstone, and A. Montanari, "Accurate prediction of phase transitions in compressed sensing via a connection to minimax denoising," *IEEE Trans. Inf. Theory*, vol. 59, no. 6, pp. 3396–3433, June 2013.

[30] J. Zhu and D. Baron, "Performance regions in compressed sensing from noisy measurements," in *Proc. 2013 Conf. Inf. Sciences Systems*, Baltimore, MD, Mar. 2013.

[31] A. Montanari, "Graphical models concepts in compressed sensing," *Arxiv preprint arXiv:1011.4328v3*, Mar. 2011.