Mismatched Estimation in Large Linear Systems

Yanting Ma,[†] Dror Baron,[†] Ahmad Beirami[*]

[†]Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27695, USA
[*]Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708, USA
Email: [†]{yma7, barondror}@ncsu.edu, [*]ahmad.beirami@duke.edu

***Abstract*—We study the excess mean square error (EMSE) above the minimum mean square error (MMSE) in large linear systems where the posterior mean estimator (PME) is evaluated with a postulated prior that differs from the true prior of the input signal. We focus on large linear systems where the measurements are acquired via an independent and identically distributed random matrix, and are corrupted by additive white Gaussian noise (AWGN). The relationship between the EMSE in large linear systems and EMSE in scalar channels is derived, and closed form approximations are provided. Our analysis is based on the decoupling principle, which links scalar channels to large linear system analyses. Numerical examples demonstrate that our closed form approximations are accurate.**

***Index Terms*—decoupling, large linear systems, mismatched estimation.**

I. Introduction

The posterior mean estimator (PME), also known as conditional expectation, plays a pivotal role in Bayesian estimation. To compute the PME, we need a prior distribution for the unknown signal. In cases where the prior is unavailable, we may compute the PME with a postulated prior, which may not match the true prior. Verdú [1] studied the mismatched estimation problem for scalar additive white Gaussian noise (AWGN) channels and quantified the excess mean square error (EMSE) above the minimum mean square error (MMSE) due to the incorrect prior. A natural extension to Verdú's result would be to quantify the EMSE due to mismatched estimation in large linear systems [2–7].

Mismatched estimation: Consider scalar estimation,

$$Y = X + \sigma W, \tag{1}$$

where X is generated by some probability density function (pdf) p_X, $W \sim \mathcal{N}(0,1)$ is independent of X, and $\mathcal{N}(\mu, \sigma^2)$ denotes the Gaussian pdf with mean μ and variance σ^2. A PME with some prior q_X, which is defined as

$$\widehat{X}_q(y;\sigma^2) = \mathbb{E}_{q_X}[X|Y=y], \tag{2}$$

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-14-1-0314.

can be applied to the estimation procedure, where in $\mathbb{E}_{q_X}[\cdot]$ expectation is calculated assuming that X is distributed as q_X. The mean square error (MSE) achieved by $\widehat{X}_q(y;\sigma^2)$ is

$$\Psi_q(\sigma^2) = \mathbb{E}_{p_X}\left[(\widehat{X}_q(\cdot;\sigma^2) - X)^2\right]. \tag{3}$$

Note that $\widehat{X}_p(y;\sigma^2)$ is the MMSE estimator, and $\Psi_p(\sigma^2)$ is the MMSE.

Having defined notation for the MSE, we can now define the EMSE above the MMSE due to the mismatched prior in the scalar estimation problem (1):

$$\text{EMSE}_s(\sigma^2) = \Psi_q(\sigma^2) - \Psi_p(\sigma^2), \tag{4}$$

where the subscript s represents scalar estimation.

Verdú [1] proved that EMSE_s is related to the relative entropy [8] as follows:

$$\text{EMSE}_s(\sigma^2) = \frac{\mathsf{d}}{\mathsf{d}\gamma} D\left(P_X \star \mathcal{N}\left(0,\frac{1}{\gamma}\right) \middle\| Q_X \star \mathcal{N}\left(0,\frac{1}{\gamma}\right)\right),$$

where $\star$ represents convolution, $\gamma = 1/\sigma^2$, P_X and Q_X are the true and mismatched probability distributions, respectively, and under the assumption that P is absolutely continuous with respect to Q, the relative entropy, also known as Kullback-Leibler divergence [8], is defined as $D(P\|Q) = \int \log\left(\frac{\mathsf{d}P}{\mathsf{d}Q}\right)\mathsf{d}P$.

Large linear system: Consider a linear system:

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{z}, \tag{5}$$

where the input signal $\mathbf{x} \in \mathbb{R}^N$ is a sequence of independent and identically distributed (i.i.d.) random variables generated by some pdf p_X, $\mathbf{A} \in \mathbb{R}^{M\times N}$ is an i.i.d. Gaussian random matrix with $A_{ij} \sim \mathcal{N}(0, 1/\sqrt{M})$, and $\mathbf{z} \in \mathbb{R}^M$ is AWGN with mean zero and variance σ_z^2. Large linear systems [2–7], which are sometimes called the large system limit of linear systems, refer to the limit that both N and M tend to infinity but with their ratio converging to a positive number, i.e., $N \to \infty$ and $M/N \to \delta$. We call δ the measurement rate of the linear system.

Decoupling principle: The decoupling principle [3] is based on replica analysis of randomly spread code-division multiple access (CDMA) and multiuser detection. It claims

the following result. Define the PME with prior q_X of the linear system (5) as

$$\widehat{\mathbf{x}}_q = \mathbb{E}_{q_X}[\mathbf{x}|\mathbf{y}]. \tag{6}$$

Denote the j-th element of the sequence $\mathbf{x}$ by $x(j)$. For any $j \in \{1, 2, ..., N\}$, the joint distribution of $(x(j), \widehat{x}_q(j))$ converges in probability to the joint distribution of $(X, \widehat{X}_q(\cdot; \sigma_q^2))$ in large linear systems, where $X \sim p_X$ and σ_q^2 is the solution to the following fixed point equation [2, 3],

$$\delta \cdot (\sigma^2 - \sigma_z^2) = \Psi_q(\sigma^2). \tag{7}$$

Note that $\widehat{X}_q(\cdot; \sigma_q^2)$ is defined in (2), and there may be multiple fixed points [3, 9, 10].

The decoupling principle provides a single letter characterization of the MSE; state evolution [11] in approximate message passing (AMP) [6] provides rigorous justification for the achievable part of this characterization. An extension of the decoupling principle to a collection of any finite number of elements in $\mathbf{x}$ is provided by Guo et al. [5].

Illustration: Figure 1 highlights our contribution, which is stated in Claim 1, using an example that compares the EMSE in scalar channels and large linear systems. The solid line with slope δ represents the linear function of σ^2 on the left-hand-side of (7). The dashed and dash-dotted curves represent $\Psi_p(\sigma^2)$ and $\Psi_q(\sigma^2)$, respectively. The intersection point a provides the solution to the fixed point equation (7) when $\widehat{X}_p$ is applied, and so the Cartesian coordinate representation of a is $a = (\sigma_p^2, \Psi_p(\sigma_p^2))$. Similarly, we have $b = (\sigma_q^2, \Psi_q(\sigma_q^2))$. Therefore, the vertical distance between a and b is the difference in MSE of the two decoupled channels for the PME with the true prior and the mismatched prior, respectively. Based on the decoupling principle, the vertical distance is equivalent to the EMSE in large linear systems, which we denote by EMSE_l and define as

$$\text{EMSE}_l = \lim_{N\to\infty} \frac{1}{N} \mathbb{E}_{p_X} \left[\|\widehat{\mathbf{x}}_q - \mathbf{x}\|_2^2 - \|\widehat{\mathbf{x}}_p - \mathbf{x}\|_2^2 \right], \tag{8}$$

where the subscript l represents large linear systems, $\widehat{\mathbf{x}}_q$ is defined in (6), and $\widehat{\mathbf{x}}_p$ is defined similarly. The pair $c = (\sigma_p^2, \Psi_q(\sigma_p^2))$ represents the MSE achieved by $\widehat{X}_q$ in the decoupled scalar channel for $\widehat{X}_p$, and so the vertical distance between a and c is $\text{EMSE}_s(\sigma_p^2)$. We can see from Figure 1 that EMSE_l is larger than $\text{EMSE}_s(\sigma_p^2)$, despite using the same mismatched prior for PME, because the decoupled scalar channel becomes noisier as indicated by the horizontal move from c to b; we call this an amplification effect. *Our contribution is to formalize the amplification of $\text{EMSE}_s(\sigma_p^2)$ to EMSE_l.*

The remainder of the paper is organized as follows. We derive the relationship between the EMSE in large linear systems and EMSE in scalar channels in Section II; closed form approximations that characterize this relationship are also provided. Numerical examples that evaluate the accuracy level of our closed form approximations are presented in Section III, and Section IV concludes the paper.

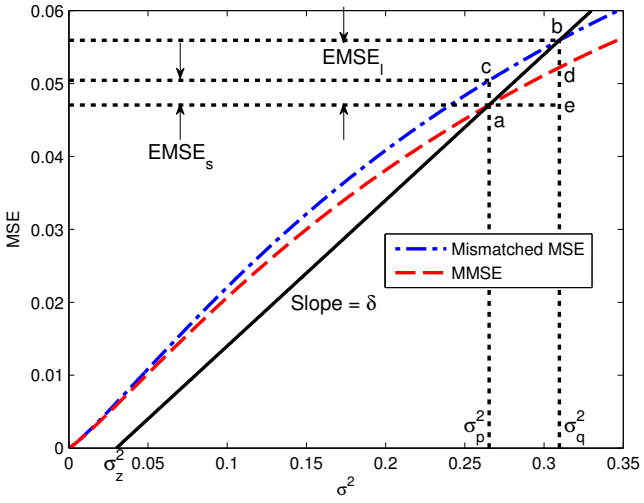

Figure 1: Mismatched estimation for Bernoulli-Gaussian input signal ($p_X(x) = (1-\theta)\delta_0(x) + \theta\mathcal{N}(0,1)$) in large linear systems. We notice that the EMSE in the scalar estimation problem is *amplified* in the large linear system, despite using the same mismatched prior for PME, because the decoupled scalar channel becomes noisier as indicated by the horizontal move from c to b. ($\delta = 0.2$, $\sigma_z^2 = 0.03$, the true sparsity parameter $\theta = 0.1$, and the mismatched sparsity parameter $\widetilde{\theta} = 0.2$.)

II. Main Results

We now characterize the relationship between EMSE_l and $\text{EMSE}_s(\sigma_p^2)$, which is the EMSE of the mismatched PME in the decoupled scalar AWGN channel for the PME with the true prior p_X.

Our main result is summarized in the following claim. We call this result a claim, because it relies on the decoupling principle [3], which is based on the replica method and lacks rigorous justification.

Claim 1. *Consider a large linear system (5). Let the EMSE in scalar channels $\text{EMSE}_s(\sigma^2)$ and EMSE in large linear systems EMSE_l be defined in (4) and (8), respectively, and let σ_p^2 be the noise variance in the decoupled scalar channel when the true prior p_X is applied. Denote $\Psi_q'(\sigma^2) = \frac{d}{d(\sigma^2)}\Psi_q(\sigma^2)$, where $\Psi_q(\sigma^2)$ is the MSE in the scalar AWGN channel with noise variance σ^2 achieved by the PME with a mismatched prior q_X (3). In large linear systems, EMSE_l and $\text{EMSE}_s(\sigma_p^2)$ satisfy the following relation:*

$$\text{EMSE}_l = \text{EMSE}_s(\sigma_p^2) + \int_{\sigma_p^2}^{\sigma_p^2 + \frac{1}{\delta}\text{EMSE}_l} \Psi_q'(\sigma^2) d\sigma^2. \tag{9}$$

Justification: The fixed point equations (7) when applying $\widehat{X}_p$ and $\widehat{X}_q$ are

$$\delta \cdot \left(\sigma_p^2 - \sigma_z^2\right) = \Psi_p(\sigma_p^2) \quad \text{and} \tag{10}$$

$$\delta \cdot \left(\sigma_q^2 - \sigma_z^2\right) = \Psi_q(\sigma_q^2), \tag{11}$$

respectively. Subtract (10) from (11):

$$\delta \cdot \left(\sigma_q^2 - \sigma_p^2\right) = \Psi_q(\sigma_q^2) - \Psi_p(\sigma_p^2). \tag{12}$$

Recall that

$$\begin{aligned} \text{EMSE}_l &= \Psi_q(\sigma_q^2) - \Psi_p(\sigma_p^2), \\ \text{EMSE}_s(\sigma_p^2) &= \Psi_q(\sigma_p^2) - \Psi_p(\sigma_p^2). \end{aligned} \tag{13}$$

Combining the two equations:

$$\begin{aligned} \text{EMSE}_l &= \text{EMSE}_s(\sigma_p^2) + \Psi_q(\sigma_q^2) - \Psi_q(\sigma_p^2) \\ &= \text{EMSE}_s(\sigma_p^2) + \int_{\sigma_p^2}^{\sigma_p^2+\left(\sigma_q^2-\sigma_p^2\right)} \Psi_q'(\sigma^2)\mathsf{d}\sigma^2, \end{aligned} \tag{14}$$

where (9) follows by noticing from (12) and (13) that $\sigma_q^2 - \sigma_p^2 = \frac{1}{\delta}\text{EMSE}_l$. □

Approximations: Consider a Taylor expansion of $\Psi_q(\sigma_q^2)$ at σ_p^2:

$$\begin{aligned} \Psi_q(\sigma_q^2) &= \Psi_q(\sigma_p^2) + \alpha(\sigma_q^2 - \sigma_p^2) + \frac{1}{2}\beta(\sigma_q^2 - \sigma_p^2)^2 \\ &\quad + o((\sigma_q^2 - \sigma_p^2)^2) \\ &= \Psi_q(\sigma_p^2) + \alpha\frac{\text{EMSE}_l}{\delta} + \frac{\beta}{2}\left(\frac{\text{EMSE}_l}{\delta}\right)^2 \\ &\quad + o\left(\Delta^2\right), \end{aligned} \tag{15}$$

where $\alpha = \Psi_q'(\sigma_p^2)$ and $\beta = \Psi_q''(\sigma_p^2)$ are the first and second derivatives, respectively, of $\Psi_q(\cdot)$ evaluated at σ_p^2, and $\Delta = \sigma_q^2 - \sigma_p^2$.[1]

Plug (15) into (14):

$$\text{EMSE}_l = \text{EMSE}_s(\sigma_p^2) + \alpha\frac{\text{EMSE}_l}{\delta} + \frac{\beta}{2}\left(\frac{\text{EMSE}_l}{\delta}\right)^2 + o\left(\Delta^2\right). \tag{16}$$

If we only keep the first order terms in (16) and solve for EMSE_l, then the first order approximation of EMSE_l is obtained in closed form:

$$\text{EMSE}_l = \frac{\delta}{\delta - \alpha}\text{EMSE}_s(\sigma_p^2) + o\left(\Delta\right). \tag{17}$$

Note that $\frac{\delta}{\delta-\alpha} > 1$. That is, the EMSE in the scalar estimation problem is *amplified* in large linear systems despite using the same mismatched prior q_X for PME. This is due to an increased noise variance in the decoupled channel for the mismatched prior beyond the variance for the correct prior.

[1] $h(\Delta) = o(g(\Delta))$ if $\lim_{\Delta\to 0}\frac{h(\Delta)}{g(\Delta)} = 0$.

TABLE I: Relative error in the Bernoulli example

$\widetilde{\theta}$	Δ	1st (17)	2nd (18)	2nd (19)
0.11	0.0008	0.13%	<0.0005%	<0.0001%
0.13	0.0070	1%	0.041%	0.017%
0.15	0.0178	2.8%	0.28%	0.11%
0.17	0.0324	5.2%	0.99%	0.35%
0.20	0.0603	10%	4%	1%

Similarly, the second order approximation is given by

$$\text{EMSE}_l = \frac{\delta\text{EMSE}_s(\sigma_p^2)}{\delta - \alpha}\left(1 + \frac{1}{2}\frac{\beta\text{EMSE}_s(\sigma_p^2)}{(\delta-\alpha)^2}\right) + o\left(\Delta^2\right), \tag{18}$$

under the condition that $\Psi_q(\sigma^2)$ is locally concave in σ^2 around σ_p^2; details in the Appendix. A more accurate approximation is also presented in the Appendix in (19).

We expect that when the mismatched distribution q_X is close to p_X, $\Psi_p(\sigma^2)$ and $\Psi_q(\sigma^2)$ are close to each other for all σ^2, and $\Delta = \sigma_q^2 - \sigma_p^2$ is smaller for minor mismatch than significant mismatch with the same slope of $\Psi_q(\sigma^2)$ at σ_p^2 and the same δ. Therefore, the first order approximation of $\Psi_q(\sigma^2)$ for $\sigma^2 \in [\sigma_p^2, \sigma_p^2 + \Delta]$ is more likely to be reasonably accurate for minor mismatch. When the mismatch is significant, we might need to include the second order term in the Taylor expansion (15) to improve accuracy. Numerical examples that show the necessity of the second order term when there is significant mismatch will be shown in Section III.

III. Numerical Examples

A. Accuracy of approximations

We begin with two examples that examine the accuracy level of our first and second order approximations given by (17)-(19).

Example 1: Bernoulli input. The input signal of the first example follows a Bernoulli distribution with $p_X(1) = \theta$ and $p_X(0) = 1 - \theta$. Let $\theta = 0.1$, and let the mismatched Bernoulli parameter $\widetilde{\theta}$ vary from 0.11 (minor mismatch) to 0.20 (significant mismatch). The linear system (5) has measurement rate $\delta = 0.2$ and measurement noise variance $\sigma_z^2 = 0.03$. Using (7), this linear system with the true Bernoulli prior yields $\sigma_p^2 = 0.34$. Table I shows the accuracy level of our three approximations (17), (18) and (19) for the Bernoulli example. The relative error of the predicted EMSE_l, which is denoted by $\text{EMSE}_l(\text{pred})$, is defined as $|\text{EMSE}_l - \text{EMSE}_l(\text{pred})|/\text{EMSE}_l$, where the first and second order approximations for $\text{EMSE}_l(\text{pred})$ are given by (17)-(19), respectively.

Example 2: Bernoulli-Gaussian input. Here the input signal follows a Bernoulli-Gaussian distribution $p_X(x) = \theta\mathcal{N}(0,1) + (1-\theta)\delta_0(x)$, where $\delta_0(\cdot)$ is the delta function [12]. Let $\theta = 0.1$, and let the mismatched parameter $\widetilde{\theta}$ vary from 0.11 to 0.20 as before. The linear system is the same as in

TABLE II: Relative error in the Bernoulli-Gaussian example

$\widetilde{\theta}$	Δ	1st (17)	2nd (18)	2nd (19)
0.11	0.0006	0.25%	0.09%	0.09%
0.13	0.0052	1.4%	0.04%	0.08%
0.15	0.0132	3.5%	0.21%	0.03%
0.17	0.0240	6.5%	0.98%	0.08%
0.20	0.0444	13%	4%	0.4%

the Bernoulli example with $\delta = 0.2$ and $\sigma_z^2 = 0.03$, and this linear system with the correct Bernoulli-Gaussian prior yields $\sigma_p^2 = 0.27$. Figure 1 compares the PME with the true parameter and the PME with the mismatched parameter $\widetilde{\theta} = 0.2$. The accuracy level of our approximations (17)-(19) for the Bernoulli-Gaussian example is shown in Table II.

It can be seen from Tables I and II that when the mismatch is minor, the first order approximation can achieve relative error below 1%. However, as the mismatch increases, we may need to include the second order term to reduce error.

B. Amplification for non-i.i.d. signals

To show that the results of this paper are truly useful, it would be interesting to evaluate whether the amplification effect can be used to characterize the performance of more complicated problems that feature mismatch. By their nature, complicated problems may be difficult to characterize in closed form, not to mention that the theory underlying them may not be well-understood. Therefore, we will pursue a setting where a large linear system is solved for non-i.i.d. signals. Note that neither state evolution [11] nor the decoupling principle [3] has been developed for such non-i.i.d. settings, and so any predictive ability would be heuristic. We now provide such an example, and demonstrate that we can predict the MSE of one ad-hoc algorithm from that of another.

Example 3: Markov-constant input. Our Markov-constant signal is generated by a two-state Markov state machine that contains states s_0 (zero state) and s_1 (nonzero state). The signal values in states s_0 and s_1 are 0 and 1, respectively. Our transition probabilities are $p(s_0|s_1) = 0.2$ and $p(s_1|s_0) = 1/45$, which yield 10% nonzeros in the Markov-constant signal. In words, this is a block sparse signal whose entries typically stay "on" with value 1 roughly 5 times, and then remain "off" for roughly 45 entries. The optimal PME for this non-i.i.d. signal requires the posterior to be conditioned on the entire observed sequence, which seems computationally intractable. Instead, we consider two sub-optimal yet practical estimators where the conditioning is limited to signal entries within windows of sizes two or three. We call these practical approaches Bayesian sliding window denoisers:

$$\widehat{x}_2(i) = \mathbb{E}\left[x(i)|(y(i-1), y(i))\right],$$
$$\widehat{x}_3(i) = \mathbb{E}\left[x(i)|(y(i-1), y(i), y(i+1))\right].$$

Increasing the window-size would capture more statistical information about the signal, resulting in better MSE performance. Unfortuntely, a larger window-size requires more computation.

To decide whether a fast and simple yet less accurate denoiser should be utilized instead of a more accurate yet slower denoiser while still achieving an acceptable MSE, we must predict the difference in MSE performance between the two. Such a prediction is illustrated in Figure 2. The dash-dotted curve represents the MSE achieved by AMP using the better denoiser $\widehat{x}_3$ as a function of the measurement rate δ, where for each δ, we obtain a solution to the fixed point equation (7), which is denoted by $\sigma^2(\delta)$. The dashed curve represents the MSE achieved by $\widehat{x}_2$ in scalar channels with noise variance $\sigma^2(\delta)$. Note, however, that $\widehat{x}_2$ uses window-size 2 whereas $\widehat{x}_3$ uses window-size 3 and can gather more information about the signal. Therefore, $\widehat{x}_2$ obtains higher (worse) MSE despite denoising a statistically identical scalar channel. The vertical distance between the dashed and dash-dotted curves is EMSE_s.

What MSE performance can be obtained by using the faster denoiser $\widehat{x}_2$ within AMP? The predicted MSE performance using $\widehat{x}_2$ within AMP is depicted with crosses; the prediction relies on our second order approximation (19). That is, the vertical distance between the crosses and dash-dotted curve is EMSE_l (19). Finally, the solid curve is the true MSE achieved by AMP with $\widehat{x}_2$. The reader can see that the predicted MSE performance may help the user decide which denoiser to employ within AMP.

Whether the successful predictive ability in this example is a coincidence or perhaps connected to a theory of decoupling for non-i.i.d. signals remains to be seen. Nevertheless, this example indicates that the formulation (9) that relates the mismatched estimation in scalar channels to that in large linear systems, as well as its closed form approximations (17), (18), and (19) can already be applied to far more complicated systems for practical uses.

IV. Conclusion

We studied the excess mean square error (EMSE) above the minimum mean square error (MMSE) in large linear systems due to the mismatched prior for the posterior mean estimator (PME). In particular, we derived the relationship between the EMSE in large linear systems and that in scalar channels; three simple approximations to this relationship were provided. Numerical examples show that our approximations are accurate, indicating that they can be used to predict the EMSE in large linear systems due to mismatched estimation.

Appendix: Second Order Approximation of EMSE_l

Solving (16) for EMSE_l:

$$\text{EMSE}_l = \frac{\delta}{\beta}\left((\delta-\alpha) \pm \sqrt{(\delta-\alpha)^2 - 2\beta\text{EMSE}_s(\sigma_p^2)}\right) + o\left(\Delta^2\right).$$

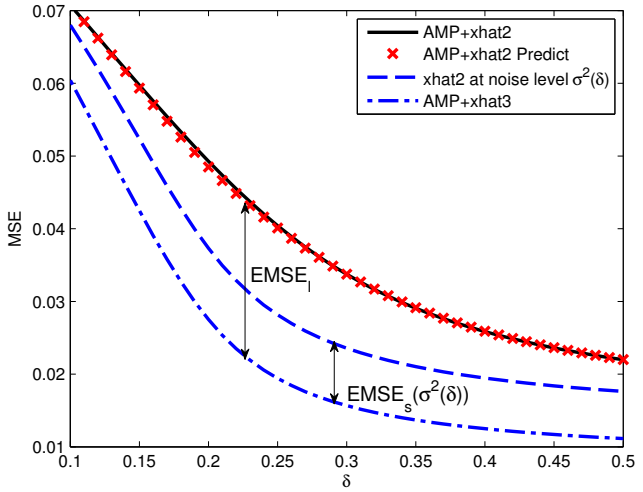

Figure 2: Prediction of MSE for Markov-constant input signal ($p(s_1|s_0) = 1/45$ and $p(s_0|s_1) = 0.2$) in large linear systems. In the legend, $\sigma^2(\delta)$ represents the noise variance that AMP with $\widehat{x}_3$ converges to at measurement rate δ. ($\sigma_z^2 = 0.1$.)

Recall that σ_p^2 is the noise variance in the decoupled scalar channel for the MMSE estimator, and $\Psi_p(\sigma_p^2)$ is the MMSE. That is, $\sigma_q^2 > \sigma_p^2$ and $\Psi_q(\sigma_p^2) > \Psi_p(\sigma_p^2)$, for all $q_X \neq p_X$. Hence, $\Delta = \sigma_q^2 - \sigma_p^2 > 0$ and $\text{EMSE}_s(\sigma_p^2) = \Psi_q(\sigma_p^2) - \Psi_p(\sigma_p^2) > 0$. Under the condition that $\Psi_q(\sigma^2)$ is locally concave around σ_p^2, the second derivative of $\Psi_q(\sigma^2)$ at σ_p^2 is negative. That is, $\beta < 0$. Therefore, $(\delta-\alpha)^2 - 2\beta\text{EMSE}_s(\sigma_p^2) \geq (\delta-\alpha)^2$, and

$$\begin{aligned}\text{EMSE}_l &= \frac{\delta}{\beta}\left((\delta-\alpha) - \sqrt{(\delta-\alpha)^2 - 2\beta\text{EMSE}_s(\sigma_p^2)}\right) \\ &\quad + o\left(\Delta^2\right) \\ &= \frac{\delta\cdot(\delta-\alpha)}{\beta}\left(1 - \sqrt{1 - \frac{2\beta\text{EMSE}_s(\sigma_p^2)}{(\delta-\alpha)^2}}\right) \\ &\quad + o\left(\Delta^2\right). \end{aligned} \tag{19}$$

A Taylor expansion of $\sqrt{1+x}$ yields

$$\sqrt{1+x} = 1 + \frac{1}{2}x - \frac{1}{8}x^2 + o(x^2).$$

Let $x = -\frac{2\beta\text{EMSE}_s(\sigma_p^2)}{(\delta-\alpha)^2}$, then

$$\begin{aligned}&\sqrt{1 - \frac{2\beta\text{EMSE}_s(\sigma_p^2)}{(\delta-\alpha)^2}} \\ &= 1 - \frac{\beta\text{EMSE}_s(\sigma_p^2)}{(\delta-\alpha)^2} - \frac{1}{2}\frac{\left(\beta\text{EMSE}_s(\sigma_p^2)\right)^2}{(\delta-\alpha)^4} \\ &\quad + o\left(\text{EMSE}_s(\sigma_p^2)^2\right). \end{aligned} \tag{20}$$

Plugging (20) into (19),

$$\begin{aligned}\text{EMSE}_l = \frac{\delta\text{EMSE}_s(\sigma_p^2)}{(\delta-\alpha)}&\left(1 + \frac{1}{2}\frac{\beta\text{EMSE}_s(\sigma_p^2))}{(\delta-\alpha)^2}\right) \\ &+ o\left(\text{EMSE}_s(\sigma_p^2)^2\right) + o\left(\Delta^2\right).\end{aligned}$$

Note that

$$\frac{\text{EMSE}_s(\sigma_p^2)^2}{\Delta^2} \leq \frac{\text{EMSE}_l^2}{\Delta^2} = \delta^2,$$

and so,

$$\lim_{\Delta\to 0}\frac{\text{EMSE}_s(\sigma_p^2)^2}{\Delta^2} \leq \delta^2 < \infty,$$

which implies that writing $o\left(\text{EMSE}_s(\sigma_p^2)^2\right) + o\left(\Delta^2\right)$ is equivalent to writing $o\left(\Delta^2\right)$.

We have proved the desired result (18).

Acknowledgements

We thank Dongning Guo, Neri Merhav, Nikhil Krishnan, and Jin Tan for informative discussions.

References

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

[2] T. Tanaka, "A statistical-mechanics approach to large-system analysis of CDMA multiuser detectors," *IEEE Trans. Inf. Theory*, vol. 48, no. 11, pp. 2888–2910, Nov. 2002.

[3] D. Guo and S. Verdú, "Randomly spread CDMA: Asymptotics via statistical physics," *IEEE Trans. Inf. Theory*, vol. 51, no. 6, pp. 1983–2010, June 2005.

[4] 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.

[5] D. Guo, D. Baron, and S. Shamai, "A single-letter characterization of optimal noisy compressed sensing," in *Proc. 47th Allerton Conf. Commun., Control, and Comput.*, Sept. 2009, pp. 52–59.

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

[7] S. Rangan, A. K. Fletcher, and V. K. Goyal, "Asymptotic analysis of MAP estimation via the replica method and applications to compressed sensing," *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1902–1923, Mar. 2012.

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

[9] 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, Aug. 2012.

[10] J. Zhu and D. Baron, "Performance regions in compressed sensing from noisy measurements," in *Proc. 2013 Conf. Inference Sci. Syst. (CISS)*, Baltimore, MD, Mar. 2013.

[11] M. Bayati and A. Montanari, "The dynamics of message passing on dense graphs, with applications to compressed sensing," *IEEE Trans. Inf. Theory*, vol. 57, no. 2, pp. 764–785, Feb. 2011.

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