

The Relationship Between Causal and Noncausal Mismatched Estimation in Continuous-Time AWGN Channels

Tsachy Weissman, *Senior Member, IEEE*

Abstract—A continuous-time finite-power process with distribution P is observed through an AWGN channel, at a given signal-to-noise ratio (SNR), and is estimated by an estimator that would have minimized the mean-square error if the process had distribution Q . We show that the causal filtering mean-square error (MSE) achieved at SNR level snr is equal to the average value of the noncausal (smoothing) MSE achieved with a channel whose SNR is chosen uniformly distributed between 0 and snr . Emerging as the bridge for equating these two quantities are mutual information and relative entropy. Our result generalizes that of Guo, Shamai, and Verdú (2005) from the nonmismatched case, where $P = Q$, to general P and Q . Among our intermediate results is an extension of Duncan's theorem, that relates mutual information and causal MMSE, to the case of mismatched estimation. Some further extensions and implications are discussed. Key to our findings is the recent result of Verdú on mismatched estimation and relative entropy.

Index Terms—AWGN channels, Brownian motion, continuous-time, I-MMSE formula, minimax estimation, minimum mean-square error estimation, mismatched estimation, mutual information, nonlinear filtering, relative entropy, Shannon theory.

I. INTRODUCTION AND MAIN RESULT

A. The Bottom Line

IN [16], a remarkable relationship between the MMSEs in causal (filtering) and noncausal (smoothing) estimation of an arbitrarily distributed signal corrupted by Additive White Gaussian Noise (AWGN) was discovered: the filtering MMSE at SNR level snr is equal to the mean value of the smoothing MMSE with SNR uniformly distributed between 0 and snr . In the present paper, we show that this equality holds also in the mismatched case, where the filters are optimized for an underlying signal distribution that differs from the true one. Bridging the two sides of this equality, up to a multiplicative constant inversely proportional to snr , is the sum of the input-output mutual information and the relative entropy between the true and mismatched output distributions. This relative entropy thus quantifies the penalty due to mismatch in continuous-time estimation over AWGN channels.

Manuscript received August 10, 2009; revised May 10, 2010. Date of current version August 18, 2010. This work was supported in part by the National Science Foundation under grants CCF-0546535 and CCF-0729119.

The author is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA, on leave from the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel (e-mail: tsachy@stanford.edu).

Communicated by H. Yamamoto, Associate Editor for Shannon Theory.

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TIT.2010.2054430

B. A Bit of Literature Context

The line of work on which we draw—developing and exploiting connections between information and estimation theoretic quantities in the presence of added Gaussian random variables—dates back to “de Bruijn’s identity” in Stam’s paper [35], which exhibits the derivative of a differential entropy as a Fisher information, the mean square score. The integral form of such a relationship, along with a demonstration of the required continuity properties at the zero Gaussian and fully Gaussian limits, was presented by Barron in [2] which also exhibits the relative entropy from a normal by integration of the mean squared difference between the score and a normal score. Among other recent uses, such results were key for the resolution of Shannon’s conjecture on the monotonicity in the CLT [1], cf. also [36] and [28]. In Gaussian noise settings, the representation of the minimal mean square error via score functions is a central part of statistical decision theory, cf. [24], [28] and references therein.

In [16], Guo, Shamai and Verdú discovered the “I-MMSE” relationship and established its equivalence to the de Bruijn identity. In turn, the “I-MMSE” relationship allowed them to discover the remarkable—purely estimation theoretic—relationship between the causal (filtering) and noncausal (smoothing) MMSEs, by combining the continuous-time version of their “I-MMSE” relationship with Duncan’s theorem [9]. It is this relationship that we extend in the present work to the case of mismatched estimation.

Our results rely heavily on the recent [37], where an intimate connection between relative entropy and mismatched estimation in additive Gaussian noise was revealed. Our main result is established by both extending the setting and results of [37] to continuous-time estimation, and extending the main result of [9] to the case of mismatch. One of our proofs of the latter result, the one relying on the Girsanov theorem, provides an alternative proof of the main result of [37], when specializing our result to DC processes.

Our work thus joins [16], and the recent body of work that it has spawned (cf., e.g., [3], [15], [17], [26], [28], [29], [31], [32], [36], [38], [40], and references therein), in illuminating and exploiting intimate connections between information and estimation theory. Whether and what counterparts of our results hold in the continuous-time Poisson channel setting of [17], and the generalized settings of [10], [11], remains for future work.

C. Problem Setting, Motivation, and Main Result

Let $X_0^T = \{X_t\}_{0 \leq t \leq T}$ be a process of finite average power $\int_0^T E[X_t^2] dt < \infty$, distributed according to P , independent of the standard Brownian motion $\{W_t\}$, and let Y_0^T be its AWGN-

corrupted observation process at signal-to-noise ratio (SNR) γ , i.e.,

$$dY_t = \sqrt{\gamma}X_t dt + dW_t. \quad (1)$$

Define the *mismatched* causal and noncausal mean-square error, at any time $t \in [0, T]$, by

$$\text{cmse}_Q(t, \gamma) = E \left[(X_t - E_Q[X_t | Y_0^t; \gamma])^2 \right] \quad (2)$$

and

$$\text{mse}_Q(t, \gamma) = E \left[(X_t - E_Q[X_t | Y_0^T; \gamma])^2 \right] \quad (3)$$

where the subscript Q denotes expectation assuming that X_0^T is distributed according to Q (and we append γ to the conditioning argument when we want to make the dependence on it explicit). Here and throughout we follow the convention of [37] that an unsubscripted expectation is with respect to X_0^T distributed according to P , independent of the standard Brownian motion $\{W_t\}$, when Y_0^T is given by (1). Denote the mismatched causal and noncausal mean-square error by

$$\text{cmse}_Q(\gamma) = \int_0^T \text{cmse}_Q(t, \gamma) dt \quad (4)$$

and

$$\text{mse}_Q(\gamma) = \int_0^T \text{mse}_Q(t, \gamma) dt \quad (5)$$

respectively. In other words, $\text{cmse}_Q(\gamma)$ and $\text{mse}_Q(\gamma)$ are the cumulative mean-square error incurred by, respectively, the causal and noncausal nonlinear filters that are optimal for Q , when the true distribution is P . In practice, filters are often designed under Q that differs from P because the latter is not exactly known and, even when it is, a filter under Q may be simpler to implement.¹ It is, therefore, of interest to study $\text{cmse}_Q(\gamma)$ and $\text{mse}_Q(\gamma)$.

Similarly as in [16], [37], we use $\text{cmmse}(\gamma)$ and $\text{mmse}(\gamma)$ to denote the minimum mean-square error (MMSE) in the corresponding problems, thus

$$\begin{aligned} \text{cmmse}(\gamma) &= \text{cmse}_P(\gamma), \quad \text{and} \\ \text{mmse}(\gamma) &= \text{mse}_P(\gamma). \end{aligned} \quad (6)$$

Recall Theorem 8 of [16], which states that the causal and noncausal MMSEs satisfy, for $\text{snr} > 0$

$$\text{cmmse}(\text{snr}) = \frac{1}{\text{snr}} \int_0^{\text{snr}} \text{mmse}(\gamma) d\gamma. \quad (7)$$

Remarkably, this fundamental relationship holds regardless of the distribution P .

What happens in the case of mismatched estimation? To note one mismatched estimation scenario to which the relationship in (7) carries over immediately, let P_G denote the Gaussian distribution with the same first and second order moments as P . The relationship

$$\text{cmse}_{P_G}(\text{snr}) = \frac{1}{\text{snr}} \int_0^{\text{snr}} \text{mse}_{P_G}(\gamma) d\gamma \quad (8)$$

¹E.g., when Q is Gaussian the associated filter is linear.

is immediate upon noting that $\text{cmse}_{P_G}(\text{snr})$ and $\text{mse}_{P_G}(\gamma)$ coincide with $\text{cmmse}(\text{snr})$ and $\text{mmse}(\gamma)$ when in the latter two quantities the true distribution is taken as P_G , and applying (7). This observation was implicit in [3].

To note another mismatched estimation scenario where (7) carries over, let Q denote the distribution governing the process formed by concatenating $X_0^{T/2}$ with $\{X_t\}_{T/2 < t \leq T}$, both governed by P , but independent of each other. We then obviously get the relationship

$$\text{cmse}_Q(\text{snr}) = \frac{1}{\text{snr}} \int_0^{\text{snr}} \text{mse}_Q(\gamma) d\gamma \quad (9)$$

by applying (7) separately on the time intervals $[0, T/2]$ and $(T/2, T]$.

To note yet another mismatched estimation scenario where (7) carries over, we recall a fundamental relationship from [37]. Consider a scalar observation G given by

$$G = \sqrt{\gamma}A + B \quad (10)$$

where $B \sim \mathcal{N}(0, 1)$ is independent of A . Suppose $A \sim P$ and let $\text{mse}_Q^{\text{scalar}}(\gamma)$ denote the mean square error in estimating A on the basis of G , using the estimator that would have been optimal if $A \sim Q$. Let $\text{mmse}^{\text{scalar}}(\gamma)$ stand for $\text{mse}_P^{\text{scalar}}(\gamma)$. Assuming that both P and Q have finite second moments, (14) of [37] presents the following relationship, which is key to our results:

$$\begin{aligned} D(P * \mathcal{N}(0, \sigma^2) \| Q * \mathcal{N}(0, \sigma^2)) \\ = \frac{1}{2} \int_0^{1/\sigma^2} [\text{mse}_Q^{\text{scalar}}(\gamma) - \text{mmse}^{\text{scalar}}(\gamma)] d\gamma. \end{aligned} \quad (11)$$

Consider now the case where X_t is a DC signal, i.e.,

$$X_t \equiv X, \quad 0 \leq t \leq T \quad (12)$$

for a random variable $X \sim P$, whereas the mismatched filter assumes $X \sim Q$. Letting $\text{mse}_Q^{\text{DC}}(\gamma)$ and $\text{mmse}^{\text{DC}}(\gamma)$ denote $\text{mse}_Q(\gamma)$ and $\text{mmse}(\gamma)$ of (5) and (6) for the case where X_t is the DC signal of (12), since the duration of the observation interval T is proportional to γ in the setting of (10), we have the obvious relationship

$$\begin{aligned} \text{mse}_Q^{\text{DC}}(\gamma) &= T \cdot \text{mse}_Q^{\text{scalar}}(\gamma T) \quad \text{and a fortiori} \\ \text{mmse}^{\text{DC}}(\gamma) &= T \cdot \text{mmse}^{\text{scalar}}(\gamma T). \end{aligned} \quad (13)$$

Consequently

$$\begin{aligned} \int_0^{\text{snr}} [\text{mse}_Q^{\text{DC}}(\gamma) - \text{mmse}^{\text{DC}}(\gamma)] d\gamma \\ = T \int_0^{\text{snr}} [\text{mse}_Q^{\text{scalar}}(\gamma T) - \text{mmse}^{\text{scalar}}(\gamma T)] d\gamma \end{aligned} \quad (14)$$

$$= \int_0^{\text{snr}T} [\text{mse}_Q^{\text{scalar}}(\gamma') - \text{mmse}^{\text{scalar}}(\gamma')] d\gamma' \quad (15)$$

$$= 2D \left(P * \mathcal{N} \left(0, \frac{1}{\text{snr}T} \right) \parallel Q * \mathcal{N} \left(0, \frac{1}{\text{snr}T} \right) \right) \quad (16)$$

where the first equality follows from (13) and the third equality is an application of (11). On the other hand, letting $\text{cmse}_Q^{\text{DC}}(\gamma)$ and $\text{cmmse}^{\text{DC}}(\gamma)$ denote $\text{cmse}_Q(\gamma)$ and $\text{cmmse}(\gamma)$ of (4) and

(6) for the case where X_t is the DC signal of (12); see (17)–(21), shown at the bottom of the page, where the last equality is another application of (11). Putting (16) and (21) together yields

$$\begin{aligned} \text{cmse}_Q^{\text{DC}}(\text{snr}) - \text{cmmse}^{\text{DC}}(\text{snr}) \\ = \frac{1}{\text{snr}} \int_0^{\text{snr}} [\text{mse}_Q^{\text{DC}}(\gamma) - \text{mmse}^{\text{DC}}(\gamma)] d\gamma \end{aligned} \quad (22)$$

implying in turn

$$\text{cmse}_Q^{\text{DC}}(\text{snr}) = \frac{1}{\text{snr}} \int_0^{\text{snr}} \text{mse}_Q^{\text{DC}}(\gamma) d\gamma \quad (23)$$

upon noting that

$$\text{cmmse}^{\text{DC}}(\text{snr}) = \frac{1}{\text{snr}} \int_0^{\text{snr}} \text{mmse}^{\text{DC}}(\gamma) d\gamma \quad (24)$$

where (24) is nothing but (7) applied to this special case of a DC process.

To recap, we have found three mismatched estimation scenarios where the relationship

$$\text{cmse}_Q(\text{snr}) = \frac{1}{\text{snr}} \int_0^{\text{snr}} \text{mse}_Q(\gamma) d\gamma \quad (25)$$

holds: In the first case, (8), $Q = P_G$. In the second case, (9), Q governs the process formed by concatenating independent copies of $X_0^{T/2}$ and $\{X_t\}_{T/2 < t \leq T}$, each $\sim P$. In the third case, (23), X_0^T is a DC process under both P and Q . Is this relationship a peculiarity of these special cases?

Our main result is that the relationship (25) holds for *arbitrary* P and Q . Further, mutual information and relative entropy emerge as the bridge for equating the two sides of (25). Specifically, let $I(\gamma)$ denote the mutual information between X_0^T and Y_0^T under the channel in (1).² Let further $P_{Y_0^T}$ and $Q_{Y_0^T}$ denote the distribution of Y_0^T [the output of the channel described in (1)] when X_0^T is distributed according to P and Q , respectively, and denote the relative entropy (divergence) between $P_{Y_0^T}$ and $Q_{Y_0^T}$ by

$$D_\gamma \left(P_{Y_0^T} \parallel Q_{Y_0^T} \right) \quad (26)$$

²Similarly as with expectation, in the notation $I(\gamma)$ we suppress the dependence on the (true) channel input process P .

where the subscript γ is to make the dependence on the SNR value explicit.³ Our main result can now be stated as follows:

Theorem 1: For any pair of distributions P and Q of finite average power, and for any $\text{snr} > 0$

$$\begin{aligned} \text{cmse}_Q(\text{snr}) &= \frac{1}{\text{snr}} \int_0^{\text{snr}} \text{mse}_Q(\gamma) d\gamma \\ &= \frac{2}{\text{snr}} \left[I(\text{snr}) + D_{\text{snr}} \left(P_{Y_0^T} \parallel Q_{Y_0^T} \right) \right]. \end{aligned} \quad (27)$$

It is evident from our proof of Theorem 2 that it carries over verbatim to the case where $X_t \in \mathbb{R}^d$, W_t is a d -dimensional standard Brownian motion, and the square of the error (in (2), (3), etc.) is replaced by the square of the Euclidean norm of the error.

D. Remaining Content

Some of the interpretations and implications of Theorem 1 are discussed in Section II. In Section III, we consider an example involving a Gaussian DC process, for which we compute the quantities that appear in Theorem 1, verify (27) directly, and present plots illustrating how some of the generic observations made in Section II play themselves out in a concrete case. Section IV is dedicated to a proof of our main result and to a discussion of the intuition behind it, alternative approaches to the proof, and how our results carry over to accommodate the presence of feedback.

II. DISCUSSION, INTERPRETATION, AND IMPLICATIONS

A. Relation to Known Results and a Pictorial Representation

Theorem 1 can be viewed as an extension, from the case of optimal to that of mismatched estimation, of the relationship

$$\begin{aligned} \frac{1}{\text{snr}} \int_0^{\text{snr}} \text{mmse}(\gamma) d\gamma &= \text{cmmse}(\text{snr}) \\ &= \frac{2}{\text{snr}} I(\text{snr}) \end{aligned} \quad (28)$$

³ $I(\gamma)$ and $D_\gamma \left(P_{Y_0^T} \parallel Q_{Y_0^T} \right)$, induced by measures on function spaces pertaining to the continuous-time AWGN channel, are well defined objects, as made explicit in Section IV-A.

$$\text{cmse}_Q^{\text{DC}}(\gamma) - \text{cmmse}^{\text{DC}}(\gamma) = \int_0^T E \left[(X - E_Q[X|Y_0^t; \gamma])^2 \right] dt - \int_0^T E \left[(X - E[X|Y_0^t; \gamma])^2 \right] dt \quad (17)$$

$$= \int_0^T \left\{ E \left[(X - E_Q[X|Y_0^t; \gamma])^2 \right] - E \left[(X - E[X|Y_0^t; \gamma])^2 \right] \right\} dt \quad (18)$$

$$= \int_0^T [\text{mse}_Q^{\text{scalar}}(\gamma t) - \text{mmse}^{\text{scalar}}(\gamma t)] dt \quad (19)$$

$$= \frac{1}{\gamma} \int_0^{\gamma T} [\text{mse}_Q^{\text{scalar}}(\alpha) - \text{mmse}^{\text{scalar}}(\alpha)] d\alpha \quad (20)$$

$$= \frac{2}{\gamma} D \left(P * \mathcal{N} \left(0, \frac{1}{\gamma T} \right) \parallel Q * \mathcal{N} \left(0, \frac{1}{\gamma T} \right) \right) \quad (21)$$

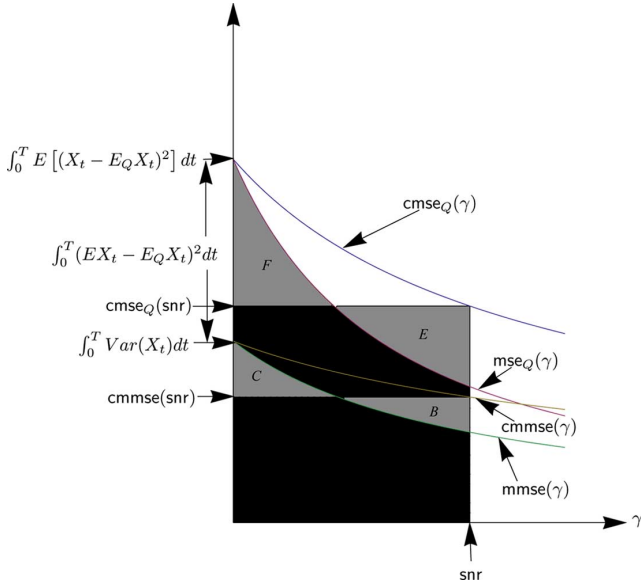


Fig. 1. Picture implied by Theorem 1.

where the first equality in (28) is Theorem 8 of [16] and the second one is Duncan's theorem [9, Theorem 3]. We illustrate the picture implied by Theorem 1, and how it relates to the one implied by (28), in Fig. 1. That the areas of Region B and Region C (in the figure) are equal follows from Theorem 8 of [16]. Duncan's theorem [9] implies that the area of the rectangle consisting of Region B and the lower black region is equal to $2 \cdot I(\text{snr})$. Theorem 1 implies that the areas of Region E and Region F are equal, and that the area of the rectangle consisting of Region C, Region E, and the black region separating them is equal to $2 \cdot D_{\text{snr}}(P_{Y_0^T} \| Q_{Y_0^T})$. Consistent with an observation made in Section II-D below, the ratio of the slopes of the non-causal and causal quantities at 0, in both the mismatched and the nonmismatched case, is seen to be 2. Notwithstanding the curves plotted in that figure, $\text{cmse}_Q(\gamma)$ and $\text{mse}_Q(\gamma)$ need not be monotone in γ in general, nor need $\text{cmse}_Q(\gamma) \geq \text{mse}_Q(\gamma)$ hold. Confer Section II-B for discussion, and Figs. 5 and 6 for examples.

Theorem 1 is an immediate consequence of (28) and the following:

Theorem 2: For any pair of distributions P and Q of finite average power, and for any $\text{snr} > 0$

$$1. \int_0^{\text{snr}} [\text{mse}_Q(\gamma) - \text{mmse}(\gamma)] d\gamma = 2D_{\text{snr}}(P_{Y_0^T} \| Q_{Y_0^T}) \quad (29)$$

$$2. \text{cmse}_Q(\text{snr}) - \text{cmmse}(\text{snr}) = \frac{2}{\text{snr}} D_{\text{snr}}(P_{Y_0^T} \| Q_{Y_0^T}). \quad (30)$$

It can be noted that Theorem 2, when specialized to the case $Q = P_G$ discussed in the Introduction, recovers the main result of [3]. Section IV is dedicated to the proof of Theorem 2.

To be explicit about how our results extend those of [9] and [16], consider the equality between the leftmost and rightmost terms in (27)

$$\text{cmse}_Q(\text{snr}) = \frac{2}{\text{snr}} \left[I(\text{snr}) + D_{\text{snr}}(P_{Y_0^T} \| Q_{Y_0^T}) \right]. \quad (31)$$

Equality (31) extends Duncan's theorem [9] to the case of mismatched estimation. It is a direct consequence of combining that original theorem with (30). On the other hand, multiplying by snr and differentiating, the second equality in (27) yields

$$\frac{1}{2} \text{mse}_Q(\text{snr}) = \frac{d}{d\text{snr}} \left[I(\text{snr}) + D_{\text{snr}}(P_{Y_0^T} \| Q_{Y_0^T}) \right] \quad (32)$$

extending the continuous-time I-MMSE relationship, [16, Theorem 6], to the mismatched case.

Finally, we recall from the Introduction, (20), that the left hand side of (30) coincides with

$$\frac{1}{\text{snr}} \int_0^{\text{snr}} [\text{mse}_Q^{\text{scalar}}(\alpha) - \text{mmse}^{\text{scalar}}(\alpha)] d\alpha \quad (33)$$

in the special case where X_0^T is a DC process (with amplitude distributed under P and Q , respectively) and $T = 1$. On the other hand, in this special case, the right hand side of (30) is nothing but

$$\frac{2}{\text{snr}} D(P * \mathcal{N}(0, 1/\text{snr}) \| Q * \mathcal{N}(0, 1/\text{snr})). \quad (34)$$

Evidently (30) yields (11) when applied in the special case of a DC process (taking $T = 1$ and $\text{snr} = 1/\sigma^2$). Our first proof of (30), given in Section IV-C, relies on (11) as a building block. In Section IV-D, we give a direct proof of (30) that uses the Girsanov theorem (cf., e.g., Section 3.5 of [21]) and does not rely on (11). Thus, when specialized to the case of a DC process, this latter proof yields a new proof of (11) and, in particular, of the main result of [37] (Theorem 1 therein).

B. Nonmonotonicity of $\text{cmse}_Q(\gamma)$ and $\text{mse}_Q(\gamma)$

The obvious low and high SNR behavior of $\text{cmse}_Q(\gamma)$ and $\text{mse}_Q(\gamma)$ is given by

$$\begin{aligned} \lim_{\gamma \downarrow 0} \text{cmse}_Q(\gamma) &= \lim_{\gamma \downarrow 0} \text{mse}_Q(\gamma) = \text{cmse}_Q(0) \\ &= \text{mse}_Q(0) \\ &= \int_0^T E[(X_t - E_Q X_t)^2] dt \end{aligned} \quad (35)$$

and, under benign conditions on Q ,⁴

$$\lim_{\gamma \rightarrow \infty} \text{cmse}_Q(\gamma) = \lim_{\gamma \rightarrow \infty} \text{mse}_Q(\gamma) = 0. \quad (36)$$

In the nonmismatched case, two things are clear from a data processing consideration:

- “The less noise the better”: both $\text{cmmse}(\gamma)$ and $\text{mmse}(\gamma)$ decrease *monotonically* to 0 with increasing γ .
- “More observations cannot hurt”: $\text{cmmse}(\gamma) \geq \text{mmse}(\gamma)$ for all $\gamma \geq 0$.

In the mismatched case, neither of these properties need hold and, in fact, Theorem 1 implies that the two are intimately related: Differentiating the relationship

$$\text{snr} \cdot \text{cmse}_Q(\text{snr}) = \int_0^{\text{snr}} \text{mse}_Q(\gamma) d\gamma \quad (37)$$

⁴E.g., that X_t , under Q , is supported on the whole real line for all $t \in [0, T]$ is readily seen to be a sufficient condition.

we obtain

$$\text{cmse}_Q(\text{snr}) - \text{mse}_Q(\text{snr}) = -\text{snr} \cdot \frac{d}{d\text{snr}} \text{cmse}_Q(\text{snr}) \quad (38)$$

i.e., the price of causality is $-\text{snr} \cdot \frac{d}{d\text{snr}} \text{cmse}_Q(\text{snr})$ (assuming this derivative exists). In the mismatched case this price can be positive or negative: (38) implies that $\text{mse}_Q(\text{snr}) = \text{cmse}_Q(\text{snr})$ if and only if $\frac{d}{d\text{snr}} \text{cmse}_Q(\text{snr}) = 0$, and that [see (39), shown at the bottom of the page]. In other words: an increase in SNR deteriorates the mismatched causal mean square error performance if and only if the latter is better than the noncausal mean square error performance.

When might an increase in SNR deteriorate performance or, conversely, when might more noise help? Qualitatively, when Q is very mismatched, more noise might serve to weaken its (erroneous) conviction on the underlying signal and, hence, to “tone down” its output in a desirable direction. This phenomenon is loosely referred to in some parts of the literature as “stochastic resonance”, cf., e.g., [13].

For an extreme example, consider the case where X_0^T is deterministically the 0 signal, i.e., $X_t \equiv 0$ for all ω and t . Let Q be an arbitrary, nondegenerate, zero mean (i.e., $E_Q[X_t] = 0 \forall t$) distribution. Then $\text{cmse}_Q(\gamma)$ and $\text{mse}_Q(\gamma)$ are both positive for all $\gamma > 0$, while clearly $\lim_{\gamma \downarrow 0} \text{cmse}_Q(\gamma) = \lim_{\gamma \downarrow 0} \text{mse}_Q(\gamma) = 0$. Figs. 5 and 6 exhibit less extreme examples.

C. Causal Versus Anticausal Mismatched MSE

That the causal and anticausal MMSEs are equal is an immediate consequence of Duncan’s theorem and the invariance of $I(\text{snr})$ to the direction of the flow of time. This equality between the causal and noncausal MMSE is remarkable considering that P is arbitrary, with no stationarity, reversibility, or any other type of restriction assumed. Even more remarkable is that this equality carries over to the case of mismatched estimation, regardless of any assumptions on P , on Q , or on the relationship between them. Indeed, let $\text{acmse}_Q(\gamma)$ denote the anticausal mismatched mse, i.e., let

$$\text{acmse}_Q(t, \gamma) = E \left[(X_t - E_Q[X_t | \{Y_s - Y_t\}_{t \leq s \leq T}; \gamma])^2 \right] \quad (40)$$

and

$$\text{acmse}_Q(\gamma) = \int_0^T \text{acmse}_Q(t, \gamma) dt. \quad (41)$$

The obvious invariance of the middle and the right terms in (27) to the direction of the flow of time implies that

$$\text{acmse}_Q(\text{snr}) = \text{cmse}_Q(\text{snr}). \quad (42)$$

Indeed, the middle and the right terms in (27) are invariant not only to the direction of the flow of time but to any “reordering”

of time. More precisely, if $\phi : [0, T] \rightarrow [0, T]$ is a one-to-one, onto, Lebesgue-measure-preserving transformation,⁵ let P_ϕ and Q_ϕ denote the distributions of the process $\{X_{\phi(t)}\}_{0 \leq t \leq T}$ when X_0^T is $\sim P$ and $\sim Q$, respectively. The middle and the right terms in (27), and consequently also the left term, remain unchanged when P and Q are replaced by P_ϕ and Q_ϕ . The implications that this fact might have on the question of the sensitivity of mismatched filtering performance to the ordering of the data are analogous to those developed in [5, Subsection III-B] for the nonmismatched case.

D. Low SNR Behavior of $\text{cmse}_Q(\text{snr})$ and $\text{mse}_Q(\text{snr})$: A Factor of 2 Relationship

As noted in Fig. 1, in the low SNR regime we have

$$\begin{aligned} \lim_{\text{snr} \rightarrow 0} \text{mse}_Q(\text{snr}) &= \lim_{\text{snr} \rightarrow 0} \text{cmse}_Q(\text{snr}) \\ &= \int_0^T E[(X_t - E_Q X_t)^2] dt. \end{aligned} \quad (43)$$

We now argue that Theorem 2 implies that, under sufficient regularity, the following relationship holds between the rates of convergence associated with the two limits in (43):

$$\lim_{\text{snr} \rightarrow 0} \frac{\int_0^T E[(X_t - E_Q X_t)^2] dt - \text{mse}_Q(\text{snr})}{\int_0^T E[(X_t - E_Q X_t)^2] dt - \text{cmse}_Q(\text{snr})} = 2. \quad (44)$$

In other words, the noncausal error approaches its low SNR limit twice as rapidly as the causal error. The relationship (44) was noted in Corollary 4.2 of [29] for Gaussian processes with non-mismatched filters, and in [16] for nonmismatched filters and general processes. To see why it holds, consider

$$\begin{aligned} \lim_{\text{snr} \rightarrow 0} \frac{\int_0^T E[(X_t - E_Q X_t)^2] dt - \text{mse}_Q(\text{snr})}{\int_0^T E[(X_t - E_Q X_t)^2] dt - \text{cmse}_Q(\text{snr})} \\ \stackrel{(a)}{=} \lim_{\text{snr} \rightarrow 0} \frac{\text{cmse}_Q(0) - \text{cmse}_Q(\text{snr}) - \text{snr} \cdot \text{cmse}'_Q(\text{snr})}{\text{cmse}_Q(0) - \text{cmse}_Q(\text{snr})} \end{aligned} \quad (45)$$

$$= 1 + \lim_{\text{snr} \rightarrow 0} \frac{\text{snr} \cdot \text{cmse}'_Q(\text{snr})}{\text{cmse}_Q(\text{snr}) - \text{cmse}_Q(0)} \quad (46)$$

$$\stackrel{(b)}{=} 1 + \text{cmse}'_Q(0) \lim_{\text{snr} \rightarrow 0} \frac{\text{snr}}{\text{cmse}_Q(\text{snr}) - \text{cmse}_Q(0)} \quad (47)$$

$$\stackrel{(c)}{=} 1 + \text{cmse}'_Q(0) \frac{1}{\text{cmse}'_Q(0)} \quad (48)$$

$$= 2 \quad (49)$$

where (a) follows from (38). Note we have assumed here differentiability of $\text{cmse}_Q(\text{snr})$ (for (a) to hold), its being *continuously* differentiable at 0 (for (b) to hold), and that $\text{cmse}'_Q(0) \neq 0$

⁵Throughout the paper, “transformations”, “functions”, “mappings”, and random objects, if not explicitly mentioned, are to be understood as measurable under a measurable space that should be clear from the context.

$$\text{mse}_Q(\text{snr}) > \text{cmse}_Q(\text{snr}) \quad \text{if and only if} \quad \frac{d}{d\text{snr}} \text{cmse}_Q(\text{snr}) > 0 \quad (39)$$

(for validity of (c)). Note also that the numerator and the denominator of the expression on the left side of (44) are either both positive or both negative in a neighborhood of $\text{snr} = 0$, according to the sign of $\text{cmse}'_Q(0)$, as is consistent with the findings of Section II-B. Fig. 1 exhibits a case where $\text{cmse}'_Q(0) < 0$ while Figs. 5 and 6 exhibit one with $\text{cmse}'_Q(0) > 0$. Regardless of the sign, the factor of 2 relation (43) is apparent in all said figures.

E. High SNR Behavior of $D_{\text{snr}}(P_{Y_0^T} \| Q_{Y_0^T})$

In the high SNR limit, one has

$$\lim_{\text{snr} \rightarrow \infty} D_{\text{snr}}(P_{Y_0^T} \| Q_{Y_0^T}) = D(P \| Q) \quad (50)$$

regardless of the finiteness of the right side of the equality. On the other hand, as in (36),

$$\lim_{\gamma \rightarrow \infty} \text{cmse}_Q(\gamma) = \lim_{\gamma \rightarrow \infty} \text{mse}_Q(\gamma) = 0 \quad (51)$$

under benign conditions on Q , e.g., that X_t , under Q , is supported on the whole real line for all $t \in [0, T]$ suffices. Combining (50) with (30) implies that, when (51) holds

$$\lim_{\text{snr} \rightarrow \infty} \frac{2}{\text{snr}} D_{\text{snr}}(P_{Y_0^T} \| Q_{Y_0^T}) = 0. \quad (52)$$

The conclusion is that, assuming Q sufficiently regular to imply (51), $D_{\text{snr}}(P_{Y_0^T} \| Q_{Y_0^T})$ exhibits one of the following possible behaviors in the high SNR regime:

1. $D_{\text{snr}}(P_{Y_0^T} \| Q_{Y_0^T}) \equiv 0$ for all $\text{snr} > 0$. This can happen if and only if $D(P \| Q) = 0$, i.e., the nonmismatched setting.
2. $D_{\text{snr}}(P_{Y_0^T} \| Q_{Y_0^T}) = \Theta(1)$ which, by (50) and the previous item, can happen if and only if $0 < D(P \| Q) < \infty$.
3. $\lim_{\text{snr} \rightarrow \infty} D_{\text{snr}}(P_{Y_0^T} \| Q_{Y_0^T}) = \infty$ but $D_{\text{snr}}(P_{Y_0^T} \| Q_{Y_0^T}) = o(\text{snr})$ which, by (50) and (52), can happen if and only if $D(P \| Q) = \infty$, i.e., when $D(P \| Q) = \infty$, $D_{\text{snr}}(P_{Y_0^T} \| Q_{Y_0^T})$ increases without bound with increasing SNR, but sub-linearly.

F. Special Cases

Mismatched Linear Estimation: Define $\text{cmse}_Q^{\text{lin}}(\gamma)$ and $\text{mse}_Q^{\text{lin}}(\gamma)$ analogously to $\text{cmse}_Q(\gamma)$ and $\text{mse}_Q(\gamma)$, for the case where the filter used is, respectively, the causal and noncausal Wiener filter designed for Q , i.e., the filter that would be optimal among linear filters if $X_0^T \sim Q$. Since obviously $\text{cmse}_Q^{\text{lin}}(\gamma) = \text{cmse}_{Q_G}(\gamma)$ and $\text{mse}_Q^{\text{lin}}(\gamma) = \text{mse}_{Q_G}(\gamma)$, Theorem 1 implies that

$$\begin{aligned} \text{cmse}_Q^{\text{lin}}(\text{snr}) &= \frac{1}{\text{snr}} \int_0^{\text{snr}} \text{mse}_Q^{\text{lin}}(\gamma) d\gamma \\ &= \frac{2}{\text{snr}} \left[I(\text{snr}) + D_{\text{snr}}(P_{Y_0^T} \| Q_{Y_0^T}^G) \right] \end{aligned} \quad (53)$$

where $Q_{Y_0^T}^G$ denotes the distribution of the Gaussian process whose first and second order moment statistics coincide with those of $Q_{Y_0^T}$.

The Price of Non-Gaussianity: Apply Theorem 2 for arbitrary P and $Q = P_G$ to obtain

$$\begin{aligned} \text{cmse}_{P_G}(\text{snr}) &= \frac{1}{\text{snr}} \int_0^{\text{snr}} \text{mse}_{P_G}(\gamma) d\gamma \\ &= \frac{2}{\text{snr}} \left[I(\text{snr}) + D_{\text{snr}}(P_{Y_0^T} \| P_{Y_0^T}^G) \right]. \end{aligned} \quad (54)$$

Note that (54) is nothing but (53) specialized to $Q = P$. On the other hand, by applying (28) with P_G playing the role of P ,⁶ we get

$$\begin{aligned} \text{cmse}_{P_G}(\text{snr}) &= \frac{1}{\text{snr}} \int_0^{\text{snr}} \text{mse}_{P_G}(\gamma) d\gamma \\ &= \frac{2}{\text{snr}} I_{P_G}(\text{snr}) \end{aligned} \quad (55)$$

where by $I_{P_G}(\text{snr})$ we denote the mutual information when the input (noise-free) process $\sim P_G$. In particular, putting (54) and (55) together yields the relationship

$$I(\text{snr}) + D_{\text{snr}}(P_{Y_0^T} \| P_{Y_0^T}^G) = I_{P_G}(\text{snr}) \quad (56)$$

confirming that a Gaussian process maximizes the mutual information between X_0^T and Y_0^T among all processes of a given correlation function, and quantifying the price of a non-Gaussian input distribution.

Degenerate Q : Let s_t , $0 \leq t \leq T$, be a deterministic signal. Applying Theorem 2, with Q assigning probability one to the noise-free signal being s_0^T , gives

$$\begin{aligned} \int_0^T E[(X_t - s_t)^2] dt \\ = \frac{2}{\text{snr}} \left[I(\text{snr}) + D_{\text{snr}}(P_{Y_0^T} \| Q_{Y_0^T}) \right] \end{aligned} \quad (57)$$

or, equivalently

$$D_{\text{snr}}(P_{Y_0^T} \| Q_{Y_0^T}) = \frac{\text{snr}}{2} \int_0^T E[(X_t - s_t)^2] dt - I(\text{snr}) \quad (58)$$

where in this case $Q_{Y_0^T}$ is the law of Brownian motion with drift function s_0^T . Specializing even further by taking $s_t \equiv 0$ we obtain

$$\int_0^T E[(X_t)^2] dt = \frac{2}{\text{snr}} \left[I(\text{snr}) + D_{\text{snr}}(P_{Y_0^T} \| W_0^T) \right] \quad (59)$$

where W_0^T is the standard Brownian motion. The relationship (59) is not really new: it follows directly from Duncan's theorem, the relationship

$$D_{\text{snr}}(P_{Y_0^T} \| W_0^T) = \frac{\text{snr}}{2} \int_0^T E[E[X_t | Y_0^t]^2] dt \quad (60)$$

(cf., e.g., [20]), and the orthogonality principle. Fig. 2 illustrates the picture implied by (59).

⁶To see why the left and middle terms in (28) coincide with those in (55) note that in both cases these quantities involve the mean square error of the same linear filters, which are the same under both P and P_G since the second order moments are the same under both measures.

Degenerate P: The ‘‘Semi-Stochastic’’ Setting: Suppose that $X_t = s_t$, $0 \leq t \leq T$, is a deterministic signal. Applying Theorem 2 with P degenerate on s_0^T gives

$$\begin{aligned} \text{cmse}_Q(\text{snr}) &= \frac{1}{\text{snr}} \int_0^{\text{snr}} \text{mse}_Q(\gamma) d\gamma \\ &= \frac{2}{\text{snr}} D_{\text{snr}} \left(P_{Y_0^T} \parallel Q_{Y_0^T} \right) \end{aligned} \quad (61)$$

where this time $P_{Y_0^T}$ is the law of a standard Brownian motion with drift s_t . Note that the invariance of relative entropy to a one-to-one transformation (of both arguments) implies that the right hand side of (61) is equal to

$$D_{\text{snr}} \left(P_{W_0^T} \parallel Q_{Y_0^T}^{s_0^T} \right) \quad (62)$$

where $Q_{Y_0^T}^{s_0^T}$ denotes that law of the output of the channel when the input is $\{X_t = s_t\}_{0 \leq t \leq T}$ and $X_0^T \sim Q$.

G. Estimation Theoretic Interpretation of a Chain Rule

Let the input and output of the channel be related as in (1), and let S be some additional information, jointly distributed with X_0^T , where the Brownian motion $\{W_t\}$ is independent of the pair (X_0^T, S) . Then

$$I_{\text{snr}}(X_0^T; Y_0^T) = I_{\text{snr}}(X_0^T, S; Y_0^T) \quad (63)$$

$$= I_{\text{snr}}(X_0^T; Y_0^T | S) + I_{\text{snr}}(Y_0^T; S) \quad (64)$$

$$= I_{\text{snr}}(X_0^T; Y_0^T | S) + D_{\text{snr}} \left(P_{Y_0^T | S} \parallel P_{Y_0^T} \mid P_S \right) \quad (65)$$

$$= \int \left[I_{\text{snr}}(P_{X_0^T | s}) + D_{\text{snr}} \left(P_{Y_0^T | s} \parallel P_{Y_0^T} \right) \right] dP_S(s) \quad (66)$$

where the first equality is due to the Markov relation $S - X_0^T - Y_0^T$, the rest are information theoretic identities, and we write $P_{X_0^T | s}$ for (a regular version of) the conditional distribution of X_0^T given S at s .⁷ Writing now $\text{cmse}(P, Q, \text{snr})$ for $\text{cmse}_Q(\text{snr})$ when we want to make the dependence on P explicit, and applying (31) to the integrand in (66) separately for every s , yields

⁷Cf. Section IV-A for the formalities.

(67), shown at the bottom of the page. On the other hand, from Duncan’s theorem [9, Theorem 3] [recall right equality in (28)], we have

$$\text{cmmse}(\text{snr}) = \frac{2}{\text{snr}} I_{\text{snr}}(X_0^T; Y_0^T). \quad (68)$$

Putting (67) and (68) together yields

$$\text{cmmse}(\text{snr}) = \int \text{cmse} \left(P_{X_0^T | s}, P, \text{snr} \right) dP_S(s). \quad (69)$$

Equality (69) could be deduced directly, from purely estimation theoretic reasoning, by noting that its right hand side is the performance of the optimum causal filter from the perspective of a genie with access to the side information. It thus provides, when coupled with (31), an estimation theoretic interpretation of the chain rule in (66).

H. Application to Minimax Causal Estimation

Suppose that the source P is known to belong to a class of possible sources \mathcal{P} . The goal is to find the causal filter that would perform best in the sense of minimizing the worst case difference between its MSE and the MMSE of the active source. Mathematically, our interest is in the minimax quantity [see (70), shown at the bottom of the page] and in the form of the achiever of the minimum on the right hand side of (70). Under benign regularity assumptions on \mathcal{P} , $\text{minimax}(\mathcal{P}, \text{snr})$ can be characterized as in (71)–(74), shown at the bottom of the next page, where:

- (a) is due to the fact that the minimum in (70) is achieved by a Bayes solution [24], namely, by the optimum filter under a Q induced by some prior on \mathcal{P}
- (b) is an application of (30)
- (c) is due to the ‘‘redundancy-capacity’’ theorem [12], [34], where we use $I_\Lambda(\text{snr})$ to denote the mutual information between X_0^T and Y_0^T when $X_0^T \sim \Lambda$.

To recap:

$$\text{minimax}(\mathcal{P}, \text{snr}) = \frac{2}{\text{snr}} C(\mathcal{P}, \text{snr}) \quad (75)$$

where $C(\mathcal{P}, \text{snr})$ is the capacity of the channel whose input is $P \in \mathcal{P}$ and whose output is the AWGN-corrupted version of

$$\begin{aligned} \int \text{cmse} \left(P_{X_0^T | s}, P, \text{snr} \right) dP_S(s) &= \int \left[I_{\text{snr}} \left(P_{X_0^T | s} \right) + D_{\text{snr}} \left(P_{Y_0^T | s} \parallel P_{Y_0^T} \right) \right] dP_S(s) \\ &= \frac{2}{\text{snr}} I_{\text{snr}}(X_0^T; Y_0^T) \end{aligned} \quad (67)$$

$$\text{minimax}(\mathcal{P}, \text{snr}) \triangleq \min_{\{\hat{X}_t(\cdot)\}_{0 \leq t \leq T}} \max_{P \in \mathcal{P}} \left\{ E_{P, \text{snr}} \left[\int_0^T (X_t - \hat{X}_t(Y^t))^2 dt \right] - \text{cmse}(P, P, \text{snr}) \right\} \quad (70)$$

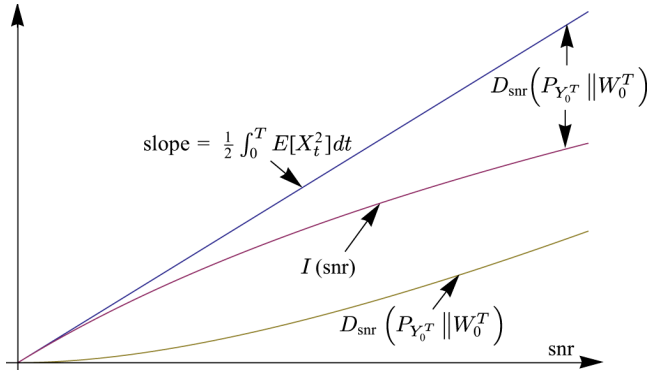


Fig. 2. Relationship between energy, mutual information, and divergence, for a generally distributed X_0^T .

the process $X^T \sim P$. Furthermore, the “strong redundancy-capacity” results of [30] are directly applicable in this context and imply that for any $\varepsilon > 0$ and any filter $\{\hat{X}_t(\cdot)\}_{0 \leq t \leq T}$

$$E_{P, \text{snr}} \left[\int_0^T (X_t - \hat{X}_t(Y^t))^2 dt \right] - \text{cmse}(P, P, \text{snr}) \geq (1 - \varepsilon) \frac{2}{\text{snr}} C(P, \text{snr}) \quad (76)$$

for all $P \in \mathcal{P}$ with the possible exception of sources in a subset $\mathcal{B} \subset \mathcal{P}$ where

$$w^*(\mathcal{B}) \leq e \cdot 2^{-\varepsilon \cdot C(P, \text{snr})} \quad (77)$$

w^* being the capacity-achieving prior, i.e., the one achieving the maximum in (73). Thus, in particular, \mathcal{B} is negligible for families of sources \mathcal{P}_T where $C(\mathcal{P}_T, \text{snr})$ is increasing with T and w^* assigns non-negligible mass to all regions of \mathcal{P}_T , which is the case with “natural” families of sources, cf. [30] for a discussion. Evidently, the minimax causal filter in this setting is optimal in a sense much stronger than the pessimistic “worst-case” one which it is optimized for.

III. EXAMPLE: A GAUSSIAN DC SIGNAL

In this section, we compute all the quantities that appear in Theorem 1, and explicitly verify (27), for the case where X_0^T is a

DC process with a standard normal amplitude whereas, under Q , its amplitude is also Gaussian but with mismatched expectation and variance.

To this end, consider first the problem of estimating the scalar Gaussian variable X based on

$$Y = \alpha X + N \quad (78)$$

where $N \sim \mathcal{N}(0, \sigma_N^2)$ is independent of X . The optimal estimator of X based on Y , under the assumption that $X \sim \mathcal{N}(\mu, \sigma^2)$, is

$$\mu \frac{\sigma_N^2}{\sigma_X^2 \alpha^2 + \sigma_N^2} + Y \frac{\sigma_X^2 \alpha}{\sigma_X^2 \alpha^2 + \sigma_N^2}. \quad (79)$$

If X is, instead, a *standard* Gaussian, the mean square error of the estimator in (79) is readily computed to be given by

$$\frac{\sigma_N^2 [\sigma_N^2 (1 + \mu^2) + \sigma^4 \alpha^2]}{(\sigma^2 \alpha^2 + \sigma_N^2)^2} \triangleq f(\mu, \sigma^2, \alpha, \sigma_N^2). \quad (80)$$

Suppose now that X_t is known to be a Gaussian DC signal, i.e., $X_t \equiv X$, where X is a standard Gaussian under P and $\sim \mathcal{N}(\mu, \sigma^2)$ under Q . In this case

$$\begin{aligned} \text{cmse}_Q(t, \gamma) &= f(\mu, \sigma^2, \sqrt{\gamma}t, t) \\ \text{mse}_Q(t, \gamma) &= f(\mu, \sigma^2, \sqrt{\gamma}T, T) \end{aligned} \quad (81)$$

so, see (82)–(86), shown at the bottom of the next page, and

$$\begin{aligned} \text{mse}_Q(\gamma) &= T \cdot f(\mu, \sigma^2, \sqrt{\gamma}T, T) \\ &= \frac{T^2 [T(1 + \mu^2) + \sigma^4 \gamma T^2]}{(\sigma^2 \gamma T^2 + T)^2}. \end{aligned} \quad (87)$$

Fig. 3 displays plots of $\text{cmse}_Q(\gamma)$ and $\text{mse}_Q(\gamma)$, for a fixed value of γ and T , as the mismatched values of σ^2 and μ vary.

Integrating (87), see (88)–(91), shown at the bottom of the next page. The expression in (91) coincides with the one in (86) evaluated at $\gamma = \text{snr}$, which checks with the first equality of (27). To check the second equality of (27), we note that in this example $I(\text{snr})$ is the mutual information between X and $Y_T = \sqrt{\text{snr}}T X + W_T$, where X is standard Gaussian and $W_T \sim \mathcal{N}(0, T)$ and independent of X , namely

$$I(\text{snr}) = \frac{1}{2} \log(1 + \text{snr}T) \quad (92)$$

$$\text{minimax}(\mathcal{P}, \text{snr}) \stackrel{(a)}{=} \min_Q \max_{P \in \mathcal{P}} [\text{cmse}(P, Q, \text{snr}) - \text{cmse}(P, P, \text{snr})] \quad (71)$$

$$\stackrel{(b)}{=} \frac{2}{\text{snr}} \min_Q \max_{P \in \mathcal{P}} D_{\text{snr}}(P_{Y_0^T} \parallel Q_{Y_0^T}) \quad (72)$$

$$\stackrel{(c)}{=} \frac{2}{\text{snr}} \max \left\{ I_\Lambda(\text{snr}) : \Lambda = \int_{\mathcal{P}} P d w(P) \text{ for some prior } w(\cdot) \text{ on } \mathcal{P} \right\} \quad (73)$$

$$\triangleq \frac{2}{\text{snr}} C(\mathcal{P}, \text{snr}) \quad (74)$$

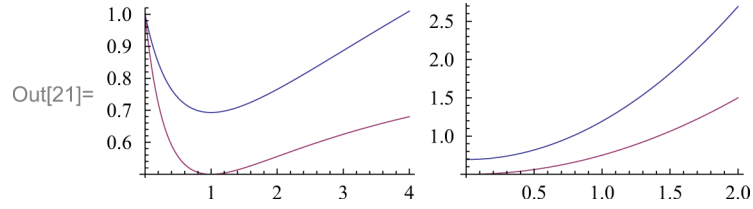


Fig. 3. Plots of $\text{cmse}_Q(\gamma)$ (blue) and $\text{mse}_Q(\gamma)$ (red), at $\gamma = T = 1$, for the Gaussian DC signal, as computed in (86) and (87), as the mismatched values of σ^2 and μ vary. In the left graph, $\mu = 0$ (the true value) and σ^2 is varied between 0 and 4. In the right graph, $\sigma^2 = 1$ (the true value) and μ is varied between 0 and 2. As is expected, the minima of both curves are attained at the true values ($\sigma^2 = 1$ and $\mu = 0$ in the respective plots).

while [see (93) and (94), shown at the bottom of the page]. Thus

$$\begin{aligned} & \frac{2}{\text{snr}} \left[I(\text{snr}) + D_{\text{snr}} \left(P_{Y_0^T} \left\| Q_{Y_0^T} \right\| \right) \right] \\ &= \frac{1}{\text{snr}} \left[\frac{\text{snr}(1 + \mu^2 - \sigma^2)T}{1 + \text{snr}\sigma^2 T} + \log(\text{snr}T\sigma^2 + 1) \right] \quad (95) \end{aligned}$$

which checks with (86) and (91).

Fig. 4 displays plots of $\text{cmse}_Q(\gamma)$, $\text{mse}_Q(\gamma)$, $\text{cmmse}(\gamma)$ and $\text{mmse}(\gamma)$ for this example, for time interval $T = 1$, where, under Q , X_t is a Gaussian DC signal with amplitude $\sim \mathcal{N}(1, 1)$. The regions are shaded corresponding to the regions of the generic Fig. 1, for $\text{snr} = 1$. Note that the curves of $\text{cmse}_Q(\gamma)$ and $\text{mse}_Q(\gamma)$ in Fig. 4 are monotonically decreasing. As discussed in Section II-B, this need not be the case in general and, in fact, need not be the case even in the present setting of a Gaussian DC signal. Figs. 5 and 6 exhibit these curves for the case where the amplitude under Q is $\mathcal{N}(1/2, 6)$, a case for

$$\text{cmse}_Q(\gamma) = \int_0^T f(\mu, \sigma^2, \sqrt{\gamma}t, t) dt \quad (82)$$

$$= \int_0^T \frac{t [t(1 + \mu^2) + \sigma^4 \gamma t^2]}{(\sigma^2 \gamma t^2 + t)^2} dt \quad (83)$$

$$= \frac{1}{\gamma} \left[\frac{\sigma^2 - (1 + \mu^2)}{\gamma t \sigma^4 + \sigma^2} + \log(\gamma t \sigma^2 + 1) \right] \Big|_{t=0}^{t=T} \quad (84)$$

$$= \frac{1}{\gamma} \left[\frac{\sigma^2 - (1 + \mu^2)}{\gamma T \sigma^4 + \sigma^2} + \log(\gamma T \sigma^2 + 1) + \frac{(1 + \mu^2)}{\sigma^2} - 1 \right] \quad (85)$$

$$= \frac{1}{\gamma} \left[\frac{\gamma(1 + \mu^2 - \sigma^2)T}{1 + \gamma \sigma^2 T} + \log(\gamma T \sigma^2 + 1) \right] \quad (86)$$

$$\frac{1}{\text{snr}} \int_0^{\text{snr}} \text{mse}_Q(\gamma) d\gamma = \frac{1}{\text{snr}} \int_0^{\text{snr}} \frac{T [T(1 + \mu^2) + \sigma^4 \gamma T^2]}{(\sigma^2 \gamma T^2 + T)^2} d\gamma \quad (88)$$

$$= \frac{1}{\text{snr}} \left[\frac{\sigma^2 - (1 + \mu^2)}{\gamma T \sigma^4 + \sigma^2} + \log(\gamma T \sigma^2 + 1) \right] \Big|_{\gamma=0}^{\gamma=\text{snr}} \quad (89)$$

$$= \frac{1}{\text{snr}} \left[\frac{\sigma^2 - (1 + \mu^2)}{\text{snr} T \sigma^4 + \sigma^2} + \log(\text{snr} T \sigma^2 + 1) + \frac{(1 + \mu^2)}{\sigma^2} - 1 \right] \quad (90)$$

$$= \frac{1}{\text{snr}} \left[\frac{\text{snr}(1 + \mu^2 - \sigma^2)T}{1 + \text{snr}\sigma^2 T} + \log(\text{snr} T \sigma^2 + 1) \right] \quad (91)$$

$$D_{\text{snr}} \left(P_{Y_0^T} \left\| Q_{Y_0^T} \right\| \right) = D \left(\mathcal{N}(0, \text{snr}T^2 + T) \left\| \mathcal{N}(\sqrt{\text{snr}}T\mu, \text{snr}T^2\sigma^2 + T) \right\| \right) \quad (93)$$

$$= \frac{\text{snr}(1 + \mu^2 - \sigma^2)T}{2 + 2\text{snr}\sigma^2 T} + \frac{1}{2} \log \frac{1 + \text{snr}\sigma^2 T}{1 + \text{snr}T} \quad (94)$$

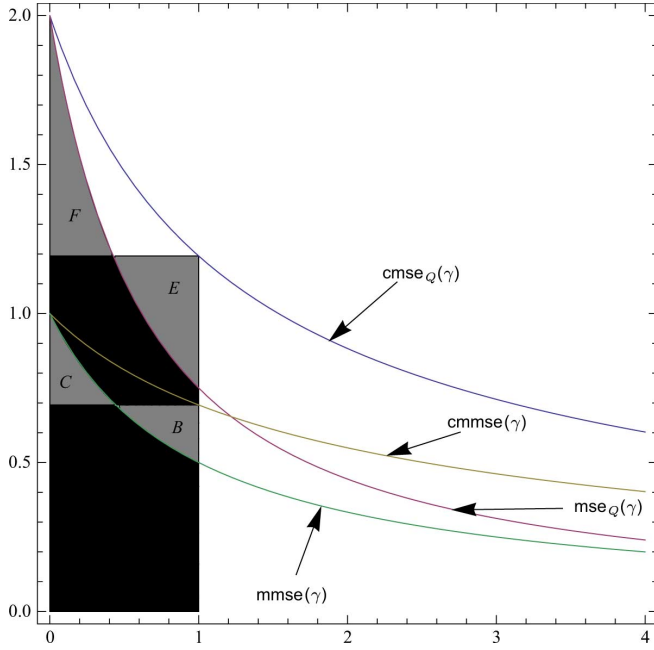


Fig. 4. Plots of $\text{cmse}_Q(\gamma)$, $\text{mse}_Q(\gamma)$, $\text{cmmse}(\gamma)$, and $\text{mmse}(\gamma)$ for the example of Section III. X_t is a DC signal with a standard normal amplitude whereas, under Q , the amplitude is distributed as $\mathcal{N}(1, 1)$. Here we have taken $T = 1$. The regions are shaded corresponding to the regions of Fig. 1, for $\text{SNR} = 1$.

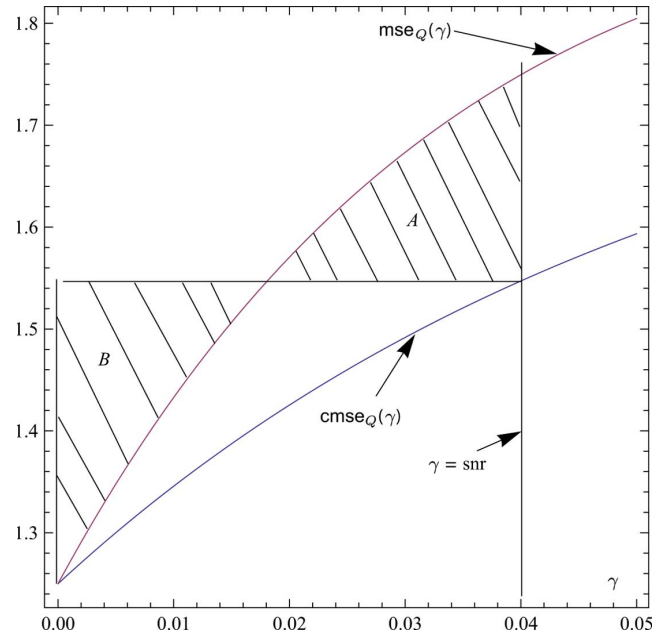


Fig. 5. Plots of $\text{cmse}_Q(\gamma)$ and $\text{mse}_Q(\gamma)$ for the example of Section III. X_t is a DC signal with a standard normal amplitude whereas, under Q , the amplitude is distributed as $\mathcal{N}(1/2, 6)$. Theorem 2 implies that the areas of Region A and Region B are equal.

which $\text{cmse}_Q(\gamma)$ and $\text{mse}_Q(\gamma)$ are increasing at low values of γ .

IV. PROOF OF MAIN RESULT

A. Some Notation

If A, B, C are three random objects taking values in $\mathcal{A}, \mathcal{B}, \mathcal{C}$ and defined on a common probability space with a probability measure P , we let $P_A, P_{A,B}$ etc. denote the probability mea-

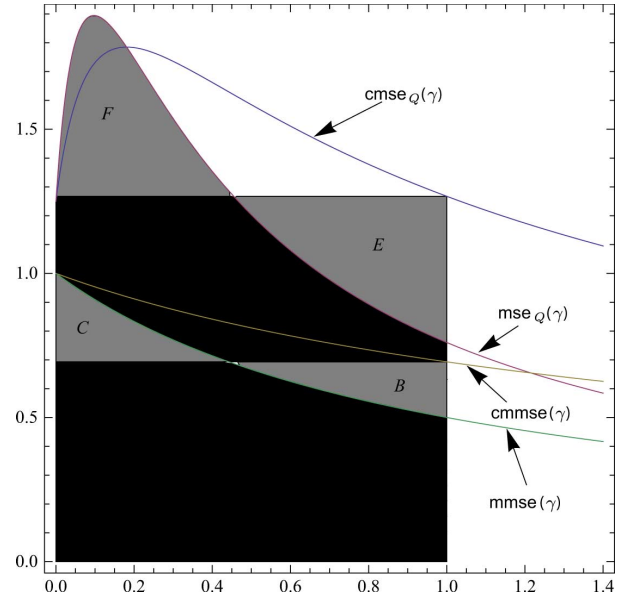


Fig. 6. Plots of $\text{cmse}_Q(\gamma)$ and $\text{mse}_Q(\gamma)$ for the example of Section III. X_t is a DC signal with a standard normal amplitude whereas, under Q , the amplitude is distributed as $\mathcal{N}(1/2, 6)$. As is consistent with the finding in Subsection II-B, the curve of $\text{mse}_Q(\gamma)$ is above that of $\text{cmse}_Q(\gamma)$ for as long as the latter is increasing. The two curves intersect at the value of γ where $\frac{d}{d\gamma}\text{cmse}_Q(\gamma) = 0$. The regions are shaded corresponding to the regions of Fig. 1, for $\text{SNR} = 1$.

asures induced on A , the pair (A, B) etc. while, e.g., $P_{A|B}$ denotes a regular version of the conditional distribution of A given B . $P_{A|b}$ is the distribution on \mathcal{A} obtained by evaluating that regular version at b . If Q is another probability measure on the same measurable space we similarly denote $Q_A, Q_{A|B}$, etc. As usual, given two measures on the same measurable space, e.g., P_A and Q_A , define their relative entropy (divergence) by

$$D(P_A||Q_A) = \int \left[\log \frac{dP_A}{dQ_A} \right] dP_A \quad (96)$$

when their Radon–Nikodym derivative $\frac{dP_A}{dQ_A}$ exists, defining $D(P_A||Q_A) = \infty$ otherwise. The logarithm in (96) is natural. Following [6], we further use the notation

$$D(P_{A|B}||Q_{A|B}|P_B) = \int D(P_{A|b}||Q_{A|b})dP_B(b) \quad (97)$$

where on the right side $D(P_{A|b}||Q_{A|b})$ is a divergence in the sense of (96) between the measures $P_{A|b}$ and $Q_{A|b}$. Similarly, we sometimes write

$$D(P_{A|B}||Q_{A|B}) \quad (98)$$

to denote $f(B)$ when $f(b) = D(P_{A|b}||Q_{A|b})$. Thus $D(P_{A|B}||Q_{A|B})$ is a random variable while $D(P_{A|B}||Q_{A|B}|P_B)$ is its expectation under P . With this notation, the chain rule for relative entropy (cf., e.g., [7, Subsection D.3]) is

$$D(P_{A,B}||Q_{A,B}) = D(P_A||Q_A) + D(P_{B|A}||Q_{B|A}|P_A) \quad (99)$$

and is valid regardless of the finiteness of both sides of the equation. The mutual information between A and B is defined as

$$I(A; B) = D(P_{A,B}||P_A \times P_B) \quad (100)$$

where $P_A \times P_B$ denotes the product measure induced by P_A and P_B . Similarly, the conditional mutual information between A and B , given C , is defined as

$$I(A; B|C) = D(P_{A,B|C} \| P_{A|C} \times P_{B|C} | P_C). \quad (101)$$

In what follows, the roles of A , B , C will primarily be played either by random variables and vectors, or by AWGN-corrupted continuous-time stochastic processes. For this case, the conditioned and unconditioned relative entropy and mutual information, as defined above, are particularly well investigated and understood objects, cf. [8], [20], [23], [33], [39], [40] and references therein.

B. Intuition

Recall from Section II-A that, given [16, Theorem 8] and [9, Theorem 3], it will suffice to prove Theorem 2. The first part of the theorem, (29), is merely an extension of (11) from estimation of random variables to estimation of random signals. We establish this part by extending the finite-dimensional vector version of (11) to random signals via finite-dimensional approximations of the latter. In the spirit of other constructions, such as that of the stochastic integral in [21], we do this by proving the result first for piecewise-constant processes, and then “lifting” to general finite-energy processes via standard limit arguments.

One way to gain some intuition as to why (30) should hold is to consider first the case where the noise-free signal, under both P and Q , is a “sample and hold” process on $[0, T + \delta]$ which is constant (in time, not ω) on $(T, T + \delta]$. In other words, $X_t \equiv X$ for $t \in (T, T + \delta]$, for a random variable X of finite second moment, arbitrarily jointly distributed with X_0^T . We would then have (102)–(105), shown at the bottom of the page, where the

last equality is simply the definition of conditional divergence [recall (97)] and the one preceding it follows by applying what we have already found in the Introduction to hold for DC processes, namely Equality (21), this time on a process in the interval $(T, T + \delta]$ whose amplitude is distributed as $P_{X|Y_0^T}$ while under the mismatched filter it is distributed as $Q_{X|Y_0^T}$. Combining (105) with the chain rule for relative entropy (99) yields (106), shown at the bottom of the page, and we obtain (107), shown at the bottom of the page. It is then plausible to expect (107), which is essentially equivalent to (30) (it is actually the derivative of (30) with respect to time T), to hold for general finite-energy processes since, for small δ , the latter are well approximated by the “sample and hold” process. As we elaborate on in Section IV-D, this intuitive path can be made rigorous and yield a formal proof of (30) via the “time-incremental” channel approach of [16]. Our formal proof, to which we now turn, follows a somewhat different path of piecewise constant process approximations.

C. Proof of Theorem 2

Vector Version of the D-MSE Relationship: Consider an M -dimensional vector observation $G^M = (G_1, G_2, \dots, G_M)$ given by

$$G_i = \sqrt{\gamma} A_i + B_i \quad (108)$$

where $B^M \sim \mathcal{N}(0, I)$ is independent of A^M . Suppose $A^M \sim P$ and let $\text{mse}_Q^{\text{vec}}(\gamma)$ denote⁸ the cumulative mean square error in estimating the components of A^M based on G^M using the

⁸The superscript “vec”, which stands for “vector”, is added to distinguish between this notation and the one already defined and used for continuous-time estimation.

$$\int_T^{T+\delta} E \left[(X_t - E_Q[X_t|Y_0^t; \text{snr}])^2 - (X_t - E[X_t|Y_0^t; \text{snr}])^2 \right] dt \quad (102)$$

$$= E \left\{ \int_T^{T+\delta} E \left[(X_t - E_Q[X_t|Y_0^t; \text{snr}])^2 - (X_t - E[X_t|Y_0^t; \text{snr}])^2 \middle| Y_0^T \right] dt \right\} \quad (103)$$

$$= \int \frac{2}{\text{snr}} D_{\text{snr}} \left(P_{Y_T^{T+\delta}|y_0^T} \parallel Q_{Y_T^{T+\delta}|y_0^T} \right) dP_{Y_0^T}(y_0^T) \quad (104)$$

$$= \frac{2}{\text{snr}} D_{\text{snr}} \left(P_{Y_T^{T+\delta}|Y_0^T} \parallel Q_{Y_T^{T+\delta}|Y_0^T} \middle| P_{Y_0^T} \right) \quad (105)$$

$$\begin{aligned} & D_{\text{snr}} \left(P_{Y_0^{T+\delta}} \parallel Q_{Y_0^{T+\delta}} \right) - D_{\text{snr}} \left(P_{Y_0^T} \parallel Q_{Y_0^T} \right) \\ &= \frac{\text{snr}}{2} \int_T^{T+\delta} E \left[(X_t - E_Q[X_t|Y_0^t; \text{snr}])^2 - (X_t - E[X_t|Y_0^t; \text{snr}])^2 \right] dt \end{aligned} \quad (106)$$

$$\begin{aligned} & D_{\text{snr}} \left(P_{Y_0^{T+\delta}} \parallel Q_{Y_0^{T+\delta}} \right) - D_{\text{snr}} \left(P_{Y_0^T} \parallel Q_{Y_0^T} \right) \\ &= \frac{\text{snr}}{2} \cdot \delta \cdot E \left[(X_T - E_Q[X_T|Y_0^T; \text{snr}])^2 - (X_T - E[X_T|Y_0^T; \text{snr}])^2 \right] + o(\delta) \end{aligned} \quad (107)$$

estimator that would have been mean-square optimal if $A^M \sim Q$. Assume that both P and Q have finite second moments. With $\text{mmse}^{\text{vec}}(\gamma)$ standing for $\text{mse}_P^{\text{vec}}(\gamma)$, we have:

Lemma 1:

$$D(P * \mathcal{N}(0, \sigma^2 I) \| Q * \mathcal{N}(0, \sigma^2 I)) = \frac{1}{2} \int_0^{1/\sigma^2} [\text{mse}_Q^{\text{vec}}(\gamma) - \text{mmse}^{\text{vec}}(\gamma)] d\gamma \quad (109)$$

where $Q * \mathcal{N}(0, \sigma^2 I)$ stands for the distribution of the M -dimensional random vector obtained by sampling from Q and then corrupting its components by an independent vector distributed $\mathcal{N}(0, \sigma^2 I)$.

Lemma 1 is nothing but the multivariate version of the ‘‘D-MSE’’ formula

$$D(P * \mathcal{N}(0, \sigma^2) \| Q * \mathcal{N}(0, \sigma^2)) = \frac{1}{2} \int_0^{1/\sigma^2} [\text{mse}_Q^{\text{scalar}}(\gamma) - \text{mmse}^{\text{scalar}}(\gamma)] d\gamma \quad (110)$$

which is (14) in [37]. That (110) carries over to the vector case to yield Lemma 1 was mentioned in [37]. For completeness, in the Appendix we provide a proof of Lemma 1, relying on its scalar version (110).

The Gist: Piecewise-Constant Processes: This subsection is dedicated to the main part of the proof, which is to establish (29) and (30) under the assumption that X_0^T is piecewise constant both under P and under Q . Thus, throughout this subsection, assume existence of M and a random vector $A^M = (A_1, \dots, A_M)$ such that

$$X_t \equiv A_i \text{ for all } \frac{i-1}{M}T < t \leq \frac{i}{M}T, \text{ and } 1 \leq i \leq M. \quad (111)$$

We use P and Q to denote either the measures governing the continuous-time piecewise-constant signal X_0^T satisfying (111), or those governing the M -dimensional random vector A^M , the distinction being clear from the context. Let

$$J^M = A^M + N^M \quad (112)$$

where $N^M \sim N(0, \frac{M}{\text{snr}T} \cdot I)$, independent of A^M . Letting P_{JM} and Q_{JM} denote the distribution of J^M in (112) when, respectively, $A^M \sim P$ and $A^M \sim Q$, we have for any $\alpha > 0$

$$D\left(P * \mathcal{N}\left(0, \frac{M}{\text{snr}T} \cdot I\right) \parallel Q * \mathcal{N}\left(0, \frac{M}{\text{snr}T} \cdot I\right)\right) = D(P_{JM} \| Q_{JM}) \quad (113)$$

$$\stackrel{(a)}{=} D(P_{\alpha \cdot J^M} \| Q_{\alpha \cdot J^M}) \quad (114)$$

$$\stackrel{(b)}{=} D\left(P_{\{Y_{iT/M} - Y_{(i-1)T/M}\}_{i=1}^M} \parallel Q_{\{Y_{iT/M} - Y_{(i-1)T/M}\}_{i=1}^M}\right) \quad (115)$$

$$\stackrel{(c)}{=} D_{\text{snr}}\left(P_{Y_0^T} \parallel Q_{Y_0^T}\right) \quad (116)$$

where:

- Equality (a) is due to the invariance of relative entropy to scaling (of both arguments by the same factor).
- In (115), $P_{\{Y_{iT/M} - Y_{(i-1)T/M}\}_{i=1}^M}$ and $Q_{\{Y_{iT/M} - Y_{(i-1)T/M}\}_{i=1}^M}$ denote the laws of $\{Y_{iT/M} - Y_{(i-1)T/M}\}_{i=1}^M$, when A^M is distributed under P and Q , respectively, X_0^T is given in (111), and Y_0^T is related to X_0^T via the channel in (1) with $\gamma = \text{snr}$. Equality (b) follows upon noting that $P_{\alpha \cdot J^M} = P_{\{Y_{iT/M} - Y_{(i-1)T/M}\}_{i=1}^M}$ and $Q_{\alpha \cdot J^M} = Q_{\{Y_{iT/M} - Y_{(i-1)T/M}\}_{i=1}^M}$ when $\alpha = \sqrt{\frac{\text{snr}T}{M}}$.
- Equality (c) is due to the fact that X_0^T is constant on the intervals $(\frac{i-1}{M}T, \frac{i}{M}T]$, and thus $\{Y_{iT/M} - Y_{(i-1)T/M}\}_{i=1}^M$ are sufficient statistics for Y_0^T (under both P and Q).

On the other hand, by (111), for any $\gamma \geq 0$

$$\text{mse}_Q(\gamma) - \text{mmse}(\gamma) = \frac{T}{M} \left[\text{mse}_Q^{\text{vec}}\left(\frac{\gamma T}{M}\right) - \text{mmse}^{\text{vec}}\left(\frac{\gamma T}{M}\right) \right]. \quad (117)$$

Consequently, we have (118)–(121), shown at the bottom of the page, where the first equality is an integration over (117), the third equality is an application of (109), and the last equality follows from (116). We have thus proven (29).

We now progress to the proof of (30). To this end, consider first the case where X_t is a DC signal, i.e.,

$$X_t \equiv X, \quad 0 \leq t \leq T \quad (122)$$

for a random variable $X \sim P$, whereas the mismatched filter assumes $X \sim Q$. Define

$$F(P, Q, \gamma, T) \triangleq \text{cmse}_Q^{\text{DC}}(\gamma) - \text{cmmse}^{\text{DC}}(\gamma) \quad (123)$$

$$\int_0^{\text{snr}} [\text{mse}_Q(\gamma) - \text{mmse}(\gamma)] d\gamma = \frac{T}{M} \int_0^{\text{snr}} \left[\text{mse}_Q^{\text{vec}}\left(\frac{\gamma T}{M}\right) - \text{mmse}^{\text{vec}}\left(\frac{\gamma T}{M}\right) \right] d\gamma \quad (118)$$

$$= \int_0^{\text{snr}T/M} [\text{mse}_Q^{\text{vec}}(\gamma') - \text{mmse}^{\text{vec}}(\gamma')] d\gamma' \quad (119)$$

$$= 2D\left(P * \mathcal{N}\left(0, \frac{M}{\text{snr}T} \cdot I\right) \parallel Q * \mathcal{N}\left(0, \frac{M}{\text{snr}T} \cdot I\right)\right) \quad (120)$$

$$= 2D_{\text{snr}}\left(P_{Y_0^T} \parallel Q_{Y_0^T}\right) \quad (121)$$

where we recall from the Introduction that $\text{cmse}_Q^{\text{DC}}(\gamma)$ and $\text{cmmse}^{\text{DC}}(\gamma)$ denote $\text{cmse}_Q(\gamma)$ and $\text{cmmse}(\gamma)$ of (4) and (6) for the case where X_t is the DC signal in (122). Thus, $F(P, Q, \gamma, T)$ is the price of mismatch in causal estimation of the DC signal for duration T using a filter that assumes the amplitude is distributed according to Q when it is actually distributed according to P . We recall from (21) that

$$F(P, Q, \gamma, T) = \frac{2}{\gamma} D \left(P * \mathcal{N} \left(0, \frac{1}{\gamma T} \right) \parallel Q * \mathcal{N} \left(0, \frac{1}{\gamma T} \right) \right). \quad (124)$$

Returning now to the generality of X_0^T as in (111), consider (125)–(138), shown at the bottom of the page, where:

- (a) follows from a switch between time-integration and expectation, justified in the standard way due to the fact that the integrand is in $L^2(dt dP)$.
- (b) is immediate from the definition of F [recall (123)].
- (c) follows from the relationship in (124). Note that the divergence inside the expectation in (133) is of the form (98), namely it is between the two random measures $P_{A_i|Y_0^{i-1}T} * \mathcal{N} \left(0, \frac{M}{\text{snr}T} \right)$ and $Q_{A_i|Y_0^{i-1}T} * \mathcal{N} \left(0, \frac{M}{\text{snr}T} \right)$, the expectation is then over the randomness in $Y_0^{i-1}T$ (distributed according to $P_{Y_0^{i-1}T}$).

$$\text{cmse}_Q(\text{snr}) - \text{cmmse}(\text{snr}) \quad (125)$$

$$= \int_0^T \left\{ E \left[(X_t - E_Q[X_t|Y_0^t; \text{snr}])^2 \right] - E \left[(X_t - E[X_t|Y_0^t; \text{snr}])^2 \right] \right\} dt \quad (126)$$

$$= \int_0^T E \left[(X_t - E_Q[X_t|Y_0^t; \text{snr}])^2 - (X_t - E[X_t|Y_0^t; \text{snr}])^2 \right] dt \quad (127)$$

$$= \sum_{i=1}^M \int_{\frac{i-1}{M}T}^{\frac{i}{M}T} E \left[(X_t - E_Q[X_t|Y_0^t; \text{snr}])^2 - (X_t - E[X_t|Y_0^t; \text{snr}])^2 \right] dt \quad (128)$$

$$= \sum_{i=1}^M \int_{\frac{i-1}{M}T}^{\frac{i}{M}T} E \left\{ E \left[(X_t - E_Q[X_t|Y_0^t; \text{snr}])^2 - (X_t - E[X_t|Y_0^t; \text{snr}])^2 \mid Y_0^{i-1}T \right] \right\} dt \quad (129)$$

$$\stackrel{(a)}{=} \sum_{i=1}^M E \left\{ \int_{\frac{i-1}{M}T}^{\frac{i}{M}T} E \left[(X_t - E_Q[X_t|Y_0^t; \text{snr}])^2 - (X_t - E[X_t|Y_0^t; \text{snr}])^2 \mid Y_0^{i-1}T \right] dt \right\} \quad (130)$$

$$= \sum_{i=1}^M E \left\{ \int_{\frac{i-1}{M}T}^{\frac{i}{M}T} E \left[\left(A_i - E_Q \left[A_i \mid Y_0^{\frac{i-1}{M}T}, Y_{\frac{i-1}{M}T}^t; \text{snr} \right] \right)^2 - \left(A_i - E \left[A_i \mid Y_0^{\frac{i-1}{M}T}, Y_{\frac{i-1}{M}T}^t; \text{snr} \right] \right)^2 \mid Y_0^{i-1}T \right] dt \right\} \quad (131)$$

$$\stackrel{(b)}{=} \sum_{i=1}^M E \left\{ F \left(P_{A_i|Y_0^{i-1}T}, Q_{A_i|Y_0^{i-1}T}, \text{snr}, T/M \right) \right\} \quad (132)$$

$$\stackrel{(c)}{=} \sum_{i=1}^M E \left\{ \frac{2}{\text{snr}} D \left(P_{A_i|Y_0^{i-1}T} * \mathcal{N} \left(0, \frac{M}{\text{snr}T} \right) \parallel Q_{A_i|Y_0^{i-1}T} * \mathcal{N} \left(0, \frac{M}{\text{snr}T} \right) \right) \right\} \quad (133)$$

$$\stackrel{(d)}{=} \frac{2}{\text{snr}} \sum_{i=1}^M D \left(P_{A_i|Y_0^{i-1}T} * \mathcal{N} \left(0, \frac{M}{\text{snr}T} \right) \parallel Q_{A_i|Y_0^{i-1}T} * \mathcal{N} \left(0, \frac{M}{\text{snr}T} \right) \mid P_{Y_0^{i-1}T} \right) \quad (134)$$

$$\stackrel{(e)}{=} \frac{2}{\text{snr}} \sum_{i=1}^M D \left(P_{A_i|\{Y_{jT/M} - Y_{(j-1)T/M}\}_{j=1}^{i-1}} * \mathcal{N} \left(0, \frac{M}{\text{snr}T} \right) \parallel Q_{A_i|\{Y_{jT/M} - Y_{(j-1)T/M}\}_{j=1}^{i-1}} * \mathcal{N} \left(0, \frac{M}{\text{snr}T} \right) \mid P_{\{Y_{jT/M} - Y_{(j-1)T/M}\}_{j=1}^{i-1}} \right) \quad (135)$$

$$\stackrel{(f)}{=} \frac{2}{\text{snr}} \sum_{i=1}^M D(P_{J_i|J^{i-1}} \parallel Q_{J_i|J^{i-1}} \mid P_{J^{i-1}}) \quad (136)$$

$$\stackrel{(g)}{=} \frac{2}{\text{snr}} D(P_{J^M} \parallel Q_{J^M}) \quad (137)$$

$$\stackrel{(h)}{=} \frac{2}{\text{snr}} D_{\text{snr}}(P_{Y_0^T} \parallel Q_{Y_0^T}) \quad (138)$$

- (d) is simply a rewriting of the expectation in (c) using the notation introduced in (97).
- (e) follows similarly as (c) in (116) due to the piecewise constancy of X_0^T on the intervals $(\frac{i-1}{M}T, \frac{i}{M}T]$.
- (f) follows since, by construction of P_{JM} and Q_{JM} in (112), for $1 \leq i \leq M$

$$P_{A_i|\{Y_{jT/M} - Y_{(j-1)T/M}\}_{j=1}^{i-1}} * \mathcal{N}\left(0, \frac{M}{\text{snr}T}\right) = P_{J_i|J^{i-1}}$$

and

$$Q_{A_i|\{Y_{jT/M} - Y_{(j-1)T/M}\}_{j=1}^{i-1}} * \mathcal{N}\left(0, \frac{M}{\text{snr}T}\right) = Q_{J_i|J^{i-1}}. \quad (139)$$

- (g) follows from the chain rule for relative entropy.
- (h) follows from (116),

proving (30). To recap, we have proven (29) and (30) for arbitrary processes of the form displayed in (111).⁹

Passing to the Limit: For general P and Q of finite average power, consider the induced stepwise process $X_0^{(n),T}$ defined by

$$X_t^{(n)} \equiv \frac{1}{2^n T} \int_{i2^{-n}T}^{(i+1)2^{-n}T} X_t dt \text{ for } t \in (i2^{-n}T, (i+1)2^{-n}T]. \quad (140)$$

The finite energy assumption implies that the integral $\int_{i2^{-n}T}^{(i+1)2^{-n}T} X_t dt$ exists and is finite P - and Q - almost surely, and is also both in $L_2(P)$ and in $L_2(Q)$. Thus, in particular, $X_0^{(n),T}$ is a finite average power process under both P and Q . Let $P^{(n)}$ and $Q^{(n)}$ be the measures governing $X_0^{(n),T}$ when X_0^T is distributed under P and Q , respectively. Note that the process $X_0^{(n),T}$ is of the form displayed in (111) (with $M = 2^n$), for which we have already proven (29) and (30). It thus remains only to establish continuity of the functionals $D_{\text{snr}}(P_{Y_0^T} \| Q_{Y_0^T})$, $\int_0^{\text{snr}} \text{mse}_Q(\gamma) d\gamma$ and $\text{cmse}_Q(\text{snr})$, at fixed $\text{snr} > 0$, in the sense that, when evaluated at $(P^{(n)}, Q^{(n)})$ instead of (P, Q) , they converge to their values on the latter as $n \rightarrow \infty$.¹⁰

To that end, note first that P and Q being of finite average power guarantees

$$X_0^{(n),T} \xrightarrow{n \rightarrow \infty} X_0^T \text{ in } L^2(dtdP) \text{ and in } L^2(dtdQ). \quad (141)$$

For any two measures P' , Q' of finite average power, the induced measures $P'_{Y_0^T}$, $Q'_{Y_0^T}$ are absolutely continuous with respect to one another (cf., e.g., [27]). Thus, in particular, the Radon–Nykodim derivatives $\frac{dP'_{Y_0^T}}{dQ'_{Y_0^T}}$ and $\frac{dP_{Y_0^T}}{dQ_{Y_0^T}}$ all exist. Furthermore, due to (121), they satisfy (cf., e.g., [20])

$$\frac{dP'_{Y_0^T}}{dQ'_{Y_0^T}} \rightarrow \frac{dP_{Y_0^T}}{dQ_{Y_0^T}} \quad P_{Y_0^T} - a.s. \quad (142)$$

⁹This actually proves (29) and (30) for arbitrary stepwise processes (a.k.a. simple processes [21]), with constancy intervals not necessarily of equal lengths. The extension to the latter is elementary but not necessary for what follows.

¹⁰We also need continuity in this same sense to hold for $\int_0^{\text{snr}} \text{mmse}(\gamma) d\gamma$ and $\text{cmmse}(\text{snr})$, but that would follow as a special case of the continuity of $\int_0^{\text{snr}} \text{mse}_Q(\gamma) d\gamma$ and $\text{cmse}_Q(\text{snr})$.

From reasoning similar to that leading to [9, Equation (20)], we obtain also

$$\sup_n \int \left[\log \frac{dP_{Y_0^T}^{(n)}}{dQ_{Y_0^T}^{(n)}} \right] dP_{Y_0^T} < \infty. \quad (143)$$

Consequently

$$D_{\text{snr}}(P_{Y_0^T}^{(n)} \| Q_{Y_0^T}^{(n)}) = \int \left[\log \frac{dP_{Y_0^T}^{(n)}}{dQ_{Y_0^T}^{(n)}} \right] dP_{Y_0^T}^{(n)} \quad (144)$$

$$= \int \left[\log \frac{dP_{Y_0^T}^{(n)}}{dQ_{Y_0^T}^{(n)}} \right] dP_{Y_0^T} \quad (145)$$

$$\xrightarrow{n \rightarrow \infty} \int \left[\log \frac{dP_{Y_0^T}}{dQ_{Y_0^T}} \right] dP_{Y_0^T} \quad (146)$$

$$= D_{\text{snr}}(P_{Y_0^T} \| Q_{Y_0^T}) \quad (147)$$

where (146) is due to the convergence (142), coupled with the uniform integrability of $\log \frac{dP_{Y_0^T}^{(n)}}{dQ_{Y_0^T}^{(n)}}$ implied by (143).

As for the functional $\text{cmse}_Q(\text{snr})$, (141) guarantees that $\{E_{Q^{(n)}}[X_t|Y_0^t]\}_{0 \leq t \leq T}$ converges to $\{E_Q[X_t|Y_0^t]\}_{0 \leq t \leq T}$ in $L^2(dtdQ_{Y_0^T})$. Coupled with the existence and boundedness of the Radon–Nykodim derivative $\frac{dQ_{Y_0^T}}{dP_{Y_0^T}}$, this implies that $\{E_{Q^{(n)}}[X_t|Y_0^t]\}_{0 \leq t \leq T}$ converges to $\{E_Q[X_t|Y_0^t]\}_{0 \leq t \leq T}$ also in $L^2(dtdP_{Y_0^T})$, in turn implying the convergence of $\text{cmse}_{Q^{(n)}}(\text{snr})$ (with true source $P^{(n)}$) to $\text{cmse}_Q(\text{snr})$ (with true source P). The convergence to $\text{mse}_Q(\text{snr})$ of $\text{mse}_{Q^{(n)}}(\text{snr})$ follows a similar line of reasoning implying, by the arbitrariness of snr and the bounded convergence theorem, that $\int_0^{\text{snr}} \text{mse}_{Q^{(n)}}(\gamma) d\gamma$ (with true source $P^{(n)}$) converges to $\int_0^{\text{snr}} \text{mse}_Q(\gamma) d\gamma$ (with true source P).

D. Alternative Proof Routes and an Extension

Proof Alternatives: The route we have taken in the proof above is to establish the result for stepwise processes, and then infer the validity for general processes via the denseness of stepwise processes in the space of finite energy processes, coupled with standard continuity arguments. An alternative of a similar spirit would have been to prove the result for processes expressible as finite sums of orthonormal functions (with random coefficients), and then pass to the limit for arbitrary finite-energy processes.

A route to proving (30) of a different spirit is to make direct use of the Girsanov theorem [4], [14]. Denoting $\pi_t = E[X_t|Y_0^t; \gamma]$ and $\pi_t^Q = E_Q[X_t|Y_0^t; \gamma]$, we note that the innovations processes

$$\bar{W}_t = Y_t - \int_0^t \sqrt{\gamma} \pi_s ds \quad (148)$$

and

$$\bar{V}_t = Y_t - \int_0^t \sqrt{\gamma} \pi_s^Q ds \quad (149)$$

are standard Brownian motions, respectively, under P and Q . Applying the Girsanov theorem under P and under Q gives, respectively

$$\log \frac{dP_{Y_0^T}}{d\mu} = \int_0^T \sqrt{\gamma} \pi_t dY_t - \frac{1}{2} \int_0^T (\sqrt{\gamma} \pi_t)^2 dt \quad (150)$$

and

$$\log \frac{dQ_{Y_0^T}}{d\mu} = \int_0^T \sqrt{\gamma} \pi_t^Q dY_t - \frac{1}{2} \int_0^T (\sqrt{\gamma} \pi_t^Q)^2 dt \quad (151)$$

where μ denotes the Wiener measure on Y_0^T . Thus, we obtain (152)–(160), shown at the bottom of the page, where:

- (a) follows from substituting from (150) and (151).
- (b) follows since, by (148), $dY_t = d\bar{W}_t + \sqrt{\gamma} \pi_t dt$.
- (c) follows since \bar{W}_t is a standard Brownian motion under P (with respect to which the expectation is taken).
- (d) follows since, by the orthogonality property

$$\begin{aligned} & E \left[\left(\pi_t^Q - X_t \right)^2 \right] \\ &= E \left[\left(\pi_t^Q - \pi_t + \pi_t - X_t \right)^2 \right] \\ &= E \left[\left(\pi_t^Q - \pi_t \right)^2 + \left(\pi_t - X_t \right)^2 \right]. \end{aligned} \quad (161)$$

Note that this constitutes a direct proof of (30) that does not rely on (11). Further, as discussed in Section II-A, when specialized to DC processes (30) yields (11) and so the proof just given yields, as a very special case, an independent proof of (11) (and thus of the main result of [37]).

Yet another alternative to a proof of Theorem 2 is to use the SNR- and time-incremental channel ideas of [16]. Specifically, for establishing (29), a line of attack analogous to that in [16, Subsection III-C] can be taken by considering the SNR-incremental channel

$$dY_{1,t} = X_t dt + \sigma_1 dW_{1,t} \quad (162)$$

$$dY_{2,t} = dY_{1,t} + \sigma_2 dW_{2,t} \quad (163)$$

where $\{W_{1,t}\}$ and $\{W_{2,t}\}$ are independent standard Brownian motions, jointly independent of $\{X_t\}$. If σ_1 and σ_2 are chosen so that the SNRs of the first channel and the composite channel are $\text{snr} + \gamma$ and snr , then

$$(\text{snr} + \gamma) dY_{1,t} = \text{snr} \cdot dY_{2,t} + \gamma X_t dt + \sqrt{\delta} dW_t \quad (164)$$

where $\{W_t\}$ is a standard Brownian motion independent of $\{X_t\}$ and $\{Y_{2,t}\}$. This now motivates characterizing the low-SNR asymptotics, analogously as in [16, Lemma 5], to obtain

$$\begin{aligned} & \lim_{\gamma \downarrow 0} \frac{1}{\gamma} D_\gamma \left(P_{Y_0^T} \parallel Q_{Y_0^T} \right) \\ &= \frac{1}{2} \left[\int_0^T E \left[(X_t - E_Q X_t)^2 \right] dt - \int_0^T \text{Var}(X_t) dt \right]. \end{aligned} \quad (165)$$

One way of establishing equality (165) that does not pass through simple process approximations is by considering the Karhunen-Loève expansion of X_0^T (with respect to P), getting arbitrarily precise finite-dimensional approximations of the

$$D_\gamma \left(P_{Y_0^T} \parallel Q_{Y_0^T} \right) \quad (152)$$

$$= E \left[\log \frac{dP_{Y_0^T}}{dQ_{Y_0^T}} \right] \quad (153)$$

$$= E \left[\log \left(\frac{\frac{dP_{Y_0^T}}{d\mu}}{\frac{dQ_{Y_0^T}}{d\mu}} \right) \right] \quad (154)$$

$$\stackrel{(a)}{=} E \left\{ \int_0^T (\sqrt{\gamma} \pi_t - \sqrt{\gamma} \pi_t^Q) dY_t - \frac{1}{2} \int_0^T \left[(\sqrt{\gamma} \pi_t)^2 - (\sqrt{\gamma} \pi_t^Q)^2 \right] dt \right\} \quad (155)$$

$$\stackrel{(b)}{=} E \left\{ \int_0^T (\sqrt{\gamma} \pi_t - \sqrt{\gamma} \pi_t^Q) d\bar{W}_t + \int_0^T (\sqrt{\gamma} \pi_t - \sqrt{\gamma} \pi_t^Q) \sqrt{\gamma} \pi_t dt - \frac{1}{2} \int_0^T \left[(\sqrt{\gamma} \pi_t)^2 - (\sqrt{\gamma} \pi_t^Q)^2 \right] dt \right\} \quad (156)$$

$$\stackrel{(c)}{=} E \left\{ \int_0^T (\sqrt{\gamma} \pi_t - \sqrt{\gamma} \pi_t^Q) \sqrt{\gamma} \pi_t dt - \frac{1}{2} \int_0^T \left[(\sqrt{\gamma} \pi_t)^2 - (\sqrt{\gamma} \pi_t^Q)^2 \right] dt \right\} \quad (157)$$

$$= \frac{\gamma}{2} \int_0^T E \left[\left(\pi_t - \pi_t^Q \right)^2 \right] dt \quad (158)$$

$$\stackrel{(d)}{=} \frac{\gamma}{2} \int_0^T E \left[\left(\pi_t^Q - X_t \right)^2 - \left(\pi_t - X_t \right)^2 \right] dt \quad (159)$$

$$= \frac{\gamma}{2} [\text{cmse}_Q(\gamma) - \text{cmmse}(\gamma)] \quad (160)$$

integrals in (165) in the “transform domain”, and invoking the differential version of Lemma 1, namely (A.3), at $\gamma \rightarrow 0$. Equipped with (165), one can now apply it to the Gaussian channel (164), conditioned on $\{Y_{2,t}\}$, to obtain (166) and (167), shown at the bottom of the page. On the other hand

$$\begin{aligned} & D\left(P_{Y_{1,0}^T|Y_{2,0}^T} \parallel Q_{Y_{1,0}^T|Y_{2,0}^T} \middle| P_{Y_{2,0}^T}\right) \\ &= D\left(P_{Y_{1,0}^T} \parallel Q_{Y_{1,0}^T}\right) \\ &\quad + D\left(P_{Y_{2,0}^T|Y_{1,0}^T} \parallel Q_{Y_{2,0}^T|Y_{1,0}^T} \middle| P_{Y_{1,0}^T}\right) \\ &\quad - D\left(P_{Y_{2,0}^T} \parallel Q_{Y_{2,0}^T}\right) \\ &= D\left(P_{Y_{1,0}^T} \parallel Q_{Y_{1,0}^T}\right) - D\left(P_{Y_{2,0}^T} \parallel Q_{Y_{2,0}^T}\right) \\ &= D_{\text{snr}+\gamma}\left(P_{Y_0^T} \parallel Q_{Y_0^T}\right) - D_{\text{snr}}\left(P_{Y_0^T} \parallel Q_{Y_0^T}\right) \end{aligned} \quad (168)$$

where the first equality follows from the chain rule for divergence and the second because, as is clear from (163), the conditional distribution of $Y_{2,0}^T$ given $Y_{1,0}^T$ does not depend on the distribution of X_0^T , so $P_{Y_{2,0}^T|Y_{1,0}^T} = Q_{Y_{2,0}^T|Y_{1,0}^T}$. Putting (167) and (168) together yields

$$\begin{aligned} & D_{\text{snr}+\gamma}\left(P_{Y_0^T} \parallel Q_{Y_0^T}\right) - D_{\text{snr}}\left(P_{Y_0^T} \parallel Q_{Y_0^T}\right) \\ &= \frac{\gamma}{2} [\text{mse}_Q(\text{snr}) - \text{mmse}(\text{snr})] + o(\gamma) \end{aligned} \quad (169)$$

establishing (the differential version of) (29).

For establishing (30) via the time-incremental channel approach of [16], one needs to characterize the behavior of $D\left(P_{Y_t^{t+\delta}|Y_0^t} \parallel Q_{Y_t^{t+\delta}|Y_0^t} \middle| P_{Y_0^t}\right)$ as $\delta \rightarrow 0$. To this end, the “time-SNR” transform of [16, Subection III-D] can be performed to convert the channel of infinitesimal duration δ (in the interval $[t, t + \delta]$) to a channel of duration $[0, 1]$ and infinitesimal SNR. Applying (165) on the latter and “transforming back” yields (170), shown at the bottom of the page. By the chain rule, the left side of (170) equals $D\left(P_{Y_0^{t+\delta}} \parallel Q_{Y_0^{t+\delta}}\right) - D\left(P_{Y_0^t} \parallel Q_{Y_0^t}\right)$ and thus (170) can equivalently be stated as (171), shown at the bottom of the

page, which is nothing but the derivative of Equality (30) with respect to time.

The Presence of Feedback: An examination of our proof of (30) reveals that it carries over to accommodate the presence of feedback. Specifically, suppose that X_0^T and Y_0^T still satisfy the input-output relationship (1), but that, under P , X_t evolves according to $X_t = a_t(Y_0^{t-\delta}, R_0^T)$, where R_0^T is an additional source of randomness, independent of the noise W_0^T . Under Q , X_t may evolve according to $X_t = b_t(Y_0^{t-\delta}, R_0^T)$. When considering the piecewise constant approximations of P and Q that, as in (111), are constant on the intervals $(\frac{i-1}{M}T, \frac{i}{M}T]$, assuming that M is large enough that $T/M < \delta$, the induced discrete-time channel is again an AWGN channel similarly as in (112), but with feedback allowed. In other words, instead of a distribution on A^M , P and Q , combined with the channel, now induce a set of conditional distributions $\{P_{A_i, J_i|A^{i-1}, J^{i-1}}\}_{i=1}^M$ and $\{Q_{A_i, J_i|A^{i-1}, J^{i-1}}\}_{i=1}^M$, with a conditional independence structure $J_i - A_i - (A^{i-1}, J^{i-1})$, and satisfying $P_{J_i|A^{i-1}, J^{i-1}} = P_{A_i|A^{i-1}, J^{i-1}} * \mathcal{N}(0, \frac{M}{\text{snr}T})$ and $Q_{J_i|A^{i-1}, J^{i-1}} = Q_{A_i|A^{i-1}, J^{i-1}} * \mathcal{N}(0, \frac{M}{\text{snr}T})$. It is readily checked that the relationship $D(P_{JM} \parallel Q_{JM}) = D_{\text{snr}}(P_{Y_0^T} \parallel Q_{Y_0^T})$ carries over to this case (by verifying that the few steps leading to (116) carry over), along with all the equalities in the chain of equalities leading to (138).

On the other hand, letting $I_\gamma(X_0^T \rightarrow Y_0^T)$ denote the directed information between X_0^T and Y_0^T , as defined in [22], under the channel in (1) when X_t is evolving according to P , Theorem 2 of [22] (cf. also [18], [19]) gives

$$\text{cmmse}(\text{snr}) = \frac{2}{\text{snr}} I_{\text{snr}}(X_0^T \rightarrow Y_0^T) \quad (172)$$

which, when combined with (30), yields

$$\text{cmse}_Q(\text{snr}) = \frac{2}{\text{snr}} \left[I_{\text{snr}}(X_0^T \rightarrow Y_0^T) + D_{\text{snr}}\left(P_{Y_0^T} \parallel Q_{Y_0^T}\right) \right]. \quad (173)$$

Equality (173) can be considered an extension of Duncan’s theorem [9] that accommodates both mismatch and the presence of feedback.

$$D\left(P_{Y_{1,0}^T|Y_{2,0}^T} \parallel Q_{Y_{1,0}^T|Y_{2,0}^T} \middle| P_{Y_{2,0}^T}\right) = \frac{\gamma}{2} \left[\int_0^T E[(X_t - E_Q[X_t|Y_{2,0}^T])^2] dt - \int_0^T \text{Var}(X_t|Y_{2,0}^T) dt \right] + o(\gamma) \quad (166)$$

$$= \frac{\gamma}{2} [\text{mse}_Q(\text{snr}) - \text{mmse}(\text{snr})] + o(\gamma). \quad (167)$$

$$\begin{aligned} & D\left(P_{Y_t^{t+\delta}|Y_0^t} \parallel Q_{Y_t^{t+\delta}|Y_0^t} \middle| P_{Y_0^t}\right) \\ &= \delta \frac{\text{snr}}{2} \left\{ E\left[(X_t - E_Q[X_t|Y_0^t; \text{snr}])^2\right] - E\left[(X_t - E[X_t|Y_0^t])^2\right] \right\} + o(\delta) \end{aligned} \quad (170)$$

$$\frac{d}{dt} D\left(P_{Y_0^t} \parallel Q_{Y_0^t}\right) = \frac{\text{snr}}{2} \left\{ E\left[(X_t - E_Q[X_t|Y_0^t; \text{snr}])^2\right] - E\left[(X_t - E[X_t|Y_0^t])^2\right] \right\} \quad (171)$$

As for (29), it does not carry over to the presence of feedback, since neither does its finite dimensional origin: Lemma 1 fails to hold in the generality of “feedback distributions” of the form $\{P_{A_i|A^{i-1}, J^{i-1}}\}_{i=1}^M$ and $\{Q_{A_i|A^{i-1}, J^{i-1}}\}_{i=1}^M$, as is easily seen already for the case $M = 2$. Consequently, neither of the equalities in (27) carry over to the presence of feedback which, as pointed out in [16, Subsection V-B], is already the case in the nonmismatched setting.

APPENDIX
PROOF OF LEMMA 1

Let us introduce the notation $G(P, Q, \gamma)$ to denote the integrand in (110), i.e.,

$$G(P, Q, \gamma) = \frac{1}{2} [\text{mse}_Q^{\text{scalar}}(\gamma) - \text{mmse}^{\text{scalar}}(\gamma)] \quad (\text{A.1})$$

thus making the dependence also on P explicit. Note that by differentiating both sides of (110) with respect to $1/\sigma^2$ we obtain the differential version

$$\frac{d}{d\gamma} D \left(P * \mathcal{N}(0, 1/\gamma) \middle\| \middle\| Q * \mathcal{N}(0, 1/\gamma) \right) = G(P, Q, \gamma). \quad (\text{A.2})$$

It will suffice to prove

$$\begin{aligned} \frac{d}{d\gamma} D \left(P * \mathcal{N} \left(0, \frac{1}{\gamma} I \right) \middle\| \middle\| Q * \mathcal{N} \left(0, \frac{1}{\gamma} I \right) \right) \\ = \frac{1}{2} [\text{mse}_Q^{\text{vec}}(\gamma) - \text{mmse}^{\text{vec}}(\gamma)] \end{aligned} \quad (\text{A.3})$$

from which (109) would follow since

$$\lim_{\gamma \downarrow 0} D \left(P * \mathcal{N} \left(0, \frac{1}{\gamma} I \right) \middle\| \middle\| Q * \mathcal{N} \left(0, \frac{1}{\gamma} I \right) \right) = 0. \quad (\text{A.4})$$

To this end, for $\gamma^M \in (0, \infty)^M$, let

$$f(\gamma^M) = D \left(P * \mathcal{N} \left(0, \frac{1}{\gamma^M} I \right) \middle\| \middle\| Q * \mathcal{N} \left(0, \frac{1}{\gamma^M} I \right) \right) \quad (\text{A.5})$$

where $\frac{1}{\gamma^M} I$ denotes the diagonal matrix whose i th diagonal term is $\frac{1}{\gamma_i}$. Let P^{γ^M} and Q^{γ^M} denote distributions under an M -dimensional vector observation Y^M given by

$$Y_i = X_i + \frac{1}{\sqrt{\gamma_i}} B_i \quad (\text{A.6})$$

where $B^M \sim \mathcal{N}(0, I)$ is independent of X^M , which is distributed, respectively, as P and Q . Using the notation $Y^{M \setminus i} = (Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_M)$, we note that

$$f(\gamma^M) = D \left(P_{Y^M}^{\gamma^M} \middle\| \middle\| Q_{Y^M}^{\gamma^M} \right) \quad (\text{A.7})$$

$$\begin{aligned} &= D \left(P_{Y^{M \setminus i}}^{\gamma^M} \middle\| \middle\| Q_{Y^{M \setminus i}}^{\gamma^M} \right) \\ &\quad + D \left(P_{Y_i|Y^{M \setminus i}}^{\gamma^M} \middle\| \middle\| Q_{Y_i|Y^{M \setminus i}}^{\gamma^M} \middle| P_{Y^{M \setminus i}}^{\gamma^M} \right) \end{aligned} \quad (\text{A.8})$$

where the first equality is immediate from the definition of f , P^{γ^M} and Q^{γ^M} , and the second equality follows from the chain rule for relative entropy. Consider now (A.9)–(A.14), shown at the bottom of the page, where:

- (a) follows from (A.8) upon noting that $P_{Y^{M \setminus i}}^{\gamma^M}$ and $Q_{Y^{M \setminus i}}^{\gamma^M}$ do not depend on γ_i and, hence, neither does the first term in (A.8).
- (b) follows from the definitions of P^{γ^M} and Q^{γ^M} which imply that $P_{Y_i|Y^{M \setminus i}}^{\gamma^M} = P_{X_i|Y^{M \setminus i}}^{\gamma^M} * \mathcal{N} \left(0, \frac{1}{\gamma_i} \right)$ and $Q_{Y_i|Y^{M \setminus i}}^{\gamma^M} = Q_{X_i|Y^{M \setminus i}}^{\gamma^M} * \mathcal{N} \left(0, \frac{1}{\gamma_i} \right)$, and from the fact that $P_{Y^{M \setminus i}}^{\gamma^M}$ does not depend on γ_i .
- (c) is an application of (A.2).
- (d) follows from the definition of $G(P, Q, \gamma)$ [recall (A.1)].

Thus

$$\frac{d}{d\gamma} D \left(P * \mathcal{N} \left(0, \frac{1}{\gamma} I \right) \middle\| \middle\| Q * \mathcal{N} \left(0, \frac{1}{\gamma} I \right) \right) \quad (\text{A.15})$$

$$= \frac{d}{d\gamma} f(\gamma, \dots, \gamma) \quad (\text{A.16})$$

$$= \sum_{i=1}^M \frac{\partial}{\partial \gamma_i} f(\gamma^M) \Big|_{\gamma^M = (\gamma, \dots, \gamma)} \quad (\text{A.17})$$

$$= \sum_{i=1}^M \frac{1}{2} E[(X_i - E_Q[X_i|Y^M; \gamma])^2 - (X_i - E[X_i|Y^M; \gamma])^2] \quad (\text{A.18})$$

$$= \frac{1}{2} [\text{mse}_Q^{\text{vec}}(\gamma) - \text{mmse}^{\text{vec}}(\gamma)] \quad (\text{A.19})$$

where the equality before last follows by substituting (A.14) for each of the summands in (A.17). This proves (A.3). \square

$$\frac{\partial}{\partial \gamma_i} f(\gamma^M) \stackrel{(a)}{=} \frac{\partial}{\partial \gamma_i} D \left(P_{Y_i|Y^{M \setminus i}}^{\gamma^M} \middle\| \middle\| Q_{Y_i|Y^{M \setminus i}}^{\gamma^M} \middle| P_{Y^{M \setminus i}}^{\gamma^M} \right) \quad (\text{A.9})$$

$$= \frac{\partial}{\partial \gamma_i} \int D \left(P_{Y_i|y^{M \setminus i}}^{\gamma^M} \middle\| \middle\| Q_{Y_i|y^{M \setminus i}}^{\gamma^M} \right) dP_{Y^{M \setminus i}}^{\gamma^M}(y^{M \setminus i}) \quad (\text{A.10})$$

$$\stackrel{(b)}{=} \int \frac{\partial}{\partial \gamma_i} D \left(P_{X_i|y^{M \setminus i}}^{\gamma^M} * \mathcal{N} \left(0, \frac{1}{\gamma_i} \right) \middle\| \middle\| Q_{X_i|y^{M \setminus i}}^{\gamma^M} * \mathcal{N} \left(0, \frac{1}{\gamma_i} \right) \right) dP_{Y^{M \setminus i}}^{\gamma^M}(y^{M \setminus i}) \quad (\text{A.11})$$

$$\stackrel{(c)}{=} \int G \left(P_{X_i|y^{M \setminus i}}^{\gamma^M}, Q_{X_i|y^{M \setminus i}}^{\gamma^M}, \gamma_i \right) dP_{Y^{M \setminus i}}^{\gamma^M}(y^{M \setminus i}) \quad (\text{A.12})$$

$$\stackrel{(d)}{=} \int \frac{1}{2} E \left[(X_i - E_Q[X_i|Y^M; \gamma^M])^2 - (X_i - E[X_i|Y^M; \gamma^M])^2 \middle| y^{M \setminus i}; \gamma^M \right] dP_{Y^{M \setminus i}}^{\gamma^M}(y^{M \setminus i}) \quad (\text{A.13})$$

$$= \frac{1}{2} E \left[(X_i - E_Q[X_i|Y^M; \gamma^M])^2 - (X_i - E[X_i|Y^M; \gamma^M])^2 \right] \quad (\text{A.14})$$

ACKNOWLEDGMENT

The author would like to thank A. Barron, P. Chigansky, E. Mayer-Wolf, H. Permuter, and S. Verdú for inspiring discussions.

REFERENCES

- [1] S. Artstein, K. Ball, F. Barthe, and A. Naor, "Solution of Shannon's problem on the monotonicity of entropy," *J. Amer. Math. Soc.*, vol. 17, pp. 975–982, 2004.
- [2] A. R. Barron, "Entropy and the central limit theorem," *Ann. Probab.*, vol. 14, pp. 336–342, 1986.
- [3] J. Binia, "Divergence and minimum mean-square error in continuous-time additive white Gaussian noise channels," *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 1160–1163, Mar. 2006.
- [4] R. H. Cameron and W. T. Martin, "Transformation of Wiener integrals under translations," *Ann. Math.*, vol. 45, pp. 386–396, 1944.
- [5] A. Cohen, N. Merhav, and T. Weissman, "Scanning and sequential decision making for multidimensional data, part II: noisy data," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 5609–5631, Dec. 2008.
- [6] I. Csiszár and J. Körner, *Information Theory: Coding theorems for discrete memoryless systems*. New York: Academic, 1981.
- [7] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed. New York: Springer, 1998.
- [8] T. E. Duncan, "Evaluation of likelihood functions," *Inf. Contr.*, vol. 13, pp. 62–74, 1968.
- [9] T. E. Duncan, "On the calculation of mutual information," *SIAM J. Appl. Math.*, vol. 19, pp. 215–220, Jul. 1970.
- [10] T. E. Duncan, "Mutual information for stochastic signals and Lévy processes," *IEEE Trans. Inf. Theory*, vol. IT-56, no. 1, pp. 18–24, Jan. 2010.
- [11] T. E. Duncan, "Mutual information for stochastic signals and fractional Brownian motion," *IEEE Trans. Inf. Theory*, vol. IT-54, no. 10, pp. 4432–4438, Oct. 2010.
- [12] R. G. Gallager, "Source coding with side information and universal coding," presented at the M.I.T. LIDS-P-937, 1979.
- [13] L. Gammaitoni, P. Hanggi, P. Jung, and F. Marchesoni, "Stochastic resonance," *Rev. Mod. Phys.*, vol. 70, pp. 223–287, 1998.
- [14] I. V. Girsanov, "On transforming a certain class of stochastic processes by absolutely continuous substitution of measures," *Theory Probab. Appl.*, vol. 5, pp. 285–301, 1960.
- [15] D. Guo, "Relative entropy and score function: New information-estimation relationships through arbitrary additive perturbation," presented at the IEEE Int. Symp. Information Theory, Seoul, Korea, Jul. 3, 2009.
- [16] D. Guo, S. Shamai, and S. Verdú, "Mutual information and minimum mean-square error in Gaussian channels," *IEEE Trans. Inf. Theory*, vol. IT-51, no. 4, pp. 1261–1283, Apr. 2005.
- [17] D. Guo, S. Shamai, and S. Verdú, "Mutual information and conditional mean estimation in Poisson channels," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 1837–1849, May 2008.
- [18] T. T. Kadota, M. Zakai, and J. Ziv, "Capacity of a continuous memoryless channel with feedback," *IEEE Trans. Inf. Theory*, vol. IT-17, pp. 372–378, 1971.
- [19] T. T. Kadota, M. Zakai, and J. Ziv, "Mutual information of the white Gaussian channel with and without feedback," *Trans. Inf. Theory*, vol. IT-17, pp. 368–371, 1971.
- [20] T. Kailath, "The structure of Radon–Nykodim derivatives with respect to Wiener and related measures," *Ann. Math. Statist.*, vol. 42, no. 3, pp. 1054–1067, 1971.
- [21] I. Karatzas and A. E. Shreve, *Brownian Motion and Stochastic Calculus*, 2nd ed. New York: Springer-Verlag, 1991.
- [22] Y. H. Kim, H. H. Permuter, and T. Weissman, "Directed information and causal estimation in continuous time," presented at the Int. Symp. Inf. Theory, Seoul, Korea, Jul. 29th, 2009.
- [23] A. Kolmogorov, "On the Shannon theory of information transmission in the case of continuous signals," *IRE Trans. Inf. Theory*, vol. 2, no. 4, pp. 102–108, Dec. 1956.
- [24] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, 2nd ed. New York: Springer, 1998.
- [25] R. S. Liptser and A. N. Shiryaev, *Statistics of Random Processes II: Applications*, 2nd ed. New York: Springer, 2001.
- [26] A. Lozano, A. M. Tulino, and S. Verdú, "Optimum power allocation for parallel Gaussian channels with arbitrary input distributions," *IEEE Trans. Inf. Theory*, vol. 52, no. 7, pp. 3033–3051, Jul. 2006.
- [27] S. Orey, "Conditions for the absolute continuity of two diffusions," *Trans. Amer. Math. Soc.*, vol. 193, pp. 413–426, 1974.
- [28] M. Madiman and A. R. Barron, "Generalized entropy power inequalities and monotonicity properties of information," *IEEE Trans. Inf. Theory*, vol. 53, no. 7, pp. 2317–2329, Jul. 2007.
- [29] E. Mayer-Wolf and M. Zakai, "Some relations between mutual information and estimation error in Wiener space," *Ann. Appl. Probab.*, vol. 17, no. 3, pp. 1102–1116, Jun. 2007.
- [30] N. Merhav and M. Feder, "A strong version of the redundancy-capacity theorem of universal coding," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 714–722, May 1995.
- [31] D. P. Palomar and S. Verdú, "Gradient of mutual information in linear vector Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 141–154, Jan. 2006.
- [32] D. P. Palomar and S. Verdú, "Representation of mutual information via input estimates," *IEEE Trans. Inf. Theory*, vol. 53, no. 2, pp. 453–470, Feb. 2007.
- [33] M. S. Pinsker, "Information and information stability of random variables and processes," presented at the Moskva: Izv. Akad. Nauk, Russian, 1960.
- [34] B. Y. Ryabko, "Encoding a source with unknown but ordered probabilities," *Probl. Inf. Transm.*, pp. 134–139, Oct. 1979.
- [35] A. Stam, "Some inequalities satisfied by the quantities of information of Fisher and Shannon," *Inf. Control*, vol. 2, pp. 101–112, Jun. 1959.
- [36] A. M. Tulino and S. Verdú, "Monotonic decrease of the non-Gaussianity of the sum of independent random variables: A simple proof," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 4295–4297, Sep. 2006.
- [37] S. Verdú, "Mismatched estimation and relative entropy," presented at the Int. Symp. Information Theory, Seoul, Korea, Jul. 29, 2009.
- [38] S. Verdú and D. Guo, "A simple proof of the entropy power inequality," *IEEE Trans. Inf. Theory*, vol. 52, no. 5, pp. 2165–2166, May 2006.
- [39] A. D. Wyner, "A definition of conditional mutual information for arbitrary ensembles," *Inf. Control*, vol. 38, pp. 51–59, 1978.
- [40] M. Zakai, "On mutual information, likelihood ratios, and estimation error for the additive Gaussian channel," *IEEE Trans. Inf. Theory*, vol. 51, no. 9, pp. 3017–3024, Sep. 2005.

Tsachy Weissman (S'99–M'02–SM'07) received the Ph.D. degree in electrical engineering from the Technion—Israel Institute of Technology, Haifa, in 2001.

During 2002–2003, he was with the Information Theory Research Group, Hewlett-Packard Laboratories. He has been on the faculty of the Electrical Engineering Department, Stanford University, Stanford, CA, since 2003, spending the two academic years 2007–2009 with the Technion Department of Electrical Engineering. His research is focused on information theory, statistical signal processing, the interplay between them, and their applications. He is the inventor of several patents and involved in a number of high-tech companies as a researcher or member of the technical board.

Dr. Weissman received a joint IT/COM Societies Best Paper, a Horev Fellowship for Leaders in Science and Technology, and a Henry Taub Prize for Excellence in Research.