

# Conservation Laws for Mutual Information

Supplemental Material for Information Theory  
Henry D. Pfister

November 12th, 2024

## 1 Introduction

EXtrinsic Information Transfer (EXIT) charts were introduced by ten Brink in 1999 as a useful tool to understand the convergence of Turbo decoding for different component codes [1]. His work led to the EXIT area theorem for erasure channels which was put on rigorous mathematical footing by Ashikhmin, Kramer, and ten Brink in 2004 [2]. A little later, this was generalized to the Gaussian channel by two groups: Guo, Shamai, and Verdú [3, 4] and Measson, Montanari, Richardson, and Urbanke [5, 6]. Measson, Montanari, and Urbanke also showed that the area theorem also allows one to upper bound the MAP decoding threshold using information from iterative decoding [7, 8]. Together these ideas highlight some fundamental connections between iterative information processing and optimal information processing.

## 2 EXIT Functions

Let  $\mathcal{X}$  be a finite alphabet and let  $\underline{X} = (X_1, \dots, X_n) \in \mathcal{X}^n$  be a random vector chosen according to  $P_{\underline{X}}(\underline{x})$ . Let  $\mathcal{Y} = \mathcal{X} \cup \{?\}$  and suppose  $Y_i \in \mathcal{X} \cup \{?\}$  is an observation of  $X_i$  through an erasure channel with erasure probability  $\epsilon_i$  so that  $\underline{Y} = (Y_1, \dots, Y_n) \in \mathcal{Y}^n$  is the channel output vector. The notation  $\underline{Y}_{\sim i}$  will be used to denote the vector  $(Y_1, \dots, Y_{i-1}, ?, Y_{i+1}, \dots, Y_n)$  and the notation  $\underline{Y}(\underline{\epsilon}) = (Y_1(\epsilon_1), \dots, Y_n(\epsilon_n))$  will be used to emphasize the dependence on  $\underline{\epsilon} = (\epsilon_1, \dots, \epsilon_n)$ . Some notation used below includes:  $[n] \triangleq \{1, 2, \dots, n\}$

**Definition 1.** The **EXIT function for the  $i$ -th variable** of  $\underline{X}$  is defined to be

$$h_i(\underline{\epsilon}) \triangleq H(X_i | \underline{Y}_{\sim i}(\underline{\epsilon}_{\sim i})).$$

If all  $\epsilon_i = \epsilon$  for  $i \in [n]$ , then the simplified notation  $h_i(\epsilon) \triangleq h_i((\epsilon, \dots, \epsilon)) = H(X_i | \underline{Y}_{\sim i}(\epsilon))$  is used. Similarly, the **average EXIT function** of  $\underline{X}$  is defined to be  $h(\underline{\epsilon}) = \frac{1}{n} \sum_{i=1}^n h_i(\underline{\epsilon})$  in the first case and  $h(\epsilon) = \frac{1}{n} \sum_{i=1}^n h_i(\epsilon)$  in the second. For all  $i, j \in [n]$ ,  $h_i(\underline{\epsilon})$  is non-decreasing in  $\epsilon_j$  by the data processing inequality because changing  $Y_j$  to an erasure can only make it harder to predict  $X_i$ . It follows that  $h(\epsilon)$  is non-decreasing in  $\epsilon$ .

**Lemma 2.** *Using the above setup, the EXIT function for the  $i$ -th variable of  $\underline{X}$  satisfies*

$$h_i(\underline{\epsilon}) = \frac{d}{d\epsilon_i} H(\underline{X} | \underline{Y}(\underline{\epsilon})).$$

*Proof.* Suppressing the explicit dependence on  $\underline{\epsilon}$ , this follows from

$$\begin{aligned}
\frac{d}{d\epsilon_i} H(\underline{X}|\underline{Y}) &= \frac{d}{d\epsilon_i} [H(X_i|\underline{Y}) + H(\underline{X}_{\sim i}|X_i, \underline{Y})] && (H \text{ chain rule}) \\
&= \frac{d}{d\epsilon_i} H(X_i|\underline{Y}) + \frac{d}{d\epsilon_i} \underbrace{H(\underline{X}_{\sim i}|X_i, \underline{Y}_{\sim i})}_{\text{ind. of } \epsilon_i} && (\underline{X}_{\sim i} \rightarrow X_i \rightarrow Y_i \text{ Markov chain}) \\
&= \frac{d}{d\epsilon_i} \left[ \mathbb{P}(Y_i = ?) H(X_i|\underline{Y}, Y_i = ?) + \mathbb{P}(Y_i \neq ?) \underbrace{H(X_i|\underline{Y}, Y_i \neq ?)}_{Y_i \neq ? \Rightarrow Y_i = X_i \Rightarrow H=0} \right] && (\text{Average over } Y_i) \\
&= \frac{d}{d\epsilon_i} \epsilon_i H(X_i|\underline{Y}_{\sim i}) = H(X_i|\underline{Y}_{\sim i}). && (\mathbb{P}(Y_i = ?) = \epsilon_i)
\end{aligned}$$

□

**Lemma 3.** Using the above setup, let  $H(\underline{X}|\underline{Y}(\underline{\epsilon}(t)))$  denote the conditional entropy evaluated along the erasure channel path  $\underline{\epsilon}(t) = (\epsilon_1(t), \dots, \epsilon_n(t))$  for  $t \in [0, 1]$ . Then,

$$H(\underline{X}|\underline{Y}(\underline{\epsilon}(1))) - H(\underline{X}|\underline{Y}(\underline{\epsilon}(0))) = \int_0^1 \underline{h}(\underline{\epsilon}(t)) \cdot \underline{\epsilon}'(t) dt = \int_0^1 \left( \sum_{i=1}^n h_i(\underline{\epsilon}(t)) \epsilon_i'(t) \right) dt,$$

where  $\underline{h}(\underline{\epsilon}) = (h_1(\underline{\epsilon}), \dots, h_n(\underline{\epsilon}))$ . If the erasure channel path satisfies  $\epsilon_i(t) = t$  for  $i \in [n]$ , then we find that

$$H(\underline{X}|\underline{Y}(\epsilon(1))) - H(\underline{X}|\underline{Y}(\epsilon(0))) = \int_0^1 \left( \sum_{i=1}^n h_i(\epsilon(t)) \epsilon_i'(t) \right) dt = n \int_0^1 h(t) dt.$$

*Proof.* These results follow directly from Lemma 2 and vector calculus. □

**Example 4.** Consider the case where  $\underline{X}$  is supported on the non-linear code  $\mathcal{C} = \{00, 10, 11\}$  and the codewords are chosen, respectively, with probabilities  $\frac{1}{2}, \frac{1}{4}, \frac{1}{4}$ . In this case,  $H(\underline{X}) = \frac{3}{2}$  and it is easy to verify that

$$\begin{aligned}
H(X_1|Y_2(\epsilon_2)) &= \epsilon_2 + \frac{3}{4}(1 - \epsilon_2)H_b(\frac{1}{3}) \\
H(X_2|Y_1(\epsilon_1)) &= \epsilon_1 H_b(\frac{1}{4}) + \frac{1}{2}(1 - \epsilon_1),
\end{aligned}$$

where  $H_b(x) = x \log_2 \frac{1}{x} + (1 - x) \log_2 \frac{1}{1-x}$  is the binary entropy function. Integrating gives

$$\begin{aligned}
\int_0^1 [H(X_1|Y_2(\epsilon)) + H(X_2|Y_1(\epsilon_1))] d\epsilon &= \int_0^1 \left[ \epsilon + \frac{3}{4}(1 - \epsilon)H_b(\frac{1}{3}) + \epsilon H_b(\frac{1}{4}) + \frac{1}{2}(1 - \epsilon) \right] d\epsilon \\
&= \left( \frac{1}{2} + \frac{1}{4} \right) + \left( \frac{3}{8}H_b(\frac{1}{3}) + \frac{1}{2}H_b(\frac{1}{4}) \right) \\
&= \frac{3}{4} + \frac{3}{4} = \frac{3}{2}.
\end{aligned}$$

One can also verify, either directly or via differentiation, that

$$H(\underline{X}|\underline{Y}(\underline{\epsilon})) = \epsilon_1 \epsilon_2 \frac{3}{2} + \frac{1}{2}(1 - \epsilon_1)\epsilon_2 + \frac{3}{4}\epsilon_1(1 - \epsilon_2)h(\frac{1}{3}).$$

## 2.1 Connection to the Chain Rule

Let  $\mathbb{S}_n$  be the symmetric group on  $n$  letters where each  $\pi \in \mathbb{S}_n$  is a permutation  $\pi: [n] \rightarrow [n]$  and the group operation is functional composition. Then, for any  $\pi \in \mathbb{S}_n$ , the chain rule implies that

$$H(\underline{X}) = \sum_{i=1}^n H(X_{\pi(i)}|X_{\pi(1)}, X_{\pi(2)}, \dots, X_{\pi(i-1)}).$$

The area theorem was originally motivated by the idea of averaging this quantity over all permutations. To see the connection, we can compute this average while grouping terms with  $\pi(i) = j$  to get

$$\begin{aligned}
H(\underline{X}) &= \frac{1}{n!} \sum_{\pi \in \mathbb{S}_n} \sum_{i=1}^n H(X_{\pi(i)} | X_{\pi(1)}, X_{\pi(2)}, \dots, X_{\pi(i-1)}) \\
&= \sum_{i=1}^n \frac{1}{n!} \sum_{j=1}^n \sum_{\pi \in \mathbb{S}_n: \pi(i)=j} H(X_{\pi(i)} | X_{\pi(1)}, X_{\pi(2)}, \dots, X_{\pi(i-1)}) \\
&= \sum_{i=1}^n \frac{1}{n} \sum_{j=1}^n \sum_{\pi \in \mathbb{S}_n: \pi(i)=j} \frac{1}{(n-1)!} H(X_j | X_{\pi(1)}, X_{\pi(2)}, \dots, X_{\pi(i-1)}) \\
&\approx \sum_{i=1}^n \frac{1}{n} \sum_{j=1}^n H\left(X_j | \underline{Y}_{\sim j} \left(\frac{i-1}{n-1}\right)\right) \\
&\approx \int_0^1 \sum_{j=1}^n H(X_j | \underline{Y}_{\sim j}(\epsilon)) d\epsilon.
\end{aligned}$$

## 2.2 Binary Linear Codes

Now, let  $\mathcal{C}$  be an  $(n, k)$  binary linear code with generator matrix  $G$  and parity-check matrix  $H$ . Assume that a random codeword  $\underline{X}$  is chosen uniformly and transmitted through BECs with erasure probabilities  $\underline{\epsilon} = (\epsilon_1, \dots, \epsilon_n)$  to get the output  $\underline{Y} \in \{0, 1, ?\}^n$ . Then, for the exit function  $h(\epsilon) = \frac{1}{n} \sum_{i=1}^n H(X_i | \underline{Y}_{\sim i}(\epsilon))$ , the exit area theorem implies that

$$\int_0^1 h(\epsilon) d\epsilon = \frac{1}{n} H(\underline{X}) = R.$$

This result has strong implications for sequences of codes achieving capacity. Let  $\mathcal{C}_j$  be an  $(n_j, k_j)$  binary linear code with rate  $R_j \rightarrow R$  and EXIT function  $h^{(j)}(\epsilon)$ . The BEC threshold of a code sequence is defined to be

$$\epsilon^* = \sup \left\{ \epsilon \in [0, 1] \mid \limsup_{j \rightarrow \infty} h^{(j)}(\epsilon) = 0 \right\}.$$

By definition, if this code sequence achieves capacity, then  $\epsilon^* = 1 - R$  and  $h^{(j)}(\epsilon) \rightarrow 0$  for all  $\epsilon < \epsilon^*$ . Since EXIT functions are non-decreasing, this implies that  $h^{(j)}(\epsilon) \rightarrow 1$  for all  $\epsilon > \epsilon^*$  in order to satisfy the area theorem. Thus, for any capacity achieving sequence, the EXIT function sequence must converge to a step function that jumps from 0 to 1 at  $\epsilon = 1 - R$ .

## 2.3 Binary Linear Codes and Duality

Using the setup of the previous section, we now consider erasure decoding in more detail. Let  $\mathcal{E}(\underline{y}) \triangleq \{i \in [n] \mid y_i = ?\}$  be set of indices where an erasure occurs. For a set  $\mathcal{E} = (e_1, e_2, \dots, e_{|\mathcal{E}|})$  with  $e_1 < e_2 < \dots < e_{|\mathcal{E}|}$  and an  $m \times n$  matrix  $A = (\underline{a}_1, \underline{a}_2, \dots, \underline{a}_n)$  whose  $i$ -th column is  $\underline{a}_i$ , we let  $A_{\mathcal{E}} = (\underline{a}_{e_1}, \underline{a}_{e_2}, \dots, \underline{a}_{e_{|\mathcal{E}|}})$ . The same rule can be applied to row vectors by choosing  $m = 1$ .

Using this notation, the a posteriori probability (APP) distribution for  $\underline{X}$  given  $\underline{Y}$  is

$$P_{\underline{X}}(\underline{x} | \underline{y}) = \begin{cases} \frac{1}{|V(\underline{y})|} & \text{if } \underline{x} \in V(\underline{y}) \\ 0 & \text{otherwise,} \end{cases}$$

where  $V(\underline{y}) = \left\{ \underline{z} \in \mathcal{C} \mid \underline{z}_{\mathcal{E}^c}(\underline{y}) = \underline{y}_{\mathcal{E}^c}(\underline{y}) \right\}$  is set of codewords that are compatible with the observations. Since  $\mathcal{C}$  is linear, the set  $V(\underline{y})$  is the affine subspace of  $\underline{x} \in \{0, 1\}^n$  satisfying

$$H_{\mathcal{E}} \underline{x}_{\mathcal{E}}^T = H_{\mathcal{E}^c} \underline{y}_{\mathcal{E}^c}^T,$$

where  $\mathcal{E}(y)$  is denoted by  $\mathcal{E}$  for simplicity and  $\underline{y}_{\mathcal{E}^c}$  is a binary vector known by the decoder. Thus, dimension of the solution space is given by  $|\mathcal{E}| - \text{rank}(H_{\mathcal{E}})$ . Similarly, affine subspace of input vectors  $\underline{u} \in \{0, 1\}^k$  compatible with  $\underline{y}$  is defined by

$$\underline{u}G_{\mathcal{E}^c} = \underline{y}_{\mathcal{E}^c}$$

and dimension of the solution space is  $k - \text{rank}(G_{\mathcal{E}^c})$ . Of course, the two spaces must have the same dimension and this implies that

$$k - \text{rank}(G_{\mathcal{E}^c}) = |\mathcal{E}| - \text{rank}(H_{\mathcal{E}}).$$

As the input distribution is uniform over  $\mathcal{C}$ , it follows that these unknown dimensions have full entropy and

$$H(\underline{X}|\underline{Y} = y) = |\mathcal{E}| - \text{rank}(H_{\mathcal{E}}) = k - \text{rank}(G_{\mathcal{E}^c}).$$

**Lemma 5.** *Using the above setup, the conditional entropy  $H(\underline{X}|\underline{Y}(\underline{\epsilon}))$  is given by*

$$\begin{aligned} H(\underline{X}|\underline{Y}(\underline{\epsilon})) &= k - \sum_{\mathcal{E} \subseteq [n]} \left( \prod_{i \in \mathcal{E}} \epsilon_i \right) \left( \prod_{i \in \mathcal{E}^c} (1 - \epsilon_i) \right) \text{rank}(G_{\mathcal{E}^c}) \\ &= \sum_{i=1}^n \epsilon_i - \sum_{\mathcal{E} \subseteq [n]} \left( \prod_{i \in \mathcal{E}} \epsilon_i \right) \left( \prod_{i \in \mathcal{E}^c} (1 - \epsilon_i) \right) \text{rank}(H_{\mathcal{E}}). \end{aligned}$$

Let  $H^\perp(\underline{X}|\underline{Y}(\underline{\epsilon}))$  denote the conditional entropy when  $\underline{X}$  is chosen uniformly from the dual code  $\mathcal{C}^\perp$ . Then,

$$H^\perp(\underline{X}|\underline{Y}(\underline{\epsilon})) = H(\underline{X}|\underline{Y}(\underline{1} - \underline{\epsilon})) - k + \sum_{i=1}^n \epsilon_i$$

and computing the derivative with respect to  $\epsilon_i$  shows that

$$h_i^\perp(\underline{\epsilon}) = 1 - h_i(\underline{1} - \underline{\epsilon}).$$

*Proof.* The first formula follows from averaging  $H(\underline{X}|\underline{Y} = y) = k - \text{rank}(G_{\mathcal{E}^c})$  over all all possible erasure patterns because the formula depends only on the erasure pattern and not on the unerased values. The second formula follows from averaging  $H(\underline{X}|\underline{Y} = y) = |\mathcal{E}| - \text{rank}(H_{\mathcal{E}})$  over all all possible erasure patterns. In this case, the expectation of  $|\mathcal{E}|$  is computed using

$$\begin{aligned} \sum_{\mathcal{E} \subseteq [n]} \left( \prod_{i \in \mathcal{E}} \epsilon_i \right) \left( \prod_{i \in \mathcal{E}^c} (1 - \epsilon_i) \right) |\mathcal{E}| &= \sum_{\mathcal{E} \subseteq [n]} \left( \prod_{i \in \mathcal{E}} \epsilon_i \right) \left( \prod_{i \in \mathcal{E}^c} (1 - \epsilon_i) \right) \sum_{j=1}^n \mathbf{1}_{\mathcal{E}}(j) \\ &= \sum_{j=1}^n \sum_{\mathcal{E} \subseteq [n]} \left( \prod_{i \in \mathcal{E}} \epsilon_i \right) \left( \prod_{i \in \mathcal{E}^c} (1 - \epsilon_i) \right) \mathbf{1}_{\mathcal{E}}(j) \\ &= \sum_{j=1}^n \epsilon_j. \end{aligned}$$

For the dual code, we note that

$$\begin{aligned} H^\perp(\underline{X}|\underline{Y}(\underline{\epsilon})) &= \sum_{i=1}^n \epsilon_i - \sum_{\mathcal{E} \subseteq [n]} \left( \prod_{i \in \mathcal{E}} \epsilon_i \right) \left( \prod_{i \in \mathcal{E}^c} (1 - \epsilon_i) \right) \text{rank}(H_{\mathcal{E}}^\perp) && \text{(Definition of } H^\perp(\underline{X}|\underline{Y}(\underline{\epsilon}))) \\ &= \sum_{i=1}^n \epsilon_i - \sum_{\mathcal{E} \subseteq [n]} \left( \prod_{i \in \mathcal{E}} \epsilon_i \right) \left( \prod_{i \in \mathcal{E}^c} (1 - \epsilon_i) \right) \text{rank}(G_{\mathcal{E}}) && (H^\perp = G) \\ &= \sum_{i=1}^n \epsilon_i - \sum_{\mathcal{E} \subseteq [n]} \left( \prod_{i \in \mathcal{E}} \epsilon_i \right) \left( \prod_{i \in \mathcal{E}^c} (1 - \epsilon_i) \right) \text{rank}(G_{\mathcal{E}^c}) && (\mathcal{E}\text{-sum invariant: } \mathcal{E} \mapsto \mathcal{E}^c) \\ &= \left( \sum_{i=1}^n \epsilon_i \right) - k + H(\underline{X}|\underline{Y}(\underline{1} - \underline{\epsilon})). && \text{(Definition of } H(\underline{X}|\underline{Y}(\underline{1} - \underline{\epsilon}))) \end{aligned}$$

Taking the derivative with  $\epsilon_i$  gives

$$\begin{aligned}
h_i^\perp(\underline{\epsilon}) &= \frac{d}{d\epsilon_i} H^\perp(\underline{X}|\underline{Y}(\underline{\epsilon})) \\
&= \frac{d}{d\epsilon_i} \left[ \left( \sum_{i=1}^n \epsilon_i \right) - k + H(\underline{X}|\underline{Y}(\underline{1} - \underline{\epsilon})) \right] \\
&= 1 - \frac{d}{d\epsilon_i} H(\underline{X}|\underline{Y}(\underline{1} - \underline{\epsilon})) \\
&= 1 - h_i(\underline{1} - \underline{\epsilon}).
\end{aligned}$$

This completes the proof.  $\square$

### 3 Mutual Information and Minimum Mean-Squared Error

Let  $X$  be a real random variable with bounded second moment (i.e.,  $\mathbb{E}[X^2] < \infty$ ) and  $Y$  be a noisy observation of  $X$ . The minimum mean-squared error (MMSE) of  $X$  given  $Y$  is given by

$$\text{mmse}(X|Y) \triangleq \mathbb{E} \left[ (X - \mathbb{E}[X|Y])^2 \right].$$

Now, we describe some general properties of estimation in Gaussian noise. Let  $Y_s = \sqrt{s}X + Z$  where  $Z \sim \mathcal{N}(0, 1)$  is standard Gaussian independent of  $X$ . Consider two hypotheses:  $H_0$  where  $Y = Y_0 = Z$  and  $H_1$  where  $Y = Y_1$ . Then, the log-likelihood ratio between these two hypotheses is given by

$$\begin{aligned}
L(y) &= \ln \frac{f_{Y_1}(y)}{f_Z(y)} \\
&= \ln \frac{\int f_X(x) e^{-(y-x)^2/2} dx}{e^{-y^2/2}} \\
&= \ln \int f_X(x) e^{yx-x^2/2} dx.
\end{aligned}$$

The following lemma shows that the derivatives of  $L(y)$  encode important information about the implied estimation problem.

**Lemma 6.** *For the given setup, we have*

$$\begin{aligned}
L'(y) &= \mathbb{E}[X|Y = y] \\
L''(y) &= \text{Var}(X|Y = y).
\end{aligned}$$

*Proof.* Computing directly, we recover the result from [9] that

$$\begin{aligned}
L'(y) &= \frac{d}{dy} \ln \int f_X(x) e^{yx-x^2/2} dx \\
&= \frac{\int f_X(x) x e^{yx-x^2/2} dx}{\int f_X(x) e^{yx-x^2/2} dx} \\
&= \mathbb{E}[X|Y = y].
\end{aligned}$$

Similarly, computing the 2nd derivative gives the result from [10] that

$$\begin{aligned}
L''(y) &= \frac{d}{dy} \frac{\int f_X(x) x e^{yx-x^2/2} dx}{\int f_X(x) e^{yx-x^2/2} dx} \\
&= \frac{\int f_X(x) x^2 e^{yx-x^2/2} dx}{\int f_X(x) e^{yx-x^2/2} dx} - \frac{\int f_X(x) x e^{yx-x^2/2} dx}{\int f_X(x) e^{yx-x^2/2} dx} \cdot \frac{\int f_X(x) x e^{yx-x^2/2} dx}{\int f_X(x) e^{yx-x^2/2} dx} \\
&= \mathbb{E}[X^2|Y = y] - \mathbb{E}[X|Y = y]^2 \\
&= \text{Var}(X|Y = y).
\end{aligned}$$

$\square$

The following theorem from [4] can be seen as a generalization of the area theorem to Gaussian noise. It relates the  $s$ -derivative of the mutual information  $I(X; Y_s)$  with the MMSE  $\text{mmse}(X|Y_s)$ .

**Theorem 7.** *The derivative of  $I(X; Y_s)$  with  $s$  is given by*

$$\frac{d}{ds} I(X; Y_s) = \frac{1}{2} \text{mmse}(X|Y_s).$$

*Proof.* To lighten notation, we will treat  $s$  as fixed and write  $Y$  instead of  $Y_s$ . First, we observe that the conditional distribution  $f_{X|Y}(x|y)$  is given by

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{f_{X,Y}(x,y)}{f_Y(y)} \\ &= \frac{f_X(x) \frac{1}{\sqrt{2\pi}} e^{-(y-\sqrt{s}x)^2/2}}{\int f_X(x) \frac{1}{\sqrt{2\pi}} e^{-(y-\sqrt{s}x)^2/2} dx} \\ &= \frac{f_X(x) e^{\sqrt{s}yx - sx^2/2}}{\int f_X(x) e^{-\sqrt{s}yx - x^2/2} dx} \\ &= \frac{f_X(x) e^{\sqrt{s}yx - sx^2/2}}{Z(y; s)}. \end{aligned}$$

The notation  $Z(y; s)$  comes from statistical physics and is based on the idea that, given the event  $Y = y$ , the posterior distribution of  $X$  is proportional to

$$f_{X|Y=y}(x) \propto e^{-E(x;y)},$$

where  $E(x; y) = -\sqrt{s}yx + sx^2/2 - \ln f_X(x)$ . Then, to get the actual posterior distribution for  $X$ , we need to divide by the normalizing constant

$$Z(y; s) \triangleq \int_{\mathcal{X}} e^{-E(x;y)} dx = \int_{\mathcal{X}} f_X(x) e^{\sqrt{s}yx - sx^2/2} dx.$$

Using this, the mutual information can be written as

$$\begin{aligned} I(X; Y) &= \mathbb{E} \left[ \ln \frac{f_{X|Y}(x|y)}{f_X(x)} \right] \\ &= \mathbb{E} \left[ \sqrt{s}YX - \frac{1}{2}sX^2 \right] - \mathbb{E} [\ln Z(Y; s)] \\ &= \mathbb{E} \left[ \sqrt{s}(\sqrt{s}X + Z)X - \frac{1}{2}sX^2 \right] - \mathbb{E} [\ln Z(Y; s)] \\ &= \frac{1}{2}s\mathbb{E}[X^2] - F(s), \end{aligned}$$

where  $F(s) = \mathbb{E} [\ln Z(Y; s)]$  is known as the free energy in statistical physics. This implies that

$$\frac{d}{ds} I(X; Y) = \frac{1}{2} \mathbb{E}[X^2] - F'(s).$$

Thus, the next step is to compute  $F'(s)$ . For this proof, one should keep in mind that the setup is based on first sampling  $X$ , then sampling  $Z$  to compute  $Y = \sqrt{s}X + Z$ , and finally sampling  $W$  from

the posterior distribution of  $X$  given  $Y$ . Now, we define  $\phi(z) \triangleq e^{-z^2/2}/\sqrt{2\pi}$  and write

$$\begin{aligned}
F'(s) &= \frac{d}{ds} \mathbb{E} [\ln Z(Y; s)] \\
&= \frac{d}{ds} \int_{\mathcal{X}} \int_{\mathcal{Y}} f_X(x) \phi(y - \sqrt{s}x) \ln Z(y; s) dy dx \\
&= \frac{d}{ds} \int_{\mathcal{X}} \int_{\mathcal{Z}} f_X(x) \phi(z) \ln Z(\sqrt{s}x + z; s) dz dx \\
&= \int_{\mathcal{X}} \int_{\mathcal{Z}} f_X(x) \phi(z) \frac{d}{ds} \frac{Z(\sqrt{s}x + z; s)}{Z(\sqrt{s}x + z; s)} dz dx \\
&= \int_{\mathcal{X}} \int_{\mathcal{Z}} \frac{f_X(x) \phi(z)}{Z(\sqrt{s}x + z; s)} \frac{d}{ds} \int_{\mathcal{X}} f_X(w) e^{\sqrt{s}(\sqrt{s}x+z)w - sw^2/2} dw dz dx \\
&= \int_{\mathcal{X}} \int_{\mathcal{Z}} f_X(x) \phi(z) \int_{\mathcal{X}} \frac{f_X(w) e^{sxw + \sqrt{s}zw - sw^2/2}}{Z(\sqrt{s}x + z; s)} \left( xw + \frac{1}{2\sqrt{s}}zw - \frac{1}{2}w^2 \right) dw dz dx \\
&= \int_{\mathcal{X}} \int_{\mathcal{Z}} f_X(x) \phi(z) \int_{\mathcal{X}} f_{X|Y}(w|\sqrt{s}x + z) \left( xw + \frac{1}{2\sqrt{s}}zw - \frac{1}{2}w^2 \right) dw dz dx \\
&= \mathbb{E} [XW] + \frac{1}{2\sqrt{s}} \mathbb{E} [ZW] - \frac{1}{2} \mathbb{E} [W^2].
\end{aligned}$$

The key observation we need is that the integral over  $w$  is computing the conditional expectation given  $Y$ . This implies that

$$\begin{aligned}
\mathbb{E} [XW] - \frac{1}{2} \mathbb{E} [W^2] &= \int_{\mathcal{X}} \int_{\mathcal{Z}} f_X(x) \phi(z) \int_{\mathcal{X}} f_{X|Y}(w|\sqrt{s}x + z) \left( xw - \frac{1}{2}w^2 \right) dw dz dx \\
&= \mathbb{E} \left[ X \mathbb{E} [X|Y] - \frac{1}{2} \mathbb{E} [X^2|Y] \right] \\
&= \mathbb{E} [\mathbb{E} [X \mathbb{E} [X|Y] | Y]] - \frac{1}{2} \mathbb{E} [X^2] \\
&= \mathbb{E} [\mathbb{E} [X|Y]^2] - \frac{1}{2} \mathbb{E} [X^2].
\end{aligned}$$

Now, it remains only to compute

$$\begin{aligned}
\mathbb{E} [ZW] &= \int_{\mathcal{X}} \int_{\mathcal{Z}} f_X(x) \phi(z) \int_{\mathcal{X}} f_{X|Y}(w|\sqrt{s}x + z) zw dw dz dx \\
&= \int_{\mathcal{X}} f_X(x) \int_{\mathcal{Z}} \phi(z) z \int_{\mathcal{X}} f_{X|Y}(w|\sqrt{s}x + z) w dw dz dx \\
&= \int_{\mathcal{X}} f_X(x) \int_{\mathcal{Z}} \phi(z) z \mathbb{E} [X|Y = \sqrt{s}x + z] dz dx \\
&= \int_{\mathcal{X}} f_X(x) \int_{\mathcal{Z}} \phi(z) z \frac{1}{\sqrt{s}} \mathbb{E} [\sqrt{s}X|Y = \sqrt{s}x + z] dz dx \\
&\stackrel{(a)}{=} \frac{1}{\sqrt{s}} \int_{\mathcal{X}} f_X(x) \int_{\mathcal{Z}} \phi(z) \frac{d}{dz} \mathbb{E} [\sqrt{s}X|Y = \sqrt{s}x + z] dz dx \\
&\stackrel{(b)}{=} \frac{1}{\sqrt{s}} \mathbb{E} [\text{Var}(\sqrt{s}X|Y)] \\
&= \sqrt{s} \mathbb{E} [X^2] - \sqrt{s} \mathbb{E} [\mathbb{E} [X|Y]^2],
\end{aligned}$$

where (a) follow from Gaussian integration by parts (i.e.,  $\phi'(z) = -z\phi(z)$ ) and (b) follows from Lemma 6.

Putting everything together, we find that

$$\begin{aligned}
F'(s) &= \mathbb{E} [\mathbb{E} [X|Y]^2] - \frac{1}{2} \mathbb{E} [X^2] + \frac{1}{2} \mathbb{E} [X^2] - \frac{1}{2} \mathbb{E} [\mathbb{E} [X|Y]^2] \\
&= \frac{1}{2} \mathbb{E} [\mathbb{E} [X|Y]^2]
\end{aligned}$$

and therefore

$$\begin{aligned}\frac{d}{ds}I(X; Y_s) &= \frac{1}{2}\mathbb{E}[X^2] - \frac{1}{2}\mathbb{E}[\mathbb{E}[X|Y_s]^2] \\ &= \frac{1}{2}\text{mmse}(X|Y_s).\end{aligned}$$

□

## References

- [1] S. ten Brink, “Convergence of iterative decoding,” *Electronic Letters*, vol. 35, pp. 806–808, May 1999.
- [2] A. Ashikhmin, G. Kramer, and S. ten Brink, “Extrinsic information transfer functions: model and erasure channel properties,” *IEEE Trans. Inform. Theory*, vol. 50, pp. 2657–2674, Nov. 2004.
- [3] D. Guo, S. Shamai, and S. Verdú, “Mutual information and MMSE in Gaussian channels,” in *Proc. IEEE Int. Symp. Inform. Theory*, pp. 349–349, IEEE, 2004.
- [4] D. Guo, S. Shamai, and S. Verdú, “Mutual information and minimum mean-square error in Gaussian channels,” *IEEE Trans. Inform. Theory*, vol. 51, pp. 1261–1282, April 2005.
- [5] C. Méasson, A. Montanari, T. Richardson, and R. Urbanke, “Life above threshold: From list decoding to area theorem and mse,” *arXiv preprint cs/0410028*, 2004.
- [6] C. Méasson, A. Montanari, T. J. Richardson, and R. Urbanke, “The generalized area theorem and some of its consequences,” *IEEE Trans. Inform. Theory*, vol. 55, pp. 4793–4821, Nov. 2009.
- [7] C. Méasson, A. Montanari, and R. Urbanke, “Maxwell’s construction: the hidden bridge between maximum-likelihood and iterative decoding,” in *Proc. IEEE Int. Symp. Inform. Theory*, (Chicago, IL, USA), p. 225, June 2004.
- [8] C. Méasson, A. Montanari, and R. L. Urbanke, “Maxwell construction: The hidden bridge between iterative and maximum a posteriori decoding,” *IEEE Trans. Inform. Theory*, vol. 54, pp. 5277–5307, Dec. 2008.
- [9] R. Esposito, “On a relation between detection and estimation in decision theory,” *Inform. and Control*, vol. 12, no. 2, pp. 116–120, 1968.
- [10] C. Hatsell and L. Nolte, “Some geometric properties of the likelihood ratio (corresp.),” *IEEE Trans. Inform. Theory*, vol. 17, no. 5, pp. 616–618, 1971.