Dérivées de l'information mutuelle dans les canaux gaussiens

Minh-Toan NGUYEN[1]

[1]GIPSA-lab, Université Grenoble-Alpes
Minh-Toan.Nguyen@gipsa-lab.grenoble-inp.fr

Résumé – La formule I-MMSE relie deux quantités importantes en théorie de l'information et en théorie de l'estimation. Plus précisément, il indique que dans un canal gaussien, la dérivée de l'information mutuelle est la moitié de l'erreur quadratique moyenne minimale. Des dérivées plus élevées de l'information mutuelle sont liées à des erreurs d'estimation de moments plus élevés, mais une formule générale est inconnue. Dans cet article, nous dérivons une formule générale pour les dérivées de l'information mutuelle. Le résultat est remarquablement similaire à la relation moment-cumulant classique.

Abstract – The I-MMSE formula connects two important quantities in information theory and estimation theory. More precisely, it states that in a gaussian channel, the derivative of the mutual information is one-half of the minimal mean-squared error. Higher derivatives of the mutual information is related to estimation errors of higher moments, however a general formula is unknown. In this paper, we derive a general formula for the derivatives of the mutual information. The result is remarkably similar to the classic moment-cumulant relation.

1 Introduction

On considère le problème d'estimation d'un signal inconnu X étant donné d'une observation Y telle que

$$Y = \sqrt{\lambda}X + Z \tag{1}$$

où Z est une variable gaussienne standard indépendante de X. Le paramètre λ est le *rapport signal sur bruit*, et ce modèle est appelé *le canal gaussien*. L'erreur quadratique moyenne d'un estimateur $\hat{X}$ pour X est définie par $\mathbb{E}[(X-\hat{X})^2]$, qui est minimisée par l'estimateur $\mathbb{E}[X|Y]$. Par conséquent, l'erreur quadratique moyenne minimale (*minimum mean squared error* ou *MMSE* en anglais) est donnée par

$$\mathrm{MMSE}_X(\lambda) = \mathbb{E}[(X - \mathbb{E}[X|Y])^2]. \tag{2}$$

D'un autre côté, du point de vue de la théorie de l'information, nous sommes intéressés par l'information mutuelle entre le signal X et l'observation Y, notée $I(X;Y)$ ou $I_X(\lambda)$. Il est évident que $I_X(\lambda)$ est une fonction croissante de λ et que $I_X(0) = 0$. La relation fondamentale I-MMSE [1] énonce que

$$I'_X(\lambda) = \frac{1}{2}\mathrm{MMSE}_X(\lambda) \tag{3}$$

Ici l'information est mesurée en *nats*. En particulier, quand $\lambda = 0$ nous avons $\mathbb{E}[X|Y] = X$ et $\mathrm{MMSE}_X(0) = \mathrm{Var}(X)$, alors

$$I'_X(0) = \frac{1}{2}\,\mathrm{Var}(X) \tag{4}$$

Malgré que (4) est déjà connut dans les années 90s [2], ce n'est que beaucoup plus tard, dans [1], que la relation I-MMSE a été découverte. L'élément clé dans le passage de (4) à (3) est l'approche de canal incrémential, qui permet de calculer les dérivées $I^{(k)}_X(\lambda)$ pour λ positif à partir de $I^{(k)}_X(0)$. Les dérivées à zéro sont obtenues à partir du processus laborieux de calcul de l'expansion de Taylor de $I_X(\lambda)$. Notons qu'il y avait une erreur dans l'article original lors du calcul de la troisième dérivée, qui a été corrigée dans [3]. Bien qu'il ait été découvert beaucoup plus tard, la relation I-MMSE est équivalent à l'identité de Bruijn [4]. Cependant, l'information mutuelle et le MMSE sont des mesures plus standard que l'entropie différentielle et l'information de Fisher, des quantités qui apparaissent dans l'identité de Bruijn.

Avec cette approche, la dérivée k-ième peut être calculée pour n'importe quelle valeur de k. Cependant, une formule générale pour tout k est actuellement inconnue. Dans ce travail, nous dériverons une formule générale pour $I^{(k)}_X(\lambda)$. Cela sera accompli en étudiant le cas multivarié suivant. Soit $\boldsymbol{X}$ une variable aléatoire dans $\mathbb{R}^n$ et $\boldsymbol{\lambda} \in \mathbb{R}^n_+$. Définissons $I_{\boldsymbol{X}}(\boldsymbol{\lambda})$ comme l'information mutuelle entre les entrées $\boldsymbol{X}$ et les sorties $\boldsymbol{Y}$ des canaux gaussiens.

$$Y_i = \sqrt{\lambda_i}X_i + Z_i, \quad i = 1,\ldots,n \tag{5}$$

où Z_i sont des bruits gaussiens standards indépendants, indépendants de $\boldsymbol{X}$. Nous dérivons une formule générale pour $\partial^{k_1}_{\lambda_1}\ldots\partial^{k_n}_{\lambda_n} I_{\boldsymbol{X}}(\boldsymbol{\lambda})$. La formule obtenue présente une similarité remarquable avec la formule cumulant-moment classique [5]. Une idée clé de cet article est que le résultat pour le canal scalaire est obtenu en étudiant le cas multivarié. Cette stratégie est inspiré par le fait que les cumulants d'ordres supérieurs d'une variable aléatoire peuvent être mieux compris à travers les cumulants multivariés, via la relation

$$\kappa_n(X) = \kappa(\underbrace{X,\ldots,X}_{n \text{ fois}})$$

Notre travail repose sur deux composantes clés. La première consiste à réduire plusieurs canaux gaussiens ayant des signaux identiques en un seul canal sans perte d'information. Cela permet d'obtenir un raisonnement plus clair et intuitif que l'approche de canal incrémential pour le passage du cas $\lambda = 0$ au cas $\lambda > 0$. Cette méthode sert également un autre but : elle permet le calcul des dérivées supérieures à l'aide de canaux gaussiens multiples. La deuxième composante clé est la méthode de répliques issue de la physique des systèmes désordonnés [6]. Bien qu'elle ne soit pas rigoureuse, la méthode permet un calcul facile des dérivées de l'information mutuelle à $\lambda = 0$. Dans un certain sens, la méthode de répliques "explique" le résultat, tandis que la méthode traditionnelle implique souvent des termes qui s'annulent mystérieusement.

Nous considérerons dans cet article des variables aléatoires de moments finis, ce qui implique que l'information mutuelle est infiniment différentiable pour tout $\lambda \geq 0$.

2 Résultats

Nous appelons *multiensemble* une collection d'éléments dans laquelle les répétitions sont autorisées. Une *partition* d'un multiensemble est une façon de le diviser en parties, ou *blocs*. Si une partition π se compose des blocs $B_1, \dots, B_k$, nous écrivons $\pi = (B_1, \dots, B_k)$. Une partition est *diverse* si les éléments de chacun de ses blocs sont distincts. Par exemple, la partition $(\{1,2\},\{1,2\})$ du multiensemble $\{1,1,2,2\}$ est diverse tandis que la partition $(\{1,1\},\{2,2\})$ ne l'est pas.

Pour toutes les variables aléatoires $X_1, \dots, X_n$ avec $n \geq 1$, définissons

$$\begin{aligned}&\tau(X_1, \dots, X_n)\\ =&\sum_\pi \frac{(-1)^{k-1}(k-2)!}{2^{s(\pi)}}\,\mathbb{E}[X_{B_1}]\dots\mathbb{E}[X_{B_k}]\end{aligned} \tag{6}$$

où $X_B = \prod_{i\in B} X_i$. La somme est prise sur toutes les partitions diverses $\pi = (B_1, \dots, B_k)$ du multiensemble $\{1,1,\dots,n,n\}$ et $s(\pi)$ est le nombre de paires identiques de blocs dans π. Par example, $s(\pi) = 0$ si $\pi = (\{1,1\},\{2,2\})$ et $s(\pi) = 1$ si $\pi = (\{1,2\},\{1,2\})$.

Soit Y une variable aléatoire ou un événement, on peut définir $\tau(\,\cdot\,|Y)$ en remplaçant les espérances $\mathbb{E}[\,\cdot\,]$ dans la définition de τ par $\mathbb{E}[\,\cdot\,|Y]$.

La forme de τ ressemble à la relation classique entre cumulants et moments [5], qui énonce que

$$\begin{aligned}&\kappa(X_1, \dots, X_n)\\ =&\sum_\pi (-1)^{k-1}(k-1)!\,\mathbb{E}[X_{B_1}]\dots\mathbb{E}[X_{B_k}]\end{aligned} \tag{7}$$

où $\kappa(X_1, \dots, X_n)$ est le cumulant des variables aléatoires $X_1, \dots, X_n$ et la somme est prise sur toutes les partitions $\pi = (B_1, \dots, B_k)$ de $\{1, \dots, n\}$. Remarquez que la valeur de τ et κ ne dépend pas de l'ordre des arguments. Le cumulant est multilinéaire et s'annule si ses arguments peut être divisé en deux parties indépendantes, autrement dit $\kappa(X_1, \dots, X_n) = 0$ s'il existe deux ensembles non vides I, J avec $I \cap J = \emptyset$, $I \cup J = \{1, \dots, n\}$ tels que $(X_i)_{i\in I}$ et $(X_j)_{j\in J}$ sont indépendants. La forme τ a également des propriétés similaires:

Proposition 2.1.
a) τ est multiquadratique.
b) $\tau(X_1, \dots, X_n) = 0$ si $\{1, \dots, n\}$ peut être divisé en deux ensembles disjoints non vides I et J tels que $(X_i)_{i\in I}$ et $(X_j)_{j\in J}$ sont indépendants.

Ici, une fonction f d'un espace vectoriel V vers $\mathbb{R}$ est *quadratique* si

$$\begin{aligned}f(\lambda x) &= \lambda^2 f(x)\\ 2f(x) + 2f(y) &= f(x-y) + f(x+y)\end{aligned}$$

pour tout $\lambda \in \mathbb{R}$ and $x, y \in V$. Une fonction multivariée est dite *multiquadratique* si elle est quadratique en chacun de ses arguments.

La conclusion principale de cette étude peut être formulée comme suit :

Théorème 2.1. *Pour les canaux gaussiens définis dans (5):*
a) Les dérivées du premier ordre de l'information mutuelle sont données par

$$\partial_{\lambda_i} I_{\boldsymbol{X}}(\boldsymbol{\lambda}) = \mathbb{E}[(X_i - \mathbb{E}[X_i|\boldsymbol{Y}])^2] \tag{8}$$

b) Pour les dérivées d'ordre supérieur,

$$\begin{aligned}\partial_{\lambda_1}^{k_1}\dots\partial_{\lambda_n}^{k_n} I(\boldsymbol{\lambda}) &= \mathbb{E}[\tau(\underbrace{X_1,\dots,X_1}_{k_1},\dots,\underbrace{X_n,\dots,X_n}_{k_n}|\boldsymbol{Y})] &(9)\\ &= \mathbb{E}[\bar\tau(\underbrace{\bar X_1,\dots,\bar X_1}_{k_1},\dots,\underbrace{\bar X_n,\dots,\bar X_n}_{k_n}|\boldsymbol{Y})] &(10)\end{aligned}$$

où $\bar X_i = X_i - \mathbb{E}[X_i|\boldsymbol{Y}]$ et la forme $\bar\tau$ est définie de la même manière que τ, sauf que la somme est sur toutes les partitions diverses avec des blocs de taille supérieure à un.

Nous donnons ici quelques exemples pour le théorème 2.1 et récupérons quelques résultats de [3].

Exemple 2.1. On a

$$\bar\tau(X_1, X_2) = -\frac{1}{2}\mathbb{E}[X_1X_2]^2$$

Donc,

$$\partial_{\lambda_i}\partial_{\lambda_j} I_{\boldsymbol{X}}(\boldsymbol{\lambda}) = -\frac{1}{2}\mathbb{E}[\,\mathbb{E}[\,\bar X_i \bar X_j|\boldsymbol{Y}\,]^2\,] \tag{11}$$

$$\partial_{\lambda_i}^2 I_{\boldsymbol{X}}(\boldsymbol{\lambda}) = -\frac{1}{2}\mathbb{E}[\,\mathbb{E}[\,\bar X_i^2|\boldsymbol{Y}\,]^2\,] \tag{12}$$

Exemple 2.2. On a

$$\begin{aligned}&\bar\tau(X_1, X_2, X_3)\\ =&\mathbb{E}[X_1X_2]\mathbb{E}[X_2X_3]\mathbb{E}[X_3X_1] - \frac{1}{2}\mathbb{E}[X_1X_2X_3]^2\end{aligned}$$

à partir de laquelle on peut calculer les dérivées des types $\partial^3_{\lambda_i}$, $\partial^2_{\lambda_i}\partial_{\lambda_j}$ et $\partial_{\lambda_i}\partial_{\lambda_j}\partial_{\lambda_k}$, pour des i, j, k distincts. En particulier, dans le cas unidimensionnel (1),

$$I^{(3)}_X(\lambda) = \frac{1}{2}\mathbb{E}\Big[2M_2^3 - M_3^2\Big]$$

où

$$M_k = \mathbb{E}[(X - \mathbb{E}[X|Y])^k|Y]$$

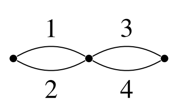
Exemple 2.3. On a

$$\begin{aligned}
&\bar{\tau}(X_1, X_2, X_3, X_4) = \\
&-2(\mathbb{E}[X_1X_2]\mathbb{E}[X_2X_3]\mathbb{E}[X_3X_4]\mathbb{E}[X_4X_1] + \text{deux autres termes}) \\
&-\frac{1}{2}(\mathbb{E}[X_1X_2]^2\mathbb{E}[X_3X_4]^2 + \text{deux autres termes}) \\
&+\mathbb{E}[X_1X_2]\mathbb{E}[X_1X_3X_4]\mathbb{E}[X_2X_3X_4] + \text{cinq autres termes} \\
&+\mathbb{E}[X_1X_2X_3X_4]\mathbb{E}[X_1X_2]\mathbb{E}[X_3X_4] + \text{deux autres termes} \\
&-\frac{1}{2}\mathbb{E}[X_1X_2X_3X_4]^2
\end{aligned}$$

d'où on peut calculer les dérivées du type ∂^4_i, $\partial^3_{\lambda_i}\partial_{\lambda_j}$, $\partial^2_{\lambda_i}\partial^2_{\lambda_j}$, $\partial^2_{\lambda_i}\partial_{\lambda_j}\partial_{\lambda_k}$ and $\partial_{\lambda_i}\partial_{\lambda_j}\partial_{\lambda_k}\partial_{\lambda_l}$, pour des i, j, k, l distincts. En particulier

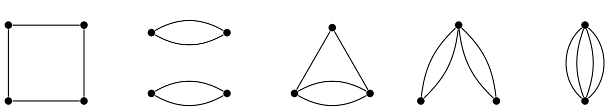
$$I^{(4)}_X(\lambda) = \frac{1}{2}\mathbb{E}\big[-15M_2^4 + 12M_3^2M_2 + 6M_4M_2^2 - M_4^2\big]$$

Remarque. Pour lister toutes les partitions diverses dans le développement de τ ou $\bar{\tau}$, il est utile de prendre en compte la bijection suivante entre les partitions diverses et les graphes sans boucle (graphes sans sommet se connectant à lui-même). Pour construire un graphe associé à une partition diverse π du multi-ensemble $\{1, 1, \ldots, n, n\}$, on considère les blocs de π comme les sommets du graphe. Chaque $i \in [n]$ doit appartenir à deux blocs différents de π, et on relie ces deux blocs par une arête étiquetée par i. Par exemple, la partition avec les blocs $\{1, 2\}$, $\{3, 4\}$, $\{1, 2, 3, 4\}$ correspond au graphe suivant



Inversement, étant donné un graphe sans boucle avec des arêtes étiquetées par $[n]$, nous pouvons récupérer la partition diverse correspondante en regardant les arêtes qui se connectent à chaque sommet.

Les lignes du développement dans l'exemple 2.3 correspondent aux graphes suivants



De plus, les termes de chaque ligne correspondent à différentes manières d'étiqueter les arêtes du graphe associé.

3 Cumulants et répliques

La méthode de réplique joue un rôle central dans la dérivation de notre résultat. Cette méthode est utilisée pour évaluer des expressions de la forme $\mathbb{E}[\log Z]$ en utilisant la formule simple suivante:

$$\mathbb{E}[\log Z] = \partial_{r=0}\mathbb{E}[Z^r] \tag{13}$$

La recette de calcul est la suivante. Tout d'abord, les moments $\mathbb{E}[Z^r]$ sont évalués pour $r \in \mathbb{N}$. Sur la base du résultat obtenu, on devine une formule pour $\mathbb{E}[Z^r]$ lorsque $r \in \mathbb{R}_+$, une étape appelée *continuation analytique*. Cela permet de prendre la dérivée à la fin.

Pour une brève introduction à la méthode, nous présentons ici une courte dérivation de la formule cumulant-moment classique à l'aide de la méthode de répliques, que nous croyons être nouvelle.

Rappelons que le cumulant des variables aléatoires $X_1, \ldots, X_n$ est défini comme

$$\kappa(X_1, \ldots, X_n) = \partial_{\lambda_1} \ldots \partial_{\lambda_n} \psi_{\boldsymbol{X}}(\boldsymbol{\lambda})|_{\boldsymbol{\lambda}=0}, \tag{14}$$

avec

$$\psi_{\boldsymbol{X}}(\boldsymbol{\lambda}) = \log \mathbb{E}e^{\langle \boldsymbol{\lambda}, \boldsymbol{X} \rangle}$$

En utilisant la méthode de réplique, nous avons

$$\begin{aligned}
\psi_{\boldsymbol{X}}(\boldsymbol{\lambda}) &= \partial_{r=0}\Big[\mathbb{E}\exp\Big(\sum_{1 \le i \le n} \lambda_i X_i\Big)\Big]^r \\
&= \partial_{r=0}\mathbb{E}\exp\Big(\sum_{1 \le i \le r} \lambda_i \sum_{1 \le a \le r} X_i^a\Big)
\end{aligned}$$

où $\boldsymbol{X}^1, \ldots, \boldsymbol{X}^r \overset{i.i.d}{\sim} \boldsymbol{X}$. De cela, nous obtenons

$$\begin{aligned}
\kappa(X_1, \ldots, X_n) &= \partial_{r=0}\mathbb{E}\prod_i \sum_a X_i^a \\
&= \partial_{r=0} \sum_{a_1, \ldots, a_n} \mathbb{E}[X_1^{a_1} X_2^{a_2} \ldots X_n^{a_n}]
\end{aligned}$$

Puisque $X_i^{a_i}$ et $X_j^{a_j}$ sont indépendants si $a_i \neq a_j$, nous avons

$$\mathbb{E}[X_1^{a_1} \ldots X_n^{a_n}] = \mathbb{E}[X_{B_1}] \ldots \mathbb{E}[X_{B_k}]$$

où $(B_1, \ldots, B_k)$ est la partition de $[n]$ telle que i et j sont dans le même bloc si $a_i = a_j$. D'autre part, chaque partition avec k blocs correspond à

$$P_{r,k} = r(r-1)\ldots(r-k+1)$$

façons de choisir $(a_1, \ldots, a_n)$. Par conséquent,

$$\kappa(X_1, \ldots, X_n) = \partial_{r=0} \sum_{\pi} P_{r,k}\, \mathbb{E}[X_{B_1}] \ldots \mathbb{E}[X_{B_k}]$$

où la somme s'étend sur toutes les partitions $\pi = (B_1, \ldots, B_k)$ de $[n]$. La relation entre cumulants et moments découle de

$$\partial_{r=0} P_{r,k} = (-1)^{k-1}(k-1)!$$

□

Table 1: Comparaison entre κ et τ

κ	τ
multilinéaire	multiquadratique
$\kappa(X_1,\ldots,X_n) = \partial_{\lambda_1}\ldots\partial_{\lambda_n}\psi_{\boldsymbol{X}}(\boldsymbol{\lambda})\|_{\boldsymbol{\lambda}=0}$	$\tau(X_1,\ldots,X_n) = \partial_{\lambda_1}\ldots\partial_{\lambda_n}I_{\boldsymbol{X}}(\boldsymbol{\lambda})\|_{\boldsymbol{\lambda}=0}$
somme sur des partitions de $\{1,\ldots,n\}$	somme sur diverses partitions de $\{1,1,\ldots,,n,n\}$
ne dépend pas de l'ordre des arguments	
s'annule si les arguments peuvent être divisés en deux parties indépendantes	

4 Esquisse de la dérivation

Nous donnons ici un aperçu de la façon dont nos résultats sont obtenus. Nous renvoyons les lecteurs à [7] pour plus de détails. Nous allons exploiter la propriété suivante des canaux gaussiens

Lemme 4.1. *Un ensemble de canaux gaussiens avec le même signal X et des bruits indépendants est équivalent à un seul canal gaussien avec le signal X et un rapport signal sur bruit égal à la somme des rapports signal sur bruit individuels.*

De plus, nous utiliserons la représentation de répliques suivante de $I_{\boldsymbol{X}}(\boldsymbol{\lambda})$:

$$\frac{1}{2}\sum_{i=1}^{n}\lambda_i\mathbb{E}[X_i^2] - \partial_{r=1}\mathbb{E}\exp\left(\sum_{i=1}^{n}\lambda_i\sum_{1\leq a<b\leq r}X_i^aX_i^b\right) \quad (15)$$

où $\boldsymbol{X}^1,\ldots,\boldsymbol{X}^r \overset{i.i.d}{\sim} \boldsymbol{X}$.

Tout d'abord, nous montrerons comment calculer $I_X^{(n)}(0)$ à partir des canaux gaussiens multivariés. Considérons les n canaux gaussiens avec des bruits indépendants, avec le même signal X et des rapports signal sur bruit $\lambda_1,\ldots,\lambda_n$. Par la Proposition 4.1, ces canaux peuvent être réduits sans perte d'information en un seul canal gaussien avec un signal X et un rapport signal sur bruit $\lambda_1+\cdots+\lambda_n$. De cette équivalence, nous avons

$$I_X(\lambda_1+\cdots+\lambda_n) = I_{X,\ldots,X}(\lambda_1,\ldots,\lambda_n) \quad (16)$$

En prenant la première dérivée de chaque λ_i à zéro, nous obtenons

$$I_X^{(n)}(0) = \partial_{\lambda_1}\ldots\partial_{\lambda_n}I_{X,\ldots,X}(\boldsymbol{\lambda})|_{\boldsymbol{\lambda}=0} \quad (17)$$

Cela nous amène à définir pour des variables aléatoires $(X_1,\ldots,X_n)$ avec $n\geq 2$,

$$\tau(X_1,\ldots,X_n) = \partial_{\lambda_1}\ldots\partial_{\lambda_n}I_{X_1,\ldots,X_n}(\boldsymbol{\lambda})_{\boldsymbol{\lambda}=0} \quad (18)$$

En suivant le même raisonnement, toutes les dérivées dans le cas multivarié peuvent être calculées grâce à la forme τ. D'autre part, il découle de la représentation de la réplique de l'information mutuelle que

$$\tau(X_1,\ldots,X_n) = -\partial_{r=1}\prod_{i=1}^{n}\sum_{1\leq a<b\leq r}X_i^aX_i^b \quad (19)$$

à partir de laquelle l'expression combinatoire de τ peut être dérivée de manière similaire à celle de la relation moment-cumulant dans la section précédente. En utilisant la connexion de cette forme avec l'information mutuelle, ses propriétés énoncés dans Proposition 2.1 peut également être prouvé.

5 Conclusion

Nous avons dérivé une formule générale pour les dérivées par rapport aux rapports signal sur bruit de l'information mutuelle entre les entrées et les sorties de plusieurs canaux gaussiens. Le calcul révèle un objet mathématique τ qui a de nombreuses propriétés similaires avec le cumulant (Table 1).

References

[1] Dongning Guo, Shlomo Shamai, and Sergio Verdú. Mutual information and minimum mean-square error in gaussian channels. *IEEE transactions on information theory*, 51(4):1261–1282, 2005.

[2] Sergio Verdú. On channel capacity per unit cost. *IEEE Transactions on Information Theory*, 36(5):1019–1030, 1990.

[3] Dongning Guo, Yihong Wu, Shlomo S Shitz, and Sergio Verdú. Estimation in gaussian noise: Properties of the minimum mean-square error. *IEEE Transactions on Information Theory*, 57(4):2371–2385, 2011.

[4] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[5] TP Speed. Cumulants and partition lattices 1. *Australian Journal of Statistics*, 25(2):378–388, 1983.

[6] Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

[7] Minh-Toan Nguyen. Derivatives of mutual information in gaussian channels. *arXiv preprint arXiv:2303.02500*, 2023.