Université de Montréal

Estimation Neuronale de L'information Mutuelle.

par

Mohamed Ishmael Belghazi

Département de mathématiques et de statistique
Faculté des arts et des sciences

Mémoire présenté en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en Informatique

Orientation Intelligence artificielle

December 4, 2020

Université de Montréal

Faculté des arts et des sciences

Ce mémoire intitulé

Estimation Neuronale de L'information Mutuelle.

présenté par

Mohamed Ishmael Belghazi

a été évalué par un jury composé des personnes suivantes :

Pascal Vincent

(président-rapporteur)

Aaron Courville

(directeur de recherche)

Emma Frejinger

(membre du jury)

Résumé

Nous argumentons que l'estimation de l'information mutuelle entre des ensembles de variables aléatoires continues de hautes dimensionnalités peut être réalisée par descente de gradient sur des réseaux de neurones. Nous présentons un estimateur neuronal de l'information mutuelle (MINE) dont la complexité croît linéairement avec la dimensionnalité des variables et la taille de l'échantillon, entrainable par retro-propagation, et fortement consistant au sens statistique.

Nous présentons aussi une poignée d'application ou MINE peut être utilisé pour minimiser ou maximiser l'information mutuelle. Nous appliquons MINE pour améliorer les modèles génératifs adversariaux. Nous utilisons aussi MINE pour implémenter la méthode du goulot d'étranglement de l'information dans un cadre de classification supervisé. Nos résultats montrent un gain substantiel en flexibilité et performance.
Mots-clés: Réseau de neurones artificiels, Théorie de l'information, Modèles génératifs.

Abstract

We argue that the estimation of mutual information between high dimensional continuous random variables can be achieved by gradient descent over neural networks. We present a Mutual Information Neural Estimator (MINE) that is linearly scalable in dimensionality as well as in sample size, trainable through back-prop, and strongly consistent. We present a handful of applications on which MINE can be used to minimize or maximize mutual information. We apply MINE to improve adversarially trained generative models. We also use MINE to implement the Information Bottleneck, applying it to supervised classification; our results demonstrate substantial improvement in flexibility and performance in these settings.
Keywords: Artifificial neural networks, Information theory, Generative models.

Table des matières

Liste des tableaux

Liste des figures

Remerciements

Ce travail n'aurait pu aboutir sans le concours de personnes auxquelles je voudrais témoigner ma reconnaissance. En premier lieu, je tiens à remercier mon superviseur, Monsieur Aaron COURVILLE, professeur à l'Institut québécois d'intelligence artificielle MILA, pour la confiance qu'il m'a accordée en encadrant ce travail, pour ses orientations, son suivi et sa grande disponibilité. Son infallible éthique scientifique, sa creativite et son souci pour le développement intellectuel de ses étudiants resteront à jamais une grande source d'inspiration pour moi. Je souhaite exprimer ma gratitude à Madame Linda PEINTHIERE, coordonnatrice aux affaires étudiantes (MILA) et à madame Céline BEGIN, technicienne à la gestion des dossiers étudiants du programmes Maitrise informatique et Doctorat en informatique de l'Université de Montréal, sans le soutien bienveillant desquelles ce travail n'aurait pas vu le jour. Ma reconnaissance va à l'ensemble des membres du staff du MILA qui n'ont lésiné sur aucun effort pour fournir aux étudiants les conditions optimales de réalisation de leurs recherches. Mes remerciements vont également à Devon Hjelm, Alex Lamb, Sai Rajeswar, Vincent Dumoulin, nos échanges ont nourri ma réflexion et stimulé ma curiosité et mes apprentissages. Je suis redevable à mes parents et à ma femme pour leur soutien inconditionnel et leurs encouragements permanents.

Chapitre 1

Introduction

L'information mutuelle de Shannon est depuis son élaboration rapidement devenue une quantité fondamentale dans divers domaines scientifiques.

L'intérêt de l'information mutuelle dans l'apprentissage machine est à la fois théorique et appliqué mais aussi épistémologique. Dans cette thèse, nous nous intéressons avant tout à l'information mutuelle comme mesure de dépendance. En effet, contrairement à la corrélation, l'information mutuelle capture les relations de dépendance statistique non-linéaire. En tant que mesure de dépendance, l'information mutuelle permet d'exprimer certains aspects fondamentaux de l'apprentissage automatique et des statistiques dans le cadre de la théorie de l'information. L'information mutuelle est utilisée dans un grand nombre de tâches d'apprentissage et statistiques. L'information mutuelle connecte la théorie de l'information aux sciences des données et, de ce fait, offre un langage uniforme rendant les méthodes d'apprentissage statistique accessibles à un grand nombre de domaines scientifiques. Ceci permet la dissémination rapide de progrès dans l'apprentissage machine aux autres domaines. Cela permet aussi d'attirer les communautés de l'apprentissage machine vers des problèmes rencontrés dans des domaines connexes ou d'application.

Cependant, l'information mutuelle est une quantité difficile à estimer. Depuis la conceptualisation de l'information mutuelle dans les travaux de Shannon, de nombreuses méthodes d'estimation de l'information mutuelle ont été proposées. Le coeur de la difficulté réside dans le fait que l'information mutuelle est une fonctionnelle non-linéaire de la densité jointe des données. L'estimation de la densité jointe étant un problème difficile et plus ardu que celui de l'estimation de l'information mutuelle, le rasoir de Vapnik suggère d'estimer l'information

mutuelle en passant outre l'estimation de la densité jointe. Nous argumentons qu'il est possible d'estimer l'information mutuelle en transformant un problème d'estimation en un problème d'optimisation sur des réseaux de neurones. Si l'on devait résumer notre travail en une phrase, nous dirions que nous tentons d'estimer l'information mutuelle en entraînant des réseaux de neurones à séparer les échantillons de la distribution jointe de ceux du produit des distributions marginales.

1.1. Plan

L'objectif de cette thèse est de fournir l'ensemble des éléments nécessaires a la compréhension de Mutual Information Neural Estimation Belghazi et al. (2018a). Cette thèse est organisée en 5 chapitres:

(1) Le chapitre 2 introduit les fondamentaux de la théorie de l'apprentissage statistiques.

(2) Le Chapitre 3 formule l'intérêt des représentations des données en apprentissage machine, leurs utilisation grâce aux méthodes à noyau et leurs apprentissages par réseaux de neurones.

(3) Le chapitre 4 présente l'ensemble des notions de théorie de l'information nécessaires à la compréhension de notre approche pour estimer l'information mutuelle.

(4) Le chapitre 5 étudie certains estimateurs de l'information mutuelle et présente certaines applications de cette dernière à l'apprentissage machine.

(5) Le chapitre 6 contient l'article Mutual Information Neural Estimation Belghazi et al. (2018a) publié lors de la conférence internationale sur l'apprentissage machine en 2018 à Stockholm.

1.2. Contributions

Nous résumons les contributions de cette thèse. Ce travail à été effectué avec la collaboration de scientifiques extraordinaires, principalement mon superviseur de thèse Aaron Courville.

(1) Nous proposons MINE, un estimateur de différentiable de l'information mutuelle adapté aux situations avec haut volume de données et en haute dimension.

(2) Nous établissons les propriétés statistiques théoriques MINE, à savoir la consistance forte et sa complexité d'échantillon.

(3) Nous appliquons MINE pour offrir une implémentation du goulot d'étranglement de l'information continue.

(4) Nous utilisons MINE comme régularisateur afin de pallier l'oblitération de modes dans les réseaux génératifs adversariels et pour améliorer la fidélité des reconstructions dans les modèles entraînés par inférence adversariellement apprise.

Chapitre 2

Éléments de théorie de l'apprentissage statistique

Dans ce chapitre nous exposerons l'ensemble des éléments nécessaires à la compréhension des propriétés statistiques de notre estimateur. Dans un premier mouvement nous procéderons à un survol des fondamentaux de la théorie de l'apprentissage statistique.

Dans un second mouvement, nous introduirions la notion de complexité de l'échantillon.

2.1. L'apprentissage comme solution à un problème inverse

Pour ancrer notre exposition dans une situation familière et utile, nous commençons par considérer un problème de prédiction. Soit X et Y deux variable aléatoire sur un espace de probabilité $(\mathcal{X} \times \mathcal{Y}, \mathcal{F}, \mathbb{P})$. Un problème de prédiction a typiquement trois composantes, une distribution de probabiliste $\mathbb{P}_{XY}$ couplant les variables X et Y. Une famille de fonctions de prédiction $f : \mathcal{X} \to \mathcal{Y}$, et une fonction de perte $l : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. La qualité d'un prédicteur f, peut être évaluée en utilisant le risque espéré, défini comme,

$$\mathcal{R}[f] := \mathbb{E}_{\mathbb{P}}[l(f(X), Y)]. \tag{2.1.1}$$

La candidat optimal dans ce scénario est donc donné par,

$$f^* = \arg\inf \mathcal{R}[f]$$

Maintenant que le cadre du problème de prédiction est défini. Il reste à le rattacher au problème d'induction. Cependant, dans les cas où la distribution jointe $\mathbb{P}_{XY}$ est connue, il n'y

a pas d'induction à faire car les relations entre X et Y y sont encodées. Le problème de prédiction devient un problème d'induction dès lors que $\mathbb{P}_{XY}$ est partiellement ou complètement indisponible.

Supposons que nous ayons accès a un ensemble d'entraînement $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d}{\sim} \mathbb{P}_{XZ}$. De manière équivalente, nous pouvons considérer l'ensemble d'entraînement comme le produit direct de n copie de $\mathbb{P}_{XZ}$, autrement dit

$$U^n \sim \mathbb{P}_{XY}^{\otimes n}.$$

Un algorithme d'apprentissage est une procédure qui prend en entrée l'ensemble d'entraînement U^n et retourne en sortie un candidat $\hat{f}_n : \mathcal{X} \mapsto \mathcal{Y}$. Dans le cadre du problème de prédiction, nous remplaçons la mesure $\mathbb{P}_{XY}$ par la mesure empirique $\mathbb{P}_{XY}^n$ dans 2.1.1, pour obtenir le risque empirique espéré,

$$\mathcal{R}_n[f] := \mathbb{E}_{\mathbb{P}^n}[l(f(X), Y)]. \tag{2.1.2}$$

Le candidat est obtenu en minimisant 2.1.2,

$$\hat{f}_n = \arg\inf \mathcal{R}_n[f].$$

Il est important de noter que le candidat $\hat{f}$, tout comme la fonctionnelle $f \mapsto \mathcal{R}_n[f] := \mathbb{E}_{\mathbb{P}_{XY}^n}[l(f(X), Y)]$, sont des fonctions de l'ensemble d'entraînement U^n.

La qualité du candidat est évaluée en considérant, l'erreur de généralisation,

$$Gen(f) := \mathbb{E}_{\mathbb{P}_{XY}}[l(f(X), Y) \mid U^n].$$

Plus l'erreur de généralisation est petite, meilleur est le candidat. Est-il suffisant de minimiser le risque empirique espéré pour minimiser l'erreur de généralisation? La réponse est un non. Pour s'en convaincre, il suffit de considérer une courbe passant par tous les points de l'ensemble d'entraînement. Deux choix d'interpolation différente des points de l'ensemble d'entraînement, par exemple, une interpolation linéaire et une autre en polynôme de Lagrange, minimise le risque empirique espéré mais n'auront pas la même erreur de généralisation tandis que cette dernière est minimisée par f^*. Pour résoudre le problème d'induction, il est nécessaire de faire des hypothèses sur la nature de la relation fonctionnelle entre X et Y, c'est ce qu'on appelle les biais inductifs, on réduit l'espace de recherche des candidats de celui de l'ensemble des fonctions mesurables à un ensemble plus petit que l'on

sait a priori pertinent à la relation entre X et Y. Nous appelons cet ensemble de fonctions, l'espace des hypothèses que nous notons $\mathcal{H}$.

Ainsi notre candidat issu de l'espace des hypothèses est donné par,

$$\hat{h}_n = \arg\inf_{h\in\mathcal{H}} \mathcal{R}_n[f]. \tag{2.1.3}$$

La qualité de $\hat{h}_n$ comme solution au problème d'induction est donc quantifiée par le risque excédentaire,

$$\mathcal{R}[\hat{h}_n] - \mathcal{R}[f^*]. \tag{2.1.4}$$

Le risque excédentaire peut être d'avantage décomposé, en effet,

$$\mathcal{R}[\hat{h}_n] - \mathcal{R}[f^*] = \underbrace{\mathcal{R}[\hat{h}_n] - \inf_{f\in\mathcal{H}} \mathcal{R}[f]}_{Erreur d'estimation} + \underbrace{\inf_{f\in\mathcal{H}} \mathcal{R}[f] - \mathcal{R}[f^*]}_{Erreur d'approximation}. \tag{2.1.5}$$

2.1.1. Complexité de l'échantillon

La complexité de l'échantillon est le nombre d'exemples d'entraînement nécessaires à un algorithme d'apprentissage de sorte que le candidat retourné par l'algorithme soit arbitrairement proche de la solution optimale, et ce, avec une probabilité arbitrairement proche de 1.

Afin d'étayer le concept, considérons un exemple simple. Soit $X \sim Bern(\theta)$. Nous supposons que le paramètre θ est inaccessible et que seul un échantillon $X^n \sim Bern(\theta)^{\otimes n}$ est disponible. Par la loi forte des grands nombres nous savons que,

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{n\to\infty} \mathbb{E}[X] = \theta, \quad \text{p.p.}$$

Ceci suggère donc un estimateur naturel du paramètre θ,

$$\hat{\theta}(X^n) = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

La question qui se pose donc est: quelle est la taille n de l'échantillon nécessaire pour assurer que $\mid \hat{\theta}(X^n) - \theta \mid$ est inférieur à un certain $\varepsilon > 0$. Nous voulons donc comprendre, pour un $\varepsilon > 0$ préalablement fixe, à quelle vitesse la probabilité de l'évènement,

$$\{\mid \hat{\theta}(X^n) - \theta \mid \leq \varepsilon\},$$

tend vers 1 lorsque n tend vers l'infini. Le paramètre θ demeurant inconnu, l'estimation de la probabilité de l'évènement est impossible. Nous chercherons donc à trouver une majoration dépendante de n et ϵ. En appliquant l'inégalité d'Hoeffding nous obtenons,

$$\mathbb{P}^{\otimes n}(\{\mid \hat{\theta}(X^n) - \theta \mid \leq \varepsilon\}) \leq 1 - 2e^{-2\varepsilon^2 n}. \tag{2.1.6}$$

Ainsi, si nous voulons que la probabilité de la magnitude de la différence entre $\hat{\theta}(X^n)$ et θ soit inférieure à ε soit d'au moins $1 - \delta$ avec $\delta > 0$, il est suffisant que la taille de n l'échantillon soit supérieure à

$$\lfloor \frac{log(\frac{2}{\delta})}{2\varepsilon^2} \rfloor + 1.$$

Chapitre 3

Représentations, leurs apprentissages et réseaux de neurones

Par représentations de données nous entendons toute quantité calculée à partir d'un échantillon de données. Dans ce sens, les représentations sont des statistiques des données. Nous obtenons ces représentations en appliquant des cartes de représentations à un échantillon. Le choix des cartes de représentations est typiquement fait avec l'objectif de répondre à une problématique définie. Par exemple, si l'objectif est d'estimer une modèle gaussien par maximum de vraisemblance le meilleur choix de représentations sont naturellement les deux premier moments empiriques car ils constituent des statistiques suffisantes des données pour ce problème d'estimation. Ainsi, le choix des représentations est fait pour répondre à des problématiques d'apprentissage. Certaines représentations rendent des problèmes de classification binaire séparables, d'autres promettent de généraliser à des données non-observées ou même à d'autres ensembles de données,alors que d'autres encore facilitent les problèmes d'optimisation sous-jacents à l'estimation, ou permettent de capturer les dépendances non-linéaires qui peuvent exister dans les données. D'un point de vue épistémologique, les représentations sont les facteurs explicatifs d'une théorie qu'une procédure d'apprentissage cherche à apprendre afin de confirmer une hypothèse. Elles sont le sujet du problème d'induction que l'apprentissage cherche à résoudre. Ce chapitre vise à présenter deux approches fondamentales pour obtenir des représentations des données. La première utilise des cartes de représentations fixes grâce aux méthodes à noyaux. La seconde, s'inscrit dans le cadre de la révolution épistémologique qu'est l'apprentissage des représentations et cherche à apprendre les cartes de représentations grâce aux réseaux de neurones.

3.1. Représentation et méthodes à noyau

Les méthodes à noyau exploitent des fonctions de similarité entre des instances de données afin d'offrir une carte de représentations plongeant les données dans un espace de représentations de dimension potentiellement infinie. L'objectif est de transformer les relations non-linéaires dans l'espace de données en relations linéaires dans l'espace des représentations. Ces fonctions de similarités sont appelées noyaux. Commençons par les définir.

Définition 3.1.1. Noyau défini positif

Une fonction $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ est dite noyau si elle est symétrique, pour tout $x_i, x_j \in \mathcal{X}$

$$k(x_i, x_j) = k(x_j, x_i),$$

et positive définie, pour tout $x_1, \ldots, x_n \in \mathcal{X}$ et $c_1, \ldots, c_n \in \mathbb{R}$

$$\sum_{i=1}^{k} c_i c_j k(x_i, x_j) \geq 0$$

Cette définition peut être comprise comme une généralisation de la définition des produit scalaires à des espaces potentiellement infinis. En effet, tout produit scalaire sur un espace vectoriel fini apparaît comme une matrice symétrique et positive de finie.

Tout noyau ainsi défini peut être caractérisé par une représentation $\phi : \mathcal{X} \to \mathcal{H}$.

Définition 3.1.2. Noyau par représentation

Une fonction $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ est un noyau s'il existe un espace d'Hilbert $\mathcal{H}$ et une fonction $\phi : \mathcal{X} \to \mathcal{H}$ tel que,

$$\forall x, x^{'} \in \mathcal{X}, \quad k(x, x^{'}) := (\phi(x), \phi(x^{'}))_{\mathcal{H}} \tag{3.1.1}$$

En général, il existe plusieurs représentations ϕ et espace d'Hilbert $\mathcal{H}$ par noyau k. Cependant, chaque noyau k est uniquement associé un espace d'Hilbert à noyau reproduisant(RKHS) $\mathcal{H}_k$, où l'indexe k met en exergue la dépendance de l'espace $\mathcal{H}_k$ sur le noyau k.

Définition 3.1.3. Espace d'Hilbert à noyau reproduisant Un espace d'Hilbert $\mathcal{H}$ est dit à noyau reproduisant si pour tout $x \in \mathcal{X}$, la fonctionnelle

$$\mathcal{H} \to \mathbb{R}$$

$$f \mapsto L_x[f] := f(x),$$

est continue.

En effet, la continuité de la fonctionnelle permet l'application du théorème de représentations de RiezRiesz (1914) affirmant qu'il est existe un unique élément $k(x,\cdot) \in \mathcal{H}$ tel que,

Définition 3.1.4. Propriété reproduisante

$$\forall f \in \mathcal{H}, \quad L_x[f] = (f, k(x,\cdot))_{\mathcal{H}} = f(x).$$

Ainsi, pour un RKHS,$\mathcal{H}_k$ nous pouvons identifier une représentation canonique, nous permettant d'écrire,

$$\forall x, x^{'} \in \mathcal{X}, k(x,x^{'}) = (k(x,\cdot), k(x^{'},\cdot))_{\mathcal{H}_k}.$$

Le théorème de Moore-AronszajnAronszajn (1950), garantit l'existence d'un unique RKHS pour tout noyau positif défini k

Théorème 3.1.5 (Moore-Aronszajn). Soit $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, alors, il existe un unique espace de Hilbert de fonctions $\mathcal{H}_k$ sur $\mathcal{X}$ pour lequel k est un noyau reproduisant.

La propriété reproduisante rend l'évaluation de représentation tractable en décomposant tout élément d'un RKHS en combinaison linéaire d'évaluation de noyau.

3.1.1. Apprentissage par méthode de noyau

Le théorème du représentant (Schölkopf et al., 1997) utilise la propriété reproduisante pour,

(1) affirmer que la solution du problème d'induction par apprentissage est possible dans les RKHS.

(2) Montrer comment calculer la solution.

Théorème 3.1.6. Théorème du représentant Soit $\Omega : [0,\infty) \to \mathbb{R}$ une fonction strictement croissante. Soit $\mathcal{X}$ un ensemble, $l : \mathcal{Z} \to \mathbb{R}$ une fonction de coût et $\mathcal{H}_k$ un espace d'Hilbert à noyau reproduisant. $z^n = (z_1, \ldots, z_n) \sim \mathbb{P}^{\otimes n}$ un échantillon. Toute fonction $f \in \mathcal{H}_k$ minimisant,

$$l(f, z^n) + \Omega(\|f\|_{\mathcal{H}_k})$$

Peut s'écrire sous la forme de

$$f(\cdot) = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot)$$

Plusieurs algorithmes d'apprentissage peuvent être exprimés en termes de produit scalaire euclidien $x_i \dot{x}_j$ ou x_i et x_j sont deux points de l'ensemble d'entraînement. Pour profiter du pouvoir expressif des représentations induites par un kernel k et son unique RKHS associé, $\mathcal{H}_k$, il est suffisant de remplacer le produit scalaire euclidien $x_i \dot{x}_j$ par l'évaluation du noyau $k(x_i, x_j)$ et d'utiliser la propriété reproduisante.

Les méthodes à noyau jouissent d'un cadre théorique formel permettant à la fois de fournir des garanties théoriques aux procédures d'apprentissage les utilisant mais aussi d'offrir une approche méthodologique permettant d'adapter plusieurs algorithmes d'apprentissage établis à leur utilisation. Il souffre, néanmoins de deux limitations fondamentales. La première est de nature computationnelle. L'utilisation de carte de représentations implicitement définies par la relation entre chaque paire de points de l'ensemble d'entraînement tel qu'évalué par le noyau, ou matrice de Gram, est exorbitante en mémoire et en calcul. En effet, si la cardinalité de l'ensemble d'entraînement est n, l'inversion de la matrice de Gram est $\mathcal{O}(n^3)$. De plus, si les données sont de dimension d la nécessité d'avoir l'ensemble d'entraînement en mémoire implique un coût de l'ordre $\mathcal{O}(dn)$. Les méthodes à noyaux sont donc inutilisables dans des régimes de haut volume et de haute dimensionnalité des données.

La seconde est due au choix du noyau. Le noyau idéal est spécifique au problème à résoudre. L'erreur de spécification du modèle peut en conséquence être élevée.

3.2. Apprentissage des représentations et réseaux de neurones

La construction de représentations en utilisant des cartes fixes indépendantes du problème d'apprentissage à résoudre pose inévitablement un problème épistémologique. Si l'objectif est la confirmation d'une hypothèse scientifique sur la base de données, le simple fait d'introduire des facteurs explicatifs émanant de cartes fixes fragilise la confirmation de l'hypothèse. En effet, nous pouvons y voir une violation du principe de Copernic. Les facteurs explicatifs sont construits par l'observateur et placent donc ce dernier au centre de la tentative de confirmation ou d'infirmation de l'hypothèse. L'apprentissage de représentations constitue en ce sens un véritable changement de paradigme. Apprendre les facteurs explicatifs à partir des données équivaut à avoir moins d'à priori sur la nature du phénomène à expliquer. Il est important de distinguer que l'apprentissage des représentations doit avoir des a priori sur la nature des facteurs explicatifs a découvrir. La différence est que ces a priori opèrent

à un niveau d'abstraction supérieur à celui des représentations émanant de cartes fixes. L'apprentissage des représentations fait des a priori sur le processus génératif des facteurs explicatifs. Bengio et al. (2013) présentes neufs a priori que de bonnes représentations devraient avoir,

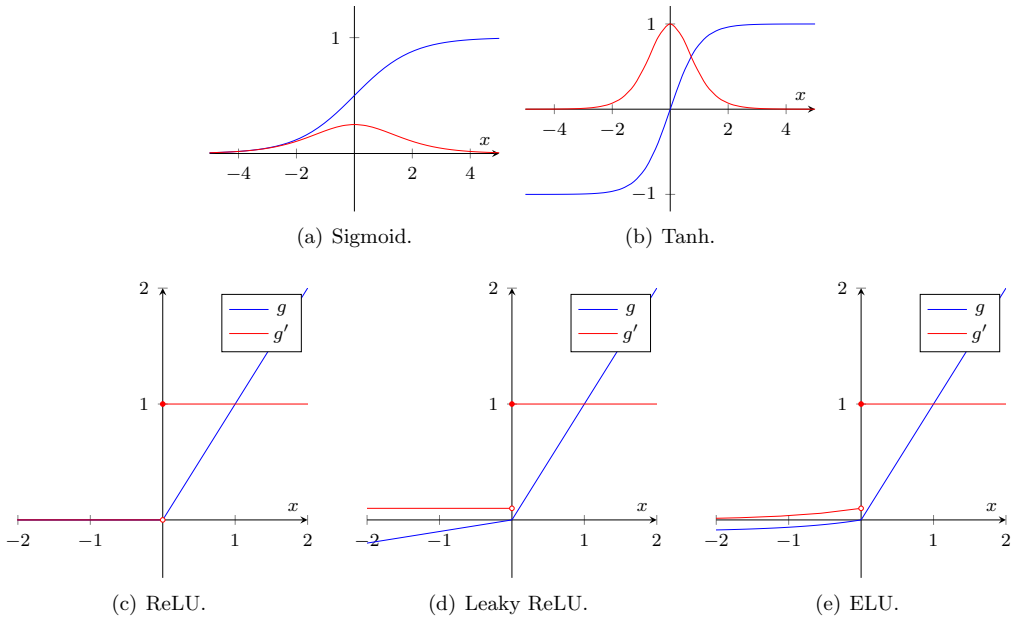
(1) Les cartes de représentations devraient être des fonctions lisses.
(2) Les représentations devraient être distribuées.
(3) Les représentations devraient être composables afin de créer une hiérarchie d'abstraction.
(4) Les représentations expliquant la structure des données devraient être utiles pour prédire à partir des données.
(5) Les représentations devraient être transférables à d'autre tâches.
(6) Les représentations devrait être distribuées sur une variété dont la dimension topologique est largement inférieure à celle des données. Cette hypothèse est communément dénommée l'hypothèse de la variété.
(7) Les représentations devraient être regroupées. Les variations de la variété au sein de chaque groupe devraient être inférieures aux variations de la variété entre les groupes.
(8) Les représentations devrait faire preuve de cohérence spatiale et temporelle.
(9) Les représentations devraient être éparses
(10) Les relations de dépendance entre les représentations devraient être simples.

3.2.1. Réseaux de neurones artificielles

Les réseaux de neurones multi-couches permettent d'apprendre des hiérarchies exhibant un niveau croissant d'abstraction. Les réseaux de neurones à convolution (LeCun et al., 1998) s'inspirent du fonctionnement du cortex primaire (V1). Zeiler & Fergus (2014) démontrent empiriquement que les représentations apprises exhibent des représentations de plus en plus abstraites. Les réseaux de neurones profonds promettent de satisfaire la majorité des des a priori mentionnés ci-dessus. L'évidence empirique montre que les réseaux de neurones sont en passe de tenir cette promesse. Il faut noter que cela n'a pas toujours été le cas. En effet, ce n'est qu'a partir de la parution de Hinton et al. (200) qu'il a été empiriquement établi que les réseaux de neurones profonds non-convolutionnel offrent de meilleures performances que leurs homologues à une ou deux couches. L'augmentation de la puissance de calcul

ainsi que la disponibilité d'ensemble de données volumineux sont certainement des facteurs expliquant le progrès dans l'entraînement des réseaux de neurones profonds. L'accumulation de savoir pratique permettant de pallier les problèmes d'estimation et d'optimisation propres aux réseaux profonds constitue une raison non-moindre de leur succès. En effet, ce savoir pratique c'est traduit par le développement de méthodes d'initialisation, de normalisation et d'optimisation ainsi que d'architecture qui ont facilité l'entraînement des réseaux de neurones. Les réseaux de neurones artificiels sont inspirés des réseaux de neurones biologiques. Un neurone artificiel est la composition d'un produit scalaire, entre les entrants et les paramètres, et d'une fonction g dite fonction d'activation. Nous envoyons le lecteur à la figure 3.1. La sortie du neurone est dénommée activation. Plus formellement, soit $x \in \mathbb{R}^d$, $w \in \mathbb{R}^d$ et $b \in \mathbb{R}$, l'activation du neurone peut s'écrire,

$$h(x) = g(\sum_{i=1}^{d} w_i x_i + b) = g(w^\top x + b).$$



(a) Sigmoid. (b) Tanh.

(c) ReLU. (d) Leaky ReLU. (e) ELU.

Fig. 3.1. Exemples de fonctions d'activation

3.2.1.1. Réseaux de neurones multi-couches.

Les réseaux de neurones artificiels sont typiquement organisés par couches. Les entrées de chaque couche correspondent aux sorties de la couche précédente, et ce, jusqu'à la première

couche qui reçoit les données. Plus formellement, soit $W^{(1)}, \ldots, W^L$ des matrices de poids avec $W_i \in \mathbb{R}^{d_{i-1} \times d_i}$, $b^1, \ldots, b^L$ des vecteurs de biais tel que $b_i \in \mathbb{R}^{d_i}$, nous pouvons écrire un réseau à propagation avant récursivement comme,

$$\begin{aligned} h^0 &= x, \\ a^l &= W^l h^{l-1} + b^l \\ h^l &= g(a^l). \end{aligned} \tag{3.2.1}$$

3.2.1.2. Théorèmes d'approximation universelle.

Pour toute fonction mesurable peut-on trouver un réseau de neurones capable de l'approximer à un degré de précision arbitraire? Les théorèmes d'approximation universelle fournissent une réponse théorique à cette question. Le théorème d'approximation universelle d'Hornick Hornik (1989), démontre que pour toute fonction continue avec support sur un ensemble compact de $\mathbb{R}^n$ il existe un réseau de neurones à propagation avant avec une couche cachée capable de l'approximer.

Théorème 3.2.1. Théorème d'Hornick Soit g une fonction, non-constante, continue et bornée. Soit K un sous-ensemble compacte de $\mathbb{R}^d$. Soit $C(K)$ l'ensemble des fonction continues sur K. Alors, pour tout $\varepsilon \geq 0$ et pour tout $f \in C(K)$, il existe $v_i, b_i \in \mathbb{R}$ et $w^i \in \mathbb{R}^d$ pour $i \in \{1, \ldots N\}$

$$\forall x \in K, \quad | f(x) - \sum_{i=1}^{N} v_i g(w_i^\top x + b_i) | < \varepsilon$$

Bien que le théorème d'Hornick fournisse une garantie théorique, il requiert que le nombre d'unités cachées soit exponentiellement large.

Récemment, Lu et al. (2017) ont montré qu'il est possible d'obtenir un théorème d'approximation universel pour les réseaux de neurones ayant des couches cachées de tailles fixes.

Théorème 3.2.2. Approximation universelle avec largeur fixe Pour toute fonction $f \in L_1(\mathbb{R}^d)$ et pour tout $\varepsilon > 0$, il existe un réseau de neurones $\mathcal{A}$, à propagation avant, doté de fonctions d'activation ReLU $\mathcal{A}$, de largeur $d_l \leq d + 4$ tel que la fonction $\mathcal{F}_\mathcal{A}$ représentant ce réseau satisfasse

$$\|f - \mathcal{F}_{\mathcal{A}}\|_{L_1(\mathbb{R}^d)} < \varepsilon.$$

3.2.1.3. Entraînement par rétro-propagation.

L'algorithme de rétro-propagation (Rumelhart et al., 1986) calcul les gradients de la fonctions de perte par rapport à chacun des paramètres du réseau à propagation avant. Ce dernier pouvant être compris comme une composition de fonction, l'algorithme de rétro-propagation combine la règle de la dérivée en chaîne et la programmation dynamique pour éviter toute redondance dans le calculs des gradients. Un des aspects les plus avantageux de l'algorithme de rétro-propagation est son efficacité computationelle. Un réseau défini par 3.2.1 contient $D := \sum_{i=1}^{L} d_{i-1}\, d_i + d_i$ paramètres, le nombre de calculs effectués pendant de les phases de propagation avant et arrière est de l'ordre $\mathcal{O}(D)$. En effet, sans une application naïve de la règle des dérivées en chaîne, on aurait une complexité de l'ordre de $\mathcal{O}(D^2$. Pour pouvoir jouir des avantages de la programmation dynamique, il est nécessaire de calculer les gradients en utilisant un algorithme de différentiation automatique. En effet, Il est possible d'approximer les gradients par différence finie, en prenant la moyenne des expansion de Taylor de second ordre de la fonction de coût et approcher chaque paramètre par la gauche et la droite. Plus formellement,

$$\frac{\partial \mathcal{R}}{\partial W_{ij}} = \frac{\mathcal{R}(W_{ij} + h) - \mathcal{R}(W_{ij} - h)}{2h} + o(h^2).$$

La nécessite de perturber chaque poids individuellement implique un coût computationel de l'ordre de $\mathcal{O}(D^2)$.

3.2.2. Entraînement de réseaux profonds

Le problème des gradients diminuant (Hochreiter, 1991; Glorot et al., 2011) est une conséquence de la profondeur[1] et de l'entraînement par rétro-propagation.

3.2.2.1. Initialisation prudente.

[1] Le problème des gradients diminuants affecte aussi les réseaux de neurones récurrents. Ces derniers étant entraînés en les déroulant, le problème peut être conçu dans ce cas aussi comme une conséquence de la profondeur.

Glorot et al. (2011) montre que la variance des gradients décroît au fur et a mesure qu'ils sont rétro-propagés vers la première couche. En faisant les hypothèses que les poids et les données sont des variables aléatoires centrées et réduites, que les fonctions d'activation du réseau sont symétriques et dans un régime linéaire lors de l'initialisation, la variance de l'activation i à la couche l est donnée par,

$$\begin{aligned} Var(z_i^l) &= Var(\sum_{j=1}^{d_{l-1}} W_{ij}^l h_j^{l-1} + b_i) \\ &= \sum_{j=1}^{d_{l-1}} Var(W_{ij}^l) Var(h_j^{l-1}) \\ &= d_{l-1} Var(W_{ij}^l) Var(h_j^l). \end{aligned}$$

La variance se décompose en somme des variances car à l'initialisation il est naturel d'assumer que les poids sont indépendants entre eux et avec les données. De plus nous assumons qu'à chaque couche les poids seront initialisés avec la même distribution. La variance des activations et des poids étant la même à chaque couche, nous omettons les indices des activations individuelles,

$$\begin{aligned} Var(h^l) &= Var(W^l h^{l-1} + b^l) \\ &= Var(W^l) Var(W^{l-1} h^{l-2} + b^{l-1}) \\ &= Var(x)\, \Pi_{j=0}^{l-1} d_i Var(W^i). \end{aligned}$$

Ainsi, la variance des activations de deux couches successives sont égales pendant la phase de propagation si et seulement si,

$$\begin{aligned} \forall lin\{1, \dots L\}, \quad \frac{Var(h^l)}{var(h^{l-1})} &= \frac{Var(x), \Pi_{i=1}^{l-1} d_i Var(W^i)}{Var(x), \Pi_{i=1}^{l-2} d_i Var(W^i)} \\ &= d_{l-1} Var(W^l) \\ &= 1. \end{aligned}$$

Pour garantir que l'égalité des variances des activations d'une couche à la suivante pendant la phase de propagation avant, il est suffisant d'initialiser les poids de chaque couche avec

une distribution ayant une variance égale à l'inverse du nombre de neurones entrants, soit égale à

$$\frac{1}{d_{l-1}}.$$

Contrôler la variance des gradients pendant la phase de rétro-propagation empêchera les gradients de diminuer prématurément. En appliquant une analyse similaire à la phase de propagation.

$$Var[\frac{\partial \mathcal{R}}{\partial W^l}] = \Pi_{n=1}^{n-1} d_k Var(W^k) \Pi_{m=l}^{L-1} Var(W^m) Var(x) Var(\frac{\partial \mathcal{R}}{\partial h^L})$$

De manière analogue, nous déduisions que le variances des gradients reste égale d'une couche à l'autre si les poids de la couche l sont initialisés avec une distribution ayant une variance égale à l'inverse du nombre de neurones sortant, soit

$$\frac{1}{d_l}$$

Pour garantir l'égalité des variances pendant les deux phases, Glorot et al. (2011) propose d'initialiser les poids de la couche l à la moyenne des deux conditions, soit

$$\frac{2}{d_l + d_{l-1}}.$$

Il faut noter que la dérivation des conditions ci-dessus pré-suppose l'utilisation de fonction d'activation centre autour de zéro. Ceci n'est pas le cas des fonctions sigmoïdes et des fonctions ReLU. Pour les premières, Glorot et al. (2011) utilise une expansion de Taylor autour de 0, $\sigma(x) = \frac{1}{2} + \frac{x}{4} + o(x^2)$ et répète l'analyse ci-dessus en incluant les termes de correction. Par conséquent, une couche l contenant des fonctions sigmoïdes devrait être initialisée avec une variance égale à

$$\frac{32}{d_l + d_{l-1}}.$$

Dans le cas des fonctions ReLU, He et al. (2015) remarque que l'analyse de Glorot et al. (2011) est inadaptée aux activations de moyennes non nulles. Cependant si les pré-activations a^l sont de moyennes nulles et centrées autour de zéro, alors

$$\mathbb{E}[(h^l)^2] = \mathbb{E}[\max(0, a^l)^2] = \mathbb{E}[(a^l)^2 \mathbb{1}[a^l \geq 0]] = 2\, Var(a^l).$$

Ainsi, pour garantir l'égalité des variances des activations pendant la phase de propagation il est suffisant d'initialiser les poids de la couche l avec une distribution dont la variance est l'inverse de la moitie du nombre de neurones entrants,

$$\frac{2}{d_{l-1}}.$$

Pour garantir l'égalité de la variance des gradients pendant la phase de rétro-propagation il est suffisant d'utiliser une distribution avec une variance égale à l'inverse de la moitie du nombre de neurones sortants,

$$\frac{2}{d_l}.$$

L'amélioration des méthodes d'initialisation couplée à l'utilisation d'activation des unités d'activation non bornées a mené a une consécration de l'apprentissage profond en permettant à He et al. (2015) de prouver empiriquement que les réseaux de neurones artificiels sont capables de surpasser les êtres humains sur une tache de classification d'images naturelles Russakovsky et al. (2014).

3.2.2.2. Normalisations des lots et autres.

Les méthodes d'initialisation cherchent à modifier l'état initial du réseau afin de faciliter son entraînement. Les méthodes de normalisation sont une continuation logique de cette idée. Au lieu de modifier l'état initial du réseau, elles cherchent à modifier l'état pendant l'entraînement. La normalisation des lots (Ioffe & Szegedy, 2015b) (BN) est à ce jour la méthode de normalisation de référence. L'idée est de standardiser les activations d'une couche l en estimant la moyenne et la variance des représentations à partir des instances présentes dans le mini- lot. Ainsi, pour un mini-mot de données de taille B, la moyenne et variance empirique sont estimées par,

$$\hat{\mu}^{(l)} = \frac{1}{B}\sum_{i=1}^{B} h^l_{i,\cdot}$$
$$\hat{\sigma^2}^{(l)} = \frac{1}{B}\sum_{i=1}^{B}(h^l_{i,\cdot} - \hat{\mu})^\top(h^l_{i,\cdot} - \hat{\mu}),$$

Pour normaliser les activations $h^{(l)}$,

$$\tilde{h}^{(l)} = \frac{h - \hat{\mu}}{\sqrt{\hat{\sigma^2} + \varepsilon}},$$

où ε est une constante strictement positive rajoutée pour des raisons de stabilité numérique. $\hat{\mu}^{(l)}$ et $\hat{\sigma^2}^{(l)}$ sont appellés statistiques de batches.

Pour pallier la perte d'expressivité induite par la normalisation deux paramètres, qui seront appris pendant l'entraînement, sont rajouter a l'expression ci-dessus. Ainsi, une couche de BN s'ecrit,

$$BN(h^l) = \gamma^{(l)}\,\tilde{h}^{(l)} + \beta^{(l)}. \tag{3.2.2}$$

Pendant la phase d'inférence, les statistiques des lots sont remplacés par des estimés sur l'ensemble des données ou des moyennes mobiles évaluées pendant la phase d'entraînement. Cette étape est essentielle pour garantir que le réseau de neurones ait un comportement déterministe et que les résultats soient indépendants des autres instances contenues dans le mini-lot.

Ioffe & Szegedy (2015b) montre empiriquement que la normalisation des lots réduit la sensibilité de l'entraînement à l'initialisation, ainsi que le temps d'entraînement nécessaire pour atteindre une performance donnée par un facteur de cinq, en plus réduire l'erreur de généralisation du réseau de neurones. Ioffe & Szegedy (2015b) motive la normalisation des lots en affirmant qu'elle permet de réduire les covariations internes des représentations. L'idée serait qu'elle agit essentiellement comme une matrice de dé-corrélations. Santurkar et al. (2018) montre empiriquement qu'au contraire la normalisation des lots augmente les covariations internes des représentations mais qu'elle facilite l'entraînement en contrôlant la constante de Lipschitz du réseau.

Il existe plusieurs autres méthodes de normalisation. Nous pouvons les diviser en deux catégories. La première est celle des méthodes qui agissent sur les activations Ba et al. (2016); Ulyanov et al. (2016); Wu & He (2018). La seconde est celle des méthodes agissant sur les poids. Salimans & Kingma (2016) propose de contrôler les variations des poids d'une couche l en divisant le poids correspondant de norme de Frobenius puis en le multipliant par un scalaire à apprendre pendant l'entrainement,

$$W^{(l)} = g\,\frac{V^{(l)}}{\left\|V^{(l)}\right\|_F}.$$

Dans le même esprit, Miyato et al. (2018) divise les poids par leur norme spectrale. Cette dernière est calculée en utilisant la méthode de puissance itérée (Mises & Pollaczek-Geiringer, 1929) pour calculer la plus grande valeur singulière de la matrice de poids.

3.2.2.3. Méthodes d'optimisation adaptatives.

Les réseaux de neurones artificiels sont typiquement entraînés par descente de gradient stochastique (SGD). La haute dimensionalité des données proscrit l'utilisation de méthodes de second ordre qui nécessitent l'inversion de matrices Hessiennes ayant un coût computionelle de l'ordre de $\mathcal{O}(d^3)$. L'utilisation des réseaux de neurones dans un régime de haut volume de données proscrit les méthodes de premier ordre stochastique car le coût en mémoire est de l'ordre de $\mathcal{O}(nd)$. De plus, le bruit inhérent au gradient stochastique améliorerait les capacités de généralisation du réseau (Goodfellow et al., 2016). L'entraînement par SGD introduit, cependant, des hyper-paramètres d'optimisation régulant le taux d'apprentissage. Les méthodes d'optimisation adaptatives visent à faciliter l'optimisation des réseaux de neurones en limitant la dépendance de l'entraînement sur les hyper-paramètres d'optimisation. Duchi et al. (2011) introduit l'algorithme Adagrad. Ce dernier, part de l'hypothèse que nombre d'ensemble de données ont une structure éparse. Dans ces conditions, les coordonnes éparses seront effectivement moins souvent mises à jour que celles qui ne le sont pas si le même taux d'apprentissage est utilisé pour toutes les coordonnées. L'algorithme Adagrad pondère chaque coordonnée par la racine carrée de la somme des carrées des gradients reçus. Plus formellement, soit g_t le gradient de la fonction de perte par rapport au paramètre θ l'itération t, la mise à jour par Adagrad s'écrit,

$$v_t = \sum_{i=1}^{2} g_t^2$$
$$\theta_{t+1} = \theta_t - \eta \frac{g_t}{\sqrt{v_t}}.$$

Un des inconvénients d'Adagrad est que le taux d'apprentissage adapté par paramètre à l'itération t dépend de l'ensemble des gradients obtenus jusque là. RMSprop (Tieleman & Hinton, 2017) corrige cette limitation en remplaçant la somme mobile d'Adagrad par une moyenne exponentielle mobile,

$$v_t = \beta v_{t-1} + (1-\beta)\, g_t^2$$
$$\theta_{t+1} = \theta_t - \eta \frac{g_t}{\sqrt{v_t}}.$$

Adam (Kingma & Ba, 2014) peut être compris comme une combinaison de RMSprop est de momentum,

$$
\begin{aligned}
m_t &= \beta_1 v_{t-1} + (1-\beta_1)\, g_t \\
v_t &= \beta_2 v_{t-1} + (1-\beta_2)\, g_t^2 \\
\hat{m}_t &= \frac{1}{1-\beta_1^t} \\
\hat{v}_t &= \frac{1}{1-\beta_2^t} \\
\theta_t &= \theta_{t-1} - \eta\, \frac{\hat{m}_t}{\sqrt{\hat{v}_t}}.
\end{aligned}
$$

Chapitre 4

Elements de theorie de l'information

Nous présentons dans ce chapitre l'ensemble des éléments de la théorie de l'information nécessaires à la compréhension, la dérivation et la construction de MINE. Nous commencerons par définir et établir les propriétés de l'entropie de Shannon dans le cas discret et différentiel. Nous procéderons ensuite à l'introduction des mesures d'information relatives communément appelées f-divergence ainsi que leur propriétés fondamentales. Nous introduirons par la suite deux approches pour obtenir la forme variationnelle des f-divergences avant de présenter la représentation variationnelle de Donsker-Varadhan pour divergence de Kullback-Leibler et de montrer comment ces deux représentations variationnelles sont liées. Enfin, nous définirons l'information mutuelle, quelque unes de ses propriétés fondamentales, ainsi que deux de ses représentations variationnelles. Ces dernières, en permettant la transformation d'un problème d'estimation de ratio de vraisemblance en problème d'optimisation sur une famille de fonction, constituent les ingrédients fondamentaux sur lesquels notre estimateur de l'information mutuelle est construit.

4.1. Entropie discrète

Définition 4.1.1 (Entropie discrète). Soit X un vecteur aléatoire discret, l'entropie de X est de finie comme,

$$H(X) = \mathbb{E}_{\mathbb{P}}[-\log \mathbb{P}(X)]. \tag{4.1.1}$$

Définition 4.1.2 (Entropie conditionnelle discrète). Soit X et Y deux vecteur aléatoires discrets, l'entropie conditionnelle de X sachant Y est défini comme,

$$H(X \mid Y) = \mathbb{E}_{\mathbb{P}_Y}[H(\mathbb{P}_{X|y})] \tag{4.1.2}$$

Théorème 4.1.3 (Propriétés de l'entropie).

(i) Positivité de l'entropie
(ii) La distribution uniforme maximise l'entropie
(iii) L'entropie est invariante aux bijections
(iv) Le conditionnement réduit l'entropie
(v) L'information en chaîne

Exemple 4.1.4 (Entropie d'une variable aléatoire Bernoulli). Soit $X \sim Bern(p)$, l'entropie de X est donnée par

$$H(X) = -p \log p - (1-p) \log(1-p) \tag{4.1.3}$$

Exemple 4.1.5. Entropie d'une variable aléatoire catégorique

Il est possible pour une variable aléatoire discrète d'avoir une entropie infinie. L'exemple qui suit montre comment construire une telle variable aléatoire. Nous sacrifions toute prétention à la rigueur afin de révéler l'intuition derrière cette construction.

Exemple 4.1.6 (Exemple d'entropie discrète infinie). Considérons une suite de variable discrète aléatoire $(X_k)_{k=1}^{\infty}$ uniformément distribue sur les intervalles disjoints $I_k = [2^k, 2^{k+1} - 1]$. Chacun de ces intervalles contient 2^k éléments. Nous avons donc, $H(X_k) = k$.

Considérons, une variable aléatoire discrète M avec support sur $[1, \infty) \cap \mathbb{N}$. Par définition, nous avons $\sum_{m=1}^{\infty} p_m = 1$.

Formons une variable aléatoire X_M, son entropie est alors,

$$H(X_M) = -\sum_{m=1}^{\infty} \sum_{k=2^m}^{2^{m+1}-1} \mathbb{P}(X_m = k \mid M = m) p_m \log(\mathbb{P}(X_m = k \mid M = m) p_m)$$

. Nous avons donc,

$$
\begin{aligned}
H(X_M) &= \sum_{m=1}^{\infty} p_m(H(X_m) - log(p_m)) \\
&= \sum_{m=1}^{\infty} m p_m + H(M) \\
&= E[M] + H[M]
\end{aligned}
$$

En prenant $p_m = \frac{1}{m} - \frac{1}{m+1}$, l'espérance de M diverge, et ce faisant l'entropie de X_M est infinie.

4.2. Entropie différentielle

L'entropie différentielle est définie par analogie à l'entropie discrète. L'entropie différentielle requiert l'existence d'une densité par rapport à la mesure de Lesbesgue.

Définition 4.2.1 (Entropie différentielle). Soit X une variable sur $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mathbb{P})$. Soit f_X la densité de X par rapport à la mesure de Lesbesgue. L'entropie différentielle de X est définie comme,

$$H(X) := \mathbb{E}[-\log(f_X(X))].$$

Similairement nous pouvons définir l'entropie conditionnelle différentielle,

Définition 4.2.2 (Entropie conditionnelle).

$$H(X \mid Y) := \mathbb{E}[-\log(f_{X|Y}(X))].$$

Établissons certaines propriétés fondamentales de l'entropie différentielle.

Théorème 4.2.3 (Propriétés de l'entropie différentielle).

Soit X une variable aléatoire sur $\mathcal{X} \subseteq \mathbb{R}^d$ $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{P})$.

(i) Si $\mathcal{X}$ est compacte alors $H(X) \leq \log(\lambda(\mathcal{X}))$ avec égalité si est seulement si $X \sim Unif(K)$

(ii) Soit $v \in \mathbb{R}^d$, alors $H(X+v) = H(X)$.

(iii) Soit A une matrice inversible, alors $H(AX) = H(X) + \log(|A|)$.

(iv) $H(X) = \sum_{i=1}^{d} H(X_i \mid X_{1:i-1})$

Démonstration. (i). Nous avons,

$$H(X) - \log(\lambda(\mathcal{X})) = \int_{\mathcal{X}} -\log(\frac{1}{\lambda(\mathcal{X}) p_X(x)}) p_X(x)\ \mathrm{d}x.$$

Par l'inégalité de Jensen, Nous avons donc,

$$H(X) - \log(\lambda(\mathcal{X})) \geq -\log(\frac{1}{\lambda(\mathcal{X})} \int_K \mathrm{d}x) = 0.$$

Ainsi nous avons $H(X) \leq \log(\lambda(K))$ avec égalité si et seulement si $p_X = frac1\lambda(K)$.
(ii). Par changement de variable, nous avons

$$H(X+v) = \int_{\mathcal{X}} -\log(p_X(y-v))p_X(y-v)\ \mathrm{d}y.$$

Par substitution $x = y - x$, nous concluons,

$$H(X+v) = H(X).$$

(iii). Par changement de variable nous avons,

$$H(AX) = \int_{\mathcal{X}} -\log(p_X(A^{-1}y\,|A|^{-1})p_X(A^{-1})y)\ \mathrm{d}y.$$

Par substitution $x = A^{-1}y$, nous concluons

$$H(AX) = H(X) + \log(|A|).$$

(iv). Par conditionnement répété nous avons,

$$p_X = p_{X_n|X_1,\ldots X_{n-1}} \cdots p_{X_2|X_1}\, p_{X_1}.$$

En composant par le la fonction $x \mapsto -\log(x)$ et en intégrant par rapport à la densité p_X sur $\mathcal{X}$, nous obtenons

$$H(X) = \sum_{i=1}^{d} H(X_i \mid X_{1:i-1}).$$

□

Calculons l'entropie différentielle d'un vecteur aléatoire Gaussien.
Exemple 4.2.4 (Entropie différentielle d'un vecteur aléatoire Gaussien).

Soit $X \sim \mathcal{N}_d(\mu, \Sigma)$. Nous avons,

$$\mathbb{E}[-\log(f_X(X))] = -\frac{1}{2}(d\log(2\pi) + \log(|\Sigma|) + \frac{1}{2}\underbrace{\int (x-\mu)^\top \Sigma^{-1}(x-\mu)\frac{\exp(\frac{(x-\mu)^\top\Sigma^{-1}(x-\mu)}{2})}{(2\pi)^{d/2}|\Sigma|^{\frac{1}{2}}}\,dx}_{:=I}.$$

En effectuant le changement de variable $y = \Sigma^{-\frac{1}{2}}(x - \mu)$, l'intégrale est réduite à

$$I = \int y^\top y \frac{\exp(y^\top y)}{(2\pi)^{\frac{d}{2}}} \, \mathrm{d}x.$$

En notant $Y \sim \mathcal{N}_d(0, I)$ et en utilisant la linéarité et la cyclicité de la trace, nous obtenons,

$$I = \mathrm{tr}(\mathbb{E}_{\mathcal{N}_d(0,I)}[y^\top y]) = \mathbb{E}_{\mathcal{N}_d(0,I)}[\mathrm{tr}(yy^\top)] = \mathrm{tr}(I) = d.$$

Nous avons donc,

$$H(X) = \frac{1}{2}\left(d \ln(2\pi e) + \log(\Sigma)\right) \qquad \text{nats,} \tag{4.2.1}$$

$$H(X) = \frac{1}{2}\left(d \log_2(2\pi e) + \log(\Sigma)\right) \qquad \text{bits.} \tag{4.2.2}$$

L'entropie jointe d'une Gaussienne croît donc linéairement avec la dimensionnalité indépendemment du choix des moments.

Il faut noter que le passage à support continu altère le comportement de l'entropie. Ainsi, l'entropie différentielle n'est pas nécessairement positive. Une distribution uniforme peut avoir une entropie négative si le volume de son support est inférieur à un.

4.2.0.1. Comportement pathologique de l'entropie différentielle.

Pour comprendre d'avantage l'entropie différentielle nous allons exposer trois exemples pathologiques. Notre optique est qu'à la résolution de ces pathologies nous augmenterons notre compréhension de l'entropie différentielle.

Exemple 4.2.5 (Entropie différentielle infinie).

Considérons une variable aléatoire X sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P})$, avec une densité $p_X(x) = \frac{1}{x(\log(x))^2}\mathbb{1}[x \geq e]$.

p_X est effectivement une densité. En effet, $\forall x \in [e, \infty)$, $\frac{1}{x(\log(x))^2} \geq 0$. De plus,

$$\begin{aligned}
\int_0^\infty p_X(x) \, \mathrm{d}x &= \int_e^\infty \frac{1}{x(\log(x))^2} \, \mathrm{d}x \\
&= \int_1^\infty \frac{1}{y^2} \, \mathrm{d}y, \quad \text{par substitution } y = \log(x) \\
&= 1.
\end{aligned}$$

L'entropie de X s'écrit après simplifications comme,

$$H(X) = \int_e^\infty \frac{1}{x \log(x)} \, \mathrm{d}x + \int_e^\infty \frac{\log(\log(x)^2)}{x \log(x)^2} \, \mathrm{d}x.$$

Le terme à droite de la somme étant positif, le terme à gauche devient une borne inférieure à $H(X)$. Or,

$$\begin{aligned}\int_e^{\infty} \frac{1}{x\log(x)}\,\mathrm{d}x &= \int_1^{\infty} \frac{1}{y}\,\mathrm{d}x, \quad \text{par substitution } y = \log(x)\\ &= +\infty.\end{aligned}$$

L'entropie de X est donc infinie.

Une condition suffisante pour éviter la divergence vers l'infini de l'entropie différentielle est un second moment borné. Toute variable aléatoire ayant une variance finie a une entropie différentielle finie par inégalité de holder. Le second moment de la variable aléatoire de l'exemple 4.2.5 diverge[1].

Une entropie infinie n'est pas nécessairement difficile à interpréter. En effet, un grand nombre de phénomènes observables sont suspectés de suivre des distributions avec fluctuation infinie: Les retours d'actifs côtés sur les marchés financiers (Fama, 1965), (Mandelbrot & Wallis, 1969) les minima des niveaux du fleuve du Nil ou le temps entre l'infection et l'apparition des symptômes du virus d'immunodéficience acquise (Mode & Sleeman, 2000).

L'entropie différentielle peut, cependant, prendre une valeur arbitrairement négative.

Exemple 4.2.6. Entropie différentielle égale à moins l'infini
Soit I une union d'intervalle $\{I_n\}_{n>1}$ disjoint et de longueur $\frac{1}{n^2\log(n)^2}$.
La longueur de I est finie. En effet, $\mid I \mid = \sum_{i=2}^{\infty} \frac{1}{n^2\log(n)^2}$. Or, la suite $n \mapsto \frac{1}{n^2\log(n^2)}$ étant strictement décroissante pour $n > 1$ et $\int_2^{\infty} \frac{1}{x^2\log(x)^2}\,\mathrm{d}x < \infty$, ainsi par comparaison série-intégrale, nous avons $\mid I \mid < \infty$.

Soit X une variable aléatoire sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ telle que sa fonction de densité soit donnée par,

$$p_X(x) = \begin{cases} \frac{n}{c} & \text{si } x \in I_n \\ 0 & \text{sinon .} \end{cases}$$

[1] Son premier moment diverge aussi ce qui est vérifiable en utilisant l'intégration par partie.

Pour que p_X soit une densité bien définie, il est nécessaire qu'elle intègre a 1 sur son domaine de définition. Déterminons donc la valeur de la constante de normalisation c. Nous avons,

$$\begin{aligned}\int_I p_X(x)\ \mathrm{d}x &= \sum_{n=2}^{\infty} \frac{n}{c} \int_{I_n} \mathrm{d}x \\ &= \sum_{n=2}^{\infty} \frac{n}{c} \frac{1}{n^2 \log(n)} \\ &= c^{-1} \sum_{n=2}^{\infty} \frac{1}{n \log(n)^2}.\end{aligned}$$

La suite $n \mapsto \frac{1}{n \log(n^2)}$ étant strictement décroissante pour $n > 1$ et $\int_2^{\infty} \frac{1}{x \log(x)^2}\ \mathrm{d}x < \infty$, encore une fois, par comparaison série-intégrale, il est suffisant donc que c soit égal à $\sum_{n=2}^{\infty} \frac{1}{n \log(n^2)}$.

L'entropie de X, s'écrit donc,

$$\begin{aligned}H(X) &= -\sum_{n=2}^{\infty} \int_{I_n} p_X(x) \log(p_X(x))\ \mathrm{d}x \\ &= -\sum_{n=2}^{\infty} \frac{n}{c} \log(\frac{n}{c}) |I_n|\ \mathrm{d}x \\ &= -c^{-1} \sum_{n=2}^{\infty} \log(\frac{n}{c}) \frac{1}{n \log(n)^2}\ \mathrm{d}x \\ &= c^{-1} \left(\log(c) \underbrace{\sum_{n=2}^{\infty} \frac{1}{n \log(n)^2}}_{=c} - \sum_{n}^{\infty} \frac{1}{n \log(n)} \right) \\ &= c^{-1} \left(\log(c) - \underbrace{\sum_{n=1}^{\infty} \frac{1}{n \log(n)}}_{=\infty} \right). \\ &= -\infty.\end{aligned}$$

Une condition suffisante pour éviter le cas pathologique d'une entropie différentielle arbitrairement négative est que la densité soit bornée. En effet, si pour tout $p_X(x) \leq K$ pour tout $x \in \mathcal{X}$ alors $H(X) \geq \log(\frac{1}{K})$.

4.3. f-Divergence

Les f-divergences introduites par Rényi et al. (1961) sont des fonctions convexes donnant une notion de proximité entre les distributions de probabilités.

Définition 4.3.1. f-Divergence

Soit $\mathbb{P}$ et $\mathbb{Q}$ deux mesures de probabilités sur un espace probabilisable $(\Omega, \mathcal{F})$, tel que $\mathbb{P} \ll \mathbb{Q}$. Pour toute fonction convexe $f : (0, \infty) \mapsto \mathbb{R}$, strictement convexe a 1, et telle que $f(1) = 0$, la f-divergence entre $\mathbb{P}$ et $\mathbb{Q}$ est définie comme,

$$D_f(\mathbb{P} \mid\mid \mathbb{Q}) = \mathbb{E}_{\mathbb{Q}}[f(d\mathbb{P}/d\mathbb{Q})]. \tag{4.3.1}$$

Dans l'éventualité la mesure $\mathbb{P}$ n'est pas absolument continue par rapport à la mesure $\mathbb{Q}$, il est souhaitable d'introduire une troisième mesure, μ sur $(\Omega, \mathcal{F})$ tel que $\mathbb{P} \ll \mu$ et $\mathbb{Q} \ll \mu$. Dans ce cas-ci, la définition ci-dessus peut-être généralisée

Définition 4.3.2. f-divergence généralise Soit $\mathbb{P}$, $\mathbb{Q}$, mu des mesures de probabilités sur un espace probabilisable $(\Omega, \mathcal{F})$, tel que $\mathbb{P} \ll \mu$ et $\mathbb{Q} \ll \mu$. Pour toute fonction convexe $f : (0, \infty) \mapsto \mathbb{R}$, strictement convexe à 1, et telle que $f(1) = 0$, la f-divergence entre $\mathbb{P}$ et $\mathbb{Q}$ est définie comme,

$$D_f(\mathbb{P} \mid\mid \mathbb{Q}) = \mathbb{E}_{\mathbb{Q}}[f(\frac{d\mathbb{P}/d\mu}{d\mathbb{Q}/d\mu})]. \tag{4.3.2}$$

Exemple 4.3.3 (Divergence de Kullblack-Leibler). En prenant comme fonction générative dans la définition 4.3.1, $u \mapsto u \log u$, Nous retrouvons la divergence de Kullblack-Leibler,

$$D_{KL}(\mathbb{P} \mid\mid \mathbb{Q}) = \mathbb{E}_{\mathbb{Q}}[d\mathbb{P}/d\mathbb{Q} \log d\mathbb{P}/d\mathbb{Q}] = \mathbb{E}_{\mathbb{P}}[\log d\mathbb{P}/d\mathbb{Q}]. \tag{4.3.3}$$

Exemple 4.3.4 (Variation totale).

En prenant comme fonction générative dans la définition 4.3.1, $u \mapsto \frac{1}{2} \mid u - 1 \mid$, nous obtenons la divergence de variation totale (TV).

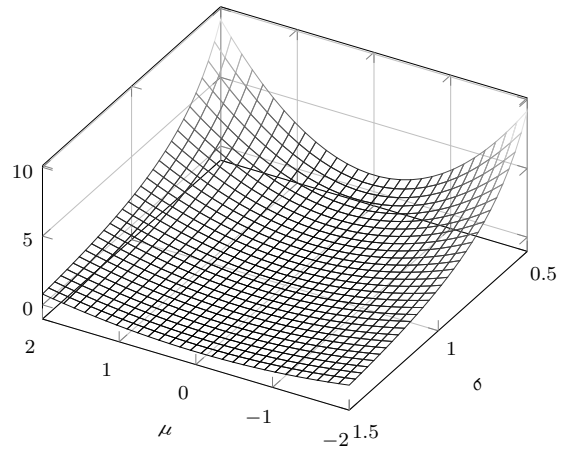


Fig. 4.1. La divergence de Kullblack-Leibler entre deux gaussiennes univariées.

$$\begin{aligned} D_{TV}(\mathbb{P} \,||\, \mathbb{Q}) &= \frac{1}{2}\,\mathbb{E}_{\mathbb{Q}}[|\ \frac{d\mathbb{P}}{d\mathbb{Q}} - 1\ |] \\ &= \frac{1}{2}\,\mathbb{E}_{\mathbb{Q}}[|\ \frac{d\ |\ \mathbb{P} - \mathbb{Q}\ |}{d\mathbb{Q}}] \end{aligned}$$

Assumons que $\mathbb{P}$ et $\mathbb{Q}$ sont absolument continues par rapport à une mesure de Lesbesque sur un ensemble $\mathcal{X} \subseteq \mathbb{R}^d$. La variation totale peut s'exprimer alors,

$$\begin{aligned} D_{TV}(\mathbb{P} \,||\, \mathbb{Q}) &= \frac{1}{2}\int [|\ p_X - q_X\ |]\,\mathrm{d}x \\ &= \frac{1}{2}\,\|p_X - q_X\|_{L_1(\mathcal{X})}\,. \end{aligned}$$

La divergence totale peut aussi être exprimée comme le minimum moyen de deux distributions sur leur domaine. En effet, en écrivant

$$\min(p_X, q_X) = \frac{p_X + q_X}{2} - \frac{|\ p_X - q_X\ |}{2},$$

et en l'intégrant sur le domaine des deux densités, nous avons

$$D_{TV}(\mathbb{P} \,||\, \mathbb{Q}) = 1 - \int \min(p_X(x), q_X(x))\,\mathrm{d}x.$$

Cette forme de la variation totale nous permet d'interpréter la distance de variation totale dans le cadre de l'erreur de Bayes et dans un test statistique à deux échantillons.

Exemple 4.3.5 (Divergence de Pearson χ^2). En prenant comme fonction générative dans la définition 4.3.1, $u \mapsto (u-1)^2$, nous obtenons la divergence χ^2,

$$D_{\chi^2}(\mathbb{P} \,||\, \mathbb{Q}) = \mathbb{E}_{\mathbb{Q}}[(d\mathbb{P}/d\mathbb{Q} - 1)^2]. \tag{4.3.4}$$

En assumant que $\mathbb{P}$ et $\mathbb{Q}$ sont absolument continues par rapport à une mesure de Lesbesque sur un ensemble $\mathcal{X} \subseteq \mathbb{R}^d$. La formule ci-dessous peut être d'avantage simplifiée,

$$D_{\chi^2}(\mathbb{P} \,||\, \mathbb{Q}) = \int_{\mathcal{X}} \frac{p_X(x)^2}{q_X(x)} \, \mathrm{d}x - 1. \tag{4.3.5}$$

Étudions certaines propriétés fondamentale des f-divergence.

Théorème 4.3.6. Propriétés élémentaires des f-divergence Soit $\mathbb{P}$ et $\mathbb{Q}$ deux mesures de probabilité sur l'espace mesurable $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$.

(i) $D_f(\mathbb{P} \,||\, QQ) \geq 0$ avec égalité si et seulement si $\mathbb{P} = \mathbb{Q}$.

(ii) La fonction $(\mathbb{P}, \mathbb{Q}) \mapsto D_f(\mathbb{P} \,||\, \mathbb{Q})$ est jointement convexe.

(iii) Soit $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ et $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ deux espaces mesurables. Soit $\mathbb{P}_X$ une mesure de probabilité sur $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. Soit $\mathbb{P}_{Y|X} : (\mathcal{X}, \mathcal{B}(X)) \mapsto$ et $\mathbb{Q}_{Y|X}$ deux opérateurs de probabilité conditionnelle (noyau Markovien). Définissons, $\mathbb{P}_Y = \mathbb{E}_{\mathbb{P}_X}[\mathbb{P}_{Y|X}]$ et $\mathbb{Q}_Y = \mathbb{E}_{\mathbb{P}_X}[\mathbb{Q}_{Y|X}]$. Alors,

$$D_f(\mathbb{P}_Y \,||\, \mathbb{Q}_Y) \leq D_f(\mathbb{P}_{Y|X} \,||\, \mathbb{Q}_{Y|X}).$$

Démonstration. (i)

La propriété suit par une simple application de l'inégalité de Jensen. En effet,

$$\begin{aligned} D_f(\mathbb{P} \,||\, \mathbb{Q}) &= \mathbb{E}_{\mathbb{Q}}[f(d\mathbb{P}/d\mathbb{Q})] \\ &\geq f(\mathbb{E}_{\mathbb{Q}}[d\mathbb{P}/d\mathbb{Q}]) \\ &= f(1) = 0. \end{aligned}$$

(iii)

Par définition nous avons,

$$D_f(\mathbb{P}_Y \,||\, \mathbb{Q}_Y) = D_f(\mathbb{E}_{\mathbb{P}_X}[\mathbb{P}_{Y|X}] \,||\, \mathbb{Q}_Y = \mathbb{E}_{\mathbb{P}_X}[\mathbb{Q}_{Y|X}]).$$

Par l'inégalité de Jensen,

$$D_f(\mathbb{P}_Y \,||\, \mathbb{Q}_Y) = \mathbb{E}_{\mathbb{P}_X}[D_f(\mathbb{P}_{Y|x} \,||\, \mathbb{Q}_{Y|x})].$$

Nous concluons alors,

$$D_f(\mathbb{P}_Y \,||\, \mathbb{Q}_Y) \leq D_f(\mathbb{P}_{Y|X} \,||\, \mathbb{Q}_{Y|X}).$$

□

4.3.0.1. Inégalités fondamentales et utiles.

Nous établirons dans le reste de cette section trois inégalités. La première est l'inégalité du traitement des données.

Nous utiliserons cette dernière pour établir la célèbre inégalité de Pinsker. Enfin, nous utiliserons l'inégalité de Pinsker pour établir l'inégalité de Györfi & van der Meulen (1987) qui sera nécessaire dans le chapitre 6 pour prouver les propriétés statistique de MINE.

Commençons par l'inégalité de traitement des données.

Théorème 4.3.7. inégalité de traitement des données

Considérons un canal générant Y sachant X au travers d'une chaîne de Markov distribue comme $\mathbb{P}_{Y|X}$, nous avons alors, Soit $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ et $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ deux espaces mesurables. Soit $\mathbb{P}_X$ et $\mathbb{Q}_X$ deux mesures de probabilité sur $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. Soit $\kappa_{Y|X} : (\mathcal{X}, \mathcal{B}(\mathcal{X})) \mapsto (\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ probabilité conditionnelle (noyau Markovien). Définissons, par marginalisation $\mathbb{P}_Y = \mathbb{E}_{\mathbb{P}_X}[\kappa_{Y|X}]$ et $\mathbb{Q}_Y = \mathbb{E}_{\mathbb{Q}_X}[\kappa_{Y|X}]$ deux mesure de probabilité sur $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$. Alors,

$$D_f(\mathbb{P}_Y \,||\, \mathbb{Q}_Y) \leq D_f(\mathbb{P}_X \,||\, \mathbb{Q}_X)$$

Démonstration.

Notons par $\mathbb{P}_{XY}$ et $\mathbb{Q}_{XY}$ les uniques mesures de probabilité sur $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}(\mathcal{X}) \otimes \mathcal{B}(\mathcal{Y}))$, tel que pour tout $A \in \mathcal{B}(\mathcal{X})$ et $B \in \mathcal{B}(\mathcal{Y})$,

$$\mathbb{P}_{XY} = \int_A \kappa(B \mid \cdot) \mathrm{d}\mathbb{P}_X$$
$$\mathbb{Q}_{XY} = \int_A \kappa(B \mid \cdot) \mathrm{d}\mathbb{Q}_X.$$

Les deux mesures ci-dessus sont définies par le même opérateur conditionnel. Nous, avons donc,

$$D_f(\mathbb{P}_X \,||\, \mathbb{Q}_X) = D_f(\mathbb{P}_{XY} \,||\, \mathbb{Q}_{XY}).$$

Par définition de la dérivée de Radon-Nikodym, nous savons que $\frac{d\mathbb{P}_Y}{d\mathbb{Q}_Y}$ est l'unique fonction dans $L_1(\mathbb{Q}_Y)$ tel que,

$$\forall A \in \mathcal{B}(\mathcal{Y}), \quad \mathbb{P}_Y(A) = \mathbb{E}_{\mathbb{Q}_Y}[\frac{d\mathbb{P}_Y}{d\mathbb{Q}_Y}\mathbb{1}_A].$$

Or, nous savons que, pour tout A dans $\mathcal{B}(\mathcal{Y})$

$$\begin{aligned}
\mathbb{E}_{\mathbb{Q}_Y}[\mathbb{E}_{\mathbb{Q}_{X|Y}}[\frac{d\mathbb{P}_{XY}}{d\mathbb{Q}_{XY}}]\mathbb{1}_A] &= \mathbb{E}_{\mathbb{Q}_A}[\mathbb{E}_{\mathbb{Q}_{X|Y}}[\frac{d\mathbb{P}_{XY}}{d\mathbb{Q}_{XY}}\mathbb{1}_A]] \quad \text{(par mesurabilté de } A\text{)} \\
&= \mathbb{E}_{\mathbb{Q}_{XY}}[\frac{d\mathbb{P}_{XY}}{d\mathbb{Q}_{XY}}\mathbb{1}_A] \\
&= \mathbb{E}_{\mathbb{P}_{XY}}[\mathbb{1}_A] \\
&= \mathbb{E}_{\mathbb{P}_Y}[\mathbb{1}_A] \\
&= \mathbb{P}_Y(A).
\end{aligned}$$

Ainsi,

$$\frac{d\mathbb{P}_Y}{d\mathbb{Q}_Y} = \mathbb{E}_{\mathbb{Q}_{X|Y}}[\frac{d\mathbb{P}_{XY}}{d\mathbb{Q}_{XY}}].$$

Par l'inégalité de Jensen nous avons,

$$\begin{aligned}
\mathbb{E}_{\mathbb{Q}_{X|Y}}[f(\frac{d\mathbb{P}_{XY}}{d\mathbb{Q}_{XY}})] &\geq f(\mathbb{E}_{\mathbb{Q}_{X|Y}}[\frac{d\mathbb{P}_{XY}}{d\mathbb{Q}_{XY}}]). \\
&= f(\frac{d\mathbb{P}_Y}{d\mathbb{Q}_Y}).
\end{aligned}$$

En intégrant par rapport à $\mathbb{Q}_Y$, nous obtenons,

$$D_f(\mathbb{P}_{XY} \,||\, \mathbb{Q}_{XY}) = \mathbb{E}_{\mathbb{Q}_Y}[f(\frac{d\mathbb{P}_Y}{d\mathbb{Q}_Y})].$$

Nous concluons donc,

$$D_f(\mathbb{P}_X \,||\, \mathbb{Q}_X) \geq D_f(\mathbb{P}_Y \,||\, \mathbb{Q}_Y)$$

□

L'inégalité de Pinsker borne la variation totale de distributions par leur divergence de Kullblack-Leibler. Remarquons que le choix d'opérateur conditionnel dans le théorème 4.3.7 peut être déterministe. En effet, pour établir l'inégalité de Pinsker nous commencerons par binariser les distributions. Il suffira par la suite d'établir la borne sur des distributions binaires.

Théorème 4.3.8 (Inégalité de Pinsker).

Soit $\mathbb{P}_X$ et $\mathbb{Q}_X$ deux mesures de probabilité sur le même espace probabilisable $\mathcal{X}, \mathcal{B}(\mathcal{X})$. Alors nous avons,

$$\sqrt{\frac{1}{2} D_{KL}(\mathbb{P}_X \,||\, \mathbb{Q}_X)} \geq D_{TV}(\mathbb{P} \,||\, \mathbb{Q})$$

Démonstration.

Soit $\mathcal{Y} = \{0, 1\}$. Construisons l'espace mesurable $(\mathcal{Y}, 2^{\mathcal{Y}})$. Soit A un élément arbitraire de $\mathcal{B}(\mathcal{X})$. Définissons un noyau Markovien,

$$\forall x \in \mathcal{X},\, \forall B \in 2^{\mathcal{Y}}, \quad \kappa(B \mid x) = \sum_{y \in B} \mathbb{1}_A(x)\delta_1(y) + \mathbb{1}_{A^c}(x)\delta_0(y). \tag{4.3.6}$$

Notons que le noyau κ dépend du choix d'évènement A. Ce noyau définit deux distributions de probabilités sur $(\mathcal{Y}, 2^{\mathcal{Y}})$

$$\forall B \in 2^{\mathcal{Y}}, \quad \mathbb{P}_Y(B) = \sum_{y \in B} \mathbb{P}_X(A)\delta_1(y) + \mathbb{P}_X(A^c)\delta_0(y)$$

$$\forall B \in 2^{\mathcal{Y}}, \quad \mathbb{Q}_Y(B) = \sum_{y \in B} \mathbb{Q}_X(A)\delta_1(y) + \mathbb{Q}_X(A^c)\delta_0(y).$$

Écrivons, $p = \mathbb{P}_X(A)$ et $q = \mathbb{Q}_X(A)$. $\mathbb{P}_Y$ et $\mathbb{Q}_Y$ définissent des variables aléatoires Bernoulli avec paramètre p et q respectivement. Par le théorème 4.3.7, nous avons,

$$D_{KL}(\mathbb{P}_X \,||\, \mathbb{Q}_X) \geq D_{KL}(\mathbb{P}_Y \,||\, \mathbb{Q}_Y).$$

Soit $g(u) = \frac{u-p}{p(1-p)}\,\mathrm{d}u$. nous avons,

$$\int_p^q g(u)\ \mathrm{d}u = p\log(\frac{p}{q}) + (1-p)\log(\frac{1-p}{1-q}).$$

Ainsi,

$$D_{KL}(\mathbb{P}_Y \,||\, \mathbb{Q}_Y) = \int_p^q g(u)\ \mathrm{d}u.$$

Considérons une expansion de deuxième ordre de g autour de p,

$$g(u) = \frac{t-p}{p(1-p)} + R_3(u)$$

$$R_3(u) = \int_p^u \frac{1}{k!} \frac{-24pt^3 + 36pt^2 - 24pt + 6p + 6t^4}{t^4(t^4 - 4t^3 + 6t^2 - 4t + 1}(u-t)^3\,\mathrm{d}t.$$

Après intégration, nous obtenons,

$$D_{KL}(\mathbb{P}_Y \,||\, \mathbb{Q}_Y) \geq \frac{1}{p(1-p)} \int_p^q (u-p)\ \mathrm{d}u.$$

En notant que la fonction $p \mapsto p(1-p)$ atteint son maximum a $p = \frac{1}{2}$. Nous obtenons,

$$\frac{1}{2} D_{KL}(\mathbb{P}_Y \mid\mid \mathbb{Q}_Y) \geq (p-q)^2.$$

En substituant, $p = \mathbb{P}_X(A)$ et $q = \mathbb{Q}_X(A)$,

$$\sqrt{\frac{1}{2} D_{KL}(\mathbb{P}_Y \mid\mid \mathbb{Q}_Y)} \geq \mid \mathbb{P}_X(A) - \mathbb{Q}_X(A) \mid .$$

et donc,

$$\sqrt{\frac{1}{2} D_{KL}(\mathbb{P}_X \mid\mid \mathbb{Q}_X)} \geq \mid \mathbb{P}_X(A) - \mathbb{Q}_X(A) \mid .$$

Le choix de l'évènement A étant arbitraire, en prenant le supremum sur $\mathcal{B}(\mathcal{X})$,

$$\sqrt{\frac{1}{2} D_{KL}(\mathbb{P}_X \mid\mid \mathbb{Q}_X)} \geq \sup_{A \in \mathcal{B}(\mathcal{X})} \mid \mathbb{P}_X(A) - \mathbb{Q}_X(A) \mid .$$

Nous concluons,

$$\sqrt{\frac{1}{2} D_{KL}(\mathbb{P}_X \mid\mid \mathbb{Q}_X)} \geq D_{TV}(\mathbb{P}_X \mid\mid \mathbb{Q}_X)$$

□

En utilisant l'inégalité de Pinsker pour établir l'inégalité suivante qui sera utiliser dans le chapitre 6 établir les propriétés statistiques de MINE.

Proposition 4.3.9 (Györfi & van der Meulen (1987)).

Soit $\mathbb{P}$ et $\mathbb{Q}$ deux mesures de probabilité sur l'espace mesurable $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, alors,

$$\mathbb{E}_{\mathbb{P}}[\mid \log(\frac{d\mathbb{P}}{d\mathbb{Q}}) \mid] \leq D_{KL}(\mathbb{P} \mid\mid \mathbb{Q}) + 2^{\frac{3}{2}} \sqrt{D_{KL}(\mathbb{P} \mid\mid \mathbb{Q})}$$

Démonstration.

Soit $A = \{x \in \mathcal{X} \mid \frac{d\mathbb{P}}{d\mathbb{Q}}(x) \geq 1\}$. Nous pouvons écrire,

$$\mathbb{E}_{\mathbb{P}}[\mid \log(\frac{d\mathbb{P}}{d\mathbb{Q}}) \mid] = \underbrace{\mathbb{E}_{\mathbb{P}}[\log(\frac{d\mathbb{P}}{d\mathbb{Q}}) \mathbb{1}A] + \mathbb{E}_{\mathbb{P}}[\log(\frac{d\mathbb{P}}{d\mathbb{Q}}) \mathbb{1}A^c]}_{= D_{KL}(\mathbb{P} \mid\mid \mathbb{Q})} - \mathbb{E}_{\mathbb{P}}[\log(\frac{d\mathbb{P}}{d\mathbb{Q}}) \mathbb{1}A^c] + \mathbb{E}_{\mathbb{P}}[-\log(\frac{d\mathbb{P}}{d\mathbb{Q}}) \mathbb{1}A^c].$$

Nous obtenons alors,

$$\mathbb{E}_{\mathbb{P}}[\mid \log(\frac{d\mathbb{P}}{d\mathbb{Q}}) \mid] = D_{KL}(\mathbb{P} \mid\mid \mathbb{Q}) + 2\, \mathbb{E}_{\mathbb{P}}[\log(\frac{d\mathbb{Q}}{d\mathbb{P}}) \mathbb{1}A^c].$$

Par la concavité de la fonction $x \mapsto log(x)$, nous avons,

$$\mathbb{E}_{\mathbb{P}}[\log(\frac{d\mathbb{Q}}{d\mathbb{P}})\mathbb{1}A^c] \leq \underbrace{\mathbb{E}_{\mathbb{P}}[|\ \log(\frac{d\mathbb{Q}}{d\mathbb{P}}) - 1\ |]}_{=D_{TV}(\mathbb{P}||\mathbb{Q})}.$$

Par le théorème 4.3.8, nous concluons,

$$\mathbb{E}_{\mathbb{P}}[|\ \log(\frac{d\mathbb{P}}{d\mathbb{Q}})\ |] \leq D_{KL}(\mathbb{P}\ ||\ \mathbb{Q}) + 2^{\frac{3}{2}}\sqrt{D_{KL}(\mathbb{P}\ ||\ \mathbb{Q})}$$

□

4.3.1. Transformation de Fenchel-Legendre

Le sujet des représentations duales des f-divergence est central pour l'élaboration de notre estimateur de l'information mutuelle. En effet, l'application d'une transformation de Fenchel au générateur f, permet de transformer un problème d'estimation de ratio de vraisemblance en un problème d'optimisation. Pour ce faire, il est nécessaire de se familiariser avec la transformation de Fenchel-Legendre.

Intuitivement, pour des fonctions suffisamment lisses, la transformée de Legendre de la fonction $f : \mathbb{R} \to \mathbb{R}$, notée f^*, est l'unique fonction, à une constante additive près, telle que les dérivées des fonctions f et f^* sont inverses l'une de l'autre,

$$Df = (Df^*)^{-1} \tag{4.3.7}$$

Définition 4.3.10. Transforme de Legendre
Soit $I \subseteq \mathbb{R}$, $f : I \to \mathbb{R}$, la transformée de Legendre de f, et la fonction $f^* : I^* \to \mathbb{R}$ tel que,

$$\forall x^* \in I^*, \quad f^*(x^*) = \sup_{x \in I} x^*x - f(x) \tag{4.3.8}$$

Avec $I^* = \{x^* \in \mathbb{R} \mid \sup_{x^* \in I} x^*x - f(x) < \infty\}$
Théorème 4.3.11. Propriétés fondamentales de la transformée de Legendre
Soit

(i) La transformée de Legendre d'une fonction convexe est convexe

(ii) La Transformation de Legendre est opération involutive, autrement dit,

$$f^{**} = f$$

4.3.1.1. Transformation de Fenchel-Legendre.

La transformation de Fenchel-Legendre peut être comprise comme une généralisation de la transformation de Legendre aux espaces de Banach.

4.3.2. Représentation dual des f-divergences

Nous présenterons ici deux approches pour obtenir la représentation duale des f-divergence. La première utilisera la transformation de Legendre du générateur f. La seconde utilisera la transformation de Fenchel-Legendre de la fonctionnelle,

$$\begin{aligned} L_1(\mathbb{Q}) &\to \mathbb{R} \\ r &\mapsto D_{f,\mathbb{Q}}(r) := \mathbb{E}_{f,\mathbb{Q}}[f(r)]. \end{aligned}$$

L'idée est de réaliser une f-divergence comme agissant sur l'espace de l'ensemble des dérivées de Radon-Nikodum de $\mathbb{Q}$.

Commençons par présenter la première approche.

Théorème 4.3.12. Représentation duale des f-Divergence (Legendre)

Soit $\mathbb{P}$ et $\mathbb{Q}$ deux mesures de probabilité sur un espace probabilisable $(\Omega, \mathcal{F})$, tel quel $\mathbb{P} \ll \mathbb{Q}$. Soit $\mathcal{T}$ en ensemble localement convexe de fonctions mesurable $T : \Omega \mapsto \mathbb{R}$ contenant $f'(d\mathbb{P}/d\mathbb{Q})$. Nous savons alors que toute f-divergence entre $\mathbb{P}$ et $\mathbb{Q}$ peut être représente comme,

$$D_f(\mathbb{P} \,||\, \mathbb{Q}) = \sup_{T \in \mathcal{T}} \mathbb{E}_{\mathbb{P}}[T] - \mathbb{E}_{\mathbb{Q}}[f^* \circ T]$$

Démonstration.

En utilisant le fait que $f^{**} = f$, nous avons,

$$D_f(\mathbb{P} \,||\, \mathbb{Q}) = \mathbb{E}_{\mathbb{Q}}[\sup_{t \in dom(f^*)} t\frac{d\mathbb{P}}{d\mathbb{Q}} - f^*(t)].$$

. Par l'inégalité de Jensen et la linéarité de l'espérance, nous obtenons,

$$D_f(\mathbb{P} \,||\, \mathbb{Q}) \geq \sup_{T \in \mathcal{T}} \mathbb{E}_{\mathbb{P}}[T] - \mathbb{E}_{\mathbb{Q}}[f^* \circ T]$$

La transformée de Legendre d'une fonction convexe étant convexe, la fonctionnelle $T \mapsto L[T] := \mathbb{E}_{\mathbb{P}}[T] - \mathbb{E}_{\mathbb{Q}}[f^* \circ T]$ est concave en T. La dérivée de Gateaux du fonctionnelle L,

satisfait les conditions de premier ordre si et seulement si,

$$\forall H \in \mathcal{T}, dF(T;H) = 0$$
$$\frac{d}{d\alpha} L(T + \alpha H)|_{\alpha=0} = 0$$
$$\mathbb{E}_{\mathbb{Q}}[H(\frac{d\mathbb{P}}{d\mathbb{Q}} - (f^*)' \circ T)] = 0$$

L'égalité devant tenir pour tout $H \in \mathcal{T}$, la condition de premier ordre n'est satisfaite que si et seulement si,

$$(f^*)' \circ T = \frac{d\mathbb{P}}{d\mathbb{Q}}$$

L'inverse de la dérivée de la transformée de Legendre, étant la dérivée de la fonction, et par la concavité de la fonctionnelle $T \mapsto L[T]$ nous avons donc,

$$T^{opt} = f'(\frac{d\mathbb{P}}{d\mathbb{Q}})$$

□

Utilisons le théorème 4.3.12 pour dériver les représentations duales de divergences de Kullblack-Leibler et de Pearson.

Présentons maintenant la seconde approche.

Théorème 4.3.13 (Représentation duale des f-Divergence). Soit $\mathbb{P}$ et $\mathbb{Q}$ deux mesures de probabilité sur un espace mesurable $(\mathcal{X}, \mathcal{B}(\mathcal{R}))$, toute f-divergence admet la représentation suivante,

$$D_f(\mathbb{P} \,||\, \mathbb{Q}) = \sup_{T \in L_\infty(\mathbb{Q})} \mathbb{E}_{\mathbb{P}}[T] - D^*_{r,\mathbb{Q}}(T)$$

Démonstration.

En rappelant que $L_1(\mathbb{Q})^* = L_\infty(\mathbb{Q})$, nous pouvons exprimer la conjuguée convexe de $D_{f,\mathbb{Q}}$ par,

$$Df, \mathbb{Q}^* = \sup_{r \in L_1(\mathbb{Q})} \mathbb{E}_{\mathbb{Q}}[rT] - D_{f,\mathbb{Q}}(r).$$

Par le Théorème de Fenchel-Moreau (Borwein & Lewis, 2010),

$$\begin{aligned} \forall r \in L_1(\mathbb{Q}), \quad D_{f,\mathbb{Q}}(r) &= D^{**}_{f,\mathbb{Q}}(r) \\ &= \sup_{T \in L_\infty(\mathbb{Q})} \mathbb{E}_{\mathbb{Q}}[rT] - D^*_{f,\mathbb{Q}}(T). \end{aligned}$$

□

Exemple 4.3.14 (Représentation duales par théorème 4.3.13). (i) Divergence de Kullblack-Leibler

Soit $\mathbb{P}$ et $\mathbb{Q}$ deux mesures de probabilité sur l'espace mesurable $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. Fixons $T \in L_\infty(\mathbb{Q})$ et $u \mapsto f(u) = u\log(u)$. Formons la fonctionnelle,

$$L_1(\mathbb{Q}) \to \mathbb{R} r \mapsto L[r] := \mathbb{E}_{\mathbb{Q}}[rT] - \mathbb{E}_{\mathbb{Q}}[r\log(r)]$$

La dérivée de Gateaux de la fonctionnelle L s'écrit,

$$\begin{aligned} \forall h \in L_1\mathbb{Q}, \quad dL(r;h) &= \frac{dL[r+\alpha h]}{d\alpha}|_{\alpha=0} \\ &= \mathbb{E}_{\mathbb{Q}}[h(T - 1 - \log(r))]. \end{aligned}$$

Par la concavité de la fonctionnelle $r \mapsto L[r]$, il est suffisante de vérifier les conditions de premier ordre pour maximiser L. $dL(r;h) = 0$ pour tout $h \in L_1(\mathbb{Q})$ si et seulement si,

$$(T - 1 - \log(r)) = 0.$$

Nous obtenons donc,

$$r^{opt} = e^{T-1}$$

Nous avons donc,

$$\begin{aligned} \forall \in L_\infty(\mathbb{Q}), \quad D^*_{f,\mathbb{Q}}(T) &= L[r^{opt}] \\ &= \mathbb{E}_{\mathbb{Q}}[Te^{T-1}] - \mathbb{E}_{\mathbb{Q}}[e^{T-1}\log(e^{T-1})] \\ &= \mathbb{E}_{\mathbb{Q}}[e^{T-1}]. \end{aligned}$$

Par le théorème 4.3.13, Pour conclure, la représentation duale de la divergence de Kullblack-Leibler est donnée par,

$$\begin{aligned} D_{KL}(\mathbb{P} \,||\, \mathbb{Q}) &= D_{f,\mathbb{Q}}(\frac{d\mathbb{P}}{d\mathbb{Q}}) \\ &= \sup_{T \in L_\infty(\mathbb{Q})} \mathbb{E}_{\mathbb{P}}[T] - \mathbb{E}_{\mathbb{Q}}[e^{T-1}]. \end{aligned}$$

(ii) divergence de Pearson χ^2.

Nous procédons de manière similaire a (i). Fixons $f(u) = (u-1)^2$. alors la fonctionnelle L s'écrit,

$$\forall r \in L_1(\mathbb{Q}), \quad L(r) = \mathbb{E}_{\mathbb{Q}}[rT] - \mathbb{E}_{\mathbb{Q}}[(r-1)^2].$$

Sa dérivée de Gateaux s'écrit,

$$\forall h \in L_1(\mathbb{Q}), dL(r;h) = \mathbb{E}_{\mathbb{Q}}[h(T - 2(r-1))].$$

Encore une fois par la concavité de $r \mapsto L[r]$, il est suffisant de vérifier les conditions de premier ordre, nous obtenons alors,

$$r^{opt} = \frac{T}{2} + 1$$

Nous avons donc,

$$\forall T \in L_\infty(\mathbb{Q}), D^*_{f,\mathbb{Q}}(T) = \mathbb{E}_{\mathbb{Q}}[\frac{T^2}{4} + T]$$

Nous concluons alors

$$\begin{aligned} D_{\chi^2}(\mathbb{P} \mid\mid \mathbb{Q}) &= D_{f,\mathbb{Q}}(\frac{d\mathbb{P}}{d\mathbb{Q}}) \\ &= \sup_{T \in L_\infty(\mathbb{Q})} \mathbb{E}_{\mathbb{P}}[T] - \mathbb{E}_{\mathbb{Q}}[\frac{T^2}{4} + T]. \end{aligned}$$

4.4. Représentations dites de Donsker-Varadhan

Nous présentons ici une différente représentation duale de la divergence de Kullblack-Leibler, dénommée représentation de Donsker-Varadhan. Suite à quoi nous procéderons à mettre en exergue les relations entre les représentations duales de Donsker-Varadhan et celles des f-Divergence avant de généraliser la représentation de Donsker-Varadhan à toute f-Divergence.

4.4.1. Représentation de Donsker-Varadhan

La représentation de Donsker-Varadhan est une représentation duale de la divergence de Kullback-Leibler populaire dans la littérature des larges déviations Donsker & Varadhan (1983).

Théorème 4.4.1. Représentation de Donsker-Varadhan de la divergence de Kullblack-Leibler Soit $\mathbb{P}$ et $\mathbb{Q}$ deux mesures de probabilité sur le même espace mesurable, $(\Omega, \mathcal{F})$. Soit $\mathcal{T} = \{T : \Omega \mapsto \mathbb{R} \mid \mathbb{E}_{\mathbb{Q}}[\exp \circ T] < \infty\}$. Nous avons alors,

$$D_{KL}(\mathbb{P} \mid\mid \mathbb{Q}) = \sup_{T \in \mathcal{T}} \mathbb{E}_{\mathbb{P}}[T] - \log \mathbb{E}_{\mathbb{Q}}[e^T]$$

Démonstration. Considérons la fonctionnelle,

$$\begin{aligned}\mathcal{T} &\mapsto \mathbb{R}\\ T &\mapsto L[T] := \mathbb{E}_{\mathbb{P}}[T] - \log \mathbb{E}_{\mathbb{Q}}[e^T].\end{aligned}$$

Par construction de $\mathcal{T}$ nous savons que $\log d\mathbb{P}/d\mathbb{Q} \in \mathcal{T}$. Ainsi, nous avons,

$$\begin{aligned}\sup_{T\in\mathcal{T}} L[T] &\geq L[d\mathbb{P}/d\mathbb{Q}]\\ &= \mathbb{E}_{\mathbb{P}}[\log d\mathbb{P}/d\mathbb{Q}]\\ &= D_{KL}(\mathbb{P} \,||\, \mathbb{Q}).\end{aligned}$$

Nous pouvons aussi écrire,

$$\begin{aligned}L[T] &= \mathbb{E}_{\mathbb{P}}[T] - \log \mathbb{E}_{\mathbb{Q}}[e^T]\\ &= \mathbb{E}_{\mathbb{P}}[\log \frac{e^T}{\mathbb{E}_{\mathbb{Q}}[e^T]}].\end{aligned}$$

Définissons une mesure sur $(\Omega, \mathcal{F})$, $\tilde{\mathbb{Q}}(A) := \mathbb{E}_Q[\frac{e^T}{\mathbb{E}_{\mathbb{Q}}[e^T]}\mathbb{1}_A]$, Nous avons donc

$$\begin{aligned}L[T] &= \mathbb{E}_{\mathbb{P}}[\log d\mathbb{P}/d\mathbb{Q}\, d\tilde{\mathbb{Q}}\mathbb{P}/d]\\ &= D_{KL}(\mathbb{P} \,||\, \mathbb{Q}) - D_{KL}(\tilde{\mathbb{Q}} \,||\, \mathbb{P}).\end{aligned}$$

Par la non-négativité de la divergence de Kullblack-Leibler, nous obtenons

$$\forall T \in \mathcal{T}, \quad L[T] \leq D_{KL}(\mathbb{P} \,||\, \mathbb{Q})$$

L'inégalité ci-dessus étant vérifiée pour tout $T \in \mathcal{T}$, nous concluons

$$D_{KL}(\mathbb{P} \,||\, \mathbb{Q}) \geq \sup_{T\in\mathcal{T}} \mathbb{E}_{\mathbb{P}}[T] - \log \mathbb{E}_{\mathbb{Q}}[e^T]$$

□

Remarquons que par une application de l'inégalité $log(x) \leq x - 1$ la représentation de duale de Kullblack-Leibler est pour tout point supérieure à celle obtenue dans l'exemple 4.3.14.

4.4.2. Généralisation aux f-divergences

La représentation de Donsker-Vardhan pour la divergence de Kullback-Leibler peut être généralisée à l'ensemble des f-divergence. Ruderman et al. (2012) montre qu'en considérant les f-divergence comme une famille de fonctionnelle agissant sur les dérivées de Radon-Nikodym d'une mesure de probabilité fixe, il est possible d'obtenir des représentations analogues à celle de Donsker-Varadhan en restreignant l'action de la divergence au sous-ensemble des dérivées de Radon-Nikodym qui s'intègre à un.

L'exemple suivant montre comment obtenir la représentation de Donsker-Varadhan pour la divergence de Kullback-Leibler en utilisant le théorème 4.4.1

Exemple 4.4.2 (Représentation restreinte de la divergence de Kullblack-Leibler). Soit $\mathbb{P}$ et $\mathbb{Q}$ deux mesures de probabilité sur l'espace mesurable $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. Fixons $T \in L_\infty(\mathbb{Q})$ et $u \mapsto f(u) = u\log(u)$. Soit $\mathcal{L} = \{r \in L_1(\mathbb{Q}) \mid \mathbb{E}_\mathbb{Q}[r] = 1\}$.

Formons la fonctionnelle $L : L_1(\mathbb{Q}) \to \mathbb{R}$, définie par,

$$L[r] := \begin{cases} \mathbb{E}_\mathbb{Q}[rT] - \mathbb{E}_\mathbb{Q}[r\log(r)], & \text{si } T \in \mathcal{L} \\ \infty, & \text{sinon.} \end{cases}$$

Notons par, $D^*_{f,\mathbb{Q}}|_\mathcal{L} : \mathcal{L} \to \mathbb{R}$ le duale restreint a $\mathcal{R}$. En utilisant les multiplicateurs de Lagrange (Lagrange, 1788), nous écrivons,

$$\forall r \in L_1(\mathbb{Q}), \quad \tilde{L}[r] := L[r] + \lambda\,(1 - \mathbb{E}_\mathbb{Q}[r]).$$

La dérive de Gateaux de la fonctionnelle L s'écrit,

$$\begin{aligned} \forall h \in L_1\mathbb{Q}, \quad d\tilde{L}(r;h) &= \frac{dL[r+\alpha h]}{d\alpha}|_{\alpha=0} \\ &= \mathbb{E}_\mathbb{Q}[h(T - 1 - \log(r) - \lambda)]. \end{aligned}$$

$d\tilde{L}(r;h) = 0$ pour tout $h \in L_1(\mathbb{Q})$ si et seulement si,

$$(T - 1 - \log(r) - \lambda) = 0.$$

Nous obtenons donc,

$$r^{opt} = e^{T-1-\lambda}.$$

Par la contrainte définie par $\mathcal{L}$, le multiplicateur de Lagrange λ vérifie,

$$1 + \lambda = \log(\mathbb{E}_\mathbb{Q}[e^T]).$$

Par la concavité de la fonctionnelle $r \mapsto \tilde{L}[r]$, nous avons,

$$\begin{aligned}\forall r \in \mathcal{L}, \quad D^*_{f,\mathbb{Q}}(r)|_{\mathcal{L}} &= (1+\lambda)\underbrace{e^{-(1+\lambda)}\mathbb{E}_{\mathbb{Q}}[e^T]}_{=1} \\ &= \log(\mathbb{E}_{\mathbb{Q}}[e^T])\end{aligned}$$

Notons que pour $\forall r \in L_1(\mathbb{Q}) \setminus \mathcal{L}, D^*_{f,\mathbb{Q}}(r) = -\infty$.

En prenant, $r = \frac{d\mathbb{P}}{d\mathbb{Q}}$, nous retrouvons la représentation de Donsker-Varadhan,

$$\begin{aligned}D_{KL}(\mathbb{P} \mid\mid \mathbb{Q}) &= \sup_{T \in L_\infty(\mathbb{Q})} \mathbb{E}_{\mathbb{Q}}[T] - D^*_{f,\mathbb{Q}}(\frac{d\mathbb{P}}{d\mathbb{Q}})|_{\mathcal{L}} \\ &= \sup_{T \in L_\infty(\mathbb{Q})} \mathbb{E}_{\mathbb{Q}}[T] - \log(\mathbb{E}_{\mathbb{Q}}[e^T])\end{aligned}$$

L'un des intérêts d'utiliser des contraintes sur l'espace des dérivées de Radon-Nikodym d'une mesure fixe est de permettre d'encoder un a priori directement dans la perte obtenue par la représentation duale de la divergence. En effet, La représentation de Donsker-Vardhan encode l'hypothèse que la dérive de Radon-Nikodym est égale en moyenne à 1 sous $\mathbb{Q}$. En réalité, cet a priori est naturellement vérifié dans la formulation primale de la divergence de Kullback-Leibler. Ce qui mène à l'égalité des représentations duales des f-divergences et de Donsker-Varadhan avec la représentation primale.

Nous pouvons, cependant, aller plus loin et définir d'autre a priori sur l'espace de dérive de Radon-Nikodym. Sous l'angle d'un problème de classification, ceci revient à spécifier des a priori sur le ratio de vraisemblance des distributions à séparer. Malheureusement, dans bien des cas, les contraintes supplémentaires mènent à des formes non-tractables de la représentation duale restreinte de la divergence. Par exemple, un a priori raisonnable serait d'imposer que le ratio de vraisemblance a une énergie borne, $\mathbb{E}_{\mathbb{Q}}[r^2] \leq K < \infty$. Dans ce cas ci, le duale restreint de la divergence Kullback-Leibler n'est pas tractable.

L'exemple suivant montre, en revanche, que la divergence de Pearson offre plus de possibilités.

Exemple 4.4.3 (Restriction avec la divergence de pearson χ^2).

Nous allons utiliser la divergence de Pearson χ^2 pour encoder l'a priori que le ratio de vraisemblance entre $\mathbb{P}$ et $\mathbb{Q}$ a un niveau d'énergie donné.

Soit $\mathbb{P}$ et $\mathbb{Q}$ deux mesures de probabilité sur l'espace mesurable $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. Fixons $T \in L_\infty(\mathbb{Q})$, $u \mapsto f(u) = (u-1)^2$, et $K > 0$. Soit $\mathcal{L} = \{r \in L_1(\mathbb{Q}) \mid \text{ et } \mathbb{E}_{\mathbb{Q}}[r^2] = K\}$.

Formons la fonctionnelle $L : L_1(\mathbb{Q}) \to \mathbb{R}$, définie par,

$$L[r] := \begin{cases} \mathbb{E}_{\mathbb{Q}}[rT] - \mathbb{E}_{\mathbb{Q}}[(r-1)^2], & \text{si } T \in \mathcal{L} \\ \infty, & \text{sinon.} \end{cases}$$

Le Lagrangien correspondant au problème ci-dessous s'écrit,

$$\tilde{L}[r] := L[r] + \lambda_2(\mathbb{E}_{\mathbb{Q}}[r^2] - K).$$

La dérive de Gateaux de $\tilde{L}$ est donnée par

$$\forall h \in L_1(\mathbb{Q}), \quad d\tilde{L}(r,h) = \mathbb{E}_{\mathbb{Q}}[h(T - 2(r-1) - 2\,\lambda r)].$$

La condition de premier ordre est vérifiée par,

$$r_{opt} = \frac{T+2}{2+2\lambda}$$

Le multiplicateur de Lagrange λ vérifie donc,

$$K = \frac{(T+2)^2}{(2\lambda+2)^2}$$

Nous avons donc, $\lambda = K^{-\frac{1}{2}}(\frac{T}{2} + 1 - K^{\frac{1}{2}})$ La conjuguée de la divergence restreinte est donc donnée par,

$$D^*_{f,\mathbb{Q}}(T)|_{\mathcal{L}} = \sqrt{K}\mathbb{E}_{\mathbb{Q}}[T] + 2\sqrt{K} - K - 1$$

. En choisissant, $r = \frac{d\mathbb{P}}{d\mathbb{Q}}$, nous obtenons,

$$D_{\chi^2}(\mathbb{P} \,||\, \mathbb{Q})|_{\mathcal{L}} = \sup_{T \in L_\infty(\mathbb{Q})} \mathbb{E}_{\mathbb{P}}[T] - \sqrt{K}\mathbb{E}_{\mathbb{Q}}[T] - 2\sqrt{K} + K + 1.$$

Il est remarquable que la divergence χ^2 restreinte à un niveau d'énergie 1 devient exactement un modèle d'énergie (LeCun et al., 2006).

4.5. Information mutuelle

Dans cette section nous définirons l'information mutuelle ainsi que ses représentations duales qui constituent les pièces maîtresses de MINE. Les résultats de cette section, découlent naturellement des sections précédentes de ce chapitre.

4.5.1. Définition et propriétés

Nous définirons l'information mutuelle en termes de divergence de Kullblack-Leibler[2]

Définition 4.5.1. Information Mutuelle par divergence Soit X et Y deux vecteurs aléatoires sur $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \mathbb{P})$ tel que $\{(x, x) \in Z\} \in \mathcal{B}(\mathcal{Z})$. Nous définissons l'information mutuelle entre X et Y comme,

(1)

$$I(X;Y) := D_{KL}(\mathbb{P}_{XY} \mid\mid \mathbb{P}_X \otimes \mathbb{P}_Y)$$

La définition permet de comprendre l'information mutuelle comme un score de force de dépendance entre deux variables aléatoires X et Y. En effet, si les variables X et Y sont indépendantes alors $\mathbb{P}_{XY} = \mathbb{P}_X \otimes \mathbb{P}_Y$ et donc la divergence de Kullblack-Leibler devient nulle.

De simples manipulations algébriques sur la définition 4.5.1 nous permettent de tirer des formes équivalentes de l'information mutuelle.

Définition 4.5.2. Définitions équivalentes de l'information Mutuelle

Soit X et Y deux vecteurs aléatoires sur $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \mathbb{P})$. De plus, nous supposons que $\mathbb{P}_{XY} << \lambda$. Nous avons donc,

(i) $I(X;Y) := H(X) - H(X \mid Y)$

(ii) $I(X;Y) := H(Y) - H(Y \mid X)$

(iii) $I(X;Y) := H(X) + H(Y) - H(X,Y)$

Considérons certaines propriétés fondamentales de l'information mutuelle.

Théorème 4.5.3. Propriétés de l'information mutuelle Soit X et Y deux variables aléatoires sur $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \mathbb{P})$.

(i) L'information mutuelle est non-négative: $I(X;Y) \geq 0$.

(ii) X et Y sont indépendantes si et seulement si $I(X;Y) = 0$.

(iii) L'information mutuelle est symétrique: $I(X;Y) = I(Y;X)$.

[2]Ceci n'est pas la définition la plus générale de l'information mutuelle. La définition la plus générale est due à Kolmogorov (Cover & Thomas, 2012). Elle utilise un supremum de partitions des supports des variables à comparer. Elle a l'avantage d'éviter tout problème de mesurabilité et de permettre la comparaison de variables discrètes et continues. Malheureusement, elle reste computationnellement intractable.

Démonstration.
(i) et (ii) suivent du fait que les f-divergence sont non-négatives et nulles si et seulement si les distributions sont égales.
(iii) est directe, □

4.5.2. Représentations duales de l'information mutuelle

Nous pouvons maintenant utiliser les théorème 4.3.12 et 4.4.1 pour offrir une représentation duale de l'information mutuelle. Ces représentations nous permettent d'envisager d'échanger l'épineux problème d'estimation du ratio de vraisemblance en un problème d'optimisation en utilisant les représentations duales de la divergence de Kullblack-Leibler.

En rappelant que le générateur correspondant à la divergence de Kullblack-Leibler est la fonction $u \mapsto f(u) := u \log u$, une application directe du théorème 4.3.12 nous permet de déduire le théorème suivant.

Théorème 4.5.4. Représentation duale de l'information mutuelle par f-divergence Soit X et Y deux vecteur aléatoires sur le même espace de probabilité $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \mathbb{P})$ tel que $\mathbb{P}_{XY} \ll \mathbb{P}_X \otimes \mathbb{P}_Y$. L'information mutuelle entre X et Y peut être écrite comme,

$$I(X;Y) = \sup_{T \in \mathcal{T}} \mathbb{E}_{\mathbb{P}_{XY}}[T] - E_{\mathbb{P}_X \otimes \mathbb{P}_Y}[e^{T-1}] \tag{4.5.1}$$

Alternativement, nous pouvons utiliser la représentation de Donsker-Varadhan de la divergence de Kullblack-Leibler. Le théorème 4.4.1 nous permet donc de conclure.

Théorème 4.5.5. Représentation duale de l'information mutuelle par Donsker-Varadhan Soit X et Y deux vecteurs aléatoires sur le même espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ tel que $\mathbb{P}_{XY} \ll \mathbb{P}_X \otimes \mathbb{P}_Y$. Soit $\mathcal{T} = T : \Omega \mapsto \mathbb{R} \mid \mathbb{E}_{\mathbb{P}_X \otimes \mathbb{P}_Y}[e^T] < \infty$ L'information mutuelle entre X et Y peut être écrite comme,

$$I(X;Y) = \sup_{T \in \mathcal{T}} \mathbb{E}_{\mathbb{P}_{XY}}[T] - \log E_{\mathbb{P}_X \otimes \mathbb{P}_Y}[e^T] \tag{4.5.2}$$

Échanger un problème d'estimation de ratio de densité par un problème d'optimisation constitue un progrès

considérable dans notre ambition d'estimer l'information mutuelle pour autant que le problème d'optimisation soit tractable. En effet, les deux théorèmes ci-dessus impliquent

un problème d'optimisation sur des espaces de fonctions non-paramétriques eux-mêmes intractables. Nous utiliserons donc le pouvoir expressif des réseaux de neurones profonds pour transformer ce problème d'optimisation sur un espace de fonctions intractable en un problème d'optimisation tractable sur l'espace des paramètres induit par un réseau de neurones.

Chapitre 5

Information mutuelle, estimation et apprentissage machine

L'objectif de ce chapitre est de présenter des estimateurs de l'information mutuelle ainsi que certaines de ses applications en apprentissage automatique.

5.1. Estimateurs de l'information mutuelle

L'estimation de l'information mutuelle est un sujet couvert par une large littérature. Pour une revue des méthodes classiques nous recommandons Paninski (2003). Notre objectif ici, n'est pas d'être exhaustif mais plutôt sélectif. Notre souci est de passer en revue certains estimateurs susceptibles de nous éclairer sur la nature des difficultés de l'estimation de l'information mutuelle mais aussi de révéler les apports que de meilleures représentations pourraient offrir afin de dépasser ces difficultés.

Une approche directe pour estimer l'information mutuelle consisterait à séparément estimer les densités jointe et marginales afin dévaluer l'information mutuelle.Fraser & Swinney (1986a) utilise l'estimation par noyaux pour estimer les densités nécessaires.

L'estimation de densité est un problème ardu en haute dimension. De plus, la division par des densités estimées favorise l'accumulation d'erreurs d'estimation et numériques.Darbellay & Vajda (1999) utilise des histogrammes avec un partitionnement dépendant des données afin d'estimer l'information mutuelle en calculant les fréquences relatives sur des partitions conditionnellement indépendantes. Wang et al. (2005) montre que cette méthode équivaut à estimer le ratio de vraisemblance entre la distribution jointe et le produit des marginales.

L'inconvénient majeur des approches à histogrammes adaptatives est le fléau de la dimensionnalité les rendant complètement inadaptées aux données en hautes dimensions.

Hulle (2005) utilise une expansion d'Edgeworth pour approximer l'entropie d'une distribution par celle d'une gaussienne à laquelle des termes de corrections de troisième et quatrième ordre sont rajoutés. L'expansion d'Edgeworth est d'utilité limitée en haute dimensionnalité car ses complexités statistique et computationnelle croissent exponentiellement avec la dimensionnalité. De plus, cette expansion devient très imprécise dès lors que la distribution de référence se retrouve trop éloignée d'une gaussienne.

L'information mutuelle peut être décomposée en une différence d'entropie, Kraskov et al. (2004) utilise cette décomposition ainsi que l'estimateur de l'entropie de Kozachenko-LeonenkoKozachenko & Leonenko (1987) pour proposer un estimateur de l'information mutuelle. L'estimateur de Kozachenko est lui-même construit autour de la méthode des k plus proches voisins, l'approche deKraskov et al. (2004) souffre donc des limites inhérentes à cette méthode. A savoir, la précision de l'estimation est fortement dépendante de l'hyperparamètre k et une complexité de l'ordre du carré de la taille de l'échantillon d'entraînement.

Bien que ce ne soit pas un un estimateur de l'information mutuelle, le critère d'indépendance de Hilbert-Schimit (HSIC) introduit dans Gretton et al. (2005) peut être certainement apparenté à l'estimation de l'information mutuelle lorsque cette dernière est considérée comme une mesure de dépendance. De plus, HSIC introduit l'importante idée que certaines représentations des données mènent à une meilleure estimation de la dépendance entre des variables aléatoires.

5.1.1. Approches directe

Nous appelons approche directe, toute estimation de l'information mutuelle utilisant des estimées des distributions jointe et marginale afin de construire un estimateur de leur ratio de vraisemblance avant de l'intégrer pour évaluer l'information mutuelle. Par exemple, si nous désirons estimer les densités jointe et marginale en utilisant un estimateur par noyaux gaussiens et des échantillons $(x_1, \ldots, x_N) \sim \mathbb{P}_X$, $(y_1, \ldots, y_N) \sim \mathbb{P}_X$ et

$((x_1, y_1), \ldots, (x_N, y_N)) \sim \mathbb{P}_{XY}$, nous obtenons

$$\hat{p}_X(x) = \frac{1}{n(2\pi\sigma^2)^{\frac{d}{2}}} \sum_{i=1}^{N} \exp(-\frac{\|x - x_i\|^2}{2\sigma^2})$$

$$\hat{p}_Y(y) = \frac{1}{n(2\pi\sigma^2)^{\frac{d}{2}}} \sum_{i=1}^{N} \exp(-\frac{\|y - y_i\|^2}{2\sigma^2})$$

$$\hat{p}_{XY}(xy) = \frac{1}{n(2\pi\sigma^2)^{d}} \sum_{i=1}^{N} \exp(-\frac{\|[x, y] - [x_i, y_i]\|}{2\sigma^2}),$$

pour estimer l'information mutuelle entre X et Y, en remplaçant les densités dans la définition de l'information mutuelle parleur estime empirique,

$$\widehat{MI}(X, Y) = \sum_{i=1}^{N} \log(\frac{\hat{p}_{XY}(x_i, y_i)}{\hat{p}_X(x_i)\hat{q}_Y(y_i)}).$$

Ce type d'approche souffre de plusieurs faiblesses. La première étant méthodologique. En essayant d'estimer la distribution jointe, les approches directes tentent de résoudre un problème plus général pour en résoudre un autre qui l'est moins. De plus, l'estimation de densité est un problème complexe dont la difficulté croît avec la dimensionnalité des données. De surcroît, la division par des estimés de densité accumule d'importantes erreurs d'estimation et numériques dans le cas où les densités ne sont bornées loin de zéro sur leurs domaines respectifs.

5.1.2. Approche par expansion d'Edgeworth

L'estimation de l'information mutuelle par expansion d'Edgeworth introduite par Hulle (2005) cherche, dans l'esprit d'une expansion de Taylor, à exprimer la densité jointe des données comme une gaussienne multipliée par des termes de correction. Il est judicieux de s'attarder sur cette méthode d'estimation car, bien qu'elle ne soit pas adaptée aux problèmes en haute dimension, elle met en exergue le caractère non-linéaire des dépendances entre les différentes variables aléatoires de la distribution jointe à travers les cumulants. Ces derniers, demeurent une quantité fondamentale en mécanique statistique. Commençons par introduire l'expansion d'Edgeworth d'une distribution multivariée.

5.1.2.1. Expansion d'Edgeworth. Nous utiliserons dans ce qui suit la notation tensorielle d'Einsten Einstein et al. (1916). Nous recommandons McCullagh (2018), un véritable tour

de force d'accessibilité et de clarté, pour une introduction aux méthodes tensorielles et à la notation d'Einstein appliquées aux statistiques multivariées.

Soit X un vecteur aléatoire sur un espace mesurable $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, Soit p_X et q_X les densités de X par rapport aux mesures de probabilités $\mathbb{P}$ et $\mathbb{Q}$ respectivement. Nous supposons que toutes les dérives partielles de p_X et q_X sont continues. Notons les cumulants de X sous $\mathbb{P}$ et $\mathbb{Q}$ par κ et λ respectivement. Nous notons la différence entre les deux cumulants par,

$$\begin{aligned}
\eta^i &:= \kappa^i - \lambda^i \\
\eta^{i,j} &:= \kappa^{i,j} - \lambda^{i,j} \\
\eta^{i,j,k} &:= \kappa^{i,j,k} - \lambda^{i,j,k} \\
&\dots
\end{aligned}$$

Les fonctions génératrices des cumulants de X sous $\mathbb{P}$ et $\mathbb{Q}$ admettent les expansions suivantes,

$$\begin{aligned}
K_{\mathbb{P}_X}(t) &= t_i\kappa^i + \frac{1}{2!}\, t_i t_j \kappa^{i,j} + \frac{1}{3!}\, t_i t_j t_k \kappa^{i,j,k} + \dots \\
K_{\mathbb{Q}_X}(t) &= t_i\lambda^i + \frac{1}{2!}\, t_i t_j \lambda^{i,j} + \frac{1}{3!}\, t_i t_j t_k \lambda^{i,j,k} + \dots
\end{aligned}$$

.

En écrivant,

$$M_{\mathbb{P}_X}(t) = \exp(K_{\mathbb{P}_X} - K_{\mathbb{Q}_X} + K_{\mathbb{Q}_X}),$$

La fonction caractéristique des moments de $\mathbb{P}_X$ peut s'écrire,

$$M_{\mathbb{P}_X}(t) = M_{\mathbb{Q}_X}(t)\,(1 + t_i\eta^i + \frac{1}{2!}t_i t_j \eta^{ij} + \frac{1}{3!}\, t_i t_j t_k \eta^{ijk} + \dots)$$

Pour obtenir une expansion de $p_X(x)$ autour de $q_X(x)$, il est suffisant d'inverser les fonction génératrices des moments. En utilisant le fait que la transformation de Fourier convertit la différenciation en multiplication de multinômes et en absorbant les changements de signe

dans les coefficients η, nous avons,

$$p_X(x) = q_X(x)(1 + \eta^i + \frac{1}{2!}\eta^{ij} + \frac{1}{3!}\eta^{ijk} + \dots) \tag{5.1.1}$$

Ainsi, la qualité de l'approximation de p_X par q_X dépend du nombre de termes inclus dans l'expansion.

5.1.3. Estimateur de l'information mutuelle

Hulle (2005) choisit une densité q_X gaussienne, est décompose l'entropie comme la différence entre l'entropie d'une gaussienne et la divergence de Kullback-Leibler entre la gaussienne et la distribution à estimer,

$$H_{\mathbb{P}}(X) = H_{\mathbb{Q}}(X) - D_{KL}(\mathbb{P} \,||\, \mathbb{Q}).$$

Le ratio des densités dans la divergence est approximé en tronquant l'expansion dans l'équation 5.1.1. En effet,

$$\frac{p_X(x)}{q_X(x)} \approx 1 + Z(x),$$

et donc,

$$D_{KL}(\mathbb{P} \,||\, \mathbb{Q}) \approx \int_{\mathbb{R}^d} (1 + Z(X)) \log(1 + Z(x)) g_x(x)\, dx$$

En simplifiant, nous obtenons et en substituant, nous obtenons

$$\begin{aligned} H_{\mathbb{P}}(X) = &\frac{1}{2}\log(|\ \Sigma\ |) + \frac{d}{2}\log(2\pi) + \frac{d}{2} \\ &- \int_{\mathbb{R}^d} (Z(x) + \frac{1}{2}Z(x)^2) g_x(x)\, dx. \end{aligned}$$

L'approximation de l'entropie ci-dessus peut être estimée par méthodes des moments, il est suffisant de remplacer les Σ ainsi que les moments présents dans Z par leur estimé empirique. L'information mutuelle est alors estimée en l'exprimant comme la différence entre la somme des entropies marginales et jointes.

L'approche par expansion d'Edgeworth a l'avantage d'exprimer explicitement les relations de dépendance mesurées par l'information mutuelle dans l'expansion du ratio de vraisemblance entre les densités et jointes et marginales. En effet, si les dépendances entre les variables aléatoires ne se produisent que deux à deux, le terme Z est nul, montrant ainsi que

la distribution gaussienne est suffisante pour capturer toute la dépendance entre les variables considérées.

Bien que théoriquement attrayante, l'approche par expansion d'Edgeworth n'est pas réaliste en haute dimension. Elle se heurte rapidement a un mur computationel et statistique. Le nombre de paramètres dans les cumulants croissant exponentiellement avec la dimensionnalité des données, elle devient rapidement impraticable pour les situations où la distribution génératrice des données s'éloigne d'une gaussienne.

5.1.4. Le critère d'indépendance de Hilbert-Schmidt

Gretton et al. (2005) introduit HSIC, une mesure d'indépendance basée sur la méthode des noyaux. Bien que HSIC ne soit pas un estimateur de l'information mutuelle, il marque néanmoins l'importance de l'utilisation de représentations de données pour évaluer les relations de dépendance statistique. HSIC cherche à calculer un analogue de la covariance, dit opérateur de covariance croisé, dans l'espace des représentations. L'espoir est que les représentations forment des statistiques suffisantes des données ce qui permettrait à l'opérateur de covariance croisée de capturer des dépendances non-linéaires, ce qui serait impossible dans l'espace des données. Commençons par comprendre l'opérateur de covariance croisée.

5.1.4.1. L'opérateur de covariance croisée. Soit $\mathcal{G}$ et $\mathcal{F}$ deux espaces d'Hilbert de fonctions à valeurs réelles sur les ensembles $\mathcal{X}$ et $\mathcal{Y}$, équipés de noyaux reproduisant Aronszajn (1950),

$$\begin{aligned} &\mathcal{X} \times \mathcal{X} \to \\ &(x, x^{'}) \mapsto K(x, x^{'}), \end{aligned}$$

et

$$\begin{aligned} &\mathcal{Y} \times \mathcal{Y} \to \mathbb{R} \\ &(y, y^{'}) \mapsto L(y, y^{'}). \end{aligned}$$

et des cartes de représentations ϕ et ψ.

Notre objectif,ici, est de généraliser l'opérateur de covariance non-centré,

$$f^{\top} \widehat{C}_{XY} g = \mathbb{E}[(f^{\top} X)(g^{\top} Y)]$$

Notons que le produit direct $\phi(X) \otimes \psi(Y)$ est une variable aléatoire sur $HS(\mathcal{G}, \mathcal{F})$. Supposons que, $\mathbb{E}_{\mathbb{P}_{XY}}[\|\phi(X) \otimes \psi(X)\|_{HS}] < \infty$. Alors, par le théorème de représentation de RiezRiesz (1914), nous savons que l'opérateur $\widehat{C}_{XY}$ peut être exprimé comme une intégrale et ce, uniquement, en considérant la fonctionnelle,

$$(\widehat{C}_{XY}, A)_{HS} = \mathbb{E}[(\phi(X) \otimes \psi(Y), A)_{HS}].$$

Ainsi, pour un élément $f \otimes g$ fixe, nous avons

$$\begin{aligned}
(f, \widehat{C}_{XY} g)_{\mathcal{F}} &= (\widehat{C}_{XY}, f \otimes g)_{HS} \\
&= \mathbb{E}[(\phi(X) \otimes \psi(Y), f \otimes g)_{HS}] \\
&= \mathbb{E}[(f, \phi(X))_{\mathcal{F}} (g, \psi(Y))_{\mathcal{G}}] \quad \text{par définition du produit directe} \\
&= \mathbb{E}[f(X) g(Y)], \quad \text{par la propriété reproduisante.}
\end{aligned}$$

En centrant l'opérateur $\widehat{C}_{XY}$, nous obtenons une expression pour l'opérateur de covariance centrée,

$$C_{XY} = \mathbb{E}[\phi(X) \otimes \psi(Y)] - \mathbb{E}[\phi(X)] \otimes \mathbb{E}[\psi(Y)].$$

Sous hypothèse que les espaces d'Hilbert $\mathcal{F}$ et $\mathcal{G}$ sont reproduisants, la plus grande valeur singulière de C_{XY}, ou de manière équivalant la norme d'Hilbert-Schmidt de l'opérateur, est nulle si et seulement si X et Y sont indépendants.

5.1.4.2. HSIC. HSIC est simplement défini comme la racine carrée de la norme d'Hilbert-Schmidt de l'opérateur C_{XY},

$$HSIC(\mathcal{F}, \mathcal{G}, \mathbb{P}_{XY}) := \sqrt{\|C_{XY}\|_{HS}}.$$

Un estimateur pour HSIC, est alors donné en passant à la mesure empirique,

$$\widehat{HSIC}(\mathcal{F}, cG, \mathbb{P}^n_{XY}) = \sqrt{\frac{1}{n^2} tr(KHLH)},$$

Avec $H := I = \frac{1}{n} 1_n 1_n^\top$. Nous noterons que cette estimateur est biaisé.

5.2. Application de l'information mutuelle à l'apprentissage machine

Nous présentons dans cette section quelques applications de l'information mutuelle à l'apprentissage automatique.

5.2.1. Le principe Infomax

Un critère naturel pour l'apprentissage de représentations est de maximiser la dépendance de ces dernières avec les données. Ce critère introduit par Linsker (1989) est le principe InfoMAX. Nous présentons dans cette section certaines application de ce principe montrant ainsi que la théorie de l'information et l'information mutuelle offre un cadre unifiant plusieurs algorithmes populaires d'apprentissage automatique.

Le principe InfoMAX permet d'interpréter l'analyse en composantes principales dans le cadre de la théorie de l'information Linsker (1988). Considérons un signal s gaussien, un source de bruit $\eta \sim \mathcal{N}_d(0, \alpha I)$ tel que la variable aléatoire $x = s + \eta$. Soit $W \in \mathcal{M}_{\mathbb{R}}(\mathbb{R}^d, \mathbb{R}^l)$ tel que $d < l$. Alors, l'information mutuelle $I(W^\top x; s)$ est maximisée par les d premières composantes principales.

L'analyse en composantes indépendantes(ICA) Hyvärinen & Oja (2000) cherche a expliquer et à décomposer un signal en combinaison de signaux indépendants.

L'information mutuelle fournit un critère naturel pour ICA . En effet, soit x et s deux vecteurs aléatoires correspondant respectivement aux observations et aux facteurs explicatifs cachés. Nous supposons que $x = Ws + \eta$, ou η est source de bruit. L'analyse en composantes indépendantes peut être exprimée comme la minimisation de l'information mutuelle $I(s_1, \ldots, s_d)$.

Plus récemment, l'information mutuelle à souvent été utilisée comme critère pour l'extraction de représentations non-supervisées. Par exemple, Hu et al. (2017) utilise l'information mutuelle entre les entrants et la sortie, caractérisée par une activation de type softmax, d'un réseau profond. En effet, la présence d'un modèle $p_\theta(y \mid x)$ permet de calculer l'information mutuelle par marginalisation et en utilisant la formule $I(X; Y) = H(Y) - H(Y \mid X)$. Cependant, ces choix limitent la topologie des variables extraites à une softmax. Hjelm et al. (2018) construit sur MINE pour proposer des critères d'apprentissage non-supervisé laissant libre choix à la topologie des variables extraites. Veličković et al. (2018) continue

dans cette lancée et utilise MINE pour offrir un critère d'apprentissage non-supervisé capable d'extraire des variables latentes ayant la topologie de graphe.

L'information mutuelle permet aussi de caractériser les autoencodeursVincent et al. (2010). En effet, considérons une variable aléatoire $X \sim \mathbb{P}_X$ ansi que sa reconstruction $\hat{X} \sim \mathbb{P}_{\hat{X}|X}$. L'objectif de l'autoencodeur est donc de maximiser l'information mutuelle entre X et sa reconstruction $\hat{X}$. Bien entendu, nous assumons que $\mathbb{P}_{\hat{X}|X}$ n'est pas déterministe sinon l'information mutuelle serait infinie,

$$\arg\max I(X;\hat{X}) = \arg\max -H(X \mid \hat{X}).$$

Or, nous avons $H(X \mid \hat{X}) \leq \mathbb{E}_{\mathbb{P}_{X\hat{X}}}[\log(p_{X\hat{X}}(X,\hat{X}))]$. Ainsi, il est suffisant de maximiser une minoration de l'information mutuelle définie par l'intégrale à droite dans l'inégalité précédente. Le choix de la distribution approximation q nous donne plusieurs critères habituels pour les autoencodeurs. A savoir, une distribution gaussienne pour l'erreur de reconstruction euclidienne ou une distribution Bernoulli multivariée pour l'erreur de reconstruction binaire.

Il est aussi possible d'interpréter les autoencodeurs variationels (Kingma & Welling, 2013) dans ce cadre.

En effet, factorisons le graphe probabiliste $X \to \hat{X}$ autour d'une variable latente Z,

$$X \to Z \to \hat{X}.$$

Une approche naturelle pour l'extraction de représentations non-triviales, adaptant le goulot d'étranglement de l'information (Tishby et al., 2000) à une tâche non-supervisée, serait de maximiser $I(X;\hat{X})$ tout en minimisant l'information $I(X;\hat{X})$.

L'autoencodeur variationel (VAE) est alors interprété comme une approximation de l'objectif suivant,

$$\arg\max I(X;\hat{X}) - \beta\, I(X;Z).$$

En effet, $I(X;\hat{X})$ serait maximisé en minimisant une erreur de reconstruction. $I(X;Z)$ peut être remarqué en observant que,

$$I(X;Z) = D_{KL}(\mathbb{P}_{Z|X} \mid\mid \mathbb{Q}_Z) - \underbrace{D_{KL}(\mathbb{P}_Z \mid\mid \mathbb{Q}_Z)}_{\geq 0}.$$

Nous obtenons donc l'inégalité variationnelle,

$$I(X;Z) \leq D_{KL}(\mathbb{P}_{Z|X} \,||\, \mathbb{Q}_Z).$$

Ce type d'approximation de l'information mutuelle est populaire dans la littérature d'apprentissage machine. Il a été utilisé par Barber & Agakov (2003) comme objectif dans des tâches de regroupement, par Alemi et al. (2016) pour proposer une implémentation variationnelle du goulot d'étranglement de l'information, ou par Mohamed & Rezende (2015) comme critère d'exploration en apprentissage par renforcement.

En définissant $\mathbb{Q}_Z$ comme gaussienne isotropique et en fixant le paramètre β a 1, nous dérivons la perte du VAE. En laissant β libre, nous dérivons le $\beta - VAE$ (Higgins et al., 2017).

Chapitre 6

Mutual Information Neural Estimation

Premier article.

Mutual Information Neural Estimation

par

Mohamed Ishmael Belghazi (MILA)[1], Aristide Baratin (MILA, McGill)[1], Sai Rajeswar (MILA)[1], Sherjil Ozair (MILA)[1], Yoshua Bengio (MILA, CIFAR, IVADO)[1], Aaron Courville (MILA, CIFAR)[1] et R Devon Hjelm (MILA, IVADO)[1]

([1]) Département d'Informatique et de Recherche Opérationnelle, Université de Montréal

Cet article a été soumis dans International Conference for Machine Learning.

Mes contributions et le rôle des coauteurs

L'idée de l'estimateur, ses applications, l'élaboration et l'implémentation des expériences sont les miennes. Les preuves des théorèmes de consistance et de complexité de l'échantillon viennent d'Aristide Baratin. Sai Rajeswar a contribué à la réalisation des expériences. Yoshua Bengio a attiré l'attention de l'équipe sur le biais des gradients de la représentation de Donsker-Varadhan et a suggéré l'utilisation de moyenne mobile comme solution. J'ai été tout au long guidé par Aaron Courville et Devon R. Hjelm. Aaron Courville, Devon R. Hjelm, et Aristide Baratin m'ont énormément aidé pour la rédaction de l'article.

Abstract. We argue that the estimation of mutual information between high dimensional continuous random variables can be achieved by gradient descent over neural networks. We present a Mutual Information Neural Estimator (MINE) that is linearly scalable in dimensionality as well as in sample size, trainable through back-prop, and strongly consistent. We present a handful of applications on which MINE can be used to minimize or maximize mutual information. We apply MINE to improve adversarially trained generative models. We also use MINE to implement the Information Bottleneck, applying it to supervised classification; our results demonstrate substantial improvement in flexibility and performance in these settings.
Keywords: Neural networks, mutual information, generative models.

1. Introduction

Mutual information is a fundamental quantity for measuring the relationship between random variables. In data science it has found applications in a wide range of domains and tasks, including biomedical sciences (Maes et al., 1997), blind source separation (BSS, e.g., independent component analysis, Hyvärinen et al., 2004), information bottleneck (IB, Tishby et al., 2000), feature selection (Kwak & Choi, 2002; Peng et al., 2005), and causality (Butte & Kohane, 2000).

Put simply, mutual information quantifies the dependence of two random variables X and Z. It has the form,

$$I(X;Z) = \int_{\mathcal{X}\times\mathcal{Z}} \log \frac{d\mathbb{P}_{XZ}}{d\mathbb{P}_X \otimes \mathbb{P}_Z} d\mathbb{P}_{XZ}, \tag{1.1}$$

where $\mathbb{P}_{XZ}$ is the joint probability distribution, and $\mathbb{P}_X = \int_{\mathcal{Z}} d\mathbb{P}_{XZ}$ and $\mathbb{P}_Z = \int_{\mathcal{X}} d\mathbb{P}_{XZ}$ are the marginals. In contrast to correlation, mutual information captures non-linear statistical dependencies between variables, and thus can act as a measure of true dependence (Kinney & Atwal, 2014).

Despite being a pivotal quantity across data science, mutual information has historically been difficult to compute (Paninski, 2003). Exact computation is only tractable for discrete variables (as the sum can be computed exactly), or for a limited family of problems where the probability distributions are known. For more general problems, this is not possible. Common approaches are non-parametric (e.g., binning, likelihood-ratio estimators based on support vector machines, non-parametric kernel-density estimators; see, Fraser & Swinney, 1986b; Darbellay & Vajda, 1999; Suzuki et al., 2008; Kwak & Choi, 2002; Moon et al., 1995;

Kraskov et al., 2004), or rely on approximate gaussianity of data distribution (e.g., Edgeworth expansion, Van Hulle, 2005). Unfortunately, these estimators typically do not scale well with sample size or dimension (Gao et al., 2014), and thus cannot be said to be general-purpose. Other recent works include Kandasamy et al. (2017); Singh & Póczos (2016); Moon et al. (2017).

In order to achieve a general-purpose estimator, we rely on the well-known characterization of the mutual information as the Kullback-Leibler (KL-) divergence (Kullback, 1997) between the joint distribution and the product of the marginals (i.e., $I(X;Z) = D_{KL}(\mathbb{P}_{XZ} \,||\, \mathbb{P}_X \otimes \mathbb{P}_Z)$). Recent work uses a dual formulation to cast the estimation of f-divergences (including the KL-divergence, see Nguyen et al., 2010) as part of an adversarial game between competing deep neural networks (Nowozin et al., 2016). This approach is at the cornerstone of generative adversarial networks (GANs, Goodfellow et al., 2014), which train a generative model without any explicit assumptions about the underlying distribution of the data.

In this paper we demonstrate that exploiting dual optimization to estimate divergences goes beyond the minimax objective as formalized in GANs. We leverage this strategy to offer a general-purpose parametric neural estimator of mutual information based on dual representations of the KL-divergence (Ruderman et al., 2012), which we show is valuable in settings that do not necessarily involve an adversarial game. Our estimator is scalable, flexible, and completely trainable via back-propagation. The contributions of this paper are as follows:

- We introduce the Mutual Information Neural Estimator (MINE), which is scalable, flexible, and completely trainable via back-prop, as well as provide a thorough theoretical analysis.
- We show that the utility of this estimator transcends the minimax objective as formalized in GANs, such that it can be used in mutual information estimation, maximization, and minimization.
- We apply MINE to palliate mode-dropping in GANs and to improve reconstructions and inference in Adversarially Learned Inference (ALI, Dumoulin et al., 2016) on large scale datasets.

- We use MINE to apply the Information Bottleneck method Tishby et al. (2000) in a continuous setting, and show that this approach outperforms variational bottleneck methods (Alemi et al., 2016).

2. Background

2.1. Mutual Information

Mutual information is a Shannon entropy-based measure of dependence between random variables. The mutual information between X and Z can be understood as the decrease of the uncertainty in X given Z:

$$I(X;Z) := H(X) - H(X \mid Z), \tag{2.1}$$

where H is the Shannon entropy, and $H(X \mid Z)$ is the conditional entropy of Z given X. As stated in Eqn. 1.1 and the discussion above, the mutual information is equivalent to the Kullback-Leibler (KL-) divergence between the joint, $\mathbb{P}_{XZ}$, and the product of the marginals $\mathbb{P}_X \otimes \mathbb{P}_Z$:

$$I(X,Z) = D_{KL}(\mathbb{P}_{XZ} \,||\, \mathbb{P}_X \otimes \mathbb{P}_Z), \tag{2.2}$$

where D_{KL} is defined as[1],

$$D_{KL}(\mathbb{P} \,||\, \mathbb{Q}) := \mathbb{E}_{\mathbb{P}} \left[\log \frac{d\mathbb{P}}{d\mathbb{Q}} \right]. \tag{2.3}$$

whenever $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{Q}$[2].

The intuitive meaning of Eqn. 2.2 is clear: the larger the divergence between the joint and the product of the marginals, the stronger the dependence between X and Z. This divergence, hence the mutual information, vanishes for fully independent variables.

2.2. Dual representations of the KL-divergence.

A key technical ingredient of MINE are dual representations of the KL-divergence. We will primarily work with the Donsker-Varadhan representation (Donsker & Varadhan, 1983),

[1] Although the discussion is more general, we can think of $\mathbb{P}$ and $\mathbb{Q}$ as being distributions on some compact domain $\Omega \subset \mathbb{R}^d$, with density p and q respect the Lebesgue measure λ, so that $D_{KL} = \int p \log \frac{p}{q} d\lambda$.

[2] and infinity otherwise.

which results in a tighter estimator; but will also consider the dual f-divergence representation (Keziou, 2003; Nguyen et al., 2010; Nowozin et al., 2016).

The Donsker-Varadhan representation. The following theorem gives a representation of the KL-divergence (Donsker & Varadhan, 1983):

Theorem 1 (Donsker-Varadhan representation). The KL divergence admits the following dual representation:

$$D_{KL}(\mathbb{P} \,||\, \mathbb{Q}) = \sup_{T:\Omega\to\mathbb{R}} \mathbb{E}_{\mathbb{P}}[T] - \log(\mathbb{E}_{\mathbb{Q}}[e^T]), \tag{2.4}$$

where the supremum is taken over all functions T such that the two expectations are finite.

Proof. See the Supplementary Material.

A straightforward consequence of Theorem 1 is as follows. Let $\mathcal{F}$ be any class of functions $T : \Omega \to \mathbb{R}$ satisfying the integrability constraints of the theorem. We then have the lower-bound[3]:

$$D_{KL}(\mathbb{P} \,||\, \mathbb{Q}) \geq \sup_{T\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}[T] - \log(\mathbb{E}_{\mathbb{Q}}[e^T]). \tag{2.5}$$

Note also that the bound is tight for optimal functions T^* that relate the distributions to the Gibbs density as,

$$d\mathbb{P} = \frac{1}{Z} e^{T^*} d\mathbb{Q}, \text{ where } Z = \mathbb{E}_{\mathbb{Q}}[e^{T^*}]. \tag{2.6}$$

The f-divergence representation. It is worthwhile to compare the Donsker-Varadhan representation to the f-divergence representation proposed in Nguyen et al. (2010); Nowozin et al. (2016), which leads to the following bound:

$$D_{KL}(\mathbb{P} \,||\, \mathbb{Q}) \geq \sup_{T\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}[T] - \mathbb{E}_{\mathbb{Q}}[e^{T-1}]. \tag{2.7}$$

Although the bounds in Eqns. 2.5 and 2.7 are tight for sufficiently large families $\mathcal{F}$, the Donsker-Varadhan bound is stronger in the sense that, for any fixed T, the right hand side of Eqn. 2.5 is larger[4] than the right hand side of Eqn. 2.7. We refer to the work by Ruderman et al. (2012) for a derivation of both representations in Eqns. 2.5 and 2.7 from the unifying perspective of Fenchel duality. In Section 3 we discuss versions of MINE based on these two representations, and numerical comparisons are performed in Section 4.

[3] The bound in Eqn. 2.5 is known as the compression lemma in the PAC-Bayes literature (Banerjee, 2006).

[4] To see this, just apply the identity $x \geq e \log x$ with $x = \mathbb{E}_{\mathbb{Q}}[e^T]$.

3. The Mutual Information Neural Estimator

In this section we formulate the framework of the Mutual Information Neural Estimator (MINE). We define MINE and present a theoretical analysis of its consistency and convergence properties.

3.1. Method

Using both Eqn. 2.2 for the mutual information and the dual representation of the KL-divergence, the idea is to choose $\mathcal{F}$ to be the family of functions $T_\theta : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ parametrized by a deep neural network with parameters $\theta \in \Theta$. We call this network the statistics network. We exploit the bound:

$$I(X;Z) \geq I_\Theta(X,Z), \tag{3.1}$$

where $I_\Theta(X,Z)$ is the neural information measure defined as

$$I_\Theta(X,Z) = \sup_{\theta \in \Theta} \mathbb{E}_{\mathbb{P}_{XZ}}[T_\theta] - \log(\mathbb{E}_{\mathbb{P}_X \otimes \mathbb{P}_Z}[e^{T_\theta}]). \tag{3.2}$$

The expectations in Eqn. 3.2 are estimated using empirical samples[5] from $\mathbb{P}_{XZ}$ and $\mathbb{P}_X \otimes \mathbb{P}_Z$ or by shuffling the samples from the joint distribution along the batch axis. The objective can be maximized by gradient ascent.

It should be noted that Eqn. 3.2 actually defines a new class information measures, The expressive power of neural network insures that they can approximate the mutual information with arbitrary accuracy.

In what follows, given a distribution $\mathbb{P}$, we denote by $\hat{\mathbb{P}}^{(n)}$ as the empirical distribution associated to n i.i.d. samples.

Definition 2 (Mutual Information Neural Estimator (MINE)). Let $\mathcal{F} = \{T_\theta\}_{\theta \in \Theta}$ be the set of functions parametrized by a neural network. MINE is defined as,

$$\widehat{I(X;Z)}_n = \sup_{\theta \in \Theta} \mathbb{E}_{\mathbb{P}^{(n)}_{XZ}}[T_\theta] - \log(\mathbb{E}_{\mathbb{P}^{(n)}_X \otimes \hat{\mathbb{P}}^{(n)}_Z}[e^{T_\theta}]). \tag{3.3}$$

Details on the implementation of MINE are provided in Algorithm 1. An analogous definition and algorithm also hold for the f-divergence formulation in Eqn. 2.7, which we refer to as MINE-f. Since Eqn. 2.7 lower-bounds Eqn. 2.5, it generally leads to a looser

[5] Note that samples $\bar{x} \sim \mathbb{P}_X$ and $\bar{z} \sim \mathbb{P}_Z$ from the marginals are obtained by simply dropping x, z from samples $(\bar{x}, z)$ and $(x, \bar{z}) \sim \mathbb{P}_{XZ}$.

Algorithm 1 MINE

$\theta \leftarrow$ initialize network parameters

repeat

Draw b minibatch samples from the joint distribution:

$(\boldsymbol{x}^{(1)}, \boldsymbol{z}^{(1)}), \ldots, (\boldsymbol{x}^{(b)}, \boldsymbol{z}^{(b)}) \sim \mathbb{P}_{XZ}$

Draw b samples from the Z marginal distribution:

$\bar{\boldsymbol{z}}^{(1)}, \ldots, \bar{\boldsymbol{z}}^{(b)} \sim \mathbb{P}_Z$

Evaluate the lower-bound:

$\mathcal{V}(\theta) \leftarrow \frac{1}{b}\sum_{i=1}^{b} T_\theta(\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(i)}) - \log(\frac{1}{b}\sum_{i=1}^{b} e^{T_\theta(\boldsymbol{x}^{(i)}, \bar{\boldsymbol{z}}^{(i)})})$

Evaluate bias corrected gradients (e.g., moving average):

$\widehat{G}(\theta) \leftarrow \widetilde{\nabla}_\theta \mathcal{V}(\theta)$

Update the statistics network parameters:

$\theta \leftarrow \theta + \widehat{G}(\theta)$

until convergence

estimator of the mutual information, and numerical comparisons of MINE with MINE-f can be found in Section 4. However, in a mini-batch setting, the SGD gradients of MINE are biased. We address this in the next section.

3.2. Correcting the bias from the stochastic gradients

A naive application of stochastic gradient estimation leads to the gradient estimate:

$$\widehat{G}_B = \mathbb{E}_B[\nabla_\theta T_\theta] - \frac{\mathbb{E}_B[\nabla_\theta T_\theta \, e^{T_\theta}]}{\mathbb{E}_B\left[e^{T_\theta}\right]}. \tag{3.4}$$

where, in the second term, the expectations are over the samples of a minibatch B, leads to a biased estimate of the full batch gradient[6].

Fortunately, the bias can be reduced by replacing the estimate in the denominator by an exponential moving average. For small learning rates, this improved MINE gradient estimator can be made to have arbitrarily small bias.
We found in our experiments that this improves all-around performance of MINE.

[6]From the optimization point of view, the f-divergence formulation has the advantage of making the use of SGD with unbiased gradients straightforward.

3.3. Theoretical properties

In this section we analyze the consistency and convergence properties of MINE. All the proofs can be found in the Supplementary Material.

3.3.1. Consistency. MINE relies on a choice of (i) a statistics network and (ii) n samples from the data distribution $\mathbb{P}_{XZ}$.

Definition 3 (Strong consistency). The estimator $\widehat{I(X;Z)}_n$ is strongly consistent if for all $\epsilon > 0$, there exists a positive integer N and a choice of statistics network such that:

$$\forall n \geq N, \quad |I(X,Z) - \widehat{I(X;Z)}_n| \leq \epsilon, \ a.e.$$

where the probability is over a set of samples.

In a nutshell, the question of consistency is divided into two problems: an approximation problem related to the size of the family, $\mathcal{F}$, and an estimation problem related to the use of empirical measures. The first problem is addressed by universal approximation theorems for neural networks (Hornik, 1989). For the second problem, classical consistency theorems for extremum estimators apply (Van de Geer, 2000) under mild conditions on the parameter space.

This leads to the two lemmas below. The first lemma states that the neural information measures $I_\Theta(X,Z)$, defined in Eqn. 3.2, can approximate the mutual information with arbitrary accuracy:

Lemma 4 (approximation). Let $\epsilon > 0$. There exists a neural network parametrizing functions T_θ with parameters θ in some compact domain $\Theta \subset \mathbb{R}^k$, such that

$$|I(X,Z) - I_\Theta(X,Z)| \leq \epsilon, \ a.e.$$

The second lemma states the almost sure convergence of MINE to a neural information measure as the number of samples goes to infinity:

Lemma 5 (estimation). Let $\epsilon > 0$. Given a family of neural network functions T_θ with parameters θ in some bounded domain $\Theta \subset \mathbb{R}^k$, there exists an $N \in \mathbb{N}$, such that

$$\forall n \geq N, \quad |\ \widehat{I(X;Z)}_n - I_\Theta(X,Z)\ | \leq \epsilon, \ a.e. \tag{3.5}$$

Combining the two lemmas with the triangular inequality, we have,

Theorem 6. MINE is strongly consistent.

3.3.2. Sample complexity. In this section we discuss the sample complexity of our estimator. Since the focus here is on the empirical estimation problem, we assume that the mutual information is well enough approximated by the neural information measure $I_\Theta(X,Z)$. The theorem below is a refinement of Lemma 5: it gives how many samples we need for an empirical estimation of the neural information measure at a given accuracy and with high confidence.

We make the following assumptions: the functions T_θ are L-Lipschitz with respect to the parameters θ, and both T_θ and e^{T_θ} are M-bounded (i.e., $|T_\theta|, e^{T_\theta} \leq M$). The domain $\Theta \subset \mathbb{R}^d$ is bounded, so that $\|\theta\| \leq K$ for some constant K. The theorem below shows a sample complexity of $\widetilde{O}\left(\frac{d \log d}{\epsilon^2}\right)$, where d is the dimension of the parameter space.

Theorem 7. Given any values ϵ,δ of the desired accuracy and confidence parameters, we have,

$$\Pr\left(|\widehat{I(X;Z)}_n - I_\Theta(X,Z)| \leq \epsilon\right) \geq 1-\delta, \tag{3.6}$$

whenever the number n of samples satisfies

$$n \geq \frac{2M^2(d\log(16KL\sqrt{d}/\epsilon) + 2dM + \log(2/\delta))}{\epsilon^2}. \tag{3.7}$$

4. Empirical comparisons

Before diving into applications, we perform some simple empirical evaluation and comparisons of MINE. The objective is to show that MINE is effectively able to estimate mutual information and account for non-linear dependence.

4.1. Comparing MINE to non-parametric estimation

We compare MINE and MINE-f to the k-NN-based non-parametric estimator found in Kraskov et al. (2004). In our experiment, we consider multivariate Gaussian random variables, X_a and X_b, with componentwise correlation, $corr(X_a^i, X_b^j) = \delta_{ij}\rho$, where $\rho \in (-1,1)$ and δ_{ij} is Kronecker's delta. As the mutual information is invariant to continuous bijective transformations of the considered variables, it is enough to consider standardized Gaussians marginals. We also compare MINE (using the Donsker-Varadhan representation in Eqn. 2.5) and MINE-f (based on the f-divergence representation in Eqn. 2.7).

Our results are presented in Figs. 6.1. We observe that both MINE and Kraskov's estimation are virtually indistinguishable from the ground truth when estimating the mutual information between bivariate Gaussians. MINE shows marked improvement over Krakov's when estimating the mutual information between twenty dimensional random variables. We also remark that MINE provides a tighter estimate of the mutual information than MINE-f.

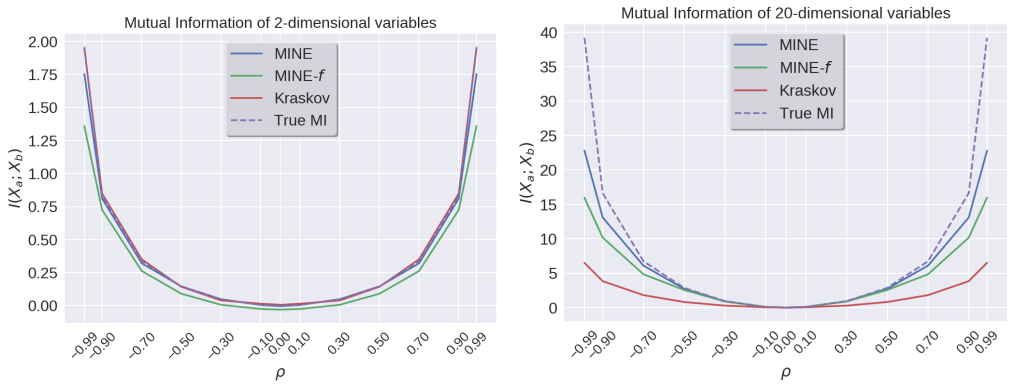


Fig. 6.1. Mutual information between two multivariate Gaussians with component-wise correlation $\rho \in (-1,1)$.

4.2. Capturing non-linear dependencies

An important property of mutual information between random variables with relationship $Y = f(X) + \sigma \odot \epsilon$, where f is a deterministic non-linear transformation and ϵ is random noise, is that it is invariant to the deterministic nonlinear transformation, but should only depend on the amount of noise, $\sigma \odot \epsilon$. This important property, that guarantees the quantification dependence without bias for the relationship, is called equitability (Kinney & Atwal, 2014). Our results (Fig. 6.2) show that MINE captures this important property.

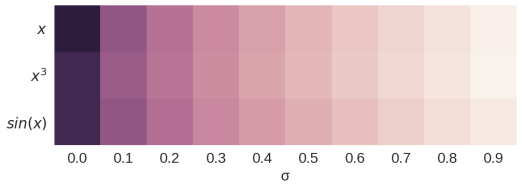


Fig. 6.2. MINE is invariant to choice of deterministic nonlinear transformation. The heatmap depicts mutual information estimated by MINE between 2-dimensional random variables $X \sim \mathcal{U}(-1,1)$ and $Y = f(X) + \sigma \odot \epsilon$, where $f(x) \in \{x, x^3, sin(x)\}$ and $\epsilon \sim \mathcal{N}(0, I)$.

5. Applications

In this section, we use MINE to present applications of mutual information and compare to competing methods designed to achieve the same goals. Specifically, by using MINE to maximize the mutual information, we are able to improve mode representation and reconstruction of generative models. Finally, by minimizing mutual information, we are able to effectively implement the information bottleneck in a continuous setting.

5.1. Maximizing mutual information to improve GANs

Mode collapse (Che et al., 2016; Dumoulin et al., 2016; Donahue et al., 2016; Salimans et al., 2016; Metz et al., 2017; Saatchi & Wilson, 2017; Nguyen et al., 2017; Lin et al., 2017; Ghosh et al., 2017) is a common pathology of generative adversarial networks (GANs, Goodfellow et al., 2014), where the generator fails to produces samples with sufficient diversity (i.e., poorly represent some modes).

GANs as formulated in Goodfellow et al. (2014) consist of two components: a discriminator, $D : \mathcal{X} \to [0, 1]$ and a generator, $G : \mathcal{Z} \to \mathcal{X}$, where $\mathcal{X}$ is a domain such as a compact subspace of $\mathbb{R}^n$. Given $Z \in \mathcal{Z}$ follows some simple prior distribution (e.g., a spherical Gaussian with density, $\mathbb{P}_Z$), the goal of the generator is to match its output distribution to a target distribution, $\mathbb{P}_X$ (specified by the data samples). The discriminator and generator are optimized through the value function,

$$\min_G \max_D V(D,G) := \mathbb{E}_{\mathbb{P}_X}[D(X)] + \mathbb{E}_{\mathbb{P}_Z}[\log(1 - D(G(Z))]. \tag{5.1}$$

A natural approach to diminish mode collapse would be regularizing the generator's loss with the neg-entropy of the samples. As the sample entropy is intractable, we propose to use the mutual information as a proxy.

Following Chen et al. (2016), we write the prior as the concatenation of noise and code variables, $Z = [\boldsymbol{\epsilon}, \boldsymbol{c}]$. We propose to palliate mode collapse by maximizing the mutual information between the samples and the code. $I(G([\boldsymbol{\epsilon}, \boldsymbol{c}]); \boldsymbol{c}) = H(G([\boldsymbol{\epsilon}, \boldsymbol{c}])) - H(G([\boldsymbol{\epsilon}, \boldsymbol{c}]) \mid \boldsymbol{c})$. The generator objective then becomes,

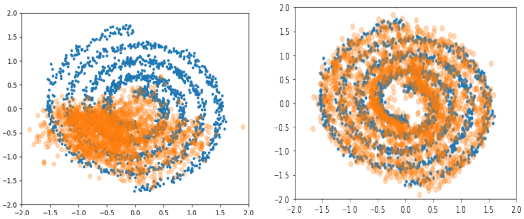
$$\arg\max_G \mathbb{E}[\log(D(G([\boldsymbol{\epsilon}, \boldsymbol{c}])))] + \beta I(G([\boldsymbol{\epsilon}, \boldsymbol{c}]); \boldsymbol{c}). \tag{5.2}$$

As the samples $G([\boldsymbol{\epsilon}, \boldsymbol{c}])$ are differentiable w.r.t. the parameters of G, and the statistics network being a differentiable function, we can maximize the mutual information using back-propagation and gradient ascent by only specifying this additional loss term. Since the mutual information is theoretically unbounded, we use adaptive gradient clipping (see the Supplementary Material) to ensure that the generator receives learning signals similar in magnitude from the discriminator and the statistics network.

Related works on mode-dropping. Methods to address mode dropping in GANs can readily be found in the literature. Salimans et al. (2016) use mini-batch discrimination. In the same spirit, Lin et al. (2017) successfully mitigates mode dropping in GANs by modifying the discriminator to make decisions on multiple real or generated samples. Ghosh et al. (2017) uses multiple generators that are encouraged to generate different parts of the target distribution. Nguyen et al. (2017) uses two discriminators to minimize the KL and reverse KL divergences between the target and generated distributions. Che et al. (2016) learns a reconstruction distribution, then teach the generator to sample from it, the intuition being that the reconstruction distribution is a de-noised or smoothed version of the data distribution, and thus easier to learn. Srivastava et al. (2017) minimizes the reconstruction error in the latent space of bi-directional GANs (Dumoulin et al., 2016; Donahue et al., 2016). Metz et al. (2017) includes many steps of the discriminator's optimization as part of the generator's objective. While Chen et al. (2016) maximizes the mutual information between the code and the samples, it does so by minimizing a variational upper bound on the conditional entropy (Barber & Agakov, 2003) therefore ignoring the entropy of the samples. Chen et al. (2016) makes no claim about mode-dropping.

Experiments: Spiral, 25-Gaussians datasets. We apply MINE to improve mode coverage when training a generative adversarial network (GAN, Goodfellow et al., 2014). We demonstrate using Eqn. 5.2 on the spiral and the 25-Gaussians datasets, comparing two models, one with $\beta = 0$ (which corresponds to the orthodox GAN as in Goodfellow et al. (2014)) and one with $\beta = 1.0$, which corresponds to mutual information maximization.

Our results on the spiral (Fig. 6.3) and the 25-Gaussians (Fig. 6.4) experiments both show improved mode coverage over the baseline with no mutual information objective. This confirms our hypothesis that maximizing mutual information helps against mode-dropping in this simple setting.



(a) GAN (b) GAN+MINE

Fig. 6.3. The generator of the GAN model without mutual information maximization after 5000 iterations suffers from mode collapse (has poor coverage of the target dataset) compared to GAN+MINE on the spiral experiment.

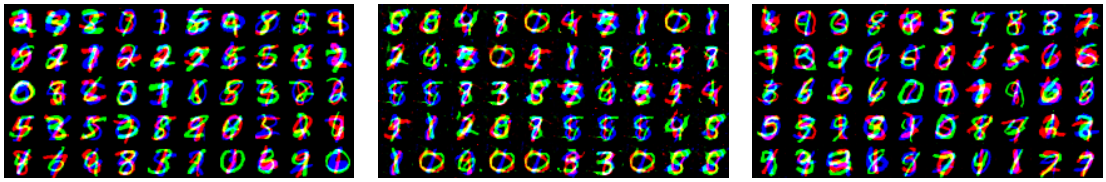
(a) Original data (b) GAN (c) GAN+MINE

Fig. 6.4. Kernel density estimate (KDE) plots for GAN+MINE samples and GAN samples on 25 Gaussians dataset.

Experiment: Stacked MNIST. Following Che et al. (2016); Metz et al. (2017); Srivastava et al. (2017); Lin et al. (2017), we quantitatively assess MINE's ability to diminish mode dropping on the stacked MNIST dataset which is constructed by stacking three randomly sampled MNIST digits. As a consequence, stacked MNIST offers 1000 modes. Using the same architecture and training protocol as in Srivastava et al. (2017); Lin et al. (2017), we train a GAN on the constructed dataset and use a pre-trained classifier on 26,000 samples to count the number of modes in the samples, as well as to compute the KL divergence between the sample and expected data distributions. Our results in Table 6.1 demonstrate the effectiveness of MINE in preventing mode collapse on Stacked MNIST.

	Stacked MNIST	
	Modes (Max 1000)	KL
DCGAN	99.0	3.40
ALI	16.0	5.40
Unrolled GAN	48.7	4.32
VEEGAN	150.0	2.95
PacGAN	1000.0 ± 0.0	$0.06 \pm 1.0\text{e}^{-2}$
GAN+MINE (Ours)	1000.0 ± 0.0	$0.05 \pm 6.9\text{e}^{-3}$

Table 6.1. Number of captured modes and Kullblack-Leibler divergence between the training and samples distributions for DCGAN (Radford et al., 2015), ALI (Dumoulin et al., 2016), Unrolled GAN (Metz et al., 2017), VeeGAN (Srivastava et al., 2017), PacGAN (Lin et al., 2017).



(a) Training set (b) DCGAN (c) DCGAN+MINE

Fig. 6.5. Samples from the Stacked MNIST dataset along with generated samples from DCGAN and DCGAN with MINE. While DCGAN only shows a very limited number of modes, the inclusion of MINE generates a much better representative set of samples.

5.2. Maximizing mutual information to improve inference in bi-directional adversarial models

Adversarial bi-directional models were introduced in Adversarially Learned Inference (ALI, Dumoulin et al., 2016) and BiGAN (Donahue et al., 2016) and are an extension of GANs which incorporate a reverse model, $F : \mathcal{X} \to \mathcal{Z}$ jointly trained with the generator. These models formulate the problem in terms of the value function in Eqn. 5.1 between two

joint distributions, $p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{z} \mid \boldsymbol{x})p(\boldsymbol{x})$ and $q(\boldsymbol{x}, \boldsymbol{z}) = q(\boldsymbol{x} \mid \boldsymbol{z})p(\boldsymbol{z})$ induced by the forward (encoder) and reverse (decoder) models, respectively[7].

One goal of bi-directional models is to do inference as well as to learn a good generative model. Reconstructions are one desirable property of a model that does both inference and generation, but in practice ALI can lack fidelity (i.e., reconstructs less faithfully than desired, see Li et al., 2017; Ulyanov et al., 2017; Belghazi et al., 2018b). To demonstrate the connection to mutual information, it can be shown (see the Supplementary Material for details) that the reconstruction error, $\mathcal{R}$, is bounded by,

$$\mathcal{R} \leq D_{KL}(q(\boldsymbol{x}, \boldsymbol{z}) \mid\mid p(\boldsymbol{x}, \boldsymbol{z})) - I_q(\boldsymbol{x}, \boldsymbol{z}) + H_q(\boldsymbol{z}) \tag{5.3}$$

If the joint distributions are matched, $H_q(\boldsymbol{z})$ tends to $H_p(\boldsymbol{z})$, which is fixed as long as the prior, $p(\boldsymbol{z})$, is itself fixed. Subsequently, maximizing the mutual information minimizes the expected reconstruction error.

Assuming that the generator is the same as with GANs in the previous section, the objectives for training a bi-directional adversarial model then become:

$$\begin{aligned} &\arg\max_{D} \mathbb{E}_{q(\boldsymbol{x},\boldsymbol{z})}[\log D(\boldsymbol{x}, \boldsymbol{z})] + \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{z})}[\log(1 - D(\boldsymbol{x}, \boldsymbol{z}))] \\ &\arg\max_{F,G} \mathbb{E}_{q(\boldsymbol{x},\boldsymbol{z})}[\log(1 - D(\boldsymbol{x}, \boldsymbol{z}))] + \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{z})}[\log D(\boldsymbol{x}, \boldsymbol{z})] \\ &\qquad + \beta I_q(\boldsymbol{x}, \boldsymbol{z}). \end{aligned} \tag{5.4}$$

Related works. Ulyanov et al. (2017) improves reconstructions quality by forgoing the discriminator and expressing the adversarial game between the encoder and decoder. Kumar et al. (2017) augments the bi-directional objective by considering the reconstruction and the corresponding encodings as an additional fake pair. Belghazi et al. (2018b) shows that a Markovian hierarchical generator in a bi-directional adversarial model provide a hierarchy of reconstructions with increasing levels of fidelity (increasing reconstruction quality). Li et al. (2017) shows that the expected reconstruction error can be diminished by minimizing the conditional entropy of the observables given the latent representations. The conditional entropy being intractable for general posterior, Li et al. (2017) proposes to augment the generator's loss with an adversarial cycle consistency loss (Zhu et al., 2017) between the observables and their reconstructions.

[7] We switch to density notations for convenience throughout this section.

Experiment: ALI+MINE. In this section we compare MINE to existing bi-directional adversarial models. As the decoder's density is generally intractable, we use three different metrics to measure the fidelity of the reconstructions with respect to the samples; (*i*) the euclidean reconstruction error, (*ii*) reconstruction accuracy, which is the proportion of labels preserved by the reconstruction as identified by a pre-trained classifier; (*iii*) the Multi-scale structural similarity metric (MS-SSIM, Wang et al., 2004) between the observables and their reconstructions.

We train MINE on datasets of increasing order of complexity: a toy dataset composed of 25-Gaussians, MNIST (LeCun, 1998), and the CelebA dataset (Liu et al., 2015). Fig. 6.6 shows the reconstruction ability of MINE compared to ALI. Although ALICE does perfect reconstruction (which is in its explicit formulation), we observe significant mode-dropping in the sample space. MINE does a balanced job of reconstructing along with capturing all the modes of the underlying data distribution.

Next, we measure the fidelity of the reconstructions over ALI, ALICE, and MINE. Tbl. 2 compares MINE to the existing baselines in terms of euclidean reconstruction errors, reconstruction accuracy, and MS-SSIM. On MNIST, MINE outperforms ALI in terms of reconstruction errors by a good margin and is competitive to ALICE with respect to reconstruction accuracy and MS-SSIM. Our results show that MINE's effect on reconstructions is even more dramatic when compared to ALI and ALICE on the CelebA dataset.

5.3. Information Bottleneck

The Information Bottleneck (IB, Tishby et al., 2000) is an information theoretic method for extracting relevant information, or yielding a representation, that an input $X \in \mathcal{X}$ contains about an output $Y \in \mathcal{Y}$. An optimal representation of X would capture the relevant factors and compress X by diminishing the irrelevant parts which do not contribute to the prediction of Y. IB was recently covered in the context of deep learning (Tishby & Zaslavsky, 2015), and as such can be seen as a process to construct an approximation of the minimally sufficient statistics of the data. IB seeks an encoder, $q(Z \mid X)$, that induces the Markovian structure $X \to Z \to Y$. This is done by minimizing the IB Lagrangian,

$$\mathcal{L}[q(Z \mid X)] = H(Y|Z) + \beta I(X,Z), \tag{5.5}$$

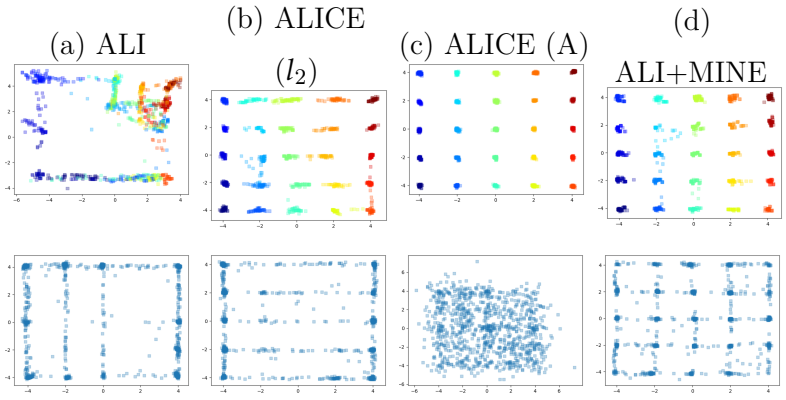


Fig. 6.6. Reconstructions and model samples from adversarially learned inference (ALI) and variations intended to increase improve reconstructions. Shown left to right are the baseline (ALI), ALICE with the l_2 loss to minimize the reconstruction error, ALICE with an adversarial loss, and ALI+MINE. Top to bottom are the reconstructions and samples from the priors. ALICE with the adversarial loss has the best reconstruction, though at the expense of poor sample quality, where as ALI+MINE captures all the modes of the data in sample space.

which appears as a standard cross-entropy loss augmented with a regularizer promoting minimality of the representation (Achille & Soatto, 2017). Here we propose to estimate the regularizer with MINE.

Related works. In the discrete setting, Tishby et al. (2000) uses the Blahut-Arimoto Algorithm Arimoto (1972), which can be understood as cyclical coordinate ascent in function spaces. While IB is successful and popular in a discrete setting, its application to the continuous setting was stifled by the intractability of the continuous mutual information. Nonetheless, IB was applied in the case of jointly Gaussian random variables in Chechik et al. (2005).

Model	Recons. Error	Recons. Acc.(%)	MS-SSIM
MNIST			
ALI	14.24	45.95	0.97
ALICE(l_2)	3.20	99.03	0.97
ALICE(Adv.)	5.20	98.17	0.98
MINE	9.73	96.10	0.99
CelebA			
ALI	53.75	57.49	0.81
ALICE(l_2)	8.01	32.22	0.93
ALICE(Adv.)	92.56	48.95	0.51
MINE	36.11	76.08	0.99

Table 6.2. Comparison of MINE with other bi-directional adversarial models in terms of euclidean reconstruction error, reconstruction accuracy, and MS-SSIM on the MNIST and CelebA datasets. MINE does a good job compared to ALI in terms of reconstructions. Though the explicit reconstruction based baselines (ALICE) can sometimes do better than MINE in terms of reconstructions related tasks, they consistently lag behind in MS-SSIM scores and reconstruction accuracy on CelebA.

In order to overcome the intractability of $I(X;Z)$ in the continuous setting, Alemi et al. (2016); Kolchinsky et al. (2017); Chalk et al. (2016) exploit the variational bound of Barber & Agakov (2003) to approximate the conditional entropy in $I(X;Z)$. These approaches differ only on their treatment of the marginal distribution of the bottleneck variable: Alemi et al. (2016) assumes a standard multivariate normal marginal distribution, Chalk et al. (2016) uses a Student-t distribution, and Kolchinsky et al. (2017) uses non-parametric estimators. Due to their reliance on a variational approximation, these methods require a tractable density for the approximate posterior, while MINE does not.

Experiment: Permutation-invariant MNIST classification. Here, we demonstrate an implementation of the IB objective on permutation invariant MNIST using MINE. We compare to the Deep Variational Bottleneck (DVB, Alemi et al., 2016) and use the same empirical setup. As the DVB relies on a variational bound on the conditional entropy, it therefore

requires a tractable density. Alemi et al. (2016) opts for a conditional Gaussian encoder $\boldsymbol{z} = \mu(\boldsymbol{x}) + \sigma \odot \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$. As MINE does not require a tractable density, we consider three type of encoders: (i) a Gaussian encoder as in Alemi et al. (2016); (ii) an additive noise encoder, $\boldsymbol{z} = enc(\boldsymbol{x} + \sigma \odot \epsilon)$; and (iii) a propagated noise encoder, $\boldsymbol{z} = enc([\boldsymbol{x}, \epsilon])$. Our results can be seen in Tbl. 6.3, and this shows MINE as being superior in these settings.

Model	Misclass. rate(%)
Baseline	1.38%
Dropout	1.34%
Confidence penalty	1.36%
Label Smoothing	1.40%
DVB	1.13%
DVB + Additive noise	1.06%
MINE(Gaussian) (ours)	1.11%
MINE(Propagated) (ours)	1.10%
MINE(Additive) (ours)	1.01%

Table 6.3. Permutation Invariant MNIST misclassification rate using Alemi et al. (2016) experimental setup for regularization by confidence penalty (Pereyra et al., 2017), label smoothing (Pereyra et al., 2017), Deep Variational Bottleneck(DVB) (Alemi et al., 2016) and MINE. The misclassification rate is averaged over ten runs. In order to control for the regularizing impact of the additive Gaussian noise in the additive conditional, we also report the results for DVB with additional additive Gaussian noise at the input. All non-MINE results are taken from Alemi et al. (2016).

6. Conclusion

We proposed a mutual information estimator, which we called the mutual information neural estimator (MINE), that is scalable in dimension and sample-size. We demonstrated the efficiency of this estimator by applying it in a number of settings. First, a term of mutual information can be introduced alleviate mode-dropping issue in generative adversarial

networks (GANs, Goodfellow et al., 2014). Mutual information can also be used to improve inference and reconstructions in adversarially-learned inference (ALI, Dumoulin et al., 2016). Finally, we showed that our estimator allows for tractable application of Information bottleneck methods (Tishby et al., 2000) in a continuous setting.

7. Acknowledgements

We would like to thank Martin Arjovsky, Caglar Gulcehre, Marcin Moczulski, Negar Rostamzadeh, Thomas Boquet, Ioannis Mitliagkas, Pedro Oliveira Pinheiro for helpful comments, as well as Samsung and IVADO for their support.

8. Appendix

In this Appendix, we provide additional experiment details and spell out the proofs omitted in the text.

8.1. Experimental Details

8.1.1. Adaptive Clipping. Here we assume we are in the context of GANs described in Sections 5.1 and 5.2, where the mutual information shows up as a regularizer in the generator objective.

Notice that the generator is updated by two gradients. The first gradient is that of the generator's loss, $\mathcal{L}_g$ with respect to the generator's parameters θ, $g_u := \frac{\partial \mathcal{L}_g}{\partial \theta}$. The second flows from the mutual information estimate to the generator, $g_m := -\frac{\partial \widehat{I(X;Z)}}{\partial \theta}$. If left unchecked, because mutual information is unbounded, the latter can overwhelm the former, leading to a failure mode of the algorithm where the generator puts all of its attention on maximizing the mutual information and ignores the adversarial game with the discriminator. We propose to adaptively clip the gradient from the mutual information so that its Frobenius norm is at most that of the gradient from the discriminator. Defining g_a to be the adapted gradient following from the statistics network to the generator, we have,

$$g_a = \min(\|g_u\|, \|g_m\|) \frac{g_m}{\|g_m\|}. \tag{8.1}$$

Note that adaptive clipping can be considered in any situation where MINE is to be maximized.

8.1.2. GAN+MINE: Spiral and 25-gaussians. In this section we state the details of experiments supporting mode dropping experiments on the spiral and 25-Gaussians dataset. For both the datasets we use 100,000 examples sampled from the target distributions, using a standard deviation of 0.05 in the case of 25-gaussians, and using additive noise for the spiral. The generator for the GAN consists of two fully connected layers with 500 units in each layer with batch-normalization (Ioffe & Szegedy, 2015a) and Leaky-ReLU as activation function as in Dumoulin et al. (2016). The discriminator and statistics networks have three fully connected layers with 400 units each. We use the Adam (Kingma & Ba, 2014) optimizer with a learning rate of 0.0001. Both GAN baseline and GAN+MINE were trained for 5,000 iterations with a mini batch-size of 100.

8.1.3. GAN+MINE: Stacked-MNIST. Here we describe the experimental setup and architectural details of stacked-MNIST task with GAN+MINE. We compare to the exact same experimental setup followed and reported in PacGANLin et al. (2017) and VEEGANSrivastava et al. (2017). We use a pre-trained classifier to classify generated samples on each of the three stacked channels. Evaluation is done on 26,000 test samples as followed in the baselines. We train GAN+MINE for 50 epochs on 128,000 samples. Details for generator and discriminator networks are given below in the table6.4 and table6.5. Specifically the statistics network has the same architecture as discriminator in DCGAN with ELU (Clevert et al., 2015) as activation function for the individual layers and without batch-normalization as highlighted in Table 6.6. In order to condition the statistics network on the z variable, we use linear MLPs at each layer, whose output are reshaped to the number of feature maps. The linear MLPs output is then added as a dynamic bias.

Generator				
Layer	Number of outputs	Kernel size	Stride	Activation function
Input $\boldsymbol{z} \sim \mathcal{U}(-1,1)^{100}$	100			
Fully-connected	2*2*512			ReLU
Transposed convolution	4*4*256	$5*5$	2	ReLU
Transposed convolution	7*7*128	$5*5$	2	ReLU
Transposed convolution	14*14*64	$5*5$	2	ReLU
Transposed convolution	28*28*3	$5*5$	2	Tanh

Table 6.4. Generator network for Stacked-MNIST experiment using GAN+MINE.

Discriminator				
Layer	Number of outputs	Kernel size	Stride	Activation function
Input x	$28 * 28 * 3$			
Convolution	14*14*64	$5 * 5$	2	ReLU
Convolution	7*7*128	$5 * 5$	2	ReLU
Convolution	4*4*256	$5 * 5$	2	ReLU
Convolution	2*2*512	$5 * 5$	2	ReLU
Fully-connected	1	1	Valid	Sigmoid

Table 6.5. Discriminator network for Stacked-MNIST experiment.

Statistics Network				
Layer	number of outputs	kernel size	stride	activation function
Input x, z				
Convolution	14*14*16	$5 * 5$	2	ELU
Convolution	7*7*32	$5 * 5$	2	ELU
Convolution	4*4*64	$5 * 5$	2	ELU
Flatten	-	-	-	-
Fully-Connected	1024	1	Valid	None
Fully-Connected	1	1	Valid	None

Table 6.6. Statistics network for Stacked-MNIST experiment.

8.1.4. ALI+MINE: MNIST and CelebA. In this section we state the details of experimental setup and the network architectures used for the task of improving reconstructions and representations in bidirectional adversarial models with MINE. The generator and discriminator network architectures along with the hyper parameter setup used in these tasks are similar to the ones used in DCGAN (Radford et al., 2015).

Statistics network conditioning on the latent code was done as in the Stacked-MNIST experiments. We used Adam as the optimizer with a learning rate of 0.0001. We trained the model for a total of 35,000 iterations on CelebA and 50,000 iterations on MNIST, both with a mini batch-size of 100.

Encoder				
Layer	Number of outputs	Kernel size	Stride	Activation function
Input $[\boldsymbol{x}, \boldsymbol{\epsilon}]$	28*28*129			
Convolution	14*14*64	$5 * 5$	2	ReLU
Convolution	7*7*128	$5 * 5$	2	ReLU
Convolution	4*4*256	$5 * 5$	2	ReLU
Convolution	256	$4 * 4$	Valid	ReLU
Fully-connected	128	-	-	None

Table 6.7. Encoder network for bi-directional models on MNIST. $\boldsymbol{\epsilon} \sim \mathcal{N}_{128}(0, I)$.

Decoder				
Layer	Number of outputs	Kernel size	Stride	Activation function
Input $\boldsymbol{z}$	128			
Fully-connected	4*4*256			ReLU
Transposed convolution	7*7*128	$5 * 5$	2	ReLU
Transposed convolution	14*14*64	$5 * 5$	2	ReLU
Transposed convolution	28*28*1	$5 * 5$	2	Tanh

Table 6.8. Decoder network for bi-directional models on MNIST. $\boldsymbol{z} \sim \mathcal{N}_{256}(0, I)$

Discriminator				
Layer	Number of outputs	Kernel size	Stride	Activation function
Input x	$28 * 28 * 3$			
Convolution	14*14*64	$5 * 5$	2	LearkyReLU
Convolution	7*7*128	$5 * 5$	2	LeakyReLU
Convolution	4*4*256	$5 * 5$	2	LeakyReLU
Flatten	-	-	-	
Concatenate $\boldsymbol{z}$	-	-	-	
Fully-connected	1024	-	-	LeakyReLU
Fully-connected	1	-	-	Sigmoid

Table 6.9. Discriminator network for bi-directional models experiments MINE on MNIST.

Statistics Network				
Layer	number of outputs	kernel size	stride	activation function
Input $\boldsymbol{x}, \boldsymbol{z}$				
Convolution	14*14*64	$5 * 5$	2	LeakyReLU
Convolution	7*7*128	$5 * 5$	2	LeakyReLU
Convolution	4*4*256	$5 * 5$	2	LeakyReLU
Flatten	-	-	-	-
Fully-connected	1	-	-	None

Table 6.10. Statistics network for bi-directional models using MINE on MNIST.

Encoder				
Layer	Number of outputs	Kernel size	Stride	Activation function
Input $[\boldsymbol{x}, \boldsymbol{\epsilon}]$	64*64*259			
Convolution	32*32*64	$5 * 5$	2	ReLU
Convolution	16*16*128	$5 * 5$	2	ReLU
Convolution	8*8*256	$5 * 5$	2	ReLU
Convolution	4*4*512	$5 * 5$	2	ReLU
Convolution	512	$4 * 4$	Valid	ReLU
Fully-connected	256	-	-	None

Table 6.11. Encoder network for bi-directional models on CelebA. $\boldsymbol{\epsilon} \sim \mathcal{N}_{256}(0, I)$.

Decoder				
Layer	Number of outputs	Kernel size	Stride	Activation function
Input $\boldsymbol{z} \sim \mathcal{N}_{256}(0, I)$	256			
Fully-Connected	4*4*512	-	-	ReLU
Transposed convolution	8*8*256	$5 * 5$	2	ReLU
Transposed convolution	16*16*128	$5 * 5$	2	ReLU
Transposed convolution	32*32*64	$5 * 5$	2	ReLU
Transposed convolution	64*64*3	$5 * 5$	2	Tanh

Table 6.12. Decoder network for bi-directional model(ALI, ALICE) experiments using MINE on CelebA.

Discriminator				
Layer	Number of outputs	Kernel size	Stride	Activation function
Input x	$64 * 64 * 3$			
Convolution	32*32*64	$5 * 5$	2	LearkyReLU
Convolution	16*16*128	$5 * 5$	2	LeakyReLU
Convolution	8*8*256	$5 * 5$	2	LeakyReLU
Convolution	4*4*512	$5 * 5$	2	LeakyReLU
Flatten	-	-	-	
Concatenate z	-	-	-	
Fully-connected	1024	-	-	LeakyReLU
Fully-connected	1	-	-	Sigmoid

Table 6.13. Discriminator network for bi-directional models on CelebA.

Statistics Network				
Layer	number of outputs	kernel size	stride	activation function
Input x, z				
Convolution	32*32*16	$5 * 5$	2	ELU
Convolution	16*16*32	$5 * 5$	2	ELU
Convolution	8*8*64	$5 * 5$	2	ELU
Convolution	4*4*128	$5 * 5$	2	ELU
Flatten	-	-	-	-
Fully-connected	1	-	-	None

Table 6.14. Statistics network for bi-directional models on CelebA.

8.1.5. Information bottleneck with MINE. In this section we outline the network details and hyper-parameters used for the information bottleneck task using MINE. To keep comparison fair all hyperparameters and architectures are those outlined in Alemi et al. (2016). The statistics network is shown, a two layer MLP with additive noise at each layer and 512 ELUs (Clevert et al., 2015) activations, is outlined in table6.15.

Statistics Network		
Layer	number of outputs	activation function
input $[\boldsymbol{x}, \boldsymbol{z}]$		
Gaussian noise(std=0.3)	-	-
dense layer	512	ELU
Gaussian noise(std=0.5)	-	-
dense layer	512	ELU
Gaussian noise(std=0.5)	-	-
dense layer	1	None

Table 6.15. Statistics network for Information-bottleneck experiments on MNIST.

8.2. Proofs

8.2.1. Donsker-Varadhan Representation.

Theorem 8 (Theorem 1 restated). The KL divergence admits the following dual representation:

$$D_{KL}(\mathbb{P} \,||\, \mathbb{Q}) = \sup_{T:\Omega\to\mathbb{R}} \mathbb{E}_{\mathbb{P}}[T] - \log(\mathbb{E}_{\mathbb{Q}}[e^T]), \tag{8.2}$$

where the supremum is taken over all functions T such that the two expectations are finite.

Proof. A simple proof goes as follows. For a given function T, consider the Gibbs distribution $\mathbb{G}$ defined by $d\mathbb{G} = \frac{1}{Z} e^T d\mathbb{Q}$, where $Z = \mathbb{E}_{\mathbb{Q}}[e^T]$. By construction,

$$\mathbb{E}_{\mathbb{P}}[T] - \log Z = \mathbb{E}_{\mathbb{P}}\left[\log \frac{d\mathbb{G}}{d\mathbb{Q}}\right] \tag{8.3}$$

Let Δ be the gap,

$$\Delta := D_{KL}(\mathbb{P} \,||\, \mathbb{Q}) - \Big(\mathbb{E}_{\mathbb{P}}[T] - \log(\mathbb{E}_{\mathbb{Q}}[e^T])\Big) \tag{8.4}$$

Using Eqn 8.3, we can write Δ as a KL-divergence:

$$\Delta = \mathbb{E}_{\mathbb{P}}\left[\log \frac{d\mathbb{P}}{d\mathbb{Q}} - \log \frac{d\mathbb{G}}{d\mathbb{Q}}\right] = \mathbb{E}_{\mathbb{P}} \log \frac{d\mathbb{P}}{d\mathbb{G}} = D_{KL}(\mathbb{P} \,||\, \mathbb{G}) \tag{8.5}$$

The positivity of the KL-divergence gives $\Delta \geq 0$. We have thus shown that for any T,

$$D_{KL}(\mathbb{P} \,||\, \mathbb{Q}) \geq \mathbb{E}_{\mathbb{P}}[T] - \log(\mathbb{E}_{\mathbb{Q}}[e^T]) \tag{8.6}$$

and the inequality is preserved upon taking the supremum over the right-hand side. Finally, the identity (8.5) also shows that this bound is tight whenever $\mathbb{G} = \mathbb{P}$, namely for optimal functions T^* taking the form $T^* = \log \frac{d\mathbb{P}}{d\mathbb{Q}} + C$ for some constant $C \in \mathbb{R}$. □

8.2.2. Consistency Proofs. This section presents the proofs of the Lemma and consistency theorem stated in the consistency in Section 3.3.1.

In what follows, we assume that the input space $\Omega = \mathcal{X} \times \mathcal{Z}$ is a compact domain of $\mathbb{R}^d$, and all measures are absolutely continuous with respect to the Lebesgue measure. We will restrict to families of feedforward functions with continuous activations, with a single output neuron, so that a given architecture defines a continuous mapping $(\omega, \theta) \to T_\theta(\omega)$ from $\Omega \times \Theta$ to $\mathbb{R}$.

To avoid unnecessary heavy notation, we denote $\mathbb{P} = \mathbb{P}_{XZ}$ and $\mathbb{Q} = \mathbb{P}_X \otimes \mathbb{P}_Z$ for the joint distribution and product of marginals, and $\mathbb{P}_n, \mathbb{Q}_n$ for their empirical versions. We will use the notation $\hat{I}(T)$ for the quantity:

$$\hat{I}(T) = \mathbb{E}_{\mathbb{P}}[T] - \log(\mathbb{E}_{\mathbb{Q}}[e^T]) \tag{8.7}$$

so that $I_\Theta(X, Z) = \sup_{\theta \in \Theta} \hat{I}(T_\theta)$.

Lemma 9 (Lemma 4 restated). Let $\eta > 0$. There exists a family of neural network functions T_θ with parameters θ in some compact domain $\Theta \subset \mathbb{R}^k$, such that

$$|I(X, Z) - I_\Theta(X,Z)| \leq \eta \tag{8.8}$$

where

$$I_\Theta(X, Z) = \sup_{\theta \in \Theta} \mathbb{E}_{\mathbb{P}_{XZ}}[T_\theta] - \log(\mathbb{E}_{\mathbb{P}_X \otimes \mathbb{P}_Z}[e^{T_\theta}]) \tag{8.9}$$

Proof. Let $T^* = \log \frac{d\mathbb{P}}{d\mathbb{Q}}$. By construction, T^* satisfies:

$$\mathbb{E}_{\mathbb{P}}[T^*] = I(X, Z), \qquad \mathbb{E}_{\mathbb{Q}}[e^{T^*}] = 1 \tag{8.10}$$

For a function T, the (positive) gap $I(X, Z) - \hat{I}(T)$ can be written as

$$I(X, Z) - \hat{I}(T) = \mathbb{E}_{\mathbb{P}}[T^* - T] + \log \mathbb{E}_{\mathbb{Q}}[e^T] \leq \mathbb{E}_{\mathbb{P}}[T^* - T] + \mathbb{E}_{\mathbb{Q}}[e^T - e^{T^*}] \tag{8.11}$$

where we used the inequality $\log x \leq x - 1$.

Fix $\eta > 0$. We first consider the case where T^* is bounded from above by a constant M. By the universal approximation theorem (see corollary 2.2 of Hornik (1989)[8]), we may choose a feedforward network function $T_{\hat{\theta}} \leq M$ such that

$$\mathbb{E}_{\mathbb{P}}|T^* - T_{\hat{\theta}}| \leq \frac{\eta}{2} \quad \text{and} \quad \mathbb{E}_{\mathbb{Q}}|T^* - T_{\hat{\theta}}| \leq \frac{\eta}{2}e^{-M} \tag{8.12}$$

Since exp is Lipschitz continuous with constant e^M on $(-\infty, M]$, we have

$$\mathbb{E}_{\mathbb{Q}}|e^{T^*} - e^{T_{\hat{\theta}}}| \leq e^M \, \mathbb{E}_{\mathbb{Q}}|T^* - T_{\hat{\theta}}| \leq \frac{\eta}{2} \tag{8.13}$$

From Equ 8.11 and the triangular inequality, we then obtain:

$$|I(X,Z) - \hat{I}(T_{\hat{\theta}})| \leq \mathbb{E}_{\mathbb{P}}|T^* - T_{\hat{\theta}}| + \mathbb{E}_{\mathbb{Q}}|e^{T^*} - e^{T_{\hat{\theta}}}| \leq \frac{\eta}{2} + \frac{\eta}{2} \leq \eta \tag{8.14}$$

In the general case, the idea is to partition Ω in two subset $\{T^* > M\}$ and $\{T^* \leq M\}$ for a suitably chosen large value of M. For a given subset $S \subset \Omega$, we will denote by $\mathbb{1}_S$ its characteristic function, $\mathbb{1}_S(\omega) = 1$ if $\omega \in S$ and 0 otherwise. T^* is integrable with respect to $\mathbb{P}$[9], and e^{T^*} is integrable with respect to $\mathbb{Q}$, so by the dominated convergence theorem, we may choose M so that the expectations $\mathbb{E}_{\mathbb{P}}[T^*\mathbb{1}_{T^*>M}]$ and $\mathbb{E}_{\mathbb{Q}}[e^{T^*}\mathbb{1}_{T^*>M}]$ are lower than $\eta/4$. Just like above, we then use the universal approximation theorem to find a feed forward network function $T_{\hat{\theta}}$, which we can assume without loss of generality to be upper-bounded by M, such that

$$\mathbb{E}_{\mathbb{P}}|T^* - T_{\hat{\theta}}| \leq \frac{\eta}{2} \quad \text{and} \quad \mathbb{E}_{\mathbb{Q}}|T^* - T_{\hat{\theta}}|\mathbb{1}_{T^*\leq M} \leq \frac{\eta}{4}e^{-M} \tag{8.15}$$

We then write

$$\begin{aligned}
\mathbb{E}_{\mathbb{Q}}[e^{T^*} - e^{T_{\hat{\theta}}}] &= \mathbb{E}_{\mathbb{Q}}[(e^{T^*} - e^{T_{\hat{\theta}}})\mathbb{1}_{T^*\leq M}] + \mathbb{E}_{\mathbb{Q}}[(e^{T^*} - e^{T_{\hat{\theta}}})\mathbb{1}_{T^*>M}] \\
&\leq e^M \mathbb{E}_{\mathbb{Q}}[|T^* - T_{\hat{\theta}}|\mathbb{1}_{T^*\leq M}] + \mathbb{E}_{\mathbb{Q}}[e^{T^*}\mathbb{1}_{T^*>M}] \\
&\leq \frac{\eta}{4} + \frac{\eta}{4} \qquad (8.16) \\
&\leq \frac{\eta}{2} \qquad (8.17)
\end{aligned}$$

[8] Specifically, the argument relies on the density of feedforward network functions in the space $L^1(\Omega, \mu)$ of integrable functions with respect the measure $\mu = \mathbb{P} + \mathbb{Q}$.

[9] This can be seen from the identity (Györfi & van der Meulen, 1987)

$$\mathbb{E}_{\mathbb{P}}\left|\log \frac{d\mathbb{P}}{d\mathbb{Q}}\right| \leq D_{KL}(\mathbb{P} \,||\, \mathbb{Q}) + 4\sqrt{D_{KL}(\mathbb{P} \,||\, \mathbb{Q})}$$

where the inequality in the second line arises from the convexity and positivity of exp. Eqns. 8.15 and 8.16, together with the triangular inequality, lead to Eqn. 8.14, which proves the Lemma.

□

Lemma 10 (Lemma 5 restated). Let $\eta > 0$. Given a family $\mathcal{F}$ of neural network functions T_θ with parameters θ in some compact domain $\Theta \subset \mathbb{R}^k$, there exists $N \in \mathbb{N}$ such that

$$\forall n \geq N, \quad \Pr\left(|\widehat{I(X;Z)}_n - I_{\mathcal{F}}(X,Z)| \leq \eta\right) = 1 \tag{8.18}$$

Proof. We start by using the triangular inequality to write,

$$|\widehat{I(X;Z)}_n - \sup_{T_\theta \in \mathcal{F}} \hat{I}(T_\theta)| \leq \sup_{T_\theta \in \mathcal{F}} |\mathbb{E}_{\mathbb{P}}[T_\theta] - \mathbb{E}_{\mathbb{P}_n}[T_\theta]| + \sup_{T_\theta \in \mathcal{F}} |\log \mathbb{E}_{\mathbb{Q}}[e^{T_\theta}] - \log \mathbb{E}_{\mathbb{Q}_n}[e^{T_\theta}]| \tag{8.19}$$

The continuous function $(\theta, \omega) \to T_\theta(\omega)$, defined on the compact domain $\Theta \times \Omega$, is bounded. So the functions T_θ are uniformly bounded by a constant M, i.e $|T_\theta| \leq M$ for all $\theta \in \Theta$. Since log is Lipschitz continuous with constant e^M in the interval $[e^{-M}, e^M]$, we have

$$|\log \mathbb{E}_{\mathbb{Q}}[e^{T_\theta}] - \log \mathbb{E}_{\mathbb{Q}_n}[e^{T_\theta}]| \leq e^M |\mathbb{E}_{\mathbb{Q}}[e^{T_\theta}] - \mathbb{E}_{\mathbb{Q}_n}[e^{T_\theta}]| \tag{8.20}$$

Since Θ is compact and the feedforward network functions are continuous, the families of functions T_θ and e^{T_θ} satisfy the uniform law of large numbers (Van de Geer, 2000). Given $\eta > 0$ we can thus choose $N \in \mathbb{N}$ such that $\forall n \geq N$ and with probability one,

$$\sup_{T_\theta \in \mathcal{F}} |\mathbb{E}_{\mathbb{P}}[T_\theta] - \mathbb{E}_{\mathbb{P}_n}[T_\theta] \leq \frac{\eta}{2} \quad \text{and} \quad \sup_{T_\theta \in \mathcal{F}} |\mathbb{E}_{\mathbb{Q}}[e^{T_\theta}] - \mathbb{E}_{\mathbb{Q}_n}[e^{T_\theta}]| \leq \frac{\eta}{2} e^{-M} \tag{8.21}$$

Together with Eqns. 8.19 and 8.20, this leads to

$$|\widehat{I(X;Z)}_n - \sup_{T_\theta \in \mathcal{F}} \hat{I}(T_\theta)| \leq \frac{\eta}{2} + \frac{\eta}{2} = \eta \tag{8.22}$$

□

Theorem 11 (Theorem 6 restated). MINE is strongly consistent.

Proof. Let $\epsilon > 0$. We apply the two Lemmas to find a a family of neural network function $\mathcal{F}$ and $N \in \mathbb{N}$ such that (8.8) and (8.18) hold with $\eta = \epsilon/2$. By the triangular inequality, for all $n \geq N$ and with probability one, we have:

$$|I(X,Z) - \widehat{I(X;Z)}_n| \quad \leq \quad |I(X,Z) - \sup_{T_\theta \in \mathcal{F}} \hat{I}(T_\theta)| + |\widehat{I(X;Z)}_n - I_{\mathcal{F}}(X,Z)| \leq \quad \epsilon \tag{8.23}$$

which proves consistency. □

8.2.3. Sample complexity proof.

Theorem 12 (Theorem 7 restated). Assume that the functions T_θ in $\mathcal{F}$ are L-Lipschitz with respect to the parameters θ; and that both T_θ and e^{T_θ} are M-bounded (i.e., $|T_\theta|, e^{T_\theta} \leq M$). The domain $\Theta \subset \mathbb{R}^d$ is bounded, so that $\|\theta\| \leq K$ for some constant K. Given any values ϵ,δ of the desired accuracy and confidence parameters, we have,

$$\Pr\left(|\widehat{I(X;Z)}_n - I_{\mathcal{F}}(X,Z)| \leq \epsilon\right) \geq 1-\delta \tag{8.24}$$

whenever the number n of samples satisfies

$$n \geq \frac{2M^2(d\log(16KL\sqrt{d}/\epsilon) + 2dM + \log(2/\delta))}{\epsilon^2} \tag{8.25}$$

Proof. The assumptions of Lemma 5 apply, so let us begin with Eqns. 8.19 and 8.20. By the Hoeffding inequality, for all function f,

$$\Pr\left(|\mathbb{E}_{\mathbb{Q}}[f] - \mathbb{E}_{\mathbb{Q}_n}[f]| \ > \frac{\epsilon}{4}\right) \leq 2\exp\left(-\frac{\epsilon^2 n}{2M^2}\right) \tag{8.26}$$

To extend this inequality to a uniform inequality over all functions T_θ and e^{T_θ}, the standard technique is to choose a minimal cover of the domain $\Theta \subset \mathbb{R}^d$ by a finite set of small balls of radius η, $\Theta \subset \cup_j B_\eta(\theta_j)$, and to use the union bound. The minimal cardinality of such covering is bounded by the covering number $N_\eta(\Theta)$ of Θ, known to satisfy(Shalev-Schwartz & Ben-David, 2014)

$$N_\eta(\Theta) \leq \left(\frac{2K\sqrt{d}}{\eta}\right)^d \tag{8.27}$$

Successively applying a union bound in Eqn 8.26 with the set of functions $\{T_{\theta_j}\}_j$ and $\{e^{T_{\theta_j}}\}_j$ gives

$$\Pr\left(\max_j |\mathbb{E}_{\mathbb{Q}}[T_{\theta_j}] - \mathbb{E}_{\mathbb{Q}_n}[T_{\theta_j}]| > \frac{\epsilon}{4}\right) \leq 2N_\eta(\Theta)\exp\left(-\frac{\epsilon^2 n}{2M^2}\right) \tag{8.28}$$

and

$$\Pr\left(\max_j |\mathbb{E}_{\mathbb{Q}}[e^{T_{\theta_j}}] - \mathbb{E}_{\mathbb{Q}_n}[e^{T_{\theta_j}}]| > \frac{\epsilon}{4}\right) \leq 2N_\eta(\Theta)\exp\left(-\frac{\epsilon^2 n}{2M^2}\right) \tag{8.29}$$

We now choose the ball radius to be $\eta = \frac{\epsilon}{8L}e^{-2M}$. Solving for n the inequation,

$$2N_\eta(\Theta)\exp\left(-\frac{\epsilon^2 n}{2M^2}\right) \leq \delta \tag{8.30}$$

we deduce from Eqn 8.28 that, whenever Eqn 8.25 holds, with probability at least $1-\delta$, for all $\theta \in \Theta$,

$$\begin{aligned}|\mathbb{E}_{\mathbb{Q}}[T_\theta] - \mathbb{E}_{\mathbb{Q}_n}[T_\theta]| &\leq |\mathbb{E}_{\mathbb{Q}}[T_\theta] - \mathbb{E}_{\mathbb{Q}}[T_{\theta_j}]| + |\mathbb{E}_{\mathbb{Q}}[T_{\theta_j}] - \mathbb{E}_{\mathbb{Q}_n}[T_{\theta_j}]| + |\mathbb{E}_{\mathbb{Q}_n}[T_{\theta_j}] - \mathbb{E}_{\mathbb{Q}_n}[T_\theta]| \\ &\leq \frac{\epsilon}{8}e^{-2M} + \frac{\epsilon}{4} + \frac{\epsilon}{8}e^{-2M} \\ &\leq \frac{\epsilon}{2} \end{aligned} \tag{8.31}$$

Similarly, using Eqn 8.20 and 8.29, we obtain that with probability at least $1-\delta$,

$$|\log \mathbb{E}_{\mathbb{Q}}[e^{T_\theta}] - \log \mathbb{E}_{\mathbb{Q}_n}[e^{T_\theta}]| \leq \frac{\epsilon}{2} \tag{8.32}$$

and hence using the triangular inequality,

$$|\widehat{I(X;Z)}_n - I_{\mathcal{F}}(X,Z)| \leq \epsilon \tag{8.33}$$

□

8.2.4. Bound on the reconstruction error. Here we clarify relationship between reconstruction error and mutual information, by proving the bound in Equ 5.3. We begin with a definition:

Definition 13 (Reconstruction Error). We consider encoder and decoder models giving conditional distributions $q(z|x)$ and $p(x|z)$ over the data and latent variables. If $q(x)$ denotes the marginal data distribution, the reconstruction error is defined as

$$\mathcal{R} = \mathbb{E}_{\boldsymbol{x}\sim q(\boldsymbol{x})}\mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z}|\boldsymbol{x})}[-\log p(\boldsymbol{x}|\boldsymbol{z})] \tag{8.34}$$

We can rewrite the reconstruction error in terms of the joints $q(\boldsymbol{x},\boldsymbol{z}) = q(\boldsymbol{z}|\boldsymbol{x})p(\boldsymbol{x})$ and $p(\boldsymbol{x},\boldsymbol{z}) = p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$. Elementary manipulations give:

$$\mathcal{R} = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{z})\sim q(\boldsymbol{x},\boldsymbol{z})} \log \frac{q(\boldsymbol{x},\boldsymbol{z})}{p(\boldsymbol{x},\boldsymbol{z})} - \mathbb{E}_{(\boldsymbol{x},\boldsymbol{z})\sim q(\boldsymbol{x},\boldsymbol{z})} \log q(\boldsymbol{x},\boldsymbol{z}) + \mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z})} \log p(\boldsymbol{z}) \tag{8.35}$$

where $q(\boldsymbol{z})$ is the aggregated posterior. The first term is the KL-divergence $D_{KL}(q \,||\, p)$; the second term is the joint entropy $H_q(\boldsymbol{x},\boldsymbol{z})$. The third term can be written as

$$\mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z})} \log p(\boldsymbol{z}) = -D_{KL}(q(\boldsymbol{z}) \,||\, p(\boldsymbol{z})) - H_q(\boldsymbol{z})$$

Finally, the identity

$$H_q(\boldsymbol{x},\boldsymbol{z}) - H_q(\boldsymbol{z}) := H_q(\boldsymbol{z}|\boldsymbol{x}) = H_q(\boldsymbol{z}) - I_q(\boldsymbol{x},\boldsymbol{z}) \tag{8.36}$$

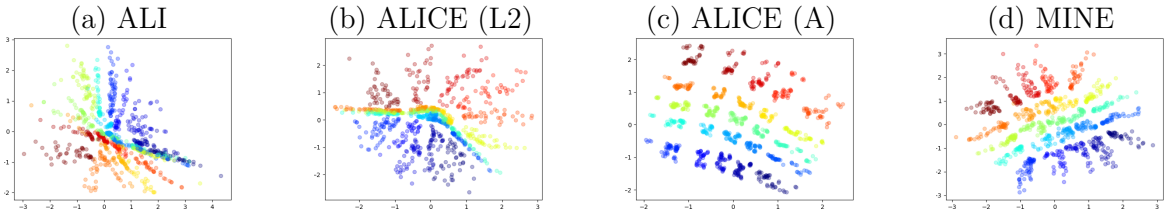


Fig. 6.7. Embeddings from adversarially learned inference (ALI) and variations intended to increase the mutual information. Shown left to right are the baseline (ALI), ALICE with the L2 loss to minimize the reconstruction error, ALI with an additional adversarial loss, and MINE.

yields the following expression for the reconstruction error:

$$\mathcal{R} = D_{KL}(q(\boldsymbol{x},\boldsymbol{z}) \mid\mid p(\boldsymbol{x},\boldsymbol{z})) - D_{KL}(q(\boldsymbol{z}) \mid\mid p(\boldsymbol{z})) - I_q(\boldsymbol{x},\boldsymbol{z}) + H_q(\boldsymbol{z}) \tag{8.37}$$

Since the KL-divergence is positive, we obtain the bound:

$$\mathcal{R} \leq D_{KL}(q(\boldsymbol{x},\boldsymbol{z}) \mid\mid p(\boldsymbol{x},\boldsymbol{z})) - I_q(\boldsymbol{x},\boldsymbol{z}) + H_q(\boldsymbol{z}) \tag{8.38}$$

which is tight whenever the induced marginal $q(\boldsymbol{z})$ matches the prior distribution $p(\boldsymbol{z})$.

8.3. Embeddings for bi-direction 25 Gaussians experiments

Here (Fig. 6.7) we present the embeddings for the experiments corresponding to Fig. 6.6.

Chapitre 7

Conclusion

Ce memoire utilise les formulations duales des divergences entre distributions et l'apprentissage de représentations par réseaux de neurones pour argumenter qu'il est possible d'offrir un estimateur de l'information mutuelle différentiable et tractable dans des régimes de données à forts volumes et en hautes dimensions. Nous concluons donc que:

L'estimation de neuronale de l'information mutuelle, ou d'autres mesures de dépendance basées sur d'autres divergences ou métrique sur les espaces de mesures de probabilités, par descente de gradient sur réseaux de neurones est prometteuse bien que loin d'être non-biaisée.

L'utilisation des estimateurs neuronaux de l'information mutuelle permet de contrôler la direction et la force de la dépendance entre des ensembles de variables aléatoires.

La differentiabilité de MINE permet de l'utiliser comme fonction de perte apprise. Ce concept a été utilisé pour régulariser les générateurs de modèles adversarielles génératifs, améliorer la qualité des reconstructions dans les modèles adversearielles à inférence jointe apprise, et d'implémenter une version continue du goulot d'étranglement d'information.

Nous espérons que l'estimation neuronale de l'information mutuelle ouvrira la voie à l'utilisation de l'apprentissage des représentations pour quantifier les dépendances entre variables des ensembles arbitraires de variables aléatoires.

Références bibliographiques

Achille, A. and Soatto, S. Emergence of invariance and disentanglement in deep representations. arXiv preprint 1706.01350v2[cs.LG], 2017.

Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. Deep variational information bottleneck. arXiv preprint arXiv:1612.00410, 2016.

Arimoto, S. An algorithm for computing the capacity of arbitrary discrete memoryless channels. IEEE Transactions on Information Theory, 18(1):14–20, 1972.

Aronszajn, N. Theory of reproducing kernels. Transactions of the American mathematical society, 68(3):337–404, 1950.

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. arXiv preprint arXiv:1607.06450, 2016.

Banerjee, A. On baysian bounds. ICML, pp. 81–88, 2006.

Barber, D. and Agakov, F. The im algorithm: a variational approach to information maximization. In Proceedings of the 16th International Conference on Neural Information Processing Systems, pp. 201–208. MIT Press, 2003.

Belghazi, M. I., Baratin, A., Rajeswar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, R. D. Mine: Mutual information neural estimation. CoRR, 2018a. URL http://arxiv.org/abs/1801.04062v4.

Belghazi, M. I., Rajeswar, S., Mastropietro, O., Mitrovic, J., Rostamzadeh, N., and Courville, A. Hierarchical adversarially learned inference. arXiv preprint arXiv:1802.01071, 2018b.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 35(8):1798–1828, 2013.

Borwein, J. and Lewis, A. S. Convex analysis and nonlinear optimization: theory and examples. Springer Science & Business Media, 2010.

Butte, A. J. and Kohane, I. S. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In Pac Symp Biocomput, volume 5, pp. 26, 2000.

Chalk, M., Marre, O., and Tkacik, G. Relevant sparse codes with variational information bottleneck. In Advances in Neural Information Processing Systems, pp. 1957–1965, 2016.

Che, T., Li, Y., Jacob, A. P., Bengio, Y., and Li, W. Mode regularized generative adversarial networks. arXiv preprint arXiv:1612.02136, 2016.

Chechik, G., Globerson, A., Tishby, N., and Weiss, Y. Information bottleneck for gaussian variables. Journal of Machine Learning Research, 6(Jan):165–188, 2005.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In Advances in Neural Information Processing Systems, pp. 2172–2180, 2016.

Clevert, D., Unterthiner, T., and Hochreiter, S. Fast and accurate deep network learning by exponential linear units (elus). CoRR, abs/1511.07289, 2015.

Cover, T. M. and Thomas, J. A. Elements of information theory. John Wiley & Sons, 2012.

Darbellay, G. A. and Vajda, I. Estimation of the information by an adaptive partitioning of the observation space. IEEE Transactions on Information Theory, 45(4):1315–1321, May 1999. ISSN 1557-9654. doi: 10.1109/18.761290.

Darbellay, G. A. and Vajda, I. Estimation of the information by an adaptive partitioning of the observation space. IEEE Transactions on Information Theory, 45(4):1315–1321, 1999.

Donahue, J., Krähenbühl, P., and Darrell, T. Adversarial feature learning. arXiv preprint arXiv:1605.09782, 2016.

Donsker, M. and Varadhan, S. Asymptotic evaluation of certain markov process expectations for large time, iv. Communications on Pure and Applied Mathematics, 36(2):183?212, 1983.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. Journal of machine learning research, 12(Jul):2121–2159, 2011.

Dumoulin, V., Belghazi, I., Poole, B., Lamb, A., Arjovsky, M., Mastropietro, O., and Courville, A. Adversarially learned inference. arXiv preprint arXiv:1606.00704, 2016.

Einstein, A. et al. The foundation of the general theory of relativity. Annalen der Physik, 49(7):769–822, 1916.

Fama, E. F. The behavior of stock-market prices. The journal of Business, 38(1):34–105, 1965.

Fraser, A. M. and Swinney, H. L. Independent coordinates for strange attractors from mutual information. Phys. Rev. A, 33:1134–1140, Feb 1986a. doi: 10.1103/PhysRevA.33.1134. URL https://link.aps.org/doi/10.1103/PhysRevA.33.1134.

Fraser, A. M. and Swinney, H. L. Independent coordinates for strange attractors from mutual information. Physical review A, 33(2):1134, 1986b.

Gao, S., Ver Steeg, G., and Galstyan, A. Efficient estimation of mutual information for strongly dependent variables. Arxiv preprint arXiv:1411.2003[cs.IT], 2014.

Ghosh, A., Kulharia, V., Namboodiri, V., Torr, P. H., and Dokania, P. K. Multi-agent diverse generative adversarial networks. arXiv preprint arXiv:1704.02906, 2017.

Glorot, X., Bordes, A., and Bengio, Y. Deep sparse rectifier neural networks. In Proceedings of the fourteenth international conference on artificial intelligence and statistics, pp. 315–323, 2011.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In Advances in neural information processing systems, pp. 2672–2680, 2014.

Goodfellow, I., Bengio, Y., and Courville, A. Deep learning. MIT press, 2016.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. Measuring statistical dependence with hilbert-schmidt norms. In International conference on algorithmic learning theory, pp. 63–77. Springer, 2005.

Györfi, L. and van der Meulen, E. C. Density-free convergence properties of various estimators of entropy. Computational Statistics and Data Analysis, 5:425?436, 1987.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE international conference on computer vision, pp. 1026–1034, 2015.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. Iclr, 2(5):6, 2017.

Hinton, G. E., Osindero, S., and Teh, Y. W. A fast learning algorithm for deep belief nets. Neural Computation, 18:1527–1554, 200.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. Learning deep representations by mutual information estimation and maximization. arXiv preprint arXiv:1808.06670, 2018.

Hochreiter, S. Untersuchungen zu dynamischen neuronalen netzen. Diploma, Technische Universität München, 91(1), 1991.

Hornik, K. Multilayer feedforward networks are universal approximators. Neural Networks, 2:359–366, 1989.

Hu, W., Miyato, T., Tokui, S., Matsumoto, E., and Sugiyama, M. Learning discrete representations via information maximizing self-augmented training. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pp. 1558–1567. JMLR. org, 2017.

Hulle, M. M. V. Edgeworth approximation of multivariate differential entropy. Neural computation, 17(9):1903–1910, 2005.

Hyvärinen, A. and Oja, E. Independent component analysis: algorithms and applications. Neural networks, 13(4-5):411–430, 2000.

Hyvärinen, A., Karhunen, J., and Oja, E. Independent component analysis, volume 46. John Wiley & Sons, 2004.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. CoRR, abs/1502.03167, 2015a. URL http://arxiv.org/abs/1502.03167.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167, 2015b.

Kandasamy, K., Krishnamurthy, A., Poczos, B., Wasserman, L., and Robins, J. Nonparametric von mises estimators for entropies, divergences and mutual informations. NIPS, 2017.

Keziou, A. Dual representation of Ï -divergences and applications. 336:857–862, 05 2003.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. CoRR, abs/1412.6980, 2014. URL http://arxiv.org/abs/1412.6980.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.

Kinney, J. B. and Atwal, G. S. Equitability, mutual information, and the maximal information coefficient. Proceedings of the National Academy of Sciences, 111(9):3354–3359, 2014.

Kolchinsky, A., Tracey, B. D., and Wolpert, D. H. Nonlinear information bottleneck. arXiv preprint arXiv:1705.02436, 2017.

Kozachenko, L. and Leonenko, N. N. Sample estimate of the entropy of a random vector. Problemy Peredachi Informatsii, 23(2):9–16, 1987.

Kraskov, A., Stögbauer, H., and Grassberger, P. Estimating mutual information. Physical review E, 69(6):066138, 2004.

Kullback, S. Information theory and statistics. Courier Corporation, 1997.

Kumar, A., Sattigeri, P., and Fletcher, P. T. Improved semi-supervised learning with gans using manifold invariances. arXiv preprint arXiv:1705.08850, 2017.

Kwak, N. and Choi, C.-H. Input feature selection by mutual information based on parzen window. IEEE transactions on pattern analysis and machine intelligence, 24(12):1667–1671, 2002.

Lagrange, J.-L. Mécanique analytique, paris. MJ Bertrand. Mallet-Bachelier.)[aPJHS], 1788.

LeCun, Y. The mnist database of handwritten digits. http://yann. lecun. com/exdb/mnist/, 1998.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.

LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., and Huang, F. A tutorial on energy-based learning. Predicting structured data, 1(0), 2006.

Li, C., Liu, H., Chen, C., Pu, Y., Chen, L., Henao, R., and Carin, L. Towards understanding adversarial learning for joint distribution matching. arXiv preprint arXiv:1709.01215, 2017.

Lin, Z., Khetan, A., Fanti, G., and Oh, S. Pacgan: The power of two samples in generative adversarial networks. arXiv preprint arXiv:1712.04086, 2017.

Linsker, R. Self-organization in a perceptual network. Computer, 21(3):105–117, 1988.

Linsker, R. An application of the principle of maximum information preservation to linear systems. In Advances in neural information processing systems, pp. 186–194, 1989.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In Proceedings of the IEEE International Conference on Computer Vision, pp. 3730–3738,

2015.

Lu, Z., Pu, H., Wang, F., Hu, Z., and Wang, L. The expressive power of neural networks: A view from the width. In Advances in neural information processing systems, pp. 6231–6239, 2017.

Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., and Suetens, P. Multimodality image registration by maximization of mutual information. IEEE transactions on Medical Imaging, 16(2):187–198, 1997.

Mandelbrot, B. B. and Wallis, J. R. Some long-run properties of geophysical records. Water resources research, 5(2):321–340, 1969.

McCullagh, P. Tensor methods in statistics. Courier Dover Publications, 2018.

Metz, L., Poole, B., Pfau, D., and Sohl-Dickstein, J. Unrolled generative adversarial networks. 2017. URL https://openreview.net/pdf?id=BydrOIcle.

Mises, R. and Pollaczek-Geiringer, H. Praktische verfahren der gleichungsauflösung. ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik, 9(2):152–164, 1929.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. arXiv preprint arXiv:1802.05957, 2018.

Mode, C. J. and Sleeman, C. K. Stochastic processes in epidemiology: HIV/AIDS, other infectious diseases, and computers. World Scientific, 2000.

Mohamed, S. and Rezende, D. J. Variational information maximisation for intrinsically motivated reinforcement learning. In Advances in neural information processing systems, pp. 2125–2133, 2015.

Moon, K., Sricharan, K., and Hero III, A. O. Ensemble estimation of mutual information. arXiv preprint arXiv:1701.08083, 2017.

Moon, Y.-I., Rajagopalan, B., and Lall, U. Estimation of mutual information using kernel density estimators. Physical Review E, 52(3):2318, 1995.

Nguyen, T., Le, T., Vu, H., and Phung, D. Dual discriminator generative adversarial nets. In Advances in Neural Information Processing Systems, pp. 2667–2677, 2017.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. Estimating divergence functionals and the likelihood ratio by convex risk minimization. IEEE Transactions on Information Theory, 56(11):5847–5861, 2010.

Nowozin, S., Cseke, B., and Tomioka, R. f-gan: Training generative neural samplers using variational divergence minimization. In Advances in Neural Information Processing Systems, pp. 271–279, 2016.

Paninski, L. Estimation of entropy and mutual information. Neural computation, 15(6): 1191–1253, 2003.

Peng, H., Long, F., and Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on pattern analysis and machine intelligence, 27(8):1226–1238, 2005.

Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł., and Hinton, G. Regularizing neural networks by penalizing confident output distributions. ICLR Workshop, 2017.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434, 2015.

Rényi, A. et al. On measures of entropy and information. In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics. The Regents of the University of California, 1961.

Riesz, F. Démonstration nouvelle d'un théorème concernant les opérations fonctionnelles linéaires. In Annales scientifiques de l'École Normale Supérieure, volume 31, pp. 9–14, 1914.

Ruderman, A., Reid, M., García-García, D., and Petterson, J. Tighter variational representations of f-divergences via restriction to probability measures. arXiv preprint arXiv:1206.4664, 2012.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning representations by back-propagating errors. nature, 323(6088):533–536, 1986.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. Imagenet large scale visual recognition challenge. CoRR, 2014. URL http://arxiv.org/abs/1409.0575v3.

Saatchi, Y. and Wilson, A. G. Bayesian gan. In Advances in Neural Information Processing Systems, pp. 3625–3634, 2017.

Salimans, T. and Kingma, D. P. Weight normalization: a simple reparameterization to accelerate training of deep neural networks. CoRR, 2016. URL http://arxiv.org/abs/1602.07868v3.

Salimans, T., Goodfellow, I. J., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. arXiv preprint arXiv:1606.03498, 2016.

Santurkar, S., Tsipras, D., Ilyas, A., and Madry, A. How does batch normalization help optimization? In Advances in Neural Information Processing Systems, pp. 2483–2493, 2018.

Schölkopf, B., Smola, A., and Müller, K.-R. Kernel principal component analysis. In International conference on artificial neural networks, pp. 583–588. Springer, 1997.

Shalev-Schwartz, S. and Ben-David, S. Understanding Machine Learning - from Theory to Algorithms. Cambridge university press, 2014.

Singh, S. and Póczos, B. Finite-sample analysis of fixed-k nearest neighbor density functional estimators. arXiv preprint 1606.01554, 2016.

Srivastava, A., Valkov, L., Russell, C., Gutmann, M., and Sutton, C. Veegan: Reducing mode collapse in gans using implicit variational learning. arXiv preprint arXiv:1705.07761, 2017.

Suzuki, T., Sugiyama, M., Sese, J., and Kanamori, T. Approximating mutual information by maximum likelihood density ratio estimation. In New challenges for feature selection in data mining and knowledge discovery, pp. 5–20, 2008.

Tieleman, T. and Hinton, G. Divide the gradient by a running average of its recent magnitude. coursera: Neural networks for machine learning. Technical Report., 2017.

Tishby, N. and Zaslavsky, N. Deep learning and the information bottleneck principle. In Information Theory Workshop (ITW), 2015 IEEE, pp. 1–5. IEEE, 2015.

Tishby, N., Pereira, F. C., and Bialek, W. The information bottleneck method. arXiv preprint physics/0004057, 2000.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022, 2016.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. Adversarial generator-encoder networks. arXiv preprint arXiv:1704.02304, 2017.

Van de Geer, S. Empirical Processes in M-estimation. Cambridge University Press, 2000.

Van Hulle, M. M. Edgeworth approximation of multivariate differential entropy. Neural computation, 17(9):1903–1910, 2005.

Veličković, P., Fedus, W., Hamilton, W. L., Liò, P., Bengio, Y., and Hjelm, R. D. Deep graph infomax. arXiv preprint arXiv:1809.10341, 2018.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of machine learning research, 11(Dec):3371–3408, 2010.

Wang, Q., Kulkarni, S. R., and Verdú, S. Divergence estimation of continuous distributions based on data-dependent partitions. IEEE Transactions on Information Theory, 51(9): 3064–3074, 2005.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13: 600–612, 2004.

Wu, Y. and He, K. Group normalization. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19, 2018.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. In European conference on computer vision, pp. 818–833. Springer, 2014.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. arXiv preprint arXiv:1703.10593, 2017.