

Université de Montréal

**Reconstruction libre de lentilles gravitationnelles de type
galaxie-galaxie avec les machines à inférence récurrentielle**

par

Alexandre Adam

Département de physique
Faculté des arts et des sciences

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

31 décembre 2022

Université de Montréal
Faculté des arts et des sciences

Ce mémoire intitulé

**Reconstruction libre de lentilles gravitationnelles de type galaxie-galaxie avec
les machines à inférence récurrentielle**

présenté par

Alexandre Adam

a été évalué par un jury composé des personnes suivantes :

René Doyon

(président-rapporteur)

Laurence Perreault-Levasseur

(directeur de recherche)

Yashar Hezaveh

(membre du jury)

Résumé

Les lentilles gravitationnelles de type galaxie-galaxie se produisent lorsque la lumière d'une galaxie en arrière-plan est déviée par le champ gravitationnel d'une galaxie en avant-plan, formant des images multiples ou même des anneaux d'Einstein selon le point de vue d'un observateur sur Terre. Ces phénomènes permettent non seulement d'étudier les galaxies lointaines, magnifiées par la galaxie-lentille, mais aussi de comprendre la distribution de masse de la galaxie-lentille et de son environnement, une opportunité unique pour sonder la matière noire contenue dans ces galaxies. Or, les méthodes traditionnelles pour analyser ces systèmes requièrent une quantité significative de temps ordinateur (de quelques heures à quelques jours), sans compter le temps des experts pour faire converger les analyses MCMC requises pour obtenir les paramètres d'intérêts. Ce problème est significatif, considérant qu'il est projeté que les grands relevés du ciel comme ceux qui seront menés aux observatoires Rubin et Euclid découvriront plusieurs centaines de milliers de lentilles gravitationnelles. De plus, le Télescope géant européen (ELT), faisant usage de la technologie d'optique adaptative, et le télescope spatial James Webb, vont nous offrir une vue sans précédent de ces systèmes, avec un pouvoir de résolution qui rendra possible certaines analyses comme la recherche de halo de matière noire froide, longtemps prédictive par le modèle cosmologique standard Λ CDM. Les approximations traditionnelles faites pour simplifier la reconstruction des lentilles gravitationnelles ne seront plus valides dans ce régime.

Dans ce mémoire, je présente un travail qui s'attaque à ces deux problèmes. Je présente une méthode d'optimisation basée sur les machines à inférence récurrente pour reconstruire deux images, soit celle d'une galaxie en arrière-plan et une image pour la distribution de masse de la galaxie en avant-plan. La représentation paramétrique choisie a le potentiel de reconstruire une classe très large de lentilles gravitationnelles, incluant des halos et sous-halos de matière noire, ce qu'on démontre dans ce travail en utilisant des profiles de densité réalistes provenant de la simulation cosmologique hydrodynamique IllustrisTNG. Nos reconstructions atteignent un niveau de réalisme jamais atteint auparavant et s'exécutent sur une fraction du temps requis pour exécuter une analyse traditionnelle, soit un pas significatif vers une méthode pouvant adresser le défi d'analyser autant de systèmes complexes et variés en un temps à l'échelle humaine.

Mots-clés : Lentilles gravitationnelles — Simulations astrophysiques — Inférence non-paramétrique

— Réseaux neuronaux convolutifs.

Abstract

Galaxy-Galaxy gravitational lenses is a phenomenon that happens when the light coming from a background galaxy is bent by the gravitational field of a foreground galaxy, producing multiple images or even Einstein ring images of the background source from the point of view of an observer on Earth. These phenomena allow us to study in detail the morphology of the background galaxy, magnified by the lens, but also study the mass density distribution of the lens and its environment, thus offering a unique probe of dark matter in lensing galaxies. Traditional methods studying these systems often need significant compute time (from hours to days), and this is without taking into account the time spent by experts to make the MCMC chains required to obtain parameters of interest converge. This problem is significant, considering that large surveys from observatories like Rubin and Euclid are projected to discover hundreds of thousands of gravitational lenses. Moreover, the Extremely Large Telescope (ELT), using adaptive optics, and the James Webb Space Telescope will offer an unprecedented glimpse of these systems, with a resolving power predicted to enable searches for cold dark matter subhalos — objects long predicted by the standard cosmological model Λ CDM. Approximations used to make analysis tractable in traditional methods will no longer be valid in that regime.

In this thesis, I present a method that aims to address these two issues. The method, based on Recurrent Inference Machines (RIM), reconstructs two pixelated maps, one for the background source and another for the mass density map of the foreground lensing galaxy. This free-form parametric representation has the potential to reconstruct a large class of gravitational lenses, including those with dark matter halos and subhalos, which we demonstrate using realistic mass density profiles from the cosmological hydrodynamic simulation IllustrisTNG. Our method can achieve an unmatched level of realism in a fraction of the time required by traditional methods, which is a significant step toward solving the challenge of studying such a large number of complex and varied systems in a human timescale.

Keywords: Gravitational lensing — Astronomical simulations — Nonparametric inference — Convolutional Neural Networks.

Table des matières

Résumé

Abstract

Liste des tableaux

Liste des figures

Acronymes

Symboles

Remerciements

1	Introduction	1
1.1	Description du mémoire	2
1.2	Déclaration de l'étudiant	3
2	Lentilles gravitationnelles fortes de type galaxie-galaxie	4
2.1	Les angles de déflexion	7
3	Introduction à l'apprentissage profond	12
3.1	Survol des statistiques bayésiennes	12
3.2	Un exemple d'apprentissage machine : la régression	15
3.3	Sélection du modèle	19
3.3.1	Compromis entre le biais et la variance	20
3.3.2	La séparabilité linéaire	21

3.4	Les réseaux de neurones	23
3.5	Les réseaux de neurones convolutifs	26
4	Apprentissage profond de distributions implicites	28
4.1	Auto-encodeur variationnel	28
4.1.1	Le truc de la reparamétrisation	29
4.1.2	Principe du goulot d'information	31
4.2	Machines à inférence récurrentielles	34
4.2.1	Formalisme bayésien des problèmes inverses	34
4.2.2	La relation de récurrence	36
4.2.3	Méta-apprentissage par rétropropagation de gradients	38
5	Pixelated Reconstruction of Foreground Density and Background Surface Brightness in Gravitational Lensing Systems using Recurrent Inference Machines	42
5.1	Introduction	44
5.2	Methods	46
5.2.1	Maximum a posteriori inference	46
5.2.2	The Forward Model	47
5.2.3	Recurrent Inference Machine	48
5.2.4	The Neural Network	49
5.2.5	Fine-Tuning	50
5.3	Data	53
5.3.1	COSMOS	53
5.3.2	IllustrisTNG	54
5.3.3	Simulated Observations	56
5.4	Training	57
5.4.1	VAE	57
5.4.2	RIM	58
5.5	Results	61
5.5.1	Goodness of Fit	61
5.5.2	Quality of the Reconstructions	64
5.6	Conclusion	64

6 Conclusion	67
Bibliographie	69
A ΛCDM	82
B Elastic Weight Consolidation	83
C VAE Architecture and optimisation	85
D RIM architecture and optimisation	88
E GRU	91
F Congrès où l'étudiant à présenté ses résultats	92

Liste des tableaux

3.1	Fonction logique XOR.	22
5.1	Physical model parameters.	56
5.2	SIE parameters.	57
5.3	Hyperparameters for fine-tuning the RIM.	58
5.4	\log_{10} -normal moments of the loss on the test set	61
A.1	Paramètres de Λ CDM ajusté avec les observations du fond diffus cosmologique par le télescope Planck (Planck Collaboration, 2020)	82
C.1	Hyperparameters for the background source VAE.	86
C.2	Hyperparameters for the convergence VAE.	87
D.1	Hyperparameters for the RIM.	90

Liste des figures

2.1	Le quasar double (QSO 0957+561 A et B) et la galaxie-lentille (G) imagée par le télescope spatial Hubble. Crédit : ESA/Hubble et NASA, élargissement et annotation par AA.	5
2.2	Lentilles gravitationnelles de type galaxie-galaxie.	6
2.3	Schéma d'une lentille gravitationnelle.	9
3.1	Exemple d'une inférence bayésienne pour le tirage au sort.	14
3.2	Exemple d'un problème de régression.	16
3.3	Contours de $\mathcal{L}_\theta(\mathcal{D})$ pour différent tirage de \mathcal{D} (rangées) et différentes taille $N = \mathcal{D} $ (colonnes) en utilisant le modèle linéaire de l'équation (3.10) et la loi générative (3.8). Une trajectoire produite par la descente de gradient (3.18) est illustré avec les flèches noires.	18
3.4	Compromis classique entre le biais et la variance d'un algorithme d'apprentissage machine pour l'ajustement d'un polynôme de degré P sur les données générées de la loi $f_{\theta^*} = 2x^5 - x$.	20
3.5	Comparaison d'un modèle linéaire et d'un modèle quadratique pour le problème XOR.	22
3.6	Illustration d'un réseau de neurones avec deux couches latentes qui constituent son espace caractéristique.	24
3.7	Illustration d'une transformation du problème XOR vers un espace caractéristique linéairement séparable par un réseau de neurones.	25
4.1	Modèle graphique d'un VAE. Les flèches pleines indiquent le processus génératif, alors que les flèches pointillées indiquent le processus d'inférence.	29
4.2	VAE comme un système de transmission d'information.	31
5.1	Rolled computational graph of the RIM. Dashed arrows represent operations not recorded for BPTT.	49

5.2	A single time step of the unrolled computation graph of the RIM. GRU units are placed in the skip connections to guide the reconstruction of the source and convergence. A schematic of the steps to compute the likelihood gradients is shown in the bottom right of the figure, including the Adam processing step of the likelihood gradient.	50
5.3	Example of a simulated lensed image in the test set that exhibits a large deflection in its eastern arc which indicates the presence of a massive object — in this case a dark matter subhalo. The fine-tuning procedure is able to recover this subhalo because of its strong signal in the lensed image and reduces the residuals to noise level.	52
5.4	Examples similar to the test task, also shown in Figure 5.3. The first column shows the ground truth used to simulate the lensed image. The second column shows the baseline prediction that is then encoded in the latent space of the VAE in order to sample the next 4 columns.	53
5.5	Examples of COSMOS galaxy images (top row) and VAE generated samples (bottom row) used as labels in \mathcal{D}	54
5.6	Examples of smoothed Illustris TNG100 convergence map (top row) and VAE generated samples (bottom row) used as labels in \mathcal{D}	55
5.7	Sample of the fine-tuned RIM reconstructions on a test set of 3000 examples. Examples are ordered from the best χ^2 (top) to the worst (bottom). The percentile rank of each example is in the leftmost column. The last example shown has SNR above the threshold defined in Figure 5.10.	59
5.8	30 reconstructions taken at random from the test set of 3000 examples simulated from COSMOS and IllustrisTNG data at high SNR. The colorscale are the same as in Figure 5.7.	60
5.9	Distribution of the goodness of fit for the baseline and fine-tuned network (right panel), as well as log-loss difference between the two network for a given example in the test set (left panel).	61
5.10	Goodness of fit as a function of SNR shows a threshold behavior where our method reaches its limit.	62
5.11	Comparison between baseline (RIM) and fine-tuned (RIM+FT) reconstructions for gravitational lensing systems from the test set (GT). From top to bottom, we increase SNR.	63
5.12	Statistics of the coherence spectrum on the test set. The solid line is the average coherence. The transparent region is the 68% confidence interval. The fine-tuning procedure yields a noticeable improvement on the coherence of the source at all frequencies.	64

6.1 Reconstruction du fer à cheval cosmique avec la machine à inférence récurrentielle décrite dans le chapitre 5.	68
D.1 Unrolled computational graph of the RIM. Operations along solid arrows are being recorded for BPTT, while operations along dashed arrows are not. The blue arrows are only used for optimisation during training. During fine-tuning or testing, the loss is computed only as an oracle metric to validate that our methods can recover the ground truth.	89

Acronymes

RIM Recurrent Inference Machine — Machine à inférence récurrentielle.

VAE Variational AutoEncoder — Auto-encodeur variationnel de Bayes.

GRU Gated Recurrent Unit — Unité récurrentielle à porte.

BPTT BackPropagation Through Time — Rétropropagation temporelle des gradients.

LSTM Long Short Term Memory unit — Unité à mémoire longue et courte.

RNN Recurrent Neural Network — Réseau de neurones récurrent.

ADAM ADaptive Momentum estimation — Estimation adaptive de l'impulsion.

RMSProp Root Mean Squared Propagation — Propagation de la moyenne quadratique.

MAP Maximum A Posteriori.

MLE Maximum Likelihood Estimate — Maximum de la vraisemblance.

ELBO Evidence Lower BOund — Limite inférieur sur l'évidence.

KL Kullback-Leibler.

BIP Bottleneck Information Principle — Principe du goulot d'information.

HST Hubble Space Telescope.

QSO Quasi-Stellar Object — Source de rayonnement quasi-stellaire.

WFC3 Wide Field Camera 3.

Symboles

$\mathbb{1}$ Matrice identité.

\mathbb{R} Ensemble des nombres réels.

π Pi.

∇ Gradient.

∇^2 Laplacien.

κ Convergence — densité surfacique de masse projetée sur l'axe de visée.

α Angles de déflexion.

β Coordonnées angulaires du plan de la source.

θ Coordonnées angulaires du plan de la lentille.

ξ Coordonnées comobiles sur le plan de la lentille.

η Coordonnées comobiles sur le plan de la source.

D_s Distance du diamètre angulaire entre l'observateur et la source.

D_ℓ Distance du diamètre angulaire entre l'observateur et la lentille.

$D_{\ell s}$ Distance du diamètre angulaire entre la lentille et la source.

$g_{\mu\nu}$ Un élément de la métrique.

$\eta_{\mu\nu}$ Un élément de la métrique de Minkowski.

\mathcal{L} Lagrangien.

z Décalage vers le rouge.

c Vitesse de la lumière.

G Constante universelle de la gravitation.

- ρ Densité.
- Σ Densité de surface.
- Σ_c Densité de surface critique.
- Φ Potentiel.
- φ Liste des paramètres pour l'algorithme d'inférence d'un problème inverse.
- ϕ Liste des paramètres pour un processus d'inférence.
- θ Liste des paramètres pour un processus génératif.
- \mathcal{L}_φ Fonction objective d'entraînement pour les paramètres φ .
- $\hat{\mathbf{x}}^{(t)}$ Estimé de vecteur des paramètres physiques après t itérations de la relation de récurrence.
- \mathbf{y} Vecteur des quantités observées.
- F Modèle physique.
- \mathcal{X} Espace vectoriel des paramètres physiques.
- \mathcal{Y} Espace vectoriel des quantités observées.
- \mathbf{z} Variable latente.
- $\mathbf{h}^{(t)}$ État latent d'une cellule mémoire après t itérations de la relation de récurrence.
- t Paramètre du temps (continu) ou indice d'une relation de récurrence (discret).
- T Nombre total d'itérations de la relation de récurrence.
- \mathcal{D} Ensemble de données d'entraînement.
- \mathcal{H} Ensemble d'hypothèses.
- \mathcal{G} Algorithme d'optimisation.
- \mathcal{T} Ensemble de données d'essai.
- \mathcal{I} Information de Fisher.
- H** Hessienne.
- $D_{\text{KL}}(\cdot \parallel \cdot)$ Distance de Kullback-Leibler.
- $\mathbb{E}[\cdot]$ Opérateur de l'espérance mathématique.
- iid Identiquement et indépendamment distribué.

$\|\cdot\|_2$ Norme euclidienne.

$I(\cdot; \cdot)$ Information mutuelle.

$H(\cdot)$ Entropie.

\mathcal{N} Loi normale.

$\mathcal{T}\mathcal{N}$ Loi normale tronquée.

\mathcal{U} Loi uniforme.

μ Moyenne.

Σ Covariance.

σ^2 Variance.

σ Déviation standard.

\oplus Concaténation.

\odot Produit d'Hadamard.

À Maman et Julia

Remerciements

Je tiens à remercier ma superviseure, Laurence Perreault-Levasseur, pour ces heures innombrables passées à discuter de ma recherche, à me pousser et m'inspirer à réfléchir aux problèmes à toutes les échelles, des petits bugs jusqu'aux problèmes les plus ambitieux.

Je tiens à remercier Yashar Hezaveh, pour ses encouragements et sa passion contagieuse pour la recherche, qui m'a infecté même après deux ans d'isolation.

Je tiens aussi à remercier mes collègues, Ronan Legin, Ève Campeau-Poirier et Charles Wilson qui ont été là durant deux ans de pandémie pour parler de la vie, de la recherche et de toutes les choses entre les deux, virtuellement ou autrement.

Et surtout, merci à mes parents, mon frère et ma copine pour leur support inconditionnel dans cette aventure que j'ai entrepris.

Chapitre 1

Introduction

Dans la quête de comprendre la naissance et l'évolution de l'Univers, la théorie de la cosmologie moderne est arrivée à une conclusion surprenante : seulement 4 ingrédients sont nécessaires pour expliquer l'ensemble des structures de l'Univers, des échelles cosmiques jusqu'à l'échelle des plus petites galaxies. De ces 4 ingrédients, soit les baryons, la radiation électromagnétique, la matière noire et l'énergie sombre, presque rien n'est connu de la matière noire et de l'énergie sombre. La particule de matière noire prédicta par le modèle standard de la cosmologie, Λ CDM, ne fait pas partie du modèle standard des particules, bien qu'elle soit maintenant un ingrédient essentiel à notre compréhension du fond diffus cosmologique ([Planck Collaboration, 2020](#)) et de la formation de la toile cosmique et des galaxies.

La seule propriété attendue et essentielle de cette particule est qu'elle n'interagit pas avec le champ électromagnétique. Les halos de matières noires peuvent alors commencer à s'effondrer par instabilité gravitationnelle avant la période du découplage du rayonnement, soit le moment où le taux de la diffusion Compton est égal au taux d'expansion de l'Univers. À ce moment, les particules chargées sont en mesure de se combiner pour former les premiers atomes neutres. Les halos de matière noire deviennent les trous de potentiel nécessaire pour accélérer l'effondrement gravitationnel du gaz baryonique primordial jusqu'à la production des larges structures observées aujourd'hui ([White et Rees, 1978](#)).

La nature élusive de la matière noire nous force à utiliser des méthodes de plus en plus sophistiquées pour l'étudier. Étant donné que cette particule n'interagit pas avec la lumière, nos télescopes ne peuvent pas détecter directement ces particules. Plutôt, on est forcé de chercher les traces gravitationnelles de ces halos de matière noires. Les lentilles gravitationnelles fortes, introduites plus en détail au chapitre 2, devraient nous permettre d'accomplir précisément cet objectif à condition qu'on soit en mesure de développer les algorithmes d'inférences nécessaire.

Les méthodes traditionnelles pour analyser les lentilles gravitationnelles requièrent une quantité significative de temps d'ordinateur (de quelques heures à quelques jours), sans compter le temps

des experts pour faire converger les analyses MCMC requises pour obtenir les paramètres d'intérêts. Ce problème est significatif, considérant qu'il est projeté que les grands relevés du ciel comme ceux qui seront menés aux observatoires Rubin et Euclid découvriront plusieurs centaines de milliers de lentilles gravitationnelles. De plus, le Télescope géant européen (ELT), faisant usage de la technologie d'optique adaptative, et le télescope spatial James Webb, vont nous offrir une vue sans précédent de ces systèmes, avec un pouvoir de résolution qui rendra possible la recherche de halo de matière noire froide (p. ex. [Coogan et al., 2020](#)), longtemps prédictive par le modèle cosmologique standard Λ CDM. Or, les approximations utilisées par les méthodes traditionnelles pour simplifier l'optimisation restreignent les distributions de masses considérées pour la galaxie-lentille à des lois de puissance (p. ex. [Nightingale et al., 2018](#); [Etherington et al., 2022](#)). Dans le régime à haute résolution des télescopes modernes, ou encore pour modéliser un ensemble aussi large de lentilles gravitationnelles, ces approximations deviennent encombrantes et peuvent biaiser l'inférence.

Dans ce mémoire, je présente une méthodologie pour l'inférence de lentilles gravitationnelles avec un niveau de réalisme sans précédent. Plutôt que de supposer que la distribution de masse possède un profil de densité simple, décrit par quelques paramètres, ou que le vrai profil est une petite perturbation autour de cette solution initiale ([Birrer et al., 2015](#); [Birrer et Amara, 2018](#)), j'entreprends de reconstruire une image d'un profile réaliste provenant de la simulation cosmologique hydrodynamique IllustrisTNG ([Nelson et al., 2019](#)). Une telle reconstruction libre du profil a le potentiel d'automatiser l'analyse de plusieurs milliers d'images de lentilles gravitationnelles à haute résolution provenant des télescope moderne et va ouvrir une fenêtre unique pour l'étude des propriétés de la matière noire.

Depuis les premières tentatives de reconstructions libres de profils de densité ([Saha et Williams, 1997](#)), les méthodes introduites dans le chapitre 3 forment le premier cadre d'analyse complet permettant de résoudre le problème de reconstruction libre de lentilles gravitationnelles, un problème inverse mal-posé et non linéaire. Finalement, les résultats présentés aux chapitres 5 montrent que notre méthode est suffisamment expressive pour résoudre des systèmes complexes, avec une précision suffisante pour produire des reconstructions statistiquement significatives étant donné des observations à haut signal sur bruit et à haute résolution.

1.1 Description du mémoire

L'objectif principal de ce mémoire est de développer une méthode permettant de modéliser la distribution de masse et la morphologie de la source dédiée à analyser un grand nombre ($> 10^3$) de lentilles gravitationnelles dans toute leur complexité et dans un temps à l'échelle humaine.

Le chapitre 2 est une introduction aux lentilles gravitationnelles fortes. Le chapitre 3 est une introduction à l'apprentissage machine avec une perspective probabiliste. Ensuite, les sections 4.1 et 4.2 du chapitre 4 introduisent les méthodes d'apprentissage profond pour résoudre le problème de reconstruction libre de lentilles gravitationnelles de type galaxie-galaxie. Finalement, le chapitre

5 est un rapport détaillé de l'application de ces méthodes appliquées à des lentilles gravitationnelles simulées avec des profils de densités et des images de galaxies réalistes.

1.2 Déclaration de l'étudiant

Je, Alexandre Adam, déclare que l'entièreté du travail présenté dans ce mémoire est le mien. J'ai effectué la revue de la littérature dans ce mémoire. Lorsque j'ai utilisé des figures provenant de sources externes, j'ai clairement identifié le titre de la figure avec la source associée.

Pour l'article présenté au chapitre 5, j'ai modifié un code et des méthodes originellement développées par Laurence Perreault-Levasseur et Yashar Hezaveh pour construire des profils de masses à partir de la simulation IllustrisTNG, simuler des lentilles gravitationnelles à partir de ces profils et faire l'inférence avec une machine à inférence récurrentielle. Ma contribution à ce projet est la production des profils de convergence à partir de la simulation IllustrisTNG, le pré-traitement d'un ensemble d'entraînement pour les images de galaxies à partir du champ large COSMOS, le développement du code d'entraînement pour la machine à inférence récurrentielle et pour les auto-encodeurs variationnels, le développement d'un code de recherche d'hyperparamètres et d'architectures pour ces modèles, la production des résultats et finalement le développement de la méthode de réglage fin de la machine à inférence récurrentielle, ainsi que son interprétation bayésienne.

Chapitre 2

Lentilles gravitationnelles fortes de type galaxie-galaxie

Fritz Zwicky (1937), suivant les calculs publiés par Einstein (1936) et la première observation de l'effet de déviation gravitationnelle de la lumière par Eddington (1919), est largement reconnu comme étant le premier à observer correctement qu'une lentille gravitationnelle, et en particulier l'anneau d'Einstein (Chwolson, 1924), est un phénomène particulièrement riche en information¹. L'article de Zwicky (1937) articule précisément deux idées centrales qui nous motivent encore aujourd'hui à étudier ces objets. En premier lieu, une lentille gravitationnelle est un télescope naturel, de sorte qu'un tel système nous permettrait en principe d'étudier l'image lentillée de la source en arrière-plan avec une résolution beaucoup plus grande que nos instruments nous le permettraient si l'effet de lentille n'avait pas eu lieu. En second lieu, la déflexion de l'image de la source est directement proportionnelle à la masse (gravitationnelle) de la lentille.

$$\theta_E = \sqrt{\frac{4GM}{c^2 D}} \simeq 3 \left(\frac{M}{M_\odot} \right)^{\frac{1}{2}} \left(\frac{D}{1 \text{ Gpc}} \right)^{-\frac{1}{2}} \mu\text{as}, \quad \left\{ D \equiv \frac{D_\ell D_s}{D_{\ell s}} \right\}. \quad (2.1)$$

Par exemple, une galaxie typique de masse $M \sim 10^{11} M_\odot$, à une distance caractéristique $D = 3 \text{ Gpc}$ produirait des images de la source séparées par $2\theta_E \sim 1''$. C'est cette observation qui intéressait particulièrement Zwicky (1937), insatisfait par les méthodes pour mesurer la masse des nébuleuses extragalactiques (galaxies) de l'époque, basées largement sur des comparaisons de la luminosité totale de ces galaxies avec L_\odot , la luminosité du Soleil, ou des courbes de rotation képlériennes.²

¹ Les travaux pionniers de František Link (1936, 1937), largement ignorés dans la littérature anglo-saxonne, offrent déjà une perspective riche et détaillée sur le phénomène des lentilles gravitationnelles au moment où Zwicky (1937) publie ses observations. En particulier, Link (1936) décrit la magnification d'une étoile lors du passage derrière un objet massif et observe que les amas globulaires et les galaxies sont des candidats idéaux pour une recherche systématique du phénomène.

² Zwicky (1937) propose d'ailleurs un estimé de la masse de l'amas de Coma à $\gtrsim 4.5 \times 10^{13} M_\odot$ avec le théorème du viriel. Cette limite inférieure est un très bon estimé de la valeur acceptée aujourd'hui, dérivée avec les effets de lentilles faibles produites par l'amas sur l'image des galaxies environnantes, soit $5_{-2.1}^{+4.3} \times 10^{14} h_{70}^{-1} M_\odot$ (Gavazzi et al., 2009).

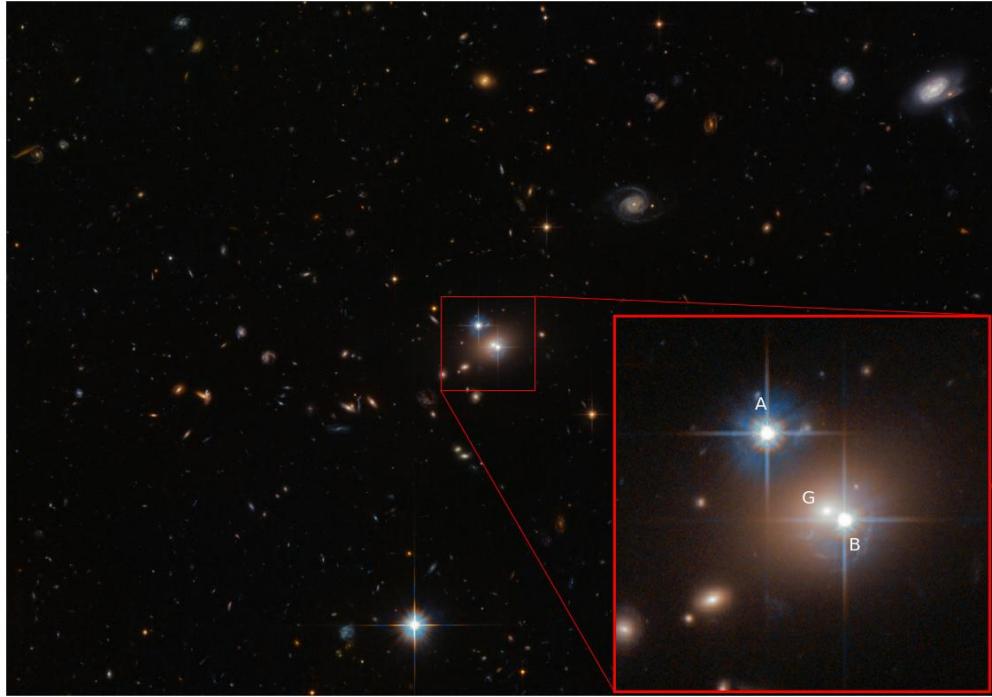
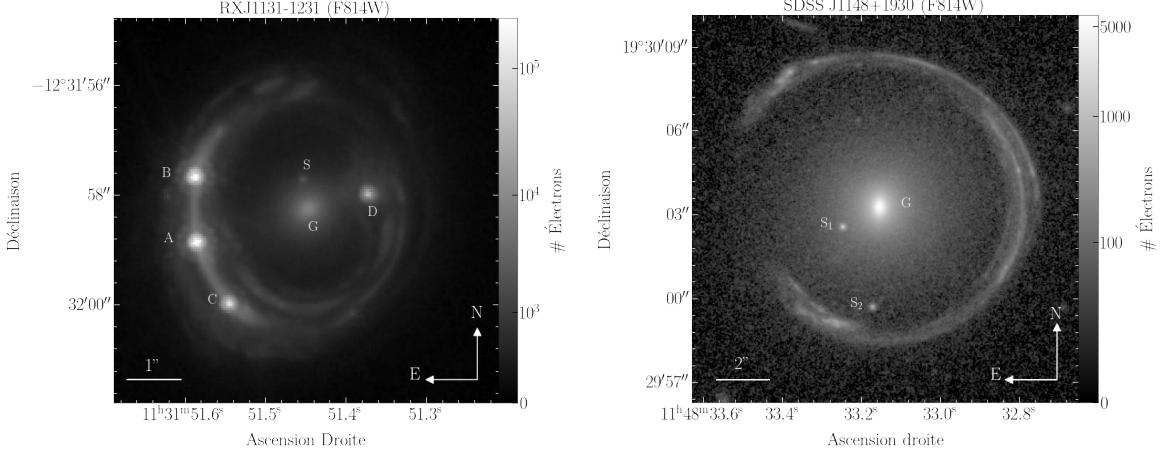


FIGURE 2.1 – Le quasar double (QSO 0957+561 A et B) et la galaxie-lentille (G) imagée par le télescope spatial Hubble. Crédit : ESA/Hubble et NASA, élargissement et annotation par AA.

Einstein (1936) considérait l'éventualité d'observer des lentilles gravitationnelles comme étant extrêmement improbable, pointant vers les limitations instrumentales de l'époque. En effet, les télescopes terrestres étaient largement limités par l'effet de *seeing* atmosphérique, soit la distorsion de l'image causée par la turbulence de l'atmosphère.

Dû à cette difficulté pratique, la première lentille gravitationnelle est découverte seulement plusieurs décennies après la prédiction de leur existence par Walsh et al. (1979), suivant l'identification de deux spectres radios de quasars identiques, QSO 0957+561 A et B (voir Figure 2.1), séparés par 5.7 secondes d'arcs et capturés avec le télescope radio Mark II à l'observatoire Jodrell Bank. Les spectres partagent la même magnitude, $m = 17$, le même décalage vers le rouge, $z = 1.405$, et possèdent des détails chimiques suspicieusement semblables. Ces coïncidences suggèrent fortement que ces deux spectres sont des copies d'un seul objet, soit un noyau actif d'une galaxie en arrière-plan, produite par l'effet de lentille gravitationnelle d'une galaxie en avant-plan, invisible dans le domaine radio à une fréquence de 966 MHz. Cette hypothèse est rapidement confirmée par l'observation optique de la galaxie-lentille ($z = 0.355$) avec l'observatoire Palomar (Young et al., 1980)³, ainsi que la modélisation de sa distribution de masse, de son environnement et des angles de déflexion qui causeraient l'apparition d'une image double du quasar (Young et al., 1981; Falco et al., 1991)

³Simultanément observé et confirmé par le télescope de 2.2 m de l'Université d'Hawaii au mont Mauna Kea (Stockton, 1980).



(a) Quasar quadruplement lentillé (A, B, C et D) par une galaxie (G). L'image de la galaxie hébergeant le quasar est déformée tangentielle-ment, formant un anneau d'Einstein. Image prise par HST avec le filtre F814W.

(b) Le fer à cheval cosmique, soit l'image d'une proto-galaxie à très haut décalage vers le rouge ($z = 2.379$) fortement magnifiée et déformée par une galaxie elliptique lumineuse en infrarouge (G) exceptionnellement massive ($5.2 \times 10^{12} h_{72}^{-1} M_\odot$, Schuldt et al., 2019). Image prise par HST avec le filtre F814W.

FIGURE 2.2 – Lentilles gravitationnelles de type galaxie-galaxie.

À la suite de cette découverte fortuite et la découverte subséquente des arcs lumineux (par exemple, voir la Figure 2.2b) dans les amas de galaxies (Lynds et Petrosian, 1989), l'étude des lentilles gravitationnelles est devenue un sujet d'étude particulièrement riche et prometteur en cosmologie (Blandford et Narayan, 1992; Bartelmann, 2010; Treu, 2010). Par exemple, les quasars lentillés comme RXJ1131-1231 (voir la Figure 2.2a) permettent de mesurer la constante de Hubble (1929), H_0 , un paramètre qui quantifie le taux de l'expansion de l'Univers au temps présent. L'idée principale derrière cette méthode est l'estimation du délai temporel entre chaque images du quasar par une surveillance décennale de la lentille gravitationnelle (e.g. Vanderriest et al., 1989; Wong et al., 2020). De façon alternative, la constante de Hubble peut être estimé par la caractérisation de la courbe de lumière des supernovas lentillées (Refsdal, 1964; Kelly et al., 2015; Goobar et al., 2017)

Les premiers relevés systématiques du ciel à la recherche de lentilles gravitationnelle (King et Browne, 1996; Muñoz et al., 1998; Myers et al., 2003), basés principalement sur l'étude des sources très brillante dans le domaine radio, ont permis de découvrir près de 80 lentilles gravitationnelles. Le programme *Sloan Lens ACS Survey* (SLACS, Bolton et al., 2005; Bolton et al., 2006), un relevé du ciel systématique basé sur l'analyse du spectre de galaxies de type ETG⁴ avec des lignes d'absorption à un décalage vers le rouge plus grand que les lignes d'émission, est un des programmes les plus réussis. Ce programme seul a mené à la découverte confirmée de plus de 150 lentilles gravitationnelles de type galaxie-galaxie (Bolton et al., 2008; Shu et al., 2017).

⁴Early-Type Galaxies

Les programmes basés sur la recherche visuelle d'images doubles, triples, d'arcs ou d'anneaux (e.g. [Faure et al., 2008](#)) dans les champs du ciel larges et profonds comme COSMOS ([Koekemoer et al., 2007; Scoville et al., 2007](#)), connaissent aujourd'hui une renaissance nourrie par les succès récents de l'apprentissage profond pour le traitement d'images ([Krizhevsky et al., 2012](#)). Cette nouvelle approche a déjà mené à la découverte de plus de 1000 lentilles gravitationnelles ([Petrillo et al., 2017; Huang et al., 2021](#)), et est projetée de découvrir plus de 10^5 systèmes grâce aux nouveaux relevés du ciels prévus dans la prochaine décennie aux observatoires Rubin ([LSST Science Collaboration et al., 2009](#)) et Euclid ([Refregier et al., 2010](#)).

Dans la section qui suit, je dérive les équations centrales qui nous permettent d'étudier les lentilles gravitationnelles de type galaxie-galaxie. Mon traitement est largement inspiré des manuels de références de [Meneghetti \(2013\)](#); [Congdon et Keeton \(2018\)](#).

2.1 Les angles de déflexion

Supposons qu'un photon est sur une trajectoire parallèle à l'axe de visée \mathbf{e}_{\parallel} d'un observateur sur Terre. Supposons de plus que la source d'un champ gravitationnel Φ est située sur l'axe de visée, ce qui a pour effet de courber la trajectoire de ce photon entre son point d'origine et son point d'arrivée. On définit l'angle de déviation comme la déviation totale de cette trajectoire dans la direction perpendiculaire à l'axe de visée de l'observateur. De façon générale, cette déviation s'écrit

$$\alpha = - \int_{\lambda_A}^{\lambda_B} \ddot{\mathbf{x}} \times \mathbf{e}_{\parallel} d\lambda, \quad (2.2)$$

où λ paramétrise la trajectoire du photon $\mathbf{x}(\lambda)$. Le signe négatif nous indique qu'on prend la perspective de l'observateur.

La trajectoire d'un photon obéit au principe de Fermat, qui stipule que la lumière suit une trajectoire qui extrémise la durée du parcours entre deux points. Dans le langage du calcul des variations, la variation de la durée s'écrit

$$\delta T = \delta \int_A^B n(\mathbf{x}(\ell)) \frac{d\ell}{c} = 0, \quad (2.3)$$

où ℓ est un élément de longueur sur la trajectoire et n est un indice de réfraction. Pour déterminer l'indice de réfraction du champ gravitationnel d'une galaxie, on doit utiliser le formalisme de la relativité générale. Selon le principe d'équivalence (fort), l'effet d'un champ gravitationnel est localement indistinguables d'une accélération causée par la courbure d'un espace-temps décrite par une métrique $g_{\mu\nu}$. La trajectoire d'un photon se trouve alors en cherchant les géodésiques de cet espace-temps. On fait l'approximation que le potentiel Φ d'une galaxie est celui d'un gaz parfait, c'est-à-dire qu'il satisfait une équation de Poisson

$$\nabla^2 \Phi = 4\pi G \rho. \quad (2.4)$$

Dans la limite où ce potentiel est faible $\frac{2\Phi}{c^2} \ll 1$, la métrique $g_{\mu\nu}$ est décrite par une expansion au premier ordre autour de la métrique de Minkowsky

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu \approx \left(1 + \frac{2\Phi}{c^2}\right) c^2 dt^2 - \left(1 - \frac{2\Phi}{c^2}\right) d\mathbf{x}^2. \quad (2.5)$$

Puisqu'un photon suit une géodésique de l'espace-temps $ds^2 = 0$, on peut déterminer l'indice de réfraction en réarrangeant l'équation (2.5)

$$n \equiv c \left(\frac{\|\mathbf{d}\mathbf{x}\|}{dt} \right)^{-1} \approx 1 - \frac{2\Phi}{c^2}. \quad (2.6)$$

En réécrivant l'élément de longueur $d\ell$ en termes du paramètre de la trajectoire $d\ell = \left\| \frac{d\mathbf{x}}{d\lambda} \right\| d\lambda$, on peut réécrire l'équation (2.3) sous la forme

$$\delta \int_{\lambda_A}^{\lambda_B} n(\mathbf{x}) \|\dot{\mathbf{x}}\| d\lambda = 0. \quad (2.7)$$

Par correspondance avec la fonctionnelle de l'action $J(x) = \int_{\lambda_0}^{\lambda_1} \mathcal{L}(\lambda, x, \dot{x}) d\lambda$, on trouve que le lagrangien de la trajectoire s'écrit $\mathcal{L} = n(\mathbf{x}) \sqrt{\dot{x}^2}$. La trajectoire qui satisfait (2.3) est une solution des équations d'Euler-Lagrange

$$\frac{d}{d\lambda} \frac{\partial \mathcal{L}}{\partial \dot{\mathbf{x}}} - \frac{\partial \mathcal{L}}{\partial \mathbf{x}} = 0. \quad (2.8)$$

On a donc

$$\frac{d}{d\lambda} n \frac{\dot{\mathbf{x}}}{\|\dot{\mathbf{x}}\|} - \|\dot{\mathbf{x}}\| \nabla n = 0, \quad (2.9)$$

Puisque le choix du paramètre λ est libre, on peut le choisir tel que $\|\dot{\mathbf{x}}\| = 1$ en tout point de la trajectoire. Ainsi,

$$\begin{aligned} \frac{d}{d\lambda} n \dot{\mathbf{x}} - \nabla n &= 0 \\ \implies n \ddot{\mathbf{x}} + (\nabla n \cdot \dot{\mathbf{x}}) \dot{\mathbf{x}} - \nabla n &= 0 \end{aligned} \quad (2.10)$$

À ce point de la dérivation, on utilise l'approximation de Born. C'est-à-dire qu'on approxime la trajectoire du photon comme une ligne droite sur l'axe de visée \mathbf{e}_{\parallel} . Cette approximation est justifiée dans le contexte des lentilles gravitationnelles de type galaxie-galaxie, puisque les angles de déviation sont généralement de l'ordre de l'arcseconde ou plus petits. Comme le vecteur $\dot{\mathbf{x}}$ est tangent à la trajectoire du photon, les termes $\propto \dot{\mathbf{x}} \times \mathbf{e}_{\parallel}$ s'annulent. En substituant l'indice de réfraction par (2.6) dans $\mathbf{e}_{\parallel} \times (2.10)$, on obtient

$$\ddot{\mathbf{x}} \times \mathbf{e}_{\parallel} = \frac{1}{n} \nabla_{\perp} n = \nabla_{\perp} \log n \approx -\frac{2}{c^2} \nabla_{\perp} \Phi, \quad (2.11)$$

où ∇_{\perp} est un gradient selon les coordonnées perpendiculaires à \mathbf{e}_{\parallel} . On note que le facteur 2 qui

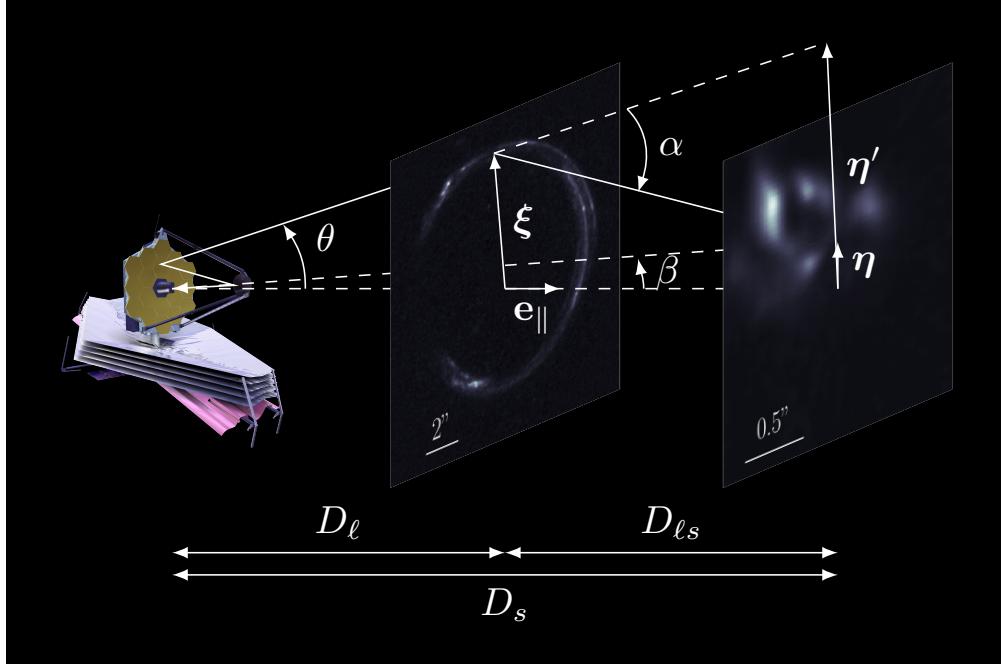


FIGURE 2.3 – Schéma d'une lentille gravitationnelle.

apparaît dans l'équation (2.11) est un effet qui vient de la relativité générale. Ce facteur corrige la solution que l'on aurait obtenue avec une dérivation classique (newtonienne).

On est maintenant en mesure de calculer l'angle de déviation. J'introduis le paramètre d'impact ξ qui est la distance perpendiculaire entre la position d'origine du photon sur le plan de la lentille et l'axe de visée (voir Figure 2.3). Dans le cas où le potentiel est généré par une masse M ponctuelle, c.-à-d. qu'on suppose $\rho = M\delta^3(\mathbf{x})$, où δ est la fonction delta de Dirac, alors le potentiel qui satisfait l'équation de Poisson (2.4) est la fonction de Green $\Phi = -\frac{GM}{\sqrt{\xi^2 + z^2}}$, où z est la coordonnée sur l'axe de visée. L'équation (2.2) se réécrit finalement comme

$$\begin{aligned} \alpha(\xi) &= -\frac{2GM}{c^2} \int_{-\infty}^{\infty} \frac{\partial}{\partial \xi} \frac{1}{(\xi^2 + z^2)^{1/2}} dz \\ \implies \alpha(\xi) &= \frac{4GM}{c^2 \xi^2} \xi \end{aligned} \quad (2.12)$$

Cette solution se généralise naturellement à un profil de masse quelconque en assumant qu'il s'exprime comme une somme d'éléments de masse $dm = \Sigma d^2\xi'$, où $\Sigma = \int \rho dz$ est une densité surfacique de masse. L'angle de déviation total mesuré à un point ξ est alors une convolution sur tout le plan de la lentille (mince) puisque l'équation (2.12) dépend linéairement de la masse M :

$$\alpha(\xi) = \frac{4G}{c^2} \int_{\mathbb{R}^2} \Sigma(\xi') \frac{\xi - \xi'}{\|\xi - \xi'\|^2} d^2\xi' \quad (2.13)$$

L'angle de déviation est une quantité cruciale pour résoudre une lentille gravitationnelle puisqu'il

décrit une transformation des coordonnées angulaires du plan de la lentille ($\boldsymbol{\theta}$) vers les coordonnées angulaires du plan de la source ($\boldsymbol{\beta}$). On assume que les distances entre l'observateur et la lentille D_ℓ , entre l'observateur et la source D_s et entre la lentille et la source $D_{\ell s}$, sont beaucoup plus grandes que les distances perpendiculaires à l'axe de visée ξ ou η (voir figure 2.3). Cette approximation est justifiée pour les objets qui nous intéressent, pour lesquels les distances parallèles à l'axe de visée sont généralement de l'ordre du Gpc, alors que les distances perpendiculaires sont généralement de l'ordre du kpc ; soit 6 ordres de grandeur de différence. Ainsi, on peut faire un argument géométrique (euclidien)

$$\begin{aligned} D_s \boldsymbol{\theta} &= \boldsymbol{\eta}' \\ D_s \boldsymbol{\beta} &= \boldsymbol{\eta} \\ D_{\ell s} \boldsymbol{\alpha} &= \boldsymbol{\eta}' - \boldsymbol{\eta} \\ \implies D_s \boldsymbol{\beta} &= D_s \boldsymbol{\theta} - D_{\ell s} \boldsymbol{\alpha} \end{aligned} \tag{2.14}$$

La dernière relation est l'équation maîtresse qui nous permet de tracer les rayons lumineux d'une source vers un détecteur fictif dans nos simulations.

On notera que cette relation reste valide pour un Univers courbe et/ou en expansion (c.-à-d. décrit par une géométrie non euclidienne), à condition qu'on utilise une notion de distance qui satisfait, par définition, la relation trigonométrique euclidienne

$$D \equiv \frac{\xi}{\theta}, \tag{2.15}$$

où ξ est la taille physique d'un objet placé à une certaine distance de l'observateur, et θ est l'angle solide sous-tendu par cet objet. Pour un Univers décrit par la métrique de Friedmann-Lemaître-Robertson-Walker, la notion de distance qui respecte la définition (2.15) est la distance du diamètre angulaire. En pratique, on peut exprimer D en termes du décalage vers le rouge des photons émis par l'objet, z . On note $a(z)$ le facteur d'échelle lorsque le photon est émis par la source et $a(0)$ le facteur d'échelle au moment présent ($z = 0$). Pour un Univers plat (voir les manuels de référence [Coles et Lucchin, 2002](#); [Dodelson et Schmidt, 2003](#); [Bartelmann, 2004](#))

$$D_z = \underbrace{ca(z) \int_{a(z)}^{a(0)} \frac{da}{a\dot{a}}}_{\text{distance comobile}}; \tag{2.16}$$

$$\begin{aligned} &= \frac{ca(z)}{H_0} \int_{a(z)}^{a(0)} \frac{da}{\sqrt{\Omega_{r,0} + \Omega_{m,0}a + \Omega_{\Lambda,0}a^4}}; \\ &= \frac{c}{H_0(1+z)} \int_0^z \frac{dz'}{\sqrt{\Omega_{r,0}(1+z')^4 + \Omega_{m,0}(1+z')^3 + \Omega_{\Lambda,0}}}. \end{aligned} \tag{2.17}$$

On a utilisé la relation entre le facteur d'échelle, a , et le décalage vers le rouge, $a = (1+z)^{-1}$, pour

obtenir l'équation (2.17) par un changement de la variable d'intégration. $\Omega_{r,0}$, $\Omega_{m,0}$ et $\Omega_{\Lambda,0}$ sont les paramètres de densités, au temps présent, de la radiation, de la matière et de l'énergie sombre respectivement. H_0 est la constante de Hubble, soit le taux d'expansion de l'Univers au temps présent. La distance $D_{\ell s}$ se trouve simplement en ajustant les bornes de l'intégrale $\int_0^z \mapsto \int_{z_\ell}^{z_s}$. La valeur des paramètres du modèle cosmologique Λ CDM obtenue par l'équipe [Planck Collaboration \(2020\)](#) est rapportée dans l'annexe A.

Il est généralement pratique de travailler avec la forme adimensionnelle de l'équation (2.14). On introduit la densité critique

$$\Sigma_c = \frac{c^2}{4\pi G} \frac{D_s}{D_{\ell s} D_\ell}, \quad (2.18)$$

qui nous permet de définir la quantité qu'on nomme convergence $\kappa(\boldsymbol{\theta}) \equiv \frac{\Sigma(\boldsymbol{\theta})}{\Sigma_c}$. On définit ainsi l'angle réduit

$$\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}) = \frac{1}{\pi} \int_{\mathbb{R}^2} \kappa(\boldsymbol{\theta}) \frac{\boldsymbol{\theta} - \boldsymbol{\theta}'}{\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|} d^2\boldsymbol{\theta}', \quad (2.19)$$

qui satisfait l'équation de la lentille adimensionnelle

$$\boldsymbol{\beta} = \boldsymbol{\theta} - \hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}). \quad (2.20)$$

Les équations (2.19) et (2.20) sont les équations centrales à la modélisation des lentilles gravitationnelles. Elles sont utilisées pour simuler des lentilles gravitationnelles au chapitre 5. Elles sont aussi utilisées pour résoudre le problème inverse qui consiste à inférer l'image non-distordue de la galaxie en arrière-plan, $I(\boldsymbol{\beta})$, et les paramètres de la distribution de masse de la galaxie-lentille, $\kappa(\boldsymbol{\theta})$, à partir des distorsions observées de l'image en arrière-plan $I(\boldsymbol{\theta})$.

Chapitre 3

Introduction à l'apprentissage profond

L'astronomie, l'astrophysique et la cosmologie sont officiellement entrées dans l'ère du *Big Data*, soit une ère dominée par le volume de données, maintenant mesuré dans l'ordre du *petabyte*. Un *petabyte* représente un million de *gigabytes*, ou encore $\sim 10^{15}$ *bytes*. En exemple particulier est l'observatoire Vera C. Rubin. Cet observatoire produira environ 1 *petabytes* de données chaque année, soit une quantité de données impossible à traiter dans son ensemble pour la plupart des algorithmes ; sans parler de la quantité accumulée par l'observatoire au travers des 10 années planifiées pour le relevé astronomique qui doit débuter en 2024 ([LSST Science Collaboration et al., 2009; Blum et al., 2022](#)).

C'est cette réalité qui nous force à développer des méthodes d'analyses plus sophistiquées pour l'inférence de quantités physiques à partir d'images, spectres et vidéos du ciel. C'est aussi ce qui motive l'étude des méthodes liées à l'apprentissage machine, et plus particulièrement l'apprentissage profond, qui promettent de simplifier énormément l'analyse statistique des grands ensembles de données à venir. Ce chapitre se veut une courte introduction aux concepts de base. La section 3.1 est un survol de la théorie des statistiques bayésiennes. Cette section introduit certains concepts cruciaux qui sous-tendent la théorie de l'apprentissage machine. Ensuite, la section 3.2 décrit en détail un exemple de régression, puis la section 3.3 discute du problème de la sélection du modèle. Les réseaux neuronaux et l'apprentissage profond sont introduits à la section 3.4 comme une généralisation possible de l'apprentissage machine classique. Finalement, on introduit les réseaux de neurones convolutifs à la section 3.5, cruciaux pour le traitement d'images.

Pour une discussion plus détaillée des concepts abordés dans ce chapitre, voir les manuels de référence [Goodfellow et al. \(2016\)](#) et [Bishop \(2007\)](#).

3.1 Survol des statistiques bayésiennes

L'inférence bayésienne est une théorie statistique qui a pour but principal de modéliser l'état de connaissance, ou le degré de croyance, associé à un évènement. De façon générale, l'inférence bayésienne est un processus de mise à jour de nos connaissances *a priori*, c.-à-d. les connaissances

acquises avant l'observation d'un évènement. On définit la vraisemblance comme la loi de probabilité qui gouverne la probabilité d'observer un évènement particulier. Étant donné un processus physique auquel on associe une loi de vraisemblance, le processus d'inférence bayésienne correspond simplement à la repondération de nos connaissances par la vraisemblance. En d'autres mots, cette procédure correspond à multiplier la loi de probabilité *a priori* par la vraisemblance d'un évènement (et renormaliser).

Par exemple, considérons un tirage au sort. On modélise la probabilité, $x \in [0, 1]$, d'observer pile ou face, $y \in \{0, 1\}$, par une loi de Bernoulli

$$p(y | x) = x^y(1 - x)^{1-y}. \quad (3.1)$$

A priori, on associe une probabilité égale (uniforme) au paramètre x , soit $p(x) = \mathcal{U}(0, 1)$. Ce choix reflète une ignorance complète du processus physique qui contrôle le tirage au sort. Supposons maintenant qu'on observe y_1 , un évènement généré de la loi physique $p(y_1 | x^*)$, où x^* est le véritable paramètre de la loi physique. Le théorème de Bayes nous indique comment ajuster notre degré de croyance, maintenant $p(x | y_1)$, soit la loi *a posteriori* du paramètre x étant donné avoir observé y_1

$$p(x | y_1) = \frac{p(y_1 | x)p(x)}{p(y_1)}. \quad (3.2)$$

La loi de probabilité $p(y)$, aussi appelée l'évidence, normalise la loi *a posteriori*, qui est proportionnelle au produit de la vraisemblance $p(y | x)$ et de la loi *a priori* $p(x)$. En pratique, on peut évaluer $p(y)$, une loi marginale, en l'exprimant en fonction de la loi jointe $p(x, y)$, soit

$$p(y) = \int_{\mathcal{X}} dx p(y, x). \quad (3.3)$$

On peut ensuite utiliser la définition d'une loi conditionnelle

$$p(y | x) = \frac{p(x, y)}{p(x)}, \quad (3.4)$$

pour exprimer l'évidence, $p(y)$, en fonction de la vraisemblance et de la loi *a priori*

$$p(y) = \int_{\mathcal{X}} dx p(y | x)p(x). \quad (3.5)$$

Supposons maintenant qu'on observe un certain nombre de tirages au sorts supplémentaires

$$y_2, y_3 \dots, y_N \stackrel{\text{iid}}{\sim} p(y | x^*),$$

indépendamment et identiquement distribués (iid) selon la même loi physique, $p(y | x^*)$. Pour mettre à jour nos connaissances, on procède itérativement. En premier lieu, on trouve la distribution *a posteriori* $p(x | y_1)$ avec le théorème de Bayes (3.2). Ensuite, on remplace la distribution *a priori*,

$p(x)$, par $p(x | y_1)$ et on remplace la loi de vraisemblance par $p(y_2 | x)$ dans le théorème de Bayes (3.2) pour obtenir la loi *a posteriori*, $p(x | y_1, y_2)$. Et ainsi de suite. Puisque les observations sont iid, alors ce processus est équivalent, par induction, à

$$p(x | y_{1:N}) = \frac{\prod_{i=1}^N p(y_i | x)p(x)}{\int_{\mathcal{X}} \prod_{i=1}^N p(y_i | x)p(x)dx} \quad (3.6)$$

Ce dernier résultat est particulièrement important pour l'inférence statistique et la dérivation des fonctions objectives pour l'apprentissage machine. La figure 3.1 montre comment nos connaissances sur le paramètre x évoluent en termes du nombre d'observations effectuées lorsqu'on applique l'équation (3.6). La figure montre que la loi *a posteriori* correspond à la loi *a priori* lorsque $N = 0$ et converge vers une loi normale centrée autour de $x^* = \frac{1}{2}$ lorsque $N \rightarrow \infty$, soit un exemple concret du théorème de la limite centrale.

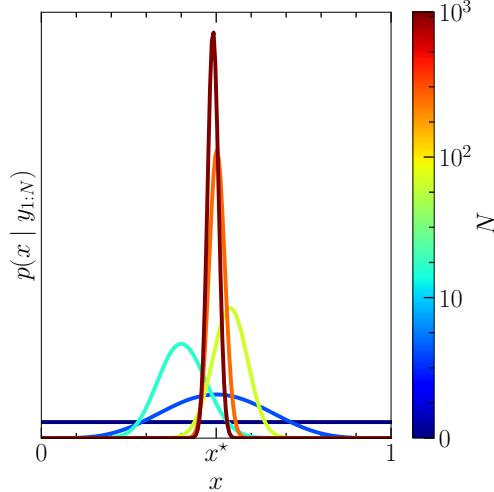


FIGURE 3.1 – Exemple d'une inférence bayésienne pour le tirage au sort.

Le lecteur aura compris qu'on va maintenant construire l'apprentissage machine sur des fondations bayésiennes. Ce n'est pas la seule façon de le faire. En particulier, il est possible de poser des fondations fréquentistes en supposant, comme fait dans cette section, que toute l'information concernant un problème est contenue dans la loi de vraisemblance. Autrement dit, on suppose une ignorance complète *a priori* sur les paramètres d'intérêts. En ce sens, et strictement dans le contexte qui nous intéresse, il est possible de passer d'un point de vue à l'autre, puisque les deux théories seront en accord sur la réponse à condition qu'on utilise une loi uniforme pour représenter nos connaissances *a priori* sur les paramètres d'intérêts. À cause de cette correspondance approximative entre les deux théories, les lois *a priori* vont parfois être abandonnées sans justification dans le texte qui suit. Le lecteur comprendra qu'on aura simplement changé de point de vue.

3.2 Un exemple d'apprentissage machine : la régression

L'apprentissage machine est un concept assez général qui décrit le processus d'extraire l'information contenue dans un ensemble de données $\mathcal{D} = \{\mathbf{y}^{(i)}\}_{i=1}^N$, soit un ensemble d'observations provenant d'une source ou d'un processus physique quelconque. Le terme *information* est utilisé de façon très vague ici. Ce terme est formellement défini dans la théorie de l'information de [Shannon \(1948\)](#) comme étant le nombre minimal de *bits* pour décrire une observation y_i , ou encore l'ensemble de données \mathcal{D} . Cette quantité est intimement liée avec l'entropie. Une discussion plus détaillée est repoussée au chapitre 4.

Il y a quatre ingrédients essentiels à l'apprentissage machine, soit

1. Un ensemble de données \mathcal{D}
2. Un ensemble d'hypothèses \mathcal{H}
3. Une fonction objective \mathcal{L}
4. Un algorithme d'optimisation \mathcal{G}

Dans cette section, je décris chacun de ces ingrédients dans le contexte d'une tâche de régression. La régression est une tâche d'apprentissage machine qui consiste à entraîner un modèle, f_θ , sur un ensemble de données augmenté pour la régression, soit $\mathcal{D} = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^N$. Chaque exemple dans l'ensemble de données est constitué d'un vecteur de paramètres physiques, \mathbf{x} , et d'une observation, \mathbf{y} . On suppose toujours que les exemples de l'ensemble de données sont générés de façon identique et indépendante par la combinaison d'une loi *a priori* sur les paramètres physiques et une loi de vraisemblance pour les observations

$$\mathbf{x} \sim p(\mathbf{x}), \quad \mathbf{y} \sim p_\theta(\mathbf{y} \mid \mathbf{x}). \quad (3.7)$$

Dans le texte, on utilisera le terme *modèle physique* pour faire référence à $p_\theta(\mathbf{y} \mid \mathbf{x})$, puisque cette loi de probabilité relie les paramètres physiques avec les observations. Elle encodera donc tous les processus physiques en jeu pour un problème d'inférence donné. On notera de plus que l'ensemble des points de données n'a pas besoin d'être un nombre fini. Dans le cas où $N \rightarrow \infty$, alors \mathcal{D} est explicitement décrit par la loi *a priori*, $p(\mathbf{x})$, et le modèle physique $p_\theta(\mathbf{y} \mid \mathbf{x})$. Dans ce cas, on dira que la loi génératrice $(\mathbf{x}, \mathbf{y}) \sim p(\mathbf{x}, \mathbf{y})$ est explicite. Dans le cas où \mathcal{D} est un nuage de points ($N < \infty$), alors on dira que le processus génératif est implicite.

Pour se fixer les idées, on considère l'exemple

$$p(x) = \mathcal{U}(a, b), \quad p_\theta(y \mid x) = \mathcal{N}(y \mid f_\theta(x), \sigma^2). \quad (3.8)$$

On a utilisé le symbole $\mathcal{N}(y \mid \mu, \sigma^2)$ pour décrire une loi normale sur y avec comme moyenne μ et

variance σ^2 . On doit supposer que les données sont générées à partir d'une solution quelconque

$$f_{\theta^*} = 2x^5 - x, \quad (3.9)$$

avec une amplitude de bruit $\sigma = 1$. Ce problème est illustré à la figure 3.2.

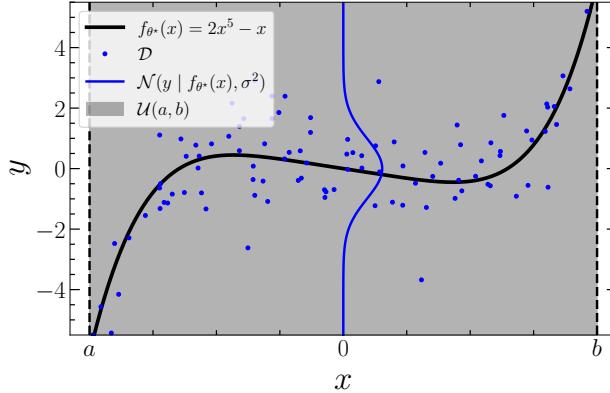


FIGURE 3.2 – Exemple d'un problème de régression.

Le second ingrédient au problème d'apprentissage machine est la famille des modèles, ou l'ensemble des hypothèses, \mathcal{H} . Il serait tentant de choisir le modèle $f_\theta = \theta_1 x^5 + \theta_2 x$, soit un modèle avec la même forme que la solution f_{θ^*} . Toutefois, on ne peut pas supposer que la forme du modèle sera connue *a priori*, ou qu'elle peut facilement être devinée. La sélection du modèle sera abordée en détail dans la section 3.3, puisque ce sujet mérite une discussion à part entière. Pour l'instant, on va simplement faire le choix le plus simple en assumant que la forme de la loi utilisée pour générer les points de données est inconnue. Le choix le plus simple est de construire une famille de modèles linéaires en termes de x et θ

$$\mathcal{H} = \{f_\theta : \mathbb{R} \rightarrow \mathbb{R} \mid f_\theta = \theta_0 + \theta_1 x\}. \quad (3.10)$$

Le troisième ingrédient à l'apprentissage machine est de construire une fonction objective pour *entraîner* notre modèle f_θ sur les points de données observés. L'objectif de l'entraînement est d'estimer le modèle f_θ , ou de façon équivalente les paramètres θ , qui maximise la loi *a posteriori*, $p(\theta|\mathcal{D})$. En appliquant le théorème de Bayes, ceci correspond à

$$\hat{\theta} = \arg \max_{f_\theta \in \mathcal{H}_\theta} \frac{p(\mathcal{D} \mid \theta)p(\theta)}{p(\mathcal{D})} \quad (3.11)$$

Pour faire correspondre cet objectif avec la littérature, on applique le logarithme au côté droit de l'équation, ce qui ne change pas la position des extrêmes de l'objectif, le logarithme étant une fonction monotone. Il est aussi convention de tourner le problème à l'envers, et de chercher plutôt

les minimas de la fonction négative. Dans ce cas, la fonction objective est

$$\hat{\theta} = \arg \min_{f_\theta \in \mathcal{H}} -\log p(\mathcal{D} | \theta) - \log p(\theta) + C(\mathcal{D}). \quad (3.12)$$

où $C(\mathcal{D})$ est une constante qui ne dépend que de l'ensemble de données. L'étape finale est d'écrire la vraisemblance $p(\mathcal{D} | \theta)$ en termes de $p_\theta(y | x)$. Dans le texte, les paramètres d'entraînements sont toujours placés en indices, par quoi on entend que les paramètres θ conditionnent la loi de probabilité, $p_\theta(x | y) = p(x | y, \theta)$. On remarque que

$$p(\mathcal{D} | \theta) = \prod_{i=1}^N p(x_i, y_i | \theta) = \prod_{i=1}^N p(y_i | x_i, \theta) p(x_i | \theta), \quad (3.13)$$

par définition de la loi conditionnelle et de la supposition que les données sont générées de façon iid. Le dernier terme peut être simplifié en termes de la loi *a priori*, $p(x | \theta) = p(x)$ puisqu'on suppose que θ et x sont deux variables aléatoires indépendantes. Un choix de modèle θ ne devrait pas changer la distribution *a priori* sur les paramètres physiques. Si c'est le cas, alors le choix du modèle doit être révisé. On trouve finalement l'objectif d'entraînement

$$\hat{\theta} = \arg \min_{f_\theta \in \mathcal{H}} -\log \prod_{i=1}^N p_\theta(y_i | x_i) - \log p(\theta) + C(\mathcal{D}), \quad (3.14)$$

où $-\log \prod_{i=1}^N p(x_i)$ est absorbé dans $C(\mathcal{D})$. On est maintenant en mesure de dériver la forme exacte de la fonction objective pour le problème posé dans cette section. Avec la supposition que le modèle physique est une loi normale, on a que

$$\log \prod_{i=1}^N p_\theta(y_i | x_i) = -\frac{1}{2} \sum_{i=1}^N \frac{(y_i - f_\theta(x_i))^2}{\sigma^2} - \frac{N}{2} \log(2\pi\sigma^2), \quad (3.15)$$

ce qui correspond à la fonction objective

$$\hat{\theta} = \arg \min_{f_\theta \in \mathcal{H}} \underbrace{\frac{1}{N} \sum_{i=1}^N (y_i - f_\theta(x_i))^2}_{\hat{\mathcal{L}}_\theta(\mathcal{D})} - \frac{2\sigma^2}{N} \log p(\theta). \quad (3.16)$$

Par convention, on a utilisé le fait que les constantes qui ne dépendent pas de θ peuvent être ignorées (elles ne changent pas les extrêmes du problème) et on a multiplié la fonction objective par $2\sigma^2/N$. L'expression obtenue fait intervenir la fonction objective approximative $\hat{\mathcal{L}}_\theta(\mathcal{D})$, soit un estimé Monte Carlo de l'espérance de l'erreur quadratique. En prenant la limite $N \rightarrow \infty$, l'estimé Monte Carlo de l'espérance devient exacte

$$\lim_{N \rightarrow \infty} \hat{\mathcal{L}}(\mathcal{D}) = \mathcal{L}_\theta(\mathcal{D}) = \mathbb{E}_{(x,y) \sim \mathcal{D}} [(y - f_\theta(x))^2]. \quad (3.17)$$

Dans la littérature, $\hat{\mathcal{L}}_\theta(\mathcal{D})$ est nommée l'erreur quadratique moyenne. Dans le reste de ce mémoire, on va laisser tomber le chapeau sur la fonction objective pour simplifier la notation. Le lecteur comprendra qu'on travaille généralement avec l'estimé Monte Carlo si l'espérance mathématique n'est pas accessible par calcul direct.

Le dernier ingrédient à l'apprentissage machine est le choix d'un algorithme d'optimisation \mathcal{G} pour résoudre l'équation (3.16), c.-à-d. déterminer $\hat{\theta}$. Les algorithmes d'optimisations récents pour l'apprentissage profond sont souvent basés sur la descente de gradient stochastique, qu'on peut décrire succinctement par la relation de récurrence

$$\hat{\theta}^{(t+1)} = \hat{\theta}^{(t)} - \gamma_t \nabla_{\hat{\theta}^{(t)}} \mathcal{L}_{\hat{\theta}^{(t)}}(\mathcal{D}), \quad (3.18)$$

γ_t est le taux d'apprentissage, généralement choisi comme étant aussi grand que possible, sans pour autant rendre instable l'algorithme d'optimisation \mathcal{G} . Une discussion plus détaillée est reportée au chapitre 4 sur ce sujet.

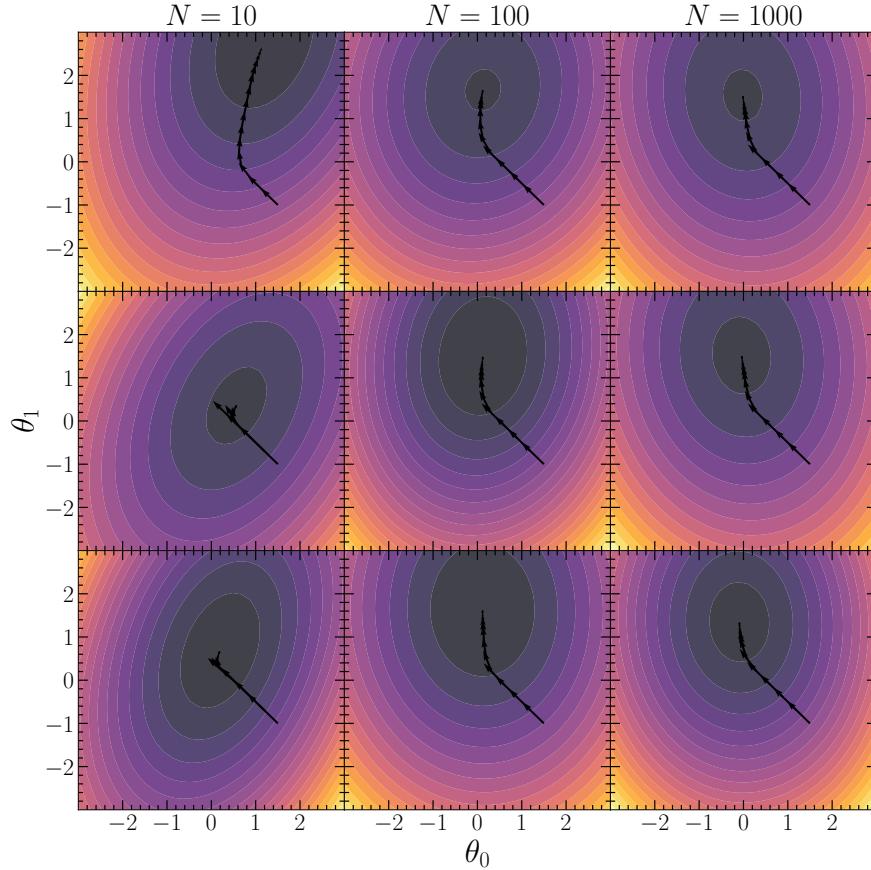


FIGURE 3.3 – Contours de $\mathcal{L}_\theta(\mathcal{D})$ pour différent tirage de \mathcal{D} (rangsées) et différentes tailles $N = |\mathcal{D}|$ (colonnes) en utilisant le modèle linéaire de l'équation (3.10) et la loi générative (3.8). Une trajectoire produite par la descente de gradient (3.18) est illustré avec les flèches noires.

Lorsqu'on travaille à basses dimensions, il est possible d'obtenir de très bons estimés pour $\mathcal{L}_\theta(\mathcal{D})$.

Par exemple, la figure 3.3 illustre comment la variance des contours de l'estimé Monte Carlo $\mathcal{L}_\theta(\mathcal{D})$ diminue en fonction de N . Les rangées montrent un estimé Monte Carlo pour différents tirages de l'ensemble de données \mathcal{D} à partir de la loi générative décrite à l'équation (3.8), alors que les colonnes montrent des ensembles de données de taille croissante. En particulier, la troisième colonne possède des contours très stables, ce qui implique que la descente de gradient va retrouver la même solution $\hat{\theta}$ systématiquement, contrairement aux estimés où N est relativement petit. Dans ce cas, la solution obtenue par \mathcal{G} dépend fortement de l'ensemble de données observé ; la première colonne est l'exemple le plus frappant de ce phénomène.

Les quelques thèmes abordés ici seront suffisants pour donner une compréhension à haut niveau des méthodes développées dans les prochains chapitres. Bien sûr, cette introduction n'est pas exhaustive. Un lecteur intéressé devrait consulter les références mentionnées au début du chapitre. Les sections qui suivent serviront à introduire et motiver l'apprentissage profond.

3.3 Sélection du modèle

Pour motiver l'apprentissage profond, on s'intéresse maintenant au problème de la sélection du modèle. La sélection d'un modèle linéaire à la section précédente n'était motivée que par un critère de simplicité. En général, ce critère est insuffisant pour extraire toute l'information qu'un ensemble de données contient, et ce, peu importe les transformations appliquées aux espaces \mathcal{X} et \mathcal{Y} . Par exemple, un polynôme avec un seul terme (monôme) comme $f_{\theta^*}(x) = Cx^\alpha$ peut naturellement être modélisé par un modèle linéaire après la transformation appropriée de l'espace des paramètres physiques $\mathcal{X} \rightarrow \log \mathcal{X}$, de sorte que le monôme dans le nouvel espace, $f'_{\theta^*} = \alpha \log x + \log C$, est une fonction de forme linéaire en termes de α , $\log C$ et x . En général, le prétraitement des données est un aspect important de l'apprentissage machine puisqu'il permet d'éplucher les couches de complexités artificielles autour d'un problème d'apprentissage machine donné.

Or, la fonction introduite à la section précédente est déjà un exemple qui ne peut pas être recouvert par un modèle linéaire puisque $f_{\theta^*} = 2x^5 - x$ est un polynôme composé de deux monômes. Minimalement, deux modèles linéaires seraient donc nécessaires pour recouvrir f_{θ^*} . Au mieux, un modèle linéaire est une bonne approximation de f_{θ^*} pour des régions spéciales du domaine de la fonction comme $|x| \ll 1$ ou $|x| \gg 1$. On doit donc considérer des modèles plus complexes que le simple modèle linéaire considéré jusqu'à maintenant.

Pour ce faire, on considère trois directions principales. La première méthode, déjà mentionnée dans la section précédente, est de construire une fonction avec la bonne forme *a priori* via l'intuition. On ne considéra pas plus longtemps cette approche, puisqu'elle est impraticable dans les cas les plus généraux comme les problèmes à haute dimensions où l'intuition humaine échoue complètement. La seconde approche, et probablement l'approche la plus intuitive considérant la façon dont le sujet a été introduit jusqu'à maintenant, est de considérer une série de puissances entières positives. Puisque c'est une approche importante en apprentissage machine, il vaut la peine de dire quelques mots sur

celle-ci avant de poursuivre vers la troisième approche considérée dans cette introduction.

3.3.1 Compromis entre le biais et la variance

Lorsque $x \in \mathbb{R}$, la série entière prend la forme très simple

$$f_\theta(x) = \sum_{p=0}^P \theta_p x^p. \quad (3.19)$$

Le modèle (3.19) est une fonction linéaire en termes des paramètres θ et non-linéaire en termes des paramètres physiques x ($P > 1$). Dans la limite où $P \rightarrow \infty$, ce modèle peut représenter l'ensemble des fonctions analytiques incluant les polynômes, e^x , les fonctions trigonométriques, hyperboliques, etc.

Avec l'hypothèse (3.19), on peut explorer comment la complexité du modèle influence le problème d'apprentissage. Strictement dans le contexte de l'ajustement d'un polynôme $f_\theta : \mathbb{R} \rightarrow \mathbb{R}$ de degré $P < \infty$, la complexité du modèle peut être quantifiée par P , soit le nombre de termes dans la série entière. En répétant l'exercice de la section précédente plusieurs fois pour des modèles d'une complexité grandissante, on peut obtenir des statistiques sur l'erreur de généralisation en fonction de P , ce qui est illustré à la figure 3.4.

L'erreur de généralisation est définie comme l'erreur quadratique moyenne d'une fonction calculée sur un ensemble test, distinct de l'ensemble d'entraînement utilisé pour ajuster la fonction. Cette métrique permet d'estimer le risque encouru lorsqu'on tente d'utiliser une fonction au-delà de son ensemble d'entraînement. On observe que notre estimé de l'erreur de généralisation atteint un minimum à la complexité $P = 5$, ce qui correspond au degré du polynôme objectif $f_{\theta^*} = 2x^5 - x$.

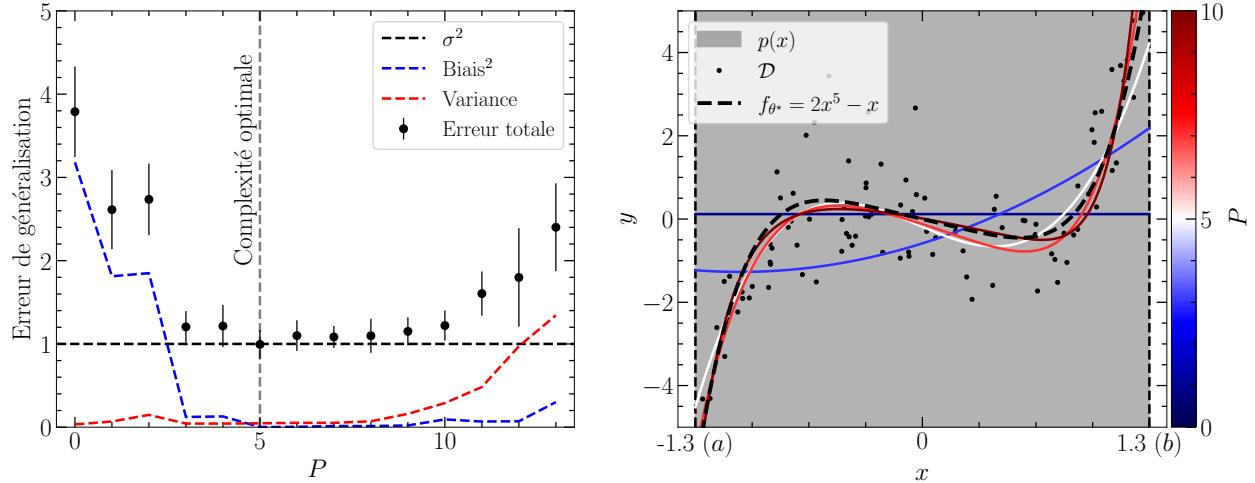


FIGURE 3.4 – Compromis classique entre le biais et la variance d'un algorithme d'apprentissage machine pour l'ajustement d'un polynôme de degré P sur les données générées de la loi $f_{\theta^*} = 2x^5 - x$.

Le comportement de cette erreur peut être compris intuitivement par une décomposition de

l'erreur totale en termes de l'erreur irréductible du problème σ^2 , du biais

$$\text{Biais}(f_\theta(x)) = \mathbb{E}_{\mathcal{D}} [f_\theta(x)] - f_{\theta^*}(x), \quad (3.20)$$

et de la variance d'un algorithme d'apprentissage machine

$$\text{Variance}(f_\theta(x)) = \mathbb{E}_{\mathcal{D}} \left[(f_\theta(x) - \mathbb{E}_{\mathcal{D}} [f_\theta(x)])^2 \right]. \quad (3.21)$$

Lorsque le modèle n'est pas suffisamment complexe pour capturer l'ensemble des points de données, la solution est biaisée par le choix du modèle (courbes bleues de la figure 3.4). Le biais est la principale cause du phénomène appelé le *sous-ajustement*, soit lorsqu'un algorithme d'apprentissage n'est pas suffisamment flexible pour modéliser une certaine distribution \mathcal{D} . Lorsque le modèle est trop complexe, le modèle est aussi trop flexible et on observe un phénomène de *sur-ajustement* (courbes rouges de la figure 3.4). C.-à-d. que les modèles entraînés se spécialisent à leur ensemble d'entraînement. L'erreur moyenne de généralisation augmente puisque la solution trouvée par l'algorithme d'apprentissage dépend fortement de l'ensemble d'entraînement \mathcal{D} , qui ne sera pas nécessairement représentatif de la loi génératrice implicite $f_{\theta^*}(x)$. La variance domine l'erreur de généralisation dans ce régime.

Dans tous les cas, l'erreur minimale correspond à σ^2 , soit le niveau d'erreur irréductible à un problème donné. Ce niveau d'erreur ne peut être atteint que lorsque la complexité du modèle est environ égale à la complexité de la loi générative. En général, on s'attend donc à ce qu'un niveau de complexité optimal existe pour un problème donné. La décomposition de l'erreur en termes du biais et de la variance est une méthode pour découvrir cette complexité. Cette procédure fonctionne certainement pour le problème de régression décrit dans cette section, mais elle n'est pas garantie de fonctionner en général. En effet, le biais et la variance d'une fonction sont des quantités presque impossibles à calculer en général, puisqu'on doit calculer l'espérance d'une fonction apprise par \mathcal{G} étant donné différentes réalisations de \mathcal{D} .

Ce qui nous concerne maintenant est une discussion sur la sélection du modèle lorsqu'on doit apprendre une fonction sur des données multidimensionnelles, soit le cas plus général d'apprentissage machine.

3.3.2 La séparabilité linéaire

C'est lorsqu'on cherche à construire une série entière pour des données multidimensionnelles qu'on réalise que la tâche est exponentiellement plus difficile que le cas unidimensionnel. Considérons le cas le plus simple, soit une fonction $f_\theta : \mathbb{R}^2 \rightarrow \mathbb{R}$. Dans ce cas, la série entière la plus générale est

$$f_\theta(x_1, x_2) = \theta^{(0)} + \begin{bmatrix} \theta_0^{(1)} & \theta_1^{(1)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} \theta_{11}^{(2)} & \theta_{12}^{(2)} \\ \theta_{21}^{(2)} & \theta_{22}^{(2)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \dots \quad (3.22)$$

Contrairement à la section précédente, on doit introduire un tenseur de rang p pour le $p^{\text{ième}}$ terme dans la série entière, soit 2^p nouveaux paramètres qui doivent être ajustés. Dans le cas général $f_\theta : \mathbb{R}^n \rightarrow \mathbb{R}^m$, on doit plutôt introduire $n^p m$ nouveaux paramètres pour chaque terme ajouté dans la série. La complexité du modèle augmente de façon exponentielle en fonction du degré P du modèle. Ce comportement est radicalement différent du cas présenté à la section précédente, où la complexité du modèle augmentait de façon linéaire en fonction de P . Dû à ce fait, une série entière n'est pas une approche valide pour construire des modèles complexes en haute dimension ; le nombre de paramètres qu'on doit introduire devient rapidement intractable.

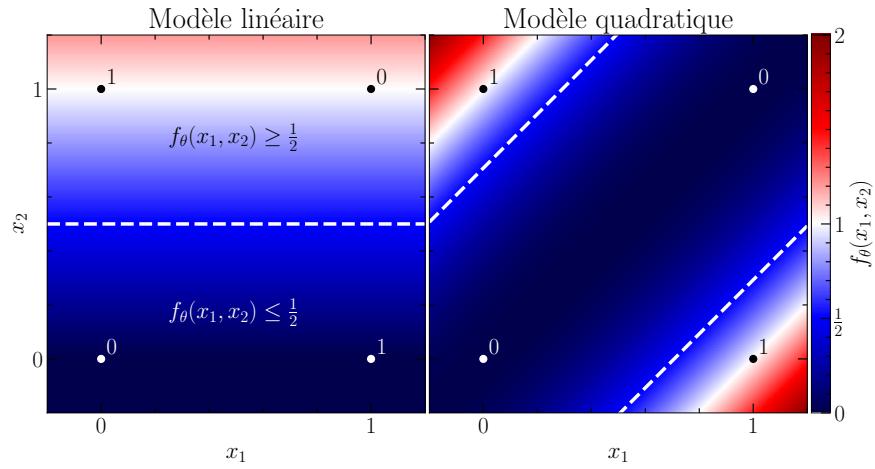


FIGURE 3.5 – Comparaison d'un modèle linéaire et d'un modèle quadratique pour le problème XOR.

Malgré cela, on peut explorer l'application d'un modèle comme (3.22) dans le but de développer une intuition géométrique qui servira à introduire la prochaine méthode pour complexifier f_θ . On considère l'exemple le plus simple qui requiert un modèle quadratique, soit la fonction logique OU exclusive (mieux connue comme XOR dans le jargon de la science informatique). Cette fonction logique peut être décrite de façon exhaustive avec la table 3.1.

x_1	x_2	$f_{\theta^*}(x_1, x_2)$
0	0	0
0	1	1
1	0	1
1	1	0

TABLE 3.1 – Fonction logique XOR.

Ce problème ne peut pas être résolu par une fonction linéaire puisque les 4 contraintes du problème ne sont pas linéairement séparables. Plus spécifiquement, on ne peut pas construire de plan qui sépare les points qui correspondent à $f_{\theta^*} = 0$ des points qui correspondent à $f_{\theta^*} = 1$. Le

panneau gauche de la figure 3.5 montre le modèle linéaire

$$f_{\hat{\theta}}(x_1, x_2) = \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = x_2 \quad (3.23)$$

pour illustrer cet argument, où on montre comment le plan $f_{\hat{\theta}}(x_1, x_2) = \frac{1}{2}$ (ligne pointillée blanche) n'est pas en mesure de séparer les 4 contraintes linéairement. De plus, aucune rotation ou translation (et en général aucune transformation affine) de ce modèle ne peut séparer les points 0 et 1 linéairement. Au mieux, on pourrait retrouver 3 contraintes avec le modèle $f_{\hat{\theta}} = x_1 - x_2$. Toutefois, on ne peut pas retrouver les 4 contraintes puisque l'ensemble des modèles linéaires ne contient pas la fonction XOR.

Pour résoudre le problème, on doit introduire le terme de second degré dans la série de puissances (3.22). Cet ensemble de fonctions contient une infinité de solutions au problème XOR. Le panneau de droite de la figure 3.5 illustre la solution particulière

$$f_{\hat{\theta}}(x_1, x_2) = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = (x_1 - x_2)^2 \quad (3.24)$$

qui recouvre parfaitement les 4 contraintes du problème XOR. Il est intéressant de noter que cette solution possède une quantité caractéristique, $h = (x_1 - x_2)^2$, qui est linéairement séparable. En effet, on peut tracer un plan $h = \frac{1}{2}$ qui sépare les points 0 et 1 dans cet espace.

Cette observation motive l'introduction d'une nouvelle méthode pour complexifier notre modèle. Cette méthode devrait minimalement être en mesure de construire un espace caractéristique (de l'anglais *feature space*) équivalent à $h = (x_1 - x_2)^2$, soit un espace caractéristique où les contraintes d'un problème donné sont linéairement séparables. C'est cette quête qui nous mène naturellement à l'apprentissage profond.

3.4 Les réseaux de neurones

Les réseaux de neurones ont été introduits par [Rosenblatt \(1958\)](#) comme des circuits analogues à des circuits biologiques de neurones. L'espoir était de mieux comprendre comment les systèmes biologiques sont en mesure de percevoir leur environnement ; comment l'information est apprise, préservée, remémorée et finalement comment ces systèmes sont en mesure de réfléchir, soit prendre en compte toute l'information à leur disposition (perception et mémoire) pour dicter ou influencer un comportement. Il s'avère toutefois que ces circuits ont des propriétés qui les rend particulièrement intéressants dans le contexte de l'apprentissage machine. Ce sont ces propriétés qui nous concernent dans cette section.

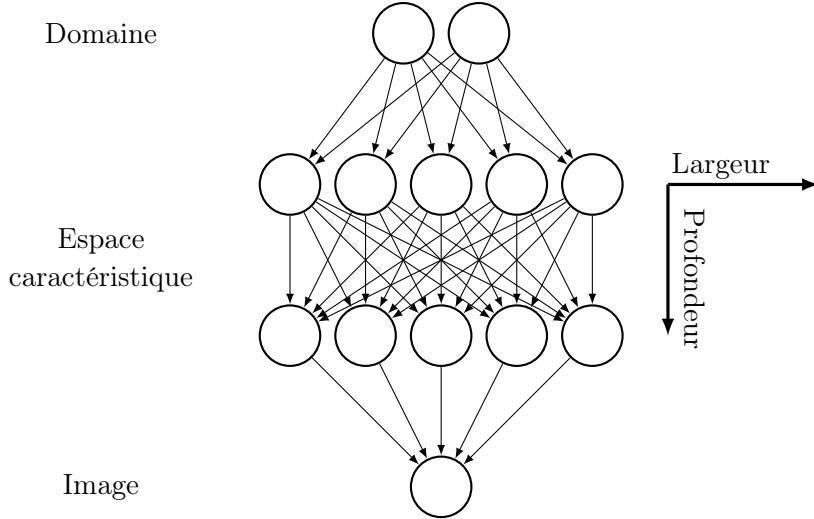


FIGURE 3.6 – Illustration d'un réseau de neurones avec deux couches latentes qui constituent son espace caractéristique.

La structure mathématique d'un réseau de neurones est généralement introduite à l'aide d'un graphe acyclique, tels qu'illustre à la figure 3.6. Une description tout à fait équivalente de ces réseaux est la composition de fonctions

$$f_{\theta}(\mathbf{x}) = (f_{\theta}^{(P)} \circ f_{\theta}^{(P-1)} \circ \cdots \circ f_{\theta}^{(1)})(\mathbf{x}). \quad (3.25)$$

Les couches du réseau (de l'anglais *layer*), $f_{\theta}^{(i)} : \mathbb{R}^n \rightarrow \mathbb{R}^m$, sont formées de deux composantes essentielles, soit une transformation affine $\mathbf{W}\mathbf{z} + \mathbf{b}$ et une fonction non linéaire $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ (aussi appelée fonction d'activation)

$$f_{\theta}^{(i)}(\mathbf{z}) = \sigma(\mathbf{W}^{(i)}\mathbf{z} + \mathbf{b}^{(i)}). \quad (3.26)$$

On note $\mathbf{W} \in \mathbb{R}^{m \times n}$, une matrice de poids (de l'anglais *weights*), et $\mathbf{b} \in \mathbb{R}^m$, un vecteur de biais (de l'anglais *bias*, à ne pas confondre avec le biais statistique mentionné précédemment). Dans cette expression, il est sous-entendu que la fonction d'activation, σ , s'applique identiquement sur les m éléments du vecteur $\mathbf{W}\mathbf{z} + \mathbf{b}$.

Individuellement, la transformation affine et la fonction d'activation ne sont pas en mesure de résoudre des problèmes comme XOR. En effet, une transformation affine préserve les lignes parallèles d'un espace vectoriel, de sorte qu'un ensemble de points qui n'est pas linéairement séparable le restera après une transformation affine. D'un autre côté, comme une fonction d'activation agit directement sur les coordonnées, ce type de fonction n'est pas en mesure de construire une quantité caractéristique qui mélange les coordonnées comme celle trouvée à la section précédente.

C'est l'application combinée de ces deux ingrédients qui permet minimalement de résoudre un problème comme XOR. Par exemple, considérons la fonction d'activation traditionnelle ReLU

(inspirée des neurones biologiques)

$$\text{ReLU}(x) = \begin{cases} x & x \geq 0 \\ 0 & x < 0 \end{cases} = \max(0, x), \quad x \in \mathbb{R} \quad (3.27)$$

Pour résoudre le problème XOR, on introduit une seule couche cachée, $f_\theta^{(1)}$, dans le réseau de neurones

$$f_\theta = \mathbf{W}(f_\theta^{(1)}(\mathbf{x})) + \mathbf{b}. \quad (3.28)$$

La dernière couche du réseau est une transformation affine seulement puisqu'on s'attend à ce que la couche cachée $f_\theta^{(1)}$ puisse construire un espace caractéristique qui sépare linéairement les contraintes du problème. Il s'avère que cette construction admet la solution suivante

$$f_{\hat{\theta}}(x_1, x_2) = [1 \ 1] \text{ReLU} \left(\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \right) = \max(0, x_1 - x_2) + \max(0, x_2 - x_1), \quad (3.29)$$

où la transformation affine de la couche cachée est la même que celle trouvée dans la section précédente. Pour comprendre intuitivement le rôle de chaque composante de ce réseau, la figure 3.7 illustre comment les opérations séquentielles du réseau de neurones (3.29) transforment l'espace du problème vers un espace caractéristique linéairement séparable (identifiable par une droite pointillée).

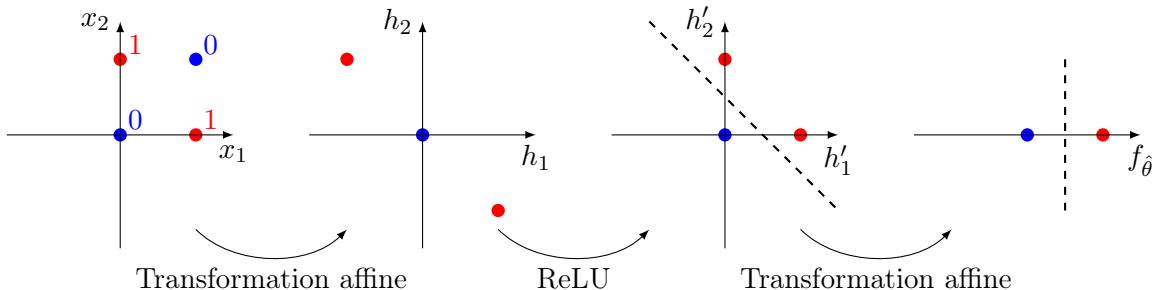


FIGURE 3.7 – Illustration d'une transformation du problème XOR vers un espace caractéristique linéairement séparable par un réseau de neurones.

L'espace caractéristique des réseaux de neurones est la source de leur flexibilité. Le théorème d'approximation universel (Cybenko, 1989; Hornik, 1991) stipule qu'un réseau avec une seule couche cachée d'une largeur suffisante (dimension de l'espace caractéristique) peut représenter n'importe quelle fonction continue. Il est possible d'obtenir une limite supérieure pour la largeur du réseau de neurones par des arguments géométriques dans des cas simples comme XOR. En général, un problème de classification binaire avec N points de données (4 dans le cas de XOR) est toujours linéairement séparable dans un espace caractéristique de dimensions D tels que $D \geq N - 1$.

Ce critère provient du théorème de Cover (1965) et de la dimension de Vapnik-Chervonenkis d'un classificateur linéaire (Vapnik et Chervonenkis, 1971; Vapnik, 1995). La probabilité qu'un

ensemble de N points soit linéairement séparable augmente considérablement en fonction de la dimension de l'espace, et atteint $p = 1$ lorsque $D \geq N - 1$. Pour le problème XOR, on a trouvé que $D = 2$ admettait une solution. Il est assez rare que la limite supérieure $D \geq N - 1$ soit atteinte puisque les données réelles tendent à se regrouper sur des variétés de basses dimensions. Néanmoins, spécifier une couche cachée suffisamment large peut devenir rapidement intractable dans certaines situations, particulièrement en hautes dimensions. C'est un problème similaire à celui rencontré à la section précédente avec la série entière. La prochaine section introduira une stratégie générale pour contourner ce problème, ainsi qu'une application spécifique aux réseaux de neurones convolutifs utilisés pour le traitement d'images.

Avant de poursuivre, on doit mentionner que les réseaux de neurones ne sont pas des fonctions linéaires en fonction de leurs paramètres (contrairement à la série entière). Ainsi, l'apprentissage profond introduit un problème d'optimisation non linéaire, soit un problème avec potentiellement plusieurs minimas secondaires et une géométrie non triviale. Les algorithmes d'apprentissage modernes sont largement équipés pour résoudre ce genre de problème grâce à l'introduction des dérivées automatiques et des algorithmes de descente de gradient stochastiques. Voir le livre de référence de [Goodfellow et al. \(2016\)](#) pour une discussion plus détaillée de ce sujet.

3.5 Les réseaux de neurones convolutifs

Le traitement d'image pose un problème sérieux à l'apprentissage machine en raison de la dimensionnalité du problème. Une image est généralement décrite par 3 dimensions qui décrivent respectivement la hauteur (H), la largeur (L) et le nombre de canaux de couleurs (C). La dimensionnalité d'une image correspond au nombre total de pixels qui décrivent cette image, soit $C \times H \times L$, un nombre qui peut facilement atteindre une valeur de l'ordre de $\mathcal{O}(10^5)$ en astronomie.

Pour construire une fonction avec un nombre raisonnable de paramètres, on utilise le principe d'équivariance. Ce principe est fortement inspiré du principe de la covariance, originellement formulé par Einstein, qui stipule qu'une théorie physique devrait être formulée en termes des quantités qui se transforment de façon covariantes par rapport à une action du groupe de symétrie de la théorie. Le principe d'équivariance stipule également qu'une fonction f_θ doit se transformer de façon équivariante sous l'action du groupe de symétrie G du problème à l'étude

$$\begin{aligned} x &\rightarrow Tx \\ f_\theta(x) &\rightarrow f'_\theta(Tx) = T' f_\theta(x). \end{aligned} \tag{3.30}$$

$T, T' \in G$ sont des éléments du groupe de symétrie. T' est la représentation de l'action T dans l'espace vectoriel de l'image de la fonction f_θ .

La convolution est une opération équivariante sous l'effet d'une translation. Pour rendre cette observation concrète, prenons l'exemple de la convolution entre le noyau de la convolution $\mathbf{W} \in \mathbb{R}^m$

et un vecteur $\mathbf{x} \in \mathbb{R}^n$

$$(\mathbf{W} * \mathbf{x})_j = \sum_{i=0}^{m-1} \mathbf{W}_i \mathbf{x}_{j-i} \quad (3.31)$$

L'effet d'une translation sur le vecteur \mathbf{x} est de modifier l'ordre des indices. On note T_w une translation qui déplace les indices de la façon suivante : $T_w \mathbf{x}_j = \mathbf{x}_{j+w}$. Il est sous-entendu que les indices du vecteur \mathbf{x} sont définis mod n . On peut alors montrer que la convolution (3.31) est une opération équivariante sous l'effet d'une translation

$$(\mathbf{W} * (T_w \mathbf{x}))_j = \sum_{i=0}^{m-1} \mathbf{W}_i \mathbf{x}_{j+w-i} = (\mathbf{W} * \mathbf{x})_{j+w} = T_w(\mathbf{W} * \mathbf{x})_j \quad (3.32)$$

Comparé à une transformation affine, le noyau de la convolution requiert beaucoup moins de paramètres pour un niveau comparable d'expressivité. En effet, il est souvent justifié de supposer que l'information à extraire du domaine est localement corrélée. En supposant que la distance de corrélation est $m < n$, on peut construire un noyau pour la convolution avec m paramètres. C'est un énorme avantage comparé à la transformation affine. Cette dernière a besoin, minimalement, de n paramètres pour chaque dimension de l'espace caractéristique.

Cet avantage devient significatif lorsqu'on considère des images $\mathbf{x} \in \mathbb{R}^{C \times H \times L}$, soit des tenseurs de rang 3. On considère une convolution sur les dimensions spatiales seulement (H et L) et on pose $W \in \mathbb{R}^{C \times c \times h \times \ell}$. On définit la convolution sur une image

$$(\mathbf{W} * \mathbf{x})_{ijk} = \sum_{m=0}^{C-1} \sum_{n=0}^{h-1} \sum_{p=0}^{\ell-1} \mathbf{W}_{minp} \mathbf{x}_{m,j-n,k-p} \quad (3.33)$$

Les dimensions spatiales du noyau sont généralement choisis tels que $h \ll H$ et $\ell \ll L$ puisque l'information pertinente est locale pour les tâches de traitement d'images typiques. Par exemple, la taille du noyau $h = \ell = 3$ est choix commun même lorsque la taille d'une image atteint $H \times L \gtrsim 10^4$. L'expressivité d'un réseau de neurones convolutifs ([Lecun et Bengio, 1995](#)) dépend principalement du nombre de canaux utilisés pour l'espace caractéristique et de la profondeur du réseau ([Krizhevsky et al., 2012](#)).

Ceci complète donc cette introduction à l'apprentissage machine. Comme l'architecture des réseaux de neurones est un sujet qui dépend du problème à l'étude, ce sujet est abordé dans les prochains chapitres qui traitent en détails des méthodes utilisées dans ce mémoire.

Chapitre 4

Apprentissage profond de distributions implicites

4.1 Auto-encodeur variationnel

Les auto-encodeurs variationnels (VAE) ont été introduits par [Kingma et Welling \(2013\)](#) comme une approche pour inférer approximativement les variables latentes (ou cachées) qui contrôlent un certain processus génératif. Leur utilité est particulièrement marquée lorsque ce processus génératif est défini implicitement par un échantillon de données, soit un cas où la forme fonctionnelle de la distribution n'est pas connue *a priori*. Dans cette section, j'introduis les concepts principaux reliés à ce type de modélisation. Le lecteur peut aussi se référer au livre blanc de [Kingma et Welling \(2019\)](#).

On définit $\mathbf{z} \in \mathbb{R}^h$ comme une variable latente et $\mathbf{x} \in \mathbb{R}^m$ ($m > h$) comme un exemple d'un échantillon de donnée $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^N$. Notre objectif est de modéliser la distribution, $p(\mathbf{x})$, implicitement décrite par notre échantillon. On définit une approximation de cette distribution, $p_\theta(\mathbf{x})$, caractérisée par une liste de paramètres θ , et on définit un processus génératif modélisé par la conditionnelle sur la variable cachée $p_\theta(\mathbf{x} | \mathbf{z})$. Déterminer p_θ directement est généralement difficile, voir impossible, si la dimensionnalité de \mathbf{x} est grande ($\dim(\mathbf{x}) \gtrsim 10^4$ pour des images). Pour contourner ce problème, on introduit une distribution variationnelle, $q_\phi(\mathbf{z} | \mathbf{x})$, dont le rôle est d'inférer la variable latente \mathbf{z} associée à $\mathbf{x} \sim p_\theta(\mathbf{x} | \mathbf{z})$. En d'autres mots, $q_\phi(\mathbf{z} | \mathbf{x})$ est une approximation variationnelle de la distribution *a posteriori* $p_\theta(\mathbf{z} | \mathbf{x})$. La notion de distance entre ces deux distributions est mesurée par la divergence de Kullback-Leibler $D_{\text{KL}}(\cdot \| \cdot) \geq 0$:

$$\begin{aligned} D_{\text{KL}}(q_\phi(\mathbf{z} | \mathbf{x}) \| p_\theta(\mathbf{z} | \mathbf{x})) &= \mathbb{E}_{q_\phi(\mathbf{z} | \mathbf{x})} \left[\log q_\phi(\mathbf{z} | \mathbf{x}) - \log p_\theta(\mathbf{z} | \mathbf{x}) \right]; \\ &= \mathbb{E}_{q_\phi(\mathbf{z} | \mathbf{x})} \left[\log q_\phi(\mathbf{z} | \mathbf{x}) - \log \frac{p_\theta(\mathbf{z}, \mathbf{x})}{p_\theta(\mathbf{x})} \right]; \end{aligned}$$

$$= \log p_\theta(\mathbf{x}) - \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log p_\theta(\mathbf{z}, \mathbf{x}) - \log q_\phi(\mathbf{z} | \mathbf{x}) \right]}_{\equiv \mathcal{L}_{\phi,\theta}(\mathbf{x})}. \quad (4.1)$$

On remarque par cette manipulation que la distance D_{KL} , en plus de mesurer la distance entre les deux distributions *a posteriori* (par définition), mesure aussi la différence entre le terme $\mathcal{L}_{\phi,\theta}(\mathbf{x})$, qu'on nomme limite inférieure sur l'évidence (de l'anglais *evidence lower bound* : ELBO), et la distribution marginale qu'on cherche à modéliser, $p_\theta(\mathbf{x})$. L'objectif d'un modèle VAE est de maximiser la ELBO, $\mathcal{L}_{\phi,\theta}$. En observant l'équation (4.1), on réalise que cet objectif nous permet d'améliorer le modèle d'inférence et le processus génératif simultanément. En effet, la divergence KL est une quantité positive, donc maximiser la ELBO a pour effet de

1. maximiser $p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x} | \mathbf{z}) p_\theta(\mathbf{z}) d\mathbf{z}$, ce qui suit de l'inégalité $\log p_\theta \geq \mathcal{L}_{\phi,\theta}(\mathbf{x})$ (améliore le processus génératif) ;
2. minimiser $D_{KL}(q_\phi(\mathbf{z} | \mathbf{x}) \| p_\theta(\mathbf{z} | \mathbf{x})) = \log p_\theta(\mathbf{x}) - \mathcal{L}_{\phi,\theta}(\mathbf{x})$ (améliore le processus d'inférence de \mathbf{z}).

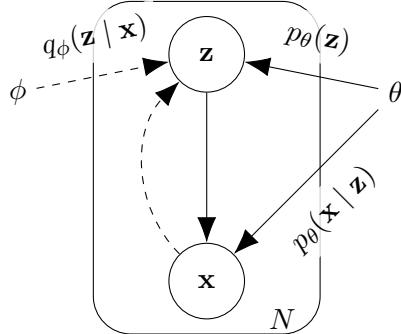


FIGURE 4.1 – Modèle graphique d'un VAE. Les flèches pleines indiquent le processus génératif, alors que les flèches pointillées indiquent le processus d'inférence.

4.1.1 Le truc de la reparamétrisation

Le gradient de la ELBO par rapport aux paramètres variationnels, $\nabla_{\phi,\theta} \mathcal{L}_{\phi,\theta}(\mathbf{x})$, est une quantité qu'on doit calculer pour faire usage d'algorithmes comme la descente de gradient stochastique pour maximiser la ELBO en termes de ϕ et θ . Or, la liste de paramètres ϕ apparaît dans la distribution de prélèvement pour calculer l'espérance mathématique $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}$ dans la ELBO (4.1). Cette opération n'a pas de dérivée formelle en termes de ϕ .

Pour résoudre ce problème, on utilise le truc de la reparamétrisation (Kingma et Welling, 2013), qui consiste à exprimer la variable aléatoire latente $\mathbf{z} \sim q_\phi(\mathbf{z} | \mathbf{x})$ comme la transformation différentiable et inversible d'une variable aléatoire auxiliaire ϵ . On considère le cas où $q_\phi(\mathbf{z} | \mathbf{x})$ et $p(\epsilon)$ font

partie de la famille gaussienne isotropique

$$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbb{1}); \quad (4.2)$$

$$\mathbf{z} = \boldsymbol{\mu}_\phi + \boldsymbol{\sigma}_\phi \odot \boldsymbol{\epsilon}, \quad (4.3)$$

de sorte que

$$\mathbf{z} \sim q_\phi(\mathbf{z} | \mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_\phi(\mathbf{x}), \mathbb{1}\boldsymbol{\sigma}_\phi^2(\mathbf{x})). \quad (4.4)$$

\odot symbolise le produit d’Hadamard, ou encore le produit élément par élément de vecteurs. La reparamétrisation fait en sorte que les paramètres variationnels ne participent plus au processus de prélèvement, maintenant pris en charge par $\boldsymbol{\epsilon}$. Cette propriété est cruciale, car elle nous permet de prendre le gradient de la ELBO (4.1). En effet, on peut maintenant échanger les opérateurs $\nabla_{\phi,\theta}$ et $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} = \mathbb{E}_{p(\boldsymbol{\epsilon})}$, ce qui nous permet d’appliquer le gradient à l’intérieur de l’espérance mathématique. De plus, ϕ décrit maintenant une fonction générique dont le rôle est d’inférer les paramètres de la distribution $q_\phi(\mathbf{z} | \mathbf{x})$

$$\begin{aligned} f_\phi : \mathbb{R}^m &\rightarrow \mathbb{R}^h \times \mathbb{R}^h \\ \mathbf{x} &\mapsto (\boldsymbol{\mu}, \log \boldsymbol{\sigma}^2). \end{aligned} \quad (4.5)$$

En pratique, on peut construire une approximation de cette fonction avec un réseau de neurones convolutif lorsque \mathbf{x} est une image, suivant le principe d’approximation universelle (Cybenko, 1989; Hornik, 1991).

L’objectif d’entraînement de la fonction f_ϕ nécessite de manipuler la ELBO pour obtenir une divergence KL

$$\mathcal{L}_{\phi,\theta}(\mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log p_\theta(\mathbf{z}, \mathbf{x}) - \log q_\phi(\mathbf{z} | \mathbf{x}) \right]; \quad (4.6)$$

$$\implies \mathcal{L}_{\phi,\theta}(\mathbf{x}) = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log p_\theta(\mathbf{x} | \mathbf{z}) \right]}_{\text{terme de reconstruction}} + \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log p_\theta(\mathbf{z}) - \log q_\phi(\mathbf{z} | \mathbf{x}) \right]}_{\equiv -D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z}))}. \quad (4.7)$$

Pour déterminer la forme fonctionnelle de la divergence KL obtenue au second terme du membre droit de l’équation (4.7), Ton stipule *a priori* que la distribution marginale des variables latentes devrait correspondre à une distribution normale isotropique

$$p_\theta(\mathbf{z}) = \mathcal{N}(0, \mathbb{1}). \quad (4.8)$$

La KL admet alors une solution fermée étant donné les familles paramétriques stipulées pour $p_\theta(\mathbf{z})$ (4.8) et $q_\phi(\mathbf{z} | \mathbf{x})$ (4.4)

$$-D_{\text{KL}}(q_\phi(\mathbf{z} | \mathbf{x}) \| p_\theta(\mathbf{z})) = \frac{1}{2} \sum_{j=1}^h (1 + [\log \boldsymbol{\sigma}_\phi^2]_j - [\boldsymbol{\mu}_\phi]_j - [\boldsymbol{\sigma}_\phi^2]_j). \quad (4.9)$$

Une dérivation de ce terme est donnée dans l'annexe B de Kingma et Welling (2013). Le premier terme du membre droit de l'équation (4.7) est nommé *terme de reconstruction* puisqu'il connecte avec l'objectif des fonctions de type auto-encodeur d'apprendre une représentation latente d'un échantillon de données. La reconstruction s'accomplit en utilisant d'abord le modèle d'inférence $\mathbf{z}^{(1:L)} \stackrel{\text{i.i.d.}}{\sim} q_\phi(\mathbf{z} \mid \mathbf{x})$ pour obtenir un échantillon de représentations latentes à partir des équations (4.2) à (4.3), puis en utilisant le modèle génératif $\hat{\mathbf{x}}^{(i)} \sim p_\theta(\mathbf{x} \mid \mathbf{z}^{(i)})$ pour obtenir un échantillon de reconstructions $\hat{\mathbf{x}}^{(1:L)}$ de l'exemple originel \mathbf{x} . Comme on a déjà une variable auxiliaire ϵ qui se charge de l'aspect génératif du modèle, on peut construire une approximation du modèle génératif avec une fonction générique des variables latentes $g_\theta : \mathbb{R}^h \rightarrow \mathbb{R}^m; \mathbf{z}^{(i)} \mapsto \hat{\mathbf{x}}^{(i)}$. Encore une fois, un réseau de neurones convolutif est un choix pratique pour modéliser cette fonction dans le cas où \mathbf{x} est une image. En général, on choisit une erreur quadratique moyenne pour modéliser le terme de reconstruction, de sorte que

$$\mathbb{E}_{q_\phi(\mathbf{z} \mid \mathbf{x})} \left[\log p_\theta(\mathbf{x} \mid \mathbf{z}) \right] = -\frac{1}{L} \sum_{i=1}^L \|\mathbf{x} - \hat{\mathbf{x}}^{(i)}\|_2^2. \quad (4.10)$$

4.1.2 Principe du goulot d'information

La fondation théorique des auto-encodeurs variationnels repose sur le principe plus général du goulot d'information (BIP, de l'anglais *bottleneck information principle* : Tishby et al., 1999). Dans cette sous-section, je décris rapidement certains concepts liés à la théorie de l'information de Shannon (1948) pour justifier l'introduction d'un multiplicateur de Lagrange β au terme de régularisation de la ELBO, $-D_{\text{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) \parallel p_\theta(\mathbf{z}))$. Pour une discussion en profondeur, voir l'excellente revue sur l'utilisation de BIP dans le contexte de l'apprentissage machine par Goldfeld et Polyanskiy (2020) et le manuel de référence sur la théorie de l'information par Cover et Thomas (2006).

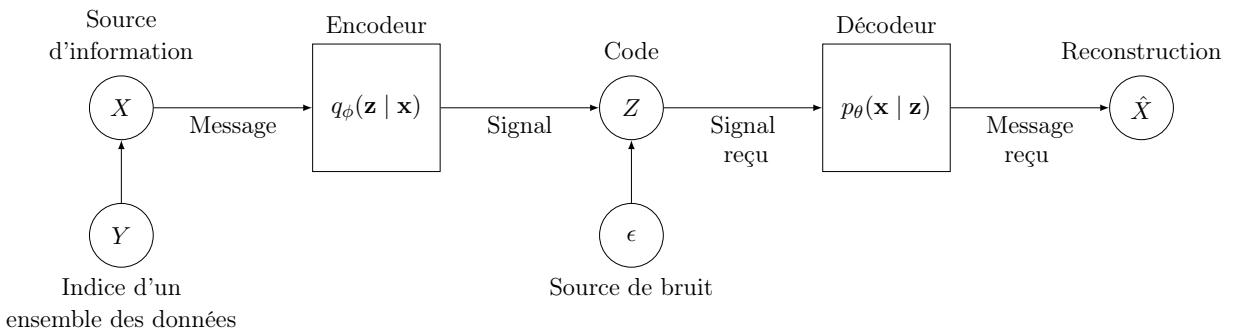


FIGURE 4.2 – VAE comme un système de transmission d'information.

L'objectif d'un auto-encodeur est de construire un code Z d'une longueur minimale qui capture un maximum d'information contenue dans un message X . Formellement, on utilise l'information

mutuelle de Shannon pour mesurer l'information capturée par Z à propos de X (et vice-versa)

$$I(Z; X) = D_{\text{KL}}(p(\mathbf{x}, \mathbf{z}) \| p(\mathbf{x})p(\mathbf{z})) . \quad (4.11)$$

Il est utile de décomposer cet objectif en termes de l'entropie du message $H(X)$ et l'entropie conditionnelle de ce message étant donné le code Z , $H(X | Z)$

$$I(Z; X) = H(X) - H(X | Z) . \quad (4.12)$$

L'entropie est une mesure de l'incertitude dans une variable aléatoire

$$H(X) = -\mathbb{E}_{p(\mathbf{x})}[\log p(\mathbf{x})] \geq 0 , \quad (4.13)$$

qu'on interprète aussi comme une mesure de la longueur minimale d'un code, Z , pour transmettre un message, X , d'un émetteur à un receveur avec le minimum de perte d'information ([Shannon, 1948](#); [Kolmogorov, 1965](#)). L'entropie conditionnelle $H(X | Z)$ est une mesure de l'incertitude résiduelle étant donné la connaissance du code Z . Dans le cas où Z détermine complètement le message X , alors $H(X | Z) = 0$ et l'information mutuelle atteint son maximum $I(Z; X) = H(X)$. Dans le cas où Z et X sont deux variables aléatoires indépendantes, $H(X | Z) = H(X)$ et l'information mutuelle devient nulle.

Déterminer l'auto-encodeur optimal, c.-à-d. celui qui maximise $I(Z; X)$, est un problème mal posé. En effet, on pourrait naïvement maximiser l'objectif (4.11) avec la fonction identité comme auto-encodeur : $f_{\phi, \theta}(X) = \mathbb{1}X$, de sorte que $Z = X$ et $I(Z; X) = H(X)$. Or, Z n'est pas une représentation pertinente dans ce cas. Pour éliminer les solutions non désirées, on introduit une contrainte sur la complexité de [Kolmogorov \(1965\)](#) du code Z , c.-à-d. qu'on cherche un auto-encodeur qui compresse le message et conserve simultanément le maximum d'information possible à propos du message. On introduit l'objectif du goulot d'information ([Tishby et al., 1999](#))

$$\begin{aligned} & \max_{\phi, \theta} I(Z; X) \\ & \text{sujet à } I(Z; Y) \leq \alpha . \end{aligned} \quad (4.14)$$

La contrainte $I(Z; Y) \leq \alpha$ impose à l'auto-encodeur une limite sur l'information mutuelle entre le code utilisé pour représenter X et l'identité $Y = i$ de chaque exemple d'un ensemble des données qu'on veut modéliser $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^N$. Il s'avère que l'objectif de comprimer l'information est intimement lié à l'objectif d'obtenir un sommaire informatif, c.-à-d. qu'un code d'une complexité de [Kolmogorov \(1965\)](#) minimale décrit la source d'information de la manière la plus informative possible. Cette connexion remarquable est une application concrète du principe du rasoir d'Occam : l'explication adéquate la plus simple est la meilleure explication.

Dans ce qui suit, je m'applique à redériver l'objectif d'un auto-encodeur variationnel suivant les approximations proposées par [Alemi et al. \(2017\)](#). Puis, je termine avec une courte interprétation des

VAE sous la lumière du principe du goulot d'information. On commence par construire une limite variationnelle inférieure sur $I(Z; X)$. On suppose d'abord la chaîne de Markov $Y \rightarrow X \rightarrow Z$ pour les variables aléatoires représentées dans la figure 4.2. La chaîne de Markov induit la factorisation de la probabilité jointe

$$p(X, Y, Z) = p(X | Z)p(Y | X)p(Z). \quad (4.15)$$

où $p(X | Z, Y) = p(X | Z)$ suit du fait que Z et Y sont des variables conditionnellement indépendantes étant donné X . On introduit ensuite un modèle pour l'encodeur, $q_\phi(Z | X)$, soit un système de transmission d'informations par compression. Il suit que

$$I(Z; X) \equiv \mathbb{E}_{p(\mathbf{z}, \mathbf{x})} \left[\log \frac{p(\mathbf{x}, \mathbf{z})}{p(\mathbf{x})p(\mathbf{z})} \right]; \quad (4.16)$$

$$= \mathbb{E}_{p(\mathbf{x})} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log \frac{q_\phi(\mathbf{x} | \mathbf{z})}{p(\mathbf{x})} \right]; \quad (4.17)$$

$$= -\mathbb{E}_{p(\mathbf{x})} \left[\log p(\mathbf{x}) \right] + \mathbb{E}_{p(\mathbf{x})} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log q_\phi(\mathbf{x} | \mathbf{z}) \right]; \quad (4.18)$$

$$\implies I(Z; X) \geq \mathbb{E}_{p(\mathbf{x})} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log p_\theta(\mathbf{x} | \mathbf{z}) \right]. \quad (4.19)$$

À la dernière ligne, on a éliminé l'entropie du message $H(X)$, puisque c'est une constante strictement positive qui ne dépend pas de ϕ . On a aussi introduit l'approximation variationnelle

$$p_\theta(\mathbf{x} | \mathbf{z}) \approx q_\phi(\mathbf{x} | \mathbf{z})$$

pour approximer la distribution *a posteriori* de l'encodeur (soit le décodeur). Cette approximation est valide dans le contexte où on cherche une limite inférieure pour $I(Z; X)$, puisque la divergence KL entre ces deux distributions est strictement positive. Ensuite, on cherche une borne supérieure au terme de compression $I(Z; Y)$. Pour ce qui suit, on remplace $p(Y = i)$ par $\mathbf{x}^{(i)} \sim \mathcal{D}$ pour illustrer avec plus de clarté comment le processus génératif de l'indice induit une sélection d'un exemple $\mathbf{x}^{(i)}$ dans l'ensemble d'entraînement \mathcal{D} . Il suit que

$$I(Z; Y) = \mathbb{E}_{\mathbf{x}^{(i)} \sim \mathcal{D}} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}^{(i)})} \left[\log \frac{q_\phi(\mathbf{z} | \mathbf{x}^{(i)})}{p_\theta(\mathbf{z})} \right]; \quad (4.20)$$

$$I(Z; Y) \leq \mathbb{E}_{\mathbf{x}^{(i)} \sim \mathcal{D}} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}^{(i)})} \left[\log \frac{q_\phi(\mathbf{z} | \mathbf{x}^{(i)})}{p_\theta(\mathbf{z})} \right], \quad (4.21)$$

où on a introduit l'approximation variationnelle $p_\theta(\mathbf{z}) \approx q_\phi(\mathbf{z})$. L'inégalité suit encore une fois du fait que la divergence KL entre ces deux distributions est strictement positive. On finit la dérivation en écrivant l'objectif du goulot d'information (4.14) en introduisant un multiplicateur de Lagrange β pour le terme $I(Z; Y)$ et en introduisant les limites variationnelles (4.19) et (4.21). On ignore

l'espérance mathématique $\mathbb{E}_{p(\mathbf{x})} = \mathbb{E}_{\mathbf{x}^{(i)} \sim \mathcal{D}}$ pour mieux illustrer la connexion avec la ELBO (4.7)

$$\mathcal{L}_{\phi,\theta,\beta}(\mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log p_\theta(\mathbf{x} | \mathbf{z}) \right] - \beta \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log q_\phi(\mathbf{z} | \mathbf{x}) - \log p_\theta(\mathbf{z}) \right]. \quad (4.22)$$

On note qu'on a laissé tomber la constante α , car elle ne dépend pas des paramètres ϕ et θ . Généralement, le paramètre β n'est pas optimisé directement, il est plutôt considéré comme un hyperparamètre. Je note de plus que l'objectif (4.22) a fait une première apparition dans le travail introduisant les β -VAE par [Higgins et al. \(2017\)](#), suivant des motivations complètement différentes de celles qu'on a suivies ici.

La dérivation accomplie dans cette sous-section montre clairement que la ELBO, $\mathcal{L}_{\phi,\theta,\beta}(\mathbf{x})$, est une limite inférieure sur $I(Z; X) - \beta I(Z; Y)$, soit l'objectif de maximiser l'information transmise par un auto-encodeur, avec une contrainte sur la complexité du code Z . En particulier, cette dérivation illumine le choix de la marginale $p_\theta(\mathbf{z})$ comme étant une approximation variationnelle de la marginale de l'encodeur $q_\phi(\mathbf{z})$, contrairement à la dérivation montrée plus haut où $p_\theta(\mathbf{z})$ apparaît comme un objectif arbitraire permettant d'améliorer l'aspect génératif du décodeur.

La dérivation commence d'un point de vue complètement différent de [Kingma et Welling \(2013\)](#). On a supposé que le modèle génératif $p_\theta(\mathbf{x} | \mathbf{z})$ est une approximation variationnelle de la distribution *a posteriori* du modèle d'inférence $q_\phi(\mathbf{x} | \mathbf{z})$. L'approche présentée par [Kingma et Welling \(2013\)](#) est exactement l'inverse. De plus, le terme de régularisation a maintenant une interprétation beaucoup plus riche, avec un paramètre β qui contrôle le niveau de compression de l'information désirée et qui s'avère à contrôler plus ou moins directement le nombre de partitions possibles dans l'espace latent ([Alemi et al., 2017](#); [Jimenez Rezende et Viola, 2018](#)). La valeur optimale de β dépend largement du problème, c.-à-d. $p(X)$, et de la capacité des modèles q_ϕ et p_θ . Pour une discussion détaillée de l'espace de phase de la ELBO en fonction de β , je réfère le lecteur au travail de [Alemi et al. \(2018\)](#).

4.2 Machines à inférence récurrentielles

4.2.1 Formalisme bayésien des problèmes inverses

Les machines à inférence récurrentielle (RIM) ont été introduites par [Putzky et Welling \(2017\)](#) pour résoudre des problèmes inverses pour lesquels le terme de régularisation est nécessaire, mais inconnu *a priori* et/ou difficile à construire, voir même calculer. Dans cette section, j'introduis le formalisme bayésien des problèmes inverses sur lequel ce modèle repose, puis j'introduis l'algorithme d'inférence et les concepts d'apprentissage machine qui motivent l'utilisation d'une RIM pour des problèmes inverses mal posés et sous-déterminés.

Les problèmes inverses en astrophysique prennent généralement la forme

$$\mathbf{y} = F(\mathbf{x}) + \boldsymbol{\eta}, \quad (4.23)$$

où $\mathbf{y} \in \mathcal{Y}$ est un vecteur d'observables (comme la valeur des pixels d'une image capturée par les capteurs photographiques CCD dans un télescope), $\mathbf{x} \in \mathcal{X}$ est un vecteur de paramètres qui gouvernent le phénomène physique et que l'on souhaite généralement obtenir, et $F : \mathcal{X} \rightarrow \mathcal{Y}$ est le modèle physique. Le vecteur $\boldsymbol{\eta}$ est une réalisation d'un bruit additif. On suppose que l'on connaît la distribution de ce bruit, de sorte qu'on peut modéliser la fonction de vraisemblance de l'observable

$$\mathbf{y} - F(\mathbf{x}) \sim p(\boldsymbol{\eta}) = p(\mathbf{y} | \mathbf{x}). \quad (4.24)$$

Le problème d'inférence est celui de déterminer les paramètres \mathbf{x} qui reproduisent l'observation \mathbf{y} , c.-à-d. l'estimé des paramètres $\hat{\mathbf{x}}_{\text{MLE}}$ qui maximisent la fonction de vraisemblance (MLE de l'anglais *maximum likelihood estimate*), ou, de façon équivalente, ceux qui maximisent le log de la vraisemblance

$$\hat{\mathbf{x}}_{\text{MLE}} = \underset{\mathbf{x} \in \mathcal{X}}{\operatorname{argmax}} \log p(\mathbf{y} | \mathbf{x}). \quad (4.25)$$

Dans le cas général, ce problème est mal posé et n'a pas de solutions. En effet, tel que l'observe Hadamard (1902), un problème aux dérivées partielles comme (4.25) ne possède une solution que si le problème est déterminé, c.-à-d. que, dans le langage de Hadamard (1902), le problème doit correspondre en entier à une situation physique. Cette connexion remarquable s'exprime en trois conditions qui déterminent si un problème inverse est bien posé

- (H_1) Une solution existe ;
- (H_2) Cette solution est unique ;
- (H_3) La fonction $G_\varphi : \mathcal{Y} \rightarrow \mathcal{X}$ qui infère les paramètres \mathbf{x} satisfait la condition de Lipschitz.

Le troisième critère (H_3) requiert que la fonction d'inférence soit stable, c.-à-d. qu'un petit changement dans le vecteur d'observations devrait correspondre à un petit changement de la solution, mesuré par la constante de Lipschitz $L \geq 0$

$$\|G_\varphi(\mathbf{y}_1) - G_\varphi(\mathbf{y}_2)\|_{\mathcal{X}} \leq L\|\mathbf{y}_1 - \mathbf{y}_2\|_{\mathcal{Y}}, \quad (4.26)$$

où $\|\cdot\|_{\mathcal{V}}$ est une métrique de distance définie pour l'espace vectoriel \mathcal{V} .

Pour un problème mal posé, on suppose *a priori* que la première condition de Hadamard (H_1) est respectée, c.-à-d. qu'on suppose que les quantités observées ou mesurées sont causées par un phénomène unique (solution physique). Toutefois, comme les problèmes qui nous intéressent sont sous-déterminés, c.-à-d. que $\dim_{\mathbb{R}}(\mathcal{X}) > \dim_{\mathbb{R}}(\mathcal{Y})$, la seconde condition de Hadamard (H_2) n'est pas respectée ; la fonction de vraisemblance ne peut pas distinguer la solution physique du nombre infini de solutions non physiques au problème (4.25).

La condition d'unicité de la solution est résolue par la construction d'une mesure de probabilité *a priori* sur l'espace des paramètres d'intérêt $p_\theta : \mathcal{X} \rightarrow \mathbb{R}$, t.q. $\int_{\mathcal{X}} p_\theta(\mathbf{x}) d\mathbf{x} = 1$, tel que les solutions non physiques sont exclues de la région de haute densité de cette distribution. On peut alors modifier

le problème (4.25) en introduisant cette distribution *a priori* comme un terme de régularisation de la vraisemblance

$$\hat{\mathbf{x}}_{\text{MAP}} = \underset{\mathbf{x} \in \mathcal{X}}{\operatorname{argmax}} \log p(\mathbf{y} \mid \mathbf{x}) + \log p_\theta(\mathbf{x}). \quad (4.27)$$

La solution $\hat{\mathbf{x}}_{\text{MAP}}$ maximise la distribution posteriori $p_\theta(\mathbf{x} \mid \mathbf{y})$, tel que définie par le théorème de Bayes

$$p_\theta(\mathbf{x} \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \mathbf{x})p_\theta(\mathbf{x})}{\int_{\mathcal{X}} p(\mathbf{y} \mid \mathbf{x})p_\theta(\mathbf{x})d\mathbf{x}}. \quad (4.28)$$

Le dénominateur est une constante qu'on nomme l'évidence bayesienne. Pour les applications qui nous intéressent, cette constante n'est pas calculée, car elle n'est pas nécessaire pour la recherche d'un maximum de la distribution *a posteriori* ou la comparaison de solutions par le ratio de la fonction de vraisemblance (ou de la distribution *a posteriori*).

On note que la stratégie la plus commune pour résoudre les problèmes inverses qui nous intéressent est plutôt de choisir judicieusement l'espace des solutions \mathcal{X} tel que $\dim_{\mathbb{R}}(\mathcal{X}) \leq \dim_{\mathbb{R}}(\mathcal{Y})$. Dans ce cas, le problème inverse est bien posé ou sur-déterminé. Par exemple, pour modéliser la masse d'une lentille gravitationnelle, il est commun de choisir un modèle singulier isotherme ou une loi de puissance elliptique (e.g. [Koopmans et al., 2006](#); [Barnabè et al., 2009](#); [Auger et al., 2010](#)), caractérisé par quelques paramètres seulement ($\dim_{\mathbb{R}}(\mathcal{X}) \sim 10$), tandis que l'observation \mathbf{y} est une image avec $\dim_{\mathbb{R}}(\mathcal{Y}) \gtrsim 10^4 \gg \dim_{\mathbb{R}}(\mathcal{X})$. Cette approche est considérablement plus stable que les méthodes sous-déterminées.

Toutefois, les modèles analytiques deviennent rapidement complexes et difficiles à construire, voir justifier, lorsque l'observation des systèmes qui nous intéressent est de haute qualité. Ceci révèle la complexité cachée de ces systèmes (e.g. [Schuldt et al., 2019](#)). De plus, ce cadre nous limite à seulement considérer les hypothèses construites par des humains ou par régression symbolique (e.g. [Lemos et al., 2022](#)), et non l'ensemble des hypothèses possibles. C'est cette observation qui nous motive à utiliser l'approche esquissée plus haut, où l'espace \mathcal{X} est construit de manière presque agnostique à la solution physique recherchée (p. ex. une grille de pixels pour modéliser une distribution de masse), de manière à contenir toutes, ou au moins la plupart, des solutions physiques. Une telle approche a le potentiel de produire des résultats surprenants ou intéressants, puisque l'exploration de l'espace des solutions physiques peut être ajustée, en principe, via la distribution *a priori*, $p_\theta(\mathbf{x})$, selon la complexité de l'observation.

4.2.2 La relation de récurrence

Pour résoudre l'équation différentielle ordinaire sous-entendue par le problème (4.27), on considère la méthode de discréétisation d'Euler

$$\hat{\mathbf{x}}^{(t+1)} = \hat{\mathbf{x}}^{(t)} + \alpha \nabla_{\hat{\mathbf{x}}^{(t)}} p_\theta(\hat{\mathbf{x}}^{(t)} \mid \mathbf{y}), \quad (4.29)$$

où α est le taux d'apprentissage dans la littérature sur l'apprentissage machine. La relation de récurrence (4.29) satisfait la condition de Lipschitz si l'erreur locale de chaque itération est proportionnelle à α^2 , ce qui est satisfait si le gradient $\nabla_{\mathbf{x}} \log p_{\theta}(\mathbf{x} | \mathbf{y})$ satisfait la condition de Lipschitz dans la région de \mathcal{X} explorée par l'algorithme (Atkinson, 1989; Butcher, 2016), ou encore si la norme de la dérivée seconde de $\log p_{\theta}(\mathbf{x} | \mathbf{y})$ est bornée dans cette région.

Putzky et Welling (2017) observent qu'on peut réécrire (4.29) en utilisant le théorème de Bayes et en absorbant le gradient de la distribution *a priori* dans une fonction $g_{\varphi(t)}$

$$\hat{\mathbf{x}}^{(t+1)} = \hat{\mathbf{x}}^{(t)} + \alpha (\nabla_{\hat{\mathbf{x}}^{(t)}} \log p(\mathbf{y} | \hat{\mathbf{x}}^{(t)}) + \nabla_{\hat{\mathbf{x}}^{(t)}} \log p_{\theta}(\hat{\mathbf{x}}^{(t)})); \quad (4.30)$$

$$\implies \hat{\mathbf{x}}^{(t+1)} = \hat{\mathbf{x}} + g_{\varphi(t)}(\hat{\mathbf{x}}^{(t)}, \nabla_{\hat{\mathbf{x}}^{(t)}} \log p(\mathbf{y} | \hat{\mathbf{x}}^{(t)})), \quad (4.31)$$

où $g_{\varphi(t)} : \mathcal{X}^2 \rightarrow \mathcal{X}$ est le modèle du gradient de la distribution *a posteriori*. On remarque que la relation de récurrence (4.29) est un cas spécial de la relation (4.31), soit le cas où on a un modèle explicite pour la distribution *a priori*, ou son gradient $\nabla_{\mathbf{x}} \log p_{\theta}(\mathbf{x})$, et le taux d'apprentissage α . Dans la relation (4.31), les paramètres α et θ sont absorbés dans les paramètres d'inférence $\varphi^{(t)}$, ce qui nous donne une plus grande liberté pour modéliser la distribution *a priori* en utilisant le théorème d'approximation universelle (Cybenko, 1989; Hornik, 1991). Selon ce nouveau point de vue, le problème de modéliser la distribution *a priori*, ou plus directement le gradient de la distribution *a priori*, est équivalent à construire un modèle pour le gradient de la distribution *a posteriori* dans une relation de récurrence.

Pour le problème de reconstruction d'image, les modèles neuronaux convolutifs avec une architecture de sablier (auto-encodeur) ou avec une architecture U-net (Ronneberger et al., 2015) sont des choix naturels pour modéliser $g_{\varphi(t)}$. Toutefois, la troisième condition d'Hadamard (H_3) n'est pas trivialement respectée lorsque $g_{\varphi(t)}$ est un réseau de neurones. Dans ce travail, cette condition n'est pas explicitement imposée au modèle. On note toutefois que l'analyse de la condition de Lipschitz pour les réseaux neuronaux est un sujet de recherche important (e.g. Miyato et al., 2018; Scaman et Virmaux, 2018; Weng et al., 2018), spécifiquement pour l'étude de méthodes robustes d'entraînement des réseaux neuronaux pour prévenir ou défendre contre les attaques antagonistes (Szegedy et al., 2013; Goodfellow et al., 2014).

Finalement, on note un aspect important du modèle $g_{\varphi(t)}$, soit la possible dépendance des paramètres $\varphi^{(t)}$ envers t . Plusieurs algorithmes d'optimisation récents comme la méthode d'accélération de Nesterov (1983), AdaGrad (Duchi et al., 2011), RMSProp¹ (Hinton, 2012) et ADAM (Kingma et Ba, 2014), utilisent explicitement l'information des gradients d'itérations antérieures à t pour calculer la mise à jour dans la relation de récurrence (4.31). Cette propriété permet à ces algorithmes de collecter de l'information par rapport à la seconde dérivée de la fonction objective, sans la calculer directement. Ainsi, il est important de considérer une classe de modèles avec une mémoire des itérations précédentes. Pour ce faire, on augmente la relation de récurrence (4.31) avec une seconde

¹L'algorithme apparaît en premier dans le cours CSC321 à l'Université de Toronto, donné par Geoffrey Hinton en 2011.

relation de récurrence sur un tenseur caché $\mathbf{h}^{(t)}$:

$$\mathbf{h}^{(t)} = g_\varphi(\hat{\mathbf{x}}^{(t)}, \nabla_{\hat{\mathbf{x}}^{(t)}} \log p(\mathbf{y} | \hat{\mathbf{x}}^{(t)}), \mathbf{h}^{(t-1)}), \quad (4.32)$$

Dans le cas de ADAM, la relation de récurrence cachée est un lissage exponentiel des deux premiers moments du gradient d'une fonction objective. La relation (4.32) est donc une généralisation de ADAM, où on a remplacé le lissage exponentiel par une relation récurrente sur un tenseur caché $\mathbf{h}^{(t)}$ quelconque, dont le rôle est de compresser l'information de la trajectoire $\hat{\mathbf{x}}^{(0:t)}$ dans un sommaire informatif. Ayant ainsi comprimé l'information dans une variable séparée, on peut réutiliser les poids du réseau de neurones φ à chaque itération temporelle t , ce qui réduit considérablement la complexité du problème d'apprentissage.

4.2.3 Méta-apprentissage par rétropropagation de gradients

Cette sous-section s'intéresse finalement au problème d'apprentissage d'une machine à inférence récurrentielle g_φ , et en particulier comment ce problème se situe dans le catégorie de méta-apprentissage. Le méta-apprentissage est un sujet de recherche qui se concentre sur la construction de règles d'apprentissage qui permettent d'accélérer l'apprentissage de fonctions pour certaines tâches spécifiques. Ce double aspect d'apprentissage, soit d'apprendre à mieux apprendre, est précisément ce qui est sous-entendu par le terme méta-apprentissage. Pour une revue du sujet, le lecteur peut consulter la revue par [Hospedales et al. \(2020\)](#). Cette sous-section se veut une introduction au méta-apprentissage basé sur l'optimisation, soit l'apprentissage profond de biais inductifs par rétropropagation de gradients.

Pour un problème de méta-apprentissage, l'ensemble des données d'entraînement est légèrement différent d'une tâche de régression ou de classification. Considérons un ensemble d'entraînement $\mathcal{D} = \{\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\}_{i=1}^N$, construit à partir d'exemples dans le domaine \mathcal{X} et l'image \mathcal{Y} , implicitement connectés par une fonction qu'on veut reconstruire ou approximer. Dans ce contexte, l'objectif est généralement d'apprendre une fonction $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$ qui minimise l'erreur quadratique moyenne, soit le risque empirique sur l'ensemble d'entraînement

$$\theta^* = \operatorname{argmin}_\theta \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}} \left[\|f_\theta(\mathbf{x}) - \mathbf{y}\|_{\mathcal{Y}}^2 \right] \quad (4.33)$$

Dans le contexte du méta-apprentissage, l'ensemble d'entraînement est plutôt constitué de tâches à performer $\mathcal{T} = \{\mathcal{D}^{(i)}, \mathcal{L}^{(i)}\}_{i=1}^N$, où $\mathcal{L}^{(i)}$ est une fonction objective pour la tâche i et $\mathcal{D}^{(i)}$ est l'ensemble des données pour résoudre ce problème. Le problème de méta-apprentissage est donc d'extraire ou encoder des biais inductifs, c.-à-d. des connaissances qui permettent d'accélérer et généraliser l'apprentissage d'une tâche spécifique, dans une liste de paramètres φ . Spécifiquement

pour le méta-apprentissage par optimisation, on a le double niveau d'optimisation

$$\begin{aligned} \varphi_{\text{méta}}^* &= \underset{\varphi}{\operatorname{argmin}} \mathbb{E}_{(\mathcal{L}, \mathcal{D}) \sim \mathcal{T}} \left[\mathcal{L}_{\theta^*(\varphi)}^{\text{méta}}(\mathcal{D}) \right]; \\ \text{sujet à } \theta^*(\varphi) &= \underset{\theta}{\operatorname{argmin}} \mathcal{L}_{\theta(\varphi)}^{\text{tâche}}(\mathcal{D}) \end{aligned} \quad (4.34)$$

Il est pertinent de considérer la notion de généralisation dans ce contexte, et en particulier faire le contraste avec la notion de généralisation dans le contexte de la régression. Dans le contexte de la régression, le concept de généralisation est synonyme avec celui d'extrapolation, c.-à-d. la mesure de la performance d'une certaine fonction, d'un algorithme ou d'une loi physique étant donné un ou plusieurs exemples tests, potentiellement similaires ou différents des exemples de l'ensemble d'entraînement utilisés pour ajuster les paramètres d'une fonction $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$. Pour plusieurs objectifs scientifiques, la capacité d'un modèle ou d'une loi physique à généraliser au-delà d'un ensemble des mesures \mathcal{D} est cruciale. Par exemple, un test important de la théorie de la relativité générale, testée et découverte dans le régime des champs faibles, est de savoir si les lois physiques restent valides dans le régime des champs forts près des trous noirs, des étoiles à neutrons ou encore au moment du Big Bang.

Or, si les biais inductifs utilisés pour entraîner (apprendre, découvrir) f_θ ne sont pas suffisamment informatifs sur la nature du problème ou du phénomène d'intérêt (p. ex. la gravité), alors un algorithme d'optimisation générique n'est pas garanti de converger vers une hypothèse universelle, c.-à-d. en mesure d'extrapoler en dehors de l'ensemble d'entraînement \mathcal{D} . Cette observation suit essentiellement le *no free lunch theorem* (théorème NFL) pour l'optimisation ([Wolpert et Macready, 1997](#)), et en particulier le théorème NFL pour l'apprentissage supervisé ([Wolpert, 1992, 1996](#)), qui stipule que tous les algorithmes d'apprentissage sont équivalents en termes du risque moyen encouru sur l'erreur de généralisation. Des biais inductifs sont nécessaires pour déterminer uniquement une hypothèse universelle, ou du moins une hypothèse qui satisfait certains critères déterminés *a priori*.

Dans le contexte du méta-apprentissage, la généralisation réfère plutôt au concept de transfert d'apprentissage, c.-à-d. le transfert des connaissances et le transfert de la structure du problème vers des tâches d'essais. Cette approche est plus appropriée dans un contexte scientifique, où la nature des hypothèses est constamment testée et jugée sur des critères comme leur capacité à reproduire et expliquer un phénomène, souvent mesurée par la simplicité de l'hypothèse. En ce sens, si l'hypothèse, f_θ , est une boîte noire comme un réseau de neurones, alors il y a très peu de moyens de soumettre cette hypothèse aux critères scientifiques requis pour convaincre une communauté sceptique. Le problème du méta-apprentissage se concerne plutôt à améliorer la méthode par laquelle l'hypothèse est obtenue, souvent par l'apprentissage explicite ou implicite de biais inductifs. Par exemple, une hypothèse peut être construite par régression symbolique de façon à satisfaire une communauté sceptique par la construction explicite d'un terme dans la fonction objective dont le rôle est de favoriser une hypothèse simple (e.g. [Lemos et al., 2022](#)). Les machines à inférences récurrentielles et

la plupart des méthodes de méta-apprentissage basées sur la rétropropagation de gradients (e.g. Finn et al., 2017) apprennent plutôt ces biais inductifs de façon implicite.

Strictement dans le contexte où on cherche à construire une seule hypothèse, et non une distribution d'hypothèses plausibles, il est possible de remplacer n'importe quel algorithme d'optimisation G_φ , qui sont aussi souvent considérés comme des boîtes noires, par une autre boîte noire $G_{\varphi_{\text{méta}}}$. L'hypothèse obtenue par l'algorithme d'optimisation $G_{\varphi_{\text{méta}}}$, qu'on peut noter comme f_θ ou encore $\hat{\mathbf{x}}^{(T)}$ dans le contexte des problèmes inverses, peut alors être soumise aux tests de validité requis par la communauté.

Le travail de Younger et al. (2001) est la première apparition concrète d'une approche pour le méta-apprentissage basée seulement sur la rétropropagation de gradients. Les travaux plus récents de Andrychowicz et al. (2016) démontrent empiriquement qu'une cellule récurrente à mémoire longue et courte (LSTM, Hochreiter et Schmidhuber, 1997) est en mesure d'apprendre un algorithme d'entraînement pour un second réseau de neurones. Contrairement à un réseau avec poids fixes, comme le perceptron (Rosenblatt, 1958), un réseau à cellules récurrentes (RNN) est Turing complet (Siegelmann et Sontag, 1992). Un RNN est en mesure de modifier son état interne, de sorte qu'un tel système peut représenter une classe d'algorithmes plus grande qu'un perceptron, incluant une descente de gradient tel que (4.29) et (4.31), soit la classe des algorithmes pouvant être représentés par les machines de Turing.

La contribution du travail de Putzky et Welling (2017) est d'appliquer ces méthodes pour les problèmes inverses, spécifiquement les problèmes inverses pour la reconstruction d'images. Dans ce contexte, la notion des biais inductifs est équivalente à celle d'une distribution *a priori* $p_\theta(\mathbf{x})$ sur l'espace des hypothèses. Le problème de méta-apprentissage devient

$$\varphi^* = \underset{\varphi}{\operatorname{argmin}} \mathbb{E}_{(\log p(\mathbf{y}|\mathbf{x}), \mathbf{x}) \sim \mathcal{D}} \left[\frac{1}{T} \sum_{t=1}^T \|\mathbf{x} - \hat{\mathbf{x}}^{(t)}\|_{\mathcal{X}}^2 \right]; \quad (4.35)$$

$$\text{ sujet à } \hat{\mathbf{x}}^{(t)} = \hat{\mathbf{x}}^{(t-1)} + g_\varphi(\hat{\mathbf{x}}^{(t-1)}, \nabla_{\hat{\mathbf{x}}^{(t-1)}} \log p(\mathbf{y} | \hat{\mathbf{x}}^{(t-1)})).$$

La fonction de vraisemblance, ou plus spécifiquement son gradient $\nabla_{\mathbf{x}} \log p(\mathbf{y} | \mathbf{x})$, encode le modèle physique ou la simulation des paramètres \mathbf{x} vers une observation \mathbf{y} . La fonction de vraisemblance nous permet d'intégrer nos connaissances sur la nature du phénomène à l'étude, p. ex. les équations (2.19) et (2.20) pour les lentilles gravitationnelles, directement dans le problème d'apprentissage. De cette façon, l'objectif de méta-apprentissage pour les paramètres φ est d'encoder seulement les connaissances manquantes aux problèmes, soit $\nabla_{\mathbf{x}} \log p_\theta(\mathbf{x})$, ou tout autre fonction de régularisation informative au problème de reconstruction.

Puisque les machines à inférences récurrentielles apprennent par méta-apprentissage, elles sont en mesure de généraliser beaucoup mieux qu'une fonction apprise par régression. En effet, le problème d'apprentissage, soit la relation (4.31), est transférable à la classe de problèmes pour laquelle les biais inductifs appris sont informatifs. Par exemple, Morningstar et al. (2019) ont découverts qu'une

machine à inférence récurrentielle entraînée à reconstruire des images de galaxies fortement lentillées est en mesure de reconstruire l'image d'une portion de paragraphe fortement lentillée, ce qui n'a rien à voir *a priori* avec des images de galaxies. Ce fait peut sembler contre-intuitif à première vue. On suppose toutefois que les biais inductifs appris par la machine à inférence récurrentielle ne sont pas simplement ceux qui permettent de reconstruire les exemples de l'ensemble d'entraînement. Plutôt, une machine à inférence récurrentielle apprend les biais inductifs transférables entre différents exemples de problème présentés en entraînement, soit précisément ce qui permet à ces machines de généraliser au-delà de leur ensemble d'entraînement.

Chapter 5

Pixelated Reconstruction of Foreground Density and Background Surface Brightness in Gravitational Lensing Systems using Recurrent Inference Machines

Alexandre Adam,^{1,2} Laurence Perreault-Levasseur,^{1,2,3} Yashar Hevazeh^{1,3}

¹*Département de physique, Université de Montréal, Montréal, H3C 3J7, Canada*

²*Mila - Quebec Artificial Intelligence Institute, Montréal, Canada*

³*Center for Computational Astrophysics, Flatiron Institute, 162 5th Avenue, 10010, New York, NY, USA*

Un résumé de cet article à été accepté à l'atelier *Machine Learning for Astrophysics Workshop at the Thirty-ninth International Conference on Machine Learning (ICML 2022)*.

Cet article sera soumis à la revue *The Astrophysical Journal* (ApJ) durant le prochains mois.

Résumé

Modéliser les lentilles gravitationnelles dans le but de quantifier les distorsions des images d’arrière-plan et de reconstruire la densité de masse de la lentille en avant-plan est encore aujourd’hui un problème difficile, posant un défi computationnel majeur. Avec le nombre croissant de lentilles découvertes et la résolution croissante des images de ces systèmes, la tâche d’exploiter complètement l’information qu’elles contiennent est présentement un problème hors d’atteinte pour les algorithmes traditionnels. Dans ce travail, on introduit un réseau neuronal récurrent basé sur les machines à inférence récurentielles (RIM) pour reconstruire simultanément une image non déformée de la source en arrière-plan et une image de la densité de masse de la lentille. La méthode que nous présentons reconstruit de façon itérative les paramètres du modèle (les pixels de la source et de la densité de la lentille) en apprenant le processus d’optimisation de la vraisemblance étant donné une observation et un modèle physique (une simulation des chemins lumineux), régularisée par des biais inductifs appris implicitement par le réseau de neurones avec les données d’entraînement. Comparée aux méthodes traditionnelles basées sur des modèles paramétriques de la densité de masse, notre approche est significativement plus expressive et peut reconstruire des distributions de masses complexes, ce qu’on démontre en utilisant des galaxies lentilles réalistes provenant de la simulation cosmologique hydrodynamique IllustrisTNG.

Mots-clés: Lentilles gravitationnelles — Simulations astrophysiques — Inférence non-paramétrique — Réseaux neuronaux convolutifs.

Abstract

Modeling strong gravitational lenses in order to quantify the distortions in the images of background sources and to reconstruct the mass density in the foreground lenses has been a difficult computational challenge. As the quality of gravitational lens images increases, the task of fully exploiting the information they contain becomes computationally and algorithmically more difficult. In this work, we use a neural network based on the Recurrent Inference Machine (RIM) to simultaneously reconstruct an undistorted image of the background source and the lens mass density distribution as pixelated maps. The method iteratively reconstructs the model parameters (the image of the source and a pixelated density map) by learning the process of optimizing the likelihood given the data using the physical model (a ray-tracing simulation), regularized by a prior implicitly learned by the neural network through its training data. When compared to more traditional parametric models, the proposed method is significantly more expressive and can reconstruct complex mass distributions, which we demonstrate by using realistic lensing galaxies taken from the IllustrisTNG cosmological hydrodynamic simulation.

Keywords: Gravitational lensing (670) — Astronomical simulations (1857) — Nonparametric inference (1903) — Convolutional Neural Networks (1938).

5.1 Introduction

Strong gravitational lensing is a natural phenomenon through which multiple, distorted images of luminous background sources are formed by the gravity of massive foreground objects along the line of sight (e.g., Vieira et al., 2013; Marrone et al., 2018; Rizzo et al., 2020; Sun et al., 2021). These distortions are tracers of the distribution of mass in foreground structures, irrespective of their light emission properties. As such, this phenomenon offers a powerful probe of the distribution of dark matter (e.g., Dalal and Kochanek, 2002; Treu and Koopmans, 2004; Hezaveh et al., 2016; Gilman et al., 2020, 2021).

Lens modeling is the process through which the parameters describing both the mass distribution in the foreground lens and the undistorted image of the background source are inferred. This has traditionally been done through explicit likelihood-based modeling methods, a time- and resource-consuming procedure. A common practice to model strong lenses is to model the light profile of the background source with a Sérsic (1963) profile and the density of the foreground lens with a power law function, $\rho \propto r^{-\gamma}$. These simple profiles allow for the exploration of their low-dimensional parameter space with non-linear samplers such as Marchov Chain Monte Carlo (MCMC) methods (e.g., Koopmans et al., 2006; Barnabè et al., 2009; Auger et al., 2010) and generally provide a good fit to low-resolution data. However, as high-resolution and high signal-to-noise ratio (SNR) images become available, lensing analysis with simple models requires the introduction of additional parameters representing the true complexity of the mass distribution in lensing galaxies and the complexity of surface brightness in the background sources (e.g., Sluse et al., 2017; Wong et al., 2017; Birrer et al., 2019; Rusu et al., 2020, 2017; Li et al., 2021). This approach becomes intractable as the complexity of the mass distribution and the quality of images increases (e.g., Schmidt et al., 2022). For example, no simple parametric model of the *Hubble Space Telescope* (*HST*) images of the Cosmic Horseshoe (J1148+1930) — initially discovered by Belokurov et al. (2007) — has been able to model the fine features of the extended arc (e.g., Bellagamba et al., 2016; Cheng et al., 2019; Schuldt et al., 2019).

Free-form methods attempt to relax the assumptions about the smoothness and symmetries of these parametric profiles using more expressive families like regular (or adaptive) grid representations and meshfree representations (Saha and Williams, 1997; Abdelsalam et al., 1998a,b; Diego et al., 2005; Birrer et al., 2015; Merten, 2016). These methods strive to model the signal contained in lensed images in a data-agnostic way, in order to place better constraints on the morphology of the source brightness or the projected mass density of the lens. However, most free-form parametrization choices make the inference problem under-constrained, meaning that imposing a prior on the reconstructed parameters becomes essential to penalize unphysical solutions and avoid overfitting the data.

In the context of traditional likelihood-based modeling, there exists a number of commonly used priors for the inference of high dimensional representations of background sources (e.g., imposing a

quadratic-log prior for linear inversion of pixellated-source models as developed by [Warren and Dye \(2003\)](#); [Suyu et al. \(2006\)](#) or iteratively specified priors for shapelets ([Birrer et al., 2015](#); [Birrer and Amara, 2018](#); [Nightingale et al., 2018](#))). However, these priors are often simplistic approximations to the actual distribution of pixel brightness in unlensed galaxies, and thus can result in important, difficult to characterize biases.

On the other hand, for lens mass reconstruction, the issue of specifying an appropriate prior is still unsolved. This has been studied extensively in the context of cluster-scale strong lensing ([Bartelmann et al., 1996](#); [Seitz et al., 1998](#); [Abdelsalam et al., 1998a,b](#); [Bradač et al., 2005](#); [Diego et al., 2005](#); [Cacciato et al., 2006](#); [Diego et al., 2007](#); [Liesenborgs et al., 2006, 2007](#); [Jee et al., 2007](#); [Coe et al., 2008](#); [Merten et al., 2009](#); [Deb et al., 2012](#); [Merten, 2016](#); [Ghosh et al., 2020](#); [Torres-Ballesteros and Castañeda, 2022](#)). Free-form approaches in the context of strong galaxy-galaxy lenses have been comparatively less studied (see however [Saha and Williams \(1997, 2004\)](#); [Birrer et al. \(2015\)](#); [Coles et al. \(2014\)](#)).

Another major challenge for these models is the issue of optimizing or sampling these high dimensional posteriors. Given the non-linear nature of the model and the existence of multiple local optima, non-linear global optimizers and samplers are needed, which often results in extremely expensive computational procedures. The high computational cost of these methods also limits the extent to which they can be thoroughly tested and validated to identify and characterize potential systematics.

Over the recent years, deep learning methods have proven extremely successful at accurate modeling of strong lensing systems ([Hezaveh et al., 2017](#); [Perreault Levasseur et al., 2017](#); [Morningstar et al., 2018](#); [Coogan et al., 2020](#); [Park et al., 2021](#); [Legin et al., 2021, 2022](#); [Wagner-Carena et al., 2021](#); [Schuldt et al., 2022](#); [Wagner-Carena et al., 2022](#); [Karchev et al., 2022](#); [Anau Montel et al., 2022](#); [Mishra-Sharma and Yang, 2022](#); [Schuldt et al., 2022](#)). More specifically, [Morningstar et al. \(2019\)](#) demonstrated that recurrent convolutional neural networks can implicitly learn complex prior distributions from their training data to successfully reconstruct pixelated undistorted images of strongly lensed sources, circumventing the need to explicitly specify a prior distribution over those parameters. Motivated by this, we propose a method that extends this framework to solve the full lensing problem and simultaneously reconstruct a pixelated mass map and a pixelated image of the undistorted background source.

The method we propose here is based on the Recurrent Inference Machine (RIM), originally developed by [Putzky and Welling \(2017\)](#). In its original version, this method proposed to solve inverse problems using a Recurrent Neural Network as a metalearner to learn the iterative process of the optimization of a likelihood. RIMs have been trained on a range of linear inverse problems both within and outside of astrophysics ([Lønning et al., 2019](#)). In [Modi et al. \(2021\)](#), this method was generalized to non-linear inference problems while using a U-net architecture ([Ronneberger et al., 2015](#)) to separate the dynamics of different scales.

In the present paper, we leverage this framework to learn an optimization process over the

highly non-convex strong lensing likelihood, and implicitly learn a data-driven prior, which allows for the reconstruction of complex mass distributions representative of realistic galaxies taken from the IllustrisTNG (Nelson et al., 2019) hydrodynamical simulations. We also introduce a fine-tuning procedure, which allows us to directly exploit the prior encoded in the neural network parameters in order to further optimize the posterior down to noise levels. We apply this to the reconstruction of high signal-to-noise galaxy-galaxy lensing systems simulated using IllustrisTNG (Nelson et al., 2019) projected density maps and background galaxy images collected from the COSMOS survey (Koekemoer et al., 2007; Scoville et al., 2007).

The paper is organised as follows. Section 5.2 details the inference pipeline. In Section 5.3, we present the data production and preprocessing for the training of the RIM and the generative models used in this paper. In Section 5.4, we report on the training strategies used. In Section 5.5, we discuss our results on a test set of gravitational lenses. We conclude in Section 5.6.

5.2 Methods

Our goal is to predict pixelated maps of both the undistorted image of the background source and the projected density in the foreground lens from noisy lensed images. Our model consists of a Recurrent Inference Machine that predicts these variables of interest. Training this model requires large number of training data, which we produce using a Variational Autoencoder (VAE) trained on density maps from the IluustrisTNG simulation and background sources from the Cosmos dataset (Section 5.4.1).

In this section, we present the structure of the lensing inference problem and provide information about our analysis method. We begin with a general introduction to maximum a posteriori (MAP) inference in Section 5.2.1. We describe the lensing simulation pipeline in Section 5.2.2. In Section 5.2.3, we motivate the use of the Recurrent Inference Machine and describe its computational graph. The architecture of the neural network is described in Section 5.2.4. Finally, we describe the fine-tuning procedure and the transfer learning technique applied to achieve noise-level reconstructions in Section 5.2.5.

5.2.1 Maximum a posteriori inference

We consider the task of reconstructing a vector of parameters of interest $\mathbf{x} \in \mathcal{X}$ given a vector of noisy observed data $\mathbf{y} \in \mathcal{Y}$, a known forward (or physical) model F , and an additive noise vector $\boldsymbol{\eta}$. In what follows, we assume this vector to be sampled from a Gaussian distribution with known covariance matrix C , such that we can write

$$\begin{aligned} \mathbf{y} &= F(\mathbf{x}) + \boldsymbol{\eta}; \\ \boldsymbol{\eta} &\sim \mathcal{N}(0, C). \end{aligned} \tag{5.1}$$

In our case study, F is a many-to-one non-linear mapping between the parameter space \mathcal{X} and the data space \mathcal{Y} . Finding physically allowed solutions for this ill-posed inverse problem requires strong priors. The maximum a posteriori (MAP) solution maximizes the product of the likelihood $p(\mathbf{y} | \mathbf{x})$ and the prior $p(\mathbf{x})$:

$$\hat{\mathbf{x}}_{\text{MAP}} = \underset{\mathbf{x} \in \mathcal{X}}{\operatorname{argmax}} \log p(\mathbf{y} | \mathbf{x}) + \log p(\mathbf{x}). \quad (5.2)$$

Assuming a Gaussian noise model for η , the log-likelihood can be written analytically as

$$\log p(\mathbf{y} | \mathbf{x}) \propto -(\mathbf{y} - F(\mathbf{x}))^T C^{-1} (\mathbf{y} - F(\mathbf{x})). \quad (5.3)$$

The prior distribution, however, is problem-dependent and encodes expert knowledge of the model domain. As such, it is typically harder to write explicitly.

5.2.2 The Forward Model

The forward model, F , is a simulation pipeline that receives a map of the surface brightness in the background source and a map of the projected density in the foreground lens to produce distorted images of background galaxies. This pipeline uses ray tracing to calculate the deflection angles, α , and maps the observed coordinates, θ , into the coordinates of the background plane, β , through the lens equation

$$\beta = \theta - \alpha(\theta). \quad (5.4)$$

The deflection angles are obtained using the projected surface density field κ — also referred to as convergence — through the integral

$$\alpha(\theta) = \frac{1}{\pi} \int_{\mathbb{R}^2} \kappa(\theta') \frac{\theta - \theta'}{\|\theta - \theta'\|^2} d^2\theta'. \quad (5.5)$$

The intensity of a pixel in a simulated observation is obtained through bilinear interpolation of the source brightness distribution at the coordinate β . Since we also use a discrete representation for the convergence, we approximate this integral by a discrete global convolution. Taking advantage of the convolution theorem, this operation can be computed in near-linear time using the Fast Fourier Transform algorithm (FFT).

Assuming the observation has M^2 pixels, the convolution kernel would have $(2M + 1)^2$ pixels. Both the convergence map and the kernel are zero-padded to a size of $(4M + 1)^2$ pixels in order to approximate a linear convolution and significantly reduce aliasing.

To produce simulated images, a blurring operator — convolution by a point spread function (PSF) — is applied to the lensed image to replicate the response of an optical system. This operator is implemented as a GPU-accelerated matrix operation since the blurring kernels used in this paper have a significant proportion of their energy distribution encircled inside a small pixel radius. Gaussian noise is then applied to the images, as described in more details in section 5.3.3.

5.2.3 Recurrent Inference Machine

Instead of handcrafting a prior distribution to solve the inverse problem (5.1), we build an inference pipeline with a data-driven implicit prior encoded in a deep neural network architecture (Bengio, 2009). The RIM (Putzky and Welling, 2017) is a form of learnt gradient-based inference algorithm, intended to solve inverse problems of the form (5.1). This framework has mainly been applied in the context of linear under-constrained inverse problems — i.e. where $F(\mathbf{x})$ can be represented as a matrix product $A\mathbf{x}$ — for which the prior on the parameters \mathbf{x} , $p(\mathbf{x})$, is either intractable or hard to compute (Morningstar et al., 2018, 2019; Lønning et al., 2019). The use of the RIM to solve non-linear inverse problems was first investigated in (Modi et al., 2021). In our case, the function representing the physical model F encodes the lens equation (5.4), which is highly non-linear.

The RIM is made up of a recurrent unit, which, given an observation \mathbf{y} , solves (5.1) for \mathbf{x} through the governing equation

$$\hat{\mathbf{x}}^{(t+1)} = \hat{\mathbf{x}}^{(t)} + g_\varphi(\hat{\mathbf{x}}^{(t)}, \mathbf{y}, \nabla_{\hat{\mathbf{x}}^{(t)}} \log p(\mathbf{y} | \hat{\mathbf{x}}^{(t)})), \quad (5.6)$$

where $\hat{\mathbf{x}}^{(t)}$ is the estimate of the parameters of interest at time t of the recursion (here, the pixel values of the image of the undistorted background source and of the density field κ) and g_φ is a neural network. In the text, we will often use the shorthand notation $\nabla_{\mathbf{y}|\mathbf{x}}$ to refer to $\nabla_{\mathbf{x}} \log p(\mathbf{y} | \mathbf{x})$, the gradient of the likelihood evaluated at \mathbf{x} . By minimizing a weighted mean squared loss backpropagated through time,

$$\mathcal{L}_\varphi(\mathbf{x}, \mathbf{y}) = \frac{1}{T} \sum_{t=1}^T \sum_{i=1}^M \mathbf{w}_i (\hat{\mathbf{x}}_i^{(t)} - \mathbf{x}_i)^2, \quad (5.7)$$

where T is the total number of time steps in the recursion, the index i labels the pixels of the reconstructions, \mathbf{w}_i is the per-pixel weight, and M is the total number of pixels in the reconstructions, the neural network g_φ learns to optimize the parameters \mathbf{x} given a likelihood function. The converged parameters of the neural network given the training set \mathcal{D} , $\varphi_{\mathcal{D}}^*$, are those that minimize the cost — or empirical risk — which is defined as the expectation of the loss over \mathcal{D}

$$\varphi_{\mathcal{D}}^* = \operatorname{argmin}_{\varphi} \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}} [\mathcal{L}_\varphi(\mathbf{x}, \mathbf{y})]. \quad (5.8)$$

Unlike previous works (Andrychowicz et al., 2016; Putzky and Welling, 2017; Morningstar et al., 2018, 2019; Lønning et al., 2019), the data vector \mathbf{y} containing the observations is fed to the neural network in order to learn a better initialization of the parameters, $\mathbf{x}^{(0)} = g_\varphi(0, \mathbf{y}, 0)$, in addition to their optimization process. Empirically, we found that this significantly improves the performance of the model for our problem and avoids situations where the model would get stuck in local minima at test time due to poor initialization.

We follow previous works in setting a uniform weight over the time steps ($\mathbf{w}^{(t)} = \frac{\mathbf{w}}{T}$). The

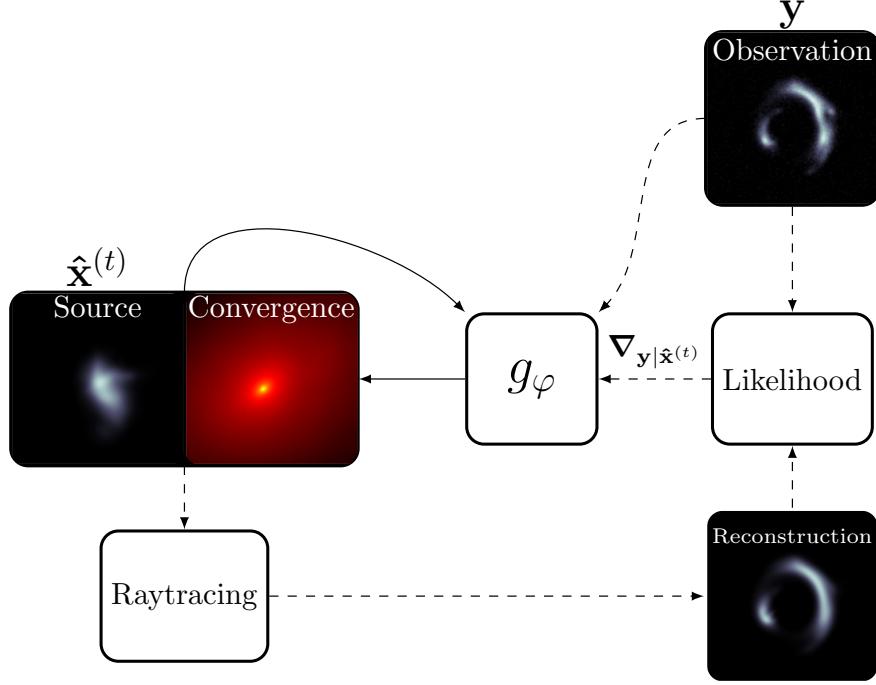


Figure 5.1 – Rolled computational graph of the RIM. Dashed arrows represent operations not recorded for BPTT.

choice of the pixel weights \mathbf{w}_i is informed by our empirical observations when training the network. Details are reported in appendix D.

In Figure 5.1, we show the rolled computational graph of the RIM. During training of the neural network g_φ , operations along the solid arrows are being recorded for backpropagation through time. The recording is stopped along the dashed arrow since these operations are part of the forward modelling process and contain no trainable parameters.

The gradient of the likelihood is computed using automatic differentiation. Following (Modi et al., 2021), we preprocess the gradients using the Adam algorithm (Kingma and Ba, 2014). For clarity, we only illustrate this step in Figure 5.2.

5.2.4 The Neural Network

The neural network architecture is illustrated in Figure 5.2, which shows a single time step of the unrolled computation graph of the RIM. We use a U-net (Ronneberger et al., 2015) architecture with Gated Recurrent Units (GRU: Cho et al., 2014) placed in each skip connection.

Each GRU cell has its own memory tensor that is updated through time at each iteration of equation 5.6. The shape of a memory tensor is set to match the feature tensor fed into it from the parent layer in the network graph. Instead of learning a compressed representation like in

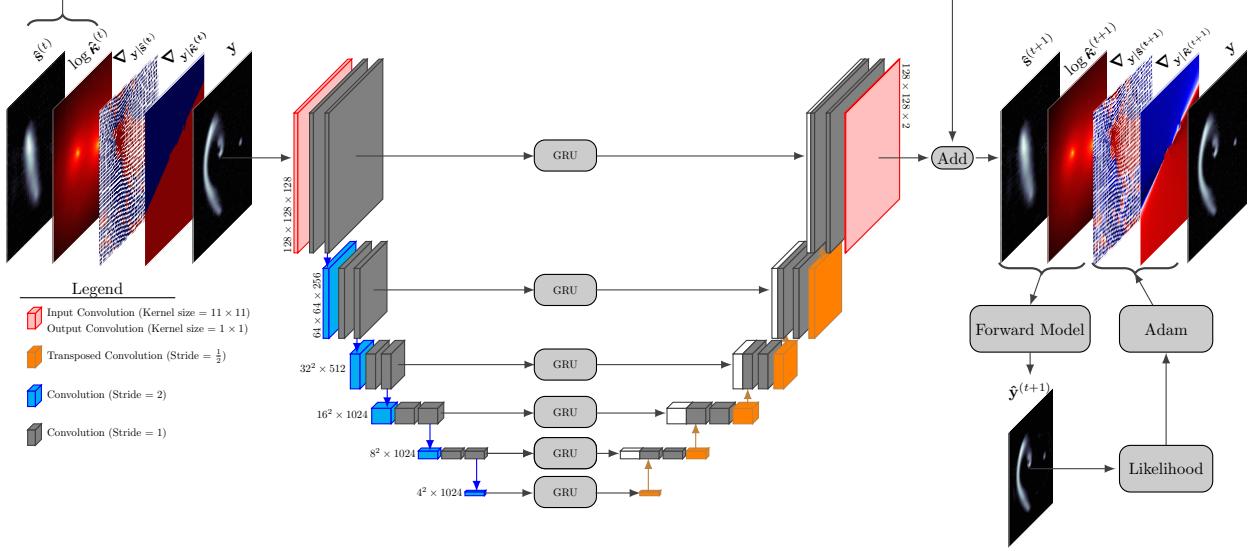


Figure 5.2 – A single time step of the unrolled computation graph of the RIM. GRU units are placed in the skip connections to guide the reconstruction of the source and convergence. A schematic of the steps to compute the likelihood gradients is shown in the bottom right of the figure, including the Adam processing step of the likelihood gradient.

the hourglass architecture (or autoencoder), the U-net architecture naturally separates the spatial frequency components of the signal into its vertical levels. The first level generally encodes high frequency features while the lower levels encode low frequency features (due to downsampling of the feature maps). Adding an independent memory unit at each level preserve this property.

Convolutional layers with a stride of 2 are used for downsampling and stride of $\frac{1}{2}$ for upsampling of the feature maps (identified in blue and orange respectively in figure 5.2). Half-stride convolutions are implemented in practice with the transposed convolution layers from [Tensorflow](#) ([Abadi et al., 2015](#)). Most layers use a kernel size of 3×3 , except the first and last layer. The first layer has larger receptive field (11×11) in order to capture more details in the input tensor. The last layer has kernels of size 1×1 . A tanh activation function is used for each convolutional layer, including strided convolutions, except for the output layer. The U-net outputs an image tensor with two channels, one dedicated for the update of the source and the other for the update of the convergence (see figure 5.2).

5.2.5 Fine-Tuning

Objective function

Once trained, the RIM produces a baseline (point) estimate of the parameters \mathbf{x} given a noisy observation \mathbf{y} , a PSF and a noise covariance matrix. We now concern ourselves with a strategy to improve this estimate. This is important for observations with high SNR, for which the estimate

must be very accurate to model all the fine features present in the arcs. The metric for the goodness of fit is the reduced chi squared $\chi_{\nu}^2 = \frac{\chi^2}{\nu}$, where ν is the total number of degrees of freedom which here corresponds to the total number of pixels in \mathbf{y} . Generally, our goal will be to reach $\chi_{\nu}^2 = 1$, or equivalently $|\chi^2 - \nu| = 0$, which suggests that the RIM's estimate has modeled all the signal to be recovered from the observations. We note that such a problem is exceedingly difficult at high SNR.

We note that we can optimize the log-likelihood directly w.r.t. the network weights given an appropriate prior on those weights (to avoid forgetting the implicit priors that have been learned during training, see section 5.2.5). The new objective function is given by

$$\hat{\varphi}_{\text{MAP}} = \underset{\varphi}{\operatorname{argmax}} \frac{1}{T} \sum_{t=1}^T \log p(\mathbf{y} \mid \hat{\mathbf{x}}^{(t)}) + \log p(\varphi), \quad (5.9)$$

where φ are the network weights, $\log p(\mathbf{y} \mid \hat{\mathbf{x}}^{(t)})$ is the log-likelihood, and $\log p(\varphi)$ is the log prior over the network weights. Unlike the loss in equation (5.7), this objective function makes no use of labels (\mathbf{x}). This allows us to use equation (5.9) at test time in order to fine-tune the RIM's weights to a specific test example.

Transfer Learning

We now address the issue of transferring knowledge from the training task defined by the loss function in equation (5.8), to a test task specific to an observation, as defined by the loss given in equation (5.9). The reader might refer to reviews on transfer learning ([Pan and Yang, 2010](#); [Zhuang et al., 2019](#)) for a broad overview of the field. The strategy we outline falls within the category of inductive transfer learning.

Optimizing the log-likelihood alone without a prior term over the weights (i.e. just the first term from the r.h.s. in (5.9)) by initializing the weights at $\varphi_{\mathcal{D}}^*$ is not strong enough to preserve the knowledge learned from the training task. This has long been observed in the literature and was coined as the catastrophic interference phenomenon in connectionist networks ([McCloskey and Cohen, 1989](#); [Ratcliff, 1990](#)). In summary, a sequential learning problem exhibits catastrophic forgetting of old knowledge when confronted with new examples (possibly from a different distribution or process), in a manner

1. proportional to the amount of learning;
2. strongly dependant to the disruption of the parameters involved in representing the old knowledge.

While introducing an early stopping condition could potentially alleviate the former issue, the latter could still remain a problem.

We therefore follow the work of [Kirkpatrick et al. \(2016\)](#) to define a prior distribution over φ

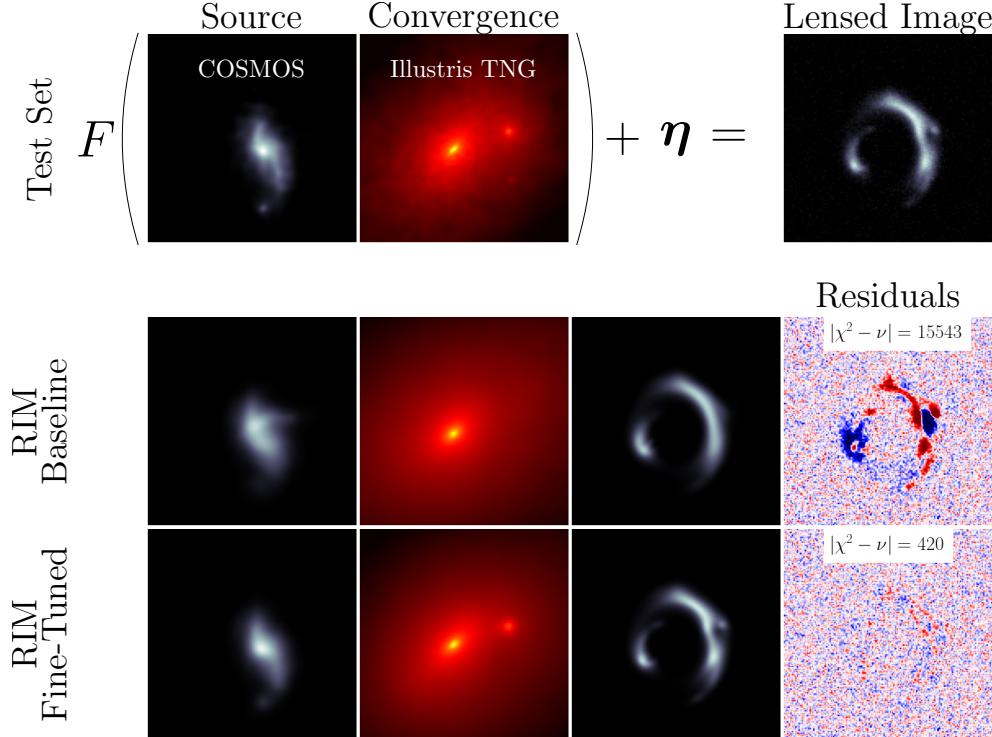


Figure 5.3 – Example of a simulated lensed image in the test set that exhibits a large deflection in its eastern arc which indicates the presence of a massive object — in this case a dark matter subhalo. The fine-tuning procedure is able to recover this subhalo because of its strong signal in the lensed image and reduces the residuals to noise level.

that address this issue

$$\log p(\varphi) \propto -\frac{\lambda}{2} \sum_j \text{diag}(\mathcal{I}(\varphi_D^*))_j (\varphi_j - [\varphi_D^*]_j)^2. \quad (5.10)$$

where $\text{diag}(\mathcal{I}(\varphi_D^*))$ is the diagonal of the Fisher information matrix encoding the amount of information that some set of gravitational lensing systems from the training set, and similar to the observed test task, carries about the baseline RIM weights φ_D^* — the parameters that minimize the empirical risk (equation 5.8). We can also understand this prior using the Cramér-Rao lower bound (Rao, 1945; Cramér, 1946). The prior can thus be framed as a multivariate Gaussian distribution characterised by a diagonal covariance matrix with $\text{diag}(\mathcal{I})$ as its inverse and by φ_D^* as its first moment. Within this view, the Lagrange multiplier is tuning our estimated uncertainty about the neural network weights for the particular task at hand. We have included a derivation of this term in the appendix B.

Examples are drawn from the set of training examples similar to the test task by sampling the latent space of two variational autoencoders (VAE) that model a distribution over the background sources and the convergence maps respectively (as described in Section 5.3.1 and 5.3.2) near the

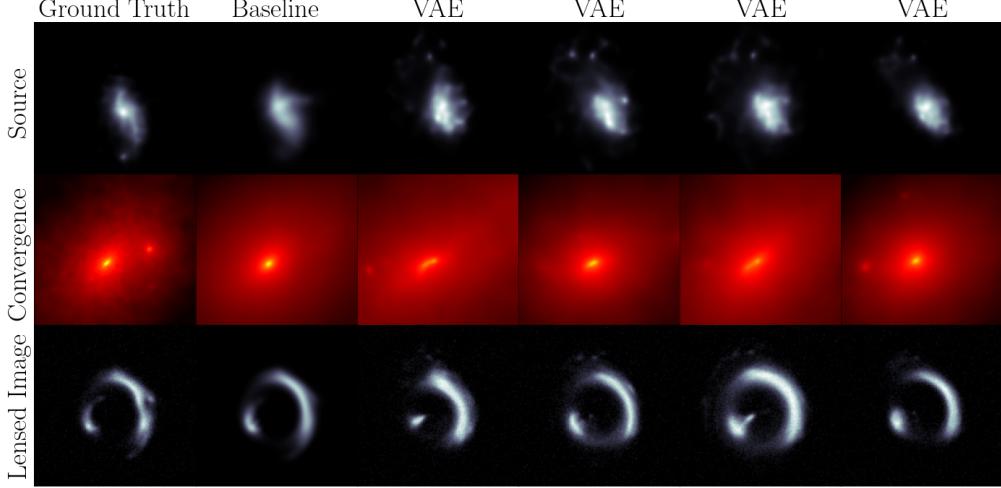


Figure 5.4 – Examples similar to the test task, also shown in Figure 5.3. The first column shows the ground truth used to simulate the lensed image. The second column shows the baseline prediction that is then encoded in the latent space of the VAE in order to sample the next 4 columns.

baseline prediction of the RIM. In practice, we choose an isotropic Gaussian distribution centered around $\hat{\mathbf{z}}^{(T)}$ — the latent code of the baseline prediction — as a sampling distribution. While we leave the possibility of improving this choice to future work, it is sufficient for our goals. Figure 5.4 illustrates examples of what is meant here by *similar*.

5.3 Data

5.3.1 COSMOS

The maps of surface brightness of background sources are taken from the *Hubble Space Telescope* (*HST*) Advanced Camera for Surveys Wide Field Channel COSMOS field (Koekemoer et al., 2007; Scoville et al., 2007), a 1.64 deg^2 contiguous survey acquired in the F814W filter. A dataset of magnitude limited ($\text{F814W} < 23.5$) deblended galaxy postage stamps (Leauthaud et al., 2007) was compiled as part of the GREAT3 challenge (Mandelbaum et al., 2014). The data is publicly available (Mandelbaum et al., 2012), and the preprocessing is done through the open-source software **GALSIM** (Rowe et al., 2015).

We apply the `marginal` selection criteria (see the `COSMOSCatalog` class) and impose a flux per image greater than 50 photons $\text{cm}^{-2} \text{s}^{-1}$. This final set has a total of 13 321 individual images. Each image is saved as a postage stamp of 158^2 pixels. We then subtract the background from each image, apply a random shift, rotate them by an angle multiple of 90° , crop them down to 128^2 pixels, and finally normalize them to pixel intensities in the range $[0, 1]$. We then train an autoencoder to denoise the galaxy images (Vincent et al., 2008, 2010). More specifically, we use the informational bottleneck principle (Tishby et al., 2000) to learn a lossy lower-dimensional representation of the

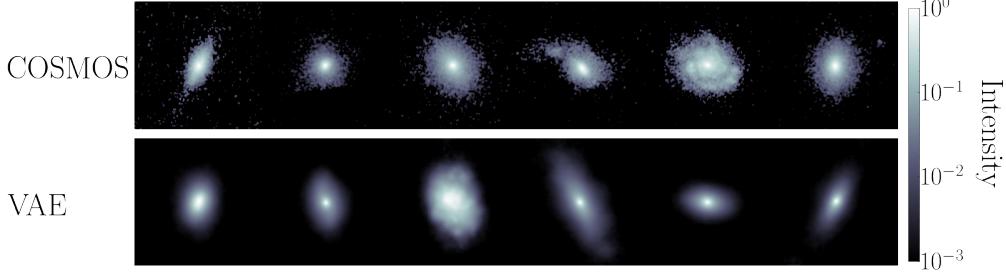


Figure 5.5 – Examples of COSMOS galaxy images (top row) and VAE generated samples (bottom row) used as labels in \mathcal{D} .

data. For a generic CNN autoencoder, this amounts to learning a low-pass frequency filter on the COSMOS dataset. Indeed, CNNs are known to exhibit a spectral bias in their learning phase (Rahaman et al., 2018), which we exploit to our advantage in order to filter pixel noise from the galaxy surface brightness. Furthermore, using an expressive CNN autoencoder produces much less artifacts than a naive implementation of such a low-pass filter — e.g. by masking Fourier modes.

We split the galaxies into a training set (90%) and a test set (10%). The augmented training set ($\sim 50\,000$ images) is then used to train a VAE, as described in Section 5.4.1, and produce simulated observations to train the RIM.

5.3.2 IllustrisTNG

Smooth Particle Lensing

To compute convergence maps from an N-body simulation, we use Kernel Density Estimation to produce smooth densities on a regular grid from discrete simulation particles. This reduces the particle noise affecting all important lensing quantities. At the same time, the choice of the kernel size is important to preserve substructures in the lens that we might potentially be interested in. Following Aubert et al. (2007); Rau et al. (2013), we use Gaussian smoothing with an adaptive kernel size determined by the distance of the 64th nearest neighbours of a given particle $D_{64,i}$.

$$\kappa(\mathbf{x}) = \frac{1}{\Sigma_{\text{crit}}} \sum_{i=1}^{N_{\text{part}}} \frac{m_i}{2\pi\hat{\ell}_i^2} \exp\left(-\frac{1}{2} \frac{(\mathbf{x} - \mathbf{x}_i)^2}{\hat{\ell}_i^2}\right) \quad (5.11)$$

$$\hat{\ell}_i = \sqrt{\frac{103}{1024}} D_{64,i}.$$

The nearest neighbours are found by fitting a k-d tree — implemented in `scikit-learn` (Pedregosa et al., 2011) — to the N_{part} particles in a cylinder centered on the centre of mass of the halo of interest. The critical surface density is defined as

$$\Sigma_{\text{crit}} = \frac{4\pi G}{c^2} \frac{D_\ell D_{\ell s}}{D_s}, \quad (5.12)$$

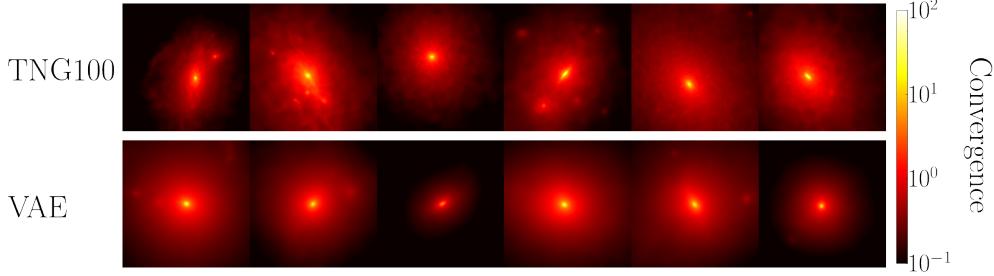


Figure 5.6 – Examples of smoothed Illustris TNG100 convergence map (top row) and VAE generated samples (bottom row) used as labels in \mathcal{D} .

where D_ℓ , D_s and $D_{\ell s}$ are angular diameter distances to the lens, source and between the lens and the source respectively, G is the gravitational constant, and c the speed of light.

Preprocessing

The projected surface density maps (convergence) of lensing galaxies were made using the redshift $z = 0$ snapshot of the IllustrisTNG-100 simulation ([Nelson et al., 2019](#)) in order to produce physically realistic realizations of density maps containing dark and baryonic matter. We selected 1604 halos with the criteria that they have a total dark matter mass of at least $9 \times 10^{11} M_\odot$. We then collected all dark matter, gas, stars and black holes particles from the data in the vicinity of the halo. We then create a smooth projected surface density map as prescribed in section 5.3.2.

We adopt the Λ CDM cosmology from [Planck Collaboration \(2020\)](#) with $h = 0.68$ to compute angular diameter distances. We also fix the source redshift to $z_s = 1.5$ and the deflector redshift to $z_\ell = 0.5$. We note that changing the redshifts or the cosmology only amount in a rescaling of the κ map by a global scalar. Thus, this choice does not change the generality of our method. The smoothed density maps from equation (5.11) are rendered into a regular grid of 188^2 pixels with a comoving field of view of $105 \text{ kpc}/h$. To avoid edge effects in the pixelated maps, we include particles outside of the field of view in the sum of equation (5.11).

Before applying augmentation or considering different projections, our dataset of halos is split into a training set (90%) and a test set (10%), in order to make sure that the test set consists only of convergence maps unseen by the RIM during training. We take 3 different projections (xy , xz and yz) of each 3D particle distribution, which amounts to a dataset with a total of 4812 individual convergence maps. Random rotations by an angle multiple of 90° and random shifts to the pixel coordinates are applied to each image. The κ maps are then rescaled by a random factor to change their estimated Einstein radius to the range $[0.5, 2.5]$ arcseconds. The Einstein radius is defined as

$$\theta_E = \sqrt{\frac{4GM(\theta_E)}{c^2} \frac{D_{\ell s}}{D_\ell D_s}} \quad (5.13)$$

where $M(\theta_E)$ is the mass enclosed inside the Einstein radius. In practice, we estimate this quantity

by summing over the mass of pixels with a value greater than the critical density ($\kappa > 1$). For data augmentation purposes, this procedure gives a good enough estimate of the lensed image separation resulting from a given κ map. We test multiple scaling factors for each κ map, then uniformly sample between those that produce an estimated Einstein radius within the desired range. This step is used to remove any bias in the Einstein radius that might come from the mass function of the simulation.

The final maps are cropped down to 128^2 pixels. Placed at a redshift $z_\ell = 0.5$, a κ map will thus span an angular field of view of $7.69''$ with a resolution similar to *HST*. With these augmentation procedures, a total of 50 000 maps are created from the training split to train a VAE, as described in Section 5.4.1, and produce simulated observations to train the RIM.

5.3.3 Simulated Observations

Having defined a source map and a convergence map, we apply the ray tracing simulation described in section 5.2.2 to produce a lensed image.

For each lensed image, a Gaussian PSF is created with a full width at half maximum (FWHM) randomly generated from a truncated normal distribution. The support of the distribution is truncated below by the angular size of a single pixel and above by the angular size of 4 pixels. White noise with a standard deviation randomly generated from a truncated normal distribution is then added to the convolved lensed image to simulate noisy observations. These noise realizations result in SNRs between 10 and 1000. For simplicity, we define $\text{SNR} = \frac{1}{\sigma}$. This definition is equivalent to the peak signal-to-noise ratio.

To ensure that the images are representative of strongly lensed source, we require a minimum flux magnification of 3. We also make sure that most pixel coordinates in the image plane are mapped inside the source coordinate system through the lens equation (5.4).

Table 5.1 – Physical model parameters.

Parameter	Distribution/Value
Lens redshift z_ℓ	0.5
Source redshift z_s	1.5
Field of view ('')	7.69
Source field of view ('')	3
PSF FWHM ('')	$\mathcal{T}\mathcal{N}(0.06, 0.3; 0.08, 0.05)$ ¹
Noise amplitude σ	$\mathcal{T}\mathcal{N}(0.001, 0.1; 0.01, 0.03)$

In total, 400 000 training observations are simulated from random pairs of COSMOS sources and IllustrisTNG convergence maps in order to train the RIM. An additional 200 000 observations are created from pairs of COSMOS sources and pixelated SIE convergence maps. The parameters for these κ maps are listed in table 5.2.

We generate 1 600 000 simulated observations from the VAE background sources and convergence

Table 5.2 – SIE parameters.

Parameter	Distribution
Radial shift ('')	$\mathcal{U}(0, 0.1)$
Azimuthal shift	$\mathcal{U}(0, 2\pi)$
Orientation	$\mathcal{U}(0, \pi)$
θ_E ('')	$\mathcal{U}(0.5, 2.5)$
Ellipticity	$\mathcal{U}(0, 0.6)$

maps as part of the training set. We apply some validation checks to each example in order to avoid configurations like a single image of the background source or an Einstein ring cropped by the field of view.

5.4 Training

5.4.1 VAE

Here, we describe the training of two VAEs that are used to produce density maps and images of unlensed background galaxies to train and test our inference model. For an introduction to VAEs we refer the reader to [Kingma and Welling \(2019\)](#).

As mentioned in [Kingma and Welling \(2019\)](#), direct optimisation of the ELBO loss can prove difficult because the reconstruction term $\log p_\theta(\mathbf{x} | \mathbf{z})$ is relatively weak compared to the Kullback Leibler (KL) divergence term. To alleviate this issue, we follow the work of [Bowman et al. \(2015\)](#) and [Kaae Sønderby et al. \(2016\)](#) in setting a warm-up schedule for the KL term, starting from $\beta = 0.1$ up to β_{\max} .

Usually, $\beta_{\max} = 1$ is considered optimal since it matches the original ELBO objective derived by [Kingma and Welling \(2013\)](#). However, we are more interested in the sharpness of our samples and accurate inference around small regions of the latent space for fine-tuning. Thus, setting $\beta_{\max} < 1$ allows us to increase the size of the information bottleneck (i.e. latent space) of the VAE and improve the reconstruction cost of the model. This is a variant of the β -VAE ([Higgins et al., 2017](#)), where $\beta > 1$ was found to improve disentangling of the latent space ([Burgess et al., 2018](#)).

The value for β_{\max} and the steepness of the schedule are grid searched alongside the architecture for the VAE. These values are found in practice by manually looking at the quality of generated samples for different VAE hyperparameters. A similar method is explored and formalized in the InfoVAE framework ([Zhao et al., 2017](#)).

A notable element of the VAE architecture is the use of a fully connected layer to reshape the features of the convolutional layer into the chosen latent space dimension. Following the work of [Lanusse et al. \(2021\)](#), we introduce an ℓ_2 penalty between the input and output of the bottleneck dense layers to encourage an identity mapping. This regularisation term is slowly removed during

training.

5.4.2 RIM

The architecture of the neural network was grid searched on a smaller dataset ($\lesssim 10\,000$ examples) in order to quickly identify a small set of valid hyperparameters. Then, the best hyperparameters were identified using a two-stage training process on the training dataset. In the first stage, we trained 24 different architectures from this small hyperparameter set for approximately 4 days (wall time using a single Nvidia A100 GPU). Different architectures would have a training time much longer than others, and this was factored in the architecture selection process. For example, adding more time steps (T) to the recurrent relation (5.6) would yield better generalisation on the test set, but this would come at great costs to training time until convergence.

Following this first stage, 4 architectures were deemed efficient enough to be trained for an additional 6 days. We only report the results for the best architectures out of these 4.

Each reconstruction is performed by fine-tuning the baseline model on a test task composed of an observation vector, a PSF, and a noise covariance. In practice, fine-tuning predictions on the test set of 3 000 examples can be accomplished in parallel so as to be done in at most a few days by spreading the computation on ~ 10 Nvidia A100 GPUs. Each reconstruction uses at most 2000 steps, corresponding to approximately 20 minutes (wall-time) per reconstruction. Early stopping is applied when the χ^2 reaches noise level. The hyperparameters for this procedure are reported in Table 5.3.

Table 5.3 – Hyperparameters for fine-tuning the RIM.

Parameter	Value
Optimizer	RMSProp
Learning rate	10^{-6}
Maximum number of steps	2 000
λ	2×10^5
ℓ_2	0
Number of samples from VAE	200
Latent space distribution	$\mathcal{N}(\mathbf{z}^{(T)}, \sigma = 0.3)^2$

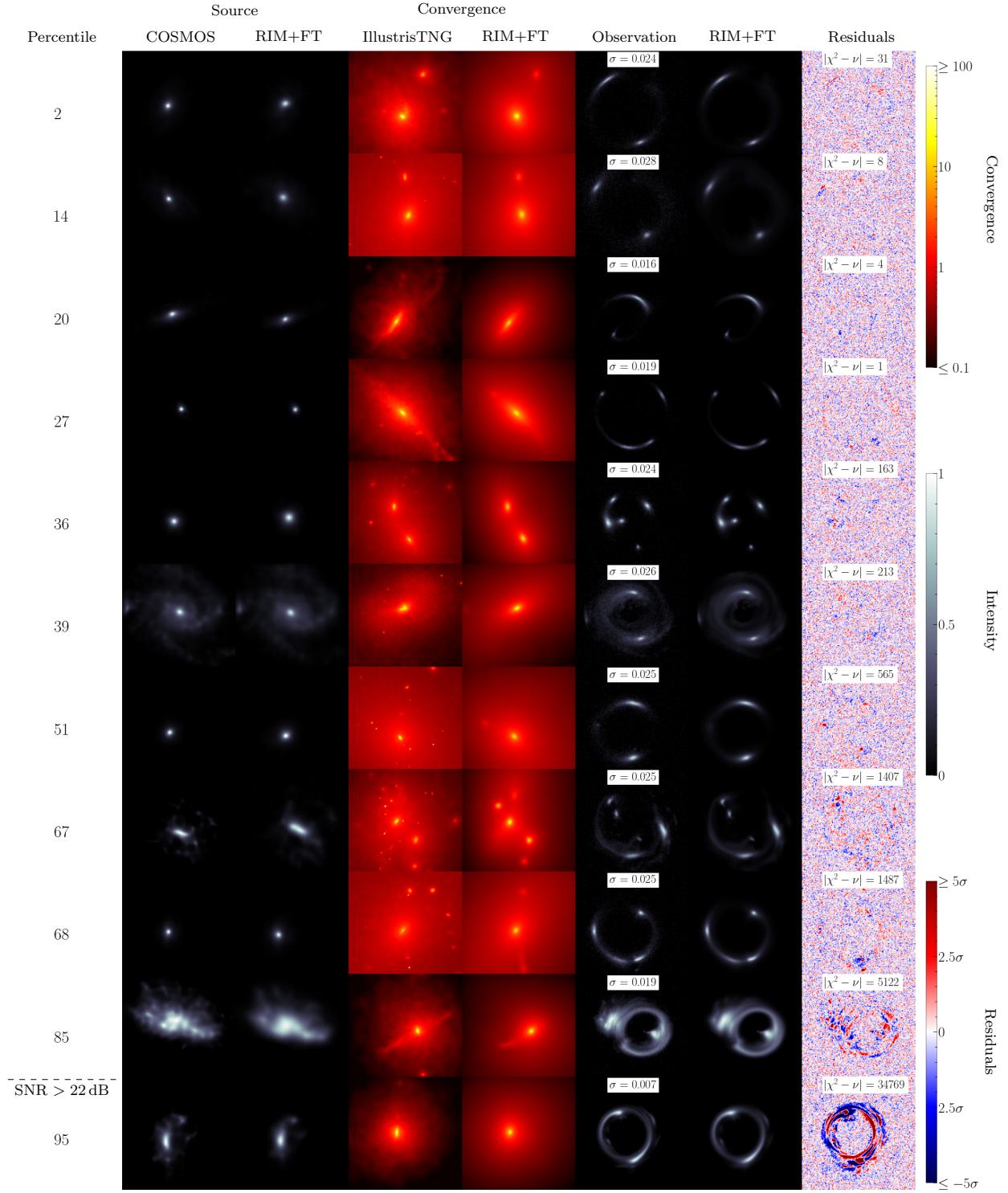


Figure 5.7 – Sample of the fine-tuned RIM reconstructions on a test set of 3000 examples. Examples are ordered from the best χ^2 (top) to the worst (bottom). The percentile rank of each example is in the leftmost column. The last example shown has SNR above the threshold defined in Figure 5.10.

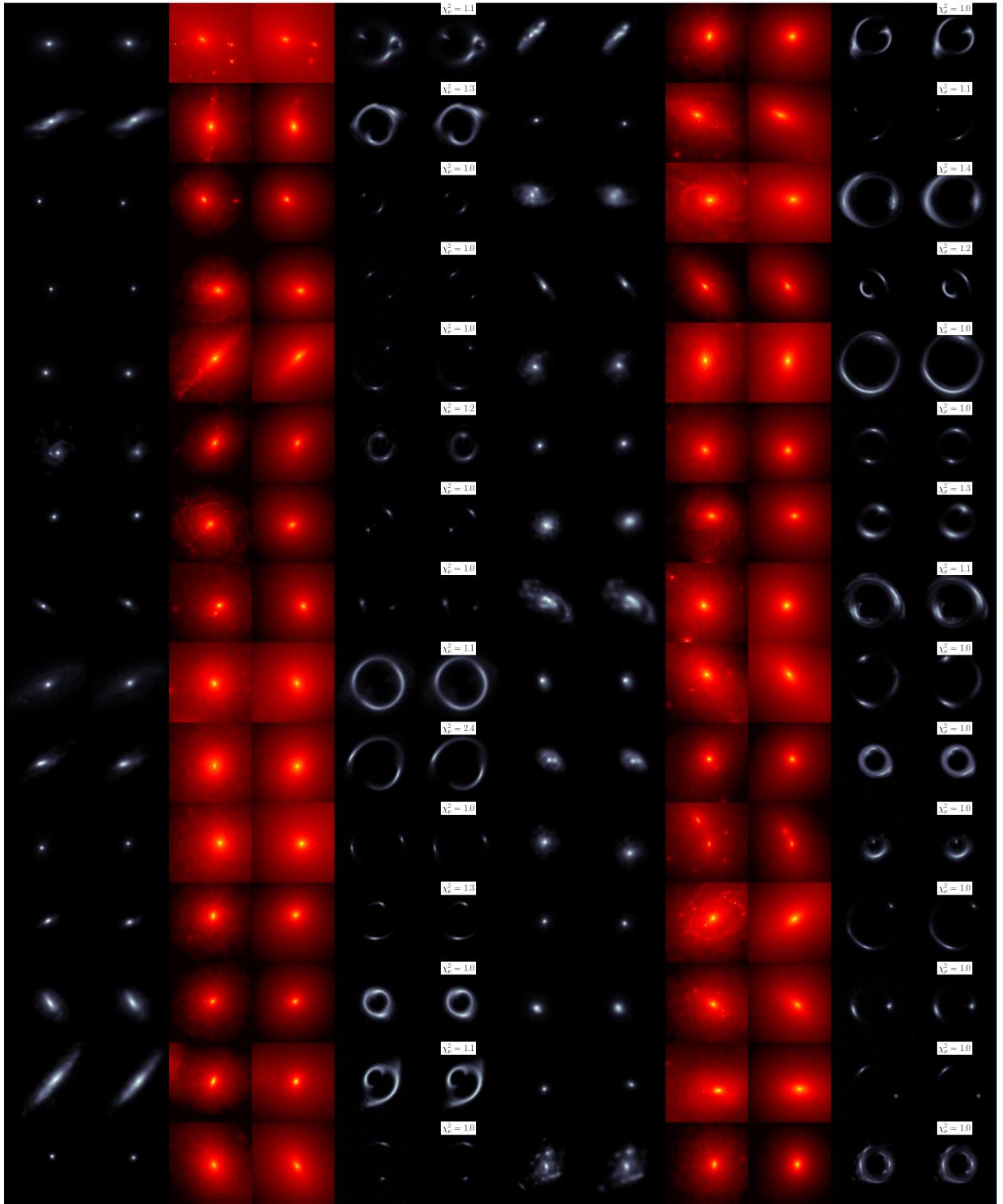


Figure 5.8 – 30 reconstructions taken at random from the test set of 3000 examples simulated from COSMOS and IllustrisTNG data at high SNR. The colorscale are the same as in Figure 5.7.

5.5 Results

In this section, we present the performance of our model on the held out test set. A sample of 3000 reconstruction problems is generated from the held-out *HST* and IllustrisTNG data with noise levels and PSFs similar to the training set.

5.5.1 Goodness of Fit

Figure 5.7 shows a sample of reconstructions for high SNR data with a wide range of lensing configurations from the test set. We select examples representative of all levels of reconstruction performance (covering the entire range of goodness of fit) for data with complex structures in their convergence map to showcase the expressivity of the approach. We also show a randomly selected sample from the test set in Figure 5.8.

Table 5.4 – \log_{10} -normal moments of the loss on the test set

Model	$\mu(\log \mathcal{L}_\varphi)$	$\sigma(\log \mathcal{L}_\varphi)$
Baseline (φ_D^*)	-1.96	0.36
Fine-tuned ($\hat{\varphi}_{MAP}$)	-2.02	0.37

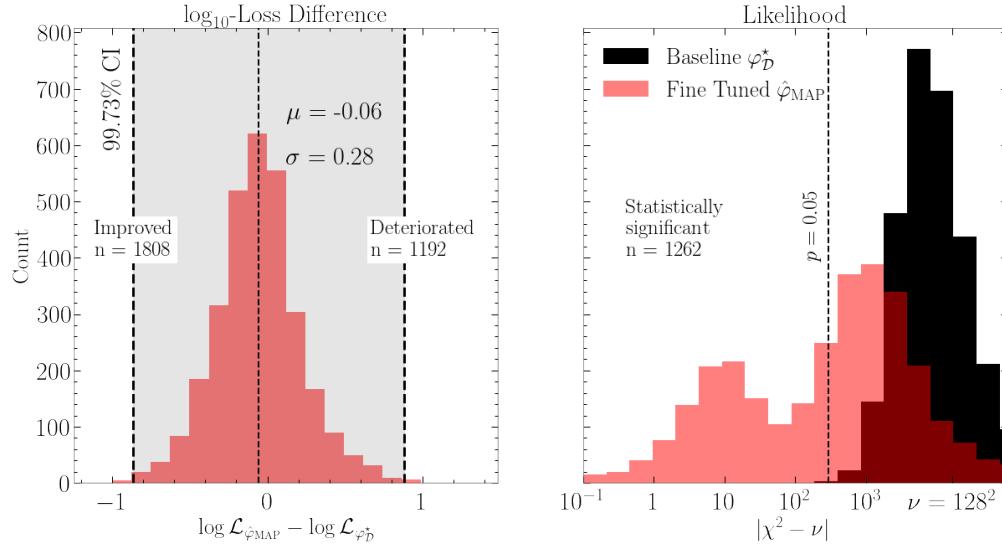


Figure 5.9 – Distribution of the goodness of fit for the baseline and fine-tuned network (right panel), as well as log-loss difference between the two network for a given example in the test set (left panel).

Figure 5.9 shows a comparison between the goodness of fit of the baseline model and the fine-tuned prediction. Since we empirically observe that the distribution of the loss on the test set (and the training set) follows a log-normal distribution, we find that it is more informative to look at the log-loss distribution to extract information about the fine-tuning procedure. The left panel

of Figure 5.9 shows the distribution of the log-loss difference between the fine-tuned prediction and the baseline model. This distribution shows that the fine-tuning procedure loss is constrained within ~ 1 order of magnitude of the original loss with a probability $> 99.73\%$. We find that the log-loss difference has a scatter of $\sigma = 0.28$, which is smaller than the scatter of the baseline log-loss over the entire test set $\sigma(\log \mathcal{L}_{\varphi_D^*}) = 0.36$ reported in Table 5.4. We note that the loss is not optimized during fine-tuning, still we notice that the fine-tuning procedure does not significantly deteriorate or improve the loss of the baseline prediction on average. We report the first 2 moments of the loss log-normal distribution for the baseline and the fine-tuned reconstructions in Table 5.4 in order to explicitly compare them. As can be seen in this table, there is no significant difference between the two distributions. This statement can be proven for the measured mean values — $\mu(\log \mathcal{L}_{\varphi_{\text{MAP}}}) = \mu(\log \mathcal{L}_{\varphi_D^*})$ — using the two-sided normal p-value test (Casella and Berger, 2001), which we find satisfy the null hypothesis with $p = 0.87288$ ($Z = -0.16$). All those observations support our claim that EWC regularisation preserves the prior learned during pretraining, or at least that it preserves the surrogate measures of the prior we reported.

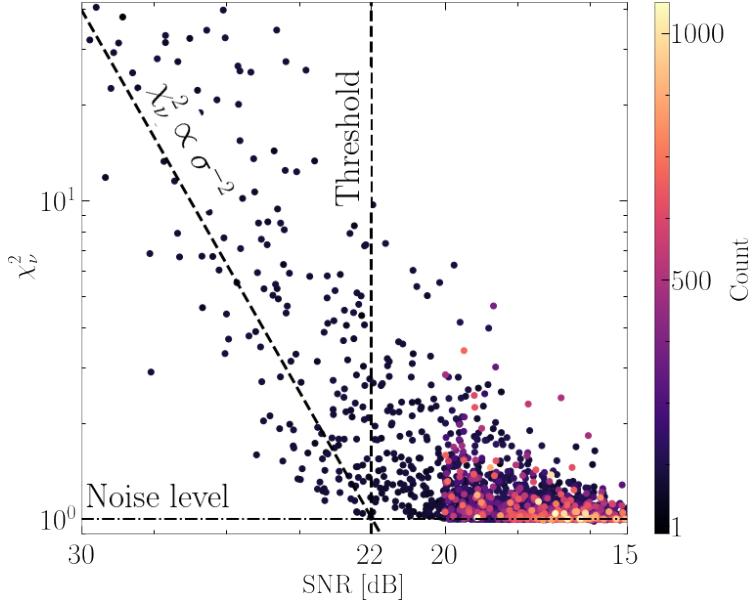


Figure 5.10 – Goodness of fit as a function of SNR shows a threshold behavior where our method reaches its limit.

The right panel of Figure 5.9 shows the distribution of χ^2 for the test set before and after the fine-tuning procedure and the theoretical χ^2 distribution corresponding to $\nu = 128^2$ degrees of freedom. We observe that the fine-tuning procedure significantly improves our χ^2 , bringing their distribution closer to that of the expected χ^2 distribution (black curve). However, the improved distribution is still far from the theoretical expectation, implying that there are statistically significant residuals in a subset of the reconstructions.

In figure 5.10, we explore how the goodness of fit of the fine-tuned RIM changes as a function of

SNR over the examples in the test set. Two behaviors can be identified. For SNR below a certain threshold, the goodness of fit of the fine-tuned model is essentially flat, with a certain scatter, around the noise level. This scatter increases as a function of SNR, which reflects the fact that above a certain SNR threshold (vertical dashed line in Figure 5.10), our reconstructions are dominated by systematics in the inference algorithm. For SNR above the threshold, the goodness of fit follows the trend $\chi^2 \propto \sigma^{-2}$ (the solid line in Figure 5.10), which means the reconstructions have stopped improving on par with the SNR.

This behavior is exhibited in a few examples of reconstructions taken from the test set in Figure 5.11, where we ordered reconstructions with increasing SNR from top to bottom and plotted the surface brightness and foreground densities in log scale. As can be seen, errors in reconstructed parameters remain of the same order of magnitude as SNR is increased from ~ 220 to 500, implying that above this SNR threshold, the reconstructions are dominated by systematics.

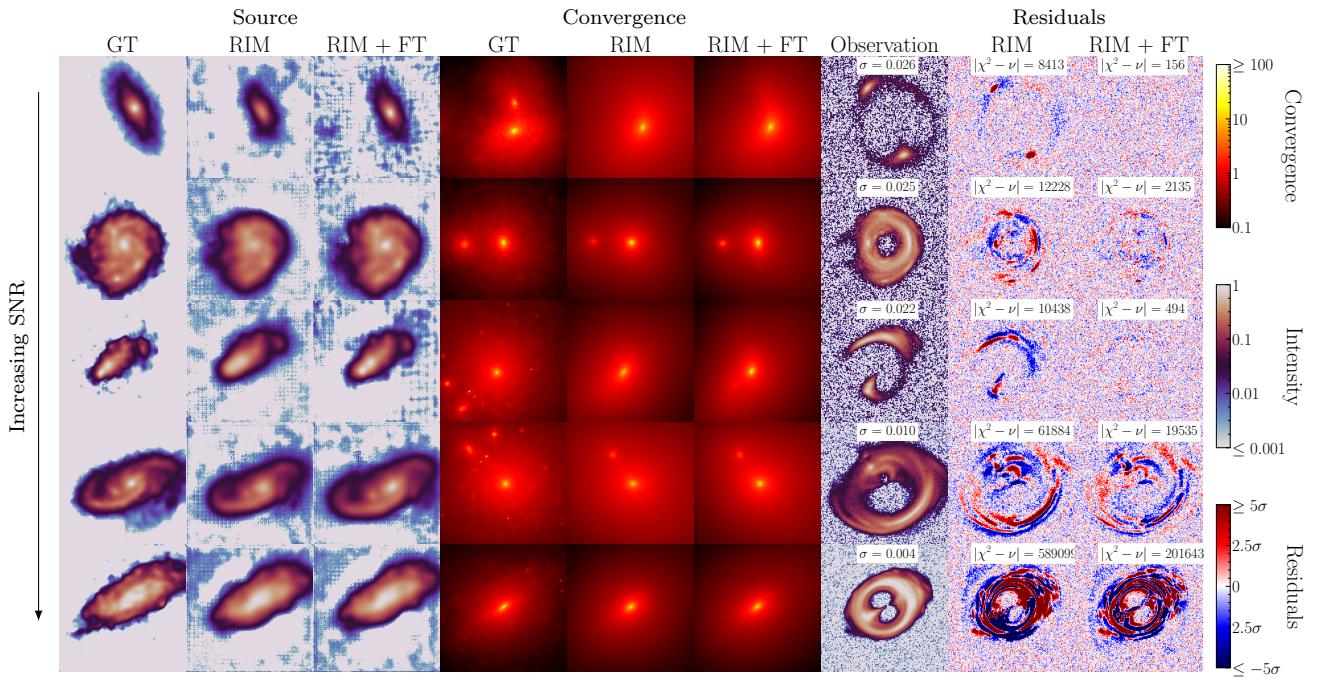


Figure 5.11 – Comparison between baseline (RIM) and fine-tuned (RIM+FT) reconstructions for gravitational lensing systems from the test set (GT). From top to bottom, we increase SNR.

5.5.2 Quality of the Reconstructions

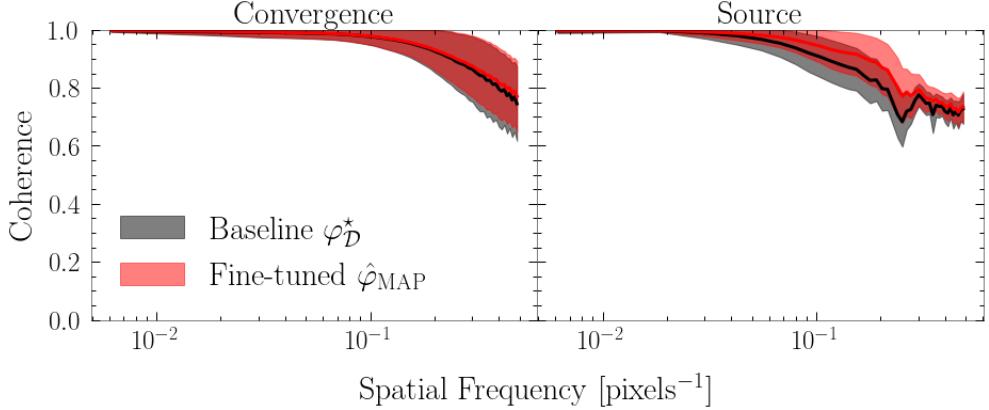


Figure 5.12 – Statistics of the coherence spectrum on the test set. The solid line is the average coherence. The transparent region is the 68% confidence interval. The fine-tuning procedure yields a noticeable improvement on the coherence of the source at all frequencies.

In addition to a visual inspection of the reconstructed sources and convergences, we compute the coherence spectrum to quantitatively assess the quality of the reconstructions

$$\gamma(k) = \frac{P_{12}(k)}{\sqrt{P_{11}(k)P_{22}(k)}}. \quad (5.14)$$

Here, $P_{ij}(k)$ is the cross power spectrum of images i and j at the wavenumber k . Figure 5.12 shows the mean value and the 68% inclusion interval of $\gamma(k)$ for the convergence and source maps in a test set of 3000 examples. The fine-tuning procedure, shown in red, is able to significantly improve the coherence of the baseline background source, shown in black, at all scales. The coherence spectrum of the convergence sees a slight improvement due to the fine-tuning procedure. Still, we note that many examples in the dataset exhibit significant improvement, which we illustrate in Figure 5.3.

5.6 Conclusion

The results obtained here demonstrate the effectiveness of machine learning methods, specifically a recurrent inference machine, for inferring pixelated maps of the distribution of mass in lensing galaxies and the distribution of surface brightness in the background galaxies. Since this is a heavily under-constrained problem, stringent priors are needed to avoid overfitting the data, a task that has traditionally been difficult to accomplish with traditional statistical models (e.g., [Saha and Williams, 1997](#)). The model proposed here can implicitly learn these priors from a set of training data.

The fine-tuning step that we propose in this work is a general procedure (i.e. not specific to our

model or problem), which enables us to exploit a diagonal second-order Laplace approximation of the implicit prior learned by a baseline estimator during pre-training. We use fine-tuning in order to significantly improve this baseline estimator (i.e., a better MAP estimate), by using the likelihood of the data and the EWC prior. In the context of our work, we find that fine-tuning has a limiting — or threshold — behavior, which we speculate is due to the limited expressivity of the neural network and its inductive biases learned during pre-training.

The flexible and expressive form of the reconstructions shown in this work means that, in principle, any lensing system (e.g., a single simple galaxy or a group of complex galaxies) could be analyzed by this model, without any need for pre-determining the model parameterization. This is of high value given the diversity of observed lensing systems, and their relevance for constraining astrophysical and cosmological parameters.

Perhaps the most important limitation of the method is the fact that, in its current form, the model only provides point estimates of the parameters of interest. Quantifying the posteriors of such high-dimensional data will require an efficient and accurate generative process (e.g., see [Adam et al., 2022](#)), which we plan to explore and develop in future works.

Software and data

The source code, as well as the various scripts and parameters used to produce the model and results is available as open-source software under the package `Censai`³. The model parameters, as well as convergence maps used to train these models and the test set examples and reconstructions results are also available as open-source datasets hosted by Zenodo⁴. This research made use of `Tensorflow` ([Abadi et al., 2015](#)), `Tensorflow-Probability` ([Dillon et al., 2017](#)), `Numpy` ([Harris et al., 2020](#)), `Scipy` ([Virtanen et al., 2020](#)), `Matplotlib` ([Hunter, 2007](#)), `Scikit-image` ([Van der Walt et al., 2014](#)), `IPython` ([Pérez and Granger, 2007](#)), `Pandas` ([Wes McKinney, 2010; pandas development team, 2020](#)), `Scikit-learn` ([Pedregosa et al., 2011](#)), `Astropy` ([Astropy Collaboration et al., 2013, 2018](#)) and `GalSim` ([Rowe et al., 2015](#)).

Acknowledgements

This research was made possible by a generous donation by Eric and Wendy Schmidt with the recommendation of the Schmidt Futures Foundation.

We would like to thank Ronan Legin for fruitful discussions and insights about training the neural network. We would also like to thank Max Welling for insightful comments on our work. The work is in part supported by computational resources provided by Calcul Quebec, Compute Canada and the Digital Research Alliance of Canada. Y.H. and L.P. acknowledge support from

³  <https://github.com/AlexandreAdam/Censai>

⁴  <https://doi.org/10.5281/zenodo.6555463>

the National Sciences and Engineering Council of Canada grant RGPIN-2020-05102, the Fonds de recherche du Québec grant 2022-NC-301305 and 300397, and the Canada Research Chairs Program. A.A. was supported by an IVADO Excellence Scholarship.

Chapitre 6

Conclusion

Dans ce mémoire, nous avons exploré deux méthodes statistiques basées sur l'apprentissage machine profond et utilisées pour la reconstruction d'image dans le contexte des lentilles gravitationnelles de type galaxie-galaxie. Les auto-encodeurs variationnels sont utilisés pour modéliser la distribution implicite marginale d'un ensemble de données. Cette modélisation nous permet non seulement d'augmenter la taille de l'ensemble de données d'entraînement d'une machine à inférence récurrentielle, mais nous permet aussi d'approximer une distribution de probabilité conditionnelle à une reconstruction approximative d'une galaxie en arrière-plan ou d'une distribution de masse d'une galaxie en avant-plan par notre modèle d'inférence. Cette propriété est utilisée pour le réglage fin de la machine à inférence récurrentielle.

Les machines à inférence récurrentielle sont utilisées pour apprendre un algorithme d'optimisation avec une convergence très rapide ($T \sim 10$ itérations) en encodant des biais inductifs dans les poids d'un réseau de neurones convolutif récurrent avec une architecture U-net. Cette approche nous permet de préserver l'information à différentes échelles spatiales et de contrôler indépendamment la reconstruction à ces différentes échelles. Pour le problème inverse étudié dans ce mémoire, cette propriété s'avère cruciale. Finalement, pour stabiliser l'entraînement du modèle g_φ , nous avons trouvé qu'ajouter l'observation en entrée au modèle nous permettait d'inclure l'inférence de la valeur initiale dans la relation de récurrence.

Ce travail est à la fois pertinent pour les objectifs scientifiques mentionnés en introduction, soit la recherche et l'étude de halos de matière noire froide et l'étude des galaxies jeunes par l'effet de grossissement des lentilles gravitationnelles, et aussi opportun étant donné le nombre grandissant de ces objets qu'on anticipe de découvrir dans la prochaine décennie et le pouvoir de résolution grandissant des grands observatoires. Mes objectifs de recherche, suivant ces succès initiaux, sont maintenant d'améliorer la méthode pour atteindre la résolution de télescopes comme le télescope spatial James Webb ou encore le télescope géant européen. Ayant démontré que notre approche peut reconstruire un large ensemble de lentilles gravitationnelles simulées dans le chapitre 5, l'étape suivante pour atteindre cet objectif consiste à appliquer notre méthode à un ensemble de données

provenant du télescope spatial Hubble. Le fer à cheval cosmique, montré à la figure 2.2b, est un exemple particulièrement intéressant pour ce test étant donné les difficultés inhérentes à modéliser cette lentille par les méthodes traditionnelles (James et al., 2018; Schuldt et al., 2019; Cheng et al., 2019).

Notre tentative récente de modéliser ce système, montrée à la figure 6.1, indique déjà des résultats prometteurs. L'image de la protogalaxie en arrière-plan reconstruite avec la machine à inférence récurrentielle possède les mêmes 4 régions brillantes avec un taux de formation d'étoile très élevées identifiées par James et al. (2018). Avec l'effet de lentille, on est en mesure d'imager cette protogalaxie avec une résolution deux fois meilleure que la résolution du télescope Hubble. De plus, la distribution de masse obtenue nous indique déjà que certaines asymétries importantes, au-delà des perturbations linéaires généralement admises autour d'un modèle elliptique, sont probablement nécessaires pour accomplir une reconstruction statistiquement significative.

Mon second objectif de recherche est de développer un cadre statistique cohérent autour de méthodes d'inférences aussi puissantes, c.-à-d. être en mesure de modéliser correctement les incertitudes aléatoires et épistémiques de nos reconstructions. Finalement, mon troisième objectif est d'appliquer les méthodes explorées dans cette thèse à d'autres problèmes inverses non linéaires, spécifiquement la reconstruction d'image dans le contexte de l'interférométrie par masque irrégulier, pour accélérer l'analyse de données et le processus de découverte scientifique en astronomie.

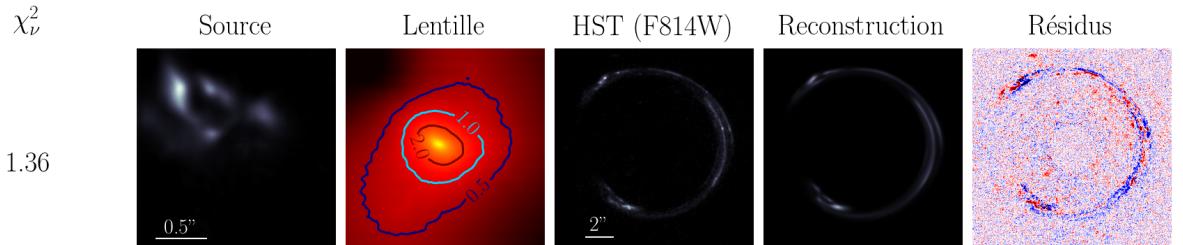


FIGURE 6.1 – Reconstruction du fer à cheval cosmique avec la machine à inférence récurrentielle décrite dans le chapitre 5.

Bibliographie

- M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, et X. Zheng. TensorFlow : Large-scale machine learning on heterogeneous systems, 2015. URL <https://www.tensorflow.org/>. Software available from tensorflow.org.
- H. M. Abdelsalam, P. Saha, et L. L. R. Williams. Nonparametric Reconstruction of Abell 2218 from Combined Weak and Strong Lensing. *AJ*, 116(4) :1541–1552, Oct. 1998a. doi : 10.1086/300546.
- H. M. Abdelsalam, P. Saha, et L. L. R. Williams. Non-parametric reconstruction of cluster mass distribution from strong lensing : modelling Abell 370. *MNRAS*, 294 :734–746, Mar. 1998b. doi : 10.1046/j.1365-8711.1998.01356.x.
- A. Adam, A. Coogan, N. Malkin, R. Legin, L. Perreault-Levasseur, Y. Hezaveh, et Y. Bengio. Posterior samples of source galaxies in strong gravitational lenses with score-based priors. *arXiv e-prints*, art. arXiv :2211.03812, Nov. 2022.
- A. Alemi, I. Fischer, J. Dillon, et K. Murphy. Deep variational information bottleneck. In *ICLR*, 2017. URL <https://arxiv.org/abs/1612.00410>.
- A. A. Alemi, B. Poole, I. Fischer, J. V. Dillon, R. A. Sauvage, et K. Murphy. Fixing a broken ELBO. In J. G. Dy et A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 159–168. PMLR, 2018. URL <http://proceedings.mlr.press/v80/alemi18a.html>.
- N. Anau Montel, A. Coogan, C. Correa, K. Karchev, et C. Weniger. Estimating the warm dark matter mass from strong lensing images with truncated marginal neural ratio estimation. *arXiv e-prints*, art. arXiv :2205.09126, May 2022.
- M. Andrychowicz, M. Denil, S. Gomez, M. W. Hoffman, D. Pfau, T. Schaul, B. Shillingford, et N. de Freitas. Learning to learn by gradient descent by gradient descent. *arXiv e-prints*, art. arXiv :1606.04474, June 2016.
- Astropy Collaboration, T. P. Robitaille, E. J. Tollerud, P. Greenfield, M. Droettboom, E. Bray, T. Aldcroft, M. Davis, A. Ginsburg, A. M. Price-Whelan, W. E. Kerzendorf, A. Conley, N. Crighton, K. Barbary, D. Muna, H. Ferguson, F. Grollier, M. M. Parikh, P. H. Nair, H. M. Unther, C. Deil, J. Woillez, S. Conseil, R. Kramer, J. E. H. Turner, L. Singer, R. Fox, B. A. Weaver, V. Zabalza, Z. I. Edwards, K. Azalee Bostroem, D. J. Burke, A. R. Casey, S. M. Crawford, N. Dencheva, J. Ely, T. Jenness, K. Labrie, P. L. Lim, F. Pierfederici, A. Pontzen, A. Ptak, B. Refsdal, M. Servillat, et O. Streicher. Astropy : A community Python package for astronomy. *A&A*, 558 :A33, Oct. 2013. doi : 10.1051/0004-6361/201322068.
- Astropy Collaboration, A. M. Price-Whelan, B. M. Sipőcz, H. M. Günther, P. L. Lim, S. M. Crawford, S. Conseil, D. L. Shupe, M. W. Craig, N. Dencheva, A. Ginsburg, J. T. VanderPlas, L. D. Bradley, D. Pérez-Suárez, M. de Val-Borro, T. L. Aldcroft, K. L. Cruz, T. P. Robitaille, E. J. Tollerud, C. Ardelean, T. Babej, Y. P. Bach, M. Bachetti, A. V. Bakanov, S. P. Bamford, G. Barentsen, P. Barmby, A. Baumbach, K. L. Berry, F. Biscani, M. Boquien, K. A. Bostroem, L. G. Bouma, G. B. Brammer, E. M. Bray, H. Breytenbach, H. Buddelmeijer, D. J. Burke, G. Calderone, J. L. Cano Rodríguez, M. Cara, J. V. M. Cardoso, S. Cheedella, Y. Copin, L. Corrales, D. Crichton, D. D’Avella, C. Deil, É. Depagne, J. P. Dietrich, A. Donath, M. Droettboom, N. Earl, T. Erben, S. Fabbro, L. A. Ferreira, T. Finethy, R. T. Fox, L. H. Garrison, S. L. J. Gibbons, D. A. Goldstein, R. Gommers, J. P. Greco, P. Greenfield, A. M. Groener, F. Grollier, A. Hagen, P. Hirst, D. Homeier, A. J. Horton, G. Hosseinzadeh, L. Hu, J. S. Hunkeler, Ž. Ivezić, A. Jain, T. Jenness, G. Kanarek, S. Kendrew,

- N. S. Kern, W. E. Kerzendorf, A. Khvalko, J. King, D. Kirkby, A. M. Kulkarni, A. Kumar, A. Lee, D. Lenz, S. P. Littlefair, Z. Ma, D. M. Macleod, M. Mastropietro, C. McCully, S. Montagnac, B. M. Morris, M. Mueller, S. J. Mumford, D. Muna, N. A. Murphy, S. Nelson, G. H. Nguyen, J. P. Ninan, M. Nöthe, S. Ogaz, S. Oh, J. K. Parejko, N. Parley, S. Pascual, R. Patil, A. A. Patil, A. L. Plunkett, J. X. Prochaska, T. Rastogi, V. Reddy Janga, J. Sabater, P. Sakurikar, M. Seifert, L. E. Sherbert, H. Sherwood-Taylor, A. Y. Shih, J. Sick, M. T. Silbiger, S. Singanamalla, L. P. Singer, P. H. Sladen, K. A. Sooley, S. Sornarajah, O. Streicher, P. Teuben, S. W. Thomas, G. R. Tremblay, J. E. H. Turner, V. Terrón, M. H. van Kerkwijk, A. de la Vega, L. L. Watkins, B. A. Weaver, J. B. Whitmore, J. Woillez, V. Zabalza, et Astropy Contributors. The Astropy Project : Building an Open-science Project and Status of the v2.0 Core Package. *AJ*, 156(3) :123, Sept. 2018. doi : 10.3847/1538-3881/aabc4f.
- K. E. Atkinson. *An Introduction to Numerical Analysis*, chapter 6, pages 341–357. John Wiley & Sons, New York, second edition, 1989. ISBN 0471500232. URL <http://www.worldcat.org/isbn/0471500232>.
- D. Aubert, A. Amara, et R. Benton Metcalf. Smooth particle lensing. *Monthly Notices of the Royal Astronomical Society*, 376(1) :113–124, 2007. ISSN 00358711. doi : 10.1111/j.1365-2966.2006.11296.x.
- M. W. Auger, T. Treu, A. S. Bolton, R. Gavazzi, L. V. E. Koopmans, P. J. Marshall, L. A. Moustakas, et S. Burles. The Sloan Lens ACS Survey. X. Stellar, Dynamical, and Total Mass Correlations of Massive Early-type Galaxies. *ApJ*, 724(1) :511–525, Nov. 2010. doi : 10.1088/0004-637X/724/1/511.
- M. Barnabè, O. Czoske, L. V. E. Koopmans, T. Treu, A. S. Bolton, et R. Gavazzi. Two-dimensional kinematics of SLACS lenses - II. Combined lensing and dynamics analysis of early-type galaxies at $z = 0.08\text{--}0.33$. *MNRAS*, 399(1) :21–36, Oct. 2009. doi : 10.1111/j.1365-2966.2009.14941.x.
- M. Bartelmann. Cosmology, 2004. URL <https://heibox.uni-heidelberg.de/f/e1e57faba9a44eb88692/>. Lecture notes from a course given at the Institut für Theoretische Astrophysik at Universität Heidelberg. Last visited 06/07/2022.
- M. Bartelmann. Gravitational lensing. *Classical and Quantum Gravity*, 27 :233001, 2010.
- M. Bartelmann, R. Narayan, S. Seitz, et P. Schneider. Maximum-likelihood Cluster Reconstruction. *ApJ*, 464 :L115, June 1996. doi : 10.1086/310114.
- F. Bellagamba, N. Tessore, et R. B. Metcalf. Zooming into the Cosmic Horseshoe : new insights on the lens profile and the source shape. *Monthly Notices of the Royal Astronomical Society*, 464(4) :4823–4834, 10 2016. ISSN 0035-8711. doi : 10.1093/mnras/stw2726. URL <https://doi.org/10.1093/mnras/stw2726>.
- V. Belokurov, N. W. Evans, A. Moiseev, L. J. King, P. C. Hewett, M. Pettini, L. Wyrzykowski, R. G. McMahon, M. C. Smith, G. Gilmore, S. F. Sanchez, A. Udalski, S. Koposov, D. B. Zucker, et C. J. Walcher. The Cosmic Horseshoe : Discovery of an Einstein Ring around a Giant Luminous Red Galaxy. *ApJ*, 671(1) :L9–L12, Dec. 2007. doi : 10.1086/524948.
- Y. Bengio. Learning deep architectures for ai. *Found. Trends Mach. Learn.*, 2(1) :1–127, jan 2009. ISSN 1935-8237. doi : 10.1561/2200000006. URL <https://doi.org/10.1561/2200000006>.
- S. Birrer et A. Amara. lenstronomy : Multi-purpose gravitational lens modelling software package. *Physics of the Dark Universe*, 22 :189–201, Dec. 2018. doi : 10.1016/j.dark.2018.11.002.
- S. Birrer, A. Amara, et A. Refregier. Gravitational Lens Modeling with Basis Sets. *ApJ*, 813(2) :102, Nov. 2015. doi : 10.1088/0004-637X/813/2/102.
- S. Birrer, T. Treu, C. E. Rusu, V. Bonvin, C. D. Fassnacht, J. H. H. Chan, A. Agnello, A. J. Shajib, G. C. F. Chen, M. Auger, F. Courbin, S. Hilbert, D. Sluse, S. H. Suyu, K. C. Wong, P. Marshall, B. C. Lemaux, et G. Meylan. H0LiCOW - IX. Cosmographic analysis of the doubly imaged quasar SDSS 1206+4332 and a new measurement of the Hubble constant. *MNRAS*, 484(4) :4726–4753, Apr. 2019. doi : 10.1093/mnras/stz200.
- C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 edition, 2007. ISBN 0387310738.
- R. D. Blandford et R. Narayan. Cosmological applications of gravitational lensing. *Annual Review of Astronomy and Astrophysics*, 30(1) :311–358, 1992. doi : 10.1146/annurev.aa.30.090192.001523. URL <https://doi.org/10.1146/annurev.aa.30.090192.001523>.

- B. Blum, S. W. Digel, A. Drlica-Wagner, S. Habib, K. Heitmann, M. Ishak, S. W. Jha, S. M. Kahn, R. Mandelbaum, P. Marshall, J. A. Newman, A. Roodman, et C. W. Stubbs. Snowmass2021 Cosmic Frontier White Paper : Rubin Observatory after LSST. *arXiv e-prints*, art. arXiv :2203.07220, Mar. 2022.
- A. S. Bolton, S. Burles, L. V. E. Koopmans, T. Treu, et L. A. Moustakas. SDSS j140228.22+632133.3 : A new spectroscopically selected gravitational lens. *The Astrophysical Journal*, 624(1) :L21–L24, apr 2005. doi : 10.1086/430440. URL <https://doi.org/10.1086/430440>.
- A. S. Bolton, S. Burles, L. V. E. Koopmans, T. Treu, et L. A. Moustakas. The Sloan Lens ACS Survey. I. A Large Spectroscopically Selected Sample of Massive Early-Type Lens Galaxies. *ApJ*, 638(2) :703–724, Feb. 2006. doi : 10.1086/498884.
- A. S. Bolton, S. Burles, L. V. E. Koopmans, T. Treu, R. Gavazzi, L. A. Moustakas, R. Wayth, et D. J. Schlegel. The Sloan Lens ACS Survey. V. The Full ACS Strong-Lens Sample. *ApJ*, 682(2) :964–984, Aug. 2008. doi : 10.1086/589327.
- S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, et S. Bengio. Generating Sentences from a Continuous Space. *arXiv e-prints*, art. arXiv :1511.06349, Nov. 2015.
- M. Bradač, P. Schneider, M. Lombardi, et T. Erben. Strong and weak lensing united. I. The combined strong and weak lensing cluster mass reconstruction method. *A&A*, 437(1) :39–48, July 2005. doi : 10.1051/0004-6361:20042233.
- C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, et A. Lerchner. Understanding disentangling in β -VAE. *arXiv e-prints*, art. arXiv :1804.03599, Apr. 2018.
- J. C. Butcher. *Numerical Methods for Ordinary Differential Equations*, chapter 2, pages 21–26. John Wiley & Sons, Hoboken, New Jersey, third edition, 2016. ISBN 9781119121503. doi : 10.1002/9781119121534. URL <https://onlinelibrary.wiley.com/doi/book/10.1002/9781119121534>.
- M. Cacciato, M. Bartelmann, M. Meneghetti, et L. Moscardini. Combining weak and strong lensing in cluster potential reconstruction. *A&A*, 458(2) :349–356, Nov. 2006. doi : 10.1051/0004-6361:20054582.
- G. Casella et R. Berger. *Statistical Inference*. Duxbury Resource Center, June 2001. ISBN 0534243126.
- J. Cheng, M. P. Wiesner, E.-H. Peng, W. Cui, J. R. Peterson, et G. Li. Adaptive Grid Lens Modeling of the Cosmic Horseshoe Using Hubble Space Telescope Imaging. *ApJ*, 872(2) :185, Feb. 2019. doi : 10.3847/1538-4357/ab0029.
- K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, et Y. Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv e-prints*, art. arXiv :1406.1078, June 2014.
- O. Chwolson. Über eine mögliche form fiktiver doppelsterne. *Astronomische Nachrichten*, 221(20) :329–330, 1924. doi : <https://doi.org/10.1002/asna.19242212003>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/asna.19242212003>.
- D. Coe, E. Fuselier, N. Benítez, T. Broadhurst, B. Frye, et H. Ford. LensPerfect : Gravitational Lens Mass Map Reconstructions Yielding Exact Reproduction of All Multiple Images. *ApJ*, 681(2) :814–830, July 2008. doi : 10.1086/588250.
- J. P. Coles, J. I. Read, et P. Saha. Gravitational lens recovery with GLASS : measuring the mass profile and shape of a lens. *MNRAS*, 445(3) :2181–2197, Dec. 2014. doi : 10.1093/mnras/stu1781.
- P. Coles et F. Lucchin. *Cosmology : The Origin and Evolution of Cosmic Structure*. Wiley, 2 edition, July 2002.
- A. Congdon et C. Keeton. *Principles of Gravitational Lensing : Light Deflection as a Probe of Astrophysics and Cosmology*. Springer Praxis Books. Springer International Publishing, 2018. ISBN 9783030021221. URL <https://books.google.ca/books?id=kt58DwAAQBAJ>.
- A. Coogan, K. Karchev, et C. Weniger. Targeted Likelihood-Free Inference of Dark Matter Substructure in Strongly-Lensed Galaxies. *arXiv e-prints*, art. arXiv :2010.07032, Oct. 2020.
- T. M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *Electronic Computers, IEEE Transactions on*, EC-14(3) :326–334, 1965. URL <http://hebb.mit.edu/courses/9.641/2002/readings/Cover65.pdf>.

- T. M. Cover et J. A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006. ISBN 0471241954.
- H. Cramér. *Mathematical methods of statistics*, volume 9. Princeton University Press, Princeton, NJ, 1946.
- G. V. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2 : 303–314, 1989.
- N. Dalal et C. S. Kochanek. Direct Detection of Cold Dark Matter Substructure. *ApJ*, 572(1) :25–33, June 2002. doi : 10.1086/340303.
- S. Deb, A. Morandi, K. Pedersen, S. Riemer-Sorensen, D. M. Goldberg, et H. Dahle. Mass Reconstruction using Particle Based Lensing II : Quantifying substructure with Strong+Weak lensing and X-rays. *arXiv e-prints*, art. arXiv :1201.3636, Jan. 2012.
- J. M. Diego, P. Protopapas, H. B. Sandvik, et M. Tegmark. Non-parametric inversion of strong lensing systems. *MNRAS*, 360 (2) :477–491, June 2005. doi : 10.1111/j.1365-2966.2005.09021.x.
- J. M. Diego, M. Tegmark, P. Protopapas, et H. B. Sandvik. Combined reconstruction of weak and strong lensing data with WSLAP. *MNRAS*, 375(3) :958–970, Mar. 2007. doi : 10.1111/j.1365-2966.2007.11380.x.
- J. V. Dillon, I. Langmore, D. Tran, E. Brevdo, S. Vasudevan, D. Moore, B. Patton, A. Alemi, M. Hoffman, et R. A. Saurous. TensorFlow Distributions. *arXiv e-prints*, art. arXiv :1711.10604, Nov. 2017.
- S. Dodelson et F. Schmidt. *Cosmology : The Origin and Evolution of Cosmic Structure*. Academic Press, 2 edition, March 2003.
- J. Duchi, E. Hazan, et Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12(null) :2121–2159, jul 2011. ISSN 1532-4435.
- A. S. Eddington. The total eclipse of 1919 May 29 and the influence of gravitation on light. *The Observatory*, 42 :119–122, Mar. 1919.
- A. Einstein. Lens-Like Action of a Star by the Deviation of Light in the Gravitational Field. *Science*, 84(2188) :506–507, 1936. doi : 10.1126/science.84.2188.506. URL <https://www.science.org/doi/abs/10.1126/science.84.2188.506>.
- A. Etherington, J. W. Nightingale, R. Massey, X. Cao, A. Robertson, N. C. Amorisco, A. Amvrosiadis, S. Cole, C. S. Frenk, Q. He, R. Li, et S.-I. Tam. Automated galaxy-galaxy strong lens modelling : no lens left behind. *arXiv e-prints*, art. arXiv :2202.09201, Feb. 2022.
- E. E. Falco, M. V. Gorenstein, et I. I. Shapiro. New Model for the 0957+561 Gravitational Lens System : Bounds on Masses of a Possible Black Hole and Dark Matter and Prospects for Estimation of H 0. *ApJ*, 372 :364, May 1991. doi : 10.1086/169984.
- C. Faure, J.-P. Kneib, G. Covone, L. Tasca, A. Leauthaud, P. Capak, K. Jahnke, V. Smolcic, S. de la Torre, R. Ellis, A. Finoguenov, A. Koekemoer, O. Le Fevre, R. Massey, Y. Mellier, A. Refregier, J. Rhodes, N. Scoville, E. Schinnerer, J. Taylor, L. Van Waerbeke, et J. Walcher. First Catalog of Strong Lens Candidates in the COSMOS Field. *ApJS*, 176(1) :19–38, May 2008. doi : 10.1086/526426.
- C. Finn, P. Abbeel, et S. Levine. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. *arXiv e-prints*, art. arXiv :1703.03400, Mar. 2017.
- R. Gavazzi, C. Adami, F. Durret, J. C. Cuillandre, O. Ilbert, A. Mazure, R. Pelló, et M. P. Ulmer. A weak lensing study of the Coma cluster. *A&A*, 498(2) :L33–L36, May 2009. doi : 10.1051/0004-6361/200911841.
- A. Ghosh, L. L. R. Williams, et J. Liesenborgs. Free-form gralé lens inversion of galaxy clusters with up to 1000 multiple images. *MNRAS*, 494(3) :3998–4014, May 2020. doi : 10.1093/mnras/staa962.
- D. Gilman, S. Birrer, A. Nierenberg, T. Treu, X. Du, et A. Benson. Warm dark matter chills out : constraints on the halo mass function and the free-streaming length of dark matter with eight quadruple-image strong gravitational lenses. *MNRAS*, 491 (4) :6077–6101, Feb. 2020. doi : 10.1093/mnras/stz3480.

- D. Gilman, J. Bovy, T. Treu, A. Nierenberg, S. Birrer, A. Benson, et O. Sameie. Strong lensing signatures of self-interacting dark matter in low-mass haloes. *MNRAS*, 507(2) :2432–2447, Oct. 2021. doi : 10.1093/mnras/stab2335.
- Z. Goldfeld et Y. Polyanskiy. The Information Bottleneck Problem and Its Applications in Machine Learning. *arXiv e-prints*, art. arXiv :2004.14941, Apr. 2020.
- A. Goobar, R. Amanullah, S. R. Kulkarni, P. E. Nugent, J. Johansson, C. Steidel, D. Law, E. Mörtsell, R. Quimby, N. Blagorodnova, A. Brandeker, Y. Cao, A. Cooray, R. Ferretti, C. Fremling, L. Hangard, M. Kasliwal, T. Kupfer, R. Lunnan, F. Masci, A. A. Miller, H. Nayyeri, J. D. Neill, E. O. Ofek, S. Papadogiannakis, T. Petrushevska, V. Ravi, J. Sollerman, M. Sullivan, F. Taddia, R. Walters, D. Wilson, L. Yan, et O. Yaron. iPTF16geu : A multiply imaged, gravitationally lensed type Ia supernova. *Science*, 356(6335) :291–295, Apr. 2017. doi : 10.1126/science.aal2729.
- I. J. Goodfellow, J. Shlens, et C. Szegedy. Explaining and Harnessing Adversarial Examples. *arXiv e-prints*, art. arXiv :1412.6572, Dec. 2014.
- I. J. Goodfellow, Y. Bengio, et A. Courville. *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016. <http://www.deeplearningbook.org>.
- J. Hadamard. Sur les problèmes aux dérivés partielles et leur signification physique. *Princeton University Bulletin*, 13 :49–52, 1902.
- C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, et T. E. Oliphant. Array programming with NumPy. *Nature*, 585(7825) :357–362, Sept. 2020. doi : 10.1038/s41586-020-2649-2. URL <https://doi.org/10.1038/s41586-020-2649-2>.
- Y. D. Hezaveh, N. Dalal, D. P. Marrone, Y.-Y. Mao, W. Morningstar, D. Wen, R. D. Blandford, J. E. Carlstrom, C. D. Fassnacht, G. P. Holder, A. Kemball, P. J. Marshall, N. Murray, L. Perreault Levasseur, J. D. Vieira, et R. H. Wechsler. Detection of Lensing Substructure Using ALMA Observations of the Dusty Galaxy SDP.81. *ApJ*, 823(1) :37, May 2016. doi : 10.3847/0004-637X/823/1/37.
- Y. D. Hezaveh, L. Perreault Levasseur, et P. J. Marshall. Fast automated analysis of strong gravitational lenses with convolutional neural networks. *Nature*, 548(7669) :555–557, Aug. 2017. doi : 10.1038/nature23463.
- I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, et A. Lerchner. beta-vae : Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.
- G. Hinton. Neural networks for machine learning. http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf, 2012. Accédé le 2022-07-10.
- S. Hochreiter et J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8) :1735–1780, 1997.
- K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2) :251–257, 1991. ISSN 0893-6080. doi : [https://doi.org/10.1016/0893-6080\(91\)90009-T](https://doi.org/10.1016/0893-6080(91)90009-T). URL <https://www.sciencedirect.com/science/article/pii/089360809190009T>.
- T. Hospedales, A. Antoniou, P. Micaelli, et A. Storkey. Meta-Learning in Neural Networks : A Survey. *arXiv e-prints*, art. arXiv :2004.05439, Apr. 2020.
- X. Huang, C. Storfer, A. Gu, V. Ravi, A. Pilon, W. Sheu, R. Venguswamy, S. Banka, A. Dey, M. Landriau, D. Lang, A. Meisner, J. Moustakas, A. D. Myers, R. Sajith, E. F. Schlafly, et D. J. Schlegel. Discovering New Strong Gravitational Lenses in the DESI Legacy Imaging Surveys. *ApJ*, 909(1) :27, Mar. 2021. doi : 10.3847/1538-4357/abd62b.
- E. Hubble. A Relation between Distance and Radial Velocity among Extra-Galactic Nebulae. *Proceedings of the National Academy of Science*, 15(3) :168–173, Mar. 1929. doi : 10.1073/pnas.15.3.168.
- J. D. Hunter. Matplotlib : A 2d graphics environment. *Computing in Science & Engineering*, 9(3) :90–95, 2007. doi : 10.1109/MCSE.2007.55.

- B. L. James, M. Auger, M. Pettini, D. P. Stark, V. Belokurov, et S. Carniani. Mapping UV properties throughout the Cosmic Horseshoe : lessons from VLT-MUSE. *MNRAS*, 476(2) :1726–1740, May 2018. doi : 10.1093/mnras/sty315.
- M. J. Jee, H. C. Ford, G. D. Illingworth, R. L. White, T. J. Broadhurst, D. A. Coe, G. R. Meurer, A. van der Wel, N. Benítez, J. P. Blakeslee, R. J. Bouwens, L. D. Bradley, R. Demarco, N. L. Homeier, A. R. Martel, et S. Mei. Discovery of a Ringlike Dark Matter Structure in the Core of the Galaxy Cluster Cl 0024+17. *ApJ*, 661(2) :728–749, June 2007. doi : 10.1086/517498.
- D. Jimenez Rezende et F. Viola. Taming VAEs. *arXiv e-prints*, art. arXiv :1810.00597, Oct. 2018.
- C. Kaae Sønderby, T. Raiko, L. Maaløe, S. Kaae Sønderby, et O. Winther. Ladder Variational Autoencoders. *arXiv e-prints*, art. arXiv :1602.02282, Feb. 2016.
- K. Karchev, A. Coogan, et C. Weniger. Strong-lensing source reconstruction with variationally optimized Gaussian processes. *MNRAS*, 512(1) :661–685, May 2022. doi : 10.1093/mnras/stac311.
- P. L. Kelly, S. A. Rodney, T. Treu, R. J. Foley, G. Brammer, K. B. Schmidt, A. Zitrin, A. Sonnenfeld, L.-G. Strolger, O. Graur, A. V. Filippenko, S. W. Jha, A. G. Riess, M. Bradac, B. J. Weiner, D. Scolnic, M. A. Malkan, A. von der Linden, M. Trenti, J. Hjorth, R. Gavazzi, A. Fontana, J. C. Merten, C. McCully, T. Jones, M. Postman, A. Dressler, B. Patel, S. B. Cenko, M. L. Graham, et B. E. Tucker. Multiple images of a highly magnified supernova formed by an early-type cluster galaxy lens. *Science*, 347(6226) :1123–1126, Mar. 2015. doi : 10.1126/science.aaa3350.
- L. J. King et I. W. A. Browne. Biases, selection effects and image multiplicities in the Jodrell Bank-VLA gravitational lens survey. *MNRAS*, 282(1) :67–76, Sept. 1996. doi : 10.1093/mnras/282.1.67.
- D. P. Kingma et J. Ba. Adam : A Method for Stochastic Optimization. *arXiv e-prints*, art. arXiv :1412.6980, Dec. 2014.
- D. P. Kingma et M. Welling. Auto-Encoding Variational Bayes. *arXiv e-prints*, art. arXiv :1312.6114, Dec. 2013.
- D. P. Kingma et M. Welling. An Introduction to Variational Autoencoders. *arXiv e-prints*, art. arXiv :1906.02691, June 2019.
- J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, et R. Hadsell. Overcoming catastrophic forgetting in neural networks. *arXiv e-prints*, art. arXiv :1612.00796, Dec. 2016.
- A. M. Koekemoer, H. Aussel, D. Calzetti, P. Capak, M. Giavalisco, J.-P. Kneib, A. Leauthaud, O. Le Fevre, H. J. McCracken, R. Massey, B. Mobasher, J. Rhodes, N. Scoville, et P. L. Shopbell. The COSMOS Survey : Hubble Space Telescope Advanced Camera for Surveys Observations and Data Processing. *The Astrophysical Journal Supplement Series*, 172(1) :196–202, sep 2007. ISSN 0067-0049. doi : 10.1086/520086.
- A. N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1 : 1–7, 1965.
- L. V. E. Koopmans, T. Treu, A. S. Bolton, S. Burles, et L. A. Moustakas. The Sloan Lens ACS Survey. III. The Structure and Formation of Early-Type Galaxies and Their Evolution since $z \sim 1$. *ApJ*, 649(2) :599–615, Oct. 2006. doi : 10.1086/505696.
- A. Krizhevsky, I. Sutskever, et G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, page 1097–1105, Red Hook, NY, USA, 2012. Curran Associates Inc.
- F. Lanusse, R. Mandelbaum, S. Ravanbakhsh, C.-L. Li, P. Freeman, et B. Póczos. Deep generative models for galaxy image simulations. *MNRAS*, 504(4) :5543–5555, July 2021. doi : 10.1093/mnras/stab1214.
- A. Leauthaud, R. Massey, J.-P. Kneib, J. Rhodes, D. E. Johnston, P. Capak, C. Heymans, R. S. Ellis, A. M. Koekemoer, O. L. Fèvre, Y. Mellier, A. Réfrégier, A. C. Robin, N. Scoville, L. Tasca, J. E. Taylor, et L. V. Waerbeke. Weak Gravitational Lensing with COSMOS : Galaxy Selection and Shape Measurements. *The Astrophysical Journal Supplement Series*, 172(1) :219, sep 2007. ISSN 0067-0049. doi : 10.1086/516598.
- Y. Lecun et Y. Bengio. *Convolutional Networks for Images, Speech and Time Series*, pages 255–258. The MIT Press, 1995.
- R. Legin, Y. Hezaveh, L. Perreault Levasseur, et B. Wandelt. Simulation-Based Inference of Strong Gravitational Lensing Parameters. *arXiv e-prints*, art. arXiv :2112.05278, Dec. 2021.

- R. Legin, C. Stone, Y. Hezaveh, et L. Perreault-Levasseur. Population-Level Inference of Strong Gravitational Lenses with Neural Network-Based Selection Correction. *arXiv e-prints*, art. arXiv :2207.04123, July 2022.
- P. Lemos, N. Jeffrey, M. Cranmer, S. Ho, et P. Battaglia. Rediscovering orbital mechanics with machine learning. *arXiv e-prints*, art. arXiv :2202.02306, Feb. 2022.
- N. Li, C. Becker, et S. Dye. The impact of line-of-sight structures on measuring H_0 with strong lensing time delays. *MNRAS*, 504(2) :2224–2234, June 2021. doi : 10.1093/mnras/stab984.
- J. Liesenborgs, S. De Rijcke, et H. Dejonghe. A genetic algorithm for the non-parametric inversion of strong lensing systems. *MNRAS*, 367(3) :1209–1216, Apr. 2006. doi : 10.1111/j.1365-2966.2006.10040.x.
- J. Liesenborgs, S. de Rijcke, H. Dejonghe, et P. Bekaert. Non-parametric inversion of gravitational lensing systems with few images using a multi-objective genetic algorithm. *MNRAS*, 380(4) :1729–1736, Oct. 2007. doi : 10.1111/j.1365-2966.2007.12236.x.
- F. Link. Sur les conséquences photométriques de la déviation d’Einstein. *Comptes rendus hebdomadaires des séances de l’Académie des sciences*, pages 917–3919, Janvier 1936. URL <https://gallica.bnf.fr/ark:/12148/bpt6k3154f/f917.item>. r=Link.
- F. Link. Sur les conséquences photométriques de la déviation d’Einstein. *Bulletin Astronomique*, pages 73–90, 1937. URL <https://gallica.bnf.fr/ark:/12148/bpt6k6544677c/f83.item>.
- K. Lønning, P. Putzky, J. J. Sonke, L. Reneman, M. W. Caan, et M. Welling. Recurrent inference machines for reconstructing heterogeneous MRI data. *Medical Image Analysis*, 53 :64–78, apr 2019. ISSN 13618423. doi : 10.1016/j.media.2019.01.005.
- LSST Science Collaboration, P. A. Abell, J. Allison, S. F. Anderson, J. R. Andrew, J. R. P. Angel, L. Armus, D. Arnett, S. J. Asztalos, T. S. Axelrod, S. Bailey, D. R. Ballantyne, J. R. Bankert, W. A. Barkhouse, J. D. Barr, L. F. Barrientos, A. J. Barth, J. G. Bartlett, A. C. Becker, J. Becla, T. C. Beers, J. P. Bernstein, R. Biswas, M. R. Blanton, J. S. Bloom, J. J. Bochanski, P. Boeshaar, K. D. Borne, M. Bradac, W. N. Brandt, C. R. Bridge, M. E. Brown, R. J. Brunner, J. S. Bullock, A. J. Burgasser, J. H. Burge, D. L. Burke, P. A. Cargile, S. Chandrasekharan, G. Chartas, S. R. Chesley, Y.-H. Chu, D. Cinabro, M. W. Claire, C. F. Claver, D. Clowe, A. J. Connolly, K. H. Cook, J. Cooke, A. Cooray, K. R. Covey, C. S. Culliton, R. de Jong, W. H. de Vries, V. P. Debattista, F. Delgado, I. P. Dell’Antonio, S. Dhital, R. Di Stefano, M. Dickinson, B. Dilday, S. G. Djorgovski, G. Dobler, C. Donalek, G. Dubois-Felsmann, J. Durech, A. Eliasdottir, M. Eracleous, L. Eyer, E. E. Falco, X. Fan, C. D. Fassnacht, H. C. Ferguson, Y. R. Fernandez, B. D. Fields, D. Finkbeiner, E. E. Figueroa, D. B. Fox, H. Francke, J. S. Frank, J. Frieman, S. Fromenteau, M. Furqan, G. Galaz, A. Gal-Yam, P. Garnavich, E. Gawiser, J. Geary, P. Gee, R. R. Gibson, K. Gilmore, E. A. Grace, R. F. Green, W. J. Gressler, C. J. Grillmair, S. Habib, J. S. Haggerty, M. Hamuy, A. W. Harris, S. L. Hawley, A. F. Heavens, L. Hebb, T. J. Henry, E. Hileman, E. J. Hilton, K. Hoadley, J. B. Holberg, M. J. Holman, S. B. Howell, L. Infante, Z. Ivezic, S. H. Jacoby, B. Jain, R. Jedicke, M. J. Jee, J. Garrett Jernigan, S. W. Jha, K. V. Johnston, R. L. Jones, M. Juric, M. Kaasalainen, Stylianis, Kafka, S. M. Kahn, N. A. Kaib, J. Kalirai, J. Kantor, M. M. Kasliwal, C. R. Keeton, R. Kessler, Z. Knezevic, A. Kowalski, V. L. Krabbendam, K. S. Krughoff, S. Kulkarni, S. Kuhlman, M. Lacy, S. Lepine, M. Liang, A. Lien, P. Lira, K. S. Long, S. Lorenz, J. M. Lotz, R. H. Lupton, J. Lutz, L. M. Macri, A. A. Mahabal, R. Mandelbaum, P. Marshall, M. May, P. M. McGehee, B. T. Meadows, A. Meert, A. Milani, C. J. Miller, M. Miller, D. Mills, D. Minniti, D. Monet, A. S. Mukadam, E. Nakar, D. R. Neill, J. A. Newman, S. Nikolaev, M. Nordby, P. O’Connor, M. Oguri, J. Oliver, S. S. Olivier, J. K. Olsen, K. Olsen, E. W. Olszewski, H. Oluseyi, N. D. Padilla, A. Parker, J. Pepper, J. R. Peterson, C. Petry, P. A. Pinto, J. L. Pizagno, B. Popescu, A. Prsa, V. Radcka, M. J. Raddick, A. Rasmussen, A. Rau, J. Rho, J. E. Rhoads, G. T. Richards, S. T. Ridgway, B. E. Robertson, R. Roskar, A. Saha, A. Sarajedini, E. Scannapieco, T. Schalk, R. Schindler, S. Schmidt, S. Schmidt, D. P. Schneider, G. Schumacher, R. Scranton, J. Sebag, L. G. Seppala, O. Shemmer, J. D. Simon, M. Sivertz, H. A. Smith, J. Allyn Smith, N. Smith, A. H. Spitz, A. Stanford, K. G. Stassun, J. Strader, M. A. Strauss, C. W. Stubbs, D. W. Sweeney, A. Szalay, P. Szkody, M. Takada, P. Thorman, D. E. Trilling, V. Trimble, A. Tyson, R. Van Berg, D. Vanden Berk, J. VanderPlas, L. Verde, B. Vrsnak, L. M. Walkowicz, B. D. Wandelt, S. Wang, Y. Wang, M. Warner, R. H. Wechsler, A. A. West, O. Wiecha, B. F. Williams, B. Willman, D. Wittman, S. C. Wolff, W. M. Wood-Vasey, P. Wozniak, P. Young, A. Zentner, et H. Zhan. LSST Science Book, Version 2.0. *arXiv e-prints*, art. arXiv :0912.0201, Dec. 2009.
- R. Lynds et V. Petrosian. Luminous Arcs in Clusters of Galaxies. *ApJ*, 336 :1, Jan. 1989. doi : 10.1086/166989.
- R. Mandelbaum, C. Lackner, A. Leauthaud, et B. Rowe. COSMOS real galaxy dataset. *Zenodo*, jan 2012. URL <https://zenodo.org/record/3242143>.

- R. Mandelbaum, B. Rowe, J. Bosch, C. Chang, F. Courbin, M. Gill, M. Jarvis, A. Kannawadi, T. Kacprzak, C. Lackner, A. Leauthaud, H. Miyatake, R. Nakajima, J. Rhodes, M. Simet, J. Zuntz, B. Armstrong, S. Bridle, J. Coupon, J. P. Dietrich, M. Gentile, C. Heymans, A. S. Jurling, S. M. Kent, D. Kirkby, D. Margala, R. Massey, P. Melchior, J. Peterson, A. Roodman, et T. Schrabback. The Third Gravitational Lensing Accuracy Testing (GREAT3) Challenge Handbook. *The Astrophysical Journal Supplement Series*, 212(1) :5, apr 2014. ISSN 0067-0049. doi : 10.1088/0067-0049/212/1/5. URL <https://iopscience.iop.org/article/10.1088/0067-0049/212/1/5><https://iopscience.iop.org/article/10.1088/0067-0049/212/1/5/meta>.
- D. P. Marrone, J. S. Spilker, C. C. Hayward, J. D. Vieira, M. Aravena, M. L. N. Ashby, M. B. Bayliss, M. Bethermin, M. Brodwin, M. S. Bothwell, J. E. Carlstrom, S. C. Chapman, C.-C. Chen, T. M. Crawford, D. J. M. Cunningham, C. De Breuck, C. D. Fassnacht, A. H. Gonzalez, T. R. Greve, Y. D. Hezaveh, K. Lacaille, K. C. Litke, S. Lower, J. Ma, M. Malkan, T. B. Miller, W. R. Morningstar, E. J. Murphy, D. Narayanan, K. A. Phadke, K. M. Rotermund, J. Sreevani, B. Stalder, A. A. Stark, M. L. Strandet, M. Tang, et A. Weiß. Galaxy growth in a massive halo in the first billion years of cosmic history. *Nature*, 553(7686) :51–54, Jan. 2018. doi : 10.1038/nature24629.
- M. McCloskey et N. J. Cohen. Catastrophic Interference in Connectionist Networks : The Sequential Learning Problem. In G. H. Bower, editor, *Psychology of Learning and Motivation - Advances in Research and Theory*, volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press, 1989. doi : 10.1016/S0079-7421(08)60536-8. URL <https://www.sciencedirect.com/science/article/pii/S0079742108605368>.
- M. Meneghetti. *Introduction to Gravitational Lensing*. Springer Cham, 2013. doi : 10.1007/978-3-030-73582-1.
- J. Merten. Mesh-free free-form lensing - I. Methodology and application to mass reconstruction. *MNRAS*, 461(3) :2328–2345, Sept. 2016. doi : 10.1093/mnras/stw1413.
- J. Merten, M. Cacciato, M. Meneghetti, C. Mignone, et M. Bartelmann. Combining weak and strong cluster lensing : applications to simulations and MS 2137. *A&A*, 500(2) :681–691, June 2009. doi : 10.1051/0004-6361/200810372.
- S. Mishra-Sharma et G. Yang. Strong lensing source reconstruction using continuous neural fields, 2022. URL <https://arxiv.org/abs/2206.14820>.
- T. Miyato, T. Kataoka, M. Koyama, et Y. Yoshida. Spectral Normalization for Generative Adversarial Networks. *arXiv e-prints*, art. arXiv :1802.05957, Feb. 2018.
- C. Modi, F. Lanusse, U. Seljak, D. N. Spergel, et L. Perreault-Levasseur. CosmicRIM : Reconstructing Early Universe by Combining Differentiable Simulations with Recurrent Inference Machines. *arXiv e-prints*, art. arXiv :2104.12864, Apr. 2021.
- W. R. Morningstar, Y. D. Hezaveh, L. P. Levasseur, R. D. Blandford, P. J. Marshall, P. Putzky, et R. H. Wechsler. Analyzing Interferometric Observations of Strong Gravitational Lenses with Recurrent and Convolutional Neural Networks. *arXiv e-prints*, 2018.
- W. R. Morningstar, L. P. Levasseur, Y. D. Hezaveh, R. Blandford, P. Marshall, P. Putzky, T. D. Rueter, R. Wechsler, et M. Welling. Data-driven Reconstruction of Gravitationally Lensed Galaxies Using Recurrent Inference Machines. *The Astrophysical Journal*, 883(1) :14, 2019. ISSN 1538-4357. doi : 10.3847/1538-4357/ab35d7.
- J. A. Muñoz, E. E. Falco, C. S. Kochanek, J. Lehár, B. A. McLeod, C. D. Impey, H. W. Rix, et C. Y. Peng. The Castles Project. *Ap&SS*, 263 :51–54, June 1998. doi : 10.1023/A:1002120921330.
- S. T. Myers, N. J. Jackson, I. W. A. Browne, A. G. de Bruyn, T. J. Pearson, A. C. S. Readhead, P. N. Wilkinson, A. D. Biggs, R. D. Blandford, C. D. Fassnacht, L. V. E. Koopmans, D. R. Marlow, J. P. McKean, M. A. Norbury, P. M. Phillips, D. Rusin, M. C. Shepherd, et C. M. Sykes. The Cosmic Lens All-Sky Survey - I. Source selection and observations. *MNRAS*, 341(1) :1–12, May 2003. doi : 10.1046/j.1365-8711.2003.06256.x.
- D. Nelson, V. Springel, A. Pillepich, V. Rodriguez-Gomez, P. Torrey, S. Genel, M. Vogelsberger, R. Pakmor, F. Marinacci, R. Weinberger, L. Kelley, M. Lovell, B. Diemer, et L. Hernquist. The IllustrisTNG simulations : public data release. *MNRAS*, 6(1), 2019. ISSN 2197-7909. doi : 10.1186/s40668-019-0028-x. URL www.tng-project.org/data.
- Y. Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. *Proceedings of the USSR Academy of Sciences*, 269 :543–547, 1983.

- J. W. Nightingale, S. Dye, et R. J. Massey. AutoLens : automated modeling of a strong lens's light, mass, and source. *MNRAS*, 478(4) :4738–4784, Aug. 2018. doi : 10.1093/mnras/sty1264.
- S. J. Pan et Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22 :1345–1359, 2010.
- T. pandas development team. pandas-dev/pandas : Pandas, Feb. 2020. URL <https://doi.org/10.5281/zenodo.3509134>.
- J. W. Park, S. Wagner-Carena, S. Birrer, P. J. Marshall, J. Y.-Y. Lin, A. Roodman, et LSST Dark Energy Science Collaboration. Large-scale Gravitational Lens Modeling with Bayesian Neural Networks for Accurate and Precise Inference of the Hubble Constant. *ApJ*, 910(1) :39, Mar. 2021. doi : 10.3847/1538-4357/abdfc4.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, et E. Duchesnay. Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12 :2825–2830, 2011.
- F. Pérez et B. E. Granger. IPython : a system for interactive scientific computing. *Computing in Science and Engineering*, 9 (3) :21–29, May 2007. ISSN 1521-9615. doi : 10.1109/MCSE.2007.53. URL <https://ipython.org>.
- L. Perreault Levasseur, Y. D. Hezaveh, et R. H. Wechsler. Uncertainties in Parameters Estimated with Neural Networks : Application to Strong Gravitational Lensing. *ApJ*, 850(1) :L7, Nov. 2017. doi : 10.3847/2041-8213/aa9704.
- C. E. Petrillo, C. Tortora, S. Chatterjee, G. Vernardos, L. V. E. Koopmans, G. Verdoes Kleijn, N. R. Napolitano, G. Covone, P. Schneider, A. Grado, et J. McFarland. Finding strong gravitational lenses in the Kilo Degree Survey with Convolutional Neural Networks. *MNRAS*, 472(1) :1129–1150, Nov. 2017. doi : 10.1093/mnras/stx2052.
- Planck Collaboration. Planck 2018 results. VI. Cosmological parameters. *A&A*, 641 :A6, Sept. 2020. doi : 10.1051/0004-6361/201833910.
- P. Putzky et M. Welling. Recurrent Inference Machines for Solving Inverse Problems. *arXiv e-prints*, 2017.
- N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. A. Hamprecht, Y. Bengio, et A. Courville. On the Spectral Bias of Neural Networks. *arXiv e-prints*, art. arXiv :1806.08734, June 2018.
- C. Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin fo the Calcutta Mathematical Society*, 1945.
- R. Ratcliff. Connectionist models of recognition memory : Constraints imposed by learning and forgetting functions. *Psychological Review*, 97(2) :285–308, 1990. doi : 10.1037/0033-295X.97.2.285.
- S. Rau, S. Vegetti, et S. D. White. The effect of particle noise in N-body simulations of gravitational lensing. *Monthly Notices of the Royal Astronomical Society*, 430(3) :2232–2248, apr 2013. ISSN 13652966. doi : 10.1093/mnras/stt043.
- A. Refregier, A. Amara, T. D. Kitching, A. Rassat, R. Scaramella, et J. Weller. Euclid Imaging Consortium Science Book. *arXiv e-prints*, art. arXiv :1001.0061, Jan. 2010.
- S. Refsdal. On the possibility of determining Hubble's parameter and the masses of galaxies from the gravitational lens effect. *MNRAS*, 128 :307, Jan. 1964. doi : 10.1093/mnras/128.4.307.
- F. Rizzo, S. Vegetti, D. Powell, F. Frernali, J. P. McKean, H. R. Stacey, et S. D. M. White. A dynamically cold disk galaxy in the early Universe. *Nature*, 584(7820) :201–204, Aug. 2020. doi : 10.1038/s41586-020-2572-6.
- O. Ronneberger, P. Fischer, et T. Brox. U-Net : Convolutional Networks for Biomedical Image Segmentation. *arXiv e-prints*, art. arXiv :1505.04597, May 2015.
- F. Rosenblatt. The perceptron : A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6) :386–408, 1958. ISSN 0033-295X. doi : 10.1037/h0042519. URL <http://dx.doi.org/10.1037/h0042519>.
- B. T. Rowe, M. Jarvis, R. Mandelbaum, G. M. Bernstein, J. Bosch, M. Simet, J. E. Meyers, T. Kacprzak, R. Nakajima, J. Zuntz, H. Miyatake, J. P. Dietrich, R. Armstrong, P. Melchior, et M. S. Gill. GalSim : The modular galaxy image simulation toolkit. *Astronomy and Computing*, 10 :121–150, apr 2015. ISSN 22131337. doi : 10.1016/j.ascom.2015.02.002.

- B. T. P. Rowe, M. Jarvis, R. Mandelbaum, G. M. Bernstein, J. Bosch, M. Simet, J. E. Meyers, T. Kacprzak, R. Nakajima, J. Zuntz, H. Miyatake, J. P. Dietrich, R. Armstrong, P. Melchior, et M. S. S. Gill. GALSIM : The modular galaxy image simulation toolkit. *Astronomy and Computing*, 10 :121–150, Apr. 2015. doi : 10.1016/j.ascom.2015.02.002.
- C. E. Rusu, C. D. Fassnacht, D. Sluse, S. Hilbert, K. C. Wong, K.-H. Huang, S. H. Suyu, T. E. Collett, P. J. Marshall, T. Treu, et L. V. E. Koopmans. H0LiCOW - III. Quantifying the effect of mass along the line of sight to the gravitational lens HE 0435-1223 through weighted galaxy counts *. *MNRAS*, 467(4) :4220–4242, June 2017. doi : 10.1093/mnras/stx285.
- C. E. Rusu, K. C. Wong, V. Bonvin, D. Sluse, S. H. Suyu, C. D. Fassnacht, J. H. H. Chan, S. Hilbert, M. W. Auger, A. Sonnenfeld, S. Birrer, F. Courbin, T. Treu, G. C. F. Chen, A. Halkola, L. V. E. Koopmans, P. J. Marshall, et A. J. Shajib. H0LiCOW XII. Lens mass model of WFI2033-4723 and blind measurement of its time-delay distance and H_0 . *MNRAS*, 498(1) :1440–1468, Oct. 2020. doi : 10.1093/mnras/stz3451.
- P. Saha et L. L. R. Williams. Non-parametric reconstruction of the galaxy lens in PG 1115+080. *MNRAS*, 292(1) :148–156, Nov. 1997. doi : 10.1093/mnras/292.1.148.
- P. Saha et L. L. R. Williams. A Portable Modeler of Lensed Quasars. *AJ*, 127(5) :2604–2616, May 2004. doi : 10.1086/383544.
- K. Scaman et A. Virmaux. Lipschitz regularity of deep neural networks : analysis and efficient estimation. *arXiv e-prints*, art. arXiv :1805.10965, May 2018.
- T. Schmidt, T. Treu, S. Birrer, A. J. Shajib, C. Lemon, M. Millon, D. Sluse, A. Agnello, T. Anguita, M. W. Auger-Williams, R. G. McMahon, V. Motta, P. Schechter, C. Spiniello, I. Kayo, F. Courbin, S. Ertl, C. D. Fassnacht, J. A. Frieman, A. More, S. Schuldt, S. H. Suyu, M. Aguena, F. Andrade-Oliveira, J. Annis, D. Bacon, E. Bertin, D. Brooks, D. L. Burke, A. Carnero Rosell, M. Carrasco Kind, J. Carretero, C. Conselice, M. Costanzi, L. N. da Costa, M. E. S. Pereira, J. De Vicente, S. Desai, P. Doel, S. Everett, I. Ferrero, D. Friedel, J. García-Bellido, E. Gaztanaga, D. Gruen, R. A. Gruendl, J. Gschwend, G. Gutierrez, S. R. Hinton, D. L. Hollowood, K. Honscheid, D. J. James, K. Kuehn, O. Lahav, F. Menanteau, R. Miquel, A. Palmese, F. Paz-Chinchón, A. Pieres, A. A. Plazas Malagón, J. Prat, M. Rodriguez-Monroy, A. K. Romer, E. Sanchez, V. Scarpine, I. Sevilla-Noarbe, M. Smith, E. Suchyta, G. Tarle, C. To, et T. N. Varga. STRIDES : Automated uniform models for 30 quadruply imaged quasars. *arXiv e-prints*, art. arXiv :2206.04696, June 2022.
- S. Schuldt, G. Chirivi, S. H. Suyu, A. Yıldırım, A. Sonnenfeld, A. Halkola, et G. F. Lewis. Inner dark matter distribution of the Cosmic Horseshoe (J1148+1930) with gravitational lensing and dynamics. *A&A*, 631 :A40, Nov. 2019. doi : 10.1051/0004-6361/201935042.
- S. Schuldt, S. H. Suyu, R. Canameras, Y. Shu, S. Taubenberger, S. Ertl, et A. Halkola. HOLISMOKES – X. Comparison between neural network and semi-automated traditional modeling of strong lenses. *arXiv e-prints*, art. arXiv :2207.10124, July 2022.
- N. Scoville, H. Aussel, M. Brusa, P. Capak, C. M. Carollo, M. Elvis, M. Giavalisco, L. Guzzo, G. Hasinger, C. Impey, J.-P. Kneib, O. LeFevre, S. J. Lilly, B. Mobasher, A. Renzini, R. M. Rich, D. B. Sanders, E. Schinnerer, D. Schminovich, P. Shopbell, Y. Taniguchi, et N. D. Tyson. The Cosmic Evolution Survey (COSMOS) : Overview. *The Astrophysical Journal Supplement Series*, 172(1) :1–8, sep 2007. ISSN 0067-0049. doi : 10.1086/516585.
- S. Seitz, P. Schneider, et M. Bartelmann. Entropy-regularized maximum-likelihood cluster mass reconstruction. *A&A*, 337 :325–337, Sept. 1998.
- J. L. Sérsic. Influence of the atmospheric and instrumental dispersion on the brightness distribution in a galaxy. *Boletín de la Asociacion Argentina de Astronomia La Plata Argentina*, 6 :41–43, Feb. 1963.
- C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27 :379–423, 1948. URL <http://plan9.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf>.
- Y. Shu, J. R. Brownstein, A. S. Bolton, L. V. E. Koopmans, T. Treu, A. D. Montero-Dorta, M. W. Auger, O. Czoske, R. Gavazzi, P. J. Marshall, et L. A. Moustakas. The Sloan Lens ACS Survey. XIII. Discovery of 40 New Galaxy-scale Strong Lenses. *ApJ*, 851(1) :48, Dec. 2017. doi : 10.3847/1538-4357/aa9794.
- H. T. Siegelmann et E. D. Sontag. On the computational power of neural nets. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, page 440–449, New York, NY, USA, 1992. Association for Computing Machinery. ISBN 089791497X. doi : 10.1145/130385.130432. URL <https://doi.org/10.1145/130385.130432>.

- D. Sluse, A. Sonnenfeld, N. Rumbaugh, C. E. Rusu, C. D. Fassnacht, T. Treu, S. H. Suyu, K. C. Wong, M. W. Auger, V. Bonvin, T. Collett, F. Courbin, S. Hilbert, L. V. E. Koopmans, P. J. Marshall, G. Meylan, C. Spinello, et M. Tewes. H0LiCOW - II. Spectroscopic survey and galaxy-group identification of the strong gravitational lens system HE 0435-1223. *MNRAS*, 470(4) :4838–4857, Oct. 2017. doi : 10.1093/mnras/stx1484.
- A. Stockton. The lens galaxy of the twin QSO 0957+561. *ApJ*, 242 :L141–L144, Dec. 1980. doi : 10.1086/183419.
- F. Sun, E. Egami, P. G. Pérez-González, I. Smail, K. I. Caputi, F. E. Bauer, T. D. Rawle, S. Fujimoto, K. Kohno, U. Dudzevičiūtė, H. Atek, M. Bianconi, S. C. Chapman, F. Combes, M. Jauzac, J.-B. Jolly, A. M. Koekemoer, G. E. Magdis, G. Rodighiero, W. Rujopakarn, D. Schaerer, C. L. Steinhardt, P. Van der Werf, G. L. Walth, et J. R. Weaver. Extensive Lensing Survey of Optical and Near-infrared Dark Objects (El Sonido) : HST H-faint Galaxies behind 101 Lensing Clusters. *ApJ*, 922(2) :114, Dec. 2021. doi : 10.3847/1538-4357/ac2578. URL <https://ui.adsabs.harvard.edu/abs/2021ApJ...922..114S>.
- S. H. Suyu, P. J. Marshall, M. P. Hobson, et R. D. Blandford. A Bayesian analysis of regularized source inversions in gravitational lensing. *MNRAS*, 371(2) :983–998, Sept. 2006. doi : 10.1111/j.1365-2966.2006.10733.x.
- C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, et R. Fergus. Intriguing properties of neural networks. *arXiv e-prints*, art. arXiv :1312.6199, Dec. 2013.
- N. Tishby, F. C. Pereira, et W. Bialek. The information bottleneck method. In *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999. URL <https://arxiv.org/abs/physics/0004057>.
- N. Tishby, F. C. Pereira, et W. Bialek. The information bottleneck method. *arXiv e-prints*, art. physics/0004057, Apr. 2000.
- D. A. Torres-Ballesteros et L. Castañeda. relensing : Reconstructing the mass profile of galaxy clusters from gravitational lensing. *arXiv e-prints*, art. arXiv :2201.10076, Jan. 2022.
- T. Treu. Strong Lensing by Galaxies. *ARA&A*, 48 :87–125, Sept. 2010. doi : 10.1146/annurev-astro-081309-130924.
- T. Treu et L. V. E. Koopmans. Massive Dark Matter Halos and Evolution of Early-Type Galaxies to $z \sim 1$. *ApJ*, 611(2) :739–760, Aug. 2004. doi : 10.1086/422245.
- S. Van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, et T. Yu. scikit-image : image processing in python. *PeerJ*, 2 :e453, 2014.
- C. Vanderriest, J. Schneider, G. Herpe, M. Chevreton, M. Moles, et G. Wlerick. The value of the time delay delta T (A,B) for the 'double' quasar 0957+561 from optical photometric monitoring. *A&A*, 215 :1–13, May 1989.
- V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., 1995. ISBN 0-387-94559-8.
- V. N. Vapnik et A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2) :264–280, 1971. doi : 10.1137/1116025. URL <https://doi.org/10.1137/1116025>.
- J. D. Vieira, D. P. Marrone, S. C. Chapman, C. De Breuck, Y. D. Hezaveh, A. Weiβ, J. E. Aguirre, K. A. Aird, M. Aravena, M. L. N. Ashby, M. Bayliss, B. A. Benson, A. D. Biggs, L. E. Bleem, J. J. Bock, M. Bothwell, C. M. Bradford, M. Brodwin, J. E. Carlstrom, C. L. Chang, T. M. Crawford, A. T. Crites, T. de Haan, M. A. Dobbs, E. B. Fomalont, C. D. Fassnacht, E. M. George, M. D. Gladders, A. H. Gonzalez, T. R. Greve, B. Gullberg, N. W. Halverson, F. W. High, G. P. Holder, W. L. Holzapfel, S. Hoover, J. D. Hrubes, T. R. Hunter, R. Keisler, A. T. Lee, E. M. Leitch, M. Lueker, D. Luong-van, M. Malkan, V. McIntyre, J. J. McMahon, J. Mehl, K. M. Menten, S. S. Meyer, L. M. Mocanu, E. J. Murphy, T. Natoli, S. Padin, T. Plagge, C. L. Reichardt, A. Rest, J. Ruel, J. E. Ruhl, K. Sharon, K. K. Schaffer, L. Shaw, E. Shirokoff, J. S. Spilker, B. Stalder, Z. Staniszewski, A. A. Stark, K. Story, K. Vanderlinde, N. Welikala, et R. Williamson. Dusty starburst galaxies in the early Universe as revealed by gravitational lensing. *Nature*, 495(7441) :344–347, Mar. 2013. doi : 10.1038/nature12001.
- P. Vincent, H. Larochelle, Y. Bengio, et P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 1096–1103, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582054. doi : 10.1145/1390156.1390294. URL <https://doi.org/10.1145/1390156.1390294>.
- P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, et P.-A. Manzagol. Stacked denoising autoencoders : Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.*, 11 :3371–3408, dec 2010. ISSN 1532-4435.

- P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, et SciPy 1.0 Contributors. SciPy 1.0 : Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17 :261–272, 2020. doi : 10.1038/s41592-019-0686-2.
- S. Wagner-Carena, J. W. Park, S. Birrer, P. J. Marshall, A. Roodman, R. H. Wechsler, et LSST Dark Energy Science Collaboration. Hierarchical Inference with Bayesian Neural Networks : An Application to Strong Gravitational Lensing. *ApJ*, 909(2) :187, Mar. 2021. doi : 10.3847/1538-4357/abdf59.
- S. Wagner-Carena, J. Aalbers, S. Birrer, E. O. Nadler, E. Darragh-Ford, P. J. Marshall, et R. H. Wechsler. From Images to Dark Matter : End-To-End Inference of Substructure From Hundreds of Strong Gravitational Lenses. *arXiv e-prints*, art. arXiv :2203.00690, Mar. 2022.
- D. Walsh, R. F. Carswell, et R. J. Weymann. 0957+561 A, B : twin quasistellar objects or gravitational lens ? *Nature*, 279 : 381–384, May 1979. doi : 10.1038/279381a0.
- S. J. Warren et S. Dye. Semilinear Gravitational Lens Inversion. *ApJ*, 590(2) :673–682, June 2003. doi : 10.1086/375132.
- T.-W. Weng, H. Zhang, P.-Y. Chen, J. Yi, D. Su, Y. Gao, C.-J. Hsieh, et L. Daniel. Evaluating the Robustness of Neural Networks : An Extreme Value Theory Approach. *arXiv e-prints*, art. arXiv :1801.10578, Jan. 2018.
- Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt et Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010. doi : 10.25080/Majora-92bf1922-00a.
- S. D. M. White et M. J. Rees. Core condensation in heavy halos : a two-stage theory for galaxy formation and clustering. *MNRAS*, 183 :341–358, May 1978. doi : 10.1093/mnras/183.3.341.
- D. Wolpert. On the connection between in-sample testing and generalization error. *Complex Systems*, 6, 01 1992.
- D. Wolpert et W. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1) :67–82, 1997. doi : 10.1109/4235.585893.
- D. H. Wolpert. The Lack of A Priori Distinctions Between Learning Algorithms. *Neural Computation*, 8(7) :1341–1390, 10 1996. ISSN 0899-7667. doi : 10.1162/neco.1996.8.7.1341. URL <https://doi.org/10.1162/neco.1996.8.7.1341>.
- K. C. Wong, S. H. Suyu, M. W. Auger, V. Bonvin, F. Courbin, C. D. Fassnacht, A. Halkola, C. E. Rusu, D. Sluse, A. Sonnenfeld, T. Treu, T. E. Collett, S. Hilbert, L. V. E. Koopmans, P. J. Marshall, et N. Rumbaugh. H0LiCOW - IV. Lens mass model of HE 0435-1223 and blind measurement of its time-delay distance for cosmology. *MNRAS*, 465(4) :4895–4913, Mar. 2017. doi : 10.1093/mnras/stw3077.
- K. C. Wong, S. H. Suyu, G. C. F. Chen, C. E. Rusu, M. Millon, D. Sluse, V. Bonvin, C. D. Fassnacht, S. Taubenberger, M. W. Auger, S. Birrer, J. H. H. Chan, F. Courbin, S. Hilbert, O. Tihhonova, T. Treu, A. Agnello, X. Ding, I. Jee, E. Komatsu, A. J. Shajib, A. Sonnenfeld, R. D. Blandford, L. V. E. Koopmans, P. J. Marshall, et G. Meylan. H0LiCOW - XIII. A 2.4 per cent measurement of H_0 from lensed quasars : 5.3σ tension between early- and late-Universe probes. *MNRAS*, 498(1) : 1420–1439, Oct. 2020. doi : 10.1093/mnras/stz3094.
- P. Young, J. E. Gunn, J. Kristian, J. B. Oke, et J. A. Westphal. The double quasar Q0957+561 A, B : a gravitational lens image formed by a galaxy at $z=0.39$. *ApJ*, 241 :507–520, Oct. 1980. doi : 10.1086/158365.
- P. Young, J. E. Gunn, J. Kristian, J. B. Oke, et J. A. Westphal. Q0957+561 : detailed models of the gravitational lens effect. *ApJ*, 244 :736–755, Mar. 1981. doi : 10.1086/158751.
- A. Younger, S. Hochreiter, et P. Conwell. Meta-learning with backpropagation. In *IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No.01CH37222)*, volume 3, pages 2001–2006 vol.3, 2001. doi : 10.1109/IJCNN.2001.938471.
- S. Zhao, J. Song, et S. Ermon. InfoVAE : Information Maximizing Variational Autoencoders. *arXiv e-prints*, art. arXiv :1706.02262, June 2017.

F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, et Q. He. A Comprehensive Survey on Transfer Learning. *arXiv e-prints*, art. arXiv :1911.02685, Nov. 2019.

F. Zwicky. Nebulae as gravitational lenses. *Phys. Rev.*, 51 :290–290, Feb 1937. doi : 10.1103/PhysRev.51.290. URL <https://link.aps.org/doi/10.1103/PhysRev.51.290>.

F. Zwicky. On the Masses of Nebulae and of Clusters of Nebulae. *ApJ*, 86 :217, Oct. 1937. doi : 10.1086/143864.

Annexe A

Λ CDM

TABLE A.1 – Paramètres de Λ CDM ajusté avec les observations du fond diffus cosmologique par le télescope Planck ([Planck Collaboration, 2020](#))

Paramètre	Description	Valeur
$\Omega_{r,0}$	Densité de la radiation	$\sim 10^{-4}$
$\Omega_{m,0}$	Densité de la matière	0.3158
$\Omega_{c,0}h^2$	Densité de la matière noire	0.12011
$\Omega_{b,0}h^2$	Densité de la matière baryonique	0.022383
$\Omega_{\Lambda,0}$	Densité de l'énergie sombre	0.6842
Ω_0	Densité totale	$\equiv 1$
h	Constante de Hubble $h \equiv \frac{H_0}{100 \text{ km s}^{-1} \text{ Mpc}^{-1}}$	0.6732

Appendix B

Elastic Weight Consolidation

Suppose we are given a training set \mathcal{D} and a test task \mathcal{T} . The posterior of the RIM parameters φ can be rewritten using the Bayes rule as

$$p(\varphi \mid \mathcal{D}, \mathcal{T}) = \frac{p(\mathcal{T} \mid \mathcal{D}, \varphi)p(\varphi \mid \mathcal{D})}{p(\mathcal{T} \mid \mathcal{D})}. \quad (\text{B.1})$$

We suppose that φ encode information about \mathcal{D} , while \mathcal{T} was unseen by φ . It follows that \mathcal{T} and \mathcal{D} are conditionally independent when given φ . We do not make the stronger assumption that \mathcal{D} and \mathcal{T} are completely independent. In fact, such an assumption would contradict the premiss of our work that building a dataset \mathcal{D} can inform a machine (RIM) about task \mathcal{T} — or that, more broadly, \mathcal{D} contains information about \mathcal{T} .

We rewrite the marginal $p(\mathcal{T} \mid \mathcal{D})$ using the Bayes rule in order to extract $p(\mathcal{D} \mid \mathcal{T})$, the sampling distribution used to compute the Fisher diagonal elements

$$p(\varphi \mid \mathcal{D}, \mathcal{T}) = \frac{p(\mathcal{T} \mid \varphi)p(\varphi \mid \mathcal{D})}{p(\mathcal{D} \mid \mathcal{T})} \frac{p(\mathcal{D})}{p(\mathcal{T})}. \quad (\text{B.2})$$

The log-likelihood $\log p(\mathcal{T} \mid \varphi)$ is equivalent to the negative of the loss function for the particular task at hand. In this work, we assign a uniform probability density to $p(\mathcal{T})$ and $p(\mathcal{D})$ in order to ignore them.

We now turn to the prior $p(\varphi \mid \mathcal{D})$, which appears as a conditional relative to the training dataset. We use the Laplace approximation around the maxima $\varphi_{\mathcal{D}}^*$ to evaluate the prior, where $\varphi_{\mathcal{D}}^*$ are the trained parameters of the RIM that minimize the empirical risk (equation (5.8)). The Taylor expansion of the prior around this maxima yields

$$\log p(\varphi \mid \mathcal{D}) \approx \log p(\varphi_{\mathcal{D}}^* \mid \mathcal{D}) + \underbrace{\frac{1}{2}(\varphi - \varphi_{\mathcal{D}}^*)^T \left(\frac{\partial^2 \log p(\varphi \mid \mathcal{D})}{\partial^2 \varphi} \Big|_{\varphi_{\mathcal{D}}^*} \right)}_{\mathbf{H}(\varphi_{\mathcal{D}}^*)} (\varphi - \varphi_{\mathcal{D}}^*). \quad (\text{B.3})$$

Since $\varphi_{\mathcal{D}}^*$ is an extrema of the prior, the linear term vanishes. The empirical estimate of the negative hessian matrix is the observed Fisher information matrix which can be written as

$$\mathcal{I}(\varphi_{\mathcal{D}}^*) = -\mathbb{E}_{\mathcal{D}|\mathcal{T}}[\mathbf{H}(\varphi_{\mathcal{D}}^*)] = \mathbb{E}_{\mathcal{D}|\mathcal{T}} \left[\left(\left(\frac{\partial \log p(\varphi | \mathcal{D})}{\partial \varphi} \right) \left(\frac{\partial \log p(\varphi | \mathcal{D})}{\partial \varphi} \right)^T \right) \Big|_{\varphi_{\mathcal{D}}^*} \right]. \quad (\text{B.4})$$

The expectation is taken over the sample space $p(\mathcal{D} | \mathcal{T})$ since the network parameters are held fixed during sampling. In order to compute the Fisher score, we apply the Bayes rule to the prior to extract a loss function, which we take to be proportional to the training loss (equation (5.7)) and the χ^2 :

$$\log p(\varphi | (\mathbf{x}, \mathbf{y}) = \mathcal{D}) \propto -\mathcal{L}_\varphi(\mathbf{x}, \mathbf{y}) + \frac{1}{T} \sum_{t=1}^T \log p(\mathbf{y} | \hat{\mathbf{x}}^{(t)}) - \frac{\ell_2}{2} \|\varphi\|_2^2 \quad (\text{B.5})$$

We find in practice the the ℓ_2 term has little effect on the Fisher diagonal and our results. Thus, we set $\ell_2 = 0$.

Since the full Fisher matrix is intractable for a neural network, we approximate the quadratic term of the prior with the diagonal of the Fisher matrix following [Kirkpatrick et al. \(2016\)](#). For an optimisation problem, the first term of (B.3) is constant. Thus, the posterior becomes proportional to

$$\log p(\varphi | \mathcal{D}, \mathcal{T}) \propto \log p(\mathcal{T} | \varphi) - \frac{\lambda}{2} \sum_j \text{diag}(\mathcal{I}(\varphi_{\mathcal{D}}^*))_j (\varphi_j - [\varphi_{\mathcal{D}}^*]_j)^2. \quad (\text{B.6})$$

The Lagrange multiplier λ is introduced to tune our uncertainty about the network parameters during fine-tuning.

Appendix C

VAE Architecture and optimisation

For the following architectures, we employ the notion of *level* to mean layers in the encoder and the decoder with the same resolution. In each level, we place a block of convolutional layers before downsampling (encoder) or after upsampling (decoder). These operations are done with strided convolutions like in the U-net architecture of the RIM.

Table C.1 – Hyperparameters for the background source VAE.

Parameter	Value
Input preprocessing	1
<i>Architecture</i>	
Levels (encoder and decoder)	3
Convolutional layer per level	2
Latent space dimension	32
Hidden Activations	Leaky ReLU
Output Activation	Sigmoid
Filters (first level)	16
Filters scaling factor (per level)	2
Number of parameters	3 567 361
<i>Optimization</i>	
Optimizer	Adam
Initial learning rate	10^{-4}
Learning rate schedule	Exponential Decay
Decay rate	0.5
Decay steps	30 000
Number of steps	500 000
β_{\max}	0.1
Batch size	20

Table C.2 – Hyperparameters for the convergence VAE.

Parameter	Value
Input preprocessing	\log_{10}
<i>Architecture</i>	
Levels (encoder and decoder)	4
Convolutional layer per level	1
Latent space dimension	16
Hidden Activations	Leaky ReLU
Output Activation	tanh
Filters (first level)	16
Filters scaling factor (per level)	2
Number of parameters	1 980 033
<i>Optimization</i>	
Optimizer	Adam
Initial learning rate	10^{-4}
Learning rate schedule	Exponential Decay
Decay rate	0.7
Decay steps	20 000
Number of steps	155 000
β_{\max}	0.2
Batch size	32

Appendix D

RIM architecture and optimisation

The notion of link function $\Psi : \Xi \rightarrow \mathcal{X}$, introduced by Putzky and Welling (2017), is an invertible transformation between the network prediction space $\xi \in \Xi$ and the forward modelling space $\mathbf{x} \in \mathcal{X}$. This is a different notion from preprocessing, discussed in section 5.3, because this transformation is applied inside the recurrent relation 5.6 as opposed to before training. In the case where the forward model has some restricted support or it is found that some transformation helps the training, then the link function chosen must be implemented as part of the network architecture as shown in the unrolled computational graph in Figure D.1. Also, the loss \mathcal{L}_φ must be computed in the Ξ space in order to avoid gradient vanishing problems when Ψ is a non-linear mapping, which happens if the non-linear link function is applied in an operation recorded for backpropagation through time (BPTT). For the convergence, we use an exponential link function with base 10: $\hat{\kappa} = \Psi(\xi) = 10^\xi$. This Ψ encodes the non-negativity of the convergence. Furthermore, it is a power transformation that leaves the linked pixel values ξ_i normally distributed, thus improving the learning through the non-linearities in the neural network.

The pixel weights \mathbf{w}_i in the loss function (5.7) are chosen to encode the fact that the pixel with critical mass density ($\kappa_i > 1$) have a stronger effect on the lensing configuration than other pixels. We find in practice that the weights

$$\mathbf{w}_i = \frac{\sqrt{\kappa_i}}{\sum_i \kappa_i}, \quad (\text{D.1})$$

encode this knowledge in the loss function and improved both the empirical risk and the goodness of fit of the baseline model on early test runs.

For the source, we found that we do not need a link function — the identity is generally better compared to other link function we tried like sigmoid and power transforms — and we found that the pixel weights can be taken to be uniform, i.e. $\mathbf{w}_i = \frac{1}{M}$.

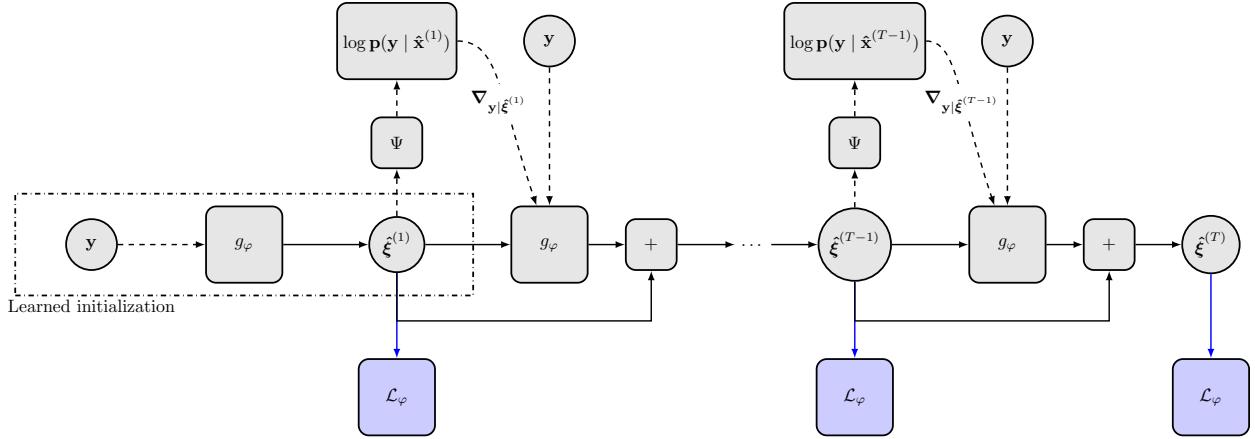


Figure D.1 – Unrolled computational graph of the RIM. Operations along solid arrows are being recorded for BPTT, while operations along dashed arrows are not. The blue arrows are only used for optimisation during training. During fine-tuning or testing, the loss is computed only as an oracle metric to validate that our methods can recover the ground truth.

Table D.1 – Hyperparameters for the RIM.

Parameter	Value
Source link function	1
κ link function	10^{ϵ}
<i>Architecture</i>	
Recurrent steps (T)	8
Number of parameters	348 546 818
<i>First Stage Optimisation</i>	
Optimizer	Adamax
Initial learning rate	10^{-4}
Learning rate schedule	Exponential Decay
Decay rate	0.95
Decay steps	100 000
Number of steps	610 000
Batch size	1
<i>Second Stage Optimisation</i>	
Optimizer	Adamax
Initial learning rate	6×10^{-5}
Learning rate schedule	Exponential Decay
Decay rate	0.9
Decay steps	100 000
Number of steps	870 000
Batch size	1

Annexe E

GRU

Une unité récurrente à porte convolutionnelles est décrite par les opérations

$$\tilde{\mathbf{x}} = S\left(\mathbf{w}_o * (\mathbf{h}^{(t-1)} \oplus \mathbf{x}^{(t)}) + \mathbf{b}_o\right) \quad \{\text{Porte d'oubli}\} \quad (\text{E.1})$$

$$\mathbf{z} = S\left(\mathbf{w}_z * (\mathbf{h}^{(t-1)} \oplus \mathbf{x}^{(t)}) + \mathbf{b}_z\right) \quad \{\text{Porte de mise à jour}\} \quad (\text{E.2})$$

$$\tilde{\mathbf{h}} = \tanh\left(\mathbf{w}_h * ((\mathbf{h}^{(t-1)} \odot \tilde{\mathbf{x}}) \oplus \mathbf{x}^{(t)})) + \mathbf{b}_h\right) \quad \{\text{État candidat}\} \quad (\text{E.3})$$

$$\mathbf{h}^{(t)} = \mathbf{h}^{(t-1)} \odot \mathbf{z} + \tilde{\mathbf{h}} \odot (1 - \mathbf{z}) \quad \{\text{Nouvel état}\} \quad (\text{E.4})$$

où $S(x) = \frac{1}{1+e^{-x}}$ est une fonction sigmoïde et $\mathbf{x}^{(t)}$ est un tenseur à l'entrée de l'unité. Les noyaux de convolution \mathbf{w} et les vecteurs de biais \mathbf{b} sont des paramètres libres appris par descente de gradient stochastique. \oplus symbolise l'opération de concatenation. Le tenseur de sortie de cette unité, soit le nouvel état latent $\mathbf{h}^{(t)}$, est une combinaison de l'état latent précédent $\mathbf{h}^{(t-1)}$ et de l'état candidat $\tilde{\mathbf{h}}$, pesée élément par élément par le vecteur à la sortie de la porte de mise à jour \mathbf{z} .

Annexe F

Congrès où l'étudiant à présenté ses résultats

IVADO Digital October

Médium : Présentateur

Titre : Pixelated Reconstruction of Gravitational Lenses using Recurrent Inference Machine

Lieu : En ligne

Année : 2021

Auteurs : A. Adam, L Perreault-Levasseur et Y. Hezaveh

Likelihood-free in Paris

Médium : Affiche

Titre : Pixelated Reconstruction of Gravitational Lenses using Recurrent Inference Machine

Lieu : Paris, France

Année : 2022

Auteurs : A. Adam, L Perreault-Levasseur et Y. Hezaveh

CRAQ Annual Meeting

Médium : Présentateur

Titre : Pixelated Reconstruction of Gravitational Lenses using Recurrent Inference Machine

Lieu : Orford, Qc

Année : 2022

Auteurs : A. Adam, L Perreault-Levasseur et Y. Hezaveh

CASCA Annual Meeting

Médium : Affiche

Titre : Pixelated Reconstruction of Gravitational Lenses using Recurrent Inference Machine

Lieu : En ligne

Année : 2022

Auteurs : A. Adam, L Perreault-Levasseur et Y. Hezaveh

Machine Learning for Astrophysics Workshop at the Thirty-ninth International Conference on Machine Learning (ICML 2022)

Médium : Affiche

Titre : Pixelated Reconstruction of Gravitational Lenses using Recurrent Inference Machine

Lieu : Baltimore, MD

Année : 2022

Auteurs : A. Adam, L Perreault-Levasseur et Y. Hezaveh

Boom! A Workshop on Explosive Transients with LSST

Médium : Présentation

Titre : Pixelated Reconstruction of Gravitational Lenses using Recurrent Inference Machine

Lieu : En ligne

Année : 2022

Auteurs : A. Adam, L Perreault-Levasseur et Y. Hezaveh

IAIFI Summer School & Workshop

Médium : Affiche

Titre : Pixelated Reconstruction of Gravitational Lenses using Recurrent Inference Machine

Lieu : Boston, MA

Année : 2022

Auteurs : A. Adam, L Perreault-Levasseur et Y. Hezaveh