

**LA RECONSTRUCTION D'IMAGE DANS LE CONTEXTE DES  
PROBLÈMES INVERSES MAL POSÉES ET NON-LINÉAIRES  
AVEC LES MACHINES À INFÉRENCE RÉCURRENTIELLES**

par

Alexandre Adam

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

Département de physique  
Université de Montréal



## Résumé

## **Abstract**

# Table des matières

<b>Résumé</b>	<b>ii</b>
<b>Abstract</b>	<b>iii</b>
<b>Liste des tableaux</b>	<b>vi</b>
<b>Liste des figures</b>	<b>vii</b>
<b>Acronymes</b>	<b>viii</b>
<b>Liste des symboles</b>	<b>ix</b>
<b>Remerciements</b>	<b>xii</b>
<b>1 Introduction</b>	<b>2</b>
1.1 Lentilles gravitationnelles fortes de type galaxie-galaxie . . . . .	2
1.1.1 Les angles de déflections . . . . .	5
1.2 Interférométrie par masque non-régulier . . . . .	9
1.2.1 Les angles de fermeture . . . . .	9
1.3 Auto-encodeur variationnel . . . . .	9
1.3.1 Description du modèle . . . . .	9
1.3.2 Le truc de reparamétrisation . . . . .	10
1.3.3 Principe du goulot d'information . . . . .	12
1.4 Machines à inférence récurrentielles . . . . .	15
1.4.1 Formalisme bayésien des problèmes inverses . . . . .	15
1.4.2 La relation de récurrence . . . . .	17

1.4.3 Méta-apprentissage . . . . .	18
<b>Bibliographie</b>	<b>21</b>
<b>A <math>\Lambda</math>CDM</b>	<b>26</b>
<b>B Elastic Weight Consolidation</b>	<b>27</b>
<b>C VAE Architecture and optimisation</b>	<b>29</b>
<b>D RIM architecture and optimisation</b>	<b>32</b>
<b>E GRU</b>	<b>36</b>

# Liste des tableaux

A.1	Paramètres de $\Lambda$ CDM ajusté avec les observations du fond diffus cosmologique par le télescope Planck (Planck Collaboration et al., 2020) . . . . .	26
C.1	Hyperparameters for the background source VAE. . . . .	30
C.2	Hyperparameters for the convergence VAE. . . . .	31
D.1	Hyperparameters for the RIM. . . . .	34

# Liste des figures

1.1	Le quasar double (QSO 0957+561 A et B) et la galaxie-lentille (G) imagé par le télescope spatial Hubble. Crédit : ESA/Hubble et NASA, enlagement par AA. . . . .	3
1.2	Lentilles gravitationnelles de type galaxie-galaxie. . . . .	4
1.3	Schéma d'une lentille gravitationnelle. . . . .	9
1.4	Modèle graphique d'un VAE. Les flèches pleines indiquent le processus génératif, alors que les flèches pointillées indiquent le processus d'inférence. . . . .	10
1.5	VAE comme un système de transmission d'information. . . . .	13
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. . . . .	33
D.2	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 ??.	35

# Acronymes

**RIM** Recurrent Inference Machine — Machine à inférence récurrentielles.

**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.

**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.

**HST** Hubble Space Telescope.

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

**WFC3** Wide Field Camera 3.

**KL** Kullback-Leibler.

# Liste des symboles

- $\mathbb{1}$  Matrice identité.
- $\mathbf{1}$  Vecteur dont chaque élément correspond à la valeur 1.
- $\mathbb{R}$  Ensemble des nombres réels.
- $\pi$  Pi.
- $\nabla$  Gradient.
- $\nabla^2$  Laplacien.
- $\kappa$  Convergence — densité surfacique de masse projeté sur l'axe de visée.
- $\alpha$  Angles de déflections.
- $\beta$  Coordonnées angulaires du plan de la source.
- $\theta$  Coordonnées angulaires du plan de la lentille.
- $\xi$  Coordonnées comobiles sur le plan de la lentille.
- $\eta$  Coordonnées comobiles sur le plan de la source.
- $D_s$  Distance du diamètre angulaire entre l'observateur et la source.
- $D_\ell$  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.
- $\rho$  Densité.
- $\Sigma$  Densité de surface.
- $\Sigma_c$  Densité de surface critique.
- $\Phi$  Potentiel.

- $\varphi$  Liste des paramètres pour l'algorithme d'inférence d'un problème inverse.  
 $\phi$  Liste des paramètres pour un processus d'inférence.  
 $\theta$  Liste des paramètres pour un processus génératif.  
 $\hat{\mathbf{x}}^{(t)}$  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.  
 $\mathcal{T}$  Ensemble de données d'essai.  
 $\mathcal{I}$  Information de Fisher.  
**H** Hessienne.  
 $D_{\text{KL}}(\cdot \parallel \cdot)$  Distance de Kullback-Leibler.  
 $\mathbb{E}_{P(X)}[\cdot]$  Opérateur de l'espérance mathématique par rapport à la variable aléatoire  $X$  distribué selon  $P(X)$ .  
i.i.d. Identiquement et indépendamment distribué.  
 $\|\cdot\|_2$  Norme euclidienne.  
 $I(X; Y)$  Information mutuelle entre les variables aléatoires  $X$  et  $Y$ .  
 $\mathcal{L}_\varphi$  Fonction objective d'entraînement pour les paramètres  $\varphi$ .  
 $\mathcal{N}$  Loi normale.  
 $\mathcal{TN}$  Loi normale tronquée.  
 $\mathcal{U}$  Loi uniforme.  
 $\boldsymbol{\mu}$  Moyenne.  
 $\boldsymbol{\Sigma}$  Covariance.  
 $\sigma^2$  Variance.  
 $\sigma$  Déviation standard.  
 $\oplus$  Concaténation.  
 $\odot$  Produit d'Hadamard.

À Maman et Julia

## **Remerciements**





# Chapitre 1

## Introduction

### 1.1 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 information<sup>1</sup>. 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 déflection de l'image de la source est directement proportionnelle à la masse (gravitationnelle) de la lentille.

$$\theta_E = \sqrt{\frac{4GM}{c^2 D}} \simeq 3 \left( \frac{M}{M_\odot} \right)^{\frac{1}{2}} \left( \frac{D}{1 \text{ Gpc}} \right)^{-\frac{1}{2}} \mu\text{as} \quad \left\{ D \equiv \frac{D_\ell D_s}{D_{\ell s}} \right\}. \quad (1.1)$$

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 masses des nébuleuses extra-galactiques (galaxies) de l'époque, basées largement sur des comparaisons de la luminosité totales de ces galaxies avec  $L_\odot$ , la luminosité du Soleil, ou des courbes de rotation képlériennes.<sup>2</sup>

---

1. Les travaux pionniers de František Link (1936, 1937), largement ignoré 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. La masse de l'amas de Coma estimée à  $\gtrsim 4.5 \times 10^{13} M_\odot$  avec le théorème du viriel par Zwicky (1937) est largement en accord avec la valeur acceptée aujourd'hui par la mesure des effets de lentilles faibles produite par l'amas sur l'image des galaxies environnantes ( $5_{-2.1}^{+4.3} \times 10^{14} h_{70}^{-1} M_\odot$ , Gavazzi et al., 2009).

Il est intéressant de noter qu'[Einstein \(1936\)](#) considérait l'éventualité d'observer ces systèmes extrêmement improbable, pointant vers les limitations instrumentales de l'époque. En effet, les télescopes terrestres étaient largement limité par l'effet de *seeing* atmosphérique, soit la distortion de l'image causé par la turbulence de l'atmosphère.

Dû à cette difficulté pratique, la première lentille gravitationnelle est découverte 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, 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, produites 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](#))<sup>3</sup> ainsi que la modélisation de sa distribution de masse, de son environnement et des angles de déflection qui causerait l'apparition d'une image double du quasar ([Young et al., 1981; Falco et al., 1991](#))

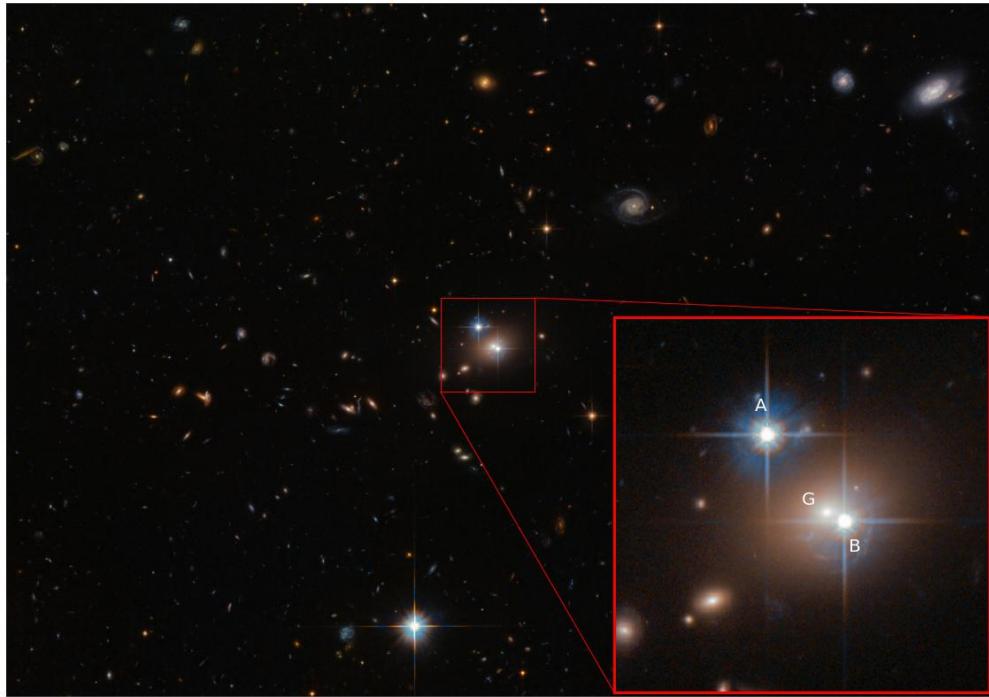
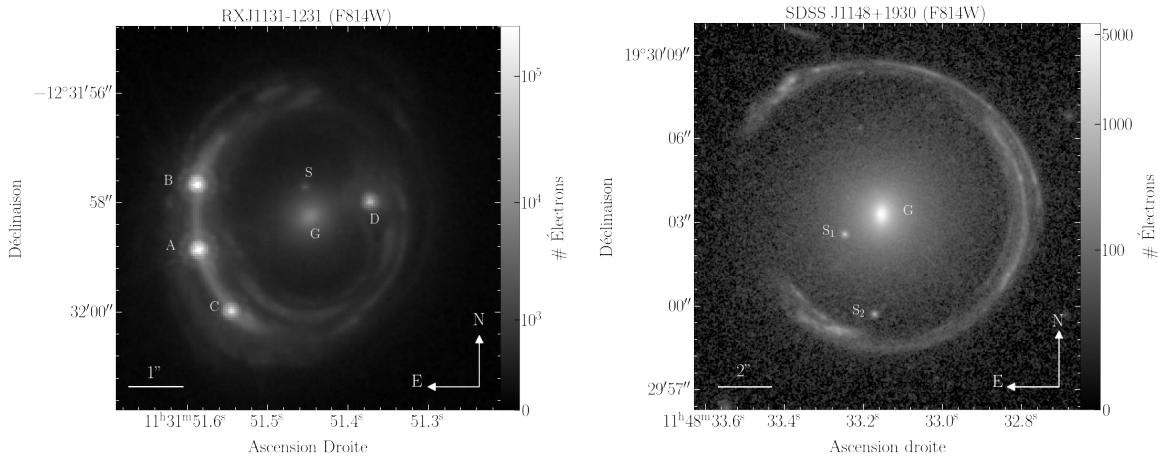


FIGURE 1.1 – Le quasar double (QSO 0957+561 A et B) et la galaxie-lentille (G) imaged par le télescope spatial Hubble. Crédit : ESA/Hubble et NASA, agrandissement par AA.

À la suite de cette découverte fortuite, l'étude des lentilles gravitationnelles est devenu un sujet

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

d'étude particulièrement riche, avec beaucoup de promesses pour la cosmologie (Blandford and Narayan, 1992; Bartelmann, 2010; Treu, 2010). De ce fait, beaucoup d'effort ont été déployés pour trouver systématiquement ces objets dans le ciel. Le programme *Sloan Lens ACS Survey* (SLAC, Bolton et al., 2005; Bolton et al., 2006), basé sur la recherche systématique de spectres de galaxies de type ETG<sup>4</sup> avec des lignes d'absorptions à un décalage vers le rouge plus grand que les lignes d'émission, est un des programmes les plus réussi, ayant mené à la découverte confirmée de plus de 150 lentilles gravitationnelles de type galaxie-galaxie (Bolton et al., 2008; Shu et al., 2017). Les programmes basés sur la recherche visuelle d'images double, triples, d'arcs ou d'anneau (e.g. Faure et al., 2008) dans les champs 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 la perception visuelle (Krizhevsky et al., 2012). Cette nouvelle approche à déjà menée 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 avec les nouvelles expéditions à champs large comme les observatoires Rubin (LSST Science Collaboration et al., 2009) et Euclid (Refregier et al., 2010).



(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 1.2 – Lentilles gravitationnelles de type galaxie-galaxie.

Le sujet du chapitre ?? se concentre sur le défi de développer un algorithme pour modéliser la distribution de masse et la morphologie de la source dédié à analyser un nombre aussi grand de lentilles gravitationnelles dans toute leur complexité et dans un temps à l'échelle humaine. Comme introduction à ce travail, je dérive les équations centrales qui nous permettent d'étudier les lentilles

4. Early-Type Galaxies

gravitationnelles de type galaxie-galaxie. Mon traitement est largement inspiré des manuels de références de [Meneghetti \(2013\)](#) et [Congdon and Keeton \(2018\)](#).

### 1.1.1 Les angles de déflections

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  $\Phi$  est situé sur l'axe de visée, ce qui a pour effet de courber la trajectoire de ce photon entre son point d'origine  $A$  et son point d'arrivée  $B$ . 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

$$\alpha = - \int_{\lambda_A}^{\lambda_B} \ddot{\mathbf{x}} \times \mathbf{e}_{\parallel} d\lambda, \quad (1.2)$$

où  $\lambda$  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 est sujette 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

$$\delta T = \delta \int_A^B n(\mathbf{x}(\ell)) \frac{d\ell}{c} = 0, \quad (1.3)$$

où  $\ell$  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 décrit 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  $\Phi$  d'une galaxie est celui d'un gaz parfait, c'est-à-dire qu'il satisfait une équation de Poisson

$$\nabla^2 \Phi = 4\pi G \rho. \quad (1.4)$$

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

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu \approx \left(1 + \frac{2\Phi}{c^2}\right) c^2 dt^2 - \left(1 - \frac{2\Phi}{c^2}\right) d\mathbf{x}^2. \quad (1.5)$$

Puisqu'un photon suit une géodésique de l'espace-temps  $ds^2 = 0$ , on peut déterminer l'indice de

réfraction en réarrangeant l'équation (1.5)

$$n \equiv c \left( \frac{\|d\mathbf{x}\|}{dt} \right)^{-1} \approx 1 - \frac{2\Phi}{c^2}. \quad (1.6)$$

En réécrivant l'élément de longueur  $d\ell$  en terme du paramètre de la trajectoire  $d\ell = \|\frac{d\mathbf{x}}{d\lambda}\| d\lambda$ , on peut réécrire l'équation (1.3) sous la forme

$$\delta \int_{\lambda_A}^{\lambda_B} n(\mathbf{x}) \|\dot{\mathbf{x}}\| d\lambda = 0. \quad (1.7)$$

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 (1.3) est une solution des équations d'Euler-Lagrange

$$\frac{d}{d\lambda} \frac{\partial \mathcal{L}}{\partial \dot{\mathbf{x}}} - \frac{\partial \mathcal{L}}{\partial \mathbf{x}} = 0. \quad (1.8)$$

On a donc

$$\frac{d}{d\lambda} n \frac{\dot{\mathbf{x}}}{\|\dot{\mathbf{x}}\|} - \|\dot{\mathbf{x}}\| \nabla n = 0, \quad (1.9)$$

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

$$\begin{aligned} \frac{d}{d\lambda} n \dot{\mathbf{x}} - \nabla n &= 0 \\ \implies n \ddot{\mathbf{x}} + (\nabla n \cdot \dot{\mathbf{x}}) \dot{\mathbf{x}} - \nabla n &= 0 \end{aligned} \quad (1.10)$$

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 petit. 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 (1.6) dans  $\mathbf{e}_{\parallel} \times (1.10)$ , on obtient

$$\ddot{\mathbf{x}} \times \mathbf{e}_{\parallel} = \frac{1}{n} \nabla_{\perp} n = \nabla_{\perp} \log n \approx -\frac{2}{c^2} \nabla_{\perp} \Phi, \quad (1.11)$$

où  $\nabla_{\perp}$  est un gradient selon les coordonnées perpendiculaires à  $\mathbf{e}_{\parallel}$ . On note que le facteur 2 qui apparaît dans l'équation (1.11) est un effet qui vient de la relativité générale. Ce facteur corrige la solution que l'on aurait obtenu avec une dérivation classique (newtonienne).

On est maintenant en mesure de calculer l'angle de déviation. J'introduit le paramètre d'impact  $\xi$  qui est la distance perpendiculaire entre la position d'origine du photon sur le plan de la lentille et l'axe de visé (voir Figure 1.3). 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  $\delta$  est la fonction delta de Dirac, alors le potentiel qui satisfait

l'équation de Poisson (1.4) est la fonction de Green  $\Phi = -\frac{GM}{\sqrt{\xi^2 + z^2}}$ , où  $z$  est la coordonné sur l'axe de visée. L'équation (1.2) se réécrit finalement comme

$$\begin{aligned}\alpha(\xi) &= -\frac{2GM}{c^2} \int_{-\infty}^{\infty} \frac{\partial}{\partial \xi} \frac{1}{(\xi^2 + z^2)^{1/2}} dz \\ \implies \alpha(\xi) &= \frac{4GM}{c^2 \xi^2} \xi\end{aligned}\quad (1.12)$$

Cette solution se généralise naturellement à un profil de masse quelconque en assumant qu'il s'exprime comme une somme d'élément de masses  $dm = \Sigma d^2 \xi'$ , où  $\Sigma = \int \rho dz$  est un densité surfacique de masse. L'angle de déviation total mesuré à un point  $\xi$  est alors une convolution sur tout le plan de la lentille (mince) puisque l'équation (1.12) dépend liniairement de la masse  $M$  :

$$\alpha(\xi) = \frac{4G}{c^2} \int_{\mathbb{R}^2} \Sigma(\xi') \frac{\xi - \xi'}{\|\xi - \xi'\|^2} d^2 \xi' \quad (1.13)$$

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 ( $\theta$ ) vers les coordonnées angulaires du plan de la source ( $\beta$ ). On assume que les distances entre l'observateur et la lentille  $D_\ell$ , 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  $\xi$  ou  $\eta$  (voir figure 1.3). Cette approximation est justifiée pour les objets qui nous intéresse, pour lesquels les distances parallèles à l'axe de visée sont généralement de l'ordre du Gpc, alors que les distances perpendiculaire sont généralement de l'ordre du kpc ; soit 6 ordres de grandeurs de différences. Ainsi, on peut faire un argument géométrique (euclidien)

$$\begin{aligned}D_s \theta &= \eta' \\ D_s \beta &= \eta \\ D_{\ell s} \alpha &= \eta' - \eta \\ \implies D_s \beta &= D_s \theta - D_{\ell s} \alpha\end{aligned}\quad (1.14)$$

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 (ç.-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

$$D \equiv \frac{\xi}{\theta}, \quad (1.15)$$

où  $\xi$  est la taille physique d'un objet placé à une certaine distance de l'observateur, et  $\theta$  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 (1.15) est la distance du diamètre angulaire. En pratique, on peut exprimer  $D$  en terme 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 and Lucchin, 2002](#); [Dodelson and Schmidt, 2003](#); [Bartelmann, 2004](#))

$$D_z = ca(z) \underbrace{\int_{a(z)}^{a(0)} \frac{da}{\dot{a}a}}_{\text{distance comobile}}; \quad (1.16)$$

$$\begin{aligned} &= \frac{ca(z)}{H_0} \int_{a(z)}^{a(0)} \frac{da}{\sqrt{\Omega_{r,0} + \Omega_{m,0}a + \Omega_{\Lambda,0}a^4}}; \\ &= \frac{c}{H_0(1+z)} \int_0^z \frac{dz'}{\sqrt{\Omega_{r,0}(1+z')^4 + \Omega_{m,0}(1+z')^3 + \Omega_{\Lambda,0}}}. \end{aligned} \quad (1.17)$$

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 (1.17) 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  $\int_0^z \mapsto \int_{z_\ell}^{z_s}$ . Par soucis de complétude, je rapporte la valeur des paramètres du modèle cosmologique  $\Lambda$ CDM obtenue par l'équipe [Planck Collaboration et al. \(2020\)](#) dans l'annexe A.

Il est généralement pratique de travailler avec la forme adimensionnelle de l'équation (1.14). On introduit la densité critique

$$\Sigma_c = \frac{c^2}{4\pi G} \frac{D_s}{D_{\ell s} D_\ell}, \quad (1.18)$$

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

$$\hat{\alpha}(\boldsymbol{\theta}) = \frac{1}{\pi} \int_{\mathbb{R}^2} \kappa(\boldsymbol{\theta}) \frac{\boldsymbol{\theta} - \boldsymbol{\theta}'}{\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|} d^2\boldsymbol{\theta}', \quad (1.19)$$

qui satisfait l'équation de la lentille adimensionnelle

$$\boldsymbol{\beta} = \boldsymbol{\theta} - \hat{\alpha}(\boldsymbol{\theta}). \quad (1.20)$$

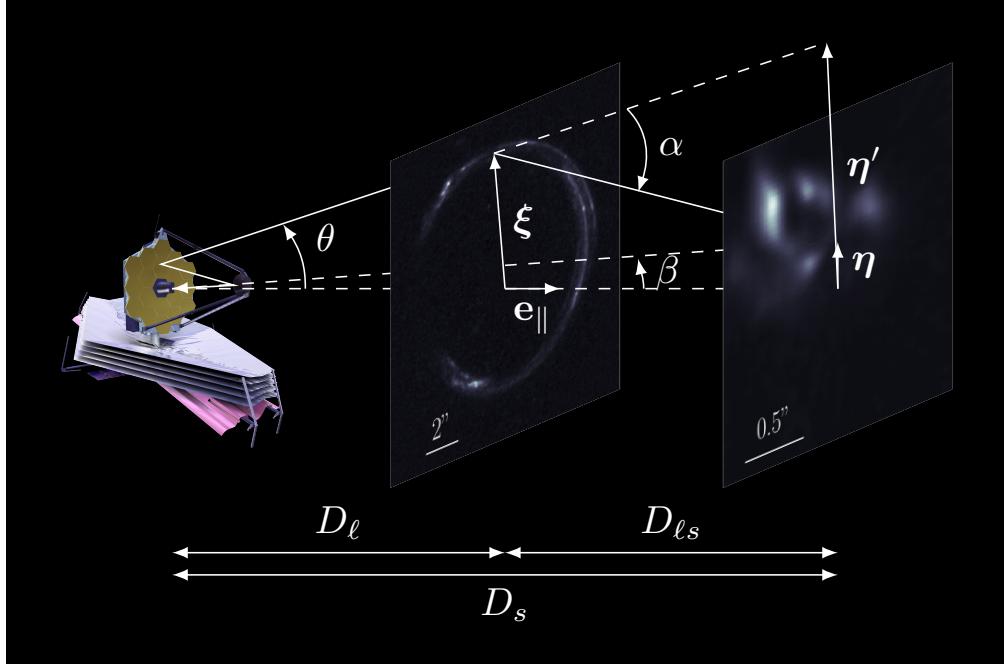


FIGURE 1.3 – Schéma d'une lentille gravitationnelle.

## 1.2 Interférométrie par masque non-régulier

### 1.2.1 Les angles de fermeture

## 1.3 Auto-encodeur variationnel

### 1.3.1 Description du modèle

Les auto-encodeurs variationnels (VAE) ont été introduits par [Kingma and 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 and 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  $\theta$ , et on définit un processus génératif modélisé par la conditionnelle sur la variable cachée  $p_\theta(\mathbf{x} | \mathbf{z})$ . Déterminer  $p_\theta$  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,  $q_\phi(\mathbf{z} \mid \mathbf{x})$ , dont le rôle est d'inférer la variable latente  $\mathbf{z}$  associé à  $\mathbf{x} \sim p_\theta(\mathbf{x} \mid \mathbf{z})$ . En d'autres mots,  $q_\phi(\mathbf{z} \mid \mathbf{x})$  est une approximation variationnelle de la distribution a posteriori  $p_\theta(\mathbf{z} \mid \mathbf{x})$ . La notion de distance entre ces deux distributions est mesurée par la divergence de Kullback-Leibler  $D_{\text{KL}}(\cdot \parallel \cdot) \geq 0$  :

$$\begin{aligned} D_{\text{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) \parallel p_\theta(\mathbf{z} \mid \mathbf{x})) &= \mathbb{E}_{q_\phi(\mathbf{z} \mid \mathbf{x})} \left[ \log q_\phi(\mathbf{z} \mid \mathbf{x}) - \log p_\theta(\mathbf{z} \mid \mathbf{x}) \right] \\ &= \mathbb{E}_{q_\phi(\mathbf{z} \mid \mathbf{x})} \left[ \log q_\phi(\mathbf{z} \mid \mathbf{x}) - \log \frac{p_\theta(\mathbf{z}, \mathbf{x})}{p_\theta(\mathbf{x})} \right] \\ &= \underbrace{\log p_\theta(\mathbf{x}) - \mathbb{E}_{q_\phi(\mathbf{z} \mid \mathbf{x})} \left[ \log p_\theta(\mathbf{z}, \mathbf{x}) - \log q_\phi(\mathbf{z} \mid \mathbf{x}) \right]}_{\equiv \mathcal{L}_{\phi, \theta}(\mathbf{x})}. \end{aligned} \quad (1.21)$$

On remarque par cette manipulation que la distance  $D_{\text{KL}}$ , 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 (1.21), 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  $p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x} \mid \mathbf{z}) p_\theta(\mathbf{z}) d\mathbf{z}$ , ce qui suit de l'inégalité  $\log p_\theta \geq \mathcal{L}_{\phi, \theta}(\mathbf{x})$  (améliore le processus génératif) ;
2. minimiser  $D_{\text{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) \parallel p_\theta(\mathbf{z} \mid \mathbf{x})) = \log p_\theta(\mathbf{x}) - \mathcal{L}_{\phi, \theta}(\mathbf{x})$  (améliore le processus d'inférence de  $\mathbf{z}$ ).

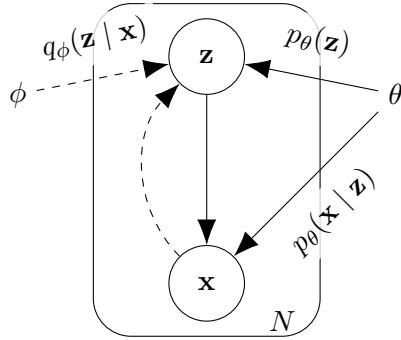


FIGURE 1.4 – Modèle graphique d'un VAE. Les flèches pleines indiquent le processus génératif, alors que les flèches pointillées indiquent le processus d'inférence.

### 1.3.2 Le truc de 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 grimpe de gradient stochastique pour

maximiser la ELBO en terme de  $\phi$  et  $\theta$ . Or, la liste de paramètres  $\phi$  apparait dans la distribution de prélevement pour calculer l'espérance mathématique  $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}$  dans la ELBO (1.21). Cette opération n'a pas de dérivée formelle en terme de  $\phi$ .

Pour résoudre ce problème, on utilise le truc de reparamétrisation (Kingma and 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  $\epsilon$ . On considère le cas où  $q_\phi(\mathbf{z} | \mathbf{x})$  et  $p(\epsilon)$  font partie de la famille gaussienne isotropique

$$\epsilon \sim \mathcal{N}(0, \mathbb{1}); \quad (1.22)$$

$$\mathbf{z} = \boldsymbol{\mu}_\phi + \boldsymbol{\sigma}_\phi \odot \epsilon, \quad (1.23)$$

de sortes que

$$\mathbf{z} \sim q_\phi(\mathbf{z} | \mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_\phi(\mathbf{x}), \mathbb{1}\boldsymbol{\sigma}_\phi^2(\mathbf{x})). \quad (1.24)$$

$\odot$  symbolise le produit d'Hadamard, ou encore le produit élément-par-élément de vecteurs. La reparamétrisation fait en sortes que les paramètres variationnelles ne participent plus au processus de prélevement, maintenant pris en charge par  $\epsilon$ . Cette propriété est cruciale, car elle nous permet de prendre le gradient de la ELBO (1.21). En effet, on peut maintenant échanger les opérateurs  $\nabla_{\phi, \theta}$  et  $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} = \mathbb{E}_{p(\epsilon)}$ , ce qui nous permet d'appliquer le gradient à l'intérieur de l'espérance mathématique. De plus,  $\phi$  décrit maintenant une fonction générique dont le rôle est d'inférer les paramètres de la distribution  $q_\phi(\mathbf{z} | \mathbf{x})$

$$\begin{aligned} f_\phi : \mathbb{R}^m &\rightarrow \mathbb{R}^h \times \mathbb{R}^h \\ \mathbf{x} &\mapsto (\boldsymbol{\mu}, \log \boldsymbol{\sigma}^2) \end{aligned} \quad (1.25)$$

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

L'objectif d'entraînement de la fonction  $f_\phi$  nécessite de manipuler la ELBO pour obtenir une divergence KL

$$\mathcal{L}_{\phi, \theta}(\mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[ \log p_\theta(\mathbf{z}, \mathbf{x}) - \log q_\phi(\mathbf{z} | \mathbf{x}) \right]; \quad (1.26)$$

$$\implies \mathcal{L}_{\phi, \theta}(\mathbf{x}) = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[ \log p_\theta(\mathbf{x} | \mathbf{z}) \right]}_{\text{terme de reconstruction}} + \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[ \log p_\theta(\mathbf{z}) - \log q_\phi(\mathbf{z} | \mathbf{x}) \right]}_{\equiv -D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z}))}. \quad (1.27)$$

Pour déterminer la forme fonctionnelle de la divergence KL obtenue au second terme du membre droit de l'équation (1.27), on stipule a priori que la distribution marginale des variables latentes

devrait correspondre à une distribution normale isotropique

$$p_\theta(\mathbf{z}) = \mathcal{N}(0, \mathbb{1}) \quad (1.28)$$

La KL admet alors une solution fermée étant donné les familles paramétriques stipulées pour  $p_\theta(\mathbf{z})$  (1.28) et  $q_\phi(\mathbf{z} | \mathbf{x})$  (1.24)

$$-D_{\text{KL}}(q_\phi(\mathbf{z} | \mathbf{x}) \| p_\theta(\mathbf{z})) = \frac{1}{2} \sum_{j=1}^h (1 + [\log \boldsymbol{\sigma}_\phi^2]_j - [\boldsymbol{\mu}_\phi]_j - [\boldsymbol{\sigma}_\phi^2]_j) \quad (1.29)$$

Une dérivation de ce terme est donnée dans l'annexe B de [Kingma and Welling \(2013\)](#). Le premier terme du membre droit de l'équation (1.27) est nommé *terme de reconstruction* puisqu'il connecte avec l'objectif des fonctions de type auto-encodeurs d'apprendre une représentation latente d'un échantillon de données. La reconstruction s'accomplit en utilisant d'abord le modèle d'inférence  $\mathbf{z}^{(1:L)} \stackrel{\text{i.i.d.}}{\sim} q_\phi(\mathbf{z} | \mathbf{x})$  pour obtenir un échantillon de représentations latentes à partir des équations (1.22) à (1.23), puis en utilisant le modèle génératif  $\hat{\mathbf{x}}^{(i)} \sim p_\theta(\mathbf{x} | \mathbf{z}^{(i)})$  pour obtenir un échantillon de reconstructions  $\hat{\mathbf{x}}^{(1:L)}$  de l'exemple originel  $\mathbf{x}$ . Comme on a déjà une variable auxiliaire  $\epsilon$  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 convolutionnelles 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

$$\mathbb{E}_{q_\phi(\mathbf{z} | \mathbf{x})} \left[ \log p_\theta(\mathbf{x} | \mathbf{z}) \right] = -\frac{1}{L} \sum_{i=1}^L \|\mathbf{x} - \hat{\mathbf{x}}^{(i)}\|_2^2 \quad (1.30)$$

### 1.3.3 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 motiver l'introduction d'un multiplicateur de Lagrange  $\beta$  au terme de régularisation de la ELBO,  $-D_{\text{KL}}(q_\phi(\mathbf{z} | \mathbf{x}) \| p_\theta(\mathbf{z}))$ . Pour une discussion en profondeur, voir l'excellente revue sur l'utilisation de BIP dans le contexte de l'apprentissage machine par [Goldfeld and Polyanskiy \(2020\)](#) et le manuel de référence sur la théorie de l'information par [Cover and Thomas \(2006\)](#).

L'objectif d'un auto-encoder est de construire un code  $Z$  d'une longueur minimale qui capture un maximum d'information contenu dans un message  $X$ . Formellement, on utilise l'information mutuelle de Shannon pour mesurer l'information capturée par  $Z$  à propos de  $X$

$$I(Z; X) = D_{\text{KL}}(p(\mathbf{x}, \mathbf{z}) \| p(\mathbf{x})p(\mathbf{z})) \quad (1.31)$$

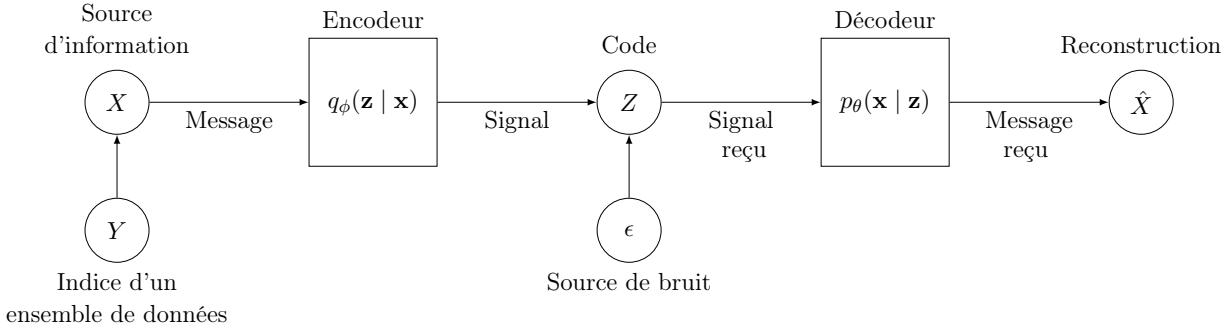


FIGURE 1.5 – VAE comme un système de transmission d’information.

Selon cet objectif, un auto-encodeur idéal atteindra la limite supérieure de l’information mutuelle, soit la capacité du canaux d’information

$$C = \max_{p(\mathbf{x})} I(Z; X) = \max_{p(\mathbf{x})} H(X) - H(Z | X) \leq \max_{p(\mathbf{x})} H(X), \quad (1.32)$$

où  $H(X)$  est l’entropie du message. L’entropie est une mesure de l’incertitude dans une variable aléatoire

$$H(X) = -\mathbb{E}_{p(\mathbf{x})} [\log p(\mathbf{x})] \geq 0, \quad (1.33)$$

qu’on interprète aussi comme une mesure de la longueur minimal 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). Selon ce point de vue,  $H(Z | X) = -\mathbb{E}_{p(\mathbf{x}, \mathbf{z})} [\log p(\mathbf{z} | \mathbf{x})]$  correspond donc à une perte d’information (augmentation de notre incertitude) dû au traitement de l’information par un encodeur et une source de bruit.

Posé de cette façon, déterminer l’auto-encodeur optimal est un problème mal posé. En effet, on pourrait naïvement maximiser l’objectif (1.31) avec la fonction identité comme auto-encodeur :  $f_{\phi, \theta}(X) = \mathbb{1}X$ , de sorte que  $\hat{X} = Z = X$ . 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’est à dire 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

$$\max_{\phi, \theta} I(Z; X) \quad (1.34)$$

sujet à  $I(Z; Y) \leq \alpha$ .

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 exemplaire d’un ensemble de données qu’on veut modéliser  $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N$ . Si  $\alpha$  est trop grand, e.g.  $\alpha > H(Y)$ , alors l’objectif redouble l’objectif mal-posé d’un auto-encodeur classique discuté plus haut : l’auto-encodeur est libre de choisir une relation un-pour-un entre  $Y$  et  $Z$  (et  $X$ ). D’un autre côté,  $\alpha < H(Y)$  force l’auto-encodeur à

comprimer le message. Finalement, de façon intéressante,  $\alpha < H(X) < H(Y)$  force l'auto-encodeur à éliminer l'information redondante dans le message  $X$ . En effet, 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 inférieure variationnelle sur  $I(Z; X)$ . On assume d'abord que la distribution jointe  $p(X, Y, Z)$  se factorise de la façon suivante

$$p(X, Y, Z) = p(X | Z)p(Y | X)p(Z) \quad (1.35)$$

où  $p(X | Z, Y) = p(X | Z)$  suivant le fait qu'on assume une chaîne de Markov  $Y \rightarrow X \rightarrow Z$  pour les variables aléatoires (voir figure 1.5). On introduit un encodeur  $q_\phi(\mathbf{z} | \mathbf{x})$  qui modélise notre système de transmission d'information par compression,  $p(X | Z)$ . Il suit que

$$I(Z; X) = \mathbb{E}_{p(\mathbf{x})} \mathbb{E}_{q_\phi(\mathbf{z} | \mathbf{x})} \left[ \log \frac{q_\phi(\mathbf{z} | \mathbf{x})}{p(\mathbf{z})} \right]; \quad (1.36)$$

$$= -\mathbb{E}_{p(\mathbf{x})} \left[ \log p(\mathbf{z}) \right] + \mathbb{E}_{p(\mathbf{x})} \mathbb{E}_{q_\phi(\mathbf{z} | \mathbf{x})} \left[ \log q_\phi(\mathbf{z} | \mathbf{x}) \right]; \quad (1.37)$$

$$\geq \mathbb{E}_{p(\mathbf{x})} \mathbb{E}_{q_\phi(\mathbf{z} | \mathbf{x})} \left[ \log p_\theta(\mathbf{z} | \mathbf{x}) \right]. \quad (1.38)$$

À la dernière ligne, on a éliminé l'entropie du message  $H(X)$  puisque c'est une constante qui ne dépend pas de  $\phi$ , et on a introduit l'approximation variationnelle  $p_\theta(\mathbf{z} | \mathbf{x}) \approx q_\phi(\mathbf{z} | \mathbf{x})$  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

$$I(Z; Y) = \mathbb{E}_{\mathbf{x}^{(i)} \sim \mathcal{D}} \mathbb{E}_{q_\phi(\mathbf{z} | \mathbf{x}^{(i)})} \left[ \log \frac{q_\phi(\mathbf{z} | \mathbf{x}^{(i)})}{p_\theta(\mathbf{z})} \right]; \quad (1.39)$$

$$I(Z; Y) \leq \mathbb{E}_{\mathbf{x}^{(i)} \sim \mathcal{D}} \mathbb{E}_{q_\phi(\mathbf{z} | \mathbf{x}^{(i)})} \left[ \log \frac{q_\phi(\mathbf{z} | \mathbf{x}^{(i)})}{p_\theta(\mathbf{z})} \right], \quad (1.40)$$

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 (1.34) en introduisant un multiplicateur de Lagrange  $\beta$  pour le terme  $I(Z; Y)$ , et en introduisant les limites variationnelles (1.38) et (1.40). On ignore l'espérance mathématique  $\mathbb{E}_{p(\mathbf{x})} = \mathbb{E}_{\mathbf{x}^{(i)} \sim \mathcal{D}}$  pour mieux illustrer la connection avec la ELBO (1.27)

$$\mathcal{L}_{\phi, \theta, \beta}(\mathbf{x}) = \mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})} \left[ \log p_{\theta}(\mathbf{x} | \mathbf{z}) \right] - \beta \mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})} \left[ \log q_{\phi}(\mathbf{z} | \mathbf{x}) - \log p_{\theta}(\mathbf{z}) \right] \quad (1.41)$$

On note qu'on a délaissé la constante  $\alpha$ , principalement dû au fait que le terme de régularisation est une limite inférieure et la valeur de  $\alpha$  est difficile à déterminer en pratique. Réintégrer la constante  $\alpha$  dans la ELBO reviendrait à choisir la fonction objective qui est le sujet du travail par [Zhao et al. \(2017\)](#).

Il est intéressant de noter que cette nouvelle dérivation commence d'un point de vue complètement différent. On a supposé que le modèle génératif  $p_{\theta}(\mathbf{x} | \mathbf{z})$  est une approximation variationnelle de l'a posteriori du modèle d'inférence  $q_{\phi}(\mathbf{x} | \mathbf{z})$ . L'approche de [Kingma and Welling \(2013\)](#) est parfaitement l'inverse. De plus, le terme de régularisation a maintenant une interprétation beaucoup plus riche, avec un paramètre  $\beta$  qui contrôle le niveau de compression de l'information désiré et qui s'avère à contrôler plus ou moins directement le nombre de partitions possibles dans l'espace latent ()

## 1.4 Machines à inférence récurrentielles

### 1.4.1 Formalisme bayésien des problèmes inverses

Les machines à inférence récurrentielles (RIM) ont été introduites par [Putzky and Welling \(2017\)](#) pour résoudre des problèmes inverses pour lesquels le terme de régularisation est nécessaire mais inconnue 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

$$\mathbf{y} = F(\mathbf{x}) + \boldsymbol{\eta}, \quad (1.42)$$

où  $\mathbf{y} \in \mathcal{Y}$  est un vecteur d'observables (comme l'image capturé par les capteurs photographiques CCD dans un télescope),  $\mathbf{x} \in \mathcal{X}$  est un vecteur de paramètres qui gouvernent le phénomène physique qui nous intéresse, modélisé par le modèle physique  $F : \mathcal{X} \rightarrow \mathcal{Y}$ . Le vecteur  $\boldsymbol{\eta}$  est une réalisation d'un bruit additif. On suppose qu'on connaît la distribution de ce bruit, de sortes qu'on peut modéliser

la fonction de vraisemblance de l'observable

$$\mathbf{y} - F(\mathbf{x}) \sim p(\boldsymbol{\eta}) = p(\mathbf{y} \mid \mathbf{x}). \quad (1.43)$$

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

$$\hat{\mathbf{x}}_{\text{MLE}} = \underset{\mathbf{x} \in \mathcal{X}}{\operatorname{argmax}} \log p(\mathbf{y} \mid \mathbf{x}). \quad (1.44)$$

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 (1.44) 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 connection 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  $L \geq 0$

$$\|G_\varphi(\mathbf{y}_1) - G_\varphi(\mathbf{y}_2)\|_{\mathcal{X}} \leq L\|\mathbf{y}_1 - \mathbf{y}_2\|_{\mathcal{Y}}, \quad (1.45)$$

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

Pour un problème mal-posé, ce qui est le cas pour le problème d'inférence des paramètres d'une lentille gravitationnelles de type galaxie-galaxie ou la reconstruction d'image dans le contexte de l'interférométrie par masque non-réguliers, on assume a priori que la première condition de Hadamard ( $H_1$ ) est respectée. C'est-à-dire qu'on assume 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 (1.44).

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'intérêts  $p_\theta : \mathcal{X} \rightarrow \mathbb{R}$ , 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 (1.44) en introduisant cette distribution a priori comme un terme de régularisation de la vraisemblance

$$\hat{\mathbf{x}}_{\text{MAP}} = \underset{\mathbf{x} \in \mathcal{X}}{\operatorname{argmax}} \log p(\mathbf{y} \mid \mathbf{x}) + \log p_\theta(\mathbf{x}). \quad (1.46)$$

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

$$p_\theta(\mathbf{x} \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \mathbf{x})p_\theta(\mathbf{x})}{\int_{\mathcal{X}} p(\mathbf{y} \mid \mathbf{x})p_\theta(\mathbf{x})d\mathbf{x}}. \quad (1.47)$$

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 (et souvent impossible à calculer) 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 de solution  $\mathcal{X}$  tel que  $\dim_{\mathbb{R}}(\mathcal{X}) \leq \dim_{\mathbb{R}}(\mathcal{Y})$ . Dans ce cas, le problème inverse est balancé 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 sont de haute qualité, ce qui 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 (e.g. une grille de pixels pour modéliser une distribution de masse), de manière à contenir toutes, ou au moins la plupart, des solutions physiques. Ce genre d'approche a le potentiel de produire des résultats surprenant ou intéressant, puisque l'exploration de l'espace des solutions physiques peut être ajustée via la distribution a priori,  $p_\theta(\mathbf{x})$ , selon la complexité de l'observation.

#### 1.4.2 La relation de récurrence

Pour résoudre l'équation différentielle ordinaire sous-entendue par le problème (1.46), on considère la méthode de discréétisation d'Euler

$$\hat{\mathbf{x}}^{(t+1)} = \hat{\mathbf{x}}^{(t)} + \alpha \nabla_{\hat{\mathbf{x}}^{(t)}} p_\theta(\hat{\mathbf{x}}^{(t)} \mid \mathbf{y}), \quad (1.48)$$

où  $\alpha$  est le taux d'apprentissage dans la littérature sur l'apprentissage machine. On est garantie d'obtenir une solution au problème à valeur initiale ( $\hat{\mathbf{x}}^{(0)} = \mathbf{x}_0$ ) si l'algorithme, après  $T$  itérations, satisfait la condition de Lipschitz. Pour la relation de récurrence (1.48), ceci revient à assumer que l'erreur locale de chaque itération est proportionnelle à  $\alpha^2$ , ce qui est satisfait si le gradient  $\nabla_{\mathbf{x}} \log p_\theta(\mathbf{x} \mid \mathbf{y})$  satisfait la condition de Lipschitz dans la région de  $\mathcal{X}$  explorée par l'algorithme (Atkinson, 1989; Butcher, 2016), en encore si la norme de la dérivée seconde de  $\log p_\theta(\mathbf{x} \mid \mathbf{y})$  est

bornée dans cette région.

[Putzky and Welling \(2017\)](#) observent qu'on peut réécrire (1.48) de la façon suivante

$$\hat{\mathbf{x}}^{(t+1)} = \hat{\mathbf{x}}^{(t)} + \alpha (\nabla_{\hat{\mathbf{x}}^{(t)}} \log p(\mathbf{y} | \hat{\mathbf{x}}^{(t)}) + \nabla_{\hat{\mathbf{x}}^{(t)}} \log p_\theta(\hat{\mathbf{x}}^{(t)})); \quad (1.49)$$

$$\implies \hat{\mathbf{x}}^{(t+1)} = \hat{\mathbf{x}} + g_{\varphi^{(t)}}(\hat{\mathbf{x}}^{(t)}, \nabla_{\hat{\mathbf{x}}^{(t)}} \log p(\mathbf{y} | \hat{\mathbf{x}}^{(t)})) \quad (1.50)$$

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 (1.48) est un cas spécial de la relation (1.50), 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  $\alpha$ . Dans la relation (1.50), les paramètres  $\alpha$  et  $\theta$  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 convolutif 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 ( $H_3$ ) est respectée seulement si  $g_{\varphi^{(t)}}$  possède une constante de Lipschitz  $L \leq 1$ , ce qui n'est pas trivialement respecté pour 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 actif (e.g. ), particulièrement dans l'étude des attaques antagonistes de réseaux de neurones (e.g. ).

Finalement, on note un aspect important du modèle  $g_{\varphi^{(t)}}$ , soit la possible dépendance des paramètres  $\varphi^{(t)}$  envers  $t$ . Cet aspect est directement inspiré des succès récents d'algorithmes d'optimisations comme la méthode d'accélération de [Nesterov \(1983\)](#), AdaGrad ([Duchi et al., 2011](#)), RMSProp<sup>5</sup> ([Hinton, 2012](#)) et ADAM ([Kingma and Ba, 2014](#)), qui utilisent explicitement l'information des gradients d'itérations antérieures à  $t$  pour calculer la mise à jour dans la relation de récurrence (1.50). 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.

### 1.4.3 Méta-apprentissage

Le méta-apprentissage est un sujet de recherche qui a une longue histoire dans le champ de recherche sur l'apprentissage machine, qu'on peut tracer jusqu'aux travaux de Marvin Minsky, puis Schmidhuber 1991 (LSTM and thesis and meta algorithm) et Bengio 1990 (). Le lecteur peut se

---

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

référer à la revue de Hospedales pour une vue moderne sur le sujet (). L'approche qui nous intéresse est classée dans la catégorie de méta-apprentissage par optimisation.

La première apparition concrète de cette méthode est Younger 2001 et Hochreiter 2001, où le théorème de l'approximation universelle est utilisée pour justifier l'utilisation de cellules à mémoire longues et courtes (LSTM, Schidhuber) pour découvrir un algorithme d'optimisation pour un classe de fonctions (e.g. un modèle neuronal). L'observation qui est faite est précisément que l'algorithme d'Euler est un cas particulier d'une classe plus générale de relations de récurrences qui permettent de résoudre des problèmes de type (??). Ainsi, un réseau de neurones récurrent est une classe de fonctions qui peuvent représenter, en principe, une large portion de cette classe de fonctions. Ce genre d'approche est motivé par le *no free lunch theorem* pour l'optimisation, qui stipule qu'il n'existe aucun algorithme général d'optimisation en mesure de résoudre toutes les classes de problèmes. Dans ce cas, la solution à ce problème est d'introduire des biais inductifs ou des connaissances a priori pour contraindre l'espace des solutions recherchées à un espace où au moins une solution existe. Le problème de méta-apprentissage est donc précisément d'apprendre ou encoder ces biais inductifs dans un modèle d'apprentissage, de sorte que les problèmes d'optimisations subséquents, sur des tâches d'essai, sont garanties d'avoir une solution.

Le travail de [Andrychowicz et al. \(2016\)](#) utilise ces idées pour construire un algorithme d'optimisation, aussi basé sur les cellules LSTM, qui performe beaucoup mieux que les algorithmes d'optimisations traditionnelles (e.g. ADAM) pour entraîner un second réseau de neurones pour les tâches spécifiques sur lesquelles l'algorithme de méta apprentissage est entraîné (style transfer etc.). Le travail de Putzky et Welling est une généralisation de cette approche aux problèmes inverses en général.

Pour un problème de méta-apprentissage, l'ensemble de donné d'entraînement est légèrement différent d'une tâche d'interpolation ou de classification, où  $\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$  est construit à partir d'exemples dans le domaine  $\mathcal{X}$  et l'image  $\mathcal{Y}$  connecté par la fonction qu'on essaie d'approximer. Pour le méta-apprentissage, l'ensemble d'entraînement est constitué de tâche à performer. Dans notre cas, la tâche à performer est l'optimisation d'une fonction de vraisemblance. On a donc

$$\mathcal{D} = \{\mathbf{x}_i, \log p_i(\mathbf{y} \mid \mathbf{x})\}_{i=1}^N \quad (1.51)$$

où  $\mathbf{x}_i$  est la solution qu'on cherche et  $\log p_i(\mathbf{y} \mid \mathbf{x})$  est la fonction de vraisemblance que l'algorithme doit optimiser pour obtenir la solution. Les paramètres d'inférence  $\varphi$  sont optimisés sur toute l'ensemble de la trajectoire construite par la relation de récurrence par une erreur quadratique moyenne

$$\mathcal{L}_\varphi(\mathbf{x}, \log p(\mathbf{y} \mid \mathbf{x})) = \sum_{t=1}^T w^{(t)} \|\mathbf{x} - \hat{\mathbf{x}}^{(t)}\|_{\mathcal{X}}^2 \quad (1.52)$$

où  $w^{(t)}$  est un poids qu'on associe à l'itération  $t$  de la relation de récurrence. Dans la plupart des travaux,  $w^{(t)} = \frac{1}{T}$ . Cet objectif est optimisé par la rétropropagation temporelle des gradients

(BPTT, de l'anglais *backpropagation through time*). Le problème de méta-apprentissage est donc un problème de minimisation du risque empirique observé de la fonction objective

$$\varphi_{\mathcal{D}}^* = \operatorname{argmin}_{\varphi} \mathbb{E}_{\mathcal{D}} [\mathcal{L}_{\varphi}] \quad (1.53)$$

Il est important de discuter de la notion de généralisation dans le contexte de méta-apprentissage. Dans le contexte où on cherche à construire une interpolation, la notion de généralisation réfère généralement au test où un point  $\mathbf{x}$  en dehors du support implicite défini par l'ensemble d'entraînement  $\mathcal{D}$  est donné en entrée à la fonction  $f_{\varphi}$  qu'on a construit. Un modèle est en mesure de généraliser si l'erreur quadratique moyenne sur la prédiction est similaire au risque empirique observé sur l'ensemble d'entraînement.

Dans notre contexte, la généralisation réfère plutôt au concept de transfert d'apprentissage, c'est à dire transférer les connaissances apprises dans un certain contexte en transférant la structure du problème pour vers des tâches d'essais. Ainsi, on comprend que la généralisation, dans notre contexte, est équivalente à la notion de transfert de connaissance. Les paramètres d'inférences  $\varphi$ , plutôt que d'encoder les détails d'une fonction, encode des biais inductifs, ou autrement des connaissances a priori sur la structure du problème qui sont transférable d'un problème à un autre. Cette réalisation est particuliè

Finalement, on note que la notion d'initialisation dans la relation de récurrence  $\mathbf{x}^{(0)} = \mathbf{x}_0$  est particulièrement importante dans notre traitement. En effet, on assume que la fonction  $g_{\varphi}$  se comporte bien dans une région de  $\mathcal{X}$  qui connecte  $\mathbf{x}^{(0)}$  à  $\mathbf{x}^{(T)}$ . Or, un mauvais choix d'initialisation fait en sorte que la troisième condition d'Hadamard est difficilement respectée. La notion d'apprendre une initialisation qui accélère l'apprentissage de la descente de gradient est un sujet actif du champ de recherche de méta-apprentissage. MAML et Reptile approchent ce problème via une boucle double d'optimisation. Or, une approche beaucoup plus simple peut être mise en place si on fait utilisation de l'observation dans notre problème d'inférence. Ici, on peut prendre le point de vue qu'une fonction approximative inverse du modèle physique  $\hat{F}_{\varphi'}^{-1}$  est un bon point de départ pour  $\mathbf{x}_0$ , en particulier si l'image de cette fonction se situe dans la région de haute densité de distribution a priori empirique déterminée par  $\mathcal{D}$ .

# Bibliographie

- A. Alemi, I. Fischer, J. Dillon, and K. Murphy. Deep variational information bottleneck. In *ICLR*, 2017. URL <https://arxiv.org/abs/1612.00410>.
- M. Andrychowicz, M. Denil, S. Gomez, M. W. Hoffman, D. Pfau, T. Schaul, B. Shillingford, and N. de Freitas. Learning to learn by gradient descent by gradient descent. *arXiv e-prints*, art. arXiv :1606.04474, June 2016.
- 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>.
- M. W. Auger, T. Treu, A. S. Bolton, R. Gavazzi, L. V. E. Koopmans, P. J. Marshall, L. A. Moustakas, and 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, and 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.
- R. D. Blandford and 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/annrev.aa.30.090192.001523>.
- A. S. Bolton, S. Burles, L. V. E. Koopmans, T. Treu, and L. A. Moustakas. SDSS j140228.22632133.3 : 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, and 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, and 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.
- 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/10.1002/9781119121534>.
- 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>.
- P. Coles and F. Lucchin. *Cosmology : The Origin and Evolution of Cosmic Structure*. Wiley, 2 edition, July 2002.
- A. Congdon and 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>.

- T. M. Cover and J. A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006. ISBN 0471241954.
- G. V. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2 : 303–314, 1989.
- S. Dodelson and F. Schmidt. *Cosmology : The Origin and Evolution of Cosmic Structure*. Academic Press, 2 edition, March 2003.
- J. Duchi, E. Hazan, and 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>.
- E. E. Falco, M. V. Gorenstein, and 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, and J. Walcher. First Catalog of Strong Lens Candidates in the COSMOS Field. *ApJS*, 176(1) :19–38, May 2008. doi : 10.1086/526426.
- R. Gavazzi, C. Adami, F. Durret, J. C. Cuillandre, O. Ilbert, A. Mazure, R. Pelló, and 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.
- Z. Goldfeld and Y. Polyanskiy. The Information Bottleneck Problem and Its Applications in Machine Learning. *arXiv e-prints*, art. arXiv :2004.14941, Apr. 2020.
- J. Hadamard. Sur les problèmes aux dérivés partielles et leur signification physique. *Princeton University Bulletin*, 13 :49–52, 1902.
- G. Hinton. Neural networks for machine learning. [http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture\\_slides\\_lec6.pdf](http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf), 2012. Accède le 2022-07-10.
- 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>.
- 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, and 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.
- D. P. Kingma and J. Ba. Adam : A Method for Stochastic Optimization. *arXiv e-prints*, art. arXiv :1412.6980, Dec. 2014.
- D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. *arXiv e-prints*, art. arXiv :1312.6114, Dec. 2013.
- D. P. Kingma and 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, and 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, and 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, and 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, and 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.
- P. Lemos, N. Jeffrey, M. Cranmer, S. Ho, and P. Battaglia. Rediscovering orbital mechanics with machine learning. *arXiv e-prints*, art. arXiv :2202.02306, Feb. 2022.
- 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>.
- 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. Ivezić, 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, and H. Zhan. LSST Science Book, Version 2.0. *arXiv e-prints*, art. arXiv :0912.0201, Dec. 2009.
- M. Meneghetti. *Introduction to Gravitational Lensing*. Springer Cham, 2013. doi : 10.1007/978-3-030-73582-1.
- 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.
- 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, and 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, N. Aghanim, Y. Akrami, F. Arroja, M. Ashdown, J. Aumont, C. Baccigalupi, M. Ballardini, A. J. Banday, R. B. Barreiro, N. Bartolo, S. Basak, R. Battye, K. Benabed, J. P. Bernard, M. Bersanelli, P. Bielewicz, J. J. Bock, J. R. Bond, J. Borrill, F. R. Bouchet, F. Boulanger, M. Bucher, C. Burigana, R. C. Butler, E. Calabrese, J. F. Cardoso, J. Carron, B. Casaponsa, A. Challinor, H. C. Chiang, L. P. L. Colombo, C. Combet, D. Contreras, B. P. Crill, F. Cuttaia, P. de Bernardis, G. de Zotti, J. Delabrouille, J. M. Delouis, F. X. Désert, E. Di Valentino, C. Dickinson, J. M. Diego, S. Donzelli, O. Doré, M. Douspis, A. Ducout, X. Dupac, G. Efstathiou, F. Elsner, T. A. Enßlin, H. K. Eriksen, E. Falgarone, Y. Fantaye, J. Fergusson, R. Fernandez-Cobos, F. Finelli, F. Forastieri, M. Frailis, E. Franceschi, A. Frolov, S. Galeotta, S. Galli, K. Ganga, R. T. Génova-Santos, M. Gerbino, T. Ghosh, J. González-Nuevo, K. M. Górski, S. Gratton, A. Gruppuso, J. E. Gudmundsson, J. Hamann, W. Handley, F. K. Hansen, G. Helou, D. Herranz, S. R. Hildebrandt, E. Hivon, Z. Huang, A. H. Jaffe, W. C. Jones, A. Karakci, E. Keihänen, R. Keskitalo, K. Kiiveri, J. Kim, T. S. Kisner, L. Knox, N. Krachmalnicoff, M. Kunz, H. Kurki-Suonio, G. Lagache, J. M. Lamarre, M. Langer, A. Lasenby, M. Lattanzi, C. R. Lawrence, M. Le Jeune, J. P. Leahy, J. Lesgourgues, F. Levrier, A. Lewis, M. Liguori, P. B. Lilje, M. Lilley, V. Lindholm, M. López-Caniego, P. M. Lubin, Y. Z. Ma, J. F. Macías-Pérez, G. Maggio, D. Maino, N. Mandolesi, A. Mangilli, A. Marcos-Caballero, M. Maris, P. G. Martin, M. Martinelli, E. Martínez-González, S. Matarrese, N. Mauri, J. D. McEwen, P. D. Meerburg, P. R. Meinhold, A. Melchiorri, A. Mennella, M. Migliaccio, M. Millea, S. Mitra, M. A. Miville-Deschénes, D. Molinari, A. Moneti, L. Montier, G. Morgante, A. Moss, S. Mottet, M. Münchmeyer, P. Natoli, H. U. Nørgaard-Nielsen, C. A. Oxborrow, L. Pagano, D. Paoletti, B. Partridge, G. Patanchon, T. J. Pearson, M. Peel, H. V. Peiris, F. Perrotta, V. Pettorino, F. Piacentini, L. Polastri, G. Polenta, J. L. Puget, J. P. Rachen, M. Reinecke, M. Remazeilles, C. Renault, A. Renzi, G. Rocha, C. Rosset, G. Roudier, J. A. Rubiño-Martín, B. Ruiz-Granados, L. Salvati, M. Sandri, M. Savelainen, D. Scott, E. P. S. Shellard, M. Shiraishi, C. Sirignano, G. Sirri, L. D. Spencer, R. Sunyaev, A. S. Suur-Uski, J. A. Tauber, D. Tavagnacco, M. Tenti, L. Terenzi, L. Toffolatti, M. Tomasi, T. Trombetti, J. Valiviita, B. Van Tent, L. Vibert, P. Vielva, F. Villa, N. Vittorio, B. D. Wandelt, I. K. Wehus, M. White, S. D. M. White, A. Zacc hei, and A. Zonca. Planck 2018 results. I. Overview and the cosmological legacy of Planck. *A&A*, 641 :A1, Sept. 2020. doi : 10.1051/0004-6361/201833880.

P. Putzky and M. Welling. Recurrent Inference Machines for Solving Inverse Problems. *arXiv e-prints*, 2017.

A. Refregier, A. Amara, T. D. Kitching, A. Rassat, R. Scaramella, and J. Weller. Euclid Imaging Consortium Science Book. *arXiv e-prints*, art. arXiv :1001.0061, Jan. 2010.

O. Ronneberger, P. Fischer, and T. Brox. U-Net : Convolutional Networks for Biomedical Image Segmentation. *arXiv e-prints*, art. arXiv :1505.04597, May 2015.

S. Schuldt, G. Chirivi, S. H. Suyu, A. Yıldırım, A. Sonnenfeld, A. Halkola, and 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.

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, and 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.

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, and 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.

A. Stockton. The lens galaxy of the twin QSO 0957+561. *ApJ*, 242 :L141–L144, Dec. 1980. doi : 10.1086/183419.

N. Tishby, F. C. Pereira, and 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>.

T. Treu. Strong Lensing by Galaxies. *ARA&A*, 48 :87–125, Sept. 2010. doi : 10.1146/annurev-astro-081309-130924.

D. Walsh, R. F. Carswell, and R. J. Weymann. 0957+561 A, B : twin quasistellar objects or gravitational lens? *Nature*, 279 :381–384, May 1979. doi : 10.1038/279381a0.

P. Young, J. E. Gunn, J. Kristian, J. B. Oke, and 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, and J. A. Westphall. Q0957+561 : detailed models of the gravitational lens effect. *ApJ*, 244 :736–755, Mar. 1981. doi : 10.1086/158751.
- S. Zhao, J. Song, and S. Ermon. InfoVAE : Information Maximizing Variational Autoencoders. *arXiv e-prints*, art. arXiv:1706.02262, June 2017.
- 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

### $\Lambda$ CDM

TABLE A.1 – Paramètres de  $\Lambda$ CDM ajusté avec les observations du fond diffus cosmologique par le télescope Planck ([Planck Collaboration et al., 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
$\Omega_0$	Densité totale	$\equiv 1$
$h$	Constante de Hubble $h \equiv \frac{H_0}{100 \text{ km s}^{-1} \text{ Mpc}^{-1}}$	0.6732

## Annexe 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  $\varphi$  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  $\varphi$  encode information about  $\mathcal{D}$ , while  $\mathcal{T}$  was unseen by  $\varphi$ . It follows that  $\mathcal{T}$  and  $\mathcal{D}$  are conditionally independent when given  $\varphi$ . 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 (??)). 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 (??)) and the  $\chi^2$  :

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

We find in practice the the  $\ell_2$  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  $\lambda$  is introduced to tune our uncertainty about the network parameters during fine-tuning.

## Annexe 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
$\beta_{\max}$	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	$\mathbb{1}$
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
$\beta_{\max}$	0.2
Batch size	32

## Annexe 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  $\boldsymbol{\xi} \in \Xi$  and the forward modelling space  $\mathbf{x} \in \mathcal{X}$ . This is a different notion from preprocessing, discussed in section ??, because this transformation is applied inside the recurrent relation ?? 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  $\mathcal{L}_\varphi$  must be computed in the  $\Xi$  space in order to avoid gradient vanishing problems when  $\Psi$  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 :  $\hat{\kappa} = \Psi(\boldsymbol{\xi}) = 10^{\boldsymbol{\xi}}$ . This  $\Psi$  encodes the non-negativity of the convergence. Furthermore, it is a power transformation that leaves the linked pixel values  $\xi_i$  normally distributed, thus improving the learning through the nonlinearities in the neural network. The pixel weights  $\mathbf{w}_i$  in the loss function (??) 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 — its performance 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.  $\mathbf{w}_i = \frac{1}{M}$ .

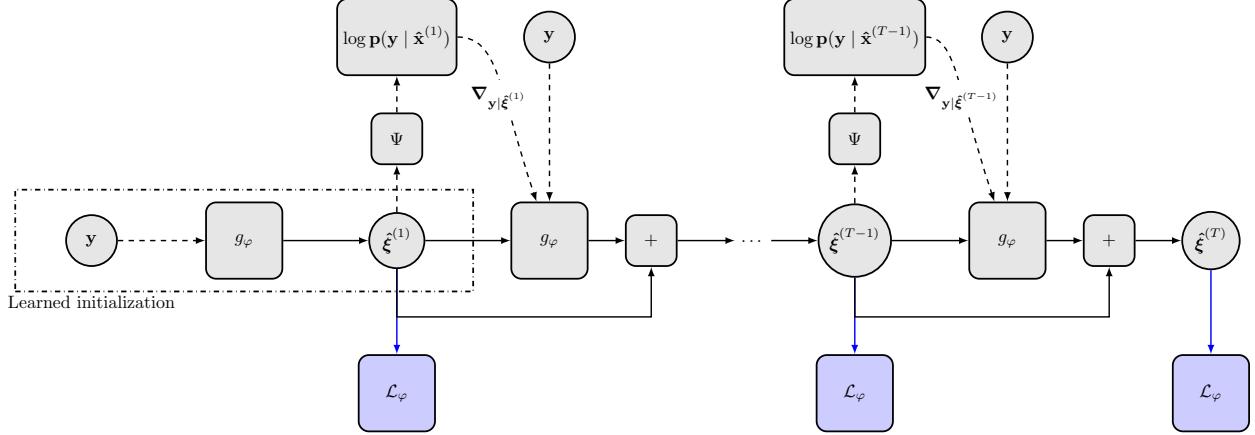


FIGURE D.1 – Unrolled computational graph of the RIM. Operations along solid arrows are being recorded for BPTT, while operations along dashed arrows are not. The blue arrows are only used for optimisation during training. During fine-tuning or testing, the loss is computed only as an oracle metric to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

TABLE D.1 – Hyperparameters for the RIM.

Parameter	Value
Source link function	1
$\kappa$ link function	$10^{\frac{1}{6}}$
<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	$6 \times 10^{-5}$
Learning rate schedule	Exponential Decay
Decay rate	0.9
Decay steps	100 000
Number of steps	870 000
Batch size	1

COSMOS RIM+FT IllustrisTNG RIM+FT Observation RIM+FT COSMOS RIM+FT IllustrisTNG RIM+FT Observation RIM+FT

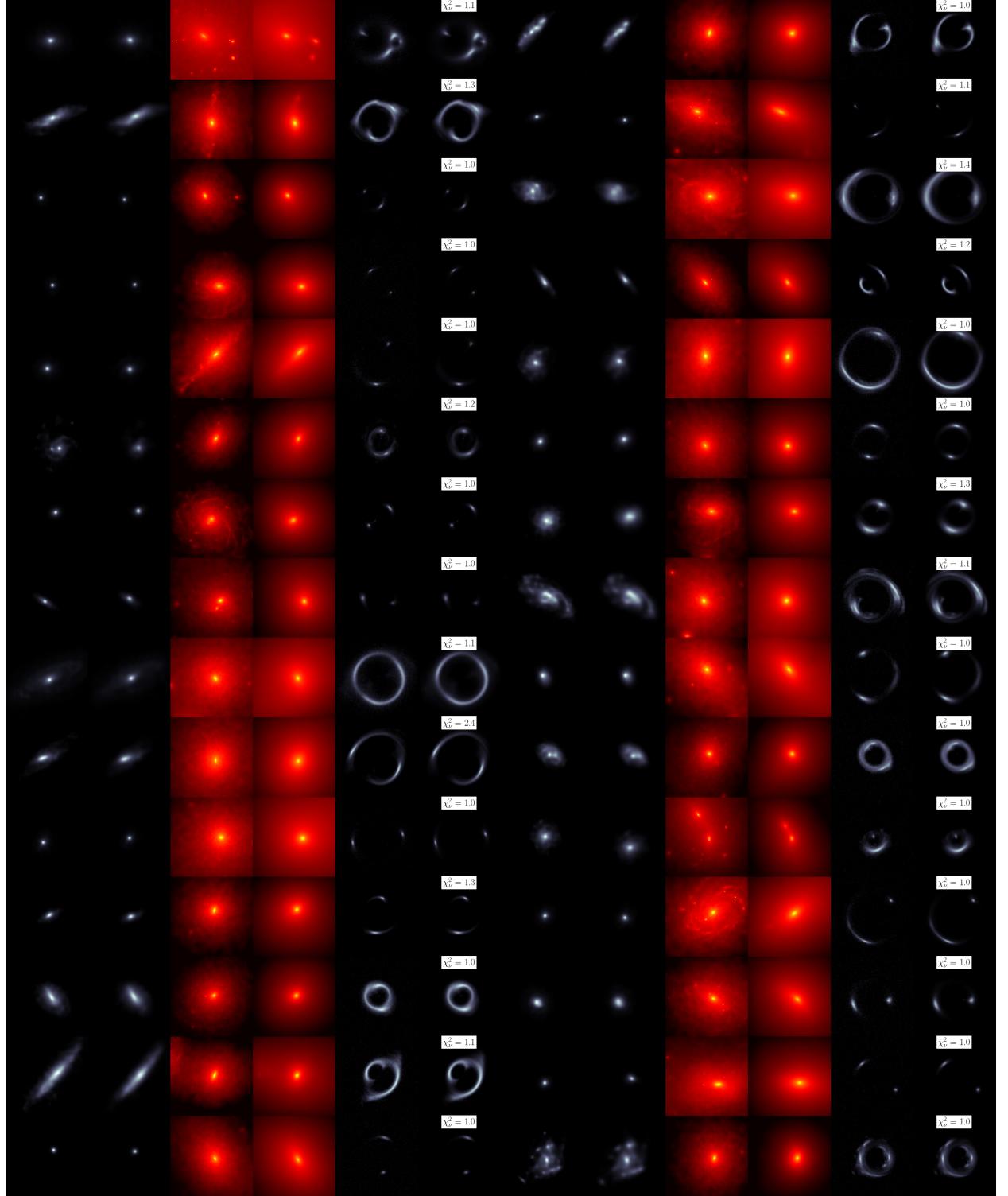


FIGURE D.2 – 30 reconstructions taken at random from the test set of 3000 examples simulated from COSMOS and IllustrisTNG data at high SNR. The colorscale are the same as in Figure ??.

## 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  $\mathbf{z}$ .