

# **RECONSTRUCTION D'IMAGES AVEC LES MACHINES À INFÉRENCES 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 de type galaxie-galaxie . . . . .	2
1.1.1 Les angles de déflections . . . . .	4
1.2 Interférométrie par masque non-régulier . . . . .	8
1.2.1 Les angles de fermeture . . . . .	8
1.2.2 Applications . . . . .	8
1.3 Auto-encodeur variationnel . . . . .	8
1.3.1 Description du modèle . . . . .	8
1.3.2 Le truc de reparamétrisation . . . . .	9
1.4 Machines à inférence récurrentielles . . . . .	11
1.4.1 Formalisme bayésien des problèmes inverses . . . . .	11
1.4.2 La relation de récurrence . . . . .	13

1.4.3 Méta-apprentissage . . . . .	15
<b>Bibliographie</b>	<b>18</b>
<b>A Elastic Weight Consolidation</b>	<b>22</b>
<b>B VAE Architecture and optimisation</b>	<b>24</b>
<b>C RIM architecture and optimisation</b>	<b>27</b>
<b>D GRU</b>	<b>31</b>

# Liste des tableaux

B.1	Hyperparameters for the background source VAE.	25
B.2	Hyperparameters for the convergence VAE.	26
C.1	Hyperparameters for the RIM.	29

# Liste des figures

1.1	Le quasar double (QSO 0957+561 A et B) et la galaxie-lentille (G). Crédit : ESA/Hubble et NASA, enlagement par AA. . . . .	3
1.2	Schéma d'une lentille gravitationnelle. . . . .	7
1.3	Modèle graphique d'un VAE. Les flèches pleines indiquent le processus génératif, alors que les flèches pointillées indiquent le processus d'inférence. . . . .	9
C.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. . . . .	28
C.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 ??.. . . . .	30

# 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  $P(X)$ .
- $\|\cdot\|_2$  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 de type galaxie-galaxie

Fritz Zwicky (1937), suivant les calculs publiés par Einstein (1936)<sup>1</sup> et la première observation de l'effet de déviation gravitationnelle de la lumière par Eddington (1919), est le premier à observer correctement qu'une lentille gravitationnelle est un objet particulièrement riche en information<sup>2</sup>. Ce phénomène se produit lorsque la lumière d'une source lointaine est fortement déviée par le champ gravitationnel d'un objet massif (e.g. une galaxie, un trou noir, une étoile, etc.) exceptionnellement bien aligné avec cette source selon le point de vue d'un observateur sur Terre. L'article de Zwicky (1937) articule précisément deux idées qui nous motivent encore aujourd'hui (plus de 85 ans plus tard) à étudier ces objets, soit

1. d'imager des galaxies trop lointaine pour que l'on puisse les résoudre avec nos télescopes ;
2. de mesurer directement la masse gravitationnelle de galaxies (appelées nébuleuse extra-galactique à l'époque) agissant comme lentilles.

Einstein (1936) considérait l'éventualité d'observer ces systèmes extrêmement improbable en raison de la minuscule taille angulaire caractéristique d'un anneau d'Einstein-Chwolson (Chwolson, 1924; Einstein, 1936)

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

Dû à cette difficulté pratique, la première lentille gravitationnelle est découverte seulement 43 ans 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 seulement 5.7 secondes d'arcs et capturés avec le télescope 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

---

1. Cacluls de Link

2. On note l' existence de plusieurs travaux spéculatifs (et possiblement Link

suspicieusement semblables. Ces coïncidences suggèrent fortement que ces deux spectres sont deux copies provenant d'un même 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 ( $z = 0.39$ ) avec l'observatoire Palomar (Young et al., 1980), simultanément observé et confirmé par le télescope de 2.2 m de l'Université d'Hawaii au mont Mauna Kea (Stockton, 1980), ainsi que la modélisation de sa distribution de masse, de son environnement et des angles de déflexion qui causerait l'apparition d'une image double du quasar (Young et al., 1981; ?)

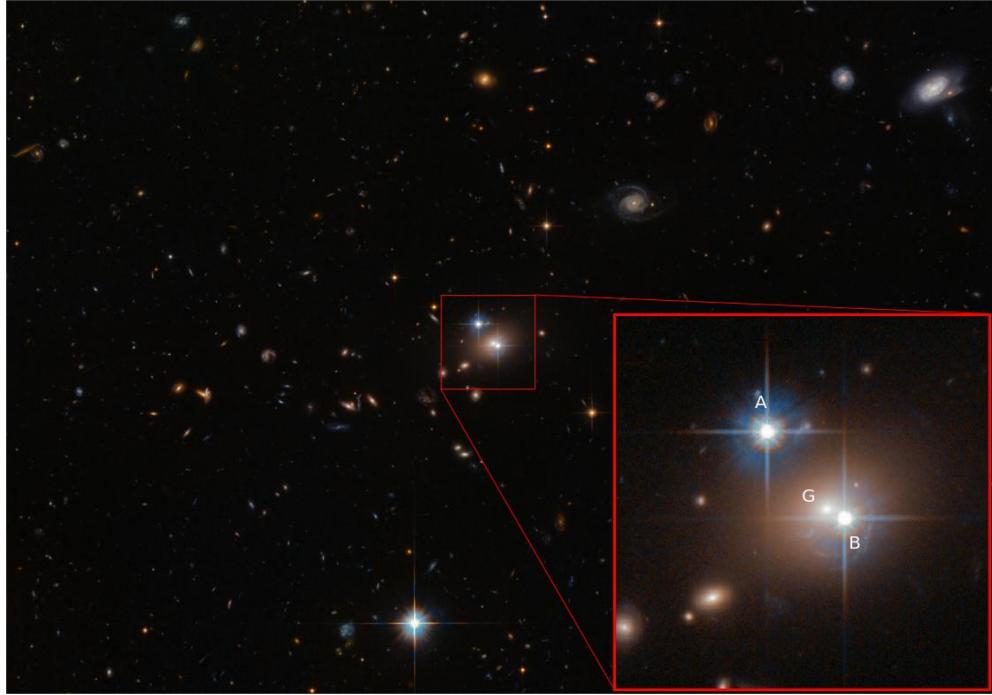


FIGURE 1.1: Le quasar double (QSO 0957+561 A et B) et la galaxie-lentille (G). Crédit : ESA/Hubble et NASA, enlagement par AA.

L'importance des lentilles gravitationnelles pour la cosmologie est subséquemment reconnue très tôt (voir la revue de Blandford and Narayan, 1992). Leur application dans ce domaine se divise principalement en 4 catégories :

1. La cosmographie par délai temporelles (Treu and Marshall, 2016; Suyu et al., 2017) En particulier, le paramètre le mieux contraint est la constante de Hubble (1929)  $H_0$ , mesurant le taux de l'expansion de l'Univers au temps présent. Les deux principales méthodes pour faire cette mesure sont la caractérisation de la courbe de lumière des supernovae Refsdal (1964), ç.-à-d. lentillé par une ou plusieurs galaxies (e.g. Kelly et al., 2015; Goobar et al., 2017), et la surveillance décennale de quasars lentillés (e.g. Vanderriest et al., 1989; ?);

Le sujet de ce travail se concentre autour des lentilles gravitationnelles fortes de type galaxie-galaxie, e.g. QSO 0957+561. De tels systèmes sont caractérisé minimalement par une distortion de

l'image source suffisamment forte pour causer l'apparition d'au moins deux images résolues de la source. Pour une discussion plus général du phénomène, le lecteur peut se référer aux manuels de références par [Meneghetti \(2013\)](#); [Congdon and Keeton \(2018\)](#) ou aux excellentes revues du sujet par [Bartelmann \(2010\)](#); [Treu \(2010\)](#).

### 1.1.1 Les angles de déflections

Dans les paragraphes qui suivent, je dérive les équations centrales qui nous permettent d'étudier les lentilles gravitationnelles de type galaxie-galaxie. Mon traitement est largement inspiré des manuels de références de [Meneghetti \(2013\)](#) et [Carroll \(2003\)](#).

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

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

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 language 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 indistinguorable 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 gas 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)

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

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, le terme  $\propto \dot{\mathbf{x}} \times \mathbf{e}_\parallel$  s'annule. 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 1.2). Dans le cas où le potentiel est généré par une masse  $M$  ponctuelle, q.-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 ( $\boldsymbol{\theta}$ ) vers les coordonnées angulaires du plan de la source ( $\boldsymbol{\beta}$ ). On assume que les distances entre l'observateur et la lentille  $D_\ell$ , 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 1.2). 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)

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

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 (q.-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

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

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

$$\Sigma_c = \frac{c^2}{4\pi G} \frac{D_s}{D_{\ell s} D_\ell}, \quad (1.16)$$

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

$$\hat{\alpha}(\boldsymbol{\theta}) = \frac{1}{\pi} \int_{\mathbb{R}^2} \kappa(\boldsymbol{\theta}) \frac{\boldsymbol{\theta} - \boldsymbol{\theta}'}{\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|} d^2 \boldsymbol{\theta}', \quad (1.17)$$

qui satisfait l'équation de la lentille adimensionnelle

$$\boldsymbol{\beta} = \boldsymbol{\theta} - \hat{\alpha}(\boldsymbol{\theta}). \quad (1.18)$$

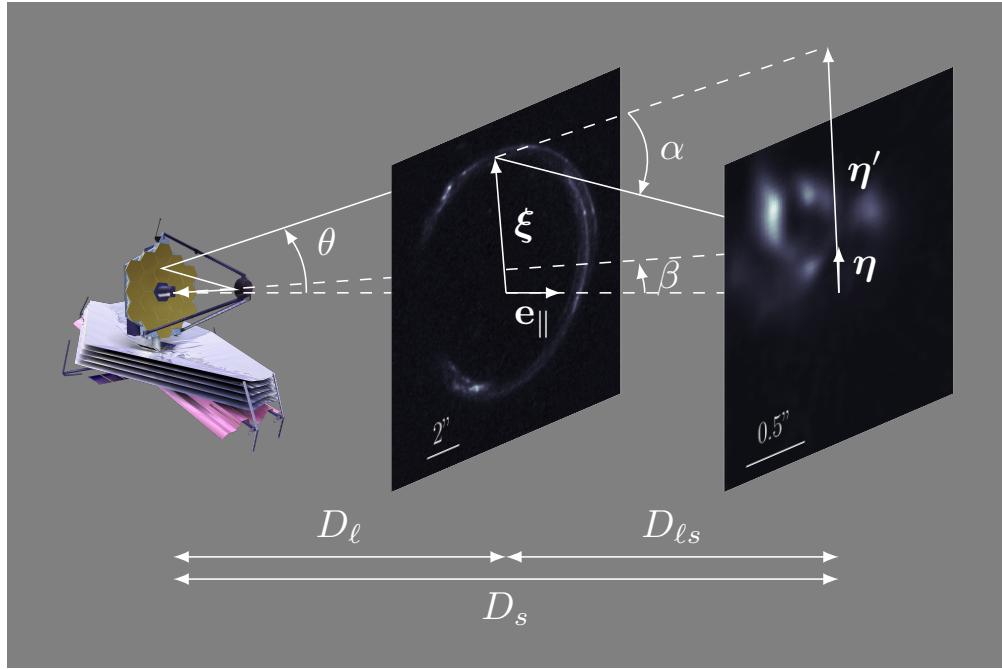


FIGURE 1.2: Schéma d'une lentille gravitationnelle.

## 1.2 Interférométrie par masque non-régulier

### 1.2.1 Les angles de fermeture

### 1.2.2 Applications

## 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 modélisent une distribution a posteriori définie implicitement via un échantillon de données. 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  $\mathbf{z} \sim q(\mathbf{z})$  comme une variable latente et  $\mathbf{x}$  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 suppose, sans perte de généralité, que la distribution de  $\mathbf{x}$  fait partie d'une famille de distribution, caractérisée par  $\theta$ , conditionnelle à la variable cachée :  $p_\theta(\mathbf{x} | \mathbf{z})$ . Déterminer  $p_\theta$  est généralement difficile, voir intractable, si la dimensionnalité de  $\mathbf{x}$  est grande, ce qui est le cas pour des images pour lesquelles on trouve facilement  $\text{dim}(\mathbf{x}) > 10^4$ . Pour résoudre cette difficulté, on introduit un modèle paramétrique d'inférence  $q_\phi(\mathbf{z} | \mathbf{x})$  dont le rôle est de modéliser la distribution a posteriori de la variable latente pour la distribution qui nous intéresse

$$q_\phi(\mathbf{z} | \mathbf{x}) \approx p_\theta(\mathbf{z} | \mathbf{x}). \quad (1.19)$$

La notion de distance entre ces deux distributions est mesurée par la divergence de Kullback-Leibler  $D_{\text{KL}}(\cdot \| \cdot) \geq 0$  :

$$\begin{aligned} D_{\text{KL}}(q_\phi(\mathbf{z} | \mathbf{x}) \| p_\theta(\mathbf{z} | \mathbf{x})) &= \mathbb{E}_{q_\phi(\mathbf{z} | \mathbf{x})} \left[ \log q_\phi(\mathbf{z} | \mathbf{x}) - \log p_\theta(\mathbf{z} | \mathbf{x}) \right] \\ &= \mathbb{E}_{q_\phi(\mathbf{z} | \mathbf{x})} \left[ \log q_\phi(\mathbf{z} | \mathbf{x}) - \log \frac{p_\theta(\mathbf{z}, \mathbf{x})}{p_\theta(\mathbf{x})} \right] \\ &= \log p_\theta(\mathbf{x}) - \underbrace{\mathbb{E}_{q_\phi(\mathbf{z} | \mathbf{x})} \left[ \log p_\theta(\mathbf{z}, \mathbf{x}) - \log q_\phi(\mathbf{z} | \mathbf{x}) \right]}_{\equiv \mathcal{L}_{\phi, \theta}(\mathbf{x})}. \end{aligned} \quad (1.20)$$

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 qui nous intéresse  $p_\theta(\mathbf{x})$ . L'objectif d'un modèle VAE est de maximiser la ELBO,  $\mathcal{L}_{\phi, \theta}$ .

En observant l'équation (1.20), on réalise que ceci accomplit deux objectifs simultanément qui suivent du fait que la divergence KL est une quantité positive :

1. Améliorer le processus génératif  $p_\theta(\mathbf{x})$  puisque  $\log p_\theta(\mathbf{x}) \geq \mathcal{L}_{\phi,\theta}(\mathbf{x})$  ;
2. Améliorer le processus d'inférence puisque  $D_{\text{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) \parallel p_\theta(\mathbf{z} \mid \mathbf{x})) = \log p_\theta(\mathbf{x}) - \mathcal{L}_{\phi,\theta}(\mathbf{x})$  est simultanément minimisé.

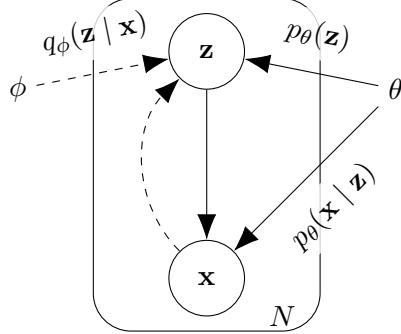


FIGURE 1.3: Modèle graphique d'un VAE. Les flèches pleines indiquent le processus génératif, alors que les flèches pointillées indiquent le processus d'inférence.

### 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  $\phi$  apparaît dans la distribution de prélevement pour calculer l'espérance mathématique  $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}$  dans la ELBO (1.20). 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 restreindre la forme fonctionnelle de  $q_\phi(\mathbf{z} \mid \mathbf{x})$  à une famille paramétrique qui s'exprime comme la transformation différentiable d'une variable aléatoire auxiliaire  $\epsilon$ . On considère le cas où  $q_\phi(\mathbf{z} \mid \mathbf{x})$  et  $p(\epsilon)$  font partie de la famille gaussienne isotropique

$$p(\epsilon) \equiv \mathcal{N}(0, \mathbb{1}); \quad (1.21)$$

$$q_\phi(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_\phi(\mathbf{x}), \mathbb{1} e^{\log \boldsymbol{\sigma}_\phi^2(\mathbf{x})}); \quad (1.22)$$

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

$\odot$  symbolise le produit d'Hadamard, ou encore le produit élément-par-élément de vecteurs. La reparamétrisation fait en sorte que les paramètres variationnelles ne participent plus au processus de prélevement, maintenant pris en charge par  $\epsilon$ . Cette propriété est cruciale dans le but de prendre le gradient de la ELBO (1.20). 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

d'une distribution gaussienne isotropique (1.22),  $f_\phi(\mathbf{x}) = (\boldsymbol{\mu}, \log \boldsymbol{\sigma}^2)$ , étant donné la valeur d'un échantillon  $\mathbf{x}$ . En pratique, on peut construire une approximation de cette fonction avec un réseau de neurones convolutionnelles.

Pour déterminer la forme fonctionnelle de la ELBO, on stipule a priori que la distribution marginale des variables latentes devrait correspondre à une distribution normale isotropique

$$p_\theta(\mathbf{z}) = \mathcal{N}(0, \mathbb{1}) \quad (1.24)$$

On est libre de faire ce choix sans pour autant limiter les formes possibles de la distribution qui nous intéresse  $p_\theta(\mathbf{x})$ . On peut alors exprimer la ELBO comme

$$\mathcal{L}_{\phi,\theta}(\mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[ \log p_\theta(\mathbf{z}, \mathbf{x}) - \log q_\phi(\mathbf{z} | \mathbf{x}) \right]; \quad (1.25)$$

$$\implies \mathcal{L}_{\phi,\theta}(\mathbf{x}) = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[ \log p_\theta(\mathbf{x} | \mathbf{z}) \right]}_{\text{terme de reconstruction}} + \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[ \log p_\theta(\mathbf{z}) - \log q_\phi(\mathbf{z} | \mathbf{x}) \right]}_{\equiv -D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z}))}. \quad (1.26)$$

La divergence de KL obtenue au second terme du membre droit de l'équation (1.26) admet une solution fermée étant donné les familles paramétriques stipulées pour  $p_\theta(\mathbf{z})$  (1.24) et  $q_\phi(\mathbf{z} | \mathbf{x})$  (1.22)

$$-D_{\text{KL}}(q_\phi(\mathbf{z} | \mathbf{x}) \| p_\theta(\mathbf{z})) = \frac{1}{2} \sum_{j=1}^{\dim(\mathbf{z})} (1 + [\log \boldsymbol{\sigma}_\phi^2]_j - [\boldsymbol{\mu}_\phi]_j - [\boldsymbol{\sigma}_\phi^2]_j) \quad (1.27)$$

Une dérivation de ce terme est donnée dans l'appendice B de [Kingma and Welling \(2013\)](#). Le premier terme du membre droit de l'équation (1.26) 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  $\mathbf{z}^{(1:L)} \stackrel{\text{i.i.d}}{\sim} q_\phi(\mathbf{z} | \mathbf{x})$ <sup>3</sup> pour obtenir un échantillon de représentations latentes à partir des équations (1.21) 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  $\hat{\mathbf{x}}^{(1:L)}$  similaire à 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  $g_\theta(\mathbf{z}^{(i)}) = \hat{\mathbf{x}}^{(i)}$ . 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

$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[ \log p_\theta(\mathbf{x} | \mathbf{z}) \right] \simeq -\frac{1}{L} \sum_{i=1}^L \|\mathbf{x} - \hat{\mathbf{x}}^{(i)}\|_2^2 \quad (1.28)$$

Je note que la fondation théorique des auto-encodeurs variationnels repose sur le principe plus

---

3. i.i.d : identiquement et indépendamment distribué.

général du goulot d'information ([Tishby et al., 1999](#)) ; un sujet qui n'est pas abordé dans ce travail, mais qui motive l'utilisation de la version  $\beta$ -VAE du modèle esquissé dans cette section. Sans rentrer dans les détails, on note qu'il est possible de dériver l'objectif de notre auto-encodeur via la théorie de l'information de [Shannon \(1948\)](#) en interprétant l'auto-encodeur comme un système de transmission d'information par compression, avec perte. Une approche naïve pour modéliser ce système serait de maximiser le taux d'information transmise par le système, c.-à-d. que le nombre de bit moyen encodé dans une variable latente aléatoire  $Z$ , mesuré par l'information mutuelle entre le message  $X$  et le code  $Z$  utilisé pour représenter le message  $I(X; Z)$ , devrait se rapprocher d'un maximum qu'on nomme la capacité du système  $C = \max_{P(X)} I(X; Z)$ . Toutefois, cet objectif ne mentionne rien sur la qualité ou la pertinence de cette information. Pour obtenir un message pertinent, on veut contraindre la complexité de Kolmogorov du message, ce qui peut être accompli en contraignant le code  $Z$  à utiliser le moins de bit possible pour encoder le message. C'est le principe de base de la théorie du taux de distortion ([Cover and Thomas, 2006](#)). [Tishby et al. \(1999\)](#) observe que la mesure du taux de distortion suivante

$$\mathcal{L}[p(\hat{\mathbf{x}} | \mathbf{x})] = I(\hat{X}; X) - \beta I(\hat{X}; Z) \quad (1.29)$$

Le paramètre  $\beta$  est un multiplicateur de Lagrange qui contrôle le niveau de compression désiré. Le lecteur est invité à se référer à la revue sur le sujet par [Goldfeld and Polyanskiy \(2020\)](#).

## 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

$$\mathbf{y} = F(\mathbf{x}) + \boldsymbol{\eta}, \quad (1.30)$$

où  $\mathbf{y} \in \mathcal{Y}$  est un vecteur d'observables (comme l'image capturé par les détecteurs CCD dans un télescope),  $\mathbf{x} \in \mathcal{X}$  est un vecteur de paramètres qui gouverne 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 sorte qu'on peut modéliser la

fonction de vraisemblance de l'observable

$$\mathbf{y} - F(\mathbf{x}) \sim p(\boldsymbol{\eta}) = p(\mathbf{y} \mid \mathbf{x}). \quad (1.31)$$

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

$$\hat{\mathbf{x}}_{\text{MLE}} = \underset{\mathbf{x} \in \mathcal{X}}{\operatorname{argmax}} \log p(\mathbf{y} \mid \mathbf{x}). \quad (1.32)$$

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 (1.32) 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 Lipshitz.

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 Lipshitz  $L \geq 0$

$$\|G_\varphi(\mathbf{y}_1) - G_\varphi(\mathbf{y}_2)\|_{\mathcal{X}} \leq L\|\mathbf{y}_1 - \mathbf{y}_2\|_{\mathcal{Y}}, \quad (1.33)$$

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 (1.32).

La condition d'unicité de la solution est résolue par la construction d'une mesure de probabilité a priori sur l'espace des paramètres  $p_\theta : \mathcal{X} \rightarrow \mathbb{R}$ , 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 (1.32) en introduisant cette distribution a priori comme un terme de régularisation de la vraisemblance

$$\hat{\mathbf{x}}_{\text{MAP}} = \underset{\mathbf{x} \in \mathcal{X}}{\operatorname{argmax}} \log p(\mathbf{y} \mid \mathbf{x}) + \log p_\theta(\mathbf{x}). \quad (1.34)$$

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

$$p_\theta(\mathbf{x} \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \mathbf{x})p_\theta(\mathbf{x})}{\int_{\mathcal{X}} p(\mathbf{y} \mid \mathbf{x})p_\theta(\mathbf{x})d\mathbf{x}}. \quad (1.35)$$

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](#)), soit une fonction de type  $f_{\mathbf{x}} : \mathbb{R}^2 \rightarrow \mathbb{R}_+$  modélisée 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, particulièrement pour les observations de basses qualités. 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 (1.34), on considère la méthode de discréétisation d'Euler

$$\hat{\mathbf{x}}^{(t+1)} = \hat{\mathbf{x}}^{(t)} + \alpha \nabla_{\hat{\mathbf{x}}^{(t)}} p_\theta(\hat{\mathbf{x}}^{(t)} \mid \mathbf{y}), \quad (1.36)$$

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 (1.36), 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 (1.36) de la façon suivante

$$\hat{\mathbf{x}}^{(t+1)} = \hat{\mathbf{x}}^{(t)} + \alpha (\nabla_{\hat{\mathbf{x}}^{(t)}} \log p(\mathbf{y} | \hat{\mathbf{x}}^{(t)}) + \nabla_{\hat{\mathbf{x}}^{(t)}} \log p_\theta(\hat{\mathbf{x}}^{(t)})); \quad (1.37)$$

$$\implies \hat{\mathbf{x}}^{(t+1)} = \hat{\mathbf{x}} + g_{\varphi^{(t)}}(\hat{\mathbf{x}}^{(t)}, \nabla_{\hat{\mathbf{x}}^{(t)}} \log p(\mathbf{y} | \hat{\mathbf{x}}^{(t)})) \quad (1.38)$$

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 (1.36) est un cas spécial de la relation (1.38), 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 (1.38), 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 convolutifs avec une architecture de sablier ou encore une architecture en forme de U ([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  $g_{\varphi^{(t)}}$  satisfait la condition de Lipschitz, 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. ). Nous reportons l'étude de la troisième condition d'Hadamard pour des travaux futurs.

Finalement, on note un aspect important du modèle  $g_{\varphi^{(t)}}$ , soit la possible dépendance envers  $t$ . Cet aspect est directement inspiré des succès récents d'algorithmes d'optimisations comme la méthode d'accélération de [Nesterov \(1983\)](#), RPROP ([Riedmiller and Braun, 1993](#)), AdaGrad ([Duchi et al., 2011](#)), RMSProp<sup>4</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 (1.38). Cette propriété permet à ces algorithmes de collecter de l'information par rapport à la seconde dérivée de la fonction objective, sans la calculer directement. Ainsi, il est important de considérer une classe de modèles avec une mémoire des itérations précédentes. Pour ce faire, on utilise des unités récurrentielles à porte (de l'anglais *gated recurrent units* : [Cho et al., 2014](#)) pour modéliser une fonction  $g_\varphi$  augmentée d'un ensemble d'états latents  $\{\mathbf{h}_i^{(t)}\}_{i=1}^H$  qui agissent comme une mémoire des activations précédentes du réseau de neurones. Les détails de cette couche neuronale sont données dans l'annexe D.

---

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

Comme ADAM est considéré comme l'algorithme le plus performant parmi ceux énumérés précédemment, une machine à inférence récurrente bénéficie énormément de son utilisation pour prétraiter le gradient de la vraisemblance  $\nabla_{\mathbf{x}} \log p(\mathbf{y} | \mathbf{x})$  avant de le passer en entrée au réseau de neurones  $g_\varphi$ . Cette idée a fait une première apparition dans les travaux de [Modi et al. \(2021\)](#), puis dans notre travail présenté au chapitre ??.

### 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 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, Schmidhuber) 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 garantis 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 données 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 tâches à performer. Dans notre

cas, la tâche à performer est l'optimisation d'une fonction de vraisemblance. On a donc

$$\mathcal{D} = \{\mathbf{x}_i, \log p_i(\mathbf{y} \mid \mathbf{x})\}_{i=1}^N \quad (1.39)$$

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 construire par la relation de récurrence par une erreur quadratique moyenne

$$\mathcal{L}_\varphi(\mathbf{x}, \log p(\mathbf{y} \mid \mathbf{x})) = \sum_{t=1}^T w^{(t)} \|\mathbf{x} - \hat{\mathbf{x}}^{(t)}\|_{\mathcal{X}}^2 \quad (1.40)$$

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

$$\varphi_D^\star = \operatorname{argmin}_\varphi \mathbb{E}_{\mathcal{D}} [\mathcal{L}_\varphi] \quad (1.41)$$

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 entré 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 transferrable 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 sortes 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 prior empirique déterminée par  $\mathcal{D}$ .

# Bibliographie

- 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. TOPICAL REVIEW Gravitational lensing. *Classical and Quantum Gravity*, 27(23) :233001, Dec. 2010. doi : 10.1088/0264-9381/27/23/233001.
- 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/annurev.aa.30.090192.001523>.
- J. C. Butcher. *Numerical Methods for Ordinary Differential Equations*, chapter 2, pages 21–26. John Wiley & Sons, Hoboken, New Jersey, third edition, 2016. ISBN 9781119121503. doi : 10.1002/9781119121534. URL <https://onlinelibrary.wiley.com/doi/book/10.1002/9781119121534>.
- S. Carroll. *Spacetime and Geometry : An Introduction to General Relativity*. Benjamin Cummings, 2003. ISBN 0805387323.
- K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv e-prints*, art. arXiv :1406.1078, June 2014.
- O. Chwolson. Über eine mögliche form fiktiver doppelsterne. *Astronomische Nachrichten*, 221(20) :329–330, 1924. doi : <https://doi.org/10.1002/asna.19242212003>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/asna.19242212003>.
- 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.
- 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>.
- Z. Goldfeld and Y. Polyanskiy. The Information Bottleneck Problem and Its Applications in Machine Learning. *arXiv e-prints*, art. arXiv :2004.14941, Apr. 2020.
- A. Goobar, R. Amanullah, S. R. Kulkarni, P. E. Nugent, J. Johansson, C. Steidel, D. Law, E. Mörtsell, R. Quimby, N. Blagorodnova, A. Brandeker, Y. Cao, A. Cooray, R. Ferretti, C. Fremling, L. Hangard, M. Kasliwal, T. Kupfer, R. Lunnan, F. Masci, A. A. Miller, H. Nayyeri, J. D. Neill, E. O. Ofek, S. Papadogiannakis, T. Petrushevska, V. Ravi, J. Sollerman, M. Sullivan, F. Taddia, R. Walters, D. Wilson, L. Yan, and O. Yaron. iPTF16geu : A multiply imaged, gravitationally lensed type Ia supernova. *Science*, 356(6335) :291–295, Apr. 2017. doi : 10.1126/science.aal2729.
- 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. Accès 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>.
- E. Hubble. A Relation between Distance and Radial Velocity among Extra-Galactic Nebulae. *Proceedings of the National Academy of Science*, 15(3) :168–173, Mar. 1929. doi : 10.1073/pnas.15.3.168.
- P. L. Kelly, S. A. Rodney, T. Treu, R. J. Foley, G. Brammer, K. B. Schmidt, A. Zitrin, A. Sonnenfeld, L.-G. Strolger, O. Graur, A. V. Filippenko, S. W. Jha, A. G. Riess, M. Bradac, B. J. Weiner, D. Scolnic, M. A. Malkan, A. von der Linden, M. Trenti, J. Hjorth, R. Gavazzi, A. Fontana, J. C. Merten, C. McCully, T. Jones, M. Postman, A. Dressler, B. Patel, S. B. Cenko, M. L. Graham, and B. E. Tucker. Multiple images of a highly magnified supernova formed by an early-type cluster galaxy lens. *Science*, 347(6226) :1123–1126, Mar. 2015. doi : 10.1126/science.aaa3350.
- 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.

- 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.
- 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.
- M. Meneghetti. *Introduction to Gravitational Lensing*. Springer Cham, 2013. doi : 10.1007/978-3-030-73582-1.
- C. Modi, F. Lanusse, U. Seljak, D. N. Spergel, and L. Perreault-Levasseur. CosmicRIM : Reconstructing Early Universe by Combining Differentiable Simulations with Recurrent Inference Machines. *arXiv e-prints*, art. arXiv :2104.12864, Apr. 2021.
- 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.
- P. Putzky and M. Welling. Recurrent Inference Machines for Solving Inverse Problems. *arXiv e-prints*, 2017.
- S. Refsdal. On the possibility of determining Hubble’s parameter and the masses of galaxies from the gravitational lens effect. *MNRAS*, 128 :307, Jan. 1964. doi : 10.1093/mnras/128.4.307.
- M. Riedmiller and H. Braun. A direct adaptive method for faster backpropagation learning : the rprop algorithm. In *IEEE International Conference on Neural Networks*, pages 586–591 vol.1, 1993. doi : 10.1109/ICNN.1993.298623.
- 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.
- 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>.
- A. Stockton. The lens galaxy of the twin QSO 0957+561. *ApJ*, 242 :L141–L144, Dec. 1980. doi : 10.1086/183419.
- S. H. Suyu, V. Bonvin, F. Courbin, C. D. Fassnacht, C. E. Rusu, D. Sluse, T. Treu, K. C. Wong, M. W. Auger, X. Ding, S. Hilbert, P. J. Marshall, N. Rumbaugh, A. Sonnenfeld, M. Tewes, O. Tihhonova, A. Agnello, R. D. Blandford, G. C. F. Chen, T. Collett, L. V. E. Koopmans, K. Liao, G. Meylan, and C. Spinelli. H0LiCOW - I.  $H_0$  Lenses in COSMOGRAIL’s Wellspring : program overview. *MNRAS*, 468(3) :2590–2604, July 2017. doi : 10.1093/mnras/stx483.
- 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.
- T. Treu and P. J. Marshall. Time delay cosmography. *A&A Rev.*, 24(1) :11, July 2016. doi : 10.1007/s00159-016-0096-8.
- C. Vanderriest, J. Schneider, G. Herpe, M. Chevreton, M. Moles, and G. Wlerick. The value of the time delay delta  $T$  (A,B) for the ‘double’ quasar 0957+561 from optical photometric monitoring. *A&A*, 215 :1–13, May 1989.

- 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.
- 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>.

## Annexe A

# 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

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

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

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

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

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

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

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

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  $\chi^2$  :

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

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 (A.3) is constant. Thus, the posterior becomes proportional to

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

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

## Annexe B

# 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 B.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 B.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 C

# 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 C.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

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

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.  $\mathbf{w}_i = \frac{1}{M}$ .

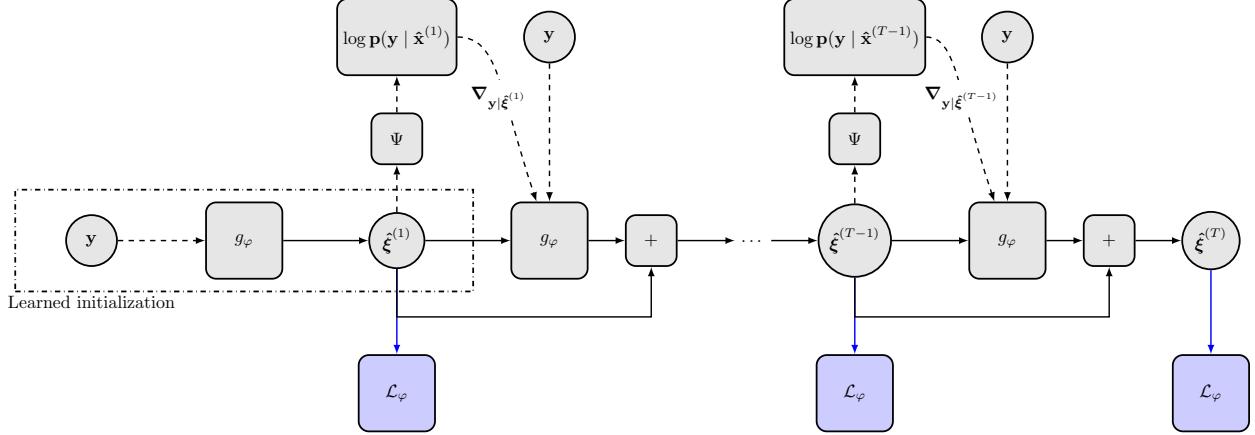


FIGURE C.1: Unrolled computational graph of the RIM. Operations along solid arrows are being recorded for BPTT, while operations along dashed arrows are not. The blue arrows are only used for optimisation during training. During fine-tuning or testing, the loss is computed only as an oracle metric to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

TABLE C.1: Hyperparameters for the RIM.

Parameter	Value
Source link function	$\text{1}$
$\kappa$ link function	$10^{\epsilon}$
<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

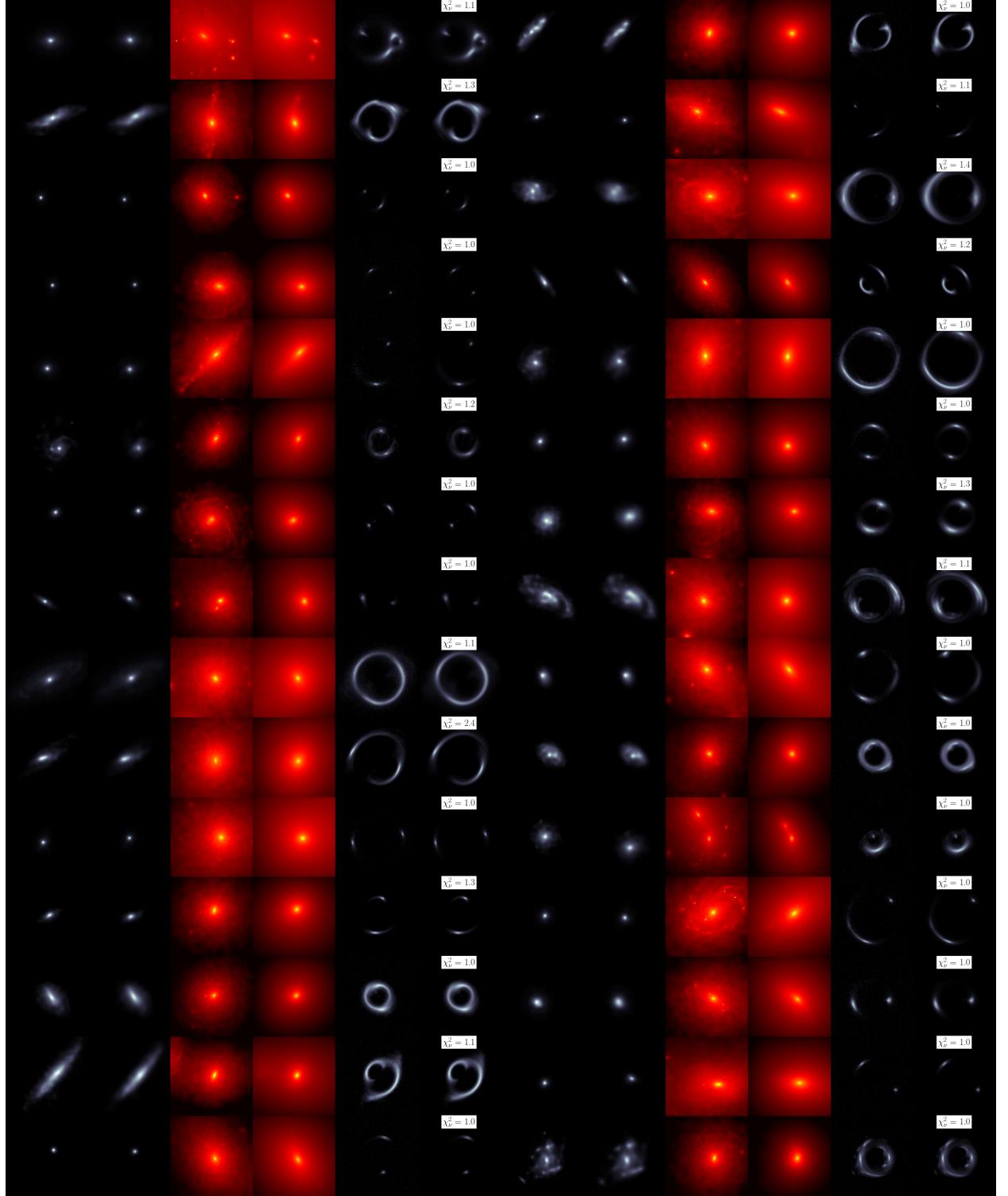


FIGURE C.2: 30 reconstructions taken at random from the test set of 3000 examples simulated from COSMOS and IllustrisTNG data at high SNR. The colorscale are the same as in Figure ??.

## Annexe D

# GRU

Une unité récurrente à porte convolutionnelles est décrite par les opérations

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

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

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

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

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  $\mathbf{z}$ .