

RECONSTRUCTION D'IMAGE 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

Résumé	ii
Abstract	iii
Liste des tableaux	vi
Liste des figures	vii
Acronymes	viii
Liste des symboles	ix
Remerciements	xii
1 Introduction	2
1.1 Lentilles gravitationnelles de type Galaxie-Galaxie	2
1.1.1 Les angles de déflections	2
1.1.2 Applications	6
1.2 Interférométrie par masque non-régulier	6
1.2.1 Les angles de fermeture	6
1.2.2 Applications	6
1.3 Auto-encodeur variationnel	6
1.3.1 Description du modèle	6
1.3.2 Le truc de reparamétrisation	8
1.4 Machines à inférence récurrentielles	10
1.4.1 Formalisme bayésien des problèmes inverses	10

1.4.2	La relation de récurrence	12
1.4.3	Méta-apprentissage et transfert de l'apprentissage	13
Bibliographie		14
A	Elastic Weight Consolidation	16
B	VAE Architecture and optimisation	18
C	RIM architecture and optimisation	21

Liste des tableaux

B.1	Hyperparameters for the background source VAE.	19
B.2	Hyperparameters for the convergence VAE.	20
C.1	Hyperparameters for the RIM.	23

Liste des figures

1.1	Schéma d'une lentille gravitationnelle.	6
1.2	Modèle graphique d'un VAE. Les flèches pleines indiquent le processus génératif, alors que les flèches pointillées indiquent le processus d'inférence.	7
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.	22
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 ??.	24

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.

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.

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.
- π Pi.
- ∇ Gradient.
- ∇^2 Laplacien.
- κ Convergence — densité surfacique de masse projeté sur l'axe de visée.
- α Angles de déflections.
- β Coordonnées angulaires du plan de la source.
- θ Coordonnées angulaires du plan de la lentille.
- ξ Coordonnées comobiles sur le plan de la lentille.
- η Coordonnées comobiles sur le plan de la source.
- D_s Distance du diamètre angulaire entre l'observateur et la source.
- D_ℓ Distance du diamètre angulaire entre l'observateur et la lentille.
- $D_{\ell s}$ Distance du diamètre angulaire entre la lentille et la source.
- $g_{\mu\nu}$ Un élément de la métrique.
- $\eta_{\mu\nu}$ Un élément de la métrique de Minkowski.
- \mathcal{L} Lagrangien.
- c Vitesse de la lumière.
- G Constante universelle de la gravitation.
- ρ Densité.
- Σ Densité de surface.
- Σ_c Densité de surface critique.
- Φ Potentiel.
- φ Liste des paramètres pour l'algorithme d'inférence d'un problème inverse.

- ϕ Liste des paramètres pour un processus d’inférence.
- θ Liste des paramètres pour un processus génératif.
- $\hat{\mathbf{x}}^{(t)}$ 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}_φ Fonction objective d’entraînement pour les paramètres φ .
- \mathcal{N} Loi normale.
- $\mathcal{T}\mathcal{N}$ Loi normale tronquée.
- \mathcal{U} Loi uniforme.
- $\boldsymbol{\mu}$ Moyenne.
- $\boldsymbol{\Sigma}$ Covariance.
- σ^2 Variance.
- σ Déviation standard.
- \oplus Concaténation.
- \odot Produit d’Hadamard.

À Maman et Julia

Remerciements

Chapitre 1

Introduction

1.1 Lentilles gravitationnelles de type Galaxie-Galaxie

L'idée des lentilles gravitationnelles est attribuée à Fritz [Zwicky \(1937\)](#) qui, suivant les calculs publié par [Einstein \(1936\)](#), est le premier à postuler correctement que l'anneau d'Einstein pourrait être observé, produit par la déflection de la lumière d'une source lointaine par le champ gravitationnel d'une galaxie en avant-plan de cette source selon le point de vue d'un observateur sur Terre, pourrait être observé. Dans le même article, Zwicky articule précisément les idées qui nous motivent encore aujourd'hui (presque 100 ans plus tard) à étudier ces objets, c'est-à-dire que les lentilles gravitationnelles permettraient

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 ces galaxies.

Il est à noté qu'Einstein considérait la possibilité d'observer ce phénomène extrêmement improbable. La résolution des télescopes optiques durant la majorité du XX^e siècle étaient limités à environ 0.5 arcsecondes par la turbulence de l'air. Avec une telle résolution, un anneau d'Einstein d'une taille caractéristique de 1 arcseconde apparaîtrait comme un point lumineux étalé, et ne serait donc pas distinguable d'une étoile.

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. Des traitements similaires peuvent être trouvé dans les 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 Φ 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'arrivé 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

$$\boldsymbol{\alpha} = - \int_{\lambda_A}^{\lambda_B} \ddot{\mathbf{x}} \times \mathbf{e}_{\parallel} d\lambda, \quad (1.1)$$

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

La trajectoire d'un photon 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

$$\delta T = \delta \int_A^B n(\mathbf{x}(\ell)) \frac{d\ell}{c} = 0, \quad (1.2)$$

où ℓ est un élément de longueur sur la trajectoire et n est un indice de réfraction. Pour déterminer l'indice de réfraction du champ gravitationnel d'une galaxie, on doit utiliser le formalisme de la relativité générale. Selon le principe d'équivalence (fort), l'effet d'un champ gravitationnel est localement 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 Φ d'une galaxie est celui d'un gaz parfait, c'est-à-dire qu'il satisfait une équation de Poisson

$$\nabla^2 \Phi = 4\pi G \rho. \quad (1.3)$$

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

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu \approx \left(1 + \frac{2\Phi}{c^2}\right) c^2 dt^2 - \left(1 - \frac{2\Phi}{c^2}\right) d\mathbf{x}^2. \quad (1.4)$$

Puisqu'un photon suit une géodésique de l'espace-temps $ds^2 = 0$, on peut déterminer l'indice de réfraction en réarrangeant l'équation (1.4)

$$n \equiv c \left(\frac{\|d\mathbf{x}\|}{dt} \right)^{-1} \approx 1 - \frac{2\Phi}{c^2}. \quad (1.5)$$

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 (1.2) sous la forme

$$\delta \int_{\lambda_A}^{\lambda_B} n(\mathbf{x}) \|\dot{\mathbf{x}}\| d\lambda = 0. \quad (1.6)$$

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 (1.2) est une solution

des équations d'Euler-Lagrange

$$\frac{d}{d\lambda} \frac{\partial \mathcal{L}}{\partial \dot{\mathbf{x}}} - \frac{\partial \mathcal{L}}{\partial \mathbf{x}} = 0. \quad (1.7)$$

On a donc

$$\frac{d}{d\lambda} n \frac{\dot{\mathbf{x}}}{\|\dot{\mathbf{x}}\|} - \|\dot{\mathbf{x}}\| \nabla n = 0, \quad (1.8)$$

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

$$\begin{aligned} \frac{d}{d\lambda} n \dot{\mathbf{x}} - \nabla n &= 0 \\ \implies n \ddot{\mathbf{x}} + (\nabla n \cdot \dot{\mathbf{x}}) \dot{\mathbf{x}} - \nabla n &= 0 \end{aligned} \quad (1.9)$$

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, on obtient

$$\ddot{\mathbf{x}} \times \mathbf{e}_{\parallel} = \frac{1}{n} \nabla_{\perp} n = \nabla_{\perp} \log n \approx -\frac{2}{c^2} \nabla_{\perp} \Phi, \quad (1.10)$$

où ∇_{\perp} est un gradient selon les coordonnées perpendiculaires à \mathbf{e}_{\parallel} . On note que le facteur 2 qui apparaît dans l'équation (1.10) 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 ξ 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 1.1). 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 δ est la fonction delta de Dirac, alors le potentiel qui satisfait l'équation de Poisson (1.3) 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 (1.1) se réécrit finalement comme

$$\begin{aligned} \alpha(\xi) &= -\frac{2GM}{c^2} \int_{-\infty}^{\infty} \frac{\partial}{\partial \xi} \frac{1}{(\xi^2 + z^2)^{1/2}} dz \\ \implies \alpha(\xi) &= \frac{4GM}{c^2 \xi^2} \end{aligned} \quad (1.11)$$

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 ξ est alors une convolution sur tout le plan de la lentille (mince) puisque l'équation (1.11) dépend linéairement de la masse M :

$$\alpha(\xi) = \frac{4G}{c^2} \int_{\mathbb{R}^2} \Sigma(\xi') \frac{\xi - \xi'}{\|\xi - \xi'\|^2} d^2 \xi' \quad (1.12)$$

L'angle de déviation est une quantité cruciale pour résoudre une lentille gravitationnelle puisqu'il décrit une transformation des coordonnées angulaires du plan de la lentille ($\boldsymbol{\theta}$) vers les coordonnées angulaires du plan de la source ($\boldsymbol{\beta}$). On assume que les distances entre l'observateur et la lentille D_ℓ , entre l'observateur et la source D_s et entre la lentille et la source $D_{\ell s}$, sont beaucoup plus grandes que les distances perpendiculaires à l'axe de visée $\boldsymbol{\xi}$ ou $\boldsymbol{\eta}$ (voir figure 1.1). Cette approximation est justifiée pour les objets qui nous intéressent, pour lesquels les distances parallèles à l'axe de visée sont généralement de l'ordre du Gpc, alors que les distances perpendiculaire sont généralement de l'ordre du kpc ; soit 6 ordres de grandeurs de différences. Ainsi, on peut faire un argument géométrique (euclidien)

$$\begin{aligned} D_s \boldsymbol{\theta} &= \boldsymbol{\eta}' \\ D_s \boldsymbol{\beta} &= \boldsymbol{\eta} \\ D_{\ell s} \boldsymbol{\alpha} &= \boldsymbol{\eta}' - \boldsymbol{\eta} \\ \implies D_s \boldsymbol{\beta} &= D_s \boldsymbol{\theta} - D_{\ell s} \boldsymbol{\alpha} \end{aligned} \tag{1.13}$$

La dernière relation est l'équation maîtresse qui nous permet de tracer les rayons lumineux d'une source vers un détecteur fictif dans nos simulations. On notera que cette relation reste valide pour un univers courbe et/ou en expansion (c.-à-d. décrit par une géométrie non-euclidienne), à condition qu'on utilise une notion de distance qui satisfait, par définition, la relation trigonométrique euclidienne

$$D \equiv \frac{\xi}{\theta} \tag{1.14}$$

Il est généralement pratique de travailler avec la forme adimensionnelle de l'équation (1.13). On introduit la densité critique

$$\Sigma_c = \frac{c^2}{4\pi G} \frac{D_s}{D_{\ell s} D_\ell}, \tag{1.15}$$

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

$$\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}) = \frac{1}{\pi} \int_{\mathbb{R}^2} \kappa(\boldsymbol{\theta}) \frac{\boldsymbol{\theta} - \boldsymbol{\theta}'}{\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|} d^2 \boldsymbol{\theta}', \tag{1.16}$$

qui satisfait l'équation de la lentille adimensionnelle

$$\boldsymbol{\beta} = \boldsymbol{\theta} - \hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}). \tag{1.17}$$

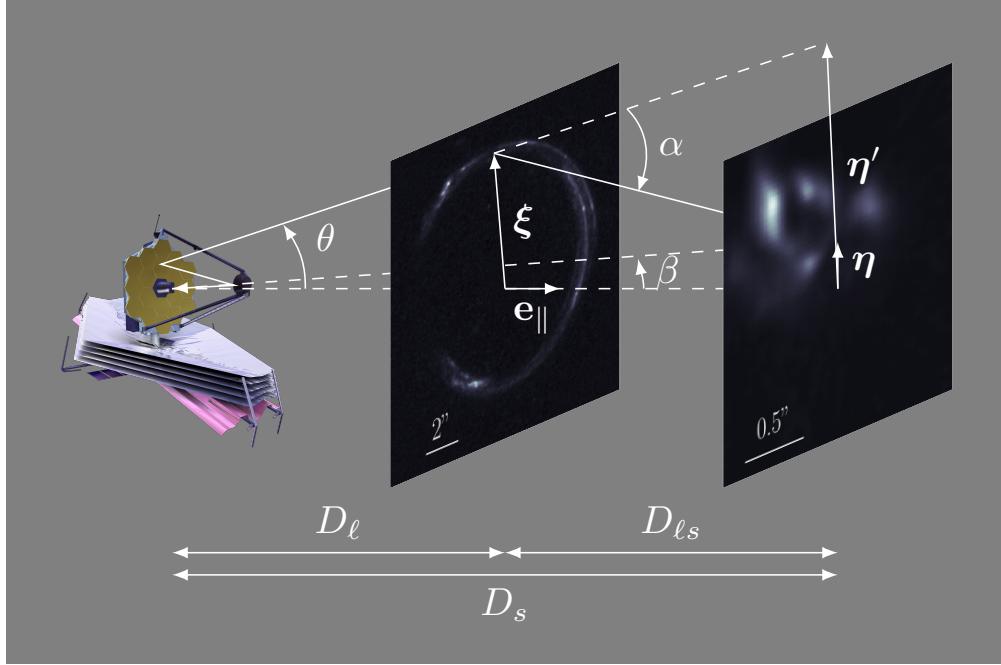


FIGURE 1.1: Schéma d'une lentille gravitationnelle.

1.1.2 Applications

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 θ , conditionnelle à la variable cachée : $p_\theta(\mathbf{x} | \mathbf{z})$. Déterminer

p_θ 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 $\dim(\mathbf{x}) > 10^4$. Pour résoudre cette difficulté, on introduit un modèle paramétrique d'inférence $q_\phi(\mathbf{z} \mid \mathbf{x})$ dont le rôle est de modéliser la distribution a posteriori de la variable latente pour la distribution qui nous intéresse

$$q_\phi(\mathbf{z} \mid \mathbf{x}) \approx p_\theta(\mathbf{z} \mid \mathbf{x}). \quad (1.18)$$

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

$$\begin{aligned} D_{\text{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) \parallel p_\theta(\mathbf{z} \mid \mathbf{x})) &= \mathbb{E}_{q_\phi(\mathbf{z} \mid \mathbf{x})} \left[\log q_\phi(\mathbf{z} \mid \mathbf{x}) - \log p_\theta(\mathbf{z} \mid \mathbf{x}) \right] \\ &= \mathbb{E}_{q_\phi(\mathbf{z} \mid \mathbf{x})} \left[\log q_\phi(\mathbf{z} \mid \mathbf{x}) - \log \frac{p_\theta(\mathbf{z}, \mathbf{x})}{p_\theta(\mathbf{x})} \right] \\ &= \log p_\theta(\mathbf{x}) - \underbrace{\mathbb{E}_{q_\phi(\mathbf{z} \mid \mathbf{x})} \left[\log p_\theta(\mathbf{z}, \mathbf{x}) - \log q_\phi(\mathbf{z} \mid \mathbf{x}) \right]}_{\equiv \mathcal{L}_{\phi, \theta}(\mathbf{x})}. \end{aligned} \quad (1.19)$$

On remarque par cette manipulation que la distance D_{KL} , en plus de mesurer la distance entre les deux distributions a posteriori (par définition), mesure aussi la différence entre le terme $\mathcal{L}_{\phi, \theta}(\mathbf{x})$, qu'on nomme limite inférieure sur l'évidence (de l'anglais *evidence lower bound* : ELBO), et la distribution 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 (1.19), on réalise que 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.

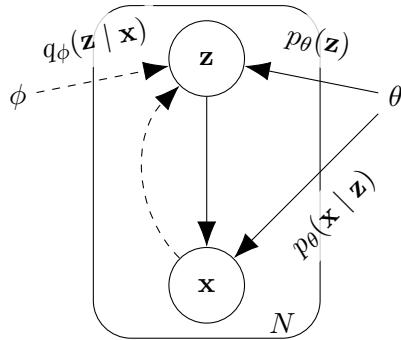


FIGURE 1.2: Modèle graphique d'un VAE. Les flèches pleines indiquent le processus génératif, alors que les flèches pointillées indiquent le processus d'inférence.

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 ϕ et θ . Or, la liste de paramètres ϕ apparaît dans la distribution de prélevement pour calculer l'espérance mathématique $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}$ dans la ELBO (1.19). Cette opération n'a pas de dérivée formelle en terme de ϕ .

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 ϵ . On considère le cas où $q_\phi(\mathbf{z} \mid \mathbf{x})$ et $p(\epsilon)$ font partie de la famille gaussienne isotropique

$$p(\epsilon) \equiv \mathcal{N}(0, \mathbb{1}); \quad (1.20)$$

$$q_\phi(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_\phi(\mathbf{x}), \mathbb{1} e^{\log \boldsymbol{\sigma}_\phi^2(\mathbf{x})}); \quad (1.21)$$

$$\mathbf{z} = \boldsymbol{\mu}_\phi + \boldsymbol{\sigma}_\phi \odot \epsilon. \quad (1.22)$$

\odot symbolise le produit d'Hadamard, ou encore le produit élément-par-élément de vecteurs. La reparamétrisation fait en sorte que les paramètres variationnels ne participent plus au processus de prélevement, maintenant pris en charge par ϵ . Cette propriété est cruciale dans le but de prendre le gradient de la ELBO (1.19). 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, ϕ décrit maintenant une fonction générique dont le rôle est d'inférer les paramètres d'une distribution gaussienne isotropique (1.21), $f_\phi(\mathbf{x}) = (\boldsymbol{\mu}, \log \boldsymbol{\sigma}^2)$, 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

$$p_\theta(\mathbf{z}) = \mathcal{N}(0, \mathbb{1}) \quad (1.23)$$

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

$$\mathcal{L}_{\phi,\theta}(\mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log p_\theta(\mathbf{z}, \mathbf{x}) - \log q_\phi(\mathbf{z} \mid \mathbf{x}) \right]; \quad (1.24)$$

$$\implies \mathcal{L}_{\phi,\theta}(\mathbf{x}) = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log p_\theta(\mathbf{x} \mid \mathbf{z}) \right]}_{\text{terme de reconstruction}} + \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log p_\theta(\mathbf{z}) - \log q_\phi(\mathbf{z} \mid \mathbf{x}) \right]}_{\equiv -D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p_\theta(\mathbf{z}))}. \quad (1.25)$$

La divergence de KL obtenue au second terme du membre droit de l'équation (1.25) admet une

solution fermée étant donné les familles paramétriques stipulées pour $p_\theta(\mathbf{z})$ (1.23) et $q_\phi(\mathbf{z} \mid \mathbf{x})$ (1.21)

$$-D_{\text{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) \parallel p_\theta(\mathbf{z})) = \frac{1}{2} \sum_{j=1}^{\dim(\mathbf{z})} (1 + [\log \boldsymbol{\sigma}_\phi^2]_j - [\boldsymbol{\mu}_\phi]_j - [\boldsymbol{\sigma}_\phi^2]_j) \quad (1.26)$$

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 (1.25) 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 $\mathbf{z}^{(1:L)} \stackrel{\text{i.i.d.}}{\sim} q_\phi(\mathbf{z} \mid \mathbf{x})$ ¹ pour obtenir un échantillon de représentations latentes à partir des équations (1.20) à (1.22), puis en utilisant le modèle génératif $\hat{\mathbf{x}}^{(i)} \sim p_\theta(\mathbf{x} \mid \mathbf{z}^{(i)})$ pour obtenir un échantillon de reconstructions $\hat{\mathbf{x}}^{(1:L)}$ similaire à l'exemple originel \mathbf{x} . Comme on a déjà une variable auxiliaire ϵ qui se charge de l'aspect génératif du modèle, on peut construire une approximation du modèle génératif avec une fonction générique des variables latentes $g_\theta(\mathbf{z}^{(i)}) = \hat{\mathbf{x}}^{(i)}$. 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

$$\mathbb{E}_{q_\phi(\mathbf{z} \mid \mathbf{x})} \left[\log p_\theta(\mathbf{x} \mid \mathbf{z}) \right] \simeq -\frac{1}{L} \sum_{i=1}^L \|\mathbf{x} - \hat{\mathbf{x}}^{(i)}\|_2^2 \quad (1.27)$$

Je note que la fondation théorique des auto-encodeurs variationnels repose sur le principe plus 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 β -VAE du modèle esquisse 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

$$\mathcal{L}[p(\hat{\mathbf{x}} \mid \mathbf{x})] = I(\hat{X}; X) - \beta I(\hat{X}; Z) \quad (1.28)$$

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

Le paramètre β 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

$$\mathbf{y} = F(\mathbf{x}) + \boldsymbol{\eta}, \quad (1.29)$$

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. Dans plusieurs situations, on peut caractériser la distribution de ce bruit, de sortes qu'on peut modéliser la fonction de vraisemblance de l'observable

$$\mathbf{y} - F(\mathbf{x}) \sim p(\boldsymbol{\eta}) = p(\mathbf{y} | \mathbf{x}). \quad (1.30)$$

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

$$\hat{\mathbf{x}}_{\text{MLE}} = \underset{\mathbf{x} \in \mathcal{X}}{\operatorname{argmax}} \log p(\mathbf{y} | \mathbf{x}). \quad (1.31)$$

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 (1.31) 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 $L \geq 0$

$$\|G_\varphi(\mathbf{y}_1) - G_\varphi(\mathbf{y}_2)\|_{\mathcal{X}} \leq L\|\mathbf{y}_1 - \mathbf{y}_2\|_{\mathcal{Y}}, \quad (1.32)$$

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

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. Comme les problèmes qui nous intéressent sont sous-déterminées, 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 ; il y a une infinité de solutions au problème (1.31).

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 (1.31) en introduisant cette distribution a priori comme un terme de régularisation de la vraisemblance

$$\hat{\mathbf{x}}_{\text{MAP}} = \underset{\mathbf{x} \in \mathcal{X}}{\operatorname{argmax}} \log p(\mathbf{y} \mid \mathbf{x}) + \log p_\theta(\mathbf{x}). \quad (1.33)$$

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

$$p_\theta(\mathbf{x} \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \mathbf{x}) p_\theta(\mathbf{x})}{\int_{\mathcal{X}} p(\mathbf{y} \mid \mathbf{x}) p_\theta(\mathbf{x}) d\mathbf{x}}. \quad (1.34)$$

Le dénominateur est une constante qu'on nomme l'évidence bayésienne. 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 (1.33), on considère la méthode de discréétisation d'Euler

$$\hat{\mathbf{x}}^{(t+1)} = \hat{\mathbf{x}}^{(t)} + \alpha \nabla_{\hat{\mathbf{x}}^{(t)}} p_\theta(\hat{\mathbf{x}}^{(t)} | \mathbf{y}), \quad (1.35)$$

où α 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 (1.35), ceci revient à assumer que l'erreur locale de chaque itération est proportionnelle à α^2 , ce qui est satisfait si le gradient $\nabla_{\mathbf{x}} \log p_\theta(\mathbf{x} | \mathbf{y})$ satisfait la condition de Lipschitz dans la région de \mathcal{X} explorée par l'algorithme (Atkinson, 1989; Butcher, 2016), en encore si la norme de la dérivée seconde de $\log p_\theta(\mathbf{x} | \mathbf{y})$ est bornée dans cette région.

Putzky and Welling (2017) observent qu'on peut réécrire (1.35) de la façon suivante

$$\hat{\mathbf{x}}^{(t+1)} = \hat{\mathbf{x}}^{(t)} + \alpha (\nabla_{\hat{\mathbf{x}}^{(t)}} \log p(\mathbf{y} | \hat{\mathbf{x}}^{(t)}) + \nabla_{\hat{\mathbf{x}}^{(t)}} \log p_\theta(\hat{\mathbf{x}}^{(t)})); \quad (1.36)$$

$$\implies \hat{\mathbf{x}}^{(t+1)} = \hat{\mathbf{x}} + g_{\varphi^{(t)}}(\hat{\mathbf{x}}^{(t)}, \nabla_{\hat{\mathbf{x}}^{(t)}} \log p(\mathbf{y} | \hat{\mathbf{x}}^{(t)})) \quad (1.37)$$

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 (1.35) est un cas spécial de la relation (1.37), soit le cas où on a un modèle explicite pour la distribution a priori (ou son gradient) $\nabla_{\mathbf{x}} \log p_\theta(\mathbf{x})$ et le taux d'apprentissage α . Les paramètres α et θ sont absorbés dans les paramètres d'inférence $\varphi^{(t)}$, ce qui nous donne une plus grande liberté pour modéliser la distribution a priori. 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 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 $g_{\varphi^{(t)}}(\cdot, \cdot)$ 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 la question de la constante 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 de Nesterov (Nesterov, 1983), RPROP (Riedmiller and Braun, 1993), AdaGrad (Duchi et al., 2011), RMSProp² (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 (1.37). Ainsi, il est important de considérer une classe de modèle avec une certaine 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_φ augmentée d'un ensemble d'états latents $\{\mathbf{h}_i^{(t)}\}_{i=1}^H$ qui agissent comme une mémoire des activations précédentes ($0 < t' < t$) du réseau de neurones. Pour un réseau convolutionnel, cette unité est représentée par les opérations suivantes :

$$\tilde{\mathbf{x}} = S\left(\mathbf{w}_o * (\mathbf{h}^{(t-1)} \oplus \mathbf{x}^{(t)}) + \mathbf{b}_o\right) \quad \{\text{Porte d'oubli}\} \quad (1.38)$$

$$\mathbf{z} = S\left(\mathbf{w}_z * (\mathbf{h}^{(t-1)} \oplus \mathbf{x}^{(t)}) + \mathbf{b}_z\right) \quad \{\text{Porte de mise à jour}\} \quad (1.39)$$

$$\tilde{\mathbf{h}} = \tanh\left(\mathbf{w}_h * ((\mathbf{h}^{(t-1)} \odot \tilde{\mathbf{x}}) \oplus \mathbf{x}^{(t)}) + \mathbf{b}_h\right) \quad \{\text{État candidat}\} \quad (1.40)$$

$$\mathbf{h}^{(t)} = \mathbf{h}^{(t-1)} \odot \mathbf{z} + \tilde{\mathbf{h}} \odot (1 - \mathbf{z}) \quad \{\text{Nouvel état}\} \quad (1.41)$$

où $S(x) = \frac{1}{1+e^{-x}}$ est une fonction sigmoïde et \mathbf{x} est un état d'activation interne au réseau de neurones. 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 concaténation.

1.4.3 Méta-apprentissage et transfert de l'apprentissage

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

Bibliographie

- 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.
- 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.
- T. M. Cover and J. A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006. ISBN 0471241954.
- 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. 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.
- J. Hadamard. Sur les problèmes aux dérivées 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, 2012. Accès le 2022-07-10.

- 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.
- 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.
- 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. Yildirim, 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>.
- 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>.
- 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 φ can be rewritten using the Bayes rule as

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

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

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

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

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 χ^2 :

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

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

Since the full Fisher matrix is intractable for a neural network, we approximate the quadratic term of the prior with the diagonal of the Fisher matrix following Kirkpatrick et al. (2016). For an optimisation problem, the first term of (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 λ 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
β_{\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
β_{\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}_φ must be computed in the Ξ space in order to avoid gradient vanishing problems when Ψ is a non-linear mapping, which happens if the non-linear link function is applied in an operation recorded for backpropagation through time (BPTT).

For the convergence, we use an exponential link function with base 10 : $\hat{\kappa} = \Psi(\boldsymbol{\xi}) = 10^{\boldsymbol{\xi}}$. This Ψ encodes the non-negativity of the convergence. Furthermore, it is a power transformation that leaves the linked pixel values ξ_i normally distributed, thus improving the learning through the 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. $\mathbf{w}_i = \frac{1}{M}$.

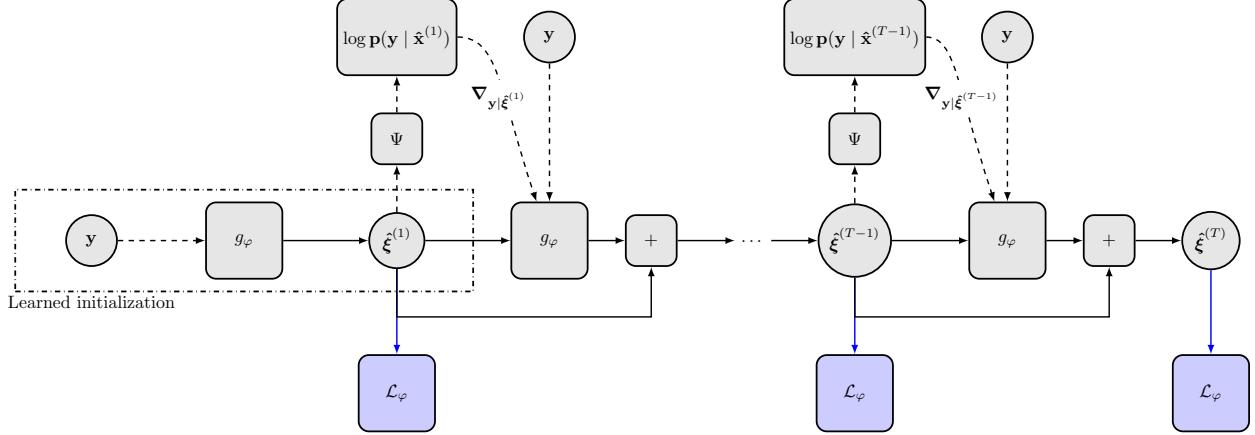


FIGURE C.1: Unrolled computational graph of the RIM. Operations along solid arrows are being recorded for BPTT, while operations along dashed arrows are not. The blue arrows are only used for optimisation during training. During fine-tuning or testing, the loss is computed only as an oracle metric to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

TABLE C.1: Hyperparameters for the RIM.

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

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

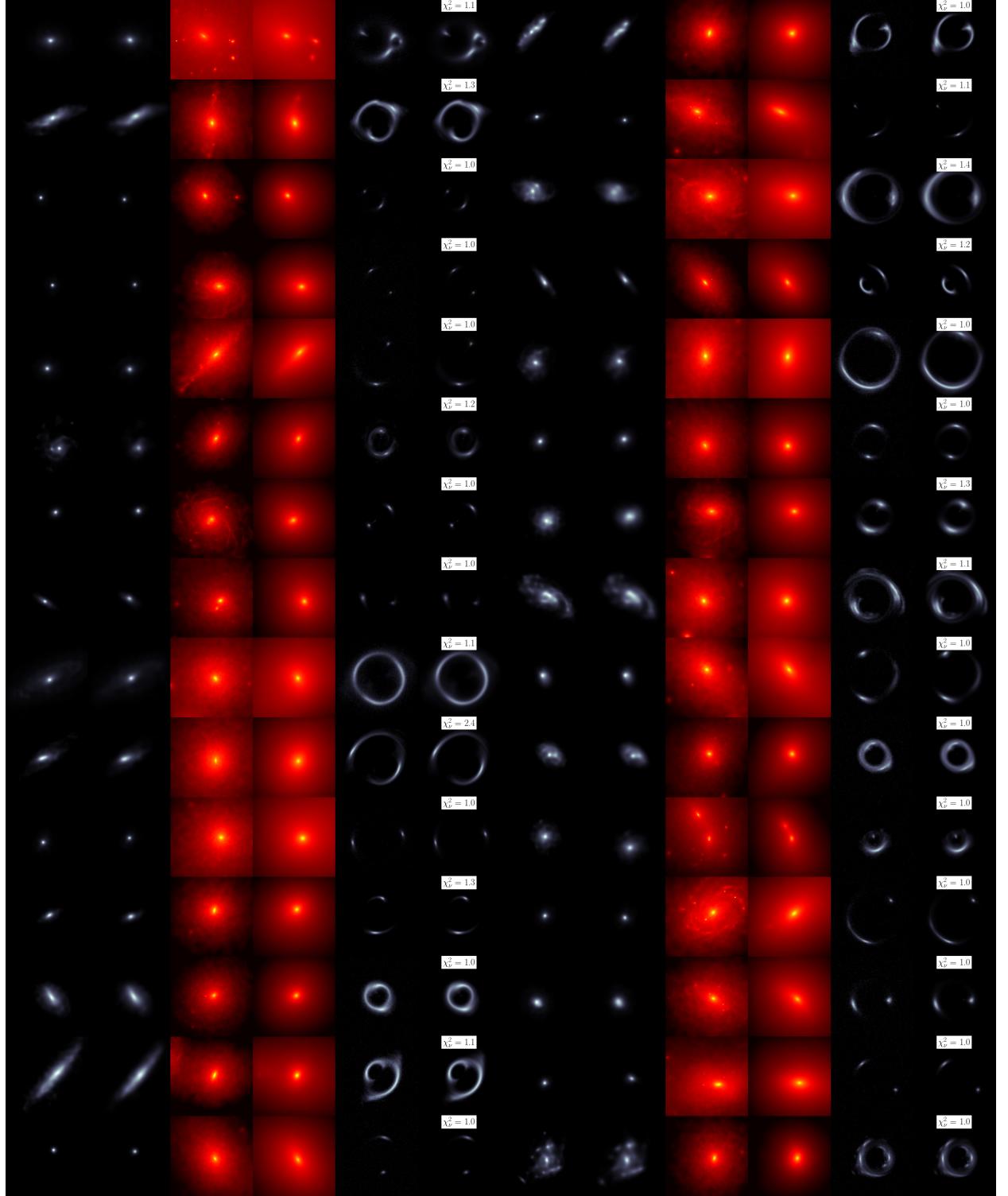


FIGURE C.2: 30 reconstructions taken at random from the test set of 3000 examples simulated from COSMOS and IllustrisTNG data at high SNR. The colorscale are the same as in Figure ??.