

Modèles génératifs profonds

Guillaume Bourmaud

PLAN

I. Introduction

II. Notion de divergence

III. Réseau inversible (NF)

IV. Auto-encodeur variationnel (VAE)

V. Réseau débruiteur (DDPM)

VI. Réseau antagoniste (GAN)

I) Introduction

1)

Principe d'un modèle génératif

Base de données : $\{\mathbf{X}_i\}_{i=1\dots N}$

↓
 $\in \mathbb{R}^n$



1)

Principe d'un modèle génératif

Base de données : $\{\mathbf{X}_i\}_{i=1\dots N}$

↓
 $\in \mathbb{R}^n$



Objectif : Optimiser les paramètres d'un réseau de neurones profond afin de générer de **nouveaux échantillons** qui **ressemblent** aux $\{\mathbf{X}_i\}_{i=1\dots N}$

1)

Principe d'un modèle génératif

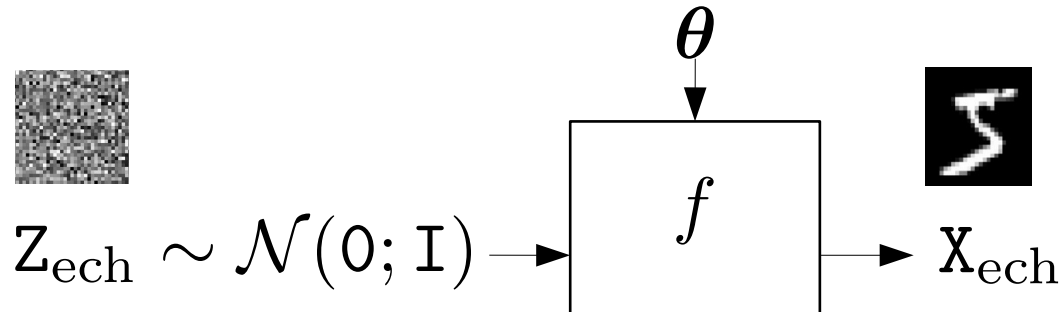
Base de données : $\{\mathbf{X}_i\}_{i=1\dots N}$

$$\downarrow$$

$$\in \mathbb{R}^n$$



Objectif : Optimiser les paramètres d'un réseau de neurones profond afin de générer de **nouveaux échantillons** qui **ressemblent** aux $\{\mathbf{X}_i\}_{i=1\dots N}$



I)

Principe d'un modèle génératif

Base de données : $\{X_i\}_{i=1\dots N}$

$$\downarrow$$

$$\in \mathbb{R}^n$$

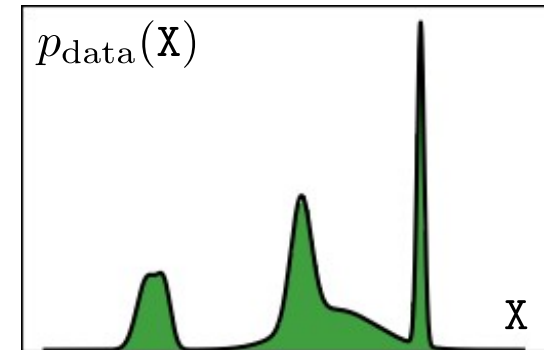


Point de vue probabiliste

“est un échantillon de”

$$X_i \sim p_{\text{data}}(X)$$

Densité de probabilité
inconnue



Exemple 1D

I)

Principe d'un modèle génératif

Base de données : $\{X_i\}_{i=1\dots N}$

↓
 $\in \mathbb{R}^n$



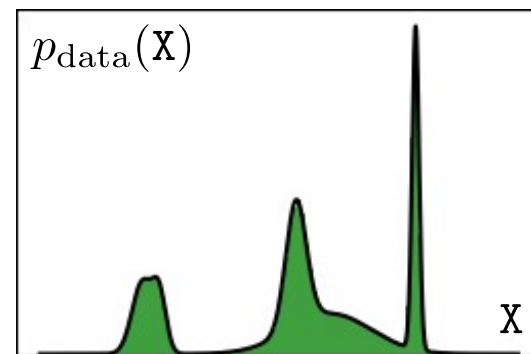
Point de vue probabiliste

“est un échantillon de”

$X_i \sim p_{\text{data}}(X)$

Densité de probabilité
inconnue

On supposera les X_i i.i.d.



Exemple 1D

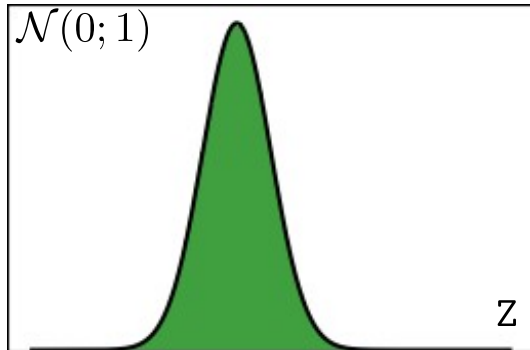
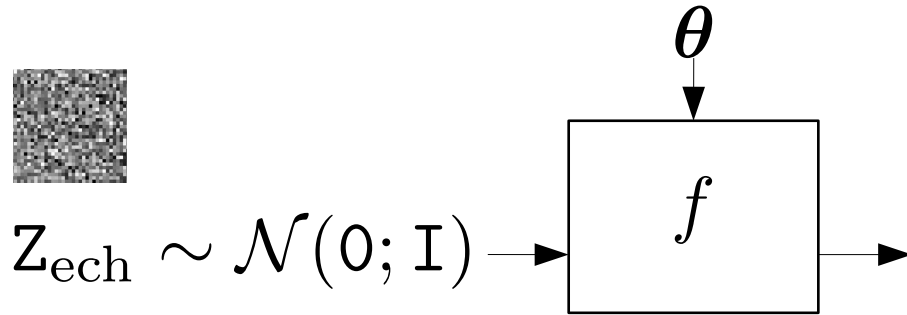
Principe d'un modèle génératif (suite)

Objectif : Optimiser les paramètres d'un réseau de neurones profond afin de générer de **nouveaux échantillons** de $p_{\text{data}}(\mathbf{X})$

1)

Principe d'un modèle génératif (suite)

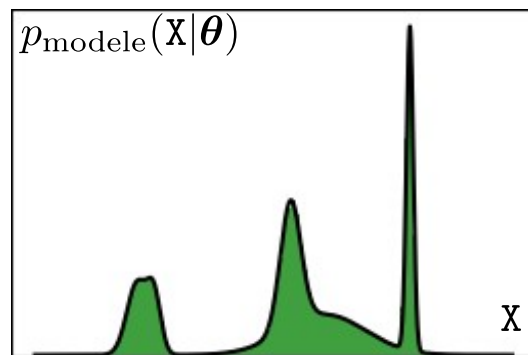
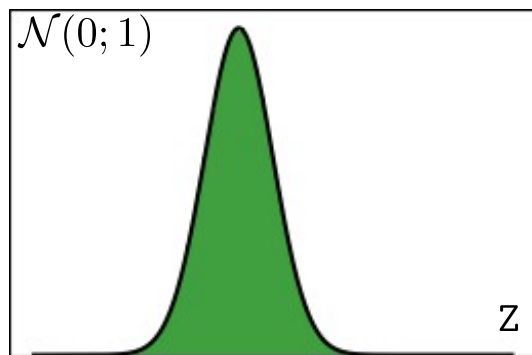
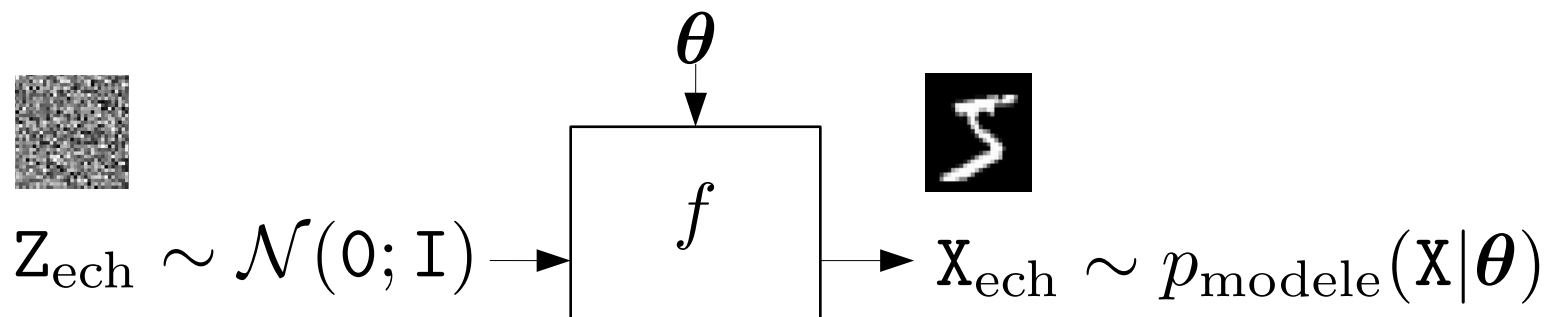
Objectif : Optimiser les paramètres d'un réseau de neurones profond afin de générer de **nouveaux échantillons** de $p_{\text{data}}(\mathbf{X})$



1)

Principe d'un modèle génératif (suite)

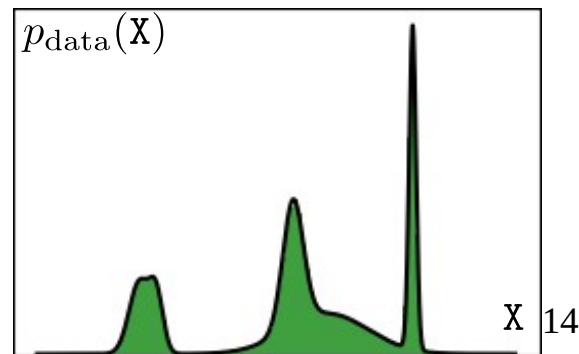
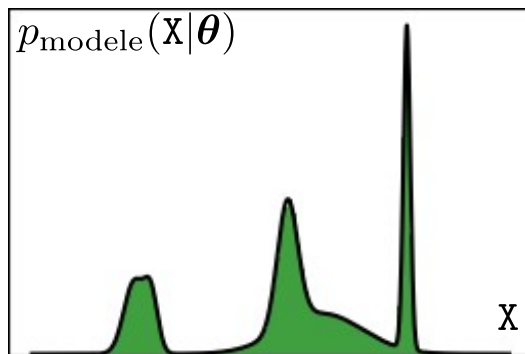
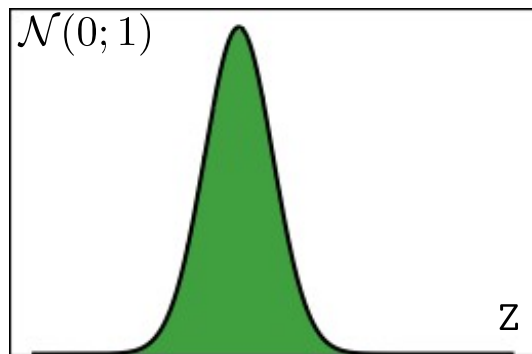
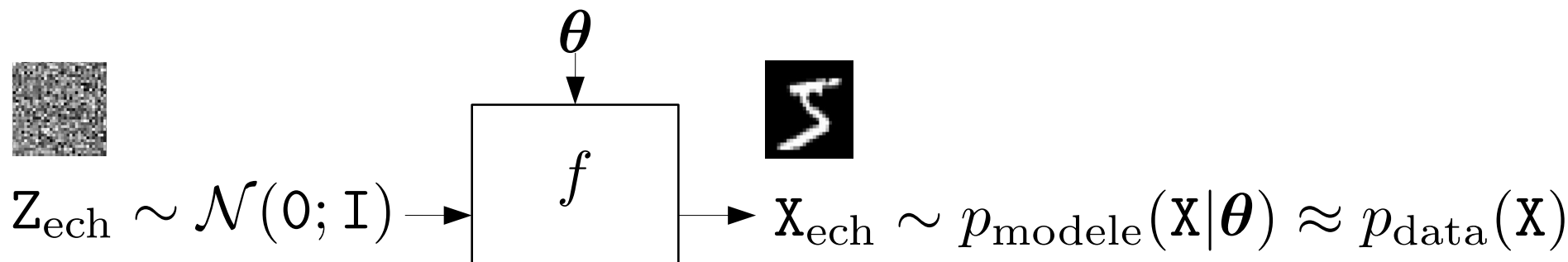
Objectif : Optimiser les paramètres d'un réseau de neurones profond afin de générer de **nouveaux échantillons** de $p_{\text{data}}(\mathbf{X})$



1)

Principe d'un modèle génératif (suite)

Objectif : Optimiser les paramètres d'un réseau de neurones profond afin de générer de **nouveaux échantillons** de $p_{\text{data}}(\mathbf{X})$



1)

Difficulté de la tâche d'apprentissage

Base de données **non-étiquetées** : $\{X_i\}_{i=1\dots N}$



Difficulté de la tâche d'apprentissage

Base de données **non-étiquetées** : $\{X_i\}_{i=1\dots N}$

On ne dispose pas de couples $\{Z_i, X_i\}_{i=1\dots N}$

→ **problème d'apprentissage non-supervisé**



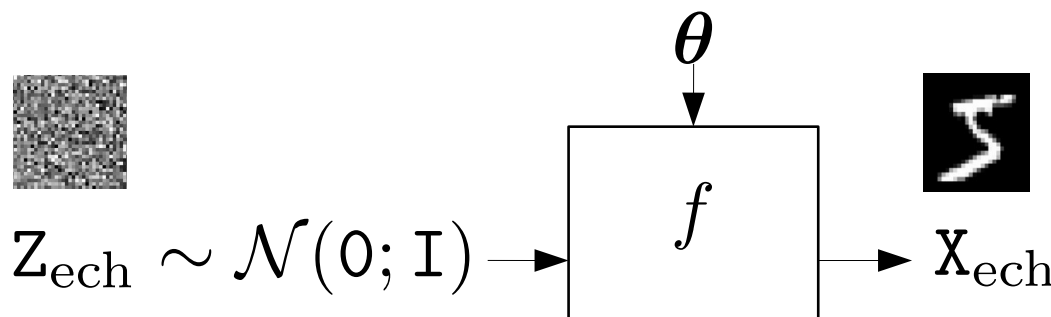
1)

Difficulté de la tâche d'apprentissage

Base de données **non-étiquetées** : $\{X_i\}_{i=1\dots N}$

On ne dispose pas de couples $\{Z_i, X_i\}_{i=1\dots N}$

→ **problème d'apprentissage non-supervisé**

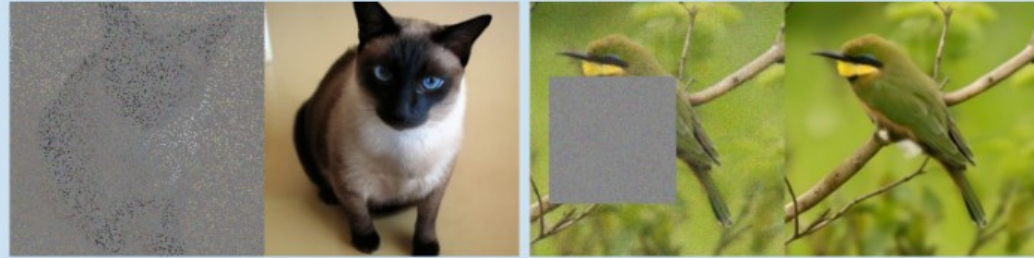


Terminologie : Z_i = variable “latente”

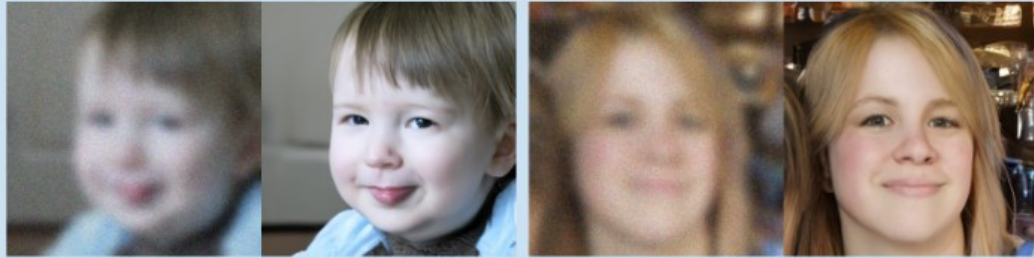
1)

Exemple d'utilisation : Apprentissage d'a priori pour des problèmes inverses

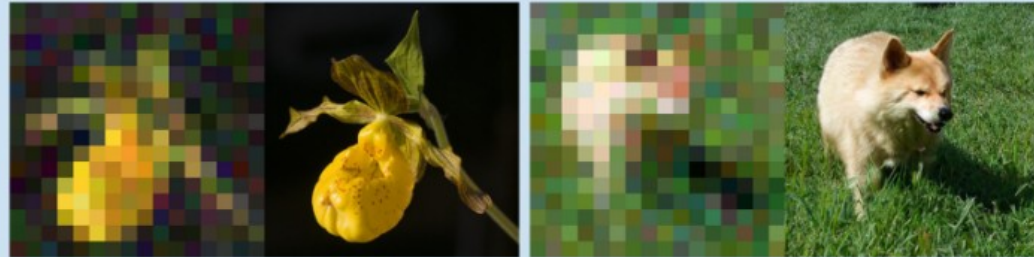
(a) Inpainting



(c) Gaussian deblur



(b) Super-resolution



(d) Motion deblur



Principe d'un modèle génératif conditionnel

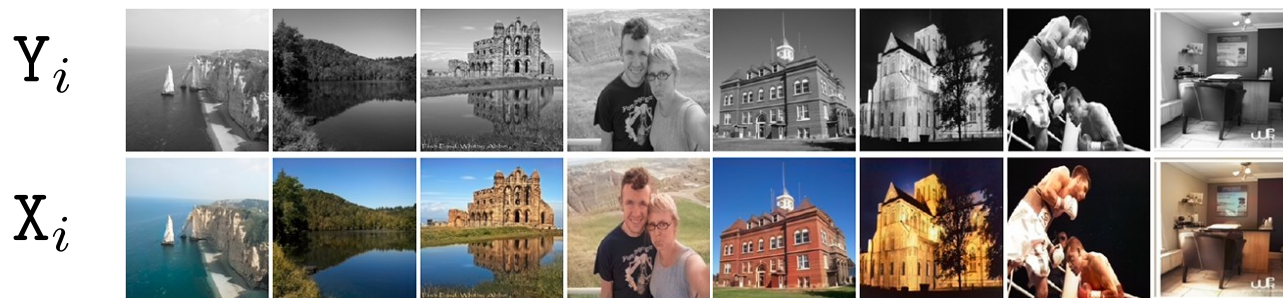
Base de données **étiquetées** : $\{Y_i, X_i\}_{i=1\dots N}$



Exemple : Colorisation d'image

Principe d'un modèle génératif conditionnel

Base de données **étiquetées** : $\{Y_i, X_i\}_{i=1\dots N}$



Exemple : Colorisation d'image

Point de vue probabiliste : $X_i \sim p_{\text{data}}(X|Y_i)$ **Inconnue**

Échantillon de la distribution a posteriori

I)

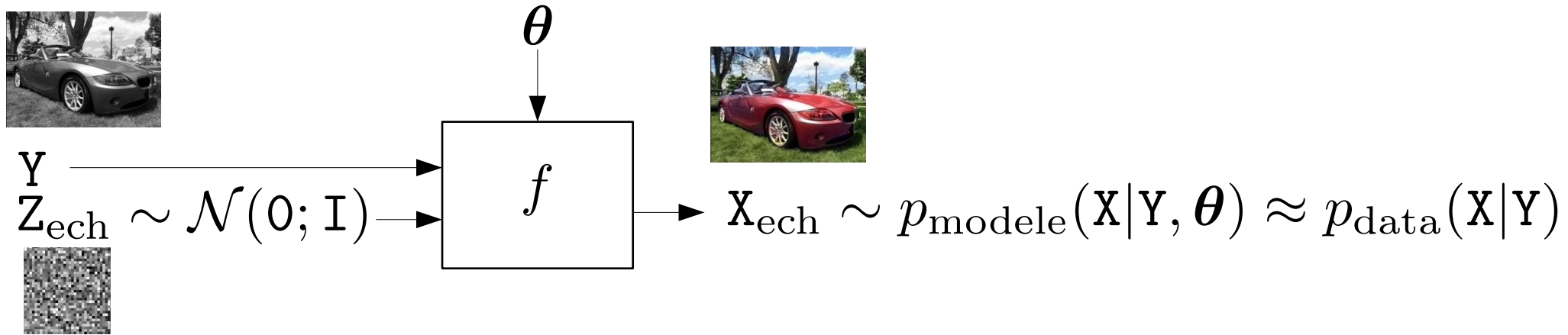
Principe d'un modèle génératif conditionnel (suite)

Objectif : Optimiser les paramètres d'un réseau de neurones afin de générer, à partir d'un Y , de **nouveaux échantillons** de $p_{\text{data}}(X|Y)$

1)

Principe d'un modèle génératif conditionnel (suite)

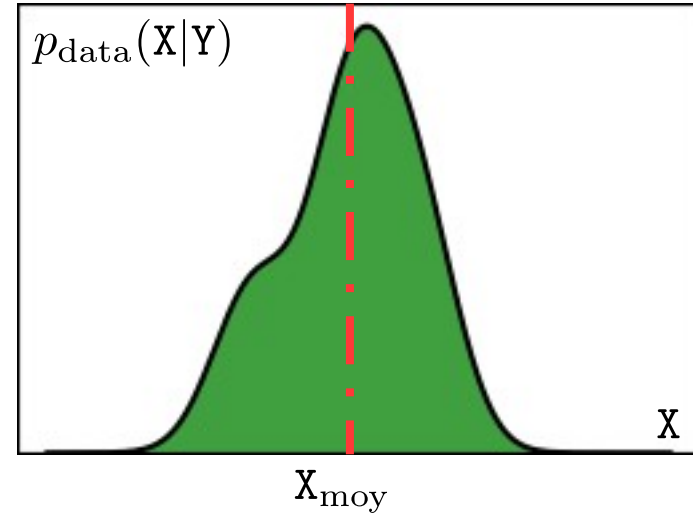
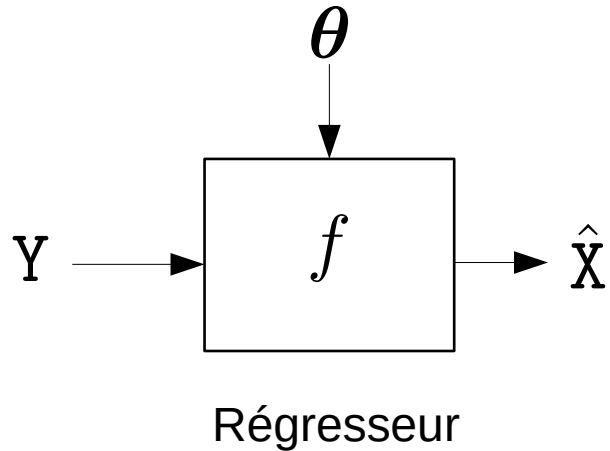
Objectif : Optimiser les paramètres d'un réseau de neurones afin de générer, à partir d'un Y , de **nouveaux échantillons** de $p_{\text{data}}(X|Y)$



1)

Différence vis-à-vis de l'apprentissage supervisé

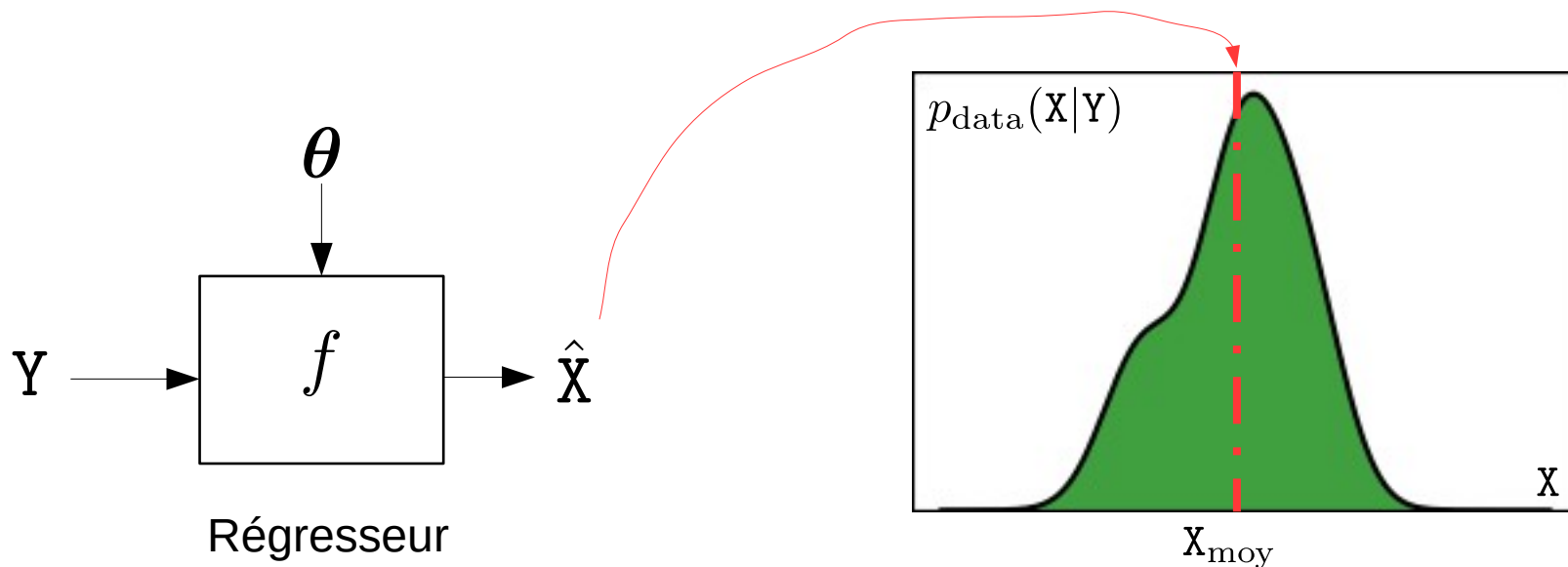
Rappel apprentissage supervisé pour de la régression



1)

Différence vis-à-vis de l'apprentissage supervisé

Rappel apprentissage supervisé pour de la régression

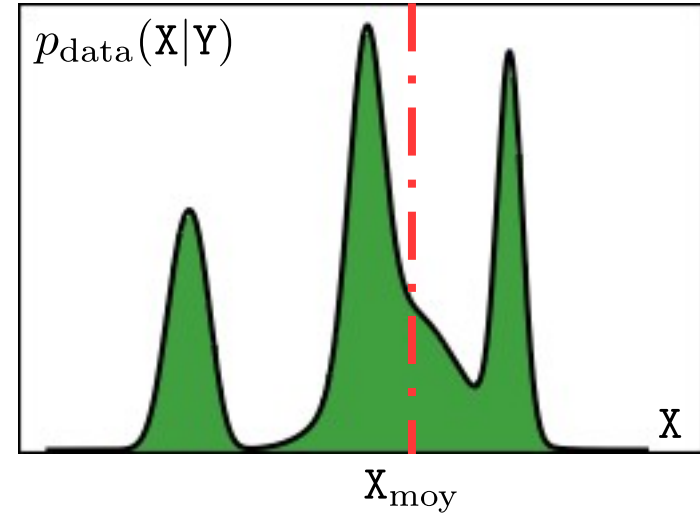
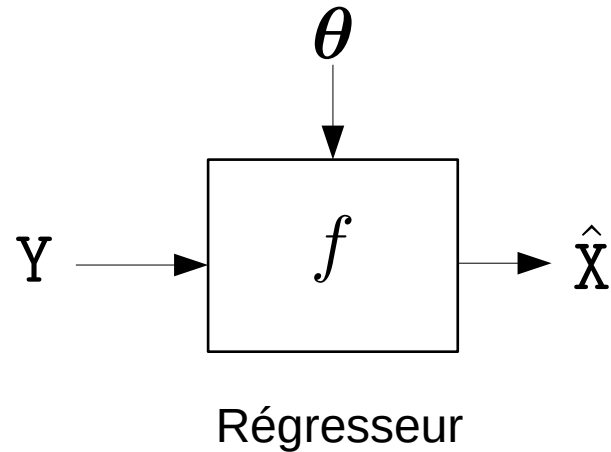


Ici, une prédiction telle que $\hat{X} \approx X_{\text{moy}}$ est satisfaisante car $p_{\text{data}}(X|Y)$ est **unimodale**.

1)

Différence vis-à-vis de l'apprentissage supervisé

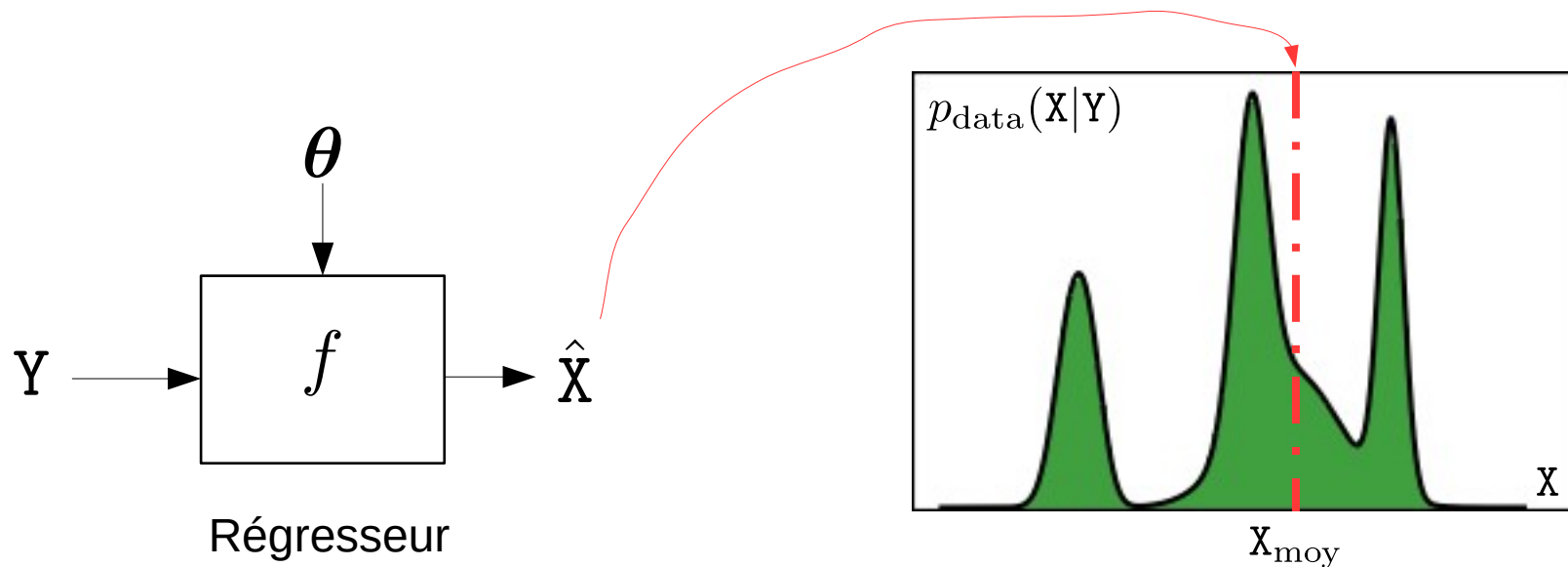
Rappel apprentissage supervisé pour de la régression



1)

Différence vis-à-vis de l'apprentissage supervisé

Rappel apprentissage supervisé pour de la régression



Ici, une prédiction telle que $\hat{X} \approx X_{\text{moy}}$ n'est pas satisfaisante car $p_{\text{data}}(X|Y)$ est **multi-modale**.

I)

Différence vis-à-vis de l'apprentissage supervisé

Question : Que se passe-t-il si on entraîne un régresseur pour colorer l'image suivante?



I)

Différence vis-à-vis de l'apprentissage supervisé

Question : Que se passe-t-il si on entraîne un régresseur pour colorer l'image suivante?



Nécessité d'apprendre à **échantillonner** $p_{\text{data}}(\mathbf{X}|\mathbf{Y})$

1)

Différence vis-à-vis de l'apprentissage supervisé

Question : Que se passe-t-il si on entraîne un régresseur pour prédire un "5" ?

"5"



1)

Différence vis-à-vis de l'apprentissage supervisé

Question : Que se passe-t-il si on entraîne un régresseur pour prédire un "5" ?

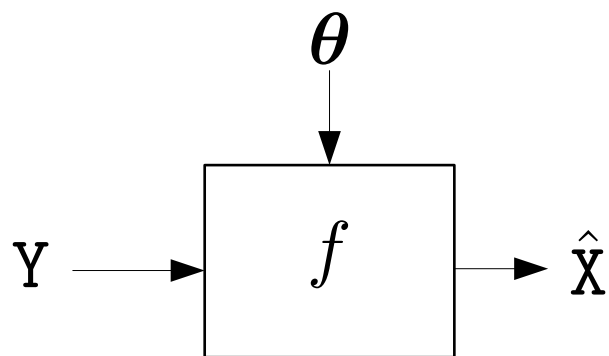
"5"



Nécessité d'apprendre à **échantillonner**

1)

Différence vis-à-vis de l'apprentissage supervisé

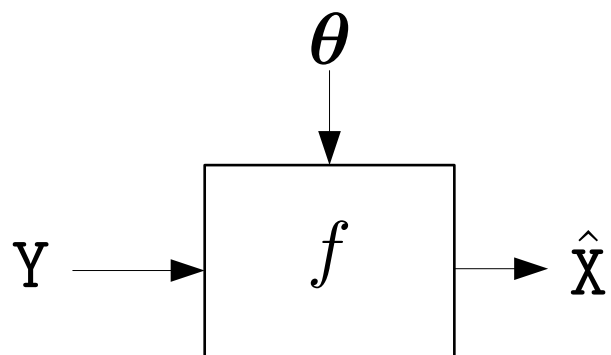


Supervisé → Régresseur

$p_{\text{data}}(\mathbf{X}|\mathbf{Y})$ de forme “simple”
(ex : unimodale)

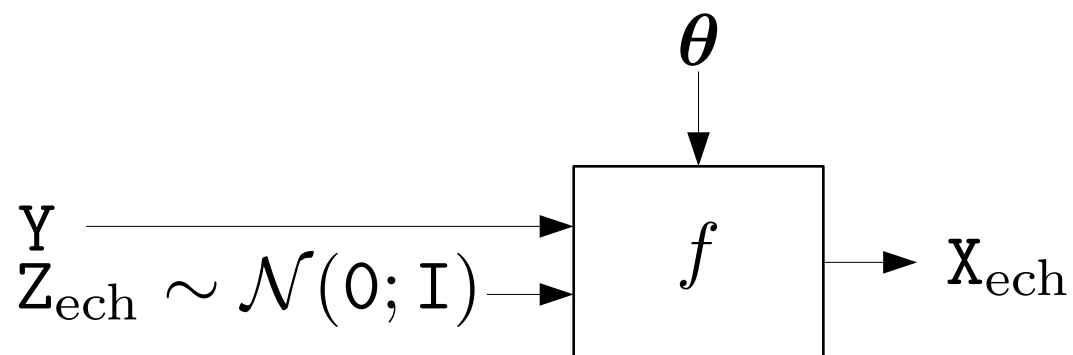
1)

Différence vis-à-vis de l'apprentissage supervisé



Supervisé → Régresseur

$p_{\text{data}}(\mathbf{X}|\mathbf{Y})$ de forme “simple”
(ex : unimodale)

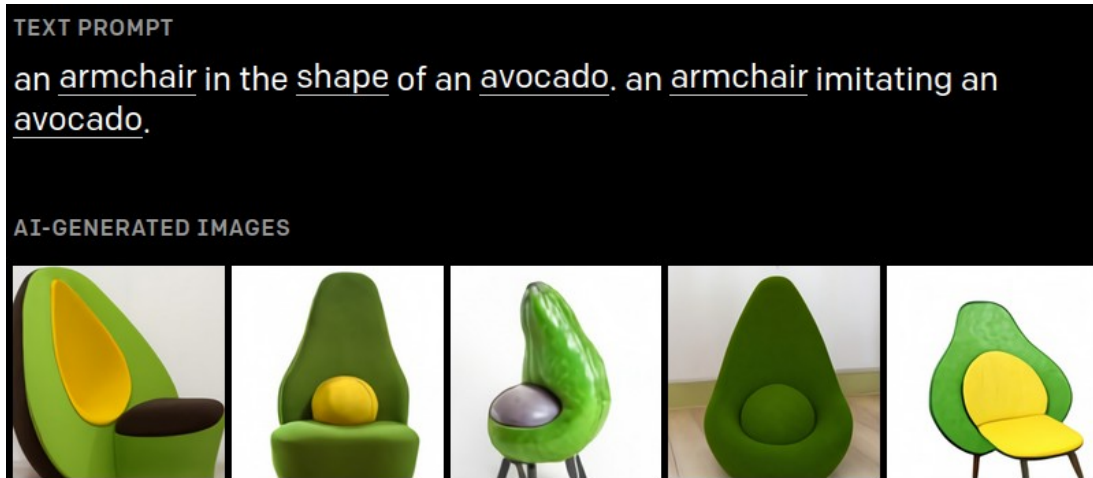


Modèle génératif → Échantillonneur

$p_{\text{data}}(\mathbf{X}|\mathbf{Y})$ de forme “complexe”

Exemple d'utilisation

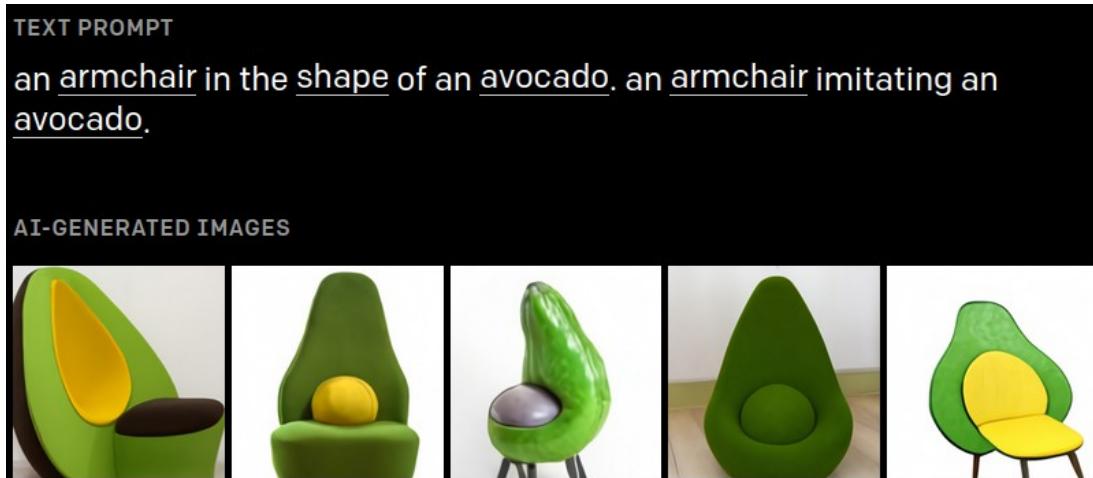
Édition/Synthèse de données (ex : retouche d'image)



<https://openai.com/blog/dall-e/>

Exemple d'utilisation

Édition/Synthèse de données (ex : retouche d'image)



<https://openai.com/blog/dall-e/>

Génération de faux contenus :

- photo
- vidéo
- signal sonore

I)

Comment entraîner un modèle génératif profond ?

Plusieurs solutions possibles, nous allons en étudier quatre :

Comment entraîner un modèle génératif profond ?

Plusieurs solutions possibles, nous allons en étudier quatre :

- le réseau inversible (NF)

Comment entraîner un modèle génératif profond ?

Plusieurs solutions possibles, nous allons en étudier quatre :

- le réseau inversible (NF)
- l'auto-encodeur variationnel (VAE)

Comment entraîner un modèle génératif profond ?

Plusieurs solutions possibles, nous allons en étudier quatre :

- le réseau inversible (NF)
- l'auto-encodeur variationnel (VAE)
- le réseau débruiteur – méthode de diffusion (DDPM)

Comment entraîner un modèle génératif profond ?

Plusieurs solutions possibles, nous allons en étudier quatre :

- le réseau inversible (NF)
- l'auto-encodeur variationnel (VAE)
- le réseau débruiteur – méthode de diffusion (DDPM)
- le réseau antagoniste (GAN)

Comment entraîner un modèle génératif profond ?

Plusieurs solutions possibles, nous allons en étudier quatre :

- le réseau inversible (NF)
- l'auto-encodeur variationnel (VAE)
- le réseau débruiteur – méthode de diffusion (DDPM)
- le réseau antagoniste (GAN)

Nous n'aurons pas le temps d'étudier les réseaux auto-régressifs.

II) Notion de divergence

II)

Notion de divergence

Nécessité de pouvoir comparer deux densités de probabilité, par exemple pour savoir si (pour une valeur de θ) $p_{\text{modele}}(\mathbf{X}|\theta)$ “ressemble” à $p_{\text{data}}(\mathbf{X})$

Notion de divergence

Nécessité de pouvoir comparer deux densités de probabilité, par exemple pour savoir si (pour une valeur de θ) $p_{\text{modele}}(\mathbf{X}|\theta)$ “ressemble” à $p_{\text{data}}(\mathbf{X})$

Densités de probabilité

$$D(p||q) \geq 0 \quad \leftarrow \text{scalaire}$$

Divergence

$$D(p||q) = 0 \text{ ssi } p = q$$

Notion de divergence (suite)

La divergence de Kullback-Leibler est souvent utilisée car elle est “pratique”.

$$KL(p(x)||q(x)) = \int p(x) \ln \frac{p(x)}{q(x)} dx$$

II)

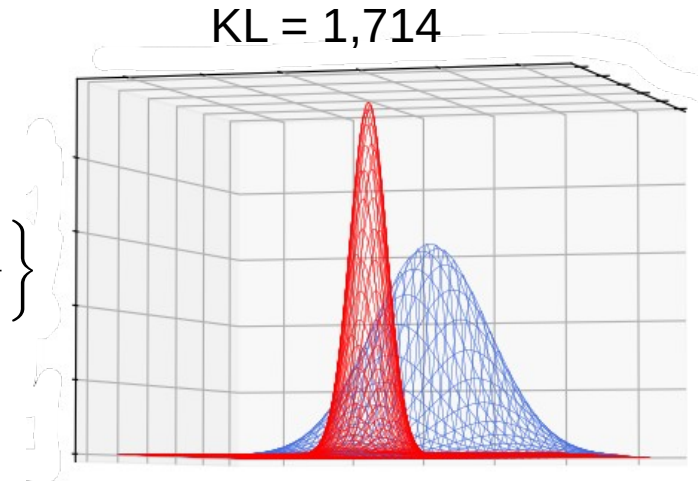
Notion de divergence (suite)

La divergence de Kullback-Leibler est souvent utilisée car elle est “pratique”.

$$KL(p(x)||q(x)) = \int p(x) \ln \frac{p(x)}{q(x)} dx$$

Exemple : Expression analytique pour deux gaussiennes
 $\mathcal{N}_0(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ et $\mathcal{N}_1(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$

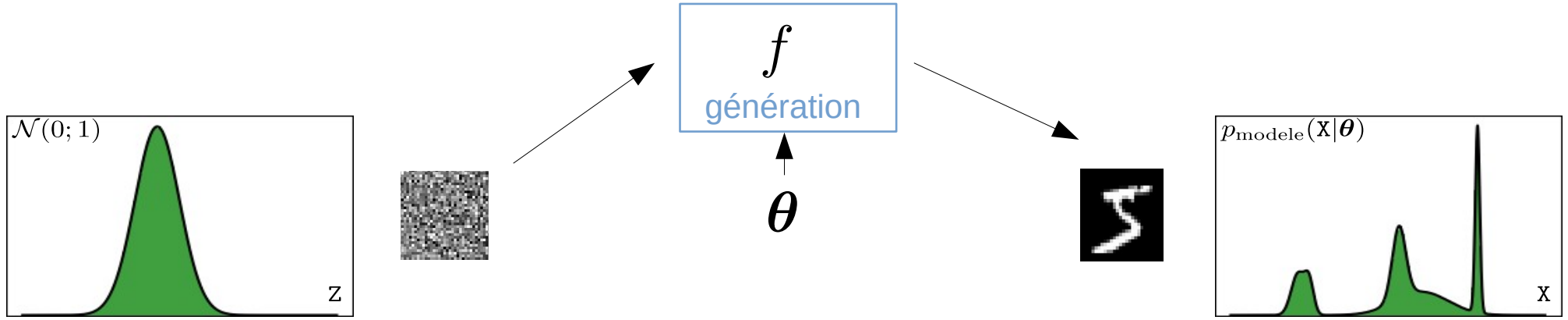
$$KL(\mathcal{N}_0 \parallel \mathcal{N}_1) = \frac{1}{2} \left\{ \text{tr}(\boldsymbol{\Sigma}_1^{-1} \boldsymbol{\Sigma}_0) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) - k + \ln \frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_0|} \right\}$$



III) Réseau inversible (NF)

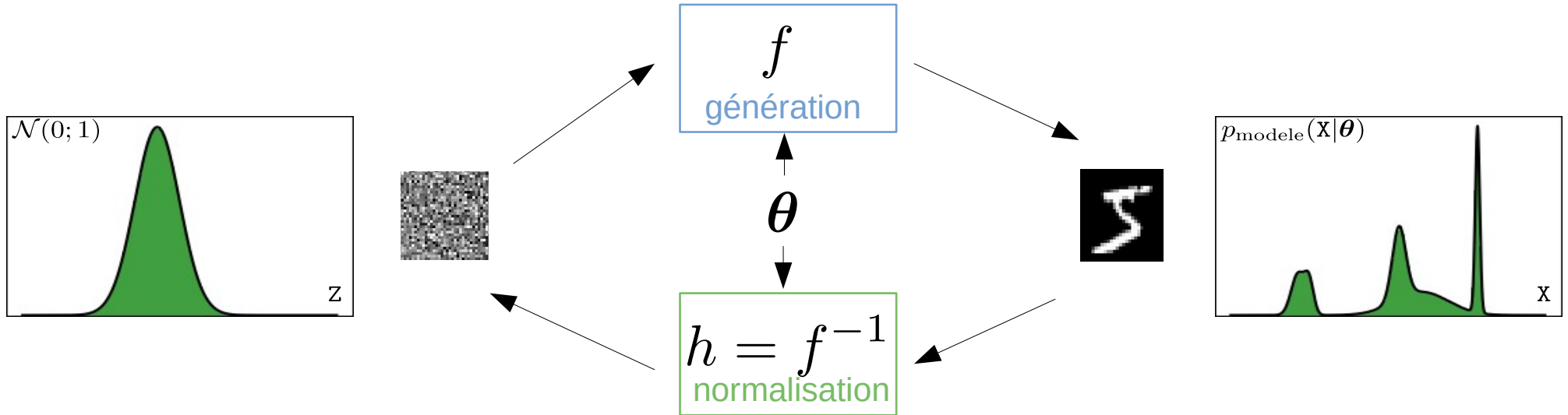
III)

Réseau inversible : idée générale

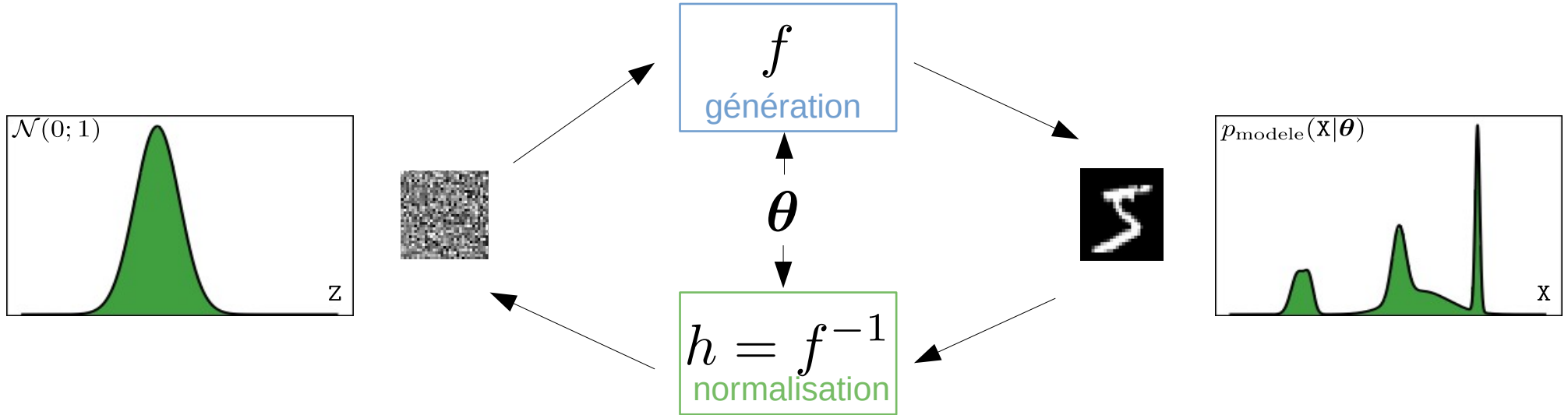


III)

Réseau inversible : idée générale



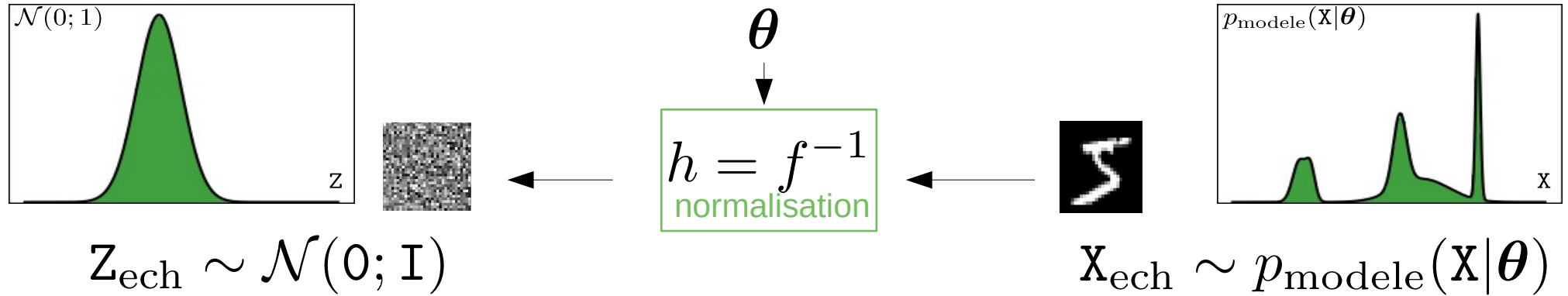
Réseau inversible : idée générale



Ne pas disposer de couples $\{Z_i, X_i\}_{i=1\dots N}$ n'est plus un problème.

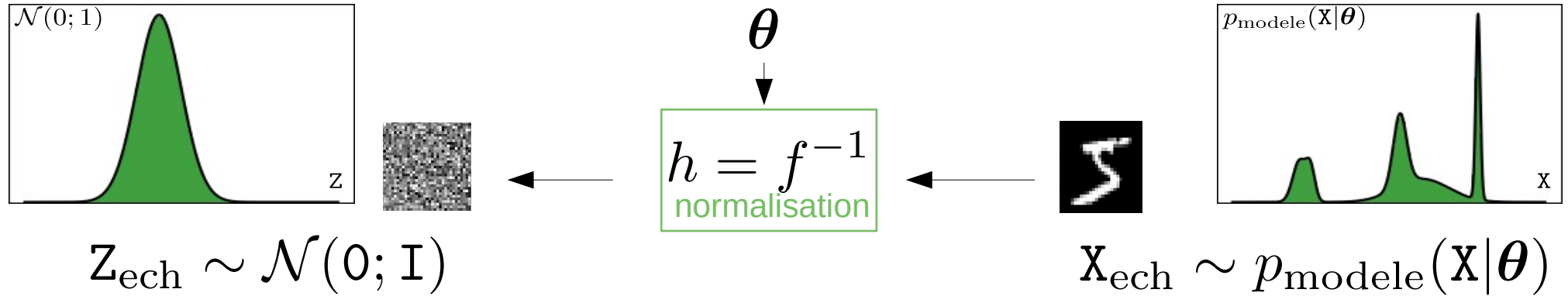
III)

Réseau inversible : formalisme



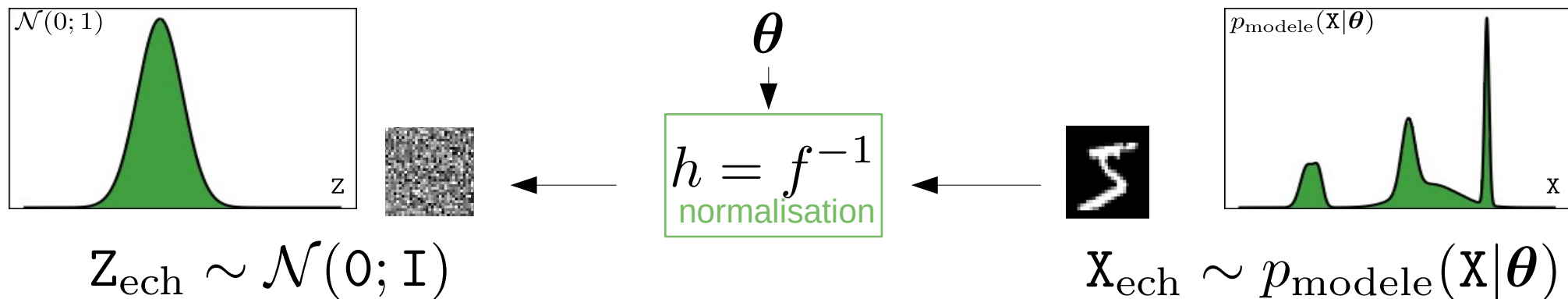
III)

Réseau inversible : formalisme



Changement de variable : $\mathbf{Z} = h(\mathbf{X}; \boldsymbol{\theta}) \longrightarrow d\mathbf{Z} = |\det J_h(\mathbf{X})| d\mathbf{X}$

Réseau inversible : formalisme

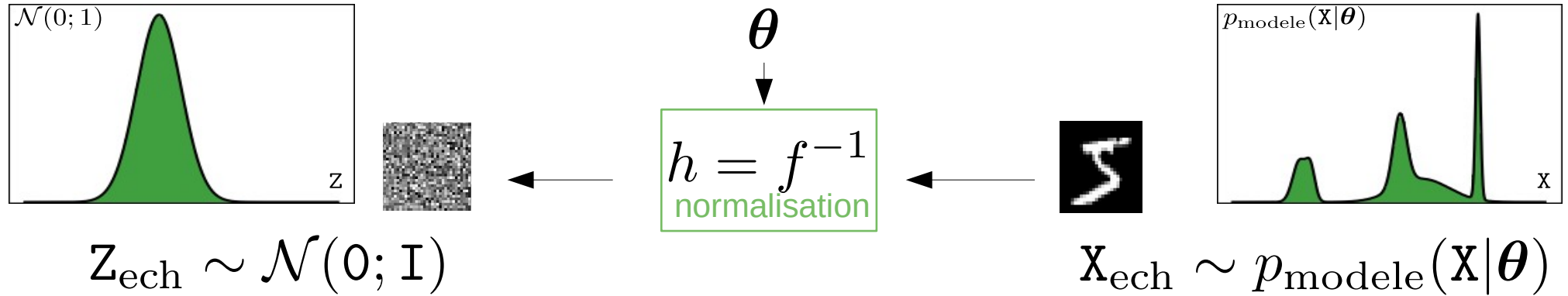


Changement de variable : $\mathbf{Z} = h(\mathbf{X}; \theta) \longrightarrow d\mathbf{Z} = |\det J_h(\mathbf{X})| d\mathbf{X}$

$$1 = \int \mathcal{N}(\mathbf{Z}; 0, \mathbf{I}) d\mathbf{Z} = \int \underbrace{\mathcal{N}(h(\mathbf{X}; \theta); 0, \mathbf{I}) |\det J_h(\mathbf{X})|}_{p_{\text{modele}}(\mathbf{X}|\theta)} d\mathbf{X}$$

III)

Réseau inversible : formalisme (suite)



$$p_{\text{modele}}(\mathbf{x}|\boldsymbol{\theta}) = \mathcal{N}(h(\mathbf{x}; \boldsymbol{\theta}); 0, \mathbf{I}) |\det J_h(\mathbf{x})|$$

$$-\ln p_{\text{modele}}(\mathbf{x}|\boldsymbol{\theta}) = ||h(\mathbf{x}; \boldsymbol{\theta})||^2 - \ln |\det J_h(\mathbf{x}; \boldsymbol{\theta})| + \text{cst}_{\boldsymbol{\theta}}$$

III)

Réseau inversible : apprentissage

Objectif : optimiser θ pour que $p_{\text{modele}}(\mathbf{X}|\theta) \approx p_{\text{data}}(\mathbf{X})$

Réseau inversible : apprentissage

Objectif : optimiser θ pour que $p_{\text{modele}}(\mathbf{X}|\theta) \approx p_{\text{data}}(\mathbf{X})$

Pour un NF on choisit généralement la divergence de Kullback-Leibler suivante :

$$\begin{aligned} KL(p_{\text{data}}(\mathbf{X})||p_{\text{modele}}(\mathbf{X}|\theta)) &= \int p_{\text{data}}(\mathbf{X}) \ln \frac{p_{\text{data}}(\mathbf{X})}{p_{\text{modele}}(\mathbf{X}|\theta)} d\mathbf{X} + \text{cst}_{\theta} \\ &= - \int p_{\text{data}}(\mathbf{X}) \ln p_{\text{modele}}(\mathbf{X}|\theta) d\mathbf{X} + \text{cst}_{\theta} \\ &= \mathbb{E}_{p_{\text{data}}(\mathbf{x})}(-\ln p_{\text{modele}}(\mathbf{X}|\theta)) + \text{cst}_{\theta} \end{aligned}$$

Réseau inversible : apprentissage

Objectif : optimiser θ pour que $p_{\text{modele}}(\mathbf{X}|\theta) \approx p_{\text{data}}(\mathbf{X})$

Pour un NF on choisit généralement la divergence de Kullback-Leibler suivante :

$$\begin{aligned} KL(p_{\text{data}}(\mathbf{X})||p_{\text{modele}}(\mathbf{X}|\theta)) &= \int p_{\text{data}}(\mathbf{X}) \ln \frac{p_{\text{data}}(\mathbf{X})}{p_{\text{modele}}(\mathbf{X}|\theta)} d\mathbf{X} + \text{cst}_{\theta} \\ &= - \int p_{\text{data}}(\mathbf{X}) \ln p_{\text{modele}}(\mathbf{X}|\theta) d\mathbf{X} + \text{cst}_{\theta} \\ &= \mathbb{E}_{p_{\text{data}}(\mathbf{x})}(-\ln p_{\text{modele}}(\mathbf{X}|\theta)) + \text{cst}_{\theta} \end{aligned}$$

Voir “Flow-GAN” pour un exemple
d’utilisation d’un apprentissage
antagoniste pour un NF

Réseau inversible : apprentissage

Objectif : optimiser θ pour que $p_{\text{modele}}(\mathbf{X}|\theta) \approx p_{\text{data}}(\mathbf{X})$

Pour un NF on choisit généralement la divergence de Kullback-Leibler suivante :

$$\begin{aligned} KL(p_{\text{data}}(\mathbf{X})||p_{\text{modele}}(\mathbf{X}|\theta)) &= \int p_{\text{data}}(\mathbf{X}) \ln \frac{p_{\text{data}}(\mathbf{X})}{p_{\text{modele}}(\mathbf{X}|\theta)} d\mathbf{X} + \text{cst}_{\theta} \\ &= - \int p_{\text{data}}(\mathbf{X}) \ln p_{\text{modele}}(\mathbf{X}|\theta) d\mathbf{X} + \text{cst}_{\theta} \\ &= \mathbb{E}_{p_{\text{data}}(\mathbf{x})} (-\ln p_{\text{modele}}(\mathbf{X}|\theta)) + \text{cst}_{\theta} \end{aligned}$$

Voir “Flow-GAN” pour un exemple d’utilisation d’un apprentissage antagoniste pour un NF

Approximation de $\mathbb{E}_{p_{\text{data}}(\mathbf{x})}$

$$L(\theta) = \frac{1}{N} \sum_{i=1}^N -\ln p_{\text{modele}}(\mathbf{x}_i|\theta) = \text{maximisation de la vraisemblance}$$

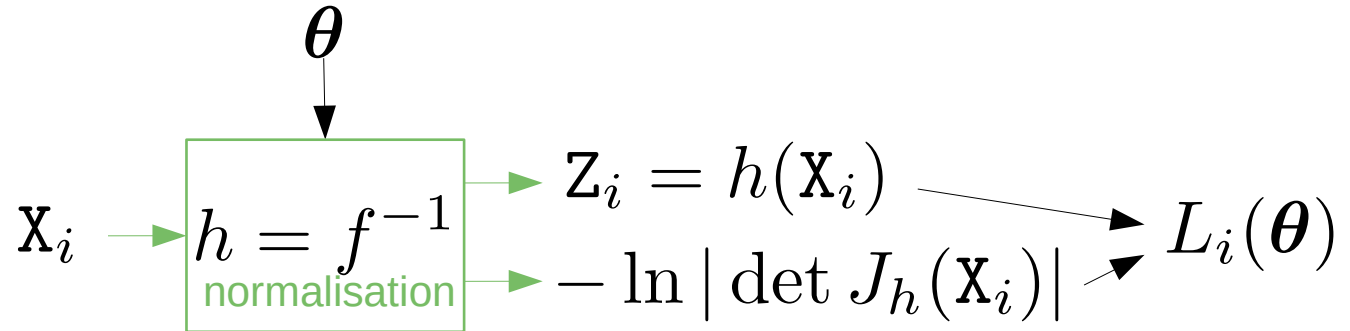
Réseau inversible : apprentissage (suite)

$$L(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^N ||h(\mathbf{x}_i; \boldsymbol{\theta})||^2 - \ln | \det J_h(\mathbf{x}_i; \boldsymbol{\theta}) |$$

Essaye de transformer \mathbf{x}_i
proche du tenseur "zéro"

Empêche que tout le monde se
transforme en un tenseur "zéro"

Descente de gradient



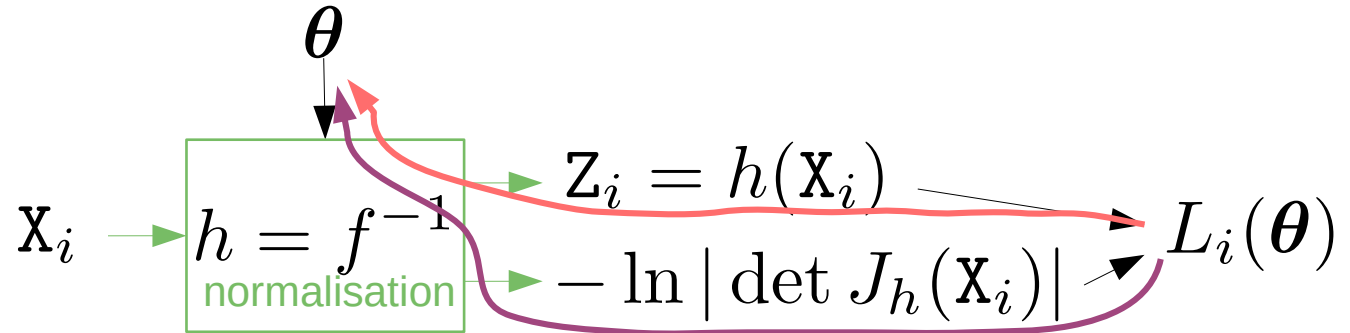
Réseau inversible : apprentissage (suite)

$$L(\theta) = \frac{1}{N} \sum_{i=1}^N ||h(\mathbf{x}_i; \theta)||^2 - \ln | \det J_h(\mathbf{x}_i; \theta) |$$

Essaye de transformer \mathbf{x}_i
proche du tenseur "zéro"

Empêche que tout le monde se
transforme en un tenseur "zéro"

Descente de gradient



En pratique : arrêt prématuré ("early stopping") sur un ensemble de validation.

Réseau inversible : architecture

Besoin d'une architecture très spécifique, chaque couche doit :

1. avoir la même taille en entrée et en sortie
2. être inversible, et pour pouvoir échantillonner efficacement cette fonction inverse doit se calculer efficacement
3. avoir comme propriété que le “logdet” de sa jacobienne se calcule efficacement et soit différentiable.

Exemple de couche : “Affine coupling layer”

$$\begin{array}{l}
 X_1 = Z_1 \\
 X_2 = \exp(s_{\theta}(Z_1)) \odot Z_2 + m_{\theta}(Z_1)
 \end{array}$$

ResNet

jacobienne triangulaire !

Réseau inversible : avantages et inconvénients

- + capable d'échantillonner efficacement
- + capable de calculer la (log-)probabilité d'une donnée
- + apprentissage possible par maximum de vraisemblance

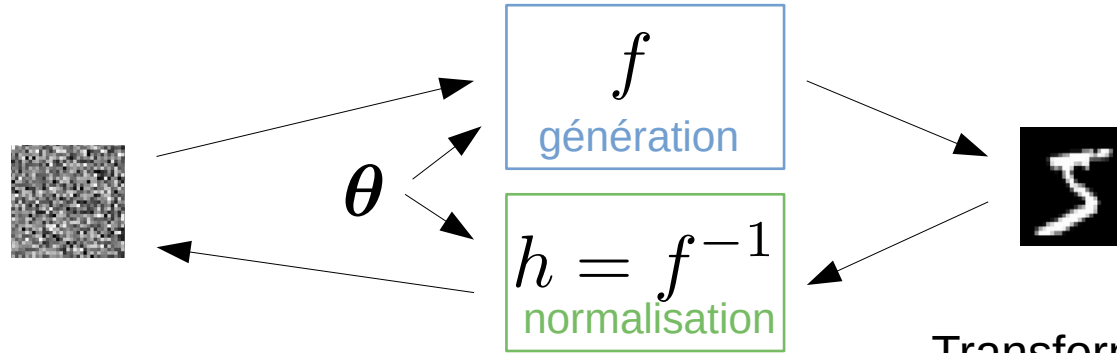
Réseau inversible : avantages et inconvénients

- + capable d'échantillonner efficacement
- + capable de calculer la (log-)probabilité d'une donnée
- + apprentissage possible par maximum de vraisemblance
- contrainte d'architecture inversible
- contrainte du même nombre de dimensions entrée/sortie
- la forme de la distribution qu'on transforme est fortement contrainte
- contrainte sur le calcul du log-déterminant

IV) Auto-encodeur variationnel (VAE)

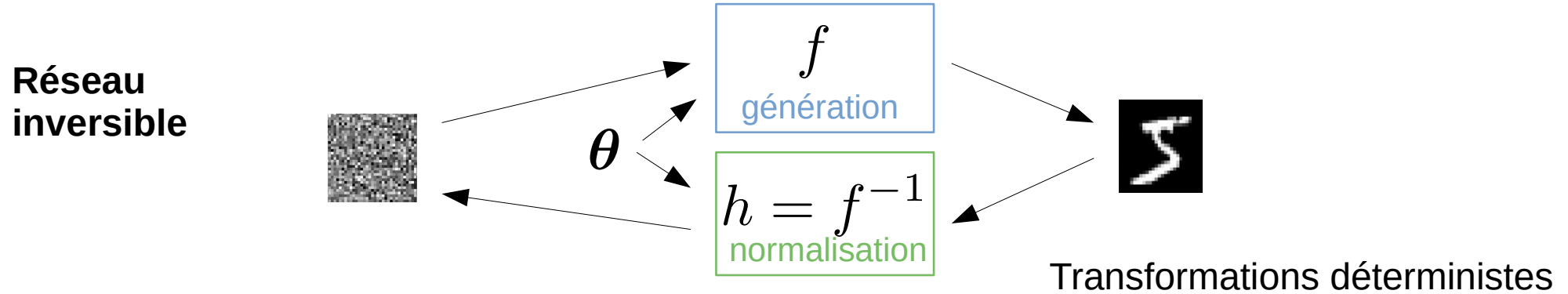
Auto-encodeur variationnel : idée générale

Réseau
inversible

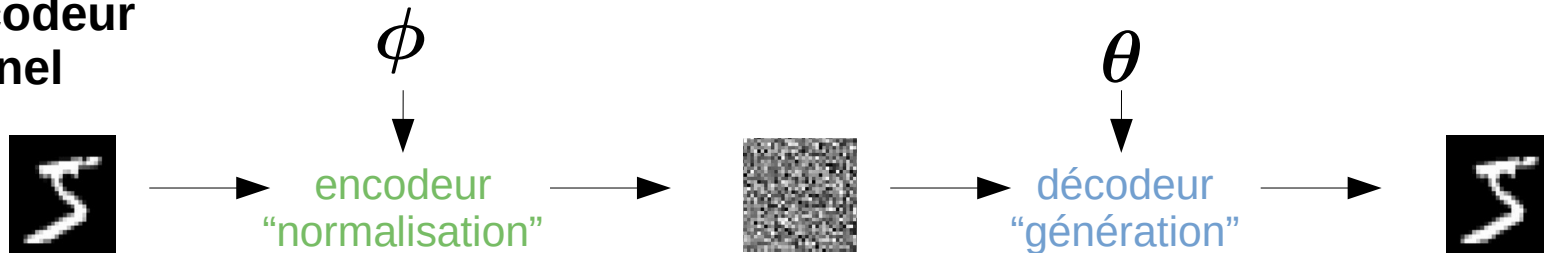


Transformations déterministes

Auto-encodeur variationnel : idée générale

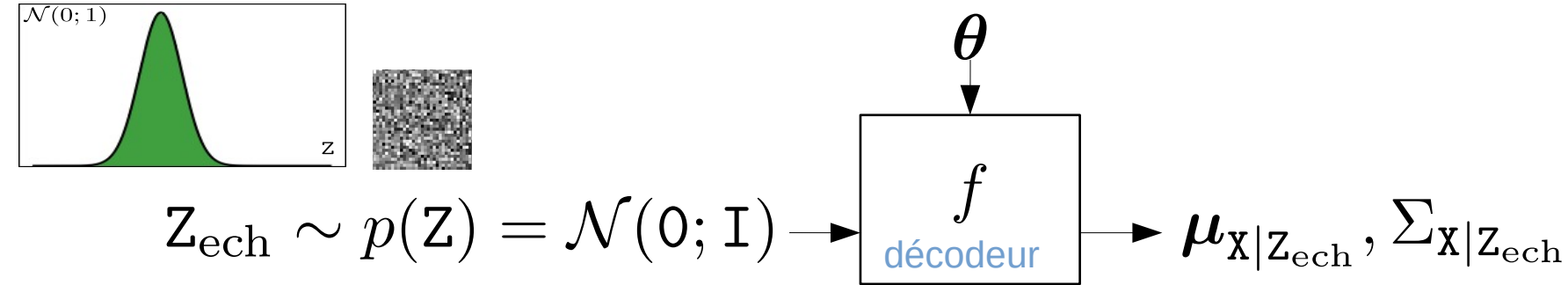


Auto-encodeur variationnel

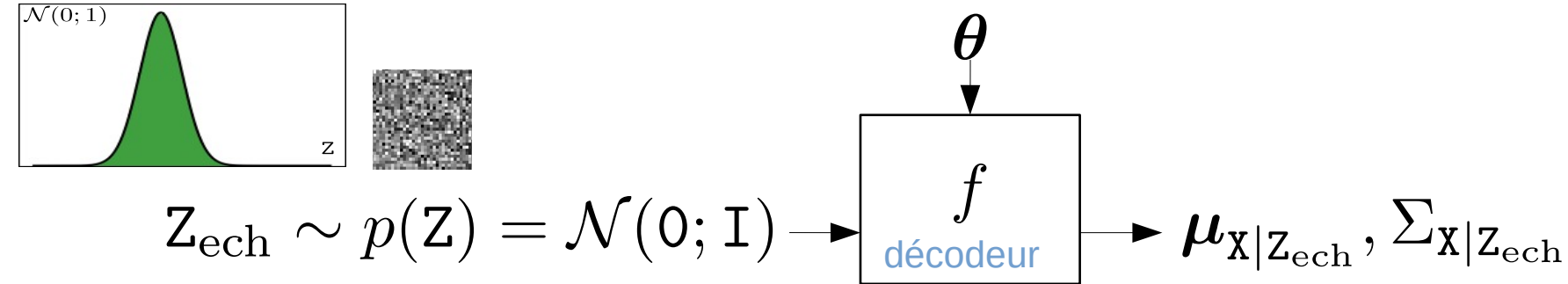


IV)

Auto-encodeur variationnel : formalisme

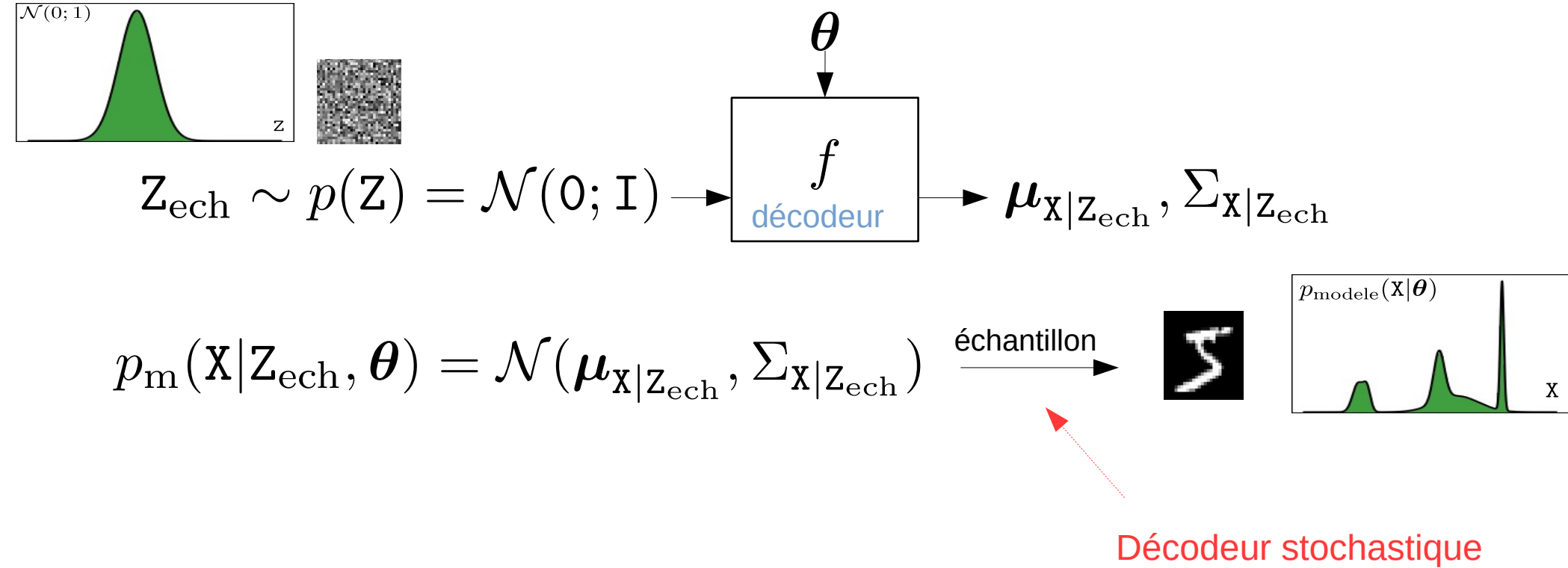


Auto-encodeur variationnel : formalisme

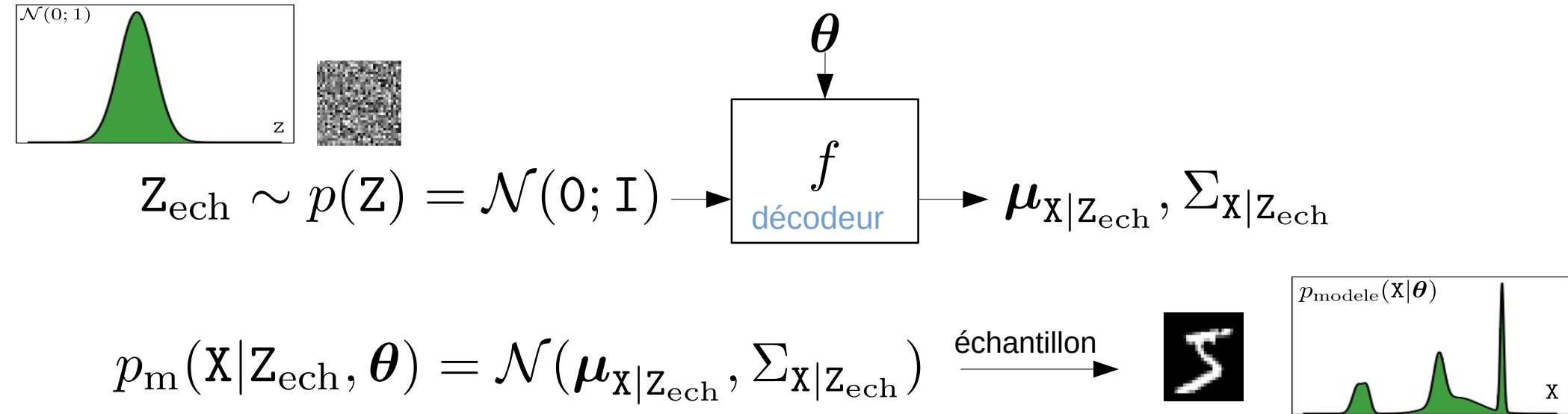


$$p_m(X|Z_{ech}, \theta) = \mathcal{N}(\mu_{X|Z_{ech}}, \Sigma_{X|Z_{ech}})$$

Auto-encodeur variationnel : formalisme

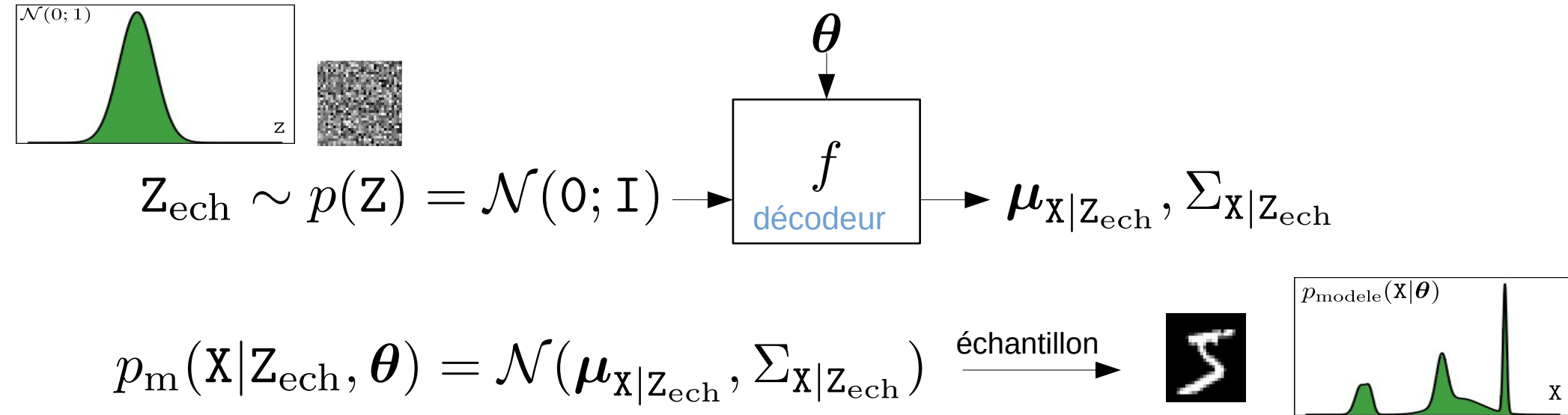


Auto-encodeur variationnel : formalisme



$$p_{\text{m}}(\mathbf{x}|\theta) = \int p_{\text{m}}(\mathbf{x}, \mathbf{z}|\theta) d\mathbf{z} = \int p(\mathbf{z}) p_{\text{m}}(\mathbf{x}|\mathbf{z}, \theta) d\mathbf{z}$$

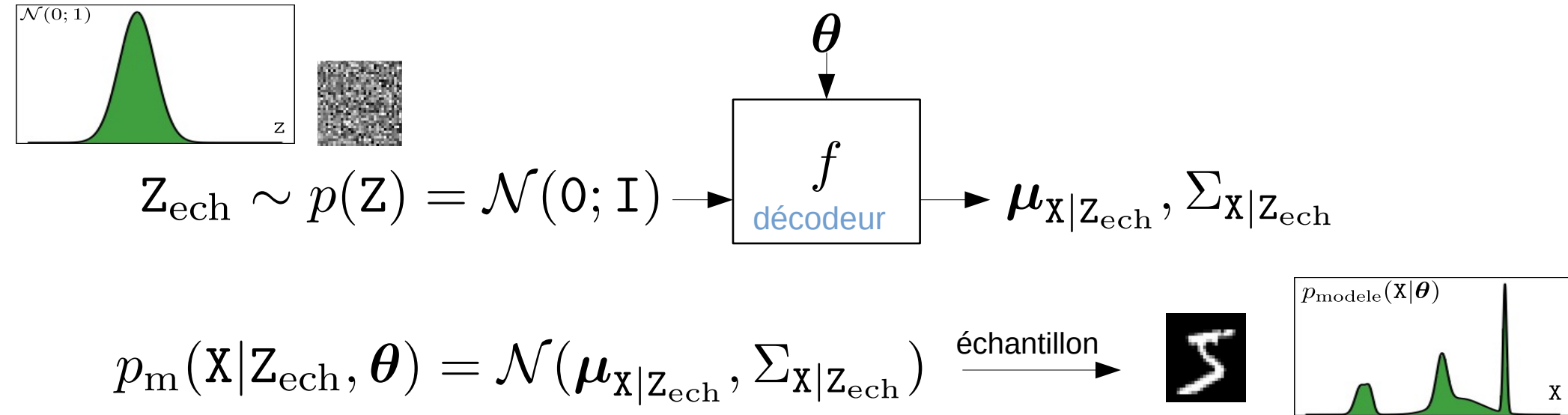
Auto-encodeur variationnel : formalisme



$$p_m(\mathbf{x}|\theta) = \int p_m(\mathbf{x}, z|\theta) dz = \int p(z) p_m(\mathbf{x}|z, \theta) dz$$

Intégrale en grande dimension, trop coûteux à calculer !

Auto-encodeur variationnel : formalisme

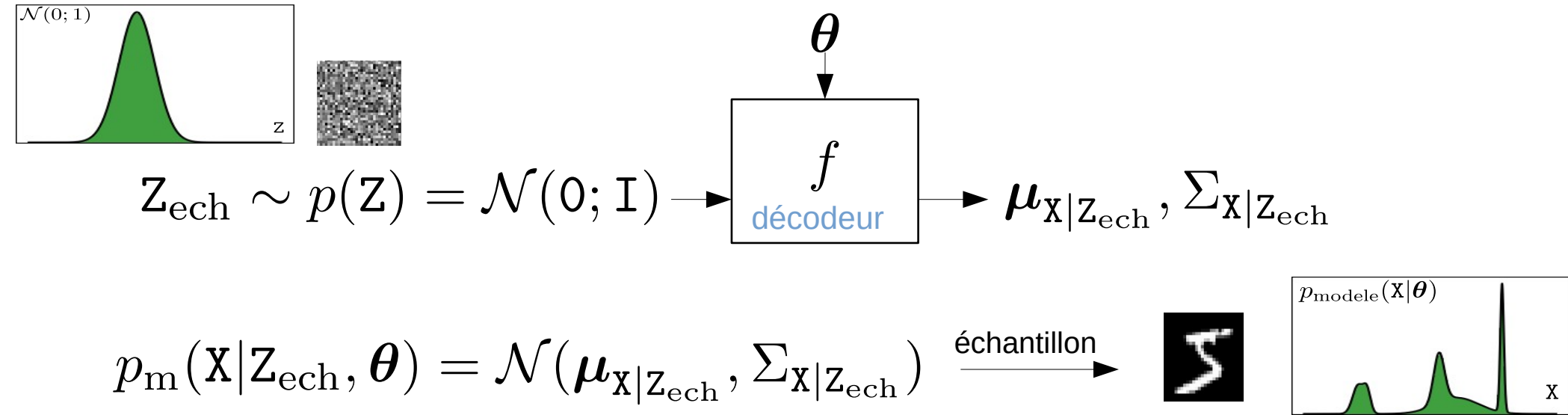


$$p_{\text{m}}(\mathbf{X}|\theta) = \int p_{\text{m}}(\mathbf{X}, \mathbf{Z}|\theta) d\mathbf{Z} = \int p(\mathbf{Z}) p_{\text{m}}(\mathbf{X}|\mathbf{Z}, \theta) d\mathbf{Z}$$

Intégrale en grande dimension, trop coûteux à calculer !

→ On ne peut pas calculer $p_{\text{m}}(\mathbf{X}|\theta)$ (on sait juste l'échantillonner)

Auto-encodeur variationnel : formalisme



$$p_m(\mathbf{X}|\theta) = \int p_m(\mathbf{X}, \mathbf{Z}|\theta) d\mathbf{Z} = \int p(\mathbf{Z}) p_m(\mathbf{X}|\mathbf{Z}, \theta) d\mathbf{Z}$$

Intégrale en grande dimension, trop coûteux à calculer !

- On ne peut pas calculer $p_m(\mathbf{X}|\theta)$ (on sait juste l'échantillonner)
- Mais on sait calculer $p_m(\mathbf{X}, \mathbf{Z}|\theta)$

Auto-encodeur variationnel : apprentissage

Objectif : optimiser θ pour que $p_{\text{modele}}(\mathbf{X}|\theta) \approx p_{\text{data}}(\mathbf{X})$

Auto-encodeur variationnel : apprentissage

Objectif : optimiser θ pour que $p_{\text{modele}}(\mathbf{X}|\theta) \approx p_{\text{data}}(\mathbf{X})$

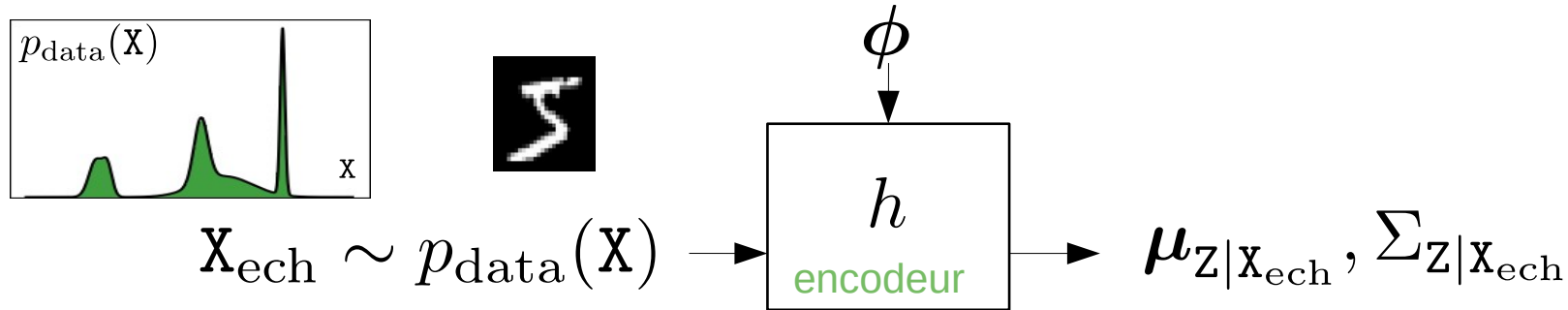
Problème : on ne peut pas calculer $p_{\text{modele}}(\mathbf{X}|\theta) \rightarrow \cancel{KL(p_{\text{data}}(\mathbf{X}) || p_{\text{modele}}(\mathbf{X}|\theta))}$

Auto-encodeur variationnel : apprentissage

Objectif : optimiser θ pour que $p_{\text{modele}}(\mathbf{X}|\theta) \approx p_{\text{data}}(\mathbf{X})$

Problème : on ne peut pas calculer $p_{\text{modele}}(\mathbf{X}|\theta) \rightarrow \cancel{KL(p_{\text{data}}(\mathbf{X}) || p_{\text{modele}}(\mathbf{X}|\theta))}$

Pour contourner ce problème, on introduit un “encodeur”

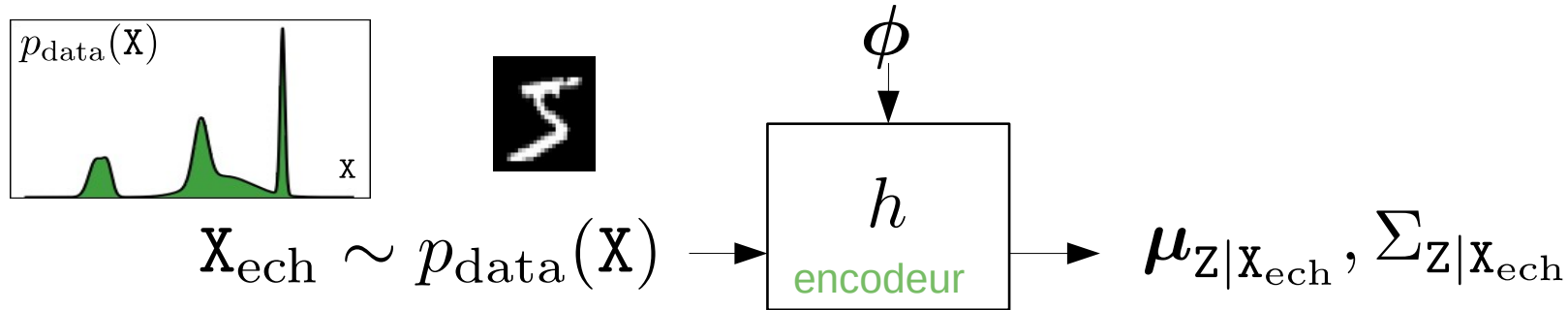


Auto-encodeur variationnel : apprentissage

Objectif : optimiser θ pour que $p_{\text{modele}}(\mathbf{X}|\theta) \approx p_{\text{data}}(\mathbf{X})$

Problème : on ne peut pas calculer $p_{\text{modele}}(\mathbf{X}|\theta) \rightarrow \cancel{KL(p_{\text{data}}(\mathbf{X}) || p_{\text{modele}}(\mathbf{X}|\theta))}$

Pour contourner ce problème, on introduit un “encodeur”



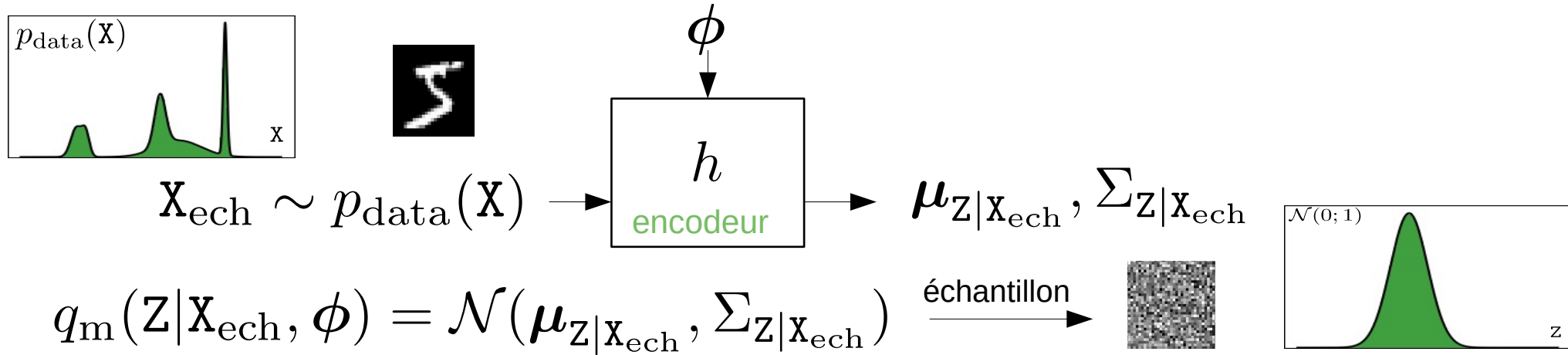
$$q_{\text{m}}(\mathbf{Z}|\mathbf{X}_{\text{ech}}, \phi) = \mathcal{N}(\mu_{\mathbf{Z}|\mathbf{X}_{\text{ech}}}, \Sigma_{\mathbf{Z}|\mathbf{X}_{\text{ech}}})$$

Auto-encodeur variationnel : apprentissage

Objectif : optimiser θ pour que $p_{\text{modele}}(\mathbf{X}|\theta) \approx p_{\text{data}}(\mathbf{X})$

Problème : on ne peut pas calculer $p_{\text{modele}}(\mathbf{X}|\theta) \rightarrow \cancel{KL(p_{\text{data}}(\mathbf{X}) || p_{\text{modele}}(\mathbf{X}|\theta))}$

Pour contourner ce problème, on introduit un “encodeur”



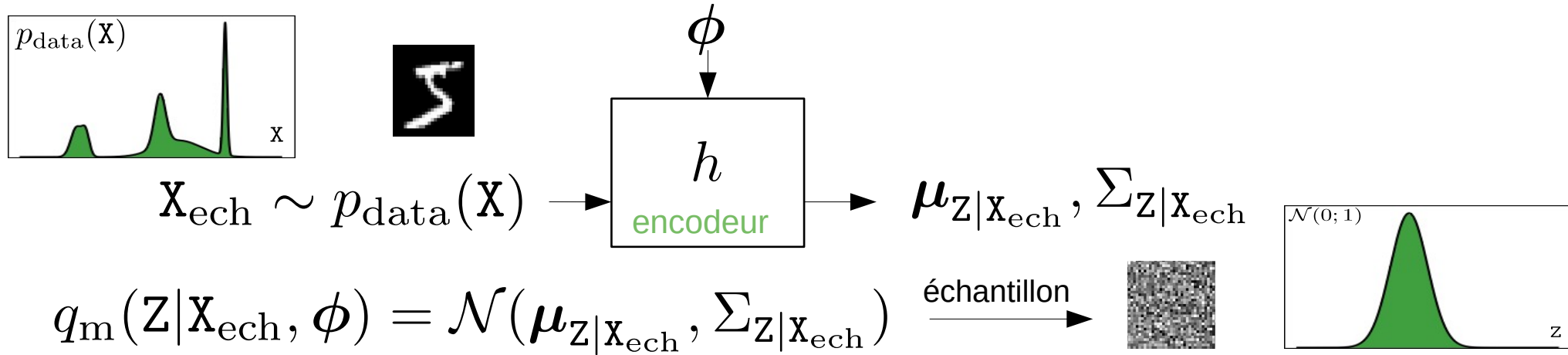
IV)

Auto-encodeur variationnel : apprentissage

Objectif : optimiser θ pour que $p_{\text{modele}}(\mathbf{X}|\theta) \approx p_{\text{data}}(\mathbf{X})$

Problème : on ne peut pas calculer $p_{\text{modele}}(\mathbf{X}|\theta) \rightarrow \cancel{KL(p_{\text{data}}(\mathbf{X}) || p_{\text{modele}}(\mathbf{X}|\theta))}$

Pour contourner ce problème, on introduit un “encodeur”



Nouvel objectif (dégradé par rapport au 1er) : optimiser θ et ϕ pour que

$$p_{\text{m}}(\mathbf{X}, \mathbf{Z}|\theta) = p(\mathbf{Z})p_{\text{m}}(\mathbf{X}|\mathbf{Z}, \theta) \approx q_{\text{m}}(\mathbf{X}, \mathbf{Z}|\phi) = p_{\text{data}}(\mathbf{X})q_{\text{m}}(\mathbf{Z}|\mathbf{X}, \phi)$$

Auto-encodeur variationnel : apprentissage (suite)

Nouvel objectif : optimiser θ et ϕ pour que

$$p_m(\mathbf{X}, \mathbf{Z}|\theta) = p(\mathbf{Z})p_m(\mathbf{X}|\mathbf{Z}, \theta) \approx q_m(\mathbf{X}, \mathbf{Z}|\phi) = p_{\text{data}}(\mathbf{X})q_m(\mathbf{Z}|\mathbf{X}, \phi)$$

Auto-encodeur variationnel : apprentissage (suite)

Nouvel objectif : optimiser θ et ϕ pour que

$$p_m(\mathbf{X}, \mathbf{Z}|\theta) = p(\mathbf{Z})p_m(\mathbf{X}|\mathbf{Z}, \theta) \approx q_m(\mathbf{X}, \mathbf{Z}|\phi) = p_{\text{data}}(\mathbf{X})q_m(\mathbf{Z}|\mathbf{X}, \phi)$$

Pour un VAE on choisit la divergence de Kullback-Leibler suivante :

$$KL(q_m(\mathbf{X}, \mathbf{Z}|\phi) || p_m(\mathbf{X}, \mathbf{Z}|\theta))$$

Auto-encodeur variationnel : apprentissage (suite)

Nouvel objectif : optimiser θ et ϕ pour que

$$p_m(\mathbf{X}, \mathbf{Z}|\theta) = p(\mathbf{Z})p_m(\mathbf{X}|\mathbf{Z}, \theta) \approx q_m(\mathbf{X}, \mathbf{Z}|\phi) = p_{\text{data}}(\mathbf{X})q_m(\mathbf{Z}|\mathbf{X}, \phi)$$

Pour un VAE on choisit la divergence de Kullback-Leibler suivante :

$$\begin{aligned} & KL(q_m(\mathbf{X}, \mathbf{Z}|\phi) || p_m(\mathbf{X}, \mathbf{Z}|\theta)) \\ = & \mathbb{E}_{p_{\text{data}}(\mathbf{x})} \mathbb{E}_{\mathcal{N}(\boldsymbol{\mu}_{\mathbf{Z}|\mathbf{x}}, \Sigma_{\mathbf{Z}|\mathbf{x}})} \left(-\ln \mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}|\mathbf{z}}, \Sigma_{\mathbf{x}|\mathbf{z}}) \right) \quad \text{“Erreur de reconstruction” :} \\ & \quad \text{Incite la sortie du d codeur  } \\ & \quad \text{ressembler aux X} \\ + & \mathbb{E}_{p_{\text{data}}(\mathbf{x})} (KL(\mathcal{N}(\boldsymbol{\mu}_{\mathbf{Z}|\mathbf{x}}, \Sigma_{\mathbf{Z}|\mathbf{x}}) || \mathcal{N}(0, \mathbf{I}))) \quad \text{“R gularisation” :} \\ & \quad \text{Incite la sortie de l’encodeur a  tre} \\ & \quad \text{centr e r duite   logique car} \\ & \quad \text{lorsqu’on g n rera des z, on les} \\ & \quad \text{tirera selon une gaussienne} \quad 90 \\ & \quad \text{centr e r duite} \end{aligned}$$

Auto-encodeur variationnel : apprentissage (suite)

Nouvel objectif : optimiser θ et ϕ pour que

$$p_m(\mathbf{X}, \mathbf{Z}|\theta) = p(\mathbf{Z})p_m(\mathbf{X}|\mathbf{Z}, \theta) \approx q_m(\mathbf{X}, \mathbf{Z}|\phi) = p_{\text{data}}(\mathbf{X})q_m(\mathbf{Z}|\mathbf{X}, \phi)$$

Pour un VAE on choisit la divergence de Kullback-Leibler suivante :

$$\begin{aligned} & KL(q_m(\mathbf{X}, \mathbf{Z}|\phi) || p_m(\mathbf{X}, \mathbf{Z}|\theta)) \\ = & \mathbb{E}_{p_{\text{data}}(\mathbf{X})} \mathbb{E}_{\mathcal{N}(\mu_{\mathbf{Z}|\mathbf{X}}, \Sigma_{\mathbf{Z}|\mathbf{X}})} \left(-\ln \mathcal{N}(\mu_{\mathbf{X}|\mathbf{Z}}, \Sigma_{\mathbf{X}|\mathbf{Z}}) \right) \quad \text{“Erreur de reconstruction” :} \\ & \quad \text{Incite la sortie du décodeur à} \\ & \quad \text{ressembler aux X} \\ + & \mathbb{E}_{p_{\text{data}}(\mathbf{X})} (KL(\mathcal{N}(\mu_{\mathbf{Z}|\mathbf{X}}, \Sigma_{\mathbf{Z}|\mathbf{X}}) || \mathcal{N}(0, \mathbf{I}))) \quad \text{“Régularisation” :} \\ & \quad \text{Incite la sortie de l’encodeur à être} \\ & \quad \text{centrée réduite} \rightarrow \text{logique car} \\ & \quad \text{lorsqu’on générera des z, on les} \\ & \quad \text{tirera selon une gaussienne} \\ & \quad \text{centrée réduite} \end{aligned}$$

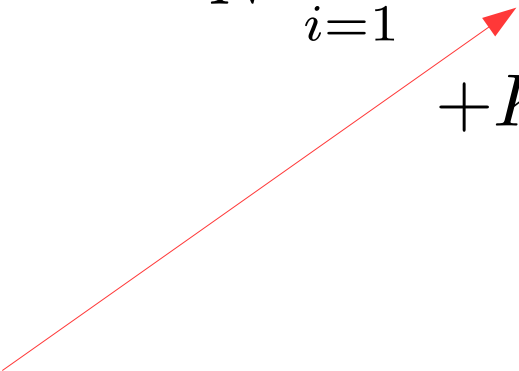
Auto-encodeur variationnel : apprentissage (suite)

Approximation de $\mathbb{E}_{p_{\text{data}}(\mathbf{x})}$

$$L(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{\mathcal{N}(\boldsymbol{\mu}_{\mathbf{z}|\mathbf{x}_i}, \Sigma_{\mathbf{z}|\mathbf{x}_i})} \left(-\ln \mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}|\mathbf{z}}, \Sigma_{\mathbf{x}|\mathbf{z}}) \right) \\ + KL(\mathcal{N}(\boldsymbol{\mu}_{\mathbf{z}|\mathbf{x}_i}, \Sigma_{\mathbf{z}|\mathbf{x}_i}) || \mathcal{N}(\mathbf{0}, \mathbf{I}))$$

Auto-encodeur variationnel : apprentissage (suite)

Approximation de $\mathbb{E}_{p_{\text{data}}(\mathbf{x})}$

$$L(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{\mathcal{N}(\boldsymbol{\mu}_{\mathbf{z}|\mathbf{x}_i}, \Sigma_{\mathbf{z}|\mathbf{x}_i})} \left(-\ln \mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}|\mathbf{z}}, \Sigma_{\mathbf{x}|\mathbf{z}}) \right) + KL(\mathcal{N}(\boldsymbol{\mu}_{\mathbf{z}|\mathbf{x}_i}, \Sigma_{\mathbf{z}|\mathbf{x}_i}) || \mathcal{N}(\mathbf{0}, \mathbf{I}))$$


Problème : ϕ intervient dans cette espérance (dans le calcul de ces moyenne et covariance)

Auto-encodeur variationnel : apprentissage (suite)

Approximation de $\mathbb{E}_{p_{\text{data}}(\mathbf{x})}$

$$L(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{\mathcal{N}(\boldsymbol{\mu}_{\mathbf{z}|\mathbf{x}_i}, \Sigma_{\mathbf{z}|\mathbf{x}_i})} \left(-\ln \mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}|\mathbf{z}}, \Sigma_{\mathbf{x}|\mathbf{z}}) \right) \\ + KL(\mathcal{N}(\boldsymbol{\mu}_{\mathbf{z}|\mathbf{x}_i}, \Sigma_{\mathbf{z}|\mathbf{x}_i}) || \mathcal{N}(\mathbf{0}, \mathbf{I}))$$

Astuce de reparamétrisation (“Reparameterization trick”)

$$= \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{\mathcal{N}(\mathbf{0}, \mathbf{I})} \left(-\ln \mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}|(\boldsymbol{\mu}_{\mathbf{z}|\mathbf{x}_i} + \Sigma_{\mathbf{z}|\mathbf{x}_i}^{1/2} \boldsymbol{\epsilon})}, \Sigma_{\mathbf{x}|(\boldsymbol{\mu}_{\mathbf{z}|\mathbf{x}_i} + \Sigma_{\mathbf{z}|\mathbf{x}_i}^{1/2} \boldsymbol{\epsilon})}) \right) \\ + KL(\mathcal{N}(\boldsymbol{\mu}_{\mathbf{z}|\mathbf{x}_i}, \Sigma_{\mathbf{z}|\mathbf{x}_i}) || \mathcal{N}(\mathbf{0}, \mathbf{I}))$$

Auto-encodeur variationnel : apprentissage (suite)

En considérant un
seul échantillon

$$\boldsymbol{\epsilon}_{\text{ech}} \sim \mathcal{N}(0; \mathbf{I})$$

$$L_i(\boldsymbol{\theta}) = KL(\mathcal{N}(\boldsymbol{\mu}_{\mathbf{Z}|\mathbf{X}_i}, \Sigma_{\mathbf{Z}|\mathbf{X}_i}) || \mathcal{N}(0, \mathbf{I})) \\ - \ln \mathcal{N}(\boldsymbol{\mu}_{\mathbf{X} | (\boldsymbol{\mu}_{\mathbf{Z}|\mathbf{X}_i} + \Sigma_{\mathbf{Z}|\mathbf{X}_i}^{1/2} \boldsymbol{\epsilon}_{\text{ech}})}, \Sigma_{\mathbf{X} | (\boldsymbol{\mu}_{\mathbf{Z}|\mathbf{X}_i} + \Sigma_{\mathbf{Z}|\mathbf{X}_i}^{1/2} \boldsymbol{\epsilon}_{\text{ech}})})$$

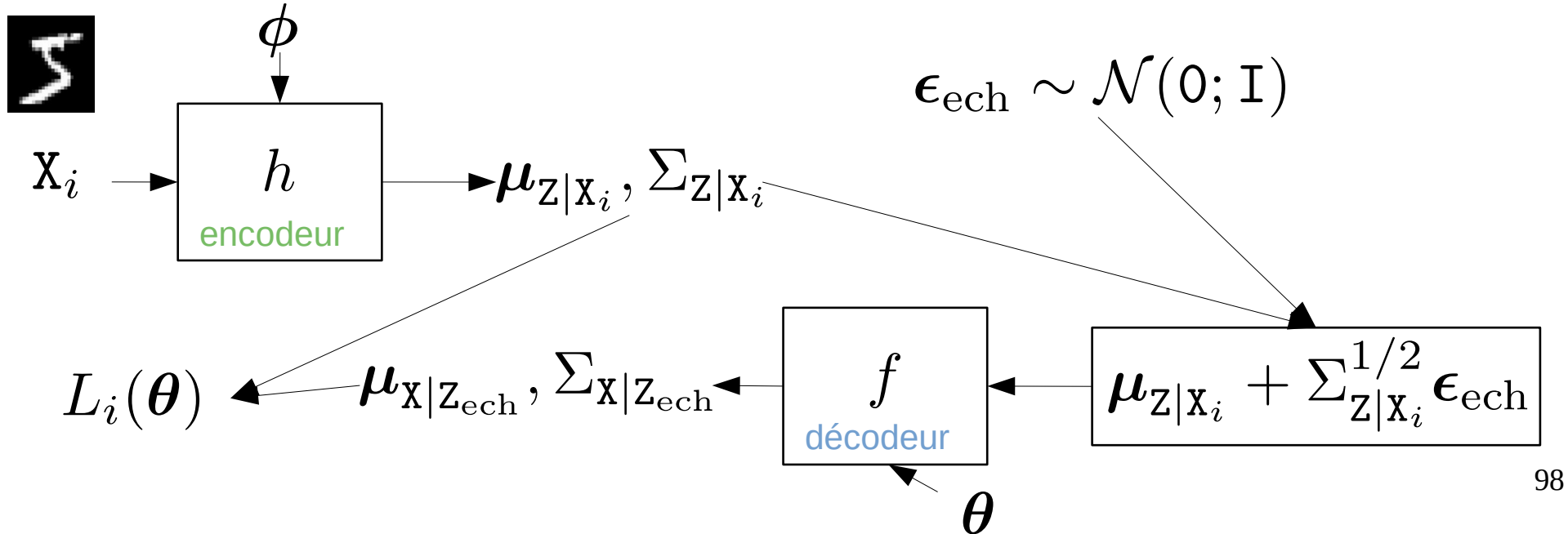
Auto-encodeur variationnel : apprentissage (suite)

En considérant un
seul échantillon

$$\epsilon_{\text{ech}} \sim \mathcal{N}(0; \mathbf{I})$$

$$L_i(\theta) = KL(\mathcal{N}(\mu_{\mathbf{Z}|\mathbf{X}_i}, \Sigma_{\mathbf{Z}|\mathbf{X}_i}) || \mathcal{N}(0, \mathbf{I}))$$

$$- \ln \mathcal{N}(\mu_{\mathbf{X} | (\mu_{\mathbf{Z}|\mathbf{X}_i} + \Sigma_{\mathbf{Z}|\mathbf{X}_i}^{1/2} \epsilon_{\text{ech}})}, \Sigma_{\mathbf{X} | (\mu_{\mathbf{Z}|\mathbf{X}_i} + \Sigma_{\mathbf{Z}|\mathbf{X}_i}^{1/2} \epsilon_{\text{ech}})})$$



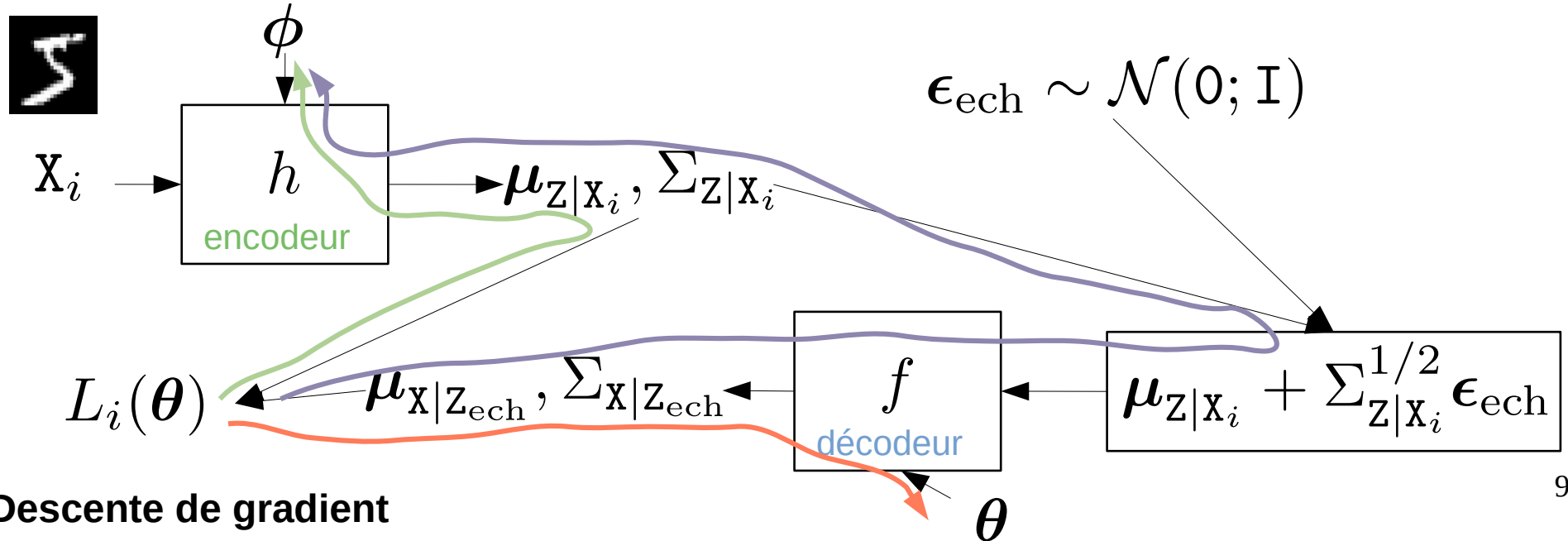
Auto-encodeur variationnel : apprentissage (suite)

En considérant un
seul échantillon

$$\epsilon_{\text{ech}} \sim \mathcal{N}(0; \mathbf{I})$$

$$L_i(\theta) = KL(\mathcal{N}(\mu_{z|x_i}, \Sigma_{z|x_i}) || \mathcal{N}(0, \mathbf{I}))$$

$$- \ln \mathcal{N}(\mu_{x|(\mu_{z|x_i} + \Sigma_{z|x_i}^{1/2} \epsilon_{\text{ech}})}, \Sigma_{x|(\mu_{z|x_i} + \Sigma_{z|x_i}^{1/2} \epsilon_{\text{ech}})})$$



Auto-encodeur variationnel : avantages et inconvénients

- + capable d'échantillonner efficacement
- + pas de contrainte particulière sur l'architecture
- + taille de Z non contrainte par celle de X

Auto-encodeur variationnel : avantages et inconvénients

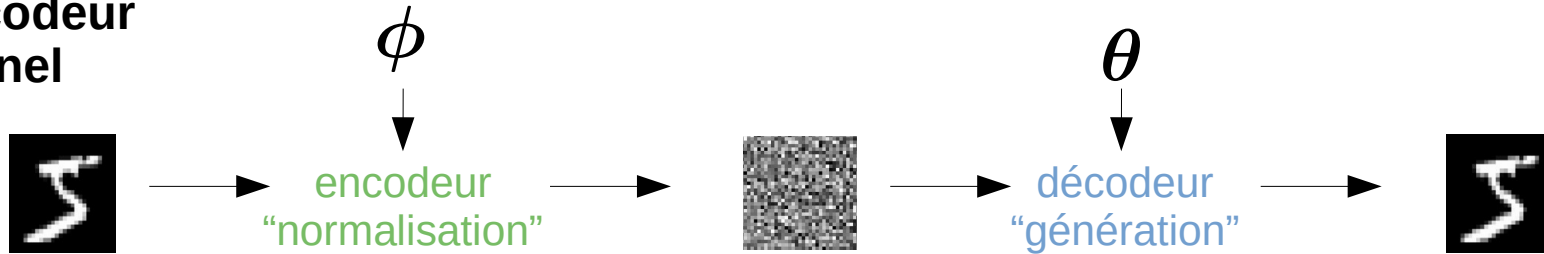
- + capable d'échantillonner efficacement
- + pas de contrainte particulière sur l'architecture
- + taille de Z non contrainte par celle de X
- pas d'expression exacte de la (log-)probabilité d'une donnée
- pas d'apprentissage par maximum de vraisemblance (mais borne inférieure quand même)
- contrainte sur la forme des distributions conditionnelles (exemple : covariance diagonale pour avoir un calcul rapide)

V) Réseau débruiteur – méthode de diffusion (DDPM)

v)

Réseau débruiteur : idée générale

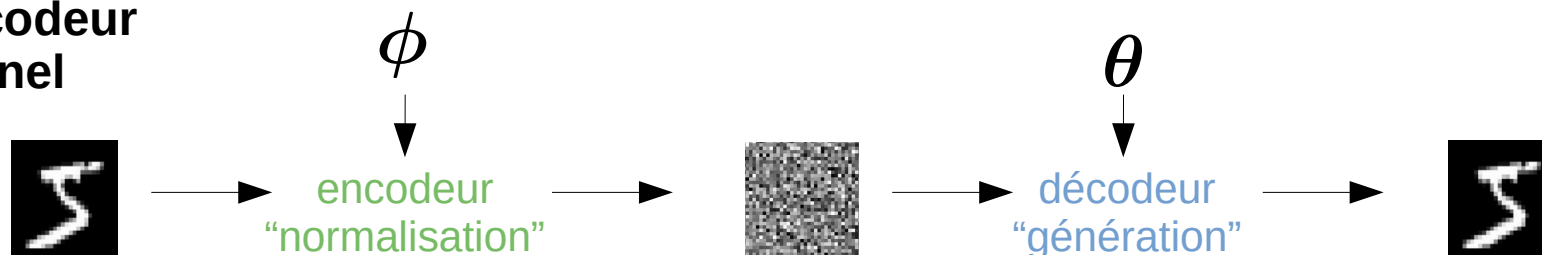
Auto-encodeur
variationnel



Transformations stochastiques

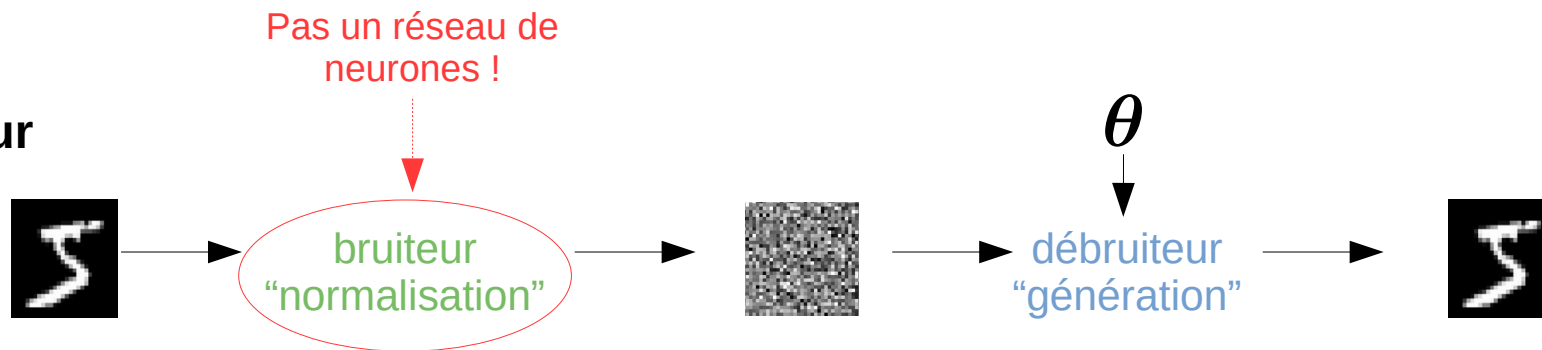
Réseau débruiteur : idée générale

Auto-encodeur
variationnel



Transformations stochastiques

Réseau
débruiteur

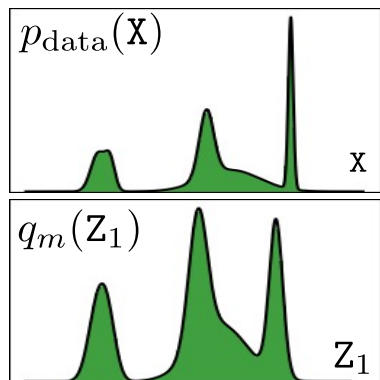


Terminologie : Modèle de diffusion ("Diffusion Models")
"Denoising Diffusion Probabilistic models"

Transformations stochastiques

v)

Bruiteur (Diffusion)



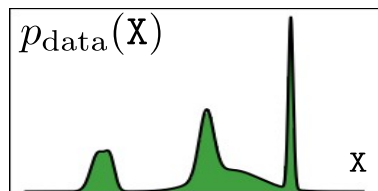
$$\mathbf{X}_{\text{ech}} \sim p_{\text{data}}(\mathbf{X})$$



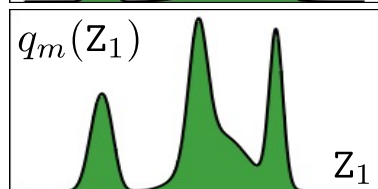
$$\mathbf{Z}_1 = \sqrt{1 - \beta_1} \mathbf{X}_{\text{ech}} + \sqrt{\beta_1} \boldsymbol{\epsilon}_1 \leftarrow \boldsymbol{\epsilon}_1 \sim \mathcal{N}(0; \mathbf{I})$$

v)

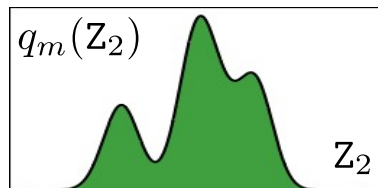
Bruiteur (Diffusion)



$$\mathbf{X}_{\text{ech}} \sim p_{\text{data}}(\mathbf{X})$$



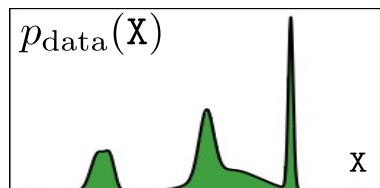
$$\mathbf{Z}_1 = \sqrt{1 - \beta_1} \mathbf{X}_{\text{ech}} + \sqrt{\beta_1} \boldsymbol{\epsilon}_1 \quad \leftarrow \quad \boldsymbol{\epsilon}_1 \sim \mathcal{N}(\mathbf{0}; \mathbf{I})$$



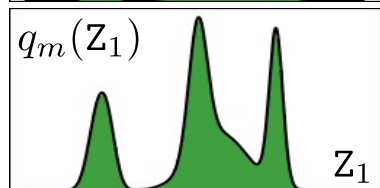
$$\mathbf{Z}_2 = \sqrt{1 - \beta_2} \mathbf{Z}_1 + \sqrt{\beta_2} \boldsymbol{\epsilon}_2 \quad \leftarrow \quad \boldsymbol{\epsilon}_2 \sim \mathcal{N}(\mathbf{0}; \mathbf{I})$$

v)

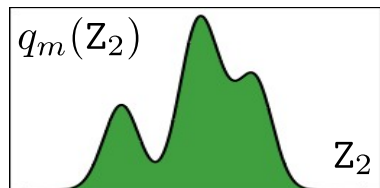
Bruiteur (Diffusion)



$$\mathbf{X}_{\text{ech}} \sim p_{\text{data}}(\mathbf{X})$$

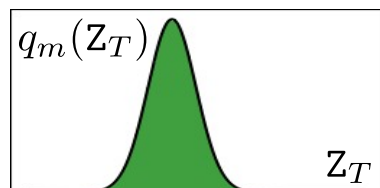


$$\mathbf{Z}_1 = \sqrt{1 - \beta_1} \mathbf{X}_{\text{ech}} + \sqrt{\beta_1} \boldsymbol{\epsilon}_1 \leftarrow \boldsymbol{\epsilon}_1 \sim \mathcal{N}(\mathbf{0}; \mathbf{I})$$



$$\mathbf{Z}_2 = \sqrt{1 - \beta_2} \mathbf{Z}_1 + \sqrt{\beta_2} \boldsymbol{\epsilon}_2 \leftarrow \boldsymbol{\epsilon}_2 \sim \mathcal{N}(\mathbf{0}; \mathbf{I})$$

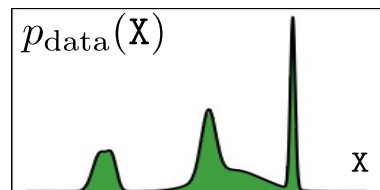
\vdots



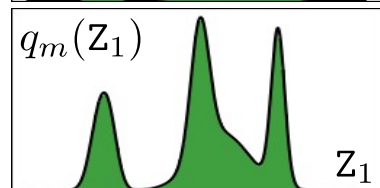
$$\mathbf{Z}_T = \sqrt{1 - \beta_T} \mathbf{Z}_{T-1} + \sqrt{\beta_T} \boldsymbol{\epsilon}_T \leftarrow \boldsymbol{\epsilon}_T \sim \mathcal{N}(\mathbf{0}; \mathbf{I})$$

v)

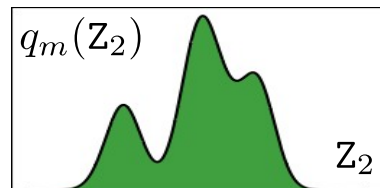
Bruiteur (Diffusion)



$$\mathbf{X}_{\text{ech}} \sim p_{\text{data}}(\mathbf{X})$$



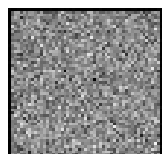
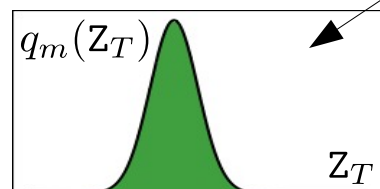
$$\mathbf{Z}_1 = \sqrt{1 - \beta_1} \mathbf{X}_{\text{ech}} + \sqrt{\beta_1} \boldsymbol{\epsilon}_1 \leftarrow \boldsymbol{\epsilon}_1 \sim \mathcal{N}(\mathbf{0}; \mathbf{I})$$



$$\mathbf{Z}_2 = \sqrt{1 - \beta_2} \mathbf{Z}_1 + \sqrt{\beta_2} \boldsymbol{\epsilon}_2 \leftarrow \boldsymbol{\epsilon}_2 \sim \mathcal{N}(\mathbf{0}; \mathbf{I})$$

\vdots

$$\approx \mathcal{N}(\mathbf{0}; \mathbf{I})$$



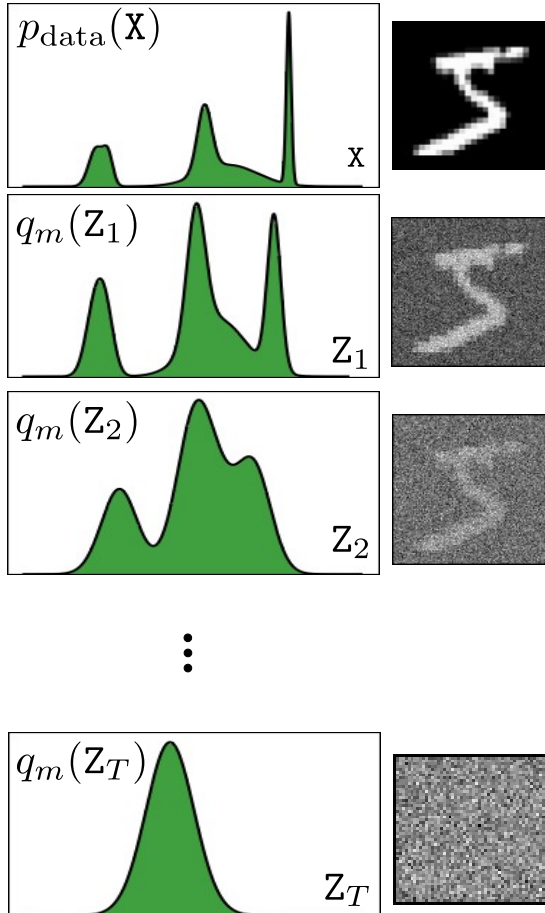
$$\mathbf{Z}_T = \sqrt{1 - \beta_T} \mathbf{Z}_{T-1} + \sqrt{\beta_T} \boldsymbol{\epsilon}_T \leftarrow \boldsymbol{\epsilon}_T \sim \mathcal{N}(\mathbf{0}; \mathbf{I})$$

v)

Bruiteur (Diffusion)

$$\mathbf{X}_{\text{ech}} \sim p_{\text{data}}(\mathbf{X})$$

$$\alpha_t = \prod_{k=1}^t (1 - \beta_k)$$



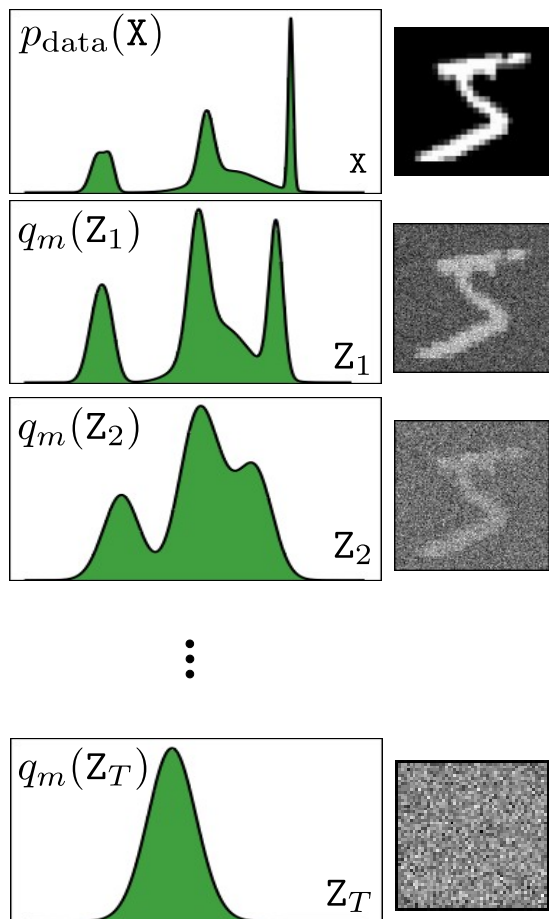
v)

Bruiteur (Diffusion)

$$\mathbf{X}_{\text{ech}} \sim p_{\text{data}}(\mathbf{X})$$

$$\alpha_t = \prod_{k=1}^t (1 - \beta_k)$$

$$\mathbf{Z}_t = \sqrt{\alpha_t} \mathbf{X}_{\text{ech}} + \sqrt{1 - \alpha_t} \epsilon_t$$



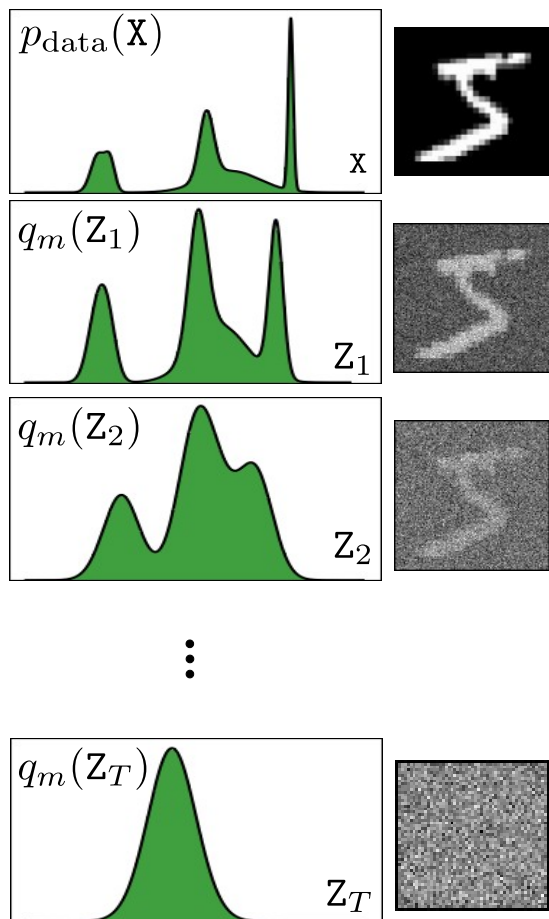
v)

Bruiteur (Diffusion)

$$\mathbf{X}_{\text{ech}} \sim p_{\text{data}}(\mathbf{X}) \quad \alpha_t = \prod_{k=1}^t (1 - \beta_k)$$

$$\mathbf{Z}_t = \sqrt{\alpha_t} \mathbf{X}_{\text{ech}} + \sqrt{1 - \alpha_t} \epsilon_t$$

$$q_m(\mathbf{Z}_t | \mathbf{X}) = \mathcal{N}(\sqrt{\alpha_t} \mathbf{X}, (1 - \alpha_t) \mathbf{I})$$



v)

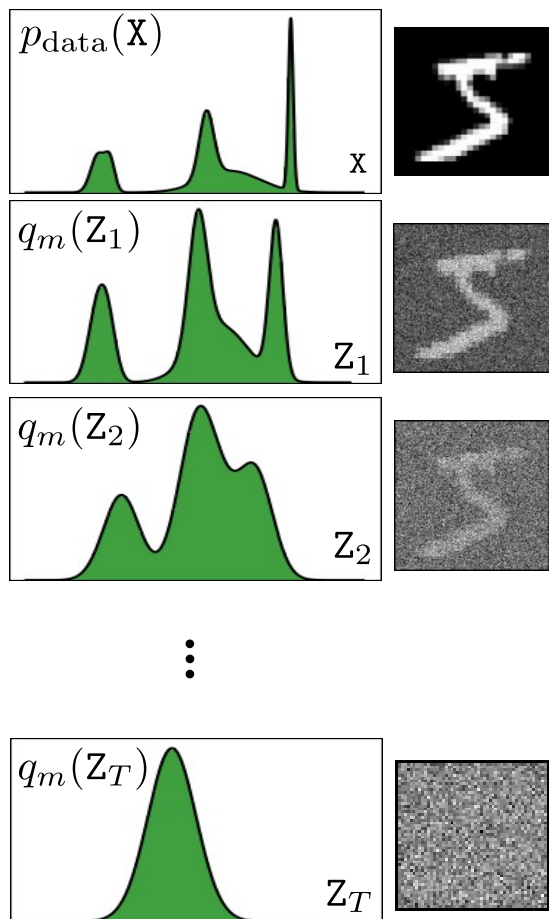
Bruiteur (Diffusion)

$$\mathbf{X}_{\text{ech}} \sim p_{\text{data}}(\mathbf{X})$$

$$\alpha_t = \prod_{k=1}^t (1 - \beta_k)$$

$$\mathbf{Z}_t = \sqrt{\alpha_t} \mathbf{X}_{\text{ech}} + \sqrt{1 - \alpha_t} \epsilon_t$$

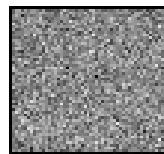
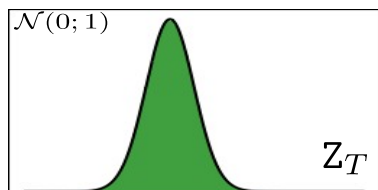
$$q_m(\mathbf{Z}_t | \mathbf{X}) = \mathcal{N}(\sqrt{\alpha_t} \mathbf{X}, (1 - \alpha_t) \mathbf{I})$$



Il faut choisir les $\{\beta_t\}_{t=1\dots T}$ de telle sorte que $\prod_{k=1}^T (1 - \beta_k) \rightarrow 0$

v)

Débruiteur



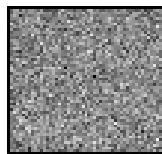
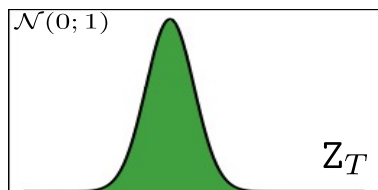
$$z_T = \mathbf{n}_T$$



$$\mathbf{n}_T \sim \mathcal{N}(0; \mathbf{I})$$

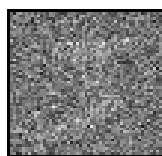
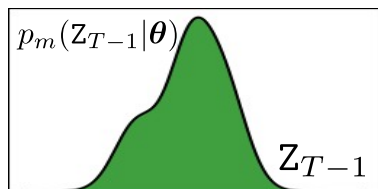
v)

Débruiteur



$$z_T = \mathbf{n}_T$$

$$\leftarrow \mathbf{n}_T \sim \mathcal{N}(0; \mathbf{I})$$



$$z_{T-1} = \mu_{\theta_{T-1}}(z_T) + \sigma_{T-1} \mathbf{n}_{T-1}$$

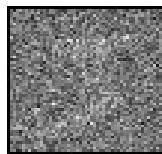
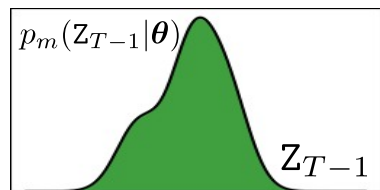
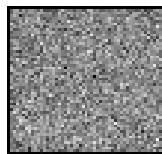
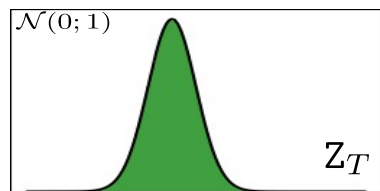
$$\mathbf{n}_{T-1} \sim \mathcal{N}(0; \mathbf{I})$$

Fonction paramétrique
(réseau de neurones)

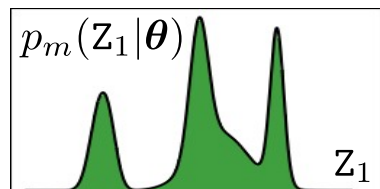
Paramètre
généralement fixé
manuellement

v)

Débruiteur



⋮



$$z_T = \mathbf{n}_T \quad \leftarrow \quad \mathbf{n}_T \sim \mathcal{N}(0; \mathbf{I})$$



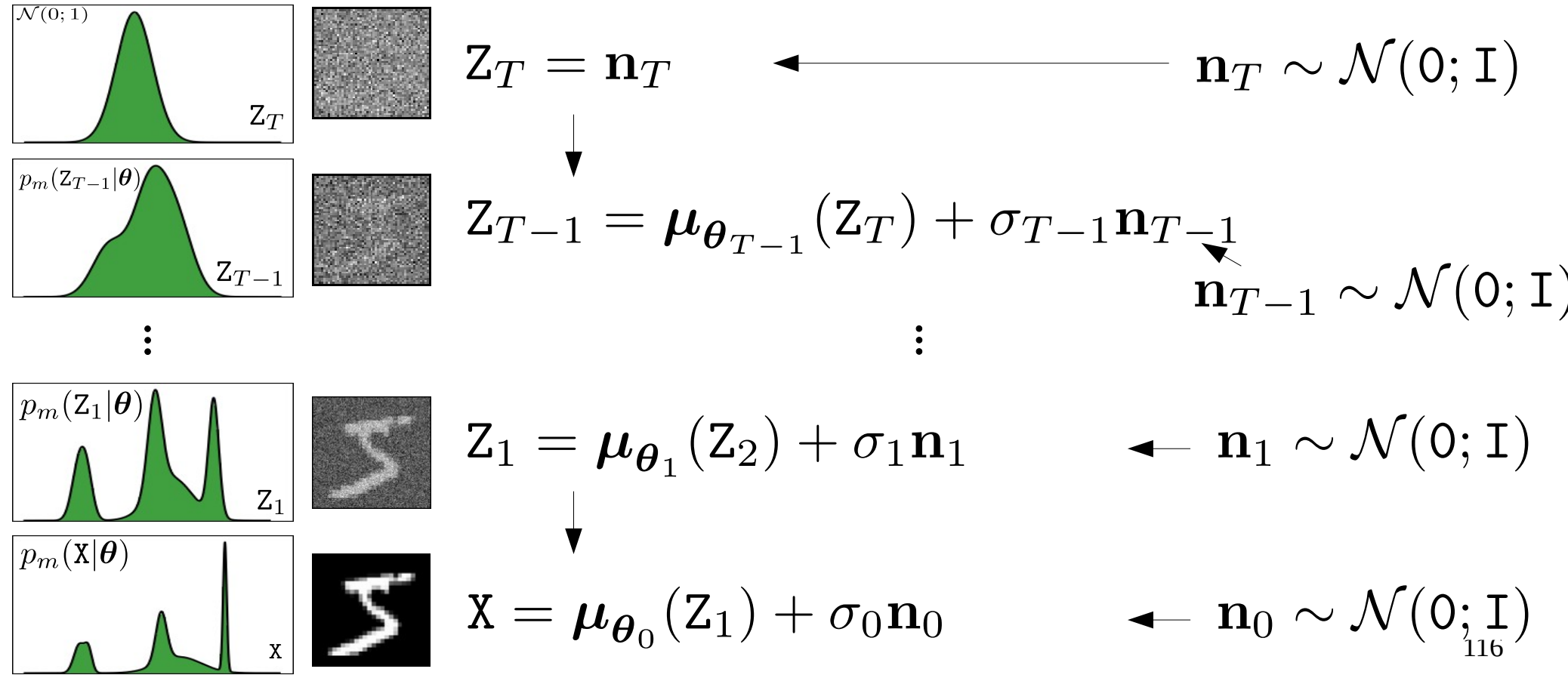
$$z_{T-1} = \mu_{\theta_{T-1}}(z_T) + \sigma_{T-1} \mathbf{n}_{T-1} \quad \leftarrow \quad \mathbf{n}_{T-1} \sim \mathcal{N}(0; \mathbf{I})$$

⋮

$$z_1 = \mu_{\theta_1}(z_2) + \sigma_1 \mathbf{n}_1 \quad \leftarrow \quad \mathbf{n}_1 \sim \mathcal{N}(0; \mathbf{I})$$

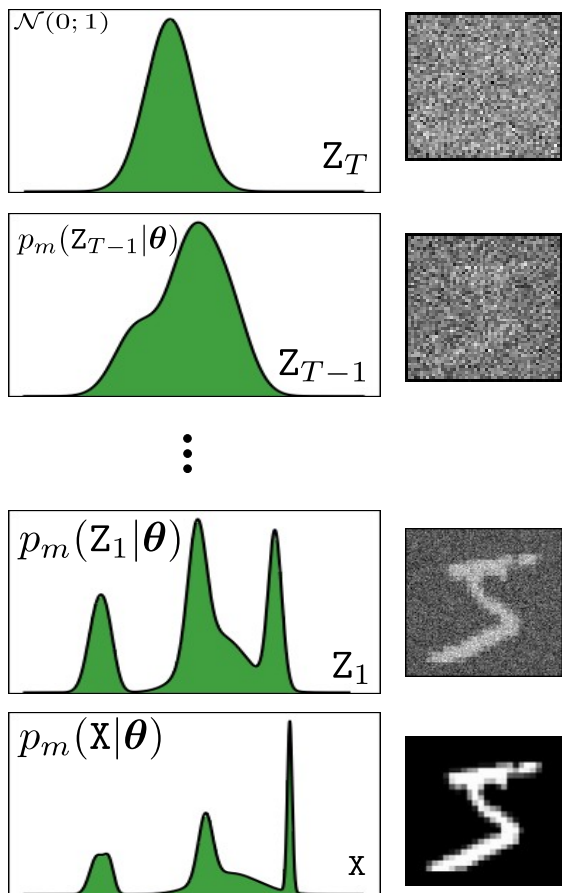
v)

Débruiteur



v)

Débruiteur



$$p_m(Z_{t-1} | Z_t) = \mathcal{N}(\mu_{\theta_{t-1}}(Z_t), \sigma_{t-1}^2 \mathbf{I})$$

v)

Réseau débruiteur : apprentissage

Comme pour le VAE, on ne peut pas calculer $p_{\text{modele}}(\mathbf{X}|\boldsymbol{\theta})$



$$~~KL(p_{\text{data}}(\mathbf{X}) || p_{\text{modele}}(\mathbf{X}|\boldsymbol{\theta}))~~$$

v)

Réseau débruiteur : apprentissage

Comme pour le VAE, on ne peut pas calculer $p_{\text{modele}}(\mathbf{X}|\boldsymbol{\theta})$



~~$$KL(p_{\text{data}}(\mathbf{X}) || p_{\text{modele}}(\mathbf{X}|\boldsymbol{\theta}))$$~~

Comme pour le VAE, on minimise la ELBO, ce qui ici correspond à minimiser (par rapport à $\boldsymbol{\theta}$)

$$KL(q_{\text{m}}(\mathbf{X}, \mathbf{Z}_1, \dots, \mathbf{Z}_T) || p_{\text{m}}(\mathbf{X}, \mathbf{Z}_1, \dots, \mathbf{Z}_T | \boldsymbol{\theta}))$$

v)

Réseau débruiteur : apprentissage (suite)

Paramétrisation de $\boldsymbol{\mu}_{\boldsymbol{\theta}_t}$:
$$\boldsymbol{\mu}_{\boldsymbol{\theta}_{t-1}}(\mathbf{z}_t) = \frac{1}{\sqrt{1 - \beta_t}} \left(\mathbf{z}_t - \frac{\beta_t}{\sqrt{1 - \alpha_t}} \hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}_{t-1}}(\mathbf{z}_t) \right)$$

v)

Réseau débruiteur : apprentissage (suite)

Paramétrisation de $\boldsymbol{\mu}_{\boldsymbol{\theta}_t}$: $\boldsymbol{\mu}_{\boldsymbol{\theta}_{t-1}}(\mathbf{Z}_t) = \frac{1}{\sqrt{1-\beta_t}} \left(\mathbf{Z}_t - \frac{\beta_t}{\sqrt{1-\alpha_t}} \hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}_{t-1}}(\mathbf{Z}_t) \right)$

Fonction de coût (obtenue pour $\sigma_{t-1} = \frac{\beta_t}{\sqrt{(1-\beta_t)(1-\alpha_t)}}$) :

$$L_i(\boldsymbol{\theta}) = \sum_{t=1}^T \left\| \boldsymbol{\epsilon}_{\text{ech},i,t} - \hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}_{t-1}}(\sqrt{\alpha_t} \mathbf{X}_i + \sqrt{1-\alpha_t} \boldsymbol{\epsilon}_{\text{ech},i,t}) \right\|_2^2$$

$$\boldsymbol{\epsilon}_{\text{ech},i,t} \sim \mathcal{N}(0; \mathbf{I})$$

v)

Réseau débruiteur : apprentissage (suite)

Paramétrisation de $\boldsymbol{\mu}_{\boldsymbol{\theta}_t}$: $\boldsymbol{\mu}_{\boldsymbol{\theta}_{t-1}}(\mathbf{z}_t) = \frac{1}{\sqrt{1 - \beta_t}} \left(\mathbf{z}_t - \frac{\beta_t}{\sqrt{1 - \alpha_t}} \hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}_{t-1}}(\mathbf{z}_t) \right)$

Fonction de coût (obtenue pour $\sigma_{t-1} = \frac{\beta_t}{\sqrt{(1 - \beta_t)(1 - \alpha_t)}}$) :

$$L_i(\boldsymbol{\theta}) = \sum_{t=1}^T \left\| \boldsymbol{\epsilon}_{\text{ech},i,t} - \hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}_{t-1}}(\sqrt{\alpha_t} \mathbf{x}_i + \sqrt{1 - \alpha_t} \boldsymbol{\epsilon}_{\text{ech},i,t}) \right\|_2^2$$

$$\boldsymbol{\epsilon}_{\text{ech},i,t} \sim \mathcal{N}(0; \mathbf{I})$$

Réseau de neurones qui apprend à prédire le bruit $\boldsymbol{\epsilon}_{\text{ech},i,t}$

Réseau débruiteur : apprentissage (suite)

Paramétrisation de $\boldsymbol{\mu}_{\boldsymbol{\theta}_t}$: $\boldsymbol{\mu}_{\boldsymbol{\theta}_{t-1}}(\mathbf{z}_t) = \frac{1}{\sqrt{1 - \beta_t}} \left(\mathbf{z}_t - \frac{\beta_t}{\sqrt{1 - \alpha_t}} \hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}_{t-1}}(\mathbf{z}_t) \right)$

Fonction de coût (obtenue pour $\sigma_{t-1} = \frac{\beta_t}{\sqrt{(1 - \beta_t)(1 - \alpha_t)}}$) :

$$L_i(\boldsymbol{\theta}) = \sum_{t=1}^T \left\| \boldsymbol{\epsilon}_{\text{ech},i,t} - \hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}_{t-1}}(\sqrt{\alpha_t} \mathbf{x}_i + \sqrt{1 - \alpha_t} \boldsymbol{\epsilon}_{\text{ech},i,t}) \right\|_2^2$$

Un réseau de neurones par pas de temps !

$$\boldsymbol{\epsilon}_{\text{ech},i,t} \sim \mathcal{N}(0; \mathbf{I})$$

→ T réseaux de neurones à apprendre ...

v)

Réseau débruiteur : apprentissage (suite)

Paramétrisation de $\boldsymbol{\mu}_{\boldsymbol{\theta}_t}$: $\boldsymbol{\mu}_{\boldsymbol{\theta}_{t-1}}(\mathbf{Z}_t) = \frac{1}{\sqrt{1 - \beta_t}} \left(\mathbf{Z}_t - \frac{\beta_t}{\sqrt{1 - \alpha_t}} \hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}_{t-1}}(\mathbf{Z}_t) \right)$

Fonction de coût (obtenue pour $\sigma_{t-1} = \frac{\beta_t}{\sqrt{(1 - \beta_t)(1 - \alpha_t)}}$) :

$$L_i(\boldsymbol{\theta}) = \sum_{t=1}^T \left\| \boldsymbol{\epsilon}_{\text{ech},i,t} - \hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}_{t-1}}(\sqrt{\alpha_t} \mathbf{X}_i + \sqrt{1 - \alpha_t} \boldsymbol{\epsilon}_{\text{ech},i,t}) \right\|_2^2$$

$$\boldsymbol{\epsilon}_{\text{ech},i,t} \sim \mathcal{N}(0; \mathbf{I})$$

En pratique, un seul réseau de la forme $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\mathbf{Z}_t, t - 1)$

v)

Réseau débruiteur : apprentissage (suite)

$$L_i(\boldsymbol{\theta}) = \sum_{t=1}^T \left\| \boldsymbol{\epsilon}_{\text{ech},i,t} - \hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\sqrt{\alpha_t} \mathbf{X}_i + \sqrt{1 - \alpha_t} \boldsymbol{\epsilon}_{\text{ech},i,t}, t - 1) \right\|_2^2$$

\uparrow
 $\boldsymbol{\epsilon}_{\text{ech},i,t} \sim \mathcal{N}(0; \mathbf{I})$

v)

Réseau débruiteur : apprentissage (suite)

$$L_i(\boldsymbol{\theta}) = \sum_{t=1}^T \left\| \boldsymbol{\epsilon}_{\text{ech},i,t} - \hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\sqrt{\alpha_t} \mathbf{X}_i + \sqrt{1 - \alpha_t} \boldsymbol{\epsilon}_{\text{ech},i,t}, t - 1) \right\|_2^2$$

\uparrow
 $\boldsymbol{\epsilon}_{\text{ech},i,t} \sim \mathcal{N}(0; \mathbf{I})$

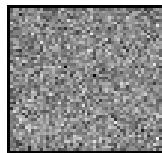
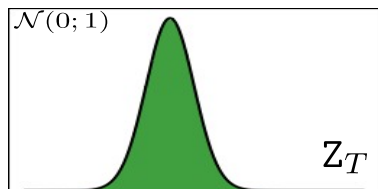
Apprentissage par descente de gradient stochastique

$$\sum_{i \in \Omega_{i,\text{data}}} \sum_{t \in \Omega_{i,\text{time}}} \left\| \boldsymbol{\epsilon}_{\text{ech},i,t} - \hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\sqrt{\alpha_t} \mathbf{X}_i + \sqrt{1 - \alpha_t} \boldsymbol{\epsilon}_{\text{ech},i,t}, t - 1) \right\|_2^2$$

Minibatch sur les données et les pas de temps → méthode “simulation free” ! 126

v)

Résumé de l'étape de génération

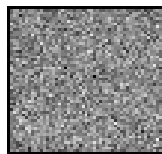
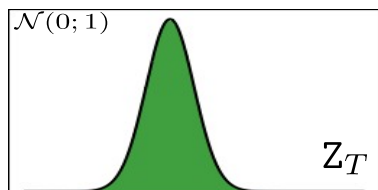


$$z_T = \mathbf{n}_T$$

$$\leftarrow \mathbf{n}_T \sim \mathcal{N}(0; \mathbf{I})$$

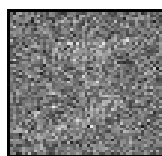
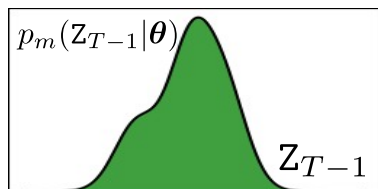
v)

Résumé de l'étape de génération



$$Z_T = \mathbf{z}_T$$

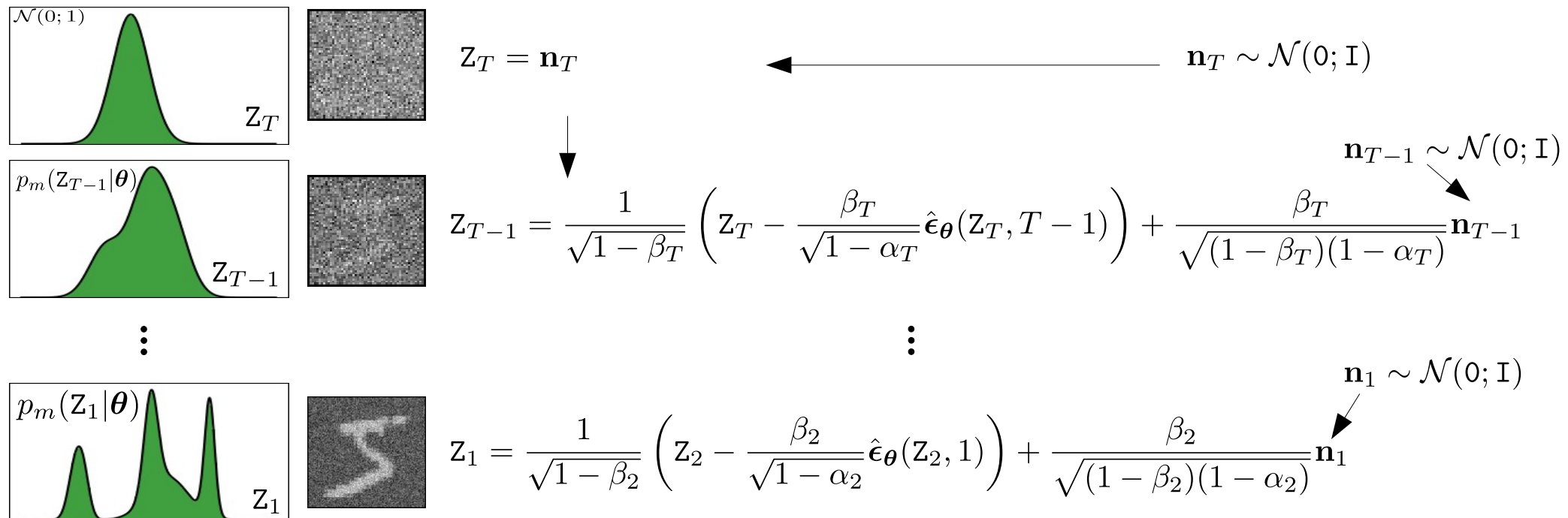
$$\longleftarrow \mathbf{z}_T \sim \mathcal{N}(0; \mathbf{I})$$



$$Z_{T-1} = \frac{1}{\sqrt{1 - \beta_T}} \left(Z_T - \frac{\beta_T}{\sqrt{1 - \alpha_T}} \hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(Z_T, T - 1) \right) + \frac{\beta_T}{\sqrt{(1 - \beta_T)(1 - \alpha_T)}} \mathbf{z}_{T-1}$$

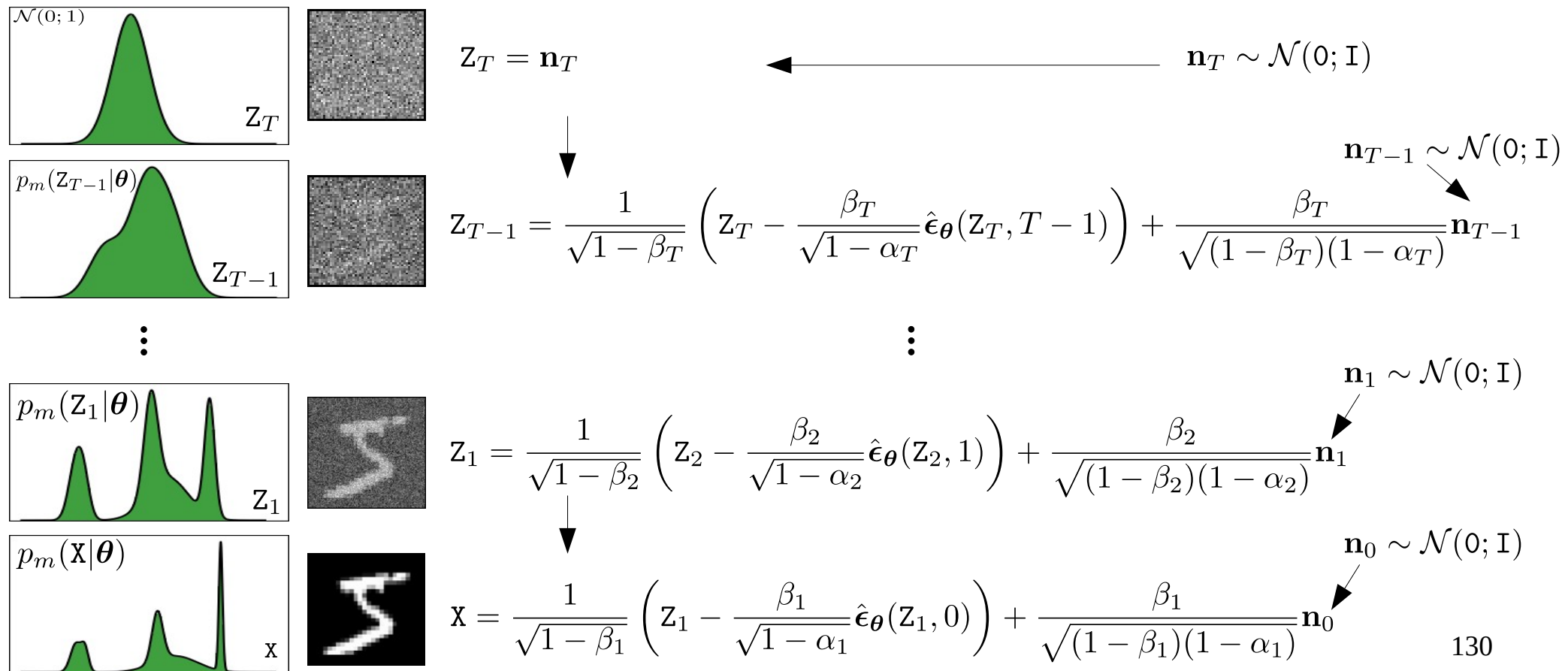
$\mathbf{z}_{T-1} \sim \mathcal{N}(0; \mathbf{I})$

Résumé de l'étape de génération



v)

Résumé de l'étape de génération



Réseau débruiteur : avantages et inconvénients

- + pas d'encodeur à apprendre
- + pas de contrainte particulière sur l'architecture
- + méthode “simulation free” (lors de l'apprentissage on ne fait pas toute l'inférence, juste des petits morceaux)

Réseau débruiteur : avantages et inconvénients

- + pas d'encodeur à apprendre
- + pas de contrainte particulière sur l'architecture
- + méthode "simulation free" (lors de l'apprentissage on ne fait pas toute l'inférence, juste des petits morceaux)
- pas d'expression exacte de la (log-)probabilité d'une donnée
- pas d'apprentissage par maximum de vraisemblance (mais borne inférieure quand même)
- échantillonnage lent (T généralement grand, Z_t de la taille de X)

Réseau débruiteur : réduction du coût calculatoire/mémoire

“Stable diffusion” - High-Resolution Image Synthesis with Latent Diffusion Models, CVPR 2022

1) entraîner un VAE pour encoder les images dans un espace latent de plus faible dimension

2) entraîner un réseau débruiteur sur cet espace latent

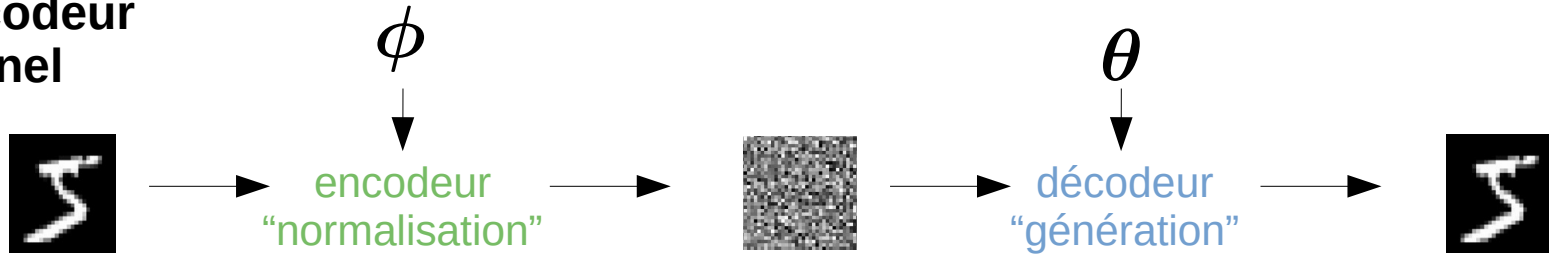


Exemple de super-résolution

VI) Réseau antagoniste (GAN)

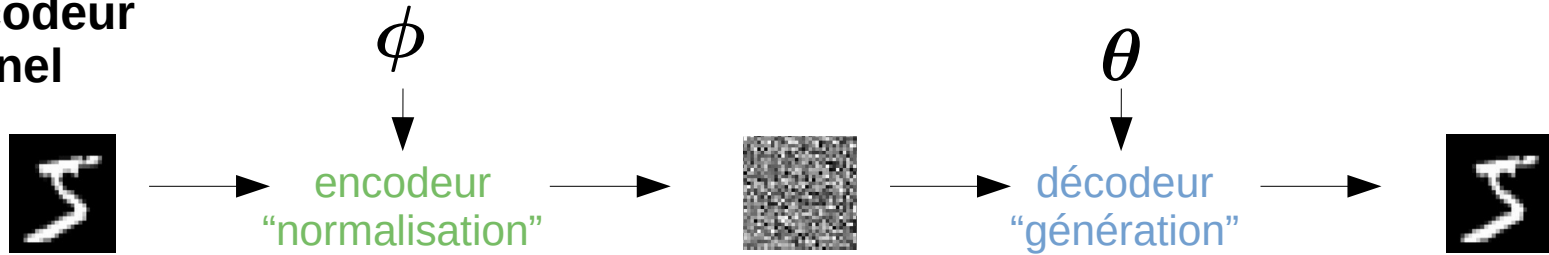
Réseau antagoniste : idée générale

Auto-encodeur
variationnel

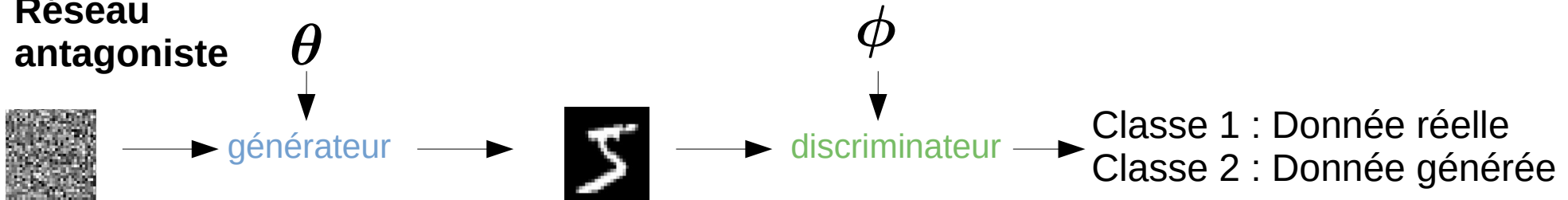


Réseau antagoniste : idée générale

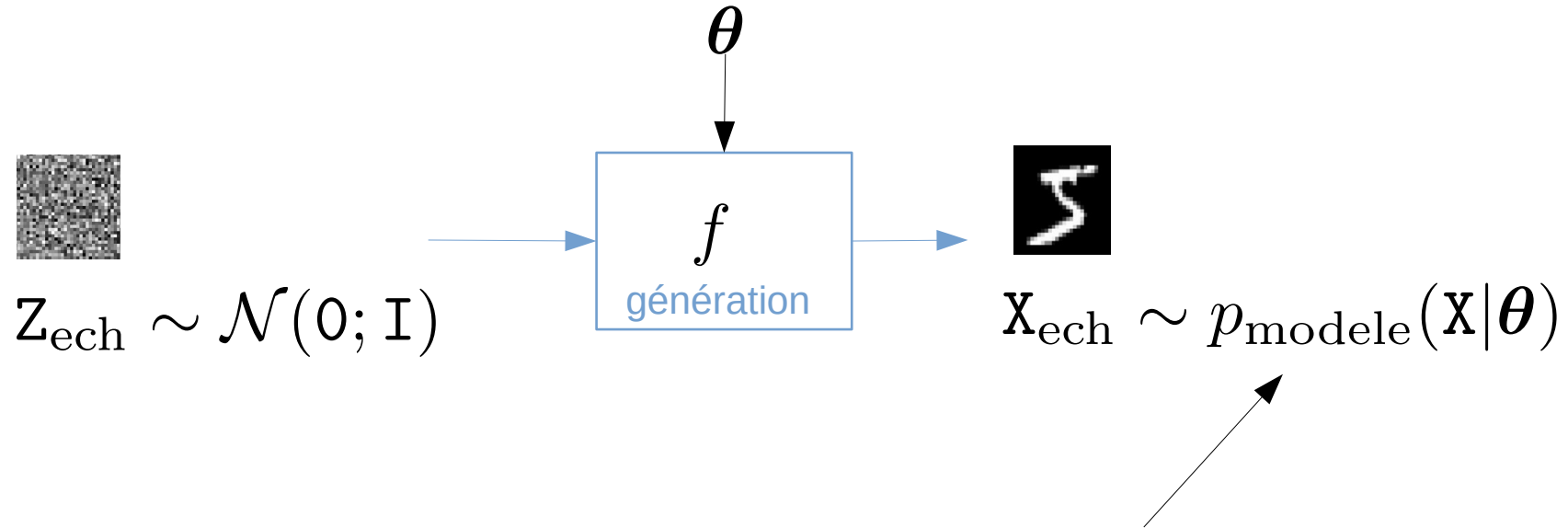
Auto-encodeur
variationnel



Réseau
antagoniste

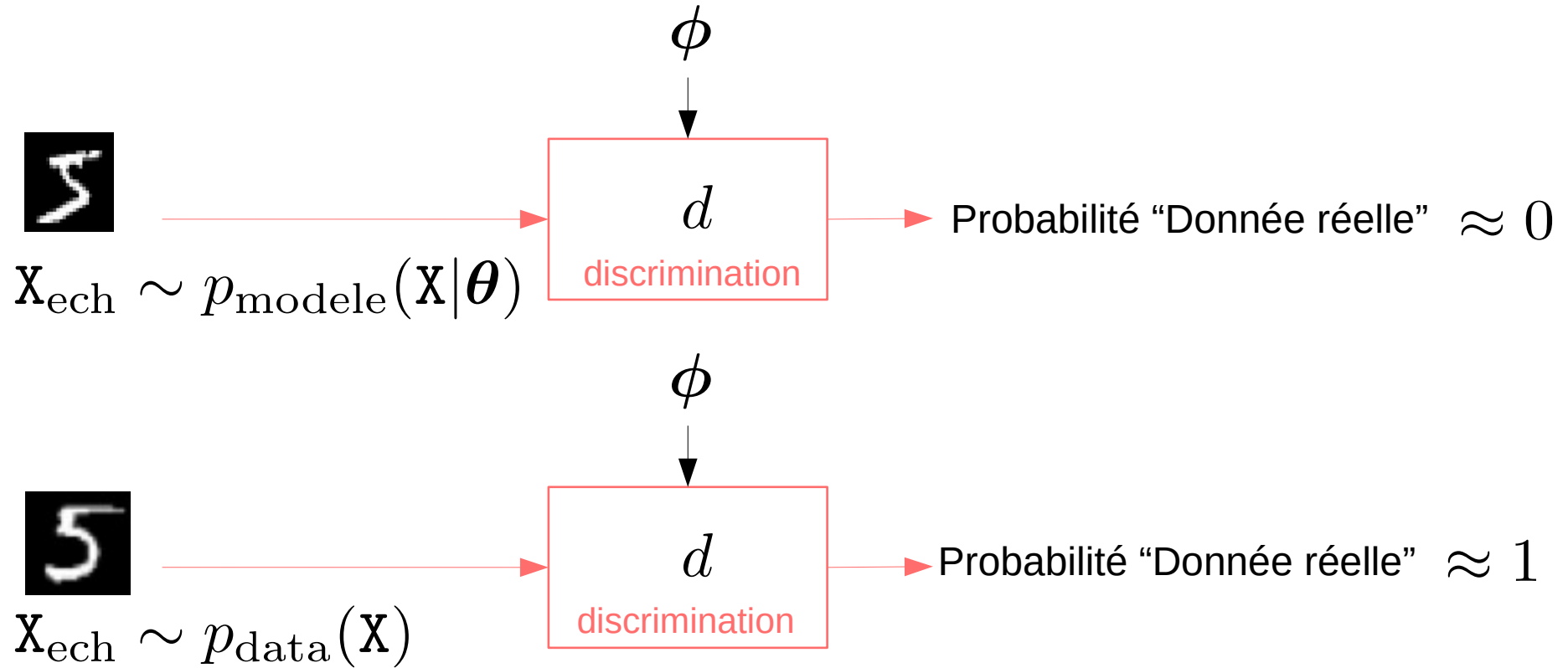


Réseau antagoniste : générateur

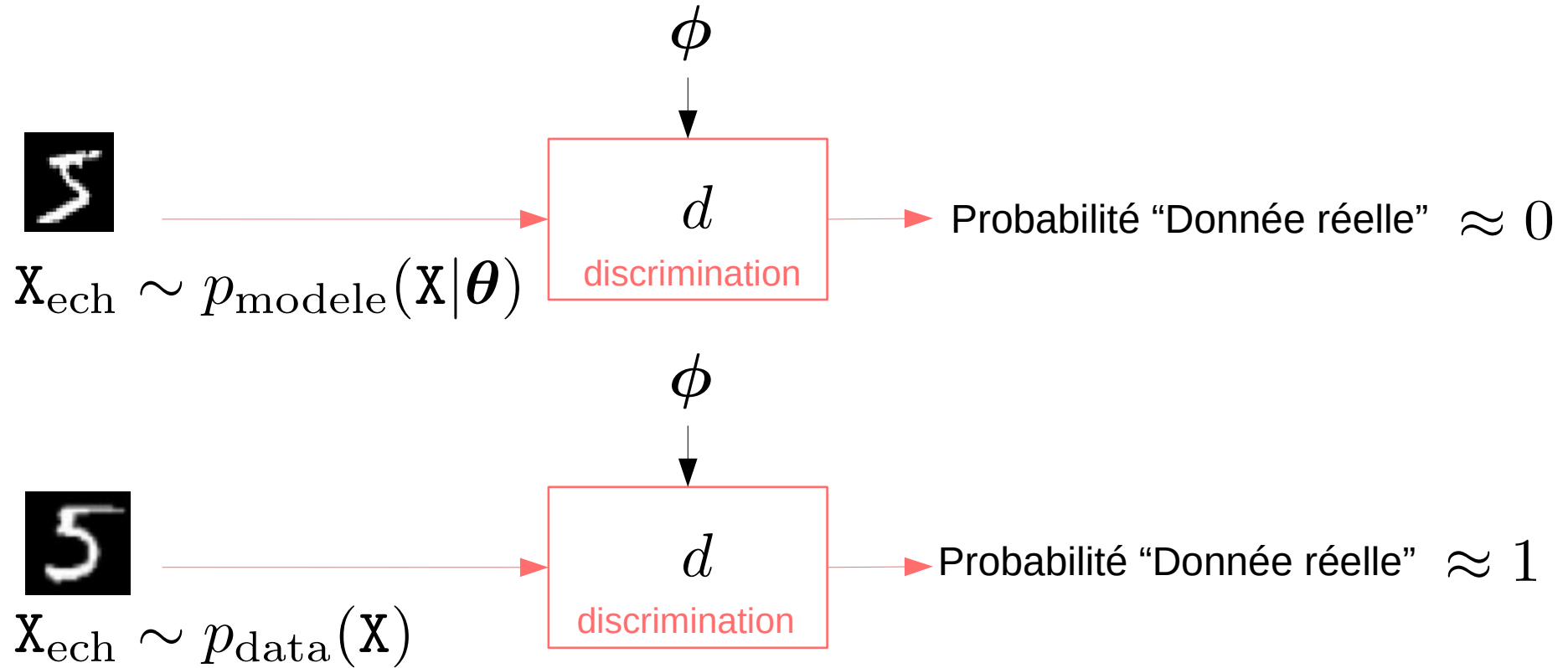


Définition implicite, le générateur n'est pas inversible

Réseau antagoniste : discriminateur

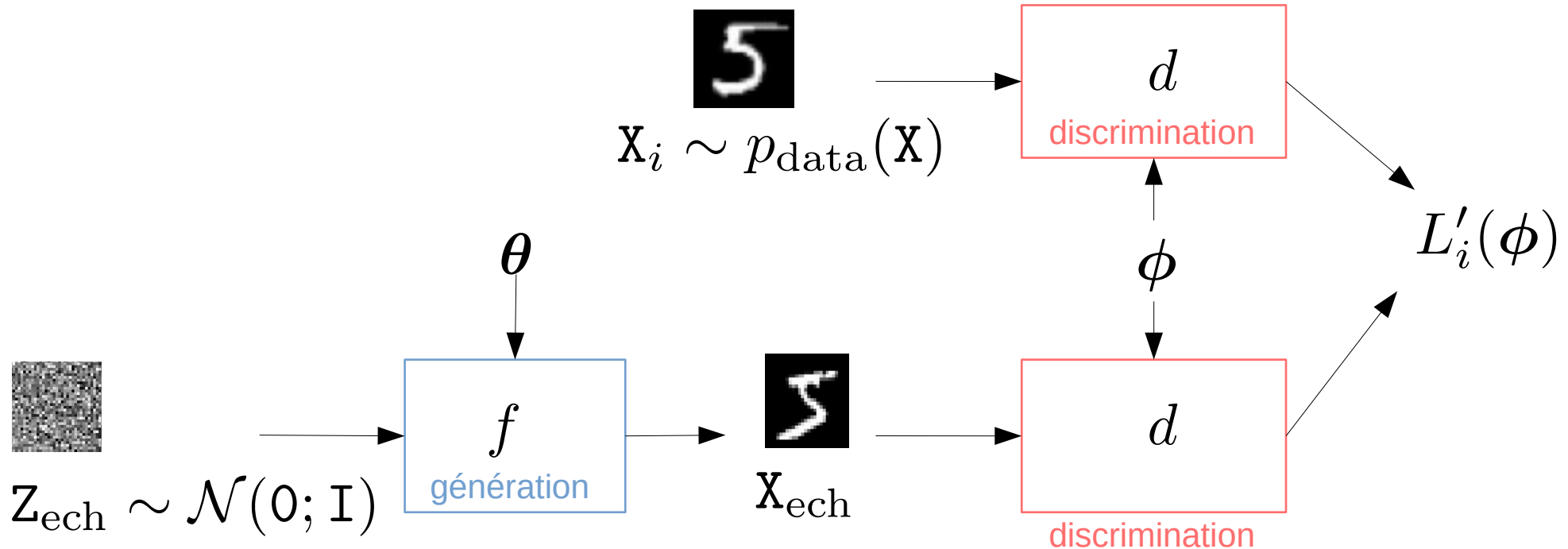


Réseau antagoniste : discriminateur



Réseau antagoniste : apprentissage

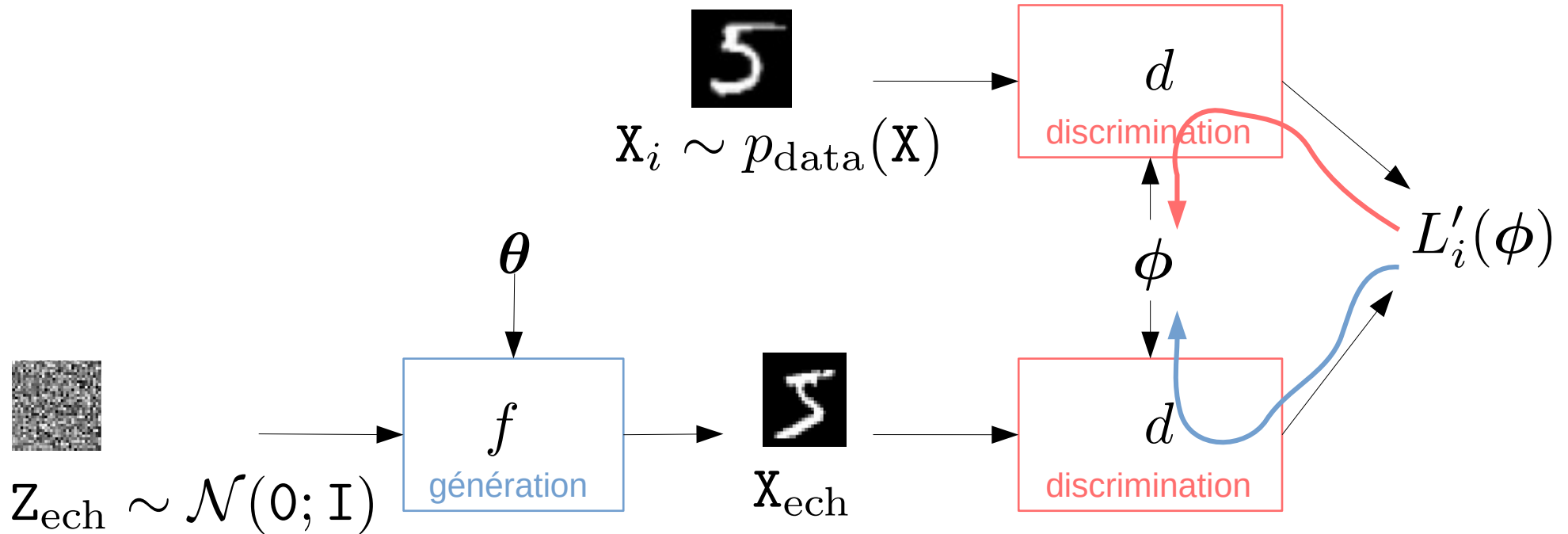
Pas de gradient sur les paramètres du discriminateur



Le discriminateur s'améliore pour mieux détecter les faux échantillons.

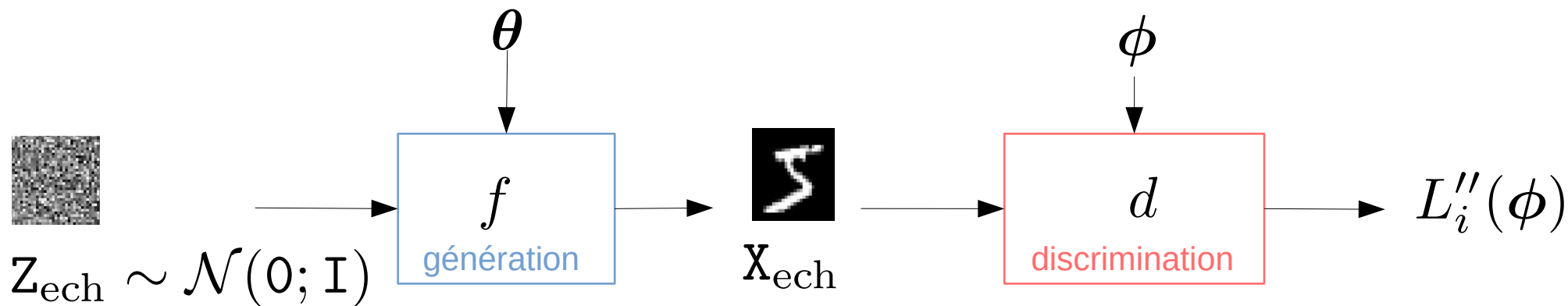
Réseau antagoniste : apprentissage

Pas de gradient sur les paramètres du discriminateur



Réseau antagoniste : apprentissage (suite)

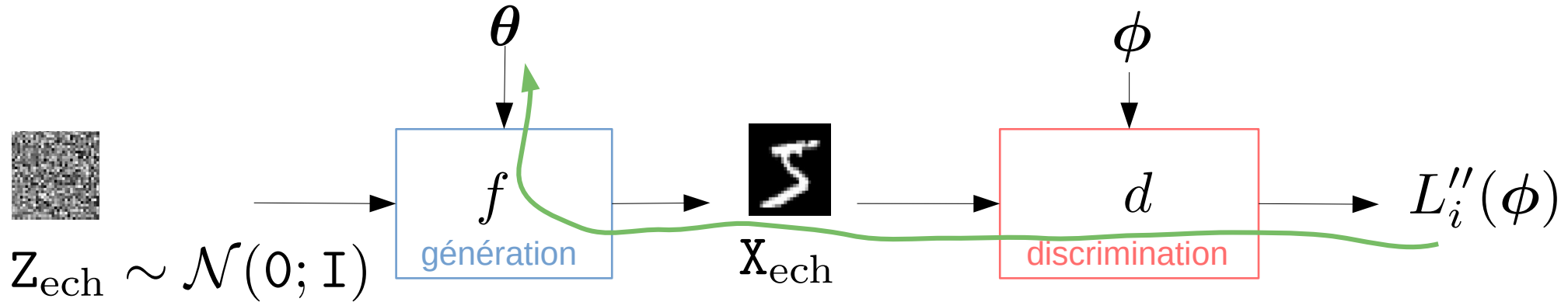
Pas de gradient sur les paramètres du générateur



Le générateur s'améliore pour mieux tromper le discriminateur.

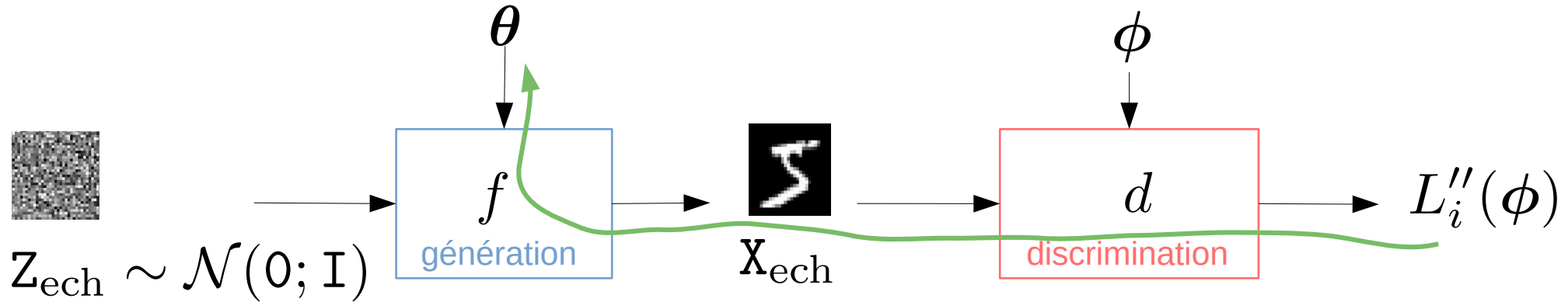
Réseau antagoniste : apprentissage (suite)

Pas de gradient sur les paramètres du générateur



Réseau antagoniste : apprentissage (suite)

Pas de gradient sur les paramètres du générateur



Problème d'apprentissage très difficile car il ne faut pas que l'un "écrase" l'autre.

Réseau antagoniste : apprentissage (formalisé)

Objectif : optimiser θ pour que $p_{\text{modele}}(\mathbf{X}|\theta) \approx p_{\text{data}}(\mathbf{X})$


Réseau antagoniste : apprentissage (formalisé)

Objectif : optimiser θ pour que $p_{\text{modele}}(\mathbf{X}|\theta) \approx p_{\text{data}}(\mathbf{X})$

Pour un GAN on choisit généralement une borne inférieure d'une divergence :

$$D(p_m(\mathbf{X}|\theta) || p_{\text{data}}(\mathbf{X}))$$

$$\geq \max_{d(.)} \mathbb{E}_{p_m(\mathbf{x}|\theta)}(\ln(1 - d(\mathbf{X}))) + \mathbb{E}_{p_{\text{data}}(\mathbf{x})}(\ln(d(\mathbf{X})))$$



discriminateur

Réseau antagoniste : apprentissage (formalisé)

Objectif : optimiser θ pour que $p_{\text{modele}}(\mathbf{X}|\theta) \approx p_{\text{data}}(\mathbf{X})$

Pour un GAN on choisit généralement une borne inférieure d'une divergence :

$$D(p_m(\mathbf{X}|\theta) || p_{\text{data}}(\mathbf{X}))$$

$$\geq \max_{d(\cdot)} \mathbb{E}_{p_m(\mathbf{x}|\theta)} (\ln(1 - d(\mathbf{X}))) + \mathbb{E}_{p_{\text{data}}(\mathbf{x})} (\ln(d(\mathbf{X})))$$

discriminateur

Uniquement besoin de savoir échantillonner selon $p_m(\mathbf{X}|\theta)$

Réseau antagoniste : apprentissage (formalisé)

Objectif : optimiser θ pour que $p_{\text{modele}}(\mathbf{X}|\theta) \approx p_{\text{data}}(\mathbf{X})$

Pour un GAN on choisit généralement une borne inférieure d'une divergence :

$$\begin{aligned} D(p_m(\mathbf{X}|\theta) || p_{\text{data}}(\mathbf{X})) \\ \geq \max_{d(\cdot)} \mathbb{E}_{p_m(\mathbf{x}|\theta)}(\ln(1 - d(\mathbf{X}))) + \mathbb{E}_{p_{\text{data}}(\mathbf{x})}(\ln(d(\mathbf{X}))) \end{aligned}$$

Astuce de reparamétrisation ("Reparameterization trick")

$$= \max_{d(\cdot)} \mathbb{E}_{\mathcal{N}(\mathbf{0}, \mathbf{I})}(\ln(1 - d(f(\mathbf{Z}, \theta)))) + \mathbb{E}_{p_{\text{data}}(\mathbf{x})}(\ln(d(\mathbf{X})))$$

Réseau antagoniste : apprentissage (formalisé)

Objectif : optimiser θ pour que $p_{\text{modele}}(\mathbf{X}|\theta) \approx p_{\text{data}}(\mathbf{X})$

Pour un GAN on choisit généralement une borne inférieure d'une divergence :

$$\begin{aligned} D(p_m(\mathbf{X}|\theta) || p_{\text{data}}(\mathbf{X})) \\ \geq \max_{d(\cdot)} \mathbb{E}_{p_m(\mathbf{x}|\theta)}(\ln(1 - d(\mathbf{X}))) + \mathbb{E}_{p_{\text{data}}(\mathbf{x})}(\ln(d(\mathbf{X}))) \end{aligned}$$

Astuce de reparamétrisation ("Reparameterization trick")

$$= \max_{d(\cdot)} \mathbb{E}_{\mathcal{N}(\mathbf{0}, \mathbf{I})}(\ln(1 - d(f(\mathbf{Z}, \theta)))) + \mathbb{E}_{p_{\text{data}}(\mathbf{x})}(\ln(d(\mathbf{X})))$$

En pratique, on va paramétrer le discriminateur par un réseau de neurones.

Réseau antagoniste : problème d'apprentissage

$$\min_{\boldsymbol{\theta}} \max_{\phi} \mathbb{E}_{\mathcal{N}(\mathbf{0}, \mathbf{I})} (\ln(1 - d(f(\mathbf{Z}, \boldsymbol{\theta}), \phi))) + \mathbb{E}_{p_{\text{data}}(\mathbf{x})} (\ln(d(\mathbf{x}, \phi)))$$

Problème minmax → potentiellement beaucoup plus difficile à optimiser qu'un VAE ou qu'un NF

Réseau antagoniste : problème d'apprentissage

En pratique on alterne :

Un pas gradient pour optimiser ϕ

$$L'_i(\phi) = -\ln(1 - d(f(\mathbf{Z}_{\text{ech}}, \boldsymbol{\theta}), \phi)) - \ln(d(\mathbf{X}_i, \phi))$$

Objectif : Améliorer les paramètres du discriminateur afin qu'il prédise une valeur proche de 0 pour une donnée générée et proche de 1 pour une donnée réelle

Réseau antagoniste : problème d'apprentissage

En pratique on alterne :

Un pas gradient pour optimiser ϕ

$$L'_i(\phi) = -\ln(1 - d(f(Z_{\text{ech}}, \theta), \phi)) - \ln(d(X_i, \phi))$$

Objectif : Améliorer les paramètres du discriminateur afin qu'il prédise une valeur proche de 0 pour une donnée générée et proche de 1 pour une donnée réelle

Suivi d'un pas de gradient pour optimiser θ

$$L''_i(\theta) = \ln(1 - d(f(Z_{\text{ech}}, \theta), \phi)) \longrightarrow -\ln(d(f(Z_{\text{ech}}, \theta), \phi))$$

En pratique remplacé par :

Objectif : Améliorer les paramètres du générateur afin que le discriminateur se trompe et prédise une valeur proche de 1

Réseau antagoniste : avantages et inconvénients

- + capable d'échantillonner efficacement
- + pas de contrainte particulière sur l'architecture
- + taille de Z non contrainte par celle de X

Réseau antagoniste : avantages et inconvénients

- + capable d'échantillonner efficacement
- + pas de contrainte particulière sur l'architecture
- + taille de Z non contrainte par celle de X
- pas d'expression de la (log-)probabilité d'une donnée
- problème minmax difficile à optimiser
- ** problèmes de convergence
- ** échantillonnage uniquement d'une partie de la distribution ("mode collapse")

