

Projet Deep Learning

Franck Chen

3 janvier 2025

Introduction

Ce rapport présente les travaux réalisés dans le cadre d'un projet sur le Deep Learning. Il est structuré comme suit :

- Présentation de la méthode étudiée : les f-GAN.
- Analyse des résultats obtenus.
- Discussion des limites rencontrées.

1 Présentation de la méthode : f-GAN

1.1 f-divergence

Étant donné deux distributions P et Q possédant respectivement des fonctions de densité absolument continues p et q par rapport à une mesure de base dx définie sur le domaine \mathcal{X} , nous définissons la f-divergence comme suit :

$$D_f(P \parallel Q) = \int_{\mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right) dx, \quad (1)$$

où $f : \mathbb{R}^+ \rightarrow \mathbb{R}$ est une fonction convexe, semi-continue inférieurement, et vérifiant $f(1) = 0$. La condition $f(1) = 0$ garantit que la f-divergence est nulle lorsque les distributions P et Q sont identiques.

1.2 Conjugué de f

Toute fonction f convexe et semi-continue inférieurement possède une fonction conjuguée convexe f^* , également appelée conjugué de Fenchel. Le conjugué de la fonction f , noté f^* , est défini par :

$$f^*(t) = \sup_{u \in \text{dom}_f} \{ut - f(u)\} \quad (2)$$

Cette fonction possède des propriétés importantes. Par exemple, f^* est convexe et semi-continue inférieurement. Étant donné que f est convexe et semi-continue inférieurement, la dualité convexe de Fenchel garantit que nous pouvons exprimer f en termes de son conjugué dual comme suit :

$$f(u) = \sup_{t \in \text{dom}_{f^*}} \{ut - f^*(t)\} \quad (3)$$

1.3 Expression et optimisation de la f-divergence

En utilisant la nouvelle définition de f , la f-divergence peut être approximée par une borne inférieure en utilisant l'inégalité de Jensen :

$$D_f(P \parallel Q) = \int_{\mathcal{X}} q(x) \sup_{t \in \text{dom}_{f^*}} \left\{ \frac{p(x)}{q(x)} t - f^*(t) \right\} dx \quad (4)$$

$$\geq \sup_{T \in \mathcal{T}} \left(\int_{\mathcal{X}} p(x) T(x) dx - \int_{\mathcal{X}} q(x) f^*(T(x)) dx \right) \quad (5)$$

$$= \sup_{T \in \mathcal{T}} (\mathbb{E}_{x \sim P}[T(x)] - \mathbb{E}_{x \sim Q}[f^*(T(x))]) \quad (6)$$

où \mathcal{T} est une classe arbitraire de fonctions $T : \mathcal{X} \rightarrow \mathbb{R}$.

Dans le cadre de l'optimisation de la borne inférieure de la f-divergence, la fonction optimale $T^*(x)$ est obtenue en prenant la dérivée de la borne inférieure par rapport à T :

$$T^*(x) = f' \left(\frac{p(x)}{q(x)} \right) \quad (7)$$

Cette fonction $T^*(x)$ permet d'atteindre le supremum dans (6). Mais en pratique, $T^*(x)$ n'est pas directement accessible.

1.4 Fonction objective des f-GAN

Les f-GAN utilisent une fonction objective définie par :

$$F(\theta, \omega) = \mathbb{E}_{x \sim P}[T_\omega(x)] - \mathbb{E}_{x \sim Q_\theta}[f^*(T_\omega(x))] \quad (8)$$

où $T_\omega(x)$ est le score du discriminateur sur les images, $\mathbb{E}_{x \sim P}[T_\omega(x)]$ est la moyenne des scores du discriminateur sur les images réelles, $\mathbb{E}_{x \sim Q_\theta}[f^*(T_\omega(x))]$ est la moyenne des scores du discriminateur sur les images générées.

Pour s'assurer que $T_\omega(x)$ renvoie des valeurs dans l'ensemble de définition de f^* , nous allons l'exprimer sous une nouvelle forme :

$$T_\omega(x) = g_f(V_\omega(x)) \quad (9)$$

où $V_\omega : \mathcal{X} \rightarrow \mathbb{R}$ représente le discriminateur et $g_f : \mathbb{R} \rightarrow \text{dom}_{f^*}$ est une fonction d'activation spécifique au choix de la f-divergence.

En utilisant cette nouvelle forme de $T_\omega(x)$, la fonction objective devient :

$$F(\theta, \omega) = \mathbb{E}_{x \sim P}[g_f(V_\omega(x))] - \mathbb{E}_{x \sim Q_\theta}[f^*(g_f(V_\omega(x)))] \quad (10)$$

1.5 Fonctions de perte

Le discriminateur cherche à maximiser F , mesurant sa capacité à distinguer les vraies données de celles générées. Le générateur cherche à minimiser F , afin de produire des données de plus en plus réalistes. Ainsi, les f-GAN utilisent les fonctions de perte suivantes :

$$L_D = -(\mathbb{E}_{x \sim P}[g_f(V_\omega(x))] - \mathbb{E}_{x \sim Q_\theta}[f^*(g_f(V_\omega(x)))] \quad (11)$$

$$L_G = -\mathbb{E}_{x \sim Q_\theta}[f^*(g_f(V_\omega(x)))] \quad (12)$$

1.6 Différentes divergences et leurs caractéristiques

Table 1: Tableau représentant les fonctions d'activation recommandées pour la couche finale, la fonction conjugué f^* associée à chaque divergence et le niveau critique du score du discriminateur défini par $f'(1)$. La valeur critique $f'(1)$ peut être interprétée comme un seuil de classification appliqué à $T_\omega(x)$ pour distinguer les échantillons réels des échantillons générés. En effet, lorsque $q(x) = p(x)$, le discriminateur est incapable de distinguer les images générées des images réelles et $T^*(x) = f' \left(\frac{p(x)}{q(x)} \right) = f' \left(\frac{p(x)}{p(x)} \right) = f'(1)$.

Nom	Fonction g_f	Domaine de f^*	Conjugué $f^*(t)$	$f'(1)$
Kullback-Leibler (KL)	v	\mathbb{R}	$\exp(t - 1)$	1
Reverse KL	$-\exp(-v)$	\mathbb{R}^-	$-1 - \log(-t)$	-1
Pearson Chi-Square	v	\mathbb{R}	$\frac{1}{4}t^2 + t$	0
Squared Hellinger	$1 - \exp(-v)$	$t < 1$	$\frac{t}{1-t}$	0
Jensen-Shannon	$\log(2) - \log(1 + \exp(-v))$	$t < \log(2)$	$-\log(2 - \exp(t))$	0
GAN	$-\log(1 + \exp(-v))$	\mathbb{R}^-	$-\log(1 - \exp(t))$	$-\log(2)$

Table 2: Tableau représentant les différentes fonctions de divergence les plus connues.

Nom	Divergence $D_f(P \parallel Q)$	Générateur $f(u)$
Kullback-Leibler (KL)	$\int_{\mathcal{X}} p(x) \log \left(\frac{p(x)}{q(x)} \right) dx$	$u \log(u)$
Reverse KL	$\int_{\mathcal{X}} q(x) \log \left(\frac{q(x)}{p(x)} \right) dx$	$-\log(u)$
Pearson Chi-Square	$\int_{\mathcal{X}} \frac{(p(x)-q(x))^2}{q(x)} dx$	$(u-1)^2$
Squared Hellinger	$\int_{\mathcal{X}} \left(\sqrt{p(x)} - \sqrt{q(x)} \right)^2 dx$	$(\sqrt{u}-1)^2$
Jensen-Shannon	$\frac{1}{2} \int_{\mathcal{X}} p(x) \log \left(\frac{2p(x)}{p(x)+q(x)} \right) + q(x) \log \left(\frac{2q(x)}{p(x)+q(x)} \right) dx$	$-(u+1) \log \left(\frac{1+u}{2} \right) + u \log(u)$
GAN	$\int_{\mathcal{X}} p(x) \log \left(\frac{2p(x)}{p(x)+q(x)} \right) + q(x) \log \left(\frac{2q(x)}{p(x)+q(x)} \right) dx - \log(4)$	$u \log(u) - (u+1) \log(u+1)$

Chaque divergence a des caractéristiques propres. L'objectif défini détermine le choix de la f-divergence : prioriser la précision, prioriser le rappel, trouver un compromis.

Quelques remarques sur ces divergences :

Vanilla GAN. Le vanilla GAN est un cas particulier des f-GANs. En effet il utilise la Binary Cross-Entropy comme fonction de perte qui s'écrit comme suit :

$$\text{BCE}(y, \hat{y}) = -[y \log(\hat{y}) + (1-y) \log(1-\hat{y})] \quad (13)$$

où :

- $y \in \{0, 1\}$ est la valeur réelle (étiquette) associée à l'échantillon. $y = 1$ pour une classe positive, $y = 0$ pour une classe négative.
- $\hat{y} \in (0, 1)$ est la probabilité prédite par le modèle pour la classe positive.
- Le premier terme $y \log(\hat{y})$ mesure la perte lorsque la prédiction pour la classe positive est incorrecte.
- Le second terme $(1-y) \log(1-\hat{y})$ mesure la perte lorsque la prédiction pour la classe négative est incorrecte.

Dans le contexte du GAN :

- Pour les données réelles ($y = 1$), $\hat{y} = D(x)$, donc la perte devient $-\log D(x)$.
- Pour les données générées ($y = 0$), $\hat{y} = D(x)$, donc la perte devient $-\log(1-D(x))$.

Par conséquent, l'objectif du Vanilla GAN, où le discriminateur maximise $\mathbb{E}_{x \sim P}[\log D_{\omega}(x)] + \mathbb{E}_{x \sim Q_{\theta}}[\log(1-D_{\omega}(x))]$, correspond exactement à minimiser la Binary Cross-Entropy sur une classification binaire entre échantillons réels et générés.

Le GAN classique n'est pas directement conçu pour minimiser une f-divergence comme les f-GAN. Au lieu de cela, il minimise une perte basée sur la Binary Cross-Entropy entre les échantillons réels et générés. Cette différence d'objectif explique pourquoi la fonction $f(u)$ du GAN classique ne satisfait pas la contrainte $f(1) = 0$ et donc $D_f(P \parallel P) \neq 0$.

Jensen-Shannon. Nous pouvons remarquer que le GAN est lié à la divergence de Jensen-Shannon car $D_{\text{GAN}} = 2 \cdot D_{\text{JS}} - \log(4)$.

Kullback Leibler. La divergence KL n'est pas symétrique : c'est-à-dire que $D_{\text{KL}}(P \parallel Q) \neq D_{\text{KL}}(Q \parallel P)$. En conséquence, ce n'est pas non plus une métrique de distance.

Afin que la divergence KL soit finie, le support de P doit être contenu dans le support de Q . Si un point x existe tel que $Q(x) = 0$ mais $P(x) > 0$, alors $D_{KL}(P \parallel Q) = \infty$.

Nous pouvons analyser la fonction de divergence pour déduire le comportement de la minimisation de celle-ci :

Lorsque $p(x) > q(x)$: La divergence $D_f(P \parallel Q)$ devient très élevée, ce qui incite le modèle génératif à augmenter considérablement $q(x)$ dans ces zones pour réduire la divergence.

Lorsque $p(x) < q(x)$: La divergence $D_f(P \parallel Q)$ ne devient pas aussi élevée, ce qui incite le modèle génératif à diminuer légèrement $q(x)$ dans ces zones pour réduire la divergence.

Ainsi, la divergence KL ne pénalise pas Q pour avoir une densité de probabilité élevée là où P n'en a pas. La divergence KL favorise ainsi le rappel.

Reverse Kullback Leibler. Un raisonnement analogue pour la reverse KL nous mène à la conclusion que celle-ci ne pénalise pas Q pour avoir une densité de probabilité faible là où P en a une élevée. La divergence reverse KL favorise ainsi la précision.

Par ce raisonnement, nous pouvons alors lister les résultats attendus pour chaque divergence et vérifier si les résultats sont conformes aux attentes.

2 Mes Résultats

Les résultats obtenus avec différentes divergences sont présentés dans le tableau ci-dessous :







Nom	Précision	Rappel	Résultats attendus	Visualisation
GAN Vanilla	0,54	0,21	Aucune préférence	
Jensen-Shannon	0,54	0,26	Aucune préférence	
Kullback-Leibler	0,52	0,29	Préférence pour le rappel	
Reverse Kullback-Leibler	0,56	0,21	Préférence pour la précision	
Pearson Chi-Squared	0,52	0,32	Préférence pour le rappel	
Squared Hellinger	0,54	0,24	Aucune préférence	

Table 3: Tableau des résultats obtenus avec différentes divergences. Ces résultats sont obtenus en effectuant 75 epochs sur un GAN classique, puis 5 epochs en entraînant que le discriminateur sur la nouvelle loss, enfin 20 epochs en entraînant le générateur sur la nouvelle loss. Nous observons que les modèles présentant une préférence produisent des résultats conformes aux attentes. En revanche, les modèles sans préférence affichent des scores de précision similaires, accompagnés de variations pour le rappel, oscillant entre 0,21 et 0,26.

3 Limites

3.1 Instabilité

L'entraînement des f-GAN peut souffrir de problèmes d'instabilité.

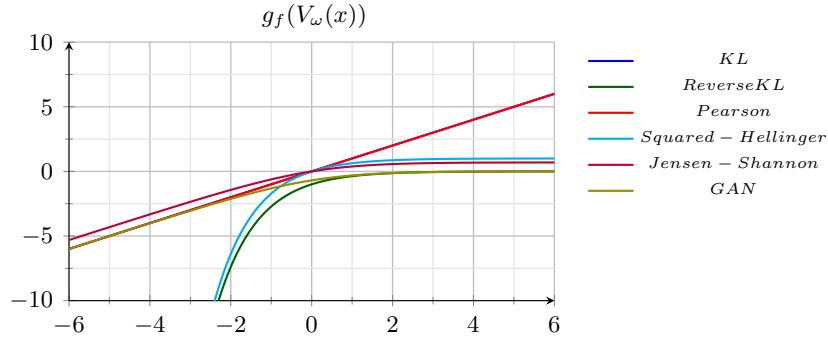


Figure 1: Représentation des différentes fonctions d'activation g_f . Nous constatons que les différentes fonctions d'activation tendent rapidement vers $-\infty$, ce qui peut entraîner des valeurs extrêmes pour la fonction de perte, que ce soit du générateur ou du discriminateur.

Ces problèmes peuvent être réduits en appliquant les stratégies suivantes :

- Limiter les scores du discriminateur sur les images générées en les bornant par le bas, car ceux-ci ont une forte probabilité d'être très faibles.
- Limiter les scores du discriminateur sur les images réelles en les bornant par le haut, car ceux-ci ont une forte probabilité d'être très élevés.
- Pré-entraîner un GAN classique, puis effectuer 20 époques avec la nouvelle fonction de perte.

Conclusion

Dans ce projet, nous avons exploré les f-GAN, une extension des GANs classiques qui utilisent des divergences f pour l'optimisation de la fonction objective. Ces divergences permettent de moduler les critères de performance du modèle génératif, en choisissant des fonctions de perte qui favorisent soit la précision, soit le rappel, en fonction de la divergence choisie. Nous avons examiné plusieurs divergences populaires telles que la divergence de Kullback-Leibler (KL), la divergence Reverse KL, et avons analysé leurs effets sur les performances du générateur et du discriminateur.

Les résultats expérimentaux montrent des comportements conformes aux attentes théoriques des différentes divergences. Par exemple, la divergence de Kullback-Leibler privilégie le rappel, tandis que la divergence Reverse KL favorise la précision. D'autres divergences, comme la Pearson Chi-Square, montrent des résultats similaires avec un avantage pour le rappel.

Cependant, l'entraînement des f-GAN peut être instable, surtout lors de l'optimisation de certaines divergences. Cela peut entraîner des oscillations dans les performances, ce qui nécessite des ajustements dans le processus d'entraînement pour garantir une convergence stable. L'instabilité dans les résultats, particulièrement lors de l'utilisation de divergences complexes, est une limitation importante qu'il convient de surmonter pour rendre les f-GANs plus robustes et applicables à une plus grande variété de tâches.

En conclusion, les f-GANs offrent une flexibilité intéressante pour les modèles génératifs en permettant de choisir la divergence qui correspond le mieux à l'objectif souhaité. Cependant, des améliorations sont nécessaires pour stabiliser l'apprentissage, ce qui pourrait faire des f-GANs une alternative viable aux GANs classiques dans des applications pratiques plus complexes.

Sources

- Sebastian Nowozin, Botond Cseke, Ryota Tomioka. *f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization*. Disponible ici.
- Dibya Ghosh. *KL Divergence for Machine Learning*. Disponible ici.