Hamiltonian Variational Auto-Encoder

Soutenance Statistiques Computationnelles - MVA

Abel Verley et Rémi Baron

Objectifs du papier

VAE :

1. Variable latente : $p_\theta(x) = \int p_\theta(x, z)dz$
2. $\mathcal{L}(\theta, x) \geq \mathcal{L}_{ELBO}(\theta, \phi, x) = \mathbb{E}_{u \sim q_\phi(u|x)}[\log \hat{p}_{\theta,\phi}(x)]$ où $\hat{p}_{\theta,\phi}(x)$ estimateur non biaisé.

Objectifs du papier

VAE :

1. Variable latente : $p_\theta(x) = \int p_\theta(x, z)dz$
2. $\mathcal{L}(\theta, x) \geq \mathcal{L}_{ELBO}(\theta, \phi, x) = \mathbb{E}_{u\sim q_\phi(u|x)}[\log \hat{p}_{\theta,\phi}(x)]$ où $\hat{p}_{\theta,\phi}(x)$ estimateur non biaisé.

Objectif de [Caterini et al., 2018] : Introduire un estimateur non biaisé $\hat{p}_{\theta,\phi}(x)$

1. De faible variance
2. Sujet au ≪ reparametrization trick ≫

$$\begin{aligned} \nabla_\phi \mathbb{E}_{z\sim p_\phi(z)}[f(z)] &= \nabla_\phi \mathbb{E}_{\varepsilon\sim p(\epsilon)}[f(h(\phi, \varepsilon)] \qquad \text{où } z \sim h(\phi, \varepsilon) \\ &= \mathbb{E}_{\varepsilon\sim p(\varepsilon)}[\nabla_\phi f(h(\phi, \varepsilon))] \\ &\approx \nabla_\phi f(h(\phi, \varepsilon)) \end{aligned}$$

Estimation par échantillonnage préférentiel

Pour estimer θ, on peut appliquer les méthodes d'estimation d'espérance :

$$p_\theta(x) = \int p_\theta(x|z)p(z)dz = \mathbb{E}_{z\sim p(z)}[p_\theta(z|x)]$$

Estimation par échantillonnage préférentiel

Pour estimer θ, on peut appliquer les méthodes d'estimation d'espérance :

$$p_\theta(x) = \int p_\theta(x|z)p(z)dz = \mathbb{E}_{z\sim p(z)}[p_\theta(z|x)]$$

Estimation par échantillonnage préférentiel [Burda et al., 2016] : Méthode pour réduire la variance d'estimateur de Monte Carlo

$$\hat{p}_{\theta,\phi}(x) = \frac{1}{L}\sum_{i=1}^{L}\frac{p_\theta(x|z_i)p(z_i)}{q_\phi(z_i|x)} \qquad z_i \sim q_\phi(z|x)$$

Estimation par échantillonnage préférentiel

Pour estimer θ, on peut appliquer les méthodes d'estimation d'espérance :

$$p_\theta(x) = \int p_\theta(x|z)p(z)dz = \mathbb{E}_{z\sim p(z)}[p_\theta(z|x)]$$

Estimation par échantillonnage préférentiel [Burda et al., 2016] : Méthode pour réduire la variance d'estimateur de Monte Carlo

$$\hat{p}_{\theta,\phi}(x) = \frac{1}{L}\sum_{i=1}^{L}\frac{p_\theta(x|z_i)p(z_i)}{q_\phi(z_i|x)} \qquad z_i \sim q_\phi(z|x)$$

Monte Carlo Séquentiel [Del Moral et al., 2006] Variante de l'échantillonnage préférentiel

$$\hat{p}_{\theta,\phi}(x) = \frac{1}{L}\sum_{i=1}^{L}\frac{p_\theta(x|z_K^i)p(z_K^i)\prod_{k=0}^{K-1} r^k(z_k^i|z_{k+1}^i)}{q_\phi^0(z_0^i|x)\prod_{k=1}^{K} q_\phi^k(z_k^i|z_{k-1}^i,x)} \qquad z_0^i,\ldots,z_K^i \sim q_\phi(z_0,\ldots,z_K|x)$$

Le papier utilise $L = 1$ mais nous avons testé l'influence de ce paramètre dans notre implémentation.

Choix des noyaux de transition

Le choix de $q_\phi(z_0, \ldots, z_K)$ correspond à la simulation de dynamique hamiltonienne avec un $\ll$ tempering $\gg$[Wolf et al., 2016]

$$\begin{cases} \frac{dz}{dt} = \nabla_\rho \mathcal{H} \\ \frac{d\rho}{dt} = -\nabla_z \mathcal{H} \\ \mathcal{H}(z, \rho|x) = -\log(p_\theta(x, z)) + \frac{1}{2}\|\rho\|^2 \end{cases}$$

Choix des noyaux de transition

Le choix de $q_\phi(z_0, \dots, z_K)$ correspond à la simulation de dynamique hamiltonienne avec un $\ll$ tempering $\gg$[Wolf et al., 2016]

$$\begin{cases} \frac{dz}{dt} = \nabla_\rho \mathcal{H} \\ \frac{d\rho}{dt} = -\nabla_z \mathcal{H} \\ \mathcal{H}(z, \rho|x) = -\log(p_\theta(x, z)) + \frac{1}{2}\|\rho\|^2 \end{cases}$$

Théorème [Caterini et al., 2018, Del Moral et al., 2006] Lorsque $q_\phi(z_0, \dots, z_K)$ est fixé, le noyau suivant minimise la variance de $\hat{p}_{\theta,\phi}(x)$:

$$r_\phi^{k,opt}(z_k|z_{k+1}, x) = \frac{q_\phi^k(z_k|x) q_\phi^{k+1}(z_{k+1}|z_k, x)}{q_\phi^{k+1}(z_{k+1}|x)}$$

Ce qui donne l'estimateur suivant :

$$\hat{p}_{\theta,\phi}(x) = \frac{p_\theta(x, z_K)}{q_\phi^K(z_K|x)}$$

Estimateur final de [Caterini et al., 2018]

Estimateur proposé [a] par [Caterini et al., 2018] :

$$\hat{p}_{\theta,\phi}(x) = \frac{p_\theta(x, z_K)\mathcal{N}(\rho_K|0, I_d)}{q^0(z_0, \rho_0)\mathcal{N}(\rho_0|0, I_d)} \tag{1}$$

Puisque z_k, ρ_k résulte de l'application d'un difféomorphisme sur z_0, ρ_0, l'estimateur est sujet au reparametrization trick.

a. avec un tempering bien choisi

Partie expérimentale : Modèle gaussien

On considère le modèle génératif suivant :

$$z \sim \mathcal{N}(0, I),$$

$$x_i \mid z \sim \mathcal{N}(z + \Delta, \Sigma) \text{ indépendants, } i \in [1, N].$$

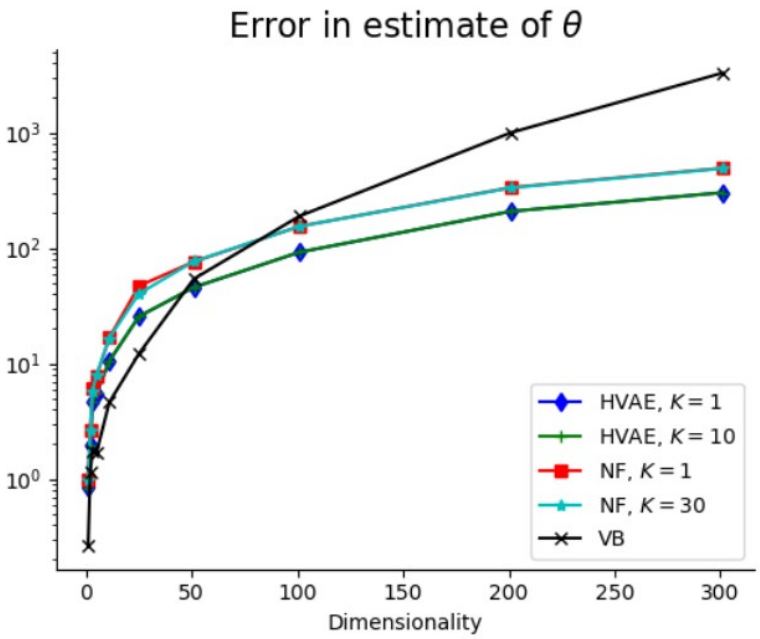
Les paramètres du modèle sont donc

$$\theta \equiv \{\Sigma, \Delta\},$$

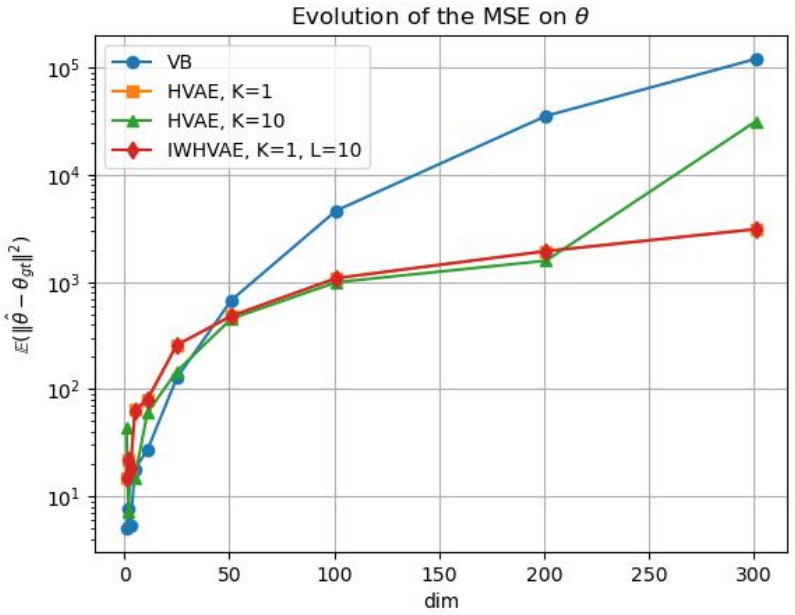
où

$$\Sigma = \mathrm{diag}\left(\sigma_1^2, \ldots, \sigma_d^2\right) \quad \text{et} \quad \Delta \in \mathbb{R}^d.$$

Partie expérimentale : Résultats



(a) Résultats du papier

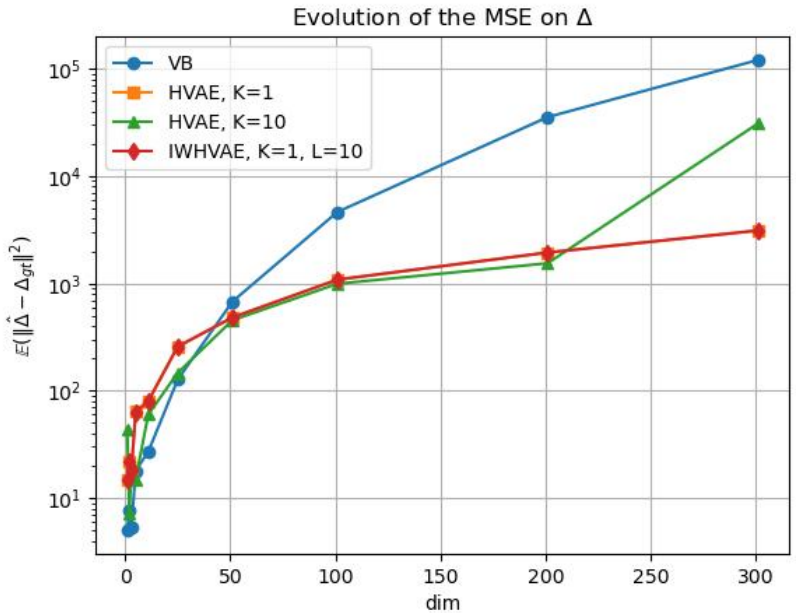


(b) Nos résultats [a]

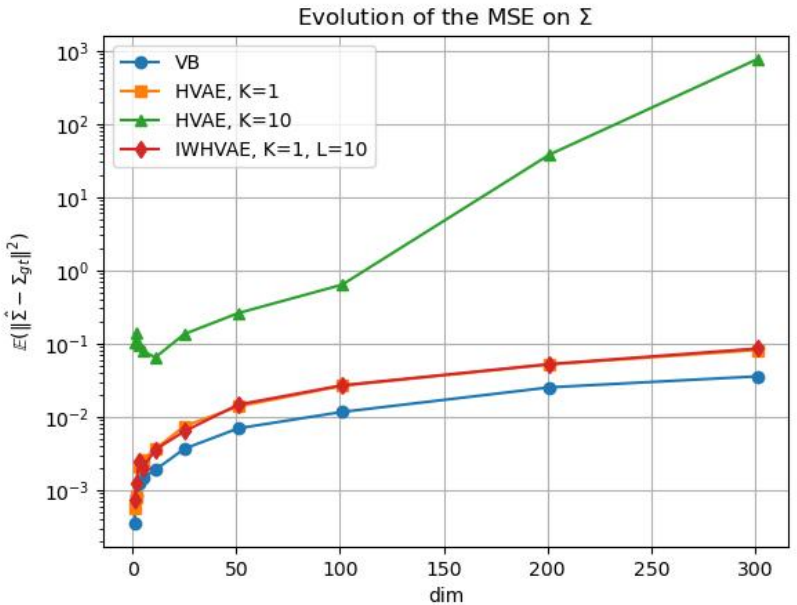
a. https ://github.com/abelmaxv/projet-CompStat

Figure 1 – Moyenne de $\|\theta - \hat{\theta}\|_2^2$ pour plusieurs méthodes variationnelles et différents choix de dimension d, où $\hat{\theta}$ désigne le maximiseur estimé de l'ELBO pour chaque méthode et θ le vrai paramètre.

Partie expérimentale : Critiques concernant Δ et Σ



(a) Moyenne de $||\Delta - \hat{\Delta}||_2^2$



(b) Moyenne de $||\Sigma - \hat{\Sigma}||_2^2$

Merci pour votre attention

Bibliographie

Bibliographie


Burda, Y., Grosse, R., and Salakhutdinov, R. (2016).
Importance Weighted Autoencoders.


Caterini, A. L., Doucet, A., and Sejdinovic, D. (2018).
Hamiltonian Variational Auto-Encoder.


Del Moral, P., Doucet, A., and Jasra, A. (2006).
Sequential monte carlo samplers.
Journal of the Royal Statistical Society Series B : Statistical Methodology, 68(3) :411–436.


Wolf, C., Karl, M., and van der Smagt, P. (2016).
Variational Inference with Hamiltonian Monte Carlo.

Annexe

Paramètres de l'implémentation

1. $n_{test} = 10$ (nombre d'expériences effectuées)
2. $n_{iter} = 30.000$ (nombre d'étapes d'optimisation)
3. $n_{data} = 10.000$ (nombre de données générées)
4. $dimensions = 1; 2; 3; 11; 25; 51; 101; 201; 301$
5. $\Delta_{gt} = \left(-\frac{d-1}{10}, \ldots, \frac{d-1}{10}\right)$
6. $\Sigma_{gt} = Diag(1, \ldots, 0.1, \ldots, 1)$

Integrateur tempéré leapfrog de la dynamique hamiltonienne

Algorithm 1 Tempered leapfrog integration of Hamiltonian dynamics [Caterini et al., 2018]

Require: $0 < \beta_0 < \cdots < \beta_k = 1$

1: Sample $z_0 \sim q_\phi^0(z_0, |x)$
2: Sample $\rho_0 \sim \mathcal{N}(0, \beta_0^{-1} I_d)$
3: **for** $k = 1$ **to** K **do**
4: $\quad \tilde{\rho}_k = \rho_{k-1} - \frac{\varepsilon}{2} \nabla U(z_{k-1}|x)$
5: $\quad z_k = z_{k-1} + \frac{\varepsilon}{2} \tilde{\rho}_k$
6: $\quad \rho'_k = \tilde{\rho}_k - \frac{\varepsilon}{2} \nabla U(z_k|x)$
7: $\quad \rho_k = \sqrt{\frac{\beta_k}{\beta_{k-1}}} \rho'_k$
8: **end for**

$$\sqrt{\beta_k} = \left(\left(1 - \frac{1}{\sqrt{\beta_0}} \right) \frac{k^2}{K^2} + \frac{1}{\sqrt{\beta_0}} \right)^{-1}$$

Preuve noyau inverse optimal

(Preuve dans [Del Moral et al., 2006]) Par la loi de la variance totale :

$$Var(\hat{p}_{\theta,\phi}(x)) = \mathbb{E}[Var(\hat{p}_{\theta,\phi}(x)|z_k)] + Var(\mathbb{E}[\hat{p}_{\theta,\phi}(x)|z_k])$$

D'une part le second terme est indépendant de r_ϕ^k :

$$\mathbb{E}[\hat{p}_{\theta,\phi}(x)|z_k] = \frac{p_\theta(x, z_K)}{q_\phi^K(z_K)}$$

D'autre part, on peut montrer que le premier terme est nul pour le noyau optimal proposé $r^{opt,k}$