階層ベイズモデルによるメタ学習

林 祐輔

2019 年 6 月 10 日

概要

昨年，様々なメタ学習手法を包括的に記述するグラフィカルモデル ML-PIP が提案された．ML-PIP の提案は，異なるタスク間に共通するバイアスをグローバル潜在変数として学習することが，メタ学習の鍵となることを示唆している．この ML-PIP をもとに，効率的に少数ショット分類問題を解くメタ学習モデル VERSA が提案され，Omniglot や Mini-Imagenet といった少数ショット分類問題のベンチマークとなるデータセットについて SOTA を達成した．もっとも，いくつかの先行研究は，推論で使う潜在変数を確率的潜在変数から決定論的潜在変数におきかえても，VERSA とほぼ同等か，それ以上の分類性能を達成できることを示しており，VERSA の確率的近似推論の枠組みが完全ではないことを示唆している．我々は，VERSA の学習過程において Information preference problem（潜在変数を無視して事後分布の尤度を最大化してしまう） が起こることを発見し，この問題を解決するため，本稿では，VERSA の目的関数に相互情報量を加えた新たな目的関数を導入することを提案する．

1 はじめに

変分オートエンコーダー（VAE）[Kingma et al., 2013] の成功を皮切りに，ニューラルネットを用いた確率的近似推論の研究が盛んに行われている．近年では，ニューラル過程（NPs）[Garnelo et al., 2018a ; b] や VERSA [Gordon et al., 2019]，Amortized Bayesian Meta Learning [Ravi et al., 2019] のように，確率的近似推論の枠組みを，メタ学習の問題に適用するアプローチも提案されている．特に，後者に挙げた VERSA や Amortized Bayesian Meta Learning は，異なるドメイン間に共通するバイアスをグローバル潜在変数 θ として学習するモデルになっており，階層的なベイズモデルとしてメタ学習をモデル化している．

本稿では，はじめに VAE とその変分下限（ELBO）に関する議論を紹介し，VAE の ELBO にあらわれる「適切でない正則化項」の存在が，教師なし表現学習において，有害な結果をもたらすことを確認する．そのあと，VERSA にも同様の議論を適用することで，VERSA の潜在変数の学習において，どこに問題があるのかを確認する，そうしたあとに，VERSA にとって，より良い正則化項を与えるフレームワークを提案する．

2 背景

2.1 変分オートエンコーダー（VAE）

変分オートエンコーダー（VAE）は，入力データ x を潜在変数 ψ に変換するエンコーダー $q_\phi(\psi|x)$ と，潜在変数 ψ から入力データ x を推測するデコーダー $p_\varphi(x|\psi)$ によって構成される．VAE は，エンコーダーとデコーダー，それぞれの同時分布

$$q_\phi(x,\psi) = q(x)\, q_\phi(\psi|x)\,. \tag{1}$$
$$p_\varphi(x,\psi) = p(\psi)\, p_\varphi(x|\psi)\,. \tag{2}$$

が一致する方向にモデルのパラメータ $\{\phi,\varphi\}$ を最適化することで，入力データの生成過程をモデリングする＜具体的には，パラメータ ϕ で定まるニューラルネット（エンコーダー）が，入力データを潜在変数に符号化し，パラメータ φ で定まるニューラルネット（デコーダー）が，潜在変数から入力データを復号する＞．

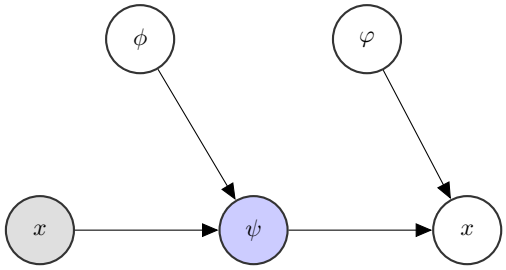

図 1　VAE のグラフィカルモデル

VAE の目的関数 $\mathcal{L}_{\mathrm{VAE-ELBO}}$ は，次のようにして導くことができる．
ELBO の導出法 1 （Jensen の不等式）：

$$\begin{aligned}
\mathbb{E}_{q(x)}\left[\log p_{\varphi}(x)\right] &= \mathbb{E}_{q(x)}\left[\log \int p_{\varphi}(x,\psi)\, d\psi\right] \\
&= \mathbb{E}_{q(x)}\left[\log \int q_{\phi}(\psi|x)\, \frac{p_{\varphi}(x,\psi)}{q_{\phi}(\psi|x)} d\psi\right] \\
&\geq \mathbb{E}_{q(x)}\left[\int q_{\phi}(\psi|x) \log \frac{p_{\varphi}(x,\psi)}{q_{\phi}(\psi|x)} d\psi\right] \\
&= \mathbb{E}_{q_{\phi}(x,\psi)}\left[\log p_{\varphi}(x|\psi)\right] - \mathbb{E}_{q(x)}\left[D_{\mathrm{KL}}\left(q_{\phi}(\psi|x) \,\|\, p(\psi)\right)\right]. \\
&=: \mathcal{L}_{\mathrm{VAE-ELBO}}
\end{aligned} \tag{3}$$

また，Eq. 3 は Jensen の不等式を使わなくても導くことができる．
ELBO の導出法 2 （同時分布の KL-ダイバージェンス）：

$$\begin{aligned}
-D_{\mathrm{KL}}\left(q_{\phi}(x,\psi) \,\|\, p_{\varphi}(x,\psi)\right) &= -\mathbb{E}_{q_{\phi}(x,\psi)}\left[\log \frac{q_{\phi}(x,\psi)}{p_{\varphi}(x,\psi)}\right] \\
&= H_{q_{\phi}}(x|\psi) + \mathbb{E}_{q_{\phi}(x,\psi)}\left[\log p_{\varphi}(x|\psi)\right] - \mathbb{E}_{q(x)}\left[D_{\mathrm{KL}}\left(q_{\phi}(\psi|x) \,\|\, p(\psi)\right)\right]. \\
&= H_{q_{\phi}}(x|\psi) + \mathcal{L}_{\mathrm{VAE-ELBO}}
\end{aligned} \tag{4}$$

ここで，Eq. 4 の第 1 項 $H_{q_{\phi}}(x|\psi)$ は定数になるため，目的関数の勾配には影響しないことに注意する（Appendix 参照）．ELBO の導出法 1, 2 に特別な優劣があるわけではないが，本稿では基本的に導出法 2 を使って ELBO を導くことにする．

さて，$\mathcal{L}_{\mathrm{VAE-ELBO}}$ の内訳をみると，まず第 1 項は，潜在変数 ψ から入力データ x をどれくらい正確に復号できるか，を測る指標（＝対数尤度関数，再構成損失）になっている．また，第 2 項は，潜在変数の学習に制限をかけるための正則化項だということがわかる．様々な先行研究 [Bowman et al., 2016, Snderby et al., 2016] から，潜在変数の正則化項 $D_{\mathrm{KL}}\left(q_{\phi}(\psi|x) \,\|\, p(\psi)\right) \geq 0$ は，入力データを無視した潜在変数の学習を促す「有害な正則化項」であると指摘されている．実際，この正則化項を相互情報量をつかって次のように書き直すと

$$\begin{aligned}
\mathcal{L}_{\mathrm{VAE-ELBO}} &= \mathbb{E}_{q_{\phi}(x,\psi)}\left[\log p_{\varphi}(x|\psi)\right] - \mathbb{E}_{q(x)}\left[D_{\mathrm{KL}}\left(q_{\phi}(\psi|x) \,\|\, p(\psi)\right)\right] \\
&= \mathbb{E}_{q_{\phi}(x,\psi)}\left[\log p_{\varphi}(x|\psi)\right] - I_{q_{\phi}}(x,\psi) - D_{\mathrm{KL}}\left(q_{\phi}(\psi) \,\|\, p(\psi)\right).
\end{aligned} \tag{5}$$

Eq. 5 より，$\mathcal{L}_{\mathrm{VAE-ELBO}}$ の最大化は，入力データと潜在変数の相互情報量 $I_{q_{\phi}}(x,\psi) \geq 0$（≒相関）を下げる方向に進むことがわかる（これは，$D_{\mathrm{KL}}\left(q_{\phi}(\psi|x) \,\|\, p(\psi)\right) = 0$ に近づくことで，$q_{\phi}(\psi|x) = p(\psi)$ が導かれ，エンコーダーが入力データを無視して潜在変数を生成するように学習が進むことを意味する）．

このような問題を解決するため，$\mathcal{L}_{\mathrm{VAE-ELBO}}$ に相互情報量を加えた新しい目的関数を導入する [Zhao et al., 2018, Zhao et al., 2018].

$$\begin{aligned}\mathcal{L}_{\mathrm{InfoVAE}} &=: \mathcal{L}_{\mathrm{VAE-ELBO}} + I_{q_\phi}(x, \psi) \\ &= \mathbb{E}_{q_\phi(x,\psi)}\left[\log p_\varphi(x|\psi)\right] - D_{\mathrm{KL}}\left(q_\phi(\psi) \| p(\psi)\right). \end{aligned} \tag{6}$$

新しい目的関数を Eq. 6 のように設定してしまえば，学習が進んでも，潜在変数が入力データを無視して生成されることはなくなる．このように，潜在変数にかかる正則化項を新しい角度から見つめ直すことで，様々な VAE の派生モデルが提案されている．

例えば，最近では，潜在変数の各成分 $\psi = (\psi_1, \psi_2, \ldots, \psi_D)$ 毎の分布 $q_\phi(\psi_1), q_\phi(\psi_2), \ldots, q_\phi(\psi_D)$ を考え

$$\begin{aligned}\mathcal{L}_{\mathrm{InfoVAE}} &= \mathbb{E}_{q_\phi(x,\psi)}\left[\log p_\varphi(x|\psi)\right] - D_{\mathrm{KL}}\left(q_\phi(\psi) \| p(\psi)\right) \\ &= \mathbb{E}_{q_\phi(x,\psi)}\left[\log p_\varphi(x|\psi)\right] - D_{\mathrm{KL}}\left(q_\phi(\psi) \| \prod_{i=1}^{D} q_\phi(\psi_i)\right) - \sum_{i=1}^{D} D_{\mathrm{KL}}\left(q_\phi(\psi_i) \| p(\psi_i)\right). \end{aligned} \tag{7}$$

のように，Eq. 6 の正則化項をも分解するアプローチが提案されている．ここであらわれる Eq. 7 の第 2 項は Total correlation と呼ばれ，成分間 $\psi = (\psi_1, \psi_2, \ldots, \psi_D)$ の相関の強さをあらわす．Total correlation を下げる方向に学習を進めることで，潜在空間の各次元の意味をより解釈しやすいものにすることができる [Gao et al., 2019].

2.2 VERSA

次に，ここまでに述べた確率的近似推論の枠組みを，メタ学習に適用するアプローチを紹介する．メタ学習とは，複数タスクの学習をとおして，統計モデルに「学習の学習（Learn to learn）」をさせることを目的としたアプローチである．通常，メタ学習の枠組みでは，訓練データ・セット $D^{(t)} = \left\{\left(x_n^{(t)}, y_n^{(t)}\right)\right\}_{n=1}^{N_t} =: \left(x^{(t)}, y^{(t)}\right)$ と，検証データ・セット $\tilde{D}^{(t)} = \left\{\left(\tilde{x}_m^{(t)}, \tilde{y}_m^{(t)}\right)\right\}_{m=1}^{M_t} =: \left(\tilde{x}^{(t)}, \tilde{y}^{(t)}\right)$ を，複数タスクについて用意する．ここで，タスク毎の区別は添字の t によってあらわすとする．このとき，メタ学習モデル ML-PIP は次の式で与えられる．

$$\begin{aligned} p\left(\left\{y^{(t)}, \tilde{y}^{(t)}, \psi^{(t)}\right\}_{t=1}^{T} | \left\{x^{(t)}, \tilde{x}^{(t)}\right\}_{t=1}^{T}, \theta\right) &= \prod_{t=1}^{T} p\left(\psi^{(t)}|\theta\right) \prod_{n=1}^{N_t} p\left(y_n^{(t)}|x_n^{(t)}, \psi^{(t)}, \theta\right) \prod_{m=1}^{M_t} p\left(\tilde{y}_m^{(t)}|\tilde{x}_m^{(t)}, \psi^{(t)}, \theta\right) \\ &= \prod_{t=1}^{T} p\left(\psi^{(t)}|\theta\right) p\left(y^{(t)}|x^{(t)}, \psi^{(t)}, \theta\right) p\left(\tilde{y}^{(t)}|\tilde{x}^{(t)}, \psi^{(t)}, \theta\right). \end{aligned} \tag{8}$$

Eq. 8 から，ある特定のタスク t に関する部分のみを取り出すと，次のようになる．

$$p\left(y^{(t)}, \tilde{y}^{(t)}, \psi^{(t)}|x^{(t)}, \tilde{x}^{(t)}, \theta\right) = p\left(\psi^{(t)}|\theta\right) p\left(y^{(t)}|x^{(t)}, \psi^{(t)}, \theta\right) p\left(\tilde{y}^{(t)}|\tilde{x}^{(t)}, \psi^{(t)}, \theta\right). \tag{9}$$

これをグラフィカルモデルであらわすと，Fig. 2 となる．

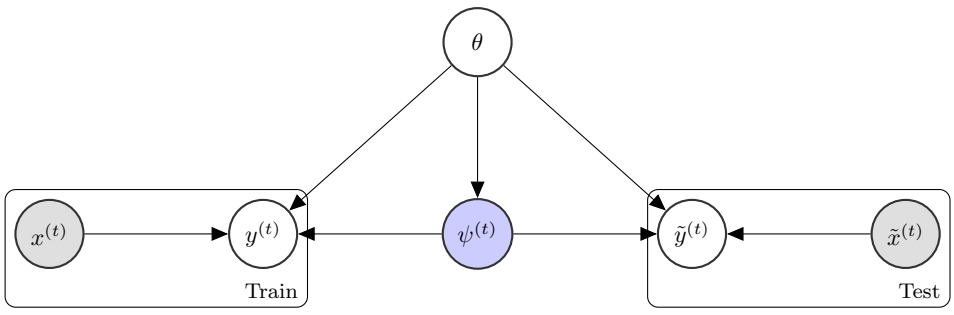

図 2　ML-PIP のグラフィカルモデル

ここで，ML-PIP を使った，$\tilde{x}^{(t)}$ および $D^{(t)}$ を所与のデータとしたときの $\tilde{y}^{(t)}$ の予測モデルを考える．まず，Eq. 9 と Fig. 2 のグラフィカルモデルより

$$\begin{aligned} p\left(y^{(t)}, \tilde{y}^{(t)}, \psi^{(t)}, x^{(t)}, \tilde{x}^{(t)} | \theta\right) &= p\left(\tilde{y}^{(t)} | \tilde{x}^{(t)}, \psi^{(t)}, D^{(t)}, \theta\right) p\left(\tilde{x}^{(t)}, \psi^{(t)}, D^{(t)} | \theta\right) \\ &= p\left(\tilde{y}^{(t)} | \tilde{x}^{(t)}, \psi^{(t)}, \theta\right) p\left(\psi^{(t)} | \tilde{x}^{(t)}, D^{(t)}, \theta\right) p\left(\tilde{x}^{(t)}, D^{(t)}\right). \end{aligned} \tag{10}$$

が導かれる．これが本稿が考察の対象とするメタ学習モデルである．

$$p\left(\tilde{y}^{(t)}, \psi^{(t)} | \tilde{x}^{(t)}, D^{(t)}, \theta\right) = p\left(\tilde{y}^{(t)} | \tilde{x}^{(t)}, \psi^{(t)}, \theta\right) p\left(\psi^{(t)} | \tilde{x}^{(t)}, D^{(t)}, \theta\right). \tag{11}$$

もっとも，潜在変数 ψ の条件付き分布 $p\left(\psi^{(t)} | \tilde{x}^{(t)}, D^{(t)}, \theta\right)$ は，実際には計算が難しいため，次式で与えるようにこれを $q_\phi\left(\psi^{(t)} | D^{(t)}, \theta\right)$ で近似したモデル（VERSA と呼ばれる）を考える．

$$q_\phi\left(\tilde{y}^{(t)}, \psi^{(t)} | \tilde{x}^{(t)}, D^{(t)}, \theta\right) = q_\phi\left(\tilde{y}^{(t)} | \tilde{x}^{(t)}, \psi^{(t)}, \theta\right) q_\phi\left(\psi^{(t)} | D^{(t)}, \theta\right). \tag{12}$$

この VERSA をグラフィカルモデルであらわすと，次の Fig. 3 のようになる．ただし，図中の $h_\theta\left(D^{(t)}\right)$ および $h_\theta\left(\tilde{x}^{(t)}\right)$ は，パラメータ θ で定まるエンコーダーによって決定論的に生成された特徴量をあらわしている．

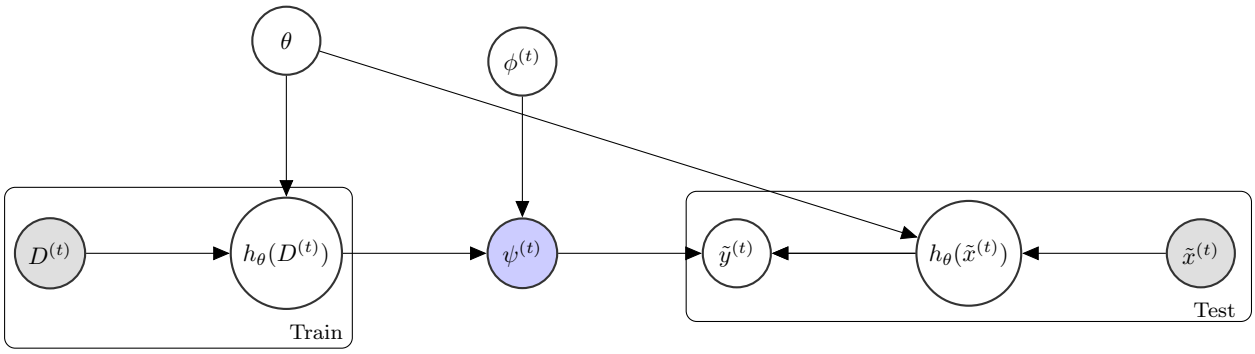

図 3　VERSA のグラフィカルモデル

なお，VERSA で確率変数として想定されているものだけを使って，Fig. 3 を描き直すと次のようになる．

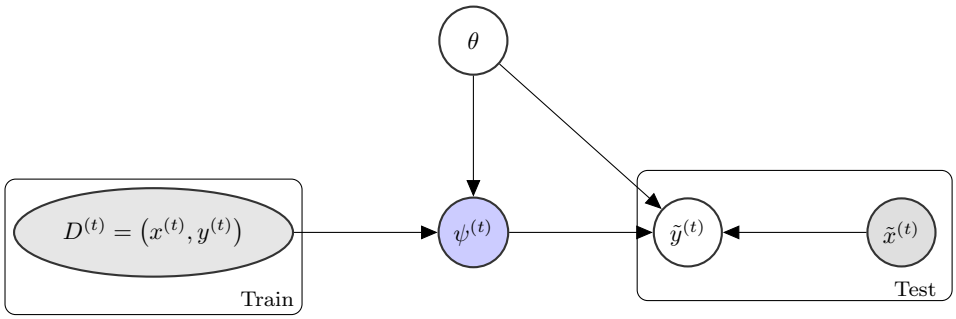

図 4　VERSA のグラフィカルモデル

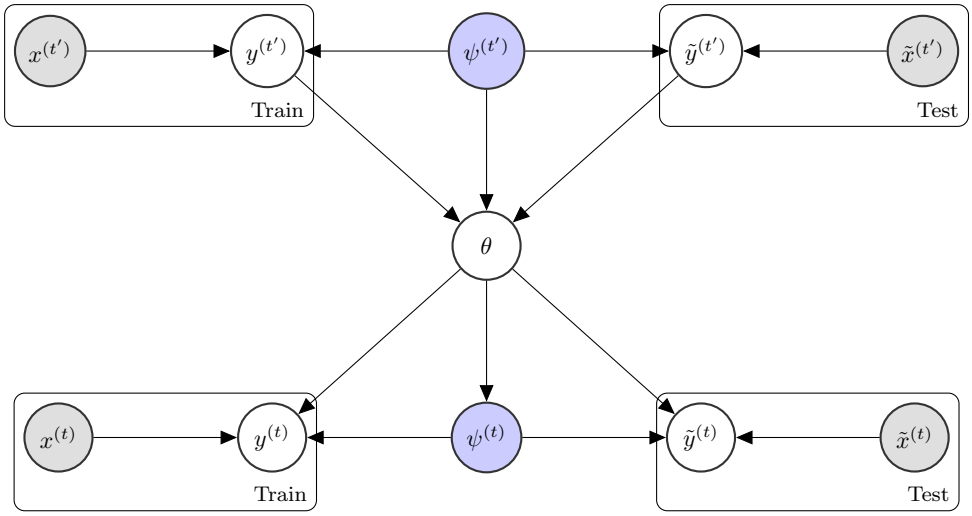

図 5　ML-PIP のグラフィカルモデル

さて，Gordon et al., 2019 の設定では，VERSA の目的関数は次のようになっていた．

$$\begin{aligned}\mathbb{E}_{p(\tilde{x},D|\theta)}\left[D_{\mathrm{KL}}\left(p\left(\tilde{y}|\tilde{x},D,\theta\right)\|q_\phi\left(\tilde{y}|\tilde{x},D,\theta\right)\right)\right] &= \mathbb{E}_{p(\tilde{y},\tilde{x},D|\theta)}\left[\log r(\tilde{y}|\tilde{x},D,\theta)\right] \\ &= -\mathbb{E}_{p(\tilde{x},D|\theta)}\left[H_p\left(\tilde{y}|\tilde{x},D,\theta\right)\right] - \mathbb{E}_{p(\tilde{y},\tilde{x},D|\theta)}\left[\log q_\phi(\tilde{y}|\tilde{x},D,\theta)\right].\end{aligned} \tag{13}$$

Eq. 13 の第 1 項は定数になる（Appendix 参照）ことを踏まえると，第 2 項の最小化が学習の目標となることがわかる．Eq. 13 の第 2 項は，実際には次のように計算される．

$$\begin{aligned}\mathcal{L}_{\mathrm{VERSA}} &=: -\mathbb{E}_{p(\tilde{y},\tilde{x},D|\theta)}\left[\log\, q_\phi(\tilde{y}|\tilde{x},D,\theta)\right] \\ &= -\mathbb{E}_{p(\tilde{y},\tilde{x},D|\theta)}\left[\log \int q_\phi(\tilde{y},\psi|\tilde{x},D,\theta)d\psi\right] \\ &= -\mathbb{E}_{p(\tilde{y},\tilde{x},D|\theta)}\left[\log \int p(\tilde{y}|\tilde{x},\psi,\theta)\; q_\phi(\psi|D,\theta)d\psi\right] \\ &\sim -\frac{1}{MT}\sum_m^M\sum_t^T \log\left(\frac{1}{L}\sum_l^L\; p\left(\tilde{y}_m^{(t)}|\tilde{x}_m^{(t)},\psi_l^{(t)},\theta\right)\right).\end{aligned} \tag{14}$$

我々は Omniglot, Mini-Imagenet を使った実験によって，学習を通じて，潜在変数 ψ の分散 σ がゼロに近づいていき，次に近いことが起こっていることを確認した．

$$\begin{aligned}q_\phi\left(\psi|D,\theta\right) &= \delta\left(\psi - h_\theta\left(D\right)\right). \\ \lim_{\sigma\to 0}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\left(\psi - h_\theta\left(D\right)\right)^2}{2\sigma^2}\right) &= \delta(\psi - h_\theta\left(D\right)).\end{aligned} \tag{15}$$

3　モデル

VERSA-ELBO の導入，および，潜在変数やグローバル潜在変数についての正則化の議論．

4　実験結果

Mnist，Omniglot

5　結論と考察

まとめ

参考文献

[1] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. In International Conference on Learning Representations (ICLR), 2014.

[2] Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and SM Ali Eslami. Conditional neural processes. In International Conference on Machine Learning (ICML), 2018a.

[3] Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J Rezende, SM Eslami, and Yee Whye Teh. Neural processes. In International Conference on Machine Learning (ICML) Workshop on Theoretical Foundations and Applications of Deep Generative Models, 2018b.

[4] Jonathan Gordon, John Bronskill, Matthias Bauer, Sebastian Nowozin, Richard E. Turner. Meta-Learning Probabilistic Inference for Prediction. In International Conference on Learning Representations (ICLR), 2019.

[5] Sachin Ravi and Alex Beatson. Amortized Bayesian Meta-Learning. In International Conference on Learning Representa-tions (ICLR), 2019

[6] Shengjia Zhao, Jiaming Song, Stefano Ermon. InfoVAE: Information Maximizing Variational Autoencoders.

[7] Tiancheng Zhao, Kyusong Lee, Maxine Eskenazi. Unsupervised Discrete Sentence Representation Learning for Interpretable Neural Dialog Generation. In proceedings of Association for Computational Linguistics (ACL), 2018.

[8] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Jozefowicz, Samy Bengio. Generating sentences from a continuous space. In Proceedings of Conference on Computational Natural Language Learning (CoNLL), 2016.

[9] Casper Kaae Snderby, Tapani Raiko, Lars Maale, Sren Kaae Snderby, and Ole Winther. How to train deep variational autoencoders and probabilistic ladder networks. In Proceedings of the 33rd International Conference on Machine Learning (ICML), 2016.

[10] Shuyang Gao, Rob Brekelmans, Greg Ver Steeg, Aram Galstyan. Auto-Encoding Total Correlation Explanation. In International Conference on Artificial Intelligence and Statistics (AISTATS), 2019.

6 Appendix

シャノン・エントロピー：

$$H_p(x) =: -\mathbb{E}_{p(x)}[\log p(x)] = -\int p(x)\log p(x)dx \tag{16}$$

$$H_{q_\phi}(x) =: -\mathbb{E}_{q_\phi(x)}[\log q_\phi(x)] = -\int q_\phi(x)\log q_\phi(x)dx \tag{17}$$

相互情報量：

$$I_p(x,y) =: \int p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy \tag{18}$$

$$I_{q_\phi}(x,y) =: \int q_\phi(x,y)\log\frac{q_\phi(x,y)}{q_\phi(x)q_\phi(y)}dxdy \tag{19}$$