林祐輔（所属: Japan Digital Design, 連絡先: bheprr@gmail.com, Twitter: @hayashiyus）

表現学習の確率熱力学的解釈（1/6）

機械学習と物理学の接点（理論の形式的類似性）を紹介します．その応用例として，物理学側のツールである「準静的過程に沿った状態操作」を機械学習に導入します．

- 確率熱力学：ゆらぐ系の熱力学．微小系が状態 X から状態 Z に遷移する過程（順過程と呼ぶ）の確率を p とし，その逆過程の確率を q とするとき，p と q との距離が，微小系の状態変化に伴うエントロピー生成量 σ_{tot} を決定する．

$$\mathcal{D}\left(p(s)||q(s)\right) = \mathcal{D}\left(p(s)||q^*(s)\right) + \mathcal{D}\left(q^*(s)||q(s)\right).$$

$$\sigma^{\mathcal{S}}_{\mathrm{tot}} := \sigma^{\mathcal{S}}_{\mathrm{sys}} + \sigma^{\mathcal{S}}_{\mathrm{bath}}.$$

特徴量の生成プロセスの学習

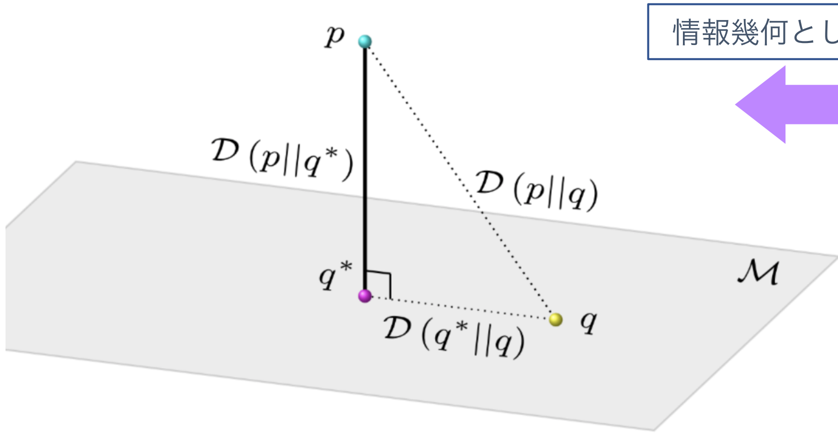


Figure 1: Information-Geometric Pythagorean theorem.

情報幾何として共通の幾何構造

ゆらぐ系のエントロピー生成

entropy production (2nd law)

$\sigma^{\mathcal{S}}_{\mathrm{tot}}\ (\geq 0)$

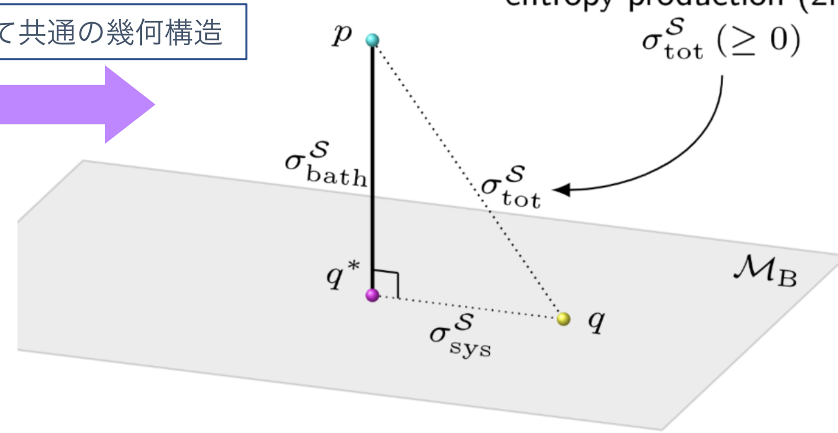


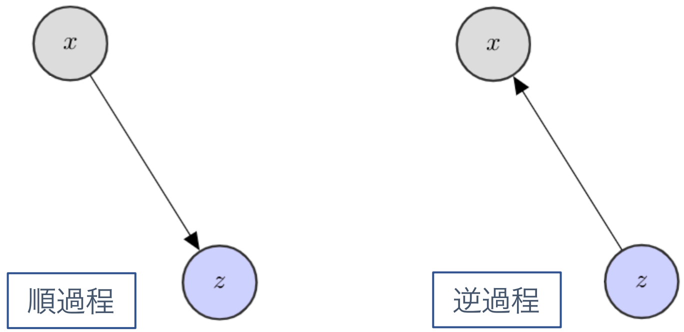
Figure 2: Total entropy production.

参考文献[1]: Ito, Oizumi and Amari (2020) arXiv:1810.09545v5 **[cond-mat.stat-mech]**

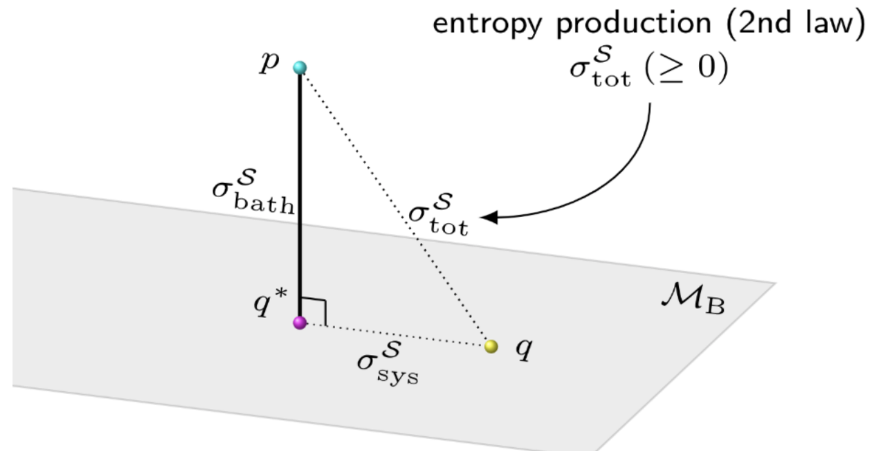
林祐輔（所属: Japan Digital Design, 連絡先: bheprr@gmail.com, Twitter: @hayashiyus）

表現学習の確率熱力学的解釈（2/6）

- 全系のエントロピー生成：順過程と逆過程とで，グラフィカルモデルの全要素について矢印の向きが逆になる．

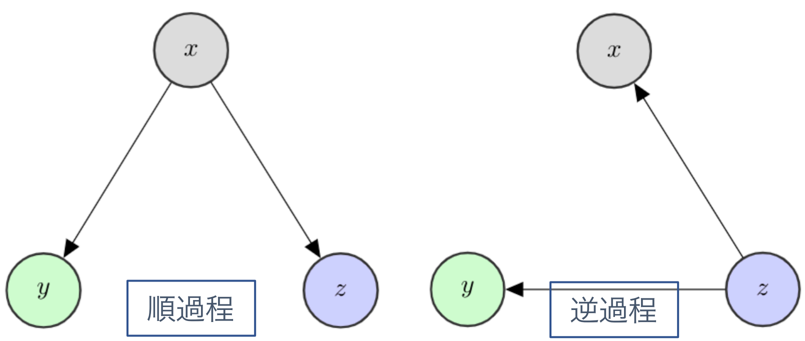


(a) (b)

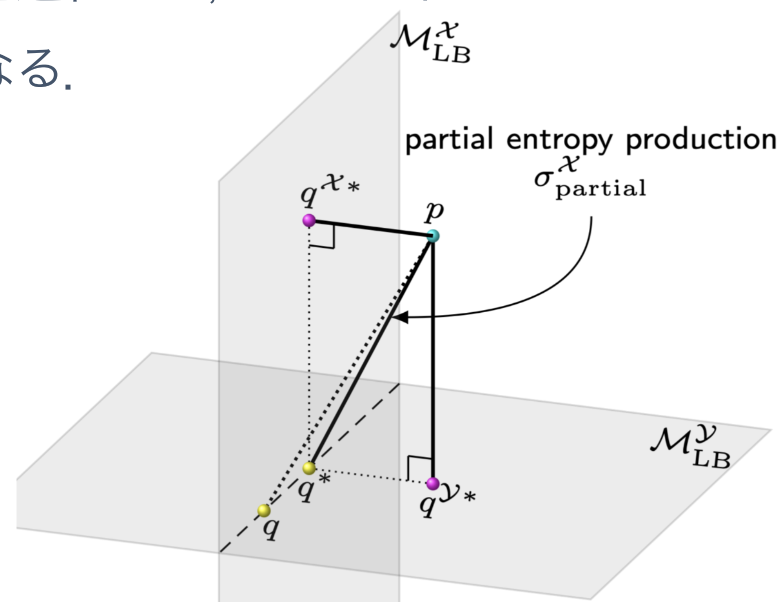


$p(\boldsymbol{s}) := T\left(\boldsymbol{z},\boldsymbol{x}\right)p(\boldsymbol{x}) = p\left(\boldsymbol{z}|\boldsymbol{x}\right)p(\boldsymbol{x})$. $q(\boldsymbol{s}) := T\left(\boldsymbol{x},\boldsymbol{z}\right)q(\boldsymbol{z}) = q\left(\boldsymbol{x}|\boldsymbol{z}\right)q(\boldsymbol{z})$ $\mathcal{M}_{\mathrm{B}} = \left\{q(\boldsymbol{s})\mid q(\boldsymbol{s}) := T\left(\boldsymbol{x},\boldsymbol{z}\right)q(\boldsymbol{z}) = q\left(\boldsymbol{x}|\boldsymbol{z}\right)q(\boldsymbol{z})\right\}$.

- 部分系のエントロピー生成：順過程と逆過程とで，グラフィカルモデルの一部の要素について矢印の向きが逆になる．



(a) (b)



$p(\boldsymbol{s}) := p(\boldsymbol{z}|\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})$ $q(\boldsymbol{s}) := q(\boldsymbol{x}|\boldsymbol{z})q\left(\boldsymbol{y}|\boldsymbol{z}\right)q(\boldsymbol{z})$

林祐輔（所属: Japan Digital Design，連絡先: bheprr@gmail.com，Twitter: @hayashiyus）

表現学習の確率熱力学的解釈（3/6）

- 表現学習の学習プロセス：データ $\boldsymbol{x}$ から潜在変数（特徴量）$\boldsymbol{z}$ の生成と教師ラベル $\boldsymbol{y}$ の予測を行う順過程の確率分布 p と，潜在変数（特徴量）$\boldsymbol{z}$ からデータ $\boldsymbol{x}$ と教師ラベル $\boldsymbol{y}$ をデコードする逆過程の確率分布 q の距離を可能な限り近づけようとするプロセス．部分系のエントロピー生成と解釈可能．

$$H := \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim p(\boldsymbol{x},\boldsymbol{y})}\left[-\log p(\boldsymbol{x},\boldsymbol{y})\right],$$

データのエントロピー

$$D := \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[-\int e(\boldsymbol{z}|\boldsymbol{x})\log d(\boldsymbol{x}|\boldsymbol{z})d\boldsymbol{z}\right],$$

再構成誤差

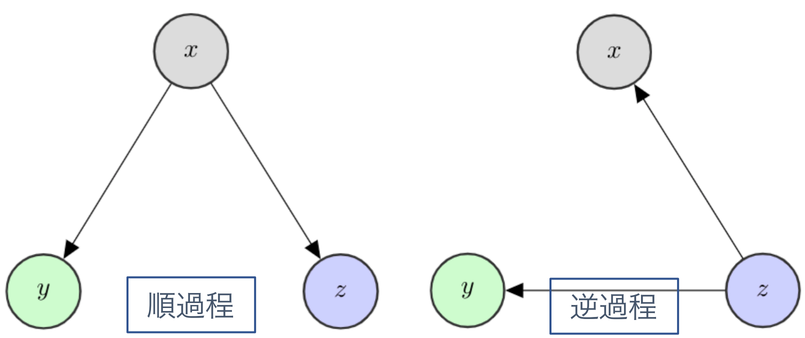
$$R := \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[\int e(\boldsymbol{z}|\boldsymbol{x})\log\frac{e(\boldsymbol{z}|\boldsymbol{x})}{q(\boldsymbol{z})}d\boldsymbol{z}\right],$$

正則化項

$$C := \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[-\int e(\boldsymbol{z}|\boldsymbol{x})\log c(\boldsymbol{y}|\boldsymbol{z})d\boldsymbol{z}\right],$$

分類誤差

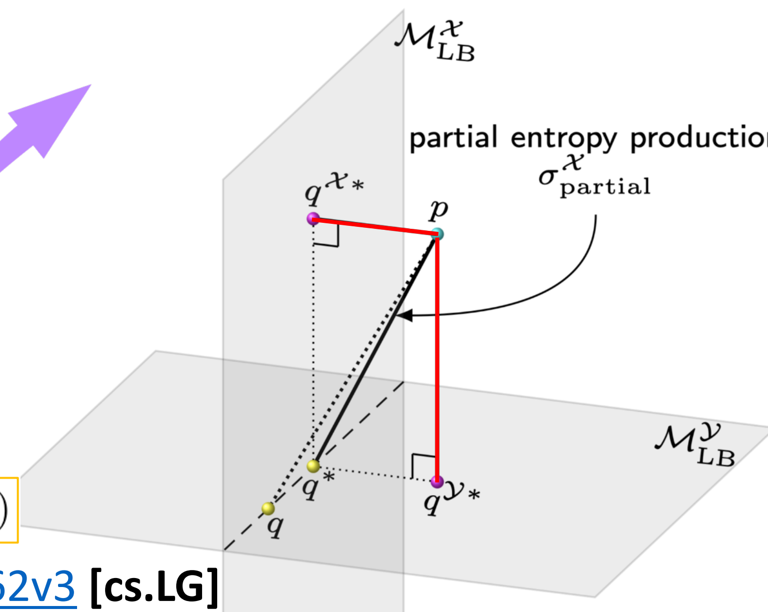
$$\begin{aligned}\mathcal{D}\left(p(\boldsymbol{s})||q(\boldsymbol{s})\right) &= -H + R + \lambda D + \gamma C \geq 0\\ &:= -H + \mathcal{J}(\lambda,\gamma) \geq 0\end{aligned}$$



(a) (b)

$$p(\boldsymbol{s}) := e(\boldsymbol{z}|\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})$$

$$q(\boldsymbol{s}) := d_\theta(\boldsymbol{x}|\boldsymbol{z})^\lambda c_\theta\left(\boldsymbol{y}_{\boldsymbol{x}}|\boldsymbol{z}\right)^\gamma q_\theta(\boldsymbol{z})$$



参考文献[2]: Alemi and Fischer (2018) arXiv:1807.04162v3 **[cs.LG]**

林祐輔（所属: Japan Digital Design, 連絡先: bheprr@gmail.com, Twitter: @hayashiyus）

表現学習の確率熱力学的解釈（4/6）

- 深層生成モデルを使った表現学習：エンコーダー，デコーダーとして深層学習モデル（例えばCNN）を使い，そのパラメータを $\boldsymbol{\theta}$ とする．

$$H := \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim p(\boldsymbol{x},\boldsymbol{y})}\left[-\log p(\boldsymbol{x},\boldsymbol{y})\right],$$

データのエントロピー

$$D := \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[-\int e(\boldsymbol{z}|\boldsymbol{x})\log d(\boldsymbol{x}|\boldsymbol{z})d\boldsymbol{z}\right],$$

再構成誤差

$$R := \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[\int e(\boldsymbol{z}|\boldsymbol{x})\log \frac{e(\boldsymbol{z}|\boldsymbol{x})}{q(\boldsymbol{z})}d\boldsymbol{z}\right],$$

正則化項

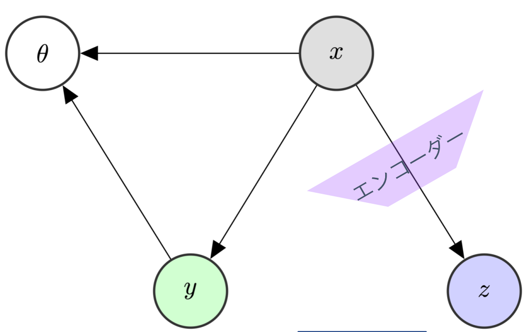
$$C := \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[-\int e(\boldsymbol{z}|\boldsymbol{x})\log c(\boldsymbol{y}|\boldsymbol{z})d\boldsymbol{z}\right],$$

分類誤差

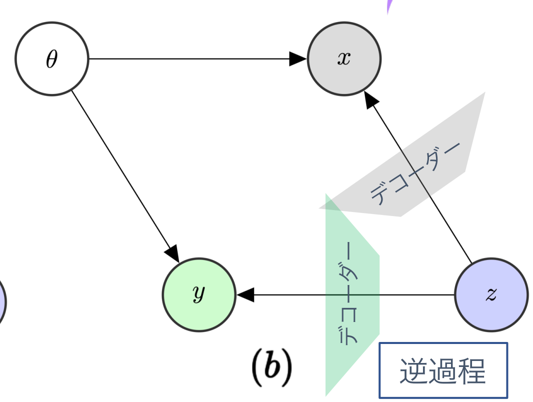
$$S := \mathbb{E}_{\theta,\boldsymbol{x},\boldsymbol{y}\sim p(\theta,\boldsymbol{x},\boldsymbol{y})}\left[\log \frac{p(\theta|\{\boldsymbol{x},\boldsymbol{y}\})}{q(\theta)}\right]$$

パラメータの相対エントロピー

$$\begin{aligned}\mathcal{D}\left(p(\boldsymbol{s})||q(\boldsymbol{s})\right) &= -H + R + \lambda D + \gamma C + \sigma S \geq 0\\ &:= -H + \mathcal{J}\left(\lambda,\gamma,\sigma\right) \geq 0\end{aligned}$$



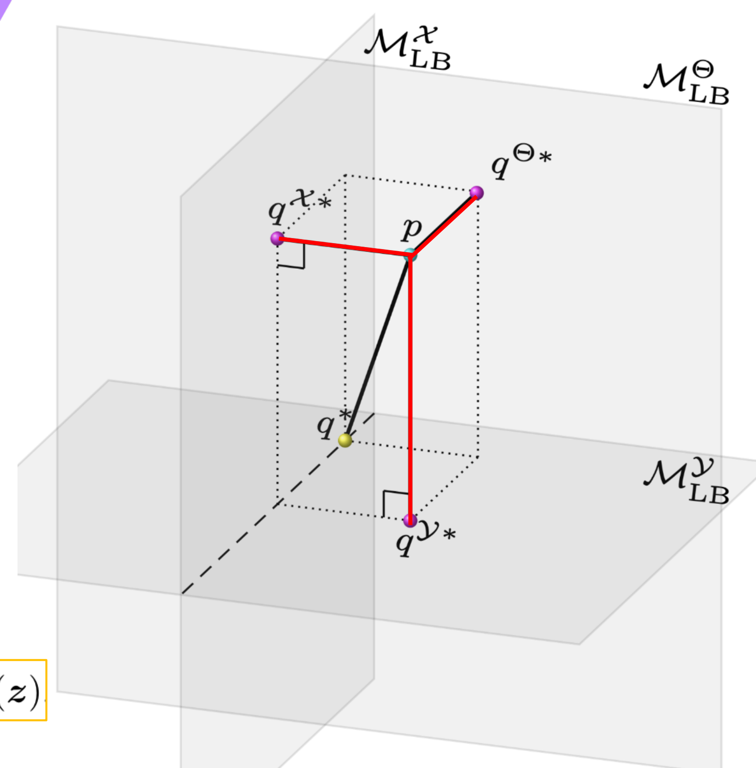
(a) 順過程



(b) 逆過程

$$p(\boldsymbol{s}) := p(\theta|\{\boldsymbol{x},\boldsymbol{y}\})^{\sigma}e(\boldsymbol{z}|\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})$$

$$q(\boldsymbol{s}) := q(\theta)^{\sigma}d_{\theta}(\boldsymbol{x}|\boldsymbol{z})^{\lambda}c_{\theta}\left(\boldsymbol{y}_{\boldsymbol{x}}|\boldsymbol{z}\right)^{\gamma}q_{\theta}(\boldsymbol{z})$$



林祐輔（所属: Japan Digital Design, 連絡先: bheprr@gmail.com, Twitter: @hayashiyus）

表現学習の確率熱力学的解釈（5/6）

- 表現学習と熱力学諸法則の形式的対応

$$p(\boldsymbol{s}) \;:=\; p(\theta|\{\boldsymbol{x},\boldsymbol{y}\})^{\sigma} e(\boldsymbol{z}|\boldsymbol{x}) p(\boldsymbol{y}|\boldsymbol{x}) p(\boldsymbol{x})$$

$$q(\boldsymbol{s}) \;:=\; q(\theta)^{\sigma} d_{\theta}(\boldsymbol{x}|\boldsymbol{z})^{\lambda} c_{\theta}\left(\boldsymbol{y}_{\boldsymbol{x}}|\boldsymbol{z}\right)^{\gamma} q_{\theta}(\boldsymbol{z})$$

$$H := \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim p(\boldsymbol{x},\boldsymbol{y})}\left[-\log p(\boldsymbol{x},\boldsymbol{y})\right],$$

データのエントロピー

$$D := \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[-\int e(\boldsymbol{z}|\boldsymbol{x}) \log d(\boldsymbol{x}|\boldsymbol{z}) d\boldsymbol{z}\right],$$

再構成誤差

$$R := \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[\int e(\boldsymbol{z}|\boldsymbol{x}) \log \frac{e(\boldsymbol{z}|\boldsymbol{x})}{q(\boldsymbol{z})} d\boldsymbol{z}\right],$$

正則化項

$$C := \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[-\int e(\boldsymbol{z}|\boldsymbol{x}) \log c(\boldsymbol{y}|\boldsymbol{z}) d\boldsymbol{z}\right],$$

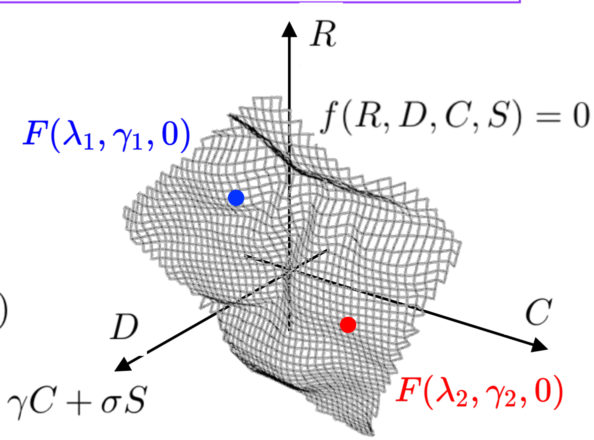
分類誤差

$$S := \mathbb{E}_{\theta,\boldsymbol{x},\boldsymbol{y}\sim p(\theta,\boldsymbol{x},\boldsymbol{y})}\left[\log \frac{p(\theta|\{\boldsymbol{x},\boldsymbol{y}\})}{q(\theta)}\right]$$

パラメータの
相対エントロピー

$$\begin{aligned}\mathcal{D}\left(p(\boldsymbol{s})||q(\boldsymbol{s})\right) &= -H + R + \lambda D + \gamma C + \sigma S \geq 0 \\ &:= -H + \mathcal{J}\left(\lambda,\gamma,\sigma\right) \geq 0\end{aligned}$$

$$\mathcal{J}(\lambda,\gamma,\sigma) \geq H$$



$$\begin{aligned}F(\lambda,\gamma,\sigma) &:= \min_{e(z|x),q(z),d(x|z),c(y|z)} \mathcal{J}\left(\lambda,\gamma,\sigma\right) \\ &= \min_{e(z|x),q(z),d(x|z),c(y|z)} R + \lambda D + \gamma C + \sigma S\end{aligned}$$

Remark 1 ("the first law" of learning).

$$\mathrm{d}R = -\lambda \mathrm{d}D - \gamma \mathrm{d}C - \sigma \mathrm{d}S$$

Remark 2 ("the second law" of learning).

$$\lambda \equiv -\left(\frac{\partial R}{\partial C}\right)_{D,S} \quad \gamma \equiv -\left(\frac{\partial R}{\partial D}\right)_{C,S} \quad \sigma \equiv -\left(\frac{\partial R}{\partial S}\right)_{C,D}$$

$$\mathcal{D}\left(p(\boldsymbol{s})||q(\boldsymbol{s})\right) = -H + R + \lambda D + \gamma C + \sigma S := -H + \mathcal{J}\left(\lambda,\gamma,\sigma\right) \geq 0$$

参考文献[3]: Gao and Chaudhari (2020) arXiv:2002.12406v1 **[cs.LG]**

林祐輔（所属: Japan Digital Design, 連絡先: bheprr@gmail.com, Twitter: @hayashiyus）

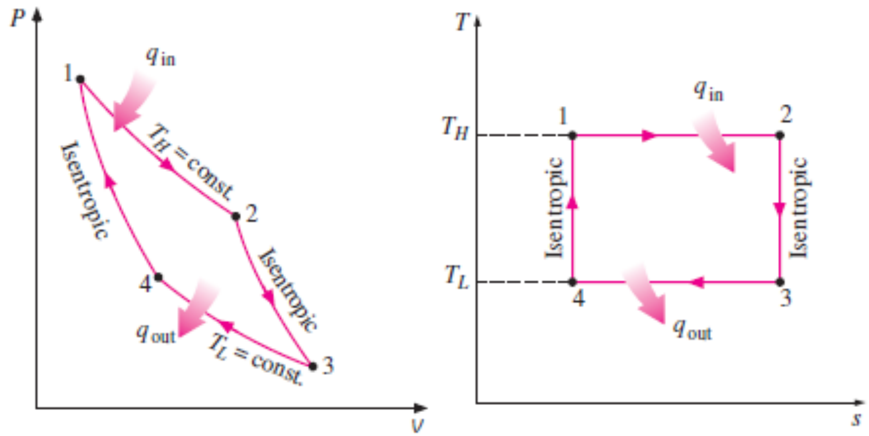
表現学習の確率熱力学的解釈（6/6）

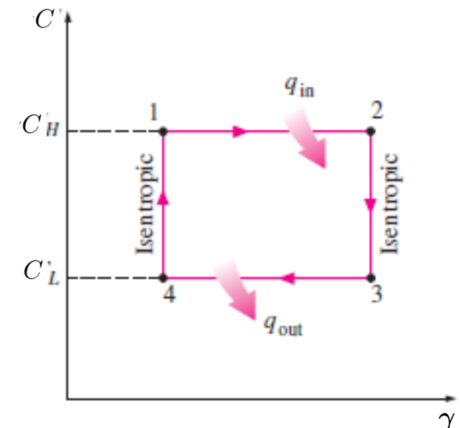
- 「準静的過程に沿った状態操作」を考えることで，統計モデルの分類誤差を低い値に抑えたまま，データのドメインを変化させることはできないか？

熱力学

$$\boxed{F(\lambda, \gamma, \sigma) = \min_{e(z|x), q(z), d(x|z), c(y|z)} R + \lambda D + \gamma C + \sigma S}$$

表現学習





- データのドメインを変化させる間，分類誤差の値が不変となることを要請する

$$p(x,t) = \underset{p(x)}{\operatorname{argmin}} (1-t) W_2^2\left(p(x^{ini}), p(x)\right) + t W_2^2\left(p(x), p(x^{fin})\right)$$

$$p(x,t) = \sum_{i=1}^{n_{ini}} \sum_{j=1}^{n_{fin}} \Gamma_{ij} \delta_{x-(1-t)x_i^{ini} - t x_j^{fin}}$$

$$y(x,t) = (1-t)\delta_{y-y_{x_i^{ini}}} + t\delta_{y-y_{x_j^{fin}}}$$

$$\boxed{\frac{dC(t)}{dt}} = 0$$