2-058: モデルベースメタ強化学習のための重み付きモデル推定

菱沼徹　泉田啓　（京都大学）

1. 設定

- メタRL（エピソード毎にMDPが変動する状況）、かつ、オフラインRL設定（事前収集データのみ利用可能な状況）
- 変動する実MDPを潜在変数を持つMDPモデルとして推定する、というモデルベース手法を議論
- 今回（IBIS2022）は、簡単のため、所与のターゲット方策挙動の予測ができるかという方差評価問題を扱う

2. 手法

- 通常のVAEの損失関数[1]

$$\sum_n \sum_t E_{z\sim q(z|D_n)} [-\ln p(s_n{}^{t+1}|s_n{}^t, a_n{}^t, z)] + \mathrm{KL\_loss}$$

- 本研究の損失関数

$$\sum_n \sum_t E_{z\sim q(z|D_n)} [-\boldsymbol{w_n\,(s_n{}^t, a_n{}^t, z; \pi)} \ln p(s_n{}^{t+1}|s_n{}^t, a_n{}^t, z)] + \mathrm{KL\_loss}$$

$$w_n(s_n{}^t, a_n{}^t, z; \pi) = \frac{\text{潜在変数}z\text{のMDPモデルで方策}\pi\text{を使う時の}sa\text{の分布}}{n\text{番目の実MDPで事前収集したデータの}sa\text{の分布}}$$

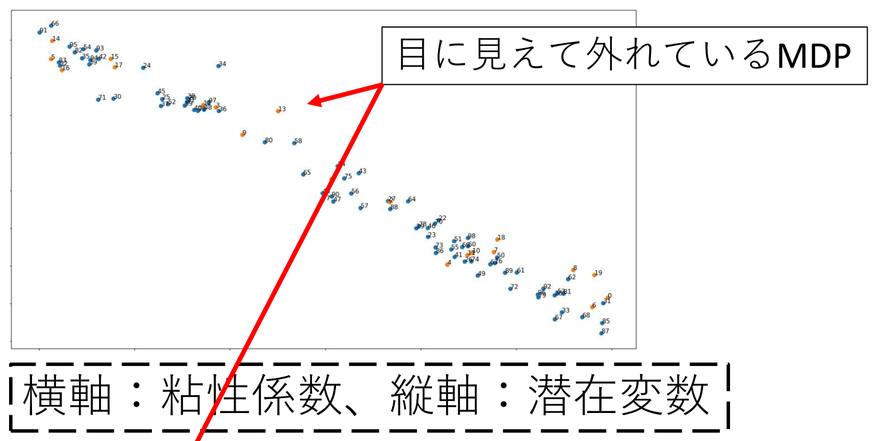
- 今回（IBIS2022）の実装方法
  - エンコーダ$q(z|D_n)$：順列不変ネット＋MLPで正規分布表現、デコーダ$p(s_n{}^{t+1}|s_n{}^t, a_n{}^t, z)$：MLPで正規分布表現
  - 重要度重みの推定（密度比メタ学習）
    - $w_n(s_n{}^t, a_n{}^t, z; \pi) \approx w(s_n{}^t, a_n{}^t, z, g_n)$とモデル化（$g_n$ は正規分布$q(z|D_n)$ のパラメータ）
    - 分母：オフラインデータ、分子：シミュレーションデータ、損失関数：ロジスティック回帰損失
  - 勾配計算の近似簡略化

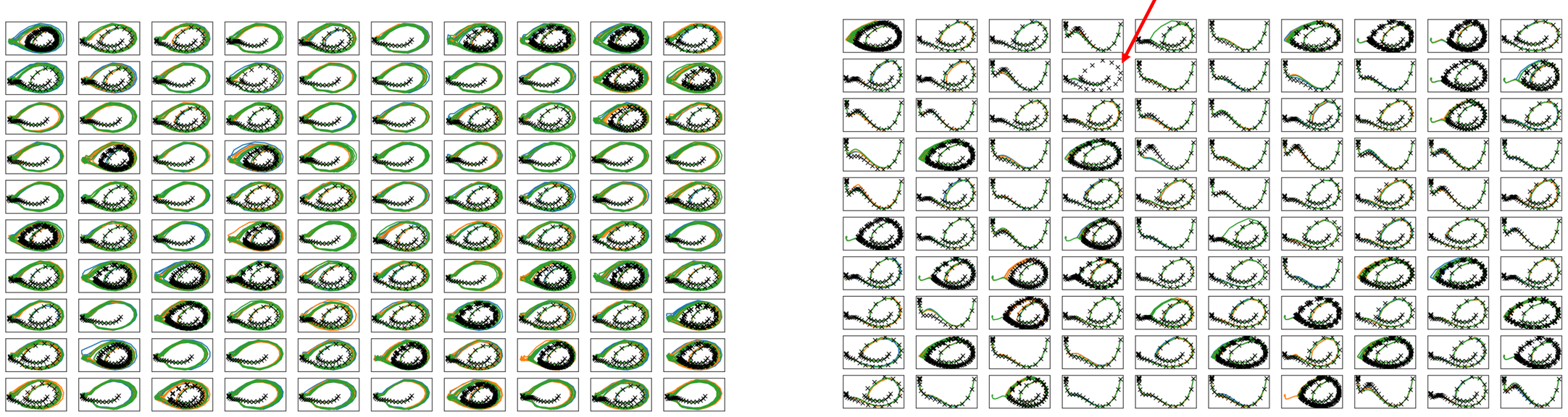
$$\nabla [w(\cdot) \ln p(\cdot)] = w(\cdot)[\nabla \ln p(\cdot) + \ln p(\cdot) \nabla \ln d(s, a; z, \pi)]$$

だが第二項を無視　←　事前検討（重み付最尤推定）は有効だった[2]ので、今回（重み付変分推論）も有効と期待

3. 数値実験（うまくいった例）

- 倒立振子タスクにおける方策挙動予測
- メタRL環境： s=（角度、角速度）、a=トルク、変動=粘性摩擦係数
- ターゲット方策：粘性摩擦ゼロのの最適方策
- オフラインデータ：100個の変動する実MDPでランダム方策で事前収集
- 80個の実MDPデータを訓練用、20個の実MDPデータを検証用





- 方策挙動予測図
  - 横軸：角速度、縦軸：角速度
  - 色線：通常のVAEによる予測（左図）と本研究による予測（右図）
  - 黒マーカ：事前収集と同じ実MDPにおけるターゲット方策の挙動（つまり真値）

4. 今後改善したい点、知りたい点（議論して頂けると嬉しいです）

- 通常VAE・本研究手法（今回実装）が両方とも失敗する例も多い（ベースとなるVAE、その実装が良くない？）
  - 最初の反復でデコーダがほぼ推定できていないと失敗（謎1）、パラメータを増やして悪化する事がある（謎2）
- メタRL関係無く、VAE×データ密度比で重要度重み付け、という研究をまず知りたい（現状はこの時点で手探り）

[1] Rakelly et al., Zintgraf et al., VariBAD: A Very Good Method for Bayes-Adaptive Deep RL via Meta-Learning, 2020.
[2] Hishinuma and Senda, Weighted model estimation for offline model-based reinforcement learning, 2021.