第二回GAIL勉強会

f-divergence最小化で学ぶGAIL

発表者: 北村俊徳 (NAIST M2)

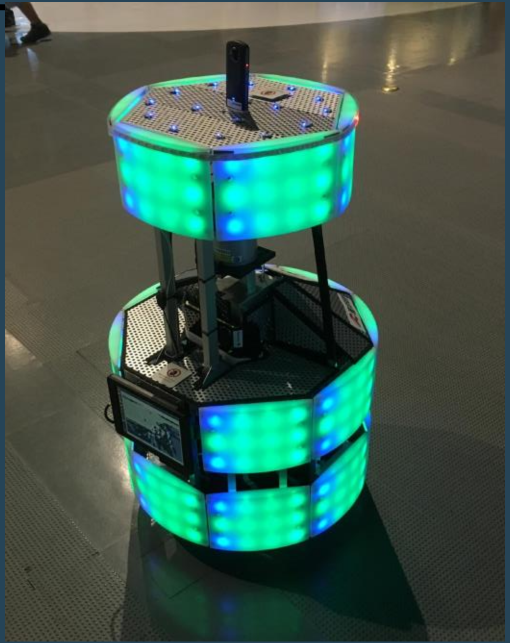
北村 俊徳（Toshinori Kitamura）

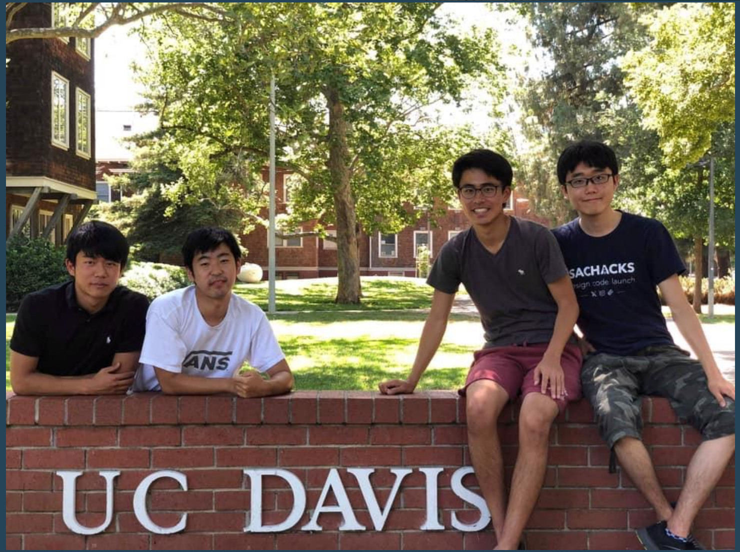
@syuntoku14

専門分野:
(深層)強化学習の理論とか, ロボティクス

学歴:
慶應義塾大学 理工学部卒
NAIST ロボットラーニング研究室 M2

その他:
2017年：MiraRobotics インターン
2018年：UC Davis留学
2019年：AIST インターン
2021年：OMRON SINIC X インターン





今日の目標

GAILについて学びますが…

- GAILの日本語の資料は割と存在するっぽい
  - 「GAIL 強化学習 スライド」でいっぱい出てきた
  - 今更僕がやっても…
- 正直原著論文(Ho+)は読みづらい

ので、今回は

1. GAILを導出し (Ke+, Nowozin+ 参考),
2. GAILがBehavior Cloningと比較してなぜ良いのか？ (Ghasemipour + 参考)

を学びます

参考文献:

- Generative Adversarial Imitation Learning (Ho+)
- Imitation Learning as f-divergence Minimization (Ke+)
- A divergence minimization perspective on Imitation Learning Methods (Ghasemipour +)
- f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization (Nowozin +)
- Estimating divergence functionals and the likelihood ratio by convex risk minimization (Nguyen +)

注意: 僕の専門は模倣学習ではないので、超詳しい話はできないです。
久しぶりにGAILの勉強したので間違いあるかも。

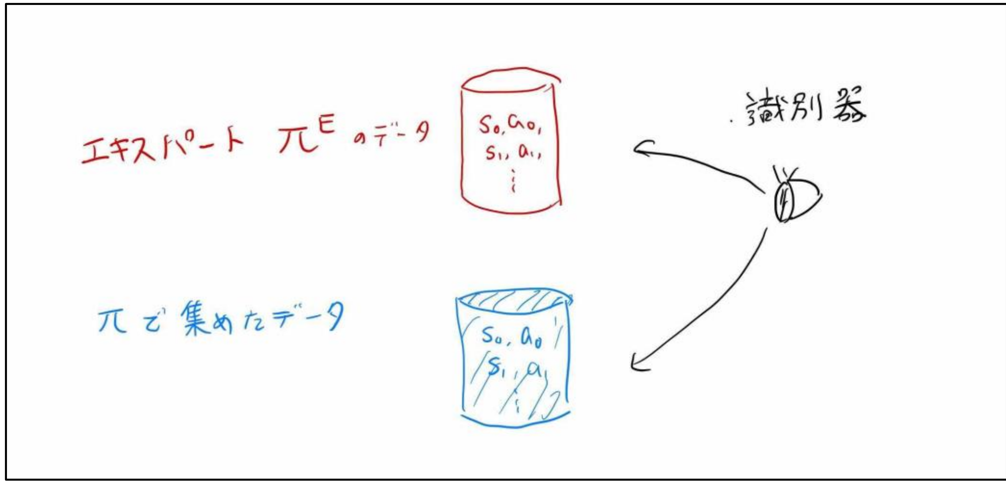
目次

GAILの概要（よくあるやつ）

ざっくりした解説 (よくあるやつ):

- GANみたいに**生成器 (方策 π)**と**識別器 (D)**の学習を繰り返す
- 識別器は状態行動系列が**エキスパート (π^E)** か**方策 π**から出ているのか識別
- **識別器の出力を使った報酬** ($r(s,a) = -\log D(s,a)$) で方策を学習
- 式: $\hat{\pi} = \operatorname{argmin}_{\pi\in\Pi} \max_{\omega} \mathbb{E}_{(s,a)\sim\rho_{\pi^E}} \log D(s,a) + \mathbb{E}_{(s,a)\sim\rho_{\pi}} \log(1 - D(s,a))$

なぜこの式が出てくるのか？を
f-ダイバージェンスの最小化 を使って導出します



GAILでやりたいこと

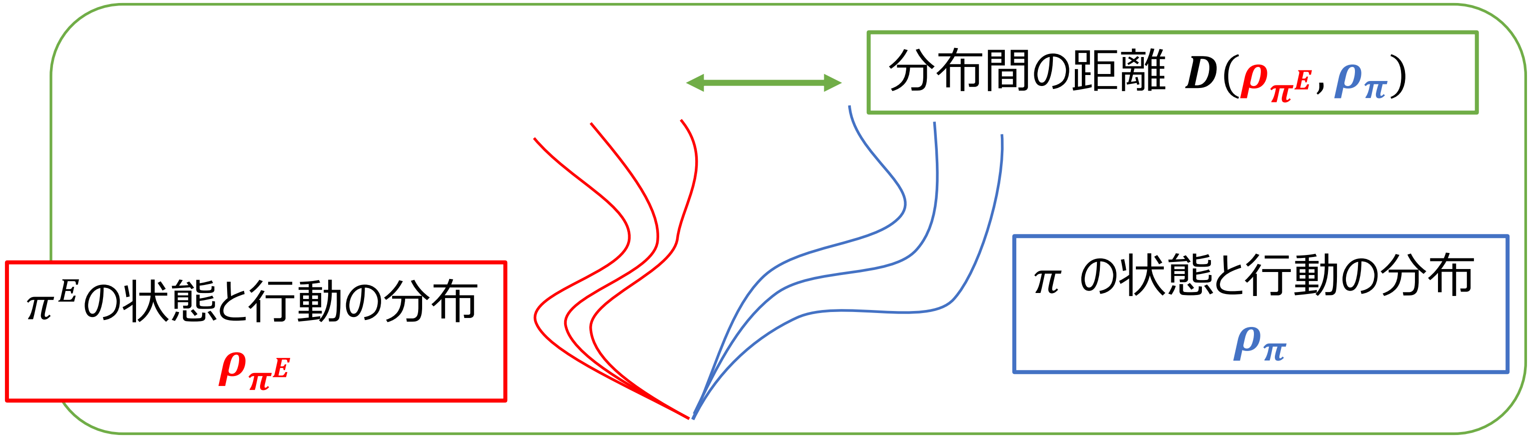
実はこれは逆強化学習 → 強化学習 による模倣学習の双対問題になってる。この証明は原著論文を見てね (Ho +)

$$\rho_\pi(s,a) = \pi(a|s) \sum_{t=0}^{\infty} \gamma^t P(s_t = s|\pi)$$

- やりたいこと:
  - **方策πが訪問する状態と行動の分布$\rho_\pi(s,a)$**を**エキスパートπ^Eの分布$\rho_{\pi^E}(s,a)$**に近づけたい
- どうやって分布同士を近づけるか？:

→ 分布間の適当な距離関数$\boldsymbol{D}$について、$\boldsymbol{D}(\boldsymbol{\rho}_{\pi^E}, \boldsymbol{\rho}_\pi)$を最小化しよう **(ダイバージェンスの最小化)**

- 距離関数Dって？:
  - 選択肢はいろいろ。KL ダイバージェンス, Jensen Shannon ダイバージェンス とか
    → **f-ダイバージェンス** を使って一般化しよう！



f-ダイバージェンス の概要

確率変数 X 上の分布 $p(x)$と$q(x)$について、p, q間の f-ダイバージェンス D_fは以下で定義される

$$\boldsymbol{D_f(p,q) = \sum_x q(x) f\left(\frac{p(x)}{q(x)}\right)}$$

ここで、$f: \mathbb{R}^+ \to \mathbb{R}$ は$f(1) = 0$を満たす凸関数。

凸関数を変えるといろいろなダイバージェンスになるよ

- $f(x) = \frac{1}{2}|x-1|$のとき全変動距離 (Total Variation): $D_f(p,q) = \frac{1}{2}\sum_x |p(x) - q(x)|$
- $f(x) = x\log x$ のときKL ダイバージェンス: $D_f(p,q) = \sum_x q(x)\frac{p(x)}{q(x)}\log\frac{p(x)}{q(x)} = \sum_x p(x)\log\frac{p(x)}{q(x)}$
- $f(x) = -(x+1)\log\frac{1+x}{2} + x\log x$ のとき Jensen-Shannon ダイバージェンス:

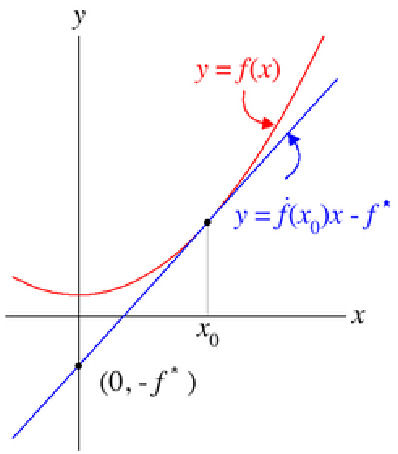
  $D_f(p,q) = \frac{1}{2}D_{KL}(p,q) + \frac{1}{2}D_{KL}(q,p)$

f-ダイバージェンスを最小化しよう (1)

やりたいこと: $\boldsymbol{D_f\left(\rho_{\pi^E}(s,a),\rho_\pi(s,a)\right)}$ を最小化したいが…

- $\rho_{\pi^E}(s,a)$や$\rho_\pi(s,a)$の値はわからないが、出てきたサンプル $(s_0,a_0,s_1,a_1,\ldots)$は持ってる

→ **サンプルだけでf-ダイバージェンスを推定したい**



https://ja.wikipedia.org/wiki/ルジャンドル変換

どうやってやるのか？

1. f-ダイバージェンスの凸関数fを**ルジャンドル変換**を使って変換
2. 出てくる期待値の部分をサンプルで近似

次ページで詳しく見るよ

ルジャンドル変換の説明は割愛（ググってね）
関数の変数をその微分に変える変換だよ

f-ダイバージェンスを最小化しよう (2)

やりたいこと: f-ダイバージェンス $D_f(p,q) = \sum_x q(x) f\left(\frac{p(x)}{q(x)}\right)$ をサンプルで推定しよう

準備: 凸関数fについてのルジャンドル変換 (f^*は凸共役) : $f\left(\frac{p(x)}{q(x)}\right) = \sup_{u \in \mathrm{dom}\, f^*} \left(\frac{p(x)}{q(x)} u - f^*(u)\right)$

ステップ1: $D_f(p,q)$をルジャンドル変換で変形 (Nguyen + の式(5) 参照)

$$D_f(p,q) = \sum_x q(x) f\left(\frac{p(x)}{q(x)}\right) = \sum_x q(x) \sup_\phi \left(\frac{p(x)}{q(x)} \phi(x) - f^*(\phi(x))\right) = \sup_\phi \sum_x \left(p(x)\phi(x) - q(x) f^*(\phi(x))\right)$$

ルジャンドル変換。supは全ての関数 $\phi: X \to \mathrm{dom}\, f^*$
について取ってる。上の準備の式とやってることは一緒

両辺のどちらでも$\frac{p(x)}{q(x)} = f^{*\prime}(\phi(x))$ を全てのxで満たすϕが$\sup_\phi$の解。
supでは全ての関数$\phi: X \to \mathrm{dom}\, f^*$を考えてるので等式が成立。

$$\sup_\phi \sum_x \left(p(x)\phi(x) - q(x) f^*(\phi(x))\right) \geq \sup_{\phi \in \Phi} \sum_x \left(p(x)\phi(x) - q(x) f^*(\phi(x))\right)$$

$\frac{p(x)}{q(x)} = f^{*\prime}(\phi(x))$で等式成立するが、$\frac{p(x)}{q(x)}$は使いたくない。supを与える$\phi$を適当な関数で近似した時を考えるため、適当な関数空間Φで制限。

f-ダイバージェンスを最小化しよう (2 つづき)

ステップ2: 期待値をサンプルで近似

$$D_f(p,q) \geq \sup_{\phi\in\Phi} \textstyle\sum_x \left(p(x)\phi(x) \ - q(x)f^*(u)\right) \ = \boldsymbol{\sup_{\phi\in\Phi}} \mathbb{E}_{\boldsymbol{x\sim p(x)}} \ \boldsymbol{\phi(x)} \ - \mathbb{E}_{\boldsymbol{x\sim q(x)}} \boldsymbol{f^*\left(\phi(x)\right)}$$

ステップ3: $\phi: X \to \mathrm{dom}\ f^*$なので、$\phi(x)$の出力が$f^*$の入力として妥当になるように変形しよう。

$V_\omega: X \to \mathbb{R}$ と 活性化関数 $g_f: \mathbb{R} \to \mathrm{dom\ f^*}$を使って、$\phi(x) = g_f\left(V_\omega(x)\right)$と表せば、

$$\boxed{D_f(p,q) \ \geq \boldsymbol{\sup_{\omega}} \mathbb{E}_{\boldsymbol{x\sim p(x)}} \ \boldsymbol{g_f\left(V_\omega(x)\right)} \ - \mathbb{E}_{\boldsymbol{x\sim q(x)}} \boldsymbol{f^*\left(g_f(V_\omega(x)\right)}}$$

が得られ、いい感じにf-ダイバージェンスがサンプルで推定できた！模倣学習に取り入れよう (次ページ)

f-ダイバージェンスを最小化しよう (3)

これでGAILの準備が整ったぞ！

- GAILでやりたいこと:　$\boldsymbol{D_f}\left(\boldsymbol{\rho_{\pi^E}(s,a)}, \boldsymbol{\rho_{\pi}(s,a)}\right)$ を最小化したい
- $D_f(p,q)$の推定は $D_f(p,q) \ \geq \sup_{\boldsymbol{\omega}} \mathbb{E}_{\boldsymbol{x\sim p(x)}}\, \boldsymbol{g_f}\left(\boldsymbol{V_\omega}(\boldsymbol{x})\right) \ - \mathbb{E}_{\boldsymbol{x\sim q(x)}} \boldsymbol{f}^*\left(\boldsymbol{g_f}(\boldsymbol{V_\omega}(\boldsymbol{x}))\right)$ でできそう

↓合体！

$$\hat{\pi} = \ \mathrm{argmin}_{\pi\in\Pi} \max_{\omega} \mathbb{E}_{(s,a)\sim\rho_{\pi^E}}\, g_f\left(V_\omega(s,a)\right) \ - \mathbb{E}_{(s,a)\sim\rho_\pi} f^*\left(g_f(V_\omega(s,a))\right)$$

これを解けば模倣学習完成！

GAILは f-ダイバージェンスが Jensen-Shannon ダイバージェンスの時に出てくるよ (次ページ)

GAILを導出しよう

$\hat{\pi} = \operatorname{argmin}_{\pi\in\Pi} \max_{\omega} \mathbb{E}_{(s,a)\sim\rho_{\pi^E}}\, g_f\left(V_\omega(s,a)\right) - \mathbb{E}_{(s,a)\sim\rho_\pi} f^*\left(g_f(V_\omega(s,a))\right)$について、

f-ダイバージェンスがJensen Shannonダイバージェンスの時、

- $f(x) = -(x+1)\log\frac{1+x}{2} + x\log x$ で $f^*(u) = -\log(2-\exp(u))$
- $f^*(u) = -\log(2-\exp(u))$なので、$u \in (-\infty, \log 2)$の範囲を取る
  $\rightarrow g_f\left(V_\omega(s,a)\right) = \log 2 + \log\frac{1}{1+\exp(-V_\omega(s,a))}$
- 代入して、$f^*\left(g_f(V_\omega(s,a))\right) = -\log 2 - \log\left(1 - \frac{1}{1+\exp(-V_\omega(s,a))}\right)$

Wikipediaみてね

よって、Jensen Shannonダイバージェンスでの模倣学習は

$$\hat{\pi} = \operatorname{argmin}_{\pi\in\Pi} \max_{\omega} \mathbb{E}_{(s,a)\sim\rho_{\pi^E}} \log\frac{1}{1+\exp\left(-V_\omega(s,a)\right)} + \mathbb{E}_{(s,a)\sim\rho_\pi} \log\left(1 - \frac{1}{1+\exp\left(-V_\omega(s,a)\right)}\right)$$

Discriminator

これはGAILの更新式と同じ！(元論文にはエントロピー正則化が入ってる)

GAILの導出まとめ

- $\boldsymbol{\rho_\pi(s,a)}$を$\boldsymbol{\rho_{\pi^E}(s,a)}$に近づけたい
  → Jensen-Shannonダイバージェンス$\boldsymbol{D_{\mathrm{JS}}(\rho_{\pi^E},\rho_\pi)}$を最小化しよう

- 密度関数の値はわからないので、サンプルだけで$\boldsymbol{D_{\mathrm{JS}}(\rho_{\pi^E},\rho_\pi)}$を推定したい
  → f-ダイバージェンスの凸関数fを**ルジャンドル変換**を使って変換
  → 出てくる期待値の部分をサンプルで近似
  → f-ダイバージェンスをJensen-Shannonダイバージェンスに置き換え

- 変形するとGAILの更新式

$$\hat{\pi} = \operatorname{argmin}_{\pi\in\Pi} \max_{\omega} \mathbb{E}_{(s,a)\sim\rho_{\pi^E}} \log\frac{1}{1+\exp\left(-V_\omega(s,a)\right)} + \mathbb{E}_{(s,a)\sim\rho_\pi} \log\left(1-\frac{1}{1+\exp\left(-V_\omega(s,a)\right)}\right)$$

が出てきて完成！

(ちなみにf-ダイバージェンスをreverse KLダイバージェンスにするとAIRL (Fu +) になるよ。シンプルに代入するだけじゃ出てこないので、証明は(Ghasemipour +) 参照。)

目次

1. GAILの導出 (数式多めな話)
2. GAILとBehavior Cloningの関係 (お気持ちの話)

GAILとBehavior Cloning

- GAIL:
  - やりたいこと: $\rho_\pi(s,a)$を$\rho_{\pi^E}(s,a)$に近づけたい
  - やりかた: Jensen-Shannonダイバージェンス$\boldsymbol{D_{\mathrm{JS}}(\rho_{\pi^E},\rho_\pi)}$を最小化する

- Behavior Cloning (BC):
  - やりたいこと: $\pi(a|s)$を$\pi^E(a|s)$に近づけたい
  - やりかた: forward KLダイバージェンス$\boldsymbol{D_{\mathrm{KL}}\left(\pi^E,\pi\right)} = \sum \pi^E \log \frac{\pi^E}{\pi}$を最小化する

$$\hat{\pi} = \operatorname{argmin}_{\pi\in\Pi} D_{\mathrm{KL}}(\pi^E,\pi) = \operatorname{argmax}_{\pi\in\Pi} \mathbb{E}_{(s,a)\sim\rho_{\pi^E}}[\log \pi(a|s)]$$

ちなみにReverse KLは

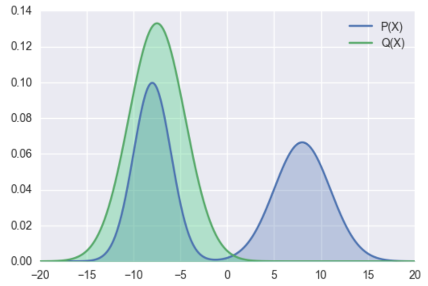
$$D_{\mathrm{KL}}(\pi,\pi^E) = \sum \pi \log \frac{\pi}{\pi^E}$$

なんでGAILの方がBCより良いのか？（Ghasemipour +)

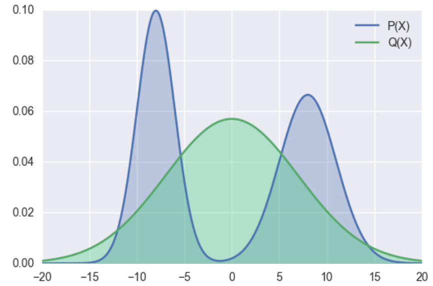
1. (やりたいことの違い): GAILは状態の分布も気にするが、BCは気にしない
2. (やることの違い): BCのforward KLダイバージェンスの最小化は mode-coveringな挙動
   GAILやAIRLはmode-seekingな挙動

次ページで詳しく見るよ

Mode-seeking



Mode-covering



https://wiseodd.github.io/techblog/2016/12/21/forward-reverse-kl/

状態の分布も模倣したほうが良い (Ghasemipour +)

仮説: RLの問題では報酬は状態に依存することが多い → 行動の分布だけ模倣してもダメ。
状態の分布も模倣したほうが性能が出るのでは？

実験: 次のFAIRLとBCを比較して検証

- FAIRL:
  - やりたいこと (GAILと同じ): $\rho_\pi(s,a)$を$\rho_{\pi^E}(s,a)$に近づけたい
  - やりかた (BCと同じ): forward KLダイバージェンス$D_{\mathrm{KL}}(\pi^E,\pi)$を最小化する

Method	Halfcheetah		Ant		Walker		Hopper	
	Det	Stoch	Det	Stoch	Det	Stoch	Det	Stoch
BC	-62 ± 182	-126 ± 218	82 ± 124	19 ± 70	1804 ± 1286	1293 ± 480	1435 ± 78	764 ± 129
AIRL	$\mathbf{8043 \pm 237}$	$\mathbf{7377 \pm 482}$	6024 ± 155	4598 ± 65	$\mathbf{3979 \pm 323}$	$\mathbf{3846 \pm 319}$	$\mathbf{3393 \pm 7}$	$\mathbf{2561 \pm 331}$
FAIRL	$\mathbf{7924 \pm 318}$	$\mathbf{7453 \pm 640}$	$\mathbf{6607 \pm 139}$	$\mathbf{5525 \pm 287}$	$\mathbf{4297 \pm 71}$	$\mathbf{4225 \pm 34}$	$\mathbf{3379 \pm 10}$	$\mathbf{3061 \pm 170}$

BCよりFAIRLの方が性能が出ている → 状態も模倣したほうが性能がでるみたい

mode-covering より mode-seekingの方が良い (Ke +)

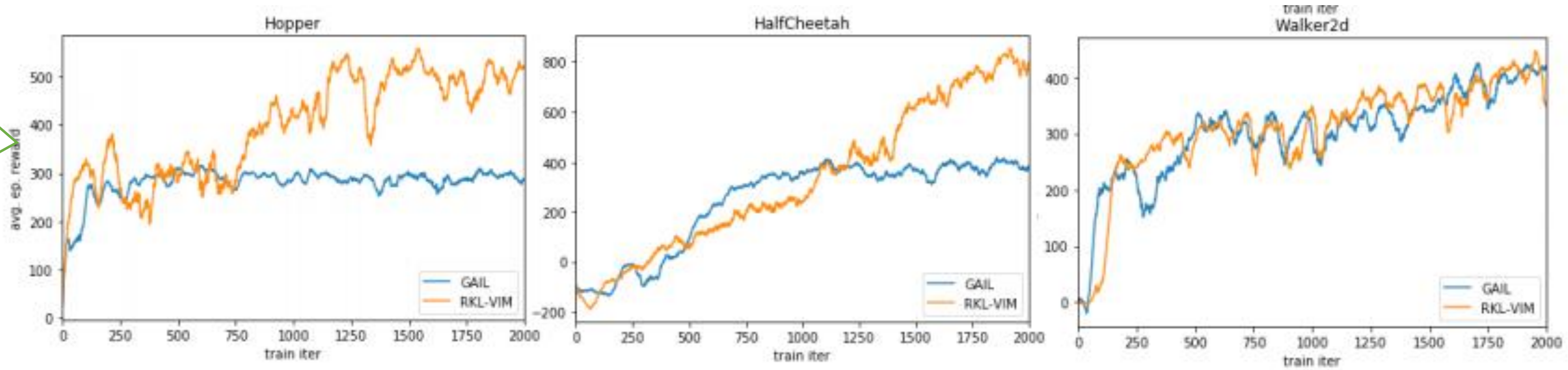
仮説: 性能を高めたいならmode-covering な模倣よりもmode-seekingな模倣の方が良い

Forward KLは2つの分布の中間をとりがち (mode-covering)
Reverse KLはどちらかの分布に偏る (mode-seeking)
仮説：制御問題では偏ったほうが有利な場面が多いのでは？

(Ke +)

実験: GAILのJensen-ShannonをReverse KLにした手法(AIRLかも？)(黄色)とGAIL (青色)を比較

黄色の方が性能が高いっぽい？
→ 純粋なmode-seekingの方
が性能出るみたい



まとめ

今回やったこと

1. GAILをf-ダイバージェンスの最小化を通じて導出
    - ルジャンドル変換が分かれば簡単にできるぞ！

2. GAILとBehavior Cloning (BC)の比較
    - GAILは状態の分布についても模倣するのでBCよりも性能がでる (かも？)
    - GAILやAIRLのダイバージェンス最小化はmode-seekingになりがち
→ mode-coveringなBCよりも性能がでる (かも？)

GAILに関する理論とかもっとあれば教えてください～