Combining Reinforcement Learning and Inverse Reinforcement Learning for Asset Allocation[1]

Igor Halperin

Fidelity Investments

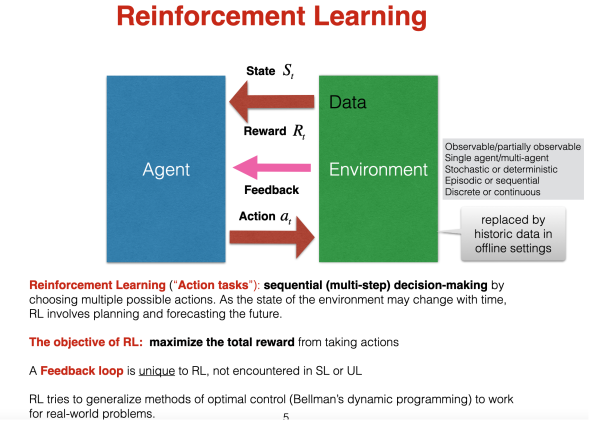
March 2023

[1] I.Halperin, J. Liu, and X. Zhang, Combining Reinforcement Learning and Inverse Reinforcement Learning for Asset Allocation Recommendations, https://arxiv.org/abs/2201.01874 (2022), M. Dixon and I. Halperin, G-Learner and GIRL: Goal Based Wealth Management with Reinforcement Learning, Risk.Net, July 2021

Background

- Asset management problems as problems of high-dimensional stochastic optimal control (SOC)
- We apply entropy regularized Reinforcement Learning (G-learning) to these SOC (noisy) problems
- Inverse Reinforcement Learning (IRL) is applied to back up the reward function of fund managers
- The combined RL/IRL scheme tries to learn from human experts, and improve over their strategies (policies)

What is Reinforcement Learning (RL)?



- We present a particular version of RL called G-learning, with applications to wealth management and asset allocation
- We also present two algorithms for for Inverse Reinforcement Learning (IRL) which recovers agents' rewards from the observed behavior

Portfolio model

Consider a simple portfolio model

- A universe of N assets (e.g. stocks) with the vector $\mathbf{p}_t$ of market prices at time t.
- In addition, can keep wealth in a risk-free bank cash account with risk-free interest rate r_f
- Vector $\mathbf{x}_t \in \mathbb{R}^N$ describes dollar amounts of positions in individual assets. $x_{it} < 0$ means a short position
- Trading has costs (fees and market impact)
- Trades $\mathbf{u}_t \in \mathbb{R}^N$ are made at the beginning of intervals t

The reward function for portfolio optmization

- For a next period pre-specified target portfolio value, $\hat{P}_{t+1}$, the expected one-step reward for time step t:

$$
\begin{aligned}
\hat{R}_t(\mathbf{x}_t, \mathbf{u}_t, c_t) &= -\mathbb{E}_t\left[\left(\hat{P}_{t+1} - (1 + \mathbf{r}_t)(\mathbf{x}_t + \mathbf{u}_t)\right)^2\right] \\
&- \lambda\left(\mathbf{1}^T\mathbf{u}_t - c_t\right)^2 - \mathbf{u}_t^T \boldsymbol{\Omega} \mathbf{u}_t. \qquad (1)
\end{aligned}
$$

- The three terms are: a penalty for underperformance against a benchmark portfolio, a soft constraints on a sum of all trades (where c_t is the flow into the portfolio), and a transaction cost term
- Trades $\mathbf{u}_t$ are considered the action variables in a dynamic portfolio optimization problem
- Note the quadratic structure of the resulting reward function!

Target portfolio

- One simple choice of the target portfolio $\hat{P}_{t+1}$ is a linear combination of a portfolio-independent benchmark B_t and the current portfolio growing with a fixed rate η:

$$\hat{P}_{t+1} = \rho B_t + \eta \, \mathbb{1}^T \mathbf{x}_t, \tag{2}$$

- ρ and η are parameters defining the relative weights of portfolio-independent and portfolio-dependent terms.
- For a sufficiently large values of B_t and η, such a target portfolio would be well above the current portfolio at all times, and thus would serve as a reasonable proxy to an asymmetric measure.
- For a benchmark B_t, we can use funds' benchmark indexes (rescaled to match the initial portfolio value)

Quadratic reward

- Asset returns as $\mathbf{r}_t = \bar{\mathbf{r}}_t + \tilde{\varepsilon}_t$ where $\bar{r}_0(t) = r_f$ is the risk-free rate (as the first asset is risk-free), and $\tilde{\varepsilon}_t = (0, \varepsilon_t)$ where ε_t is an idiosyncratic noise with covariance $\mathbf{\Sigma}_r$ of size $(N-1) \times (N-1)$.
- The one-step reward in Eq.(1) is computed more explicitly as follows:

$$
\begin{aligned}
R_t(\mathbf{x}_t, \mathbf{u}_t) &= -\hat{P}_{t+1}^2 + 2\hat{P}_{t+1}(\mathbf{x}_t + \mathbf{u_t})^{\mathbf{T}}(\mathbf{1} + \bar{\mathbf{r}}_{\mathbf{t}}) - (\mathbf{x_t} + \mathbf{u_t})^{\mathbf{T}}\, \hat{\mathbf{\Sigma}}_{\mathbf{t}}\, (\mathbf{x_t} + \mathbf{u_t}) - \lambda \left(\mathbb{1}^{\mathbf{T}} \mathbf{u_t} - \mathbf{c_t} \right)^{\mathbf{2}} - \omega \mathbf{u_t^T} \mathbf{u_t} \\
&= \mathbf{x}_t^T \mathbf{R}_t^{(xx)} \mathbf{x}_t + \mathbf{u}_t^T \mathbf{R}_t^{(ux)} \mathbf{x}_t + \mathbf{u}_t^T \mathbf{R}_t^{(uu)} \mathbf{u}_t + \mathbf{x}_t^T \mathbf{R}_t^{(x)} + \mathbf{u}_t^T \mathbf{R}_t^{(u)} + R_t^{(0)}
\end{aligned}
$$

where

$$
\hat{\mathbf{\Sigma}}_t = \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_r \end{bmatrix} + (\mathbb{1} + \bar{\mathbf{r}}_t)(\mathbb{1} + \bar{\mathbf{r}}_t)^T
$$

- The vector of free parameters defining the reward function is thus $\theta := (\lambda, \eta, \rho, \omega)$.
- The quadratic reward specification gives rise to semi-analytic optimal policies.

Stochastic policies

- For any parametrized deterministic policy $\pi_\theta(\cdot|\mathbf{x}_t)$, parameters θ are found from data, and hence are random themselves.
- Example: Markowitz portfolio model: allocations depend on expected returns that are estimated from data, thus random.
- A measure of uncertainly in recommended allocations is highly desirable in view of an **uncertain** world.
- Any sub-optimal behavior have probability zero under a deterministic policies.
- Conclusion: we need to work with **stochastic** policies.

Stochastic policies

- A stochastic policy is any valid probability distribution for actions $\mathbf{a}_t$:
$$\pi_\theta = \pi_\theta(\mathbf{a}_t | \mathbf{x}_t)$$
(Will also depend on expected returns $\bar{\mathbf{r}}_t$).
- If we have a stochastic policy, we have a **generative** model of action and dynamics - can be used for both past and *future* simulated data.

RL with stochastic policies

$$
\begin{aligned}
&\text{maximize } \mathbb{E}_{q_\pi}\left[\sum_{t'=t}^{T-1} \gamma^{t'-t} \hat{R}_{t'}(\mathbf{x}_{t'}, \mathbf{a}_{t'})\right] \\
&\qquad\qquad \text{w.r.t. } q_\pi(\bar{x}, \bar{a}|\mathbf{x}_0) = \pi(\mathbf{a}_0) \prod_{t=1}^{T-1} \pi(\mathbf{a}_t|\mathbf{x}_t) p\left(\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{a}_t\right) \\
&\qquad\qquad \text{subject to } \int d\mathbf{a}_t \, \pi\left(\mathbf{a}_t|\mathbf{x}_t\right) = 1
\end{aligned}
$$

Here $\mathbb{E}_{q_\pi}[\cdot]$ stands for expectations with respect to path probabilities defined according to the third line - driven by **stochastic** policies.

Reference policy

We assume that we are given a probabilistic **reference** (or **"prior"** policy $\pi_0(\mathbf{a}_t|\mathbf{x}_t)$.

It can be based on a parametric model, past historic data, etc.

We will use a simple Gaussian reference policy

$$\pi_0(\mathbf{a}_t|\mathbf{x}_t) = \frac{e^{-\frac{1}{2}(\mathbf{a}_t - \hat{\mathbf{a}}(\mathbf{x}_t))^T \Sigma_a^{-1} (\mathbf{a}_t - \hat{\mathbf{a}}(\mathbf{x}_t))}}{\sqrt{(2\pi)^N |\Sigma_a|}} \tag{3}$$

where

$$\hat{\mathbf{a}}(\mathbf{x}_t) = \hat{\mathbf{A}}_0 + \hat{\mathbf{A}}_1 \mathbf{x}_t \tag{4}$$

Bellman Optimality Equation

Let

$$V_t^\star(\mathbf{x}_t) = \max_{\pi(\cdot|x)} \mathbb{E}_t \left[\sum_{t'=t}^{T-1} \gamma^{t'-t} \hat{R}_{t'}(\mathbf{x}_{t'}, \mathbf{a}_{t'}) \right] \quad (5)$$

The optimal state value function $V_t^\star(\mathbf{x}_t)$ satisfies the Bellman optimality equation

$$V_t^\star(\mathbf{x}_t) = \max_{\mathbf{a}_t} \hat{R}_t(\mathbf{x}_t, \mathbf{a}_t) + \gamma \mathbb{E}_{t,\mathbf{a}_t} \left[V_{t+1}^\star(\mathbf{x}_{t+1}) \right] \quad (6)$$

The optimal policy $\pi^\star$ can be obtained from $V^\star$ as follows:

$$\pi_t^\star(\mathbf{a}_t|\mathbf{x}_t) = \arg\max_{\mathbf{a}_t} \hat{R}_t(\mathbf{x}_t, \mathbf{a}_t) + \gamma \mathbb{E}_{t,\mathbf{a}_t} \left[V_{t+1}^\star(\mathbf{x}_{t+1}) \right] \quad (7)$$

When $V_t(\mathbf{x}_t)$ is found, solving for π takes another optimization problem in Eq.(7) (a policy improvement step).

Bellman Optimality Equation: a reformulation

Reformulate the Bellman optimality equation:

$$V_t^\star(\mathbf{x}_t) = \max_{\pi(\cdot|x)\in\mathcal{P}} \sum_{\mathbf{a}_t\in\mathcal{A}_t} \pi(\mathbf{a}_t|\mathbf{x}_t)\left(\hat{R}_t(\mathbf{x}_t,\mathbf{a}_t) + \gamma\mathbb{E}_{t,\mathbf{a}_t}\left[V_{t+1}^\star(\mathbf{x}_{t+1})\right]\right) \tag{8}$$

Here $\mathcal{P} = \left\{\pi : \pi \geq 0, \mathbf{1}^T\pi = 1\right\}$ is a set of all valid distributions. Eq.(8) is equivalent to the original Bellman equation (5), because for any $x \in \mathbb{R}^n$, we have $\max_{i\in\{1,\ldots,n\}} x_i = \max_{\pi\geq 0,||\pi||\leq 1} \pi^T x$.

Information cost of a policy

The one-step *information cost* of a learned policy $\pi(\mathbf{a}_t|\mathbf{x}_t)$ relative to a reference policy $\pi_0(\mathbf{a}_t|\mathbf{x}_t)$ is (Tishby *et. al.*, 2015)

$$g^{\pi}(\mathbf{x}, \mathbf{a}) = \log \frac{\pi(\mathbf{a}_t|\mathbf{x}_t)}{\pi_0(\mathbf{a}_t|\mathbf{x}_t)} \tag{9}$$

Its expectation with respect to π is the KL divergence of $\pi(\cdot|\mathbf{x}_t)$ and $\pi_0(\cdot|\mathbf{x}_t)$:

$$\begin{aligned} \mathbb{E}_{\pi}\left[g^{\pi}(\mathbf{x}, \mathbf{a}) \middle| \, \mathbf{x}_t \right] &= KL[\pi || \pi_0](\mathbf{x}_t) \\ &\equiv \sum_{\mathbf{a}_t} \pi(\mathbf{a}_t|\mathbf{x}_t) \log \frac{\pi(\mathbf{a}_t|\mathbf{x}_t)}{\pi_0(\mathbf{a}_t|\mathbf{x}_t)} \end{aligned} \tag{10}$$

The total discounted information cost for a trajectory is

$$I^{\pi}(\mathbf{x}) = \sum_{t'=t}^{T} \gamma^{t'-t} \mathbb{E}\left[g^{\pi}(\mathbf{x}_{t'}, \mathbf{a}_{t'}) \middle| \, \mathbf{x}_t = \mathbf{x} \right] \tag{11}$$

Free energy

The *free energy* function $F_t^\pi(\mathbf{x}_t)$ is entropy-regularized value function (with the information cost penalty):

$$\begin{aligned} F_t^\pi(\mathbf{x}_t) &= V_t^\pi(\mathbf{x}_t) - \frac{1}{\beta} I^\pi(\mathbf{x}_t) \qquad (12) \\ &= \sum_{t'=t}^{T} \gamma^{t'-t} \mathbb{E}\left[\hat{R}_{t'}(\mathbf{x}_{t'}, \mathbf{a}_{t'}) - \frac{1}{\beta} g^\pi(\mathbf{x}_{t'}, \mathbf{a}_{t'}) \right] \end{aligned}$$

β is the regularization parameter that controls a trade-off between reward optimization and proximity to the reference policy.

Bellman equation for free energy

A Bellman equation for the free energy function $F_t^\pi(\mathbf{x}_t)$ is obtained from (12):

$$F_t^\pi(\mathbf{x}_t) = \mathbb{E}_{\mathbf{a}|x}\left[\hat{R}_t(\mathbf{x}_t, \mathbf{a}_t) - \frac{1}{\beta} g^\pi(\mathbf{x}_t, \mathbf{a}_t) + \gamma \mathbb{E}_{t,\mathbf{a}}\left[F_{t+1}^\pi(\mathbf{x}_{t+1})\right]\right] \tag{13}$$

Eq.(13) can be viewed as a soft probabilistic relaxation of the Bellman optimality equation for the value function, with the KL information cost penalty (11) as regularization controlled by the inverse temperature β.

G-function: an entropy-regularized Q-function

Define the state-action free energy function $G^{\pi}(\mathbf{x}, \mathbf{a})$ as

$$G_t^{\pi}(\mathbf{x}_t, \mathbf{a}_t) = \hat{R}_t(\mathbf{x}_t, \mathbf{a}_t) + \gamma \mathbb{E}\left[F_{t+1}^{\pi}(\mathbf{x}_{t+1}) \middle| \mathbf{x}_t, \mathbf{a}_t \right] \tag{14}$$

$$= \mathbb{E}_{t,\mathbf{a}} \left[\sum_{t'=t}^{T} \gamma^{t'-t} \left(\hat{R}_{t'}(\mathbf{x}_{t'}, \mathbf{a}_{t'}) - \frac{1}{\beta} g^{\pi}(\mathbf{x}_{t'}, \mathbf{a}_{t'}) \right) \right]$$

In the last equation we used the fact that the first action $\mathbf{a}_t$ in the G-function is fixed, and hence $g^{\pi}(\mathbf{x}_t, \mathbf{a}_t) = 0$ when we condition on $\mathbf{a}_t = \mathbf{a}$.

Compare this expression with Eq.(12) to get a relation between the G-function and F-function":

$$F_t^{\pi}(\mathbf{x}_t) = \sum_{\mathbf{a}_t} \pi(\mathbf{a}_t | \mathbf{x}_t) \left[G_t^{\pi}(\mathbf{x}_t, \mathbf{a}_t) - \frac{1}{\beta} \log \frac{\pi(\mathbf{a}_t | \mathbf{x}_t)}{\pi_0(\mathbf{a}_t | \mathbf{x}_t)} \right] \tag{15}$$

Optimal policy

We obtained

$$F_t^{\pi}(\mathbf{x}_t) = \sum_{\mathbf{a}_t} \pi(\mathbf{a}_t|\mathbf{x}_t) \left[G_t^{\pi}(\mathbf{x}_t, \mathbf{a}_t) - \frac{1}{\beta} \log \frac{\pi(\mathbf{a}_t|\mathbf{x}_t)}{\pi_0(\mathbf{a}_t|\mathbf{x}_t)} \right]$$

This is maximized by the following distribution: $\pi(\mathbf{a}_t|\mathbf{x}_t)$:

$$\pi(\mathbf{a}_t|\mathbf{x}_t) = \frac{1}{Z_t} \pi_0(\mathbf{a}_t|\mathbf{x}_t) e^{\beta G_t^{\pi}(\mathbf{x}_t, \mathbf{a}_t)} \tag{16}$$

$$Z_t = \sum_{\mathbf{a}_t} \pi_0(\mathbf{a}_t|\mathbf{x}_t) e^{\beta G_t^{\pi}(\mathbf{x}_t, \mathbf{a}_t)}$$

Optimal free energy

The free energy (15) evaluated at the optimal solution (16):

$$F_t^{\pi}(\mathbf{x}_t) = \frac{1}{\beta} \log Z_t = \frac{1}{\beta} \log \sum_{\mathbf{a}_t} \pi_0(\mathbf{a}_t|\mathbf{x}_t) e^{\beta G_t^{\pi}(\mathbf{x}_t, \mathbf{a}_t)} \tag{17}$$

Can use this to re-write the optimal policy:

$$\pi(\mathbf{a}_t|\mathbf{x}_t) = \pi_0(\mathbf{a}_t|\mathbf{x}_t) e^{\beta \left(G_t^{\pi}(\mathbf{x}_t, \mathbf{a}_t) - F_t^{\pi}(\mathbf{x}_t) \right)} \tag{18}$$

Putting all together: G-learning

We now have a set of three equations that have to be solved self-consistently for $t = T-1, \ldots, 0$:

$$
\begin{aligned}
G_t^\pi(\mathbf{x}_t, \mathbf{a}_t) &= \hat{R}_t(\mathbf{x}_t, \mathbf{a}_t) + \gamma \mathbb{E}_{t,\mathbf{a}} \left[F_{t+1}^\pi(\mathbf{x}_{t+1}) \middle| \mathbf{x}_t, \mathbf{a}_t \right] \\
F_t^\pi(\mathbf{x}_t) &= \frac{1}{\beta} \log \sum_{\mathbf{a}_t} \pi_0(\mathbf{a}_t | \mathbf{x}_t) e^{\beta G_t^\pi(\mathbf{x}_t, \mathbf{a}_t)} \\
\pi(\mathbf{a}_t | \mathbf{x}_t) &= \pi_0(\mathbf{a}_t | \mathbf{x}_t) e^{\beta (G_t^\pi(\mathbf{x}_t, \mathbf{a}_t) - F_t^\pi(\mathbf{x}_t))}
\end{aligned} \tag{19}
$$

with

$$
\begin{aligned}
G_T^\pi(\mathbf{x}_T, \mathbf{a}_T) &= \hat{R}_T(\mathbf{x}_T, \mathbf{a}_T) \\
F_T^\pi(\mathbf{x}_T) &= G_T^\pi(\mathbf{x}_T, \mathbf{a}_T) = \hat{R}_T(\mathbf{x}_T, \mathbf{a}_T)
\end{aligned} \tag{20}
$$

G-learning with a quadratic reward

- For quadratic rewards, the general equations of G-learning can be solved semi-analytically for Gaussian priors π_0 with the mean $\hat{u}_t$ given by a linear function of the state:

$$\pi_0(\mathbf{u}_t|\mathbf{x}_t) = \frac{1}{\sqrt{(2\pi)^N \left|\boldsymbol{\Sigma}\right|}} e^{-\frac{1}{2}(\mathbf{u}_t - \hat{u}_t)^T \boldsymbol{\Sigma}^{-1}(\mathbf{u}_t - \hat{u}_t)}, \quad \hat{u}_t := \bar{\mathbf{u}}_t + \bar{v}_t \mathbf{x}_t \tag{21}$$

- We start by specifying a functional form of the value function as a quadratic form of $\mathbf{x}_t$:

$$F_t^{\pi}(\mathbf{x}_t) = \mathbf{x}_t^T \mathbf{F}_t^{(xx)} \mathbf{x}_t + \mathbf{x}_t^T \mathbf{F}_t^{(x)} + F_t^{(0)}, \tag{22}$$

where parameters $\mathbf{F}_t^{(xx)}$, $\mathbf{F}_t^{(x)}$, $F_t^{(0)}$ can depend on time via their dependence on $\hat{P}_{t+1}$ and $\bar{\mathbf{r}}_t$.

- The dynamic equation takes the form[2]:

$$\mathbf{x}_{t+1} = \mathbf{A}_t (\mathbf{x}_t + \mathbf{u}_t) + (\mathbf{x}_t + \mathbf{u}_t) \circ \tilde{\varepsilon}_t, \quad \mathbf{A}_t := \text{diag}\left(1 + \bar{\mathbf{r}}_t\right), \;\; \tilde{\varepsilon}_t := (0, \varepsilon_t) \tag{23}$$

[2] Note that the only features used here are the expected asset returns $\bar{\mathbf{r}}_t$ for the current period t. We assume that the expected asset returns are available as an output of a separate statistical model using e.g., a factor model framework.

Putting it all together

- Coefficients of the value function (22) are computed backward in time starting from the last maturity $t = T - 1$.
- For $t = T - 1$, the quadratic reward (3) can be optimized analytically by the following action:

$$\mathbf{u}_{T-1} = -\frac{1}{2}\left[R_t^{(uu)}\right]^{-1}\left(\mathbf{R}_t^{(u)} + \mathbf{R}_t^{(ux)}\mathbf{x}_{T-1}\right) := \frac{1}{2}\left(\mathbf{M}_{T-1}\mathbf{x}_{T-1} + \mathbf{K}_{T-1}\right) \tag{24}$$

  where we defined

$$\mathbf{M}_t := -\left[R_t^{(uu)}\right]^{-1}\mathbf{R}_t^{(ux)}, \quad \mathbf{K}_t := -\left[R_t^{(uu)}\right]^{-1}\mathbf{R}_t^{(u)}$$

- As for the last time step we have $F_{T-1}^{\pi}(\mathbf{x}_{T-1}) = \hat{R}_{T-1}$

Optimal policy

- The optimal policy for the given step is given by

$$\pi(\mathbf{u}_t|\mathbf{x}_t) = \pi_0(\mathbf{u}_t|\mathbf{x}_t)e^{\beta(G_t^\pi(\mathbf{x}_t,\mathbf{u}_t) - F_t^\pi(\mathbf{x}_t))}. \tag{25}$$

- Using the quadratic action-value function produces a new Gaussian policy $\pi(\mathbf{u}_t|\mathbf{x}_t)$:

$$\pi(\mathbf{u}_t|\mathbf{x}_t) = \frac{1}{\sqrt{(2\pi)^n \left|\tilde{\Sigma}_p\right|}} e^{-\frac{1}{2}(\mathbf{u}_t - \tilde{\mathbf{u}}_t - \tilde{\mathbf{v}}_t\mathbf{x}_t)^T \tilde{\Sigma}_p^{-1}(\mathbf{u}_t - \tilde{\mathbf{u}}_t - \tilde{\mathbf{v}}_t\mathbf{x}_t)} \tag{26}$$

where $\tilde{\Sigma}_p^{-1} = \mathbf{\Sigma}_p^{-1} - 2\beta\mathbf{Q}_t^{(uu)}$, $\tilde{\mathbf{u}}_t = \tilde{\Sigma}_p\left(\mathbf{\Sigma}_p^{-1}\bar{\mathbf{u}}_t + \beta\mathbf{Q}_t^{(u)}\right)$ and $\tilde{\mathbf{v}}_t = \tilde{\Sigma}_p\left(\mathbf{\Sigma}_p^{-1}\bar{\mathbf{v}}_t + \beta\mathbf{Q}_t^{(ux)}\right)$.

Optimal policy

- Therefore, policy optimization for G-learning with quadratic rewards and Gaussian reference policy amounts to the Bayesian update of the prior distribution with parameters updates $\bar{\mathbf{u}}_t$, $\bar{\mathbf{v}}_t$, Σ_p to the new values $\tilde{\mathbf{u}}_t$, $\tilde{\mathbf{v}}_t$, $\tilde{\Sigma}_p$.
- These quantities depend on time via their dependence on the targets $\hat{P}_t$ and expected asset returns $\bar{\mathbf{r}}_t$.

The final scheme: RL case

RL case: rewards are **observed**.

Initiate a trajectory $(\bar{\mathbf{x}}_1^{(0)}, \bar{\mathbf{u}}_1^{(0)}), \ldots, (\bar{\mathbf{x}}_T^{(0)}, \bar{\mathbf{u}}_T^{(0)})$
Repeat until convergence:
For $t = T - 1, \ldots, 0$:
1. Compute the expected value at time t of the F-function at time $t + 1$
2. Use this value and observed rewards to update the Q-function.
3. Compute the value of the F-function at time t.
4. Recompute the policy distribution $\pi(\mathbf{u}_t | \mathbf{x}_t)$ by updating its mean and variance

Unobservable rewards: IRL

- Inverse Reinforcement Learning (IRL): states and actions are observed, but rewards are **not** observed.
- IRL in our model is easy, as it amounts to Maximum Likelihood:
  The negative log-likelihood of data is

$$LL(\theta) = \sum_{t \in \zeta} \left(\beta \left(G_t^{\pi}(\mathbf{x}_t, \mathbf{u}_t) - F_t^{\pi}(\mathbf{x}_t) \right) - \frac{1}{2} \log |\mathbf{\Sigma}_r| - \frac{1}{2} \mathbf{\Delta}_t^T \mathbf{\Sigma}_r^{-1} \mathbf{\Delta}_t \right) \tag{27}$$

  where $\mathbf{x}_t, \mathbf{u}_t$ are *observed* optimal state-action sequences and $\mathbf{\Delta}_t := \frac{\mathbf{x}_{t+1}^{(r)}}{\mathbf{x}_t^{(r)} + \mathbf{u}_t^{(r)}} - \mathbf{A}_t^{(r)}$.
- All unknown parameters $\Theta = (\lambda, \mu_i, \beta)$ can then be computed using Gradient Descent or Stochastic Gradient Descent.
- This produces the GIRL (G-learning for IRL) algorithm (see M.Dixon and IH 2021)

T-REX algorithm for IRL

T-REX algorithm for IRL

- T-REX (Trajectory-ranked Reward EXtrapolation) is an IRL algorithm that is able to **extrapolate** beyond the demonstrated behavior (Brown et. al 2019)

- Based on externally provided ranking of demonstrated trajectories. This creates preference relations such as $\tau_i \prec \tau_j$ that suggests that trajectory j is preferred to trajectory i

cumulative rewards computed with this function should match the rank-ordering relation:

$$\sum_{(s,a)\in\tau_i} \hat{r}_\theta(s,a) < \sum_{(s,a)\in\tau_j} \hat{r}_\theta(s,a) \quad \text{if} \quad \tau_i \prec \tau_j. \tag{11.129}$$

Let $\hat{J}_\theta(\tau_i) = \sum_t \gamma^t \hat{r}_\theta(s_t, a_t)$ be a discounted cumulative rewards on trajectory τ_i. We train T-REX by minimizing the following loss function:

$$\mathcal{L}(\theta) = \mathbb{E}_{\tau_i,\tau_j\sim\Pi}\left[\xi\left(P\left(\hat{J}_\theta(\tau_i) < \hat{J}_\theta(\tau_j)\right), \tau_i \prec \tau_j\right)\right], \tag{11.130}$$

where Π is a distribution over pairs of demonstrations, and ξ is a binary loss function. The binary event probability P in Eq. (11.130) is modeled as a softmax distribution

$$P\left(\hat{J}_\theta(\tau_i) < \hat{J}_\theta(\tau_j)\right) = \frac{\exp\sum_{s,a\in\tau_j}\hat{r}_\theta(s,a)}{\exp\sum_{s,a\in\tau_i}\hat{r}_\theta(s,a) + \exp\sum_{s,a\in\tau_j}\hat{r}_\theta(s,a)}. \tag{11.131}$$

For the loss function $\xi(\cdot)$, a cross-entropy loss is used, so that the loss function becomes

$$\mathcal{L}(\theta) = -\sum_{\tau_i\prec\tau_j} \log \frac{\exp\sum_{s,a\in\tau_j}\hat{r}_\theta(s,a)}{\exp\sum_{s,a\in\tau_i}\hat{r}_\theta(s,a) + \exp\sum_{s,a\in\tau_j}\hat{r}_\theta(s,a)}. \tag{11.132}$$

T-REX can not only mimic, but surpass a teacher!

Combined IRL-RL framework

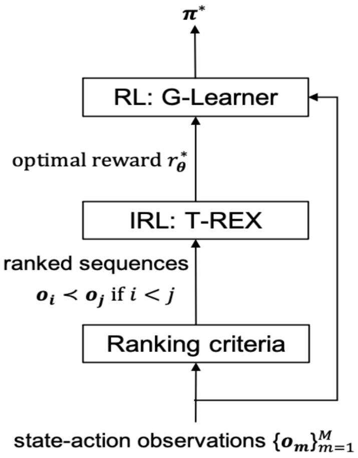


Figure 1. The flowchat of our IRL-RL framework

Convergence of the T-REX algorithm for IRL

G-learner and T-REX for optimization of funds' performance

- RL-IRL for optimization of funds' performance (Halperin, Liu, Zhang 2022, https://arxiv.org/pdf/2201.01874.pdf).
- Analyze a few groups of mutual funds with different benchmark indexes (S&P500 or Russell 3000).

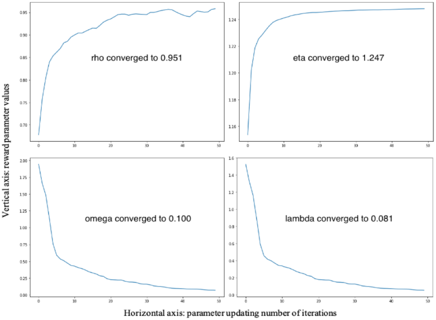


Figure 3. T-REX: reward parameter learning curve

Table 1. Inferred reward parameter values and T-REX model accuracy from all three fund groups

Fund Group	$\{S_i\}_{i=1}^{6}$	$\{RG_i\}_{i=1}^{7}$	$\{RV_i\}_{i=1}^{5}$
ρ^*	0.951	0.186	0.584
η^*	1.247	1.532	1.210
λ^*	0.081	0.009	0.009
ω^*	0.100	0.012	0.009
acc (train/test)	0.911/0.832	0.878/0.796	0.906/0.832
cor (train/test)	0.988/0.954	0.925/0.884	0.759/0.733

-

Classification accuracy forT-REX

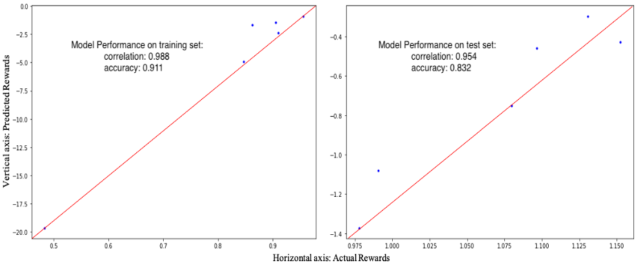


Figure 2. T-REX: classification accuracy and ranking order preservation measured by correlation scores

Can T-REX outperform fund managers?

G-learner optimization of funds' performance

- RL-IRL for optimization of funds' performance (Halperin, Liu, Zhang 2022, https://arxiv.org/pdf/2201.01874.pdf).

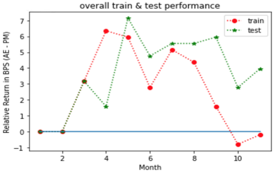


Figure 5. G-learner: overall trading performance with growth funds benchmarked by Russell 3000

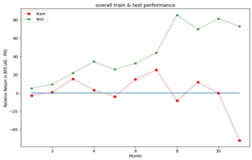


Figure 4. G-learner: overall trading performance with funds benchmarked by S&P500

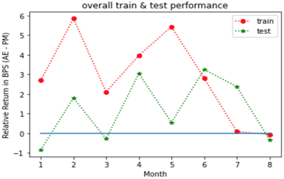


Figure 6. G-learner: overall trading performance with value funds benchmarked by Russell 3000

Summary: IRL with RL for asset management

Summary

- The IRL-RL framework works in a two-step setting: IRL is used to infer the reward function, and RL is used to optimize asset allocation
- Usage for asset allocation: our model is able to learn from the collective intelligence of individual fund manages, and outperform most of them
- Usage for trade recommendation: asset allocation recommendations can be converted to recommendation to buy/sell individual stocks
- Usage for fund analysis: outputs of the IRL model (parameters of the reward function) can be used to cluster different funds according to both their declared and perceived investment philosophies
- Can use other dimension reduction methods (e.g. factor-based)
- Extras on RL and IRL: the MLF book