## Reinforcement Learning for Stock Trading

- Stock trading as dynamic optimization of stock portfolios
- **Single stock case:** optimal execution
- **Portfolio problems:** optimal investment portfolio, optimal portfolio liquidation, index tracking
- A one-step reward should include expected cost of trade, plus risk penalties
- Need to keep stock positions as state variables
- Need to take into account a feedback loop (market impact)
- May be a very high-dimensional problem!
- May be a Big Data problem (for HFT tasks)!

## Optimal Execution: Single Stock Case

**Task:** sell or buy $N$ stocks of a given company within time $T$ with minimal costs

- An agency (broker) problem (or a problem for traders who executes such trades themselves)
- **Time scale:** minutes
- Solved hundred or thousand times a day
- Should split a large order into smaller chunks and execute them sequentially to minimize market impact from trades
- Amounts to a dynamic optimization problem with a feedback loop
- Market orders vs limit orders
- A high-dimensional and Big Data problem, when working with the limit order book (LOB) data

## Optimal Portfolio Liquidation

**Task:** sell *N different* stocks within time *T* with minimal costs

- An agency (broker) problem (or a problem for traders who executes such trades themselves)
- **Time scale:** minutes/hours/
- Similar to a single-stock case, should split a large order into smaller chunks and execute them sequentially to minimize market impact from trades
- Amounts to a dynamic *portfolio* optimization problem with a feedback loop
- Market orders vs limit orders
- Even higher-dimensional and Bigger Data problem, when working with the limit order book (LOB) data

## Optimal Investment

**Task:** optimally manage a portfolio of *N different* stocks for a planning horizon *T* with minimal costs

- An investor (trader) problem
- **Objective:** optimization of dynamic portfolio allocation to maximize risk-adjusted and cost-adjusted portfolio return at time *T*
- **Time scale:** days/weeks/months/years
- Different set of predictors (macro-variables, long-term forecasts, etc.)
- **Non-dynamic, one-step case:** Markowitz portfolio model

# Index Tracking

**Task:** optimally track a market index (e.g. the S&P 500 index) with a smaller portfolio of *N different* stocks for a planning horizon *T* with minimal costs

- An investor (trader) problem
- **Time scale:** weeks/months/continuous
- **Objective: A**s one cannot directly invest in the S&P 500, can try to track it with a smaller portfolio for period *T* at a minimal cost
- Exchange-Traded Funds (ETF) do such index tracking, one can 'invest in the market' by investing in SPY
- **'Custom' (bespoke) index tracking problem:** A dynamic portfolio optimization with an index-matching terminal condition at *T*
- A high-dimensional and Big Data problem

## What We Will D

- Most of these problems are problems of high-dimensional Stochastic Optimal Control (SOC)
- We apply Reinforcement Learning to these SOC problems
- We will *not* discretize a state space, as we need to work in high dimensions
- Ideally, want a framework that works for all of the above settings
- Start with a general portfolio model
- Work with similar convex (quadratic) objective functions for all settings

## 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
- Trades $\mathbf{u}_t \in \mathbb{R}^N$ are made at the beginning of intervals $t$, so that the portfolio vector $\mathbf{x}_t^+$ right after trades are deterministic:

$$\mathbf{x}_t^+ = \mathbf{x}_t + \mathbf{u}_t \qquad \text{Eq: 1}$$

- Trading has costs (fees and market impact)

## Portfolio Value Pre- and Post-trade

The total portfolio value is

$$v_t = \mathbf{1}^T \mathbf{x}_t + b_t \qquad \text{Eq: 2}$$

where $\mathbf{1}$ is a vector of ones. The post-trade portfolio is therefore

$$\begin{aligned} v_t^+ &= \mathbf{1}^T \mathbf{x}_t + b_t^+ = \mathbf{1}^T (\mathbf{x}_t + \mathbf{u}_t) + b_t^+ \\ &= v_t + \mathbf{1}^T \mathbf{u}_t + b_t^+ - b_t \end{aligned} \qquad \text{Eq: 3}$$

## Self-financing Condition

Assume that all re-balancing of stock positions are financed from the bank cash account (additional costs will be introduced later). This imposes the following 'self-financing' constraint:

$$\mathbf{1}^T \mathbf{u}_t + b_t^+ - b_t = 0 \qquad \text{Eq: 4}$$

**Meaning:** the portfolio value remains instantaneously unchanged upon a trade:

$$v_t^+ = v_t \qquad \text{Eq: 5}$$

## One-period Investment

The post-trade portfolio $v_t^+$ and cash are invested at the beginning of period $t$ until the beginning of the next period. The return of asset $i$ over period $t$ is defined as

$$(r_t)_i = \frac{(p_{t+1})_i - (p_t)_i}{(p_t)_i}, \quad i = 1, \ldots, n \qquad \text{Eq: 6}$$

Asset positions at the next time period are

$$\mathbf{x}_{t+1} = \mathbf{x}_t^+ + \mathbf{r}_t \odot \mathbf{x}_t^+ \qquad \text{Eq: 7}$$

where $\odot$ stands for an element-wise (Hadamard) product, and $\mathbf{r}_t \in \mathbb{R}^n$ is the vector of asset returns from period $t$ to period $t+1$.

## Asset Returns Model

Use a linear specification of one-period excess asset returns:

$$\mathbf{r}_t - r_f \mathbf{1} = \mathbf{W}_t \mathbf{z}_t - \mathbf{M}_t^T \mathbf{u}_t + \varepsilon_t \qquad \text{Eq: 8}$$

where:

$\mathbf{z}_t$ is a vector of predictors

$\mathbf{W}_t$ is a factor loading matrix

$\mathbf{M}_t$ is a matrix of permanent market impacts within a linear impact specification

$\varepsilon_t$ is a vector of residuals with

$$\mathbb{E}\left[\varepsilon_t\right] = 0, \ \mathrm{Var}_t\left[\varepsilon_t\right] = \Sigma_t \qquad \text{Eq: 9}$$

## The Next-period Portfolio

The next-period portfolio value is then obtained as follows:

$$\begin{aligned} v_{t+1} &= \mathbf{1}^T \mathbf{x}_{t+1} = (1 + \mathbf{r}_t)^T \mathbf{x}_t^+ \\ &= (1 + \mathbf{r}_t)^T (\mathbf{x}_t + \mathbf{u}_t) \end{aligned} \qquad \text{Eq: 10}$$

Can also compute the change of the portfolio value in excess of a risk-free growth $r_f$:

$$\begin{aligned} \Delta v_t &\equiv v_{t+1} - (1 + r_f) v_t \\ &= (\mathbf{1} + \mathbf{r}_t)^T (\mathbf{x}_t + \mathbf{u}_t) + (1 + r_f) b_t^+ \\ &- (1 + r_f) \mathbf{1}^T \mathbf{x}_t - (1 + r_f) b_t \\ &= (\mathbf{r}_r - r_f \mathbf{1})^T (\mathbf{x}_t + \mathbf{u}_t) \end{aligned} \qquad \text{Eq: 11}$$

where in the second equation, we used Eq: 4

## Terminal Conditions

Boundary conditions for an index/benchmark tracking problem: $\mathbf{x}_T = \mathbf{x}_T^B$ where $\mathbf{x}_T^B$ is a benchmark portfolio at time $T$ (e.g. the S&P 500 index). This fixes the action $\mathbf{u}_T$:

$$\mathbf{u}_T = \mathbf{x}_T^M - \mathbf{x}_{T-1} \qquad \boxed{\text{Eq: 12}}$$

Therefore, action $\mathbf{u}_T$ at the last step is deterministic and is not subject to optimization that should be applied to $T$ remaining actions $\mathbf{u}_{T-1}, \ldots, \mathbf{u}_0$.

Other terminal conditions can be used for other formulations (portfolio liquidation, optimal investment portfolios).

## Initial Conditions

Initial conditions depend on the problem:

- **Optimal stock execution:** an initial value of a block of stock
- **Optimal portfolio liquidation:** an initial stock portfolio
- **Optimal portfolio management:** initial cash $b_0$, or initial portfolio $\mathbf{x}_0$.
- **Index tracking:** initial cash $b_0$, or initial portfolio $\mathbf{x}_0$.

## Instantaneous Rewards

Substitute the excess return equation

$$\mathbf{r}_t - r_f \mathbf{1} = \mathbf{W}_t \mathbf{z}_t - \mathbf{M}_t^T \mathbf{u}_t + \varepsilon_t$$

into portfolio change equation

$$\delta v_t = (\mathbf{r}_r - r_f \mathbf{1})^T (\mathbf{x}_t + \mathbf{u}_t)$$

This produces an instantaneous random reward received upon taking action $\mathbf{u}_t$:

$$R_t^{(0)}(\mathbf{x}_t, \mathbf{u}_t) = \left(\mathbf{W}_t \mathbf{z}_t - \mathbf{M}_t^T \mathbf{u}_t + \varepsilon_t\right)^T (\mathbf{x}_t + \mathbf{u}_t) \qquad \text{Eq: 13}$$

To this, we have to add (negative) rewards received due to instantaneous market impact and transaction fees, and a risk penalty.

## Risk Penalty

A risk penalty should be added as a negative reward.

We choose the variance of instantaneous reward $R_t^{(0)}(\mathbf{x}_t, \mathbf{u}_t)$ as a simple quadratic risk measure.
A negative risk-penalty contribution to a one-step reward:

$$\begin{aligned} R_t^{(risk)}(\mathbf{x}_t, \mathbf{u}_t) &= -\lambda \mathrm{Var}_t \left[ R_t^{(0)}(\mathbf{x}_t, \mathbf{u}_t) \middle| \mathbf{x}_t + \mathbf{u}_t \right] \\ &= -\lambda (\mathbf{x}_t + \mathbf{u}_t)^T \Sigma_t (\mathbf{x}_t + \mathbf{u}_t) \end{aligned} \qquad \text{Eq: 14}$$

Here $\lambda$ is a risk aversion parameter, and $\Sigma_t$ is the noise covariance matrix.

## Fee Penalty

Fees (transaction costs) depend on a sign of $u_t$.

We represent each action $u_{ti}$ as a difference of two non-negative action variables $u_{ti}^+, u_{ti}^- \geq 0$:

$$u_{ti} = u_{ti}^+ - u_{ti}^- , \quad |u_{ti}| = u_{ti}^+ + u_{ti}^- , \quad u_{ti}^+, u_{ti}^- \geq 0,$$

so that, Eq: 16

$$u_{ti} = u_{ti}^+ \text{ if } u_{ti} > 0 \text{ and } u_{ti} = -u_{ti}^- \text{ if } u_{ti} < 0.$$

The instantaneous fee (transaction costs) penalty:

$$R_t^{(fee)}(\mathbf{x}_t, \mathbf{u}_t) = -\kappa_t^{+T} \mathbf{u}_t^+ - \kappa_t^{-T} \mathbf{u}_t^- \qquad \text{Eq: 17}$$

## Market Impact Penalty

The market impact penalty describes an additional cost due to an impact of trades on stock prices

$$R_t^{(impact)}(\mathbf{x}_t, \mathbf{u}_t) = -\mathbf{x}_t^T \left( \Theta_t^{+T} \mathbf{u}_t^+ + \Theta_t^{-T} \mathbf{u}_t^- + \Phi_t^T \mathbf{z}_t \right)$$

Note that we assume a proportional market impact which can also depend on signals $\mathbf{z}_t$.

## Risk- and Cost-adjusted One-step Reward

The final one-step reward

$$R_t(\mathbf{x}_t, \mathbf{u}_t) = R_t^{(0)}(\mathbf{x}_t, \mathbf{u}_t) + R_t^{(risk)}(\mathbf{x}_t, \mathbf{u}_t) + R_t^{(impact)}(\mathbf{x}_t, \mathbf{u}_t) + R_t^{(fee)}(\mathbf{x}_t, \mathbf{u}_t) \quad \text{Eq: 18}$$

The *expected* one-step reward given action $\mathbf{u}_t = \mathbf{u}_t^+ - \mathbf{u}_t^-$ is

$$\hat{R}_t(\mathbf{x}_t, \mathbf{u}_t) = \hat{R}_t^{(0)}(\mathbf{x}_t, \mathbf{u}_t) + R_t^{(risk)}(\mathbf{x}_t, \mathbf{u}_t) + R_t^{(impact)}(\mathbf{x}_t, \mathbf{u}_t) + R_t^{(fee)}(\mathbf{x}_t, \mathbf{u}_t) \quad \text{Eq: 19}$$

where

$$\hat{R}_t^{(0)}(\mathbf{x}_t, \mathbf{u}_t) = \mathbb{E}_{t,u}\left[R_t^{(0)}(\mathbf{x}_t, \mathbf{u}_t)\right] \quad \text{Eq: 20}$$

where $\mathbb{E}_{t,u}[\cdot] = \mathbb{E}[\cdot|\mathbf{x}_t, \mathbf{u}_t]$ stands for averaging over next-periods realizations of market returns.

## One-step Reward as a Quadratic Functional

We can write the one-step expected reward as a quadratic functional of states and actions:

$$\hat{R}_t(\mathbf{x}_t, \mathbf{u}_t) = \mathbf{a}_t^T \mathbf{R}_{aat} \mathbf{a}_t + \mathbf{x}_t^T \mathbf{R}_{xxt} \mathbf{x}_t + \mathbf{a}_t^T \mathbf{R}_{axt} \mathbf{x}_t + \mathbf{a}_t^T \mathbf{R}_{at} + \mathbf{x}_t^T \mathbf{R}_{xt} \quad \text{Eq: 21}$$

Where

$$\mathbf{a}_t = \begin{bmatrix} \mathbf{u}_t^+ \\ \mathbf{u}_t^- \end{bmatrix}, \ \mathbf{R}_{aat} = \begin{bmatrix} -\mathbf{M}_t - \lambda\Sigma_t & \mathbf{M}_t + \lambda\Sigma_t \\ \mathbf{M}_t + \lambda\Sigma_t & -\mathbf{M}_t - \lambda\Sigma_t \end{bmatrix},$$

$$\mathbf{R}_{xxt} = -\lambda\Sigma_t, \ \mathbf{R}_{axt} = \begin{bmatrix} -\mathbf{M}_t - 2\lambda\Sigma_t - \Theta_t^+ \\ \mathbf{M}_t + 2\lambda\Sigma_t - \Theta_t^- \end{bmatrix}$$

$$\mathbf{R}_{at} = \begin{bmatrix} \mathbf{W}_t \mathbf{z}_t - \kappa_t^+ \\ \mathbf{W}_t \mathbf{z}_t - \kappa_t^- \end{bmatrix}, \ \mathbf{R}_{xt} = [\mathbf{W}_t - \Phi_t]\, \mathbf{z}_t \quad \text{Eq: 22}$$

## Forward Portfolio Optimization Problem

A multi-period risk and cost-adjusted reward maximization problem:

$$\text{maximize} \quad \mathbb{E}_t \left[ \sum_{t'=t}^{T-1} \gamma^{t'-t} \hat{R}_{t'}(\mathbf{x}_{t'}, \mathbf{a}_{t'}) \right] \qquad \text{Eq: 23}$$

$$\hat{R}_t(\mathbf{x}_t, \mathbf{a}_t) = \mathbf{a}_t^T \mathbf{R}_{aat} \mathbf{a}_t + \mathbf{x}_t^T \mathbf{R}_{xxt} \mathbf{x}_t + \mathbf{a}_t^T \mathbf{R}_{axt} \mathbf{x}_t + \mathbf{a}_t^T \mathbf{R}_{at} + \mathbf{x}_t^T \mathbf{R}_{xt}$$

$$\text{w.r.t.} \quad \mathbf{a}_t = \begin{pmatrix} \mathbf{u}_t^+ \\ \mathbf{u}_t^- \end{pmatrix} \geq 0,$$

$$\text{subject to } \mathbf{u}_t \in \mathcal{A}_t,\ \mathbf{x}_t + \mathbf{u}_t \in \mathcal{Z}_t$$

Here $0 < \gamma \leq 1$ is a discount factor, and $\mathcal{A}_t$, $\mathcal{Z}_t$ are sets of constraints. The sum over $t' = [t, \ldots, T-1]$ does not include the last period $t' = T$, because the last action is fixed by Eq.(12).

## Convex Optimization

$$\text{maximize} \quad \mathbb{E}_t \left[ \sum_{t'=t}^{T-1} \gamma^{t'-t} \hat{R}_{t'}(\mathbf{x}_{t'}, \mathbf{a}_{t'}) \right]$$

$$\hat{R}_t(\mathbf{x}_t, \mathbf{a}_t) = \mathbf{a}_t^T \mathbf{R}_{aat} \mathbf{a}_t + \mathbf{x}_t^T \mathbf{R}_{xxt} \mathbf{x}_t + \mathbf{a}_t^T \mathbf{R}_{axt} \mathbf{x}_t + \mathbf{a}_t^T \mathbf{R}_{at} + \mathbf{x}_t^T \mathbf{R}_{xt}$$

$$\text{w.r.t.} \quad \mathbf{a}_t = \begin{pmatrix} \mathbf{u}_t^+ \\ \mathbf{u}_t^- \end{pmatrix} \geq 0,$$

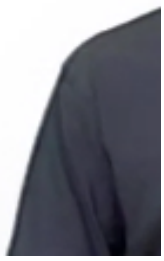
$$\text{subject to } \mathbf{u}_t \in \mathcal{A}_t,\ \mathbf{x}_t + \mathbf{u}_t \in \mathcal{Z}_t$$

This is a convex optimization problem if constraints are convex (see S. Boyd et. al. "Multi-Period Trading via Convex Optimization" (2017)).

# Examples of Constraints

Different trading constraints $\mathcal{A}_t$ and holding constraints $\mathcal{Z}_t$ can be imposed, depending on a problem, e.g.

- **Long only :** $\mathbf{x}_t + \mathbf{u}_t \geq 0$
- **Limits on asset holdings:** $\mathbf{x}_{min} \leq \mathbf{x}_t + \mathbf{u}_t \leq \mathbf{x}_{max}$
- **Leverage constraint:** $\sum_{i=1}^{N} |x_{it} + u_{it}| \leq L_{max} v_t$
- **Minimum cash balance:** $b_t \geq b_{min}$



# Forward Portfolio

**Forward** (conventional) portfolio optimization:

- **The task:** build an optimal trading strategy for a portfolio
- **Given:** an objective function, terminal and initial conditions, constraints
- **Needs:** a dynamic model for prices $\mathbf{p}_t$ and signals $\mathbf{z}_t$
- Sensitive to details of dynamics and forecast for signals $\mathbf{z}_t$

# One-step Portfolio

**Special case:** one-step optimization:

- **The task:** find an optimal single step strategy (asset allocations)
- **Given:** an objective function, terminal and initial conditions, constraints
- A variance-adjusted reward is equivalent to the Markowitz portfolio model
- Sensitive to details of forecast for signals $\mathbf{z}_t$

## Inverse Optimization

Invert the one-step optimization:

- **Given:** an optimal asset allocation
- **The task:** find an objective (reward) function, or parameters defining it
- For the Markowitz portfolio model, this produces a "market view" of signals $\mathbf{z}_t$
- This is an approach of Black-Litterman (BL) (1992)
- The BL model was re-interpreted as inverse optimization by Bertsimas *et. al.* (2012)
- **Usage:** assess a value of 'private' signals $\mathbf{z}'_t$

## Dynamic Inverse Portfolio Optimization

The same as above, but for a multi-step optimization:

- **Given:** an optimal *sequential* asset allocation (actions)
- **The task:** find a reward function, or parameters defining it, and find an action policy
- **Two possible settings:** a proprietary portfolio, or a market portfolio
- For a market portfolio, this produces a *dynamic* "market view" of signals $\mathbf{z}_t$, similar to the BL
- For a proprietary portfolio, this produces a model of a trader
- **Usage:** assess a value of 'private' signals $\mathbf{z}'_t$
- A proper problem for RL or IRL!

## Dynamic Inverse Portfolio Optimization

The same as above, but for a multi-step optimization:

- **Given:** an optimal *sequential* asset allocation (actions)
- **The task:** find a reward function, or parameters defining it, and find an action policy
- **Two possible settings:** a proprietary portfolio, or a market portfolio
- For a market portfolio, this produces a *dynamic* "market view" of signals $\mathbf{z}_t$, similar to the BL
- For a proprietary portfolio, this produces a model of a trader
- **Usage:** assess a value of 'private' signals $\mathbf{z}'_t$
- A proper problem for RL or IRL!

## Deterministic Policies

The multi-period optimization problem (Eq: 23) assumes that an optimal policy that determines actions $\mathbf{a}_t$ is a deterministic policy
We can write it as a delta-like probability distribution

$$\pi(\mathbf{a}_t|\mathbf{x}_t) = \delta\left(\mathbf{a}_t - \mathbf{a}_t^\star(\mathbf{x}_t)\right) \qquad \boxed{\text{Eq: 24}}$$

where the optimal deterministic action $\mathbf{a}_t^\star(\mathbf{x}_t)$ is obtained by maximization of the objective (Eq: 23) with respect to controls $\mathbf{a}_t$.

## Stochastic Policies

**Problems with deterministic policies:** they hardly exist, and therefore are hardly relevant!

- For any parametrized deterministic policy $\pi_\theta(\cdot|\mathbf{x}_t)$, parameters $\theta$ 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 uncertainty in recommended allocation is highly desirable in view of an uncertain world
- Any sub-optimal behavior should have probability zero under deterministic policies
- Conclusion: we need to work with stochastic policies

## Portfolio Optimization with Stoc

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

$$\hat{R}_t(\mathbf{x}_t,\mathbf{a}_t)=\mathbf{a}_t^T\mathbf{R}_{aat}\mathbf{a}_t+\mathbf{x}_t^T\mathbf{R}_{xxt}\mathbf{x}_t+\mathbf{a}_t^T\mathbf{R}_{axt}\mathbf{x}_t+\mathbf{a}_t^T\mathbf{R}_{at}+\mathbf{x}_t^T\mathbf{R}_{xt}$$

$$\text{w.r.t. } q_\pi(\bar{x},\bar{a}|\mathbf{x}_0)=\pi(\mathbf{a}_0)\prod_{t=1}^{T-1}\pi(\mathbf{a}_t|\mathbf{x}_t)P(\mathbf{x}_{t+1}|\mathbf{x}_t,\mathbf{a}_t)$$

$$\text{subject to } \int d\mathbf{a}_t\,\pi(\mathbf{a}_t|\mathbf{x}_t)=1$$

Here $\mathbb{E}_{q_\pi}[\cdot]$ stands for expectations with respect to path probabilities defined according to the third line

## Reference Policy

We assume that we are given a probabilistic reference "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

$$\pi_0(\mathbf{a}_t|\mathbf{x}_t) = \frac{e^{-\frac{1}{2}(\mathbf{a}_t-\hat{\mathbf{a}}(\mathbf{x}_t))^T\Sigma_a^{-1}(\mathbf{a}_t-\hat{\mathbf{a}}(\mathbf{x}_t))}}{\sqrt{(2\pi)^N|\Sigma_a|}} \quad \text{Eq: 25}$$

where

$$\hat{\mathbf{a}}(\mathbf{x}_t) = \hat{\mathbf{A}}_0 + \hat{\mathbf{A}}_1\mathbf{x}_t \quad \text{Eq: 26}$$

## Bellman Optimality Equation

Let

$$V_t^\star(\mathbf{x}_t) = \max_{\pi(\cdot|x)} \mathbb{E}_t\left[\sum_{t'=t}^{T-1}\gamma^{t'-t}\hat{R}_{t'}(\mathbf{x}_{t'},\mathbf{a}_{t'})\right] \quad \text{Eq: 27}$$

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

$$V_t^\star(\mathbf{x}_t) = \max_{\mathbf{a}_t} \hat{R}_t(\mathbf{x}_t,\mathbf{a}_t) + \gamma\mathbb{E}_{t,\mathbf{a}_t}\left[V_{t+1}^\star(\mathbf{x}_{t+1})\right] \quad \text{Eq: 28}$$

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

$$\pi_t^\star(\mathbf{a}_t|\mathbf{x}_t) = \arg\max_{\mathbf{a}_t} \hat{R}_t(\mathbf{x}_t,\mathbf{a}_t) + \gamma\mathbb{E}_{t,\mathbf{a}_t}\left[V_{t+1}^\star(\mathbf{x}_{t+1})\right] \quad \text{Eq: 29}$$

When $V_t(\mathbf{x}_t)$ is found, solving for $\pi$ takes another optimization problem in Eq.(29) (a policy improvement step).

## Bellman Optimality Equation: A Reformulation

Reformulate the Bellman optimality equation (a Fenchel-type representation):

$$V_t^\star(\mathbf{x}_t) = \max_{\pi(\cdot|x)\in\mathcal{P}} \sum_{\mathbf{a}_t\in\mathcal{A}_t} \pi(\mathbf{a}_t|\mathbf{x}_t)\left(\hat{R}_t(\mathbf{x}_t,\mathbf{a}_t) + \gamma \mathbb{E}_{t,\mathbf{a}_t}\left[V_{t+1}^\star(\mathbf{x}_{t+1})\right]\right) \quad \text{Eq: 30}$$

Here $\mathcal{P} = \{\pi : \pi \geq 0, \mathbf{1}^T\pi = 1\}$ is a set of all valid distributions. Eq.(30) is equivalent to the original Bellman equation (Eq: 27), 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)

$$g^\pi(\mathbf{x},\mathbf{a}) = \log \frac{\pi(\mathbf{a}_t|\mathbf{x}_t)}{\pi_0(\mathbf{a}_t|\mathbf{x}_t)} \quad \text{Eq: 30}$$

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

$$\begin{aligned}\mathbb{E}_\pi\left[g^\pi(\mathbf{x},\mathbf{a})|\,\mathbf{x}_t\right] &= KL[\pi||\pi_0](\mathbf{x}_t) \quad \text{Eq: 32}\\ &\equiv \sum_{\mathbf{a}_t} \pi(\mathbf{a}_t|\mathbf{x}_t)\log\frac{\pi(\mathbf{a}_t|\mathbf{x}_t)}{\pi_0(\mathbf{a}_t|\mathbf{x}_t)}\end{aligned}$$

The total discounted information cost for a trajectory is

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

## Free Energy

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

$$F_t^\pi(\mathbf{x}_t) = V_t^\pi(\mathbf{x}_t) - \frac{1}{\beta} I^\pi(\mathbf{x}_t) \qquad \text{Eq: 34}$$

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

$\beta$ is the "inverse temperature" 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 (Eq: 34):

$$F_t^\pi(\mathbf{x}_t) = \mathbb{E}_{\mathbf{a}|x}\left[\hat{R}_t(\mathbf{x}_t, \mathbf{a}_t) - \frac{1}{\beta} g^\pi(\mathbf{x}_t, \mathbf{a}_t) + \gamma \mathbb{E}_{t,\mathbf{a}}\left[F_{t+1}^\pi(\mathbf{x}_{t+1})\right]\right] \qquad \text{Eq: 35}$$

(Eq.35) can be viewed as a soft probabilistic relaxation of the Bellman optimality equation for the value function, with the **KL** information cost penalty (Eq: 32) as regularization controlled by the inverse temperature $\beta$.

## G-Function: An Entropy-Regularized Q-Function

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

$$G_t^\pi(\mathbf{x}_t, \mathbf{a}_t) = \hat{R}_t(\mathbf{x}_t, \mathbf{a}_t) + \gamma \mathbb{E}\left[F_{t+1}^\pi(\mathbf{x}_{t+1}) | \mathbf{x}_t, \mathbf{a}_t\right] \quad \text{Eq: 36}$$

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

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.34) to get a relation between the G-function and F-function:

$$F_t^\pi(\mathbf{x}_t) = \sum_{\mathbf{a}_t} \pi(\mathbf{a}_t|\mathbf{x}_t)\left[G_t^\pi(\mathbf{x}_t, \mathbf{a}_t) - \frac{1}{\beta}\log\frac{\pi(\mathbf{a}_t|\mathbf{x}_t)}{\pi_0(\mathbf{a}_t|\mathbf{x}_t)}\right] \quad \text{Eq: 37}$$

## 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: $\pi(\mathbf{a}_t|\mathbf{x}_t)$:

$$\pi(\mathbf{a}_t|\mathbf{x}_t) = \frac{1}{Z_t}\pi_0(\mathbf{a}_t|\mathbf{x}_t)e^{\beta G_t^\pi(\mathbf{x}_t, \mathbf{a}_t)} \quad \text{Eq: 38}$$

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

## Optimal Free Energy

The free energy (Eq: 37) evaluated at the optimal solution (Eq: 38):

$$F_t^\pi(\mathbf{x}_t) = \frac{1}{\beta} \log Z_t = \frac{1}{\beta} \log \sum_{\mathbf{a}_t} \pi_0(\mathbf{a}_t|\mathbf{x}_t) e^{\beta G_t^\pi(\mathbf{x}_t, \mathbf{a}_t)} \quad \boxed{\text{Eq: 39}}$$

Can use this to re-write the optimal policy:

$$\pi(\mathbf{a}_t|\mathbf{x}_t) = \pi_0(\mathbf{a}_t|\mathbf{x}_t) e^{\beta(G_t^\pi(\mathbf{x}_t, \mathbf{a}_t) - F_t^\pi(\mathbf{x}_t))} \quad \boxed{\text{Eq: 40}}$$

## Putting All Together

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

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

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

$$\pi(\mathbf{a}_t|\mathbf{x}_t) = \pi_0(\mathbf{a}_t|\mathbf{x}_t) e^{\beta(G_t^\pi(\mathbf{x}_t, \mathbf{a}_t) - F_t^\pi(\mathbf{x}_t))} \quad \boxed{\text{Eq: 41}}$$

With

$$G_T^\pi(\mathbf{x}_T, \mathbf{a}_T) = \hat{R}_T(\mathbf{x}_T, \mathbf{a}_T)$$

$$F_T^\pi(\mathbf{x}_T) = G_T^\pi(\mathbf{x}_T, \mathbf{a}_T) = \hat{R}_T(\mathbf{x}_T, \mathbf{a}_T)$$

## State Dynamics for the Port

The dynamics of the state vector $\mathbf{x}_t$ is obtained using Eqs. (7) and (8):

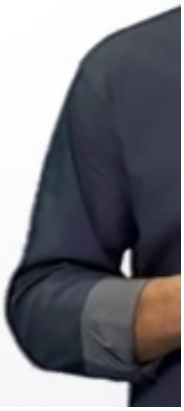
$$\begin{aligned}\mathbf{x}_{t+1} &= \mathbf{x}_t + \mathbf{u}_t + \mathbf{r}_t \odot (\mathbf{x}_t + \mathbf{u}_t) \\ &= \mathbf{x}_t + \mathbf{u}_t + \Big( r_f \mathbf{1} + \mathbf{W}_t \mathbf{z}_t - \mathbf{M}_t^T \mathbf{u}_t \\ &+ \varepsilon_t \Big) \odot (\mathbf{x}_t + \mathbf{u}_t) \\ &= \mathbf{A}_t \mathbf{x}_t + \mathbf{A}_t \mathbf{u}_t - \mathbf{u}_t^T \mathbf{M}_t \mathbf{u}_t - \mathbf{u}_t^T \mathbf{M}_t \mathbf{x}_t \\ &+ \varepsilon(\mathbf{x}_t, \mathbf{u}_t)\end{aligned}$$

Here we assumed that the matrix M of market impacts is diagonal with elements $\mu_i$, and set

$$\mathbf{A}_t = \mathbf{1} + \text{diag}\left(\mathbf{1} r_f + \mathbf{W}_t \mathbf{z}_t\right), \quad \mathbf{M}_t = \text{diag}\left(\mu_i\right)$$

$$\varepsilon(\mathbf{x}_t, \mathbf{u}_t) \equiv \varepsilon_t \odot (\mathbf{x}_t + \mathbf{u}_t)$$

Eq: 42



## Linearization of Dynamics

The dynamics are non-linear due to the market impact $\sim \mathbf{M}$.
Assume we are given a trajectory $(\bar{x}_1, \bar{u}_1), \ldots, (\bar{x}_T, \bar{u}_T)$
(a way to do it will be described later). Define increments
$\delta \mathbf{x}_t$ and $\delta \mathbf{u}_t$ as follows:

$$\mathbf{x}_t = \bar{\mathbf{x}}_t + \delta \mathbf{x}_t, \;\; \mathbf{u}_t = \bar{\mathbf{u}}_t + \delta \mathbf{u}_t \qquad \boxed{\text{Eq: 43}}$$

The dynamics equation (Eq: 42) can now be linearized
by keeping linear terms in increments $\delta \mathbf{x}_t$ and $\delta \mathbf{u}_t$:

$$\delta \mathbf{x}_{t+1} = \hat{\mathbf{A}}_t \delta \mathbf{x}_t + \hat{\mathbf{B}}_t \delta \mathbf{u}_t + \hat{\mathbf{x}}_{t+1}$$
$$+ \varepsilon_t \odot (\bar{\mathbf{x}}_t + \delta \mathbf{x}_t + \bar{\mathbf{u}}_t + \delta \mathbf{u}_t) \qquad \boxed{\text{Eq: 44}}$$

## Free Energy Function Spe

We parametrize the free energy function as follows:

$$F_t^{\pi}(\mathbf{x}_t) = \delta\mathbf{x}_t^T \mathbf{D}_t \delta\mathbf{x}_t + \delta\mathbf{x}_t^T \mathbf{H}_t + f_t(\bar{\mathbf{x}}_t) \qquad \text{Eq: 46}$$

As positions $\mathbf{x}_T$ are fixed by (12) , we can use Eqs.(41) and (43) to get

$$F_T^{\pi}(\mathbf{x}_t) = \hat{R}_T(\bar{\mathbf{x}}_T + \delta\mathbf{x}_T, \bar{\mathbf{a}}_T + \delta\mathbf{a}_T) \qquad \text{Eq: 47}$$

We use this to fix $\mathbf{D}_T$, $\mathbf{H}_T$, $f_T(\bar{\mathbf{x}}_T)$ in Eq.(46).

For values $t = T-1, \ldots, 0$, the expectation of the the next-period F-function is

$$\mathbb{E}_{t,\mathbf{a}}\left[F_{t+1}^{\pi}(\mathbf{x}_{t+1})\right] = f_t(\bar{\mathbf{x}}_t) + \widehat{\delta\mathbf{x}}_{t+1}^T \mathbf{H}_t$$
$$+\widehat{\delta\mathbf{x}}_{t+1}^T \mathbf{D}_{t+1} \widehat{\delta\mathbf{x}}_{t+1} + Tr\left[\mathbf{D}_{t+1} \mathrm{Cov}\left(\delta\mathbf{x}_{t+1}\right)\right]$$

## Free Energy From the Next Time

The previous formulation can be simplified

$$\mathbb{E}_{t,\mathbf{a}}\left[F_{t+1}^{\pi}(\mathbf{x}_{t+1})\right] = \delta\mathbf{a}_t^T \mathbf{F}_{aat} \delta\mathbf{a}_t + \delta\mathbf{x}_t^T \mathbf{F}_{xxt} \delta\mathbf{x}_t$$
$$+\delta\mathbf{a}_t^T \mathbf{F}_{axt} \delta\mathbf{x}_t + \delta\mathbf{a}_t^T \mathbf{F}_{at} + \delta\mathbf{x}_t^T \mathbf{F}_{xt} + f_{t+1}(\bar{\mathbf{x}}_t, \bar{\mathbf{u}}_t)$$

## Parametrization of the G-Function

For the G-function, we introduce a similar parametrization to Eq.(45):

$$G_t^{\pi}(\mathbf{x}_t, \mathbf{a}_t) = \delta\mathbf{a}_t^T \mathbf{G}_{aat}\delta\mathbf{a}_t + \delta\mathbf{x}_t^T \mathbf{G}_{xxt}\delta\mathbf{x}_t + \delta\mathbf{a}_t^T \mathbf{G}_{axt}\delta\mathbf{x}_t + \delta\mathbf{a}_t^T \mathbf{G}_{at} + \delta\mathbf{x}_t^T \mathbf{G}_{xt} + g_t(\bar{\mathbf{x}}_t, \bar{\mathbf{u}}_t) \quad \text{Eq: 46}$$

Can also express the reward Eq.(21) in terms of $\delta\mathbf{x}_t$ and $\delta\mathbf{a}_t$:

$$\begin{aligned}\hat{R}_t(\mathbf{x}_t, \mathbf{u}_t) &= \mathbf{a}_t^T \mathbf{R}_{aat}\mathbf{a}_t + \mathbf{x}_t^T \mathbf{R}_{xxt}\mathbf{x}_t \\ &+ \mathbf{a}_t^T \mathbf{R}_{axt}\mathbf{x}_t + \mathbf{a}_t^T \mathbf{R}_{at} + \mathbf{x}_t^T \mathbf{R}_{xt} \\ &= \delta\mathbf{a}_t^T \hat{\mathbf{R}}_{aat}\delta\mathbf{a}_t + \delta\mathbf{x}_t^T \hat{\mathbf{R}}_{xxt}\delta\mathbf{x}_t + \delta\mathbf{a}_t^T \hat{\mathbf{R}}_{axt}\delta\mathbf{x}_t \\ &+ \delta\mathbf{a}_t^T \hat{\mathbf{R}}_{at} + \delta\mathbf{x}_t^T \hat{\mathbf{R}}_{xt} + r_t(\bar{\mathbf{x}}_t, \bar{\mathbf{u}}_t)\end{aligned} \quad \text{Eq: 47}$$

## Computing the G-Function

Next we take the Bellman equation for the G-function

$$G_t^{\pi}(\mathbf{x}_t, \mathbf{a}_t) = \hat{R}_t(\mathbf{x}_t, \mathbf{a}_t) + \gamma \mathbb{E}_{t,\mathbf{a}}\left[F_{t+1}^{\pi}(\mathbf{x}_{t+1}\right] \quad \text{Eq: 48}$$

substitute Eqs.(46), (47) and (46), and equate coefficients in front of like powers of $\delta\mathbf{x}_t$ and $\delta\mathbf{a}_t$ in the left-hand side and the right-hand side of the resulting equation:

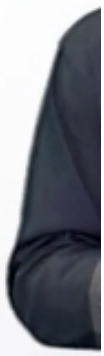
$$\mathbf{G}_{aat} = \mathbf{R}_{aat} + \mathbf{F}_{aat}, \quad \mathbf{G}_{xxt} = \mathbf{R}_{xxt} + \mathbf{F}_{xxt},$$

$$\mathbf{G}_{axt} = \mathbf{R}_{axt} + \mathbf{F}_{axt}$$

$$\mathbf{G}_{at} = \mathbf{R}_{at} + \mathbf{F}_{at}$$

$$\mathbf{G}_{xt} = \mathbf{R}_{xt} + \mathbf{F}_{xt},$$

$$g_t(\bar{\mathbf{x}}_t, \bar{\mathbf{u}}_t) = r_t(\bar{\mathbf{x}}_t, \bar{\mathbf{u}}_t) + f_{t+1}(\bar{\mathbf{x}}_t, \bar{\mathbf{u}}_t)$$



## Computing the F-Func

Use Eq.(39) to compute F-function for time t from the G-function:

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

Computing the Gaussian integral and simplifying, we obtain

$$F_t^{\pi}(\mathbf{x}_t) = \delta\mathbf{x}_t^T \mathbf{D}_t \delta\mathbf{x}_t + \delta\mathbf{x}_t^T \mathbf{H}_t + f_t(\bar{\mathbf{x}}_t) \qquad \text{Eq: 49}$$

where

$$\tilde{\Sigma}_a = \Sigma_a^{-1} - 2\beta\mathbf{G}_{aat}, \quad \sigma_a = \Sigma_a^{-1}\widehat{\delta\mathbf{a}}_t + \beta\mathbf{G}_{at}$$

$$\mathbf{D}_t = \mathbf{G}_{xxt} + \frac{\beta}{2}\mathbf{G}_{axt}^T \tilde{\Sigma}_a^{-1} \mathbf{G}_{axt}$$

$$\mathbf{H}_t = \mathbf{G}_{xt} + \mathbf{G}_{axt}^T \tilde{\Sigma}_a^{-1} \left(\Sigma_a^{-1}\widehat{\delta\mathbf{a}}_t + \beta\mathbf{G}_{axt}\right)$$

$$f_t(\bar{\mathbf{x}}_t) = -\frac{\log\left(|\Sigma_a|\,|\tilde{\Sigma}_a|\right) + \sigma_a^T \Sigma_a^{-1} \sigma_a}{2\beta} + g_t(\bar{\mathbf{x}}_t, \bar{\mathbf{u}}_t)$$

## Optimal Policy

As the reference distribution $\pi_0$ is Gaussian and the Q-function is quadratic, the optimal action policy $\pi$ is again Gaussian with a new mean and covariance:

$$\pi(\delta\mathbf{a}_t|\mathbf{x}_t) = \pi_0(\delta\mathbf{a}_t|\mathbf{x}_t)\, e^{\beta(G_t^\pi(\mathbf{x}_t,\mathbf{a}_t)-F_t^\pi(\mathbf{x}_t))}$$
$$= \mathcal{N}\left(\widehat{\delta\mathbf{a}}_t', \Sigma_a'\right) \qquad \text{Eq: 50}$$

where $\mathcal{N}(\cdot)$ is a multivariate Gaussian distribution with the following mean and covariance matrix:

$$\widehat{\delta\mathbf{a}}_t' = \Sigma_a'\left(\Sigma_a^{-1}\widehat{\delta\mathbf{a}}_t + \beta\mathbf{G}_{axt}\delta\mathbf{x}_t + \beta\mathbf{G}_{at}\right)$$
$$\Sigma_a' = \left[\Sigma_a^{-1} - 2\beta\mathbf{G}_{aat}\right]^{-1} \qquad \text{Eq: 51}$$

These relations can be viewed as Bayesian updates for the mean and variance of the optimal action policy.

## Trajectory Optimization

Iterative scheme that computes a sequence of trajectories that converges to a true solution of the dynamics of the system. (Similar to the Iterative Linear Quadratic-Gaussian Regulator (ILQG) method of Todorov and Li (2005)).

**The method:**
Start with an initial trajectory determined by the mean of the prior policy distribution.
Repeat until convergence:

**Step 1 (Backward pass):** Compute the G-function and update the mean using Eq.(51)

**Step 2 (Forward pass):** Generate a new trajectory using the new mean value

## Updated Policy

The updated policy now takes the form

$$\pi(\mathbf{a}_t|\mathbf{x}_t) = \mathcal{N}\left(\hat{\mathbf{A}}_0 + \hat{\mathbf{A}}_1\mathbf{x}_t, \Sigma_M\right) \qquad \text{Eq: 52}$$

where variance $\Sigma_M$ is found from Eq.(51):

$$\Sigma_M = \left[\Sigma_a^{-1} - 2\beta\mathbf{G}_{aat}\right]^{-1} \qquad \text{Eq: 53}$$

## 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 $\pi(\mathbf{a}_t|\mathbf{x}_t)$ by updating its mean and variance
5. Construct a new trajectory using the updated mean

## 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

$$LL(\Theta) = -\log \prod_{t=0}^{T-1} \frac{e^{-\frac{1}{2}\left(\mathbf{a}_t - \hat{\mathbf{A}}_0 - \hat{\mathbf{A}}_1 \mathbf{x}_t\right)^T \Sigma_M^{-1} \left(\mathbf{a}_t - \hat{\mathbf{A}}_0 - \hat{\mathbf{A}}_1 \mathbf{x}_t\right)}}{\sqrt{(2\pi)^N |\Sigma_M|}} \quad \text{Eq: 54}$$

where $\mathbf{x}_t$ are observed optimal investments.

$\hat{\mathbf{A}}_0 = \hat{\mathbf{A}}_0^{(k)}$ and $\hat{\mathbf{A}}_1 = \hat{\mathbf{A}}_1^{(k)}$ are computed from the backward pass on iteration $k$, and $\Sigma_M$ is given by Eq.(56).

All unknown parameters $\Theta = (\lambda, \mu_i, \beta)$ can then be computed using Gradient Descent or Stochastic Gradient Descent.

## The Final Scheme: IRL Case

**RL case:** rewards are **not** 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. **IRL: estimate the reward** at time $t$ using Maximum Likelihood
3. Use the expected next-period F-function and **estimated** rewards to update the Q-function
4. Compute the value of the F-function at time $t$
5. Recompute the policy distribution $\pi(\mathbf{a}_t|\mathbf{x}_t)$ by updating its mean and variance
6. Construct a new trajectory using the updated mean

The model can be used in two IRL settings:

- As a model of a particular trader - needs proprietary data
- As a model for the market portfolio - uses only public data
- As a model for the market portfolio with private signals, similar to the Black-Litterman model.

"Market-implied" optimal policy is:

$$\pi(\mathbf{a}_t|\mathbf{x}_t) = \mathcal{N}\left(\hat{\mathbf{A}}_0 + \hat{\mathbf{A}}_1\mathbf{x}_t, \Sigma_M\right)$$

with

$$\Sigma_M = \left[\Sigma_a^{-1} - 2\beta\mathbf{G}_{aat}\right]^{-1}$$

# Summary

## In this course:

- We analyzed two most fundamental problems of Quantitative Finance, and learnt that not only Reinforcement Learning can be used for these problems, but rather these problems themselves can be formulated in terms of Reinforcement Learning tasks!
- We studied about the methods of Reinforcement Learning such as Q-Learning and Fitted Q-Iteration, and noted how they apply to this model.

## In this course:

- We analyzed very large class of portfolio optimization problems
  - We learnt Stochastic Policies
  - We learnt Entropy-Regularized Reinforcement Learning
- We applied Inverse Reinforcement Learning with the same dynamic portfolio model
- We used Inverse Reinforcement Learning to infer market views and values of private signals