Reinforcement Learning for DVA Hedging

Giorgio VIT

DVA

Debt Value Adjustment (DVA)

The DVA is the risk that the bank defaults and does not pay the derivative to the client. At time t, the DVA is given by:

$$\mathrm{DVA}(t) = \mathbb{E}_t^Q[\mathrm{LGD}_B \, 1_{\{\tau_B \le T\}} \, 1_{\{\tau_B < \tau_C\}} \, D(t, \tau_B) \, (V_0(\tau_B))^-]$$

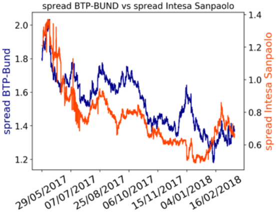
DVA generated by the liability represented by a single cash flow N that the bank must pay at time T (5Y rolling) $\Rightarrow V_0(\tau_B)^- = N \, D(\tau_B, T)$

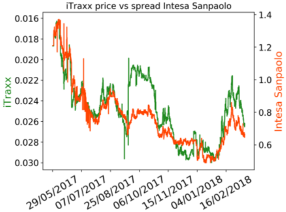
$$\boxed{DVA(t) = LGD_B \cdot N \cdot \left(1 - e^{-\frac{spread_t^{5y}}{LGD_B}(T-t)}\right)}$$

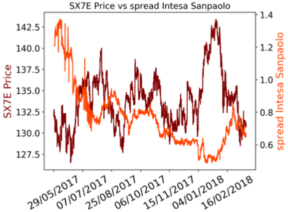
DVA Hedging

Possible trades:

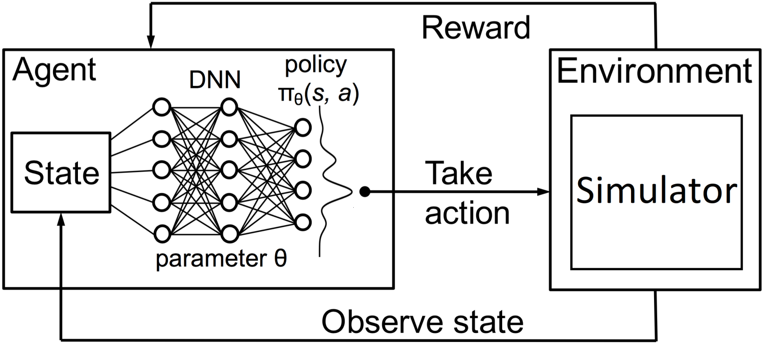
1. *BTP Spread Trade:* purchase/sale of 10y BTP futures and simultaneous sale/purchase of 10y Bund Futures.
2. 5y iTraxx Financial Senior (FinSen) CDS index.
3. Futures on the Eurostoxx Banks SX7E.







Agent-Environment Interaction



Natural Policy Optimization (NPO)

input: initial policy parameterization θ_0.
return: optimal policy parameters $\theta^* = \theta_{m+1}$.

while *policy parameterization* $\theta_m \approx \theta_{m+1}$ *converges* **do**
  obtain policy gradient $\nabla J(\theta_m)$ from estimator
  update policy $\theta_{m+1} = \theta_m + \alpha_m \, \mathbf{F}_\theta^{-1} \, \nabla J(\theta_m)$
end

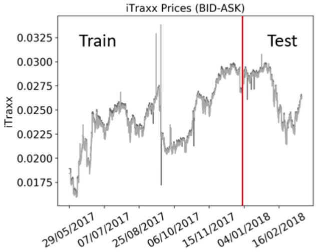
$$\nabla J(\theta) = \int_{\mathcal{T}} \nabla_\theta \mathcal{P}_\theta(\tau) \, r(\tau) \, d\tau = \mathbb{E}[\nabla_\theta log \mathcal{P}_\theta(\tau) \, r(\tau)] \approx \left\langle \sum_{k=0}^{H} \left(\sum_{l=k}^{H} \nabla_\theta log \pi_\theta(A_k|s_k) \right) (r_l - b) \right\rangle$$

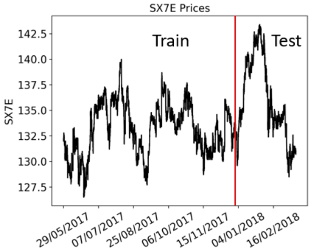
$$\mathbf{F}_\theta = \mathbb{E}_{d^\pi(s)} \left[\mathbb{E}_{\pi(a;s,\theta)} \left[\frac{\partial log \pi(a;s,\theta)}{\partial \theta_i} \frac{\partial log \pi(a;s,\theta)}{\partial \theta_j} \right] \right] \approx \langle \sum_{k=0}^{H} (\sum_{l=0}^{k} \nabla_\theta log \pi_\theta(A_l|s_l)) \nabla_\theta log \pi_\theta(A_k|s_k)^{\mathbf{T}} \rangle$$

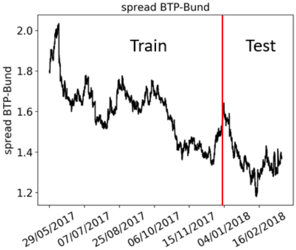
Algorithm: General setup for NPO.

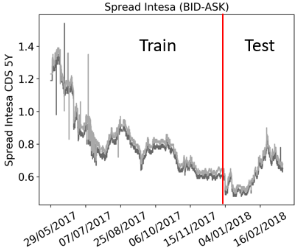
Dataset

n records per day = 96
n days train = 130
n days test = 43
n records train = 12480
n records test = 4128









Baseline

Baseline

A baseline is a simple strategy that is used to measure the performance of our RL agent's policy.

BTP Baseline

$$\begin{cases} A_{k+1}^{BTP} = -\dfrac{D_k^{DVA}}{d_k^{BTP}} - L_k^{BTP} \\ A_{k+1}^{iTraxx} = 0 \\ A_{k+1}^{SX7E} = 0 \end{cases}$$

iTraxx Baseline

$$\begin{cases} A^{BTP} = 0 \\ A^{iTraxx} = -2\dfrac{D^{DVA}}{d^{iTraxx}} - L^{iTraxx} \\ A^{SX7E} = 0 \end{cases}$$

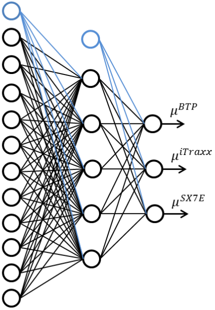
BTP-iTraxx Baseline

$$\begin{cases} A^{BTP} = -\dfrac{1}{2}\dfrac{D^{DVA}}{d^{BTP}} - L^{BTP} \\ A^{iTraxx} = -\dfrac{D^{DVA}}{d^{iTraxx}} - L^{iTraxx} \\ A^{SX7E} = 0 \end{cases}$$

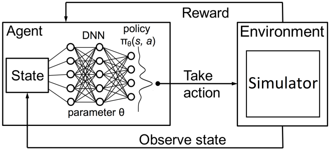
State - Action

$s_i = [baseline\_features_i,\ total\_allocation_i,\ price_i]$

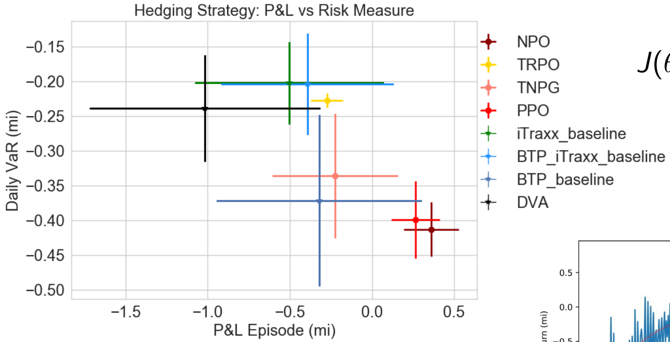
1. $baseline\_features_i$:
$\left[\frac{D_i^{DVA}}{d_i^{BTP}}, \frac{D_i^{DVA}}{d_i^{iTraxx}}\right]$
2. $total\_allocation_i$:
$\left[L_i^{SX7E}, L_i^{BTP}, L_i^{Bund}, L_i^{ITRAXX}, \psi_i^0\right]$
3. $prices_i$:
$\left[X_i^{SX7E}, s_i^{BTP-Bund}, X_i^{iTraxx}, \pi_i^{5y}\right]$



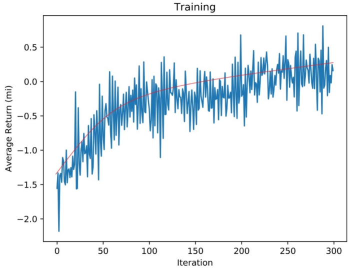
$num\ param = 11 \cdot 5 + 5 \cdot 3 + 11 + 3 = 84$



NPO, TRPO, PPO, TNPG - Train



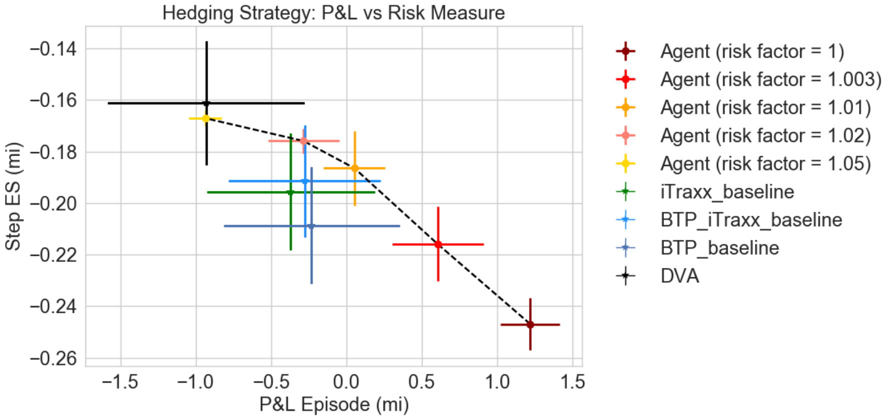
$$J(\theta) = \mathbb{E}_{\pi_\theta}[\sum_{k=0}^{H} P\&L_k]$$



NPO. Average return during training.

Efficient Frontier - NPO - Train

$$R(x) = \begin{cases} x & \text{if } x \geq 0 \text{ } or \text{ } x < -0.5\, mi \\ \left(1 - (1-x)^{rf}\right) & \text{if } x < 0 \end{cases} \Rightarrow J(\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{k=0}^{H} R(P\&L_k)\right]$$



Efficient Frontier - NPO - Test

$$R(x) = \begin{cases} x & \text{if } x \geq 0 \text{ or } x < -0.5\, mi \\ \left(1-(1-x)^{rf}\right) & \text{if } x < 0 \end{cases} \Rightarrow J(\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{k=0}^{H} R(P\&L_k)\right]$$