A ADDITIONAL MATERIAL FOR SECTION 3

A.1 PROOF OF PROPOSITION 3.1

We start by proving first that given any π^* that satisfies (3), it must also satisfy

$$\pi^* \in \arg\min_{\pi} \mathbb{E}_{(t,s)\sim\beta}[Q_t^{\pi^*}(s, \pi_t(s))], \tag{5}$$

where β captures the distribution of $(\tilde{t}, s_{\tilde{t}})$ used in (3). We do so by contradiction. Let's assume that there exists a $\bar{\pi}$ such that

$$\mathbb{E}_{(t,s)\sim\beta}[Q_t^{\pi^*}(s, \bar{\pi}_t(s))] < \mathbb{E}_{(t,s)\sim\beta}[Q_t^{\pi^*}(s, \pi_t^*(s))].$$

Then, one can design the following policy:

$$\bar{\pi}_t^*(s) := \begin{cases} \bar{\pi}_t(s) & \text{if } Q_t^{\pi^*}(s, \bar{\pi}_t(s)) < Q_t^{\pi^*}(s, \pi_t^*(s)) \\ \pi_t^*(s) & \text{otherwise.} \end{cases}$$

Using a recursive argument, one can show that $Q_t^{\bar{\pi}^*}(s_t, a_t) \leq Q_t^{\pi^*}(s_t, a_t)$ for all t and (s_t, a_t) pair. In this recursion, we start with:

$$Q_T^{\bar{\pi}^*}(s_T, a_T) = -r_T(s_T, a_T, s_T) = Q_T^{\pi^*}(s_T, a_T).$$

Moreover, for all $t < T$, and (s_t, a_t) pairs, we have that:

$$\begin{aligned} Q_t^{\bar{\pi}^*}(s_t, a_t) &= \bar{\rho}(-r_t(s_t, a_t, s_{t+1}) + Q_{t+1}^{\bar{\pi}^*}(s_{t+1}, \bar{\pi}^*(s_{t+1}))|s_t) \\ &\leq \bar{\rho}(-r_t(s_t, a_t, s_{t+1}) + Q_{t+1}^{\pi^*}(s_{t+1}, \bar{\pi}^*(s_{t+1}))|s_t) \\ &\leq \bar{\rho}(-r_t(s_t, a_t, s_{t+1}) + Q_{t+1}^{\pi^*}(s_{t+1}, \pi^*(s_{t+1}))|s_t) = Q_t^{\pi^*}(s_t, a_t), \end{aligned}$$

where we first used $Q_{t+1}^{\bar{\pi}^*}(s_t, a_t) \leq Q_{t+1}^{\pi^*}(s_t, a_t)$, then exploited the definition of $\bar{\pi}_t^*$. With this result in hand we can obtain that for all t and s_t

$$Q_t^{\bar{\pi}^*}(s_t, \bar{\pi}_t^*(s_t)) \leq Q_t^{\pi^*}(s_t, \bar{\pi}_t^*(s_t)) \leq Q_t^{\pi^*}(s_t, \bar{\pi}_t(s_t)),$$

where we again used the definition of $\bar{\pi}^*$. Finally, we must therefore have that:

$$\mathbb{E}_{(t,s)\sim\beta}[Q_t^{\bar{\pi}^*}(s, \bar{\pi}_t^*(s))] \leq \mathbb{E}_{(t,s)\sim\beta}[Q_t^{\pi^*}(s, \bar{\pi}_t(s))] < \mathbb{E}_{(t,s)\sim\beta}[Q_t^{\pi^*}(s, \pi_t^*(s))]$$

which leads to a contradiction, hence (5) must hold.

Next, applying the interchangeability property (see Shapiro (2017)) to equation (5) and using the fact that the β distribution puts positive probability on all time periods and all sub-regions of $\mathcal{S} \times \mathcal{A}$, we know that the following necessarily hold:

$$\pi_t^*(s) \in \arg\min_{a} Q_t^{\pi^*}(s, a), \ \forall s \in \mathcal{S}, \forall t \in \{0, \ldots, T\}.$$

Our last step involves using recursion to show that $\pi^* \in \arg\min_\pi Q_t^\pi(s_t, \pi_t(s_t))$ for all t and all s_t. We start once more at $t = T$ where for all s_T:

$$Q_T^{\pi^*}(s_T, \pi_T^*(s)) = \min_{a_T} Q_T^{\pi^*}(s_T, a_T) = \min_{a_T} -r_T(s_T, a_T, s_T) \leq Q_T^{\pi}(s_T, \pi_T(s_T)), \quad \forall \pi.$$

And then recursively for all $t < T$ and all s_t,

$$\begin{aligned} Q_t^{\pi^*}(s_t, \pi_t^*(s_t)) &= \min_{a_t} Q_t^{\pi^*}(s_t, a_t) \\ &= \min_{a_t} \bar{\rho}(-r_t(s_t, a_t, s_{t+1}) + Q_{t+1}^{\pi^*}(s_{t+1}, \pi_{t+1}^*(s_{t+1}))|s_t) \\ &\leq \min_{a_t} \bar{\rho}(-r_t(s_t, a_t, s_{t+1}) + Q_{t+1}^{\pi}(s_{t+1}, \pi_{t+1}(s_{t+1}))|s_t) \ \forall \pi \\ &\leq \bar{\rho}(-r_t(s_t, \pi_t(s_t), s_{t+1}) + Q_{t+1}^{\pi}(s_{t+1}, \pi_{t+1}(s_{t+1}))|s_t) \ \forall \pi \\ &\leq \min_{\pi} Q_t^{\pi}(s_t, \pi_t(s_t)). \quad \square \end{aligned}$$

A.2 ADAPTING DDPG TO HANDLE DYNAMIC EXPECTILE RISK MEASURES

We include below the extension of deep deterministic policy gradient (DDPG) algorithm to a risk averse MDP that employs a dynamic expectile risk measure. In **bold** we highlight the modification to DDPG that is due to the use of a dynamic expectile risk measure. Note that after assuming that the information about t is included in the state, we drop the subscript t notation to increase similarity with Lillicrap et al. (2015). For completeness, we make precise that the original DDPG uses $\partial\ell(y) := 2y$ while this risk averse DDPG, with risk level τ, uses $\partial\ell(y) := 2(\tau \max(0, y) - (1-\tau)\max(0, -y))$.

Algorithm 2: Risk averse deep deterministic policy gradient

Randomly initialize the main actor and critic networks' parameters θ^{π} and θ^{Q};
Initialize the target actor, $\theta^{\pi'} \leftarrow \theta^{\pi}$, and critic, $\theta^{Q'} \leftarrow \theta^{Q}$, networks;
Initialize replay buffers R;
for $j = 1 : \#Episodes$ **do**
  Initialize a random process $\mathcal{N}$ for action exploration;
  Receive initial observation state s_0;
  for $t = 0 : T-1$ **do**
    Select action $a_t = \pi_t(s_t|\theta^{\pi}) + \mathcal{N}_t$;
    Execute a_t and observe reward r_t and new state s_{t+1};
    Store transition (s_t, a_t, r_t, s_{t+1}) in R;
    Sample a minibatch of N transitions $\{(s_j, a_j, r_j, s_{j+1})\}_{j=1}^{N}$ in R;
    Set the realized losses $y_j^i := -r_j^i + Q(s_{j+1}^i, \pi(s_{j+1}^i|\theta^{\pi'})|\theta^{Q'})$;
    Update the main critic network:

$$\boldsymbol{\theta^{Q} \leftarrow \theta^{Q} - \alpha \frac{1}{N} \sum_{i=1}^{N} \partial\ell(Q(s_j^i, a_j^i|\theta^{Q}) - y_j^i) \nabla_{\theta^{Q}} Q(s_j^i, a_j^i|\theta^{Q})}$$

    where $\partial\ell(y) := \tau \max(0, y) - (1-\tau)\max(0, -y)$;
    Update the main actor network:

$$\theta^{\pi} \leftarrow \theta^{\pi} - \alpha \frac{1}{N} \sum_{i=1}^{N} \nabla_a Q(s_j^i, a|\theta^{Q})|_{a=\pi(s_j^i|\theta^{\pi})} \nabla_{\theta^{\pi}} \pi(s_j^i|\theta^{\pi}) ;$$

    Update the target networks:

$$\theta^{Q'} \leftarrow \alpha\theta^{Q} + (1-\alpha)\theta^{Q'}, \qquad \theta^{\pi'} \leftarrow \alpha\theta^{\pi} + (1-\alpha)\theta^{\pi'} ;$$

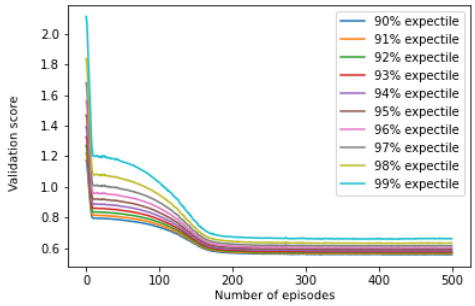
  end
end

B ADDITIONAL MATERIAL FOR SECTION 4

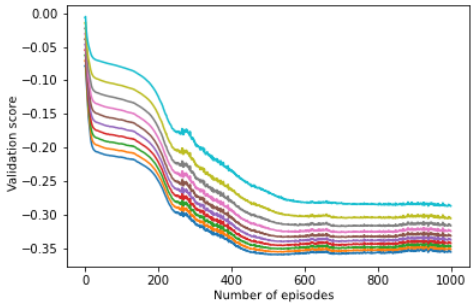
Table 3: Stock data including the mean, standard deviation, and the correlation matrix

	AAPL	AMZN	FB	JPM	GOOGL
S_0	78.81	1877.94	221.77	137.25	1450.16
μ	-0.0015	-0.0017	-0.0001	0.0006	-0.0004
σ	0.0298	0.0243	0.0295	0.0345	0.0246
AAPL	1.0000	0.7133	0.7744	0.5383	0.7680
AMZN	0.7133	1.0000	0.6903	0.2685	0.6837
FB	0.7744	0.6903	1.0000	0.4807	0.8054
JPM	0.5383	0.2685	0.4807	1.0000	0.6060
GOOGL	0.7680	0.6837	0.8054	0.6060	1.0000

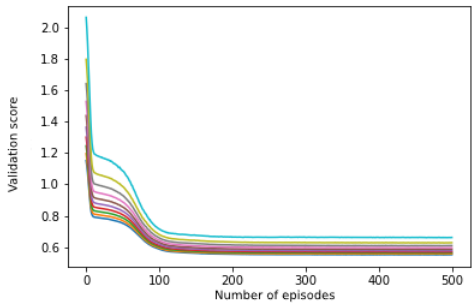
B.1 ADDITIONAL MATERIAL FOR SECTION 4.2



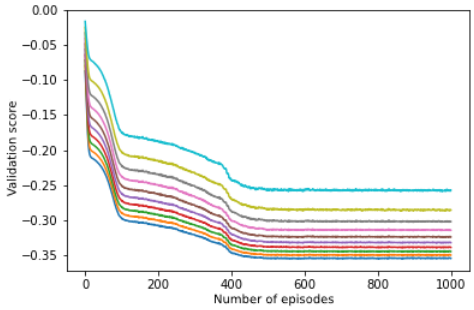
(a) ACRL for DRM's writer



(b) ACRL for DRM's buyer

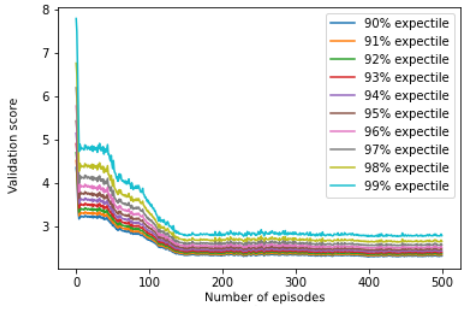


(c) AORL for SRM's writer

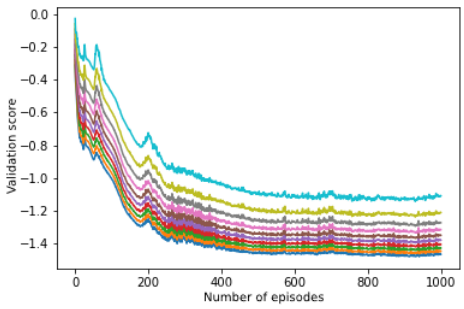


(d) AORL for SRM's buyer

Figure 2: Learning curves of the DRM and SRM for an at-the-money vanilla call option on AAPL when a 90% expectile measure is used. The graphs show the validation scores for a range of static expectile measures with risk level ranging from 90% to 99%.

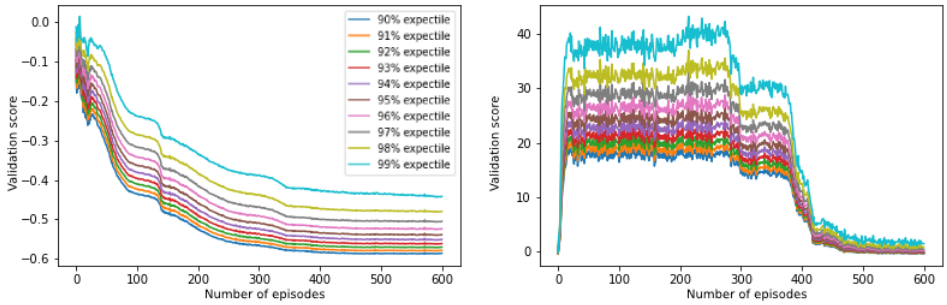


(a) Writer



(b) Buyer

Figure 3: Learning curves of the ACRL algorithm for the writer and buyer's DRM for a basket at-the-money call option over AAPL, AMZN, FB, JPM, and GOOGL at the risk level $\tau = 90\%$. The graphs show the validation scores for a range of static expectile measures with risk level ranging from 90% to 99%.

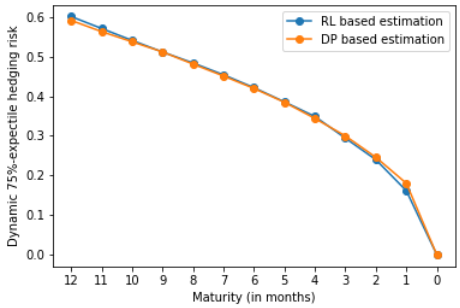


(a) With immediate rewards

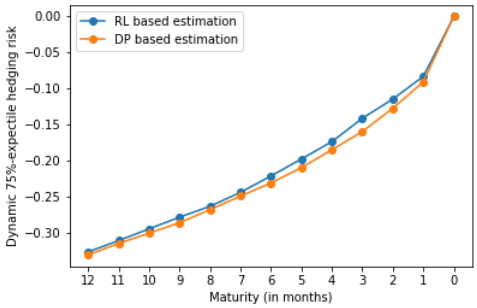
(b) With delayed rewards

Figure 4: Learning curves of the ACRL algorithm for the buyer's DRM when using (a) the immediate rewards versus (b) delayed rewards in the hedging of a vanilla call at-the-money option.

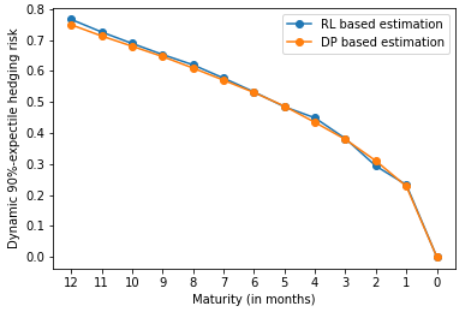
B.2 ADDITIONAL MATERIAL FOR SECTION 4.3



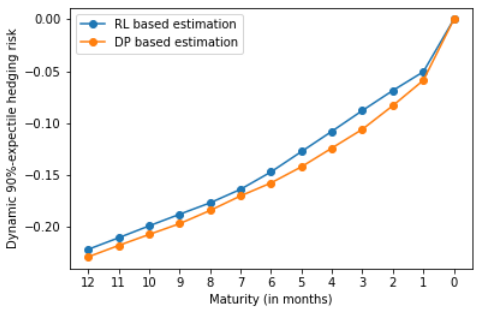
(a) Writer, $\tau = 75\%$



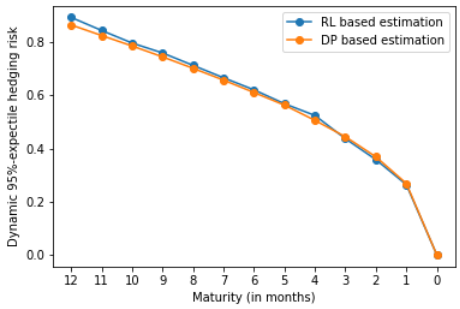
(b) Buyer, $\tau = 75\%$



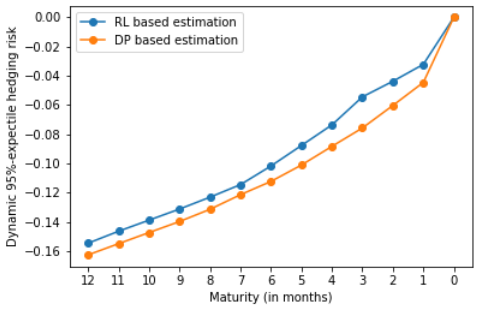
(c) Writer, $\tau = 90\%$



(d) Buyer, $\tau = 90\%$

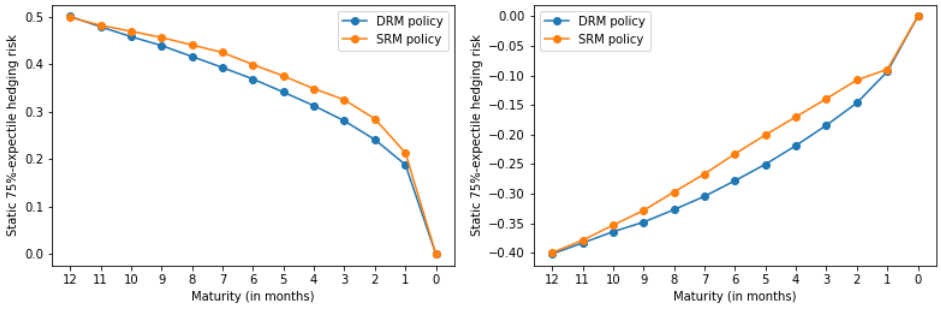


(e) Writer, $\tau = 95\%$



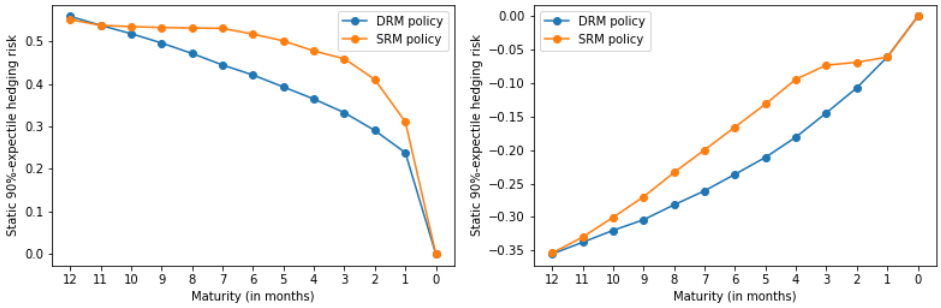
(f) Buyer, $\tau = 95\%$

Figure 5: The out-of-sample dynamic risk imposed to the two sides of a vanilla at-the-money call option over AAPL (with maturity ranging from 12 months to 0 months) under the DRM policy trained for a 12 months maturity and at different risk levels $\tau \in \{75\%, 90\%, 95\%\}$.



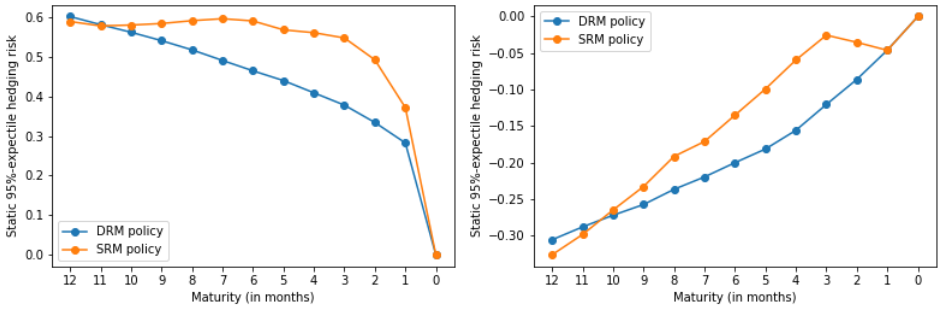
(a) Writer, $\tau = 75\%$

(b) Buyer, $\tau = 75\%$



(c) Writer, $\tau = 90\%$

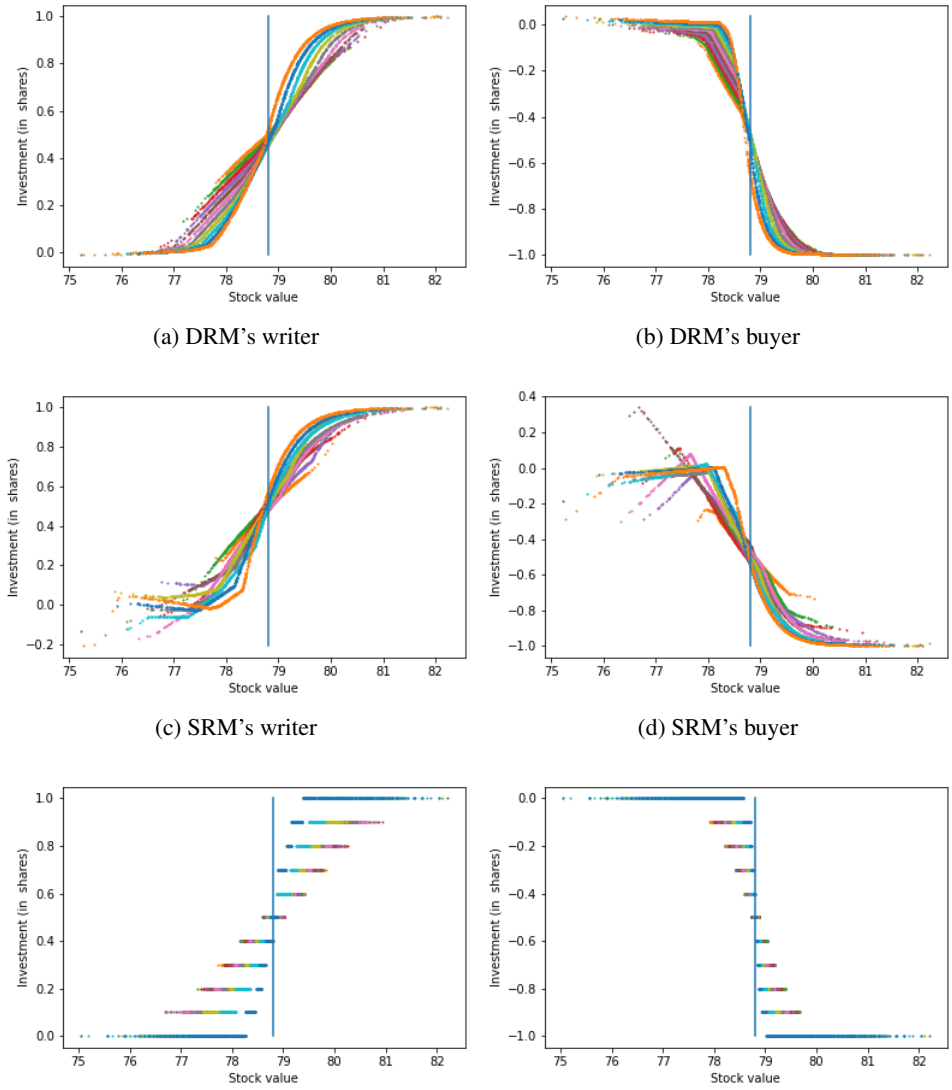
(d) Buyer, $\tau = 90\%$



(e) Writer, $\tau = 95\%$

(f) Buyer, $\tau = 95\%$

Figure 6: The out-of-sample static risk imposed to the two sides of a vanilla at-the-money call option over AAPL (with maturity ranging from 12 months to 2 months) under the DRM and SRM policies trained for a 12 months maturity and at different risk levels $\tau \in \{75\%, 90\%, 95\%\}$.



(a) DRM's writer

(b) DRM's buyer

(c) SRM's writer

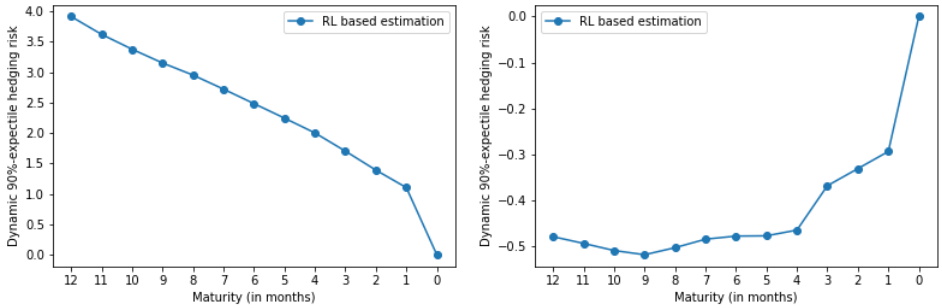
(d) SRM's buyer

(e) DP-DRM's writer

(f) DP-DRM's buyer

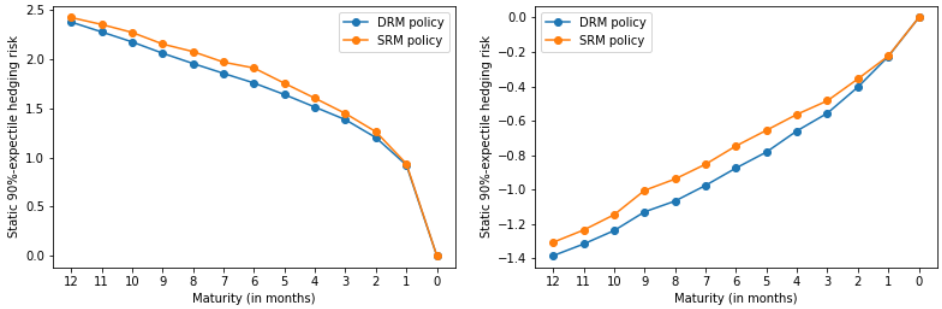
Figure 7: Comparison of the optimal DRL policies obtained for DRM and SRM (with 90% expectile measures) to the discretized DP solution (DP-DRM) for an at-the-money vanilla call option on AAPL with a one year maturity. Each figure presents the sampled actions in our simulated trajectories as a function of the AAPL stock value. The strike price is marked at 78.81.

B.3 ADDITIONAL MATERIAL FOR SECTION 4.4



(a) Writer (b) Buyer

Figure 8: The out-of-sample dynamic risk imposed to the two sides of a basket at-the-money call option over AAPL, AMZN, FB, JPM, and GOOGL at the risk level $\tau = 90\%$ (as maturity ranges from 12 to 0 months) under a DRM policy trained for a 12 months maturity.



(a) Static risk, writer (b) Static risk, buyer

Figure 9: The out-of-sample static risk imposed to the two sides of a basket at-the-money call option over AAPL, AMZN, FB, JPM, and GOOGL at the risk level $\tau = 90\%$ (as maturity ranges from 12 to 0 months) under the DRM and SRM policies trained for a 12 months maturity.