A. Proofs

A.1. Main Lemma

The following lemma will be helpful in the proofs of all our Theorems.

Lemma 2. *Suppose we have a family of stochastic processes $(M_t(m))_{t=0}^{\infty}$ indexed by $m \in [0,1]$ and further assume the process $(M_t(\mu))_{t=0}^{\infty}$ for some $\mu \in [0,1]$ is a non-negative martingale with respect to a filtration $\mathcal{F}_t$ (i.e. $\mathbb{E}[M_t|\mathcal{F}_{t-1}] = M_{t-1}$ for $t \geq 1$) with initial value $M_0 = 1$. Then for any given $\alpha \in [0,1]$ the sequence of sets $C_t = \left\{m : M_t(m) \leq \frac{1}{\alpha}\right\}$ is a $(1-\alpha)$ confidence sequence for μ and so is its running intersection $\bigcap_{i=1}^{t} C_i$.*

Proof. For the first part, by the definition of a CS it suffices to show that $\Pr\left(\exists t \in \mathbb{N} : \mu \notin C_t\right) \leq \alpha$ or

$$\Pr\left(\exists t \in \mathbb{N} : \mu \notin \left\{m : M_t(m) \leq \frac{1}{\alpha}\right\}\right) \leq \alpha.$$

An error occurs only if $M_t(\mu)$ exceeds $1/\alpha$ at any point. This means that it suffices to show that

$$\Pr\left(\exists t \in \mathbb{N} : M_t(\mu) \geq \frac{1}{\alpha}\right) \leq \alpha,$$

which is true by Ville's inequality (Ville, 1939) since $M_t(\mu)$ is a non-negative martingale with initial value 1.

For the second part, we need to show that

$$\Pr\left(\exists t \in \mathbb{N} : \mu \notin \bigcap_{s=1}^{t} \left\{m : M_s(m) \leq \frac{1}{\alpha}\right\}\right) \leq \alpha.$$

This reduces to showing

$$\Pr\left(\exists t \in \mathbb{N} : \exists s \in \{1,\ldots,t\} : M_s(\mu) \geq \frac{1}{\alpha}\right) \leq \alpha,$$

which further simplifies to

$$\Pr\left(\exists t \in \mathbb{N} : M_t(\mu) \geq \frac{1}{\alpha}\right) \leq \alpha,$$

and this is again implied by Ville's inequality. □

A.2. Proof of Theorem 1

Proof. Consider the filtration $(\mathcal{F}_t)_{t=0}^{\infty}$ generated by the sequence of sigma-fields $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \ldots$ with $\mathcal{F}_0$ the trivial sigma-field and $\mathcal{F}_t = \sigma((w_0,r_0),(w_1,r_1),\ldots,(w_t,r_t))$. It suffices to show that our betting ensures that $K_t(V(\pi))$ is a non-negative martingale with initial value 1 as we can then apply lemma 2. $K_0(v) = 1$ is by the definition of the process (we start with a wealth of 1), and $K_t(v) \geq 0$ for all $v \in [0,1]$ because our bets are in the set $\mathcal{D}_v^0$ (c.f. eq (4)). Thus it remains to show $\mathbb{E}\left[K_t\left(V(\pi)\right)|\mathcal{F}_{t-1}\right] = K_{t-1}(V(\pi))$. We have the following chain of equalities

$$\begin{aligned}
\mathbb{E}\left[K_t\left(V(\pi)\right)|\mathcal{F}_{t-1}\right] &= \mathbb{E}\left[K_{t-1}\left(1+\lambda_{1,t}(w_t-1)+\lambda_{2,t}(w_t r_t - V(\pi))\right)|\mathcal{F}_{t-1}\right] \\
&= K_{t-1}\mathbb{E}\left[1+\lambda_{1,t}(w_t-1)+\lambda_{2,t}(w_t r_t - V(\pi))|\mathcal{F}_{t-1}\right] \\
&= K_{t-1}\left(1+\mathbb{E}\left[\lambda_{1,t}(w_t-1)|\mathcal{F}_{t-1}\right]+\mathbb{E}\left[\lambda_{2,t}(w_t r_t - V(\pi))|\mathcal{F}_{t-1}\right]\right) \\
&= K_{t-1}\left(1+\lambda_{1,t}\mathbb{E}\left[(w_t-1)|\mathcal{F}_{t-1}\right]+\lambda_{2,t}\mathbb{E}\left[(w_t r_t - V(\pi))|\mathcal{F}_{t-1}\right]\right) \\
&= K_{t-1}\left(1+\lambda_{1,t}\cdot 0+\lambda_{2,t}\cdot 0\right) = K_{t-1}
\end{aligned}$$

where we have used that K_{t-1}, $\lambda_{1,t}$, $\lambda_{2,t}$ are measurable with respect to $\mathcal{F}_{t-1}$ and that $\mathbb{E}[w] = 1$ and $\mathbb{E}[wr] = V(\pi)$. □

A.3. Proof of Lemma 1

Proof. Consider the function $f(x) = \ln(1+x) - x - \psi x^2$ with domain $\left[-\frac{1}{2},\infty\right)$. Note that $f\left(-\frac{1}{2}\right) = 0$ and $\lim_{x\to\infty} f(x) = \infty$. Furthermore f has two critical points: 0 and $-\frac{2\psi+1}{2\psi}$. But $f(0) = 0$ and $f\left(-\frac{2\psi+1}{2\psi}\right) > 0$ so we conclude that $f(x) \geq 0$ for all $x \geq -\frac{1}{2}$. □

A.4. Proof of Theorem 2

Proof. We will first show that $K_t^{\pm}(V(\pi))$ is a non-negative martingale with initial value 1. Consider the same filtration as for Theorem 1. Note that $K_0^{\pm}(v) = 1$ is by the definition of the process (we start with a wealth of 1). We analyze $K_t^+(v)$ and $K_t^-(v)$ separately. Note that $K_t^+(v) \geq 0$ for all $v \in [0,1]$ because our bets are in the set $\mathcal{C} \subset \mathcal{D}_v^0$ (c.f. eq (4)). For $K_t^-(v)$ we note that the process is isomorphic to a process similar to $K_t^+(v)$ but with the reward and v redefined. Thus our bets keep $K_t^-(v) \geq 0$. We now show the equality

$$\mathbb{E}\left[K_t^-\left(V(\pi)\right)|\mathcal{F}_{t-1}\right] = K_{t-1}^-(V(\pi)),$$

as the equality $\mathbb{E}\left[K_t^+\left(V(\pi)\right)|\mathcal{F}_{t-1}\right] = K_{t-1}^+(V(\pi))$ is exactly what was shown in Theorem 1. We have

$$\begin{aligned}
\mathbb{E}\left[K_t^-\left(V(\pi)\right)|\mathcal{F}_{t-1}\right] &= \mathbb{E}\left[K_{t-1}^-\left(1+\lambda_{1,t}^-(w_t-1)+\lambda_{2,t}^-(w_t(1-r_t)(1-V(\pi)))\right)|\mathcal{F}_{t-1}\right] \\
&= K_{t-1}^-\mathbb{E}\left[1+\lambda_{1,t}^-(w_t-1)+\lambda_{2,t}^-(w_t-1+V(\pi)-w_tr_t)|\mathcal{F}_{t-1}\right] \\
&= K_{t-1}^-\left(1+(\lambda_{1,t}^-+\lambda_{2,t}^-)\mathbb{E}\left[(w_t-1)|\mathcal{F}_{t-1}\right]+\lambda_{2,t}^-\mathbb{E}\left[(V(\pi)-w_tr_t)|\mathcal{F}_{t-1}\right]\right) \\
&= K_{t-1}^-\left(1+(\lambda_{1,t}^-+\lambda_{2,t}^-)\cdot 0+\lambda_{2,t}^-\cdot 0\right) = K_{t-1}^-.
\end{aligned}$$

Therefore $\frac{1}{2}\left(K_t^+\left(V(\pi)\right)+K_t^-\left(V(\pi)\right)\right)$ is also a non-negative martingale with initial value 1. Applying Lemma 2 finishes the proof of the theorem. □

A.5. Proof of Theorem 3

Proof. We note that the proof below works for a sequence of predictable functions $q_t(x,a)$ but to reduce notation we use $q(x,a)$. Consider the filtration $(\mathcal{F}_t)_{t=0}^{\infty}$ generated by the sequence of sigma-fields $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \ldots$ with $\mathcal{F}_0$ the trivial sigma-field and $\mathcal{F}_t = \sigma((x_1,a_1,r_1),\ldots,(x_t,a_t,r_t))$. Note that $K_0^q(v) = 1$ and $K_t^q(v) \geq 0$ for all $v \in [0,1]$ because our bets are in the set $\mathcal{C}^q$. Thus it remains to show $\mathbb{E}\left[K_t^q\left(V(\pi)\right)|\mathcal{F}_{t-1}\right] = K_{t-1}(V(\pi))$. We have the following chain of equalities

$$\begin{aligned}
\mathbb{E}\left[K_t^q\left(V(\pi)\right)|\mathcal{F}_{t-1}\right] &= \mathbb{E}\left[K_{t-1}^q\left(1+\lambda_{1,t}(w_t-1)+\lambda_{2,t}(w_tr_t-c_t-V(\pi))\right)|\mathcal{F}_{t-1}\right] \\
&= K_{t-1}^q(1+\lambda_{1,t}\mathbb{E}\left[w_t-1|\mathcal{F}_{t-1}\right]+\lambda_{2,t}\mathbb{E}\left[w_tr_t-V(\pi)|\mathcal{F}_{t-1}\right]-\lambda_{2,t}\mathbb{E}\left[c_t|\mathcal{F}_{t-1}\right]) \\
&= K_{t-1}^q\left(1-\lambda_{2,t}\mathbb{E}\left[w_tq(x_t,a_t)-\sum_{a'}\pi(a';x_t)q(x_t,a')\middle|\mathcal{F}_{t-1}\right]\right) = K_{t-1}^q,
\end{aligned}$$

where we have used that K_{t-1}, $\lambda_{1,t}$, $\lambda_{2,t}$ are measurable with respect to $\mathcal{F}_{t-1}$ and that $\mathbb{E}[w]=1$ and $\mathbb{E}[wr]=V(\pi)$ as well as $\mathbb{E}_{x_t\sim D,a_t\sim h}\left[w_tq(x_t,a_t)\right] = \mathbb{E}_{x_t}\left[\sum_{a'}\pi(a';x_t)q(x_t,a')\right]$. Thus the claim for C_t^q and its running intersection can be shown by applying lemma 2. The claim for $C_t^{\pm q}$ is completely analogous using the ideas here and in the proof of Theorem 2. □

A.6. Proof of Theorem 4

Proof. Consider the same filtration as for Theorem 1. Note that $K_0^{gd}(v) = 1$ is by the definition of the process and that $K_t^{gd}(v) \geq 0$ for all $v \in [0,1]$ because our bets are in the set $\mathcal{G}_v^0$ (c.f. eq (17)). Finally, we have

$$\begin{aligned}
\mathbb{E}\left[K_t^{gd}\left(V(\pi)-V(h)\right)|\mathcal{F}_{t-1}\right] &= \mathbb{E}\left[K_{t-1}^{gd}\Big(1+\lambda_{1,t}(w_t-1)+\lambda_{2,t}\big(w_tr_t-r_t-(V(\pi)-V(h))\big)\Big)|\mathcal{F}_{t-1}\right] \\
&= K_{t-1}^{gd}\left(1+\lambda_{1,t}\mathbb{E}\left[w_t-1|\mathcal{F}_{t-1}\right]+\lambda_{2,t}\mathbb{E}\left[w_tr_t-V(\pi)|\mathcal{F}_{t-1}\right]-\lambda_{2,t}\mathbb{E}\left[r_t-V(h)|\mathcal{F}_{t-1}\right]\right) \\
&= K_{t-1}^{gd}\left(1+\lambda_{1,t}\cdot 0+\lambda_{2,t}\cdot 0-\lambda_{2,t}\cdot 0\right) = K_{t-1}^{gd}.
\end{aligned}$$

Therefore $K_t^{gd}(V(\pi)-V(h))$ is a non-negative martingale with initial value 1. Applying lemma 2 finishes the proof of the theorem. □

B. Avoiding Grid Search

We first lower bound each process separately, then lower bound the hedged process. We denote the bets for K^+ (respectively K^-) as λ^+, (resp. λ^-). From lemma 1 we have

$$\ln(K_t^+(v)) \geq \sum_{i=1}^{t-1} {\lambda_i^+}^\top b_i(v) + \psi \sum_i {\lambda_i^+}^\top A_i(v) \lambda_i^+,$$

and

$$\ln(K_t^-(v)) \geq \sum_{i=1}^{t-1} {\lambda_i^-}^\top b_i'(v') + \psi \sum_i {\lambda_i^-}^\top A_i'(v') \lambda_i^-,$$

where $v' = 1 - v$, $b_i'(v) = \begin{bmatrix} w_i - 1 \\ w_i(1 - r_i) - v \end{bmatrix}$ and $A_i'(v) = b_i'(v) b_i'(v)^\top$. For the Hedged process, using that for any a, b

$$\ln\left(\exp(a) + \exp(b)\right) \geq \max(a, b)$$

to first establish

$$\ln(K^\pm(v)) \geq \max(\ln(K^+(v)) - \ln(2), \ln(K^-(v)) - \ln(2))$$

and further bound each term in the maximum by the respective quadratic lower bound. We conclude that if a v achieves

$$\sum_{i=1}^{t-1} {\lambda_i^+}^\top b_i(v) + \psi \sum_i {\lambda_i^+}^\top A_i(v) \lambda_i^+ = \ln\left(\frac{2}{\alpha}\right),$$

or a $v' = 1 - v$ achieves

$$\sum_{i=1}^{t-1} {\lambda_i^-}^\top b_i'(v') + \psi \sum_i {\lambda_i^-}^\top A_i'(v') \lambda_i^- = \ln\left(\frac{2}{\alpha}\right),$$

then we also achieve $K_t^\pm(v) \geq \frac{1}{\alpha}$. In terms of v and v' these expressions are second degree equations and thus their real roots in $[0, 1]$ (if any) provide a safe bracketing of the confidence region $\{v : K_t^\pm(v) \leq 1/\alpha\}$. For K_t^+ let

$$C_t = \sum_{i=1}^{t-1} {\lambda_i^+}^\top \begin{bmatrix} w_i - 1 \\ w_i r_i \end{bmatrix}, \tag{18}$$

$$S_t = \sum_{i=1}^{t-1} {\lambda_i^+}^\top \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \tag{19}$$

$$Q_t = \sum_{i=1}^{t-1} \psi {\lambda_i^+}^\top \begin{bmatrix} (w_i - 1)^2 & (w_i - 1) w_i r_i \\ (w_i - 1) w_i r_i & w_i^2 r_i^2 \end{bmatrix} \lambda_i^+, \tag{20}$$

$$T_t = \sum_{i=1}^{t-1} \psi {\lambda_i^+}^\top \begin{bmatrix} 0 & -(w_i - 1) \\ -(w_i - 1) & -2 w_i r_i \end{bmatrix} \lambda_i^+, \tag{21}$$

$$U_t = \sum_{i=1}^{t-1} \psi {\lambda_i^+}^\top \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \lambda_i^+, \tag{22}$$

and define $C_t', S_t', Q_t', T_t', U_t'$ similarly by using λ_i^- instead of λ_i^+ and $1 - r_i$ instead of r_i. Then the largest real root v^+ of

$$C_t - S_t v + Q_t + T_t v + U_t v^2 - \ln\left(\frac{2}{\alpha}\right) = 0,$$

if it exists, satisfies $K_t^\pm(v^+) \geq \frac{1}{\alpha}$. Similarly we can obtain v' as the largest real root of the quadratic with $C_t', S_t', Q_t', T_t', U_t'$ in place of C_t, S_t, Q_t, T_t, U_t, if it exists. Then $v^- = 1 - v'$ satisfies $K_t^\pm(v^-) \geq \frac{1}{\alpha}$.

C. Details of the Scalar Betting Strategy

C.1. Elimination of a Bet

Since in the long term λ_1 should be 0 its purpose can only be as a hedge in the short-term. We formulate this by considering the worst case wealth reduction among three outcomes : $(w, r) = (w_{\max}, 1)$, $(w, r) = (w_{\max}, 0)$ and $w = 0$ with any reward. We choose λ_1 to maximize the wealth in the worst of these outcomes. Thus we set up a family of Linear Programs (LPs) parametrized by λ_2 and v and with optimization variables α and λ_1:

$$\begin{array}{lll} \text{maximize} & \alpha & \\ \text{subject to} & \alpha \le 1 + \lambda_1(w_{\max} - 1) + \lambda_2(w_{\max} - v) & (z_1) \\ & \alpha \le 1 + \lambda_1(w_{\max} - 1) - \lambda_2 v & (z_2) \\ & \alpha \le 1 - \lambda_1 - \lambda_2 v & (z_3), \end{array}$$

where the variable z_i in parentheses next to each constraint is the corresponding dual variable.

Theorem 5. *For any $v \in [0, 1]$ and any $\lambda_2 \in \mathbb{R}$, the optimal value of λ_1 in the above LP is $\lambda_1^* = \max(-\lambda_2, 0)$.*

Proof. The dual program is

$$\begin{array}{lll} \text{minimize} & (1 + \lambda_2(w_{\max} - v))z_1 + (1 - \lambda_2 v)z_2 + (1 - \lambda_2 v)z_3 & \\ \text{subject to} & z_i \ge 0 & i = 1, 2, 3 \\ & -(w_{\max} - 1)(z_1 + z_2) + z_3 = 0 & \\ & z_1 + z_2 + z_3 = 1. & \end{array}$$

Consider the following two dual feasible settings:

$$z_1 = 0, z_2 = \frac{1}{w_{\max}}, z_3 = \frac{w_{\max} - 1}{w_{\max}},$$

and

$$z_1 = \frac{1}{w_{\max}}, z_2 = 0, z_3 = \frac{w_{\max} - 1}{w_{\max}},$$

with corresponding dual objectives: $1 - \lambda_2 v$ and $1 - \lambda_2 v + \lambda_2$. From here we see that if $\lambda_2 > 0$ the former attains a better dual objective and is thus a better bound for the primal objective. When $\lambda_2 < 0$ the latter is better.

When $\lambda_2 > 0$, a primal feasible setting is $\alpha = 1 - \lambda_2 v, \lambda_1 = 0$. Furthermore this setting achieves the same objective as the first dual feasible setting so we conclude that these are the optimal primal and dual solutions when $\lambda_2 > 0$.

When $\lambda_2 < 0$, a primal feasible setting is $\alpha = 1 - \lambda_2 v + \lambda_2, \lambda_1 = -\lambda_2$. Furthermore this setting achieves the same objective as the second dual feasible setting so we conclude that these are the optimal primal and dual solutions when $\lambda_2 < 0$.

Finally when $\lambda_2 = 0$ the two cases give the same value for λ_1 so we conclude $\lambda_1 = \max(-\lambda_2, 0)$ for all $\lambda_2 \in \mathbb{R}$ (and $v \ge 0$). □

The theorem suggests that in a hedged strategy the wealth process eliminating low values of $V(\pi)$ should set $\lambda_1^> = 0$ because $\mathbb{E}[wr - v] > 0$ and thus $\lambda_2^> > 0$. The wealth process that eliminates high values of $V(\pi)$ on the other hand should have $\lambda_1 = -\lambda_2$ because $\mathbb{E}[wr - v] < 0$ and thus $\lambda_2 < 0$. Thus the two processes look like

$$\begin{aligned} K_t^>(v) &= \prod_{i=1} \left(1 + \lambda_{2,i}^>(w_i r_i - v)\right), \\ K_t^<(v) &= \prod_{i=1} \left(1 - \lambda_{2,i}^<(w_i - 1) + \lambda_{2,i}^< (w_i r_i - v)\right) = \prod_{i=1} \left(1 - \lambda_{2,i}^< (w_i(1 - r_i) - (1 - v))\right). \end{aligned}$$

In the main text we have redefined $\lambda_{2,i}^< := -\lambda_{2,i}^<$ for symmetry.

C.2. A Technical Lemma

The following result can be extracted from the proof of Proposition 4.1 in (Fan et al., 2015).

Lemma 3. *For $\xi \geq -1$ and $\lambda \in [0,1)$ we have*

$$\ln(1+\lambda\xi) \geq \lambda\xi + (\ln(1-\lambda)+\lambda)\cdot\xi^2. \tag{23}$$

Proof. Note that $\lambda\xi \geq -\lambda > -1$. For $x > -1$ the function $f(x) = \frac{\ln(1+x)-x}{x^2}$ is increasing in x, therefore $f(\lambda\xi) \geq f(-\lambda)$. Rearranging leads to the statement of the lemma. □

We will be using this lemma with bets $\lambda \in [0,1)$ and $\xi_i = w_i r_i - v$ or $\xi_i = w_i(1-r_i) - (1-v_i)$. In either case $\xi_i \geq -1$. This lemma provides a stronger lower bound than that of Lemma 1. The reason we use the latter for vector bets is that the natural extension of (23) to the vector case does not lead to a convex problem.

C.3. Avoiding Grid Search

Suppose that our bets $\lambda^+_{2,i}$ and $\lambda^-_{2,i}$ do not depend on v. We have the individual lower bounds

$$\ln(K^+(v)) \geq \sum_i \lambda^+_{2,i}(w_i r_i - v) + \sum_i (\ln(1-\lambda^+_{2,i}) + \lambda_{2,i})(w_i r_i - v)^2$$

and

$$\ln(K^-(v)) \geq \sum_i \lambda^-_{2,i}(w_i r'_i - v') + \sum_i (\ln(1-\lambda^-_{2,i}) + \lambda^-_{2,i})(w_i r'_i - v')^2,$$

where $r' = 1-r$, $v' = 1-v$. For the Hedged process, using that for any a, b

$$\ln(\exp(a)+\exp(b)) \geq \max(a,b)$$

to first establish

$$\ln(K^\pm(v)) \geq \max(\ln(K^+(v)) - \ln(2), \ln(K^-(v)) - \ln(2))$$

and further bound each term in the maximum by the respective quadratic lower bound. We conclude that if a v achieves

$$\sum_i \lambda^+_{2,i}(w_i r_i - v) + \sum_i (\ln(1-\lambda^+_{2,i}) + \lambda^+_{2,i})(w_i r_i - v)^2 = \ln\left(\frac{2}{\alpha}\right)$$

or a $v' = 1-v$ achieves

$$\sum_i \lambda^-_{2,i}(w_i r'_i - v') + \sum_i (\ln(1-\lambda^-_{2,i}) + \lambda^-_{2,i})(w_i r'_i - v')^2 = \ln\left(\frac{2}{\alpha}\right)$$

then we also achieve $K^\pm(v) > \frac{1}{\alpha}$. Thus, a valid confidence interval can be obtained by considering the roots of these quadratics. Let

$$C = \sum_i \lambda^+_{2,i} w_i r_i \qquad C' = \sum_i \lambda^-_{2,i} w_i r'_i$$

$$S = \sum_i \lambda^+_{2,i} \qquad S' = \sum_i \lambda^-_{2,i}$$

$$Q = \sum_i \left(\ln(1-\lambda^+_{2,i}) + \lambda^+_{2,i}\right) w_i^2 r_i^2 \qquad Q' = \sum_i \left(\ln(1-\lambda^-_{2,i}) + \lambda^-_{2,i}\right) w_i^2 r_i'^2$$

$$T = \sum_i \left(\ln(1-\lambda^+_{2,i}) + \lambda^+_{2,i}\right) w_i r_i \qquad T' = \sum_i \left(\ln(1-\lambda^-_{2,i}) + \lambda^-_{2,i}\right) w_i r'_i$$

$$U = \sum_i \left(\ln(1-\lambda^+_{2,i}) + \lambda^+_{2,i}\right) \qquad U' = \sum_i \left(\ln(1-\lambda^-_{2,i}) + \lambda^-_{2,i}\right)$$

We obtain:

$$v_{\min} = \frac{2T + S - \sqrt{(2T+S)^2 - 4U(Q + C - \ln(2/\alpha))}}{2U}$$

or $v_{\min} = 0$ if the discriminant is negative, and

$$v_{\max} = 1 - v' = 1 - \frac{2T' + S' - \sqrt{(2T'+S')^2 - 4U'(Q' + C' - \ln(2/\alpha))}}{2U'}$$

or $v_{\max} = 1$ if the discriminant is negative.

D. Reward Predictors

D.1. Betting

We describe betting for $K_t^{+q}(v)$. Betting for $K_t^{-q}(v)$ is analogous. We overload the log wealth at step i when betting against v as $\ell_i^v(\lambda) = \ln(1 + \lambda_{1,i}(w_i - 1) + \lambda_{2,i}(w_i r_i - c_i - v)$. We use lemma 1 to obtain that for any $\lambda \in \mathcal{E}_v^{1/2}$, we have

$$\ln(K_t^{+q}(v)) = \sum_{i=1}^{t-1} \ell_i^v(\lambda) \geq \lambda^\top \sum_{i=1}^{t-1} b_i(v) + \psi \lambda^\top \left(\sum_{i=1}^{t-1} A_i(v) \right) \lambda,$$

where now $b_i(v) = \begin{bmatrix} w_i - 1 \\ w_i r_i - c_i - v \end{bmatrix}$ and $A_i(v) = b_i(v) b_i(v)^\top$. As in the case without reward predictor we have that the wealth lower bound is a polynomial in v with

$$\sum_{i=1}^{t-1} A_i(v) = A_t^{(0)} + v A_t^{(1)} + v^2 A_t^{(2)},$$
$$\sum_{i=1}^{t-1} b_i(v) = b_t^{(0)} + v b_t^{(1)},$$

and the coefficients can be maintained as

$$A_t^{(0)} = \sum_{i=1}^{t-1} \begin{bmatrix} (w_i - 1)^2 & (w_i - 1)(w_i r_i - c_i) \\ (w_i - 1)(w_i r_i - c_i) & (w_i r_i - c_i)^2 \end{bmatrix},$$
$$A_t^{(1)} = \sum_{i=1}^{t-1} \begin{bmatrix} 0 & -(w_i - 1) \\ -(w_i - 1) & -2(w_i r_i - c_i) \end{bmatrix},$$
$$A_t^{(2)} = \sum_{i=1}^{t-1} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix},$$
$$b_t^{(0)} = \sum_{i=1}^{t-1} \begin{bmatrix} w_i - 1 \\ w_i r_i - c_i \end{bmatrix},$$
$$b_t^{(1)} = \sum_{i=1}^{t-1} \begin{bmatrix} 0 \\ -1 \end{bmatrix}.$$

Given a v we compute concrete values for these coefficients and then solve

$$\lambda_t = \operatorname*{argmax}_{\lambda \in \mathcal{E}^{1/2}} \lambda^\top \sum_{i=1}^{t-1} b_i(v) + \psi \lambda^\top \left(\sum_{i=1}^{t-1} A_i(v) \right) \lambda.$$

A similar procedure like the one in Algorithm 1 can then be used for solving this problem.

D.2. Avoiding Grid Search

To find the value of v that we can plug in to the above optimization problem we proceed as in section 4.4, and further explained in Appendix B. To find a v such that $K_t^{\pm q}(v) \geq \frac{1}{\alpha}$ it suffices to solve

$$\sum_{i=1}^{t-1} \lambda_i^\top b_i(v) + \psi \sum_{i=1}^{t-1} \lambda_i^\top A_i(v) \lambda_i = \ln\left(\frac{2}{\alpha}\right),$$

given the previous bets $\lambda_1, \ldots, \lambda_{t-1}$. This is a second degree equation which can be solved by maintaining the quantities

$$\begin{aligned}
C_t &= \sum_{i=1}^{t-1} {\lambda_i}^\top \begin{bmatrix} w_i - 1 \\ w_i r_i - c_i \end{bmatrix}, \\
S_t &= \sum_{i=1}^{t-1} {\lambda_i}^\top \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \\
Q_t &= \sum_{i=1}^{t-1} \psi {\lambda_i}^\top \begin{bmatrix} (w_i-1)^2 & (w_i-1)(w_i r_i - c_i) \\ (w_i-1)(w_i r_i - c_i) & (w_i r_i - c_i)^2 \end{bmatrix} \lambda_i, \\
T_t &= \sum_{i=1}^{t-1} \psi {\lambda_i}^\top \begin{bmatrix} 0 & -(w_i-1) \\ -(w_i-1) & -2(w_i r_i - c_i) \end{bmatrix} \lambda_i, \\
U_t &= \sum_{i=1}^{t-1} \psi {\lambda_i}^\top \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \lambda_i,
\end{aligned}$$

and finding the largest real root v of

$$C_t - S_t v + Q_t + T_t v + U_t v^2 - \ln\left(\frac{2}{\alpha}\right) = 0,$$

if it exists, otherwise setting $v = 0$.

D.3. Double Hedging

Double Hedging boils down to running four processes: $K_t^{+q}, K_t^{-q}, K_t^+$, and K_t^-. Note that the wealth is split in 4 so anywhere we used $\ln\left(\frac{2}{\alpha}\right)$ in a hedged process now we need to use $\ln\left(\frac{4}{\alpha}\right)$. Note that both $K_t^{+q}(v)$ and $K_t^+(v)$ are trying to establish bounds for the same random variable and in principle they could communicate about values that have been eliminated. However we keep things simple and just run the four processes without sharing any information. The wealth of the doubly hedged process can then be lower bounded by the wealth of the most successful betting strategy starting from a wealth of $\frac{1}{4}$.

E. Gated Deployment

E.1. Hedging

Since we don't typically know whether π is better or worse that h we can hedge our bets via the process

$$K_t^{\pm gd}(v) = \frac{1}{2}(K_t^{+gd}(v) + K_t^{-gd}(v)),$$

where

$$\begin{aligned}
K_t^{+gd}(v) &= \prod_{i=1}^{t} \left(1 + \lambda_{1,i}^+(w_i - 1) + \lambda_{2,i}^+(w_i r_i - r_i - v)\right), \\
K_t^{-gd}(v) &= \prod_{i=1}^{t} \left(1 + \lambda_{1,i}^-(w_i - 1) + \lambda_{2,i}^-(w_i r_i' - r_i' - v')\right),
\end{aligned}$$

for predictable $\lambda_{1,i}^+, \lambda_{2,i}^+, \lambda_{1,i}^-, \lambda_{2,i}^-$ subject to $\lambda_i^+, \lambda_i^- \in \mathcal{G}_v^0$. As before, $r_i' = 1 - r_i$ and $v' = 1 - v$.

E.2. Betting and Avoiding Grid Search

Betting and avoiding grid search can be obtained using the same equations as for reward predictors but replacing all occurrences of c_i with r_i.

A key difference we spell out is the feasible region. In order to use common bets and to be able to use the quadratic lower bound of the log wealth we need to specify the set $\bigcap_{v\in[0,1]} \mathcal{G}_v^m$. This set is equivalent to

$$\mathcal{G}^m = \left\{ \lambda : \begin{bmatrix} -1 & -2 \\ -1 & 0 \\ W & -1 \\ W & W \end{bmatrix} \lambda \geq m-1 \right\},$$

where $W = w_{\max} - 1$. If we further restrict $\lambda_2 \geq 0$ for each of the subprocesses because we expect each to eliminate v such that $\mathbb{E}[wr - v] > 0$ and v' such that $\mathbb{E}[wr' - v'] > 0$ then the feasible region further simplifies to

$$\mathcal{G} = \{\lambda : \lambda_2 \geq 0, W\lambda_1 - \lambda_2 \geq m-1, -\lambda_1 - 2\lambda_2 \geq m-1\}.$$

Placing bets in this region can be done using the same ideas as Algorithm 1.

F. Reproducibility Checklist

Assumptions: The contextual bandit data is iid. The policy π is absolutely continuous with respect to behavior policy h.

Complexity: MOPE and the scalar Betting Strategy are streaming algorithms. They require constant time per sample and constant memory independent of number of samples. The exact wealth ablation requires memory that scales linearly with the number of samples and time per step that scales at least linearly with the number of samples. The ablation that solves a QP per value v requires at least $\frac{1}{\epsilon}$ times more memory and computation that MOPE and provides results that are accurate up to ϵ. We used $\epsilon = 0.005$ in the experiments.

Code: Available at `https://github.com/n17s/mope`

Data: synthetic environments are part of the code. Instructions for getting the mnist8m data are in the "Mnist-Policies" notebook.

Hyperparameters: There are no hyperparameters. The confidence level is an input and is stated in each experiment description or the corresponding figure.

Computing infrastructure: Off-the-shelf workstation running Linux (Code works on Windows as well).