Optimal performance of a tontine overlay subject to withdrawal constraints

Peter A. Forsyth a Kenneth R. Vetzal b Graham Westmacott c
November 16, 2022

Abstract

We consider the holder of an individual tontine retirement account, with maximum and minimum withdrawal amounts (per year) specified. The tontine account holder initiates the account at age 65, and earns mortality credits while alive, but forfeits all wealth in the account upon death. The holder desires to maximize total withdrawals, and minimize the expected shortfall, assuming the holder survives to age 95. The investor controls the amount withdrawn each year and the fraction of the investments in stocks and bonds. The optimal controls are determined based on a parametric model fitted to almost a century of market data. The optimal control algorithm is based on dynamic programming and solution of a partial integro differential equation (PIDE) using Fourier methods. The optimal strategy (based on the parametric model) is tested out of sample using stationary block bootstrap resampling of the historical data. In terms of an expected total withdrawal, expected shortfall (EW-ES) efficient frontier, the tontine overlay greatly outperforms an optimal strategy (without the tontine overlay), which in turn outperforms a constant weight strategy with withdrawals based on the ubiquitous four per cent rule.

Keywords: tontine, decumulation, expected shortfall, optimal stochastic control

JEL codes: G11, G22

AMS codes: 91G, 65N06, 65N12, 35Q93

1 Introduction

It is now commonplace to observe that defined benefit (DB) plans are disappearing. A recent OECD study (OECD 2019) observes that less than 50% of pension assets in 2018 were held in DB plans in over 80% of countries reporting. Of course, the level of assets in defined contribution (DC) plans is a lagging indicator, since historically, many employees were covered by traditional DB plans. These traditional DB plans still have a sizeable share of pension assets, simply because these plans have accumulated contributions over a longer period of time.

Consider the typical case of a DC plan investor upon retirement. Assuming that the investor has managed to accumulate a reasonable amount in her DC plan, the investor now faces the problem

---

aDavid R. Cheriton School of Computer Science, University of Waterloo, Waterloo ON, Canada N2L 3G1, paforsyt@uwaterloo.ca, +1 519 888 4567 ext. 34415.
bSchool of Accounting and Finance, University of Waterloo, Waterloo ON, Canada N2L 3G1, kvetzal@uwaterloo.ca, +1 519 888 4567 ext. 46518.
cRichardson Wealth, 540 Bingemans Centre Drive, Suite 100, Kitchener, ON, Canada N2B 3X9, Graham.Westmacott@RichardsonWealth.com, +1 800 451 4748.

1
of determining a decumulation strategy, i.e. how to invest and spend during retirement. It is often suggested that retirees should purchase annuities, but this is an unpopular strategy (Peijnenburg et al., 2016). MacDonald et al. (2013) note that this avoidance of annuities can be entirely rational.

A major concern of DC plan investors during the decumulation phase is, naturally, running out of savings. Possibly the most widely cited benchmark strategy is the 4% rule (Bengen, 1994). This rule posits a retiree who invests in a portfolio of 50% stocks and 50% bonds, rebalanced annually, and withdraws 4% of the original portfolio value each year (adjusted for inflation). This strategy would have never depleted the portfolio over any rolling thirty year historical period tested by Bengen on US data. This rule has been revisited many times. For example, Guyton and Klinger (2006) suggest several heuristic modifications involving withdrawal amounts and investment strategies.

Another approach has been suggested by Waring and Siegel (2015), which they term an Annually Recalculated Virtual Annuity (ARVA) strategy. The idea here is that the amount withdrawn in any given year should be based on the cash flows from a virtual (i.e. theoretical) fixed term annuity that could be purchased using the existing value of the portfolio. In this case, the DC plan can never run out of cash, but the withdrawal amounts can become arbitrarily small.

Turning to asset allocation strategies, Irlam (2014) used dynamic programming methods to conclude that deterministic (i.e. glide path) allocation strategies are sub-optimal. Of course, the asset allocation strategy and the withdrawal strategy are intimately linked. A more systematic approach to the decumulation problem involves formulating decumulation strategies as a problem in optimal stochastic control. The objective function for this problem involves a measure of risk and reward, which are, of course, conflicting measures. Forsyth (2022b) uses the withdrawal amount and the asset allocation (fraction in stocks and bonds) as controls. The measure of reward is the total (real) accumulated withdrawal amounts over a thirty year period. The withdrawal amounts have minimum and maximum constraints, hence there is a risk of depleting the portfolio. The measure of risk is the expected shortfall at the 5% level, of the (real) value of the portfolio at the thirty year mark. Utilizing both withdrawal amounts and asset allocation as controls considerably reduces the risk of portfolio depletion compared to fixed allocation or fixed withdrawal strategies.

A recent innovation in retirement planning involves the use of modern tontines (Donnelly et al., 2014; Donnelly, 2015; Milevsky and Salisbury, 2015; Fullmer, 2019; Weinert and Grundi, 2021; Winter and Planchet, 2022; Milevsky, 2022). In a tontine, the investor makes an irrevocable investment in a pooled fund for a fixed time frame. If the investor dies during the time horizon of the investment, the investor’s portfolio is divided amongst the remaining (living) members of the fund. If the investor survives until the end of the time horizon, then she will earn mortality credits from those members who have passed away. Unlike an annuity, there are no guaranteed cash flows, since typically the funds are invested in risky assets. Since there are no guarantees, the expected cash flows from a tontine are larger than for an annuity (with the same initial investment). Some authors have argued that the annuity puzzle should be replaced by the tontine puzzle, i.e. since tontines seem to very efficient products for pooling longevity risk, it is puzzling that the tontine market is still in its infancy (Chen and Rach, 2022). Pooled funds with tontine characteristics have been in use for some time.¹ We should also mention that retail investors may find the concept of a tontine appealing, simply due to the peer-to-peer model for managing longevity risk, which is also

¹The variable annuity funds offered by TIAA [https://www.tiaa.org/public/](https://www.tiaa.org/public/) the University of British Columbia pension plan [https://faculty.pensions.ubc.ca/](https://faculty.pensions.ubc.ca/) and the Australian Q-super fund [https://qsuper.qld.gov.au/](https://qsuper.qld.gov.au/) can all be viewed as having tontine characteristics. However, the Q-super fund takes the approach of averaging mortality credits over the entire pool, giving age independent mortality credits, which would appear to violate actuarial fairness. [https://i3-invest.com/2021/04/behind-qaupers-retirement-design/](https://i3-invest.com/2021/04/behind-qaupers-retirement-design/) The Q-super fund is perhaps more properly termed a collective defined contribution (CDC) fund. CDCs [https://www.ft.com/content/10448b2c-1141-4d2e-943c-70cce2caec52](https://www.ft.com/content/10448b2c-1141-4d2e-943c-70cce2caec52) have been criticized for lack of transparency and fairness.
consistent with the trend towards financial disintermediation. However, tontines may also require changes to existing legislation in some jurisdictions (MacDonald et al., 2021). There have also been suggestions for government management of tontine accounts (Fullmer and Forman, 2022; Fuentes et al., 2022). The attractiveness of tontines, from a behavioral finance perspective, is discussed in Chen et al. (2021). For an overview comparison of modern tontines to existing decumulation products, we refer the reader to Bar and Gatzert (2022).

Our focus in this article is on individual tontine accounts (Fullmer, 2019), whereby the investor has full control over the asset allocation in her account. We also allow the investor to control the withdrawal amount from the account, subject to maximum and minimum constraints. Usually it is suggested that withdrawal amounts from a tontine account cannot be increased, to avoid moral hazard issues. However, we view the maximum withdrawal as the desired withdrawal, allowing temporary reductions in withdrawals to minimize sequence of return risk and probability of ruin.

Consider an investor whose objective function uses reward as measured by total expected accumulated (real) withdrawals (EW) over a thirty year period. As a measure of risk, the investor uses the expected shortfall (ES) of the portfolio at the thirty year point. In this work, we define the expected shortfall as the mean of the worst 5% of the outcomes at year thirty. The investor’s controls are the amount withdrawn each year, and the allocations to stocks and bonds. The investor follows an optimal strategy to maximize this objective function.

Alternatively, the investor can use the same objective function, with the same controls, but this time add a tontine overlay (i.e. the investor is part of a pooled tontine). The investor has control over the withdrawals (subject of course to the same maximum and minimum constraints), and the allocation strategy.

Of course, we expect that the investor who uses the tontine overlay would achieve a better result than without the overlay, due to the mortality credits earned (we assume that the investor does not pass away during the thirty year investment horizon). However, this does not come without a cost. If the investor passes away, then her portfolio is forfeited.

Therefore, the investor must be compensated with a sizeable reduction in the risk of portfolio depletion, compared to the no-tontine overlay case. The objective of this article is to quantify this reduction, assuming optimal policies are followed in each case.

More precisely, we consider a 65-year old retiree, who can invest in a portfolio consisting of a stock index and a bond index, with yearly withdrawals, and rebalancing. The investor desires to maximize the multi-objective function in terms of the risk and reward measures described above, evaluated at the thirty year horizon (i.e. when the investor is 95).

We calibrate a parametric stochastic model for real (i.e. inflation adjusted) stock and bond returns, to almost a century of market data. We then solve the optimal stochastic control problem numerically, using dynamic programming. Robustness of the controls is then tested using block bootstrap resampling of the historical data.

Our main conclusion is that for a reasonable specification of acceptable tail risk (i.e. expected shortfall), the expected total cumulative withdrawals (EW) are considerably larger with the tontine overlay, compared to without the overlay. This conclusion holds even if the tontine overlay has fees of the order of 50-100 bps per year. Consequently, if the retiree has no bequest motive, and is primarily concerned with the risk of depleting her account, then a tontine overlay is an attractive solution.

It is also interesting to note that the optimal control for the withdrawal amount is (to a good approximation) a bang-bang control, i.e. it is only optimal to withdraw either the maximum or

---

2 See van Benthem et al. (2018) for an experiment with setting up a tontine using blockchain techniques.

3 An obvious case would be if an investor was given a medical diagnosis with a high probability of a poor outcome, at which point the investor would withdraw all remaining funds in her account.
minimum amount in any year. The allocation control essentially starts off with 40-50\% allocation to stocks. The median allocation control then rapidly reduces the fraction in equities to a very small amount after 5 – 10 years. The median withdrawal control starts off at the minimum withdrawal amount, and then rapidly increases withdrawals to the maximum after 2 – 5 years. The precise timing of the switch from minimum withdrawal to maximum withdrawal is simply a function of how much depletion risk (ES) the investor is prepared to take.

2 Overview of Individual Tontine Accounts

2.1 Intuition

We give a brief overview of modern tontines in this section. We restrict attention to the case of an individual tontine account (Fullmer, 2019), which is a constituent of a perpetual tontine pool. Consider a pool of \( m \) investors, who are alive at time \( t_{i-1} \). Let \( v_{ij} \) be the balance in the portfolio of investor \( j \) at time \( t_i \). In a tontine, if investor \( j \) participates in a tontine pool in time interval \((t_{i-1}, t_i)\), and investor \( j \) dies in that interval, then her portfolio \( v_{ij} \) is forfeited and given to the surviving members of the pool in the form of mortality credits (gains). Suppose that the probability that \( j \) dies in \((t_{i-1}, t_i)\) is \( q_{j_{i-1}} \). Let the tontine gain (mortality credit) for investor \( j \), for the period \((t_{i-1}, t_i)\), paid out at time \( t_i \), be denoted by \( c_{ij} \). The tontine will be a fair game if, for each player \( j \), the expected gain from participating in the tontine is zero,

\[
-q_{j_{i-1}} v_{ij} + (1 - q_{j_{i-1}}) c_{ij} = 0 ,
\]

and solving for the tontine gain \( c_{ij} \) gives

\[
c_{ij} = \frac{\text{Gain rate}}{1 - q_{j_{i-1}}} v_{ij} .
\]

For notational convenience, we define the tontine gain rate at \( t_i \) for investor \( j \) as

\[
(Tg^j_i = \left( \frac{q_{j_{i-1}}}{1 - q_{j_{i-1}}} \right) ,
\]

In our optimal control formulation, we will typically drop the superscript \( j \) from equation 2.3,

\[
Tg^j_i = \left( \frac{q_{j_{i-1}}}{1 - q_{j_{i-1}}} \right) ,
\]

since we will consider a given investor \( j \) with conditional mortality probability of \( q_{j_{i-1}} \) in \((t_{i-1}, t_i)\).

2.2 Group Gain

Consider tontine members \( j = 1,\ldots,m \) who are are alive at \( t_{i-1} \). Let

\[
1^j_i = \begin{cases} 1 & \text{Investor } j \text{ is alive at } t_{i-1} \text{ and alive at } t_i \\ 0 & \text{Investor } j \text{ is alive at } t_{i-1} \text{ and dead at } t_i \end{cases}
\]

\[
E[1^j_i] = 1 - q_{j_{i-1}} ,
\]
where $E[\cdot]$ is the expectation operator. Note that the total tontine gain for all members at $t_i$ is

$$\text{Total Tontine Gains} = \sum_{j=1}^{m} 1_j^i c_j^i = \sum_{j=1}^{m} 1_j^i \left( \frac{q_{i-1}^j}{1-q_{i-1}^j} \right) v_i^j, \quad (2.6)$$

and that the total amount forfeited by members who have died in $(t_{i-1}, t_i)$ is

$$\text{Total Forfeited} = \sum_{j=1}^{m} (1 - 1_j^i) v_i^j. \quad (2.7)$$

Then

$$E[ \text{Total Tontine Gains} ] = \sum_{j=1}^{m} q_{i-1}^j v_i^j$$

$$E[ \text{Total Forfeited} ] = \sum_{j=1}^{m} q_{i-1}^j v_i^j, \quad (2.8)$$

implying that the expected total tontine gain is balanced by the expected total amount forfeited.

In practice, of course, the expected number of deaths in period $(t_{i-1}, t_i)$ may not be equal to the actual number of deaths. To compensate for this, a practical implementation method has been suggested in (Sabin and Forman, 2016; Denuit and Vernic, 2018; Fullmer and Sabin, 2019; Fullmer, 2019; Winter and Planchet, 2022). Denote the realized group gain at $t_i$ by $G_i$

$$G_i = \frac{\sum_{j=1}^{m} (1 - 1_j^i) v_i^j}{\sum_{j=1}^{m} 1_j^i c_j^i}. \quad (2.9)$$

Observe that the values of $1_j^i$ in equation (2.9) are the realized values, not expectations. The actual tontine gain (mortality credit) $\hat{c}_j^i$ earned by investor $j$ (assuming investor $j$ is alive at $t_i$) is then

$$\hat{c}_j^i = G_i \left( \frac{q_{i-1}^j}{1-q_{i-1}^j} \right) v_i^j. \quad (2.10)$$

Essentially, we are scaling equation (2.2) by the factor $G_i$, which ensures that the the total amount forfeited by the observed deaths in $(t_{i-1}, t_i)$ is exactly equal to the total mortality credits disbursed to the survivors. In the following, we will refer to $\hat{c}_j^i$ as the actual mortality gain, while $c_j^i$ will be termed the nominal tontine gain.

Remark 2.1 ($\sum_{j=1}^{m} 1_j^i c_j^i = 0$). Note that equation (2.9) is undefined if all members die in $(t_{i-1}, t_i)$. We assume that the tontine is large (in terms of members) and perpetual, i.e. open to new members, so that the probability of all members dying is negligible. For mathematical completeness, we can suppose that if all members die in $(t_{i-1}, t_i)$, we collapse the tontine, and distribute all remaining account values $v_i^j$ to the estates of members $j$.

While use of equation (2.10) looks like a reasonable approach, it turns out that this is not strictly fair in the actuarial sense, i.e. there is some bias that favors some members over others. Informally,
this bias exists since the total amount forfeited in a period depends on the binary state of each
member of the pool, i.e. alive or dead. Sabin and Forman (2016) show that the bias is negligible
under the following conditions

**Condition 2.1 (Small bias condition).** Suppose that the following conditions hold

(a) the pool of participants in the tontine is sufficiently large;

(b) the expected amount forfeited by all members is large compared to any member’s nominal gain,
    i.e.

\[ \left( v_i^j \frac{q_{i-1}^j}{1 - q_{i-1}^j} \right) \ll \sum_k q_{i-1}^k v_i^k \right) ; \quad j = 1, \ldots, m , \tag{2.11} \]

then the bias is negligibly small (Sabin and Forman, 2016).

Condition (2.11) is essentially a diversification requirement: no member of the pool has an
abnormally large share of the total pool capital. In addition, of course, if the pool is sufficiently
large, then the actual number of deaths in \((t_{i-1}, t_i)\) will converge to the expected number of deaths.

Note that it does not matter what investment strategy is followed by any given investor in period
\((t_{i-1}, t_i)\). Each investor can choose whatever policy they like, since only the observed final portfolio
value at \(t_i\) matters. Somewhat counterintuitively, \(G_i\) is very close to unity, even if the participants
in the tontine pool are very heterogeneous, i.e. with different ages, genders, invested amounts, and
asset allocations (assuming Condition 2.1 holds).

In Fullmer and Sabin (2019), simulations were carried out to determine the magnitude of the
volatility of \(G_i\) under practical sizes of tontine pools. Given a tontine pool of 15,000 members,
with varying ages, initial capital, and randomly assigned investment policies (i.e. the bond/stock
split), the simulations showed that \(E[G_i] \approx 1\) and that the standard deviation was about 0.1. This
standard deviation at each \(t_i\) actually resulted in a smaller effect over a long term (assuming that
the tontine member lived long enough). This is simply because everybody dies eventually, so that
if fewer deaths than expected are observed in a year, then more deaths will be observed in later
years, and vice versa.

In the following, we will assume that the pool is sufficiently large and that it satisfies the
diversity condition (2.11), so that there is no error in assuming that \(G \equiv 1\), i.e we will assume that
the nominal tontine gain is the actual tontine gain. As a sanity check, we also carry out a test
whereby we simulate the effect of randomly varying \(G_i\), based on the statistics of the simulations in
Fullmer and Sabin (2019). Our results show that effect of randomness in \(G\) can be safely ignored
for a reasonably sized tontine pool. To be more precise here, we will modify equation (2.4) so that

\[ T_{g_i} = \left( \frac{q_{i-1}}{1 - q_{i-1}} \right) G_i , \tag{2.12} \]

for a numerical example showing the effects of randomness of \(G_i\), in Monte Carlo simulations. Our
computation of the optimal strategy will always assume \(G_i \equiv 1\).

\[ ^4 \text{This is illustrated in Winter and Planchet (2022), using an example with pool consisting of a large number of}
\[ \text{young investors (with small individual portfolios), and a single elderly member with a large portfolio. The elderly}
\[ \text{member effectively subsidizes the younger members.} \]
2.3 Variable withdrawals

We will allow the individual tontine member to withdraw variable amounts, subject to minimum and maximum constraints. We remind the reader that if a tontine pool is strictly actuarially fair, then, in theory, there are no constraints on withdrawals and injections of cash \cite{Braughtigam2017}. However, in practice, since pools are finite sized, heterogeneous, and mortality credits are not distributed at infinitesimal intervals, we do not allow arbitrarily large withdrawals. This avoids moral hazard issues.

Since we have a minimum withdrawal amount in each time period, there is a risk of running out of cash. We assume that if the tontine member’s account becomes negative, than all trading in this account ceases, and debt accumulates at the borrowing rate. In practice, once the tontine account becomes zero, the retiree has to fund expenses from another source. We implicitly assume that the tontine member has other assets which can be used to fund this minimum consumption level (e.g. real estate). Of course, we aim to make this a very improbable event. In fact, this is the reason why we allow variable withdrawals. We can regard the upper bound on the withdrawals as the desired consumption level, but we allow the tontine member to reduce (hopefully only temporarily) their withdrawals, to minimize risk of depletion of their tontine account.

2.4 Money back guarantees

In practice, we observe that many tontine funds offer a money back guarantee\footnote{https://qsuper.qld.gov.au/}. This is usually specified as a return of the initial (nominal) investment less any withdrawals (if the sum is non-negative) at the time of death. We do not consider such guarantees in this work, focusing on the pure tontine aspect, which has no guarantees, and presumably the highest possible expected total withdrawals. A money back guarantee would have to be hedged, which would reduce returns. In practice, this guarantee could be priced separately, and added as overlay to the tontine investment if desired.

2.5 Survivor Benefits

Many DB plans have survivor benefits which are received by a surviving spouse. A typical case would involve the surviving spouse receiving 60–75% of the yearly pension after the DB plan holder dies.

Consider the following case of a male, same-sex couple, both of whom are exactly the same age. As an extreme case, suppose the survivor benefit is 100% of the tontine cash flows, which continue until the survivor dies. From the CPM2014 table from the Canadian Institute of Actuaries\footnote{www.cia-ica.ca/docs/default-source/2014/214013e.pdf}, the probability that an 85-year old Canadian male dies before reaching the age of 86 is about .076. Assuming that the mortality probabilities are independent for both spouses, then the probability that both 85-year old spouses die before reaching age 86, conditional on both living to age 85 is \((.076)^2 \approx .0053\). The tontine gain rate per year (from equation (2.3)) is

\[
\text{tontine gain rate} = \frac{.0053}{1 - .0053} \approx .0053 .
\]  

(2.13)

We will assume in our numerical examples that the base case fee charged for managing the tontine is 50 bps per year. This means that, net of fees, there are essentially no tontine gains for our
hypothetical couple, for the first 20 years of retirement, which is surely undesirable. Once one of 
the partners passes away, the tontine gain rate will, of course, take a jump in value.

As another extreme case, suppose that the surviving spouse receives 50\% of the tontine cash 
flows. In this case, the total cash flows accruing this couple are exactly the same as dividing the 
original total wealth into two, and then having each spouse invest in their own individual tontine. 

It is possible to determine the distribution of the cash flows for a survivor benefit which is 
intermediate to these edge cases. However, this requires additional state variables in our optimal 
control problem, and is probably best tackled using a machine learning approach \cite{LiForsyth2019,Nietal2022}. We will leave this case for future work, and focus attention on the individual 
tontine case, with no survivor benefit. Note that in the tontine context, survivor benefits are 
typically provided by a separate insurance overlay\footnote{https://i3-invest.com/2021/04/behind-qsupers-retirement-design/}

\section{Formulation}

We assume that the investor has access to two funds: a broad market stock index fund and a 
constant maturity bond index fund.

The investment horizon is $T$. Let $S_t$ and $B_t$ respectively denote the real (inflation adjusted) 
amounts invested in the stock index and the bond index respectively. In general, these amounts 
will depend on the investor’s strategy over time, as well as changes in the real unit prices of the 
assets. In the absence of an investor determined control (i.e. cash withdrawals or rebalancing), all 
changes in $S_t$ and $B_t$ result from changes in asset prices. We model the stock index as following a 
jump diffusion.

In addition, we follow the usual practitioner approach and directly model the returns of the 
constant maturity bond index as a stochastic process, see for example \cite{Linetal2015,MacMinnetal2014}. As in \cite{MacMinnetal2014}, we assume that the constant maturity bond index follows 
a jump diffusion process as well.

Let $S_{t^\epsilon} = S(t - \epsilon)$, $\epsilon \rightarrow 0^+$, i.e. $t^\epsilon$ is the instant of time before $t$, and let $\xi^s$ be a random 
number representing a jump multiplier. When a jump occurs, $\xi^s S_{t^\epsilon}$. Allowing for jumps 
permits modelling of non-normal asset returns. We assume that $\log(\xi^s)$ follows a double exponential 
distribution \cite{Kou2002,KouWang2004}. If a jump occurs, $u^s$ is the probability of an upward 
jump, while $1 - u^s$ is the chance of a downward jump. The density function for $y = \log(\xi^s)$ is 

\begin{equation}
 f^s(y) = u^s \eta_1^s e^{-\eta_1^s y} \mathbf{1}_{y \geq 0} + (1-u^s) \eta_2^s e^{\eta_2^s y} \mathbf{1}_{y < 0} .
\end{equation}

We also define 

\begin{equation}
 \gamma^s = E[\xi^s - 1] = \frac{u^s \eta_1^s}{\eta_1^s - 1} + \frac{(1-u^s) \eta_2^s}{\eta_2^s + 1} - 1 .
\end{equation}

In the absence of control, $S_t$ evolves according to 

\begin{equation}
 \frac{dS_t}{S_{t^\epsilon}} = (\mu^s - \lambda^s \gamma^s) dt + \sigma^s dZ^s + d \left( \sum_{i=1}^{\pi^s_t} (\xi^s_t - 1) \right) ,
\end{equation}

where $\mu^s$ is the (uncompensated) drift rate, $\sigma^s$ is the volatility, $dZ^s$ is the increment of a Wiener 
process, $\pi^s_t$ is a Poisson process with positive intensity parameter $\lambda^s_t$, and $\xi^s_t$ are i.i.d. positive
random variables having distribution (3.1). Moreover, $\xi_i^s$, $\pi_i^s$, and $Z^s$ are assumed to all be mutually independent.

Similarly, let the amount in the bond index be $B_{t-} = B(t-\epsilon), \epsilon \to 0^+$. In the absence of control, $B_t$ evolves as

$$
\frac{dB_t}{B_{t-}} = \left( \mu^b - \gamma^b \xi_{t-}^b + \mu^c_b \{B_{t-} < 0\} \right) dt + \sigma^b dZ^b + d \left( \sum_{i=1}^{\pi^b_t} (\xi^b_i - 1) \right), \quad (3.4)
$$

where the terms in equation (3.4) are defined analogously to equation (3.3). In particular, $\pi^b_t$ is a Poisson process with positive intensity parameter $\lambda_b \xi_{t-}^b$, and $\xi^b_i$ has distribution

$$
f^b(y = \log \xi^b_i) = u^b_i \eta^b_1 e^{-\eta^b_1 y} 1_{y \geq 0} + (1 - u^b_i) \eta^b_2 e^{\eta^b_2 y} 1_{y < 0}, \quad (3.5)
$$

and $\gamma^b = E[\xi^b - 1]$. $\xi^b_i$, $\pi^b_t$, and $Z^b$ are assumed to all be mutually independent. The term $\mu^c_b \{B_{t-} < 0\}$ in equation (3.4) represents the extra cost of borrowing (the spread).

The diffusion processes are correlated, i.e. $dZ^s \cdot dZ^b = \rho_{sb} dt$. The stock and bond jump processes are assumed mutually independent. See Forsyth (2020b) for justification of the assumption of stock-bond jump independence.

We define the investor’s total wealth at time $t$ as

$$
\text{Total wealth} \equiv W_t = S_t + B_t. \quad (3.6)
$$

We impose the constraints that (assuming solvency) shorting stock and using leverage (i.e. borrowing) are not permitted. In the event of insolvency (due to withdrawals), the portfolio is liquidated, trading ceases and debt accumulates at the borrowing rate.

4 Notational Conventions

Consider a set of discrete withdrawal/rebalancing times $\mathcal{T}$

$$
\mathcal{T} = \{t_0 = 0 < t_1 < t_2 < \ldots < t_M = T\} \quad (4.1)
$$

where we assume that $t_i - t_{i-1} = \Delta t = T/M$ is constant for simplicity. To avoid subscript clutter, in the following, we will occasionally use the notation $S_t \equiv S(t), B_t \equiv B(t)$ and $W_t \equiv W(t)$. Let the inception time of the investment be $t_0 = 0$. We let $\mathcal{T}$ be the set of withdrawal/rebalancing times, as defined in equation (4.1). At each rebalancing time $t_i$, $i = 0, 1, \ldots, M - 1$, the investor (i) withdraws an amount of cash $q_i$ from the portfolio, and then (ii) rebalances the portfolio. At $t_M = T$, the portfolio is liquidated and no cash flow occurs. For notational completeness, this is enforced by specifying $q_M = 0$.

In the following, given a time dependent function $f(t)$, then we will use the shorthand notation

$$
f(t_i^+) \equiv \lim_{\epsilon \to 0^+} f(t_i + \epsilon) \quad ; \quad f(t_i^-) \equiv \lim_{\epsilon \to 0^+} f(t_i - \epsilon). \quad (4.2)
$$

Let

$$
(\Delta t)_i = \begin{cases} \Delta t & i = 1, \ldots, M, \\ 0 & i = 0. \end{cases} \quad (4.3)
$$
We assume that a tontine fee of $T^f$ per unit time is charged at $t \in \mathcal{T}$, based on the total portfolio value at $t^-$, after tontine gains but before withdrawals. Recalling the definition of tontine gain rate $T^g_i$ in equation (2.4), we modify this definition to enforce no tontine gain at $t_0$,

$$T^g_i = \begin{cases} \frac{q_{i-1}}{1 - q_{i-1}} & i = 1, \ldots, M \\ 0 & i = 0 \end{cases}. \quad (4.4)$$

Then, $W(t^+_i)$ is given by

$$W(t^+_i) = \left( S(t^-_i) + B(t^-_i) \right) \left( 1 + T^g_i \right) \exp(-\Delta t) - q_i ; i \in \mathcal{T}, \quad (4.5)$$

where we recall that $q_M \equiv 0$ and $(\Delta t)_0 \equiv 0$. With some abuse of notation, we define

$$W(t^-_i) = \left( S(t^-_i) + B(t^-_i) \right) \left( 1 + T^g_i \right) \exp(-\Delta t) \quad (4.6)$$

as the total portfolio value, after tontine gains and tontine fees, the instant before withdrawals and rebalancing at $t_i$.

Typically, DC plan savings are held in a tax advantaged account, with no taxes triggered by rebalancing. With infrequent (e.g. yearly) rebalancing, we also expect other transaction costs, apart from the tontine fees, to be small, and hence can be ignored. It is possible to include transaction costs, but at the expense of increased computational cost \cite{vanStaden18}.

We denote by $X(t) = (S(t), B(t))$, $t \in [0,T]$, the multi-dimensional controlled underlying process, and by $x = (s,b)$ the realized state of the system. Let the rebalancing control $p_i(\cdot)$ be the fraction invested in the stock index at the rebalancing date $t_i$, i.e.

$$p_i \left( X(t^-_i) \right) = p \left( X(t^-_i), t_i \right) = \frac{S(t^+_i)}{S(t^+_i) + B(t^+_i)}. \quad (4.7)$$

Let the withdrawal control $q_i(\cdot)$ be the amount withdrawn at time $t_i$, i.e. $q_i \left( X(t^-_i) \right) = q \left( X(t^-_i), t_i \right)$. Formally, the controls depend on the state of the investment portfolio, before the rebalancing occurs, i.e. $p_i(\cdot) = p \left( X(t^-_i), t_i \right) = p \left( X^-_i, t_i \right)$, and $q_i(\cdot) = q \left( X(t^-_i), t_i \right) = q \left( X^-_i, t_i \right)$, $t_i \in \mathcal{T}$, where $\mathcal{T}$ is the set of rebalancing times.

However, it will be convenient to note that in our case, we find the optimal control $p_i(\cdot)$ amongst all strategies with constant wealth (after withdrawal of cash). Hence, with some abuse of notation, we will now consider $p_i(\cdot)$ to be function of wealth after withdrawal of cash

$$p_i(\cdot) = p(W(t^+_i), t_i)$$

$$W(t^+_i) = W(t^-_i) - q_i(\cdot)$$

$$W(t^-_i) = \left( S(t^-_i) + B(t^-_i) \right) \left( 1 + T^g_i \right) \exp(-\Delta t) \quad (4.8)$$

$$S(t^+_i) = S^+_i = p_i(W^+_i) W^+_i$$

$$B(t^+_i) = B^+_i = (1 - p_i(W^+_i)) W^+_i \quad (4.8)$$

Note that the control for $p_i(\cdot)$ depends only $W^+_i$. Since $p_i(\cdot) = p_i(W^-_i - q_i)$, then it follows that

$$q_i(\cdot) = q_i(W^-_i), \quad (4.9)$$
which we discuss further in Section [7].

A control at time \( t_i \), is then given by the pair \((q_i(\cdot), p_i(\cdot))\) where the notation \((\cdot)\) denotes that the control is a function of the state.

Let \( Z \) represent the set of admissible values of the controls \((q_i(\cdot), p_i(\cdot))\). We impose no-shorting, no-leverage constraints (assuming solvency). We also impose maximum and minimum values for the withdrawals. We apply the constraint that in the event of insolvency due to withdrawals \((W(t_i^+) < 0)\), trading ceases and debt (negative wealth) accumulates at the appropriate borrowing rate of return (i.e. a spread over the bond rate). We also specify that the stock assets are liquidated at \( t = t_M \).

More precisely, let \( W_i^+ \) be the wealth after withdrawal of cash, and \( W_i^- \) be the total wealth before withdrawals (but after fees and tontine cash flows), then define

\[
Z_q = \begin{cases} 
[q_{\min}, q_{\max}] & t \in \mathcal{T} ; t \neq t_M ; W_i^- \geq q_{\max} \\
[q_{\min}, \max(q_{\min}, W_i^-)] & t \in \mathcal{T} ; t \neq t_M ; W_i^- < q_{\max} \\
\{0\} & t = t_M 
\end{cases} 
\]  
(4.10)

\[
Z_p(W_i^+, t_i) = \begin{cases} 
[0,1] & W_i^+ > 0 ; t_i \in \mathcal{T} ; t_i \neq t_M \\
\{0\} & W_i^+ \leq 0 ; t_i \in \mathcal{T} ; t_i \neq t_M \\
\{0\} & t_i = t_M 
\end{cases} 
\]  
(4.11)

The rather complicated expression in equation (4.10) imposes the assumption that, as wealth becomes small, the retiree (i) tries to avoid insolvency as much as possible and (ii) in the event of insolvency, withdraws only \( q_{\min} \).

The set of admissible values for \((q_i, p_i), t_i \in \mathcal{T} \), can then be written as

\[
(q_i, p_i) \in Z(W_i^-, W_i^+, t_i) = Z_q(W_i^-, t_i) \times Z_p(W_i^+, t_i) .
\]  
(4.13)

For implementation purposes, we have written equation (4.13) in terms of the wealth after withdrawal of cash. However, we remind the reader that since \( W_i^- = W_i^- - q_i \), the controls are formally a function of the state \( X(t_i^-) \) before the control is applied.

The admissible control set \( \mathcal{A} \) can then be written as

\[
\mathcal{A} = \left\{ (q_i, p_i)_{0 \leq i \leq M} : (p_i, q_i) \in Z(W_i^-, W_i^+, t_i) \right\} 
\]  
(4.14)

An admissible control \( \mathcal{P} \in \mathcal{A} \), where \( \mathcal{A} \) is the admissible control set, can be written as,

\[
\mathcal{P} = \{ (q_i(\cdot), p_i(\cdot)) : i = 0, \ldots, M \} .
\]  
(4.15)

We also define \( \mathcal{P}_n \equiv \mathcal{P}_{t_n} \subset \mathcal{P} \) as the tail of the set of controls in \( [t_n, t_{n+1}, \ldots, t_M] \), i.e.

\[
\mathcal{P}_n = \{ (q_n(\cdot), p_n(\cdot)), \ldots, (q_{M}(\cdot), p_{M}(\cdot)) \} .
\]  
(4.16)

For notational completeness, we also define the tail of the admissible control set \( \mathcal{A}_n \) as

\[
\mathcal{A}_n = \left\{ (q_i, p_i)_{0 \leq i \leq M} : (q_i, p_i) \in Z(W_i^-, W_i^+, t_i) \right\} 
\]  
(4.17)

so that \( \mathcal{P}_n \in \mathcal{A}_n \).
5 Risk and Reward

5.1 Risk: Definition of Expected Shortfall (ES)

Let \( g(W_T) \) be the probability density function of wealth \( W_T \) at \( t = T \). Suppose

\[
\int_{-\infty}^{W_\alpha^*} g(W_T) \, dW_T = \alpha, \tag{5.1}
\]

i.e. \( Pr[W_T > W_\alpha^*] = 1 - \alpha \). We can interpret \( W_\alpha^* \) as the Value at Risk (VAR) at level \( \alpha \). The Expected Shortfall (ES) at level \( \alpha \) is then

\[
ES_\alpha = \frac{\int_{-\infty}^{W_\alpha^*} W_T \, g(W_T) \, dW_T}{\alpha}, \tag{5.2}
\]

which is the mean of the worst \( \alpha \) fraction of outcomes. Typically \( \alpha \in \{.01, .05\} \). The definition of ES in equation (5.2) uses the probability density of the final wealth distribution, not the density of loss. Hence, in our case, a larger value of ES (i.e. a larger value of average worst case terminal wealth) is desired. The negative of ES is commonly referred to as Conditional Value at Risk (CVAR).

Define \( X_0^+ = X(t_0^+) \), \( X_0^- = X(t_0^-) \). Given an expectation under control \( P \), \( E_P[\cdot] \), as noted by Rockafellar and Uryasev (2000), \( ES_\alpha \) can be alternatively written as

\[
ES_\alpha(X_0^-, t_0^-) = \sup_{W^*} E_{P_0}[t_0^-] \left[ W^* + \frac{1}{\alpha} \min(W_T - W^*, 0) \right]. \tag{5.3}
\]

The admissible set for \( W^* \) in equation (5.3) is over the set of possible values for \( W_T \).

The notation \( ES_\alpha(X_0^-, t_0^-) \) emphasizes that \( ES_\alpha \) is as seen at \( (X_0^-, t_0^-) \). In other words, this is the pre-commitment \( ES_\alpha \). A strategy based purely on optimizing the pre-commitment value of \( ES_\alpha \) at time zero is time-inconsistent, hence has been termed by many as non-implementable, since the investor has an incentive to deviate from the time zero pre-commitment strategy at \( t > 0 \). However, in the following, we will consider the pre-commitment strategy merely as a device to determine an appropriate level of \( W^* \) in equation (5.3). If we fix \( W^* \forall t > 0 \), then this strategy is the induced time consistent strategy (Strub et al., 2019), hence is implementable. We delay further discussion of this subtle point to Appendix A.

5.2 A measure of reward: expected total withdrawals (EW)

We will use expected total withdrawals as a measure of reward in the following. More precisely, we define \( EW \) (expected withdrawals) as

\[
EW(X_0^-, t_0^-) = E_{P_0}[t_0^-] \left[ \sum_{i=0}^{M} q_i \right]. \tag{5.4}
\]

Note that there is no discounting term in equation (5.4) (recall that all quantities are real, i.e. inflation adjusted). It is straightforward to introduce discounting, but we view setting the real discount rate to zero to be a reasonable and conservative choice. See Forsyth (2022b) for further comments.

\[\text{In practice, the negative of } W_\alpha^* \text{ is often the reported VAR.}\]
6 Problem EW-ES

Since expected withdrawals (EW) and expected shortfall (ES) are conflicting measures, we use a scalarization technique to find the Pareto points for this multi-objective optimization problem. Informally, for a given scalarization parameter $\kappa > 0$, we seek to find the control $P_0$ that maximizes

$$\text{EW}(X_0^-, t_0^-) + \kappa \text{ES}_\alpha(X_0^-, t_0^-).$$

(6.1)

More precisely, we define the pre-commitment EW-ES problem ($PCES_{t_0}(\kappa)$) problem in terms of the value function $J(s,b,t_0^-)$

$$(PCES_{t_0}(\kappa)) : J(s,b,t_0^-) = \sup_{P_0 \in A} \sup_{W^*} \left\{ E_{P_0}^{X_0^+, t_0^+} \left[ \sum_{i=0}^{M} q_i + \kappa \left( W^* + \frac{1}{\alpha} \min(W_T - W^*, 0) \right) \right] + \epsilon W_T \big| X(t_0^-) = (s,b) \right\}$$

(6.2)

subject to

$$\begin{align*}
(S_t,B_t) & \text{ follow processes } (3.3) \text{ and } (3.4); \ t \notin T \\
W^+_t &= W^-_t - q_t; X^+_t = (S^+_t, B^+_t) \\
W^-_t &= (S(t^-_i) + B(t^-_i)) \left( 1 + T^g_i \right) \exp(-(\Delta t)_i T^f) \\
S^+_t &= p_t(\cdot)W^+_t; B^+_t = (1 - P_t(\cdot))W^+_t \\
(q_t(\cdot),p_t(\cdot)) & \in Z(W^-_t, W^+_t,t_t) \\
\ell & = 0, \ldots, M; \ t_t \in T
\end{align*}$$

(6.3)

Note that we have added the extra term $E_{P_0}^{X_0^+, t_0^+} [\epsilon W_T]$ to equation (6.2). If we have a maximum withdrawal constraint, and if $W_t \gg W^*$ as $t \to T$, the controls become ill-posed. In this fortunate state for the investor, we can break the investment policy ties by setting $\epsilon < 0$, which will force investments in bonds, while if $\epsilon > 0$, then this will force investments into stocks. Choosing $|\epsilon| \ll 1$ ensures that this term only has an effect if $W_t \gg W^*$ and $t \to T$. See Forsyth (2022b) for more discussion of this.

Interchange the sup in equation (6.2), so that value function $J(s,b,t_0^-)$ can be written as

$$J(s,b,t_0^-) = \sup_{W^*} \sup_{P_0 \in A} \left\{ E_{P_0}^{X_0^+, t_0^+} \left[ \sum_{i=0}^{M} q_i + \kappa \left( W^* + \frac{1}{\alpha} \min(W_T - W^*, 0) \right) + \epsilon W_T \big| X(t_0^-) = (s,b) \right] \right\}.$$

(6.4)

Noting that the inner supremum in equation (6.4) is a continuous function of $W^*$, and noting that the optimal value of $W^*$ in equation (6.4) is bounded, then define

$$W^*(s,b) = \arg \max_{W^*} \sup_{P_0 \in A} \left\{ E_{P_0}^{X_0^+, t_0^+} \left[ \sum_{i=0}^{M} q_i + \kappa \left( W^* + \frac{1}{\alpha} \min(W_T - W^*, 0) \right) + \epsilon W_T \big| X(t_0^-) = (s,b) \right] \right\}.$$

(6.5)

This is the same as noting that a finite value at risk exists. This easily shown, assuming $0 < \alpha < 1$, since our investment strategy uses no leverage and no-shorting.
We refer the reader to Forsyth (2020a) for an extensive discussion concerning pre-commitment and time consistent ES strategies. We summarize the relevant results from that discussion in Appendix A.

7 Formulation as a Dynamic Program

We use the method in Forsyth (2020a) to solve problem (6.4). We write equation (6.4) as

\[ J(s, b, t_0) = \sup_{W^*} V(s, b, W^*, 0^-) , \]  

where the auxiliary function \( V(s, b, W^*, t) \) is defined as

\[ V(s, b, W^*, t_n) = \sup_{p_n \in A_n} \left\{ E_{p_n}^{\hat{X}_{n}^+, t_n^+} \left[ \sum_{i=n}^M q_i + \kappa \left( W^* + \frac{1}{\alpha} \min((W_T - W^*), 0) \right) + \epsilon W_T \right| \hat{X}(t_n) = (s, b, W^*) \right\} . \]  

subject to

\[
\begin{align*}
(S_t, B_t) & \text{ follow processes } (3.3) \text{ and } (3.4); \quad t \notin \mathcal{T} \\
W^+_t & = W^+_t - q_t; \quad X^+_t = (S^+_t, B^+_t, W^*) \\
W^-_t & = \left( S(t^-_t) + B(t^-_t) \right) \left( 1 + T^q_t \right) \exp(- (\Delta t)_t T^f) . \end{align*}
\]

We have now decomposed the original problem (6.4) into two steps

- For given initial cash \( W_0 \), and a fixed value of \( W^* \), solve problem (7.2) using dynamic programming to determine \( V(0, W_0, W^*, 0^-) \).
- Solve problem (6.4) by maximizing over \( W^* \)

\[ J(0, W_0, 0^-) = \sup_{W^*} V(0, W_0, W^*, 0^-) . \]  

7.1 Dynamic Programming Solution of Problem (7.2)

We give a brief overview of the method described in detail in Forsyth (2022b). Apply the dynamic programming principle to \( t_n \in \mathcal{T} \)

\[ V(s, b, W^*, t_n) = \sup_{q \in Z_q(w^-, t_n)} \left\{ \sup_{p \in Z_p(w^- - q, t_n)} \left[ q + V((w^- - q)p, (w^- - q)(1 - p), W^*, t_n^+) \right] \right\} \]

\[ = \sup_{q \in Z_q(w^-, t_n)} \left\{ q + \sup_{p \in Z_p(w^- - q, t_n)} V((w^- - q)p, (w^- - q)(1 - p), W^*, t_n^+) \right\} \]

\[ w^- = (s + b)(1 + T^q_t) \exp(- (\Delta t)_t T^f) . \]  

For computational purposes, we define

\[ \hat{V}(w, t_n, W^*) = \left[ \sup_{p \in Z_p(w, t_n)} V(wp, w(1 - p), W^*, t_n^+) \right] . \]
Equation (7.5) now becomes

\[ V(s,b,W^*, t_n^-) = \sup_{q \in \mathbb{Z}(w^-, t_n^-)} \left\{ q + \left[ \hat{V}((w^-, q), W^*, t_n^+) \right] \right\} \]

\[ w^- = (s + b) \left( 1 + T^g_i \right) \exp(-\Delta t_i \mathbb{T}^f) . \]

(7.7)

This approach effectively replaces a two dimensional optimization for \((q_n, p_n)\), to two sequential one dimensional optimizations. From equations (7.6\textsuperscript{7} 7.7), it is clear that the optimal pair \((q_n, p_n)\) is such that

\[ q_n = q_n(w^-, W^*) \]

\[ w^- = (s + b) \left( 1 + T^g_i \right) \exp(-\Delta t_i \mathbb{T}^f) \]

\[ p_n = p_n(w, W^*) \]

\[ w = w^- - q_n . \]

(7.8)

In other words, the optimal withdrawal control \(q_n\) is only a function of total wealth (after tontine gains and fees) before withdrawals. The optimal control \(p_n\) is a function only of total wealth after withdrawals, tontine gains, and fees.

At \(t = T\), we have

\[ V(s, b, W^*, T^+) = \kappa \left( W^* + \frac{\min((s + b - W^*), 0)}{\alpha} \right) + \epsilon(s + b) . \]

(7.9)

At points in between rebalancing times, i.e. \(t \notin T\), the usual arguments (from SDEs (3.3\textsuperscript{3} 3.4), and Forsyth (2022b)) give

\[ V_t + \frac{(\sigma^s)^2 e^s}{2} V_{ss} + (\mu^s - \lambda^s \gamma^s) s V_s + \lambda^s \int_{-\infty}^{+\infty} V(s, e^{\gamma} s, b, t) f^s(y) \, dy + \frac{(\sigma^b)^2 b^2}{2} V_{bb} \]

\[ + (\mu^b + \mu^b \mathbf{1}_{b<0} - \lambda^b \gamma^b) b V_b + \lambda^b \int_{-\infty}^{+\infty} V(s, e^{\gamma} b, t) f^b(y) \, dy - (\lambda^s + \lambda^b V + \rho_{sb} \sigma^s \sigma^b sb V_{sb} = 0 , \]

\[ s \geq 0 ; b \geq 0 . \]

(7.10)

In case of insolvency\textsuperscript{10} \(s = 0, b < 0\)

\[ V_t + \frac{(\sigma^b)^2 b^2}{2} V_{bb} + (\mu^b + \mu^b \mathbf{1}_{b<0} - \lambda^b \gamma^b) b V_b + \lambda^b \int_{-\infty}^{+\infty} V(0, e^{\gamma} b, t) f^b(y) \, dy - \lambda^b V = 0 , \]

\[ s = 0 ; b < 0 . \]

(7.11)

It will be convenient, for computational purposes, to re-write equation (7.11) in terms of debt \(\hat{b} = -b\) when \(b < 0\). Now let \(\hat{V}(\hat{b}, t) = V(0, b, t), b < 0, \hat{b} = -b\) in equation (7.11) to give

\[ \hat{V}_t + \frac{(\sigma^b)^2 b^2}{2} \hat{V}_{\hat{b} \hat{b}} + (\mu^b + \mu^b - \lambda^b \gamma^b \hat{b} V_{\hat{b}} + \lambda^b \int_{-\infty}^{+\infty} \hat{V}(e^{\gamma} \hat{b}, t) f^b(y) \, dy - \lambda^b \hat{V} = 0 , \]

\[ s = 0 ; b < 0 ; \hat{b} = -b . \]

(7.12)

\textsuperscript{10} Insolvency can only occur due to the minimum withdrawals specified.
Note that equation (7.12) is now amenable to a transformation of the form \( \hat{x} = \log \hat{b} \) since \( \hat{b} > 0 \), which is required when using a Fourier method (Forsyth and Labahn, 2019; Forsyth, 2022b) to solve equation (7.12).

After rebalancing, if \( b \geq 0 \), then \( b \) cannot become negative, since \( b = 0 \) is a barrier in equation (7.11). However, \( b \) can become negative after withdrawals, in which case \( b \) remains in the state \( b < 0 \), where equation (7.12) applies, unless there is an injection of cash to move to a state with \( b > 0 \). The terminal condition for equation (7.12) is

\[
\hat{V}(\hat{b}, W^*, T^+) = \kappa \left( W^* + \frac{\min((-\hat{b} - W^*), 0)}{\alpha} \right) + \epsilon(-\hat{b}) ; \hat{b} > 0.
\]

(7.13)

A brief overview of the numerical algorithms is given in Appendix B along with a numerical convergence verification.

8 Data

We use data from the Center for Research in Security Prices (CRSP) on a monthly basis over the 1926:1-2020:12 period. Our base case tests use the CRSP US 30 day T-bill for the bond asset and the CRSP value-weighted total return index for the stock asset. This latter index includes all distributions for all domestic stocks trading on major U.S. exchanges. All of these various indexes are in nominal terms, so we adjust them for inflation by using the U.S. CPI index, also supplied by CRSP. We use real indexes since investors funding retirement spending should be focused on real (not nominal) wealth goals.

We use the threshold technique (Mancini, 2009; Cont and Mancini, 2011; Dang and Forsyth, 2016) to estimate the parameters for the parametric stochastic process models. Since the index data is in real terms, all parameters reflect real returns. Table 8.1 shows the results of calibrating the models to the historical data. The correlation \( \rho_{sb} \) is computed by removing any returns which occur at times corresponding to jumps in either series, and then using the sample covariance. Further discussion of the validity of assuming that the stock and bond jumps are independent is given in Forsyth (2020b).

<table>
<thead>
<tr>
<th>CRSP</th>
<th>( \mu^s )</th>
<th>( \sigma^s )</th>
<th>( \lambda^s )</th>
<th>( u^s )</th>
<th>( \eta_{1}^s )</th>
<th>( \eta_{2}^s )</th>
<th>( \rho_{sb} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.08912</td>
<td>0.1460</td>
<td>0.3263</td>
<td>0.2258</td>
<td>4.3625</td>
<td>5.5335</td>
<td>0.08420</td>
</tr>
<tr>
<td>30-day T-bill</td>
<td>( \mu^b )</td>
<td>( \sigma^b )</td>
<td>( \lambda^b )</td>
<td>( u^b )</td>
<td>( \eta_{1}^b )</td>
<td>( \eta_{2}^b )</td>
<td>( \rho_{sb} )</td>
</tr>
<tr>
<td></td>
<td>0.0046</td>
<td>0.0130</td>
<td>0.5053</td>
<td>0.3958</td>
<td>65.801</td>
<td>57.793</td>
<td>0.08420</td>
</tr>
</tbody>
</table>

Table 8.1: Estimated annualized parameters for double exponential jump diffusion model. Value-weighted CRSP index, 30-day T-bill index deflated by the CPI. Sample period 1926:1 to 2020:12.

Remark 8.1 (Choice of 30-day T-bill for the bond index). It might be argued that the bond index should hold longer dated bonds, e.g. ten-year treasuries, which would allow the investor to harvest

11More specifically, results presented here were calculated based on data from Historical Indexes, ©2020 Center for Research in Security Prices (CRSP), The University of Chicago Booth School of Business. Wharton Research Data Services (WRDS) was used in preparing this article. This service and the data available thereon constitute valuable intellectual property and trade secrets of WRDS and/or its third-party suppliers.
Data series | Optimal expected block size $\hat{b}$ (months)
--- | ---
Real 30-day T-bill index | 50.6
Real CRSP value-weighted index | 3.42

Table 9.1: Optimal expected blocksize $\hat{b} = 1/v$ when the blocksize follows a geometric distribution $Pr(b = k) = (1-v)^{k-1}v$. The algorithm in Patton et al. (2009) is used to determine $\hat{b}$. Historical data range 1926:1-2020:12.

the term premium. Long term bonds have enjoyed high real returns over the last thirty years, due to decreasing real interest rates during that time period. However, it is unlikely that this will continue to be true in the next thirty years. Hatch and White (1985) study the real returns of equities, short term T-bills, and long term corporate and government bonds, over the period 1950-1983, and conclude that, in both Canada and the US, only equities and short term T-bills had non-negative real returns. Inflation (both US and Canada) averaged about 4.75% per year over the period 1950-1983. If one imagines that the next thirty years will be a period with inflationary pressures, then this suggests that the defensive asset should be short term T-bills. However, there is nothing in the methodology in this paper which prevents us from using other underlying bonds in the bond index.

9 Historical Market

We compute and store the optimal controls based on the parametric model (3.3-3.4) as for the synthetic market case. However, we compute statistical quantities using the stored controls, but using bootstrapped historical return data directly. We remind the reader that all returns are inflation adjusted. We use the stationary block bootstrap method (Politis and Romano, 1994; Politis and White, 2004; Patton et al., 2009; Cogneau and Zakalmouline, 2013; Dichtl et al., 2016; Cavaglia et al., 2022; Simonian and Martirosyan, 2022; Anarkulova et al., 2022). A key parameter is the expected blocksize. Sampling the data in blocks accounts for serial correlation in the data series. We use the algorithm in Patton et al. (2009) to determine the optimal blocksize for the bond and stock returns separately, see Table 9.1. We use a paired sampling approach to simultaneously draw returns from both time series. In this case, a reasonable estimate for the blocksize for the paired resampling algorithm would be about 2.0 years. We will give results for a range of blocksizes as a check on the robustness of the bootstrap results. Detailed pseudo-code for block bootstrap resampling is given in Forsyth and Vetzal (2019).

10 Investment Scenario

Table 10.1 shows our base case investment scenario. We will use thousands of dollars as our units of wealth in the following. For example, a withdrawal of 40 per year corresponds to $40,000 per year (all values are real, i.e. inflation adjusted), with an initial wealth of 1000 ($1,000,000). This would correspond to the use of the four per cent rule (Bengen, 1994). Our base case scenario assumes a fee of 50 bps per year. We refer to Chen et al. (2021) for a discussion of tontine fees.

As a motivating example, we consider a 65-year old Canadian retiree who has a pre-retirement salary of $100,000 per year, with $1,000,000 in a DC savings account. Government benefits are assumed to amount to about $20,000 per year (real). The retiree wishes the DC plan to generate
at least $40,000 per year (real), so that the DC plan and government benefits replace 60% of pre-retirement income. We assume that the retiree owns mortgage-free real estate worth about $400,000.

In an act of mental accounting, the retiree plans to use the real estate as a longevity hedge, which could be monetized using a reverse mortgage. In the event that the longevity hedge is not needed, the real-estate will be a bequest. Of course, the retiree would like to withdraw more than $40,000 per year from the DC plan, but has no use for withdrawals greater than $80,000 per year. We make the further assumption that the real-estate holdings can generate $200,000 through a reverse mortgage. Hence, as a rough rule of thumb, any expected shortfall at $T = 30$ years greater than $-200,000$ is an acceptable level of risk.

Our view that personal real estate is not fungible with investment assets (unless investment assets are depleted) is consistent with the behavioral life cycle approach originally described in Shefrin and Thaler [1988] and Thaler [1990]. In this framework, investors divide their wealth into separate “mental accounts” containing funds intended for different purposes such as current spending or future need.

We take the view of a 65-year old retiree, who wants to maximize her total withdrawals, and minimize the risk of running out of savings, assuming that she lives to the age of 95. We also assume that the retiree has no bequest motive.

Recall that Bengen [1994] attempted to determine a safe real withdrawal rate, and constant allocation strategy, such that the probability of running out of cash after 30 years of retirement was small. In other words, Bengen [1994] maximized total withdrawals over a 30 year period, assuming that the retiree survived for the entire 30 years. This is, of course, a conservative assumption.

In our case, we are essentially answering the same question. The key difference here is that we (i) allow for dynamic asset allocation, (ii) allow variable withdrawals (within limits) and (iii) assume a tontine overlay.

<table>
<thead>
<tr>
<th>Retiree</th>
<th>65-year old Canadian male</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tontine Gain $T^g$</td>
<td>equation (2.2)</td>
</tr>
<tr>
<td>Group Gain $G$ ( see equation (2.12) )</td>
<td>1.0</td>
</tr>
<tr>
<td>Mortality table</td>
<td>CPM 2014</td>
</tr>
<tr>
<td>Investment horizon $T$ (years)</td>
<td>30.0</td>
</tr>
<tr>
<td>Equity market index</td>
<td>CRSP Cap-weighted index (real)</td>
</tr>
<tr>
<td>Bond index</td>
<td>30-day T-bill (US) (real)</td>
</tr>
<tr>
<td>Initial portfolio value $W_0$</td>
<td>1000</td>
</tr>
<tr>
<td>Cash withdrawal/rebalancing times</td>
<td>$t = 0, 1.0, 2.0, \ldots, 29.0$</td>
</tr>
<tr>
<td>Maximum withdrawal (per year)</td>
<td>$q_{\text{max}} = 80$</td>
</tr>
<tr>
<td>Minimum withdrawal (per year)</td>
<td>$q_{\text{min}} = 40$</td>
</tr>
<tr>
<td>Equity fraction range</td>
<td>$[0, 1]$</td>
</tr>
<tr>
<td>Borrowing spread $\mu_c^b$</td>
<td>0.02</td>
</tr>
<tr>
<td>Rebalancing interval (years)</td>
<td>1.0</td>
</tr>
<tr>
<td>$\alpha$ (EW-ES)</td>
<td>.05</td>
</tr>
<tr>
<td>Fees $T^f$ ( see equation (4.5) )</td>
<td>50 bps per year</td>
</tr>
<tr>
<td>Stabilization $\epsilon$ ( see equation (6.2) )</td>
<td>$-10^{-4}$</td>
</tr>
<tr>
<td>Market parameters</td>
<td>See Table 8.1</td>
</tr>
</tbody>
</table>

Table 10.1: Input data for examples. Monetary units: thousands of dollars. CPM2014 is the mortality table from the Canadian Institute of Actuaries.
11 Constant Withdrawal, Constant Equity Fraction.

As a preliminary example, in this section we present results for the scenario in Table 10.1 except that a constant withdrawal of 40 per year is specified, along with a constant weight in stocks at each rebalancing date.

Table 11.1 gives the results for various values of the constant weight equity fraction, in the synthetic market. The best result\(^\text{12}\) for $ES$ (the largest value) occurs at the rather low constant equity weight of $p = 0.1$, with an $ES = -239$. Table 11.2 gives similar results, this time using bootstrap resampling of the historical data (the historical market). This time, the best value of $ES = -305$ occurs for a constant equity fraction of $p = 0.4$. Consequently, in both the historical and synthetic market, the constant weight, constant withdrawal strategy fails to meet our risk criteria of $ES > -200$.

These simulations indicate that there is significant depletion risk for the constant withdrawal, constant weight strategy suggested in Bengen (1994).

| Equity fraction $p$ | $E[\sum q_i]/T$ | ES (5%) | Median $|W_T|$ |
|---------------------|------------------|---------|------------|
| 0.0                 | 40               | -302.57 | -150.56    |
| 0.1                 | 40               | -238.62 | -6.82      |
| 0.2                 | 40               | -245.48 | 168.10     |
| 0.3                 | 40               | -280.27 | 386.05     |
| 0.4                 | 40               | -330.37 | 649.58     |
| 0.5                 | 40               | -391.61 | 958.33     |
| 0.6                 | 40               | -461.54 | 1312.17    |
| 0.7                 | 40               | -538.04 | 1706.49    |
| 0.8                 | 40               | -619.31 | 2135.24    |

Table 11.1: Constant weight, constant withdrawals, synthetic market results. No tontine gains. Stock index: real capitalization weighted CRSP stocks; bond index: real 30-day T-bills. Parameters from Table 8.1. Scenario in Table 10.1. Units: thousands of dollars. Statistics based on $2.56 \times 10^6$ Monte Carlo simulation runs. $T = 30$ years.

| Equity fraction $p$ | $E[\sum q_i]/T$ | ES (5%) | Median $|W_T|$ |
|---------------------|------------------|---------|------------|
| 0.0                 | 40               | -508.44 | -155.04    |
| 0.1                 | 40               | -418.02 | -10.98     |
| 0.2                 | 40               | -350.00 | 164.75     |
| 0.3                 | 40               | -312.24 | 382.16     |
| 0.4                 | 40               | -305.52 | 649.04     |
| 0.5                 | 40               | -326.40 | 966.61     |
| 0.6                 | 40               | -370.18 | 1336.31    |
| 0.7                 | 40               | -432.55 | 1759.66    |
| 0.8                 | 40               | -509.00 | 2232.29    |


\(^{12}\)Recall that $ES$ is defined in terms of the left tail mean of final wealth (not losses) hence a larger value is preferred.
12 Synthetic Market Efficient Frontiers

Figure 12.1 shows the efficient EW-ES frontiers, computed in the synthetic market, for the following cases:

**Tontine:** the case in Table 10.1. The control is computed using the algorithm in Section 7 and then stored, and used in Monte Carlo simulations. The detailed frontier is given in Table D.1.

**No Tontine:** the case in Table 10.1, but without any tontine gains. The control is computed and stored, and then used in Monte Carlo simulations. The detailed frontier is given in Table D.2.

**Const q=40, Const p:** The best single point from Table 11.1, based on Monte Carlo simulations.

Note that all these strategies produce a minimum withdrawal of 40 per year (i.e. 4% real of the initial investment) for thirty years. However, the best result for the constant weight strategies was \((EW,ES) = (40, -239)\) This can be improved significantly by optimizing over withdrawals and asset allocation, but with no tontine gains. For example, from Table D.2 the nearest point with roughly the same level of risk is \((EW,ES) = (58, -242)\). However, the improvement with optimal controls and tontine gains is remarkable. For example, it seems reasonable to target a value of \(ES \simeq 0\). From Table D.1 we note the point \((EW,ES) = (69, 47)\), which is dramatically better than any No Tontine Pareto point. This can also be seen from the large outperformance in the EW-ES frontier compared to the No Tontine case in Figure 12.1.

![Figure 12.1: Frontiers generated from the synthetic market. Parameters based on real CRSP index, real 30-day T-bills (see Table 8.1). Tontine case is as in Table 10.1. The No Tontine case uses the same scenario, but with no tontine gains, and no fees. The Const q, Const p case has q = 40, p = 0.10, with no tontine gains, which is the best result from Table 11.1 assuming no tontine gains, and no fees. Units: thousands of dollars.](image)

12.1 Effect of Fees

Figure 12.2 shows the effect of varying the annual fee in the synthetic market, for the scenario in Table 10.1. Recall that the base case specified a fee of 50 bps per year. Assuming a shortfall target of \(ES \simeq 0\), then the effect of fees in the range 0 – 100 bps is quite modest. Even with annual fees of 100 bps, the Tontine case still significantly outperforms the No Tontine case (which is assumed to have no fees).
12.2 Effect of Random G

Recall the definition of the group gain $G_i$ at time $t_i$ in equation (2.9). Basically, the group gain is used to ensure that the total amount of mortality credits disbursed is exactly equal to the total amount forfeited by tontine participants who have died in $(t_{i-1}, t_i)$.

If Condition 2.1 holds, then we expect that the effect of randomly varying $G_i$ to have a small cumulative effect. In [Fullmer and Sabin (2019)] and [Winter and Planchet (2022)], the authors create synthetic tontine pools, where the investors have different initial wealth, ages, genders, and investment strategies. These pools are perpetual, i.e. new members join as original members die. It is assumed that the investors can only select an asset allocation strategy from a stock index and a bond index, both of which follow a geometric Brownian motion (GBM).

The payout rules are different from those suggested in this paper, however, it is instructive to observe the following. In [Fullmer and Sabin (2019)], the perpetual tontine pool has 15,000 investors at steady-state. After the initial start-up period, the expected value of the group gain $G_i$ at each rebalancing time is close to unity, with a standard deviation of about 0.1. [Fullmer and Sabin (2019)] also note that there is essentially no correlation between investment returns, and the group gain.

Figure 12.3 shows the effect of randomly varying $G_i$. The curve labeled $G = 1.0$ is the base case EW-ES curve from the scenario in Table 10.1 in the synthetic market (parameters in Table 8.1). The controls from this base case are stored, and then used in Monte Carlo simulations, where $G$ is assumed to have a normal distribution with mean one, and standard deviation of 0.1. The EW-ES curves for both cases essentially overlap, except for very large values of ES, which are not of any practical interest. We get essentially the same result if we use a uniform distribution for $G$ with $E[G] = 1$, with the same standard deviation. This is not surprising, since, assuming that the value function is smooth, then a simple Taylor series argument shows that, for any assumed distribution of $G$ with mean one, the effect of randomness of $G$ is a second order effect in the standard deviation.

Of course, we cannot determine the actual distribution of $G$ without a detailed knowledge of the characteristics of the tontine pool. In fact, if we knew the distribution, we could include it in the formulation of the optimal control problem. However, knowledge of the distribution of $G$ is unlikely to be available to pool participants in practice.

Nevertheless, the simulations in [Fullmer and Sabin 2019] [Winter and Planchet 2022], coupled...
with our results as shown in Figure 12.3, suggest that, for a sufficiently large, diversified pool of investors that the effects of randomly varying $G$ are negligible.

![Diagram](image)

**Figure 12.3:** Effect of randomly varying group gain $G$ (Section 2.2). Frontiers generated from the synthetic market. Parameters based on real CRSP index, real 30-day T-bills (see Table 8.1). Base case Tontine ($G = 1.0$) is as in Table 10.1. Random G case uses the control computed for the base case, but in the Monte Carlo simulation, $G$ is normally distributed with mean one and standard deviation 0.1. Units: thousands of dollars.

13 Bootstrapped Results

As discussed in Section 9, a key parameter in the stationary block bootstrap technique is the expected blocksize. In Figure 13.1(a), we show the results of the following test. We compute and store the optimal controls, based on the synthetic market. Then we use these controls, but carry out tests on bootstrapped historical data. The efficient frontiers in Figure 13.1(a) for $ES < 1000$ essentially overlap for all expected blocksizes in the range $0.5 - 5.0$. Since it is probably not of interest to aim for an ES of 1000 (which is one million dollars) at age 95, this indicates that the computed strategy is robust to parameter uncertainty.

Figure 13.1(b) compares the efficient frontier tested in the historical market (expected blocksize 2 years), with the efficient frontier in the synthetic market. We observe that the synthetic and historical curves overlap for $ES < 1000$, which again verifies that the controls are robust to data uncertainty. The efficient frontiers/points for the No Tontine case and the constant weight, constant withdrawal strategy (computed in the historical market) are also shown. The Tontine overlay continues to outperform the No Tontine case by a wide margin.

14 Detailed Historical Market Results: EW-ES Controls

In this section, we examine some detailed characteristics of the optimal EW-ES strategy, tested in the historical market for the scenario in Table 10.1. Figure 14.1 shows the percentiles of fraction in stocks, wealth, and withdrawals versus time, for the EW-ES control with $\kappa = 0.18$, with $(EW, ES) = (69, 204)$. To put this in perspective, recall that this strategy never withdraws less than 40 per year.

Compare this to the best case for a constant withdrawal, constant weight strategy (no tontine) from Table 11.2, which has $(EW, ES) = (40, -306)$, or to the optimal EW-ES strategy, but with no tontine, from Table E.2 which has $(EW, ES) = (70, -806)$.
Figure 13.1: Optimal strategy determined by solving Problem 6.2 in the synthetic market, parameters in Table 8.1. Control stored and then tested in bootstrapped historical market. Inflation adjusted data, 1926:1-2020:12. Non-Pareto points eliminated. Expected blocksize (Blk, years) used in the bootstrap resampling method also shown. Units: thousands of dollars. The const q, const p case had $(p,q) = (0.4, 40)$ (no tontine gains). This is the best result for the constant $(p,q)$ case, shown in Table 11.2.

(a) Efficient frontiers, historical market, effect of varying expected blocksize (Blk).

(b) Efficient frontiers, historical market, expected blocksize $= 2.0$ years. Synthetic market frontier (tontine) also shown.

Figure 14.1(a) shows that the median optimal fraction in stocks starts at about 0.60, then drops to about 0.20 at 15 years, finally ending up at zero in year 26. Figure 14.1(b) indicates that for the years in the span of 20 – 30, the median and fifth percentiles of wealth are fairly tightly clustered, with the fifth percentile being well above zero at all times. The optimal withdrawal percentiles are shown in Figure 14.1(c). The median withdrawal starts at 40 per year, then increases to the maximum withdrawal of 80 in years 3 – 4, and remains at 80 for the remainder of the thirty year time horizon.

Figure 14.2 shows the optimal control heat maps for the fraction in stocks and withdrawal amounts, for the scenario in Table 10.1. Figure 14.2(a) shows a smooth behavior of the optimal fraction in stocks as a function of $(W,t)$. This can be compared with the equivalent heat map for the EW-ES control in Forsyth (2022b) (no tontine gains), which is much more aggressive at changing the asset allocation in response to changing wealth amounts. The smoothness of the controls in Figure 14.2(a) appears to be due to the rapid de-risking of a strategy which includes tontine gains, which provides a natural protection against sudden stock index drops. The upper blue zone in Figure 14.2(a) is de-risking due to the fact that, with sufficiently large wealth, there is essentially no probability of running out of cash even at the maximum withdrawal amount. The use of the stabilization factor $\epsilon = -10^{-4}$ forces the strategy to increase the weight in bonds for large values of wealth (see equation (6.2)). The lower red zone is in response to extremely poor wealth outcomes, which means that the optimal strategy is to invest 100% in stocks and hope for the best. However, this is an extremely unlikely outcome, as can be verified from Figure 14.1(b).

From Figure 14.2(b) we can observe that the optimal withdrawal strategy is essentially a bang-bang control, i.e. withdraw at either the maximum or minimum amount per year. This is not unexpected, as discussed in Appendix C. We also note that this type of strategy has been suggested previously, based on heuristic reasoning.14

13"When you have won the game, stop playing,” William Bernstein.
14“If we have a good year, we take a trip to China....if we have a bad year, we stay home and play canasta,” retired professor Peter Ponzo, discussing his DC plan withdrawal strategy https://www.theglobeandmail.com/
Figure 14.1: Scenario in Table 10.1. EW-ES control computed from problem EW-ES Problem (6.2). Parameters based on the real CRSP index, and real 30-day T-bills (see Table 8.1). Control computed and stored from the Problem (6.2) in the synthetic market. Control used in the historical market, $10^6$ bootstrap samples. $q_{\min} = 40, q_{\max} = 80$ (per year), $\kappa = 0.18$. $W^* = 385$. Units: thousands of dollars.

Figure 14.2: Optimal EW-ES. Heat map of controls: fraction in stocks and withdrawals, computed from Problem EW-ES (6.1). Real capitalization weighted CRSP index, and real 30-day T-bills. Scenario given in Table 10.1. Control computed and stored from the Problem 6.2 in the synthetic market. $q_{\min} = 40, q_{\max} = 80$ (per year). $\kappa = 0.18$. $W^* = 385$. $\epsilon = -10^{-4}$. Normalized withdrawal $(q - q_{\min})/(q_{\max} - q_{\min})$. Units: thousands of dollars.

15 Discussion

Traditional annuities with true inflation protection are unavailable in Canada\footnote{\text{report-on-business/math-prof-tests-investing-formulas-strategies/article22397218/}.} Since inflation is expected to be a major factor in the coming years, inflation protection is a valuable aspect of the optimal EW-ES strategy, with a tontine overlay\footnote{Some providers advertise annuities with inflation protection, however this is simply an escalating nominal payout, based on a fixed escalation rate.} This strategy has an expected real withdrawal\footnote{Examination of historical periods of high inflation suggests that a portfolio of short term T-bills and an equal weight stock index generates significant positive real returns, see Forsyth (2022a).}
rate, over thirty years, of about 7% of the initial capital (per annum), never withdraws less than
4% of initial capital per annum, and a positive ES (expected shortfall) at the 5% level after thirty
years.

Consequently, if we consider a retiree with no bequest motive, then joining a tontine pool, and
following an optimal EW-ES strategy, is certainly an excellent alternative to a life annuity. Hence,
it could be argued that, going forward, the EW-ES optimal tontine pool strategy has less risk than
a conventional annuity.

Of course, there is no free lunch here. The reason that the tontine approach has a higher
mean (and median) payout is that it is not guaranteed. There is some flexibility in the withdrawal
amounts, and the portfolio contains risky assets. However, the ultimate risk, as measured by the
expected shortfall at year thirty, is very small. We can also see that the median payout rises rapidly
to the maximum withdrawal rate (8% real of the initial investment) within 3-4 years of retirement,
and stays at the maximum for the remainder of the thirty year horizon.

As well, the investor forfeits the entire portfolio in the event of death. Although this is often
considered a drawback, we remind the reader that annuities and defined benefit (DB) plans have
this same property (restricting attention to a single retiree with no guarantee period). Of course,
it is possible to overlay various guarantees on to the tontine pool, e.g. a guarantee period, a money
back guarantee, or joint and survivor benefits. The cost of these guarantees would, of course, reduce
the expected annual withdrawals.

These results are robust to fees in the range of 50-100 bps per year. The long term results are
also insensitive to random group gains.\footnote{The randomness of the group gain is due to fact that real tontine pools will be finite and heterogeneous.}

However, the tontine gains (after fees) are comparatively small for retirees in the 65-70 age range.
This suggests that it may be optimal to delay joining a tontine until the investor has attained an
age of 70 or more.

Although we have explicitly excluded a bequest motive from our considerations, note that the
median withdrawal strategy rapidly ramps up to the maximum withdrawal within a few years of
retirement, and remains there for the entire remaining retirement period. Although it is commonly
postulated that retirement consumption follows a \textit{U-shaped} pattern, recent studies indicate that
real retirement consumption falls with age (at least in countries which do not have large end of life
expenses)\footnote{Moshe Milevsky, an advocate of modern tontines, is quoted in the Toronto Star (April 13, 2021) as noting that "If you give up some of your money when you die, you can get more when you are alive."}. In this case, the withdrawals which occur towards the end of the
retirement period may exceed consumption. This allows the retiree to use these excess cash flows
as a living bequest to relatives or charities.

\section{16 Conclusions}

DC plan decumulation strategies are typically based on some variant of the \textit{four per cent rule}
\cite{Bengen_1994}. However, bootstrap tests of these rules using historical data show a significant risk
of running out of savings at the end of a thirty year retirement planning horizon.

This risk can be significantly reduced by using optimal stochastic control methods, where the
controls are the asset allocation strategy and the withdrawal amounts (subject to maximum and
minimum constraints)\cite{Forsyth_2022b,Forsyth_2020}.

However, if we assume the retiree couples an optimal allocation/withdrawal strategy with par-
ticipation in a tontine fund, then the risk of portfolio depletion after 30 years is virtually eliminated.
At the same time, the cumulative total withdrawals are considerably increased compared with the
previous two strategies. Of course, this comes at a price: the retiree forfeits her portfolio upon death. Hence the tontine overlay is most appealing to investors who have no bequest motivation, or who manage bequests using other funds.

We should also note that individual tontine accounts allow for complete flexibility in asset allocation strategies and do not require purchase of expensive investment products. These accounts are essentially peer-to-peer longevity risk management tools. Consequently, the custodian of these accounts bears no risk, and incurs only bookkeeping costs. Hence the fees charged by the custodian of these accounts can be very low. If desired, the retiree can pay for additional investment advice in a completely transparent manner.

17 Acknowledgements

Forsyth’s work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) grant RGPIN-2017-03760. Vetzal’s work was supported by a Canadian Securities Institute Research Foundation grant 2021-2022.

18 Declaration

The authors have no conflicts of interest to report.
Appendix

A Induced Time Consistent Strategy

Denote the investor’s initial wealth at \( t_0 \) by \( W_0^- \). Then we have the following result:

**Proposition A.1** (Pre-commitment strategy equivalence to a time consistent policy for an alternative objective function). The pre-commitment EW-ES strategy \( P^* \) determined by solving \( J(0, W_0, t_0^-) \) \( \) (with \( W^*(0, W_0^-) \) from equation (6.5)) is the time consistent strategy for the equivalent problem \( TCEQ \) (with fixed \( W^*(0, W_0^-) \)), with value function \( \tilde{J}(s,b,t) \) defined by

\[
(TCEQ_{tn}(\kappa/\alpha)) : \quad \tilde{J}(s,b,t_n) = \sup_{P_n \in A} \left\{ \int_{t_n}^{X(t_n)} \sum_{i=n}^{M} q_i + \frac{\kappa}{\alpha} \min(W_T - W^*(0, W_0^-), 0) \right\} .
\] (A.1)

**Proof.** This follows similar steps as in Forsyth (2020a), proof of Proposition 6.2, with the exception that the reward in Forsyth (2020a) is expected terminal wealth, while here the reward is total withdrawals.

**Remark A.1** (An Implementable Strategy). Given an initial level of wealth \( W_0^- \) at \( t_0 \), then the optimal control \( P^* \) for the pre-commitment problem (6.2) is the same optimal control for the time consistent problem (A.1), \( \forall t > 0 \). Hence we can regard problem \( (TCEQ_{tn}(\kappa/\alpha)) \) as the EW-ES induced time consistent strategy. Thus, the induced strategy is implementable, in the sense that the investor has no incentive to deviate from the strategy computed at time zero, at later times Forsyth (2020a).

**Remark A.2** (EW-ES Induced Time Consistent Strategy). In the following, we will consider the actual strategy followed by the investor for any \( t > 0 \) as given by the induced time consistent strategy \( (TCEQ_{tn}(\kappa/\alpha)) \) in equation (A.1), with a fixed value of \( W^*(0, W_0^-) \), which is identical to the EW-ES strategy at time zero. Hence, we will refer to this strategy in the following as the EW-ES strategy, with the understanding that this refers to strategy \( (TCEQ_{tn}(\kappa/\alpha)) \) for any \( t > 0 \).

B Numerical Techniques

We solve problems (6.2) using the techniques described in detail in Forsyth and Labahn (2019); Forsyth (2020a, 2022b). We give only a brief overview here.

We localize the infinite domain to \((s,b) \in [s_{\text{min}}, s_{\text{max}}] \times [b_{\text{min}}, b_{\text{max}}]\), and discretize \([b_{\text{min}}, b_{\text{max}}]\) using an equally spaced log \( b \) grid, with \( n_b \) nodes. Similarly, we discretize \([s_{\text{min}}, s_{\text{max}}]\) on an equally spaced log \( s \) grid, with \( n_s \) nodes. Localization errors are minimized using the domain extension method in Forsyth and Labahn (2019).

At rebalancing dates, we solve the local optimization problem (7.7) by discretizing \((q(\cdot), p(\cdot))\) and using exhaustive search. Between rebalancing dates, we solve the two dimensional partial integro-differential equation (PIDE) (7.10) using Fourier methods Forsyth and Labahn (2019) Forsyth.

\[\text{To be perfectly precise here, in the event that the control is non-unique, we impose a tie-breaking strategy to generate a unique control.}\]

\[\text{Assuming that the same tie breaking strategy is used as for the pre-commitment problem.}\]
Finally, the optimization problem (7.4) is solved using a one-dimensional optimization technique. We used the value $\epsilon = -10^{-4}$ in equation (7.2), which forces the investment strategy to be bond heavy if the remaining wealth in the investor’s account is large, and $t \to T$. Using this small value of gave the same results as $\epsilon = 0$ for the summary statistics, to four digits. This is simply because the states with very large wealth have low probability. However, this stabilization procedure produced smoother heat maps for large wealth values, without altering the summary statistics appreciably.

B.1 Convergence Test: Synthetic Market

We compute and store the optimal controls from solving Problem 6.2 using the parametric model of the stock and bond processes. We then use the stored controls in Monte Carlo simulations to generate statistical results. As a robustness check, we also use the stored controls and simulate results using bootstrap resampling of historical data.

Table B.1 shows a detailed convergence test for the base case problem given in Table 10.1, for the EW-ES problem. The results are given for a sequence of grid sizes, for the dynamic programming algorithm in Section 7 and Appendix B. The dynamic programming algorithm appears to converge at roughly a second order rate. The optimal control computed using dynamic programming is stored, and then used in Monte Carlo computations. The MC results are in good agreement with the dynamic programming solution.

For all the numerical examples, we will use the $2048 \times 2048$ grid, since this seems to be accurate enough for our purposes.

<table>
<thead>
<tr>
<th>Algorithm in Section 7 and Appendix B</th>
<th>Monte Carlo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grid</td>
<td>ES (5%)</td>
</tr>
<tr>
<td>512 $\times$ 512</td>
<td>108.13</td>
</tr>
<tr>
<td>1024 $\times$ 1024</td>
<td>158.88</td>
</tr>
<tr>
<td>2048 $\times$ 2048</td>
<td>201.88</td>
</tr>
<tr>
<td>4096 $\times$ 4096</td>
<td>206.56</td>
</tr>
</tbody>
</table>

Table B.1: Convergence test, real stock index: deflated real capitalization weighted CRSP, real bond index: deflated 30 day T-bills. Scenario in Table 10.1. Parameters in Table 8.1. The Monte Carlo method used $2.56 \times 10^6$ simulations. The MC method used the control from the algorithm in Section 7. $\kappa = 0.185, \alpha = 0.05$ Grid refers to the grid used in the Algorithm in Section 7: $n_x \times n_b$, where $n_x$ is the number of nodes in the log $s$ direction, and $n_b$ is the number of nodes in the log $b$ direction. Units: thousands of dollars (real). $M$ is the total number of withdrawals (rebalancing dates).

C Continuous Withdrawal/Rebalancing Limit

In order to develop some intuition about the nature of the optimal controls, we will examine the limit as the rebalancing interval becomes vanishingly small.

Proposition C.1 (Bang-bang withdrawal control in the continuous withdrawal limit). Assume that

- the stock and bond processes follow (3.3) and (3.4),
Table B.2: No tontine case. Convergence test, real stock index: deflated real capitalization weighted CRSP, real bond index: deflated 30 day T-bills. Scenario in Table 10.1, but no tontine. Parameters in Table 8.1. The Monte Carlo method used $2.56 \times 10^6$ simulations. The MC method used the control from the algorithm in Section 7. $\kappa = 3.75, \alpha = .05$ Grid refers to the grid used in the Algorithm in Section B: $n_x \times n_b$, where $n_x$ is the number of nodes in the log $s$ direction, and $n_b$ is the number of nodes in the log $b$ direction. Units: thousands of dollars (real). $M$ is the total number of withdrawals (rebalancing dates). $W^* = -106.476$ on the finest grid.

<table>
<thead>
<tr>
<th>Grid</th>
<th>ES (5%)</th>
<th>$E[\sum_i q_i]/T$</th>
<th>Value Function</th>
<th>ES (5%)</th>
<th>$E[\sum_i q_i]/M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>512 x 512</td>
<td>-203.31</td>
<td>54.08</td>
<td>860.033</td>
<td>-191.99</td>
<td>53.96</td>
</tr>
<tr>
<td>1024 x 1024</td>
<td>-191.40</td>
<td>53.58</td>
<td>889.613</td>
<td>-188.07</td>
<td>53.53</td>
</tr>
<tr>
<td>2048 x 2048</td>
<td>-188.91</td>
<td>53.57</td>
<td>898.712</td>
<td>-188.14</td>
<td>53.55</td>
</tr>
<tr>
<td>4096 x 4096</td>
<td>-188.04</td>
<td>53.54</td>
<td>901.106</td>
<td>-187.95</td>
<td>53.53</td>
</tr>
</tbody>
</table>

- the portfolio is continuously rebalanced, and withdrawals occur at a continuous (finite) rate $\hat{q} \in [\hat{q}_{\min}, \hat{q}_{\max}]$.
- the HJB equation for the EW-ES problem in the continuous rebalancing limit has bounded derivatives w.r.t. total wealth,
- in the event of ties for the control $\hat{q}$, the smallest withdrawal is selected,

then the optimal withdrawal control $\hat{q}^*(\cdot)$ for the EW-ES problem ($PCES_t(\kappa)$) and for the EW-LS problem ($EWLS_t(\hat{\kappa})$) is bang-bang, $\hat{q}^* \in \{\hat{q}_{\min}, \hat{q}_{\max}\}$.

Proof. This follows the same steps as in Forsyth (2022b).

Remark C.1 (Bang-bang control for discrete rebalancing/withdrawals). Proposition C.1 suggests that, for sufficiently small rebalancing intervals, we can expect the optimal $q$ control (finite withdrawal amount) to be bang-bang, i.e. it is only optimal to withdraw either the maximum amount $q_{\max}$ or the minimum amount $q_{\min}$. However, it is not clear that this will continue to be true for the case of yearly rebalancing (which we specify in our numerical examples), and finite amount controls $q$.
In fact, we do observe that the finite amount control $q$ is very close to bang-bang in our numerical experiments, even for yearly rebalancing.

D Detailed Efficient Frontiers: Synthetic Market
\[
\kappa \quad E\left[\sum_i q_i/T\right] \quad \text{ES (5\%)} \quad \text{Median}[W_T] \quad W^* \\
0.15 \quad 70.06 \quad -309.569 \quad 189.48 \quad .490 \\
0.170 \quad 70.04 \quad -270.13 \quad 185.19 \quad .489 \\
.185 \quad 68.51 \quad 46.77 \quad 599.42 \quad 385.28 \\
.20 \quad 67.56 \quad 203.87 \quad 820.65 \quad 585.97 \\
.25 \quad 66.41 \quad 384.76 \quad 1058.40 \quad 802.40 \\
.30 \quad 62.22 \quad 912.29 \quad 1754.40 \quad 1439.83 \\
.50 \quad 58.48 \quad 732.34 \quad 1517.04 \quad 1220.33 \\
1.0 \quad 54.81 \quad 1372.46 \quad 2327.42 \quad 2021.22 \\
10.0 \quad 48.96 \quad 1457.52 \quad 2484.58 \quad 2151.79 \\
\infty \quad 40 \quad 1460.76 \quad 2885.85 \quad 2173.04 \\
\]

Table D.1: EW-ES synthetic market results for optimal strategies, assuming the scenario given in Table 10.1. Tontine gains assumed. Stock index: real capitalization weighted CRSP stocks; bond index: real 30-day T-bills. Parameters from Table 8.1. Units: thousands of dollars. Statistics based on \(2.56 \times 10^6\) Monte Carlo simulation runs. Control is computed using the Algorithm in Section 7 and Section B, stored, and then used in the Monte Carlo simulations. \(q_{\min} = 0.40, \ q_{\max} = 80\) (annually). \(T = 30\) years, \(\epsilon = -10^{-4}\).

\[
\kappa \quad E\left[\sum_i q_i/T\right] \quad \text{ES (5\%)} \quad \text{Median}[W_T] \quad W^* \\
0.180 \quad 69.17 \quad -823.76 \quad -2.51 \quad -691.81 \\
1.0 \quad 61.38 \quad -319.66 \quad -39.47 \quad -229.18 \\
1.5 \quad 58.98 \quad -260.92 \quad -65.88 \quad -179.60 \\
1.75 \quad 57.97 \quad -242.34 \quad -74.74 \quad -161.25 \\
2.5 \quad 55.86 \quad -211.03 \quad -81.44 \quad -132.87 \\
3.75 \quad 53.55 \quad -188.14 \quad -81.11 \quad -107.00 \\
5.0 \quad 52.08 \quad -177.88 \quad -78.39 \quad -90.10 \\
6.25 \quad 51.29 \quad -173.59 \quad -79.08 \quad -89.03 \\
7.5 \quad 50.72 \quad -171.05 \quad -79.3 \quad -88.25 \\
10.0 \quad 49.89 \quad -168.16 \quad -78.77 \quad -87.18 \\
100.0 \quad 46.41 \quad -162.86 \quad -68.28 \quad -77.47 \\
\infty \quad 40 \quad -162.67 \quad +5.72 \quad -76.0 \\
\]

Table D.2: EW-ES synthetic market results for optimal strategies, assuming the scenario given in Table 10.1. No tontine gains assumed. Stock index: real capitalization weighted CRSP stocks; bond index: real 30-day T-bills. Parameters from Table 8.1. Units: thousands of dollars. Statistics based on \(2.56 \times 10^6\) Monte Carlo simulation runs. Control is computed using the Algorithm in Section 7 and Appendix B, stored, and then used in the Monte Carlo simulations. \(q_{\min} = 0.40, \ q_{\max} = 80\) (annually). \(T = 30\) years, \(\epsilon = -10^{-4}\).
### Table E.1: EW-EW Historical Market Results for Optimal Strategies

| $\kappa$ | $E[ \sum_i q_i ]/T$ | $ES$ (5%) | Median $|W_T|$ |
|----------|---------------------|------------|----------------|
| 0.15     | 71.25               | -165.23    | 157.16         |
| 0.17     | 71.01               | -138.15    | 153.13         |
| 0.18     | 68.94               | 204.20     | 573.29         |
| 0.185    | 67.99               | 369.26     | 769.96         |
| 0.20     | 66.64               | 546.98     | 1038.07        |
| 0.25     | 63.84               | 863.20     | 1500.51        |
| 0.30     | 62.08               | 1011.55    | 1739.21        |
| 0.5      | 58.13               | 1211.18    | 2115.22        |
| 1.0      | 54.50               | 1285.93    | 2330.33        |
| 10.0     | 49.42               | 1275.98    | 2485.58        |
| $\infty$| 40                  | 1280.97    | 2892.41        |

Table E.1: EW-EW historical market results for optimal strategies, assuming the scenario given in Table 10.1. Tontine gains assumed. Stock index: real capitalization weighted CRSP stocks; bond index: real 30-day T-bills. Parameters from Table 8.1. Units: thousands of dollars. Statistics based on $10^6$ bootstrap simulation runs. Expected blocksize = 2 years. Control is computed using the Algorithm in Section 7 and Appendix B, stored, and then used in the bootstrap simulations. $q_{\min} = 40$, $q_{\max} = 80$ (annually). $T = 30$ years, $\epsilon = -10^{-4}$.

### Table E.2: EW-EW Historical Market Results for Optimal Strategies

| $\kappa$ | $E[ \sum_i q_i ]/T$ | $ES$ (5%) | Median $|W_T|$ |
|----------|---------------------|------------|----------------|
| 0.18     | 69.91               | -805.65    | -31.84         |
| 1.0      | 61.77               | -290.03    | -40.87         |
| 1.5      | 59.21               | -248.15    | -77.26         |
| 1.75     | 58.16               | -235.46    | -78.50         |
| 2.5      | 6.02                | -219.00    | -81.84         |
| 3.75     | 53.78               | -209.9     | -80.68         |
| 5.0      | 52.43               | -207.15    | -77.25         |
| 6.25     | 51.74               | -209.02    | -78.11         |
| 7.5      | 51.26               | -210.38    | -78.48         |
| 10.0     | 50.58               | -212.41    | -77.95         |
| 100.0    | 47.72               | -217.82    | -67.91         |
| $\infty$| 40.0                | -219.16    | +17.34         |

Table E.2: EW-EW historical market results for optimal strategies, assuming the scenario given in Table 10.1. No Tontine gains assumed. Stock index: real capitalization weighted CRSP stocks; bond index: real 30-day T-bills. Parameters from Table 8.1. Units: thousands of dollars. Statistics based on $10^6$ bootstrap simulation runs. Expected blocksize = 2 years. Control is computed using the Algorithm in Section 7 and Appendix B, stored, and then used in the bootstrap simulations. $q_{\min} = 40$, $q_{\max} = 80$ (annually). $T = 30$ years, $\epsilon = -10^{-4}$.
References


