BANK OF ENGLAND

Staff Working Paper No. 765

Macroprudential margins: a new countercyclical tool?

Cian O'Neill and Nicholas Vause

November 2018

BANK OF ENGLAND

Staff Working Paper No. 765
Macroprudential margins: a new countercyclical tool?

Cian O'Neill[1] and Nicholas Vause[2]

Abstract

We quantify the size of a fire-sale externality in the derivatives market in the absence of a macroprudential buffer on top of microprudential initial margin requirements. We show how this varies over the financial cycle with market volatility. We then assess the ability of a macroprudential buffer to reduce this externality. We find this depends critically on the release conditions of the buffer. A buffer could reduce or, if set appropriately, even eliminate the externality, as long as it was released when investors faced any significant collateral calls, regardless of whether these related to variation or initial margins. However, it could be harmful if it was released only with calls for additional initial margin. Predicated on ideal release conditions, we test the performance of macroprudential buffers based on 'anti-procyclicality' mechanisms in current regulations. These mechanisms can reduce the fire-sale externality in some market conditions, but not all. Conceptually, we devise alternative mechanisms that eliminate the externality, although it may be difficult for policymakers to specify these in practice. Finally, as an alternative to quantity-based solutions, we investigate the ability of taxes to reduce the externality. We find that such a price-based solution could also eliminate the externality if set appropriately, but this would require a high tax rate and the redistribution of significant tax revenues.

Key words: Collateral, derivatives, externality, fire sales, macroprudential policy.

JEL classification: G18, G23.

(1) Bank of England. Email: cian.o'neill@bankofengland.co.uk
(2) Bank of England. Email: nicholas.vause@bankofengland.co.uk

The views expressed in this paper are those of the authors, and not necessarily those of the Bank of England or its committees. We are grateful for comments from Franklin Allen, Stephen Cecchetti, Anil Kashyap, Yvan Lengwiler, David Murphy, Martin Oehmke, Jean-Pierre Zigrand and members of the European Systemic Risk Board's Expert Group on Margins and Haircuts and participants at seminars at the Bank of England, Bank for International Settlements and European Central Bank.

The Bank's working paper series can be found at www.bankofengland.co.uk/working-paper/staff-working-papers

Publications and Design Team, Bank of England, Threadneedle Street, London, EC2R 8AH
Telephone +44 (0)20 7601 4030 email publications@bankofengland.co.uk

ISSN 1749-9135 (on-line)

1. Introduction

Regulatory reforms following the 2007-08 global financial crisis will result in the vast majority of derivative exposures in the core of the financial system being backed by collateral. First, mandates have been introduced in major jurisdictions requiring financial institutions to clear new trades in many of the most popular over-the-counter (OTC) derivatives with central counterparties (CCPs). CCPs collect collateral to cover both the current and potential future value of derivative exposures with their counterparties, where the former is known as 'variation margin' (VM) and the latter as 'initial margin' (IM). This is shifting clearing of OTC contracts towards that of exchange-traded derivatives, which have been cleared with CCPs since well before the financial crisis. Second, the same jurisdictions have introduced requirements for financial counterparties to exchange both variation and initial margins on any new OTC derivative trades that are not centrally cleared.

The use of collateral in derivative markets greatly reduces systemic risk. It does so by preventing the spread of potential losses between counterparties through derivative exposures. In contrast, tens of billions of dollars of losses spread from monoline insurance companies to derivative dealers during the financial crisis. This happened as losses incurred by the monolines reduced their credit worthiness, which forced the dealers to make downward credit valuation adjustments (CVAs) to derivatives with positive market value but held with monolines. If the dealers had held more collateral against these exposures, they would not have needed to revalue them as counterparty credit risk declined.[1]

The flipside is that financial institutions can expect routinely to face larger margin calls. With few exceptions, they will have to post variation margin to their financial counterparties whenever the value of their derivatives moves against them. They will also have to top up initial margins exchanged after completing derivative trades if their riskiness should subsequently increase. This is because the expected volatility of derivative positions is a key determinant of IM requirements. In either case, margin calls must be settled in cash or liquid securities, so they would erode unencumbered liquid-asset buffers.[2]

There is a limited range of defensive actions available to financial institutions should margin calls materially erode these liquid-asset buffers. First, they may liquidate some of their derivatives, thereby reducing IM requirements. Second, they may bolster their liquid assets by selling less-liquid securities or entering repo agreements to sell and later repurchase them. Third, they may be able to borrow from central banks, although this option is not available to all types of institution. The first two options involve trading, whether in derivative, security or repo markets, which can move prices against the institution taking the defensive action.

This can amplify the financial cycle. When volatility is low, initial margins are low, so derivative users can take relatively large positions while still maintaining healthy buffers of liquid assets on top of those immediately encumbered by margining. These buffers reduce the likelihood of needing to take defensive actions, which reinforces low volatility. However, when volatility rises, initial margins rise, squeezing unencumbered liquid-asset buffers. Large prices moves, which are

[1] See, for instance, FSA (2010) and ISDA (2011).

[2] For centrally cleared derivatives, variation margin typically must be paid in cash, while initial margin may alternatively be settled in high-quality liquid securities. In practice, market participants use these two collateral types in roughly equal measure (ISDA, 2015). For non-centrally cleared derivatives, international rules allow variation and initial margins to be settled in cash or liquid securities (BCBS-IOSCO, 2015), though in practice the former is usually settled in cash and the latter in securities (ISDA, 2017).

already more probable under elevated levels of volatility, are then more likely to be amplified because the resulting variation margin calls are more likely to lead to defensive actions and further price movements. This amplification of changes in volatility over the financial cycle by margin calls is often referred to as 'procyclicality'.

In the absence of policy intervention, the level of procyclicality resulting from a given margining regime might be too high. While any individual derivative user should recognise that taking a large position relative to its liquid-asset holdings would undermine its profitability if it had to take defensive actions, it would not recognise the effects of these actions on the profitability of other participants in affected markets. Thus, there can be an externality. If so, an omniscient social planner could raise risk-adjusted expected returns (RAERs) across the financial system by choosing smaller positions relative to liquid assets for each derivative user, as this would reduce spillovers to other market participants from price movements due to defensive actions. However, it may not be optimal for the planner to go so far in this direction that it eliminated the possibility of such spillovers and reduced procyclicality to zero.

In this paper we develop a model of a derivative market with an externality that stems from defensive actions and investigate the effectiveness of different policies at mitigating it. We focus on the defensive action of liquidating derivative positions, as this avoids the need to model a security or repo market in addition to a derivative market. For the same reason, the investors in our derivative market take outright positions, rather than positions that hedge a security holding. We do not think our results would be qualitatively affected if the price consequences of defensive actions were recorded in a security or repo market instead of a derivative market. Similarly, we think our focus on speculative rather than hedging positions is qualitatively inconsequential. While the motivation for holding these alternative types of position differs, they would generate the same margin calls (as offsetting changes in the value of a hedged security do not affect margin requirements), which would have the same implications for unencumbered liquid-asset buffers and liquidations.

Equipped with our model, we first investigate the effects of a set of quantity-based policy tools. These add a macroprudential buffer to IM requirements. The idea is to discourage investors from taking as large derivative positions as otherwise by forcing them to hold more low-yielding liquid assets against each unit of these positions. Moreover, the buffer, which could be posted to the same custodial account as the pre-buffer initial margins, could be released in the event of large margin calls, making previously encumbered liquid assets available to help meet these demands. Such a policy could decrease the amplification of price movements by reducing the need to fire-sell derivatives.

We experiment with different ways to set and release such a macroprudential buffer. Our setting policies include a discretionary approach, in which the buffer is reset over the financial cycle based on market conditions prevailing at the time. They also include less information-intensive rules that set the buffer as a simple function of pre-buffer (or 'microprudential') IM requirements, which vary with volatility over the financial cycle. These include functions based on the anti-procylicality (APC) mechanisms in European Market Infrastructure Regulations (EMIR).[3] They also include two new ideas motivated by how an ideal buffer set under the discretionary

[3] See Article 28 of European Union (2013).

approach tends to vary with microprudential IM requirements. These are a constant buffer and a buffer that is a variable percentage of the microprudential IM requirement, with the percentage varying inversely with volatility (*i.e.* a countercyclical buffer). The release policies we consider include releasing the buffers only in response to increases in IM requirements, as with the EMIR mechanisms, or additionally with calls for variation margin.

We show that a discretionary approach could replicate the social optimum, reducing fire-sales and raising the aggregate RAER to levels that a social planner would target. However, this would require policymakers to set the macroprudential buffer perfectly at each point of the financial cycle. This, in turn, would require them to always have full information on market participants' positions as well as the various risks that might affect them. Acknowledging that this may not be the case in practice, we focus on our results for the different rule-based approaches to setting the macroprudential buffer.

We find that the EMIR-based tools can reduce the fire-sale externality at certain points of the financial cycle, but they are ineffective at other points and can occasionally increase the size of the externality by demanding too much initial margin. The countercyclical buffer performs much better and is the best of all our rule-based approaches. It virtually eliminates the externality at certain points of the financial cycle and reduces it substantially at others. The constant buffer performs almost as well, just leaving slightly more of the externality in place at the extremes of the cycle. Due to its ease of implementation, however, it may be most preferable.

We also investigate a price-based policy tool, which is essentially a tax on derivative positions. Such a policy would raise the marginal cost of positions, which could nudge investors towards the position sizes that a social planner would choose. However, this would also generate tax revenues, which would need to be redistributed to market participants in a manner that did not affect these incentives (*e.g.* as lump sums) to achieve the social optimum.

2. Related literature

As outlined, post-crisis reforms have mandated that many derivative transactions are collateralised, increasing the importance of margin requirements for many types of firm. This has brought with it much scrutiny on the extent of procyclicality inherent in the models used to calculate margins. Research in the area has broadly focussed on two questions. Firstly, to what degree procyclicality in margin models can exacerbate stress, for example by placing participants under liquidity pressures? And secondly, whether low levels of margin during periods of low volatility can contribute to the build-up of leverage in the financial system? This paper makes a contribution to the literature in both of these areas.

Although there are relatively few studies that look at macroprudential margin buffers, there are an increasing number examining the theoretical and empirical evidence of procyclicality in margin requirements. In one influential paper, Brunnermeier and Pederson (2009) create a model that shows that a margin spiral can emerge when margins are increasing in a time of market illiquidity. They show that funding liquidity (ease of obtaining funding) and market liquidity (ease of trading assets) are intertwined: when a funding shock hits, market liquidity is lower, leading to higher margins, which tightens funding further. Added to this is a loss spiral, whereby market illiquidity leads to speculator losses, causing asset sales and further price falls. These spirals

reinforce each other meaning that they are larger than the sum of their parts. This forms a theoretical basis for how procyclical margins can be destabilising.

Certain studies have also sought to uncover an empirical link between margin setting and market stress, in an attempt to uncover whether margin setting is procyclical. In a recent study, Lewandowska and Glaser (2017) use ten years of data from a large CCP, but do not confirm that CCP margin setting is procyclical. Their research shows only a low average level of correlation between price volatility and the level of margins or haircuts in the investigated time frame. They say that this places doubt on whether regulatory action would be effective.

In another recent empirical study, Glasserman and Wu (2017) analyse whether margin levels need to be higher "through the cycle" in order to avoid unnecessary procylicality. They use a GARCH framework, which they say offers insights into the heavy-tailed distribution of long-run volatility, which is present even when short-run volatility is low. They find that current procylicality mitigation techniques do not encapsulate this long-run heavy tail, and argue it should be accounted for in future policy analysis as it governs the size of the buffer needed to counter procyclicality. Another empirical study by Abruzzo and Park (2014) shows that margin can rise very quickly following volatility increases, based on empirical evidence at CME and ICE.

As well as impacting liquidity stress, excessive procyclicality may be harmful as it can allow market participants to increase leverage during the upswing of the cycle when margins are low. For example, Geanakoplos (2010) describes a model of a 'leverage cycle', in which good times are categorised by low volatility, rising asset prices and low margins. Low margin allows agents to borrow using only a small amount of collateral, while rising asset prices serve to loosen borrowing constraints further by freeing up more collateral. In this scenario some market participants are constantly increasing leverage by borrowing more, or are increasing synthetic leverage through derivatives. This changes when market participants receive some bad news that increases uncertainty, volatility of asset prices and margin on transactions. Overly leveraged agents then have to sell in order to meet margin increases, causing prices to fall and losses to be realised. Reversing this process once it has started is difficult, so the author claims that the best way to stop a crash is to act long before it occurs by restricting leverage or making borrowers more resilient to shocks. Margin requirements may be one way to achieve the former since they are a key part of the leverage cycle. Other papers provide example of similar dynamics whereby procyclicality in margin requirements and leverage can reinforce each other (see Cont and Schaanning (2017), Shleifer and Vishny (2011), Kiyotaki and Moore (1997)).

Leverage is not only important in that it makes agents more susceptible to sudden margin calls, price movements and painful deleveraging, but it also has a broader implication in the literature on systemic crises. For example, Jordà et al. (2013) use a dataset on advanced economies since 1870 and show that recessions are deeper when accompanied by a period of high credit growth, while it also takes longer to recover. They also show that higher leverage significantly impacts the growth path of other financial variables such as real investment, real lending, government rates and the current account. The results suggest that the link between low margin and leverage should be investigated further, and policymakers should consider tools that could curtail leverage or increase participants' resilience through margin setting.

As mentioned in the introduction, EMIR already has in place three tools which are intended to limit procyclicality of margin requirements. The usefulness of these tools has been assessed by

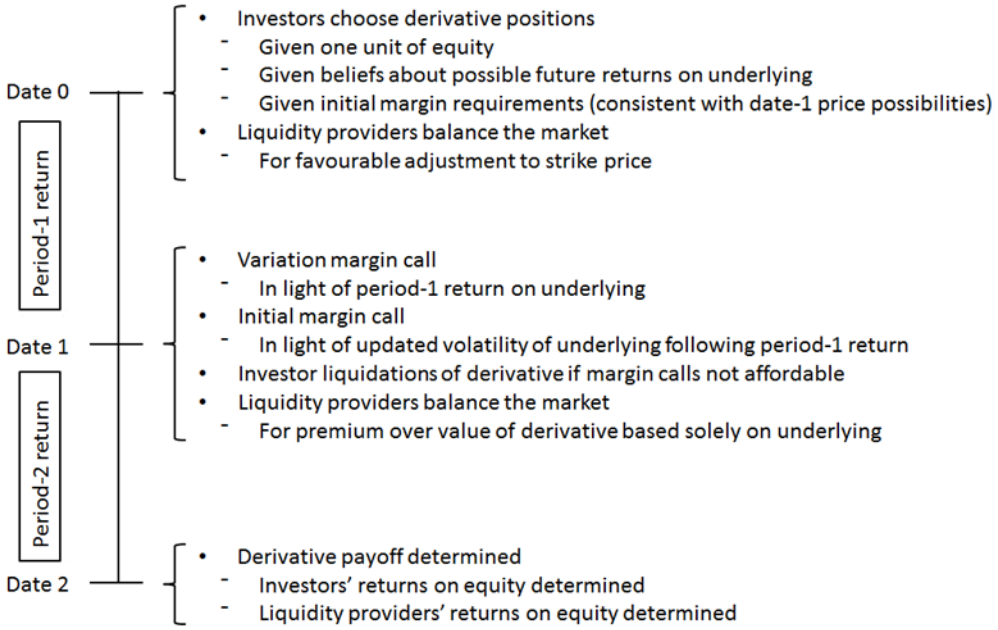
Murphy et al. (2016). They show that all tools are somewhat useful in mitigating procyclicality, but the preferred option depends on the weight you place on preventing sharp changes in margin following stress versus over-margining in benign times. However, the tools are not assessed with regards to their impact on agents' leverage decisions. Our paper explicitly addresses this factor.

In one similar study, Brumm et al. (2015) seek to quantify the impact of margin regulations on aggregate volatility. The authors use "Regulation T", which introduced margin regulations in the United States following the stock market crash of 1929, as a case study. They find that minimum margin requirements are ineffective in moderating aggregate volatility if some asset classes are left unregulated. The regulation is only effective if it is applied to all markets. Their model differs to ours in that it does not identify an externality, or address how EMIR tools or other macroprudential tools can reduce or eliminate the externality.

3. The model

The model focuses on investors who, perhaps as a result of individual research efforts, have different views about the prospects for returns on a particular financial instrument. They seek to profit from these views by trading a derivative that references the instrument. This allows them to establish leveraged positions, equivalent to multiples of their equity, thereby boosting their expected returns. Until the derivative matures, however, these positions need to be serviced. This involves posting collateral to meet calls from counterparties for initial and variation margin. If investors run short of collateral they must liquidate at least some of their positions in the derivative, which has a 'fire-sale' impact on its price. This is the source of the externality that we will later seek to quantify and mitigate. Figure 1 provides a schematic overview of the model.

Figure 1: Model schematic



3.1. Model ingredients

There are three types of agent in the model. These are optimistic investors (OI), pessimistic investors (PI) and liquidity providers (LP). The investors may be thought of as hedge funds, while the liquidity providers may be considered derivative dealers. There are n^{OI}, n^{PI} and n^{LP} of these agent types respectively. Each agent of each type begins with one unit of equity.

In addition, there are two financial instruments. These are cash and a derivative. Cash serves as a store of value, generating a fixed return of zero. As such, it may be used to collateralise positions in the derivative. The derivative is a futures contract. At maturity, it pays off an amount equal to the difference between the prevailing (or 'spot') value (s) of an 'underlying' reference variable, such as a stock price or exchange rate and a 'strike price' (k). Following market convention, the initial price of the derivative is zero. However, in lieu of an initial charge, the strike price adjusts to balance the derivative's supply and demand. This works by affecting its prospective payoff.

Finally, there are two time periods. These might represent, for instance, six months. In each period (t), changes in the underlying (Δs_t) are uncertain, though they have some structure, as summarised by the model in equations 1-6 below. We normalise the initial value of the underlying by setting $s_0 = 1$, so Δs_t can be interpreted as returns on the underlying.[4]

$$\Delta s_1 = \sigma_1 \varepsilon_1 \tag{1}$$

$$\Delta s_2 = \sigma_2 \varepsilon_2 \tag{2}$$

$$\sigma_2^2 = (\omega + \alpha(\Delta s_1)^2 + \beta\sigma_1^2)(1 + \delta\varepsilon_\sigma)^2 \tag{3}$$

$$\omega = (1 - \alpha - \beta)\sigma_{LR}^2 \tag{4}$$

$$\varepsilon_1, \varepsilon_2\ \varepsilon_\sigma \sim iid\ N(0,1) \tag{5}$$

$$\alpha, \beta, \delta, \sigma_{LR} > 0 \tag{6}$$

We take the initial level of underlying return volatility (σ_1) as given. This has a relatively low value in the 'boom' phase of the financial cycle and a relatively high value in the 'bust' phase. Regardless of whether volatility starts high or low, the structure above implies that it has a tendency to revert towards its long-run average value (σ_{LR}) in the second period. In addition, realisation of an extreme return in the first period (Δs_1), whether positive or negative, boosts volatility in the second period (σ_2). This raises the probability of subsequent returns (Δs_2) being extreme. Thus, the model generates both volatility clustering and a fat-tailed distribution of returns over time, as are often observed in in financial markets in practice. Note that if δ were zero our returns structure would reduce to a Generalised Autoregressive Conditional Heteroskedasticity (GARCH) model. However, we set $\delta > 0$ to introduce some randomness to the relationship between Δs_1 and σ_2, as these two variables are not always highly correlated in practice. The final value of the underlying ($s_2 = s_0 + \Delta s_1 + \Delta s_2$), along with the strike price, then determines the derivative's payoff ($\theta_2 = s_2 - k$) when it matures at the end of the second period. However, as an alternative to holding the derivative to maturity, investors may trade it at a market-determined price at the end of the first period.

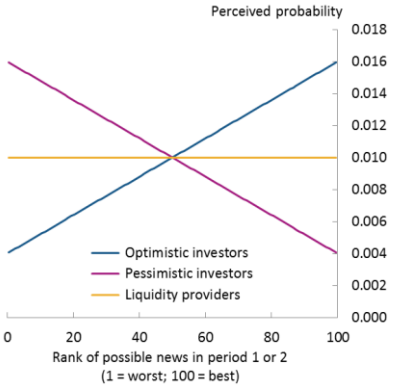
[4] As an alternative to modelling Δs_t, we could have modelled $\Delta \ln(s_t)$. However, this would have generated an asymmetric distribution of returns on the underlying, and exposition of the model is simpler with a symmetric distribution. Moreover, switching to $\Delta \ln(s_t)$ would have had little quantitative effect on our results, given that returns only cumulate over two periods.

3.2. Model timeline

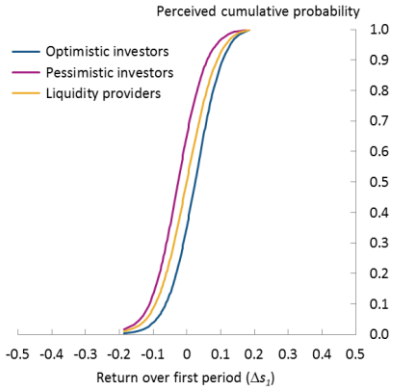
At date 0, at the start of the first period, investors, who are risk averse, establish positions in the derivative (w_0^{OI}, w_0^{PI}), given their unit of equity, which they hold as cash.[5] They do this based on beliefs about the likelihood of good and bad news ($\varepsilon_1, \varepsilon_2$) for returns on the underlying. Optimistic investors perceive higher chances of good news than bad news, while pessimistic investors have the opposite view. The precise beliefs of each type of agent are shown in Figure 2. Reflecting these beliefs, optimistic investors will establish long positions in the derivative ($w_0^{OI} > 0$), while pessimistic investors will establish short positions ($w_0^{PI} < 0$).

Figure 2: Beliefs about returns on the underlying

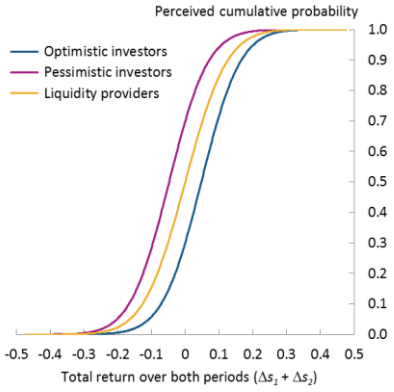
Perceived probabilities of news about returns



Perceived cumulative period-1 return probabilities



Perceived cumulative total return probabilities



The magnitudes of these positions are constrained by IM requirements. Initial margins are collected on new derivative trades to provide some protection against losses that may be incurred in finding replacement contracts should the original counterparty default. Hence, IM requirements are set to cover a high percentile (m_p) of possible losses due to the price of the derivative moving in the time it may take to find replacement contracts. Indeed, regulations specify that m_p must be at least 99%.[6] In our stylised model, with only two periods, investors must post as initial margin m_0^L units of cash for each unit of long position and m_0^S units of cash for each unit of short position, where $Pr(-\Delta p_1 > m_0^L) = m_p$ and $Pr(\Delta p_1 > m_0^S) = m_p$, with Δp_1 denoting the change in the market price of the derivative in period 1.[7] In the absence of liquidations, these valuation changes would be equal to Δs_1. However, as we will see below, liquidations amplify changes in the market value of the derivative beyond those of the underlying. Given that investors have only one unit of cash, their initial investments in the derivative are subject to the following constraints:

[5] Since the only alternative investment, *i.e.* the derivative has an initial price of zero.

[6] For centrally-cleared derivatives, for instance, EU regulations (European Union, 2013) require "for the calculation of initial margins the CCP shall at least respect the following confidence intervals: for OTC derivatives, 99.5%; …". Similarly, 'Margin requirements for non-centrally cleared derivatives' (BCBS-IOSCO, 2015) state that initial margins must be "consistent with a one-tailed 99 per cent confidence interval".

[7] As the date-0 price is zero, the change in price during period 1 is equal to the date-1 price. Also, the probabilities of period-1 price changes used to compute IM requirements are neutral, rather than those perceived by optimistic or pessimistic speculators. This is consistent with IM requirements being set in practice by non-speculators, either CCPs or derivative dealers.

$$\left|w_0^{OI}\right| m_0^L \leq 1 \quad \text{(for long positions, as taken by optimistic investors)} \tag{7}$$

$$\left|w_0^{PI}\right| m_0^S \leq 1 \quad \text{(for short positions, as taken by pessimistic investors)} \tag{8}$$

Thus, higher IM requirements reduce the maximum leverage available to investors. As we will see below, with exception of some asymmetric markets (studied in Section 4.3), these constraints will not bind in equilibrium. In other words, investors will keep some spare cash for possible future margin calls.

Liquidity providers balance the derivative market at date 0. They do this by taking on an equal-sized but opposite position to the aggregate position of investors. As detailed below, some adjustment to the strike price may be necessary to persuade liquidity providers, who are risk averse, to adopt this position. Thus, we set $k = 1 + k_0$, where k_0 is the adjustment needed to balance the market. Liquidity providers are always capable of playing this balancing role, as we assume they have sufficient cash for margin requirements to never bind.[8]

At date 1, derivative holders must service their trades. This means posting additional cash as collateral to meet any variation or initial margin calls. VM calls are determined by the preceding change in the market value of the derivative (Δp_1). Positive values of Δp_1 generate marked-to-market (MTM) profits for long-position holders, prompting them to call for additional collateral from their short counterparts to protect this value. Conversely, negative values of Δp_1 generate MTM profits for short-position holders, prompting them to call for additional collateral from their long counterparts. In either case, date-1 VM calls are equal to $-w_0 \Delta p_1$. At the same time, IM calls are determined by changes in potential losses on the derivative looking ahead to the next period. In particular, with σ_2 known at this time, date-1 IM requirements (m_1^L and m_1^S) are set such that $Pr(-\Delta s_2 > m_1^L) = m_p$ and $Pr\left(\Delta s_2 > m_1^S\right) = m_p$. Note that these IM requirements only reflect possible values of Δs_2, whereas date-0 IM requirements reflect possible values of Δs_1 and Δp_1. This is because the derivative matures at the end of period 2, so, in contrast to period 1, there is no end-of-period trading that could drive the return on the derivative away from that of the underlying. Given the resulting requirements, IM calls at date 1 are equal to $\Delta m_1^L = m_1^L - m_0^L$ for each unit of long position and $\Delta m_1^S = m_1^S - m_0^S$ for each unit of short position.

If combined variation and initial margin calls at date-1 would exhaust an investor's unencumbered cash, its initial derivative position would be too large for it to maintain beyond the first period. This happens if

$$\left|w_0^{OI}\right|(m_0^L + \Delta m_1^L) > 1 + w_0^{OI} \Delta p_1 \quad \text{(for long positions, as taken by optimistic investors)} \tag{9}$$

$$\left|w_0^{PI}\right|\left(m_0^S + \Delta m_1^S\right) > 1 + w_0^{PI} \Delta p_1 \quad \text{(for short positions, as taken by pessimistic investors)} \tag{10}$$

In this case, the maximum supportable position at date 1 (w_1) may be inferred by solving

$$\left|w_1^{OI}\right|(m_0^L + \Delta m_1^L) = 1 + w_0^{OI} \Delta p_1 \quad \text{(for long positions, as taken by optimistic investors)} \tag{11}$$

$$\left|w_1^{PI}\right|\left(m_0^S + \Delta m_1^S\right) = 1 + w_0^{PI} \Delta p_1 \quad \text{(for short positions, as taken by pessimistic investors)} \tag{12}$$

Otherwise, we assume investors retain their date-0 positions at date 1. This means date-1 derivative positions for optimistic and pessimistic investors are given by

[8] For example, they may borrow additional cash. In contrast, we assume this option is not available to investors, who may only obtain leverage synthetically through their derivative positions.

$$w_1^{OI} = w_0^{OI} \qquad \text{if } w_0^{OI} m_1^L < 1 + w_0^{OI} \Delta p_1 \tag{13}$$

$$w_1^{OI} = \frac{1 + w_0^{OI} \Delta p_1}{m_1^L} \qquad \text{otherwise} \tag{14}$$

$$w_1^{PI} = w_0^{PI} \qquad \text{if } -w_0^{PI} m_1^S < 1 + w_0^{PI} \Delta p_1 \tag{15}$$

$$w_1^{PI} = -\frac{1 + w_0^{PI} \Delta p_1}{m_1^S} \qquad \text{otherwise} \tag{16}$$

As at date 0, liquidity providers balance the market at date 1. So, if optimistic or pessimistic investors are forced to liquidate some of their derivative positions at this time, liquidity providers will act as counterparts to these trades. However, they will only do so at an advantageous price from their point of view. The larger the size of positions that liquidity providers are asked to take on, the more favourable pricing they will require.

This means liquidations amplify changes in the market value of the derivative beyond those consistent with returns on the underlying. To see this, first note that optimistic and pessimistic investors never liquidate at the same time following changes in the value of the underlying. This is due to their opposing positions. For instance, while a negative Δs_1 could force optimistic investors to liquidate some of their long positions, there would be no offsetting demand from pessimistic investors wanting to liquidate short positions in such circumstances. Instead, optimistic investors would have to trade with liquidity providers. As liquidity providers would be taking on long positions in this case, they would require a price discount to do so. This would pull p_1 down, reducing Δp_1 below (the already negative) Δs_1.

While investors gaining from a change in the value of the underlying would not offset any pressure on the price of the derivative from those forced to liquidate it, they could potentially add to it. Continuing with the example above, the pessimistic investors, who profit and receive VM payments following the decline in the value of the underling, could use these resources to further express their pessimistic view by increasing their short positions.

However, we assume that these 'winning' investors simply hold their positions constant at date 1. This makes the model much more tractable. In addition, we offer two arguments in support of this assumption. First, in practice, we would expect 'losing' investors with unaffordable margin calls looming to liquidate some of their positions in advance of those calls. This is because failing to meet those calls would trigger default proceedings. Meanwhile, winning investors could not bolster their positions, assuming they would wish to do so, until they had actually received margin payments. In the interim, losing investors would trade with liquidity providers, as in our model. Second, even if these asynchronous effects of margin calls and payments could be eliminated, any extra demand for positions from winning investors would be modest relative to those liquidated by losing investors. This is because the former is driven by maximisation of a utility function, which is concave in position sizes, while the latter is driven by a margin constraint, which is linear in position sizes. Of course, this second reason allows for some increase in position sizes for winning investors, whereas we have assumed none. Hence, depending on the reader's view of the importance of our first argument, the amplification of price moves we report below in our results section may be regarded as a lower bound

Finally, at date 2, the derivative payoff is realised. This is

$$\theta_2 = s_2 - k = s_2 - (s_0 + k_0) = \Delta s_1 + \Delta s_2 - k_0 \quad (17)$$

This profit per unit of derivative held to maturity compares with a profit of Δp_1 for each unit liquidated prematurely. Thus, p_1 only affects profits if agents need to trade some of their initial positions.

It is with a view to these potential profits that agents choose their positions. In particular, they choose positions to maximise risk-adjusted expected profits given their beliefs. As they each begin with one unit of equity, this also amounts to maximising risk-adjusted expected returns (Ω^k). In other words, their 'utility' functions are

$$\Omega_t^k = E_t^k\left(\pi_2^k\right) - \frac{\gamma_k}{2} var_t^k\left(\pi_2^k\right) \qquad k \in \{OI, PI, LP\} \quad (18)$$

where π_2^k denotes portfolio profits ($\pi_2^k = w_1^k\theta_2 + \left(w_0^k - w_1^k\right)\Delta p_1$) and γ_k is a parameter that measures each agent type's aversion to risk. The *k*-superscripts on the expectation and variance operators indicate that these functions operate over different probability distributions for different agent types (as shown in Figure 2).

3.3. Liquidity providers' demand function

Given their utility function and the possibilities for derivative profits, we can now derive the demand function of liquidity providers at date 1. As we assume liquidity providers have ample cash to cover margin requirements, they solve an unconstrained optimisation:

$$\max_{w_1^{LP}} E_1^{LP}\left(w_1^{LP}(\Delta s_1 + \Delta s_2 - k_0) - w_1^{LP}\Delta p_1\right) - \frac{\gamma_{LP}}{2} var_1^{LP}\left(w_1^{LP}(\Delta s_1 + \Delta s_2 - k_0) - w_1^{LP}\Delta p_1\right) \quad (19)$$

At date 1, Δs_1 and k_0 are known, and Δp_1 is also taken as given by liquidity providers as the small size of their individual desired positions has essentially no bearing on aggregate market demand and, hence, the price. As a result, these variables do not contribute to the variance term in equation 19. In addition, the expected value of Δs_2 is zero. The optimisation consequently simplifies and solves to give

$$\rho_1 = \Delta p_1 - \Delta s_1 = -w_1^{LP}\gamma_{LP}\sigma_2^2 - k_0 \quad (20)$$

Thus, if investors require liquidity providers to take on short positions, the price must rise by more than the change in the underlying, and *vice versa*. The magnitude of this premium, ρ_1, (or 'basis') depends on the size of positions taken on by liquidity providers (w_1^{LP}). Since they clear the market, these are

$$w_1^{LP} = -\left(n^{OI}w_1^{OI} - n^{PI}w_1^{PI}\right)/n^{LP} \quad (21)$$

The rate at which the premium changes with the magnitude of positions depends on the liquidity providers' risk aversion (γ_{LP}) and remaining uncertainty about the payoff of the derivative (σ_2^2). Since σ_2 is related to σ_1 through equation 3, this means the price impact of liquidations varies over the financial cycle. This measure of market liquidity is relatively low in the boom (low σ_1) and relatively high in the bust (high σ_1).

Moreover, the premium may rise further if the pessimistic investors liquidating short positions at date 1 had generated a negative adjustment to the strike price at date 0 (k_0) by outnumbering and trading a greater volume of the derivative than optimistic investors. This is

because liquidity providers do not share their pessimism, so the price would have to jump to a level consistent with their beliefs for them to trade. This accounts for the final term in equation 21. This term is zero in a symmetric market, with an equal number of optimistic and pessimistic investors.

In a similar manner, we can derive the liquidity providers' demand function at date-0. They again solve an unconstrained optimisation:

$$\max_{w_0^{LP}} E_0^{LP}\left(w_0^{LP}(\Delta s_1 + \Delta s_2 - k_0)\right) - \frac{\gamma_{LP}}{2} var_0^{LP}\left(w_0^{LP}(\Delta s_1 + \Delta s_2 - k_0)\right) \quad (22)$$

Since, at date 0, the expected values of Δs_1 and Δs_2 are zero and k_0 is taken as given by individual liquidity providers, this simplifies and solves to give:

$$p_0 = w_0^{LP}\gamma_{LP}(\sigma_1^2 + \sigma_2^2) \quad (23)$$

The positions taken on by liquidity providers again balance the market, *i.e.* $w_0^{LP} = -\left(n^{OI}w_0^{OI} - n^{PI}w_0^{PI}\right)/n^{LP}$. So, when investors have net demand for long positions, for instance, liquidity providers will take on short positions. But they will do so only at a strike price above s_0, *i.e.* $k_0 > 0$. This raises their expected profits, compensating them for the risk they are taking on: the higher the risk ($\sigma_1^2 + \sigma_2^2$) or the greater liquidity providers' aversion to it (γ_{LP}), the higher the strike price. Conversely, the strike price falls below s_0 when investors have net demand for short positions and liquidity providers have to be induced to take on long positions.

All that remains is to consider how the initial demands of the investors are determined. These follow from a constrained optimisation:

$$\max_{w_0^k} \Omega_0^k = E_0^k(\pi_2^k) - \frac{\gamma_k}{2} var_0^k(\pi_2^k) \quad (24)$$

$$\text{subject to} \quad \left|w_0^{OI}\right| m_0^L \leq 1 \quad (\text{for } w_0^{OI} > 0) \quad (25)$$

$$\left|w_0^{PI}\right| m_0^S \leq 1 \quad (\text{for } w_0^{PI} < 0) \quad (26)$$

$$\text{where} \quad \pi_2^k = w_1^k\theta_2 + \left(w_0^k - w_1^k\right)\Delta p_1 \quad (27)$$

This says that investors choose their date-0 positions to maximise RAERs given their beliefs, subject to compliance with initial margin requirements at the outset. Profits are determined by the derivative payoff for positions held through to maturity and by the interim price for positions liquidated prior to maturity. Liquidation volumes are still governed by equations 14 and 16.

3.4. Private and social optima

Solving the model outlined above involves each agent maximising its individual RAER, without any coordination. In particular, there is no coordination of position choices to take into account that potential fire-sales resulting from these choices would amplify changes in the price of the derivative during period 1, thereby creating greater uncertainty about returns for all position-holders at date-0, which reduces RAERs. We refer to this no-coordination solution as the 'private optimum'.

In addition, we solve a version of the above model that allows for coordination of position choices. To do this, we replace equation 24 with:

$$\max_{w_0^{OI}, w_0^{PI}, w_0^{LP}} \Omega_0 = \sum_k n^k \left(E_0^k \left(\pi_2^k \right) - \frac{\gamma_k}{2} var_0^k \left(\pi_2^k \right) \right) \tag{28}$$

Maximising the sum of RAERs over all agents means that any increments in position size that would increase an individual agent's RAER but reduce the aggregate RAER (Ω_0) would no longer be optimal. We refer to the resulting solution as the 'social optimum'. This could be attained by an omniscient social planner choosing positions on behalf of all the agents in the model. Comparing RAERs in the private and social optima allow us to quantify the fire-sale externality in our model.

4. Benchmark results

Before we turn to policy measures, we report some benchmark results for our model without policy tools.

4.1. Symmetrical market at a point in time

First, we report results for a symmetrical market ($n^{OI} = n^{PI}$), at an intermediate point in the financial cycle with volatility equal to its long-run average ($\sigma_1 = \sigma_{LR}$). The full set of parameter values underpinning these results is detailed in Table 1. Here, the values of n^{OI} and n^{PI} were set so the investors in our model collectively have approximately the same equity as the global hedge fund industry. Similarly, the value of n^{LP} was set so our liquidity providers have roughly the same equity as global derivative dealers. Liquidity providers are more risk averse than investors. This both seems realistic and implies that welfare in our derivative market, as measured by the aggregate RAER across agents, benefits if investors hold the derivative instead of liquidity providers. The percentile of derivative losses covered by initial margins was set at a very high level, similar to those used in practice in both the cleared and non-cleared market. The remaining parameters, which relate to returns on the underlying, were fitted to historical data on the GBP/USD exchange rate, as detailed in the appendix.

Table 1: Parameter settings

Parameter	Symbol	Value
Length of each period (*years*)	t	0.5
Number of optimistic investors (each with $1 of equity)	n^{OI}	2 trillion
Number of pessimistic investors (each with $1 of equity)	n^{PI}	2 trillion
Number of liquidity providers (each with $1 of equity)	n^{LP}	2 trillion
Risk aversion coefficient for optimistic and pessimistic speculators	γ	1
Risk aversion coefficient for liquidity providers	γ_{LP}	3
Percentile of derivative losses covered by initial margins	m_p	0.995
Initial volatility of returns on underlying (*% p.a.*)	σ_1	10
Long-run volatility of returns on underlying (*% p.a.*)	σ_{LR}	10
Coefficient relating past returns to contemporary variance in equation 3	α	0.10
Coefficient relating past variance to contemporary variance in equation 3	β	0.55
Coefficient adding randomness to returns-variance relationship in equation 3	δ	0.15

Table 2 shows the values of some key variables for this first set of results. Reflecting the symmetry of the market, optimistic and pessimistic investors take equal but opposite positions in the derivative at date 0, equivalent in size to a few times their equity, while liquidity providers do not establish a position at this time. The magnitudes of investor positions are smaller in the social optimum than in the private optimum, and their RAERs are higher. This reflects the negative externality generated by position-taking by private investors.

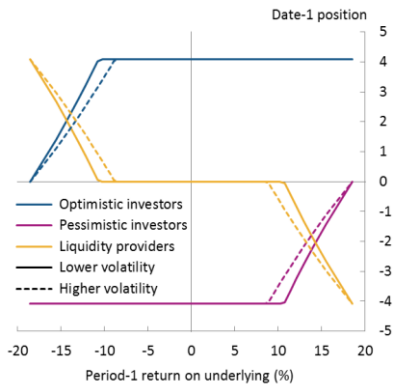
Table 2: Results for symmetrical market at intermediate point in financial cycle

	Social optimum	Private optimum
Date-0 position of optimistic investors (*multiple of equity*)	3.91	4.08
Date-0 position of pessimistic investors (*multiple of equity*)	-3.91	-4.08
Date-0 position of liquidity providers (*multiple of equity*)	0	0
Period-1 volatility of return on derivative (*% p.a.*)	11.5	12.0
Date-0 initial margin on unit long position (*%*)	23.7	24.5
Date-0 initial margin on unit short position (*%*)	23.7	24.5
Date-0 RAER of optimistic investors (*% p.a.*)	11.20	11.08
Date-0 RAER of pessimistic investors (*% p.a.*)	11.20	11.08
Date-0 RAER of liquidity providers (*% p.a.*)	0.26	0.37
Date-0 aggregate RAER (*% p.a.*)	7.55	7.47

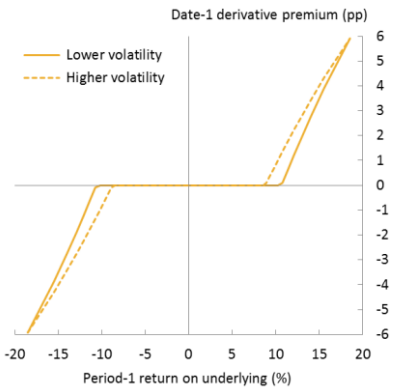
The emergence of this externality can be seen in Figure 3. First, the left-hand panel shows how adverse combinations of changes in the value of the underlying (which lead to VM calls) and its volatility (which lead to IM calls) cause investors to reduce their positions at date 1 compared with date 0. The middle panel then shows how this affects the derivative premium (ρ_1) as liquidity providers take on the positions shed by investors. Finally, the right-hand panel shows how, as a result of this premium, possible period-1 returns on the derivative cover a wider range than those of the underlying. This additional risk undermines RAERs for all market participants.

Figure 3: Basic model results

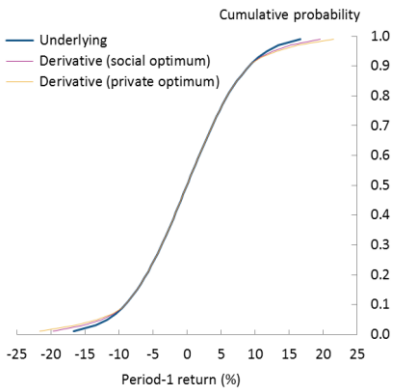
Date-1 positions



Date-1 derivative premium



Period-1 return distribution



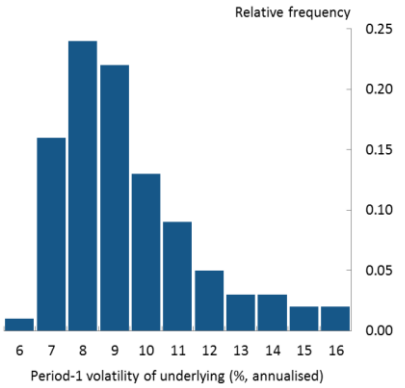
Returning to Table 2, we see how potential liquidations drive the volatility of the derivative above that of the underlying (10%) in the first period. This is even the case in the social optimum, which has some fire-selling, though not as much as in the private optimum. The possibility of fire-selling affects the tails of the derivative return distributions even more than their volatilities. As a result, equilibrium IM requirements, which must cover 99.5% of these 'amplified' returns, are materially higher than the same percentile of returns on the underlying (16.7%) in the social optimum, and they are higher still in the private optimum (Figure 3, right panel).

The final four rows of Table 2 quantify the externality. They show RAERs for each agent type and the aggregate RAER across all the agents in the market. RAERs for investors are a little over 10% per annum. They are much lower for liquidity providers, who do not take positions at date 0 and only establish them at date 1 if investors liquidate, which is a low-probability event in equilibrium. Liquidity providers earn higher RAERs in the private optimum, as fire-selling is more likely, which means they are more likely to hold a position in the second period. Looking across agents, the difference between the aggregate RAER in the social optimum and the private optimum is only 8 basis points or about one-hundredth of the social-optimum level. With $6 trillion of equity in this calibration, it follows that the fire-sale externality costs $4.8 billion a year. In Section 6, we discuss some reasons why this cost might be higher in practice than in our model, including because the extra losses suffered by derivative market participants due to fire-selling might affect their ability to supply onward services to firms and households.

4.2. Symmetrical market through the cycle

In this section, we show how results for the symmetrical market vary with the volatility of returns on the underlying over the financial cycle. We examine eleven different volatility levels, which occur with different frequencies, as shown in Figure 4. The relative frequencies were derived from the same historical data used to estimate the parameters in Table 1.

Figure 4: Volatility over the financial cycle



Changes in volatility affect the fire-sale externality in three ways. First, investors reduce the size of their positions as volatility rises, with position sizes in the private optimum remaining slightly larger than in the social optimum (Figure 5, left panel). This keeps the distribution of possible portfolio returns and, hence, their RAERs relatively constant in both the private and social optimums. Indeed, in the absence of the other two effects, which are relatively minor, they would be perfectly constant. This would hold the externality constant. Intuitively, this first effect determines the average height of the blue shaded area in Figure 5 (right panel).

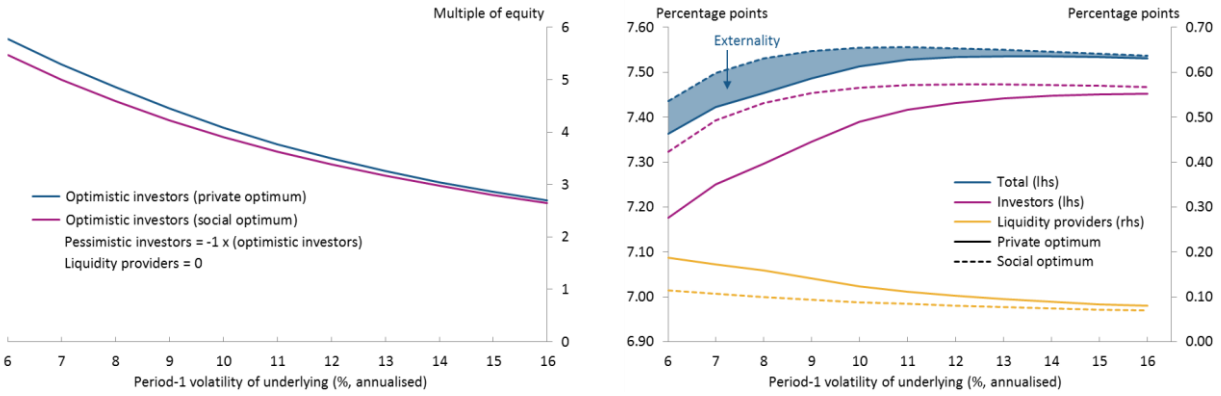
A second effect arises because higher volatility increases the price impact of date-1 liquidations. As shown in equation 20, this is because liquidity providers require more price compensation to establish positions in the derivative when it is riskier.[9] This extra cost of liquidations leads to smaller position sizes, thereby reducing their incidence in both the private and social optimum. As collateral shortfalls become more of a tail event, the probability of

[9] This equation shows that the price impact depends directly on σ_2, but σ_2 is positively related to σ_1 through equation 3.

incurring them in the private and social optimum becomes a more-similar small number. Hence, the externality tends to get smaller with rising volatility. This explains the general narrowing of the blue shaded area from left to right in Figure 5 (right panel).

Figure 5: Effects of changes in volatility over the financial cycle

Date-0 position sizes *Contributions to aggregate RAER*



The third effect arises because higher volatility raises the probability of simultaneous large IM and VM calls. This comes about through the second term in equation 3, and makes the distribution of total margin calls per unit of position more fat-tailed. With different position sizes in the social optimum and the private optimum, this makes the incidence of fire-sales less similar as volatility rises, which widens the externality. This effect is usually dominated by the second effect, though there are exceptions at the lowest volatilities. This is why the blue shaded area is slightly narrower at 6% and 7% volatility than for 8% volatility in Figure 5 (right panel).

4.3. Asymmetric market

Next, we examine how our results are affected by asymmetric investor positions in our derivative market. This is motivated by several historic episodes in which one or more large hedge funds held substantial directional positions in particular derivatives, while leveraged interest on the other side of these markets was relatively modest. During these episodes, large margin calls significantly amplified price movements. These episodes include the failures of Long Term Capital Management in 1998 and Amaranth in 2006, as well as the 'Quant Quake' of 2007.[10]

We vary the degree of asymmetry in our derivative market by changing the balance of optimistic and pessimistic investors, while holding their total number fixed. We define this balance (b) as

$$b = 1 - \frac{n^{OI} - n^{PI}}{n^{OI} + n^{PI}} \tag{29}$$

It ranges from +1 (when all investors are optimistic) to -1 (when all investors are pessimistic), via 0 (when the numbers of optimistic and pessimistic investors are equal).

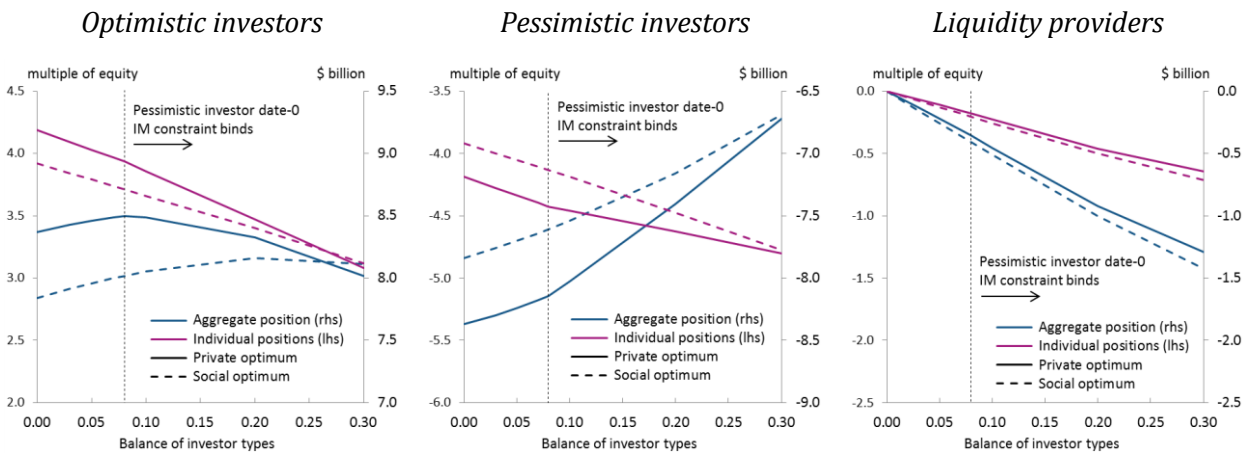
Figures 6 and 7 respectively show equilibrium derivative positions and risk-adjusted expected returns for each type of agent in the model as the balance moves above zero. Results are

[10] See Khandani and Lo (2007) and Mallaby (2010) for descriptions of these events.

symmetrical for balances below zero. These results were generated with the volatility of the underlying fixed at its long-run average level ($\sigma_1 = \sigma_{LR}$).

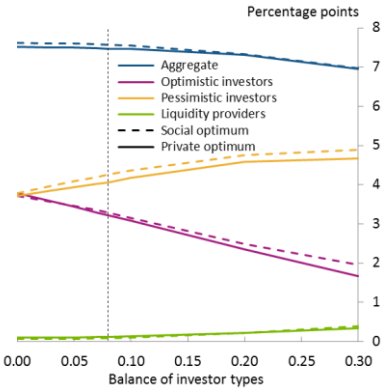
Figure 6 shows that the aggregate position of optimistic investors initially rises (left-hand panel), while the aggregate position of pessimistic investors initially falls in magnitude (centre panel), as the balance of investor types moves above zero. However, the proportionate changes in the magnitudes of these aggregate positions are smaller than the corresponding changes in investor numbers. This is because a majority of optimistic investors pulls up the strike price at date 0 (k_0), which prompts each optimistic investor to reduce its holdings of the derivative and each pessimistic investor to increase its holdings. This is despite these choices skewing the distribution of possible date-1 derivative price changes against pessimistic investors, since their relatively large initial position-to-cash ratios are more likely to lead to subsequent fire-selling. As a result, pessimistic investors face relative high date-0 IM requirements in equilibrium. Although aggregate derivative holdings are less imbalanced than investor numbers, they still do not balance, which means liquidity providers must hold non-zero positions from date-0 in asymmetric markets (right-hand panel).

Figure 6: Equilibrium initial positions in an asymmetric market



As liquidity providers, who have weaker risk appetites than investors, hold non-zero date-0 derivative positions in asymmetric markets, these markets have lower RAERs than balanced markets (Figure 7). Asymmetry also brings a substitution of RAERs from the majority-class of investors to the minority-class, *i.e.* from optimistic to pessimistic investors in Figure 7, which shows the effects of optimists increasingly outnumbering pessimists. This mainly reflects the increase in the strike price, which transfers expected returns between the two types of investor.

Figure 7: RAERs in an asymmetric market



The externality, which was small in symmetric markets, remains essentially unchanged as the balance starts to tilt in favour of one type of investor. At a particular point, however, marked by the dotted vertical line in Figure 7, the minority class of investors would like to establish larger positions than they can

afford in terms of date-0 IM requirements in the private equilibrium. This constrains them to put on relatively small positions, which reduces potential fire-selling in the private equilibrium. This reduces the externality even further.

5. Policy Measures

In this section, we consider policy tools that may nudge private choices of derivative holdings towards those of the social optimum, thereby boosting the aggregate RAER. First, we consider a quantity-based tool, which is a macroprudential buffer added to date-0 IM requirements. Then, we study a price-based tool, which is a tax on investors' date-0 derivative positions.

5.1. Quantity-based policy tools

We introduce a macroprudential buffer to our model by replacing the date-0 margin constraints in equation 7 and 8 with

$$\left|w_0^{OI}\right|(m_0^L + b_0^L) \leq 1 \quad \text{(for optimistic investors)} \tag{30}$$

$$\left|w_0^{PI}\right|\left(m_0^S + b_0^S\right) \leq 1 \quad \text{(for pessimistic investors)} \tag{31}$$

where b_0^L and b_0^S are the macroprudential add-ons for long and short positions respectively. In addition, we replace the date-1 margin constraints in equations 13-16 with

$$w_1^{OI} = w_0^{OI} \qquad \text{if } w_0^{OI}(m_1^L + b_0^L - r_1^L) < 1 + w_0^{OI}\Delta p_1 \tag{32}$$

$$w_1^{OI} = \frac{1+w_0^{OI}\Delta p_1}{m_1^L + b_0^L - r_1^L} \qquad \text{otherwise} \tag{33}$$

$$w_1^{PI} = w_0^{PI} \qquad \text{if } -w_0^{PI}\left(m_1^S + b_0^S - r_1^S\right) < 1 + w_0^{OI}\Delta p_1 \tag{34}$$

$$w_1^{PI} = -\frac{1+w_0^{OI}\Delta p_1}{m_1^S + b_0^S - r_1^S} \qquad \text{otherwise} \tag{35}$$

where r_1^L and r_1^S are the amounts of buffer released on long and short positions respectively at date 1. These release amounts depend on the size and nature of margin calls, as discussed below, and, of course, cannot exceed the initial size of the buffer, *i.e.* $r_1^L \leq b_0^L$ and $r_1^S \leq b_0^S$.

5.1.1. Optimal macroprudential buffer

With full knowledge of the parameters of the model, including the prevailing level of volatility, a policymaker could introduce a buffer policy that would replicate the social optimum. This would require two aspects of the policy to be calibrated correctly.

First, the long and short IM buffers must be set at the right levels. In our baseline parameterisation, for example, we found the social planner would select position sizes of 3.91 for investors, whereas they would choose 4.08 if the decision was their own. In other words, they would choose to hold 0.245 (=1/4.08) units of cash per unit of position, whereas this would ideally be 0.256 (=1/3.91) units. A policymaker could force investors to adopt this socially optimal ratio by boosting IM requirements from 23.7% to 25.6% by imposing a 1.9 percentage point macroprudential buffer.

Second, the macroprudential buffer must always be released to help meet margin calls that would lead to fire-selling. While the optimal buffer setting policy forces investors to hold ideal amounts of cash relative to the size of their derivative positions, it also encumbers more cash. To ensure fire-selling is reduced to social-optimum levels, this extra cash needs to be made available to help meet significant margins calls as if it were an unencumbered resource. This needs to happen regardless of whether the calls are for additional initial margin or variation margin.

In contrast a buffer released only with IM calls would not recover the social optimum. Indeed, in our baseline parameterisation we find that such a buffer would be harmful to welfare (Table 3, final column). This arises because investors would reduce the size of their derivative positions relative to their cash holdings by even more than the buffer forced upon them, since they would want some unencumbered cash to help meet VM calls. That is, the buffer would require them to hold 25.6% as liquid assets, but they would choose to hold 25.8%. As this would mean earning a yield of zero on a higher proportion of their portfolios, investors would be prepared to accept a greater risk of needing to fire-sell the derivative, rather than reduce positions even further relative to cash holdings. This would raise period-1 volatility compared with the social optimum. It would also drive date-0 IM requirements higher. These would not increase proportionately as much as volatility because the most intense fire-selling, which generates the extreme price changes that determine margin requirements, occurs when there are both VM and IM calls, and this buffer policy would still help with the latter. Nevertheless, the aggregate RAER would fall quite significantly from 7.55% to 7.33%. This is because investors would end up holding fewer return-generating derivative positions that would also be subject to greater volatility.

Table 3: Effects of release condition on optimally sized macroprudential buffer

% of social optimum	No buffer	Buffer released with …	
		IM & VM calls	IM calls only
Date-0 position of investors (*multiple of equity*)	104.5	100.0	99.4
Period-1 volatility of return on derivative (*% p.a.*)	103.9	100.0	104.4
Date-0 initial margin on unit position (*%*)	103.2	100.0	100.7
Date-0 RAER of investors (*% p.a.*)	99.0	100.0	96.0
Date-0 aggregate RAER (*% p.a.*)	99.5	100.0	97.1

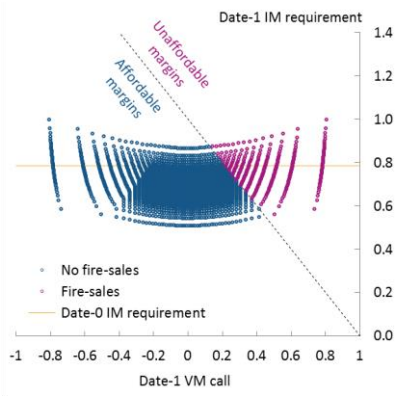
The effects of a macroprudential buffer under different release conditions can also be seen in Figure 8. In this figure, each panel shows 10,000 pairs of simulated VM calls and IM requirements at date 1 for optimistic investors, generated under our baseline parameter settings. Equivalent panels for pessimistic investors would be symmetrical in the x-axes. The potential margin demands shown in these panels depend on the magnitude of derivative positions chosen by the investors at date 0. The black dotted lines then separate margin demands that the investors could meet from their cash holdings (points to the southwest of the lines) from those that could not be met and therefore would lead to fire-selling (points to the northeast). These latter points stretch out to the right because fire-selling amplifies price movements and, hence, VM calls. Equivalent

points on the left are also stretched because of the same effects on pessimistic investors. This gives the U-shaped envelopes of potential margin demands curved, rather than straight, sides.

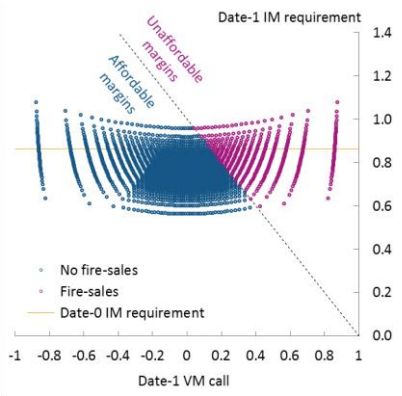
The top row of the figure helps explain why the social optimum is superior to the private optimum. In the social optimum, smaller position sizes mean the simulated margin demands fall inside a relatively small envelope. This leaves fewer instances of margin demands beyond the affordability boundary, so fire-selling occurs less frequently in the social optimum (purple points). Moreover, each purple point in the social optimum has a corresponding one in the private optimum for which the combined IM and VM margin demand, and consequent fire-selling, is stronger.[11]

Figure 8: Effects of release condition on optimally sized macroprudential buffer

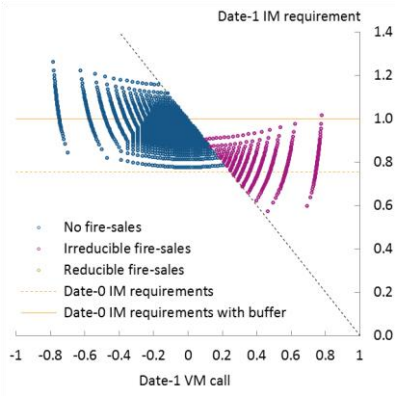
Social optimum



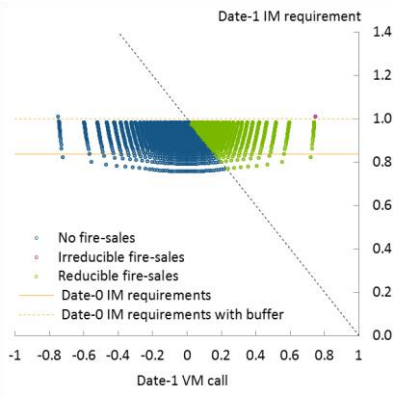
Private optimum



Released with IM or VM



Released only with IM



The bottom row of the figure compares policies of releasing an optimally sized macroprudential buffer with VM or IM calls or only with IM calls. In both panels, the increase in the height of the orange line to 1 with the addition of the buffer to date-0 IM requirements confirms that the resulting requirements form a binding constraint, with no cash left unencumbered. This forces investors to reduce the size of their positions. In the left-hand panel, the buffer is then either left in place at date 1 if margin demands are affordable (blue points), or it is released as much as necessary up to the full-size of the buffer to help meet otherwise unaffordable combined VM and

[11] Except if margin demands are so large that investors must liquidate their entire date-0 position in both the social and private optimums. This is extremely rare in our baseline calibration.

IM calls (purple points). This leaves no instances of fire-selling that could have been reduced in scale by releasing more of the buffer. In contrast, in the right-hand panel, releasing the buffer only to meet IM calls leaves any VM component of margin demands requiring finance. With cash endowments fully allocated to date-0 initial margins (including the buffer), this requires IM requirements to fall at date-1, releasing sufficient cash to finance the VM calls. Otherwise, fire-selling will be necessary. This happens often, and in almost every instance the scale of fire-selling could have been reduced if more of the buffer had been released (green points). This is despite further reductions in position sizes relative to initial cash holdings by investors in anticipation of this effect. Only rarely is the IM call large enough to fully release the buffer, meaning that fire-selling could not have been further reduced (purple points). Releasing the buffer with IM calls when the overall margin demand is still affordable (blue points) is inconsequential, but explains why the envelope in this final panel has a flat top.

Having established the optimal setting and release policies for a macroprudential buffer, we acknowledge that these would be very difficult to implement in practice. In particular, without full knowledge of the parameters in Table 1, a policymaker could not optimally set a macroprudential buffer at a given point in the financial cycle on a discretionary basis. Hence, we next consider alternative rules-based approaches to setting a macroprudential buffer. To evaluate these different settings in the best possible light, we assume the buffers may be released with IM or VM calls.

5.1.2. EMIR-based anti-procyclicality tools

First, we consider three 'anti-procyclicality' (APC) tools based on European Market Infrastructure Regulation (EMIR). As stated in Article 28 of EMIR, the aim of its APC tools is to avoid disruptive changes in initial margin requirements for market participants. To that end, CCPs must adopt at least one of the three available tools when calculating IM requirements. These are:

a) *Stress-weight tool*. Assign at least 25% weight to an IM requirement that reflects stressed observations and the remaining weight to the current IM requirement.
b) *Floor tool*. A fixed floor, which IM requirements may not fall below.
c) *Buffer tool*. Apply a margin buffer of at least 25% on top of the current IM requirement, which can be temporarily exhausted in periods when those requirements are rising significantly.

We implement these APC tools in our model through the following IM buffers:

Stress-weight tool $$b_0^s = w^s m^s + (1 - w^s) m_0^{soc} - m_0^{soc} \quad (35)$$

Floor tool $$b_0^f = \max\left(m^f, m_0^{soc}\right) - m_0^{soc} \quad (36)$$

Buffer tool $$b_0^b = \max\left(\min\left(m^b, (1 + p^b) m_0^{soc}\right)\right) - m_0^{soc} \quad (37)$$

where m_0^{soc} is the pre-tool IM requirement, for which we use the social optimum requirement. This helps us to see the APC tools in the best possible light, since if they happened to set a buffer on top of this requirement at the right level, it would reproduce the social optimum. In addition, in equation 35, w^s is the weight in the stress-weight tool given to stressed IM requirements, which we set at 0.25, and m^s is the value of the stressed requirements themselves, which we set at the 99.99th percentile of possible losses per unit position in period 1 over all volatility levels. Then, in

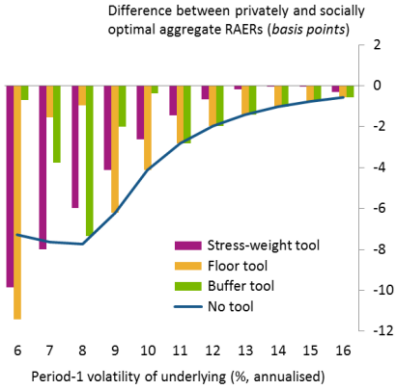
equation 36, m^f is the IM floor, which we set at 0.21 for each unit of position. Finally, in equation 37, p^b is a proportionate add-on to microprudential margins, which we set at 0.25. This add-on is reduced, potentially until it is eliminated, if the resulting IM requirements should rise above a threshold, m^b, which we set at 0.24 of each unit of position.

Figure 9 shows the size of the externality under these three tools at different levels of volatility over the financial cycle. As a benchmark, the blue line shows the externality with no policy tools in operation. Each of these tools performs well at some levels of volatility, but none reduces the externality across the financial cycle.

Starting with the stress-weight tool (purple bars), as calibrated, this adds too large a buffer at low volatilities. These induce investors to cut positions even beyond those of the social optimum. For higher volatilities, however, the stress-weight tool adds a less significant buffer, which induces investors to cut positions only towards those of the social optimum, bringing the aggregate RAER towards the maximum achievable.

Secondly, the floor tool (orange bars), again as calibrated, also adds too high a buffer at the lowest level of volatility. At 8% volatility, it then adds a buffer close to the value that induces the social optimum. At higher volatilities, however, the floor does not bind, so the outcome is the same as in the no-tool equilibrium.

Figure 9: Externality under EMIR-based APC tools



Finally, the buffer tool (green bars) has a more complicated profile. As calibrated, the 25% buffer added to initial margins when volatility is at its lowest (6%) is close to the optimal add-on. Then, at 8% volatility, a 25% add-on is larger than the optimal buffer. At 10% volatility, with the release threshold now crossed, a buffer of less than 25% is added, and this again happens to be close to the optimum. Then, at higher levels of volatility, the tool does not bind (*i.e.* the buffer is fully released) and IM requirements and aggregate RAERs are again those of the no-tool equilibrium.

Figure 10: Risk-adjusted expected returns under the EMIR-based anti-procyclicality tools

Stress-weight tool *Floor tool* *Buffer tool*

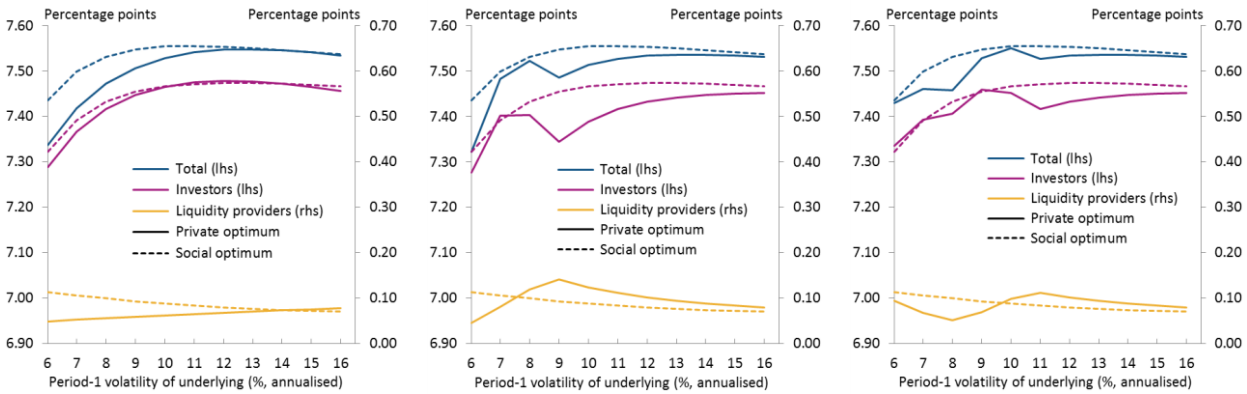
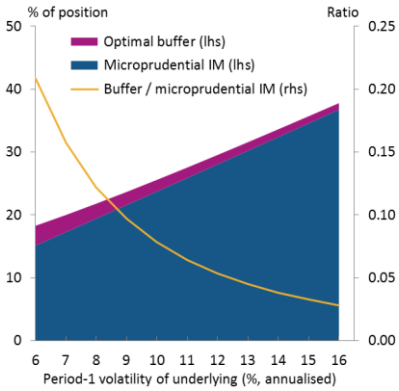


Figure 10 shows how the RAERs of individual agents are affected by the EMIR-based APC tools. Essentially, it shows that whenever the add-ons are too high (*e.g.* the floor tool at 6% volatility, the buffer tool at 8% volatility and the stress-weight tool at some of the lower volatilities), investors are induced to hold small derivative positions that generate even less fire-selling than in the social optimum. This reduces the role for liquidity providers, which reduces their RAERs. It also reduces the RAERs of investors, despite lowering the risk per unit of position due to reduced fire-selling, as expected returns also fall with position sizes, which are reduced too far.

5.1.3. Other rule-based buffers

We now assess two alternative buffer policies. These are motivated by Figure 11, which shows how an optimal macroprudential buffer would be set at each point in the financial cycle by a policymaker acting at its discretion with full information.

Figure 11: Optimal buffer

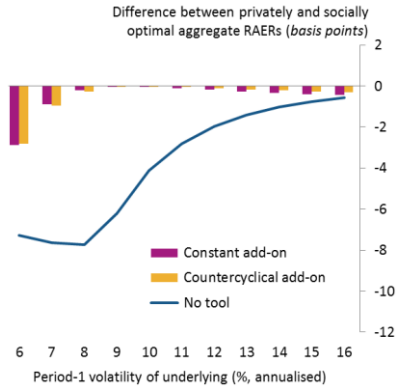


The size of the buffer appears to (i) be close to a constant amount (purple bar) and (ii) vary inversely with volatility when expressed as a proportion of the microprudential IM requirement (orange line). Hence, we study these two additional buffer policies:

$$\textit{Constant buffer} \qquad b_0^c = k_1 \quad \text{where } k_1 = 0.025 \tag{38}$$

$$\textit{Countercyclical buffer} \qquad b_0^y = {}^{k_2}/_{\sigma_1}\, m_0^{soc} \quad \text{where } k_2 = 0.02 \tag{39}$$

Figure 12: Externality under alternative buffer policies



The impact of these policies is shown in Figure 12. In contrast to the EMIR-based APC tools, they both reduce the externality compared with the no-tool equilibrium at all volatility levels. Moreover, they virtually eliminate it at several intermediate levels, which are the ones that occur most frequently during the financial cycle.

The countercyclical buffer slightly outperforms the constant buffer at extreme volatilities, where its ability to vary pays off. This allows it to get closer to the optimal buffer shown in Figure 11, which increases a little when volatility is low and contracts when it is high. However, these extreme volatility levels occur relatively less frequently. As a result, the additional complexity of a countercyclical buffer brings little gain on average over the financial cycle compared with the simpler alternative of a constant add-on.

5.2. Price-based policy tools

A textbook solution to externalities is a tax. Here, we introduce a tax on investors' positions as an incentive to reduce the size of these positions, which would reduce the fire-sales externality. In

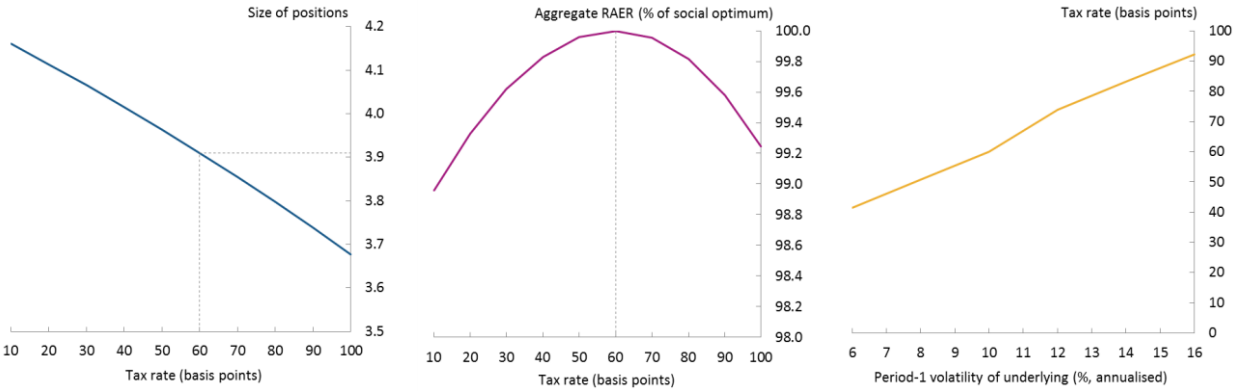
particular, we introduce a tax that is equal to a certain proportion (τ) of investor's initial positions. That is, we supplement equation 27, so it becomes

$$\pi_2^k = w_1^k \theta_2 + \left(w_0^k - w_1^k\right) p_1 - w_0^k \tau \quad \text{for } k \in \{OI, PI\} \tag{40}$$

Figure 13 illustrates how the tax rate can be set to reproduce the social optimum. This is based on our symmetrical market calibration described in Section 4 above. The left panel shows how raising the tax rate from zero causes investors to cut the size of their initial positions. At first, this raises the aggregate RAER (centre panel), as smaller positions reduce the likelihood of fire-sales, and this reduces the volatility of investors' derivative holdings. At higher tax rates, however, these holdings are reduced more materially, which dents expected returns by more than the reduced risk can compensate. When this first happens, the aggregate RAER starts to fall. In this calibration, the aggregate RAER is maximised with a tax rate of 60 basis points. This is the tax rate that reproduces the social optimum. Finally, the right panel shows that this optimal tax rate varies with the level of volatility over the financial cycle.

Figure 13: Optimal tax rates

Position size vs. tax rate *Aggregate RAER vs. tax rate* *Optimal tax rate vs. volatility*



Thus, one implementation challenge to a macroprudential derivatives tax is that it would have to be reset at different stages of the financial cycle. Moreover, to do this, the policymaker would need the rich information set summarised in Table 1. While this was also the case for the optimal macroprudential IM buffer, it turned out that a constant buffer was a good proxy for this first-best solution (Figure 11). In contrast, a constant tax rate would significantly misrepresent the optimal rate at some points in the financial cycle (Figure 13, right-hand panel).

A second implementation challenge is that the amount of tax revenue collected would be very large relative to the size of the externality. In our symmetrical market calibration, it would be about 30 times larger. In our model, these revenues would have to be returned to market participants, as they would otherwise represent a deadweight loss. Moreover, they would have to be returned in a manner that did not undo the incentive effects of the proportional tax, for instance as lump sums. In practice, this problem would diminish to the extent that a proportional tax on derivative positions could contribute to the collection of total desired tax revenues from market participants.

6. Conclusion and discussion

In the previous sections we have identified a fire-sale externality in the derivatives market and shown, in theory, how both quantity and price-based policy interventions could eliminate it, thus replicating the first-best outcome. The quantity-based approach adds a buffer to IM requirements that would be released in the event of significant margin calls, regardless of whether these related to initial or variation margins. The price-based policy is to apply a proportional tax to derivative positions. To replicate the first-best outcome, calibration of both of these policies would need to vary with changes in volatility over the financial cycle, although not by much in the case of IM buffers. Implementation of either of these policies in practice, however, would encounter some significant challenges.

First, the information requirements for successful policy interventions are very high. To infer the benefits of policy interventions in our model we needed information on all market participants' derivative positions and liquid-asset holdings as well as a known distribution of potential market shocks. In reality, further information would also be required on credit lines or securities that could be used to raise funds in the repo market if either of these could help investors to meet margin calls and thus avoid the need to liquidate derivative positions. Moreover, updates of all this information would be required over the financial cycle, so the scale of policy intervention could be recalibrated.

This begs the question of whether a discretionary or rules-based approach to policy calibration would be superior. In theory, a discretionary approach that adjusted policy settings on the basis of all information at each point in time would be better, as it could maintain the first-best outcome. In practice, however, this appears especially challenging to implement. Corresponding to the global nature of the derivatives market, a group of international policymakers would need to agree on policy settings over the financial cycle, and it may be impractical for such a group to do this as frequently as the changing data may require.

These data challenges should be evaluated against the magnitude of the externality. This is because errors in calibrating policy interventions due to imperfect information would erode their effectiveness. Our modelling work finds only a small externality, which suggests a significant risk of counterproductive policy given the imperfect information available to policymakers. Intuitively, our externality is small because it reflects the *ex ante* cost of fire-sales, and this is small because fire-sales do not happen very often and, when they do, the ideal response is only reduce – rather than eliminate – them. Of course, the externality may be larger in reality than in our model. For instance, our simplifying assumption that liquidity providers have unlimited access to cash at zero marginal cost means the price impact of fire-sales could in fact be higher. In addition, fire-sale losses incurred by derivatives users could affect their ability to supply other services to firms and households that we have not captured in the model.

There are also practical implementation challenges specific to our IM buffer policy. These relate to the need to release the buffer with calls for variation margin as well as initial margin. In contrast, the debate on macroprudential margins to date has focussed on buffers that are released only with calls for initial margin. The latter would be relatively straightforward to implement as released buffers could be netted against IM calls, resulting in smaller IM calls for settlement. The former would be more complicated as, in place of the 'losing' counterparty paying variation margin

to the 'winning' counterparty, the losing counterparty's IM custodian would have to release some of the buffer to the winning counterparty, possibly transforming it from securities to cash (*e.g.* via a repo) if initial and variation margins have different settlement criteria.

Moreover, it may be more difficult in practice than in our model to identify when aggregate VM calls would lead to significant fire-sales, which should be mitigated by releasing the buffer. This is because we only modelled two types of investor and all investors of a given type faced significant margin calls together. In practice, however, investors have diversified derivatives portfolios, which may or may not overlap. Conceptually, policymakers would need to infer the aggregate VM calls these portfolios would generate if certain market shocks occurred, and stand ready to release IM buffers should any of the shocks associated with large aggregate VM calls crystallise.

Finally, the proportional tax rate also has a specific implementation challenge, given the need to redistribute the revenues collected. We suggested in Section 5.2 that this may be less of a problem in practice than in the model, as the derivatives tax could substitute for other corporate taxes and the distribution of tax revenues left unaffected. However, since the derivatives tax would need to be set internationally, while other taxes are national, this would generate a cumbersome balancing act for national governments.

References

Nicole Abruzzo and Yang-Ho Park (2014), “An Empirical Analysis of Futures Margin Changes: Determinants and Policy Implications”, *Board of Governors of the Federal Reserve System Discussion Paper*, No. 2014-86

BCBS-IOSCO (2015), “Margin requirements for non-centrally cleared derivatives”

Johannes Brumm, Michael Grill, Felix Kubler and Karl Schmedders (2015), "Margin regulation and volatility", *Journal of Monetary Economics, Elsevier*, Vol. 75(C), pp. 54-68

Markus Brunnermeier and Lasse Pedersen (2009), "Market Liquidity and Funding Liquidity", *Review of Financial Studies*, Vol. 22, pp. 2201-2238

Rama Cont and Eric Schaanning (2017), “Fire sales, indirect contagion and systemic stress testing”, *Norges Bank Working Paper*, No. 2/2017

ESRB (2017), “The macroprudential use of margins and haircuts”

European Union (2013), “Regulation EU No. 648/2012 on OTC derivatives, central counterparties and trade repositories”

Financial Services Authority (2010), “The prudential regime for trading activities: A fundamental review”, *Discussion Paper*

Paul Glasserman and Qi Wu (2017), “Persistence and Procyclicality in Margin Requirements”, *The Office of Financial Research (OFR) Working Paper Series*, No. 17-01

Jean Geanakoplos (2010), “The leverage cycle”, *NBER Macroeconomics Annual 2009*, Vol. 24, pp. 1-65, University of Chicago Press

International Swaps and Derivatives Association (2011), “Counterparty Credit Risk Management in the US OTC Derivatives Market, Part II”

International Swaps and Derivatives Association (2015), *ISDA Margin Survey 2015*

International Swaps and Derivatives Association (2017), *ISDA Margin Survey 2017*

Òscar Jordà, Moritz Schularick and Alan M. Taylor (2013), "When Credit Bites Back," *Journal of Money, Credit and Banking*, Blackwell Publishing, Vol. 45(s2), pp. 3-28, December

Nobuhiro Kiyotaki and John Moore (1997), "Credit Cycles", *Journal of Political Economy*, University of Chicago Press, Vol. 105(2), pp. 211-248

Olga Lewandowska and Florian Glaser (2017), “The recent crises and central counterparty risk practices in the light of procyclicality: empirical evidence”, *Journal of Financial Market Infrastructures*, Vol. 5(3), 1–24

David Murphy, Michalis Vasios and Nicholas Vause (2016), “A comparative analysis of tools to limit the procyclicality of initial margin requirements”, *Bank of England Staff Working Paper*, No. 597

Andrei Shleifer and Robert Vishny (2011), "Fire Sales in Finance and Macroeconomics", *Journal of Economic Perspectives*, American Economic Association, Vol. 25(1), pp. 29-48

Appendix: calibration of GARCH model parameters

In this appendix we explain how we derive values for the parameters in equations 1-6, which govern the possible changes in the level and volatility of the derivative's underlying. Our chosen underlying for this calibration is the GBP/USD exchange rate.

Our first step is to estimate the GARCH model subsumed in equations 1-6. As we assumed elsewhere in our calibration that each holding period for the derivative lasts for six months, it would be natural to estimate this model using data on six-monthly returns. However, we do not have sufficient data at this frequency to make robust parameter estimates. Hence, we instead estimate the model using weekly returns and adjust the parameter estimates to the desired frequency as necessary. Thus, we estimate the model in equations A1-A2 below using Wednesday-to-Wednesday returns between start-1980 and end-2017 as reported by Bloomberg.

$$\Delta s_t = \sigma_t \varepsilon_t \tag{A1}$$

$$\sigma_t^2 = (\omega_w + \alpha_w(\Delta s_{t-1})^2 + \beta_w \sigma_{t-1}^2) \tag{A2}$$

The *w* subscripts in equation A2 denote that the parameters correspond to weekly data.

As well as estimates of the three parameter values in equation A2, this model fitting delivers a time series of conditional volatilities, σ_t. A histogram showing the relative frequency of these volatilities leads to Figure 4, which is another key aspect of our calibration. As some of the volatilities in this histogram have almost no chance of occurring, we discard the volatilities with the lowest probabilities of occurrence and rescale the remaining probabilities until all remaining volatilities have occurrence probabilities of at least 1%. This results in Figure 4.

Our estimates of the three parameters in equation A2 lead to an estimate of the long-run volatility, σ_{LR}. This is given by

$$\sigma_{LR}^2 = \frac{\omega_w}{1 - \alpha_w - \beta_w} \tag{A3}$$

Plugging in our estimates of α_w (0.08), β_w (0.90) and ω_w (4.3x10^{-6}), we find that $\sigma_{LR} = 1.4\%$, which is close to 10% on an annualised basis.[12]

Next, we map our estimates of α_w and β_w to α and β. The former is straightforward, as α appears invariant to the frequency of our data. Certainly, re-estimating the GARCH model using daily (*i.e.* higher frequency) and monthly (*i.e.* lower frequency) data generates similar estimates of α_w. Hence, we set $\alpha = \alpha_w$. In contrast, β_w requires some adaptation, which we base on the expected speed of adjustment of volatility towards its long-run average. This is ϑ_w in equations A4 and A5, which we derive by substituting equation A3 into equation A2 and taking expectations.

$$\Delta\sigma_t^2 = \vartheta_w(\sigma_{LR}^2 - \sigma_{t-1}^2) \tag{A4}$$

$$\vartheta_w = (1 - \alpha_w - \beta_w) \tag{A5}$$

Inputting our estimates of α_w and β_w into equation A5, we find $\vartheta_w = 0.02$, which means we can expect 2% of any gap between volatility and its long-run average to close each week. Using the left-hand equality in the formula below, we find this corresponds to 42% of any gap being closed over six months, *i.e.* $\vartheta = 0.42$.

[12] We annualise volatility using the square-root-of-time rule, which gives $1.4 * \sqrt{52} \approx 10$.

$$\vartheta = 1 - (1 - \vartheta_w)^{26} = 1 - \alpha - \beta \tag{A6}$$

Then, using the right-hand equality in equation A6, we find $\beta = 0.50$.

Finally, to calibrate the 'noise' parameter (δ) in equation 3, which makes volatility uncertain even with known returns, we compare the volatility estimates from our GARCH model with an alternative set of estimates. These alternative volatility estimates (φ_t) come from exponentially weighted moving average (EWMA) model, as described below.

$$\varphi_t^2 = \lambda\varphi_{t-1}^2 + (1 - \lambda)(\Delta s_t)^2 \tag{A7}$$

We set $\lambda = 99.2\%$, which is the same value used by some major clearing services. We then compute the proportional difference between the two volatilities ($\delta\epsilon_t$), as in the equation below, and find the normal distribution that best fits this variable.

$$\sigma_t = (1 + \delta\epsilon_t)\varphi_t \quad \text{where } \delta\epsilon_t = \frac{\sigma_t - \varphi_t}{\varphi_t} \tag{A8}$$

The standard deviation of this distribution is δ, and our estimate of it is 0.11.