Optimisation Solutions in Initial Margin for Non-cleared OTC Derivatives

Donal Gallagher, Roland Lichters, Sharyn O'Halloran, Roland Stamm

August 4, 2016

1 Executive Summary

The regulatory response to the financial Crisis of 2008-2010 has seen the introduction of a broad swathe of new regulation designed to both increase the quality and quantity of bank capital and simultaneously reduce the amount of aggregate risk in the financial system, in particular the amount of counterparty credit risk. Counterparty credit risk, most publicly visible in the proliferation of uncollateralised derivative transactions relating to mortgage lending, was seen to be the proximate cause of the mass bail-outs and taxpayer funded rescues of banks in both the US and Europe. A well founded desire to minimise the probability of a recurrence of this mode of failure in the banking system has given rise to a wave of change and tightening in regulatory requirements including the introduction of Basel III and CCAR in Europe and the US respectively, new requirements in leverage ratios and net stable funding ratios, the requirement to clear OTC derivatives where possible and most recently the requirement to hold initial margin against uncleared derivatives. This has been accompanied by a tightening of the standards applied to internal models (e.g. SR 11-7) and a change in the standard rules approach (e.g. SACCR and FRTB). In addition to the regulatory change, the financial crisis has radically altered the canonical models for pricing derivatives in the presence of risk. The assumption of a single risk free discount curve has been replaced by a new spectrum of discounting and forwarding curves as well as the introduction of risky valuation adjustments including CVA (Credit Valuation Adjustment), MVA (Margin Valuation Adjustment) and others, collectively known as XVA.
These measures have simultaneously forced banks to invest in new programs of work to respond to these regulatory requirements and pricing innovations but of course also raise the question of how a bank should alter its operating model to respond to tightening capital, margin and leverage requirements while tak-

ing advantage of the new opportunities presented by the change in the banking landscape.
In particular, from a technical perspective, one could view the problem of banking as maximising the return on shareholder capital subject to the many new constraints on capital, margin, leverage and funding, to name but a few.
There are two large classes of optimisation problems that are highlighted by this new landscape, the first being the problem of optimising a bank balance sheet in full knowledge of the details of all positions of the bank and subject to all the regulatory constraints of the institution, which are of course a function of regulatory regime. The second and newer type of problem is that of collective optimisation, which can only be tackled with knowledge of information of several institutions' positions (suitably anonymised). It is this newer form of optimisation that is the subject of this paper and for which we present some approaches and examples.

2 Initial Margin and BCBS-IOSCO

An example of a collective optimisation problem which arises as a result of the new regulatory regime is the requirement for the posting of bi-lateral initial margin for non-cleared OTC derivatives.
One of the responses by the regulatory authorities, in order to reduce the aggregate amount of counterparty credit risk, was to introduce mandatory clearing for over the counter (OTC) derivatives. Clearing via a central counterparty (CCP) offers the benefit of both transparency and enhanced collateralisation through the joint standards of *variation* and *initial margin*. Variation margin captures the value of the derivative portfolio today, i.e. current exposure, and is exchanged both ways depending on the value of the portfolio. Initial margin is only posted by the CCP's clients, and is an additional safety net that protects against further adverse moves of the portfolio's value over a 10 day horizon, designed to give an ample buffer for re-hedging in the event of default of one of the CCP's clients. Initial margin is based on the position's market risk and has to be posted regardless of the position's value. It is thus a cost that cannot be evaded.
In practice, not all derivatives can be standardised and made subject to central clearing and so – in part to ensure consistency of treatment between derivative subject to clearing and those that are not subject to clearing – the Basel Banking Committee on Banking Supervision (BCBS) and the International Organization of Securities Commissions (IOSCO) jointly drafted a standard, released as BCBS 261 (insert link). This regulation covers the posting of bi-lateral initial margin and covers the scope and timeframe for the roll-out of the regulation starting with the 17 largest financial institions in September 2016 and ultimately covering most banks with meaningful derivative exposures by 2021.
In the absence of a central counterparty, both parties are obliged to post initial

margin to each other. Initial margin received cannot be rehypothecated, i.e. re-used as collateral in other transactions, which is an important difference to variation margin. Therefore, the obligation to post initial margin imposes a funding requirement on both parties of a derivative transaction which cannot be netted against similar income.

The industry response to BCBS 261 has been two-fold: first to agree on a standard implementation of the core calculation of initial margin, an initiative coordinated by ISDA and giving rise to a standard for initial margin, ISDA SIMM™. The second response of the industry was to jointly invest in a utility firm, Acadiasoft, to coordinate the calculation, reconciliation and posting of SIMM™, between financial institutions on a daily basis.

The existence of a central utility with access to partial (anonymised) position data from a consortium of financial institutions makes it possible to optimise the risk positions between financial institutions in order to collectively minimise the amount of initial margin (and hence financial risk) in the system. This outcome is desirable both from a financial institution and regulatory standpoint.

3 The Optimisation Problem

Initial margin is in practice a value at risk measure (VaR). In the case of ISDA SIMM™ (insert link or reference) it is a modified delta/gamma parametric VaR, the calculation of which requires a covariance matrix and information about the portfolio sensitivities. The contribution of ISDA SIMM™ is, among other things, to set a standard covariance matrix and define a bucketing scheme for sensitivities to make it easier for diverse financial institutions to agree on a single quantity that represents the initial margin[1]. Without going into too much detail of the actual definition of the intial margin, we can write it simplified as

$$IM^2 = \mathbf{S}^\top \mathbf{\Sigma} \mathbf{S} = \sum_{ij} \Sigma_{ij} s_i s_j,$$

where $\mathbf{\Sigma}$ is a pre-defined covariance matrix (including scaling factors) and $\mathbf{S} = s_i$, $i = 1 \dots m$ is the sensitivity vector of the portfolio in question. In practice, there is a threshold amount, T, of 50M EUR below which it is not necessary to post initial margin, hence the initial marging actually posted, M, is given by

$$M = [IM - T]^+. \tag{1}$$

We are going to concern ourselves with n institutions, each of which is required to post initial margin according to equation (**??**). To define some notation, the

[1]Any two parties will still have to reconcile their sensitivities, which may be difficult if they use different pricing models, data sources and snapshot times. However, the level of difficulty is not higher than that of agreeing the variation margin, which is nowadays a smooth process in most banks.

amount that institution i posts to institution j is M_{ij}, likewise the sensitivity vector of institution i facing institution j is $\mathbf{S}_{ij} = -\mathbf{S}_{ji},$. Our goal is to minimise

$$F = \sum_{i \neq j} M_{ij}, \tag{2}$$

subject to the constraints, by changing the sensitivity vectors $\mathbf{S}_{ij}$ by quantities $\boldsymbol{\Delta}_{ij}$. The simplest constraint, and the one we start with, is that the aggregate risk position of each institution remains unchanged as a result of the optimisation, that is

$$\mathbf{S}_i = \sum_j \mathbf{S}_{ij} = \sum_j (\mathbf{S}_{ij} + \boldsymbol{\Delta}_{ij}) \qquad \Leftrightarrow \qquad \sum_j \boldsymbol{\Delta}_{ij} = 0 \quad \forall i. \tag{3}$$

4 Solution Approach

It is generally conceivable that by collapsing offsetting positions and redistribution of positions to utilise thresholds the sum of all SIMM's may be reduced. But how to find such beneficial redistributions of risk? For example, it is straight forward to find un-utilised thresholds, and it is tempting to move risk from high SIMM combinations to these, but what is the optimal way of doing this from the entire consortium's perspective, i.e. in terms of the cost function? If we find one opportunity to collapse positions which reduces party A's overall SIMM, how does this affect other parties' SIMM when we introduce additional risk transfers necessary to satisfy the strict constraints? Which position collapses are then beneficial from the overall objective's perspective, and which collapses are optimal or near optimal choices? Rather than trying any heuristic and intuitive searches down this route we aim at developing a systematic approach in the following.

Three Entities

To start, let us consider the simplest case of three parties dealing with each other and exploring SIMM reduction potential. We label the parties 1, 2, and 3, respectively. We further consider a change in the sensitivity of positions that party 1 faces with party 2 by an amount Δ^k. Sensitivity is a vector, and Δ^k means a change in sensitivity vector component k. To satisfy the constraint that party 1's total sensitivity must not change, we have to make the offsetting change of party 1's sensitivity facing party 3. And in order to keep party 2's and 3's overall sensitivity position unchanged as well, we need to add further offsetting sensitivity changes between 2 and 3. This chain of risk transfers is illustrated in figure **??**.

This shows that for a group of three entities (a triple), only a single vector of shifts Δ^k determines the shifts in sensitivities among all parties in the triple

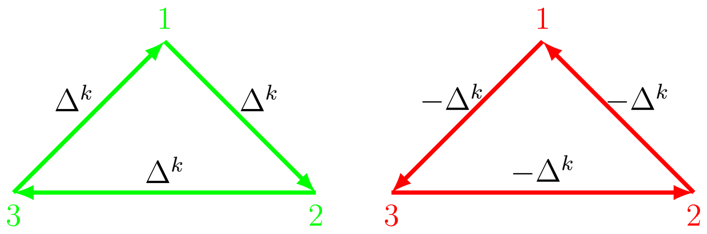

Figure 1: Risk transfers between three parties which satisfy constraints (**??**).

(three sensitivity vectors, or six including the reverse perspectives). Rather than considering each sensitivity vector's changes as 'problem parameters' with rigid constraints we therefore consider the single vector of shifts $\mathbf{\Delta}$ as problem parameters, with straight forward translation into sensitivity changes for all parties, and in particular *without any further constraints*. This is useful since unconstrained optimisations are generally more efficient than constrained optimisations.

More Entities

The situation becomes more complex when moving to more than three entities. So let us consider four next. Looking for closed loops – closed chains of three or more parties where the same change in sensitivity is passed on between all members of the chain – along which we can apply sensitivity moves satisfying all constraints by construction, we easily see that there are five such loops in total, four of them involving three parties, and one involving all four. The loop involving four entities can be seen as a concatenation of two suitable loops of three entities, so that the four-loop appears to be redundant. Clearly we need to consider more than a single loop here, but before we jump ahead and conclude that we need the four three-loops in this case, let us analyse the problem systematically and determine the *degrees of freedom* of the problem. This will tell us how many loops we will have to take into account. And as a next step we will then determine how to select these loops for the general case of N entities. We see that we can always restrict our search to three-loops, because any k-loop with $k \geq 3$ can be composed from $k-2$ three-loops.

Assuming N entities, what are the degrees of freedom of the problem, i.e. how many unconstrained model parameters do we need? We further assume that the dimension of the sensitivity vector is S, i.e. the sensitivity vector has S components and represents the 'union' of sensitivity vector compositions (including vegas) among all parties involved. If we include all parties' perspectives, then we count $S \times N \times (N-1)$ sensitivities in the system. Following Equation (**??**), we label sensitivities $\mathbf{S_{ij}} = (s^k_{ij})$ where i, j denote entities, k denotes the sensitivity vector component, and s^k_{ij} may differ from s^k_{ji} if we 'receive' sensitivity input from all parties independently[2]. However, we would generally expect $s_{ij}^k \approx -s_{ji}^k$. The *changes* to sensitivities $\mathbf{\Delta_{ij}} = (\Delta s_{ij}^k)$ are then constrained in two ways:

- Symmetry: $\Delta s_{ij}^k = -\Delta s_{ji}^k$ though absolute sensitivities may differ; these comprise $N \times (N-1)/2$ constraints per component k;
- Constant net sensitivity for each party: For any fixed i, $\sum_{i \neq j}^N \Delta s_{ij}^k = 0$; these yield additional $(N-1)$ independent constraints per component k.

In summary, the problem dimension or degrees of freedom D are

$$\begin{aligned} D &= S \times N \times (N-1) - S \times N \times (N-1)/2 - S \times (N-1) \\ &= S \times (N-1) \times (N-2)/2. \end{aligned}$$

Let us check:

- For $N = 3$ we have $D = S$, this agrees with the number of shift parameters applied along our single triple loop.
- For $N = 4$ we get $D = S \times 3$, i.e. three loops should be sufficient. Since we have seen that the four-loop is redundant, there must be one more redundancy. In fact, the triples (1,2,3), (1,2,4) and (1,3,4) are sufficient to cover all cases (generate all possible three- and four-loops by concatenation). Note that this choice is not unique, one could for example also use (2,3,4) instead of (1,3,4) etc.

Generally, we can build a valid set of triples as follows

$$\begin{array}{rl} (1,2,x) & \quad x = 3,\dots,N \\ (1,3,x) & \quad x = 4,\dots,N \\ (1,4,x) & \quad x = 5,\dots,N \\ \vdots & \\ (1,N-1,N) & \end{array}$$

This yields in total $\binom{n-1}{2} = (N-1) \times (N-2)/2$ triples, matching the problem dimension. Elements from the list above can be concatenated to form any triple (a,b,c) with $1 \leq a < b < c \leq N$ and in turn any 'longer' loop (for a proof see Appendix **??**). These triples hence form both a minimal and complete set of loops, i.e. which allow covering all states of the optimisation problem.

Let us summarise what we have achieved so far:

[2]For instance, this may be due to different pricing models or market data inputs that the two respective parties are using, which will necessarily lead to different sensitivities.

- For N entities we can identify a minimal number of three-loops or triples in a simple way

- The consecutive application of independent shifts along all three-loops above generates all permissible sensitivity states

- These sensitivity scenarios satisfy the boundary constraints (**??**) by construction

- The remaining optimisation problem is unconstrained when formulated in terms of *shifts along loops*.

Optimiser

So far we have just formulated the optimisation problem in terms of independent/unconstrained parameters. It is still to be determined how we 'walk' through this D-dimensional parameter space to find configurations that reduce the objective function value. This may be challenging, because

- the problem dimension D is large in realistic cases: 200 sensitivity components (only) and 5 entities (only) means $D = 1200$, 10 entities means $D = 7200$;

- we cannot exclude that the problem has various sub-optimal local minima in which an optimiser can 'get stuck'[3]

The latter point means that we generally have to resort to an optimisation procedure that does not simply 'walk' straight downhill into the next available local minimum, but which explores the parameter space globally. Many such *meta-heuristic schemes* are described in the literature (add references). We have found the *Differential Evolution* genetic algorithm [**?**] to be useful in this context. In summary, the algorithm generates a sequence of generations, each containing many alternative populations that are initially randomly generated. Each population represents one sensitivity scenario, and for each population we evaluate the objective function. Populations are then sorted by function value. The populations of the subsequent generation are then generated from the previous generation's populations by *crossover* and *mutation* where the latter random modifications give the optimiser the potential to 'walk away' from local minima with some probability. The algorithm's implementation has various configuration options such as the population size which determine convergence speed and proximity of the result to the global optimum after a given number of generations.

[3]For example, we could chose to keep all sensitivities fixed except for one risk factor whose sensitivity we modify along a single loop; along this one-dimensional path we will generally find a minimum of the objective function which is most likely just a local one; the same can be done for any other risk factor or any other loop.

Obviously, this approach involves many repeated computations of SIMM for all parties so that an efficient implementation of the SIMM calculation itself is crucial. Moreover, the *Differential Evolution* optimiser is parallelised, distributing the evaluation of the population in each generation over multiple threads.

Additional Constraints

It is conceivable that additional constraints have to be taken into account. We briefly consider two likely constraints here.

1. In the consortium of N parties, some pairs are not allowed to deal with each other at all. This can be solved by reviewing the list of three-loops for N entities, removing those which contain forbidden connections, and adding alternative three-loops to cover the problem dimension (original number of loops minus number of 'forbidden trading' constraints) such that all permissible pairs are covered at least once. This preserves the algorithm above.

2. Some of the entities do not accept a SIMM increase: The optimisation aims for the overall reduction of the objective function which - so far - allows for individual SIMM increase if the overall SIMMs decreases[4]. However, if a particular party does not accept a SIMM increase, this can be implemented conveniently in the objective function: If the party's current SIMM (for a particular state during the optimisation process) exceeds the initial SIMM then add a prohibitively large contribution to the objective function value. This will ensure that this state will not be considered in subsequent generations.

Further constraints such as risk limits can be similarly incorporated into the algorithm via modifications of the objective function.

Ongoing Research

We are currently investigating several options to further enhance the algorithm's convergence speed, including

- dimension reduction by selecting a subset of the sensitivity vector for optimisation
- PCA with truncation to further reduce number of independent variables
- local optimisation after global optimisation
- heuristics to find good starting points, followed by global optimisation

[4]This may require establishing a compensation mechanism between the parties post optimisation

5 Examples

To illustrate the approach outlined above, we consider two examples with three respectively five entities. The sensitivity vector has 19 entries in the 'IR/FX' class, randomly generated and distributed over the entity pairs such that a full collapse to zero sensitivity and hence SIMM for all parties is possible.

5.1 Three Entities

Table **??** shows a comparison of initial and final SIMM and and excerpt of the sensitivities for this three-entity example, after optimisation. This demonstrates

Entity1	Entity2			InitialSIMM	FinalSIMM
0	1			11918155	18201
0	2			11583549	21467
1	0			11583549	21467
1	2			11918155	18201
2	0			11918155	18201
2	1			11583549	21467
Entity1	Entity2	Factor	Active	InitialSensi	FinalSensi
0	1	Risk_IRCurve/RatesFX/GBP/0/2/2	1	-91864	-6
0	1	Risk_IRCurve/RatesFX/GBP/0/4/2	1	101705	5
0	1	Risk_IRCurve/RatesFX/GBP/0/8/2	1	-81382	7
0	1	Risk_IRCurve/RatesFX/GBP/0/9/2	1	-110039	-10
0	1	Risk_IRCurve/RatesFX/USD/0/1/0	1	-78464	34
0	1	Risk_IRCurve/RatesFX/USD/0/1/3	1	97010	-34
0	1	Risk_IRCurve/RatesFX/USD/0/10/0	1	97334	6
0	1	Risk_IRCurve/RatesFX/USD/0/10/1	1	-98494	2
0	1	Risk_IRCurve/RatesFX/USD/0/10/3	1	110801	-6
0	1	Risk_IRCurve/RatesFX/USD/0/12/1	1	-88351	2
0	1	Risk_IRCurve/RatesFX/USD/0/3/0	1	-98068	-34
0	1	Risk_IRCurve/RatesFX/USD/0/3/3	1	121256	34
0	1	Risk_IRCurve/RatesFX/USD/0/5/1	1	87051	1
0	1	Risk_IRCurve/RatesFX/USD/0/5/3	1	-99166	-1
0	1	Risk_IRCurve/RatesFX/USD/0/7/1	1	104796	-1
0	1	Risk_IRVol/RatesFX/GBP/0/1/0	1	83070	-753
0	1	Risk_IRVol/RatesFX/GBP/0/11/0	1	93913	65074
0	1	Risk_IRVol/RatesFX/GBP/0/3/0	1	95472	9589
0	1	Risk_IRVol/RatesFX/GBP/0/6/0	1	97172	-97138

Table 1: SIMM and sensitivities after optimisation with finite number of generations and population size 1000. Convergence is almost perfect with the exception of the vega risk factors.

that the approach is in principle capable of identifying SIMM reduction potential (down to less than 0.2% of initial SIMM in this case) and to obtain a result close to the global optimum. Note that the algorithm was not 'directed' in any way towards collapsing sensitivity positions. The snapshot in table **??** is rather the result of the unbiased procedure described in section **??**. The speed of convergence and quality of the result depends on the parametrisation of the *Differential Evolution* algorithm, in particular the chosen population size. Figure **??** shows the evolution of the objective function with generations for different choices of the population size.

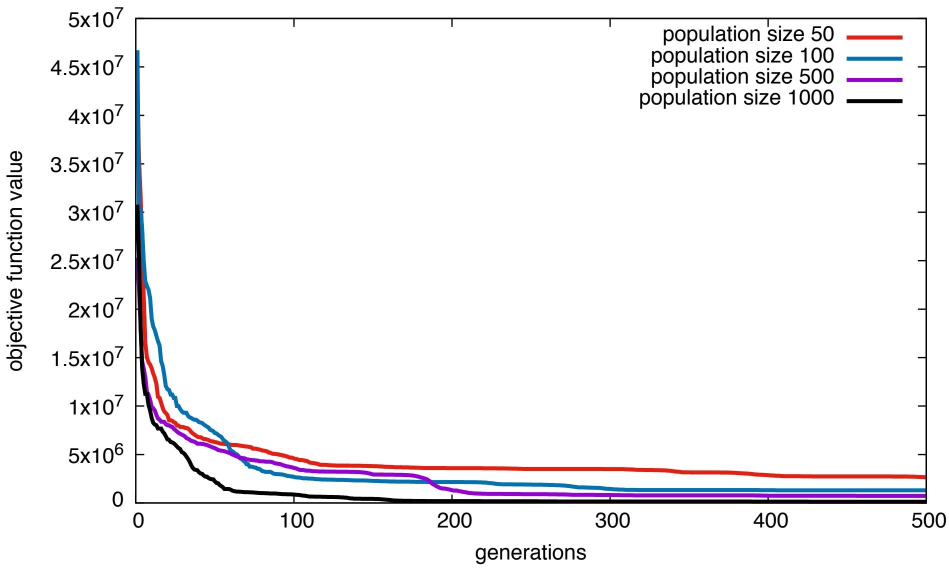

Figure 2: Convergence of the optimisation procedure for various choices of the population size. Increasing the size speeds up convergence and yields results closer to the global optimum (zero SIMM in this case).

5.2 Five Entities

6 Summary

Blah

A Proof

Given the set $\{1, \ldots, n\}$ of the first n natural numbers, there are $\binom{n}{2}$ ordered pairs of numbers (a, b) with $a < b$. We use the convention of signing arbitrary pairs as follows: If $a < b$, the sign is positive, and we write (a, b). Otherwise, we write $-(a, b)$.

Likewise, there are $\binom{n}{3}$ ordered triples (a, b, c) of numbers $a < b < c$. We give an arbitrary triple an orientation by giving a sign to each of the three tuples it contains. For instance, the triples $(1,2,3) = (1,2) + (2,3) - (1,3)$ and $(1,6,2) = (1,6) - (2,6) - (1,2)$.

For $n \geq 3$, we are looking for a minimal subset of triples from which we can construct all triples. A very simple set of triples is

$$T_1^n = \{(1,a,b) \,|\, 2 \leq a < b \leq n\},$$

which contains $\binom{n-1}{2}$ elements. More generally, we can define

$$T_k^n = \{(k,a,b) \,|\, a < b, a, b \neq k\}$$

for $k \leq n$.

Lemma. Any triple $(a,b,c), a < b < c$ can be written as a finite sum of signed triples from T_1^n.

Proof. Induction by n.

The claim is trivial for $n = 3$.

Assume the claim is true for $n \geq 3$. Because of the induction assumption and $T_1^n \subset T_1^{n+1}$, we only have to consider triples of the form $(a,b,n+1)$. We note that

$$\begin{aligned}(a,b,n+1) &= (a,b) + (b,n+1) - (a,n+1)\\ &= (1,a) + (a,b) - (1,b)\\ &\quad - (1,a) - (a,n+1) + (1,n+1)\\ &\quad + (1,b) + (b,n+1) - (1,n+1)\\ &= (1,a,b) - (1,a,n+1) + (1,b,n+1),\end{aligned}$$

which concludes the proof. ■

Since there is nothing special about index 1, it is clear that any of the T_k^n spans the set of all triples.

References

[1] Price, K., Storn, R., 1997. Differential Evolution - A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. Journal of Global Optimization, Kluwer Academic Publishers, 1997, Vol. 11, pp. 341 - 359.