

Global valuation and dynamic risk management

Claudio Albanese, Guillaume Gimonet and Steve White

November 5, 2010

1 Market standards

Some instruments such as interest rate swaps and variance swaps, admit low-risk, fairly robust replication strategies that do not rely on probabilistic models. As a consequence, these instruments are traded with confidence and they are among the most liquid pillars of financial markets. Vice-versa, instruments which cannot be easily replicated are not liquid and require model-based valuation.

To arrive at a consensus around valuations of illiquid instruments, the finance industry has structured itself around standards. Valuation methodologies have historically followed cycles whereby innovations were promoted, disseminated and finally established into industry-wide standards. Rating methodologies developed into standards for the sake of maintaining consistency. Risk management methodologies developed significantly in the 1990s after the introduction of value-at-risk (VaR) measures [29] and mark-to-market and mark-to-model accounting. Capital adequacy directives subsequently encoded the industry practice into a regulatory framework.

Standards are double-edged swords. Markets benefit from standards as these facilitate communication and mutual risk assessments, thus enhancing confidence and trading volumes. Standards are beneficial because they provide a measure of stability, facilitate the commoditisation of business functions and spur the growth and efficiency of the financial system.

But standards tend to fall victims of their own success. The complexity of derivatives markets has increased at a rapid pace over the last few decades, creating a global financial system that is vastly different from the one prevailing in the 1990s and upon which existing standards were developed. As scales change, standard methodologies confer rigidity to the system and potentially give rise to systematic misspricings, faulty risk assessments and systemic risk.

In parallel with the growth of the financial industry, we also witnessed vast improvements in computing technologies with the advent of massively parallel multi-core architectures. Nowadays, high-throughput computing microchips and high-density multi-core CPU boards enable highly multi-threaded designs and shared memory applications that could not possibly have

been conceived in the grid computing environments introduced in the 1990s. However, these latter technologies are still prevailing in the financial services industry.

In principle, one could hope that improved technologies would allow for a greater degree of efficiency in information processing so that risk management systems could cope with market growth by leveraging on technology innovations. However, the existence of financial standards represents an obstruction to this process. Standards embed numerous technical assumptions and mathematical approximations that are validated by market practices and regulatory frameworks. Technology innovations can be absorbed only gradually and insofar as they recognise and accelerate existing processes, providing cost reduction benefits.

Technology shocks enabling qualitative innovations are much harder to absorb. Computing technologies are characterised by thresholds. Once processing power takes one across a critical threshold, balance in the ecosystem of algorithms is upset, new designs become viable and entirely new approaches, which are theoretically more rigorous and less dependent on faulty approximations, are possible. New developments on the computer engineering front potentially give rise to new and more efficient forms of financial organization. However, the resilience of established standards in the marketplace hampers the process of development and adoption of new standards.

When investigated in sufficient depth, standardised methodologies reveal their roots as tightly linked to the limitations of computing technologies prevailing at the moment standards were proposed. Regulatory provisions validate methodologies and approximation schemes without, however, providing an incentive for radical innovation.

Major crises unveil opportunities for change on a scale that is seldom seen. The financial storms that have rocked markets since the summer of 2007 are no exception in this regard, as they brought unprecedented focus on the structure and internal workings of financial markets. In this paper, we give a new look at this from our own angle, and imagine guiding principles of renewal, leading to innovative and possibly more robust practices in the financial industry.

This paper provides a general discussion and does not dive into a detailed discussion of specific models with Mathematical formulas. The reader interested in a more technical discussion is referred to [1] and references therein. See also [2] for a more concise presentation.

2 de Finetti's Fundamental Theorem of Finance

Valuation and risk management methodologies rely on numerous mathematical approximations and estimation strategies. Market standards reflect generally accepted approximations and methodologies, and regulations freeze them into enforceable directives. But what if there were no technology limitations and approximations were not needed? How would we ideally structure valuation systems, risk management practices and banking operations?

The pivotal theoretical result for asset valuation is the Fundamental Theorem of Finance, invented in 1931 by Bruno de Finetti, [14]. The result takes the premise from the ancient principle of arbitrage freedom, according to which riskless profits are not legitimate. In medieval times, the prohibition of riskless profits was stipulated in law. As free financial markets developed, the principle of arbitrage freedom began to be regarded as a self-enforcing reality. The argument is that the very observation of trading opportunities which give rise to riskless profits prompts actions by market participants aimed at exploiting them. But in so doing, the actions cause price shifts that eliminate the opportunity itself.

The mathematical expression of the principle of arbitrage freedom can be stated as a set of linear inequalities. The Fundamental Theorem indicates how to find all solutions to these equations and how to express them via transition probabilities which can be estimated, or implied, from market prices.

For the sake of this argument, consider the example of a portfolio composed of a number of positions in various financial contracts. Each contract is written in a subset of a natural language such as English. The subset is sufficiently restricted and rigorous that disputes on the meanings and implications of contractual clauses can safely be decided in a court of law with no room for ambiguities and misunderstandings. Decidability is in fact a concept of paramount importance for all facets of finance practice and technology.

Legally valid financial contracts envision a number of scenarios that may possibly occur, and stipulate exchanges of assets or monetary instruments between the contractual parties at one or more times in the future, in amounts contingent to the realization of future scenarios. What matters is that possible scenarios are contractually identified and that the occurrence of future events which can potentially affect an asset exchange between the parties can be verified or falsified with certainty at the time the exchange is contractually stipulated to occur. The language constructs of financial contracts are thus ruled by temporal modal logic, the branch of Aristotelian predicate logic including also the notions of time and the distinction between possible, impossible and certain future events.

Asset valuation is a form of measurement that necessitates a meter. Since only the relative value of two assets is meaningful, a valuation methodology requires the identification of a numeraire asset. In principle, any asset would do. Gold has traditionally been privileged as a value numeraire due to the relative ease of secure storage and transferability. But cash deposited in money market accounts, accruing interest on a daily basis, are a far more practical way of expressing and transferring value.

To mathematise precisely the notion of arbitrage freedom, one can contrast trading strategies involving a given portfolio. Consider a strategy consisting of simply holding the portfolio over a certain time period and compare this with the strategy of liquidating the portfolio in terms of the selected numeraire and holding the numeraire over the same period. If the liquidation of the

portfolio at a given point in time can possibly give rise to a gain at a future time with respect to the strategy of just holding it, then there ought to be another possible scenario under which the strategy of holding gives rise to a profit instead.

Mathematically, the events of loss and gain can be expressed through systems of linear inequalities. These systems have been studied by various mathematicians. In 1826, Fourier proposed to consider any given system of inequalities along with a dual system a technique which he proves is instrumental to finding solutions, [19]. In 1903, Farkas elaborated along this line and found a useful technical Lemma that established necessary and sufficient conditions of solvability of a system and its dual, [18]. By applying Farkas Lemma to the systems of linear inequalities expressing the condition of no arbitrage, one arrives at de Finetti's Fundamental Theorem of Finance, which is predicated on the duality between prices and discounted transition probabilities.

According to the Fundamental Theorem, all valuation schemes consistent with arbitrage freedom can be expressed by assigning probabilities to all future scenarios. The relative value of any asset with respect to the fixed numeraire of choice can then be expressed in terms of probabilistic expectations of the future payoffs stipulated in the contract.

From a valuation theory angle, probability theory is a mathematical invention to express all acceptable solutions to a system of constraints for absence of arbitrage. In de Finetti's words, the probability theory that emerges in valuation theory through the Fundamental Theorem of Finance is subjective as it proposes to capture expectations of market participants, as opposed to intrinsic frequencies of occurrence of random events. To relate these probabilities to real-world statistical observables, one needs to include risk adjustments that model the discrepancy between the willingness of market participants to accept bets on future events and the forecasted probability of the actual events occurring, [15].

The Fundamental Theorem yields a mathematical framework for probability theory based on temporal modal logic. This approach is very close to the Philosophy of Mathematics elaborated by Wittgenstein in the 1930s, who proposed to root mathematics into language [35] [36]. Both de Finetti and Wittgenstein insisted in upholding the Aristotelic principle according to which the objects of discourse in Mathematics and Finance need to be exclusively finite sets. The notion of infinite sets can only be intended in the potential sense of passage to the limit, not as representing a property of actual objects of mathematical discourse.

The use of infinite sets leads to statements that could not possibly be settled in court. One could produce many examples of such statements but, just to mention one with which Mathematical Finance students are confronted, consider the following: There are subsets of the real numbers which are not Lebesgue measurable. All students aspiring to quantitative analyst jobs repeat this statement as part of the curriculum development. However, this theorem depends on the so-called axiom of choice, a social convention regarding infinitary logic that most math-

ematicians at the moment find intuitive but, as Cohen showed in 1964 [10], is independent of the other standard axioms. Further to that, Solovay in 1970 showed that if one was willing to eliminate the axiom of choice and replace it with another axiom stating that all sets are Lebesgue measurable, one would still obtain an equally consistent axiomatic framework, [32].

If two parties had to agree on a payoff to be exchanged at a given date in the future in case there exists a set that is not Lebesgue measurable, no court of law would possibly be able to arbitrate and decide the case. In other words, this is an instance of a sentence which is not decidable. The reasons why students of Finance are asked to accept a language with such ambiguities and obvious pitfalls are rather arcane. Unfortunately, this is more than an academic oddity but rather the tip of a far greater and unsettling iceberg.

If we step back for a moment and imagine a theoretically sound valuation and risk management process without regard for technical viability, we would envisage a methodology along the following lines:

- (i) Make a list of all possible future events as dynamic scenarios. This involves identifying all sources of financial uncertainties as a list of risk factors and a specification of possible scenarios that said factors can realize in the future.
- (ii) Run lexical analyzers to map contractual specifications into descriptions of future payoffs contingent to the realization of risk-factor scenarios. This involves capturing instrument descriptors as classes embedding not only attributes but also payoff specifications, exercise conditions, etc.
- (iii) Select a set of risk factors spanning the entire risk space but not over-complete. For instance, not all FX crosses are independent as they form triangular relations; hence only a spanning set of crosses can be included, [1].
- (iv) For each risk factor, identify the most comprehensive portfolio of contracts dependent only on that individual risk factor and whose market valuations are known with some degree of confidence. Based on this information, one can then calibrate, i.e. infer probabilities of future scenarios for that particular factor marginal. Successful calibration for single-factor processes requires a parsimonious and economically realistic specification involving as few parameters as possible to achieve realism, but not fewer.
- (v) Value all contracts dependent on a single risk factor by means of the marginal processes inferred at step (iv).
- (vi) Structure a general dynamic correlation model among all risk factors and estimate correlations based on pricing information for contracts depending on several risk factors.
- (vii) Value all assets depending on several factors using the probabilities inferred in step (vi).

- (viii) Analyze historically realized risk-factor series and model the systematic discrepancies from the inferred probabilistic models with the objective of finding model adjustments to achieve consistencies. These adjustments express the market price of risk. The adjusted probability measure is called historical, to contrast it with the implied or pricing measure obtained in the previous steps. The two are in general different, but need to be intimately related and share the same sets of possible scenarios, or else the hedging problem is ill posed.
- (ix) Identify relevant single-factor sub-portfolios to compute and store risk factor-dependent pricing information at future points in time using the estimated pricing measure. Sub-portfolios of exotic exposures depending on more than one factor are not to be tabulated.
- (x) At all times in the future and for all scenarios, maintain the value of all sub-portfolios of interest.
- (xi) Generate scenarios for the future evolution of risk factors using the historical measure containing also a reasonably estimated risk adjustment.
- (xii) At each epoch date in the simulation, evaluate the pre-computed tables giving future prices of single-factor portfolios. For the remainder exotic portfolio of multi-factor exposures, carry out an embedded Monte Carlo pricing step.
- (xiii) The time horizon for scenario generation should be at least as long as the longest expiry in the book, in order to generate a comprehensive lifetime assessment and risk profile for all sub-portfolios of interest.
- (xiv) Identify state-contingent hedging strategies to mitigate the risk of bank portfolios held for the purpose of financial intermediation.
- (xv) Quantify the impact of model uncertainties.
- (xvi) Identify optimal allocation strategies for client portfolios held for the purpose of investment by assessing suitable risk-reward metrics and applying portfolio theory.

3 The Black-Scholes-Merton methodology of local valuation

The previous section is based on the status of Theoretical Finance in the 1950s. At that stage, all the foundational concepts had already been elaborated at the theoretical level. The practices that ensued afterwards were governed by the necessities of implementing a form of the theory with limited technology and deviated substantially from this theoretically sound framework.

The first option pricing formula in analytic closed form was proposed by Bachelier in this doctoral thesis [4] written in 1900 under the supervision of Poincaré. This piece was one of the

pioneering works on the mathematical theory of Brownian motion, preceding by five years the Physics paper by Einstein [17] and [30].

The drawback of the Bachelier model from the financial modelling viewpoint was that, under the Brownian motion dynamics, the underlying can take negative values. In 1973, Black, Scholes [6] and Merton [28] introduce a variant of the Bachelier formula based on geometric Brownian motion. Under this modified dynamic specification, the underlying is constrained to be positive. Positivity is achieved by postulating that volatility falls linearly to zero as a function of the underlying asset price. This is a radical assumption whose drawbacks sometimes outweigh the benefit of positivity. Nowadays, the Bachelier formula and the Black-Scholes formula are both widely used as a convenient way of parametrizing option prices in terms of an implied volatility parameter. In the foreign exchange domain, there is generally a preference for Black-Scholes implied volatilities as cross rates are never zero. For interest rate derivatives instead, Bachelier implied volatilities are most used because they better reflect the volatility process in situations where the underlying is near zero. In the case of equity derivatives, the situation is mixed and both formulas are used.

The contribution by Black, Scholes and Merton however went far beyond the proposal of a useful new formula to parameterize option prices: they introduced a far reaching deformation and re-interpretation of de Finetti's Fundamental Theorem of Finance revolving around the notions of *replication* and *local valuation*.

The original version of the Fundamental Theorem stated that, assuming absence of arbitrage, there exists a *global* probability measure over the set of all possible future events which can possibly be realized in the future such that current prices can be written as discounted expectations of future cash flow streams. Emphasis here is on the adjective *global* as the same measure should be used to price all assets. In general, it is not possible to replicate a price process by means of a dynamic trading strategy. Replication is of course highly desirable but can realistically be achieved only in approximate fashion within a truly global model. In other words, when using global models we should be deeply aware of model risk and its repercussions on the ability to hedge dynamically.

Black, Scholes and Merton turned the argument on its head. They restricted the attention to imaginary financial micro-universes in which only a handful of assets were traded. The canonical example involves a non-dividend paying stock, an option and a cash account. Within such a micro-universe, they then postulated that the dynamics follows geometric Brownian motion. Assuming also absence of arbitrage, they demonstrate that option replication is theoretically feasible. Unsurprisingly, the prices obtained by the replication argument also obey de Finetti's Fundamental Theorem as the class of financial models admitting replication are simply a particular case.

Nowadays, Finance students are taught that pricing by replication is the key central concept

of derivative valuation theory. The ability to express prices as discounted expectations is mistakenly presented as a consequence of the ability to replicate. To reinforce the message, the overwhelming majority of textbooks follows John Hull's bestseller [23] and go to great lengths to narrow down the focus of attention on models which theoretically admit exact replication: namely, either variations on the binomial lattice model or models based on diffusion processes in the continuum.

The situation was further complicated by technology limitations. In fact, in the 1970s it would have been technically very difficult to use models with no analytical tractability. Since stochastic calculus provides examples of analytically tractable models which also admit dynamic replication, the two requirements were linked and fused with one another. Stochastic calculus adds further layers of complexity through the use of infinitary arguments that are rife with logical paradoxes. Most students have no chance to ever master stochastic calculus beyond learning the mechanical use of the ubiquitous Ito's Lemma.

By restricting the emphasis on the class of models which lend themselves to dynamic replication, one conveys the erroneous impressions that robust replication of all derivatives is indeed possible. To achieve it, all one needs is to dream of a restricted financial universe and select a model out of the very narrow class of dynamic specifications which theoretically admit perfect replication and one is magically going to obtain theoretically sound hedge ratios. On the contrary, models which admit dynamic replication are at odds with econometric evidence.

To compensate for the use of poor models, market participants developed methodologies based on the notion of *local valuation*, a concept that in its very essence goes against the core of de Finetti's Fundamental Theorem. According to the local valuation paradigm, instruments need to be valued individually using deal specific model specifications which are consistent in a miniature financial universe entailing only the instrument itself, its hedging vehicles and a cash account. From this standpoint, models need to calibrate to very limited market information given by the prices of selected few hedging vehicles, an easily achievable task even with poor models.

The RiskMetrics [29] methodology extends local valuation to the portfolio level by arguing that each portfolio position should be priced by means of a deal specific model and sensitivities should then be aggregated on the presumption that model risk diversifies away. Financial regulators [5] quickly validated the proposal and pushed it industry-wide. From the technology viewpoint, grid computing was ideally suited to handle the task: if valuations need to be instrument specific, pricing tasks include instrument specific calibration and can be dispatched to individual nodes without any need of mutual communication. To further refine the portfolio methodology in Ptolemaic style with additional layers of cycles and epicycles, empirical "adjusters" methodologies [21] were developed to magically "turn good prices into great prices". Data providers thought well to patent adjuster alchemies that guarantee "great" prices, see [34]:

according to patent law, mathematical theorems cannot be patented, but alchemy can, so why not? CDOs were priced with models that ignored the basket constituents, except that a broad variety of "mapping methodologies" were invented to empirically relate models across different baskets, all without the benefit of a theoretical understanding based on the Fundamental Theorem.

Counterparty credit risk is a new development along these lines. In this area, the metric for counterparty credit risk on which an agreement was reached among the industry and regulators is the Credit Valuation Adjustment (CVA), see [7], [9], [8], [11], [25], [27] and [33]. The CVA of a portfolio of netting sets is defined as total expected loss. Since expected loss is a linear function, the CVA is equal to the sum of the expected losses for each counterparty. The CVA owes its existence to this property which brings it within the reach of local valuation methodologies on grid computing equipment. However, linearity is precisely what makes the CVA immune to credit correlation risk. Arguably, the correlation between default arrival times of counterparties is the single most important risk factor to be assessed and nearly all risk metrics one could possibly imagine are sensitive to it. The CVA is the only measure which is not sensitive to credit correlation risk and it is precisely this odd property which, far from making it useless and unappealing, motivated instead its selection as the risk metrics of choice to assess capital adequacy requirements!

The flaws of local valuation can be summarized as follows:

- (i) A financial model for any given instrument needs to identify a restricted universe of financial derivatives to be used as hedging vehicles for dynamic replication of the same, and calibrate exactly to them, with no a priori margin allowance for errors. Substantial discrepancies can still exist with respect to other observed prices, but as long as these are not used as hedges, mismatches are tolerated, even if sizeable.
- (ii) In local valuation there can be potentially as many models as there are instrument positions in a given portfolio. This implies in particular that inferred valuation measures are not attributed any forecasting power, as to formulate a forecast one would obviously need to identify a single measure.
- (iii) The price of risk adjustment cannot be estimated because of the proliferation of starting points used for valuation. As a consequence, historical measures need to be estimated separately.
- (iv) The standard statistical methodology to estimate historical measures involves assuming that there is no model uncertainty and one has an infinitely long time series, so that one is theoretically justified in using the maximum likelihood principle. A more primitive alternative allowed by regulators is to generate future scenarios over a short time horizon,

such as ten business days, directly on the basis of historical returns over a time window of about 500 days.

- (v) Individual instruments in a portfolio are valued along with endogenous and exogenous model sensitivities. These two types of sensitivities are not differentiated from each other on the basis of the Black-Scholes analogy. In particular, model risk is not budgeted on an actuarial basis. Sensitivities of both types are freely aggregated at the portfolio level to reconstruct portfolio sensitivities with respect to risk factors.
- (vi) Risk-factor scenarios are modelled by means of measures estimated historically. In the estimation process, no consideration is given to ensure that the set of possible scenarios for the historical measure is related to those of the many processes underlying pricing models. As a rule, major inconsistencies are allowed, and the implied risk is not budgeted.
- (vii) Valuation models are selected out of a family of extensions of the Black-Scholes model which allow for analytic solvability, at least to the extent needed for calibration purposes. Calibration is then carried out separately on a deal-by-deal basis. Analytic solvability is a straightjacket to model flexibility that hampers the ability to calibrate against a broad spectrum of targets. Again, pricing mismatches result in non-linear model risk which is not budgeted or accounted for in any way.
- (viii) Models which theoretically allow for dynamic replication, such as continuous time models with continuous paths, are privileged as they are erroneously considered theoretically more sound.
- (ix) In a risk management implementation with full valuation, one generates scenarios for each underlying risk factor. For each scenario, one recalibrates each individual model used to value the portfolio positions. At the portfolio level, risk information is often aggregated by a mix of endogenous and exogenous sensitivities. Simulations then use an unrelated historical process.
- (x) Model risk is believed to be diversified into a marginal effect. However, this approximation is not controlled or even controllable quantitatively.

The theoretical foundation of the methodology is greatly limited by the extensive use of extrapolations by analogy. Analogy is the weakest form of logic and the most common trap for faulty reasoning. Empirical experience is processed by the human mind on the basis of analogy, and humans are particularly apt at pattern recognition. But the use of analogy has serious limitations. Engineers would never dream of building a bridge or designing a plane relying on long chains of arguments by analogy and without rigorous comprehensive and consistent modelling and assessments based on the laws of Physics and mathematical deductions based on

predicate logic. If they did, bridges would crack and planes would fall. But in Finance, analogy rules and faulty reasoning prevails.

Finance is not ruled by the laws of Physics, but instead there are firm regulatory requirements. This sometimes complicates matters. Consequences and implications of model risk are assessed empirically based on past performance and case studies carried out by regulators. Regulatory documents place the final seal of approval and in so doing they induce banking institutions to adopt uniform standards. Trading strategies are then designed to exploit the inefficiencies generated by the practice of using a myriad of mutually inconsistent models. The net effect is that, insofar as regulatory and market standards embed many mathematical approximations and empirical constructs without a sound theoretical basis, model risk compounds and correlates on a global scale in uncontrollable ways.

4 Mathematics

Validation by regulators and widespread market practice favored the acceptance of local valuation methodologies by recognizing and standardizing practices. But the real motivation that historically spurred this development laid originally in computing technology and Mathematics.

Mathematics in the 20th century did not evolve along the Aristotelian, finitist directions indicated by Wittgenstein and de Finetti. Instead, Mathematics developed into a Platonist discipline, defining itself as a science of the infinite. In fact, in a departure from tradition, in 20th century Mathematics actually infinite sets are considered as legitimate objects of mathematical discourse. This happened to resonate well with Black-Scholes-like modelling schemes of local valuation. Most valuation models are based on analytic solvability and the notions that time and prices are both continuous. These notions promise the other Platonist ideal of market completeness and dynamic replication, the cornerstones of model Finance whereby banks position themselves as market-neutral financial intermediaries.

Academic programmes for quantitative analysts are organized around this scientific paradigm. Analysts are trained in stochastic calculus to learn about stochastic processes defined directly in the continuum, because the consideration of this limit allows one to understand Black-Scholes theory of replication in its purest form, whereby replication is riskless. The technicalities behind these theoretical constructs are compounded by layers of formalism which obscure much of the substance. This indirectly grants scientific status to uncertain modelling constructs through the use of arguments by analogy, without firm application of theoretical rigor.

Using the rhetoric of analogy, as opposed to rigorous theoretical deductions as contained in the Fundamental Theorem to justify the empirical practices of local valuation, is an operation quite common in the social sciences. In their 2004 book [31], Bricmont and Sokal draw attention to the fashion in certain philosophical circles to use analogies from quantum physics or relativity

in support of otherwise shaky arguments. Since modern philosophers mostly address a non-technical readership alien to rigorous studies in Theoretical Physics and who are keen to use analogy in their thought process, the rhetoric paradigm turns out to be quite effective and pervasive. Key to the use of analogy is the abandonment of rigorous predicate logic in favor of a semiotic approach that leverages on the natural ability of the human brain to interpret and manipulate symbols. Engineers are trained to proceed much more carefully, as they reason in terms of datasets admitting a hardware representation, and with which one can effectively compute using the rules of arithmetic, not free associations by analogy, to achieve a faithful representation of physical processes. In the realm of philosophical discourse and Finance instead, freewheeling chains of arguments by analogy are the rule.

Similarly, the mathematical apparatus of probability and stochastic calculus was built around complex layers of formalism that obscure otherwise very simple concepts. Infinitary constructs are ubiquitous and used to create a semiotic environment, whereby symbols do not admit a hardware representation as datasets and whose meaning can be captured and manipulated only by human pattern recognition abilities. The step from here to an improper use of analogy is very short. Many Finance graduates have only a limited understanding of the formalism, no grasp of the underlying basic concepts, and find themselves under peer pressure to just go along with the crowd, pushing arguments by analogy for the sake of wording conclusions.

A layer of social complexity derives from the high degree of entropy of a social organization of labor, where a myriad of different engineering solutions are wired around a myriad of different mathematical shortcuts. This induces hyper-specialization and turf protection strategies across the many business niches in derivative markets, each implementing its own separate modelling framework and engineering solutions.

We are now all deeply aware of the consequences of this. Faulty pseudo-mathematical reasoning, combined with unjustifiable empirical constructs, give rise to systemic instabilities in the global economy.

Pressed by the imperative to move ahead, proposals arise. One possibility that is attracting attention is the one indicated by Haug and Taleb in [22]. His argument is centred on the unreliability of models of *all* models because of the occurrence of rare and unpredictable black swan events over which we have no control. The conclusion drawn is that one should give up pursuing the effort of modelling altogether and leave traders unconstrained.

Traders do not perform valuation with some pricing kernel until the expiration of the security, but, rather, produce a price of an option compatible with other instruments in the markets, with a holding time that is stochastic. They do not need top-down science. [22]

This sentence does reflect reality. Local valuation methodologies allow traders to mark their own books with a certain degree of freedom. By using models which are instrument-specific,

they do not need to coordinate their valuation with all other market information correlating to the same risk factors, or even reconcile valuations consistently within their own holdings. At the root of it lies the neglect of the Fundamental Theorem of Finance. Valuations are carried out by traders themselves using intuition, analogy, empirical interpolation and sometimes a good dose of wishful thinking. Pricing models are used more as a pro-forma than as a real tool. In the above quotation, Haug and Taleb conclude that one might as well honestly admit the state of affairs and eliminate the pretence that there is any scientific content in all this. They then continue their line of criticism of the Black-Scholes model and derived pricing schemes, saying that the portrayed efficiency of delta hedging is illusory because it is obtained under unrealistic assumptions. Mathematical models are very abstract and they are improperly used by analogy to validate the use of otherwise theoretically unjustifiable practices.

We agree with the diagnosis, but differ on the prognosis. Unfortunately, Talebs "black swan" events are neither rare nor unpredictable in financial markets. They historically have occurred about once in a decade, which is significant. Black swan events are not acts of God or caused by force majeure. They are predictable to the point that Taleb himself was able to forecast the 2007 event and allow his clients to take advantage of it by his own telling.

We maintain that black swan events in financial markets are instead the consequence of the unruly behaviour of market participants who are allowed to hide behind local valuation methodologies to create local bubbles in derivative markets. We agree that abstract mathematical models based on entirely unrealistic assumptions and shrouded in formalism are seriously damaging. But they are so because they are used by analogy and illicit extrapolation to validate practices without sound theoretical justification.

The fact that we do not need these failed models does not imply we do not need models at all. Models have been of little use in volatile conditions, if not damaging, and effectively left individual market participants either unconstrained or misguided because they were not consistent and theoretically sound. To move forward, we need to go back to first principles and build a theoretically consistent valuation framework with the help of the technology we now have access to.

Part of the solution to move ahead involves reformulating the mathematical apparatus of Finance on the basis of objects which admit faithful hardware representations as datasets. Finance needs to incorporate solid engineering practices. The infinitary abstractions of certain mathematical formalisms are to be avoided, as they deal with symbols which cannot be objectified as computable datasets, do not provide faithful representations of reality and can ultimately only be used in the framework of rhetoric exercises in analogy and uncontrolled extrapolations.

It is also necessary to reconstruct probability theory from the foundations as a discipline intimately intertwined with numerical analysis. One of the pivotal mandates of probability theory is to explain why certain numerical algorithms to compute probabilistic quantities are

proved effective and others are not. The other focus is to build computational strategies to broaden as much as possible the range of decidable questions arising in valuation.

In Finance, decidability is a concept of paramount importance. If a mathematical framework is weak, the range of questions that can be decided is inadequate. Since decisions are taken on the basis of algorithms that execute on computers, mathematicians are tasked with the goal of devising optimal formal frameworks for the given technology environment. In other words, one should recognize that Mathematics is signed by history: as Wittgenstein insisted, Mathematics is a collection of inventions, not of statements with eternal intrinsic value. Inventions are relevant only insofar that they achieve a goal on existing hardware better than all other inventions aimed at the same purpose. If uncompetitive, inventions are assigned to the dustbin of history and bad ideas. This dustbin should be well advertised to students because it is of great educational value and because an invention which is bad today may turn out to be excellent tomorrow. But still, at any given point in history, mathematical inventions are not just divided between true and wrong theorems: they are divided between inventions that broaden the set of technologically decidable relevant questions and inventions that are less efficient at that task.

As technology changes, Mathematics must adapt to stay relevant. This is a controversial statement, as most mathematicians nowadays believe they deal with absolutes and reject the notion of relevance. They interpret Mathematics as a science of the infinite, effectively turning it into a cult of which they are the high priests. Freedom of thought is of paramount importance in the sciences, but when these views are commoditised and passed on to students on an industrial scale, the social consequences are severe. Black swans are not an act of God, they are an act of the priesthood!

What we are advocating is a return of Mathematics to its engineering roots, a rediscovery of the spirit dominant in the 19th century that seeded the scientific and technology revolution. The detachment of Mathematics from engineering was initiated by Cantor, who introduced the notion that legitimate objects of mathematical discourse could be infinite. Actually, infinite objects cannot be represented as datasets, and results in this context can thus only be interpreted by analogy the root of all illusions of scientific rigor. Opposition to this trend was substantial from the start. Hilbert took an open view and attempted a compromise. He proposed, on general grounds, that whatever result can be proved by means of infinitary logic can also be derived using finitely means. This programme was later proven wrong by Godet, but still affected Lomonosovs foundational work on probability theory.

In 1932, Kolmogorov rooted probability in infinitary formalism by introducing the axiom of countable additivity [26]. He himself justified this as an expedient, saying it would facilitate calculations. He was motivated by his own work on the proof of the law of large numbers, whereby he provided two proofs: a constructive one which took many pages of lengthy estimates and a non-constructive one based on infinitary logic and the Borel-Cantelli Lemmas, which was shorter.

As he laid down the axioms for probability theory and extrapolating by analogy, he proposed that infinitary logic would provide useful shortcuts without counter-indications. The appellative expedient in Kolmogorov's original article however, was rapidly forgotten. As a result, mathematical Finance students are now confronted with the task of expressing themselves by means of infinitary abstractions they do not master, and their computers cannot represent, further weakening any rigorous thinking they may happen to start with. Nevertheless, Kolmogorov's proposal succeeded to an extent that perhaps went even beyond his own intention.

Dissent among mathematicians is still alive. V.I. Arnold, a Fields medal laureate for his work in classical mechanics and one of the most prominent mathematicians alive, in a visionary 1996 interview [3] says the following:

It is awful to think what kind of pressure the Bourbakists put on (evidently non-silly) students to reduce them to formal machines! This kind of formalized education is completely useless for any practical problem and even dangerous, leading to Chernobyl-type events. Unfortunately, this plague of formal deduction is propagating in many countries, and the future of the Mathematics infected by it is rather bleak.

It would seem that the Chernobyl-type event predicted by Arnold, actually occurred in 2007.

5 Technology

Computing technology in the 1970s and until the end of the 1990s was severely limited, if seen with modern eyes. Processors consisted of a single core and were able to process a single thread at a time. In addition, memory was a scarce resource. This motivated the development of algorithms that (i) can be parallelized by running separate processes on different nodes without communication and sharing memory and (ii) use memory sparingly.

Today's systems are vastly different. They are heterogeneous boards linking together multi-core MIMD processors and SIMD multi-processors. The acronym MIMD stands for multiple-instruction-multiple-data and denotes processors that are capable of executing threads independently of each other. SIMD stands for single-instruction-multiple-data and denotes multi-processors whereby a single instruction stream is shared among 16 or 32 different cores handling different data. These different designs are best at accomplishing different tasks and optimal configurations therefore combine both breeds.

MIMD boards perform best at Monte Carlo scenario generation and payoff valuation. Each random scenario has individual specificities and necessitates the execution of instruction streams that depend on the scenario itself. A SIMD processor would not accomplish this task efficiently, as all threads in the same block of 16 or 32 would be forced to coordinate and share the same instruction register, thus incurring inefficiencies and idle times.

SIMD multi-processors are strong at executing tasks which can be carefully planned and where it does not hurt to synchronize blocks of threads executing the same logic. Thread synchronization is advantageous for operations like reading and writing entire pages to and from memory. Scenario generation cannot take advantage of this, and as a rule requires random memory access. However, matrix manipulations such as multiplication are ideal for SIMD multi-processors. Instructions can be shared by multiple threads operating on contiguous memory addresses and access to memory can be coalesced into read and write operations of whole pages in global memory.

Current hardware capabilities thus provide heterogeneous systems comprising high-density MIMD boards with as many cores as possible and several SIMD multi-processors. The SIMD multi-processors provide matrix multiplication functionality and other linear algebra manipulations of similar type, while MIMDs are tasked to generate scenarios and conduct global orchestration and coordination. Efficient programming leverages on heavy multi-threading and exploits as much asynchronous execution as possible.

This sets the hardware backdrop against which probability theory and mathematical Finance need to be updated. Recognizing explicit hardware dependencies of the mathematical formalism is a condition for relevance. Along with a renewed foundational work in these areas of Mathematics, one needs to engage in the identification and optimization of a small set of hardware-dependent core library interfaces to which one can concentrate all numerical bottlenecks. The mathematical formalism needs to be structured around this core library to enable one to accomplish all necessary tasks in valuation and risk management.

There is, of course, room left for the development of mathematical abstractions. The most fundamental kind of abstraction is in the use of algebraic methods as opposed to the exclusive use of objects with a direct probabilistic content.

Historically, algebraisation of geometry proved extremely powerful. Power series expansions for trigonometric functions discovered by the Kerala school gave a way to tabulate these functions and produced pioneering examples of modern integral and differential calculus. These ideas were then algebraised and generalized by Taylor who discovered a general rule to obtain power series expansions, thus giving birth to modern calculus.

Much of the Mathematics in the last three centuries is based around the acquired ability by mankind of evaluating power series expansions. Technology involved the first use of computers in the 18th century, where computer was interpreted as an appellation for graduate students in Cambridge dedicated to the hand tabulation of special functions. Successively, mechanical devices were designed to accomplish the same task, culminating in the Babbage machine, a mechanical artifact to churn out power series calculations. These were the precursors to modern computers built with electronic parts.

The ability of evaluating power series expansions gave rise to the notion of a special function.

The special functions which are by far most common are the hypergeometric ones, written as expansions with coefficients of a special algebraic form depending on three parameters. Special cases of hypergeometric functions include trigonometric functions, exponentials and logarithms, error functions, Bessel functions, etc. Mathematicians in the 18th and 19th centuries tasked themselves with inferring equations admitting analytic closed form solutions which can be expressed in terms of special functions. Jacobi said [24]:

The main difficulty in integrating a given differential equation lies in introducing convenient variables, which there is no rule for finding. Therefore we must travel the reverse path and after finding some notable substitution, look for problems to which it can be successfully applied.

Jacobi was perfectly legitimate in taking this position in 1847. Acknowledging that only hypergeometric functions were effectively computable in the sense that one could obtain their value for any choice of parameters, the mathematician Jacobi tasks himself to find models which, by means of a change of variables, reduce the calculation to such hypergeometric functions. The fact that Jacobis work and that of his colleagues was relevant is linked to their ability to conform to the prevailing technology at their time. However, now technology has changed and, to follow in their footsteps, we need to reconsider those technical issues in a new light.

18th and 19th century Mathematics was largely aimed at physics and engineering applications. Mathematical Finance has developed more recently. Analytically solvable models being in limited supply, derivative pricing models were built around equations that were first discovered in a physics context as solvable in terms of special functions. The joint demands posed to hardware manufacturers by physics and Finance clients motivated the introduction of features aimed at optimising the evaluation of special functions this primarily means double-precision support. The valuation of special functions is often delicate, and accuracy of results tends to be fragile with respect to floating point errors.

Innovations in computing technology in the 1970s motivated the consideration of new classes of tractable equations. Early architectures were characterized by fairly inefficient ALUs (arithmetic logic units). Moving data from memory to registers was relatively quick compared with the time required to execute arithmetic on it. However, memory was a very costly and scarce resource. This technology environment motivated the introduction of a new class of algorithms, leveraging on the ability to apply a sparse matrix stored in compact format to a vector. Combined with growing educational programmes, the new ability gave rise to numerical analysis as an academic discipline. Once again, the pioneering applications of the new technology were in physics and engineering and were afterwards adapted to Finance by the creation of suitable model classes.

Coincidentally, numerical analysis as a discipline developed a dependence on double precision arithmetic. To solve an equation robustly with strongly stable methods, one needs to discretise

time very finely, respecting the so-called CFL condition, from the names of Courant, Friedrichs and Lewy. As many textbooks of numerical analysis explain, the CFL condition was considered impractical to the point of being called a curse. Numerical analysis as a discipline thus developed around the task of avoiding the CFL condition by constructions of discretisation schemes that would enjoy some form of numerical stability, notwithstanding the choice of a time step longer than the Courant-bound. These efforts gave rise to a theory for so-called unconditionally stable methods. Unconditionally means that the CFL condition is not postulated. What the nomenclature does not reflect though, is that the methods are only marginally stable, not strongly stable. In particular, high-frequency noise compounds rapidly, especially near singularities where the solutions show high convexity. Unfortunately, high-convexity regions are often the most interesting ones for Finance applications. They occur, for instance, near the strike of a call option or near the barrier of a barrier option, where risk management is particularly delicate. To stabilise marginally stable [or nearly unstable] methods, these mainstream algorithms were conceived specifically for double-precision ALUs.

It was not until the advent of GPU computing based on SIMD microprocessors that entirely new kinds of considerations prevailed in microchip design. GPU stands for graphic processing unit, signalling that graphics rendering was initially intended as the primary objective. As the human eye is not able to perceive differences in colour palettes defined with 32-bit representations as opposed to 64-bit, single floating point arithmetic was implemented as the favourite representation on SIMD microprocessors.

GPUs also introduce a new element in the landscape: they come with abundant memory and are capable of multiplying matrices very effectively. Processing speed for matrix multiplication in single precision is in the teraflop order of magnitude, an unprecedented level of performance in a shared memory architecture. This ability opens the door to another acronym: GPGPU, which stands for general purpose GPU, a GPU used for tasks other than rendering graphics.

The invention of GPUs was not prompted by mathematicians, although it could have been. In Fourier's publications two centuries ago, he clearly indicates his goal: to evaluate the exponential of a Laplace operator. This is the operation required to evaluate a transition probability kernel or pricing kernel. Matrix exponentiation can be achieved once one has access to high performance matrix multiplication engines. A very efficient algorithm to reduce the calculation of exponentials to a sequence of multiplications was already known to the Greeks and is called fast exponentiation. Although Fourier knew what he wanted, he didn't have the means to store and directly multiply matrices. He thus invented Fourier series, a tool that only works for a restricted class of matrices characterised by the property of translation invariance and known as Toeplitz matrices. He restricted himself to dynamic specifications of this particular form and envisaged a clever use of trigonometric functions to reduce the calculation of matrix exponentials to power series expansions. In turn, power series expansions are reduced to lookups of tables

prepared by human computers, the technology of his time.

GPUs were devised for the requirements of the games market. Now that this technology is available however, we unexpectedly have a game-changer in Mathematics to factor into the formalism. Perhaps for the first time in history, Mathematics is exposed to an exogenous and unexpected technology shock.

The key to creating a mathematical framework for probability theory which is suitable to the current hardware environment is algebraisation. Algebraisation of geometry was immensely successful as it gave rise to calculus. With probability, algebraisation is also possible as long as one reduces all calculations to matrix manipulations. Fortunately, this is not an entirely unexplored direction. Quantum physics is based on a form of probability theory that is different from the one required in Finance. Of course, the most radical differences are at the level of interpretation. But there are also deep similarities at the algebraic level. Under some circumstances, one can map a quantum dynamics specification into a probabilistic process by multiplying the time coordinate by an imaginary unit. This way, the Schrodinger equations for a free quantum particle are reduced to the Bachelier option pricing model. Complex numbers are an eminently algebraic concept which proved very useful in all sorts of engineering applications. The formalism of quantum physics is largely algebraic and deals with problems which are orders of magnitude more complex and challenging mathematically than any problem in the theory of stochastic processes which is relevant to mathematical Finance. Quantum physics manages the complexity by relying on algebraic means, i.e. operator methods.

Even without accounting for the role of modern technology, the complexity gap between quantum physics and mathematical Finance is by itself a good indication that algebraisation of probability can redefine the decidability boundary between answerable and unanswerable questions. Curiously enough, physicists who entered Finance in the early 1990s did not push this direction further and instead were channelled into the standardised probabilistic formalism of Finance.

Operator methods from quantum physics can be adapted to mathematical Finance, as long as the language of probability is properly algebraised. Stochastic calculus is based on calculations on bridges, whereby one considers a process with an initial condition, and looks at all paths, ending up with a given value within a certain time frame. The key is to observe that such bridge-conditional expectations can be expressed as matrices for which two indices are indicators of the initial condition and the final state variable. Further, these matrices can be evaluated by fast exponentiation on GPUs. Having realised this, stochastic calculus can be stripped of the layers of formalism it is coated with and can be reduced to a branch of linear algebra.

Algebraisation of probability was, however, not pushed as a research direction until recently, and it is still considered unorthodox. Objects of mathematical discourse in orthodox probability are given by infinitary symbols without direct hardware representations as datasets. The

significance of these symbols is perceived only by biological human brains using analogies with everyday experience and a form of infinitary mathematical logic fraught with paradoxes. The grounding of probability theory into human intuition is so deep that manipulations of probabilistically meaningful objects into ones not admitting a direct interpretation is not considered good style, or even acceptable for publication in academic journals.

Finite datasets represented as matrices naturally admit a large class of algebraic morphisms without direct probabilistic interpretation. It is not only natural, but practically very useful to apply general algebraic transformations. Even if intermediate steps involve complex numbers without direct interpretation, the ontological relevance is of course not lost, as long as the arithmetic deductions are logically correct and numerically efficient. In the context of geometry and calculus, engineering students are now educated to perceive this statement as a triviality and freely use complex numbers. Curiously, in the context of probability theory, this is still considered as an unorthodox innovation. Modern technology is, however, bound to change the situation, as its effective use depends on restructuring the entire mathematical framework of probability and Finance around a key set of powerful algebraic manipulations.

6 Modelling

Generality and modelling flexibility are of paramount importance to achieve economically realistic representations of the global financial system.

The extensions to the Black-Scholes model that are currently used as part of generally accepted valuation standards are either (i) analytically solvable in closed form in terms of special functions, or (ii) admit asymptotic expansions, or else (iii) can be calibrated against options without having to be solved explicitly. Many models of the first kind were built by reinterpreting physics models of the type Jacobi refers to in the aforementioned quotation. For instance, the Cox-Ingersol-Ross (CIR) interest rate model [12] is obtained by re-interpreting heat flow equations with radial symmetry in two dimensions. Affine models are a class pursued by Duffie, Pan and Singleton [16] on the basis that if the characteristic functions of certain stochastic integrals have a particular analytical structure, then a range of analytical tools such as Fourier transforms can be deployed. The models admitting asymptotic expansions, such as Hagan's SABR [20], are typically obtained through small volatility expansions reminiscent of the quasi-classical expansion methods in quantum physics and optics. The third category comprises models such as Dupire's local volatility [13]. These are models characterized by a state dependent local volatility function. In theory, they admit dynamic replication, the local volatility can be inferred directly from option data and in the asymptotic limit of short maturities they can reproduce the effects of any (possibly multi-factor) diffusion model. In practice, local volatility dynamics are quite unrealistic over long time horizons as the volatility is locked to state variables and does not relax

as it should.

Analytic solvability, or existence of asymptotic expansions or other properties that facilitate calibration or estimation, pose severe constraints on model flexibility. This leads to the inability to calibrate consistently against the entire spectrum of liquid securities within the horizon of price discovery.

Current computing technologies enable one to take economic realism to a higher level. To achieve realistic modelling, one has to combine several of the effects traditionally included in separate models and unify pricing and historical estimation methodologies. In the case of equity and foreign exchange processes for instance, it is desirable to have local volatility and a substantial diffusion component in the process, along with small frequent jumps and a form of stochastic volatility, whereby volatility increases in response to the occurrence of occasional large-size jumps. In the case of a short-term interest rate model, one would also want local volatility and a stochastic drift correlated to jumps in the short rate in order to model changing monetary policy. Commodity processes require seasonality patterns and, depending on the assets, a special form of transient behaviour and regime switching.

Realistic modelling is necessary in order to achieve global fitting to a broad array of calibration instruments. However, realism is not compatible with analytic tractability of any of the types mentioned beforehand. Hence, one needs to calibrate the brute force way, by actually pricing a large basket with a generic model. This is an operation that in a real-life situation can take resources in the multi-exaflop range for a complete set of risk factors. (One exaflop is one thousand petaflops, one milion teraflops and one bilion gigaflops. As a comparison, the single-core performance of a standard PC is in the one gigaflop range for efficient algorithms). The fact that this strategy is technologically possible with low-cost, mass-produced equipment should be recognised as a game-changer.

At the moment, analytic solutions come first, and engineering implementations are hardwired around them. With the proliferation of models, virtually hundreds of different and separate engineering solutions are created. Organizations are then molded around this technology structure, creating insulated cells that are limited in their ability to pass information to each other. Once analytic solvability is no longer a requirement, the very same base software libraries can be designed to support valuation and risk management functions across all asset classes including interest rates, foreign exchange rates, credit-equity asset price processes and commodity prices. Thus organizations become porous, risk management horizontal, and it becomes possible to aggregate and transfer information at a global level.

Theoretical principles for statistical estimation should also be updated. Global valuation is about choosing a unique model for each risk factor and a unique correlation scheme. Hence, the estimation procedure needs to confront model parameter choices against option data of a variety of different types, along with historical data.

For instance, it is plainly wrong to keep equity derivative model estimation separate from credit model calibration. If the underlying reference name is the same, modelling can, and should be, coalesced into a single defaultable credit-equity model for both sides of the market. This task is feasible as long as a sufficient degree of modelling flexibility is allowed.

Had credit and equity markets been tied together at the modelling level, the crisis would have unfolded differently. Instead, equity models are radically different from credit models. Equity models focus on volatility dynamics, while credit models focus on default arrival times, ignoring volatility altogether. If in 2005-2006 the majority of market participants had been using volatility-sensitive credit models, they would have realized that the extreme volatility observed through the spread-tightening period was inconsistent with the low level of credit spreads and the low level of equity volatility. As the credit crisis unfolded and, notwithstanding the unprecedented spread widening after the Bear-Stearns crisis, equity markets remained unusually tranquil. It was only several months into the crisis that equity markets picked up volatility and market participants ran to the exit all at once. Not being able to cover their short exposure fast enough, the rush exacerbated the fall in equity markets. Model inconsistencies give rise to uncontrollable bubbles and price gyrations, impairing the ability to hedge dynamically.

To achieve global valuation, one needs to reconcile as much data as possible within a unified modelling specification. Options prices are a good place to start, as they capture forward-looking market expectations on price dynamics. This information is crucial to use during transient times such as a post-crisis period, where the immediate history contains gyrations of unusual severity and singular features.

The historical process does not necessarily have to obey the same dynamic specification of the process used for valuation, because the two are distinguished through the price of risk adjustments. One often hears statements aimed at justifying freedom in using whatever pricing model fits the pricing data, coupled with a specification altogether different for the historical process for risk management on the basis that the two need not coincide. We maintain that this is not acceptable, and that the price of risk should be directly and consistently modelled. For consistent modelling, it is necessary that the scenarios which are possible under the simulation measure are also possible under the pricing measure. Using a more technical vocabulary, Radon-Nikodym derivatives between the two measures need to be regular, non-zero and not infinite. They also need to be estimated reasonably and entail economically understandable information. A good procedure is to first start from the determination of a global pricing measure to satisfy the Fundamental Theorem of Finance, and then correct this measure to incorporate systematic price of risk biases and reconcile with historical information.

7 System architecture

System architecture for valuation and risk management is intimately tied up with the choice of base algorithms and the distinction of roles in the organization between quantitative analysts and IT infrastructure developers.

In the prevailing organization of labor, quantitative analysts are modelers primarily dedicated to inventing or implementing variations over the Black-Scholes theme, consisting of models with some degree of analytic tractability to be used to rapidly calibrate against selected targets. Analytic closed-form solutions are utilized for the purpose of risk-factor scenario generation and, whenever possible, for valuation. Valuation is also accomplished on the basis of numerical methods which use memory sparingly, such as semi-implicit backward induction schemes. These methods reduce computational requirements at the cost of introducing marginal numerical instabilities which require the use of double-precision floating point arithmetic combined with smoothing schemes.

According to common practice, models developed by quantitative analysts are often compiled and handed over in binary format to IT infrastructure developers, who are tasked to maintain the middleware layers managing the distribution of analytic functionality over grids with (typically) tens of thousands of compute nodes. In a standard application, when valuing a portfolio, individual instrument models are invoked by assigning jobs to a queue which is published on a shared drive. Jobs are executed concurrently on a number of compute nodes that query the queue for a task and deposit the resulting price and risk sensitivities on the same shared resource.

Within the standard implementations, concurrency is achieved by means of the middleware layer that distributes tasks to independent processes which otherwise do not communicate with each other directly in memory. Multi-threading techniques are not used for valuation modules. Also, pricing models typically have very modest memory requirements, well below the 2GB mark. Hence 32-bit environments are still the norm in data centers.

Recent system architectures built with low-cost, volume components allow for up to 48 MIMD cores on the same board, along with half a terabyte in physical memory. These specs have been increasing at a steady pace, doubling up on an 18-month cycle, a process one can expect to continue over the next decade. To benefit from these configurations, a minimum requirement is to use 64-bit technology and heavily multi-threaded applications designed for using shared memory. Porting non-thread-safe 32-bit code to this new technology is likely to be far more costly than a ground-up rewrite based on new principles.

On the software side, current configurations see a proliferation of dozens of different model concepts. Each model is wrapped around some specific analytic approximation or solution scheme which affects the entire system infrastructure built around the model, from data provision to optimization and usage modalities. Resources to maintain systems are thus allocated across a large number of specific application domains, each wired in a way that is specific to its own

analytic prowess.

A major benefit of eliminating the straightjacket of analytic tractability and insisting on flexible model specifications, is that the same base libraries can be recycled across all application domains. One can institute a Chinese wall between system developers and modelers, so that system design is entirely independent of modelling decisions. A benefit is that modelers will be free to experiment with a variety of specifications. Another benefit is that new asset classes can be covered while reusing existing engines for valuation, calibration and simulation. A third is that whenever the base libraries are enhanced and optimized, the benefits are realized across all application domains.

At the lowest level of system architecture design, one should identify a very limited number of core interfaces for basic routines which capture the computational bottlenecks. Once these routines are standardized, it is worth dedicating specialized resources to their optimization in hand-written assembler language, tailoring them to the microprocessors in use and establishing a hardware abstraction layer to insulate the key functionality.

At mid levels in the architecture, one needs to orchestrate the basic Finance algorithms in an asset-independent and model-independent fashion, i.e. (i) forward induction for term structure fitting, (ii) possibly multi-pass backward induction to price single-factor assets, (iii) optimization algorithms for calibration and (iv) Monte Carlo scenario generation algorithms for multi-factor assets and portfolio risk calculations.

Mid-level functionality executes on dedicated heterogeneous CPCs, (Central Processing Clusters). A CPC is based on a combination of high-density CPU and high-density GPU boards. Cluster management is ideally localized using current technologies. Being able to transfer objects asynchronously across application domains and operate on them is a key ability that goes beyond traditional message passing paradigms, and is essential to manage the level of complexity required for portfolio analysis.

CPCs also need to be endowed with a local database to store pre-processed data such as model specifications for calibrated models and the corresponding transition probability matrices. The pre-processed data will also contain valuation sheets consisting of tables of state-dependent future values of single-factor sub-portfolios, which can be evaluated after a calibration process by running backward induction algorithms. Valuation sheets contain aggregate pricing information, which greatly facilitates dynamic portfolio simulation, as discussed below.

Top-level functionality mediated by middleware focuses on user interaction, booking and risk reporting at the aggregate sub-portfolio level. Unlike current systems that treat instruments separately from one another, concurrent pricing and the use of globally defined models across all instruments allows aggregation of individual exposures at the middleware level and provides holistic cross-sectional views on information essential to risk management. This deviates from current middleware designs which are based on an instrument-by-instrument view of portfolios

and work around system latency with local pre-processing loops. In an architecture based on CPCs, the middleware designs have to reflect the changed landscape by handling aggregate data instead, while pre-processing is centralized and broadcasted by local data mirroring.

8 The decidability boundary and information processing

The global financial system is highly complex. To navigate through it, market participants need to ask questions that require complex analysis. Honest answers combine theoretical understanding and computational processing ability. However, not all questions can be answered if one insists on an honest answer. Depending on the technology available and the efficiency of the mathematical framework deployed upon it, there is a decidability boundary separating questions that are answerable, from questions that are not. This is a meta-theoretical and meta-technological question whose analysis heavily depends on the technology environment and the adopted theoretical framework. It is crucial to understand where the decidability boundary lies.

The alternatives to understanding the decidability boundary are either to (i) use analogy to extrapolate answers that give a false sense of confidence to decision-makers or (ii) declare that all theory is of limited use anyway and accept periodic catastrophic events. Regulatory frameworks tend to lean towards the first route, while Talebs black swans theory hinges on the second strategy. There are also intermediate approaches, where theoretically inconsistent modelling is used with fine-print disclaimers saying that they actually do not work. What we propose is to go down a third route and determine which questions are honestly decidable within a theoretically consistent framework and within a given technology environment.

Decidability is one of the most fundamental concepts, but also one of the most deeply misunderstood. A stream of philosophical discussions in the 20th century was prompted by Godels finding that within infinitary logic there are undecidable propositions. This result is framed in a very Platonist setting, with axiomatic systems enabling the formation of propositions which are true, false or undecidable. The truth value descends immediately from the axioms and has eternal valence. Godels examples were very convoluted variations on Cantor argument about the non-denumerability of real numbers, but a large number of concrete examples have been found more recently. These examples typically involve sequences of finite combinatorial problems labelled by an integer N and of growing complexity as N tends to infinite. Because of the inner workings of infinitary mathematical logic, if the growth rate of complexity outpaces the growth rate of all arithmetic functions that can be explicitly represented, then the truth value of these propositions for all values of N is undecidable. Notice here that the propositions which are undecidable do not make statements over finite sets. Each individual proposition for a fixed N involves only a finite number of cases, and is thus decidable within this logic. What is undecidable is the question of whether all propositions for all values of N are true or not.

With analogical reasoning, undecidability results, the quantum mechanics uncertainty principle and a good dose of impostures intellectuelles [31], one concludes that there's a lot we do not know and we will never know.

What is missing from this debate is the role of technology in determining the decidability boundary. To be relevant and useful, the mathematical question of decidability needs to be posed within a technology context. A travelling salesman problem with 50 cities is decidable on modern computers, but was not decidable 200 years ago. The same problem with one million cities is not decidable today. In a Finance context, there are questions that are decidable today and were not decidable two decades ago. It is important to understand how the boundary has shifted.

A related concept which has historic relevance, but is often misunderstood, is that of information content. Within 20th century probability theory, information is revealed through filtered probability spaces. One can ask whether a CDO is hedgeable in the filtration engendered by CDSs, and one may even come out with an answer subject to technical hypothesis. But from an operational viewpoint, information cannot be abstracted from the technological framework that actually processes it in a format suitable for analysis and supporting operative decisions.

Risk management has traditionally been predicated on the VaR methodology. VaR combines the local valuation methodology with a definition of risk measure defined as a percentile of the distribution of gains and losses over a daily time frame. Oddly enough, VaR attracted much academic criticism because of the perceived inconsistencies of the risk measure. Percentile levels are indeed problematic when applied to multimodal distributions with bumps in the loss tails, as it may well be that one perceives the risk of a portfolio to be greater than the combined risk of two of its sub-portfolios. This reflects an inconsistency from a regulatory standpoint, as it would induce users to split portfolios as opposed to aggregate them, while in principle aggregation and diversification should always favor the ability to manage risk. But in reality, loss distributions are often uni-modal and this problem does not arise. When the problem does arise, suitable averages of VaR measures eliminate the pathology.

Unfortunately, the most significant problems with the VaR methodology are not in the choice of risk measure. If that were the only issue, the fix would be quick. More serious issues include:

- (i) Lack of theoretical consistency deriving from the neglect of the Fundamental Theorem of Finance in favor of local valuation methodologies
- (ii) Limitation to a fixed short-term time horizon, as opposed to a dynamic simulation over the portfolio lifetime
- (iii) Confusion between endogenous and exogenous model sensitivities
- (iv) Use of a very limited number of historical shocks

We have already commented on the first key point (i), whose resolution hinges on the ability to calibrate models globally.

Simulations over a fixed 24-hour or ten-day time horizon are insufficient. The Black-Scholes model promises that dynamic hedging is possible based on single period information, but extrapolating this conclusion by analogy to complex real-life portfolios is unrealistic. Prudent risk management is only possible by understanding the dynamics of portfolio valuations over time frames comparable to their own lifetime. Hedging positions should be classified as a function of their renewal schedule. Some are short term and other are semi-static and to be held over weeks or months. To even pose the question and find optimal hedging strategies, one needs to be able to accurately simulate a consistent measure dynamically over timescales commensurable with portfolio lifetimes.

From a regulatory standpoint, there are two possibilities. Either one applies a regime of mark-to-market on the premise that risks are hedged correctly, i.e. dynamically, in which case capital charges are proportional to the risk of hedge slippage. Or one marks-to-risk in a similar fashion to the insurance industry and capital is allocated on the premise that risks cannot be hedged but only diversified. The two approaches lead to entirely different capital requirements.

Regarding point (iii), exogenous risks linked to changes in model parameters lead to inconsistencies in the process definition and cannot be considered hedgeable on par with endogenous risks. There are two possibilities: either models endogenise risk factors and be consistently used for simulation and forecasting; or else exogenous risks should be treated as un-hedgeable and capital reserves should be assigned accordingly.

The widespread use of 250-500 historical shocks to carry out a portfolio simulation is particularly weak. Historical shocks are taken out of context and applied to both endogenous and exogenous sensitivities, so they are inconsistent in uncontrolled and uncontrollable ways. Additionally, the size of the scenario set is too small for the purpose of hedge optimization. We estimate that at least 50,000 to one million fully dynamic scenarios are needed for a thorough optimization analysis. Furthermore, the simulation process used needs to be consistent with the valuation process for derivative positions.

Finally, there is the issue of whether one is carrying out VaR analysis only as a pro-forma requirement to assess risk exposure and capital requirement, or whether this is a reasonable basis for hedging. If the quality of data is poor, it may not be useful for hedging purposes. Hedging is thus carried out heuristically and not systematically in a process disjoint from risk assessment, thus defeating the purpose of the exercise.

In other words, the question of what is the optimal hedging strategy is not decidable on the basis of the standard VaR scenario analysis because this methodology does not process the available information efficiently. To make the question decidable and the information useful, the analytical framework needs to change radically.

To summarize, we propose to:

- (i) Insist on theoretical consistency based on the Fundamental Theorem of Finance. Unless a framework is demonstrably consistent, it should not be used for the purpose of assessing capital adequacy on a mark-to-market basis, but instead should use a more prudent insurance style, mark-to-risk analysis.
- (ii) Recognizing that hedging strategies are articulated across an array of positions extending over multiple time horizons, simulations need to be performed dynamically and consistently over the lifetime of the portfolio.
- (iii) Only endogenous risks can be treated as hedgeable. Model risk should be recognized as unhedgeable and provisioned separately.
- (iv) One needs to carry out simulations involving a number of dynamic scenarios in the 50,000 to one million range, over time horizons as long as the lifetime of the portfolio, to achieve sufficient and meaningful resolution.

9 Organisation of banking operations

One may take the view that organizations exist and are managed independently of technology used to implement basic functions. But there is an alternative view: organizations grow out of the need to optimize the use of a given underlying technology. Evidently, the choice between local and global valuation schemes has far-reaching impacts on the organization of banking operations.

Within the standard local valuation approach to financial engineering, business units are separate and naturally aggregate around a model implementation. Each unit is responsible for data sourcing, model calibration, portfolio mark-to-market, risk assessment, hedging and, of course, trading. Communication between business units involves sharing metrics for risk which are abstracted as sensitivities to a standard risk-factor set. There is no consistency in the algorithms used to determine these sensitivities between business units. Even within the same business unit, often each instrument is priced with a different model specification. Using two models of the same class with different choices of parameters is arguably less of an inconsistency as opposed to using models from two entirely separate classes.

Model risk is highly correlated for a simple reason: the liberty of choosing inconsistent model specifications on an instrument-by-instrument basis enables trading strategies that create and burst local bubbles. Bubble bursting in a mark-to-market environment is highly correlated, as realized losses of one market participant induces unwinds in another, which in turn induces price shifts and causes other bubbles to burst. As in a nuclear chain reaction, a sufficiently high density of local bubbles can give rise to a nuclear explosion. Arnolds predicted Chernobyl-style

event caused by inappropriate mathematical modelling has again proved to be prophetic and tragically accurate.

In addition to model risk, there is also an impediment to information transfer across business units. If information about the future evolution of asset values and correlations is not reflected in a unified modelling exercise, it cannot be transferred among units. If information is not common, coordinated and concerted actions cannot take place and business optimization stops at the lowest level.

In a global valuation setting, two functions are centralized: engineering of valuation and risk management systems and risk-factor modelling. Once analytic tractability is no longer a business priority, or even considered desirable, the same engine library and computing infrastructure can be used for all valuation and risk management tasks across all asset classes. System architecture design can thus be decoupled from the modelling exercise. Engineers specialism in one and economists in the latter. The professional figure of a quant working for a trader, developing their model, implementing it in code, providing it with data and delivering it to a general-purpose computing grid, disappears. Traditional grid computing itself disappears and is replaced by a single CPC, a general-purpose cross-asset central processing cluster. Local users can still rely on a grid infrastructure but nodes of this grid are designed differently: they embed a database mirroring device that queries preprocessed model information from the CPC and executes user-specific risk analytic functions on that basis. These nodes will thus be configured as high-density CPU boards with as many MIMD cores as one can pack within shared memory architecture to enable multi-factor dynamic simulations.

Without a centralized infrastructure, information gathering is mediated by a myriad of conventional mapping procedures, often designed on analogy and common sense, but without the rigor of theoretical consistency and manageable model risk. In a global valuation setting, model calibration is centralized and detached from trading functions. Ownership of process definitions, i.e. probability distributions, and book mark-to-market functions are concentrated in a single team with a global mandate. By removing barriers between models, one transforms banking organizations into open systems, whereby information flows horizontally and management and control becomes possible. Only a centralized valuation and data analysis infrastructure makes questions relevant to firm-wide risk management decidable.

10 The proper design of an investment bank

If the organizational design of investment banks were prevalent in the automotive industry, we would observe each automotive model being independently manufactured from scratch. Each product line for each type of car would assemble its own headlights, windscreen wipers, electronics and engine, and then subsequently have them approved by the firms risk management

and senior management teams.

Because investment banking is not dealing with physical products in the way that other industries do, it has so far not needed to apply the same level of discipline and rigor in designing process flows. Mainstream management thinking opines that capital for an organization is a raw material for production, and can therefore be thought of in the same way.

Any industry can be decomposed into a chain of competencies and capabilities, and these are defined in terms of the value that they create and the effort that a competitor would need to expend in order to replicate it.

A competency is hard for competitors to replicate because it requires specific knowledge that can be protected, and typically takes the form of a body of expertise and intellectual property. A capability is hard to replicate because it takes time to establish and often takes the form of infrastructure or social networks.

The risk management function of a bank can thus be categorized as a competency, and since risk is more fully defined as being risk to the balance sheet, the risk management function logically ought to be positioned as the first step in the production line that processes the raw material of capital.

But this cannot be done with a common approach to estimating and forecasting risk it requires global valuation.

The next step in the investment bank production line is the productization of this risk-treated capital the function of the front office, and this also meets the definition of a competency. This competency is not about assessing the risk of the capital deployed, but rather in understanding the market in which it may be consumed. To do this, the product designers have to understand the market participants, the demands that they make and their purpose for the products. In this way, investment banking becomes socially useful.

The third and final value-generating aspect of an investment bank is its distribution franchise the business lines, the markets, the venues and the services in which it is active. This is a capability, rather than a competency.

The universal consequence of local valuation is that organizations are poorly understood, highly intertwined with vast dependency graphs and very cumbersome and sensitive to change. The back-to-front organizational design that is locked in to current practitioners is prohibiting their progress.

Global valuation presents an opportunity for investment banks to redesign themselves and streamline their organization. As a minimum, this design will reap organization cost savings and improve risk management and controls.

But the benefits mentioned so far only consider the current commercial activities of these organizations. One redesigned as a well-delineated production line process will be more scalable, flexible and adaptable and can bring new products and services to market more quickly. It will

be able to innovate and new businesses can leapfrog competitors who do not adopt and fully exploit this technology.

11 Conclusions

In this article we discuss our view of the future of risk management. The combination between the openness to innovation in the aftermath of the credit crisis and the game-changing technologies that are surfacing, prompts one to reconsider existing practices and consider redesigning valuation and risk management infrastructures.

If there is one thing that the crisis taught us, it is that patched-up solutions do not work. Incremental enhancements will never enable decidable questions that are vital for firm-wide risk management and that require honest, theoretically consistent and sound answers. We need to acquire the ability to process asset price information in a global data mining effort and to render a realistic and consistent representation of the financial universe. Question decidability and information flow depend on technology and need to be qualitatively improved throughout the financial system at large. This involves rethinking mathematical and numerical frameworks, along with business and management practices.

The momentum is there and we have a unique opportunity to pursue it. Complexity in the global financial system has reached a level that a ground-up rethink of industry practices is the only viable path. The good news is that this path to innovation is now technologically possible.

References

- [1] C. Albanese, T. Bellaj, G. Gimonet, and G. Pietronero, *Coherent Global Market Simulations for Counterparty Credit Risk*, (2010).
- [2] C. Albanese, G. Gimonet, and S. White, *An Introduction to Global Valuation*, Risk Magazine, May 2010 (2010).
- [3] V.I. Arnold, Notices of the American Mathematical Society **44** (1996).
- [4] L. Bachelier, *Thorie de la speculation*, Annales Scientifiques de l'cole Normale Suprieure **3** (1900), 2186.
- [5] Basle Committee on Banking Supervision, *Amendment to the capital accord to Incorporate Market Risks*, Tech. report, Bank for International Settlements, 1996.
- [6] F. Black and M. Scholes, *The pricing of options and corporate liabilities*, Journal of Political Economy **81** (1973), 637–59.

- [7] D. Brigo and A. Capponi, *Bilateral Counterparty Risk Valuation with Stochastic Dynamical Models and Application to Credit Default Swaps*, arXiv:0812.3705v4 [q-fin.RM] 18 Nov 2009 (2010).
- [8] D. Brigo and M. Masetti, *A Formula for Interest Rate Swap Valuation under Counterparty Risk in Presence of Netting Agreements*, Counterparty Credit Risk Modelling: Risk Management, Pricing and Regulation **London : Risk Books, 2005** (2010).
- [9] D. Brigo and A. Pallavicini, *Counterparty Risk and Contingent CDS under Correlation*, Risk **February** (2010).
- [10] P. J. Cohen, *Set theory and the continuum hypothesis.*, Addison-Wesley, 1966.
- [11] Ian A. Cooper and Antonio S. Mello, *The Default Risk of Swaps*, Journal of Finance **46** (1991).
- [12] J.C. Cox, J.E. Ingersoll, and S.A. Ross, *A theory of the term structure of interest rates*, Econometrica **53** (1985), 385-407.
- [13] ———, *Pricing with a smile*, Risk Magazine (1994), 18–20.
- [14] B. de Finetti, *Sul Significato Soggettivo della Probabilità.*, (1931), 298–329.
- [15] ———, *Sulla preferibilità.*, (1952), 685–709.
- [16] D. Duffie, J. Pan, and K. Singleton, *Transform analysis and asset pricing for affine jump-diffusions*, Econometrica **68** (2000), 1343–1376.
- [17] A. Einstein, *Ueber die von der molekularkinetischen Theorie der Waerme geforderte Bewegung von in ruhenden Fluessigkeiten suspendierten Teilchen.*, Annalen der Physik **17** (1905), 549-560.
- [18] J. Farkas, *Theorie der einfachen ungleichungen*, (1902), 1–27.
- [19] J.B.J. Fourier, *Solution d'une question particuliere du calcul des inegalites.*, (1826), 99–100.
- [20] P. Hagan, D. Kumar, A. Lesniewski, and D. Woodward, *Managing Smile Risk*, Wilmott Magazine **September** (2002), 84–108.
- [21] Patrick S. Hagan, *Adjusters: Turning good prices into great prices*, **1** (2005).
- [22] G. Haug and N. Taleb, *Why we have never used the black-scholes-merton option pricing formula*, (2010).
- [23] J. Hull, *Options, futures, and other derivatives*, Pearson, Upper Saddle River, New Jersey, USA, 1988-2008.

- [24] C. Jacobi, *Vorlesungen ueber dynamik*, Encyclopedia Britannica Online Edition, 2010, 1847.
- [25] Robert A. Jarrow and Fan Yu, *Counterparty risk and the pricing of defaultable securities*, Journal of Finance **LVI** (2001).
- [26] A. N. Kolmogorov, *Grundlagen der wahrscheinlichkeitsrechnung*, Grattan-Guinness, ed, Landmarks in Western Mathematics: Case Studies 1640-1940, 2005 (1933), 960–969.
- [27] A. Lipton and Artur Sepp, *Credit Value Adjustment for Credit Default Swaps via the Structural Default Model*, The Journal of Credit Risk **5** (2009).
- [28] R. Merton, *Theory of Rational Option Pricing*, Bell Journal of Economics and Management Science **4** (1973), 141–183.
- [29] J.P. Morgan/Reuters, *Riskmetrics—technical document*, 4th edition ed., J.P. Morgan, 1996.
- [30] M. Smoluchowski, *Sur le chemin moyen parcouru par les molecules dun gaz et sur son rapport avec la theorie de la diffusion*, Bulletin International de l'Academie des Sciences de Cracovie (1906), 202213.
- [31] A. Sokal and J. Bricmont, *Impostures intellectuelles*, Odile Jacob, Livre de Poche, dition, 1997.
- [32] R. M. Solovay, *A model of set-theory in which every set of reals is lebesgue measurable*, Annals of Mathematics **92** (1970), 156.
- [33] H. Stein and Kin Pong Lee, *Counterparty Valuation Adjustments*, The Handbook of Credit Derivatives, Bielecki, Tomasz; Damiano Brigo, and Frederic Patras, Eds. (2009).
- [34] Superderivatives, *Method and system for pricing options*, US Patent 7315838 (2008).
- [35] L. Wittgenstein, *Tractatus logico-philosophicus*, Routledge and Kegan Paul, 1961; translated by D.F. Pears and B.F. McGuinness., 1922.
- [36] ———, *Wittgenstein's lectures on the foundations of mathematics, diamond, cora, (ed.)*, Cornell University Press, Ithaca, N.Y., 1976.