

Adjoint Algorithmic Differentiation Masterclass

Luca Capriotti

Quantitative Strategies - Credit Suisse
Department of Mathematics - University College London

London, 8 March 2017

The opinions and views expressed in this presentation are uniquely those of the author, and do not necessarily represent those of Credit Suisse Group.

Disclaimer

This material has been prepared by the Quantitative Strategy Group ("Quant Strategists") which is part of the institutional trading desk of Credit Suisse and/or its affiliates (collectively "CS"). Quant Strategists are not part of the Research Department and the written materials disseminated by Quant Strategists are not research reports. The views of CS' Quant Strategists may differ materially from the views of the Research Department and other divisions at CS. CS has a number of policies in place designed to ensure the independence of CS' Research Department from CS' Trading Desks, including policies relating to the front-running of research reports. These policies do not apply to the materials provided by Quant Strategists. This material may have previously been communicated to the CS trading desk or other CS clients. You should assume that the CS' trading desk makes markets and/or currently maintains positions in the securities (or related securities) or derivatives thereof that are the subject of these materials, and such positions may be inconsistent with the materials provided by Quant Strategists.

This material is being provided to you solely for information purposes, is intended for your use and does not constitute an offer or commitment, a solicitation of an offer or commitment, or any advice or recommendation to enter into a transaction (whether on the indicative terms or any other). This has been prepared based on assumptions and parameters determined in good faith. Calculations in these materials may reflect CS's theoretical (i.e., model) prices only and may not take into account other potentially significant factors. It is important that you (recipient) understand that those assumptions and parameters are not the only ones that might have reasonably been selected or that could apply in the preparation of these materials or in an assessment of any transaction. A variety of other assumptions or parameters, or other market factors or conditions, could result in different contemporaneous good faith analysis or assessment of a transaction. Past performance should not be taken as an indication or guarantee of future performance, and no warranty or representation, expressed or implied is made regarding future performance. Opinions and estimates may be changed without notice. The information set forth above has been obtained from or based upon sources believed by CS to be reliable, but CS does not represent or warrant its accuracy or completeness. This material does not purport to contain all of the information that an interested party may desire. In all cases, interested parties should conduct their own investigation and analysis of the data set forth in these materials. Each person receiving these materials should make an independent assessment of the merits of pursuing a transaction described in these materials and should consult their own professional advisors. CS may from time to time, participate or invest in other financing transactions with the issuers of securities referred to herein, perform services for or solicit business from such issuers, and/or have a position or effect transactions in the securities or derivatives thereof. The market value of any security may be affected by changes in economic, financial, and political factors (including, but not limited to, spot and forward interest and exchange rates), time to maturity, market conditions and volatility and the credit quality of any issuer or reference issuer. Any investor interested in purchasing a product should conduct its own investigation and analysis of the product and consult with its own professional advisors as to the risk involved in making such a purchase.

CS is not qualified to give tax or accounting advice. This document is not to be relied upon in substitution for the exercise of independent judgment and for consultation of an external tax or accounting advisor. CS may have issued other documents that are inconsistent with, and reach different conclusions from, the information presented in this document. Those documents may reflect different assumptions, views and analytical methods of the analysts who prepared them. This document may not be reproduced in whole or in part or made available without the written consent of CS. The distribution of this information may be restricted by local law or regulation in certain jurisdictions.

CS may provide various services to US municipal entities or obligated persons ("municipalities"), including suggesting individual transactions or trades and entering into such transactions. Any services CS provides to municipalities are not viewed as "advice" within the meaning of Section 975 of the Dodd-Frank Wall Street Reform and Consumer Protection Act. CS is providing any such services and related information solely on an arm's length basis and not as an advisor or fiduciary to the municipality. In connection with the provision of the any such services, there is no agreement, direct or indirect, between any municipality (including the officials, management, employees or agents thereof) and CS for CS to provide advice to the municipality. Municipalities should consult with their financial, accounting and legal advisors regarding any such services provided by CS. In addition, CS is not acting for direct or indirect compensation to solicit the municipality on behalf of an unaffiliated broker, dealer, municipal securities dealer, municipal advisor, or investment adviser for the purpose of obtaining or retaining an engagement by the municipality for or in connection with Municipal Financial Products, the issuance of municipal securities, or of an investment adviser to provide investment advisory services to or on behalf of the municipality.

Section 1

Outline

Computing Risk in Monte Carlo

Option Pricing Problems

Pathwise Derivative Method

Adjoint Algorithmic Differentiation (AAD)

Algebraic Adjoint Approaches

Algorithmic Adjoint Approaches (AAD)

AAD as a Software Design Principle

AAD and the Pathwise Derivative Method

First Applications

Case Study: Adjoint Greeks for the Libor Market Model

Correlation Greeks and Binning Techniques

Case Study: Correlation Greeks for Basket Default Contracts

Case Study: Real Time Counterparty Credit Risk Management

Application to Partial Differential Equations

Calibration and Implicit Function Theorem

Conclusions

Section 2

Computing Risk in Monte Carlo

Option Pricing Problems

- ▶ Option pricing problems can be typically formulated in terms of the calculation of expectation values of the form

$$V = \mathbb{E}_{\mathbb{Q}} \left[P(X(T_1), \dots, X(T_M)) \right].$$

- ▶ Here $X(t)$ is a N -dimensional vector and represents the value of a set of underlying market factors (e.g., stock prices, interest rates, foreign exchange pairs, etc.) at time t .
- ▶ $P(X(T_1), \dots, X(T_M))$ is the discounted payout function of the priced security, and depends in general on M observations of those factors.
- ▶ In the following, we will indicate the collection of such observations with a $d = N \times M$ dimensional state vector

$$X = (X(T_1), \dots, X(T_M))^T.$$

Monte Carlo Sampling of the Payoff Estimator

- ▶ The expectation value above can be estimated by means of Monte Carlo (MC) by sampling a number N_{MC} of random replicas of the underlying state vector $X[1], \dots, X[N_{MC}]$, sampled according to the distribution $\mathbb{Q}(X)$, and evaluating the payout $P(X)$ for each of them.
- ▶ This leads to the estimate of the option value V as

$$V \simeq \frac{1}{N_{MC}} \sum_{i_{MC}=1}^{N_{MC}} P(X[i_{MC}]),$$

with standard error $\Sigma/\sqrt{N_{MC}}$, where

$$\Sigma^2 = \mathbb{E}_{\mathbb{Q}}[P(X)^2] - \mathbb{E}_{\mathbb{Q}}[P(X)]^2$$

is the variance of the sampled payout.

Pathwise Derivative Method

- ▶ The Pathwise Derivative Method allows the calculation of the sensitivities of the option price V with respect to a set of N_θ parameters $\theta = (\theta_1, \dots, \theta_{N_\theta})$, with a single set of N_{MC} simulations.
- ▶ This can be achieved by noticing that, whenever the payout function is regular enough, e.g., Lipschitz-continuous, and under additional conditions that are often satisfied in financial pricing (see, e.g., [1]), one can write the sensitivity $\langle \bar{\theta}_k \rangle \equiv dV/d\theta_k$ as

$$\langle \bar{\theta}_k \rangle = \mathbb{E}_{\mathbb{Q}} \left[\frac{dP_\theta(X)}{d\theta_k} \right].$$

- ▶ In the context of MC simulations, this equation can be easily understood by thinking the random sampling of the state vector X as performed in terms of a mapping of the form, $X = X(\theta; Z)$, where Z is a random vector *independent* of θ . In fact, after this mapping, the expectation value $\mathbb{E}_{\mathbb{Q}}[\dots]$ can be expressed as an average over the probability distribution of Z , $\mathbb{Q}(Z)$, which is independent of θ .

Pathwise Derivative Method: Interpretation

- ▶ The calculation of $\langle \bar{\theta}_k \rangle$ can be performed by applying the chain rule, and averaging on each MC sample the so-called Pathwise Derivative Estimator

$$\bar{\theta}_k \equiv \frac{dP_\theta(X)}{d\theta_k} = \sum_{j=1}^d \frac{\partial P_\theta(X)}{\partial X_j} \times \frac{\partial X_j}{\partial \theta_k} + \frac{\partial P_\theta(X)}{\partial \theta_k}.$$

- ▶ The matrix of derivatives of each state variable, or *Tangent state vector*, is by definition given by

$$\frac{\partial X_j}{\partial \theta_k} = \lim_{\Delta \theta \rightarrow 0} \frac{X_j(\theta_1, \dots, \theta_k + \Delta \theta, \dots, \theta_{N_\theta}) - X_j(\theta)}{\Delta \theta}.$$

- ▶ This gives the intuitive interpretation of $\partial X_j / \partial \theta_k$ in terms of the difference between the sample of the j -th component of the state vector obtained after an infinitesimal 'bump' of the k -th parameter, $X_j(\theta_1, \dots, \theta_k + \Delta \theta, \dots, \theta_{N_\theta})$, and the base sample $X_j(\theta)$, both calculated on *the same random realization*.

Pathwise Derivative Method: Diffusions

- ▶ Consider the case for instance in which the state vector X is a path of a N -dimensional diffusive process,

$$dX(t) = \mu(X(t), t, \theta) dt + \sigma(X(t), t, \theta) dW_t,$$

with $X(t_0) = X_0$. Here the drift $\mu(X, t, \theta)$ and volatility $\sigma(X, t, \theta)$ are N -dimensional vectors and W_t is a N -dimensional Brownian motion with instantaneous correlation matrix $\rho(t)$ defined by $\rho(t) dt = \mathbb{E}_{\mathbb{Q}} [dW_t dW_t^T]$.

- ▶ The Pathwise Derivative Estimator may be rewritten as

$$\bar{\theta}_k = \sum_{l=1}^M \sum_{j=1}^N \frac{\partial P(X(T_1), \dots, X(T_M))}{\partial X_j(T_l)} \frac{\partial X_j(T_l)}{\partial \theta_k} + \frac{\partial P_{\theta}(X)}{\partial \theta_k}$$

where we have relabeled the d components of the state vector X grouping together different observations $X_j(T_1), \dots, X_j(T_M)$ of the same (j -th) asset.

Pathwise Derivative Method: Diffusions

- In particular, the components of the Tangent vector for the k -th sensitivity corresponding to observations at times (T_1, \dots, T_M) along the path of the j -th asset, say,

$$\Delta_{jk}(T_l) = \frac{\partial X_j(T_l)}{\partial \theta_k}$$

with $l = 1, \dots, M$, can be obtained by solving a stochastic differential equation

$$\begin{aligned} d\Delta_{jk}(t) = & \sum_{i=1}^N \left[\frac{\partial \mu_j(X(t), t; \theta)}{\partial X_i(t)} dt + \frac{\partial \sigma_j(X(t), t; \theta)}{\partial X_i(t)} dW_t^i \right] \Delta_{ik}(t) \\ & + \left[\frac{\partial \mu_j(X(t), t; \theta)}{\partial \theta_k} dt + \frac{\partial \sigma_j(X(t), t; \theta)}{\partial \theta_k} dW_t^j \right], \end{aligned}$$

with the initial condition $\Delta_{jk}(0) = \partial X_j(0)/\partial \theta_k$.

Pathwise Derivative Method: Is it worth the trouble?

- ▶ The Pathwise Derivative Estimators of the sensitivities are mathematically equivalent to the estimates obtained by standard finite differences approaches when using the same random numbers in both simulations and for a vanishing small perturbation. In this limit, the Pathwise Derivative Method and finite differences estimators provide exactly the same estimators for the sensitivities, i.e., estimators with the same expectation value, *and* the same MC variance.
- ▶ As a result, the implementation effort associated with the Pathwise Derivative Method is generally justified if the computational cost of the Pathwise Estimator is significantly less than the corresponding finite differences one.
- ▶ This is the case for instance in very simple models but difficult to achieve for those used in the financial practice.

Section 3

Adjoint Algorithmic Differentiation (AAD)

'Algebraic' Adjoint Methods

- ▶ In 2006 Mike Giles and Paul Glasserman published a ground breaking 'Smoking Adjoints' in Risk Magazine [6].
- ▶ They proposed a very efficient implementation of the Pathwise Derivative Method in the context of the Libor Market Model for European payouts (generalized to Bermudan options by Leclerc *et al.* [3] and extended by Joshi *et al.* [5]).
- ▶ In a nutshell:
 1. Concentrate on the Tangent process and formulate its propagation in terms of Linear Algebra operations.
 2. Optimize the computation time by rearranging the order of the computations.
 3. Implement the rearranged sequence of operations.
- ▶ In the following we denote these Adjoint approaches as *algebraic*.

Libor Market Model

- ▶ Let's indicate with T_i , $i = 1, \dots, N + 1$, a set of $N + 1$ bond maturities, with spacings $\delta = T_{i+1} - T_i$ (constant for simplicity).
- ▶ In a Lognormal setup the dynamics of the forward Libor rates as seen at time t for the interval $[T_i, T_{i+1})$, $L_i(t)$, takes the form

$$\frac{dL_i(t)}{L_i(t)} = \mu_i(L(t))dt + \sigma_i(t)^T dW_t,$$

$0 \leq t \leq T_i$, and $i = 1, \dots, N$, where W_t is a d_W -dimensional standard Brownian motion, $L(t)$ is the N -dimensional vector of Libor rates, and $\sigma_i(t)$ the d_W -dimensional vector of volatilities, at time t .

Libor Market Model: Euler Discretization

- ▶ The dynamics of the forward Libor rates can be simulated by applying a Euler discretization to the logarithms of the forward rates.
- ▶ By dividing each interval $[T_i, T_{i+1})$ into N_s steps of equal width, $h = \delta/N_s$. This gives

$$\frac{L_i(t_{n+1})}{L_i(t_n)} = \exp \left[\left(\mu_i(L(t_n)) - \|\sigma_i(t_n)\|^2/2 \right) h + \sigma_i^T(n) Z(t_n) \sqrt{h} \right],$$

for $i = \eta(nh), \dots, N$, and $L_i(t_{n+1}) = L_i(t_n)$ if $i < \eta(nh)$. Here Z is a d_W -dimensional vector of independent standard normal variables and t_0 is today.

Swaption Payout

- ▶ The standard test case are contracts with expiry T_m to enter in a swap with payments dates T_{m+1}, \dots, T_{N+1} , at a fixed rate K

$$V(T_m) = \sum_{i=m+1}^{N+1} B(T_m, T_i) \delta(S_m(T_m) - K)^+,$$

where $B(T_m, T_i)$ is the price at time T_n of a bond maturing at time T_i

$$B(T_m, T_i) = \prod_{l=m}^{i-1} \frac{1}{1 + \delta L_l(T_m)},$$

and the swap rate reads

$$S_m(T_m) = \frac{1 - B(T_m, T_{N+1})}{\delta \sum_{l=m+1}^{N+1} B(T_m, T_l)}.$$

- ▶ Here we consider European style payouts. It is simple to generalize to Bermudan options (see [3]).

Pathwise Derivative Estimator for Delta

- ▶ The Pathwise Estimator for the Delta,

$$\bar{L}_k(t_0) = \frac{\partial V(T_m)}{\partial L_k(t_0)},$$

reads:

$$\bar{L}_k(t_0) = \sum_{j=1}^N \frac{\partial V(T_m)}{\partial L_j(T_m)} \frac{\partial L_j(T_m)}{\partial L_k(t_0)} = \frac{\partial V(T_m)}{\partial L(T_m)}^T \Delta(T_m),$$

where the Tangent process is

$$\Delta_{jk}(t) = \frac{\partial L_j(t)}{\partial L_k(t_0)}.$$

Euler Evolution of the Tangent Process

- ▶ By differentiating the Euler discretization for the Libor dynamics one obtains the Euler discretization of the Tangent process dynamics:

$$\Delta_{ik}(t_{n+1}) = \Delta_{ik}(t_n) \frac{L_i(t_{n+1})}{L_i(t_n)} + L_i(t_{n+1}) \sum_{j=1}^N \frac{\partial \mu_i(t_n)}{\partial L_j(t_n)} \Delta_{jk}(t_n),$$

where $\Delta_{ik}(t_0) = \partial L_i(t_0) / \partial L_k(t_0) = \delta_{jk}$.

- ▶ The evolution of the Tangent process can be expressed as the matrix recursion:

$$\Delta(t_{n+1}) = B(t_n) \Delta(t_n)$$

where $B(t_n)$ is an $N \times N$ matrix.

Standard (Forward) Implementation of the Pathwise Derivative Estimator

- ▶ A standard implementation for the calculation of the Pathwise Estimator

$$\bar{L}_k(T_0) = \frac{\partial V(T_m)}{\partial L(T_m)}^T \Delta(T_m),$$

where $T_m = t_M$, with $M = N_s \times m$, involves:

1. Apply the matrix recursion

$$\Delta(t_{n+1}) = B(t_n) \Delta(t_n),$$

M times starting from $\Delta_{ik}(t_0) = \delta_{jk}$ in order to compute $\Delta(T_m)$. The total cost in the general case is $O(MN^3)$.

2. Compute analytically the derivatives of the payoff

$$\frac{\partial V(T_m)}{\partial L(T_m)},$$

and multiply it by $\Delta(T_m)$, at a cost $O(N^2)$.

Standard (Forward) Implementation of the Pathwise Derivative Estimator

- ▶ This involves proceeding from right to left (i.e., forward in time):

$$\bar{L}_k(T_0) = \frac{\partial V(T_m)}{\partial L(T_m)}^T B(t_{M-1}) \dots B(t_0) \Delta(t_0)$$

at a total computational cost $O(MN^3)$ in the general case.

- ▶ However, a simple observation allows a much more efficient implementation...

Adjoint (Backward) Implementation

- After completing the evolution of the Libor path up to T_m the right hand side of

$$\bar{L}_k(T_0) = \frac{\partial V(T_m)}{\partial L(T_m)}^T B(t_{M-1}) \dots B(t_0) \Delta(t_0)$$

can be computed from left to right (i.e., backward in time) by taking the transpose (i.e., the 'Adjoint')

$$\bar{L}_k(T_0)^T = \Delta(t_0) B(t_0)^T \dots B(t_{M-1})^T \frac{\partial V(T_m)}{\partial L(T_m)},$$

or equivalently as

$$\bar{L}_k(T_0)^T = \Delta(t_0) A(t_0)^T$$

where the $A(t_0)$ is the N dimensional vector given by the matrix-vector recursion

$$A(t_n) = B(t_n)^T A(t_{n+1}) \quad A_k(t_M) = \frac{\partial V(T_m)}{\partial L_k(T_m)}.$$

Forward vs Adjoint: Computational Complexity

Compare:

- ▶ The forward computation of the Pathwise Estimator

$$\bar{L}_k(T_0) = \frac{\partial V(T_m)}{\partial L(T_m)}^T B(t_{M-1}) \dots B(t_0) \Delta(t_0)$$

which consists of M matrix-matrix products and a final matrix-vector product for an overall cost of $O(MN^3)$ in the general case.

- ▶ The Adjoint computation of the Pathwise Estimator

$$\bar{L}_k(T_0)^T = \Delta(t_0) B(t_0)^T \dots B(t_{M-1})^T \frac{\partial V(T_m)}{\partial L(T_m)},$$

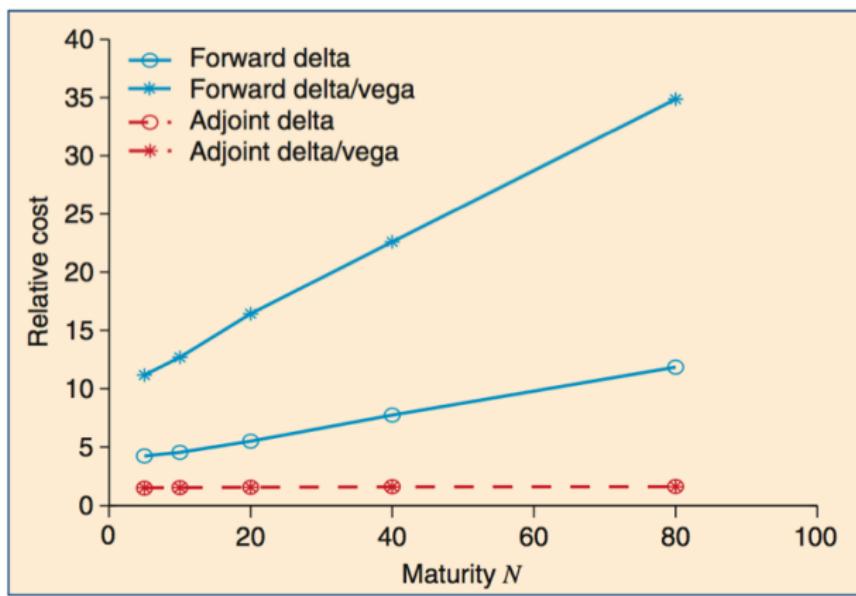
which consists of $M + 1$ matrix-vector products with an overall computational cost of $O(MN^2)$.

- ▶ The Adjoint implementation is $O(N)$ cheaper than the forward one.

Forward vs Adjoint: Computational Complexity

- ▶ As a result, computing the Pathwise Derivative Estimators for Delta has the same computational complexity of propagating the forward Libor rates and evaluating the payout.
- ▶ This means that we can get all the Delta sensitivities at a cost that is of the same order of magnitude than computing the payout (rather than $O(N)$ larger if we were computing the Deltas by bumping).
- ▶ The same results holds also for Vega.

Algebraic Adjoint Methods



From Ref. [6]

Arbitrary number of sensitivities at a **fixed small** cost.

Limitations of Algebraic Adjoint Methods

The Libor Market Model is bit of an ad-hoc application:

- ▶ Difficult to generalize to Path Dependent Options or multi asset simulations.
- ▶ The required algebraic analysis is in general cumbersome.
- ▶ Not general enough for all the applications in Finance.
- ▶ The derivatives required are often not available in closed form.
- ▶ What about the derivatives of the payout?

Algorithmic Adjoint Approaches: AAD

- ▶ Adjoint implementations can be seen as instances of a programming technique known as Adjoint Algorithmic Differentiation (AAD) [4].
- ▶ In general AAD allows the calculation of the gradient of an algorithm at a cost that is a small constant (~ 4) times the cost of evaluating the function itself, independent of the number of input variables.
- ▶ Given that for each random realization the Payoff estimator can be seen as a map

$$\theta_k \rightarrow P(X(\theta_k)),$$

AAD allows the calculation of the Pathwise Derivative Estimators for **any** number of sensitivities

$$\bar{\theta}_k = \frac{\partial P(X(\theta_k))}{\partial \theta_k},$$

at a **small fixed cost**, similarly to the Algebraic Adjoint applications of the Libor Market Model, but now in complete generality.

Algorithmic Differentiation

- ▶ Algorithmic Differentiation (AD) is a set of programming techniques first introduced in the early 60's aimed at computing accurately and efficiently the derivatives of a function given in the form of a computer program.
- ▶ The main idea underlying AD is that any such program can be interpreted as the composition of functions each of which is in turn a composition of basic arithmetic (addition, multiplication etc.), and intrinsic operations (logarithm, exponential, etc.).
- ▶ Hence, it is possible to calculate the derivatives of the outputs of the program with respect to its inputs by applying mechanically the rules of differentiation.
- ▶ This makes it possible to generate *automatically* a computer program that evaluates efficiently and with machine precision accuracy the derivatives of the function [4].

Algorithmic Differentiation

- ▶ What makes AD particularly attractive when compared to standard (e.g., finite difference) methods for the calculation of the derivatives, is its computational efficiency.
- ▶ In fact, AD aims at exploiting the information on the structure of the computer function, and on the dependencies between its various parts, in order to optimize the calculation of the sensitivities.
- ▶ AD comes in two main flavors, Tangent and Adjoint mode, which are characterized by different properties in different (complementary) computational complexity.

Algorithmic Differentiation: Tangent mode

- ▶ Consider a function

$$Y = \text{FUNCTION}(X)$$

mapping a vector X in \mathbb{R}^n in a vector Y in \mathbb{R}^m .

- ▶ The execution time of its Tangent counterpart

$$\dot{X} = \text{FUNCTION_d}(X, \dot{X})$$

(with suffix `_d` for “dot”) calculating the linear combination of the columns of the Jacobian of the function:

$$\dot{Y}_j = \sum_{i=1}^n \dot{X}_i \frac{\partial Y_j}{\partial X_i},$$

with $j = 1, \dots, m$, is bounded by

$$\frac{\text{Cost}[\text{FUNCTION_d}]}{\text{Cost}[\text{FUNCTION}]} \leq \omega_T$$

with $\omega_T \in [2, 5/2]$.

Algorithmic Differentiation: Adjoint mode

- ▶ The execution time of the Adjoint counterpart of

$$Y = \text{FUNCTION}(X),$$

namely,

$$\bar{X} = \text{FUNCTION_b}(X, \bar{Y})$$

(with suffix `_b` for “backward” or “bar”) calculating the linear combination of the rows of the Jacobian of the function:

$$\bar{X}_i = \sum_{j=1}^m \bar{Y}_j \frac{\partial Y_j}{\partial X_i},$$

with $i = 1, \dots, n$, is bounded by

$$\frac{\text{Cost}[\text{FUNCTION_b}]}{\text{Cost}[\text{FUNCTION}]} \leq \omega_A$$

with $\omega_A \in [3, 4]$.

Algorithmic Differentiation: Tangent vs Adjoint mode

Given the results above:

- ▶ The Tangent mode is particularly well suited for the calculation of (linear combinations of) the columns of the Jacobian matrix.
- ▶ Instead, the Adjoint mode is particularly well-suited for the calculation of (linear combinations of) the rows of the Jacobian matrix .
- ▶ In particular, the Adjoint mode provides the full gradient of a scalar ($m = 1$) function at a cost which is just a small constant times the cost of evaluating the function itself. Remarkably such relative cost is *independent* of the number of components of the gradient.
- ▶ When the full Jacobian is required, the Adjoint mode is likely to be more efficient than the Tangent mode when the number of independent variables is significantly larger than the number of the dependent ones ($m \ll n$). Or viceversa.

Tangent mode: Propagating Forwards

- ▶ Imagine that the function $Y = \text{FUNCTION}(X)$ is implemented by means of a sequence of steps

$$X \rightarrow \dots \rightarrow U \rightarrow V \rightarrow \dots \rightarrow Y,$$

where the real vectors U and V represent intermediate variables used in the calculation and each step can be a distinct high-level function or even an individual instruction.

- ▶ Define the Tangent of any intermediate variable U_k as

$$\dot{U}_k = \sum_{i=1}^n \dot{X}_i \frac{\partial U_k}{\partial X_i}.$$

Tangent mode: Propagating Forwards

- ▶ Using the chain rule we get,

$$\dot{V}_j = \sum_{i=1}^n \dot{X}_i \frac{\partial V_j}{\partial X_i} = \sum_{i=1}^n \dot{X}_i \sum_k \frac{\partial V_j}{\partial U_k} \frac{\partial U_k}{\partial X_i} = \sum_k \frac{\partial V_j}{\partial U_k} \dot{U}_k,$$

which corresponds to the Tangent mode equation for the intermediate step represented by the function $V = V(U)$

$$\dot{V}_j = \sum_k \dot{U}_k \frac{\partial V_j}{\partial U_k},$$

namely a function of the form $\dot{V} = \dot{V}(U, \dot{U})$.

Tangent mode: Propagating Forwards

- ▶ Hence the computation of the Tangents can be executed in the same direction of the original function

$$\dot{X} \rightarrow \dots \rightarrow \dot{U} \rightarrow \dot{V} \rightarrow \dots \rightarrow \dot{Y}.$$

This can be executed simultaneously with the original function, since at each intermediate step $U \rightarrow V$ one can compute the derivatives $\partial V_j(U)/\partial U_k$ and execute the Tangent forward propagation

$$\dot{U} \rightarrow \dot{V} \quad \dot{V}_j = \sum_k \dot{U}_k \frac{\partial V_j}{\partial U_k}.$$

- ▶ The Tangent of the output obtained with this forward recursion is by definition:

$$\dot{Y}_k = \sum_{i=1}^n \dot{X}_i \frac{\partial Y_k}{\partial X_i},$$

i.e., in a single forward sweep one can produce a linear combination of the columns of the Jacobian $\partial Y/\partial X$.

Tangent mode: Propagating Forwards

- ▶ The Tangent mode produces the linear combination of columns of the Jacobian

$$\dot{Y}_k = \sum_{i=1}^n \dot{X}_i \frac{\partial Y_k}{\partial X_i},$$

where \dot{X} is an arbitrary vector in \mathbb{R}^n .

- ▶ By initializing in turn \dot{X} with each vector of the canonical basis in \mathbb{R}^n , (e_1, \dots, e_n) with

$$e_j = (\underbrace{0, \dots, 1, 0, \dots, 0}_j)$$

one can obtain the partial derivatives of all the outputs with respect to each of the inputs $\dot{Y}_k = \partial Y_k / \partial X_i$, thus resulting in a cost that is n times the cost of a single forward Tangent sweep.

Tangent mode: Propagating Forwards

- ▶ It is not difficult to realize that the cost of computing each single step $\dot{U} \rightarrow \dot{V}$ is just a small multiple of the cost of executing $U \rightarrow V$.
- ▶ Going at the level of a single instruction:

$$V_1 = V_1(U_1, U_2)$$
$$\dot{V}_1 = \frac{\partial V_1}{\partial U_1} \dot{U}_1 + \frac{\partial V_1}{\partial U_2} \dot{U}_2$$

For instance:

$$V_1 = U_1 U_2$$
$$\dot{V}_1 = U_2 \dot{U}_1 + U_1 \dot{U}_2$$

Key Observation: One multiply in the original nonlinear calculation turns into 2 multiply and one add operations in forward mode.

Tangent mode: Propagating Forwards

- ▶ Extending to the whole computation one can see how keeping into account of the relative cost of different types of operation one can arrive to the result [4]:

$$\frac{\text{Cost[FUNCTION_d]}}{\text{Cost[FUNCTION]}} \leq \omega_T$$

with $\omega_T \in [2, 5/2]$.

- ▶ By performing simultaneously the calculation of all the components of the gradient one can optimize the calculation by reusing a certain amount of computations (for instance the arc derivatives). This leads to a more efficient implementation also known as *Tangent Multimode*. The constant ω_T for these implementations is generally smaller than in the standard Tangent mode.

Adjoint mode: Propagating Backwards

- ▶ Let's consider again the function $Y = \text{FUNCTION}(X)$ implemented by means of a sequence of steps

$$X \rightarrow \dots \rightarrow U \rightarrow V \rightarrow \dots \rightarrow Y.$$

- ▶ Define the Adjoint of any intermediate variable V_k as

$$\bar{V}_k = \sum_{j=1}^m \bar{Y}_j \frac{\partial Y_j}{\partial V_k},$$

where \bar{Y} is vector in \mathbb{R}^m .

Adjoint mode: Propagating Backwards

- ▶ Using the chain rule we get,

$$\bar{U}_i = \sum_{j=1}^m \bar{Y}_j \frac{\partial Y_j}{\partial U_i} = \sum_{j=1}^m \bar{Y}_j \sum_k \frac{\partial Y_j}{\partial V_k} \frac{\partial V_k}{\partial U_i},$$

which corresponds to the Adjoint mode equation for the intermediate step represented by the function $V = V(U)$

$$\bar{U}_i = \sum_k \bar{V}_k \frac{\partial V_k}{\partial U_i},$$

namely a function of the form $\bar{U} = \bar{V}(U, \bar{V})$.

Adjoint mode: Propagating Backwards

- ▶ Starting from the Adjoint of the outputs, \bar{Y} , we can apply this rule to each step in the calculation, working from right to left,

$$\bar{X} \leftarrow \dots \leftarrow \bar{U} \leftarrow \bar{V} \leftarrow \dots \leftarrow \bar{Y}$$

until we obtain \bar{X} , i.e., the following linear combination of the rows of the Jacobian $\partial Y / \partial X$

$$\bar{X}_i = \sum_{j=1}^m \bar{Y}_j \frac{\partial Y_j}{\partial X_i},$$

with $i = 1, \dots, n$.

- ▶ Contrary to the Tangent mode, the backward propagation can start only after the calculation of the function and the intermediate variables have been computed and stored.

Adjoint mode: Propagating Backwards

- ▶ It is not difficult to realize that the cost of computing each single step $\bar{U} \leftarrow \bar{V}$ is just a small multiple of the cost of executing $U \rightarrow V$.
- ▶ Going at the level of a single instruction:

$$V_1 = V_1(U_1, U_2)$$

$$\bar{U}_1 = \bar{U}_1 + \frac{\partial V_1}{\partial U_1} \bar{V}_1$$

$$\bar{U}_2 = \bar{U}_2 + \frac{\partial V_1}{\partial U_2} \bar{V}_1$$

For instance:

$$V_1 = U_1 U_2$$

$$\bar{U}_1 = \bar{U}_1 + U_2 \bar{V}_1$$

$$\bar{U}_2 = \bar{U}_2 + U_1 \bar{V}_1$$

Key Observation: One multiply in the original nonlinear calculation turns into 2 multiply-add operations in adjoint mode.

Adjoint mode: Propagating Backwards

- ▶ Extending to the whole computation one can see how keeping into account of the relative cost of different types of operation one can arrive to the result [4]:

$$\frac{\text{Cost[FUNCTION_b]}}{\text{Cost[FUNCTION]}} \leq \omega_A$$

with $\omega_A \in [3, 4]$.

- ▶ This result is based on the number of arithmetic operations which must be performed. It also includes the cost of memory operations, but assumes a uniform cost for these, irrespective of the total amount of memory used. This assumption is violated in practice due to the cache hierarchy in modern computers.
- ▶ Nevertheless, it remains true in practice that one can obtain the sensitivity of a single output, or a linear combination of outputs, to an unlimited number of inputs for only a little more work than the original calculation.

Recap: Tangent vs Adjoint

- When computing the Jacobian of a composition of N functions:
 $\mathbb{R}^n \rightarrow \mathbb{R}^m$

$$y = f_N \circ f_{N-1} \circ \dots \circ f_1(x)$$

$$J_i = \frac{\partial f_i}{\partial z}$$

we have 2 main ways of applying the chain rule:

Tangent

Adjoint

$$\dot{y} = \frac{\partial y}{\partial x} \dot{x}$$

$$\bar{x} = \left(\frac{\partial y}{\partial x} \right)^T \bar{y}$$

$$\dot{y} = J_N \dots J_1 \dot{x} \quad \bar{x} = J_1^T \dots J_N^T \bar{y}$$

- The calculation of the Jacobian is $O(n)$ in the tangent mode and $O(m)$ in the adjoint mode.
- The adjoint mode is particularly well suited for the computation of gradients.

First Examples: Derivatives of Payoff Functions

- ▶ As a first example let's consider the Payoff of a Basket Option

$$P(X(T)) = e^{-rT} \left(\sum_{i=1}^N w_i X_i(T) - K \right)^+,$$

where $X(T) = (X_1(T), \dots, X_N(T))$ represent the value of a set of N underlying assets, say a set of equity prices, at time T , w_i , $i = 1, \dots, N$, are the weights defining the composition of the basket, K is the strike price, and r is the risk free yield for the considered maturity.

- ▶ For this example, we are interested in the calculation of the sensitivities with respect to r and the N components of the state vector X so that the other parameters, i.e., strike and maturity, are seen here as dummy constants.

Pseudocode of the Basket Option

```
(P)= payout (r, X[N]){

    B = 0.0;
    for (i = 1 to N)
        B += w[i]* X[i];

    x = B - K;
    D = exp(-r * T);
    P = D * max(x, 0.0);
};
```

From Ref. [2]

Pseudocode of the Tangent Payoff for the Basket Option

```
(P, P_d) = payout_d(r, X[N], r_d, x_d[N]){

    B = 0.0;
    for (i = 1 to N) {
        B += w[i]*X[i];
        B_d += w[i]*X_d[i];
    }

    x = B - K;
    x_d = B_d;

    D = exp(-r * T);
    D_d = -T * D * r_d;

    P = D * max(x, 0.0);
    P_d = 0;
    if(x > 0)
        P_d = D_d*x + D*x_d;

};

}
```

From Ref. [2]

- ▶ The computational cost of the Tangent payoff is of the same order of the original Payoff.
- ▶ To get all the components of the gradient of the payoff, the Tangent payoff code must be run $N + 1$ times, setting in turn one component of the Tangent input vector $I = (\dot{r}, \dot{X})^t$ to one and the remaining ones to zero.

Pseudocode of the Adjoint Payoff for the Basket Option

```

(P, r_b, X_b[N]) = payout_b(r, X[N], P_b){

                                // Forward sweep
B = 0.0;
for (i = 0 to N)
    B += w[i] * X[i];

x = B - K;
D = exp(-r * T);
P = D * max(x, 0.0);
                                // Backward sweep
D_b = max(x, 0.0) * P_b;

x_b = 0.0;
if (x > 0)
    x_b = D * P_b;

r_b = - D * T * D_b;
B_b = x_b;

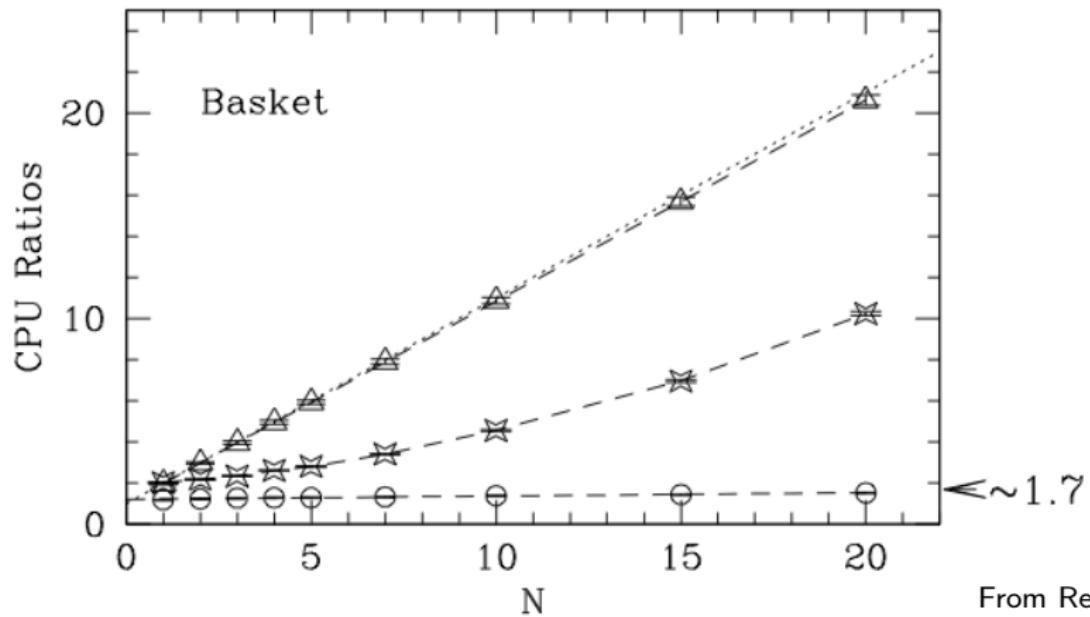
for (i = 0 to N)
    X_b[i] = w[i] * B_b;
};

```

From Ref. [2]

- ▶ The Adjoint payoff contains a forward sweep.
- ▶ The computational cost of the Adjoint payoff is of the same order of the original Payoff.
- ▶ All the components of the gradient of the payoff, are obtained by running the Adjoint payoff only once setting $\bar{P} = 1$.

Tangent vs Adjoint



From Ref. [2]

- ▶ The Tangent payoff performs similarly to bumping (much better for the Multimode version) and has a computational complexity that scales with the number of inputs.
- ▶ In the Adjoint mode the calculation of all the derivatives of the payoff requires an extra overhead of just 70% with respect to the calculation of the payoff itself for any number of inputs.

Tangent vs Adjoint

- ▶ In general we are interested in computing the sensitivities of a derivative or of a portfolio of derivatives with respect to a large number of risk factors.
- ▶ The Adjoint model of Algorithmic Differentiation is therefore the one best suited for the task.
- ▶ In some applications, however, one is also interested in computing the sensitivities of a multiplicity of derivatives individually. In those cases one can effectively combine the Adjoint and Tangent mode. See e.g. [2].
- ▶ In the following we will concentrate on the Adjoint mode of Algorithmic Differentiation (AAD) as it is the one of wider applicability.

Section 4

AAD as a Software Design Principle

AAD as a Software Design Principle

- ▶ The propagation of the Adjoints according to the steps, being mechanical in nature, can be automated.
- ▶ Several AD tools are available that given a procedure of the form:

$$Y = \text{FUNCTION}(X),$$

generate the Adjoint function:

$$\bar{X} = \text{FUNCTION_b}(X, \bar{Y}).$$

- ▶ An excellent source of information can be found at www.autodiff.org.

AAD as a Software Design Principle

- ▶ The principles of AD can be used as a programming paradigm for any algorithm.
- ▶ An easy way to illustrate the Adjoint design paradigm is to consider again the arbitrary computer function

$$Y = \text{FUNCTION}(X),$$

and to imagine that this represents a certain high level algorithm that we want to differentiate.

- ▶ By appropriately defining the intermediate variables, any such algorithm can be abstracted in general as a composition of functions like

$$X \rightarrow \dots \rightarrow U \rightarrow V \rightarrow \dots \rightarrow Y.$$

AAD as a Software Design Principle

- ▶ However, the actual calculation graph might have a more complex structure. For instance the step $U \rightarrow V$ might be implemented in terms of two computer functions of the form

$$\begin{aligned} V^1 &:= v1(U^1), \\ V^2 &:= v2(U^1, U^2), \end{aligned}$$

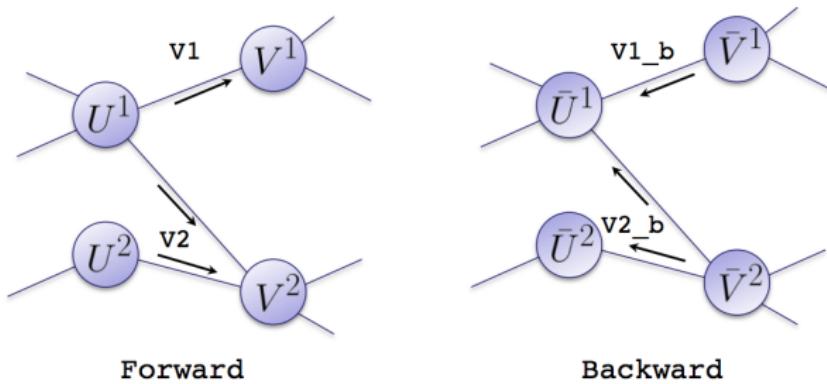
with $U = (U^1, U^2)^t$ and $V = (V^1, V^2)^t$. Here the notation $W = (W^1, W^2)^t$ simply indicates a specific partition of the components of the vector W in two sub-vectors.

- ▶ A natural way to represent the step $\bar{U} \leftarrow \bar{V}$ in

$$\bar{X} \leftarrow \dots \leftarrow \bar{U} \leftarrow \bar{V} \leftarrow \dots \leftarrow \bar{Y}$$

i.e., the function $\bar{U} = \bar{V}(U, \bar{V})$, can be given in terms of an Adjoint calculation graph.

AAD as a Software Design Principle



- ▶ The Adjoint graph has the same structure of the original graph with each node/variable representing the Adjoint of the original node/variable, and it is executed in opposite direction with respect to the original one.

Forward and Backward Sweeps

- ▶ The Adjoint instructions

$$\begin{aligned} (\bar{U}^1, \bar{U}^2)^t &:= \text{v2_b}(U^1, U^2, \bar{V}^2), \\ \bar{U}^1 &:= \bar{U}^1 + \text{v1_b}(U^1, \bar{V}^1). \end{aligned}$$

depend on the variables U^1 and U^2 .

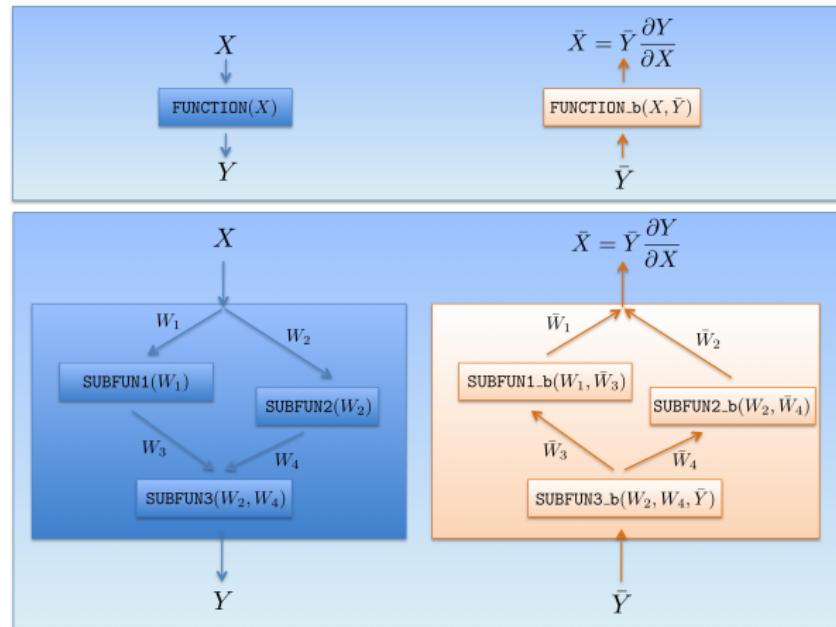
- ▶ As a result, the Adjoint algorithm can be executed only after the original instructions

$$X \rightarrow \dots \rightarrow U \rightarrow V \rightarrow \dots \rightarrow Y.$$

have been executed and the necessary intermediate results have been computed and stored.

- ▶ This is the reason why, as noted before, the Adjoint of a given algorithm generally contains a *forward sweep*, which reproduces the steps of the original algorithm, plus a *backward sweep*, which propagates the Adjoints.

AAD as a Software Design Principle



- The principles of AD can be used as a programming paradigm for any algorithm.
- The Adjoint graph has the same structure of the original graph with each node/variable representing the Adjoint of the original node/variable, and it is executed in opposite direction with respect to the original one.

Section 5

AAD and the Pathwise Derivative Method

AAD and the Pathwise Derivative Method

- ▶ AAD provides a general design and programming paradigm for the efficient implementation of the Pathwise Derivative Method.
- ▶ This stems from the observation that the Pathwise Estimator in

$$\bar{\theta}_k \equiv \frac{dP_\theta(X)}{d\theta_k} = \sum_{j=1}^d \frac{\partial P_\theta(X)}{\partial X_j} \times \frac{\partial X_j}{\partial \theta_k} + \frac{\partial P_\theta(X)}{\partial \theta_k},$$

is a l.c. of the rows of the Jacobian of the map $\theta \rightarrow X(\theta)$, with weights given by the X gradient of the payout function $P_\theta(X)$, plus the derivatives of the payout function with respect to θ .

- ▶ Both the calculation of the derivatives of the payout and of the linear combination of the rows of $\partial X / \partial \theta$ are tasks that can be performed efficiently by AAD.
- ▶ We know now that we can compute all the Pathwise sensitivities with respect to θ , $\bar{\theta}$, at a cost that is at most roughly 4 times the cost of calculating the payout estimator itself.

AAD enabled Monte Carlo Engines: Forward Sweep

- ▶ In a typical MC simulation, in order to generate each sample $X[i_{MC}]$, the evolution of the process X is usually simulated, possibly by means of an approximate discretization scheme, by sampling $X(t)$ on a discrete grid of points, $0 = t_0 < t_1 < \dots < t_n < \dots < t_{N_s}$, a superset of the observation times (T_1, \dots, T_M) .
- ▶ The state vector at time t_{n+1} is obtained by means of a function of the form

$$X(t_{n+1}) = \text{PROP}_n[\{X(t_m)\}_{m \leq n}, Z(t_n), \theta],$$

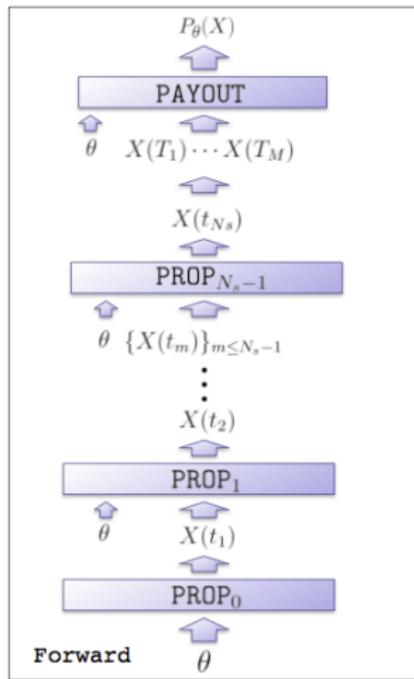
mapping the set of state vector values on the discretization grid up to t_n , $\{X(t_m)\}_{m \leq n}$, into the value of the state vector at time $t_n + 1$.

AAD enabled Monte Carlo Engines: Forward Sweep

- ▶ Note that in $X(t_{n+1}) = \text{PROP}_n[\{X(t_m)\}_{m \leq n}, Z(t_n), \theta]$:
 - a The propagation method is a function of the model parameters θ and of the particular time step considered.
 - b $Z(t_n)$ indicates the vector of uncorrelated random numbers which are used for the MC sampling in the step $n \rightarrow n + 1$.
 - c The initial values of the state vector $X(t_0)$ are known quantities and they can be considered as components of θ so that the $n = 0$ step is of the form, $X(t_1) = \text{PROP}_0[Z(t_0), \theta]$.
- ▶ Once the full set of state vector values on the simulation time grid $\{X(t_m)\}_{m \leq N_s}$ is obtained, the subset of values corresponding to the observation dates is passed to the payout function, evaluating the payout estimator $P_\theta(X)$ for the specific random sample $X[i_{\text{MC}}]$

$$(X(T_1), \dots, X(T_M)) \rightarrow P_\theta(X(T_1), \dots, X(T_M)).$$

AAD enabled Monte Carlo Engines: Forward Sweep



Schematic illustration of the orchestration of a MC engine.

AAD enabled Monte Carlo Engines: Backward Sweep

- ▶ The evaluation of a MC sample of a Pathwise Estimator can be seen as an algorithm implementing a function of the form $\theta \rightarrow P(\theta)$.
- ▶ As a result, it is possible to design its Adjoint counterpart $(\theta, \bar{P}) \rightarrow (P, \bar{\theta})$ which gives (for $\bar{P} = 1$) the Pathwise Derivative Estimator $dP/d\theta_k$.
- ▶ The backward sweep can be simply obtained by reversing the flow of the computations, and associating to each function its Adjoint counterpart.

AAD enabled Monte Carlo Engines: Backward Sweep

- In particular, the first step of the Adjoint algorithm is the Adjoint of the payout evaluation $P = P(X, \theta)$. This is a function of the form

$$(\bar{X}, \bar{\theta}) = \bar{P}(X, \theta, \bar{P}),$$

where $\bar{X} = (\bar{X}(T_1), \dots, \bar{X}(T_M))$ is the Adjoint of the state vector on the observation dates, and $\bar{\theta}$ is the Adjoint of the model parameters vector, respectively (for $\bar{P} = 1$)

$$\begin{aligned}\bar{X}(T_m) &= \frac{\partial P_\theta(X)}{\partial X(T_m)}, \\ \bar{\theta} &= \frac{\partial P_\theta(X)}{\partial \theta},\end{aligned}$$

for $m = 1, \dots, M$. The Adjoint of the state vector on the simulation dates corresponding to the observation dates are initialized at this stage. The remaining ones are initialized to zero.

AAD enabled Monte Carlo Engines: Backward Sweep

- ▶ The Adjoint state vector is then propagated backwards in time through the Adjoint of the propagation method, namely

$$(\{\bar{X}(t_m)\}_{m \leq n}, \bar{\theta}) += \text{PROP_b}_n[\{X(t_m)\}_{m \leq n}, Z(t_n), \theta, \bar{X}(t_{n+1})],$$

for $n = N_s - 1, \dots, 1$, giving

$$\bar{X}(t_m) += \sum_{j=1}^N \bar{X}_j(t_{n+1}) \frac{\partial X_j(t_{n+1})}{\partial X(t_m)},$$

with $m = 1, \dots, n$,

$$\bar{\theta} += \sum_{j=1}^N \bar{X}_j(t_{n+1}) \frac{\partial X_j(t_{n+1})}{\partial \theta}.$$

AAD enabled Monte Carlo Engines: Backward Sweep

- ▶ Here, according to the principles of AAD, the Adjoint of the propagation method takes as arguments the inputs of its forward counterpart, namely the state vectors up to time t_n , $\{X(t_m)\}_{m \leq n}$, the vector of random variates $Z(t_n)$, and the θ vector. The additional input is the Adjoint of the state vector at time t_{n+1} , $\bar{X}(t_{n+1})$.
- ▶ The return values of PROP_b_n are the contributions associated with the step $n+1 \rightarrow n$ to the Adjoints of
 - the state vector $\{\bar{X}(t_m)\}_{m \leq n}$;
 - the model parameters $\bar{\theta}_k$, $k = 1, \dots, N_\theta$.
- ▶ The final step of the backward propagation corresponds to the Adjoint of $X(t_1) = \text{PROP}_0[Z(t_0), \theta]$, giving

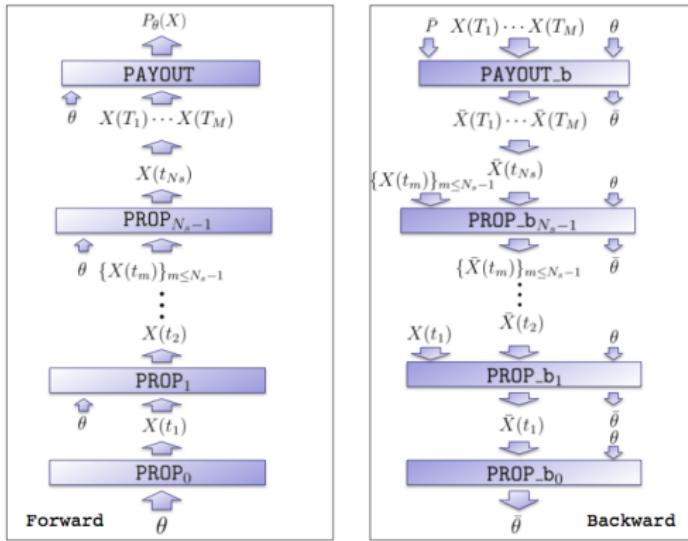
$$\bar{\theta} += \text{PROP_b}_0[X(t_0) Z(t_0), \theta, \bar{X}(t_1)],$$

i.e., the final contribution to the Adjoints of the model parameters.

- ▶ It is easy to verify that the final result is the Pathwise Derivative Estimator $dP/d\theta_k$ for all k 's on the given MC path.

AAD enabled Monte Carlo Engines: The complete blueprint

- ▶ The resulting algorithm can be illustrated as follows:



Schematic illustration of the orchestration of an AAD enabled MC engine.

Diffusion Processes and Euler Discretization

- ▶ As a first example, consider the case in which the underlying factors follow multi dimensional diffusion processes introduced in slide 3

$$dX(t) = \mu(X(t), t, \theta) dt + \sigma(X(t), t, \theta) dW_t.$$

- ▶ In this case, the evolution of the process X is usually approximated by sampling $X(t)$ on a discrete grid of points by means, for instance, of an Euler scheme, so that the propagation function

$$X(t_{n+1}) = \text{PROP}_n[\{X(t_m)\}_{m \leq n}, Z(t_n), \theta]$$

implements the rule

$$X(t_{n+1}) = X(t_n) + \mu(X(t_n), t_n, \theta) h_n + \sigma(X(t_n), t_n, \theta) \sqrt{h_n} Z(t_n),$$

where $h_n = t_{n+1} - t_n$, and $Z(t_n)$ is a N -dimensional vector of correlated unit normal random variables.

Diffusion Processes and Euler Discretization: Forward Sweep

- ▶ In particular, given the state vector at time t_n , $X(t_n)$, and the vector $Z(t_n)$, one can implement the method PROP_n according to the following steps:

Step 1. Compute the drift vector, by evaluating the function:

$$\mu(t_n) = \mu(X(t_n), t_n, \theta) .$$

Step 2. Compute the volatility vector, by evaluating the function:

$$\sigma(t_n) = \sigma(X(t_n), t_n, \theta) .$$

Step 3. Compute the function

$$(X(t_n), \mu(t_n), \sigma(t_n), Z(t_n), \theta) \rightarrow X(t_{n+1}) ,$$

defined by

$$X(t_{n+1}) = X(t_n) + \mu(t_n)h_n + \sigma(t_n)\sqrt{h_n}Z(t_n) .$$

Diffusion Processes and Euler Discretization: Backward Sweep

- ▶ The corresponding Adjoint method `PROP_bn` is executed from time step t_{n+1} to t_n and consists of the Adjoint counterpart of each of the steps above executed in reverse order, namely:

Step 3. Compute the Adjoint of the function defined by Step 3. This is a function

$$(X(t_n), \mu(t_n), \sigma(t_n), Z(t_n), \bar{X}(t_{n+1})) \rightarrow \bar{X}(t_n)$$

defined by the instructions

$$\begin{aligned} \bar{X}(t_n) &+= \bar{X}(t_{n+1}), \\ \bar{\mu}(t_n) &= 0, & \bar{\mu}(t_n) &+= \bar{X}(t_{n+1})h_n, \\ \bar{\sigma}(t_n) &= 0, & \bar{\sigma}(t_n) &+= \bar{X}(t_{n+1})\sqrt{h_n} Z(t_n), \\ \bar{Z}(t_n) &= 0, & \bar{Z}(t_n) &+= \bar{X}(t_{n+1})\sqrt{h_n} \sigma(t_n). \end{aligned}$$

Diffusion Processes and Euler Discretization: Backward Sweep

► And:

Step 2. Compute the Adjoint of the volatility function in Step 2, namely

$$\bar{X}_i(t_n) += \sum_{j=1}^N \bar{\sigma}_j(t_n) \frac{\partial \sigma_j(t_n)}{\partial X_i}, \quad \bar{\theta}_k += \sum_{j=1}^N \bar{\sigma}_j(t_n) \frac{\partial \sigma_j(t_n)}{\partial \theta_k},$$

for $i = 1, \dots, N$ and $k = 1, \dots, N_\theta$.

Step 1. Compute the Adjoint of the drift function in Step 1, namely

$$\bar{X}_i(t_n) += \sum_{j=1}^N \bar{\mu}_j(t_n) \frac{\partial \mu_j(t_n)}{\partial X_i(t_n)}, \quad \bar{\theta}_k += \sum_{j=1}^N \bar{\mu}_j(t_n) \frac{\partial \mu_j(t_n)}{\partial \theta_k},$$

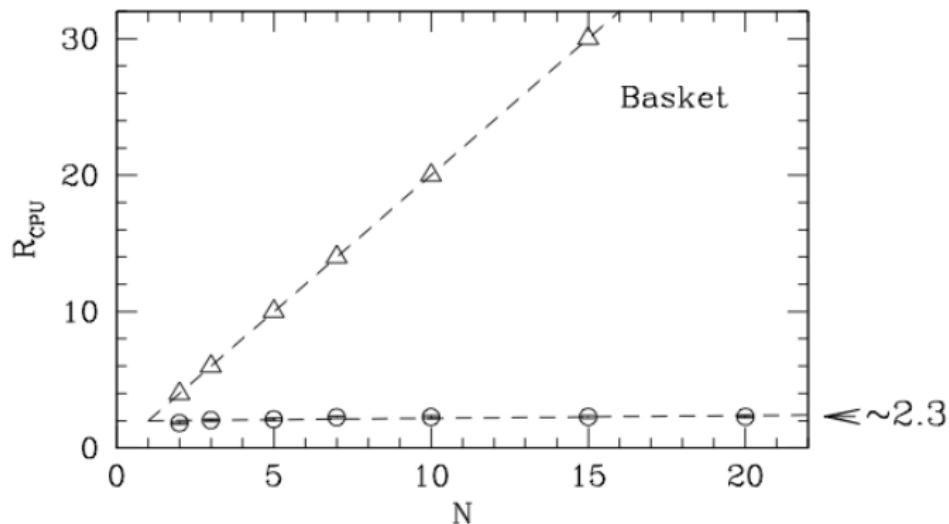
for $i = 1, \dots, N$ and $k = 1, \dots, N_\theta$.

Diffusion Processes and Euler Discretization: Backward Sweep

- ▶ Note that, the variables $\bar{X}(t_{n+1})$, $\bar{X}(t_n)$ and $\bar{\theta}$ typically contain on input the derivatives of the payout function. During the backward propagation $\bar{X}(t_n)$ (resp. $\bar{\theta}$) accumulate several contributions, one for each Adjoint of an instruction in which $X(t_n)$ (resp. θ) is on the right hand side of the assignment operator in the forward sweep (Steps 1-3).
- ▶ The implementation of the Adjoint of the drift and volatility functions in Step $\bar{2}$ and Step $\bar{1}$ is problem dependent. In many cases, the drift and volatility may be represented by computer routines self-contained enough to be processed by means of an automatic differentiation tool, thus facilitating the implementation.

Basket Options: Results

- ▶ Let's consider again the Basket Option example introduced earlier for the Payoff.



CPU time ratios for the calculation of Delta and Vega Risk as a function of the number of underlying assets, N : circles (AAD), triangles (Bumping).

Basket Options: Comments

- ▶ The performance of the AAD implementation of the Pathwise Derivative Method in this setup is well within the expected bounds.
- ▶ In particular, the computation of the $2 \times N$ sensitivities for the N assets requires a very small overhead (of about 130%) with respect to the calculation of the option value itself. This is true for any number of underlying assets.
- ▶ This is in stark contrast with the relative cost of evaluating the same sensitivities by means of finite-differences, scaling linearly with the number of assets.
- ▶ For typical applications this clearly results in remarkable speedups with respect to bumping.

Libor Market Model

- ▶ Let's consider again the application of the seminal paper by Giles and Glasserman [6] See slide 2 and ff. Recall that:
- ▶ T_i , $i = 1, \dots, N + 1$, is a set of $N + 1$ bond maturities, with spacings $\delta = T_{i+1} - T_i$ (constant for simplicity).
- ▶ In a Lognormal setup the dynamics of the forward Libor rates as seen at time t for the interval $[T_i, T_{i+1})$, $L_i(t)$, takes the form

$$\frac{dL_i(t)}{L_i(t)} = \mu_i(L(t))dt + \sigma_i(t)^T dW_t,$$

$0 \leq t \leq T_i$, and $i = 1, \dots, N$, where W_t is a d_W -dimensional standard Brownian motion, $L(t)$ is the N -dimensional vector of Libor rates, and $\sigma_i(t)$ the d_W -dimensional vector of volatilities, at time t .

Libor Market Model: Euler Discretization

- ▶ The dynamics of the forward Libor rates can be simulated by applying a Euler discretization to the logarithms of the forward rates.
- ▶ By dividing each interval $[T_i, T_{i+1})$ into N_s steps of equal width, $h = \delta/N_s$. This gives

$$\frac{L_i(t_{n+1})}{L_i(t_n)} = \exp \left[\left(\mu_i(L(t_n)) - \|\sigma_i(t_n)\|^2/2 \right) h + \sigma_i^T(n) Z(t_n) \sqrt{h} \right],$$

for $i = \eta(nh), \dots, N$, and $L_i(t_{n+1}) = L_i(t_n)$ if $i < \eta(nh)$. Here Z is a d_W -dimensional vector of independent standard normal variables and t_0 is today.

Swaption Payout

- ▶ The standard test case are contracts with expiry T_m to enter in a swap with payments dates T_{m+1}, \dots, T_{N+1} , at a fixed rate K

$$V(T_m) = \sum_{i=m+1}^{N+1} B(T_m, T_i) \delta(S_n(T_m) - K)^+,$$

where $B(T_m, T_i)$ is the price at time T_m of a bond maturing at time T_i

$$B(T_m, T_i) = \prod_{l=m}^{i-1} \frac{1}{1 + \delta L_l(T_m)},$$

and the swap rate reads

$$S_m(T_m) = \frac{1 - B(T_m, T_{N+1})}{\delta \sum_{l=m+1}^{N+1} B(T_m, T_l)}.$$

- ▶ Here we consider European style payouts. It is simple to generalize to Bermudan options (see [3]).

Libor Market Model: Forward Sweep

```

PROP(n, L[,], Z, lambda[], L0[])

if(n=0)
  for(i= 1 .. N)
    L[i,n] = L0[i]; Lhat[i,n] = L0[i];

  for (i = 1 .. eta[n]-1)
    L[i,n+1] = L[i,n]; // settled rates

  sgez = sqrt(h)*Z;
  v = 0.; v_pc = 0.;
  for (i = eta[n] .. N)
    lam = lambda[i-eta[n]+1];
    c1 = del*lam; c2 = h*lam;
    v += (c1*L[i,n])/(1.+del*L[i,n]);
    vrat = exp(c2*(-lam/2.+v)+lam*sgez);
    // standard propagation with the Euler drifts
    Lhat[i,n+1] = L[i,n]*vrat;
    // (n + 1) drift term
    v_pc += (c1*Lhat[i,n+1])/(1.+del*Lhat[i,n+1]);
    vrat_pc = exp(c2*(-lam/2.+(v_pc+v)/2.)+lam*sgez);
    // actual propagation using the average drift
    L[i,n+1] = L[i,n]*vrat_pc;
    // store what is needed for the reverse sweep
    hat_scra[i,n+1] = vrat*((v-lam)*h+sgez);
    scra[i,n+1] = vrat_pc*(((v_pc+v)/2.-lam)*h+sgez);

```

Pseudocode implementing the propagation method $PROP_n$ for the Libor Market Model for $d_W = 1$, under the predictor corrector Euler approximation.

Libor Market Model: Backward Sweep

```

PROP_b(n, L[], z, lambda[], L0[], lambda_b[], L0_b[])

v_b = 0.; v_pc_b = 0.;

for (i=NM .. eta[n])
    lam = lambda[i-eta[n]+1];
    c1 = del*lam; c2 = lam*h;
    //L[i,n+1] = L[i,n]*vrat_pc
    vrat_pc = L[i,n]/L[i,n];
    vrat_pc_b = L[i,n]*L_b[i,n+1];
    L_b[i,n] = vrat_pc*L_b[i,n+1];
    // vrat_pc = exp(c2*(-lam/2.+(v_pc+v)/2.)+lam*sqez)
    lambda_b[i-eta[n]+1] += scra[i,n+1]*vrat_pc_b;
    v_pc_b += vrat_pc*lam*h*vrat_pc_b/2.;
    v_b += vrat_pc*lam*h*vrat_pc_b/2.;
    // v_pc += (c1*Lhat[i,n+1])/(1.+del*Lhat[i,n+1])
    rpip = 1. / (del*Lhat[i,n+1]+1.);

    Lhat_b[i,n+1] += (c1-c1*Lhat[i,n+1]*del*rpip)*rpip*v_pc_b;
    c1_b = Lhat[i,n+1]*rpip*v_pc_b;
    // Lhat[i,n+1] = L[i,n]*vrat
    vrat_b = L[i,n]*Lhat_b[i,n+1];
    vrat = Lhat[i,n+1]/L[i,n];
    L_b[i,n] += vrat*Lhat_b[i,n+1];
    // vrat = exp(lam*h*(-lam/2.+v)+lam*sqez)
    lambda_b[i-eta[n]+1] += hat_scra[i,n+1]*vrat_b;
    v_b += vrat*lam*h*vrat_b;
    // v += (c1*L[i,n])/(1.+del*L[i,n])
    rpip = 1. / (del*L[i,n]+1.);

    L_b[i,n] += (c1-c1*L[i,n]*del*rpip)*rpip*v_b;
    c1_b += L[i,n]*rpip*v_b;
    // lam = lambda[i-eta[n]+1]; c1 = del*lam
    lambda_b[i-eta[n]+1] += del*c1_b;

for (i=eta[n]-1 .. 1)
    // L[i,n+1] = L[i,n]
    L_b[i,n] += L_b[i,n];

if(n=0)
    for(i=1 .. N)
        // L[i,n] = L0[i]
        L0_b[i] = L0[i,0];

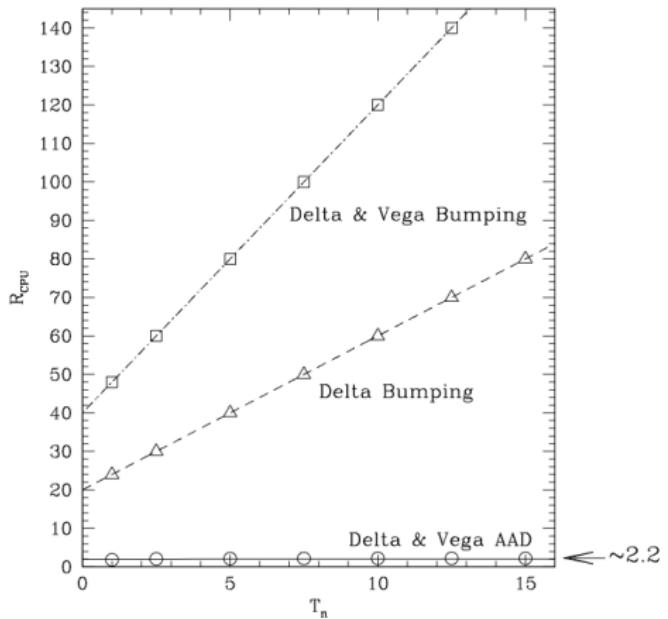
```

Adjoint of the propagation method $PROP_b_n$ [7].

Libor Market Model: Comments on the Code

- ▶ The algebraic formulation discussed in [5] comes with a significant analytical effort. Instead, as illustrated in the Figure above, the AAD implementation is quite straightforward.
- ▶ According to the general design of AAD, this simply consists of the Adjoint of the instructions in the forward sweep executed in reverse order.
- ▶ In this example, the information computed by PROP that is required by PROP_b is stored in the vectors `scra` and `hat_scra`.
- ▶ By inspecting the structure of the pseudocode it also appears clear that the computational cost of PROP_b is of the same order as evaluating the original function.

Libor Market Model: Results



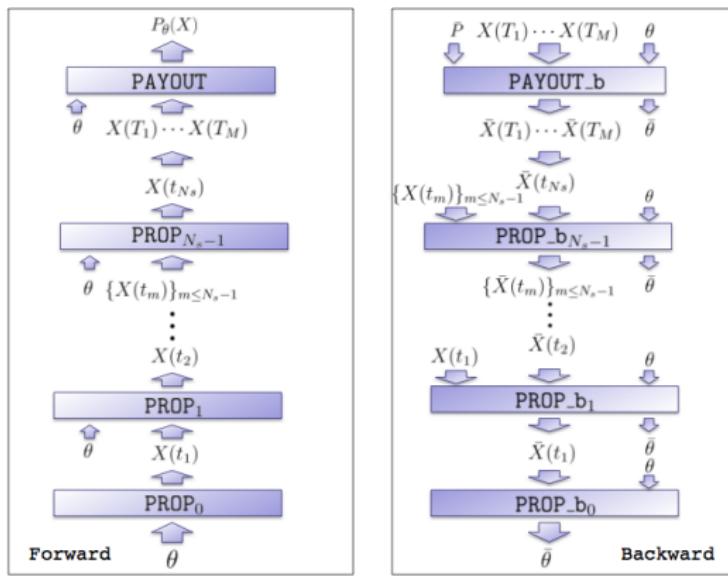
Ratio of the CPU time required for the calculation of Delta and Vega and the time to calculate the option value for the Swaption as a function of the option expiry T_n .

Section 6

Correlation Greeks and Binning Techniques

Correlation Structure of the Random Variates

Recall the general AAD MC design for the computation of the estimators on each MC path:



Correlation Structure of the Random Variates

- ▶ In the AAD MC design we have assumed for simplicity that the random variates $Z(t_n)$ entering in the propagation method:

$$X(t_{n+1}) = \text{PROP}_n[\{X(t_m)\}_{m \leq n}, Z(t_n), \theta],$$

are dummy variables carrying no interesting sensitivities.

- ▶ As a result, in the corresponding adjoint propagation methods:

$$(\{\bar{X}(t_m)\}_{m \leq n}, \bar{\theta}) += \text{PROP_b}_n[\{X(t_m)\}_{m \leq n}, Z(t_n), \theta, \bar{X}(t_{n+1})],$$

the adjoint of the random variates $\bar{Z}(t_n)$ do not appear among the outputs.

- ▶ If we want to compute the sensitivities with respect to the correlation structure of the random variates, this scheme needs to be extended.

Correlation Structure of the Random Variates

- ▶ In a typical setup, the random variates Z_i driving the random processes are correlated.
- ▶ For instance, assume that the random variates $Z(t_n)$ are jointly normal, and denote with $\rho_{ij}(t_m) = \mathbb{E}[Z_i(t_m)Z_j(t_m)]$ the correlation matrix.
- ▶ Uncorrelated random variates $Z'(t_n)$ are therefore mapped into their correlated counterparts $Z(t_n)$ and then used to implement the propagation step $X(t_n) \rightarrow X(t_{n+1})$ so that the propagation step is modified as

$$\begin{aligned} Z(t_n) &= \text{CORRELATE}(Z'(t_n), \theta) \\ X(t_{n+1}) &= \text{PROP}_n[\{X(t_m)\}_{m \leq n}, Z(t_n), \theta] , \end{aligned}$$

where we have included the correlation parameters defining the correlation matrix ρ in the vector θ .

Modified Adjoint of the Propagation Step

- ▶ The adjoint of the Propagation Step

$$X(t_{n+1}) = \text{PROP}_n[\{X(t_m)\}_{m \leq n}, Z(t_n), \theta],$$

is modified as

$$(\{\bar{X}(t_m)\}_{m \leq n}, \bar{\theta}, \bar{Z}(t_n)) += \text{PROP_b}_n[\{X(t_m)\}_{m \leq n}, Z(t_n), \theta, \bar{X}(t_{n+1})],$$

where

$$\bar{X}(t_m) += \sum_{j=1}^N \bar{X}_j(t_{n+1}) \frac{\partial X_j(t_{n+1})}{\partial X(t_m)} \quad \bar{\theta} += \sum_{j=1}^N \bar{X}_j(t_{n+1}) \frac{\partial X_j(t_{n+1})}{\partial \theta},$$

with $m = 1, \dots, n$. Here the additional output is given by the adjoint of the correlated variates:

$$\bar{Z}(t_n) += \sum_{j=1}^N \bar{X}_j(t_{n+1}) \frac{\partial X_j(t_{n+1})}{\partial Z(t_n)}.$$

Adjoint of the Correlation Step

- ▶ The adjoint of the Correlation Step

$$Z(t_n) = \text{CORRELATE}(Z'(t_n), \theta),$$

reads

$$\bar{\theta} += \text{CORRELATE_b}(Z'(t_n), \theta, \bar{Z}(t_n)),$$

corresponding to the operation

$$\bar{\theta} += \sum_{j=1}^N \bar{Z}'_j(t_n) \frac{\partial Z_j(t_n)}{\partial \theta}$$

updating the components of the vector θ corresponding to the adjoint of the correlation parameters.

Example: Cholesky Factorization

- ▶ In a simple setup the method CORRELATE generally involves the so-called Cholesky factorization of an $N \times N$ correlation matrix ρ .
- ▶ Recall that the Cholesky factorization of a Hermitian positive-definite matrix ρ produces a lower triangular $N \times N$ matrix L such that $\rho = LL^T$.
- ▶ Given the Cholesky factor L , and a vector of N uncorrelated normal Z' , it is immediate to verify that $Z = LZ'$ are correlated normal such that $\mathbb{E}[Z_i Z_j] = \rho_{ij}$.

Adjoint of the Cholesky Factorization

- When implemented in terms of the Cholesky factorization, the method CORRELATE reads

Step 1 Perform Cholesky factorization, say $L = \text{CHOLESKY}(\rho)$.

Step 2 Compute: $Z = LZ'$.

- The corresponding method CORRELATE_b reads

Step 2 Compute: $\bar{L} = \bar{Z}Z'^t$.

Step 1 Compute: $\bar{\rho} = \text{CHOLESKY_b}(\rho, \bar{L})$, where

$$\bar{\rho}_{ij} = \sum_{l,m=1}^N \frac{\partial L_{l,m}}{\partial \rho_{ij}} \bar{L}_{lm},$$

providing the sensitivities with respect to the entries of the correlation matrix. These are copied in the appropriate components of the vector $\bar{\theta}$.

- Note that Z' are now dummy integration variables (sampled stochastically). Therefore their adjoints \bar{Z}' are not computed.

Adjoint of the Cholesky Factorization (Pseudocode)

```

Cholesky_b(rho, L_b,rho_b)

// Forward Sweep
for (i=0 .. n-1)
  for (j=i .. n-1)
    sum[i,j] = rho[i,j];
    for (k=i-1 .. 0)
      sum[i,j] -= L[i,k] * L[j,k];
    if (i == j)
      L[i,i] = sqrt(sum[i,j]);
    else
      L[j,i] = sum[i,j] / L[i,i];

// Backward Sweep
for (i=n-1 .. 0)
  for (j=n-1 .. i)
    sum_b = 0.0;

    if (i == j)
      if (sum[i,j] == 0.0)
        sum_b = 0.0;
      else
        sum_b = L_b[i,j]/( 2.0 * L[i,j]);
      L_b[i,j] = 0.0;
    else
      sum_b = L_b[j,i]/L[i,i];
      L_b[i,i] -= sum[i,j] * sum_b / L[i,i];
      L_b[j,i] = 0.0;

    for (k=i-1 .. 0)
      L_b[i,k] -= L[j,k]*sum_b;
      L_b[j,k] -= L[i,k]*sum_b;

    rho_b[i,j] += sum_b;
  
```

The adjoint algorithm contains the original Cholesky factorization plus a backward sweep with the same complexity and a similar number of operations [8].

Hence, as expected, the computational cost is just a small multiple (of order 2, in this case) of the cost of evaluating the original factorization.

Adjoint of the Cholesky Factorization

- ▶ The Cholesky factorization $L = \text{CHOLESKY}(\rho)$ does not depend on the random variates Z therefore it can be performed before the first Monte Carlo path is performed. As a result, CORRELATE consists of the matrix multiplication $Z = LZ'$, only.
- ▶ Similarly the Adjoint of CORRELATE_b consists only of the step $\bar{L} = \bar{Z}'Z^t$, ($\bar{\theta}$ will contain the adjoint of the Cholesky factors L rather than the entries of the correlation matrix ρ) and the Adjoint of the Cholesky factorization

$$\bar{\rho} = \text{CHOLESKY_b}(\rho, \bar{L})$$

can be performed after the end of the backward sweep after the last MC path.

Statistical Uncertainties

- Given the MC estimators for the Cholesky factors sensitivities $\langle \bar{L} \rangle = \langle \partial V(X) / \partial L \rangle$ and their statistical uncertainties

$$\langle \bar{L} \rangle = \frac{1}{N_{MC}} \sum_{i_{MC}=1}^{N_{MC}} \bar{L}(X[i_{MC}]) \quad \sigma_{\bar{L}} = \sqrt{\frac{1}{N_{MC}} \sum_{i_{MC}=1}^{N_{MC}} (\bar{L}(X[i_{MC}])^2 - \langle \bar{L} \rangle)^2}$$

- One can compute the estimator for the correlation sensitivities via the Cholesky factorization

$$\langle \bar{\rho} \rangle = \text{CHOLESKY_b}(\rho, \langle \bar{L} \rangle)$$

but not their sensitivities:

$$\sigma_{\bar{\rho}} \neq \text{CHOLESKY_b}(\rho, \sigma_{\bar{L}})$$

- Performing the adjoint of the Cholesky decomposition once per simulation does not allow the calculation of a confidence interval for the correlation sensitivities.

Path by Path Adjoint Cholesky Factorization

- ▶ An alternative approach would be to convert \bar{L} to $\bar{\rho}$ for each individual path $i_{\text{MC}} = 1, \dots, N_{\text{MC}}$

$$\bar{\rho}(X[i_{\text{MC}}]) = \text{CHOLESKY_b}(\rho, \bar{L}(X[i_{\text{MC}}]))$$

and then compute the average and standard deviation of $\bar{\rho}[i_{\text{MC}}]$ in the usual way:

$$\langle \bar{\rho} \rangle = \frac{1}{N_{\text{MC}}} \sum_{i_{\text{MC}}=1}^{N_{\text{MC}}} \bar{\rho}(X[i_{\text{MC}}]) \quad \sigma_{\bar{\rho}} = \sqrt{\frac{1}{N_{\text{MC}}} \sum_{i_{\text{MC}}=1}^{N_{\text{MC}}} (\bar{\rho}(X[i_{\text{MC}}])^2 - \langle \bar{\rho} \rangle)^2}$$

However, this is rather costly.

Binning

- ▶ An excellent compromise between these two extremes is to divide the N_{MC} paths into N_B 'bins' of equal size $n = N/N_B$.
- ▶ For each bin $j_B = 1, \dots, N_B$, an average value of $\langle \bar{L} \rangle_{j_B}$ is computed

$$\langle \bar{L} \rangle_{j_B} = \frac{1}{n} \sum_{i_{MC}=1}^n \bar{L}(X[i_{MC}])$$

and converted into a corresponding value for

$$\langle \bar{\rho} \rangle_{j_B} = \text{CHOLESKY_b}(\rho, \langle \bar{L} \rangle_{j_B}).$$

Binning

- ▶ These N_B estimates for $\bar{\rho}$ can then be combined in the usual way to form an overall estimate of the correlation risk:

$$\langle \bar{\rho} \rangle = \frac{1}{N_B} \sum_{j_B=1}^{N_B} \langle \bar{\rho} \rangle_{j_B} = \frac{1}{N_{MC}} \sum_{i_{MC}=1}^{N_{MC}} \bar{\rho}(X[i_{MC}]),$$

where the second equality follows from the linearity of the adjoint functions, and the associated confidence interval:

$$\sigma_{\bar{\rho}} = \sqrt{\frac{1}{N_B} \sum_{j_B=1}^{N_B} \left(\langle \bar{\rho} \rangle_{j_B}^2 - \langle \bar{\rho} \rangle \right)^2}.$$

Binning

- ▶ In the standard evaluation, the cost of the Cholesky factorization is $O(N^3)$ (where N is the number of random factors), and the cost of the MC sampling is $O(N_{MC}N^2)$, so the total cost is $O(N^3 + N_{MC}N^2)$. Since N_{MC} is always much greater than N , the cost of the Cholesky factorization is usually negligible.
- ▶ The cost of the adjoint steps in the MC sampling is also $O(N_{MC}N^2)$, and when using N_B bins the cost of the adjoint Cholesky factorization is $O(N_B N^3)$.
- ▶ To obtain an accurate confidence interval, but with the cost of the Cholesky factorisation being negligible, requires that N_B is chosen so that $1 \ll N_B \ll N_{MC}/N$.
- ▶ Without binning, i.e., using $N_B = N_{MC}$, the cost to calculate the average of the estimators for $\langle \rho \rangle$ is $O(N_{MC}N^3)$, and so the relative cost compared to the evaluation of the option value is $O(N)$.

Binning and Risk Transforms

- ▶ We have presented Binning in the context of the calculation of correlation risk, but there is nothing specific to correlation. In fact these ideas can be applied everytime some computational preprocessing is performed before the MC simulation, and we need to transform the adjoint MC estimators and their confidence interval into the corresponding quantities for the inputs of such preprocessing.
- ▶ This is the case for instance when a calibration routine performed before the MC simulation transforms some market inputs $M = (M_1, \dots, M_{N_M})$, corresponding to the observable prices of securities which the model is calibrated to, into the set of internal model parameters that are used in the MC simulation θ :

$$\theta = \text{CALIBRATION}(M).$$

Binning and Risk Transforms

- ▶ The binned MC estimators of the adjoint of the internal model parameters $\langle \bar{\theta} \rangle_{j_B}$ can be transformed into binned MC estimators of the market inputs

$$\langle \bar{M} \rangle_{j_B} = \text{CALIBRATION_B}(M, \langle \bar{\theta} \rangle_{j_B}).$$

- ▶ Then their distribution can be used to construct the overall MC estimator and the associated statistical uncertainty

$$\langle \bar{M} \rangle = \frac{1}{N_B} \sum_{j_B=1}^{N_B} \langle \bar{M} \rangle_{j_B},$$

$$\sigma_{\bar{M}} = \sqrt{\frac{1}{N_B} \sum_{j_B=1}^{N_B} \left(\langle \bar{M} \rangle_{j_B}^2 - \langle \bar{M} \rangle \right)^2}.$$

Credit Basket Contracts

- ▶ Credit basket contracts are derivatives that are contingent on credit events (defaults for short) of a pool of reference entities typically sovereign, financial or corporate. Generally the credit event is defined as failure to pay a specific liability, say a coupon on a specific bond or category of bonds referenced by the contract, but it can include other events not involving a proper default, like a restructuring of the debt, or regulatory action on a financial institution.
- ▶ n -th to default, Collateralized Debt Obligations (CDO) and their variations are examples of credit basket products.
- ▶ In the context of basket credit default products the random factors X_i are the time of default τ_i of the i -th reference entity in a basket of N names and the payoff is of the form:

$$P = P(\tau_1, \dots, \tau_N)$$

Example: n -th to default Basket Default Swap

- ▶ In a n -th to default Basket Default Swap one party (protection buyer) makes regular payments to a counterparty (protection seller) at time $T_1, \dots, T_M \leq T$ provided that less than n defaults events among the components of the basket are observed before time T_M .
- ▶ If n defaults occur before time T , the regular payments cease and the protection seller makes a payment to the buyer of $(1 - R_i)$ per unit notional, where R_i is the normalized recovery rate of the i -th asset.
- ▶ The value at time zero of the Basket Default Swap on a given realization of the default times τ_1, \dots, τ_N , i.e., the Payout function, can be expressed as

$$P(\tau_1, \dots, \tau_N) = P_{\text{prot}}(\tau_1, \dots, \tau_N) - P_{\text{prem}}(\tau_1, \dots, \tau_N)$$

i.e., as the difference between the so-called *protection* and *premium* legs.

Example: n -th to default Basket Default Swap

- ▶ The value leg is given by

$$P_{prot}(\tau_1, \dots, \tau_N) = (1 - R_n) D(\tau) \mathbb{I}(\tau \leq T),$$

where R_n and τ are the recovery rate and default time of the n -th to default, respectively, $D(t)$ is the discount factor for the interval $[0, t]$ (here we assume for simplicity uncorrelated default times and interest rates), and $\mathbb{I}(\tau \leq T)$ is the indicator function of the event that the n -th default occurs before T .

- ▶ The premium leg reads instead, neglecting for simplicity any accrued payment,

$$P_{prem}(\tau_1, \dots, \tau_N) = \sum_{k=1}^{T_M} c_k D(T_k) \mathbb{I}(\tau \geq T_k)$$

where c_k is the premium payment (per unit notional) at time T_k .

Copula Models

- ▶ Credit Basket Products are also known as *correlation products* because their value depends not only on the marginal distribution of the default times but also on their correlation structure.
- ▶ Such correlation structure is typically captured by means of a copula model. For instance, in a Gaussian copula, the cumulative joint distribution of default times is assumed of the form:

$$\mathbb{P}(\tau_1 \leq t_1, \dots, \tau_N \leq t_N) = \Phi_N(\Phi^{-1}(F_1(t_1)), \dots, \Phi^{-1}(F_N(t_N)); \rho)$$

where $\Phi_N(Z_1, \dots, Z_N; \rho)$ is a N -dimensional multivariate Gaussian distribution with zero mean, and a $N \times N$ positive semidefinite correlation matrix ρ ; Φ^{-1} is the inverse of the standard normal cumulative distribution, and $F_i(t) = \mathbb{P}(\tau_i \leq t)$, $i = 1, \dots, N$, are the marginal distributions of the default times of each reference entity, depending on a set of model parameters θ .

Hazard Rate Model

- ▶ The key concept for the valuation of credit derivatives, in the context of the models generally used in practice, is the *hazard rate*, λ_u , representing the probability intensity of default of the reference entity between times u and $u + du$, conditional on survival up to time u . The hazard rate function λ_u is commonly parameterized as piece-wise constant with M knot points at time (t_1, \dots, t_M) , $\lambda = (\lambda_1, \dots, \lambda_M)$, .
- ▶ By modelling the default event of a reference entity i as the first arrival time of a Poisson process with intensity λ_u^i , the survival probability, $\mathbb{P}(\tau_i > t)$, is given by

$$\mathbb{P}(\tau_i > t) = \exp \left[- \int_0^t du \lambda_u^i \right],$$

so that the marginal cumulative distribution of default times reads

$$F_i(t; \lambda^i) = \mathbb{P}(\tau \leq t) = 1 - \exp \left[- \int_0^t du \lambda_u^i \right],$$

Forward Simulation Algorithm

The simulation of a Gaussian Copula model can be seen as a single time-step instance of the general approach, consisting of the following steps:

Step 0 Perform a Cholesky factorization of the matrix ρ , say $L = \text{CHOLESKY}(\rho)$.

For each MC replication:

Step 1 Generate an N dimensional vector of uncorrelated normal Gaussian variates Z' .

Step 2 Correlate the random variates: $Z = \text{CORRELATE}(Z', L)$, where as previously discussed the correlation step consist of a single matrix vector multiplication $Z = LZ'$.

Step 3 Perform the 'propagation step' $\tau = \text{PROP}_0[Z, \theta]$.

Step 4 Evaluate the payout function: $P = P(\tau)$.

Forward Simulation Algorithm

- ▶ From the form of the cumulative joint distribution of default times

$$\mathbb{P}(\tau_1 \leq t_1, \dots, \tau_N \leq t_N) = \Phi_N(\Phi^{-1}(F_1(t_1; \lambda^1)), \dots, \Phi^{-1}(F_N(t_N; \lambda^N)); \rho)$$

it follows that the random variates $\Phi^{-1}(F_1(\tau_i, \lambda^i))$ are distributed according to a multivariate normal distribution.

- ▶ Hence the propagation step $\tau = \text{PROP}_0[Z, \theta]$ consists in turn of the following sub-steps:

Step 3a Set $U_i = \Phi(Z_i)$, $i = 1, \dots, N$.

Step 3b Set $\tau_i = F_i^{-1}(U_i; \lambda_i)$, $i = 1, \dots, N$.

where $F_i^{-1}(U_i; \lambda_i)$ is the root τ_i of the equation

$$\exp \left[- \int_0^{\tau_i} du \lambda_u^i \right] = 1 - U_i.$$

Adjoint Simulation Algorithm

- ▶ The corresponding adjoint algorithm consists of the following steps:

Step 4 Evaluate the adjoint Payout $\bar{\tau}_i = \partial P / \partial \tau_i$, for $i = 1, \dots, N$.

Step 3 Evaluate the adjoint of the propagation step:

$$(\bar{\lambda}, \bar{Z}) = \text{PROP_b}_0[Z, \theta, \bar{\tau}].$$

Step 2 Calculate the adjoint of the correlation step:

$$\bar{L} = \text{CORRELATE_b}(Z', \bar{Z}),$$

implemented as

$$\bar{L} = \bar{Z} Z'^t.$$

Adjoint Simulation Algorithm

- In turn, the adjoint of the correlation step reads:

Step 3b Calculate:

$$\bar{U}_i = \bar{\tau}_i \frac{\partial F_i^{-1}(U_i; \lambda^i)}{\partial U_i} = \bar{\tau}_i \frac{1}{f_i(F_i^{-1}(U_i; \lambda^i); \lambda)},$$

$$\bar{\lambda}_j^i = \bar{\tau}_i \frac{\partial F_i^{-1}(U_i; \lambda^i)}{\partial \lambda_j^i},$$

for $i = 1, \dots, N$ and $j = 1, \dots, M$.

Step 3a Calculate: $\bar{Z}_i = \bar{U}_i \phi(Z_i)$, $i = 1, \dots, N$.

where $f_i(t; \lambda) = \partial F(t; \lambda)/\partial t$ is the p.d.f. of the default time of the i -th reference entity and $\phi(x)$ is the standard normal p.d.f. Note that computing the derivative $\partial F_i^{-1}(U_i; \lambda^i)/\partial \lambda_j^i$ involves differentiating the root searching algorithm used to determine the default time τ_i . However, a much better implementation is possible by means of the so-called implicit function theorem [11].

Payout Smoothing

- ▶ In order to apply the Pathwise Derivative method to the payout above, the indicator functions in the premium and protection legs

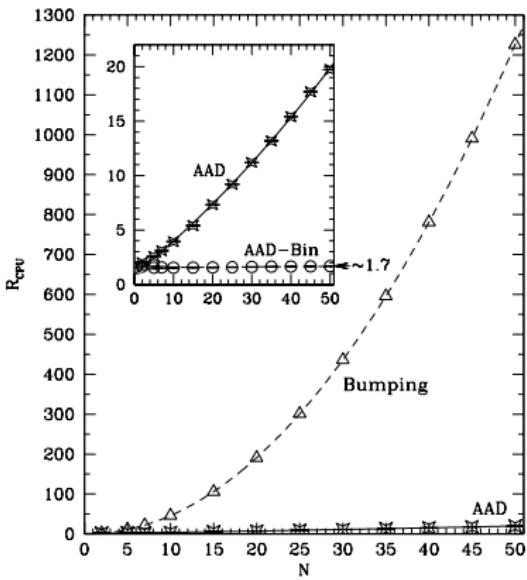
$$P_{\text{prem}}(\tau_1, \dots, \tau_N) = \sum_{k=1}^{T_M} c_k D(T_k) \mathbb{I}(\tau \geq T_k),$$

$$P_{\text{prot}}(\tau_1, \dots, \tau_N) = (1 - R_n) D(\tau) \mathbb{I}(\tau \leq T),$$

need to be regularized.

- ▶ As seen before, one simple and practical way of doing that is to replace the indicator functions with their smoothed counterpart, at the price of introducing a small amount of bias in the Greek estimators.
- ▶ For the problem at hand, as it is also generally the case, such bias can be easily reduced to be smaller than the statistical errors that can be obtained for any realistic number of MC iteration N_{MC} .

Results



Ratios of the CPU time required for the calculation of the option value, and correlation Greeks, and the CPU time spent for the computation of the value alone, as functions of the number of names in the basket. Symbols: Bumping (one-sided finite differences) (triangles), AAD without binning (i.e. $N_B = N_{MC}$) (stars), AAD with binning ($N_B = 20$) (empty circles) [9].

Results

- ▶ As expected, for standard finite-difference estimators, such ratio increases quadratically with the number of names in the basket. Already for medium sized basket ($N \simeq 20$) the cost associated with Bumping is over 100 times more expensive than the one of AAD.
- ▶ Nevertheless, at a closer look (see the inset) the relative cost of AAD without binning is $O(N)$, because of the contribution of the adjoint of the Cholesky decomposition.
- ▶ However, when using $N_B = 20$ bins the cost of the adjoint Cholesky computation is negligible and the numerical results show that all the Correlation Greeks can be obtained with a mere 70% overhead compared to the calculation of the value of the option.
- ▶ This results in over 2 orders of magnitude savings in computational time for a basket of over 40 Names.

Section 7

Case Study: Real Time Counterparty Credit Risk Management

Counterparty Credit Risk Problem

- ▶ Credit Valuation Adjustment (CVA):

$$V_{\text{CVA}} = \mathbb{E} \left[\mathbb{I}(\tau_c \leq T) D(\tau_c) \times L_{\text{GD}}(\tau_c) \left(NPV(\tau_c) - C(R(\tau_c^-)) \right)^+ \right],$$

where τ_c is the default time of the counterparty, $NPV(t)$ is the net present value of the portfolio at time t , $C(R(t))$ is the collateral outstanding, typically dependent on the rating R of the counterparty, $L_{\text{GD}}(t)$ is the loss given default, $D(t)$ is the discount factor, and T is the longest deal maturity in the portfolio.

- ▶ Here for simplicity we consider the unilateral CVA, the generalization to bilateral CVA and Debt Valuation Adjustment (DVA) or Funding Valuation Adjustment (FVA) is straightforward.

Counterparty Credit Risk Problem

- ▶ The expectation above is typically computed on a discrete time grid of 'horizon dates' $T_0 < T_1 < \dots < T_{N_O}$ as, for instance,

$$V_{\text{CVA}} \simeq \sum_{i=1}^{N_O} \mathbb{E} \left[\mathbb{I}(T_{i-1} < \tau_c \leq T_i) D(T_i) \times L_{\text{GD}}(T_i) \left(NPV(T_i) - C(R(T_i^-)) \right)^+ \right].$$

- ▶ Risk manage CVA/DVA is challenging because all the trades facing the same counterparty must be valued at the same time, typically with Monte Carlo.

A New Challenge: Rating Dependent Payoffs

- We are dealing expectation values of the form

$$V = \mathbb{E}_{\mathbb{Q}} \left[P(R, X) \right],$$

with 'payout' given by

$$P = \sum_{i=1}^{N_O} P(T_i, R(T_i), X(T_i)),$$

where

$$P(T_i, R(T_i), X(T_i)) = \sum_{r=0}^{N_R} \tilde{P}_i(X(T_i); r) \delta_{r, R(T_i)}.$$

- The Rating variable is discrete so the Payoff is not Lipschitz-continuous.

Rating Transition

- ▶ We consider the rating transition Markov chain model of Jarrow, Lando and Turnbull [10]:

$$R(T_i) = \sum_{r=1}^{N_R} \mathbb{I} \left(\tilde{Z}_i^R > Q(T_i, r) \right),$$

where \tilde{Z}_i^R is a standard normal variate, and $Q(T_i, r)$ is the quantile-threshold corresponding to the transition probability from today's rating to a rating r at time T_i .

- ▶ Note that the discussion below is not limited to this particular model, and it could be applied with minor modifications to other commonly used models describing the default time of the counterparty, and its rating.

Singular Pathwise Derivative Estimator

- ▶ Due to the discreteness of the state space of the rating factor, the pathwise estimator for its related sensitivities is not well defined.
- ▶ This can be easily seen by expressing the Payoff as

$$P(T_i, \tilde{Z}_i^R, X(T_i)) = \tilde{P}_i(X(T_i); 0)$$

$$+ \sum_{r=1}^{N_R} \left(\tilde{P}_i(X(T_i); r) - \tilde{P}_i(X(T_i); r-1) \right) \mathbb{I}(\tilde{Z}_i^R > Q(T_i, r; \theta)),$$

so that the singular contribution reads

$$\begin{aligned} \partial_{\theta_k} P(T_i, \tilde{Z}_i, X(T_i)) &= - \sum_{r=1}^{N_R} \left(\tilde{P}_i(X(T_i); r) - \tilde{P}_i(X(T_i); r-1) \right) \\ &\quad \times \delta(\tilde{Z}_i^R = Q(T_i, r; \theta)) \partial_{\theta_k} Q(T_i, r; \theta). \end{aligned}$$

- ▶ This cannot be sampled with Monte Carlo.

Singular Pathwise Derivative Estimator

- ▶ The singular contribution can be integrated out using the properties of Dirac's delta, giving after straightforward computations,

$$\bar{\theta}_k = - \sum_{r=1}^{N_R} \frac{\phi(Z^*, Z_i^X, \rho_i)}{\sqrt{i} \phi(Z_i^X, \rho_i^X)} \partial_{\theta_k} Q(T_i, r; \theta) \\ \times \left(\tilde{P}_i(X(T_i); r) - \tilde{P}_i(X(T_i); r-1) \right),$$

where Z^* is such that $(Z^* + \sum_{j=1}^{i-1} Z_j^R)/\sqrt{i} = Q(T_i, r; \theta)$, and $\phi(Z_i^X, \rho_i^X)$ is a N_X -dimensional standard normal probability density function with correlation matrix ρ_i^X obtained by removing the first row and column of ρ_i ; here $\partial_{\theta_k} Q(T_i, r; \theta)$ is not stochastic, and can be evaluated (e.g., using AAD) once per simulation.

- ▶ The final result is rather intuitive as it is given by the probability weighted sum of the discontinuities in the payout.

Test Application: CVA of a portfolio of swaps on commodity Futures

- We consider a simple one factor lognormal model for the Futures curve of the form

$$\frac{dF_T(t)}{F_T(t)} = \sigma_T \exp(-\beta(T-t)) dW_t,$$

where W_t is a standard Brownian motion; $F_T(t)$ is the price at time t of a Futures contract expiring at T ; σ_T and β define a simple instantaneous volatility function that increases approaching the contract expiry, as empirically observed for many commodities.

- As underlying portfolio, we consider a set of commodity swaps, paying on a strip of Futures (e.g., monthly) expiries t_j , $j = 1, \dots, N_e$ the amount $F_{t_j}(t_j) - K$. The net present value for this portfolio reads

$$NPV(t) = \sum_{j=1}^{N_e} D(t, t_j) (F_{t_j}(t) - K).$$

Forward and Backward Propagation

- ▶ The propagation (PROP) step for the Futures price reads:

$$F_T(T_i) = F_T(T_{i-1}) \exp \left(\sigma_i \sqrt{\Delta T_i} Z - \frac{1}{2} \sigma_i^2 \Delta T_i \right),$$

where $\Delta T_i = T_i - T_{i-1}$, and

$$\sigma_i^2 = \frac{\sigma_T^2}{2\beta\Delta T_i} e^{-2\beta T} \left(e^{2\beta T_i} - e^{2\beta T_{i-1}} \right).$$

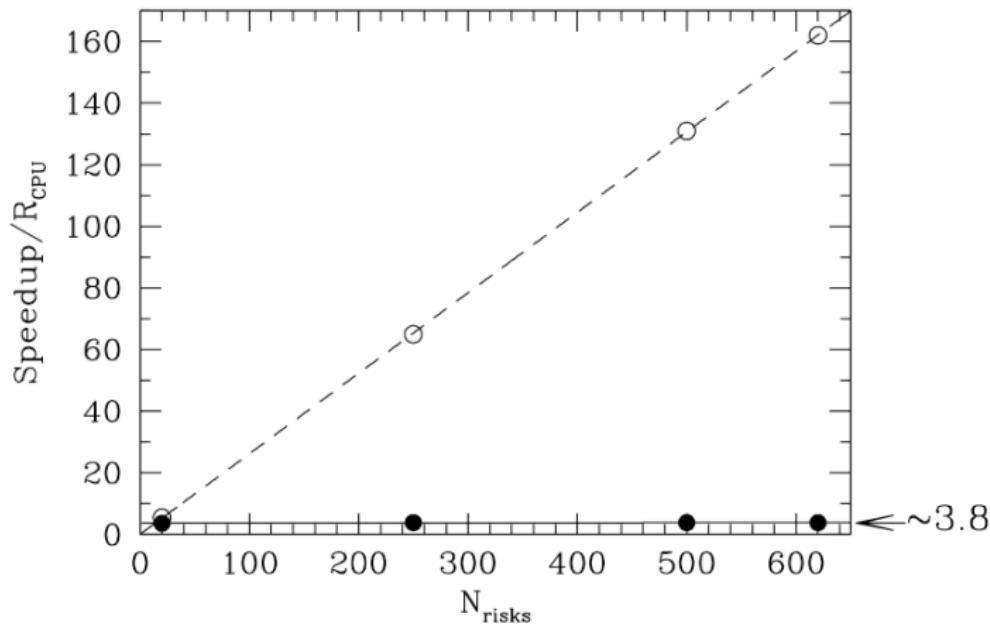
- ▶ The associated adjoint (PROP_b) reads:

$$\begin{aligned} \bar{F}_T(T_i-1) &+= \bar{F}_T(T_i) \exp \left(\sigma_i \sqrt{\Delta T_i} Z - \frac{1}{2} \sigma_i^2 \Delta T_i \right), \\ \bar{\sigma}_i &= \bar{F}_T(T_i) F(T_i) (\sqrt{\Delta T_i} Z - \sigma_i \Delta T_i) \end{aligned}$$

with

$$\bar{\sigma}_T + = \frac{\bar{\sigma}_i}{\sqrt{2\beta\Delta T_i}} \sqrt{e^{-2\beta T} \left(e^{2\beta T_i} - e^{2\beta T_{i-1}} \right)}.$$

Results



Portfolio of 5 commodity swaps over a 5 years horizon. Bumping (empty dots), AAD (full dots).

Total time: 1h 40 min (Bumping); 10 sec (AAD). From Ref.[10].

Variance Reduction

- ▶ Because of the analytic integration of the singularities, the AAD risk is typically less noisy than the one produced by Bumping.

δ	VR[Q(1,1)]	VR[Q(1,2)]	VR[Q(1,3)]
0.1	24	16	12
0.01	245	165	125
0.001	2490	1640	1350

Table: Variance reduction (VR) on the sensitivities with respect to the thresholds $Q(1, r)$ ($N_R = 3$). δ indicates the perturbation used in the finite-differences estimators of the sensitivities.

- ▶ The variance reduction can be thought of as a further speedup factor because it corresponds to the reduction in the computation time for a given statistical uncertainty on the sensitivities. This diverges as the perturbation δ tends to zero, and may be very significant even for a fairly large value of δ .

Section 8

Application to Partial Differential Equations

Beyond Monte Carlo Applications: Partial Differential Equations

- ▶ Option pricing problems can be often formulated in terms of the solution of a parabolic (backward) PDE of the form

$$\frac{\partial V}{\partial t} + \mu(x, t; \theta) \frac{\partial V}{\partial x} + \frac{1}{2} \sigma^2(x, t; \theta) \frac{\partial^2 V}{\partial x^2} - \nu(x, t; \theta) V = 0,$$

where

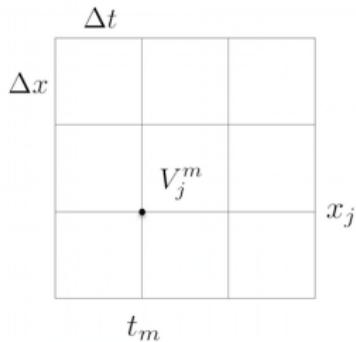
$$V = V(x_t, t; \theta) \equiv \mathbb{E} \left[e^{-\int_t^T \nu(x_u, u; \theta) du} P(x_T; \theta) \right],$$

is the value of a derivative contract at time t , with payoff at expiry $P(x_T; \theta)$. Here the risk factor x_t follows a diffusion of the form

$$dx_t = \mu(x_t, t; \theta) dt + \sigma(x_t, t; \theta) dW_t.$$

- ▶ As before, $\theta = (\theta_1, \dots, \theta_{N_\theta})$ represents the vector of N_θ model parameters the model is dependent on.

Numerical Solution by Finite-Difference Discretization



The solution $V_0(\theta) = V(x_{t_0}, t_0; \theta)$ of the backward PDE can be found numerically by discretization on the rectangular domain $(t, x) \in [t_0, T] \times [x_{\min}, x_{\max}]$:

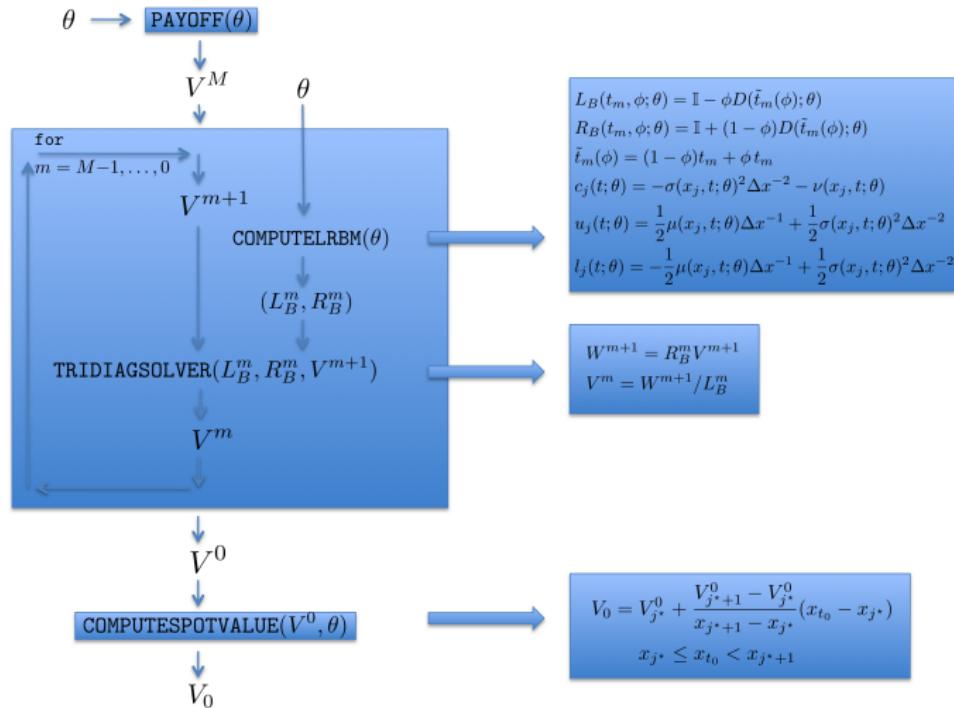
$$V^m(\theta) = (V(x_1, t_m; \theta), \dots, V(x_N, t_m; \theta))^t.$$

- Given the value of the option at expiry, $V_j^M(\theta) = P(x_j; \theta)$, the value of the option at time t_0 can be found by iterating for $m = M-1, \dots, 0$ the matrix recursion

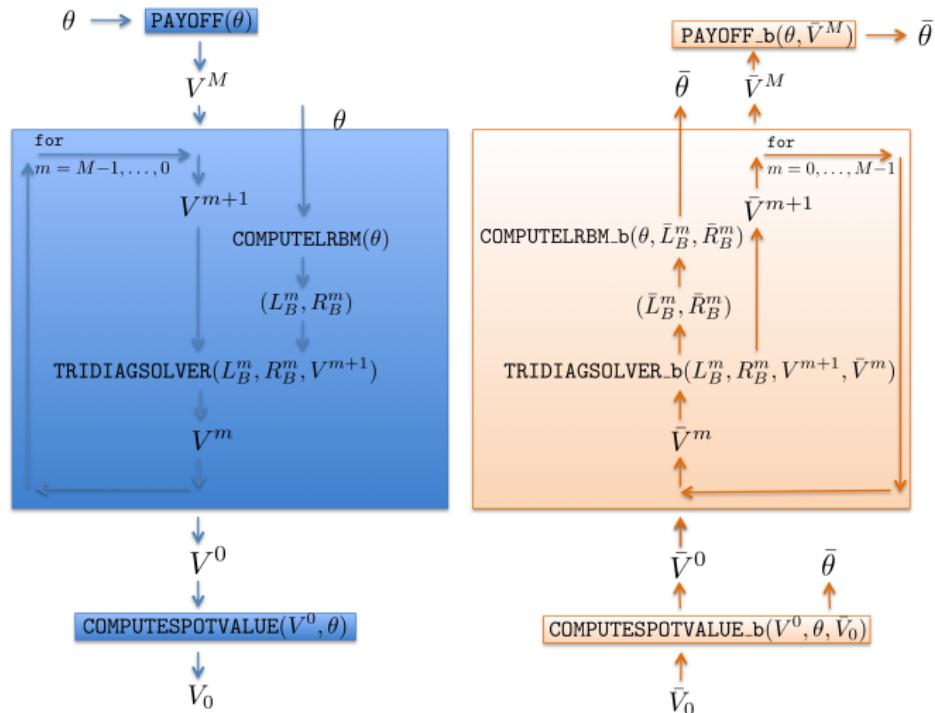
$$L_B(t_m, \phi; \theta) V^m(\theta) = R_B(t_m, \phi; \theta) V^{m+1}(\theta)$$

obtained with finite-difference approximations of the first and second derivatives in the backward PDE.

Blueprint of a Backward PDE Solver



Blueprint of an AAD Backward PDE Solver



- Using the AAD rules it is easy to pin down the structure of the Adjoint PDE solver.

The heart of the algorithm: the tridiagonal solver

$$V^m = \text{TRIDIAGSOLVER}(L_B^m, R_B^m, V^{m+1})$$

$$W^{m+1} = R_B^m V^{m+1}$$

$$V^m = W^{m+1} / L_B^m$$

$$y = Bx \quad \begin{cases} \bar{x} = A^t \bar{y} \\ \bar{A} = \bar{y} x^t \\ \bar{A} = -A^{-t} \overline{A^{-1}} A^{-t} \end{cases}$$

$$\bar{V}^{m+1} = \text{TRIDIAGSOLVER_b}(L_B^m, R_B^m, V^{m+1}, \bar{V}^m)$$

$$\bar{W}^{m+1} = [L_B^m]^{-t} \bar{V}^m$$

$$\overline{[L_B^m]^{-1}} = \bar{V}^m [W^{m+1}]^t$$

$$\bar{L}_B^m = -[L_B^m]^{-t} \overline{[L_B^m]^{-1}} [L_B^m]^{-t}$$

$$\bar{R}_B^m = \bar{W}^{m+1} [V^{m+1}]^t$$

$$\bar{V}^{m+1} = [R_B^m]^t \bar{W}^{m+1}$$

- ▶ A collection of results on the Adjoint of linear algebra operations (see M. Giles' 'Collected Matrix Derivative Results for Forward and Reverse Mode Algorithmic Differentiation') is very useful when dealing with code implementing linear algebra.
- ▶ The computational cost of the naïve AAD algorithm above is $O(N^3)$. In order to reduce the computational cost to $O(N)$, as in the original algorithm, we need to avoid the matrix inversion by using all the information that is available to us (including the forward sweep!).

The heart of the algorithm: the tridiagonal solver (cont'd)

$V^m = \text{TRIDIAGSOLVER}(L_B^m, R_B^m, V^{m+1})$

$$W^{m+1} = R_B^m V^{m+1}$$

$$V^m = W^{m+1} / L_B^m$$

$$y = Bx \quad \left\{ \begin{array}{l} \bar{x} = A^t \bar{y} \\ \bar{A} = \bar{y} x^t \\ \bar{A} = -A^{-t} \overline{A^{-1}} A^{-t} \end{array} \right.$$

$\bar{V}^{m+1} = \text{TRIDIAGSOLVER_b}(L_B^m, R_B^m, V^{m+1}, \bar{V}^m)$

$$\bar{W}^{m+1} = [L_B^m]^{-t} \bar{V}^m = \bar{V}^m / [L_B^m]^t$$

$$\overline{[L_B^m]^{-1}} = \bar{V}^m [W^{m+1}]^t$$

$$\bar{L}_B^m = -[L_B^m]^{-t} \overline{[L_B^m]^{-1}} [L_B^m]^{-t}$$

$$= -[L_B^m]^{-t} \bar{V}^m [W^{m+1}]^t [L_B^m]^{-t}$$

$$= -\bar{W}^{m+1} [[L_B^m]^{-1} W^{m+1}]^t$$

$$= -\bar{W}^{m+1} [V^m]^t$$

$$\bar{R}_B^m = \bar{W}^{m+1} [V^{m+1}]^t$$

$$\bar{V}^{m+1} = [R_B^m]^t \bar{W}^{m+1}$$

- ▶ Only the elements on the three main diagonals of \bar{L}_B^m and \bar{R}_B^m contribute to the sensitivities, so that only $3N$ multiplications are required for their computation.
- ▶ The overall computational cost of the adjoint tridiagonal solver is $O(N)$, exactly as for the forward counterpart and as expected from the general result on the computational efficiency of AAD.

Some Results: the Black-Karasinski model for default intensities

- ▶ To illustrate the efficiency of the AAD-PDE approach the Black-Karasinski (BK) model for the stochastic instantaneous hazard rate $h_t = \exp x_t$, namely

$$d \log h_t = \kappa(t)(\mu(t) - \log h_t)dt + \sigma(t)dW_t.$$

Here, we will fix the mean reversion rate $\kappa = 0.01$ and assume $\mu(t)$ and $\sigma(t)$ to be left-continuous piecewise constant functions.

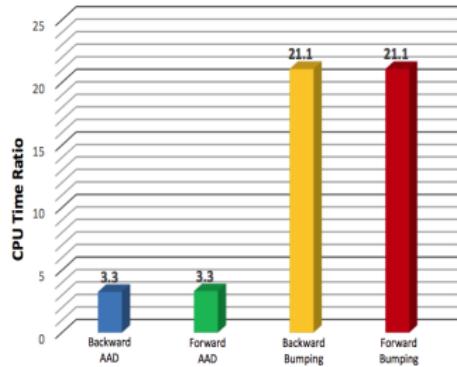
- ▶ The conditional probability of the obligor surviving up to time T is given by

$$Q(h_t, t, T) = \mathbb{E} \left[\exp \left[- \int_t^T du h_u \right] \middle| h_t, \tau > t \right].$$

- ▶ Any credit derivative whose payoff at time T is a function of the hazard rate h_T , such as defaultable bonds, CDS, bond options and CDS options can be valued within the PDE approach.

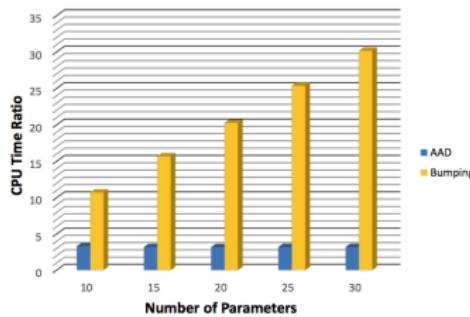
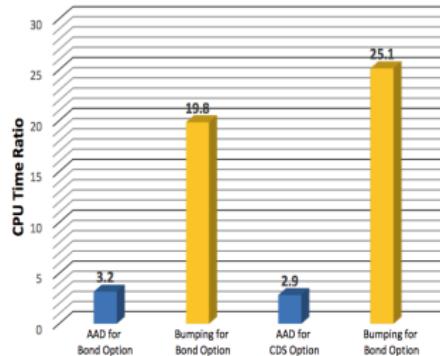
Warming up: Defaultable Zero Bond

	Bwd PDE (AAD)	Fwd PDE (AAD)	Bwd PDE (FD)	Fwd PDE (FD)
μ_1	-1.8e-4	-1.8e-4	-1.8e-4	-1.8e-4
μ_2	-1.7e-4	-1.6e-4	-1.7e-4	-1.6e-4
μ_3	-1.6e-4	-1.6e-4	-1.6e-4	-1.6e-4
μ_4	-1.5e-4	-1.5e-4	-1.5e-4	-1.5e-4
μ_5	-2.7e-4	-2.7e-4	-2.7e-4	-2.7e-4
μ_6	-2.3e-4	-2.3e-4	-2.3e-4	-2.3e-4
μ_7	-2.0e-4	-2.0e-4	-2.0e-4	-2.0e-4
μ_8	-1.6e-4	-1.6e-4	-1.6e-4	-1.6e-4
μ_9	-1.3e-4	-1.3e-4	-1.3e-4	-1.3e-4
μ_{10}	-9.1e-5	-9.1e-5	-9.2e-5	-9.1e-5
μ_{11}	-5.6e-5	-5.6e-5	-5.6e-5	-5.5e-5
μ_{12}	-1.8e-5	-1.8e-5	-1.9e-5	-1.8e-5
σ_1	-7.8e-3	-7.8e-3	-7.5e-3	-7.8e-3
σ_2	-0.011	-0.01	-0.011	-0.01
σ_3	-0.016	-0.016	-0.017	-0.016
σ_4	-0.015	-0.015	-0.015	-0.015
σ_5	-0.013	-0.013	-0.013	-0.013
σ_6	-0.012	-0.011	-0.012	-0.011
σ_7	-9.9e-3	-9.9e-3	-0.01	-9.9e-3
σ_8	-0.019	-0.019	-0.02	-0.019



- As expected, the results obtained with both the AAD version of the backward and forward PDE are consistent with the ones obtained by bumping.
- For both the AAD version of the backward and the forward PDE scheme the calculation of the sensitivities can be performed in about 3.3 times the cost of computing the value of the option, i.e., well within the theoretical bound of 4.

Bond and CDS Options



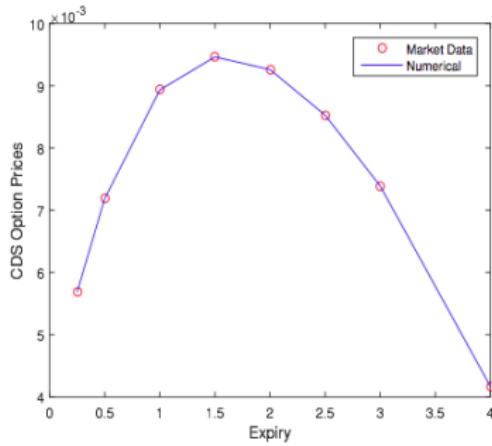
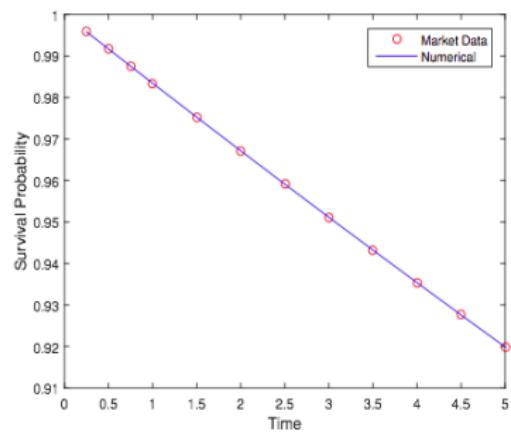
- ▶ The computational cost of the AAD algorithm is well within the theoretical bound of 4.
- ▶ The overall cost of computing all the sensitivities by means of AAD, relative to the cost of a single valuation of the option, is independent on the number of sensitivities.

Calibration: Model Parameter and Market Parameters

- ▶ The sensitivities with respect to the internal model parameters θ are generally of limited utility because they do not correspond directly to financially meaningful quantities.
- ▶ The sensitivities we need for hedging are the sensitivities with respect to the market observables M that have been used to calibrate the model.
- ▶ It is useful to think in terms of two distinct steps:

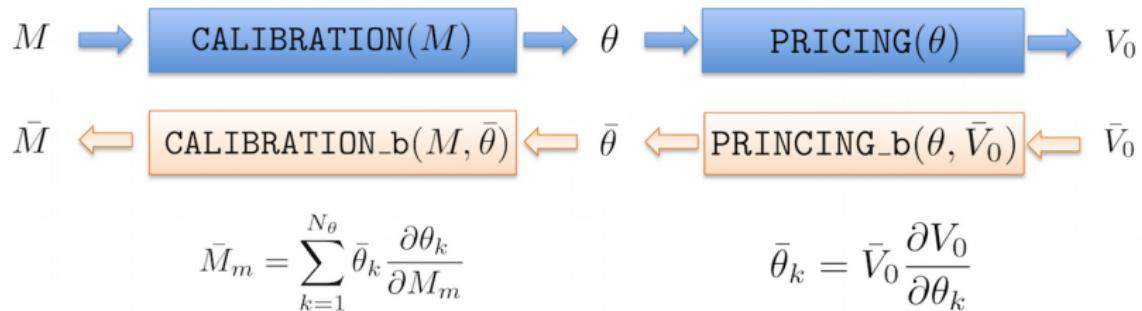


Calibration



- As it is customary we have used a combination of the forward and backward PDE algorithm to bootstrap the survival probabilities and calibrate to the CDS option prices.

Getting the Market Parameters Sensitivities



- ▶ The adjoint of the calibration step $M \rightarrow \theta(M)$ can be produced following the general rules of AAD.
- ▶ The computational cost can be expected to be of the order of the cost of performing the calibration algorithm a few (less than 4) times.
- ▶ This in itself is generally much better than bumping, involving repeating the calibration algorithm as many times as sensitivities required.
- ▶ However, we can do better thanks to ...

Ulisse Dini and the Implicit Function Theorem



Implicit Function Theorem

- ▶ The calibration algorithm consists of the numerical solution of a system of equations of the form

$$G_i(M, \theta) = 0,$$

with $M \in \mathcal{R}^{N_M}$, $\theta \in \mathcal{R}^{N_\theta}$ and $i = 1, \dots, N_\theta$, where the function $G_i(M, \theta)$ is of the form

$$G_i(M, \theta) = T_i(M) - V_i(\theta)$$

where $V_i(\theta)$ is the price of the i -th calibration instrument as produced by the model we want to calibrate, and $T_i(M)$ are the prices of the target instruments.

- ▶ By differentiating with respect to M

$$\frac{\partial G_i}{\partial M_m} + \sum_{j=1}^{N_\theta} \frac{\partial G_i}{\partial \theta_j} \frac{\partial \theta_j}{\partial M_m} = 0$$

for $m = 1, \dots, N_M$.

Implicit Function Theorem (cont'd)

- Or equivalently

$$\frac{\partial \theta_k}{\partial M_m} = - \left[\left(\frac{\partial G}{\partial \theta} \right)^{-1} \frac{\partial G}{\partial M} \right]_{km}$$

with $[\partial G / \partial M]_{ij} = \partial G_i / \partial M_j$.

- This relation allows the computation of the sensitivities of $\theta(M)$, locally defined in an implicit fashion by the calibration equation, in terms of the sensitivities of the function G . These can be computed by implementing the corresponding adjoint function

$$(\bar{M}, \bar{\theta}) = \bar{G}(M, \theta, \bar{G})$$

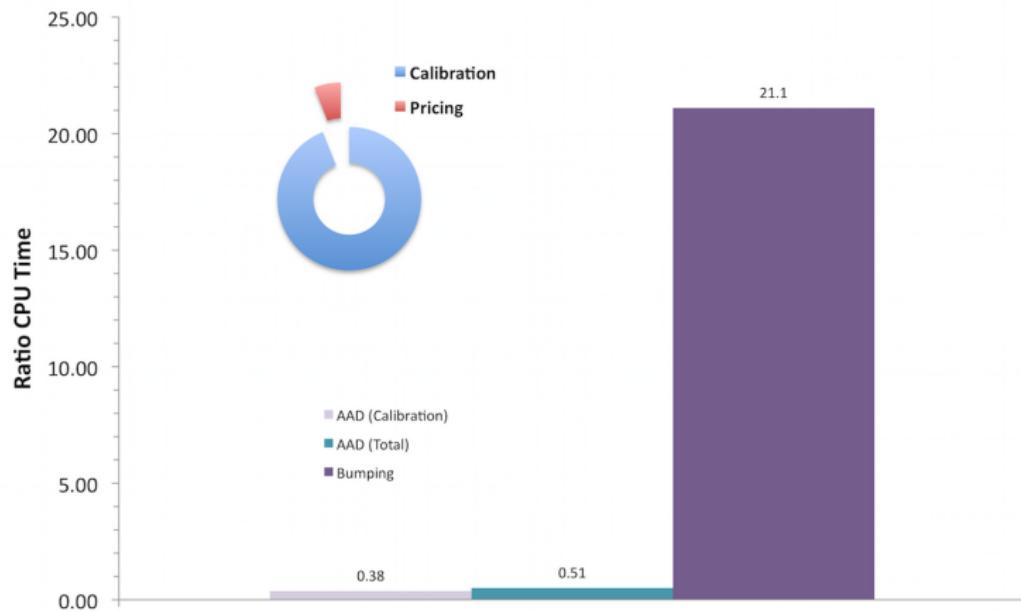
giving according to the general rule

$$\bar{M}_m = \sum_{i=1}^{N_\theta} \bar{G}_i \frac{\partial G_i}{\partial M_m} \quad \bar{\theta}_k = \sum_{i=1}^{N_\theta} \bar{G}_i \frac{\partial G_i}{\partial \theta_k}.$$

Implicit Function Theorem (cont'd)

- ▶ The Implicit Function Theorem method is significantly more stable and efficient than of calculating the derivatives of the implicit functions $M \rightarrow \theta(M)$ by applying AAD to the calibration step.
- ▶ This is because $G_i(M, \theta) = T_i(M) - V_i(\theta)$ are *explicit* functions of the model and market parameters that are easy to compute and differentiate, e.g., using the AAD version of the combination of the forward and backward PDE for the calculation of $V_i(\theta)$ and the AAD version of the algorithm for the computation of $T_i(M)$.
- ▶ Combining the implicit function theorem with AAD results in extremely efficient risk computations.

AAD and the Implicit Function Theorem: Results



- ▶ Combining AAD with the Implicit Function Theorem allows the computation of risk in 50% *less* than the cost of computing the option value, resulting in remarkable savings in computational time.

Section 9

Conclusions

Conclusions

- ▶ We have shown how Adjoint Algorithmic Differentiation (AAD) can be used to implement the Adjoint calculation of price sensitivities in a straightforward manner and in complete generality.
- ▶ In contrast to algebraic Adjoint methods, the algorithmic approach can be straightforwardly applied to both path dependent options and multi asset Monte Carlo simulations. It also eliminates altogether the need for the sometimes cumbersome analytical work required by algebraic formulations.
- ▶ AAD can be applied to any numerical technique. In particular, it can be used to implement efficiently and in full generality the calculation of sensitivities of option prices computed by means of the numerical solution of Partial Differential Equations (PDE).

Conclusions

- ▶ By combining the adjoint version of the pricing algorithm, and the Implicit Function Theorem one can avoid the necessity of repeating multiple times the calibration algorithm or implementing the AAD version of the calibration routine.
- ▶ This allows the calculation of all price sensitivities for an additional computational cost that is a small multiple of the cost of computing the P&L of the portfolio, thus typically resulting in orders of magnitudes savings in computational time with respect to standard finite-difference approach.

References I

- [1] P. Glasserman, *Monte Carlo Methods in Financial Engineering*, Springer, New York (2004).
- [2] L. Capriotti, *Fast Greeks by Algorithmic Differentiation*, J. of Computational Finance, **14**, 3 (2011).
- [3] M. Leclerc, Q. Liang and I. Schneider, *Fast Monte Carlo Bermudan Greeks*, Risk **22**, 84 (2009).
- [4] A. Griewank, *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*, Frontiers in Applied Mathematics, Philadelphia, 2000.
- [5] N. Denson and M.S. Joshi, *Fast and Accurate Greeks for the Libor Market Model*, J. of Computational Finance **14**, 115 (2011).
- [6] M. Giles and P. Glasserman, *Smoking Adjoints: Fast Monte Carlo Greeks*, Risk **19**, 88 (2006).
- [7] L. Capriotti and M. Giles, *Algorithmic Differentiation: Adjoint Greeks Made Easy*, Risk **25**, 92 (2012).
- [8] S. P. Smith, *Differentiation of Cholesky Algorithm*, J. of Computational and Graphic Statistics, **4**, 134 (1995).

References II

- [9] L. Capriotti and M. Giles, *Fast Correlation Greeks by Adjoint Algorithmic Differentiation*, Risk **23**, 79 (2010).
- [10] L. Capriotti, Jacky Lee, and Matthew Peacock, *Real Time Counterparty Credit Risk Management in Monte Carlo*, Risk **24**, 86 (2011).
- [11] L. Capriotti and J. Lee, *Adjoint Credit Risk Management*, Risk Magazine, **27**, 90 (2014).
- [12] M. Henrard, *Calibration and Implicit Function Theorem*, OpenGamma Quantitative Research, 1 (2011).
- [13] L. Capriotti, Y. Jiang, A. Macrina, *Real-Time Risk Management: An AAD-PDE Approach*, preprint (2015).

See also:

► My Publications' Page