Auto-Differentiation and Sparsity with Tomlab and Julia

Jesse Perla
University of British Columbia

November 14, 2017

Auto-Differentiation(AD)

Derivatives and Numerical Methods

There are two general types of algorithms for optimizers/solvers/etc.:

1. Derivative-free:
   - e.g. Simplex and Nelder-Mead. This is Matlab's `fminsearch`
   - Also, "costly global function" optimization and "black box" optimization where derivatives are not possible
   - Avoid at all costs (though sometimes don't have a choice)

2. Derivatives
   - Pretty much every other algorithm, especially for large number of variables/constraints
   - Including global optimization techniques (which use derivatives locally)
   - Sometimes naive use of software is generating derivatives with finite differences without your knowledge

Key derivatives to calculate are:

- Gradient of objective
- Hessian of objective (nonlinear least squares only uses gradient)
- Jacobian, and sometimes Lagrangian, of constraints

Calculating Derivatives

How to calculate derivatives for the objective and constraints?

1. Calculate by hand
   - Sometimes, though not always, the most accurate and fastest option
   - But algebra is error prone for non-trivial setups
   - (note: many optimizers have a way to check your analytical derivatives)

2. Finite-differences:
   - $\partial_{x_i} f(x_1, \ldots x_N) \approx \frac{f(x_1, \ldots x_i + \Delta, \ldots x_N) - f(x_1, \ldots x_i, \ldots x_N)}{\Delta}$
   - Evaluates function at least N extra times to get a gradient
   - $\#$ evaluations for Jacobians with M constraints even worse
   - Large Δ is numerically stable but inaccurate, small Δ is unstable
   - **Avoid like the plague!** (and is what matlab does out of the box)

3. Auto-differentiation
   - Not a form of finite-differences or numeric differentiation
   - Essentially analytical. Repeated use of the chain-rule
   - Does not work for every function, but only evaluates $f(\cdot)$ once if it works—i.e. $O(1)$ not $O(N \times M)$ for $f : \mathbb{R}^N \to \mathbb{R}^M$

Auto-differentiation (adapted from Wikipedia)

- Remember the chain rule: $\frac{dy}{dx} = \frac{dy}{dw}\frac{dw}{dx}$
- Consider functions composed of calculations with fundamental operations (with known analytical derivatives)
- For example, consider function: $f(x_1, x_2) = x_1 x_2 + \sin(x_1)$

Operations to compute value	Operations to compute $\frac{df(x_1,x_2)}{dx_1}$
$w_1 = x_1$	$\frac{dw_1}{dx_1} = 1$ (seed)
$w_2 = x_2$	$\frac{dw_2}{dx_1} = 0$ (seed)
$w_3 = w_1 \cdot w_2$	$\frac{dw_3}{dx_1} = w_2 \cdot \frac{dw_1}{dx_1} + w_1 \cdot \frac{dw_2}{dx_1}$
$w_4 = \sin w_1$	$\frac{dw_4}{dx_1} = \cos w_1 \cdot \frac{dw_1}{dx_1}$
$w_5 = w_3 + w_4$	$\frac{dw_5}{dx_1} = \frac{dw_3}{dx_1} + \frac{dw_4}{dx_1}$

- Generalizes to multiple variables. AD takes source code and generates the derivatives at the same time it calculates the function value.

Auto-differentiation With Dual Numbers

- Augment number with a second part (like complex numbers)
  - $x \to x + x'\epsilon$ where $\epsilon^2 = 0$
  - Can represent as a tuple, $\langle x, x' \rangle$
- General rule for $g(x, y)$:
  $g(\langle x, x' \rangle, \langle y, y' \rangle) = \langle g(x, y), \boldsymbol{\partial}_x g(x, y)x' + \boldsymbol{\partial}_y g(x, y)y' \rangle$
  - i.e., the chain rule for a total derivative
  - Note: concurrently calculates function and derivatives
- Example rules:

$$x + y \to \langle x, x' \rangle + \langle y, y' \rangle = \left\langle x + y, \underbrace{x' + y'}_{\boldsymbol{\partial}(x+y)=\boldsymbol{\partial}x+\boldsymbol{\partial}y} \right\rangle$$

$$xy \to \langle x, x' \rangle \times \langle y, y' \rangle = \left\langle xy, \underbrace{x'y + y'x}_{\boldsymbol{\partial}(xy)=y\boldsymbol{\partial}x+x\boldsymbol{\partial}yy} \right\rangle$$

$$\exp(x) \to \exp(\langle x, x' \rangle) = \left\langle \exp(x), \underbrace{x' \exp(x)}_{\boldsymbol{\partial}(\exp(x))=\exp(x)\boldsymbol{\partial}x} \right\rangle$$

Seeding and Operator Overloading

- Hence, an arbitrary sequence of "basic" operations can be composed
- How to start? Need to start with a dual number
  - $\langle x, 1 \rangle$: note the derivative wrt itself is 1
  - Then can apply whatever sequence of operations you want
  - Constants are implemented as $\langle c, 0 \rangle$, i.e. $\boldsymbol{\partial} c = 0$
- Operator overloading: version of basic operations for dual numbers

```
%matlab with fmad libary
x = fmad(x_val, 1); %Seeds to create a dual, since dx/dx = 1
y = exp(x); %Actually using the exp for dual numbers
z = x + y; %Actually using the + for dual numbers.
function out = f(x)
        out = sin(x); %no problem for dual numbers if sin is defined
end
q = f(y); %q still has the derivative and value calculated
```

One of the Julia Packages

```
#Auto-differentiation with ForwardDiff in Julia
using ForwardDiff
h(x) = sin(x[1]) + x[1] * x[2] + sinh(x[1] * x[2]) #multivariate.
x = [1.4 2.2]
ForwardDiff.gradient(h,x) #uses AD, seeds from x

#Or, can use complicated functions of many variables
f(x) = sum(sin, x) + prod(tan, x) * sum(sqrt, x);
g = (x) -> ForwardDiff.gradient(f, x); #New gradient function
x2 = rand(20)
g(x2) #gradient at a random 20 dim point
ForwardDiff.hessian(f,x2) #Or the hessian

#Practical note: for high dimensions(N > 100), use Reverse-Mode AD with spa
```

Implementations of AD

- A field unto itself. Do not implement directly
- Implementation is language dependent. Several approaches:
  - *Source code transformation:* utility (outside of the language itself) reads in the code for your function, and generates new source
  - *Operator Overloading:* Writes rules for dual-numbers (or equivalent). Can be magical, or can be infuriating
  - *Reverse-Mode AD:* A more complicated approach required for big problems. Can't do easily with operator overloading
- Implementation depends on the language:
  - *Fortran:* usually needs SCT. Many choices: e.g. `http://tapenade.inria.fr:8080/tapenade/index.jsp`
  - *Python:* `https://github.com/LowinData/pyautodiff` and `https://pythonhosted.org/algopy/`
  - *Matlab:* open source SCT (e.g. AdiMat) not very good. Use Tomlab/MAD instead, coupled with the Tomlab optimizer
  - *Julia:* ForwardDiff, ReverseDiff, ReverseDiffSparse (used in JuMP)

Sparsity

Sparse Matrices and Methods

- Many algorithms are specialized for matrices (or Jacobians or Hessians) with many 0s—e.g. Gaussian elimination
- Only store non-zero values, but $0 \neq 0.0$ for optimizers
- Not (usually) for storage, but rather specialized algorithms
- For Jacobians and Hessians, can solve enormous (e.g. hundreds of thousands or millions) of variable systems
  - But the more non-zeros, the more likely dense methods are preferable.
- For example, $f : \mathbb{R}^N \to \mathbb{R}^N$ with $f(x) = \sqrt{x}$ point-wise
  - Jacobian has N non-zeros, while dense has N^2
  - Optimizers/solvers can use this to step in the right direction
  - Auto-differentiation will figure out the sparsity pattern of derivatives—i.e., which values are always 0 for all inputs

Sparse Matrices in Matlab

```
%First, can convert dense matrix, and it drops the 0's.
X = [1.0 0   0;
     2.0 1.0 0];
S = sparse(X)
%S =
%(1,1)        1
%(2,1)        2
%(2,2)        1

%Or can take lists of indices and values,
x_indices = [1; 2; 2];
y_indices = [1; 1; 2];
values = [1; 2; 1];
S2 = sparse(x_indices, y_indices, values)

%Or can preallocate and just reference in loops/etc.
S3 = sparse(0,3);
S3(1,1) = 1;
S3(2,1) = 2;
S3(2,2) = 1;
```

Tomlab Examples

What is in Tomlab?

- Sadly, the Operations Research community keeps the best implementations closed-source
- Collection of sparse/dense linear/nonlinear local/global constrained/unconstrained continuous/mixed-integer optimizers
- Nonlinear methods have built in auto-differentiation
- Repackages and resells state-of-the-art commercial products, and adds a few of its own (which tend to be high quality)
- Several methods to solve the same type of problem, because you never know which one will work best. Easy to swap

What Types of Problems?

See `http://tomopt.com/docs/TOMLAB_QUICKGUIDE.pdf`.

- Programming = Optimizer in OR
- Linear Programming (LP) and Mixed-Integer LP (MILP)
- Constrained Nonlinear Programming (NLP)
- Unconstrained Global Optimization (glb)
- Linear Least Squares (LLS)
- Nonlinear Least Squares (NLLS)
- Solving systems of equations generally uses NLLS
- . . . and many others (semi-definite, quadratic, etc.)

Most have sparse vs. dense algorithms, and constrained vs. unconstrained

- Read docs to find best fit for your particular problem
- Always use appropriate constraints (none, box-bounded, linear, etc.)
- For borderline sparse problems, sometimes dense works better

Some Packages

See `http://tomopt.com/tomlab/products/`

- Stanford Systems Optimization Laboratory (SOL): SNOPT, NPSOL, NLSSOL, LSSOL...
- Knitro. Good for big problems, and complementarity conditions
- MAD (auto-differentiation)
- LGO and CGO (global optimizers, costly and otherwise)

After installing tomlab, open up matlab and type `tomRun` to get a list of all licensed (and recommended) solvers by problem type

Linear Least Squares (i.e. Regression)

$$\begin{aligned} \min_x & \frac{1}{2} \|Cx - d\|_2 \\ \text{s.t. } & x_L \leq x \leq x_U \\ & b_L \leq Ax \leq b_U \end{aligned}$$

- Little benefit over stata until problems get large or sparse
- Though if there are linear constraints, A, may be helpful
- Major benefit for large, sparse C
- See "Section 11. LLS Problem" in `http://tomopt.com/docs/TOMLAB_QUICKGUIDE.pdf`
- Matlab also has a built in sparse linear least squares solver

Example: Two-way fixed Effect

- Use student, i, and instructor, j fixed effects with observables

$$\text{grade}_{ij} = \text{observables}_{ij} + \text{student}_i + \text{instuctor}_j + \epsilon_{ij}$$

- 300K observations for 36K students and 945 instructors
  - ≈ 37K variables, **but only** 2 **non-zero** (plus observables)
- Took about a day to solve in Stata
- See `teacher_student_fixed_effect.m` example for generating sparse matrix. Given `id1` student id, and `id1` instructor id. Key code:

```
%Preallocate a sparse matrix
%Total number of observables with indicators for the two types.
X = sparse(N_observations, N_observables + N_students + N_teachers);

%Filling in indicators for the matches
for i=1:N_observations
X(i, N_observables + id1(i)) = 1; %set student indicator
X(i, N_observables + N_students + id2(i)) = 1; %sets instructor indicator
end
```

Solving LLS in Tomlab

- Given X and y such that $\min_\beta \|X\beta - y\|_2$
- Ensure X loaded sparse, with each row having 2 indicators

```
%linear least squares, can pass in sparse matrices or use dense
help llsAssign %Can see options, if you wish.  Or use manual

% tomlab convention, call XXXAssign for problem type XXX
Prob = llsAssign(X, y, [], [], 'LLS Example');

%Can change settings.  See tomlab documentation
Prob.optParam.MaxIter = 5000; %optional, increase iterations
Prob.PriLevOpt  = 1; %optional, gives more information if higher

%Tomlab convention: tomRun, passing in the algorithm type and problem
%Takes about 10-20 seconds to run, instead of a day.  Not even tweaked
Result = tomRun('Tlsqr', Prob); %intended for sparse unconstrained LLS
beta = Result.x_k;

%Tried alternative methods.  Easy to swap
%Result = tomRun('snopt', Prob); %Tlsqr works much better here
```

Constrained Optimization

Nonlinear Programming (NLP)

Given $f : \mathbb{R}^N \to \mathbb{R}$ and $c : \mathbb{R}^N \to \mathbb{R}^M$

$$\begin{aligned} &\min_x \{f(x)\} \\ &\text{s.t. } x_L \le x \le x_U \\ &\quad b_L \le Ax \le b_U \\ &\quad c_L \le c(x) \le c_U \end{aligned}$$

- See "Section 7. NLP Problem" in `http://tomopt.com/docs/TOMLAB_QUICKGUIDE.pdf`
- Can solve large problems in general with dense solvers (e.g. tens, hundreds, thousands of variables depending on structure)
- Can solve enormous problems if Hessian of $f(x)$, Jacobian of $c(x)$ sparse (e.g. hundreds of thousands or millions of variables)
- If constraints are bounds, use x_L, x_U. If linear, use A. Equality set $b_L(m) = b_U(m)$ where appropriate. No bounds, leave unconstrained

Example: Porfolio Choice under Rational Inattention

- Entropy constraint, κ and signal variances, σ_n, weights α_n
- Choose precision x_n

$$\min_{\{x_n\}} \sum_{n=1}^{N} \alpha_n^2 x_n^2$$

$$\text{s.t. } \frac{1}{2} \sum_{n=1}^{N} \left(\log \sigma_n^2 - \log x_n^2 \right) \leq \kappa$$

$$0 < x_n < \infty$$

- Generate some $\{\alpha_n, \sigma_n\}$ and solve
- How large for N? Using `fminsearch` naively, got to $N \approx 30$. Sample code I give you solves with $N = 100,000$ in ≈ 20 seconds
- Since the hessian is sparse, could potentially use specialized methods (but didn't even bother to play with that, just using AD)

Solving in Tomlab (with Anonymous Functions)

```
%...define kappa, alpha vector of length N, sigma vector of length n...
objective = @(x) sum((x.^2).*(alpha.^2)); %Objective
constraint = @(x) (1/2)*sum( log(sigma.^2) - log(x.^2) ); %Constraint

%Bounds on x, and initial guess
x_L = 1E-10 * ones(N,1); %Bounding a little above 0 since it takes logs.
x_U = inf(N,1); %Upper bounds for x.
x_0 = 1.1*ones(N,1); %Some initial conditions

%Generate problem.  Use `help conAssign' to see arguments, or use manual
madinitglobals; %Need to run for auto-differentiation to work.
Prob = conAssign(objective, [], [], [], x_L, x_U, 'Portfolio example',...
   x_0, [], [], [], [], [], constraint, [], [], [], kappa, kappa);
Prob.ADObj = 1; % Gradient with AD. ADObj = -1 for Hessian
Prob.ADCons = 1; % Jacobian with AD. ADCons = -1 for constraint Lagrangian

%Run type
Result = tomRun('knitro', Prob); %Choose algorithm. See tomlab for options
x_optimal = Result.x_k;

%Result = tomRun('npsol', Prob, 1) %Could try another algorithm
```

Systems of Equations and NLLS

Nonlinear Least Squares (NLLS)

Given residual $r : \mathbb{R}^N \to \mathbb{R}^M$ and $c : \mathbb{R}^N \to \mathbb{R}^P$

$$\min_x \left\{ \frac{1}{2} r(x)^T r(x) \right\}$$
$$\text{s.t. } x_L \leq x \leq x_U$$
$$b_L \leq Ax \leq b_U$$
$$c_L \leq c(x) \leq c_U$$

Also for solving system of equations (potentially with inequalities):

$$r(x) = \mathbf{0}$$

- See "Section 13. NLLS Problem" in `http://tomopt.com/docs/TOMLAB_QUICKGUIDE.pdf`
- Can solve very large problems, with/without constraints

Solving NLLS

```
%Residual
r = @(x) x.^2 - [2; 1];
x_0 = [5; 10];
%Create object
Prob = clsAssign(r, [], [], [], [], 'NLLS example', x_0, zeros(2,1),[]);
Prob.ADObj = 1; % Use AD

%Solve it.
Result = tomRun('nlssol', Prob, 1);
```

Julia Examples

Auto-differentiation (Forward)

```
#Auto-differentiation with ForwardDiff in Julia
using ForwardDiff
h(x) = sin(x[1]) + x[1] * x[2] + sinh(x[1] * x[2]) #multivariate.
x = [1.4 2.2]
ForwardDiff.gradient(h,x) #uses AD, seeds from x

#Or, can use complicated functions,
f(x) = sum(sin, x) + prod(tan, x) * sum(sqrt, x);
g = (x) -> ForwardDiff.gradient(f, x); #New gradient function
x2 = rand(20)
g(x2) #gradient at a random 20 dim point
ForwardDiff.hessian(f,x2) #Or the hessian

#See https://github.com/JuliaDiff/ReverseDiff.jl
```

Optimization

$$\min_{x,y} \left\{ (1-x)^2 + 100(y-x^2)^2 \right\}$$
$$\text{s.t. } x + y = 10$$

```
using JuMP,Ipopt
m = Model(solver = IpoptSolver())
@variable(m, x, start = 0.0)
@variable(m, y, start = 0.0)

@NLobjective(m, Min, (1-x)^2 + 100(y-x^2)^2)

solve(m)
println("x = ", getvalue(x), " y = ", getvalue(y))

# adding a (linear) constraint
@constraint(m, x + y == 10)
solve(m)
println("x = ", getvalue(x), " y = ", getvalue(y))
```

Complicated Function with AD

$$\max_{x \in \mathbb{R}^2} \{x(1) + x(2)\}$$

$$\text{s.t. } \sqrt{x(1)^2 + x(2)^2} \leq 1$$

```
using JuMP, Ipopt
#Can auto-differentiate complicated functions with embedded iterations
function squareroot(x) #pretending we don't know sqrt()
        z = x # Initial starting point for Newton's method
        while abs(z*z - x) > 1e-13
                z = z - (z*z-x)/(2z)
        end
        return z
end
m = Model(solver = IpoptSolver())
JuMP.register(m,:squareroot, 1, squareroot, autodiff=true) # For user defin
@variable(m, x[1:2], start=0.5)
@objective(m, Max, sum(x))
@NLconstraint(m, squareroot(x[1]^2+x[2]^2) <= 1)
solve(m)
```

Solving Functional Equations

- One important use of this is solving functional equations of the form

$$\Phi(f) = \mathbf{0}$$

  - Where $f : \mathbb{R}^N \to \mathbb{R}^M$ and Φ is an operator $\Phi : \mathbb{C}(\mathbb{R}^N) \to \mathbb{R}^P$
  - Could include differential equations, difference equations, etc.
- To solve, can approximate f with a basis. Collocation methods, etc.
  - e.g. $f(x) \approx \sum_{q=1}^{Q} d_q P_q(x)$
  - Where $P_q(x)$ is a polynomial/spline/finite-element basis
  - d_q are the unknown coefficients to solve for
- Then, problem is to find the coefficients

$$\min_d \left\{ \frac{1}{2} \Phi(d)^T \Phi(d) \right\}$$

- Since for fixed x_n nodes, can usually evaluate P_q through linear algebra can use AD on $\Phi(d)$
- Note: if $\Phi(d)$ is a linear operator (e.g. linear PDEs) can use sparse LLS. How huge numerical PDEs with millions of points are solved.

Example: Joint HJBE and KFE

For example, solving a problem like

$$rv(z) = \pi(z) + \mu v'(z) + \frac{\sigma^2}{2} v''(z)$$

With the KFE

$$0 = -\mu f'(z) + \frac{\sigma^2}{2} f''(z) + \text{stuff}$$

With boundary values, integral constraints, etc.

$$1 = \int f(z) \mathrm{d}z$$

Approach (though in this case, since linear, use finite-differences):

- Use chebyshev basis for $v(z)$ and $f(z)$
- Define a system of equations in coefficients. Naively stack
- Rely on auto-differention to find jacobians of big system
- Throw a commerical solver at it supporting big sparse systems

Additional Information and Hints

AD with Tomlab

Using AD Directly in Tomlab

- Keep in mind that optimizers/solvers in Tomlab do this automatically
- But if having trouble with optimizer calls, can test function separately

```
%Example function.  Also works fine with separate files/function defs
f = @(x) 3*x + exp(x);

%Evaluating function
x_val = 2.1;
f(x_val)

%Evaluating with derivative at x_val
x = fmad(x_val, 1); %Seed, since dx/dx = 1
f_val = f(x)

%Extract (both calculated at same time)
getvalue(f_val)
getderivs(f_val)
```

Black Magic? Is it Always so Easy?

- Auto-differentiation works seamlessly for functions composed of an arbitrarily complicated graph of simple functions
  - Just need analytical derivatives for the lowest-level functions
  - Functions of vector and matrices are no problem at all. In fact, the field was designed for large numbers of variables/constraints and sparsity
- Can you call other functions (with operator overloading)?
  - Depends on how they were written. Often no problem at all
  - If the functions assume arguments are numbers, there can be problems
  - Sometimes can fix the underlying code to make more generic
- Verboten: Iterations and fixed-points *within a function*
  - e.g. it can't differentiate a nested optimization step within a function
  - However, many algorithms can be re-written without nesting (e.g. nested fixed-point vs. MPEC for discrete-choice estimation)
  - Possible that simulation could be embedded (e.g. mixed-logit) but have never tried it

Keep Functions Generic

- Remember, MAD replaces arguments with things that look like variables. Keep everything generic, don't overwrite with other types
- Some internal matlab functions do this sort of thing
- Sometimes can copy/paste others sourcecode and tweak

```
madinitglobals; %Need to run for auto-differentiation to work
x_val = [2.1;3.0];
x = fmad(x_val, eye(2,2)); %Seeds with derivatives

%Extract (both calculated at same time)
f_val = f_func(x)

function y = f_func(x)
  y = x.^2; %This leaves x, y generic
  %x = 1; %Don't do this!!!!!!
  %y = zeros(1,1) %Don't do this!!!!!
  %y(1) = x.^2; %indexing is fine (as long as you do not preallocate)
  %One trick is to allocate as something like: y = x, then index
end
```

Missing Function (with Analytical Derivative)

- See MAD manual, `http://tomopt.com/docs/TOMLAB_MAD.pdf`
- Section: `"Adding Functions to the fmad Class"`
- See extensions in in `https://github.com/econtoolkit/tomlab/MAD`
- Example, `normcdf` isn't there, could add something like:

```
%Add to the appropriate location in tomlab
%This should work for vectors/matrices since using .*
%This has not been sufficiently tested!
function y=normcdf(x)
        y = x; %Needs to copy
        y.value= normcdf(x.value); %Evaluate given double
        y.deriv = normpdf(x.value) .* x.deriv; %Note chain rule
end
```

Using Tomlab External Functions and Fixed Parameters

```
%Alternatively, the objective/constraint can be in separate files
%Assumption is that vectors alpha/sigma/kappa attached to problem

%File: portfolio_objective.m
function f = portfolio_objective(x, Prob)
f = sum((x.^2).*(Prob.alpha .^2));
end

%File: portfolio_constraint.m
function c = portfolio_constraint(x, Prob)
        c = (1/2)*sum( log(Prob.sigma.^2) - log(x.^2) ) - Prob.kappa;
end
```

Calling with External Functions

```
%Changed to have c(x) = 0 for the constraint since in c(x)
Prob = conAssign(@portfolio_objective, [], [], [], x_L, x_U, 'Example',...
x_0, [], [], [], [], [], @portfolio_constraint, [], [], [], 0, 0);

%Put any constants into the Prob, available in the function
%Can throw anything you want only Prob (e.g. vectors, cells, etc.)
Prob.alpha = alpha;
Prob.sigma = sigma;
Prob.kappa = kappa;

%Setup AD
Prob.ADObj = 1; % Gradient with AD. ADObj = -1 for Hessian
Prob.ADCons = 1; % Jacobian with AD. ADCons = -1 for constraint Lagrangian

%Run the optimizer
Result = tomRun('knitro', Prob); %The last
```