SANDIA REPORT

SAND2016-12293
Unlimited Release
Printed Dec 2016

A harmonic balance method for PDEs

Andy C. Huang

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-mission laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Reviewed by: Lawrence C Musson _____
Date: 30 November 2016

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

SAND2016-12293
Unlimited Release
Printed Dec 2016

A harmonic balance method for PDEs

Andy C. Huang
Electrical Models and Simulation
Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185-9999
ahuang@sandia.gov
MS 1177

Abstract

In this report, we describe some approaches to calculate the non-linear system of equations prescribed by the harmonic balance method (HB), a frequency domain analysis technique for modelling a non-linear system of partial differential equations (PDEs). The approach which we ultimately pursue can be seen as a time-collocation approach, except that the harmonic balance equations are obtained weakly (in the sense used in the calculus of variations). This weak formulation allows us to adapt existing transient or stationary PDEs models in the Panzer/Trilinos framework for frequency domain analysis via the harmonic balance method.

We begin with a motivatation for the harmonic balance method and outline its mathematical formulation. We then describe some approaches to calculate the harmonic balance formulae, and their means of implementation through the modification of a Panzer tutorial problem - a *s*tationary Helmholtz equation with a constant Dirichlet boundary condition and a non-linear source. For each of these approaches, we outline the necessary adaptations to solve the corresponding (periodically) transient Helmholtz equation with a (temporally) periodic Dirichlet boundary condition and non-linear source.

Acknowledgment

The author would like the especially thank the Charon team - Suzey Gao, Gary Hennigan, Larry Musson, and Mihai Negoita - for their encouragement and guidance. They have all been excellent buddies and mentors. Thanks are also due to many Trilinos developers from whom the author has shared patient, helpful, and insightful conversations: Eric Phipps, Roger Pawlowski, Stephen Bond, Kara Peterson, and Jason Gates.

Contents

Figures

Tables

This page intentionally left blank.

1 Background

The Harmonic Balance method (HB) is a technique applied to approximate periodic steady state solutions of transient partial differential equations (PDEs). Like a standard finite element method (FEM), it approximates the PDE spatially by an algebraic system of equations. Whereas a standard time-domain FEM uses, say, an implicit Backward-Euler method to march forward in time, the HB method incorporates the time dimension as part of the FEM scheme by resolving the temporal dimension with periodic sine and cosine functions at each spatial node. HB is particularly efficient when the forcing term is quasi-periodic (in time) with few modes, and also when the degree of non-linearity in the PDE is not too great. We note these considerations below, comparing its advantages to alternative approaches such as the shooting method and the standard FEM.

1.1 Motivation for the HB method

A standard transient FEM can be used to approximate the periodic steady state solution of a PDE by allowing the simulation to run until the transient behavior all but disappears. However, this approach must overcome some major difficulties:

1. Too many time steps may be required in order for the system to (nearly) reach its periodic steady-state. If this is the case, simulating an adequte number of time steps can be prohibitively expensive.
2. Simulation of a transient system with a mixed frequency input may require an extremely small time step size to resolve great oscillations. This is the case, for example, if a multi-tone input signal contains widely spaced frequencies f_1 and f_2, for which a time step size of $\frac{1}{f_1 f_2}$ is recommended.

Troyanovsky implements a harmonic balance method to address these points for the problem of simulating distortion in RF/microwave semiconductor devices [Tro98]. The key advantages of HB over a standard time domain FEM with (e.g., Backward-Euler) transient approximation are that HB:

1. directly captures the steady state response,
2. performs equally robustly for two spectral tone inputs with different spacing, and
3. efficiently handles distributed linear components.

There is no free lunch, though. HB requires a larger number of degrees of freedom. For a standard piecewise linear FEM on N spatial nodes (and hence N degrees of freedom per dependent variable), and for which there are H frequencies to be accounted for, the HB method results in a system of equations involving $N \cdot (2H + 1)$ degrees of freedom - one cosine and one sine per fundamental harmonic, plus the constant offset (the average over the period) - per dependent variable. Nonetheless, Troyanovsky claims HB has advantages in certain situations, and he optimizes the non-linear solver method to combat this larger number of degrees of freedom.

1.2 Mathematical overview of the HB method

The commonplace harmonic balance method is also appropriately called the harmonic Newton method. The HB principle is that of representing all functions in a truncated Fourier expansion (hence "harmonic") and then equating the sine and cosine coefficients (hence "balance"). This typically results in a non-linear algebraic system which is then solved using the Newton method. We note that finite Fourier series are usually called *quasi-periodic* signals in this context.

Suppose we are solving a non-linear (real-valued) PDE

$$\frac{\partial u}{\partial t} = \mathfrak{F}(\nabla u, u, \mathbf{x}, t)$$

for a function $u(\mathbf{x}, t)$ which we expect to be periodic. For example, consider the drift-diffusion operator

$$\mathfrak{F} \equiv \nabla \cdot (D\nabla u) + \boldsymbol{\beta} \cdot \nabla u + R(u) + F(t)$$

with a periodic forcing term $F(t)$ and a non-linear reaction term $R(u)$.

The HB principle adopts the ansatz that the solution is well-approximated by the real part of a finite Fourier series, expressible in a basis of sines and cosines:

$$\begin{aligned} u(\mathbf{x}, t) &= \mathfrak{Re}\left(\sum_{h=-H}^{H} \widehat{u}_{\omega_n}(\mathbf{x}) \cdot e^{i\omega_n t} \right) \\ &= u_0(\mathbf{x}) + \sum_{h=1}^{H} u^R_{\omega_n}(\mathbf{x}) \cdot cos(\omega_n t) + \sum_{h=1}^{H} u^I_{\omega_n}(\mathbf{x}) \cdot sin(\omega_n t) \end{aligned}$$

for some collection of *non-negative* frequencies $\Omega \equiv \{\omega_0 = 0, \omega_1, \ldots, \omega_H\}$ satisfying $\omega_1 > \omega_2 > \cdots > \omega_H$. The choice of a collection of frequencies is called the *truncation scheme* of the harmonic balance method.

The HB technique exploits the following algebraic and differential trigonometric identities:

$$\text{resolves spatial non-linearities} \begin{cases} \sin(\omega t)\sin(\eta t) = \frac{1}{2}\left[\cos\left((\omega - \eta)\, t\right) - \cos\left((\omega + \eta)\, t\right)\right] \\ \sin(\omega t)\cos(\eta t) = \frac{1}{2}\left[\sin\left((\omega - \eta)\, t\right) + \sin\left((\omega + \eta)\, t\right)\right] \\ \cos(\omega t)\cos(\eta t) = \frac{1}{2}\left[\cos\left((\omega - \eta)\, t\right) + \cos\left((\omega + \eta)\, t\right)\right] \end{cases}$$

$$\text{periodicity linearizes temporal derivative} \begin{cases} \frac{d}{dt} cos(\omega t) = -\omega \cdot sin(\omega t) \\ \frac{d}{dt} sin(\omega t) = \omega \cdot cos(\omega t) \end{cases}$$

Of course, the temporal derivative operator (and any of its powers) becomes a linear operator in the frequency domain. The spatial non-linearities in the system of PDEs are expressible linearly in a sine and cosine basis by applying the product-to-sum identities.

For example, if the reaction term is $R(u) = u^2$ and if we adopt a solution ansatz includin an ω harmonic, then the reaction term contains a $sin(\omega t)cos(\omega t)$ as a summand. This term can be captured by the ansatz (i.e., can itself be expressed in the form of the ansatz) by the inclusion of $sin(2\omega t)$ in the HB basis. This suggests that, if we'd like to resolve a second-order interaction and if we have ω in the frequency basis of the forcing term $F(t)$, we should at least include 2ω as well.

Furthermore, we can restrict the frequency basis to be entirely non-negative because the evenness and odd-ness of cosine and sine provide the useful identities

$$cos(\omega) = cos(-\omega)$$
$$sin(\omega) = -sin(-\omega)$$

whenever ω is negative, so that the result is expressible in positive frequency harmonics.

Because of these properties, it is common to take Ω of the truncation scheme to be a subset of the integer linear combinations of some fundamental frequencies. For example, if the forcing term is given by

$$F(t) = \sum_{i=1}^{M} F^R_{\omega_i} \left(cos(\omega_i) + F^I_{\omega_i} sin(\omega_i) \right),$$

then it is reasonable to take Ω to be a subset of $\{k_1\omega_1 + k_2\omega_2 + \cdots + k_M\omega_M | k_i \in \mathbb{Z}, \sum_{i=i}^{M} k_i\omega_i \geq 0\}$. With respect to this truncation scheme, we say that the frequency $\eta = k_1\omega_1 + k_2\omega_2 + \cdots + k_M\omega_M$ has:

$$\begin{aligned} &order \quad ord(\omega) := |k_1| + |k_2| + \cdots + |k_M| \\ &index \quad ind(\omega) := (k_1, k_2, \ldots, k_M) \end{aligned}$$

We say that η comes from *intermodulation* if more than one k_i is non-zero. The *index set of the truncation scheme* is the set of indices which record the frequencies in the truncation scheme, $\mathcal{I} = \{(k_1, k_2, \ldots, k_M) | \sum_{i=1}^{M} k_i\omega_i \in \Omega\}$. The *order of the truncation scheme* is equal to $\max_{\omega \in \Omega} ord(\omega)$, and the *fundamental frequencies of the truncation scheme* are $\omega_1, \omega_2, \ldots, \omega_M$ - which themselves have indices $(1,0,\ldots,0)$, $(0,1,0,\ldots,0)$, $\ldots$, $(0,\ldots,0,1)$. We will adopt a vector notation for indices and the fundamental frequencies, so that $\vec{\omega} := (\omega_1, \omega_2, \ldots, \omega_M)$.

With this notation, it is sensible to alternatively express the ansatz concisely as

$$u(\mathbf{x}, t) = u_0(\mathbf{x}) + \sum_{\vec{k} \in \mathcal{I}} u^R_{\vec{k} \cdot \vec{\omega}}(\mathbf{x}) \cdot cos(\vec{k} \cdot \vec{\omega} t) + \sum_{\vec{k} \in \mathcal{I}} u^I_{\vec{k} \cdot \vec{\omega}}(\mathbf{x}) \cdot sin(\vec{k} \cdot \vec{\omega} t).$$

1.3 Example: Lotka-Volterra system

Consider the simple first-order, non-linear Lotka-Volterra predator-prey model, a system of transient ODEs with a periodic solution (for appropriate parameters):

$$\begin{aligned} \frac{dx}{dt} &= \alpha x - \beta xy \\ \frac{dy}{dt} &= -\gamma y + \delta xy \end{aligned} \tag{1}$$

Suppose we are only interested in a single mode ω of the system, truncating at $H = 1$ and hence only considering first-order interactions. Applying the HB principle, we arrive at the quasi-periodic forms:

$$x(t) = X_0 + X^R_\omega cos(\omega t) + X^I_\omega sin(\omega t)$$
$$y(t) = Y_0 + Y^R_\omega cos(\omega t) + Y^I_\omega sin(\omega t)$$

We compute the terms appearing in the Lotka-Volterra equations. The temporal derivatives are simply:

$$\frac{dx}{dt} = -\omega X_\omega^R sin(\omega t) + \omega X_\omega^I cos(\omega t)$$
$$\frac{dy}{dt} = -\omega Y_\omega^R sin(\omega t) + \omega Y_\omega^I cos(\omega t)$$

Whereas the interaction term is given by

$$\begin{aligned} x \cdot y = {} & X_0 Y_0 + X_0 Y_\omega^R cos(\omega t) + X_0 Y_\omega^I sin(\omega t) \\ & + X_\omega^R Y_0 cos(\omega t) + X_\omega^R Y_\omega^R cos(\omega t) cos(\omega t) + X_\omega^R Y_\omega^I cos(\omega t) sin(\omega t) \\ & + X_\omega^I Y_0 sin(\omega t) + X_\omega^I Y_\omega^R sin(\omega t) cos(\omega t) + X_\omega^I Y_\omega^I sin(\omega t) sin(\omega t), \end{aligned}$$

and becomes (by applying the product-to-sum rules)

$$\begin{aligned} x \cdot y = {} & X_0 Y_0 + X_0 Y_\omega^R cos(\omega t) + X_0 Y_\omega^I sin(\omega t) \\ & + X_\omega^R Y_0 cos(\omega t) + X_\omega^R Y_\omega^R \frac{1}{2} \left[cos(2\omega t) + 1 \right] + X_\omega^R Y_\omega^I \frac{1}{2} sin(2\omega t) \\ & + X_\omega^I Y_0 sin(\omega t) + X_\omega^I Y_\omega^R \frac{1}{2} sin(2\omega t) + X_\omega^I Y_\omega^I \frac{1}{2} \left[1 - cos(2\omega t) \right], \end{aligned}$$

which is ultimately expressible in the $\{1, cos(1\omega t), sin(1\omega t), cos(2\omega t), sin(2\omega t)\}$ basis as

$$\begin{aligned} x \cdot y = {} & \left[X_0 Y_0 + \frac{1}{2} X_\omega^R Y_\omega^R + \frac{1}{2} X_\omega^I Y_\omega^I \right] + \left[X_0 Y_\omega^R + X_\omega^R Y_0 \right] cos(\omega t) + \left[X_0 Y_\omega^I + X_\omega^I Y_0 \right] sin(\omega t) \\ & + \left[\frac{1}{2} X_\omega^R Y_\omega^R - \frac{1}{2} X_\omega^I Y_\omega^I \right] cos(2\omega t) + \left[\frac{1}{2} X_\omega^R Y_\omega^I + \frac{1}{2} X_\omega^I Y_\omega^R \right] sin(2\omega t). \end{aligned}$$

Finally, by realizing the Lotka-Volterra system as

$$\text{prey equation} \left\{ \frac{dx}{dt} - \alpha x + \beta xy = 0 \right.$$
$$\text{predator equation} \left\{ \frac{dy}{dt} + \gamma y - \delta xy = 0 \right.$$

we arrive at the following system of non-linear equations (by balancing the amplitudes of like harmonics):

$$\begin{aligned} \text{prey equation} & \begin{cases} cos(0\omega t) \text{ term} & 0 = -\alpha X_0 + \beta \left[X_0 Y_0 + \frac{1}{2} X_\omega^R Y_\omega^R + \frac{1}{2} X_\omega^I Y_\omega^I \right] \\ sin(0\omega t) \text{ term} & 0 = 0 \\ cos(1\omega t) \text{ term} & 0 = \omega X_\omega^I - \alpha X_\omega^R + \beta \left[X_0 Y_\omega^R + X_\omega^R Y_0 \right] \\ sin(1\omega t) \text{ term} & 0 = -\omega X_\omega^R - \alpha X_\omega^I + \beta \left[X_0 Y_\omega^I + X_\omega^I Y_0 \right] \\ cos(2\omega t) \text{ term} & 0 = \beta \left[\frac{1}{2} X_\omega^R Y_\omega^R - \frac{1}{2} X_\omega^I Y_\omega^I \right] \\ sin(2\omega t) \text{ term} & 0 = \beta \left[\frac{1}{2} X_\omega^R Y_\omega^I + \frac{1}{2} X_\omega^I Y_\omega^R \right] \end{cases} \\ \text{predator equation} & \begin{cases} cos(0\omega t) \text{ term} & 0 = \gamma Y_0 - \delta \left[X_0 Y_0 + \frac{1}{2} X_\omega^R Y_\omega^R + \frac{1}{2} X_\omega^I Y_\omega^I \right] \\ sin(0\omega t) \text{ term} & 0 = 0 \\ cos(1\omega t) \text{ term} & 0 = \omega Y_\omega^I + \gamma Y_\omega^R - \delta \left[X_0 Y_\omega^R + X_\omega^R Y_0 \right] \\ sin(1\omega t) \text{ term} & 0 = -\omega Y_\omega^R + \gamma Y_\omega^I - \delta \left[X_0 Y_\omega^I + X_\omega^I Y_0 \right] \\ cos(2\omega t) \text{ term} & 0 = -\delta \left[\frac{1}{2} X_\omega^R Y_\omega^R - \frac{1}{2} X_\omega^I Y_\omega^I \right] \\ sin(2\omega t) \text{ term} & 0 = -\delta \left[\frac{1}{2} X_\omega^R Y_\omega^I + \frac{1}{2} X_\omega^I Y_\omega^R \right] \end{cases} \end{aligned} \tag{2}$$

However, since we assume the solution is real (so that the balancing of $sin(0\omega t)$ is trivial) and that the solution is only to be approximated to the first order in the frequency basis, we discard the equations resulting from balancing the coefficients of harmonics not expressible by the ansatz. Thus, we arrive at

$$\begin{aligned}
\text{prey equation} \begin{cases} cos(0\omega t)\text{ term} & 0 = -\alpha X_0 + \beta\left[X_0 Y_0 + \frac{1}{2}X_\omega^R Y_\omega^R + \frac{1}{2}X_\omega^I Y_\omega^I\right] \\ cos(1\omega t)\text{ term} & 0 = \omega X_\omega^I - \alpha X_\omega^R + \beta\left[X_0 Y_\omega^R + X_\omega^R Y_0\right] \\ sin(1\omega t)\text{ term} & 0 = -\omega X_\omega^R - \alpha X_\omega^I + \beta\left[X_0 Y_\omega^I + X_\omega^I Y_0\right] \end{cases} \\
\text{predator equation} \begin{cases} cos(0\omega t)\text{ term} & 0 = \gamma Y_0 - \delta\left[X_0 Y_0 + \frac{1}{2}X_\omega^R Y_\omega^R + \frac{1}{2}X_\omega^I Y_\omega^I\right] \\ cos(1\omega t)\text{ term} & 0 = \omega Y_\omega^I + \gamma Y_\omega^R - \delta\left[X_0 Y_\omega^R + X_\omega^R Y_0\right] \\ sin(1\omega t)\text{ term} & 0 = -\omega Y_\omega^R + \gamma Y_\omega^I - \delta\left[X_0 Y_\omega^I + X_\omega^I Y_0\right] \end{cases}
\end{aligned} \tag{3}$$

Note that this results in 6 equations in the 6 unknowns (the coefficients of the ansatz). At this stage, the HB method proceeds to solve the resulting system of $N \cdot (2H+1)$ non-linear equations, and transforms the solution (solved in the frequency domain) back into a time-domain expression. Here, $N = 2$ because there are two quantities whose values vary over time.

We note a few observations. The linear terms of the differential equation produce linear summands in the harmonic balance. The non-linear terms of the differential equation produce non-linear summands in the harmonic balance at orders prescribed exactly by the product-to-sum trigonometric identities (which includes higher- and lower- order frequencies).

To demonstrate a higher-order truncation scheme and to better understand the structure of the harmonic balance equations, we now consider a secord order truncation HB method in which the solution ansatz takes the form

$$\begin{aligned}
x(t) &= X_0 + X_{\omega_1}^R cos(\omega_1 t) + X_{\omega_1}^I sin(\omega_1 t) + X_{\omega_2}^R cos(\omega_2 t) + X_{\omega_2}^I sin(\omega_2 t) \\
y(t) &= Y_0 + Y_{\omega_1}^R cos(\omega_1 t) + Y_{\omega_1}^I sin(\omega_1 t) + Y_{\omega_2}^R cos(\omega_2 t) + Y_{\omega_2}^I sin(\omega_2 t)
\end{aligned}$$

We will not explicitly carry out the procedure of the harmonic balancing as above, but we arrange the terms of the resulting equations in Table 1.

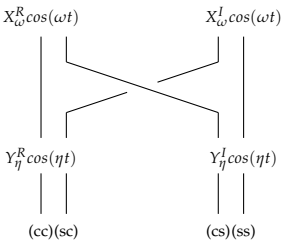
Here, we made use of the following *symbolic* formulae (which ensures the expressions on the right hand side are expressible the harmonic truncation scheme):

Formulae for $\omega > \eta$

$$\cos(\omega t)\cos(\eta t) = \frac{1}{2}\left[\cos\left((\omega-\eta)t\right) + \cos\left((\omega+\eta)t\right)\right]$$
$$\cos(\omega t)\sin(\eta t) = \frac{1}{2}\left[-\sin\left((\omega-\eta)t\right) + \sin\left((\omega+\eta)t\right)\right]$$
$$\sin(\omega t)\cos(\eta t) = \frac{1}{2}\left[+\sin\left((\omega-\eta)t\right) + \sin\left((\omega+\eta)t\right)\right]$$
$$\sin(\omega t)\sin(\eta t) = \frac{1}{2}\left[\cos\left((\omega-\eta)t\right) - \cos\left((\omega+\eta)t\right)\right]$$

Formulae for $\omega < \eta$

$$\cos(\eta t)\cos(\omega t) = \frac{1}{2}\left[\cos\left((\omega-\eta)t\right) + \cos\left((\omega+\eta)t\right)\right]$$
$$\cos(\eta t)\sin(\omega t) = \frac{1}{2}\left[+\sin\left((\omega-\eta)t\right) + \sin\left((\omega+\eta)t\right)\right]$$
$$\sin(\eta t)\cos(\omega t) = \frac{1}{2}\left[-\sin\left((\omega-\eta)t\right) + \sin\left((\omega+\eta)t\right)\right]$$
$$\sin(\eta t)\sin(\omega t) = \frac{1}{2}\left[\cos\left((\omega-\eta)t\right) - \cos\left((\omega+\eta)t\right)\right]$$



Note that the last four bands in Table 1 record the product interactions $\omega - \eta$ in the frequency bins $(1,0)-(1,0)$, $(0,1)-(0,1)$, $(1,0)-(0,1)$, and $(0,1)-(1,0)$, respectively and in order. Furthermore, within each of these four bands are four rows, recording the cos-cos, cos-sin, sin-cos, and sin-sin interactions, respectively and in order. In general, higher-order interactions (products)

Table 1. Second-order harmonic balance of $\frac{dx}{dt} - \alpha x + \beta xy = 0$ with $(\omega_1 > \omega_2 > 0)$

Order	0^{th}	1^{st}				2^{nd}							
Freq	$0 = \omega_0$	ω_1		ω_2		$\begin{bmatrix}2\\0\end{bmatrix}\cdot\vec{\omega}$		$\begin{bmatrix}1\\1\end{bmatrix}\cdot\vec{\omega}$		$\begin{bmatrix}1\\-1\end{bmatrix}\cdot\vec{\omega}$		$\begin{bmatrix}0\\2\end{bmatrix}\cdot\vec{\omega}$	
Term	cos | sin	cos	sin	cos	sin	cos	sin	cos	sin	cos	sin	cos	sin
Prey equation: $\frac{dx}{dt} - \alpha x + \beta xy = 0$													
$\frac{dx}{dt}$		$\omega_2 X^I_{(1,0)}$	$-\omega_r X^R_{(1,0)}$	$\omega_1 X^I_{(0,1)}$	$-\omega_1 X^R_{(0,1)}$	$\begin{bmatrix}2\\0\end{bmatrix}^T\vec{\omega}\,X^I_{(2,0)}$	$-\begin{bmatrix}2\\0\end{bmatrix}^T\vec{\omega}\,X^R_{(2,0)}$	$\begin{bmatrix}2\\0\end{bmatrix}^T\vec{\omega}\,X^I_{(2,0)}$	$-\begin{bmatrix}1\\1\end{bmatrix}^T\vec{\omega}\,X^R_{(1,1)}$	$\begin{bmatrix}1\\-1\end{bmatrix}^T\vec{\omega}\,X^I_{(1,-1)}$	$-\begin{bmatrix}1\\-1\end{bmatrix}^T\vec{\omega}\,X^R_{(1,-1)}$	$\begin{bmatrix}0\\2\end{bmatrix}^T\vec{\omega}\,X^I_{(0,2)}$	$-\begin{bmatrix}0\\2\end{bmatrix}^T\vec{\omega}\,X^R_{(0,2)}$
$-\alpha x$	$-\alpha X_{(0,0)}$												
$+\beta xy$	$\beta X_{(0,0)} Y_{(0,0)}$												
		$\beta X^R_{(1,0)} Y_{(0,0)}$ $\beta X^I_{(1,0)} Y^R_{(1,0)}$ $\beta X_{(0,0)} Y^R_{(1,0)}$ $\beta X_{(0,0)} Y^I_{(1,0)}$											
				$\beta X^R_{(0,1)} Y_{(0,0)}$ $\beta X^I_{(0,1)} Y_{(0,0)}$ $\beta X_{(0,0)} Y^R_{(0,1)}$ $\beta X_{(0,0)} Y^I_{(0,1)}$									
	$\frac{1}{2}\beta X^R_{(1,0)} Y^R_{(1,0)}$ $\frac{1}{2}\beta X^R_{(1,0)} Y^R_{(1,0)}$					$\frac{1}{2}\beta X^R_{(1,0)} Y^R_{(1,0)}$ $\frac{1}{2}\beta X^R_{(1,0)} Y^I_{(1,0)}$ $\frac{1}{2}\beta X^I_{(1,0)} Y^R_{(1,0)}$ $-\frac{1}{2}\beta X^R_{(1,0)} Y^R_{(1,0)}$							
	$\frac{1}{2}\beta X^R_{(0,1)} Y^R_{(0,1)}$ $\frac{1}{2}\beta X^R_{(0,1)} Y^R_{(0,1)}$											$\frac{1}{2}\beta X^R_{(0,1)} Y^R_{(0,1)}$ $\frac{1}{2}\beta X^R_{(0,1)} Y^I_{(0,1)}$ $\frac{1}{2}\beta X^I_{(0,1)} Y^R_{(0,1)}$ $-\frac{1}{2}\beta X^R_{(0,1)} Y^R_{(0,1)}$		
								$\frac{1}{2}\beta X^R_{(1,0)} Y^R_{(0,1)}$ $\frac{1}{2}\beta X^R_{(1,0)} Y^I_{(0,1)}$ $\frac{1}{2}\beta X^I_{(1,0)} Y^R_{(0,1)}$ $-\frac{1}{2}\beta X^I_{(1,0)} Y^I_{(0,1)}$		$\frac{1}{2}\beta X^R_{(1,0)} Y^R_{(0,1)}$ $-\frac{1}{2}\beta X^R_{(1,0)} Y^I_{(0,1)}$ $+\frac{1}{2}\beta X^I_{(1,0)} Y^R_{(0,1)}$ $\frac{1}{2}\beta X^I_{(1,0)} Y^I_{(0,1)}$			
						$\frac{1}{2}\beta X^R_{(0,1)} Y^R_{(1,0)}$ $\frac{1}{2}\beta X^R_{(0,1)} Y^I_{(1,0)}$ $\frac{1}{2}\beta X^I_{(0,1)} Y^R_{(1,0)}$ $-\frac{1}{2}\beta X^I_{(0,1)} Y^I_{(1,0)}$		$\frac{1}{2}\beta X^R_{(0,1)} Y^R_{(1,0)}$ $+\frac{1}{2}\beta X^R_{(0,1)} Y^I_{(1,0)}$ $-\frac{1}{2}\beta X^I_{(0,1)} Y^R_{(1,0)}$ $\frac{1}{2}\beta X^I_{(0,1)} Y^I_{(1,0)}$			

are captured in the frequency domain (as sums) by the following *analytic* identities:

$$\prod_{j=1}^{m} cos(\omega_j) = \frac{1}{2^m} \sum_{\vec{e}\in\{1,-1\}^m} cos(\vec{e}\cdot\vec{\omega})$$

$$\prod_{k=1}^{n} sin(\eta_k) = \frac{1}{2^n} \sum_{\vec{f}\in\{1,-1\}^n} cos\left(\vec{f}\cdot\vec{\eta} - \vec{f}\cdot\vec{1}\cdot\frac{\pi}{2}\right)$$

$$\left[\prod_{j=1}^{m} cos(\omega_j)\right]\cdot\left[\prod_{k=1}^{n} sin(\eta_k)\right] = \frac{1}{2^{m+n+1}} \sum_{\substack{\vec{e}\in\{1,-1\}^m \\ \vec{f}\in\{1,-1\}^n}} cos\left(\begin{pmatrix}\vec{\omega}\\+\vec{\eta}\end{pmatrix}\cdot\begin{pmatrix}\vec{e}\\\vec{f}\end{pmatrix} - \begin{pmatrix}\vec{0}_m\\\vec{f}\end{pmatrix}\cdot\begin{pmatrix}\vec{0}_m\\\vec{1}\end{pmatrix}\cdot\frac{\pi}{2}\right)$$

$$+ \frac{1}{2^{m+n+1}} \sum_{\substack{\vec{e}\in\{1,-1\}^m \\ \vec{f}\in\{1,-1\}^n}} cos\left(\begin{pmatrix}\vec{\omega}\\-\vec{\eta}\end{pmatrix}\cdot\begin{pmatrix}\vec{e}\\\vec{f}\end{pmatrix} + \begin{pmatrix}\vec{0}_m\\\vec{f}\end{pmatrix}\cdot\begin{pmatrix}\vec{0}_m\\\vec{1}\end{pmatrix}\cdot\frac{\pi}{2}\right)$$

However, it should be noted that although these product-to-sum formulae are analytically correct, but are not meaningful in the adopted harmonic truncation scheme. In particular, when $\begin{pmatrix}\vec{\omega}\\-\vec{\eta}\end{pmatrix}\cdot\begin{pmatrix}\vec{e}\\\vec{f}\end{pmatrix} < 0$, then the coefficient of this interaction should be recorded in the the $\begin{pmatrix}\vec{\omega}\\-\vec{\eta}\end{pmatrix}\cdot\begin{pmatrix}\vec{e}\\\vec{f}\end{pmatrix} \geq 0$ frequency bin (with the opposite sign for sine bases, and the same sign for cosine bases).

2 HB implementation in Troyanovsky's thesis

Troyonovsky implemented HB in the Stanford PISCES TCAD simulator, which uses a control volume method (CVM) for spatial discretization. The equations we seek to model are the isothermal lattice semiconductor electron and hole drift-diffusion system coupled with the Pois-

son equation, called the van Roosebroeck semiconductor equations:

$$\begin{aligned}
\nabla \cdot (-\epsilon \nabla \Psi) &= q\left(p - n + N_D^+ - N_A^-\right) \\
\frac{\partial n}{\partial t} &= \frac{1}{q}\nabla \cdot J_n - R & J_n &= qD_n \nabla n - q\mu_n n \nabla \Psi \\
\frac{\partial p}{\partial t} &= -\frac{1}{q}\nabla \cdot J_p - R & J_p &= -qD_p \nabla p - q\mu_p p \nabla \Psi
\end{aligned}$$

Here, Ψ is the electric potential, n and p are the electron and hole concentrations, N_D^+ and N_A^- are the donor and aceptor doping concentrations, q is the fundamental electron charge, ϵ is the electric permittivity of the semiconductor, R is a regeneration term, μ_n and μ_p are the carrier mobilities, and D_n and D_p are the carrier diffusivity coefficients.

Troyanovsky presents and compares a few harmonic balance methods in his thesis [Tro98], the efficiency of each hinging on accelerating the linear solver required in the Newton method for the solution of the non-linear system of equations obtained by discretization of the semiconductor equations. The approaches resemble that of [FML96], in which a block-diagonal preconditioner (with some off-diagonal blocks for highly non-linear systems) is advocated. This, in essence, makes implemented harmonic balance approach a GMRES/Krylov subspace method assisted Newton method. The present section details each component of the implementation.

We remark that the van Roosbroeck semiconductor equations proposed in 1950 [VR50] is based on the assumption of Maxwell-Boltzmann statistics and the Einstein relation (to determine the diffusivity $D = \mu k_B T$ as a function of the carrier mobilities μ and the Boltzmann constant k_B). Since its formulation, this model has been extensively analyzed. Typically in the analyses, the physically appropriate charge-neutrality and thermal equalibrium assumptions are imposed.

The existence of a stationary solution was established in [Jer85]. The transient, mixed Dirichlet/homogeneous Neumann boundary value problem has been analyzed [Jer87], even for the assumption of Fermi-Dirac statistics [GG89]. In these analyses, weak (and physically realistic) assumptions are placed on the diffusion coefficient functions. Recently, the transient model was formulated as a Boltzmann equation with a low–density collision operator, in which the existence of a solution follows readily from an application of the Fredholm alternative [Jün09].

2.1 Discretization scheme in the frequency domain

Since PISCES implements a control volume method, the equations above are re-expressed and evaluated at each control volume Ω_k (and its boundary $\partial\Omega_k$) as

$$\begin{aligned}
\oiint_{\partial\Omega_k} \epsilon \mathbf{E} d\mathbf{S} &= \iiint_{A_k} q\left(p - n + N_D^+ - N_A^-\right) dA \\
\oiint_{\partial\Omega_k} \mathbf{J}_n d\mathbf{S} &= \iiint_{A_k} q\left(R + \frac{\partial n}{\partial t}\right) dA \\
\oiint_{\partial\Omega_k} \mathbf{J}_p d\mathbf{S} &= \iiint_{A_k} -q\left(R + \frac{\partial p}{\partial t}\right) dA
\end{aligned}$$

The left-hand side integrals are evaluated by assuming that the current densities $\mathbf{J}_p$ and $\mathbf{J}_n$ along with the electrostatic potential $\mathbf{E} \equiv -\nabla\Psi$ are constant along each edge, and a finite-difference

approximate is used for the integration over each component of $\partial\Omega_k$. The right-hand side integrals are integrated over the control volume with the assumption that the integrands are constant within each control volume.

Stabilization

Troyanovsky adopts the Scharfetter-Gummel stabilization scheme. Whenver the potential difference between the endpoints of an edge e_{km} exceeds $\frac{2k_BT}{q}$, we replace the flux with

$$J_n \cdot \mathbf{s}_{km} = \frac{q\overline{\mu}_{km}\overline{\beta}_{km}}{d_{km}} \left[n_k \frac{\frac{\Psi_k - \Psi_m}{\overline{\beta}_{km}}}{e^{\frac{\Psi_k - \Psi_m}{\overline{\beta}_{km}}} - 1} - n_m \frac{\frac{\Psi_m - \Psi_k}{\overline{\beta}_{km}}}{e^{\frac{\Psi_m - \Psi_k}{\overline{\beta}_{km}}} - 1} \right]$$

where $\mathbf{s}_{km}$ is the (directed) unit tangent vector from node v_k to v_m, $\overline{\mu}_{km}$ is the effective average mobiility along the edge, and

$$\overline{\beta}_{km} = \frac{\beta_k - \beta_m}{ln(\frac{\beta_k}{\beta_m})} \text{ with } \beta = \frac{D_n}{\mu_n}.$$

Harmonic truncation frequency basis

Let $V = \{x_n\}_{n=0}^{N-1}$ be the collection of vertices of the mesh. A carrier concentration $u(\mathbf{x},t)$ is evaluated at these N spatial points. Applying the HB principle at each node x_n, the approximating quasiperiodic solution is taken to have the form

$$u(x_n,t) = U_n(t) = U_{n0} + \sum_{h=1}^{H} \left[U_{nh}^R cos(\omega_h t) - U_{nh}^I sin(\omega_h t) \right].$$

Some common and simple truncation schemes are

1. *Box Method*: Choose $B_1, \dots B_H > 0$. Demand each index to satisfy $|k_i| < B_i$.
2. *Diamond Method*: Choose $D > 0$. Demand $|k_1| + \cdots + |k_H| < D$
3. *Combination of Box and Diamond*

We depict the truncation schemes by marking subsets of $\mathbb{Z}^M$. When $M = 2$, this is easy to visualize. We produce an example in Figure 2.1. Note that all the indices are contained in the hyperspace $\vec{k} \cdot \vec{\omega} \geq 0$, and that the number of harmonics H can be readily computed in each of the Box and Diamond schemes:

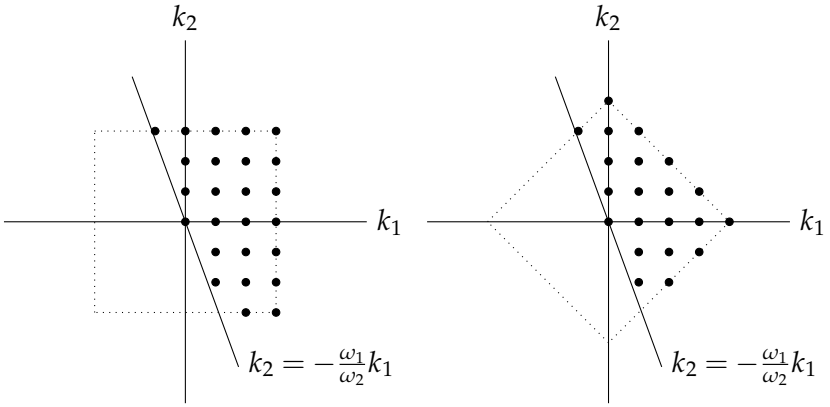
Proposition 2.1. *Let $\vec{k} = (k_1, \dots, k_M) \in \mathbb{Z}^M$ denote the lattice points which correspond to the indices of the truncation scheme. Suppose $\vec{\omega} = (\omega_1, \dots, \omega_M)$ are the fundamental frequencies. Let $B_1, \dots, B_M$ be the Box method pure harmonic cut-offs, and let P be the Diamond method pure harmonic cut-off. Then the number*

of frequencies in the Box and Diamond truncation schemes are given by

$$H_{Box} = \frac{(B_1+1)(B_2+1)\cdots(B_M+1)}{2} + \frac{1}{2}$$

$$H_{Diamond} = \frac{1}{2}\sum_{\ell=0}^{min\{M,P\}} 2^\ell \binom{M}{\ell}\binom{P}{\ell} + \frac{1}{2}$$

Figure 1. Example Box and Diamond truncation schemes for $M = 2$, $\vec{\omega} = (30Hz, 11Hz)$



Box (order 3, $B_1 = B_2 = 3$) Diamond (order 4, $D = 5$)

Proof of Proposition 2.1. Geometrically, H_{Box} is the number of lattice points inside the intersection of the closed hyperspace $\mathcal{H}^+ = \{\vec{k} | \vec{k} \cdot \vec{\omega} \geq 0\}$ with the closed box $\mathcal{B} = [-B_1, B_1] \times \cdot [-B_M, B_M]$. Let us first count the total number of points in $\mathcal{B}$. Since the fundamental frequencies are not commensurate, i.e., are linearly independent over $\mathbb{Q}$, the hyperplane $\vec{k} \cdot \vec{\omega} = 0$ contains only one point, $\vec{0}$. Note that there are $(B_1+1)(B_2+1)\cdots(B_M+1) - 1$ integer points in $\mathcal{B} \setminus \{\vec{0}\}$. By symmetry, the hyperplane exactly divides the number of points in $\mathcal{B} \setminus \{\vec{0}\}$ in half. Accounting for the point $\vec{0}$ in the hyperplane $\vec{k} \cdot \vec{\omega} = 0$, we arrive at the formula for H_{Box}.

Similarly, $H_{Diamond}$ is the number of lattice points contained in the intersection of the closed hyperspace $\mathcal{H}^+ = \{\vec{k} | \vec{k} \cdot \vec{\omega} \geq 0\}$ with the closed diamond $\mathcal{D} = \{\vec{k} | \sum_{i=1}^M |k_i| \leq P\}$. Let us first count the total number of points in $\mathcal{D}$. Observe that a point $\vec{k}_0 = (k_1, \ldots, k_M) \in \mathcal{D}$ defines a non-negative integer partition of an integer no greater than P, since $|k_1| + \cdots + |k_M| \leq P$ and $|k_i| \geq 0$ for all i.

We briefly discuss such partitions before counting. Suppose that exactly ℓ of the k_i are non-zero, so that $M - \ell$ of the k_i are zero, and that $|k_1| + \cdots + |k_M| = r$. We will call the sum $|k_1| + \cdots + |k_M|$ of M non-negative integers $|k_i|$, which consists of exactly ℓ non-zero summands, an ℓ-partition of the integer r. We say that two ℓ-partitions of r are different if the ordering of the summands or the values of the non-zero summands are different. Observe that there are 2^ℓ points of $\mathcal{D}$ (including $\vec{k}_0$) which define the same ℓ-partition of r as $\vec{k}_0$; this is because the point $(\pm k_1, \ldots, \pm k_M)$ for any choice of signs also has exactly ℓ non-zero terms, and 2^ℓ signs (in front of the non-zero k_i) can be alternated to account for a different point in $\mathcal{D}$. Second, the number of integer partitions of $r < P$ into exactly ℓ positive numbers is equal to $\binom{P}{\ell}$; we can see this be choosing ℓ numbers $p_1, \ldots, p_\ell$ from the list $1, 2, 3, \ldots, P$, and uniquely correspond to it the partition

$p_1 + (p_2 - p_1) + \cdots + (p_\ell - p_{\ell-1})$ of the integer $p_\ell < P$. Finally, observe that ℓ can be any integer between 0 and $min\{P, N\}$, inclusive. Thus, there are exactly $\sum_{\ell=0}^{min\{M,P\}} 2^\ell \binom{M}{\ell} \binom{P}{\ell}$ points in $\mathcal{D}$.

Accounting for the point $\vec{0}$ as in counting H_{Box} (which is counted in the summation at index $\ell = 0$ corresponding to zero non-zero entries), we arrive at the formula for $H_{Diamond}$. □

Aside: The symmetry of the formula for $H_{Diamond}$ reveals a surprising quality of the Diamond truncation scheme. Adopting M fundamental frequencies and truncating at order P results in the same number of degrees of freedom in the harmonic balance as from adopting P fundamental frequencies and truncating at order M. There is a generating function proof of this fact, as well. Let $L(P, M)$ denote the number of points $\vec{k} = (k_1, \ldots, k_M) \in \mathbb{Z}^M$ such that $|k_1| + \cdots + |k_M| \leq P$. It has been shown that

$$\sum_{P=1}^{\infty} \sum_{M=1}^{\infty} L(P, M) x^P y^M = \frac{1}{1 - x - y - xy},$$

as a generating function identity, in which the generating function is obviously symmetric in x and y.

Large signal distortion characterization

For a quasi-periodic input signal

$$\begin{aligned} v(t) &= \sum_{n=1}^{H} \left(V_n^R cos(\omega_n t) + V_n^I sin(\omega_n t) \right) \\ &= \mathcal{R} \left(\sum_{n=1}^{H} V_n e^{i\omega_n} \right), \end{aligned}$$

a common metric for quantifying its large signal distortion is its n^{th} harmonic distortion:

$$HD_n := \frac{|V_n|}{|V_1|}$$

for $n \geq 2$. Its total harmonic distortion is defined as

$$THD := \frac{\sqrt{\sum_{n=2}^{\infty} |V_n|^2}}{|V_1|}.$$

In the presence of great harmonic distortion, more frequencies should be included in the HB basis.

Temporal collocation points

The harmonic balance method results in a non-linear system of equations in $2H + 1$ unknowns, which are the amplitudes of the quasi-periodic ansatz expansion of the state variable. Observe that a non-linear term in the PDE will most likely produce a necessary condition for the exact solution to truly have the form of the ansatz, resulting in an over-determined system of equations. These

additional conditions come from balancing amplitudes at higher frequencies, and are discarded in the non-linear system of equations. For example, in the predator-prey system (1), the harmonic balancing resulted in an over-determined system of equations (2) which was reduced to a system of equations (3) by balancing only the frequencies present in the solution ansatz.

Some non-linear terms like e^u, $sin(\omega t) \cdot u$, u^2 (as functions of the state variable $u(x,t)$) expressed in the frequency domain are not representable in the harmonic balance ansatz since higher-order harmonics appear in their Fourier expansion. However, their Fourier series expansion can be easily truncated, and heuristically, the higher order harmonics have negigible amplitudes. Other terms non-linear terms like $f(u) \cdot u$, where $f(u)$ does not have a simple Fourier expansion, must be transformed approximated in the frequency domain differently.

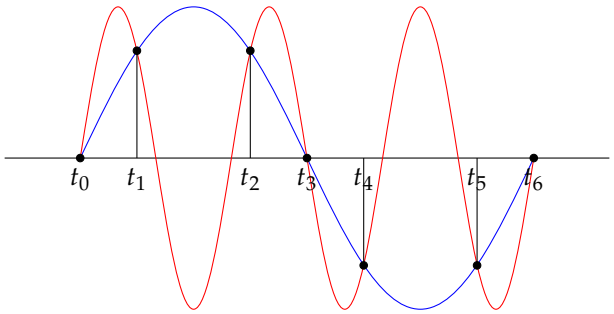
In order to arrive at a system of $2H+1$ equations in $2H+1$ unknowns, we choose *temporal collocation points* $\vec{t} := (t_0, t_1, \dots t_{2H})$ at which to calculate the Discrete Fourier transform matrix $\Gamma = \Gamma(\vec{t})$ and its inverse Γ^{-1}. We use this to evaluate the non-linear terms like $f(u) \cdot u$ in the harmonic balancing, resulting in a summand of $\Gamma^{-1}\left(f(u(\vec{t})) \cdot u(\vec{t})\right)$. The harmonic balancing then doesn't (analytically) come from balancing the coefficients of the $cos(\omega_h t)$ and $sin(\omega_h t)$ terms, but from (symbolically) balancing the coefficients of $cos(\omega_h t_i)$ and $sin(\omega_h t_i)$ for all $h = 0, \dots, H$ and $i = 0, \dots 2H+1$. Thus, we can drive the residual as close to zero as possible *at these points in time*. The collocation points $\vec{t}$ should be chosen so the matrices Γ and Γ^{-1} are well-conditioned for the DFT and the IDFT.

Aside: The Nyquist Sampling Theorem states that a quasi-periodic signal

$$f(t) = \sum_{h=0}^{H} F_h e^{i\omega_h t}$$

can be recovered by sampling at a rate of $2\omega_H$ (if the ω_i are given in Hz). Improperly imposing a frequency band limit, say, if we are attempting to reconstruct $f(t)$ without knowledge of its frequencies, results in a poor reconstruction. The numerical inaccuracy of the solution obtained from solving the truncated harmonic balancing system of non-linear equations is related to the aliasing effect of inadequate sampling.

Figure 2. Caution for choosing temporal collocation points



This phenomenon is illustrated in Figure 2. Even with the 7 time collocation points $t_0, \dots, t_6$, the two depicted harmonics are indistinguisable. In other words, a signal that has been sampled in the time domain at these collocation points cannot be reliably expressed in a frequency basis which includes both of these harmonics. So, more time collocation points are required to distinguish

them. From a different perspective, if the lower frequency is included but the higher frequency is not included in the truncation scheme, then the harmonic balance method may inaccurately attribute too much energy to the lower frequency harmonic (since it omits the larger frequency harmonic).

Frequency Mapping Method

When very few frequencies are included in the harmonic truncation, it may be beneficial to perform a change of basis in the frequency domain so that all of the frequencies can be realized as multiples of a single fundamental harmonic. This would allow for the use of a Discrete Fast Fourier Transform. Frequency remapping methods are referred to as Artificial Frequency Mapping, Frequency Mapping Method, and Frequency-remapping Method in the literature.

A frequency remapping replaces the H harmonics

$$k_1\omega_1 + \cdots + k_M\omega_M$$

and with another set of H harmonics

$$k_1\widehat{\omega_1} + \cdots + k_M\widehat{\omega_M}.$$

We define the *remapping function* to be the assignment

$$\sigma : \omega_i \mapsto \widehat{\omega_i}.$$

A robust frequency remapping method should be easily invertible and has a densely distributed spectrum in its entire bandwidth. For example, the order P, two-tone box truncation harmonic balance method has a commonly used remapping function $\sigma : (\omega_1, \omega_2) \mapsto (2P+1, 1)$.

An immediate benefit of a FMM for the weak formulation of the HB equations is the *reduction in magnitude of the largest truncated frequency*. This allows us weakly obtain the harmonic balance equations readily via numerical integration over the time domain.

2.2 Newton method

Since the harmonic balance method arrives at a non-linear system of equations, it is customary to solve it via the Newton method. Starting with an initial guess $\mathbf{x}_0$, we calculate the residual $\mathbf{r}_0$. Then, we iterate via

$$\mathbf{x}_n = \left[\left. \frac{\partial \mathfrak{F}}{\partial \mathbf{x}} \right|_{\mathbf{x}_{n-1}} \right]^{-1} \mathfrak{F}(\mathbf{x}_{n-1}) + \delta_{n-1}.$$

The linearized problem may still be difficult to solve due to the sheer size of the matrix involved. Where the standard backward Euler method deals with a $N \times N$ matrix, we are now faced with a $N(2H+2) \times N(2H+2)$ matrix. More will be said of this linearized problem, shortly.

HB Jacobian

The residual vector is of length $N(2H+2)$ (we include the $sin(0t)$ term for consistency in the assembly of the HB Jacobian). It is formed by concatenation of the N vectors $\mathbf{F}_n$, each of length

$2H+2$ and corresponding to the $2H+2$ frequency coefficients at the n^{th} node:

$$\mathbf{F}_n = \begin{bmatrix} F^R_{n0} \\ 0 \\ F^R_{n1} \\ F^I_{n1} \\ \vdots \\ F^R_{nH} \\ F^I_{nH} \end{bmatrix} \quad \Rightarrow \quad \mathbf{F} = \begin{bmatrix} \mathbf{F}_0 \\ \mathbf{F}_1 \\ \vdots \\ \mathbf{F}_{N-1} \end{bmatrix} \qquad \mathbf{X}_n = \begin{bmatrix} X^R_{n0} \\ 0 \\ X^R_{n1} \\ X^I_{n1} \\ \vdots \\ X^R_{nH} \\ X^I_{nH} \end{bmatrix} \quad \Rightarrow \quad \mathbf{X} = \begin{bmatrix} \mathbf{X}_0 \\ \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_{N-1} \end{bmatrix}$$

Observe that the HB Jacobian $\frac{\partial \mathbf{F}}{\partial \mathbf{X}}$ is of size $N(2H+2) \times N(2H+2)$. It has a block form, realized as an $N \times N$ matrix of $(2H+2) \times (2H+2)$ blocks. The state vector itself is a spectral expansion of the amplitudes in the frequency domain, recording the amplitudes of each of the $2H+2$ harmonics. For consistency, we can arrange the state vector of these spectral coefficients in the product lexicographical order on the set $\{nodes\} \times \{harmonics\}$.

Explicitly, we have a $(2H+2) \times (2H+2)$ block

$$\frac{\partial \mathbf{F}_n}{\partial \mathbf{X}_m} = \begin{bmatrix} \begin{bmatrix} \frac{\partial \mathbf{F}^R_{n0}}{\partial \mathbf{X}^R_{m0}} & \frac{\partial \mathbf{F}^R_{n0}}{\partial \mathbf{X}^I_{m0}} \\ \frac{\partial \mathbf{F}^I_{n0}}{\partial \mathbf{X}^R_{m0}} & \frac{\partial \mathbf{F}^I_{n0}}{\partial \mathbf{X}^I_{m0}} \end{bmatrix} & \begin{bmatrix} \frac{\partial \mathbf{F}^R_{n0}}{\partial \mathbf{X}^R_{m1}} & \frac{\partial \mathbf{F}^R_{n0}}{\partial \mathbf{X}^I_{m1}} \\ \frac{\partial \mathbf{F}^I_{n0}}{\partial \mathbf{X}^R_{m1}} & \frac{\partial \mathbf{F}^I_{n0}}{\partial \mathbf{X}^I_{m1}} \end{bmatrix} & \cdots & \begin{bmatrix} \frac{\partial \mathbf{F}^R_{n0}}{\partial \mathbf{X}^R_{mH}} & \frac{\partial \mathbf{F}^R_{n0}}{\partial \mathbf{X}^I_{mH}} \\ \frac{\partial \mathbf{F}^I_{n0}}{\partial \mathbf{X}^R_{mH}} & \frac{\partial \mathbf{F}^I_{n0}}{\partial \mathbf{X}^I_{mH}} \end{bmatrix} \\ \begin{bmatrix} \frac{\partial \mathbf{F}^R_{n1}}{\partial \mathbf{X}^R_{m0}} & \frac{\partial \mathbf{F}^R_{N1}}{\partial \mathbf{X}^I_{m0}} \\ \frac{\partial \mathbf{F}^I_{n1}}{\partial \mathbf{X}^R_{m0}} & \frac{\partial \mathbf{F}^I_{N1}}{\partial \mathbf{X}^I_{m0}} \end{bmatrix} & \begin{bmatrix} \frac{\partial \mathbf{F}^R_{n1}}{\partial \mathbf{X}^R_{m1}} & \frac{\partial \mathbf{F}^R_{N1}}{\partial \mathbf{X}^I_{m1}} \\ \frac{\partial \mathbf{F}^I_{n1}}{\partial \mathbf{X}^R_{m1}} & \frac{\partial \mathbf{F}^I_{N1}}{\partial \mathbf{X}^I_{m1}} \end{bmatrix} & \cdots & \begin{bmatrix} \frac{\partial \mathbf{F}^R_{n1}}{\partial \mathbf{X}^R_{mH}} & \frac{\partial \mathbf{F}^R_{N1}}{\partial \mathbf{X}^I_{mH}} \\ \frac{\partial \mathbf{F}^I_{n1}}{\partial \mathbf{X}^R_{mH}} & \frac{\partial \mathbf{F}^I_{N1}}{\partial \mathbf{X}^I_{mH}} \end{bmatrix} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \begin{bmatrix} \frac{\partial \mathbf{F}^R_{nH}}{\partial \mathbf{X}^R_{m0}} & \frac{\partial \mathbf{F}^R_{nH}}{\partial \mathbf{X}^I_{m0}} \\ \frac{\partial \mathbf{F}^I_{nH}}{\partial \mathbf{X}^R_{m0}} & \frac{\partial \mathbf{F}^I_{nH}}{\partial \mathbf{X}^I_{m0}} \end{bmatrix} & \begin{bmatrix} \frac{\partial \mathbf{F}^R_{nH}}{\partial \mathbf{X}^R_{m1}} & \frac{\partial \mathbf{F}^R_{nH}}{\partial \mathbf{X}^I_{m1}} \\ \frac{\partial \mathbf{F}^I_{nH}}{\partial \mathbf{X}^R_{m1}} & \frac{\partial \mathbf{F}^I_{nH}}{\partial \mathbf{X}^I_{m1}} \end{bmatrix} & \cdots & \begin{bmatrix} \frac{\partial \mathbf{F}^R_{nH}}{\partial \mathbf{X}^R_{mH}} & \frac{\partial \mathbf{F}^R_{nH}}{\partial \mathbf{X}^I_{mH}} \\ \frac{\partial \mathbf{F}^I_{nH}}{\partial \mathbf{X}^R_{mH}} & \frac{\partial \mathbf{F}^I_{nH}}{\partial \mathbf{X}^I_{mH}} \end{bmatrix} \end{bmatrix}$$

and from which we can assemble the $N \cdot 2 \cdot (H+1) \times N \cdot 2 \cdot (H+1)$ HB Jacobian

$$\frac{\partial \mathbf{F}}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial \mathbf{F}_1}{\partial \mathbf{X}_1} & \frac{\partial \mathbf{F}_1}{\partial \mathbf{X}_2} & \cdots & \frac{\partial \mathbf{F}_1}{\partial \mathbf{X}_N} \\ \frac{\partial \mathbf{F}_2}{\partial \mathbf{X}_1} & \frac{\partial \mathbf{F}_2}{\partial \mathbf{X}_2} & \cdots & \frac{\partial \mathbf{F}_2}{\partial \mathbf{X}_N} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mathbf{F}_N}{\partial \mathbf{X}_1} & \frac{\partial \mathbf{F}_N}{\partial \mathbf{X}_2} & \cdots & \frac{\partial \mathbf{F}_N}{\partial \mathbf{X}_N} \end{bmatrix}$$

The assembly of the HB Jacobian is efficiently constructed by the block structure

$$\frac{\partial \mathbf{F}_n}{\partial \mathbf{X}_m} = \begin{bmatrix} \Phi_{nm}^{00} & \Phi_{nm}^{01} & \cdots & \Phi_{nm}^{0H} \\ \Phi_{nm}^{10} & \Phi_{nm}^{11} & \cdots & \Phi_{nm}^{1H} \\ \vdots & & \ddots & \vdots \\ \Phi_{nm}^{H0} & \Phi_{nm}^{H1} & \cdots & \Phi_{nm}^{HH} \end{bmatrix} = \left[\Phi_{nm}^{hi}\right]_{hi} \quad \text{where} \quad \Phi_{nm}^{hi} = \begin{bmatrix} \frac{\partial \mathbf{F}_{nh}^R}{\partial \mathbf{X}_{mi}^R} & \frac{\partial \mathbf{F}_{nh}^R}{\partial \mathbf{X}_{mi}^I} \\ \frac{\partial \mathbf{F}_{nh}^I}{\partial \mathbf{X}_{mi}^R} & \frac{\partial \mathbf{F}_{nh}^I}{\partial \mathbf{X}_{mi}^I} \end{bmatrix}$$

and Φ_{nm}^{hi} is nothing other than the partial derivative of the h^{th} spectral expansion coefficient at the n^{th} spatial node, with respect to the i^{th} spectral expansion coefficient at the m^{th} spatial node. Of course, the 2×2 size of Φ_{nm}^{hi} is necessary to account for the derivative of the complex function $F_{nh}(\mathbf{X}) = F_{nh}^R(\mathbf{X}) + jF_{nh}^I(\mathbf{X})$ with respect to the complex state variable $X_{mi} = X_{mi}^R + jX_{mi}^I$.

2.3 Preconditioner

The diagonal block preconditioner is effective for systems with few to moderate non-linearities. This is because the diagonal block preconditioner can be obtained in the small signal limit (sometimes referred to as the low-level distorition approximation). Though it is simple to implement, drawback of a block diagonal preconditioner is that it loses its sparsity when higher orders of harmonic interactions are included in the truncation scheme. Some authors have proposed a sectioned preconditioner [TYD00].

2.4 Linear solver: MGS-GMRES

Troyanovsky uses the Generalized Minimum RESidual algorithm (GMRES), a Krylov subspace method involving non-symmetric matrices, to solve the linear system $A\mathbf{x} = \mathbf{b}$. At the k^{th} iteration, we produce an approximate solution $\mathbf{x}_k$ and corresponding k^{th} residual

$$\mathbf{r}_k := A\mathbf{x}_k - \mathbf{b}.$$

Each iterate $\mathbf{x}_k$ belongs to the affine space $\mathbf{x}_0 + K_k$, where

$$K_k := span\{\mathbf{r}_0, A\mathbf{r}_0, A^2\mathbf{r}_0, \ldots, A^k\mathbf{r}_k\}$$

is the k^{th} Krylov subspace of A with the initial vector $\mathbf{r}_0$. It is chosen (uniquely) to minimize the Euclidean norm of $\mathbf{r}_k$, implemented as a least squares problem

$$\min_{\mathbf{x}_k \in \mathbf{x}_0 + K_k} ||\mathbf{r}_k||_2 = \min_{\mathbf{x}_k \in \mathbf{x}_0 + K_k} ||A\mathbf{x}_k - \mathbf{b}||_2.$$

Note that formulating the k^{th} least squares problem requires defining a basis for K_k. The common implementation of GMRES adopts the Arnoldi process to calculate a basis. The Arnoldi process is the modified Gram-Schmidt method applied to the (linearly independent) set

$$\{\mathbf{r}_0, A\mathbf{r}_0, A^2\mathbf{r}_0, \ldots, A^{N-1}\mathbf{r}_0\},$$

resulting in an orthonormal basis

$$\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$$

for the Krylov suubspace K_k. The k^{th} least squares problem is thus determining the k coefficients $a_k^1, a_k^2, \ldots, a_k^k$ to minimize the quantity

$$||\mathbf{r}_k||_2 := \left|\left|\left(\mathbf{x}_0 + a_k^1 v_1 + \cdots + a_k^k v_k\right) - \mathbf{b}\right|\right|_2.$$

Modified Gram-Schmidt

During the process of obtaining an orthonormal basis for the Krylov subspaces, an appropriate algorithm should be adopted to minimize the accumulation of numerical error.

(Classical) Gram-Schmidt	Modified Gram-Schmidt
for $i = 1$ to n **do**	**for** $i = 1$ to n **do**
$w \leftarrow u_i$	$w \leftarrow u_i$
for $j = 1$ to $i - 1$ **do**	**for** $j = 1$ to $i - 1$ **do**
$r_{ji} = \langle v_j, w \rangle$	$r_{ji} = \langle v_j, w \rangle$
end for	$w = w - r_{ji} v_j$
for $j = 1$ to $i - 1$ **do**	**end for**
$w = w - r_{ji} v_j$	$r_{ii} = \|\|w\|\|_2$
end for	$v_i \leftarrow \frac{w}{r_{ii}}$
$r_{ii} = \|\|w\|\|_2$	**end for**
$v_i \leftarrow \frac{w}{r_{ii}}$	
end for	

The only difference between the classical and modified Gram-Schmidt processes are: in classical Gram-Schmidt, at each (outer-most) i^{th} step, the projection of the i^{th} original vector to the $i-1$ previously constructed vectors are substracted from the i^{th} original vector; in Modified Gram-Schmidt, at each (outer-most) i^{th} step, the projections are subtracted so as not to accumulate numerical error.

Krylov subspace methods

The GMRES method exploits the Cayley-Hamilton theorem from linear algebra: if A is an $n \times n$ invertible, square matrix, then A is a root of its characteristic polynomial χ_A

$$\chi_A(A) = A^n + c_{n-1}A^{n-1} + \cdots + c_1 A + c_0 I_{n\times n} = 0_{n\times n}.$$

Note that the coefficient c_0 is the product of the eigenvalues of A, since the roots of χ_A are exactly the eigenvalues of A. Invertibility of A ensures $c_0 \neq 0$. Manipulating this equation, we easily see that the inverse A^{-1} is expressible as a linear combination of powers of A itself:

$$\begin{aligned} 0_{n\times n} &= A^n + c_{n-1}A^{n-1} + \cdots + c_1 A + c_0 I_{n\times n} \\ \Rightarrow 0_{n\times n} &= A^{n-1} + c_{n-1}A^{n-2} + \cdots + c_1 I_{n\times n} + c_0 A^{-1} \\ \Rightarrow A^{-1} &= -\tfrac{1}{c_0}\left[A^{n-1} + c_{n-1}A^{n-2} + \cdots + c_1 I_{n\times n}\right] \end{aligned}$$

So, the solution $\mathbf{x}_0$ to $A\mathbf{x} = \mathbf{b}$ is readily given by

$$\mathbf{x}_0 = -\frac{1}{c_0}A^{n-1}\mathbf{b} - \frac{c_{n-1}}{c_0}A^{n-2}\mathbf{b} - \cdots - \frac{c_2}{c_0}A^1\mathbf{b} - \frac{c_1}{c_0}I_{n\times n}\mathbf{b}$$

2.5 Reverting from frequency domain to time domain

The matrix Γ is the Discrete Fourier Transform (DFT) matrix. Its numerical representation depends on the choice of the $2H+1$ time collocation points. Its inverse, Γ^{-1}, is the Inverse Discrete Fourier Transform (IDFT) matrix.

$$\Gamma = \begin{bmatrix} 1 & \cos(\omega_1 t_1) & \sin(\omega_1 t_1) & \cdots & \cos(\omega_1 t_S) & \sin(\omega_1 t_S) \\ 1 & \cos(\omega_2 t_1) & \sin(\omega_2 t_1) & \cdots & \cos(\omega_2 t_S) & \sin(\omega_2 t_S) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ 1 & \cos(\omega_H t_1) & \sin(\omega_H t_1) & \cdots & \cos(\omega_H t_S) & \sin(\omega_H t_S) \end{bmatrix}$$

3 Considerations

3.1 Factors for numerical accuracy

We discuss modifications which may affect numerical accuracy.

Tuning the Box and Diamond harmonic trunctation

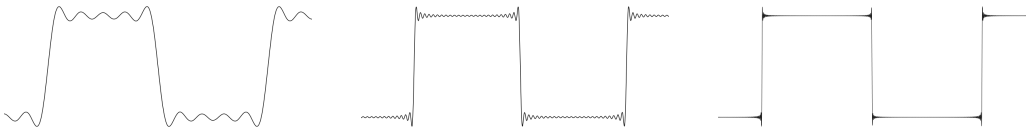
The Box and Diamond methods for harmonic truncation can be equivalently thought of as a choice of a ℓ^p metric on $\mathbb{Z}^M$ and a cut-off R-ball in the ℓ^p-norm on the space $\mathbb{Z}^H$. For, the linear combinations $k_1\omega_1 + \cdots + k_M\omega_M$ of the M harmonics correspond to a $\mathbb{Z}^M$ tuple of numbers $(k_1,\ldots,k_M)$. Then, the Box Method simply considers the tuples which lie within an ℓ^∞ distance at most R from the origin, whereas the Diamond Method only considers the tuples which lie within an ℓ^1 distance at most R from the origin. Interpolation between the Box and Diamond methods can be achieved by choosing $p \in (0,1)$ for the ℓ^p-norm on $\mathbb{Z}^H$.

Gibbs phenomenon

If the input signals are continuously piecewise differentiable, but with discrete jumps at countably many points of discontinuity, then their Fourier truncations poorly estimate their values near the discontinuity. See Figure 3.1 (borrowed from `https://en.wikipedia.org/wiki/Gibbs_phenomenon`) for a depiction of the poor approximation of a square wave.

For the square wave, it is known that any quasi-periodic approximation (with large bandwidth) under- and over-shoots the true value (depending on whether the discontinuity is approached

Figure 3. Truncation of a square wave at the first 10, 50, and 250 positive integer harmonics.



from the left or the right) by a value of approximately 0.089489872236 times the value of the jump at the discontinuity.

If any of the input signals are discontinuous, then their quasi-periodic truncation will experience Gibbs penomena. In this case, the temporal collocation points should be chosen away from the discontinuities to avoid the numerical inaccuracies of truncation coming from the over- and under-shooting.

The collocation points can also be chosed to resolve extremely large gradients (an by analogy we can think of the discontinuous jumps as infinite gradients). A time-mapped harmonic balance has been explored [NW99] in which a non-linear change of variable $t = \lambda(\hat{t})$ of the time variable t scales the temporal derivative via the product rule:

$$\frac{d}{dt} = \frac{1}{\lambda'(\hat{t})}\frac{d}{d\hat{t}}.$$

One advantage of the time collocation point remapping is that a uniform grid of points can be used in the $\hat{t}$ variable, and a Discrete Fourier Transform can be applied.

Aliasing from insufficient harmonics

The Plancherel theorem relates the $\mathcal{L}^2$-norm of a T-periodic function to the ℓ^2-norm of its frequency spectrum (given by its Fouerier expansion):

$$f(t) = \sum_{n=-\infty}^{\infty} \widehat{f}(n) e^{in\frac{2\pi}{T}t}$$
$$\Rightarrow ||f||_{L^2(\mathbb{R})} = \sum_{h=-\infty}^{\infty} |\hat{f}(n)|^2$$

where $\hat{f}(n)$ is the projection of f onto $e^{in\frac{2\pi}{T}t}$ given by

$$\hat{f}(n) = \frac{1}{T}\int_{s=0}^{s=T} f(s) e^{-in\frac{2\pi}{T}s} ds.$$

In general,

Proposition 3.1. *We have the following a priori estimates on the Fourier coefficients of f:*

$$\begin{aligned}
f \text{ absolutely continuous } &\Rightarrow \left|\widehat{f}(n)\right| \leq \frac{K(f)}{|n|} \\
f \text{ bounded variation } &\Rightarrow \left|\widehat{f}(n)\right| \leq \frac{var(f)}{2\pi|n|} \\
f \text{ } p\text{-times continuously differentiable } &\Rightarrow \left|\widehat{f}(n)\right| \leq \frac{\left|\left|f^{(p)}\right|\right|_{\mathcal{L}^1}}{|n|^p} \\
f \in \mathcal{C}^p, \omega_p = \text{modulus of cty of } f^{(p)} &\Rightarrow \left|\widehat{f}(n)\right| \leq \frac{\omega_p(\frac{2\pi}{n})}{|n|} \\
f \text{ } \alpha\text{-Hölder continuous } &\Rightarrow \left|\widehat{f}(n)\right| \leq \frac{K(f)}{|n|^\alpha}
\end{aligned}$$

Such a priori estimates of Fourier coefficients inform a choice for the cut-off frequency ω_H, ensuring negigible $\mathcal{L}^2$ energy in the tail

$$\sum_{n \geq \omega_H} |\widehat{f}(n)|^2.$$

In an effort to make our harmonic balance implementation as paramter-free as possible, we may, for example, ask the user for a tolerance on the solution approximation. Then, these a priori Fourier coefficient estimates can be used to determine the truncation order of the harmonic balance method.

Proof. We will prove the estimate for $f \in \mathcal{C}^p$.

$$\begin{aligned}
\left|\widehat{f}(n)\right| &= \left|\int_0^T f(t)e^{int}dt\right| \\
&= \left|f(s)e^{ins}\Big|_{s=0}^{s=T} - \int_0^T f(t)\frac{1}{in}e^{int}dt\right| \text{ by integration by parts} \\
&= \left|0 - \frac{1}{in}\int_0^T f'(t)e^{int}dt\right| \text{ since } T\text{-periocity of } f \text{ gives } f(T) = f(0) \\
&\leq \frac{1}{|n|}\left|\int_0^T f'(t)e^{int}dt\right| \\
&\vdots \\
&\leq \frac{1}{|n|^p}\left|\int_0^T f^{(p)}e^{int}dt\right| \text{ by re-iterating the above } (p-1) \text{ times} \\
&\leq \frac{1}{|n|^p}\left|\left|f^{(p)}\right|\right|_{\mathcal{L}^1}
\end{aligned}$$

□

3.2 Factors for numerical efficiency

Collocation point sampling affects conditioning of Γ

A *near-orthogonal selection* method is applied in [KSSV88]. They are motivated by the observation: a well-conditioned matrix has nearly orthogonal rows. The method was originally proposed by [KSV86] as an over-sampling-and-optimize approach.

3.3 Alternatives and modifications for modelling periodic responses

Shooting method

The shooting method is an approach to solve the equation

$$\mathbf{g}(\mathbf{u}(t)) + \frac{d}{dt}\mathbf{q}(\mathbf{u}(t)) = \mathbf{w}(t)$$

when $\mathbf{w}$ is a periodic driving function with period T, and when we expect the solution u to be periodic *with the same period* T. Mathematically, this imposes the constraint

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

The shooting method has some advantages over HB when HB experieinces difficulties. The presence of great non-linearities (i.e., expressions that must be written in many modes) depreciates the efficiency of HB. This can be detected by a more populated HB Jacobian, in which case the Jacobian loses its diagonally dominant form. This makes the Jacobian more difficult to precondition (the commonly used block Jacobian is no longer a sufficient preconditioner, heuristaclly so because the block Jacobian can be obtained in the small-signal limit).

Shooting method to calibrate HB

Bertazzi-Bonani-Guerrieri-Ghione compare the shooting method and HB method for determining the steady-state solution of a physics-based device model in [BBGG08]. They observe that "the shooting method appears more robust from a numerical standpoint, at least for strongly nonlinear conditions, while the HB approach yields better accuracy but is significantly more sensitive to the initial condition". In light of this, they propose the following strategy: apply the shooting method with a limited number of time points to arrive at a good initial condition with which to start the HB method.

4 Adapting a transient PDE for HB

We briefly summarize the essential terminology for the harmonic balance method in order to formulate/set notation for a PDE analysis.

We assume the PDE we would like to numerically solve takes the form

$$\mathcal{F}(u(x,t),f(x,t))=0$$

on a domain $\Omega \subset \mathbb{R}^3$, with solution $u(x,t)$, and where $f(x,t)$ is a quasi-periodic function band-limited to a specified set of *fundamental frequencies* $\omega_1 > \ldots > \omega_M > 0$, whose frequency spectrum is supported at positive integer linear combinations of these fundamental frequencies. For a successful harmonic balance method, the solution $u(x,t)$ should itself be a quasi-periodic function. We call $\vec{\omega} = (\omega_1,\ldots,\omega_M)$ the *vector of fundamental frequencies*, and M the *number of fundamental harmonics*. A harmonic balance method must adopt $\vec{\omega}$ and a *truncation scheme* $\mathcal{T} \subsetneq \mathbb{Z}_*^M$ which indexes linear combinations of the fundamental frequencies without loss of information, i.e. $|\mathcal{T}| = |\vec{\omega}\cdot\mathcal{T}|$. The truncation scheme dictates the *truncated frequency basis* $\vec{\omega}\cdot\mathcal{T}$ and the solution ansatz

$$u(x,t) = U_0(x) + \sum_{\alpha\in\mathcal{T}} \left[U_{\vec{\alpha}}^c(x)cos(\vec{\alpha}\cdot\vec{\omega}t) + U_{\vec{\alpha}}^s(x)sin(\vec{\alpha}\cdot\vec{\omega}t)\right].$$

We will assume f takes the corresponding form, following the truncation scheme,

$$f(x,t) = F_0(x) + \sum_{\vec{\alpha}\in\mathcal{T}} \left[F_{\vec{\alpha}}^c(x)cos(\vec{\alpha}\cdot\vec{\omega}t) + F_{\vec{\alpha}}^s(x)sin(\vec{\alpha}\cdot\vec{\omega}t)\right].$$

We call $P := max_{\vec{\alpha}\in\mathcal{T}}||\vec{\alpha}||_{\ell^\infty}$ the *order of the truncation scheme*, and $H := |\mathcal{T}|$ the *total number of harmonics* of the truncation scheme. When necessary, we will order $\mathcal{T}$ using the (left-to-right) lexicographical order $\succ$ on $\mathbb{Z}_*^M$, so that $\mathcal{T} = \{\vec{\alpha}_1 \succ \vec{\alpha}_2 \succ \cdots \succ \vec{\alpha}_H\}$. Note that this lexicographical ordering reasonably yields:

$$\vec{\alpha}_1\cdot\vec{\omega} > \cdots > \vec{\alpha}_H\cdot\vec{\omega}.$$

A successful harmonic balance method first creates a system of equations whose degrees of freedom are the coefficients of the solution ansatz, $\{U_0(x), U_k^c(x), U_k^s(x) | k = 1,\ldots,2H\}$, and then minimizes $\mathcal{F}(u)$ (or some functional of $\mathcal{F}$, in the method of weighted residuals) as much as possible (in a sense dictated by the FEM, say by GMRES). Of course, when it is known a priori that the solution u is band limited, then a reasonable implementation of the harmonic balance method with a truncated frequency basis containing the frequency spectrum of u should recover u exactly (up to machine precision). In this note, our focus lies solely in obtaining the system of harmonic balance equations.

5 Mathematical description of approaches

5.1 Method of Undetermined Coefficients

By expressing $\mathcal{F}(u)$ analytically in the frequency domain according to the truncation scheme $\mathcal{T}$, we have the expression

$$\begin{aligned}\mathcal{F}(u(x,t),f(x,t),t) =& \mathcal{F}_0(x)\left(U_0(x),U^c_{\vec{\gamma}}(x),U^s_{\vec{\gamma}}(x),F_0(x),F^c_{\vec{\gamma}}(x),F^s_{\vec{\gamma}}(x)\right)\\ &+\sum_{\vec{\alpha}\in\mathcal{T}}\mathcal{F}^c_{\vec{\alpha}}(x)\left(U_0(x),U^c_{\vec{\gamma}},U^s_{\vec{\gamma}},F_0,F^c_{\vec{\gamma}},F^s_{\vec{\gamma}}\right)cos(\vec{\alpha}\cdot\omega t)\\ &+\sum_{\vec{\alpha}\in\mathcal{T}}\mathcal{F}^s_{\vec{\alpha}}(x)\left(U_0(x),U^c_{\vec{\gamma}}(x),U^s_{\vec{\gamma}}(x),F_0(x),F^c_{\vec{\gamma}}(x),F^s_{\vec{\gamma}}(x)\right)sin(\vec{\alpha}\cdot\omega t)\\ &+\sum_{\vec{\beta}\in\left(\mathbb{Z}^M_*\setminus\mathcal{T}\right)}\mathcal{F}^c_{\vec{\beta}}(x)\left(U_0(x),U^c_{\vec{\gamma}}(x),U^s_{\vec{\gamma}}(x),F_0(x),F^c_{\vec{\beta}}(x),F^s_{\vec{\beta}}(x)\right)cos(\vec{\beta}\cdot\omega t)\\ &+\sum_{\vec{\beta}\in\left(\mathbb{Z}^M_*\setminus\mathcal{T}\right)}\mathcal{F}^s_{\vec{\beta}}(x)\left(U_0(x),U^c_{\vec{\gamma}}(x),U^s_{\vec{\gamma}}(x),F_0(x),F^c_{\vec{\beta}}(x),F^s_{\vec{\beta}}(x)\right)sin(\vec{\beta}\cdot\omega t)\end{aligned}$$

We adopt the convention that a subscript vector $\vec{\gamma}$, especially in the argument of a function, indicates the dependence of that function on all $\gamma\in\mathcal{T}$. In particular, this helps us suppress notation:

$$\mathcal{F}_0(U_0,U^c_{\vec{\gamma}},U^s_{\vec{\gamma}},F_0,F^c_k,F^s_k) := \mathcal{F}_0\left(U_0,U^c_{\vec{\alpha}_1},U^s_{\vec{\alpha}_1},\ldots,U^c_{\vec{\alpha}_H},U^s_{\vec{\alpha}_H},F_0,F^c_{\vec{\alpha}_1},F^s_{\vec{\alpha}_1},\ldots,F^c_{\vec{\alpha}_H},F^s_{\vec{\alpha}_H}\right)$$

The first three summands yields a system of $2H+1$ coefficients $\mathcal{F}_0,\mathcal{F}^c_{\vec{\gamma}},\mathcal{F}^s_{\vec{\gamma}}$ in the $2H+1$ undetermined coefficients $U_0(x),U^c_{\vec{\gamma}}(x),U^s_{\vec{\gamma}}(x)$. Equating these to zero yields the coupled, non-linear system of harmonic balance equations. We discard the last two summands for a few reasons: the are not captured by the ansatz form, their equation to zero would lead to an over-determined system of equations, and they heuristically have very small/numerically insignificant amplitudes because of the Plancherel theorem (assuming the solution is $\mathcal{L}^1$). Thus, in order to perform the harmonic balance, we must collect the system of equations

$$\begin{aligned}0 &= \mathcal{F}_0(x)(U_0(x),U^c_{\vec{\gamma}}(x),U^s_{\vec{\gamma}}(x),F_0(x),F^c_{\vec{\gamma}}(x),F^s_{\vec{\gamma}}(x))\\ 0 &= \mathcal{F}^c_{\vec{\alpha}}(x)(U_0(x),U^c_{\vec{\gamma}}(x),U^s_{\vec{\gamma}}(x),F_0(x),F^c_{\vec{\gamma}}(x),F^s_{\vec{\gamma}}(x))\\ 0 &= \mathcal{F}^s_{\vec{\alpha}}(x)(U_0(x),U^c_{\vec{\gamma}}(x),U^s_{\vec{\gamma}}(x),F_0(x),F^c_{\vec{\gamma}}(x),F^s_{\vec{\gamma}}(x))\end{aligned}$$

for all $\vec{\alpha}\in\mathcal{T}$ and for each $x\in\Omega$. In the discretization, of course, only finitely many $\{x_n|0\le n\le N\}\subset\Omega$ are considered. We can then enforce this system of equations weakly:

$$\begin{aligned}0 &= \int_V \mathcal{F}_0(x)(U_0(x),U^c_{\vec{\gamma}}(x),U^s_{\vec{\gamma}}(x),F_0(x),F^c_{\vec{\gamma}}(x),F^s_{\vec{\gamma}}(x))\cdot\phi_n\,dx\\ 0 &= \int_V \mathcal{F}^c_{\vec{\alpha}}(x)(U_0(x),U^c_{\vec{\gamma}}(x),U^s_{\vec{\gamma}}(x),F_0(x),F^c_{\vec{\gamma}}(x),F^s_{\vec{\gamma}}(x))\cdot\phi_n\,dx\\ 0 &= \int_V \mathcal{F}^s_{\vec{\alpha}}(x)(U_0(x),U^c_{\vec{\gamma}}(x),U^s_{\vec{\gamma}}(x),F_0(x),F^c_{\vec{\gamma}}(x),F^s_{\vec{\gamma}}(x))\cdot\phi_n\,dx\end{aligned}\tag{4}$$

for all volumes V in the mesh of Ω, for all nodal basis functions ϕ_n (for a first-order FEM), and arrive at a system of non-linear equations of the harmonic balance.

5.2 Time Collocation Method

Suppose that $\vec{\alpha} \cdot \vec{\omega}$ divides T for all $\alpha \in \mathcal{T}$. Let V be a volume in the mesh of Ω, and again let $\{x_n | 0 \leq n \leq N\}$ be the mesh vertices with nodal basis functions ϕ_n. Observe that the integrals (over space and time) of $\mathcal{F}(u,t)$ against the time-space tensor basis (Fourier-Lagrange) yields:

$$\int_V \left[\int_0^T \mathcal{F}(u,t) \cdot 1 \, dt \right] \cdot \phi_n \, dx = T \int_V \mathcal{F}_0(x)(U_0(x), U^c_{\vec{\gamma}}(x), U^s_{\vec{\gamma}}(x), F_0(x), F^c_{\vec{\gamma}}(x), F^s_{\vec{\gamma}}(x)) \cdot \phi_n \, dx$$

$$\int_V \left[\int_0^T \mathcal{F}(u,t) \cdot cos(\vec{\alpha} \cdot \vec{\omega} t) \, dt \right] \cdot \phi_n \, dx = \frac{T}{4\pi} \int_V \mathcal{F}^c_{\vec{\alpha}}(x)(U_0(x), U^c_{\vec{\gamma}}(x), U^s_{\vec{\gamma}}(x), F_0(x), F^c_{\vec{\gamma}}(x), F^s_{\vec{\gamma}}(x)) \cdot \phi_n \, dx$$

$$\int_V \left[\int_0^T \mathcal{F}(u,t) \cdot sin(\vec{\alpha} \cdot \vec{\omega} t) \, dt \right] \cdot \phi_n \, dx = \frac{T}{4\pi} \int_V \mathcal{F}^s_{\vec{\alpha}}(x)(U_0(x), U^c_{\vec{\gamma}}(x), U^s_{\vec{\gamma}}(x), F_0(x), F^c_{\vec{\gamma}}(x), F^s_{\vec{\gamma}}(x)) \cdot \phi_n \, dx$$

the right-hand sides of which coincide (up to a constant scalar) with the right-hand sides of the system of equations from the Method of Undetermined Coefficients, above. Note that the time-space tensor product basis is $\{\phi_n \cdot 1, \phi_n \cdot cos(\vec{\alpha} \cdot \vec{\omega} t), sin(\vec{\alpha} \cdot \vec{\omega} t)\}$. In other words, the weak form of the harmonic balance equations arises from the $\mathcal{L}^2(V, [0,T])$-projection of $\mathcal{F}(u,t)$ onto the time-space tensor basis, and performing only the temporal integral. So, we can alternatively equate the left-hand sides to 0, and form a superficially different system of harmonic balance equations.

By applying Clenshaw-Curtis quadrature, we can approximate the temporal integration:

$$\int_0^T \mathcal{F}(u,t) \cdot 1 \, dt = \frac{1}{L} \left[\sum_{k=0}^{L} w_k \mathcal{F}(u, t_k) \right]$$

$$\int_0^T \mathcal{F}(u,t) \cdot cos(\vec{\alpha} \cdot \vec{\omega} t) \, dt = \frac{1}{L} \left[\sum_{k=0}^{L} w_k \mathcal{F}(u, t_k) \cdot cos(\vec{\alpha} \cdot \vec{\omega} t_k) \right]$$

$$\int_0^T \mathcal{F}(u,t) \cdot sin(\vec{\alpha} \cdot \vec{\omega} t) \, dt = \frac{1}{L} \left[\sum_{k=0}^{L} w_k \mathcal{F}(u, t_k) \cdot sin(\vec{\alpha} \cdot \vec{\omega} t_k) \right]$$

Note that if $\mathcal{F}(u,t)$ actually has a frequency spectrum supported on $\vec{\omega} \cdot \mathcal{T}$, then this approximation is exact (even if we use the trapezoidal rule) when $L > 2H$. Hence, we arrive at

$$\begin{aligned} 0 &= \frac{1}{L} \sum_{k=0}^{L} w_k \left[\int_V \mathcal{F}(u(x,t_k), t_k) \cdot \phi_n \, dx \right] \\ 0 &= \frac{1}{L} \sum_{k=0}^{L} w_k cos(\vec{\alpha} \cdot \vec{\omega} t_k) \left[\int_V \mathcal{F}(u(x,t_k), t_k) \cdot \phi_n \, dx \right] \\ 0 &= \frac{1}{L} \sum_{k=0}^{L} w_k sin(\vec{\alpha} \cdot \vec{\omega} t_k) \left[\int_V \mathcal{F}(u(x,t_k), t_k) \cdot \phi_n \, dx \right] \end{aligned} \tag{5}$$

an alternative formulation to the system of non-linear harmonic balance equations.

6 Implementation of approaches in Panzer/Trilinos

In this section we outline means for obtaining the harmonic balance equations corresponding to an arbitrary PDE model implemented in the Panzer/Trilinos framework, with respect to the

approaches described in the previous section. By minimally modifying the existing time-domain (transient or stationary) FEM model, and by externally handling the harmonic balance calculations (with respect to the time-domain equation set), we seek to gain the capability to model any time-domain equation set in the frequency domain with few adaptations. For the sake of concreteness, we will work with a Panzer tutorial problem: a stationary Helmholtz equation on a 2-D domain, with a constant Dirichlet boundary condition and a non-linear source term. The tutorial files themselves can be found in the Trilinos repository at:

```
{Trilinos directory}/packages/panzer/adapters-stk/tutorial/step0*/
```

However, the Helmholtz equation and the contribution of a non-linear source term are exercises to be completed in the tutorial. The resulting files we work with are:

```
main.cpp
Step01_LinearFunction.cpp
Step01_SinXSinYFunction.cpp
Step01_BCStrategy_Dirichlet_Constant_decl.hpp
Step01_BCStrategy_Dirichlet_Constant.hpp
Step01_BCStrategy_Dirichlet_Constant_impl.hpp
Step01_BCStrategy_Factory.hpp
Step01_BCStrategy_FreqDom_Dirichlet_decl.hpp
Step01_BCStrategy_FreqDom_Dirichlet.hpp
Step01_BCStrategy_FreqDom_Dirichlet_impl.hpp
Step01_ClosureModel_Factory.hpp
Step01_ClosureModel_Factory_impl.hpp
Step01_ClosureModel_Factory_TemplateBuilder.hpp
Step01_EquationSetFactory.hpp
Step01_EquationSet_Helmholtz.hpp
Step01_EquationSet_Helmholtz_impl.hpp
Step01_EquationSet_Projection.hpp
Step01_EquationSet_Projection_impl.hpp
Step01_LinearFunction.hpp
Step01_LinearFunction_impl.hpp
Step01_SinXSinYFunction.hpp
Step01_SinXSinYFunction_impl.hpp
input.xml
```

We will refer to an arbitrary transient PDE model as `EquationSet_TimeDomain`. For our purposes, these are derived from `panzer::EquationSet_DefaultImpl`. With the aforementioned goals in mind, we create a new equation set

```
EquationSet_FreqDom :  panzer::EquationSet_DefaultImpl
```

which constructs the harmonic balance residuals inside its

```
EquationSet_FreqDom :  buildAndRegisterEquationSetEvaluators()
```

member function by instantiating a `EquationSet_TimeDomain` object and calling its corresponding member function

```
EquationSet_TimeDomain :  buildAndRegisterEquationSetEvaluators()
```

appropriately. The precise details and construction of the harmonic balance residuals depends on the approach. Concretely, we create these new files for the frequency domain equation set and also for a library of calculators of the various harmonic balance parameters (calculating the truncation order, truncation scheme, ordering the harmonics, dealing with the multi-indices, etc.)

```
Step01_EquationSet_FreqDom.hpp
Step01_EquationSet_FreqDom_impl.hpp

Step01_FreqDom_Parameters.cpp
```

In the following, the descriptions of the implementations can be almost completely contained within these new files. This is in accordance with our goal of minimally affecting the existing `EquationSet_TimeDomain` files.

6.1 Method of Undetermined Coefficients

Observe that the finite element numerical integration of $\int_V \mathcal{F}(u,t)\phi_n dx$ in the system of equations (4) provides us with a formula relating the coefficients of u over space, for a fixed point in time. The structure of this formula determines the formulae $\mathcal{F}_0$, $\mathcal{F}^c_{\vec{\gamma}}$, $\mathcal{F}^s_{\vec{\gamma}}$.

Record the transient residual formula

Create degrees of freedom which are the amplitudes $U_0, U^c_{\vec{\gamma}}, U^s_{\vec{\gamma}}$ of the solution ansatz. If the formula can actually be recovered, then we can create a new "Fourier transform" evaluator which takes that formula and performs the appropriate evaluations on the $U_0, U^c_{\vec{\gamma}}, U^s_{\vec{\gamma}}$ to calculate the method of undetermined coefficients approach to the harmonic balance formulae $\mathcal{F}_0(U_0, U^c_{\vec{\gamma}}, U^s_{\vec{\gamma}})$, $\mathcal{F}^c_{\vec{\alpha}}(U_0, U^c_{\vec{\gamma}}, U^s_{\vec{\gamma}})$, $\mathcal{F}^s_{\vec{\alpha}}(U_0, U^c_{\vec{\gamma}}, U^s_{\vec{\gamma}})$.

Overloading operators (á la Sacado)

Create a new `Sacado::Fourier` class analogous to the `Sacado::Tay` Taylor polynomial automatic differention class. With operator overloading, we can re-use the code for the evaluators from the time domain equation set because the calculation of the residual will automatically perform calculations in the truncated Fourier basis.

6.2 Time Collocation Method

Observe that each integral of the summations in the time collocation approach to the harmonic balance formulae (5) can be seen as the weighted summand, whose value is the residual at time

t_k. So, the system of non-linear equations is simply a weighted linear combination of the FEM residuals at the $L+1$ points in time. This is the approach we are currently pursuing.

Note that, for each of the time collocation points t_k, we need a time domain degree of freedom $U_k(x_n)$ with which to evaluate the `EquationSet_TimeDomain` residual evaluators. Note that

$$U_k(x_n) = U(x_n, t_k) = U_0(x_n) + \sum_{\vec{\gamma} \in \mathcal{I}} \left[U^c_{\vec{\gamma}}(x_n) \cdot cos\left(\vec{\gamma} \cdot \vec{\omega} \cdot t_k\right) + U^s_{\vec{\gamma}}(x_n) \cdot sin\left(\vec{\gamma} \cdot \vec{\omega} \cdot t_k\right) \right]$$

Now, each of the factors $cos\left(\vec{\gamma} \cdot \vec{\omega} \cdot t_k\right)$ and $sin\left(\vec{\gamma} \cdot \vec{\omega} \cdot t_k\right)$ are constants, so $U_k(x_n)$ is nothing other than a linear combination of the frequency domain degrees of freedom, U_0, $U^c_{\vec{\gamma}}$, and $U^s_{\vec{\gamma}}$.

Solving this system of equations then gives us $L+1$ snapshots in time of the solution. By the Shannon-Nyquist Sampling theorem, if

$$L > 2 \max_{\vec{\gamma} \in \mathcal{I}} \{\vec{\gamma} \cdot \vec{\omega}\},$$

then we can recover the time domain expression from the coefficients $U_0, U^c_{\vec{\gamma}}, U^s_{\vec{\gamma}}$. By this, we mean that, mathematically speaking, if it is known a priori that the solution is truly band-limited with a frequency spectrum supported within the harmonic balance truncation basis, then this can recover the true solution.

Note that this method sums over many time collocation points. A smaller value of L can be used if we apply a frequency remapping method. Furthermore, in order to efficiently perform this large summation, we can employ a new Panzer/Phalanx capability - the "contributes to" property from `PHX::EvaluatorWithBaseImpl`. Whereas the tutorial problem's evaluators are registered via:

```
{
    Teuchos::ParameterList p;
        p.set("Sum Name",      dof_name_+"timedom_res_at_time_coll_pt");
        p.set("Values Names",  residual_operator_names_RCP);
        p.set("Data Layout",   basis->functional);

        RCP< PHX::Evaluator<panzer::Traits> >
          op = rcp(new panzer::Sum<EvalT,panzer::Traits>(p));

    fm.template registerEvaluator<EvalT>(op);
}
```

we now instead register them via:

```
{
    Teuchos::ParameterList p;
        p.set(''Sum Name'',      dof_name_+''timedom_res_at_time_coll_pt'');
        p.set(''Values Names'',  residual_operator_names_RCP);
        p.set(''Data Layout'',   basis->functional);
```

```
            // begin: use the ``contributes to'' Panzer capability
            RCP< PHX::EvaluatorWithBaseImpl<panzer::Traits> >
              op = rcp(new panzer::Sum<EvalT,panzer::Traits>(p));

            PHX::Tag<typename EvalT::ScalarT>
              tag0(``FreqDom Residual'',basis->functional);
            op->addContributedField(tag0);
            // end: use the ``contributes to'' Panzer capability

        fm.template registerEvaluator<EvalT>(op);
    }
```

Note that we must appropriately zero out the fields which are contributing.

Intrepid-Fourier tensor basis

Another benefit of the time collocation approach is that, if `EquationSet_TimeDomain` uses Intrepid's compatible function spaces

$$\begin{aligned} H(grad,\Omega) &= HGRAD \\ H(curl,\Omega) &= HCURL \\ H(div,\Omega) &= HDIV \\ \mathcal{L}^2(\Omega) &= HVOL \end{aligned}$$

then the time collocation approach inherits all of these advantages. In essence, the time collocaiton approach is an Intrepid-Fourier basis finite element method whose basis functions are tensored over space and time. This approach essentially manually integrates out the time dimension, and evaluates each value of the residual in time as a linear combination of the frequency domain degrees of freedom.

References

[BBGG08] Francesco Bertazzi, Fabrizio Bonani, Simona Donati Guerrieri, and Giovanni Ghione. Large-signal device simulation in time- and frequency-domain: a comparison. *COMPEL - The international journal for computation and mathematics in electrical and electronic engineering*, 27(6):1319–1325, 2008.

[FML96] P. Feldmann, B. Melville, and D. Long. Efficient frequency domain analysis of large nonlinear analog circuits. In *Proceedings of Custom Integrated Circuits Conference*, pages 461–464, May 1996.

[GG89] Herbert Gajewski and Konarad Grger. Semiconductor equations for variable mobilities based on boltzmann statistics or fermi-dirac statistics. *Mathematische Nachrichten*, 140(1):7–36, 1989.

[Jer85] Joseph W. Jerome. Consistency of semiconductor modeling: An existence/stability analysis for the stationary van roosbroeck system. *SIAM Journal on Applied Mathematics*, 45(4):565–590, 1985.

[Jer87] Joseph W. Jerome. Evolution systems in semiconductor device modeling: A cyclic uncoupled line analysis for the gummel map. *Mathematical Methods in the Applied Sciences*, 9(1):455–492, 1987.

[Jün09] Ansgar Jüngel. *Drift-Diffusion Equations*, pages 1–29. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[KSSV88] K. S. Kundert, G. B. Sorkin, and A. Sangiovanni-Vincentelli. Applying harmonic balance to almost-periodic circuits. *IEEE Transactions on Microwave Theory and Techniques*, 36(2):366–378, Feb 1988.

[KSV86] Kenneth S Kundert and Alberto Sangiovanni-Vincentelli. Simulation of nonlinear circuits in the frequency domain. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 5(4):521–535, 1986.

[NW99] Ognen J. Nastov and Jacob K. White. Time-mapped harmonic balance. *Design Automation Conference*, pages 641–646, 1999.

[Tro98] Boris Troyanovsky. *Frequency domain algorithms for simulating large signal distortion in semiconductor devices*. PhD thesis, Stanford University, 1998.

[TYD00] Boris Troyanovsky, Zhiping Yu, and Robert W. Dutton. Physics-based simulation of nonlinear distortion in semiconductor devices using the harmonic balance method. *Computer Methods in Applied Mechanics and Engineering*, 181(4):467 – 482, 2000.

[VR50] W. Van Roosbroeck. Theory of the flow of electrons and holes in germanium and other semiconductors. *Bell System Technical Journal*, 29(4):560–607, 1950.

DISTRIBUTION:

1	MS 1081	Paiboon Tangyunyong, 01755
1	MS 1168	Steven Wix, 01356
1	MS 1177	Joseph Castro, 01355
1	MS 1177	Xujiao (Suzey) Gao, 01355
1	MS 1177	Gary Hennigan, 01355
1	MS 1177	Eric Keiter, 01355
1	MS 1177	Lawrence C. Musson, 01355
1	MS 1177	Mihai Negoita, 01355
1	MS 1177	Heidi K. Thornquist, 01355
1	MS 1179	Harry P. Hjalmarson, 01341
1	MS 1318	Roger P. Pawlowski, 01355
1	MS 1318	Kara J. Peterson, 01442
1	MS 1318	Eric T. Phipps, 01441
1	MS 1320	Eric C. Cyr, 01442
1	MS 1321	Jason M. Gates, 01442
1	MS 1352	Lorena I. Basilio, 01152
1	MS 1352	Matthew Bettencourt, 01168
1	MS 1352	Keith Cartwright, 01152
1	MS 0899	Technical Library, 9536 (electronic copy)

v1.40