A Model for Self-Phase Modulation Noise in Fiber-link Communication Systems

by

Alireza Hosseinzadeh

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2019

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Recent tele-communication hugely relies on fiber-optic systems. Linear fiber-optic impairments can be modeled with chromatic and polarization-mode dispersion. However, nonlinear impairments such as SPM and XPM noise are much more complicated to model and compensate. In this thesis, we try to manipulate an original model which is used by Ciena Corporation -called the C-model- to achieve a simpler form which provides lower time-complexity and performs efficiently on random data stream during online transmission.

This simplified model is defined based on C-matrix which is used to calculate fiber-link Self-Phase Modulation (SPM) noise. Our model, which is fitted on random training data sets, is evaluated with respect to different performance measures such as Minimum Mean Square Error (MMSE). Different analytic and iterative algorithms such as gradient Descent and Moore-Penrose pseudo-inverse method are proposed for model fitting and it is shown that these algorithms perform quite efficiently according to error calculations. Also, we propose a method that enables us to track the perturbations in model parameters which are caused by the changes in channel characteristics that can happen frequently in online fiber-link transmission.

Acknowledgements

I would like to thank my supervisor, Professor Amir Keyvan Khandani for his knowledge, guidance and unconditional support throughout my program. I would also thank my friends Nima, Shayan, Ali, Takin and Saber who made this possible. Also, I would like to thank Dr. Shahab Oveis Gharan and Dr. Masoud Ebrahimzadeh from Ciena Corporation whose contributions have been essential for conducting this thesis.

Dedication

This is dedicated to my family for their unconditional love.

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Overview

In fiber-optic communications, carrier signal is the light beam which is an electromagnetic wave that can be modulated for information transmission. The basic phenomenon which guides the light beam inside of a fiber is Total Internal Reflection and it was known from nineteenth century. Although uncladded glass fibers were fabricated in 1920s, the main invention which led to considerable improvement in fiber characteristics and gave birth to fiber optics field was the use of a cladding layer in the fabrication of fibers in 1950s [2].

In a fiber-optic lightwave system, several types of optical fibers may consist a fiber link that optical bit stream propagates through. The length of the link varies from a few to thousands of kilometers. The quality of the signal degrades during the propagation and it's desired for any lightwave system to control these degradations as much as possible to prevent the loss of original information when the signal is recovered at the receiver[1].

Our main objective in this thesis is to propose a model -based on C-model from Ciena Corporation- that can predict the impairments of the transmitted signal so we can mitigate them. Next section discusses a brief literature review about this topic.

In chapter 2, first we introduce the C-model from Ciena Corporation. Then, we propose our simplified model and provide the derivations resulted from it.

Chapters 3, 4 and 5 discuss different model fitting methods with their respective error analysis, such as pseudo-inverse, gradient descent and Hermitian inner product.

In chapter 6, we provide some methods which are based on previous model fitting algorithms to be able to update model parameters when channel characteristics change over time. Finally, some other performance measures are discussed in chapter 7.

I should also mention that this work and [21] have been part of a joint project for Ciena Corporation.

1.2 Literature Review

Digital coherent detection in wavelength-division multiplexing (WDM) standard technology is considered as an irreplaceable method for high capacity transmissions in 100 Gb/s and beyond. The linear impairments, such as Chromatic Dispersion (CD) and Polarization Mode Dispersion (PMD) can be compensated by Digital Signal Processing (DSP) methods[31][1].

Optical fiber refractive index depends on the intensity of transmitted signal. This dependence gives rise to Kerr effect which is the source of the nonlinearity of this medium. For signals with small enough power, the nonlinearities can be neglected throughout the link. This approximation is valid for many short-haul systems (distance $<$ 100 km). However, for long-haul systems where optical amplifiers are utilized in a cascaded chain, nonlinearities become much more significant and can not be excluded. This originates from two main reasons. First The degradation in SNR caused by the noise added at each amplifier builds up to the extent that the signal can only be recovered at high launching powers ($>$ 1mw). Second, nonlinearities accumulate at each amplifier and distortion increases as length of the transmission link increases. Self-phase modulation (SPM), cross-phase modulation (XPM) and four-wave mixing (FWM) are some of these nonlinearities[1]. For the system without in-line dispersion compensation, the nonlinear distortions are expected to be the major complication [31].

Further on, we provide a simple form of deriving the SPM noise from Kerr effect to gain a perspective of the nature of this noise.

Nonlinear Schrodinger equation gives rise to evolution along distance z of the electric field $E(z)$ [27]. Neglecting dispersion we will have:

$$\frac{dE(z)}{dz} = \gamma |E(z)|^2(-jE(z))$$

where γ is the nonlinear coefficient and $|E(z)|^2 E(z)$ is resulted by Kerr effect.

Since the power is constant, the Kerr effect comes into effect only as a phase rotation:

$$\frac{d|E|e^{j\varphi}}{dz} = \frac{d|E|}{dz}e^{j\varphi} + j|E|e^{j\varphi}\frac{d\varphi}{dz} = -j\gamma|E(z)|^3 e^{j\varphi}$$
$$\frac{d\varphi}{dz} = -\gamma|E|^2$$

Therefore, the phase φ at coordinate z is:

$$\varphi(z) = \varphi(0) - \gamma|E(0)|^2 z \tag{1.1}$$

Equation 1.1 denotes that SPM noise is originated from the power of the field[27][2]. So, this nonlinearity is self-induced and it causes a frequency shift, known as frequency chirping, and results in the spectral broadening of the optical pulse. Spectral broadening and chirping effect is proportional to the input power[16][3]. Some techniques that are used for mitigation of nonlinear impairments of a fiber-link transmission system are briefly mentioned in the following.

Digital back-propagation (DBP) is a method that can be implemented either at transmitter or receiver side. This method is based on split-step Fourier method (SSFM) which is widely used to solve the Manakov equation (nonlinear Schrodinger equation (NLSE))[8][13]. Analytical solution to the Manakov equation is only known for some particular cases, such as zero-dispersion transmission. Therefore, numerical solutions such as DBP are proposed. The basic idea is to find a solution for inverse Manakov equation and use it to mitigate the nonlinear effects. DBP uses the idea of transmitting the signal through an imaginary fiber with inverse parameters to compensate impairments. The fiber link is divided to several steps with small length and considering the order of linear and nonlinear sections, different implementations of DBP are proposed[26]. At small steps, DBP has a high performance. However, real-time implementation suffers a high computational complexity as the number steps increases.

Another powerful tool for solving the Manakov equation is Volterra series transfer function (VSTF)[24]. After modeling the channel, Volterra series based nonlinear equalizer (VNLE) uses the p-th order theory developed by Schetzen [30] to find the inverse VSTF kernels. These kernels characterize the equalizer to compensate nonlinearities and chromatic dispersion. Similar to DBP, VNLE attempts to find the parameters of the inverse channel for mitigation. The capability of NLVE to implement the compensation of linear and nonlinear sections in parallel results in lower computational load compared to DBP. NLVE can be processed in time and frequency domain[28][23].

In Nyquist WDM super-channel, neighboring subcarriers affect transmission quality due to interference. Former techniques such as DBP and NLVE don't perform efficiently because of this effect. Therefore, inter-subcarrier nonlinear interference canceler (INIC) method based on Volterra series is proposed[4]. Using the knowledge of neighboring subcarrier detection to cancel the interference of the desired subcarrier is the basic idea behind this method. It detects the neighboring subcarrier, uses Volterra series to regenerate the model and deduct the effect from the desired subcarrier. The mentioned method uses decision feedback equalizer (DEF)[9]. INIC mitigates both intra-subcarrier and inter-subcarrier linear and nonlinear effects in Nyquist WDM systems.

Perturbation-based compensation techniques is also widely used for modeling the fiber and mitigating impairments. These techniques provide an approximate numerical solution to the Manakov equations and can be implemented either at transmitter or receiver side[32][17]. In this method, nonlinear distortion is modeled as a first-order perturbation added to original solution for linear case. Pulse shape and model parameters determine the number of perturbation coefficients. Perturbation-based compensation can be implemented on a single stage which reduces the complexity in comparison to DBP and VLNE. However, it needs a large number of perturbation terms which makes its practical implementation challenging[11].

Optical phase conjugation (OPC) and digital phase conjugated twin waves (PCTW) are some of phase conjugation techniques to mitigate nonlinearities. OPC inverts the spectrum of data signal in optical domain in the middle of the link. The basic idea is to cancel out the nonlinearity generated in the first segment with the second one. Positioning and symmetry of the link are some of the challenges in implementation of this method[23]. Also, OPC-based transmission is sensitive to nonlinear effect which acts similar to positive feedback for nonlinearity.

PCTW is a digital signal processing method which is implemented at the receiver side. In dual polarization format, PCTW transmits the signal on polarization x and its conjugate on polarization y. Therefore, the first-order phase shift can be cancelled out with superposition of the two signals[19][18]. Although PC techniques are computationally efficient, one drawback is the loss of half of spectral efficiency due to transmission of conjugate signal on polarization y. Some recent approaches try to resolve this problem using a modified method[34].

Ciena Corporation utilizes split-step Fourier method (SSFM) to model the nonlinearities. We won't go through the details of the nonlinear Schrodinger equations, instead the result which is the C-matrix is used to model SPM noise. A careful derivation of the C-model is provided in [29].

Chapter 2

Models for SPM noise

2.1 Original C-model from Ciena Corporation

Nonlinear propagation in optical fiber is well described by the Nonlinear Schrodinger Equation (NLSE). A careful derivation follows from Maxwell's equations for the propagating fiber mode. Most fiber nonlinearities originate through nonlinear refraction (a dependence of the refractive index on the field intensity). Due to Kerr effect, refractive index of the glass is modulated by the intensity of the optical field [27]. The procedure to derive the solutions from these equations are not the main focus of this thesis. Therefore we concentrate on the results derived by Ciena Corporation to continue our main objective which is manipulating the model to achieve a simpler form and implement model fitting methods. The solution to these equations results in a model named C-model which puts the nonlinearities into effect as mentioned in the following.

The nonlinear noise in X-pole in the time zero is approximated as:

$$\Delta \mathrm{A_x} = \mathrm{SPM_1} + \mathrm{SPM_2} + \sum_{\mathrm{w}} \left(\mathrm{XPM_{1w}} + \mathrm{XPM_{2w}} + \mathrm{XPM_{3w}} + \mathrm{XPM_{4w}} \right) \tag{2.1}$$

where $\Delta \mathrm{A_x}$ indicates nonlinear noise in X-pole of a transmitted symbol $\mathrm{A_x}$ and W indicates the neighboring channels and the summation is over all of them.

The effective nonlinear noise can be written as:

$$\Delta \mathrm{A_{ex}} = \Delta \mathrm{A_x} - \mathrm{E} \left[\Delta \mathrm{A_x} | \mathrm{A_x}(0) \right] \tag{2.2}$$

Equation 2.2 suggest to deduct the conditional mean of the noise from calculations.

Each of the terms in Equation 2.1 are calculated as follows:

$$
\begin{aligned}
SPM_1 &= \sum_{m,n} C_{m,n}^{spm} A_x(m) A_x(n) A_x^c(m+n) \\
SPM_2 &= \sum_{m,n} C_{m,n}^{spm} A_x(m) A_y(n) A_y^c(m+n) \\
XPM_{1w} &= \sum_{m,n} C_{m,n}^{xpmw} A_x(m) B_x(n) B_x^c(m+n) \\
XPM_{2w} &= \sum_{m,n} C_{m,n}^{xpmw} A_x(m) B_y(n) B_y^c(m+n) \\
XPM_{3w} &= \sum_{m,n} C_{m,n}^{xpolmw} B_x(m) A_x(n) B_x^c(m+n) \\
XPM_{4w} &= \sum_{m,n} C_{m,n}^{xpolmw} B_x(m) A_y(n) B_y^c(m+n)
\end{aligned}
\tag{2.3}
$$

From equations 2.3 we can conclude:

- As the nature of the noises suggests, Self-Phase Modulation (SPM) noise is not dependent on symbols from other channels, however Cross-Phase Modulation (XPM) is.
- SPM and XPM noises added to a single symbol in X-pole are dependent on itself and neighboring symbols from both X-pole and Y-pole.
- C^{spm} , C^{xpmw} and C^{xpolmw} are the matrices used to calculate these noises[7].

C^{spm} is called the C-matrix and it's used to calculate the SPM noise added to transmitted signal. C-matrix size and the values of its elements depend on the type of the fiber and number of the spans. However, for almost all fiber-link types and lengths, C-matrix is a considerably large matrix which can cause computational complexity[29]. This fact gives us an incentive to try to define a simpler model which can provide sufficiently accurate results and have lower complexity.

The output of C-matrix are depicted in figures 2.1 and 2.2. In figure 2.2, we can see the raw output of SPM C-matrix. The asymmetric form of the output in figure 2.1 is resulted from 90-degree phase shift which is an inherent property of C-matrix.

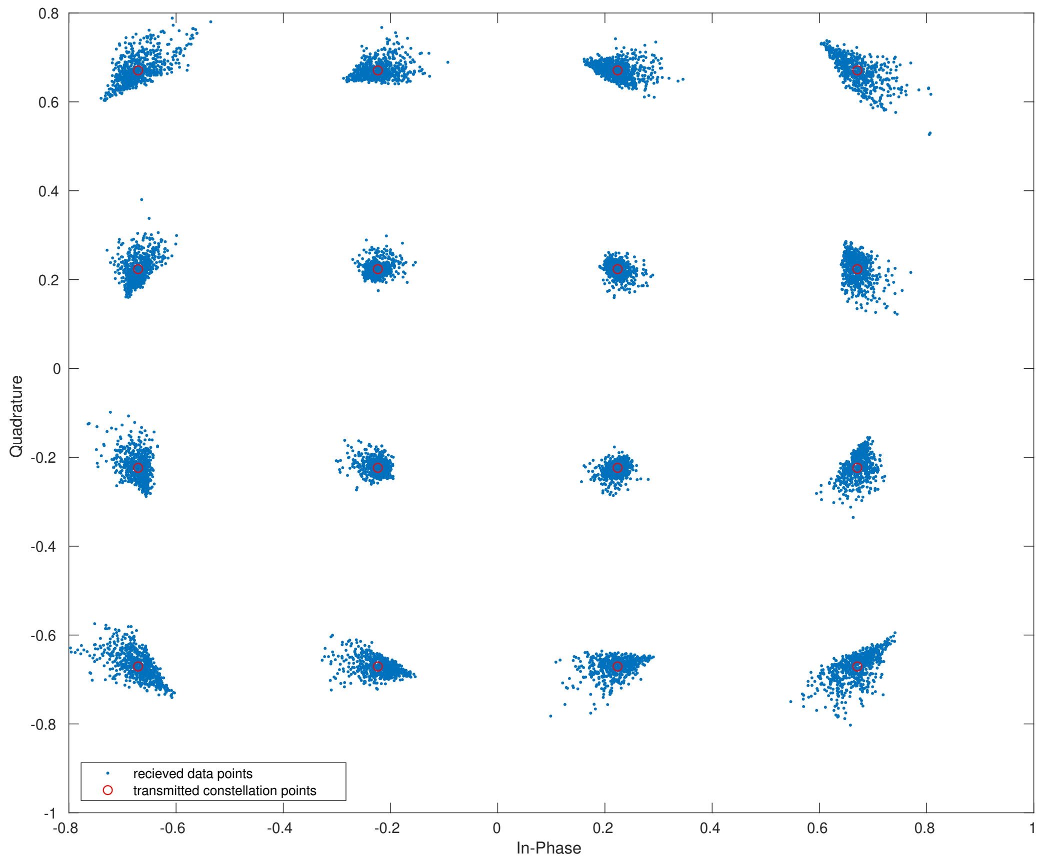


Fig. 2.1. SPM-induced Received points for a 10 Span fiber with random input data stream of size 10k and 16QAM constellation design after compensation of conditional mean , input data point is incremented with effective SPM and mean constellation energy= 0.5

2.2 Simplified C-model

Closest substitute for real data we have at hand for fiber SPM noise is the C matrix. If we slide the boundaries of the summation and redefine the C-matrix such that $C_{0,0}$ would be the element at the center of the matrix, the first equation in 2.3 can be rewritten as:

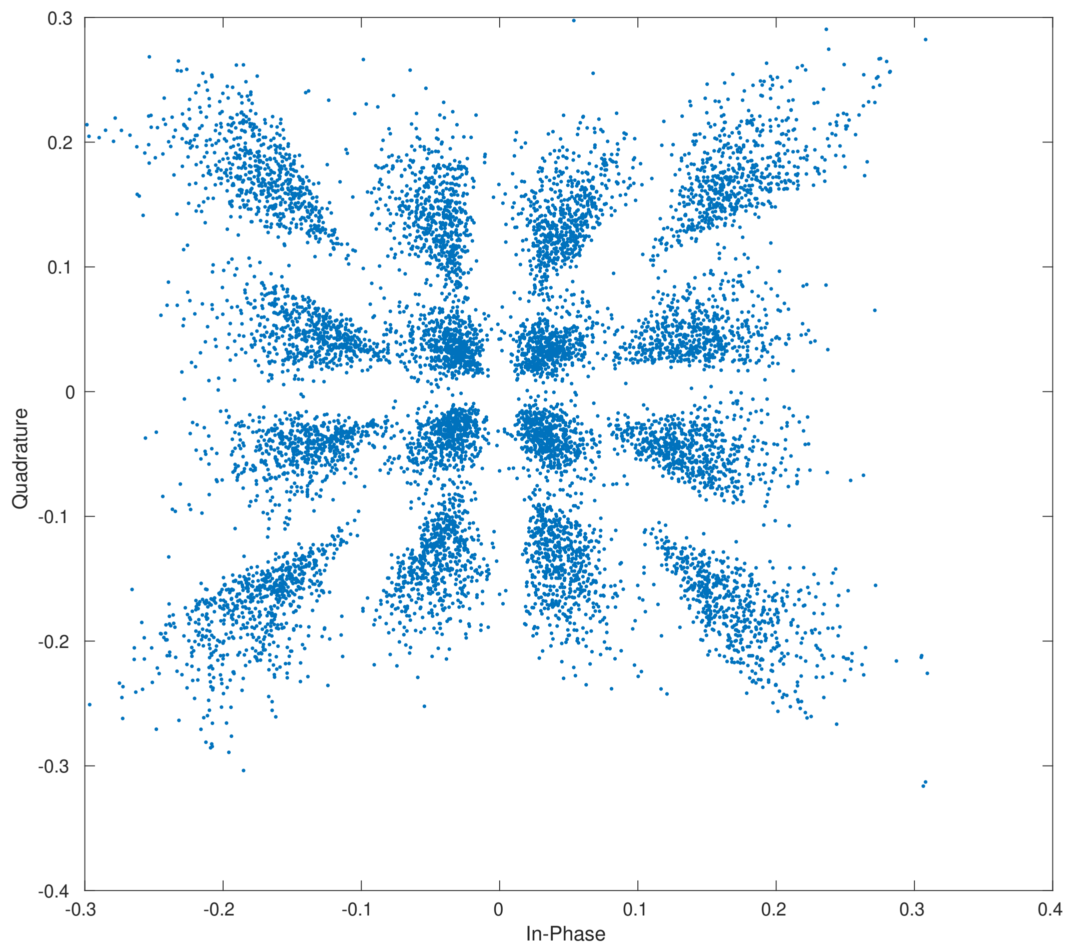


Fig. 2.2. raw output of SPM C-matrix for a 10 Span fiber with input data stream of size 10k and 16QAM constellation design

$$spm_i = \sum_{m=-M}^{M} \sum_{n=-M}^{M} C_{m,n} x_{i+m} x_{i+n} x^*_{i+m+n} \tag{2.4}$$

where spm_i is the noise added to a data sample at position i and the stream of data is $x_1, x_2, x_3, ...$ (x_i's are constellation points). $C_{m,n}$ is a C-matrix element, x_i^* is the complex conjugate of x_i and M depends on the size of the C-matrix and the accessible memory.

C-matrix can be very large in size (97*97 for 10 Span fiber). Therefore, it can add computational complexities to the problem. However, most of the dominant elements of the C matrix can be found near the center of it. This fact gives us an intuition to model the fiber with a matrix of smaller size. To have a sense of the dominance of the central elements of C-matrix, there is a comparison between 5×5 matrix at the center with the original C-matrix in figure 2.3.

These observations provide a good perspective of the simplified model that we can define. Since using a small matrix consisted of the central and most dominant elements of the C-matrix provides a considerably acceptable result without any model fitting, we propose that our model should have a similar structure. Due to symmetry and observations from the structure of the C-matrix, matrices of size $n \times n$ where n is odd can be used as the model. It is obvious that using a larger matrix will provide more precise results due to the increase in number of parameters.

Using $C = \begin{bmatrix} C_{-1,1} & C_{0,1} & C_{1,1} \\ C_{-1,0} & C_{0,0} & C_{1,0} \\ C_{-1,-1} & C_{0,-1} & C_{1,-1} \end{bmatrix}$ as a substitute for C-matrix and equation 2.4 we have:

$$\begin{aligned} spm_i &= \sum_{m=-1}^{1} \sum_{n=-1}^{1} C_{m,n} x_{i+m} x_{i+n} x^*_{i+m+n} \\ &= C_{0,0} x_i \left\| x_i \right\|^2 + C_{0,1} x_i \left\| x_{i+1} \right\|^2 + C_{1,0} x_i \left\| x_{i+1} \right\|^2 + C_{0,-1} x_i \left\| x_{i-1} \right\|^2 \\ &+ C_{-1,0} x_i \left\| x_{i-1} \right\|^2 + C_{-1,1} x_{i-1} x_{i+1} x^*_i + C_{-1,-1} x^2_{i-1} x^*_{i-2} + C_{1,1} x^2_{i+1} x^*_{i+2} + C_{1,-1} x_{i-1} x_{i+1} x^*_i \end{aligned} \tag{2.5}$$

From equation 2.5 we can observe:

- SPM noise added to a single data sample at position i of input data stream depends on matrix elements and neighboring data samples.
- The number of neighboring points for calculating SPM depends on the size of the matrix used to model C-matrix. In this case, a 3×3 matrix uses a window of size 5 centered at position i, which means two data samples before and two data samples after x_i are going to be needed to calculate spm_i.

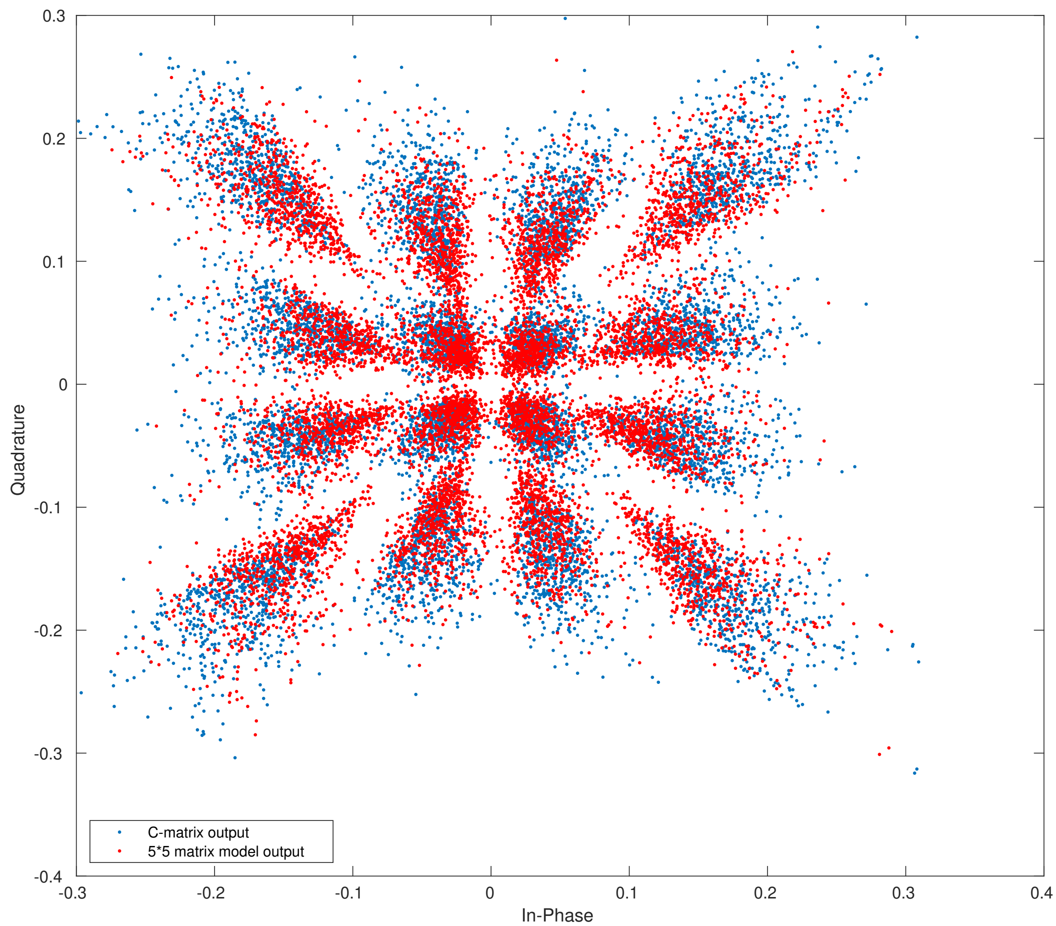


Fig. 2.3. Comparison between raw output of C-matrix and 5×5 model with no model fitting for 10 span fiber

- The summation consists of quadruple multiplications (triple multiplications of data points multiplied by a matrix element).
- If enough neighboring data points are known, spm_i only depends on matrix elements. In fact, it is a linear combination of $C_{i,j}$'s with known coefficients.

The linearity of the model with respect to model parameters is the basis that this research builds up on in the following chapters. We will go through some model fitting methods which exploits this property to obtain the best set of parameters to predict SPM noise.

Similar derivations to 2.5 can be obtained for a matrix of bigger size. Equations have more terms and use a bigger window of input data, but the linearity of the model is still preserved.

2.3 Summary

Solving Nonlinear Schrodinger Equation (NLSE) in fiber medium by Ciena Corporation has led to a model called the C-model. SPM and XPM noises are calculated with C-matrix which is considerably large and computationally expensive to work with. However, our observations help to define a simpler model with less number of parameters which is linear with respect to model parameters when a big enough window of data points are known.

This fact provides a new perspective to the problem which helps us to use different tools to train the model and obtain a considerably precise approximation of SPM noise.

Chapter 3

Model Fitting by Moore-Penrose Inverse

3.1 Moore-Penrose Inverse

A set of equations in which the number of equations is larger than the number of unknowns is called an over-determined set. From linear algebra, there are multiple methods one can approach an over-determined set[5]. Considering the linearity of the model discussed in previous chapter, a common method is Linear Least Squares (LLS) in which the objective function that should be minimized is the summation of all Euclidean norms of the distance between observations and model predictions [12].

Suppose there are L observations and M unknown parameters. If equations are linear with respect to unknown parameters, this optimization problem can be written in matrix form as follows.

$$\min_{x} \|A\mathbf{x} - \mathbf{y}\| \tag{3.1}$$

where A is a matrix of size $L \times M$, $\mathbf{x}$ is the vector of parameters of size M and $\mathbf{y}$ is the vector of observations of size L. Based on previous results from optimization and linear algebra, a solution to problem 3.1 which itself has minimum weight would be :

$$\mathbf{x} = (A^{\star}A)^{-1} A^{\star}\mathbf{y} \tag{3.2}$$

where A^* is the conjugate transpose of A and $A^+ = (A^\star A)^{-1} A^\star$ is Moore-Penrose Inverse of matrix A [25] [20][14]. This is a left inverse since $A^+A = I$. A proof of this theorem is provided in Appendix A.

3.2 Model Fitting and Error Analysis Results

As discussed in chapter 2, SPM noise added to a data point when neighboring points are known is a linear combination of model parameters. Therefore, we can use the Moore-Penrose Inverse to estimate the parameters of the model. For each observation, there is a prediction for SPM noise from equation 2.5. Our objective is to estimate the parameters $C_{i,j}$ such that square-norm of error would be minimized. From equation 3.1, A is the matrix of coefficients in a way that each row of it is the coefficients of the linear combination in 2.5 ($x_i \|x_i\|^2$, $x_i \|x_{i+1}\|^2$, ...) and the number of rows is equal to the number of observed data points. $\mathbf{x}$ is the parameter vector which for 3×3 model would be $[C_{-1,1}\ C_{0,1}\ C_{1,1} ...\ C_{1,-1}]^T$ and $\mathbf{y}$ is the vector of SPM noise observations.

The summation of all square-norms of the distances is the objective function which we want to minimize. This summation also represents the final error of the model. However, to have a better perspective of the result, we have to compare this summation with SPM values. So we calculate the ratio of this error and the variance of the raw SPM noise which is calculated from C-matrix and use it as our point of reference for error analysis.

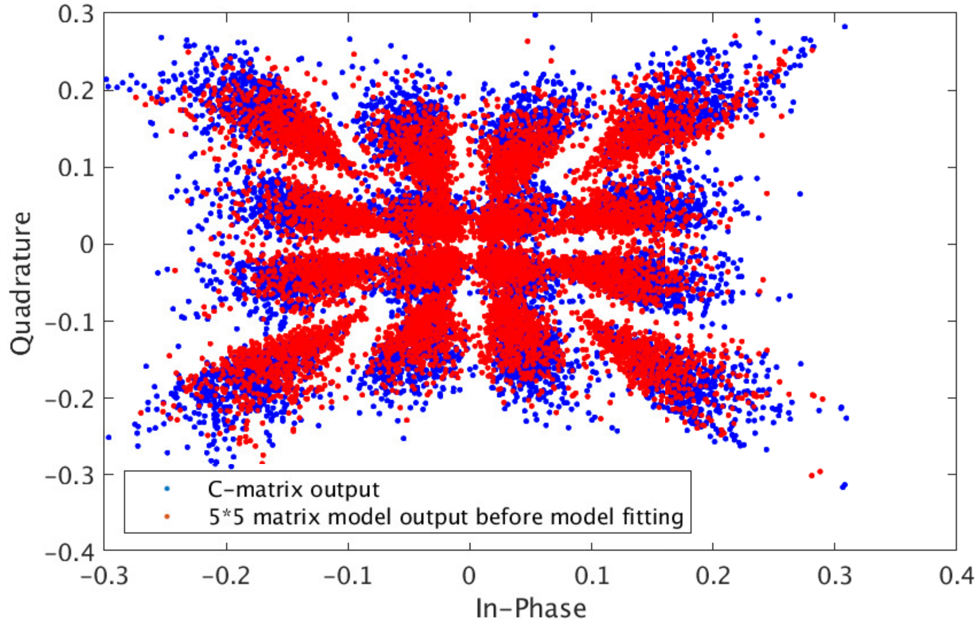
$$ratio = \frac{\|error\|^2}{\|SPM_{Cmatrix}\|^2} = \frac{\|distances\|^2}{\|SPM_{Cmatrix}\|^2} \tag{3.3}$$

In figure 3.1, we provide results of model fitting with Moore-Penrose Inverse method for 5×5 model. As it can be observed from the figures, after model fitting with Moore-Penrose method, the shape of the model prediction cloud is closer to C-matrix output and error ratio is decreased considerably.

An important note to mention about this method is that calculation of Pseudo-Inverse Matrix for a large number of observed points is computationally expensive. Hence, applying alternative methods which optimize the cost function in an iterative manner is the subject of next chapters of this thesis.

3.3 Summary

In this chapter, we used the linearity of our proposed model to obtain a well-defined optimization problem in matrix form. Results from linear algebra provides Moore-Penrose Inverse as a solution to the Linear Least Squares problem for this over-determined set of equations. Observing the results from our simulations suggests that this method performs efficiently good with mean-square error ratio as a performance measure.



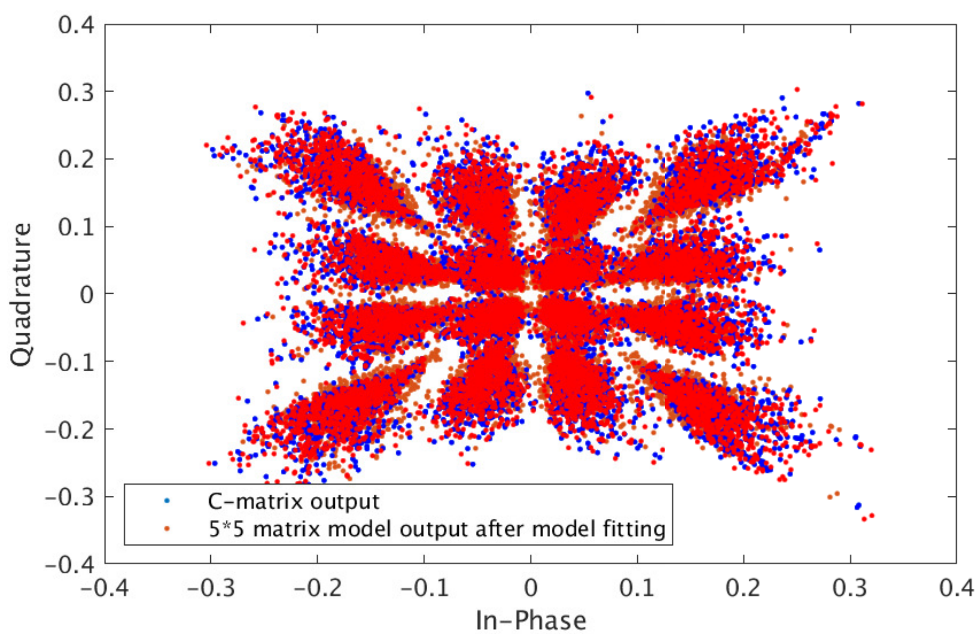


Fig. 3.1. Output of 5×5 model for input size=$10k$,
before optimization $error\ ratio = 0.0189$ and after optimization $error\ ratio = 0.0023$

Chapter 4

Model Fitting by Gradient Descent Method

4.1 Gradient Descent Optimization Method

Gradient descent is an optimization algorithm used to minimize an objective function by iteratively moving in the direction of local minimum as defined by the negative of the gradient. In optimization problems, we use gradient descent to update the parameters of our model. At each step, the update in parameters is dependent to learning rate and partial derivative of objective function.

If the multi-variable function $F(x)$ is defined and differentiable in a neighborhood of point x, the objective function F decreases fastest if we move towards the the negative of the gradient. Starting from $\mathbf{x}_0$ as a guess for local minimum:

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \lambda_n \nabla F\left(\mathbf{x}_n\right), n > 0 \tag{4.1}$$

with some assumptions on F, one can choose a suitable learning rate at each step that guarantees converging to a local minimum. If F is convex and the assumptions check out, $\mathbf{x}_n$ will converge to global minimum of the objective function. Due to the Barzilai-Borwein method[6] , a suitable learning rate could be:

$$\lambda_n = \frac{\left|\left(\mathbf{x}_n - \mathbf{x}_{n-1}\right)^T \left[\nabla F\left(\mathbf{x}_n\right) - \nabla F\left(\mathbf{x}_{n-1}\right)\right]\right|}{\left\|\nabla F\left(\mathbf{x}_n\right) - \nabla F\left(\mathbf{x}_{n-1}\right)\right\|^2}. \tag{4.2}$$

4.2 Model Fitting Algorithm and Results

There are some observations from the structure of original C-matrix that help us define a variation of Gradient Descent method for our problem. For example, $C_{0,1}$ and $C_{-1,0}$ are exactly equal and have zero real value. In an ideal scenario, we should allow all the imaginary and real parts of the parameters to take any value resulted from Gradient Descent algorithm, but due to our simulation results and the symmetry of the system derived by Ciena, we let these properties to be preserved in our model exactly like the original C-matrix.

To properly define an objective function, we need to separate the real and imaginary parts of the parameters and match the corresponding properties from C-matrix. For example, $C_{0,1}$ is purely imaginary in C-matrix, so we use $C_{0,1} = C_{imag}i$ as an element of our matrix model and C_{imag} as the parameter that we try to optimize.

To gain a better perspective, equation 2.5 is written considering the above mentioned assumption for $C_{0,1}$ in the following. For simplicity, all other terms of 2.5 including SPM noise observation are replaced with a complex value $a + bi$ and complex coefficient of $C_{0,1}$ is shown with $c + di$.

$$\begin{aligned} \|distance\|^2 &= \|(C_{imag}i)(c + di) + (a + bi)\|^2 \\ \|distance\|^2 &= \|(-dC_{image}) + (cC_{imag} + b)i\|^2 \\ \|distance\|^2 &= d^2 C_{image}^2 + (cC_{imag} + b)^2 \end{aligned} \tag{4.3}$$

Similar derivations can be observed for non-purely imaginary parameters as well. Equation 4.3 shows that for each observation square norm of distance is a quadratic function of real and imaginary parts of model parameters. Also, the final objective function F is the summation of all square distances, hence it's quadratic.

$$F = \sum_{observations} \|SPM_{model\ prediction} - SPM_{observed}\|^2 \tag{4.4}$$

This fact allows us to use the Gradient descent method and define the following iterative algorithm.

The results of applying algorithm 1 is provided in figure 4.1 .

4.3 Model Fitting with Presence of AWGN

Previous simulations have been conducted in absence of noise. However, there are multiple resources of noise in a real communication channel. Our simulations show that Algorithm

Algorithm 1 Modified Gradient Descent Algorithm

1: **Start** original C-matrix values as a guess $\mathbf{x}_0 = [C_{-1,1}\ C_{0,1}\ C_{1,1} \ldots\ C_{1,-1}]^T$
2: **Set** number of iteration **n**
3: **for** i=1: **n do**
4: Find ∇F w.r.t. separate real and imaginary parts
5: Find appropriate *learning rate* λ using 4.2
6: $\mathbf{x}_i \leftarrow \mathbf{x}_{i-1} - \lambda \nabla F\left(\mathbf{x}_{i-1}\right)$
7: **If** $F(\mathbf{x}_i) > F(\mathbf{x}_{i-1})$ **then** $\lambda \leftarrow \lambda/2$, **go to** line 5
8: **return** $\mathbf{x}_n$

1 is able to cancel out the effect of Gaussian noise to a considerable amount when input data point is noisy at different SNR values. A note of significance to mention is that all the model fittings are performed when input data stream is a training data set, which means the stream of data is considered to be known at the receiver. Hence, no decoding is needed at the receiver. Results of model error ratio values after some number of iterations at different SNRs are provided in figure 4.2. It can be observed that performance of the algorithm is improved at higher SNRs as expected.

4.4 Summary

A very common iterative method that can be used for minimization of convex quadratic functions is Gradient Descent. The basic idea behind this method is to start with a guess for the optimum value of the parameter and move gradually towards the local minimum by calculating the partial derivatives and a proper *learning rate*. Based on our observations from original C-matrix, we separate the real and imaginary parts of the parameters and match the corresponding similarities. After each iteration, the parameters are closer to the desired optimum.

A note of significance is checking that whether $F(\mathbf{x}_n)$ decreases with each iteration or not. If the new $\mathbf{x}_n$ does not move towards the optimum, we have to decrease the *learning rate* and calculate a new $\mathbf{x}_n$.

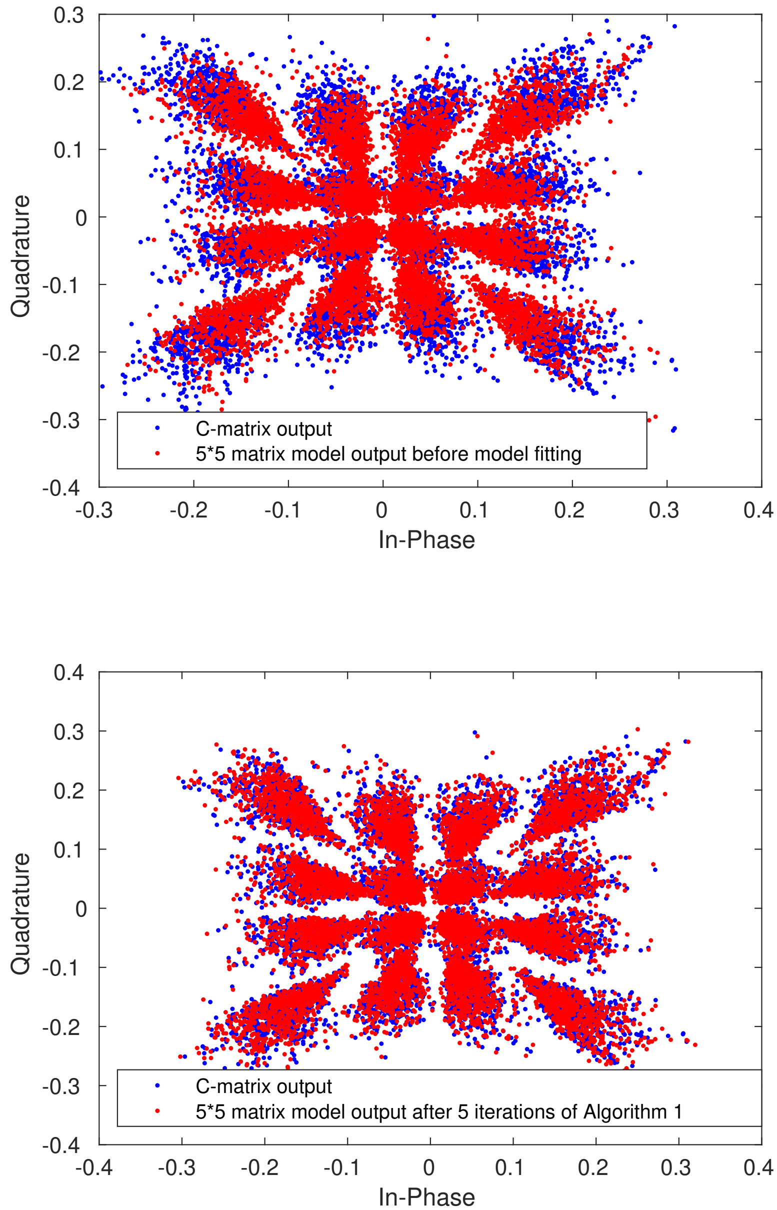


Fig. 4.1. Output of 5×5 model for input size=$10k$,
before optimization $error\ ratio = 0.0189$ and after optimization $error\ ratio = 0.0037$

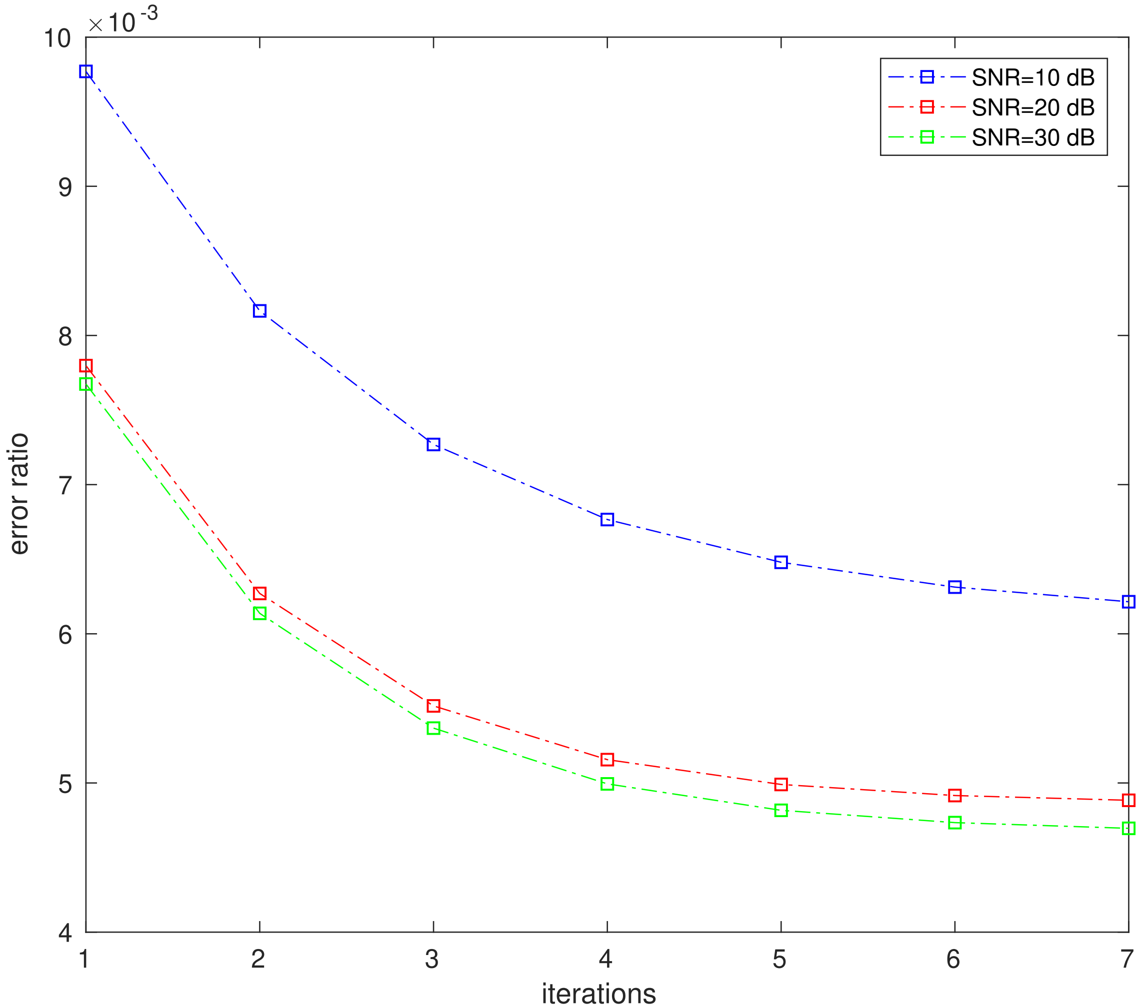


Fig. 4.2. error ratio of 5×5 model after some iterations of Algorithm 1 for input size=$10k$, at different SNRs

Chapter 5

Alternative Iterative Method Using Hermitian Inner Product

5.1 Algorithm Definition

Another approach to the problem is to represent the coefficients which are multiplied to a single parameter with a vector. Bearing in mind that all the combinations are linear, we rewrite the matrix form of the problem as follows:

$$\mathbf{A}\mathbf{x} = \sum_{i=1}^{n} \mathbf{A}_i x_i \ , \ \textit{observation vector } \mathbf{y} \tag{5.1}$$

where n is the number of parameters (for the case of 3×3 model $n = 9$), A_i is the vector of coefficients for parameter x_i.

Supposing the number of observations is L, $\mathbf{y}$ and each of $\mathbf{A}_i$'s are vectors in L-dimensional space (If we consider the complex nature of elements, each vector will have $2L$ dimensions). In this framework, each of the parameter x_i's are coefficients of L-dimensional vectors $\mathbf{A}_i$ and our goal is to find the coefficients x_i in a way which the summation is as close as possible to observation vector $\mathbf{y}$.

In set of real vector space, one of the methods which is used to find the projection of a vector on another vector is inner product. In complex vector space, Hermitian inner product helps us achieve the same property. Hermitian inner product of two complex vectors $\mathbf{v}$ and $\mathbf{w}$ is:

$$< \mathbf{v}, \mathbf{w} >= \sum_i v_i w_i^* \tag{5.2}$$

where w_i^* is the complex conjugate of w_i [33]. Applying Hermitian inner product, the projection of observation vector $\mathbf{y}$ on the coefficients vector $\mathbf{A}_i$ can be found. If we deduct the inner product result from original vector, the remnant would be orthogonal to projected vector[10]. Using the projection, we can derive an estimation for each of x_i's and update them in an iterative manner. Assuming we initiate with all x_i's equal to zero:

$$x_1^* = \frac{< \mathbf{y}, \mathbf{A}_1 >}{\|\mathbf{A}_1\|^2} \tag{5.3}$$

is the new estimation for x_1. To estimate the next parameter, we have to subtract the term that first parameter contributes to the summation and find the projection on $\mathbf{A}_2$:

$$x_2^* = \frac{< \mathbf{y} - x_1^* \mathbf{A}_1, \mathbf{A}_2 >}{\|\mathbf{A}_2\|^2} \tag{5.4}$$

One iteration is finished when all parameters are updated once. Algorithm 2 depicts this method when initial point for parameter vector is the original C-matrix parameters.

Algorithm 2 Iterative Hermitian Inner Product Algorithm

1: **Start** original C-matrix values as initial parameter $\mathbf{x} = [C_{-1,1}\ C_{0,1}\ C_{1,1} \ldots\ C_{1,-1}]^T$
2: **Set** number of iteration $\mathbf{n}$ and number of parameters $\mathbf{m}$
3: **for** j=1: $\mathbf{n}$ **do**
4:     **for** i=1: $\mathbf{m}$ **do**
5:         $x_i \leftarrow \dfrac{< \mathbf{y} - \sum_{k=1,k\neq i}^{\mathbf{m}} x_k \mathbf{A}_k, \mathbf{A}_i >}{\|\mathbf{A}_i\|^2}$
6: **return x**

5.2 Model Fitting Results

The results of model fitting using this method is provided in figures 5.1 and 5.2.

As it was expected, performance of the algorithm is enhanced at higher SNRs.

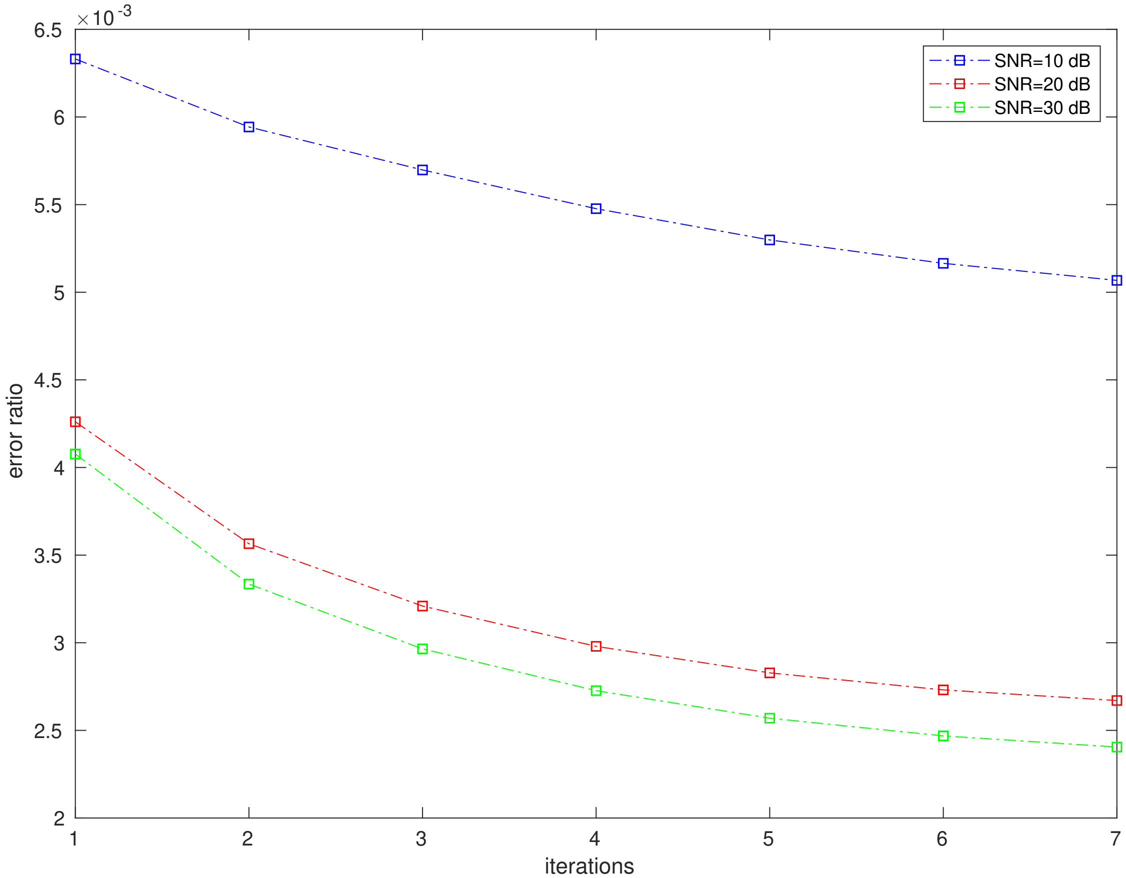


Fig. 5.1. error ratio of 5×5 model after some iterations of Algorithm 2 for input size=$10k$, at different SNRs

5.3 Summary

Hermitian inner product is a powerful tool for finding the projection of a complex vector on another. In this chapter, using this method, we defined an algorithm to iteratively update the parameters of the model with finding the projection of observation vector on the vector of coefficients for each parameter. According to the results, this algorithm is able to reduce the error ratio drastically after a few iterations. However, at higher number of iteration, error ratio decreases with slower rate.

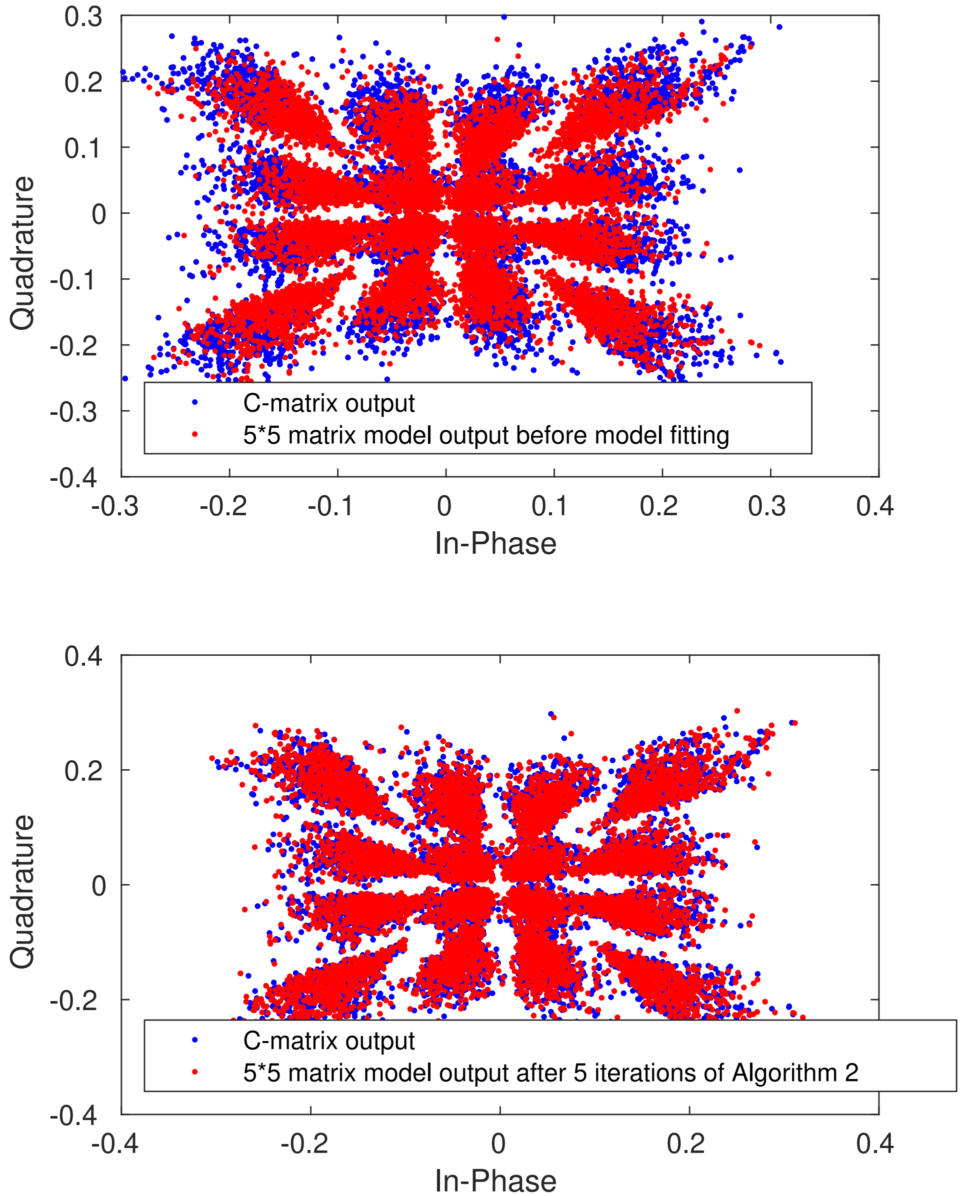


Fig. 5.2. Output of 5×5 model for input size=$10k$,
before optimization $error\ ratio = 0.0189$ and after optimization $error\ ratio = 0.0027$

Chapter 6

Online Tracking of Model Parameters

Numerous factors contribute to channel characteristics of a fiber-optic communication system. As channel characteristics change, nonlinear noises such as SPM and XPM change and this fact denotes the significance of time-domain variations of model parameters. The procedure in practical implementation is that when the characteristics of channel change, the system detects an increase in error vector norm which implies that the parameters are not predicting the SPM noise accurately enough. So the parameters need to be updated with the information that new observation points provide about channel characteristic perturbations. In this work, our closest approximation for real SPM is calculated by C-matrix. So in order to generate updated SPM noise, we change the elements of C-matrix and investigate that how efficiently our online tracking method works.

6.1 Online Tracking by Pseudo-Inverse

In chapter 3 , model fitting using Moore-Penrose pseudo-inverse method was discussed. Suppose model parameters are estimated after observing a batch of received data with length M. So optimum set of parameters are

$$\mathbf{x}_{opt} = \left(A^{\star}A\right)^{-1} A^{\star}\mathbf{y}$$

where objective function that needs to be minimized is

$$F_{\mathbf{x}} = \|A\mathbf{x} - \mathbf{y}\|.$$

Suppose a new batch of data with size N is received and we detect an error vector which has a norm greater than a threshold e_{th} due to different channel characteristics. In other words,

$$\begin{aligned} \mathbf{y}_{predicted} &= B\mathbf{x}_{opt} \\ \Delta\mathbf{y} = \mathbf{y}_{obs} &- \mathbf{y}_{predicted} \\ \|\Delta\mathbf{y}\| &> e_{th} \end{aligned}$$

where $B_{N\times|x|}$ is the coefficient matrix for new batch of data and $\mathbf{y}_{obs}$ is vector of new N number of observations. This format helps us to define a new objective function

$$G_{\Delta\mathbf{x}} = \|B\Delta\mathbf{x} - \Delta\mathbf{y}\|. \tag{6.1}$$

Using Pseudo-Inverse the solution that minimizes 6.3 is

$$\Delta\mathbf{x}_{opt} = (B^{\star}B)^{-1}B^{\star}\Delta\mathbf{y}.$$

So our method considers a perturbation vector $\Delta\mathbf{x}$ and optimizes it to minimize the error-norm of the new batch of data. New parameter vector will be the summation of previous optimum vector and new optimum perturbation vector:

$$\mathbf{x}_{opt}^{new} = \mathbf{x}_{opt} + \Delta\mathbf{x}_{opt}$$

Results for a simulation are provided in the following.

3×3 **trained model on input of size**$= 10k$ **by pseudo-inverse:**

$$C_{opt} = \begin{bmatrix} 0.0191 + 0.0122i & 0.0000 - 0.0569i & -0.0193 + 0.0121i \\ 0.0000 - 0.0569i & -0.0001 - 0.1679i & 0.0001 - 0.0571i \\ -0.0193 + 0.0121i & 0.0001 - 0.0571i & 0.0191 + 0.0117i \end{bmatrix} \quad \frac{\|error\|^2}{\|SPM\|^2} = 0.0140$$

To model channel variations, a Gaussian noise with power of 20 percent is added to C-matrix elements and a new batch of data with size$= 1k$ is transmitted. Perturbation vector ΔC_{opt} is trained on the new batch of data.

$$\Delta C_{opt} = \begin{bmatrix} -0.0038 - 0.0011i & 0.0005 - 0.0053i & -0.0007 - 0.0029i \\ 0.0005 - 0.0053i & -0.0024 + 0.0035i & 0.0011 + 0.0022i \\ -0.0007 - 0.0029i & 0.0011 + 0.0022i & -0.0005 - 0.0018i \end{bmatrix}$$

Finally, new model matrix is calculated as the summation of previous optimum and optimum perturbation vector.

$$C_{opt}^{new} = C_{opt} + \Delta C_{opt}, \quad \frac{\|error\ C_{opt}^{new}\|^2}{\|SPM\|^2} = 0.0141$$

As we can see the new optimum model parameter set provides an accurate estimation of SPM.

6.2 Online Tracking by Stochastic Gradient Descent

In chapter 4, we used Gradient Descent to estimate model parameters. In this section, we introduce a slightly different method named Stochastic Gradient Descent and showcase how it can be useful for online tracking. In Gradient Descent, vector $\mathbf{x}$ moves towards the negative of partial derivative of objective function F which is the summation of all error functions for each observation. Stochastic Gradient Descent suggests that we can use the gradient of only **one** of these error functions to update the estimation.

In our case, our goal is to update parameters with respect to new batch of observed data. Therefore, objective function F is the summation of error functions of new N number of observations.

$$F = \sum_{i=1}^{N} F_i$$

$$F_i = \| \sum_{j=1}^{K} B_{i,j} x_j - y_i \|$$

where B is the coefficient matrix of new batch. Initial value of the parameters is the previous optimum that was trained before the new batch is received. So after each iteration:

$$\mathbf{x}_{opt}^{new} = \mathbf{x}_{opt} - \lambda \nabla F_i \left(\mathbf{x}_{opt} \right). \tag{6.2}$$

Learning rate λ can be determined with 4.2.

To implement this method, we initiate with the same optimum in previous section and optimize the perturbation vector by using stochastic gradient descent.

After 5 iterations of stochastic gradient descent:

$$\Delta C_{opt} = \begin{bmatrix} -0.0037 - 0.0009i & 0.0011 + 0.0523i & 0.0166 - 0.0090i \\ -0.0006 - 0.0398i & -0.0020 + 0.0101i & 0.0015 + 0.0173i \\ -0.0178 + 0.0035i & 0.0005 - 0.0371i & -0.0007 - 0.0019i \end{bmatrix}$$

The summation of previous optimum and new perturbation vector is:

$$C_{opt}^{new} = C_{opt} + \Delta C_{opt}, \quad \frac{\|error\ C_{opt}^{new}\|^2}{\|SPM\|^2} = 0.0171$$

Due to the iterative manner of this method and considering that pseudo-inverse provides the most accurate solution for this problem, it is expected that error ratio of stochastic gradient descent method would be greater than pseudo-inverse method.

6.3 Online Tracking by Hermitian Inner Product

As an alternative method, we use Hermitian inner product to find the projection of error vector on the coefficient vector of each parameter. As equation 6.3 states, the matrix format of the minimization problem is:

$$G_{\Delta \mathbf{x}} = \|B\Delta \mathbf{x} - \Delta \mathbf{y}\|. \tag{6.3}$$

We initiate with $\Delta \mathbf{x} = 0$. With finding the projection of Δy on the first coefficient vector, we calculate a new value for the perturbation of the first parameter:

$$\Delta x_1^* = \frac{< \Delta \mathbf{y}, B_1 >}{\|B_1\|^2} \tag{6.4}$$

To update the second parameter perturbation, we deduct the effect of multiplications of first parameter and it's coefficients. Meaning:

$$\Delta x_2^* = \frac{< \Delta \mathbf{y} - \Delta x_1^* B_1, B_2 >}{\|B_2\|^2} \tag{6.5}$$

After one iteration all the parameters are updated. We repeat the same procedure until the decrease in error vector is less than a predetermined threshold value which means iteration is not significantly helpful in order to approach the minimum. Using an algorithm similar to Algorithm 2 we obtain the optimum Δx and

$$\mathbf{x}_{opt}^{new} = \mathbf{x}_{opt} + \Delta \mathbf{x}_{opt}$$

is the new optimum.

Starting with the same optimum as previous sections, the result of training to obtain the optimum perturbation vector by using Hermitian inner product is provided in the following.

After 5 iterations of Hermitian inner product:

$$\Delta C_{opt} = \begin{bmatrix} -0.0037 - 0.0008i & 0.0014 + 0.0811i & 0.0253 - 0.0121i \\ -0.0011 - 0.0571i & -0.0017 + 0.0135i & 0.0017 + 0.0249i \\ -0.0264 + 0.0067i & 0.0002 - 0.0568i & -0.0008 - 0.0019i \end{bmatrix}$$

$$C_{opt}^{new} = C_{opt} + \Delta C_{opt}, \quad \frac{\|error\ C_{opt}^{new}\|^2}{\|SPM\|^2} = 0.0164$$

The performance is not as accurate as the pseudo-inverse method which is expected since pseudo-inverse method provides the global optimum and any iterative method ideally can converge to that optimum when the number of iterations increases.

6.4 Summary

Since channel characteristics can change over time, tracking the model parameters in time domain is of great significance in a communication system. In this chapter, we tried to propose methods based on our previous model fitting algorithms to track these perturbations. Channel characteristic variation was modeled in original C-matrix and it was shown that our model can update the parameters after a new batch of data which is passed through the new channel arrives at the receiver side.

Chapter 7

Alternative Performance Measures

7.1 Sensitivity to Randomness of Input Data Stream

Since the data stream that we use as training test is generated randomly, there are no specific statistical characteristic associated with different data points and their correlation. To put this notion to test, our model-trained on a set of random test data- should have acceptable error when it predicts SPM for any other random set of data since there are no assumptions about the structure of the training data-set. Results from simulations that support this claim are provided in the following.

Random Training Data-set 1: size=$10k$

3×3 model trained using Pseudo-inverse:

$$C_1 = \begin{bmatrix} 0.0191 + 0.0122i & 0.0000 - 0.0569i & -0.0193 + 0.0121i \\ 0.0000 - 0.0569i & -0.0001 - 0.1679i & 0.0001 - 0.0571i \\ -0.0193 + 0.0121i & 0.0001 - 0.0571i & 0.0191 + 0.0117i \end{bmatrix} \quad \frac{\|error\|^2}{\|SPM\|^2} = 0.0140$$

Random Training Data-set 2: size=$10k$

3×3 model trained using Pseudo-inverse:

$$C_2 = \begin{bmatrix} 0.0197 + 0.0118i & -0.0001 - 0.0572i & -0.0197 + 0.0119i \\ -0.0001 - 0.0572i & 0.0000 - 0.1678i & 0.0001 - 0.0575i \\ -0.0197 + 0.0119i & 0.0001 - 0.0575i & 0.0197 + 0.0124i \end{bmatrix} \quad \frac{\|error\|^2}{\|SPM\|^2} = 0.0141$$

In order to test our claim, we use C_1 to predict the SPM of data-set 2 and C_2 to predict the SPM of data-set 1.

Predicted SPM of Data-set 2 using C_1

$$\frac{\|error\|^2}{\|SPM\|^2} = 0.0147$$

Predicted SPM of Data-set 1 using C_2

$$\frac{\|error\|^2}{\|SPM\|^2} = 0.0143$$

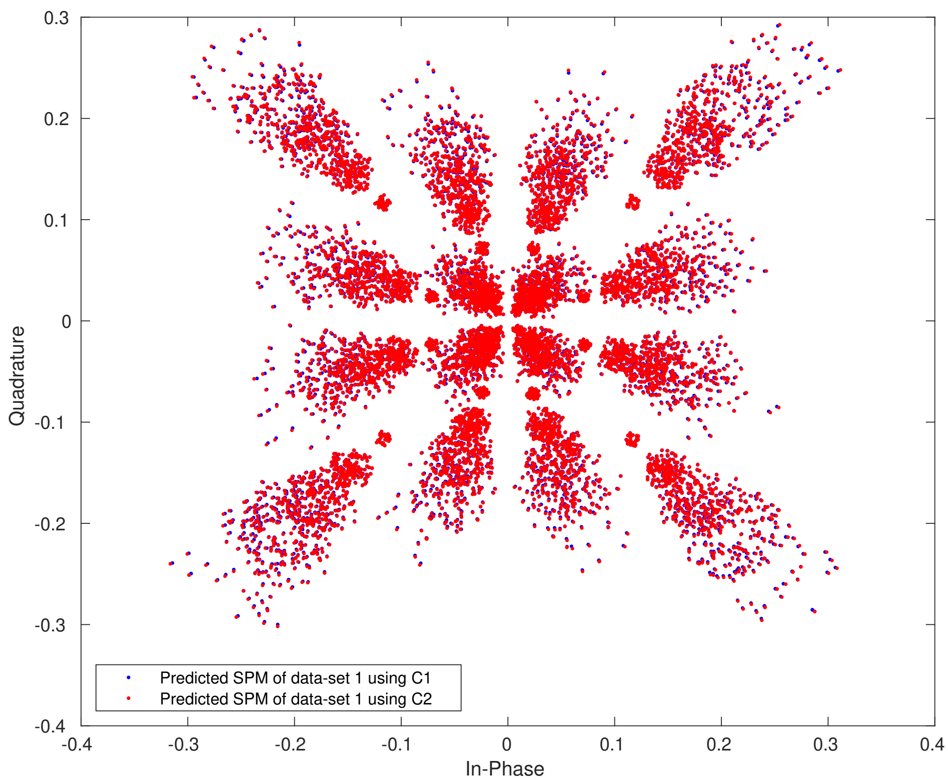


Fig. 7.1. Comparison between SPM cloud of Data-set 1 predicted by C_1 and C_2

7.2 Dominance of Most Significant Parameters

In previous chapters we discussed that most dominant elements of C-matrix are located around the center of the large C-matrix. Whether this is correct or not, after the optimization is done, is going to be discussed in this section. The procedure is that first we train a 5×5 model on a data-set which results in $M = 25$ optimum parameters. Then we choose a subset of size $K = 9$ of these parameters located at the center of 5×5 matrix. On the other hand, we can have a 3×3 model where these parameters can be trained separately. A comparison between the performance of these two cases can provide a good sense of how significant the parameters located near the center of the matrix are.

5×5 **model** $M = 25$ **and** $K = 9$ **on Data-set 1:**

$$C_M =$$
$$\begin{bmatrix} 0.0009 - 0.0016i & -0.0011 + 0.0080i & -0.0002 - 0.0206i & 0.0010 + 0.0081i & -0.0010 - 0.0017i \\ -0.0011 + 0.0080i & 0.0195 + 0.0120i & 0.0002 - 0.0488i & -0.0195 + 0.0122i & 0.0010 + 0.0081i \\ -0.0002 - 0.0206i & 0.0002 - 0.0488i & -0.0002 - 0.1342i & 0.0001 - 0.0488i & 0.0000 - 0.0207i \\ 0.0010 + 0.0081i & -0.0195 + 0.0122i & 0.0001 - 0.0488i & 0.0195 + 0.0120i & -0.0011 + 0.0082i \\ -0.0010 - 0.0017i & 0.0010 + 0.0081i & 0.0000 - 0.0207i & -0.0011 + 0.0082i & 0.0010 - 0.0016i \end{bmatrix}$$

$$\frac{\|error\|^2}{\|SPM\|^2} = 0.0023 \tag{7.1}$$

$$C_K \subset C_M, \quad C_K = \begin{bmatrix} 0.0195 + 0.0120i & 0.0002 - 0.0488i & -0.0195 + 0.0122i \\ 0.0002 - 0.0488i & -0.0002 - 0.1342i & 0.0001 - 0.0488i \\ -0.0195 + 0.0122i & 0.0001 - 0.0488i & 0.0195 + 0.0120i \end{bmatrix}$$

using C_K as the new model we will obtain:

$$\frac{\|error\|^2}{\|SPM\|^2} = 0.0426 \tag{7.2}$$

From previous section, we observed that a 3×3 model which is trained on data-set 1 by Pseudo-inverse method has error ratio $\frac{\|error\|^2}{\|SPM\|^2} = 0.0140$ while using the most dominant parameters of 5×5 model yields error ratio $\frac{\|error\|^2}{\|SPM\|^2} = 0426$. Considering this simulation, even though parameters which are close to the center of the matrix are the most dominant ones, optimizing the model with a larger set of parameters does not provide the dominant parameters with sufficient accuracy to predict the SPM noise.

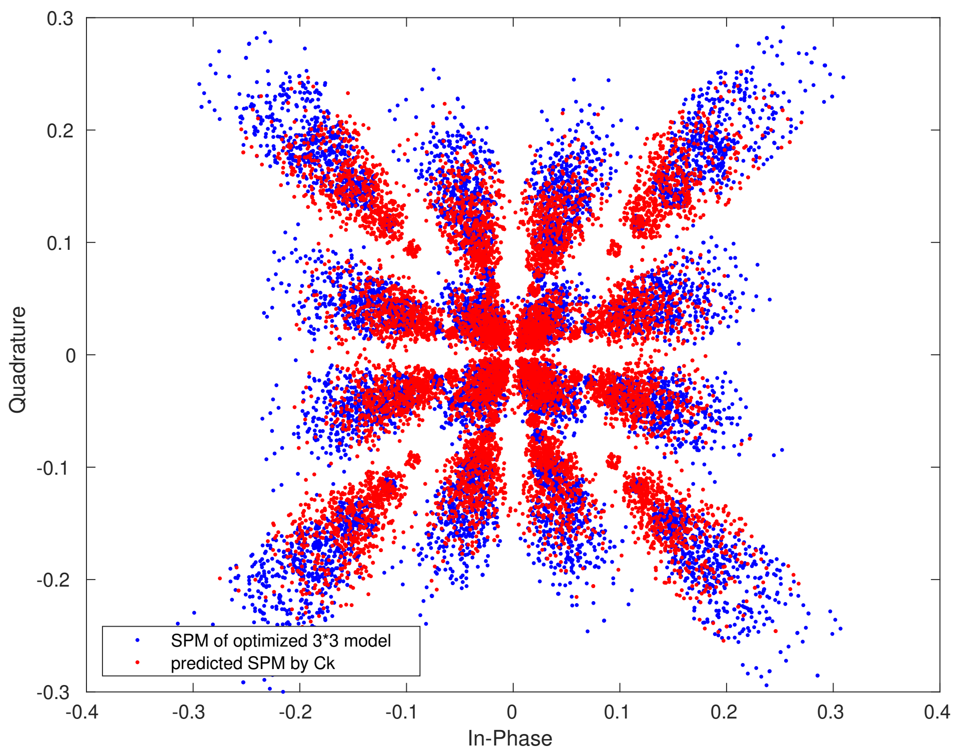


Fig. 7.2. Comparison between SPM cloud of C_k and optimized 3×3 model

References

[1] Govind P Agrawal. *Lightwave technology: telecommunication systems*. John Wiley & Sons, 2005.

[2] Govind P Agrawal. *Nonlinear Fiber Optics*. Elsevier, 5th edition, 2007.

[3] Govind P Agrawal and N Anders Olsson. Self-phase modulation and spectral broadening of optical pulses in semiconductor laser amplifiers. *IEEE Journal of quantum electronics*, 25(11):2297–2306, 1989.

[4] Abdelkerim Amari, Philippe Ciblat, and Yves Jaouën. Inter-subcarrier nonlinear interference canceler for long-haul nyquist-wdm transmission. *IEEE Photonics Technology Letters*, 28(23):2760–2763, 2016.

[5] Howard Anton and Chris Rorres. *Elementary Linear Algebra, Binder Ready Version: Applications Version*. John Wiley & Sons, 2013.

[6] Jonathan Barzilai and Jonathan M Borwein. Two-point step size gradient methods. *IMA journal of numerical analysis*, 8(1):141–148, 1988.

[7] H. Ebrahimzad. Nonlinear variance formula. Technical report, Ciena Corporation, 2017.

[8] R-J Essiambre and Peter J Winzer. Fibre nonlinearities in electronically pre-distorted transmission. In *2005 31st European Conference on Optical Communication, ECOC 2005*, volume 2, pages 191–192. IET, 2005.

[9] David D Falconer. Adaptive equalization of channel nonlinearities in qam data transmission systems. *The Bell System Technical Journal*, 57(7):2589–2611, 1978.

[10] Daniel T Finkbeiner. *Introduction to matrices and linear transformations*. Courier Corporation, 2013.

[11] Ying Gao, John C Cartledge, Abdullah S Karar, Scott S-H Yam, Maurice O?Sullivan, Charles Laperle, Andrzej Borowiec, and Kim Roberts. Reducing the complexity of perturbation based nonlinearity pre-compensation using symmetric edc and pulse shaping. *Optics express*, 22(2):1209–1219, 2014.

[12] James E Gentle. *Numerical linear algebra for applications in statistics*. Springer Science & Business Media, 2012.

[13] Ezra Ip and Joseph M Kahn. Compensation of dispersion and nonlinear impairments using digital backpropagation. *Journal of Lightwave Technology*, 26(20):3416–3425, 2008.

[14] M James. The generalised inverse. *The Mathematical Gazette*, pages 109–114, 1978.

[15] SL Jansen, D Van den Borne, B Spinnler, S Calabro, H Suche, PM Krummrich, W Sohler, G-D Khoe, and H De Waardt. Optical phase conjugation for ultra long-haul phase-shift-keyed transmission. *Journal of Lightwave Technology*, 24(1):54.

[16] Nobuhiko Kikuchi and Shinya Sasaki. Analytical evaluation technique of self-phase-modulation effect on the performance of cascaded optical amplifier systems. *Journal of Lightwave Technology*, 13(5):868–878, 1995.

[17] Xiaojun Liang and Shiva Kumar. Multi-stage perturbation theory for compensating intra-channel nonlinear impairments in fiber-optic links. *Optics express*, 22(24):29733–29745, 2014.

[18] Xiang Liu, S Chandrasekhar, and PJ Winzer. Phase-conjugated twin waves and fiber nonlinearity compensation. In *2014 OptoElectronics and Communication Conference and Australian Conference on Optical Fibre Technology*, pages 938–940. IEEE, 2014.

[19] Xiang Liu, AR Chraplyvy, PJ Winzer, RW Tkach, and S Chandrasekhar. Phase-conjugated twin waves for communication beyond the kerr nonlinearity limit. *Nature Photonics*, 7(7):560, 2013.

[20] E. H. Moore. On the reciprocal of the general algebraic matrix. In *Bulletin of the American Mathematical Society*. Cambridge University Press.

[21] Hadi Nekoei Qachkanloo. Adaptive compensation of nonlinear impairments in fiber-optic systems. Master's thesis, University of Waterloo, 2019.

[22] Shahab Oveis Gharan. Nonlinearity modeling and compensation. Technical report, 2016.

[23] Jie Pan and Chi-Hao Cheng. Nonlinear electrical predistortion and equalization for the coherent optical communication system. *Journal of Lightwave Technology*, 29(18):2785–2789, 2011.

[24] Kumar V Peddanarappagari and Maite Brandt-Pearce. Volterra series transfer function of single-mode fibers. *Journal of lightwave technology*, 15(12):2232–2241, 1997.

[25] Roger Penrose. A generalized inverse for matrices. In *Mathematical proceedings of the Cambridge philosophical society*, volume 51, pages 406–413. Cambridge University Press, 1955.

[26] Danish Rafique, Marco Mussolin, Marco Forzati, Jonas Mårtensson, Mohsan N Chugtai, and Andrew D Ellis. Compensation of intra-channel nonlinear fibre impairments using simplified digital back-propagation algorithm. *Optics express*, 19(10):9453–9460, 2011.

[27] Michael Reimer. Propagation nonlinearities: Nonlinear manakov equations. Technical report, 2016.

[28] Jacklyn D Reis and António L Teixeira. Unveiling nonlinear effects in dense coherent optical wdm systems with volterra series. *Optics express*, 18(8):8660–8670, 2010.

[29] Ali Saheb Pasand. Methods for nonlinear impairments compensation in fiber-optic communication systems. Master's thesis, University of Waterloo, 2018.

[30] Martin Schetzen. Theory of pth-order inverses of nonlinear systems. *IEEE Transactions on Circuits and Systems*, 23(5):285–291, 1976.

[31] Zhenning Tao, Liang Dou, Weizhen Yan, Yangyang Fan, Lei Li, Shoichiro Oda, Yuichi Akiyama, Hisao Nakashima, Takeshi Hoshida, and Jens C Rasmussen. Complexity-reduced digital nonlinear compensation for coherent optical system. In *Next-Generation Optical Communication: Components, Sub-Systems, and Systems II*, volume 8647, page 86470K. International Society for Optics and Photonics, 2013.

[32] Zhenning Tao, Liang Dou, Weizhen Yan, Lei Li, Takeshi Hoshida, and Jens C Rasmussen. Multiplier-free intrachannel nonlinearity compensating algorithm operating at symbol rate. *Journal of Lightwave Technology*, 29(17):2570–2576, 2011.

[33] Nicholas Young. *An introduction to Hilbert space*. Cambridge university press, 1988.

[34] Yukui Yu, Wei Wang, Paul D Townsend, and Jian Zhao. Modified phase-conjugate twin wave schemes for spectral efficiency enhancement. In *2015 European Conference on Optical Communication (ECOC)*, pages 1–3. IEEE, 2015.

[35] Qunbi Zhuge, Michael Reimer, Andrzej Borowiec, Maurice O'Sullivan, and David V Plant. Aggressive quantization on perturbation coefficients for nonlinear pre-distortion. IEEE, 2014.

APPENDICES

Appendix A

Proof of Moore-Penrose Pseudo-inverse Theorem

Theorem: Every linear system $Ax = b$, where A is an $m \times n$-matrix, has a unique least-squares solution x^+ of smallest norm.

Proof: Geometry offers a nice proof of the existence and uniqueness of x^+. Indeed, we can interpret b as a point in the Euclidean (affine) space $\mathbb{R}^m$, and the image subspace of A (also called the column space of A) as a subspace U of $\mathbb{R}^m$ (passing through the origin). Then, we claim that x minimizes $\|Ax - b\|^2$ iff Ax is the orthogonal projection p of b onto the subspace U, which is equivalent to $\mathbf{pb} = b - Ax$ being orthogonal to U .

First of all, if $U^\perp$ is the vector space orthogonal to U , the affine space $b + U^\perp$ intersects U in a unique point p.

Next, for any point $y \in U$, the vectors $\mathbf{py}$ and $\mathbf{bp}$ are orthogonal, which implies that

$$\|\mathbf{by}\|^2 = \|\mathbf{bp}\|^2 + \|\mathbf{py}\|^2. \tag{A.1}$$

Thus, p is indeed the unique point in U that minimizes the distance from b to any point in U. To show that there is a unique x^+ of minimum norm minimizing $\|Ax - b\|^2$, we use the fact that

$$\mathbb{R}^n = \operatorname{Ker} A \oplus (\operatorname{Ker} A)^\perp. \tag{A.2}$$

Indeed, every $x \in \mathbb{R}^n$ can be written uniquely as $x = u + v$, where $u \in \operatorname{Ker} A$ and $v \in (\operatorname{Ker} A)^\perp$, and since u and v are orthogonal,

$$\|x\|^2 = \|u\|^2 + \|v\|^2 \tag{A.3}$$

Furthermore, since $u \in \operatorname{Ker} A$, we have $Au = 0$, and thus $Ax = p$ iff $Av = p$, which shows that the solutions of $Ax = p$ for which x has minimum norm must belong to $\operatorname{Ker} A^{\perp}$.

However, the restriction of A to $\operatorname{Ker} A^{\perp}$ is injective. This shows that there is a unique x of minimum norm minimizing $\|Ax - b\|^2$, and that it must belong to $\operatorname{Ker} A^{\perp}$. $\square$

The proof also shows that x minimizes $\|Ax - b\|^2$ iff $\mathbf{pb} = b - Ax$ is orthogonal to U , which can be expressed by saying that $b - Ax$ is orthogonal to every column of A. However, this is equivalent to

$$A^{\top}(b - Ax) = 0, \quad \text{i.e.} \quad A^{\top}Ax = A^{\top}b. \tag{A.4}$$

Finally, it turns out that the minimum norm least squares solution x^+ can be found in terms of the pseudo-inverse A^+ of A, which is itself obtained from the SVD of A.

If $A = VDU^{\top}$, with

$$D = \operatorname{diag}(\lambda_1, \ldots, \lambda_r, 0, \ldots, 0), \tag{A.5}$$

where D is an $m \times n$ matrix and $\lambda_i > 0$, letting

$$D^+ = \operatorname{diag}(1/\lambda_1, \ldots, 1/\lambda_r, 0, \ldots, 0), \tag{A.6}$$

an $n \times m$ matrix, the pseudo-inverse of A is defined as

$$A^+ = UD^+V^{\top}. \tag{A.7}$$

assume that A is a (rectangular) diagonal matrix D, as above. Then, since x minimizes $\|Dx - b\|^2$ iff Dx is the projection of b onto the image subspace F of D, it is fairly obvious that $x^+ = D + b$.Otherwise, we can write

$$A = VDU^{\top}, \tag{A.8}$$

where U and V are orthogonal. However, since V is an isometry,

$$\|Ax - b\| = \left\|VDU^{\top}x - b\right\| = \left\|DU^{\top}x - V^{\top}b\right\|. \tag{A.9}$$

Letting $y = U^{\perp}x$, we have $\|x\| = \|y\|$ since U is an isometry, and since U is surjective, $\|Ax - b\|$ is minimized iff $\left\|Dy - V^{\top}b\right\|$ is minimized, and we showed that the least solution is

$$y^{+} = D^{+}V^{\top}b \tag{A.10}$$

Since $y = U^{\perp}x$, with $\|x\| = \|y\|$, we get

$$x^{+} = UD^{+}V^{\top}b = A^{+}b. \tag{A.11}$$

Thus, the pseudo-inverse provides the optimal solution to the least-squares problem.$\square$