7. Yıl İzni Çalışmaları Sonuç Raporu

Prof. Dr. Çağatay Candan

ODTU – EEMB

21 Eylül 2020

Özet

Bu rapor 09.09.2019 – 09.09.2020 tarihleri arasında 7. yıl iznim süresinde Boğaziçi Üniversitesi Elektrik ve Elektronik Mühendisliği bölümünde gerçekleştirdiğim çalışmalar hakkında bilgi vermektedir. İzin süresince yapılan çalışmalar sonucu oluşan yayınlar şu şekilde sıralanabilir: ODTU - Q1 listesinde yer alan bir dergide Ağustos 2020 tarihinde basılmış olan bir yayın; Ekim 2020'de çevrimiçi olarak gerçekleşecek olan ulusal bir konferansta sunulacak olan bir bildiri; halen değerlendirme aşamasında olan bir makale ve hazırlığı tamamlanmış, yakın zamanda değerlendirme için dergiye gönderilecek olan bir makale. Raporda çalışmaların içeriği kısaca anlatılmakta ve ilgili yayınlar rapor eklerinde sunulmaktadır. Ayrıca izin süresince doktora öğrencilerimle uzaktan devam ettirdiğim çalışmalar hakkında da bilgi verilmektedir.

Çalışma 1: Chebyshev Center Computation on Probability Simplex with α-divergence Measure, IEEE Signal Processing, vol. 27, p. 1515-1519, 2020.

Bu çalışma kapsamında olasılık kütle fonksiyonların birleştirilmesi için α-ıraksaklık metriği ile verilen olasılık kütle fonksiyonların Chebyshev merkezinin hesaplanması anlatılmaktadır. Bazı uygulamalarda yerel kestirim sonuçların birleştirilmesi için Bayesci yaklaşımı gerçekleştirmek yerel kestirimlere ait bileşik olasılık kütle fonksiyonu bilinmediği için mümkün olmamaktadır. Örneğin radar sistemleri özelinde, D ve K şehirlerinde yer alan radarların kapsama alanları örtüştüğü için aynı hedefi tespit ve takip edebilmektedirler. Bu iki radarın ürettiği kestirim bilgisinin birleştirilmesi pratik bir ihtiyaçtır. Bayes kuralı uygulanarak yapılacak bir birleştirme işlemi için hedef ve iki kestirime ait bileşik yoğunluk fonksiyonuna ihtiyaç vardır. Bu uygulamada bileşik yoğunluk fonksiyonu bilinmediği için Bayesci birleştirme yapılamamaktadır. Bu tür uygulamalarda eldeki marjinal olasılık yoğunluk fonksiyonlarına (verilen örnekte D ve K radar istasyonlarının ürettiği kestirime ait yoğunluk fonksiyonları) min-max yaklaşımı ile eldeki fonksiyonlara en yakın olan olasılık yoğunluk fonksiyonu bulunabilir. Buradaki min-max yaklaşımı ifade edilen, verilen olasılık düzlemindeki (simplex) nokta kümesine olan uzaklıkların en büyük değerini küçültmektir. Bu çalışmada olasılık düzlemindeki mesafeler $\alpha-$ ıraksaklık ölçütü ($\alpha-$ divergence measure) ile belirlenmektedir. Daha fazla bilgi için Ek-1A'ya bakabilirsiniz.

Bu çalışmanın daha basit bir hali olan N-boyutlu düzlem üzerinde verilen bir nokta kümesinin, Öklid mesafe metriği altında Chebyshev merkezinin hesaplanması problemi Boğaziçi Universitesi öğretim üyelerinden Prof. Bülent Sankur ile ortak danışmanlığı yaptığımız bir öğrenciye (Ömer Avcı) lisans tezi dersi kapsamında verilmiştir. Ek-1B'de Ömer Avcı'nın final raporu bilgi olarak sunulmaktadır. Ömer Avcı'nın çalışmasının da ileride makaleye dönüşmesi mümkün gözükmektedir.

Çalışma 2: Patlamalı-Kesikli Gözlemler için Parametre Kestirimi, SIU 2020 Konferansı (5-7 Ekim)

Her yıl düzenlenen Sinyal İşleme ve İletişim Uygulamaları Kurultayı (SIU) için hazırlanmış olan bu bildiride gizli Markov modeli ile modellenen patlamalı-kesikli (bursty-intermittent) gürültü ve sinyal ortamı için parametre kestirim problemi çalışılmaktadır. Problem algılayıcı sistemin düşük hatalı ve yüksek hatalı olarak adlandırılan iki çalışma modu olduğu ve bu modlardan birine girdiği zaman, yeterince uzun süre aynı çalışma modunda kaldığı algılayıcı için veya bu özelliklere sahip gözlem toplama ortamında parametre kestirimi yapmak olarak düşünülebilir. Örneğin hareketli bir algılayıcı sistemin A noktası civarında olduğunda yüksek girişim etkisi altına girmesi ve algılayıcının kontrol edilemeyen şekilde A noktası civarında bulunması durumunda çalışmadaki kestirici kullanılabilir. Daha fazla bilgi için Ek-2'ye bakabilirsiniz.

Bu çalışmada ilginç olabilecek unsur problemin kuramsal olarak zengin olan gizli Markov modeli ile parametre kestirim konusunu içermesi ve bu problemin tam veya yaklaşık kestirim yöntemleri ile çözülebilir olmasıdır. Bahsedilen problemi $\alpha\beta$ yinelemeleri ve beklenti eniyileme (Expectation maximization) yöntemi ile tam olarak çözmek mümkündür. Ayrıca bu problemi "mean-field" ve "Expectation Propagation" yöntemleri ile yaklaşık olarak çözmek de mümkündür. Konferans çalışmasında sadece tam çözüm verilmiş ve diğer yaklaşık kestirim yöntemlerinden bahsedilmemiştir. Bahsi geçen gizli Markov yapısının çok az değişmesi durumunda (loopy graph) tam çözüm yapmak mümkün olmamaktadır. Yaklaşık çözümler bu durumlar için önem kazanmaktadır. İleride bu konferans çalışmasının devamı niteliğinde olan ve konferans çalışmasını özel durum olarak içerecek olan kismi ve kontrolü olarak "loopy graph" içerecek bir çalışma yapılması düşünülmektedir.

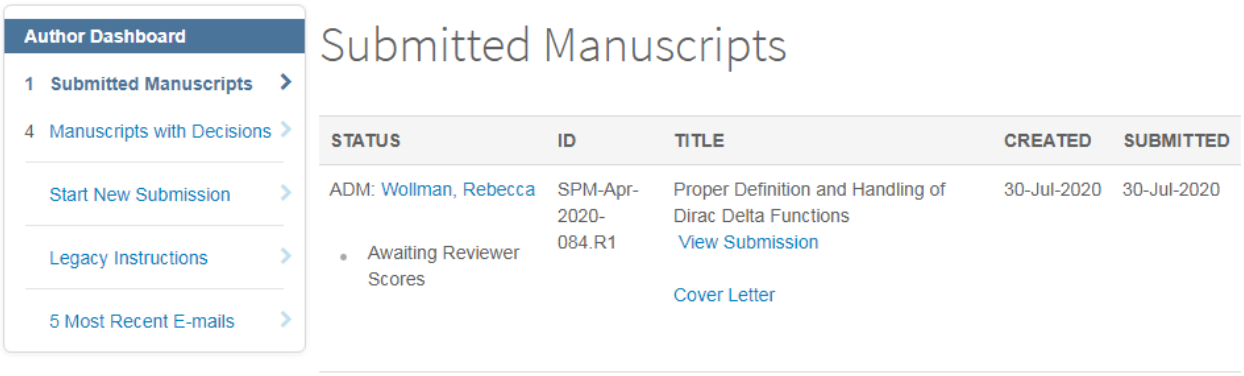
Çalışma 3: Proper Definition and Handling of Dirac Delta Functions, IEEE Signal Processing Magazine, Lecture Notes Column (değerlendirme aşamasında)

Bu çalışma Elektrik ve Elektronik Mühendisliği bölümleri lisans programı 2. sınıfta ilk kez ve 3. sınıfta daha detaylıca anlatılan Dirac delta fonksiyonları üzerinedir. Bu çalışma IEEE Signal Processing Magazine

isimli işaret işleme dalının en yüksek etki değerli dergide yer alan Ders Notları (Lecture Notes) köşesi için hazırlanmıştır. Okuyucuda ilgi uyandırdığını düşündüğüm bu köşede daha önceden yayınlanmış olan iki farklı ders notum bulunmaktadır.

Dergiye gönderilen son çalışmada Dirac delta fonksiyonları ve yoğun şekilde kullanıldıkları Fourier analizi konularıyla ilgili olarak mühendislik literatüründe pek yer almayan, ileri düzeyde matematik bilgisiyle tam olarak açıklanabilecek bazı konula, yine biraz basitleştirmeyle ama özüne daha sadık kalarak anlatılmaktadır. Özellikle Fourier dönüşümü sonucu genelleştirilmiş fonksiyonla (generalized function) ifade edilen, genelde tepeden inme şekilde verilen bazı ilişkilerin ispatları verilmektedir. Bu çalışmanın amacı mühendislik literatüründe pek yer almayan ama doğru şekilde kullanımı ile sonuç alınan ama temel kalkülüs kurallarıyla çelişkili gözüken bazı eşitlikleri açıklamaktır. Daha detaylı bilgi için Ek-3'e bakabilirsiniz.

Bu çalışma şu anda değerlendirme aşamasındadır. İlk revizyon sonrasındaki makale 30 Temmuz 2020'de tekrar gönderilmiştir. Aşağıda gönderi hakkında yer alan en güncel bilgiler verilmektedir.



Çalışma 4: Covariance Matrix Estimation of Compound-Gaussian Vectors With Texture Correlation, (hazırlanmakta olan makale)

Bu çalışmada gizli Markov modeli ile üretilen ilgilenmediğimiz ama performansı etkileyen bilinmeyen bir parametre (nuisance parameter) olduğu durumda parametre kestirimi problemi çalışılmaktadır. Bu çalışma yapı olarak SIU 2020'de sunulan çalışmaya benzer ama içerik olarak farklıdır. Bu çalışmada bileşik-Gauss (compound-Gaussian) vektörleri için kovaryans matrisi kestirimi çalışılmaktadır. Bu problem özellikle radar işaret işlemede yer alan uyarlamalı hedef tespiti için önemlidir. Çalışma kapsamında Markov modeli ile üretilen değeri bilinmeyen bir çarpan (texture parameter) kovaryans matrisi bilinmeyen Gauss dağılımlı vektör ile çarpılmakta ve çarpım sonucu olan vektör gürültü altında gözlenmektedir. Çalışmanın hedefi

kovaryans matrisinin gürültülü gözlemlerden, bilinmeyen çarpan değerine rağmen kestirilmesidir. Detaylar için Ek-4'e bakabilirsiniz.

Bu çalışmada CentraleSupélec University Paris-Saclay'de öğretim üyesi olan Prof. Frederic Pascal ile işbirliği yapılmaktadır. Frederic Pascal özellikle kovaryans matrisi kestirimi konusunda bir otorite olan, çok prestijli bir araştırma universite sistemi olan Saclay Universite sisteminde yer alan bir öğretim üyesidir. Salgın hastalık nedeniyle karantinada bulunduğumuz zaman diliminde, kendisiyle uzaktan erişimle (zoom programı) ortaklaşa olarak yaptığımız bu çalışmayı Elsevier Signal Processing veya IEEE Transactions of Aerospace and Electronic Systems dergilerinden birine yakın zamanda göndermeyi planlıyoruz.

Çalışma 5: Diğer Çalışmalar

Tez danışmanlığını yaptığım doktora öğrencim **Ömer Çayır** ile

- Ö. Çayir, C. Candan, "Transmit beamformer design with a PAPR constraint to trade-off between beampattern shape and power efficiency," Elsevier Digital Signal Processing, 102674, Nisan 2020.

yaptığımız çalışma izin süresinde revize edilmiş, tekrar gönderilmiş ve kabul olmuştur. Bu yayının ardından Ömer Çayır'ın tezinde yer almasını hedeflediğimiz "Maximum Likelihood Estimator for Noisy Autoregressive Model Parameter Estimation Problem with Multiple Snapshots" başlıklı ikinci bir makale hazırlanmış ve Eylül ayı içinde Elsevier Signal Processing dergisine değerlendirme için gönderilecektir.

Tez danışmanlığını yaptığım doktora öğrencim **Şafak Bilgi Akdemir** ile yaptığımız "Maximum-likelihood Direction of Arrival Estimation via Expectation Maximization Algorithm Under Intermittent Jamming" başlıklı yayın Elsevier Digital Signal Processing dergisine Nisan 2020'de gönderilmiş ve halen ilk hakem raporları beklenmektedir. Aşağıda bu yayın başvurusu hakkında yayıncı kuruluşun sitesinde yer alan yazarlara açık olan en güncel bilgiler verilmektedir.

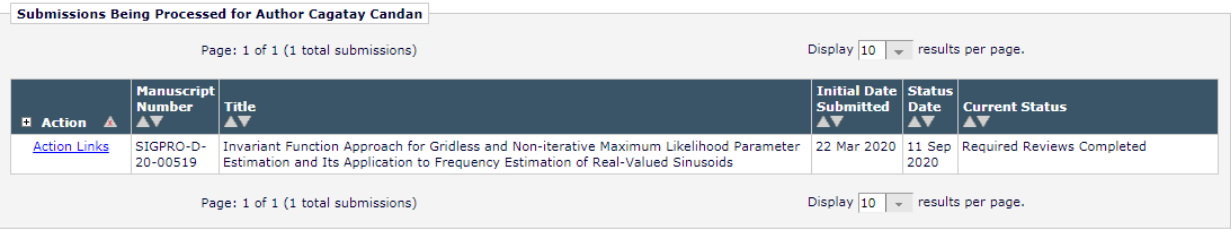
Page: 1 of 1 (1 total submissions) Display 10 results per page.

Action	Manuscript Number	Title	Initial Date Submitted	Status Date	Current Status
View Submission Send E-mail	DSP-D-20-00217	Maximum-likelihood Direction of Arrival Estimation via Expectation Maximization Algorithm Under Intermittent Jamming	Apr 10, 2020	Aug 23, 2020	Under Review

Page: 1 of 1 (1 total submissions) Display 10 results per page.

Tez danışmanlığını yaptığım yüksek lisans öğrencim **Utku Çelebi** ile SIU 2020 konferansı için yaptığımız "Harmonik İşaretlerin Hassas Frekans Kestirimi İçin Düşük İşlem Yüklü Bir Yöntem" bildiri sözlü sunum (zoom programı ile) için kabul olmuştur.

Utku Çelebi'nin ikinci yazar olduğu "Invariant Function Approach for Gridless and Non-iterative Maximum Likelihood Parameter Estimation and Its Application to Frequency Estimation of Real-Valued Sinusoids" başlıklı yayın Elsevier Signal Processing dergisine Mart 2020'de gönderilmiştir. Yayıncı kuruluşun sisteminde yer alan, aşağıda ilettiğimiz bilgiye göre, bu çalışmanın ilk değerlendirme sonuçlarının yakın zamanda bizlere iletilecek olarak gözükmektedir. Bu çalışmanın detaylarına Ek-5'ten bakabilirsiniz.



Submissions Being Processed for Author Cagatay Candan

Page: 1 of 1 (1 total submissions) Display 10 results per page.

Action	Manuscript Number	Title	Initial Date Submitted	Status Date	Current Status
Action Links	SIGPRO-D-20-00519	Invariant Function Approach for Gridless and Non-iterative Maximum Likelihood Parameter Estimation and Its Application to Frequency Estimation of Real-Valued Sinusoids	22 Mar 2020	11 Sep 2020	Required Reviews Completed

Page: 1 of 1 (1 total submissions) Display 10 results per page.

Chebyshev Center Computation on Probability Simplex With α-Divergence Measure

Çağatay Candan, *Senior Member, IEEE*

***Abstract*—Chebyshev center computation problem, i.e. finding the point which is at minimum distance to a set of given points, on the probability simplex with α-divergence distance measure is studied. The proposed solution generalizes the Arimoto-Blahut (AB) algorithm utilizing Kullback-Leibler divergence to α-divergence, and reduces to the AB method as $\alpha \to 1$. Similar to the AB algorithm, the method is an ascent method with a guarantee on the objective value (α-mutual information or Chebyshev radius) improvement at every iteration. A practical application area for the method is the fusion of probability mass functions lacking a joint probability description. Another application area is the error exponent calculation.**

***Index Terms*—Arimoto-Blahut algorithm, minimax-redundancy, redundancy-capacity theorem, alpha-divergence, Rényi-divergence, information fusion, error exponent calculation.**

I. Introduction

TO ILLUSTRATE the problem of interest, we consider a set of probability mass functions (pmf) $\mathbf{p}_j$, $j = \{1, \ldots, M\}$ each of which is said to represent a locally generated posterior distribution of a random variable of interest. The goal is to fuse the local posteriors to a set-representative pmf $\mathbf{q}$. In the absence of joint distribution information on $\mathbf{p}_j$, $j = \{1, \ldots, M\}$; the Bayesian approach can not be pursued. Instead, a minimax formulation $\widehat{\mathbf{q}} = \arg\min_{\mathbf{q}} \max_j D(\mathbf{p}_j||\mathbf{q})$ where $D(\mathbf{p}_j||\mathbf{q})$ is a distance measure between two distributions (possibly in a loose sense) can be suggested. The minimax solution $\widehat{\mathbf{q}}$ can be interpreted as the point on the probability simplex which minimizes the worst-case distance to the set members. This study presents an efficient method for the solution of the minimax problem with α-divergence distance measure.

The minimax problem $\widehat{\mathbf{q}} = \arg\min_{\mathbf{q}} \max_j D(\mathbf{p}_j||\mathbf{q})$ is also known as the Chebyshev center problem. Its optimizer $\widehat{\mathbf{q}}$ and optimal value $r = \min_{\mathbf{q}} \max_j D(\mathbf{p}_j||\mathbf{q})$ are called Chebyshev center and radius, respectively. The solution of minimax problem with Kullback-Leibler (KL) divergence is of major concern in information theory [1]. Source code design problem with minimax redundancy is identical to the Chebyshev center problem with KL-divergence, [1, Ch. 13]. Furthermore, Gallager and Ryabko [2], [3] have shown that the solution of the minimax problem coincides with the capacity calculation (mutual information maximization over all input distributions) for the special case of discrete memoryless channels with finite input/output alphabets [1, Theorem 13.1.1]. Capacity-redundancy theorem is considered as one of the cornerstone results of information theory and extended to infinite alphabets, probability density functions and measures under different divergences [4]–[7].

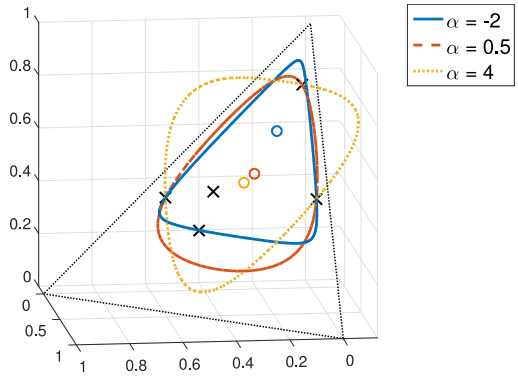


Fig. 1. 5 points (crosses) on probability simplex, Chebyshev centers (round markers) and Chebyshev circles for different orders of α-divergence.

Extending the capacity-redundancy discussions from KL divergence to α-divergence can be motivated by the desire of adjusting inclusiveness or exclusiveness of the solution [5], [8]. Fig. 1 shows five 3-dimensional probability vectors (crosses) and their Chebyshev centers (circles) according to different α-divergence orders. Loosely speaking, the divergence order affects how to measure the distance between points. Chebyshev circle for $\alpha = -2$ in Fig. 1 can be said to be risk-avoiding, that is showing a tendency not to include as many points as other circles. The circle for $\alpha = 0.5$ can be considered to have a balanced risk. The capability of adjusting the degree of inclusiveness/exclusiveness is especially important in approximate inference problems [8].

This study presents a method for the computation of Chebyshev center with α-divergence for orders $\alpha \in (0, 1)$ which is the interval in which risk balancing is possible [8]. The suggested method is an alternating minimization-maximization method with a Chebyshev-radius improvement guarantee at every iteration. The method approaches Arimoto-Blahut (AB) algorithm as $\alpha \to 1$ [9], [10] and can be interpreted as its generalization to α-divergence measure. Utilized proximal point based approach is an extension of earlier similar efforts by Chretien and Hero, in the context of Expectation-Maximization (EM) algorithm [11]; Matz and Duhamel, in the context of AB algorithm [12], to α-mutual information maximization problem. The method can

Manuscript received June 21, 2020; accepted August 17, 2020. Date of publication August 21, 2020; date of current version September 7, 2020. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Qing Ling.

The author is with the Department of Electrical, and Electronics Engineering, Middle East Technical University (METU), Ankara 06800, Turkey (e-mail: ccandan@metu.edu.tr).

Digital Object Identifier 10.1109/LSP.2020.3018661

be utilized in information fusion [13], error-exponent calculation [14]–[16] and α-divergence applications [5], [17].

II. Preliminaries

The column vectors $\mathbf{p_j}$, $j=\{1,\dots,M\}$ of dimension N with nonnegative entries, $p_j(i)\geq 0$, $i=\{1,\dots,N\}$, denote probability mass functions (pmf) defined over the alphabet $\mathcal{Y}=\{1,2,\dots,N\}$. The set of all pmf's for alphabet $\mathcal{Y}$ is shown with $\mathcal{P}^{\mathcal{Y}}$. For $\mathbf{p}\in\mathcal{P}^{\mathcal{Y}}$ and $\mathbf{q}\in\mathcal{P}^{\mathcal{Y}}$, f-divergence is defined as $D_f(\mathbf{p}||\mathbf{q})=\sum_{i=1}^{N} q(i)f(p(i)/q(i))$, where $f(r)$ is a convex function with $f(1)=0$, [18], [19]. Our focus is on α-divergence $D_\alpha(\mathbf{p}||\mathbf{q})$ which is a special case of f-divergence for $f_\alpha(r)=\frac{1-r^\alpha}{\alpha(1-\alpha)}$, $\alpha\in\mathbb{R}\setminus\{0,1\}$:

$$D_\alpha(\mathbf{p}||\mathbf{q})=\sum_{i=1}^{N} q(i) f_\alpha\left(\frac{p(i)}{q(i)}\right)=\frac{1-\sum_{i=1}^{N} p^\alpha(i)q^{1-\alpha}(i)}{\alpha(1-\alpha)}. \tag{1}$$

It can be shown that as $\alpha\to 1$, $D_\alpha(\mathbf{p}||\mathbf{q})$ approaches KL-divergence $\mathrm{KL}(\mathbf{p}||\mathbf{q})=\sum_i p(i)\log(p(i)/q(i))$ [8]. For additional properties of α-divergence and general properties of f-divergence, one can examine [8], [20].

The Chebyshev center problem can be expressed as

$$P_1:\quad \min_{\mathbf{q}\in\mathcal{P}^{\mathcal{Y}}}\max_{j} D_\alpha(\mathbf{p}_j||\mathbf{q}). \tag{2}$$

The outer minimization in (2) is a convex problem due to i. joint convexity of f-divergences over both arguments [20, Theorem 6.1], ii. convexity preservation by maximum function. Problem P_1 can be reparameterized as:

$$P_2:\ \min_{z,\mathbf{q}} z\quad \text{s.t.}\ \begin{cases} D_\alpha(\mathbf{p}_j||\mathbf{q})\leq z,\ j=\{1,\dots,M\} \\ \mathbf{q}\in\mathcal{P}^{\mathcal{Y}} \end{cases}. \tag{3}$$

The equivalent formulation P_2 has linear objective with convex constraints. Lagrangian function for problem P_2 is

$$L(z,\mathbf{q},\mathbf{p}_x)=z+\sum_{j=1}^{M} p_x(j)(D_\alpha(\mathbf{p}_j||\mathbf{q})-z), \tag{4}$$

along with the constraint of $\mathbf{q}\in\mathcal{P}^{\mathcal{Y}}$ not noted in (4) for the sake of expression clarity. In (4), $p_x(j)$ are nonnegative valued Lagrange multipliers which are the unknowns of dual problem [21]. Differentiating $L(z,\mathbf{q},\mathbf{p}_x)$ with respect to z, immediately yields the stationarity condition of $\sum_{j=1}^{M} p_x(j)=1$. Hence, we reach the important conclusion that elements of $M\times 1$ dimensional Lagrange multiplier vector $\mathbf{p}_x$ are nonnegative valued with a cumulative sum of 1. Therefore, we consider Lagrange multiplier vector $\mathbf{p}_x$ as a probability vector lying in M-dimensional probability simplex $\mathcal{P}^{\mathcal{X}}$, $\mathcal{X}=\{1,\dots,M\}$.

The dual problem $\mathcal{D}:\max_{\mathbf{p_x}}\min_{z,\mathbf{q}} L(z,\mathbf{q},\mathbf{p}_x)$ can be stated as

$$\mathcal{D}:\max_{\mathbf{p}_x\in\mathcal{P}^{\mathcal{X}}}\min_{\mathbf{q}\in\mathcal{P}^{\mathcal{Y}}}\sum_{j=1}^{M} p_x(j)D_\alpha(\mathbf{p}_j||\mathbf{q}). \tag{5}$$

In this study, we use of the dual problem statement in (5) to formulate an alternating maximization-minimization solution.

To establish connections with other works in the literature, we introduce $N\times M$ dimensional probability transition matrix (channel) $\mathbf{P}_{Y|X}=[\mathbf{p}_1\,\mathbf{p}_2\,\dots\,\mathbf{p}_M]$ which is formed by the juxtaposition of column vectors $\mathbf{p}_j$. Stated differently, the j'th column of $\mathbf{P}_{Y|X}$ is $\mathbf{P}_{Y|X=j}=\mathbf{p}_j$. With this definition, the objective of dual problem $\mathcal{D}$ coincides with conditional α-divergence definition in [6]:

$$D_\alpha(\mathbf{P}_{Y|X}||\mathbf{q}\,|\,\mathbf{p}_x)\triangleq E_{\mathbf{x}\sim\mathbf{p}_x}\{D_\alpha(\mathbf{P}_{Y|X=j}||\mathbf{q}).\}. \tag{6}$$

With this definition, the dual problem (5) can be expressed as:

$$\mathcal{D}:\max_{\mathbf{p}_x\in\mathcal{P}^{\mathcal{X}}}\underbrace{\min_{\mathbf{q}\in\mathcal{P}^{\mathcal{Y}}} D_\alpha(\mathbf{P}_{Y|X}||\mathbf{q}\,|\,\mathbf{p}_x)}_{I_\alpha(\mathbf{P}_{Y|X},\mathbf{p}_x)}. \tag{7}$$

In [22], Sibson introduced a definition for α-mutual information, through information radius considerations, as follows:

$$I_\alpha(\mathbf{P}_{Y|X},\mathbf{p}_x)\triangleq\min_{\mathbf{q}\in\mathcal{P}^{\mathcal{Y}}} D_\alpha(\mathbf{P}_{Y|X}||\mathbf{q}\,|\,\mathbf{p}_x). \tag{8}$$

This definition generalizes the conventional mutual-information definition.[1] In fact, the conventional mutual information is the special case of Sibson's definition as $\alpha\to 1$. In this study, we recognize the fact that the dual problem of minimax redundancy for α-divergence measure, say for the sake of information fusion, given in (7) is the capacity maximization problem with Sibson's α-mutual information definition, that is $C_\alpha=\max_{\mathbf{p}_x\in\mathcal{P}^{\mathcal{X}}} I_\alpha(\mathbf{P}_{Y|X},\mathbf{p}_x)$.

As a final remark, we note that Slater's strong duality condition [21] is satisfied for the given problem. Hence, with the equality of primal (P_1) and dual optimal values, we have:

$$\begin{aligned} C_\alpha &\triangleq \max_{\mathbf{p}_x\in\mathcal{P}^{\mathcal{X}}}\min_{\mathbf{q}\in\mathcal{P}^{\mathcal{Y}}} D_\alpha(\mathbf{P}_{Y|X}||\mathbf{q}\,|\,\mathbf{p}_x) \\ &= \min_{\mathbf{q}\in\mathcal{P}^{\mathcal{Y}}}\max_{j} D_\alpha(\mathbf{p}_j||\mathbf{q}). \end{aligned}$$

The optimizer $\mathbf{q}^*=\mathrm{argmin}_{\mathbf{q}\in\mathcal{P}^{\mathcal{Y}}}\max_j D_\alpha(\mathbf{p}_j||\mathbf{q})$ and its value $C_\alpha=\max_{\mathbf{p}_x\in\mathcal{P}^{\mathcal{X}}} I_\alpha(\mathbf{P}_{Y|X},\mathbf{p}_x)$ are denoted as the Chebyshev center and radius, respectively. An efficient method for the Chebyshev center computation is the topic of this study.

III. Proposed Method

Proposed method iteratively solves the optimization problem given in (5) by fixing either $\mathbf{p}_x$ or $\mathbf{q}$, alternatively and optimizing over the other variable.

Minimization over $\mathbf{q}$ for a fixed $\mathbf{p}_x^{(k)}$: When $\mathbf{p}_x$ is fixed to $\mathbf{p}_x^{(k)}$, the dual problem in (5) reduces to an optimization problem involving a weighted average of α-divergences,

$$\begin{aligned} \mathbf{q}^{(k)} &= \mathrm{argmin}_{\mathbf{q}\in\mathcal{P}^{\mathcal{Y}}}\sum_{j=1}^{M} p_x^{(k)}(j)D_\alpha(\mathbf{p}_j||\mathbf{q}) \\ &\overset{(b)}{=} \mathrm{argmax}_{\mathbf{q}\in\mathcal{P}^{\mathcal{Y}}}\sum_{j=1}^{M} p_x^{(k)}(j)\sum_{i=1}^{N} p_j^\alpha(i)q^{1-\alpha}(i). \end{aligned} \tag{9}$$

Line-(b) is obtained based on (1). By calculating the gradient of the objective function in (9), the elements of $\mathbf{q}^{(k)}$ can be written as

$$q^{(k)}(i)=\frac{1}{c\left(\mathbf{p}_x^{(k)}\right)}\left(\sum_{j=1}^{M} p_x^{(k)}(j)P_{Y|X=j}^\alpha(i)\right)^{1/\alpha}. \tag{10}$$

Here $P_{Y|X=j}^\alpha(i)$ is the (i,j) entry of $\mathbf{P}_{Y|X}$ raised to the power α, that is $P_{Y|X=j}^\alpha(i)=([\mathbf{P}_{Y|X}]_{i,j})^\alpha=p_j^\alpha(i)$ and

[1] Sibson's definition is given for Rényi divergence, which is not an f-divergence; but in one-to-one correspondence with α-divergence in (1). Also see [7, Section I] for more connections with existing works in the literature.

$c(\mathbf{p}_x^{(k)})$ is a normalization constant defined as $c(\mathbf{p}_x^{(k)}) = \sum_{i=1}^{N}(\sum_{j=1}^{M} p_x^{(k)}(j) P_{Y|X=j}^{\alpha}(i))^{1/\alpha}$. The normalization constant $c(\mathbf{p}_x^{(k)})$ can also be expressed in terms of Gallager function [23] as $c(\mathbf{p}_x) = \exp(-E_0(\frac{1-\alpha}{\alpha}, \mathbf{p}_x))$ where

$$E_0(\rho, \mathbf{p}_x) = -\log\left(\sum_{i=1}^{N}\left(\sum_{j=1}^{M} p_x(j) P_{Y|X=j}^{\frac{1}{1+\rho}}(i)\right)^{1+\rho}\right). \tag{11}$$

The optimal vector $\mathbf{q}^{(k)}$ in (10) is denoted as the α-response to $\mathbf{p}_x^{(k)}$ in the literature [6]. It is easy to see that as $\alpha \to 1$, $\mathbf{q}^{(k)} \to \mathbf{P}_{Y|X}\mathbf{p}_x^{(k)}$, which is the output distribution of channel $\mathbf{P}_{Y|X}$ for the input $\mathbf{p}_x^{(k)}$. Normalization constant $c(\mathbf{p}_x)$ is upper bounded by 1 for $\alpha \in (0,1)$ and $c(\mathbf{p}_x) \to 1$ as $\alpha \to 1$.

Inserting α-response to $\mathbf{p}_x^{(k)}$ into (8) and simplifying, we get the α-mutual information induced by $\mathbf{p}_x^{(k)}$ as

$$I_\alpha(\mathbf{P}_{Y|X}, \mathbf{p}_x^{(k)}) = \frac{1 - c^\alpha(\mathbf{p}_x^{(k)})}{\alpha(1-\alpha)} = \frac{1 - e^{-\alpha E_0\left(\frac{1-\alpha}{\alpha}, \mathbf{p}_x^{(k)}\right)}}{\alpha(1-\alpha)}. \tag{12}$$

Here $I_\alpha(\mathbf{P}_{Y|X}, \mathbf{p}_x^{(k)})$ is the value of the dual problem objective, given in (7), at the k'th iteration.

Maximization over $\mathbf{p}_x$ for a fixed $\mathbf{q}^{(k)}$: Exact solution of the dual problem in (7) requires the solution of $\max_{\mathbf{p}_x \in \mathcal{P}^{\mathcal{X}}} I_\alpha(\mathbf{P}_{Y|X}, \mathbf{p}_x)$ which is guaranteed to be a concave maximization problem by duality [21]. An iterative solution is given by Arimoto in the context of Gallager function maximization in [14]. We present a novel method based on proximal point iterations [24]. Approach is in principle similar to the ones given for EM and AB algorithms in [11], [12].

As a naive attempt, we may try to fix $\mathbf{q}$ to $\mathbf{q}^{(k)}$ in (5), as in earlier sub-problem, and optimize over $\mathbf{p}_x$. This attempt leads to a linear program without any interior point solution and not particularly suitable for an iterative optimization. Instead, we suggest to modify the problem to

$$\mathbf{p}_x^{(k+1)} = \underset{\mathbf{p}_x \in \mathcal{P}^{\mathcal{X}}}{\operatorname{argmax}}\, f(\mathbf{p}_x, \mathbf{p}_x^{(k)}), \tag{13}$$

and $f(\mathbf{p}_x, \mathbf{p}_x^{(k)}) \triangleq \sum_{j=1}^{M} p_x(j) D_\alpha(\mathbf{p}_j || \mathbf{q}^{(k)}) - \mu^{-1}\mathrm{KL}(\mathbf{p}_x || \mathbf{p}_x^{(k)})$. The modified problem aims to update the solution in the proximity of the previous primal variable estimate $\mathbf{p}_x^{(k)}$. The deviation from earlier iteration $\mathbf{p}_x^{(k)}$ is penalized with μ^{-1}. It is shown that if the step-size parameter μ is chosen properly, the algorithm monotonically converges to the optimum.

The problem in (13) is additively separable. By evaluating $\frac{\partial f(\mathbf{p}_x, \mathbf{p}_x^{(k)})}{\partial p_x(j)} = D_\alpha(\mathbf{p}_j || \mathbf{q}^{(k)}) - \frac{\log(p_x(j)) + 1 - \log(p_x^{(k)}(j))}{\mu}$ and optimizing over the simplex, we get the update equation as:

$$p_x^{(k+1)}(j) = p_x^{(k)}(j) \frac{e^{\mu D_\alpha(\mathbf{p}_j || \mathbf{q}^{(k)})}}{\sum_j e^{\mu D_\alpha(\mathbf{p}_j || \mathbf{q}^{(k)})}}. \tag{14}$$

Selection of step-size parameter: It is shown below that the step-size μ can be selected to guarantee the monotonic increase of dual problem objective, $I_\alpha(\mathbf{P}_{Y|X}, \mathbf{p}_x^{(k)})$ in (7), at every iteration.

From (13), $f(\mathbf{p}_x, \mathbf{p}_x^{(k)})|_{\mathbf{p}_x = \mathbf{p}_x^{(k+1)}} \geq f(\mathbf{p}_x, \mathbf{p}_x^{(k)})|_{\mathbf{p}_x = \mathbf{p}_x^{(k)}}$; since $\mathbf{p}_x^{(k+1)}$ is the maximizer of the problem. The evaluation of right hand side, $f(\mathbf{p}_x^{(k)}, \mathbf{p}_x^{(k)}) = f(\mathbf{p}_x, \mathbf{p}_x^{(k)})|_{\mathbf{p}_x = \mathbf{p}_x^{(k)}}$, is immediate, $f(\mathbf{p}_x^{(k)}, \mathbf{p}_x^{(k)}) = I_\alpha(\mathbf{P}_{Y|X}, \mathbf{p}_x^{(k)})$; since $\mathrm{KL}(\mathbf{p}_x^{(k)} || \mathbf{p}_x^{(k)}) = 0$. For the evaluation of left hand side, $f(\mathbf{p}_x, \mathbf{p}_x^{(k)})|_{\mathbf{p}_x = \mathbf{p}_x^{(k+1)}} = f(\mathbf{p}_x^{(k+1)}, \mathbf{p}_x^{(k)})$, we examine the summation in (13), $S = \sum_{j=1}^{M} p_x^{(k+1)}(j) D_\alpha(\mathbf{p}_j || \mathbf{q}^{(k)})$ first:

$$\begin{aligned} S &\overset{(a)}{=} \sum_{j=1}^{M} p_x^{(k+1)}(j) \sum_{i=1}^{N} q^{(k)}(i) f_\alpha\left(\frac{p_j(i)}{q^{(k)}(i)}\right) \\ &\overset{(b)}{=} \frac{1 - \sum_{j=1}^{M} p_x^{(k+1)}(j) \sum_{i=1}^{N} (q^{(k)}(i))^{1-\alpha} p_j^\alpha(i)}{\alpha(1-\alpha)} \\ &\overset{(c)}{=} \frac{1 - c^\alpha(\mathbf{p}_x^{(k+1)}) \sum_{i=1}^{N} q^{(k)}(i) \left(\frac{q^{(k+1)}(i)}{q^{(k)}(i)}\right)^\alpha}{\alpha(1-\alpha)} \end{aligned} \tag{15}$$

Line-(a) follows the α-divergence definition in (1). In line-(b), $f_\alpha(r) = \frac{1-r^\alpha}{\alpha(1-\alpha)}$ is substituted. In line-(c), summation over j is recognized from (10) as $(q^{(k+1)}(i) c(\mathbf{p}_{\mathbf{x}}^{(k+1)}))^\alpha$.

The summation over i on the numerator of line-(c) of (15) is in the form of $\sum_{i=1}^{N} q^{(k)}(i) g_\alpha(\frac{q^{(k+1)}(i)}{q^{(k)}(i)})$, where $g_\alpha(r) = r^\alpha$. In order to write this summation in terms α-divergence with $f_\alpha(r) = \frac{1-r^\alpha}{\alpha(1-\alpha)}$, the function $g_\alpha(r)$ is expressed as $g_\alpha(r) = 1 - f(r)\alpha(1-\alpha)$:

$$\begin{aligned} S &= \frac{1 - c^\alpha(\mathbf{p}_x^{(k+1)}) \sum_{i=1}^{N} q^{(k)}(i) g_\alpha\left(\frac{q^{(k+1)}(i)}{q^{(k)}(i)}\right)}{\alpha(1-\alpha)} \\ &= \frac{1 - c^\alpha(\mathbf{p}_x^{(k+1)})}{\alpha(1-\alpha)} + c^\alpha(\mathbf{p}_x^{(k+1)}) D_\alpha(\mathbf{q}^{(k+1)} || \mathbf{q}^{(k)}) \\ &\overset{(d)}{=} I_\alpha(\mathbf{P}_{Y|X}, \mathbf{p}_x^{(k+1)}) + c^\alpha(\mathbf{p}_x^{(k+1)}) D_\alpha(\mathbf{q}^{(k+1)} || \mathbf{q}^{(k)}). \end{aligned} \tag{16}$$

In line-(d), α-mutual information definition from (12) is recognized. With these results, the inequality of $f(\mathbf{p}_x^{(k+1)}, \mathbf{p}_x^{(k)}) \geq f(\mathbf{p}_x^{(k)}, \mathbf{p}_x^{(k)})$ is equivalent to

$$I_\alpha(\mathbf{P}_{Y|X}, \mathbf{p}_x^{(k+1)}) - I_\alpha(\mathbf{P}_{Y|X}, \mathbf{p}_x^{(k)}) \geq \gamma \tag{17}$$

with $\gamma \triangleq \mu^{-1}\mathrm{KL}(\mathbf{p}_x^{(k+1)} || \mathbf{p}_x^{(k)}) - c^\alpha(\mathbf{p}_x^{(k+1)}) D_\alpha(\mathbf{q}^{(k+1)} || \mathbf{q}^{(k)})$. We consider the value of γ as the margin. If margin is positive at an iteration, the objective value (α-mutual information) is improved at that iteration. In fact, setting

$$\mu \leq \frac{\mathrm{KL}(\mathbf{p}_x^{(k+1)} || \mathbf{p}_x^{(k)})}{c^\alpha(\mathbf{p}_x^{(k+1)}) D_\alpha(\mathbf{q}^{(k+1)} || \mathbf{q}^{(k)})} \triangleq \bar{\mu}^{(k)} \tag{18}$$

guarantees the positivity of the margin at that iteration. Yet, this step-size selection rule is inadmissible; since the step-size depends on the divergence values after the update. A rather pessimistic approach can be the calculation of a lower bound for the right side of (18):

$$\begin{aligned} \frac{\mathrm{KL}(\mathbf{p}_x^{(k+1)} || \mathbf{p}_x^{(k)})}{c^\alpha(\mathbf{p}_x^{(k+1)}) D_\alpha(\mathbf{q}^{(k+1)} || \mathbf{q}^{(k)})} &\overset{(a)}{\geq} \frac{\mathrm{KL}(\mathbf{p}_x^{(k+1)} || \mathbf{p}_x^{(k)})}{D_\alpha(\mathbf{q}^{(k+1)} || \mathbf{q}^{(k)})} \\ &\overset{(b)}{\geq} \frac{\mathrm{KL}(\mathbf{p}_x^{(k+1)} || \mathbf{p}_x^{(k)})}{\mathrm{KL}(\mathbf{q}^{(k+1)} || \mathbf{q}^{(k)})} \\ &\overset{(c)}{\geq} 1. \end{aligned} \tag{19}$$

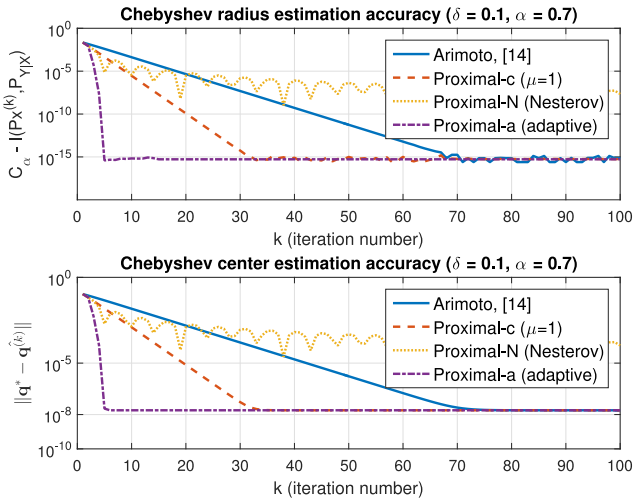


Fig. 2. Example 1 - Convergence of methods to Chebyshev circle parameters.

Line-(a) is due to the fact that $c(\mathbf{p}_x^{(k+1)}) = \exp(-E_0(\frac{1-\alpha}{\alpha}, \mathbf{p}_x)) \leq 1$ for $\alpha \in (0,1)$. Line-(b) follows from the monotonic increase of α-divergence with divergence order α, [6, Property 9]. Line-(c) is the data-processing inequality for KL-divergence [20], [25]. Hence, the choice of $\mu = 1$ guarantees the monotonic convergence. (Also, for $\mu = 1$, the method reduces to AB algorithm as $\alpha \to 1$.)

IV. Numerical Results

Example 1: Probability vectors $\mathbf{p}_j$, $j = \{1,2,3\}$ of dimension 4×1 are concatenated to form the columns of channel $\mathbf{P}_{Y|X}$, given below, which is also studied in [6, Ex. 6]:

$$\mathbf{P}_{Y|X} = \begin{bmatrix} \frac{1}{2}-\delta & \delta & \frac{1}{2}-\delta \\ \delta & \frac{1}{2}-\delta & \delta \\ \delta & \frac{1}{2}-\delta & \delta \\ \frac{1}{2}-\delta & \delta & \frac{1}{2}-\delta \end{bmatrix}, \quad \delta \in \left[0, \frac{1}{2}\right] \tag{20}$$

Due to the symmetry in the problem, the Chebyshev center, or α-capacity achieving output distribution, is $[\frac{1}{4}\ \frac{1}{4}\ \frac{1}{4}\ \frac{1}{4}]$. The Chebyshev radius, or α-capacity expression, is $C_\alpha = [1 - 2^{2\alpha-1}(\delta^\alpha + (\frac{1}{2}-\delta)^\alpha)]/[\alpha(1-\alpha)]$.

Fig. 2 shows the accuracy improvement for Chebyshev center and radius estimates versus iterations for $\delta = 0.1$ and α-divergence order $\alpha = 0.7$. Arimoto's error exponent calculation algorithm [14] and different versions of proximal point algorithm are compared. The curve with "Proximal-c" label shows the case with $\mu = 1$ at every iteration. "Proximal-N" shows the version with Nesterov's acceleration as given in [26]. "Proximal-a" is the adaptive version of the suggested scheme where μ changes at every iteration.

In "Proximal-a," the step-size of current iteration is taken as the bound $\bar{\mu}$ in (18) calculated from the data of the previous iteration, $\mu^{(k)} = \bar{\mu}^{(k-1)}$. It is observed from Fig. 2 that "Proximal-a" converges rapidly to the accuracy of the computing platform with this policy. Table I shows the number of iterations required to get 10 digit accuracy in the capacity estimate. Main conclusions of this experiment are i.) "Proximal-c" with a constant step-size of $\mu = 1$ has a monotonic performance as expected and outperforms Arimoto's method, ii.) Nesterov's acceleration brings some improvements over "Proximal-c"; but requires more effort for its adaptation to the probability simplex, iii.) "Proximal-a" with suggested step-size policy yields very rapid convergence.

TABLE I
Number of Iterations Required for 10 Digit Accuracy at α-Capacity in Example 1 Setting ($\delta = 0.1$)

	Arimoto [14]	Proximal-N	Proximal-c	Proximal-a
$\alpha = 0.1$	2034	119	17	5
$\alpha = 0.3$	228	72	18	5
$\alpha = 0.5$	84	79	19	5
$\alpha = 0.7$	44	101	21	5
$\alpha = 0.9$	28	10	22	5

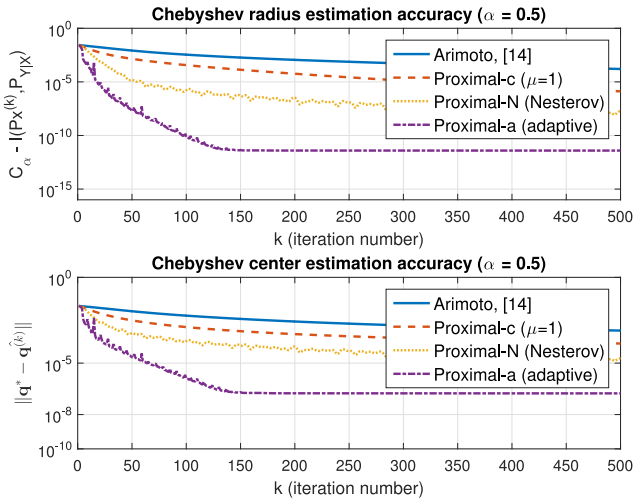


Fig. 3. Example 2 - Convergence of methods to Chebyshev circle parameters.

Example 2: This example presents a higher dimensional comparison with $M = 25$ vectors on $N = 100$ dimensional probability simplex. Fig. 3 shows the results for randomly sampled probability vectors on the simplex. Analytical expressions for the Chebyshev center and radius are not available and calculated by the general purpose numerical optimization routines. We see that Arimoto's method and "Proximal-c" present monotonic performance improvements; while "Proximal-a" suffers intermittent performance losses. In this example, we limit the adaptive step size range to $\mu \in [1, 50]$ by setting $\mu^{(k)} = \min(50, \bar{\mu}^{(k-1)})$. Here, the factor 50 indicates that acceleration over "Proximal-c" can be up to 50 fold and observed sporadic performance losses are due to over-acceleration.

Computational Complexity Discussion: The sum for the update in (10) corresponds to a matrix and vector product, with a complexity $\mathcal{O}(NM)$ multiplications, assuming that $P^\alpha_{Y|X=j}(i)$ is pre-computed and stored. The update in (14) requires M fold α-divergence calculation, an operation of complexity $\mathcal{O}(NM)$ multiplications, and an additional $\mathcal{O}(M)$ multiplications. Hence, the overall complexity of suggested scheme is $\mathcal{O}(NM)$ multiplications per iteration.

V. Conclusion

An efficient method for the computation of Chebyshev center and radius (maximum α-mutual information) with α-divergence measure for finite samples spaces is given. The scheme generalizes the celebrated Arimoto-Blahut algorithm to α-divergence measure. An important open problem is the extension of the suggested method to continuous random variables. Ready-to-use Matlab codes reproducing the results in paper (and more) are available in [27].

REFERENCES

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.

[2] R. G. Gallager, "Source coding with side information and universal coding," Laboratory for Information Decision Systems, MIT, Cambridge, MA, USA, Tech. Rep. LIDS-P-937, 1979.

[3] B. Y. Ryabko, "Coding a source with unknown but ordered probabilities," *Probl. Inf. Transmission*, vol. 15, no. 2, pp. 134–138, Oct. 1979.

[4] J. Kemperman, "On the Shannon capacity of an arbitrary channel," *Indagationes Mathematicae (Proc.)*, vol. 77, no. 2, pp. 101–115, 1974.

[5] S. Yagli, Y. Altug, and S. Verdú, "Minimax Rényi redundancy," *IEEE Trans. Inf. Theory*, vol. 64, no. 5, pp. 3715–3733, May 2018.

[6] C. Cai and S. Verdú, "Conditional Rényi divergence saddlepoint and the maximization of α-mutual information," *Entropy*, vol. 21, no. 10, 2019, Art. no. 969. [Online]. Available: https://www.mdpi.com/1099-4300/21/10/969

[7] B. Nakiboglu, "The Rényi capacity and center," *IEEE Trans. Inf. Theory*, vol. 65, no. 2, pp. 841–860, Feb. 2019.

[8] T. Minka, "Divergence measures and message passing," Tech. Rep. MSR-TR-2005-173, Jan. 2005. [Online]. Available: https://www.microsoft.com/en-us/research/publication/divergence-measures-and-message-passing/

[9] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 1, pp. 14–20, Jan. 1972.

[10] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 4, pp. 460–473, Jul. 1972.

[11] S. Chretien and A. O. Hero, "Kullback proximal algorithms for maximum-likelihood estimation," *IEEE Trans. Inf. Theory*, vol. 46, no. 5, pp. 1800–1810, Aug. 2000.

[12] G. Matz and P. Duhamel, "Information geometric formulation and interpretation of accelerated Blahut–Arimoto-type algorithms," in *Proc. Inf. Theory Workshop*, 2004, pp. 66–70.

[13] S. J. Julier, "An empirical study into the use of Chernoff information for robust, distributed fusion of Gaussian mixture models," in *Proc. Int. Conf. Inf. Fusion*, 2006, pp. 1–8.

[14] S. Arimoto, "Computation of random coding exponent functions," *IEEE Trans. Inf. Theory*, vol. IT-22, no. 6, pp. 665–671, Nov. 1976.

[15] Y. Polyanskiy and S. Verdú, "Arimoto channel coding converse and Rényi divergence," in *Proc. 48th Annu. Allerton Conf. Commun., Control, Comput.*, 2010, pp. 1327–1333.

[16] Y. Jitsumatsu and Y. Oohama, "A new iterative algorithm for computing the correct decoding probability exponent of discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. 66, no. 3, pp. 1585–1606, Mar. 2020.

[17] I. Sason and S. Verdú, "Arimoto–Rényi conditional entropy and Bayesian M-Ary hypothesis testing," *IEEE Trans. Inf. Theory*, vol. 64, no. 1, pp. 4–25, Jan. 2018.

[18] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *J. Royal Statist. Soc. B (Methodological)*, vol. 28, no. 1, pp. 131–142, 1966.

[19] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observations," *Studia Scientiarum Mathematicarum Hungarica*, vol. 2, pp. 299–318, 1967.

[20] Y. Polyanskiy and Y. Wu, "Lecture notes on information theory," MIT, Cambridge, MA, USA, 2019.

[21] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[22] R. Sibson, "Information radius," *Zeitschrift Wahrscheinlichkeitstheorie Verwandte Gebiete*, vol. 14, no. 2, pp. 149–160, 1969.

[23] R. G. Gallager, *Information Theory and Reliable Communication*. Hoboken, NJ, USA: Wiley, 1968.

[24] R. T. Rockafellar, "Monotone operators and the proximal point algorithm," *SIAM J. Control Optim.*, vol. 14, no. 5, pp. 877–898, 1976.

[25] F. Liese and I. Vajda, "On divergences and informations in statistics and information theory," *IEEE Trans. Inf. Theory*, vol. 52, no. 10, pp. 4394–4412, Oct. 2006.

[26] R. Tibshirani, "Lecture notes on convex optimization," CMU, Pittsburgh, PA, USA, 2019.

[27] C. Candan. (2020). "Chebyshev center computation on probability simplex with α-divergence measure [source code]." [Online]. Available: https://doi.org/10.24433/CO.3654345.v1

EK-1B

EE491-EE492

Senior Design Project Final Report

OPTIMAL DENSITIES, GEODESICS AND DISTANCE MEASURES FOR PROBABILITY DISTRIBUTIONS

Submitted by: Omer AVCI

Principal Investigator: Prof. Bulent Sankur, Prof. Çağatay Candan

13.05.2020

1. Introduction

The Smallest Enclosing Ball problem, is a mathematical problem of computing the smallest ball that contains all of a given set of points in the Euclidean plane. This problem arises in applications such as location analysis and military operations and it is itself of interest as a problem in computational geometry. The Enclosing Ball problem in the plane is an example of a facility location problem (the 1-center problem) in which the location of a new facility must be chosen to provide service to a number of customers, minimizing the farthest distance that any customer must travel to reach the new facility [1]. Also the problem has wide usage areas like Machine Learning. It can also be used for training Support Vector Machines (SVMs). In this Project we aim to learn about Probability and Information Theory alongside the Convex Optimization. We aim to learn about the algorithms that are currently used to solve Smallest Enclosing Ball problem and hopefully developing a new robust algorithm.

2. Objectives

One of our objective is learning the properties of linear mixture density and exponential mixture density of a given set of probability distributions. We can define linear mixture density of a given set of probability distributions $f_1, f_2, \dots, f_n$ as:

$$\alpha_1 f_1 + \alpha_2 f_2 + \cdots + \alpha_n f_n$$

for some non-negative real number $\alpha_1, \alpha_2, \dots, \alpha_n$ such that:

$$\alpha_1 + \alpha_2 + \cdots + \alpha_n = 1.$$

We can define exponential mixture density of some given probability distributions $f_1, f_2, \dots, f_n$ as:

$$\frac{1}{M}(f_1^{\alpha_1} f_2^{\alpha_2} \dots f_n^{\alpha_n})$$

for some non-negative real number $\alpha_1, \alpha_2, \dots, \alpha_n$ such that:

$$\alpha_1 + \alpha_2 + \cdots + \alpha_n = 1$$

$$M = \int_{-\infty}^{\infty} f_1^{\alpha_1}(x) f_2^{\alpha_2}(x) \dots f_n^{\alpha_n}(x) dx$$

Similary it can be defined as follows for discrete probability distributions:

$$M = \sum_{x \in S} f_1^{\alpha_1}(x) f_2^{\alpha_2}(x) \dots f_n^{\alpha_n}(x) .$$

Probability distributions arising from linear or exponential averaging play a role in our project because later we will prove that optimal densities for some optimization problems turn out to be exactly these types of probability

distributions, especially when we are using Kullback-Leibler divergence to measure distance between probability distributions.

In a second goal in this project we focus on solving minimum distance problems for different distance measures. More specifically, we address the problem of minimum circle problem, i.e., the Chebyshev center and radius problem, for geometric points as well as for probability distributions. One can consider the probability distributions as points on the n-dimensional space (probability simplex) and try to find the geometric locus of the distributions which satisfies desired equalities. In fact, this constitutes our main objective: solving Smallest Enclosing Ball Problem under different distance functions. We will try to solve them analytically in general and also find numerical solutions whenever we fail to find analytical solutions. We will use Euclidean Metric for probability distributions to Generalize Apollonius Circle for Probability distributions. We can also analyze some special distributions types (Gauss, Binomial, Poisson etc.) and the features of them under certain metrics.

As an important byproduct of this project, we learnt substantially about the mathematical tools as convex optimization, information theory, distance metrics and their properties, geodesics in finite spaces, locus points of solutions of some problem.

3. Preliminaries

3.1 *f-Divergence*

Let P and Q be two probability distributions over a space Ω such that P is absolutely continous with respect to Q. Then for a convex function f such that $f(1) = 0$, the $f - Divergence$ of P from Q defined as

$$D_f(P||Q) \coloneqq \int_\Omega f(\frac{dP}{dQ})dQ$$

If P and Q are both absolutely continuous with respect to a reference distribution μ on Ω then their probability densities p and q satisfy $dP = p\, d\mu$ and $dQ = q\, d\mu$. In this case the f-divergence can be written as [2]

$$D_f(P||Q) = \int_\Omega f\left(\frac{p(x)}{q(x)}\right) q(x) d\mu(x)$$

If p and q are both discrete distributions then f-divergence of p from q is equal to

$$D_f(p||q) = \sum_{x \in X} f\left(\frac{p(x)}{q(x)}\right) q(x)$$

Instances of f-Divergence:
Many common divergences, such as KL-divergence, Hellinger distance, and total variation distance, are special cases of *f*-divergence, coinciding with a particular choice of *f*. The following table lists many of the common divergences between probability distributions and the *f* function to which they correspond

Divergence	**Corresponding $f(t)$**
KL-divergence	$tlogt$
Reverse KL-divergence	$-logt$
squared Hellinger distance	$\left(\sqrt{t}-1\right)^2, 2(1-\sqrt{t})$
Total Variation Distance	$\frac{1}{2}\lvert t-1\rvert$
Pearson χ^2-Divergence	$(t-1)^2, t^2-1, t^2-t$
Neyman χ^2-Divergence (reverse Pearson)	$\frac{1}{t}-1, \frac{1}{t}-t$
α −divergence	$\frac{4}{1-\alpha^2}(1-t^{(1+\alpha)/2})$
α −divergence (other designation)	$\frac{t^\alpha - t}{\alpha(\alpha-1)}$

Properties of $f - divergence$:

- **Non-negativity:** the f-divergence is always positive; it's zero if and only if the measures P and Q coincide. This follows immediately from Jensen's inequality:

$$D_f(P||Q) = \int f\left(\frac{dP}{dQ}\right) dQ \geq f\left(\int \frac{dP}{dQ} dQ\right) = f(1) = 0$$

- **Monotonicty:** if κ is an arbitrary transition probability that transforms P and Q into P_κ and Q_κ correspondingly, then we have:

$$D_f(P||Q) \geq D_f(P_\kappa||Q_\kappa)$$

The equality here holds if and only if the transition is induced from a sufficient statistic with respect to $\{P, Q\}$

- **Joint Convexity:** for any $0 \leq \lambda \leq 1$

  $D_f(\lambda P_1 + (1-\lambda)P_2 || \lambda Q_1 + (1-\lambda)Q_2) \leq \lambda D_f(P_1||Q_1) + +(1-\lambda)D_f(P_2||Q_2)$

  This follows from the convexity of the mapping $(p, q) \to qf(\frac{p}{q})$ on $\mathbb{R}_+^2$
- **Symmetry:** if for some function f induces an $f - divergence\ D_f(P||Q)$ then $D_f(Q||P)$ is also an $f - divergence$ of P from Q. Because if we choose the convex function $g(t) = tf(\frac{1}{t})$ as divergence function, then we have:

$$D_f(Q||P) = D_g(P||Q)$$

3.2 *Chebyshev Radius-Chebyshev Center*

Chebyshev Radius of a bounded set M in metric space (X, d) if

$$r = \inf_{x \in X} \sup_{y \in M} d(x, y)$$

then we call that r, Chebyshev Radius of the set M.

If for some point $x_0 \in X$ satisfies:

$$\sup_{y \in M} d(x_0, y) = r$$

then we call that point x_0 a Chebyshev Center [3].

Also we can state the Chebyshev Center as $\arg\min_{x \in X}\{\max_{y \in M} d(x, y)\}$. If a normed linear space is dual to some normed linear space, then any bounded set M has at least one Chebyshev center.

3.3 Continuous Optimization Problem

The standard form of a continuous optimization problem [4] is

$$\textbf{minimize}_{x}\ f(x)$$

subject to $g_i(x) \leq 0,\ i = 1, \ldots, m$ and $h_j(x) = 0, j = 1, \ldots, p$ where

- $f: \mathbb{R}^n \to \mathbb{R}^n$ is the objective function to be minimized over the n-variable vector x,
- $g_i(x) \leq 0$ are called inequality constraints
- $h_i(x) = 0$ are called equality constraints, and
- $m \geq 0$ and $p \geq 0$

If $m = p = 0$, the problem is an unconstrained optimization problem. By convention, the standard form defines a **minimization problem**.
A **maximization problem** can be treated by negating the objective function.

3.4 KKT Conditions for Optimization Problems:

Consider the following nonlinear optimization problem:

$$\textbf{optimize}_{\boldsymbol{x}}\ f(\boldsymbol{x})$$
$$\textbf{subject to}\ g_i(\boldsymbol{x}) \leq 0 \text{ and } h_j(\boldsymbol{x}) = 0$$

where $\boldsymbol{x} \in \boldsymbol{X}$ is the optimization variable chosen from a convex subset of $\mathbb{R}^n$,
f is the objective or utility function,
$g_i(i = 1, \ldots, m)$ are the inequality constraint functions
and $h_j(j = 1, \ldots, p)$ are the equality constraint functions.
The number of inequalities and equalities are denoted by m and p respectively.
Corresponding to the constraint optimization problem one can form the Lagrangian function:

$$L(\boldsymbol{x}, \boldsymbol{\mu}, \lambda) = f(x) + \sum_{i=1}^{m} \mu_i g_i(\boldsymbol{x}) + \sum_{j=1}^{p} \lambda_i h_i(\boldsymbol{x}).$$

The **Karush-Kuhn-Tucker Theorem** states the following [5], [6].

Theorem: If $(\boldsymbol{x}^*, \boldsymbol{\mu}^*)$ is a saddle point of $L(\boldsymbol{x}, \boldsymbol{\mu})$ in $\boldsymbol{x} \in \boldsymbol{X}, \boldsymbol{\mu} \geq \boldsymbol{0}$, then $\boldsymbol{x}^*$ is an optimal vector for the above optimization problem. Suppose that $f(\boldsymbol{x})$ and $g_i(\boldsymbol{x}), i = 1, \ldots, m$ are concave in $\boldsymbol{x}$ and that there exists $\boldsymbol{x_0} \in \boldsymbol{X}$ such that $\boldsymbol{g(x_0)} > \boldsymbol{0}$. Then with an optimal vector $\boldsymbol{x}^*$ for the above optimization problem there is associated a non-negative vector $\boldsymbol{\mu}^*$ such that $(\boldsymbol{x}^*, \boldsymbol{\mu}^*)$ is a saddle point of $L(\boldsymbol{x}, \boldsymbol{\mu})$.

Necessary Conditions:

1. **Stationarity**

   For maximizing $f(\boldsymbol{x})$:
   $\nabla f(\boldsymbol{x}^*) - \sum_{i=1}^{m} \mu_i g_i(\boldsymbol{x}^*) - \sum_{j=1}^{p} \lambda_i h_i(\boldsymbol{x}^*) = 0$
   For minimizing $f(\boldsymbol{x})$:

$\nabla f(\boldsymbol{x}^*) + \sum_{i=1}^{m} \mu_i g_i(\boldsymbol{x}^*) + \sum_{j=1}^{p} \lambda_i h_i(\boldsymbol{x}^*) = 0$

2. **Primal feasibility**

   $g_i(\boldsymbol{x}^*) \leq 0$, for $i = 1, \dots, m$
   $h_j(\boldsymbol{x}^*) = 0$, for $j = 1, \dots, p$

3. **Dual feasibility**

   $\mu_i \geq 0$, for $i = 1, \dots, m$

4. **Coplementary Slackness**

   $\mu_i g_i(\boldsymbol{x}^*) = 0$, for $i = 1, \dots, m$

In the particular case $m = 0$, i.e., when there are no inequality constraints, the KKT conditions turn into the Lagrange conditions, and the KKT multipliers are called Lagrange multipliers.
If some of the functions are non-differentiable, subdifferential versions of Karush–Kuhn–Tucker (KKT) conditions are available [7].

3.5 CVX: Matlab Software for Disciplined Convex Programming

CVX is a Matlab-based modeling system for convex optimization. CVX turns Matlab into a modeling language, allowing constraints and objectives to be specified using standard Matlab expression syntax [8].

3.6 von Neumann's Minimax Theorem

A minimax theorem is a theorem providing conditions that guarantee that the max–min inequality is also an equality. The first theorem in this sense is von Neumann's minimax theorem from 1928, which was considered the starting point of game theory. Since then, several generalizations and alternative versions of von Neumann's original theorem have appeared in the literature.
The minimax theorem was first proven and published in 1928 by John von Neumann, who is quoted as saying *"As far as I can see, there could be no theory of games … without that theorem … I thought there was nothing worth publishing until the Minimax Theorem was proved"*.
Formally, von Neumann's minimax theorem states [9]:

Let $\boldsymbol{X} \subset \mathbb{R}^n$ and $\boldsymbol{Y} \subset \mathbb{R}^m$ be compact convex sets. If $f: \boldsymbol{X} \times \boldsymbol{Y} \to \mathbb{R}$ is a continuous function that is concave-convex, i.e.
$\boldsymbol{f}(\cdot, \boldsymbol{y})$: $\boldsymbol{X} \to \mathbb{R}$ is concave for fixed $\boldsymbol{y}$, and
$\boldsymbol{f}(\boldsymbol{x}, \cdot)$: $\boldsymbol{Y} \to \mathbb{R}$ is concave for fixed $\boldsymbol{x}$.
Then we have that:

$$\max_{x \in X} \min_{y \in Y} f(x, y) = \min_{y \in Y} \max_{x \in X} f(x, y)$$

3.7 Convex Hull

A set of points in a Euclidean space is defined to be convex if it contains the line segments connecting each pair of its points. The convex hull of a given set $\boldsymbol{X}$ may be defined as [11]

1. The (unique) minimal convex set containing **X**
2. The intersection of all convex sets containing **X**
3. The set of all convex combinations of points in **X**
4. The union of all simplices with vertices in **X**

3.8 Bregman Divergence

Let $F: \boldsymbol{\Omega} \to \mathbb{R}$ be a continuously differentiable, strictly convex function defined on a closed convex set $\boldsymbol{\Omega}$.
The Bregman distance associated with F for points, $\boldsymbol{p}, \boldsymbol{q} \in \boldsymbol{\Omega}$ is the difference between the value F at point $\boldsymbol{p}$ and the value of the first order Taylor Expansion of F around point $\boldsymbol{q}$ evaluated at point $\boldsymbol{p}$: [12], [13]

$$D_F(p, q) = F(p) - F(q) - \langle \nabla F(q), p - q \rangle$$

Properties:

- **Non-negativity**: $D_F(p, q) \geq 0$ for all p, q. This is a consequence of the convexity of F.

- **Convexity**: $D_F(p, q)$ is convex in its first argument, but not necessarily in the second argument [14].

- **Linearity**: If we think of the Bregman distance as an operator on the function F, then it is linear with respect to non-negative coefficients. In other words, for F_1, F_2 strictly convex and differentiable, and $\lambda_1, \lambda_2 \geq 0$,

$$D_{\lambda_1 F_1 + \lambda_2 F_2}(p, q) = \lambda_1 D_{F_1}(p, q) + \lambda_2 D_{F_2}(p, q)$$

for all $p, q \in \boldsymbol{\Omega}$

- **Duality**: The function F has a convex conjugate F^*. The Bregman distance defined with respect to F^* has an interesting relationship to $D_F(p, q)$ which is:

$$D_{F^*}(p^*, q^*) = D_F(q, p)$$

Here, $p^* = \nabla F(q)$ and $q^* = \nabla F(q)$ are the dual points corresponding to p and q.

- **Mean as minimizer**: A key result about Bregman divergences is that, given a random vector, the mean vector minimizes the expected Bregman divergence from the random vector. This result generalizes the textbook result that the mean of a set minimizes total squared error to elements in the set. This result was proved for the vector case by (Banerjee et al. 2005), and extended to the case of functions/distributions by (Frigyik et al. 2008). This result is important because it further justifies using a mean as a representative of a random set, particularly in Bayesian estimation.

Examples:

- Squared Euclidean Distance $D_F(x, y) = \|x - y\|^2$ is the canonical example of a Bregman distance, generated by the convex function:

$$F(x) = \|x\|^2$$

- The squared Mahalonbais distance, $D_F(x, y) = \frac{1}{2}(x - y)^T Q(x - y)$ which is generated by the convex function $F(x) = \frac{1}{2}x^T Q(x)x$. This can be thought of as a generalization of the above squared Euclidean distance.

- The generalized Kullback-Leibler divergence:

$$D_F(p, q) = \sum_i p(i) \log \frac{p(i)}{q(i)} - \sum_i p(i) + \sum_i q(i)$$

is generated by the negative entropy function:

$$F(p) = \sum_i p(i) \log p(i)$$

- The *Itakura-Saito* distance,

$$D_F(p, q) = \sum_i \left(\frac{p(i)}{q(i)} - \log \frac{p(i)}{q(i)} - 1 \right)$$

is generated by the convex function

$$F(p) = -\sum_i \log p(i)$$

3.9 Proximal Point Method:

Let f be a closed convex function defined on some closed convex region $X \subset \mathbb{R}^n$
To minimize f over X we choose some starting point $x_0 \in X$ then we iteratively choose other points in such manner [15]:

$$x_{k+1} = prox_{t_k} f(x_k) = \arg\min_{u \in X} \left(f(u) + \frac{1}{2t_k} \|u - x_k\|_2^2 \right)$$

For $k \geq 0$ and here the $t_k > 0$ is the *step size* which we choose
We will show the sequence of points $\{x_n\}_{n \geq 0}$ is converging to global minimum if we choose step sizes correctly. Here the advantage of this method comes when if the $prox$ evaluations are much easier than minimizing f directly. We can see that $prox$ function is convex for any value of $t_k > 0$ because the f is convex itself.

4. Approach and Methodology

In general we will consider probability distributions with finite sample space. When we trying to find closest probability distributions to some finitely many given probability distributions, we are trying to extrema points of distance functions in a convex region. Therefore we are generally trying to solve convex optimization problems. To solve this problems our most powerful tool is Lagrange Multipliers Method and it's generalization with KKT conditions. We are also using some tool as CVX to implement our works analytically on MATLAB.

5. Results and Accomplishments So Far

So far, we have been studying Smallest Enclosing Ball problem from many different point of views. Before studying about this problem, we learned about different distance measurements for probability distributions. We studied about $f-divergences$ and their properties. Then we continued with simple cases of the problem. The most simple case of the problem is plane version where the problem simple called Smallest Circle problem. Then we tried to solve further versions and tried to find robust algorithms that solves the problem.

The Smallest Enclosing Ball Problem:

Let $\boldsymbol{X} \subset \mathbb{R}^n$ be a set consisting of the points $x_1, x_2, \dots, x_m$.
Let $d: \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^+$ be our distance function.
We call $r = \min\limits_{\boldsymbol{y}\in\mathbb{R}^n} \max\limits_{\boldsymbol{x}\in\boldsymbol{X}} d(\boldsymbol{x},\boldsymbol{y})$ the Chebyshev Radius of the set $\boldsymbol{X}$.
We call $x^* = \arg\min\limits_{y\in\mathbb{R}^n} \max\limits_{x\in X} d(x,y)$ the Chebyshev Center of the set X [16].
The smallest Enclosing Ball problem is finding the Chebyshev Center $\boldsymbol{x}^*$ and Chebyshev Radius r.

Our main objective for this project is solving Smallest Enclosing Ball Problem with using different distance functions. We are trying to solve it theoretically if possible or we try to find analytical algorithms to approximately find them. We are searching for different algorithms that satisfies our objectives. After dealing with many distance functions we focused to develop a robust algorithm that solves Smallest Enclosing Ball Problem using Euclidean Metric.

The Smallest Enclosing Ball Problem in $\mathbb{R}^2$:

We started studying about "Smallest Enclosing Ball" problem in two dimensional Euclidean space as an introduction to general Smallest Enclosing Ball problem [17]. Dealing with this special case first is rather easy and gave us intuition about the general problem. This special case often called as "Smallest Circle" problem. We proved the existence and uniqueness of the Chebyshev Center which is the name of the center of the Smallest Enclosing Ball. We proved the following lemma:

Lemma:
The minimum covering circle of a set *S* can be determined by at most three points in *S* which lie on the boundary of the circle. If it is determined by only two points, then the line segment joining those two points must be a diameter of the minimum circle. If it is determined by three points, then the triangle consisting of those three points is not obtuse [18].

Direct Approach:
We studied about "Smallest Enclosing Ball" problem in $\mathbb{R}^n$. Firstly, we studied about the algorithms that uses the upper Lemma to find Chebyshev Center. We learned about Welzl's Algorithm [19] which solves the problem in $O(m)$ complexity where m is the points in our set. Welzl's Algorithm uses Linear Programming to find the Chebyshev Center.

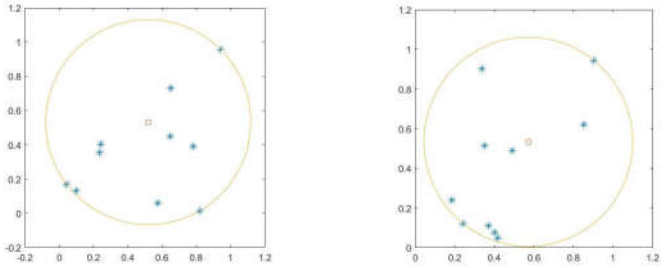
The Smallest Enclosing Ball Problem in $\mathbb{R}^n$:

After dealing with Smallest Circle problem we started to work on general version of it. We started to search for other type of algorithms to solve this more general problem.

Optimization Approach:

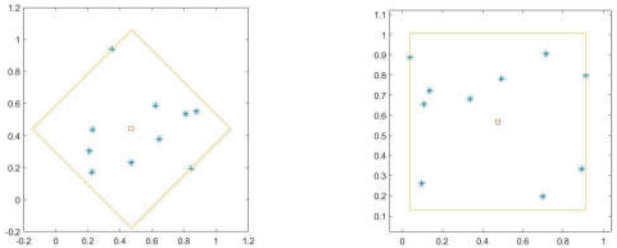
After a while we studied a different approach which is finding a dual problem and using convex optimization to find Chebyshev Center. To manage this, we started learning about Convex Optimization Methods, Duality Concept, and KKT conditions. Then we found a dual to Smallest Enclosing Ball problem. We

gave a detailed explanation for dual problem in Appendix. Then we started to think about solving the optimization problem in MATLAB. We started to learn about CVX complex optimization program to solve our problem. We got fruitful results with this program.



Some Chebyshev Circle Examples found with CVX

Also we experimented with other metrics. We tried to find the Chebyshev Center with ℓ_1 and ℓ_∞ metrics. Again we used CVX program to manage this.



Chebyshev Circle using ℓ_1 norm

Chebyshev Circle using ℓ_∞ norm

The Smallest Enclosing Ball Problem for Gaussian Distributions:

Our project's main objective is solving the Smallest Enclosing Ball Problem but solving this problem for probability distributions is our main focus. So we tried to solve this problem for given Gaussian Distributions. But here the problem is, we can not think the Gaussian Distributions as points directly or else they needed to be points of a space with infinite dimensions. So we started learning about Bregman Divergences and then we learned that the Smallest Enclosing Ball Problem can be generalized for Bregman Divergences. Then using KL-distance we managed to think to Gaussian Distributions as points therefore we can find their Chebyshev Center. We actually found similar finding when dealing with Lemma 1 and later on we discovered a paper [20] which solves this problem using Bregman Divergences.

6. Expected Results and Accomplishments

We expect to find robust algorithms that solves Smallest Enclosing Ball Problem with Euclidean Distance as distance measure. We are trying to develop a descent algorithm that converges to Chebyshev Center. We will learn more about Convex Optimization methods to make our algorithm more robust. Our focus is implementing a version of Proximal Point Method equipped with KL-distance as distance measure for error function in proximal point method. We consider KL-distance here as a suitable distance measure because we observed that our dual problem is actually equivalent to making convex optimization on probability simplex. We will learn more about different algorithms that solves Smallest Enclosing Ball Problem to increase our vision.

$$\sum_{i=1}^{m} \lambda_i \|\boldsymbol{x_i} - \boldsymbol{y}\|^2$$

Proposed Method

Our suggested method tries to solve the dual problem iteratively. This method is a modification of proximal point method equipped with KL-distance instead of Euclidean Distance. The dual problem becomes maximizing:

$$\sum_{i=1}^{m} \lambda_i \|\boldsymbol{x_i} - \boldsymbol{x}\|^2$$

In terms of the nonnegative vector $\boldsymbol{\lambda}$ with $\|\boldsymbol{\lambda}\|_1 = 1$ and

$$\boldsymbol{x} = \sum_{i=1}^{m} \lambda_i \boldsymbol{x_i}$$

But this task is rather difficult so instead we developed the following algorithm. We start with an initial vector $\boldsymbol{\lambda^{(0)}}$ and in each step k we try to maximize the new objective function in order to get next coefficient vector:

$$\boldsymbol{\lambda^{(k+1)}} = \arg\max_{\boldsymbol{\lambda \in \mathcal{P}^m}} \sum_{i=1}^{M} \lambda_i \left\| \boldsymbol{x_i} - \boldsymbol{x^{(k)}} \right\|^2 - \mu_k^{-1} KL(\boldsymbol{\lambda} || \boldsymbol{\lambda^{(k)}})$$

Where $\boldsymbol{\lambda^{(k)}}$ is the coefficient vector in k-th step and $\boldsymbol{x^{(k)}} = \boldsymbol{A\lambda^{(k)}}$.
Recall that:

$$A = [x_1\, x_2 \ldots x_m]$$

is the matrix of the points $\boldsymbol{x_i}$ for $i = 1,2,\ldots,m$

Also the μ_k is the k-th step size which will be determined in order to make the sequence of vector $\boldsymbol{\lambda^{(k)}}$ converge to the optimal vector $\boldsymbol{\lambda^*}$

This approach makes easier to determine next coefficient vector because it is easier to find

$$\arg\max_{\boldsymbol{\lambda \in \mathcal{P}^m}} \sum_{i=1}^{M} \lambda_i \left\| \boldsymbol{x_i} - \boldsymbol{x^{(k)}} \right\|^2 - \mu_k^{-1} KL(\boldsymbol{\lambda} || \boldsymbol{\lambda^{(k)}})$$

Then this dual problem.

$$\arg\max_{\boldsymbol{\lambda \in \mathcal{P}^m}} \sum_{i=1}^{M} \lambda_i \left\| \boldsymbol{x_i} - \boldsymbol{x} \right\|^2$$

Let's show how to find next coefficient vector analytically.

Determining Next Coefficient Vector in Proposed Method:

For $k \geq 0$ we have:

$$\boldsymbol{\lambda^{(k+1)}} = \arg\max_{\boldsymbol{\lambda \in \mathcal{P}^m}} \sum_{i=1}^{M} \lambda_i \left\| \boldsymbol{x_i} - \boldsymbol{x^{(k)}} \right\|^2 - \mu_k^{-1} KL(\boldsymbol{\lambda} || \boldsymbol{\lambda^{(k)}})$$

Where $\boldsymbol{x^{(k)}} = \boldsymbol{A\lambda^{(k)}}$

If we take the gradient of the desired function to maximize, we get:

$$\left\| \boldsymbol{x_i} - \boldsymbol{x^{(k)}} \right\|^2 - \mu_k^{-1} \log \frac{\lambda_i^{(k+1)}}{\lambda_i^{(k)}} - \mu_k^{-1} \lambda_i^{(k)} = constant$$

For all $i = 1,2,\ldots,m$

From this equation we can see the relation:

$$\lambda_i^{(k+1)} \propto \lambda_i^{(k)} \exp\left(\mu_k \left\| \boldsymbol{x_i} - \boldsymbol{x^{(k)}} \right\|^2\right)$$

Then we can derive the following equation:

$$\lambda_i^{(k+1)} = \lambda_i^{(k)} \frac{\exp\left(\mu_k \left\| \boldsymbol{x_i} - \boldsymbol{x^{(k)}} \right\|^2\right)}{\sum_{j=1}^{m} \lambda_j^{(k)} \exp\left(\mu_k \left\| \boldsymbol{x_j} - \boldsymbol{x^{(k)}} \right\|^2\right)}$$

This concludes the calculation of the next coefficient vector. According to these calculations we can summarize our algorithm as follows:

Algorithm 1: Proposed Method

Input: $A = [x_1\, x_2 \ldots x_M]$, μ, maxiter (maximum number of iterations),

Output: x^* (Chebyshev center), ρ (Chebyshev radius)

1 Initial Conditions: $k = 0$, $\boldsymbol{\lambda^{(0)}} = \frac{1}{M} \times [1\ 1 \ldots 1]$

2 $\boldsymbol{x^{(k)}} = \boldsymbol{A\lambda^{(0)}}$

3 $D_i^{(k)} = \left\| x_i - \boldsymbol{x^{(k)}} \right\|^2$

4 $\lambda_i^{(k+1)} = \lambda_i^{(k)} \frac{\exp\left(\mu_k D_i^{(k)}\right)}{\sum_{j=1}^{m} \lambda_j^{(k)} \exp\left(\mu_k D_j^{(k)}\right)}$

5 $k \leftarrow k + 1$

6 **if** $k < maxiter$, go to step-2

7 Return $\boldsymbol{x^*} = \boldsymbol{x^{(k)}}$, $\rho = \sum_{i=1}^{M} \lambda_i D_i^{(k)}$

Step Size Selection for Monotonic Convergence

As we described before we rewrite the minimax problem as follows:

$$\boldsymbol{\lambda}^{(\boldsymbol{k+1})} = \arg\max_{\boldsymbol{\lambda}\in\mathcal{P}^{\boldsymbol{m}}} \sum_{i=1}^{M} \lambda_i \left\|\boldsymbol{x_i} - \boldsymbol{x}^{(\boldsymbol{k})}\right\|^2 - \mu_k^{-1} KL(\boldsymbol{\lambda}|\boldsymbol{\lambda}^{(\boldsymbol{k})}) = \arg\max_{\boldsymbol{\lambda}\in\mathcal{P}^{\boldsymbol{m}}} \tilde{g}(\boldsymbol{\lambda}, \boldsymbol{x}^{(\boldsymbol{k})})$$

Here we introduce a new modified objective function:

$$\tilde{g}(\boldsymbol{\lambda}, \boldsymbol{x}^{(\boldsymbol{k})}) = \sum_{i=1}^{M} \lambda_i \left\|\boldsymbol{x_i} - \boldsymbol{x}^{(\boldsymbol{k})}\right\|^2 - \mu_k^{-1} KL(\boldsymbol{\lambda}|\boldsymbol{\lambda}^{(\boldsymbol{k})})$$

Instead of the following main objective function:

$$g(\boldsymbol{\lambda}) = \sum_{i=1}^{M} \lambda_i \left\|\boldsymbol{x_i} - \boldsymbol{x}^{(\boldsymbol{k})}\right\|^2$$

The primal variable $\boldsymbol{x}^{(\boldsymbol{k})}$ appearing in the Lagrange dual function:

$$g(\boldsymbol{\lambda}^{(\boldsymbol{k})}) = \sum_{i=1}^{M} \lambda_i^{(k)} \left\|\boldsymbol{x_i} - \boldsymbol{x}^{(\boldsymbol{k})}\right\|^2$$

is the primal variable induced by $\boldsymbol{\lambda}^{(\boldsymbol{k})}$. Hence, $g(\boldsymbol{\lambda}^{(\boldsymbol{k})})$ is a function of a single variable, as the notation suggests. Yet, the modified objective function $\tilde{g}(\boldsymbol{\lambda}, \boldsymbol{x}^{(\boldsymbol{k})})$ is a function of two variables. It is easy to see that when $\boldsymbol{x}^{(\boldsymbol{k})}$ is induced by $\boldsymbol{\lambda}^{(\boldsymbol{k})}$, the modified objective function is identical to the dual function, $\tilde{g}(\boldsymbol{\lambda}^{(\boldsymbol{k})}, \boldsymbol{x}^{(\boldsymbol{k})}) = g(\boldsymbol{\lambda}^{(\boldsymbol{k})})$

Considering upper equality, since $\boldsymbol{\lambda}^{(\boldsymbol{k+1})}$ is the maximizer of the optimization problem, we have the trivial fact that:

$$\tilde{g}(\boldsymbol{\lambda}^{(\boldsymbol{k+1})}, \boldsymbol{x}^{(\boldsymbol{k})}) \geq \tilde{g}(\boldsymbol{\lambda}^{(\boldsymbol{k+1})}, \boldsymbol{x}^{(\boldsymbol{k})}) = g(\boldsymbol{\lambda}^{(\boldsymbol{k})})$$

Our main goal in this section is to express $\tilde{g}(\boldsymbol{\lambda}^{(\boldsymbol{k+1})}, \boldsymbol{x}^{(\boldsymbol{k})})$ (left side of the inequality) in terms of dual function $g(\boldsymbol{\lambda}^{(\boldsymbol{k+1})})$ and other terms in order to provide a converge guarantee.

The term on the left side of the inequality $\tilde{g}(\boldsymbol{\lambda}^{(\boldsymbol{k+1})}, \boldsymbol{x}^{(\boldsymbol{k})}) \geq g(\boldsymbol{\lambda}^{(\boldsymbol{k})})$ can be expressed as:

$$\tilde{g}(\boldsymbol{\lambda}^{(\boldsymbol{k+1})}, \boldsymbol{x}^{(\boldsymbol{k})}) = \sum_{i=1}^{M} \lambda_i^{(k+1)} \left\|\boldsymbol{x_i} - \boldsymbol{x}^{(\boldsymbol{k})}\right\|^2 - \mu_k^{-1} KL\,(\boldsymbol{\lambda}^{(\boldsymbol{k+1})}||\boldsymbol{\lambda}^{(\boldsymbol{k})})$$

$$\overset{(a)}{=} \sum_{i=1}^{M} \lambda_i^{(k+1)} \left\|\boldsymbol{x_i} - \boldsymbol{x}^{(\boldsymbol{k+1})} + \Delta\boldsymbol{x}\right\|^2 - \mu_k^{-1} KL\,(\boldsymbol{\lambda}^{(\boldsymbol{k+1})}||\boldsymbol{\lambda}^{(\boldsymbol{k})})$$

$$\overset{(b)}{=} \textstyle\sum_{i=1}^{M} \lambda_i^{(k+1)} \left(\left\|\boldsymbol{x_i} - \boldsymbol{x}^{(\boldsymbol{k+1})}\right\|^2 + \|\Delta\boldsymbol{x}\|^2 + 2\left(\boldsymbol{x_i} - \boldsymbol{x}^{(\boldsymbol{k+1})}\right)^{\boldsymbol{T}} \Delta\boldsymbol{x}\right) - \mu_k^{-1} KL\,(\boldsymbol{\lambda}^{(\boldsymbol{k+1})}||\boldsymbol{\lambda}^{(\boldsymbol{k})})$$

$$\overset{(c)}{=} g(\boldsymbol{\lambda}^{(\boldsymbol{k+1})}) + \|\Delta\boldsymbol{x}\|^2 + 2\left(\sum_{i=1}^{M} \lambda_i^{(k+1)} \left(\boldsymbol{x_i} - \boldsymbol{x}^{(\boldsymbol{k+1})}\right)\right)^{T} \Delta\boldsymbol{x} - \mu_k^{-1} KL\,(\boldsymbol{\lambda}^{(\boldsymbol{k+1})}||\boldsymbol{\lambda}^{(\boldsymbol{k})})$$

$$\overset{(d)}{=} g(\boldsymbol{\lambda}^{(\boldsymbol{k+1})}) + \|\Delta\boldsymbol{x}\|^2 - \mu_k^{-1} KL\,(\boldsymbol{\lambda}^{(\boldsymbol{k+1})}||\boldsymbol{\lambda}^{(\boldsymbol{k})})$$

In line-(a) $\Delta\boldsymbol{x} = \boldsymbol{x}^{(\boldsymbol{k+1})} - \boldsymbol{x}^{(\boldsymbol{k})}$ is introduced. Line-(b) is the expansion of the quadratic term. In line- (c), the first of the sum is recognized as the dual function evaluated at $\boldsymbol{\lambda}^{(\boldsymbol{k+1})}$. In line-(d), definition of $\boldsymbol{x}^{(\boldsymbol{k+1})}$ from the definition $\boldsymbol{x}^{(\boldsymbol{k+1})} = \boldsymbol{A}\boldsymbol{\lambda}^{(\boldsymbol{k+1})}$ is used to show that the bracketed term in line-(c) is equal to zero vector. When the inequality of $\tilde{g}(\boldsymbol{\lambda}^{(\boldsymbol{k+1})}, \boldsymbol{x}^{(\boldsymbol{k})}) \geq g(\boldsymbol{\lambda}^{(\boldsymbol{k})})$ is combined with the equation in line-(d), we get:

$$g(\boldsymbol{\lambda}^{(\boldsymbol{k+1})}) \geq g(\boldsymbol{\lambda}^{(\boldsymbol{k})}) + \mu_k^{-1} KL\,(\boldsymbol{\lambda}^{(\boldsymbol{k+1})}|\boldsymbol{\lambda}^{(\boldsymbol{k})}) - \left\|\boldsymbol{x}^{(\boldsymbol{k+1})} - \boldsymbol{x}^{(\boldsymbol{k})}\right\|^2$$

If the term denoted as margin in the equation can be guaranteed to be positive, the value of the dual optimization problem is guaranteed to increase at every update cycle, that is $g(\boldsymbol{\lambda}^{(\boldsymbol{k+1})}) \geq g(\boldsymbol{\lambda}^{(\boldsymbol{k})})$

Next, we present a choice of μ guaranteeing the non-negativity of the margin. The margin is guaranteed to be non-negative if

$$\mu_k \leq \frac{KL\,(\boldsymbol{\lambda}^{(\boldsymbol{k+1})}||\boldsymbol{\lambda}^{(\boldsymbol{k})})}{\|\boldsymbol{x}^{(\boldsymbol{k+1})} - \boldsymbol{x}^{(\boldsymbol{k})}\|^2} \triangleq \bar{\mu}$$

We give a lower bound for $\bar{\mu}$ by optimizing over four variables $\boldsymbol{\lambda}^{(\boldsymbol{k+1})}, \boldsymbol{\lambda}^{(\boldsymbol{k})}, \boldsymbol{x}^{(\boldsymbol{k+1})}, \boldsymbol{x}^{(\boldsymbol{k})}$ and suggest to select a step-size μ smaller than this lower bound:

$$\min_{\boldsymbol{\lambda}^{(k+1)},\boldsymbol{\lambda}^{(k)},\boldsymbol{x}^{(k+1)},\boldsymbol{x}^{(k)}} \bar{\mu} \overset{(a)}{\geq} \frac{\min_{\boldsymbol{\lambda}^{(k+1)},\boldsymbol{\lambda}^{(k)}} KL\left(\boldsymbol{\lambda}^{(k+1)} || \boldsymbol{\lambda}^{(k)}\right)}{\max_{\boldsymbol{x}^{(k+1)},\boldsymbol{x}^{(k)}} \|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\|^2}$$

$$\overset{(b)}{\geq} \frac{\min_{\boldsymbol{\lambda}^{(k+1)},\boldsymbol{\lambda}^{(k)}} \frac{1}{2}\|\boldsymbol{\lambda}^{(k+1)} - \boldsymbol{\lambda}^{(k)}\|^2}{\max_{\boldsymbol{x}^{(k+1)},\boldsymbol{x}^{(k)}} \|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\|^2}$$

$$\overset{(c)}{\geq} \frac{\min_{\boldsymbol{\lambda}^{(k+1)},\boldsymbol{\lambda}^{(k)}} \frac{1}{2}\|\boldsymbol{\lambda}^{(k+1)} - \boldsymbol{\lambda}^{(k)}\|^2}{\max_{\boldsymbol{\lambda}^{(k+1)},\boldsymbol{\lambda}^{(k)}} (\boldsymbol{\lambda}^{(k+1)} - \boldsymbol{\lambda}^{(k)})^T A^T A (\boldsymbol{\lambda}^{(k+1)} - \boldsymbol{\lambda}^{(k)})^2}$$

$$\overset{(d)}{\geq} \frac{\min_{\boldsymbol{\lambda}^{(k+1)},\boldsymbol{\lambda}^{(k)}} \frac{1}{2}\|\boldsymbol{\lambda}^{(k+1)} - \boldsymbol{\lambda}^{(k)}\|^2}{\max_{\boldsymbol{\lambda}^{(k+1)},\boldsymbol{\lambda}^{(k)}} \frac{1}{2}\|\boldsymbol{\lambda}^{(k+1)} - \boldsymbol{\lambda}^{(k)}\|_2^2 \, \lambda_{max}(A^T A)^2}$$

$$= \frac{1}{2\lambda_{max}(A^T A)}$$

Line-(a) is due to independent minimization and maximization of numerator and denominator, respectively. Line-(b) is due to Pinsker's inequality, [5, p.370] where $||\cdot||_1$ refers to the L_1 norm. Also in this line, $\boldsymbol{x}^{(k)}$ vectors in the denominator are recognized to be induced from $\boldsymbol{\lambda}^{(k)}$vectors via the relation we found earlier. Line-(c) is due to the fact that $||\cdot||_2 \leq ||\cdot||_1$. In line-(d), the maximum eigenvalue of $A^T A$ is expressed as $\lambda_{max}(A^T A)$.

Hence, the choice of step-size μ smaller than $\frac{1}{2\lambda_{max}(A^T A)}$ guarantees the monotonic convergence of the iterations to the dual problem optima. It should be observed that this choice of μ guaranteing convergence is highly pessimistic (due to several bounding arguments), that is much higher step-size values in comparison to $\mu = \frac{1}{2\lambda_{max}(A^T A)}$ can provide a similar monotonic ascent of dual objective value. An adaptive step-size selection method is given in Numerical Results section to illustrate this opportunity.

7. Realistic Constraints

The time dedicated to acquire the theoretical background may curtail the time I can dedicate to the analysis part because before getting into advancing the project I need to increase my theoretical and operational knowledge in mathematical areas as Probability Theory, Information Theory and Optimization Theory.

a. Social, Enviornmental and Economic Impact

The project can be used at any domain consisting of a time series signal since it has no tight bound with specific applications. Although the proposed method can be used to making deductions with multiple information sources which give information about the same event in terms of probability distributions. Our project can find the most occasional event according to supplied probabilistic data.

b. Cost Analysis

The Project mainly based on theoretical thinking and mathematical proofs, it does not require much technological power so it does have very low cost. Altough we will try to numerically solve some of the problems so a moderate computer that can handle programs as MATLAB will be sufficient for our Project.
Since this project is a senior design project done by myself, an appropriate salary would be something close to a fresh graduate working at a company. Since I did my internship computer vision this summer, a full-time employee with master's degree and experience earns around 7000TL, so a fresh graduate might earn about 4000TL. But we can think this project process as a part-time job therefore 1500T would be a suitable salary.

c. Standards

We will follow mathematical rigor and correctness since our project mainly rely on mathematical proofs of our conjectures.

7. References

[1] Francis, R. L.; McGinnis, L. F.; White, J. A. (1992), Facility Layout and Location: An Analytical Approach (2nd ed.), Englewood Cliffs, N.J.: Prentice–Hall, Inc..

[2] Csiszár, I.; Shields, P. (2004). "Information Theory and Statistics: A Tutorial" (PDF). Foundations and Trends in Communications and Information Theory. 1 (4): 417–528. doi:10.1561/0100000004. Retrieved 2009-04-08.

[3] A. L. Garkavi, "The theory of best approximation in normed linear spaces", Itogi Nauki. Ser. Matematika. Mat. Anal. 1967, VINITI, Moscow (1969) 75–132; Progr. Math., 8 (1970) 83–150 Zbl 0258.41019

[4] Boyd, Stephen P.; Vandenberghe, Lieven (2004). Convex Optimization (pdf). Cambridge University Press. p. 129. ISBN 978-0-521-83378-3.

[5] Kuhn, H. W.; Tucker, A. W. (1951). "Nonlinear programming". Proceedings of 2nd Berkeley Symposium. Berkeley: University of California Press. pp. 481–492. MR 0047303.

[6] W. Karush (1939). Minima of Functions of Several Variables with Inequalities as Side Constraints (M.Sc. thesis). Dept. of Mathematics, Univ. of Chicago, Chicago, Illinois.

[7] Ruszczyński, Andrzej (2006). Nonlinear Optimization. Princeton, NJ: Princeton University Press. ISBN 978-0691119151. MR 2199043.

[8] M. Grant, "CVX: Matlab Software for Disciplined Convex Programming," CVX Research, Inc. [Online]. Available: http://cvxr.com/cvx/. [Accessed: 11-May-2020].

[9] Von Neumann, J. (1928). "Zur Theorie der Gesellschaftsspiele". Math. Ann. 100: 295–320. doi:10.1007/BF01448847.

[10] Rockafellar, R. Tyrrell (1970), Convex Analysis, Princeton Mathematical Series, 28, Princeton, N.J.: Princeton University Press, MR 0274683

[11] C. Candan, "Generation of Exponential Mixtures and Their Optimality Properties."

[12] Bregman, L. M. (1967). "The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming". USSR Computational Mathematics and Mathematical Physics. 7 (3): 200–217. doi:10.1016/0041-5553(67)90040-7.

[13] "Bregman divergence," Wikipedia, 13-Feb-2020. [Online]. Available: https://en.wikipedia.org/wiki/Bregman_divergence. [Accessed: 13-May-2020].

[14] "Joint and separate convexity of the Bregman Distance", by H. Bauschke and J. Borwein, in D. Butnariu, Y. Censor, and S. Reich, editors, Inherently Parallel Algorithms in Feasibility and Optimization and their Applications, Elsevier 2001

[15] Boyd, Stephen P.; Vandenberghe, Lieven (2004). Convex Optimization (pdf). Cambridge University Press. p. 129. ISBN 978-0-521-83378-3.

[16] Elzinga, J.; Hearn, D. W. (1972), "The minimum covering sphere problem", Management Science, 19: 96–104, doi:10.1287/mnsc.19.1.96

[17] Sylvester, J. J. (1857), "A question in the geometry of situation", Quarterly Journal of Mathematics, 1: 79.

[18] Welzl, Emo (1991), "Smallest enclosing disks (balls and ellipsoids)", in Maurer, H. (ed.), New Results and New Trends in Computer Science, Lecture Notes in Computer Science, 555, Springer-Verlag, pp. 359–370, CiteSeerX 10.1.1.46.1450, doi:10.1007/BFb0038202, ISBN 978-3-540-54869-0.

[19] Welzl, Emo (1991), "Smallest enclosing disks (balls and ellipsoids)", in Maurer, H. (ed.), New Results and New Trends in Computer Science, Lecture Notes in Computer Science, 555, Springer-Verlag, pp. 359–370, CiteSeerX 10.1.1.46.1450, doi:10.1007/BFb0038202, ISBN 978-3-540-54869-0.

[20] F. Nielsen and R. Nock, "On the smallest enclosing information disk," Information Processing Letters, vol. 105, no. 3, pp. 93–97, 2008.

Appendix A:

Dual Problem of the Smallest Circle Problem

Let's say $\boldsymbol{X} \subset \mathbb{R}^n$ be a set of m points $x_1, x_2, \dots, x_m$ of the points of $\boldsymbol{X}$.
Let $r = \min_{\boldsymbol{y} \in \mathbb{R}^n} \max_{i \in \{1,\dots m\}} \|\boldsymbol{x_i} - \boldsymbol{y}\|$ be our Chebyshev Radius.
Instead of dealing with this, let's deal with the problem of maximizing:

$$\sum_{i=1}^{m} \lambda_i \|\boldsymbol{x_i} - \boldsymbol{y}\|^2$$

for the vector:

$$\boldsymbol{\lambda} \geq 0 \text{ (i.e. } \lambda_i \geq 0 \text{ for all } i = 1, \dots, m)$$

and also with the condition:

$$\textstyle\sum_{i=1}^{m} \lambda_i = 1.$$

Now let's call:

$$f(\boldsymbol{\lambda}, \boldsymbol{y}) = \sum_{i=1}^{m} \lambda_i \|\boldsymbol{x_i} - \boldsymbol{y}\|^2$$

Now we will apply von Neumann's Minimax Theorem to this function to show it is actually same problem with our smallest circle problem.
$\boldsymbol{\lambda}$ is an element of the region bounded by:

$$\lambda_i \geq 0 \text{ for all } i = 1, \dots, m \text{ and}$$

$$\sum_{i=1}^{m} \lambda_i = 1$$

therefore it is a convex subset of $\mathbb{R}^m$. (The compactness is easy to show, both regions are bounded and closed which can be easily seen.) Then from minimax Theorem we have:

$$\min_{\boldsymbol{y}} \max_{\boldsymbol{\lambda}} f(\boldsymbol{x}, \boldsymbol{y}) = \max_{\boldsymbol{\lambda}} \min_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y})$$

which is easy to optimize. In order to maximize:

$$\sum_{i=1}^{m} \lambda_i \|\boldsymbol{x_i} - \boldsymbol{y}\|^2$$

we should choose the coefficients such that:

$$\lambda_i = 0 \text{ when } \|\boldsymbol{x_i} - \boldsymbol{y}\|^2 < \max_{i=1,\dots,m} \|\boldsymbol{x_i} - \boldsymbol{y}\|^2$$

This is indeed the Complementary Slackness Property coming from KKT conditions. We found the conditions on $\boldsymbol{\lambda}$, now let's find the minimizing $\boldsymbol{y}$.
Now we need to find

$$\min_{\boldsymbol{y}} \sum_{i=1}^{m} \lambda_i \|\boldsymbol{x_i} - \boldsymbol{y}\|^2$$

where we don't have any limiting conditions.
Therefore we can directly apply gradient to see that:

$$\sum_{i=1}^{m} \lambda_i (\boldsymbol{x_i} - \boldsymbol{y}) = 0$$

Since we have:

$$\sum_{i=1}^{m} \lambda_i = 1$$

we can see that:

$$\boldsymbol{y} = \sum_{i=1}^{m} \lambda_i \boldsymbol{x_i}$$

Since we found **y** in terms of **x** and $\boldsymbol{\lambda}$ we can put it into the equation directly to make maximization with the vector $\boldsymbol{\lambda}$.
Now we need to find [10]:

$$\max_{\boldsymbol{\lambda}} \sum_{i=1}^{m} \lambda_i \left\| \boldsymbol{x_i} - \sum_{i=1}^{n} \lambda_i \boldsymbol{x_i} \right\|^2$$

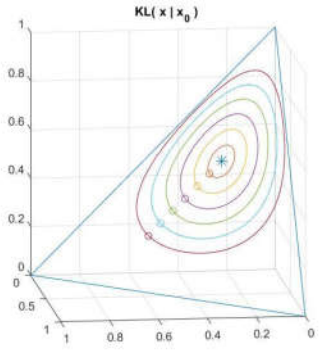
Also we can present this problem as in matrix form such that:

$$\min_{\boldsymbol{\lambda}} \boldsymbol{\lambda^T Q \lambda} - \boldsymbol{b^T \lambda}$$

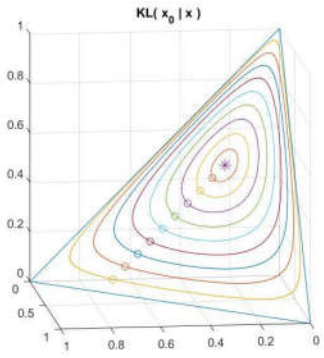
where $A = [x_1\ x_2 \dots x_m]$ which are the points of the set and all x_i are column vectors which are the coordinates of the points. We define the matrix $\boldsymbol{Q} = \boldsymbol{A^T A}$ and vector $\boldsymbol{b}$ as the vector which has the elements $\|x_i\|^2$ for $i = 1, \dots, m$.

Appendix B:

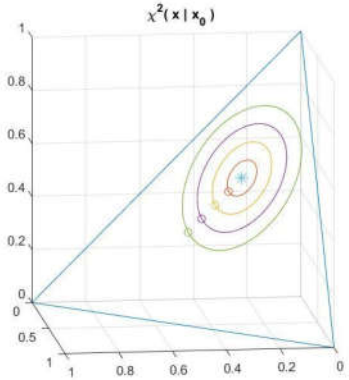
When we ar studying about $f - divergences$ we also studied about level curves equipped with different $f - divergences$ and how to draw them in 3-dimensional probability simplex. We used Gradient Method to draw this level curves. Here are some level curve examples:



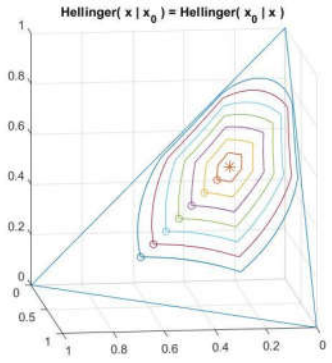
Different Level Curves with KL-distance



Different Level Curves with reverse KL-distance



Different Level Curves with χ^2 distance



Different Level Curves with Hellinger distance

EK-2

Patlamalı-Kesikli Gözlemler için Parametre Kestirimi
Parameter Estimation For Bursty-Intermittent Observations

Çağatay Candan
Elektrik ve Elektronik Mühendisliği Bölümü
Orta Doğu Teknik Üniversitesi
ccandan@metu.edu.tr

***Özetçe* — Gözlem gürültüsü güç seviyesinin örnekten örneğe değişken olduğu veri toplama ortamındaki parametre kestirim problemi incelenmektedir. Bu çalışmada gürültü sürecine ait parametrelerin değişimi Markov süreç olarak modellenmiş ve Markov sürecin gözlemlenemeyen durum vektörü gizli değişken olarak kestirim problemine eklenmiştir. Beklenti-enbüyütme yöntemi ile hem gizli değişken vektörü hem de ilgilendiğimiz işaret yinelemeli olarak kestirilmektedir. Önerilen yöntem gürültü güç seviyesinin gözlem toplama süresi boyunca değiştiği, patlamalı gürültü ve/veya kesikli işaretin olduğu uygulamalarda işaret değerini kestirme amacıyla kullanılabilir.**

Anahtar Kelimeler—Parametre kestirimi, Gizli Markov model, Beklenti-enbüyütme yöntemi, Cramer-Rao sınırı.

***Abstract*—Parameter estimation problem is examined in the setting where the noise power is allowed to change from sample to sample. Parameters of the noise source is assumed to be generated by a Markov chain whose state sequence is not known by the observation system. Expectation-maximization algorithm is applied for the estimation of desired parameter with the inclusion of unknown state vector of the Markov chain realization as a latent variable. The suggested scheme can be utilized in applications with bursty noise and/or intermittent signals.**

Keywords—Parameter estimation, Hidden Markov models, Expectation-maximization method, Cramer-Rao bound.

I. GİRİŞ

Toplanır gürültü altında parametre kestirimi işaret işlemenin temel problemlerinden biridir. Gürültü dağılımının bilindiği durumda en büyük olabilirlik kestirim yöntemi ile işaret kestirimi yapılabilmekte ve bu kestirim sonucu Cramer-Rao alt sınırı gibi başarım sınırlarıyla karşılaştırılarak kestirimci değerlendirilebilmektedir [1]. Birçok işaret işleme uygulamasında en büyük olabilirlik yöntemini gerçeklemek pratik olarak mümkün olmadığı için alternatif yöntemler geliştirilmesi gerekmektedir. Bu çalışmada gürültü kaynağı parametrelerinin gizli Markov modeli uyarınca değiştiği varsayılmış ve bu model altında kestirim problemi incelenmiştir.

Kalman filtreleme işlemi parametreleri bilinen, doğrusal bir sistem vasıtasıyla üretildiği varsayılan Gauss sürece ait durum vektörünü bağımsız toplanır Gauss gürültüsü altındaki gözlemlerden kestirme işlemini gerçekleştirir, [1, Bölüm 13]. Bu işlem Gauss dağılımlı gürültü altında Gauss dağılımlı süreç kestirimi için ortalama karesel hatayı (OKH) enküçülten kestirimcidir. Gürültü dağılımlarının Gauss olmaması durumunda aynı işlem OKH'yı enküçülten *doğrusal* kestirimcidir. Kalman filtreleme işleminin dayandığı klasik varsayımlardan (ilgilendiğimiz Gauss sürecin Gauss dağılımlı gürültü altında gözlenmesi) uzaklaşıldığı bazı durumlar için benzer eniyileme özelliklerine sahip kestirimciler literatürde bulunmaktadır. Örneğin,

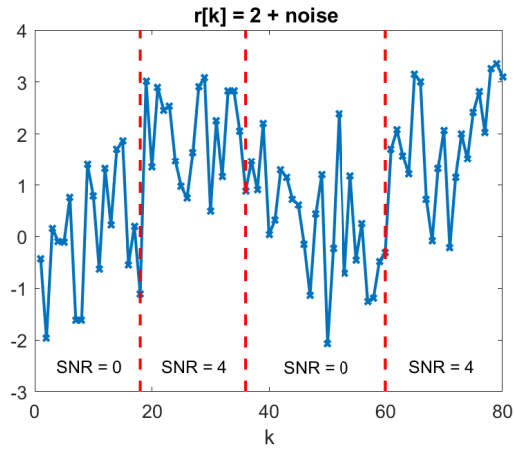
$$\begin{aligned} \boldsymbol{x}[k+1] &= \boldsymbol{A}\boldsymbol{x}[k] + \mathbf{B}\mathbf{u}[k] \\ \boldsymbol{y}[k] &= \gamma(k)\mathbf{C}\boldsymbol{x}[k] + \mathbf{D}\mathbf{v}[k], \end{aligned}$$

modelinde $\gamma(k)$ değişkenleri yerine 1 değerinin yazıldığı durum klasik Kalman filtreme durumu, $\gamma(k)$ değişkeninin $\{0,1\}$ değerlerini rastgele şekilde aldığı durum ise belirsiz gözlem durumunu olarak adlandırılmaktadır [2]. Modelden görüleceği gibi $\gamma(k)$ değişkeninin 0 değerini alması durumunda $\boldsymbol{y}[k]$ gözlemi ilgilendiğimiz süreç olan $\boldsymbol{x}[k]$'dan bağımsız hale gelmekte ve $\boldsymbol{x}[k]$'nın kestirimi için $\boldsymbol{y}[k]$ gözlemi bir bilgi taşımamaktadır. [2] numaralı çalışmada $\gamma(k)$ rastgele değişkenlerinin bağımsız türdeş Bernoulli dağılımlı olduğu varsayılmış ve bu varsayım altında filtreleme uygulaması için en iyi *doğrusal* kestirimci türetilmiştir. Bu çalışmanın devamı niteliğinde olan [3]'te $\gamma(k)$ değişkeninin bağımsız türdeş olma koşulunun genişletildiği durum için eniyi doğrusal kestirici geliştirilmiştir. [4]'de ise kestirim için kıymet taşıyan gözlem toplama olasılığına ait değerin ($\gamma(k) = 1$ olayı için olasılık değeri) kestiricinin asimptotik başarımına olan etkisi incelenmiş ve bu değerinin belli bir eşik değerinin altında olması durumunda hata kovaryans değerinin hudutsuz şekilde büyüdüğü gösterilmiştir. Bu çalışmada daha basit model olan

$$r[k] = \begin{cases} s + w_0[k] & \text{eğer } \gamma_k = 0 \\ s + w_1[k] & \text{eğer } \gamma_k = 1 \end{cases}, \; k = \{0, 1, \ldots, K-1\}$$

gözlem modeli için kestirim problemi çalışılmaktadır. Burada s ilgilendiğimiz rastgele olmayan değişken, $w_0[k]$ ve $w_1[k]$ Gauss dağılımlı rastgele değişkenler, γ_k ise 2-durumlu Markov zinciri ile üretilmiş olan $\{0,1\}$ değerlerini alan rastgele değişkendir. İncelenen problem $r[k]$ gözlemlerinden s parametresinin kestirimidir. Yukarıda bahsedilen literatürden temel fark γ_k değerinin Markov zinciri ile üretilmiş olması ve kestirici olarak enbüyük olabilirlik yönteminin kullanılmasıdır. Öte yandan incelenen problem sıçramalı Markov doğrusal sistemlere ait süreç kestirimi probleminin özel bir hali olarak değerlendirilebilir [5]. Şekil 1'de bu özel durum gösterilmektedir. Burada $r[k]$ gözlemlerinin farklı zaman dilimlerinde, farklı işaret-gürültü oranı (SNR) seviyelerinde toplanma durumu gösterilmektedir. Algılayıcı sistem bazı zaman dilimlerinde işaretten bağımsız

978-1-7281-7206-4/20/$31.00



Şekil 1: Veri toplama sisteminde patlamalı hatalar olmasından kaynaklı olarak farklı SNR değerlerinde toplanan veriye ait bir gösterim

şekilde sadece gürültü üretmekte (SNR $=0$ durumu), diğer zaman dilimlerinde ise SNR $=4$ koşullarında çalışmaktadır. Algılayıcı sistemin gözlem toplama anındaki durumu (sağlıklı/sağlıksız çalışma durumu) ve bu durumlara ait SNR seviyelerin bilinmediği gözlem toplama ortamında işaret kestirimi (örnekteki s değişkeninin kestirimi) bu çalışmanın hedefidir.

II. Problem Tanımı ve Önerilen Çözüm

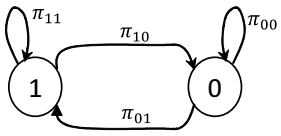
Elimizde aşağıdaki işaret toplama modeli ile elde edilmiş olan K adet gözlem verisi olsun:

$$r[k] = \begin{cases} \alpha_0 s + \beta_0 w[k], & \text{if } \gamma_k = 0 \\ s + \beta_1 w[k], & \text{if } \gamma_k = 1 \end{cases}, \; k = \{0, 1, \ldots, K-1\}. \tag{1}$$

Yukarıda yer alan s ilgilendiğimiz değişkeni, $w[k]$ bağımsız türdeş sıfır ortalama ve birim değişinti değerli Gauss dağılımlı gürültüyü, $w[k] \sim N(w[k]; 0, 1)$; iki değer alan γ_k değişkeni ise veri toplama sisteminin durumunu göstermektedir. Veri toplama sisteminin sağlıksız çalıştığı $\gamma_k = 0$ durumunda, işaret-gürültü-oranı $\text{SNR}_0 = s^2\alpha_0^2/\beta_0^2$, diğer durumda ise (sağlıklı çalışma durumu) $\text{SNR}_1 = s^2/\beta_1^2$ olmaktadır. $\text{SNR}_0 \ll \text{SNR}_1$ olduğu varsayılmıştır.

Veri toplama sisteminin durum dizisi Şekil 2'de gösterilen Markov zinciri yapısı ile modellenmektedir. Bu modelde, örneğin, $k-1$ zamanında sağlıklı çalışan sistemin, bir sonraki anda sağlıklı durumda kalma olasılığı $P(\gamma_k = 1|\gamma_{k-1} = 1) = \pi_{11}$ ile gösterilmektedir. Bu yapıda π_{11} ve π_{00} değerleri 0.5'den çok daha büyük seçilerek, sistemin bulunduğu durumu koruma olasılığının yüksek olması; böylelikle ölçüm hataların yüksek olduğu bir dizi "kötü" örneğin (patlamalı gürültü) ardından bir dizi doğruluğu yüksek "iyi" örneğin toplandığı çalışma ortamı modellenmektedir. Nümerik bir örnek olarak $\pi_{11} = 0.95$ ise sağlıklı veri toplama durumu ortalama olarak $1/(1-\pi_{11}) = 20$ örnek sürmektedir. Markov zinciri ilk durumu (γ_0 değişkeni) zincirin kalıcı değer olasılığına sahip bir rastgele değişken olarak alınmıştır, daha farklı olarak da alınabilir.

İncelenen problem denklem (1) ile verilen model altında s değişkeninin enbüyük olabilirlik yöntemi ile kestirimidir. Problemde sistem durumunu gösteren γ_k rastgele değişkenleri gizli rastgele değişkenler; $\{s, \alpha_0, \beta_0, \beta_1\}$ değişkenleri ise değerleri bilinmeyen *rastgele olmayan* diğer değişkenlerdir.



Durum Geçiş Şeması

$$P(\gamma_k = 1|\gamma_{k-1} = 1) = \pi_{11}$$
$$P(\gamma_k = 0\,|\gamma_{k-1} = 1) = \pi_{10}$$
$$P(\gamma_k = 1|\gamma_{k-1} = 0) = \pi_{01}$$
$$P(\gamma_k = 0|\gamma_{k-1} = 0) = \pi_{00}$$

Durum Geçiş Olasılıkları

Şekil 2: Algılayıcı sistemin durum geçişleri hakkında bilgi

Beklenti-Enbüyütme Yöntemi ile Kestirim: Elimizdeki gözlemlerin alt alta yazılmasıyla oluşturulan $K \times 1$ boyutlu $\boldsymbol{r}$ vektörü ile γ_k değerlerinin benzer şekilde yazılmasıyla oluşturulan $\boldsymbol{\gamma}$ vektörünün (gizli değişkenler vektörü) birleşime eksiksiz gözlem vektörü adı verilmekte ve $\boldsymbol{x} = [\boldsymbol{r}; \boldsymbol{\gamma}]$ ile gösterilmektedir. Beklenti-enbüyütme yöntemi eksiksiz gözlem vektörünün işlendiği iki adımdan oluşur:

1. (Beklenti) $Q(\theta^{\text{yeni}}) = E\{\log(p(\boldsymbol{r}, \boldsymbol{\gamma}; \theta^{\text{yeni}})) \mid \boldsymbol{r}, \theta^{\text{eski}}\}$
2. (Enbüyütme) $\theta^{\text{yeni}} = \text{argmax}_{\theta^{\text{yeni}}} Q(\theta^{\text{yeni}})$

İlk adımın birinci aşamasında eksiksiz gözlem vektörüne ait log-olabilirlik ifadesi $\log(p(\boldsymbol{r}, \boldsymbol{\gamma}; \theta^{\text{yeni}}))$ yazılır. Bu ifadede geçen θ değişkeni problemdeki bilinmeyen rastgele olmayan değişkenleri göstermektedir, $\theta = [s\ \alpha_0\ \beta_0\ \beta_1]$. Yazılan log-olabilirlik ifadesinin gizli değişkenler üzerinden ortalaması hesaplanarak veya daha doğru bir ifade ile eksiksiz log-olabilirlik fonksiyonun $\boldsymbol{\gamma}$'ya ait ardıl olasılık dağılımı, $p(\boldsymbol{\gamma}|\boldsymbol{r}; \theta^{\text{eski}})$, üzerinden beklenti değeri hesaplanarak ilk adım tamamlanır. Burada dikkat edilmesi gereken önemli bir konu $\boldsymbol{\gamma}$'ya ait ardıl olasılık dağılımı θ^{eski} ile gösterilen bilinmeyen θ vektörüne bir takım sabit nümerik değerler atandıktan sonra hesaplanmasıdır. İkinci adımda, birinci adımın çıktısı olan $Q(\theta^{\text{yeni}})$ fonksiyonu analitik veya nümerik yöntemlerle enbüyütülür. İkinci adımın sonucu olan θ^{yeni} değerleri birinci adımda yer alan θ^{eski} değerleri yerine yerleştirilir ve beklenti-enbüyütme adımları yinelemeli şekilde tekrarlanır. Birçok problemde yöntemin başarısı θ vektörünün ilk seçimine hassasiyet göstermektedir.

Beklenti Adımı: $p(\boldsymbol{r}, \boldsymbol{\gamma}; \theta)$ ifadesi $p(\boldsymbol{r}, \boldsymbol{\gamma}; \theta) = p(\boldsymbol{r}\,|\,\boldsymbol{\gamma}; \theta)p(\boldsymbol{\gamma})$ şeklinde yazılabilir. Bu problemde $p(\boldsymbol{\gamma})$ fonksiyonu bilinmeyen parametrelere bağlı değildir. Bu durumda $\log(p(\boldsymbol{r}, \boldsymbol{\gamma}; \theta)) = \log(p(\boldsymbol{r}\,|\,\boldsymbol{\gamma}; \theta)) + c$ olarak yazılabilir. Son eşitlikteki c değeri bilinmeyen parametrelere bağlı olmayan terimleri içermektedir. Beklenti hesabı için ilk olarak $p(\boldsymbol{r}|\boldsymbol{\gamma}; \theta)$ ifadesini

$$p(\boldsymbol{r}|\boldsymbol{\gamma}; \theta) = \prod_{k=0}^{K-1} N(r[k]; s\gamma_k + \alpha_0 s\gamma_k^c, \beta_1^2\gamma_k + \beta_0^2\gamma_k^c) \tag{2}$$

şeklinde yazalım. Burada $\gamma_k^c = 1 - \gamma_k$ şeklinde tanımlanmıştır ve γ_k, γ_k^c değişkenleri sadece 0 ve 1 değerlerini alan ve birbirlerini tümleyen değişkenler olarak düşünülebilir. Eksiksiz gözlem vektörüne ait log-olabilirlik ifadesini $\Lambda(\boldsymbol{r}) = \log(p(\boldsymbol{r}\,|\,\boldsymbol{\gamma}; \theta))$ ile gösterirsek, bu ifade

$$\Lambda(\boldsymbol{r}) \stackrel{c}{=} -\sum_{k=0}^{K-1} \frac{\log(\beta_1^2\gamma_k + \beta_0^2\gamma_k^c)}{2} + \frac{(r[k] - s\gamma_k - \alpha_0 s\gamma_k^c)^2}{2(\beta_1^2\gamma_k + \beta_0^2\gamma_k^c)} \tag{3}$$

şeklinde yazılabilir. Bu eşitlikte yer alan $\stackrel{c}{=}$ sembolu eşitliğin sağ tarafında sonucu etkilemeyen daha önce c ile gösterilmiş olan bazı terimlerin yazılmadığını işaret etmektedir. Beklenti

adımı $\Lambda(\boldsymbol{r})$ ifadesinin $p(\boldsymbol{\gamma}|\boldsymbol{r};\theta^{\text{eski}})$ üzerinden beklenti hesabıyla tamamlanır:

$$Q(\theta) = -\sum_{k=0}^{K-1} p_k \left(\frac{\log(\beta_1^2)}{2} + \frac{(r[k]-s)^2}{2\beta_1^2} \right) - \sum_{k=0}^{K-1} (1-p_k) \left(\frac{\log(\beta_0^2)}{2} + \frac{(r[k]-\overbrace{\alpha_0 s}^{\mu_0})^2}{2\beta_0^2} \right). \quad (4)$$

Son ifadede yer alan p_k, gizli değişken γ_k'ya ait ardıl dağılımın 1 değerini alma olasılığını göstermektedir, $p_k = p(\gamma_k = 1\,|\,\boldsymbol{r};\theta^{\text{eski}})$. Ardıl dağılım hesabı enbüyütme adımı sonrasında verilecektir.

Enbüyütme Adımı: Enbüyütme adımı (4)'de verilen ifadenin türev hesabı ile enbüyütülmesidir. Bu ifadede yer alan $\alpha_0 s$ çarpımı analitik çözümü zorlaştırmaktadır. Algılayıcı sistemin kötü çalıştığı durumdaki $\text{SNR}_0 = s^2\alpha_0^2/\beta_0^2$ değerinin iyi çalışma koşullarındaki $\text{SNR}_1 = s^2/\beta_1^2$ çok daha kötü olması beklendiğinden $\alpha_0 s$ çarpımı yerine μ_0 ile gösterilen yeni bir bağımsız değişken atanabilir. Böylelikle enbüyütme işleme basitleştirilmiş olur ve aşağıdaki sonuçlar elde edilir:

$$\begin{aligned}
s^{\text{yeni}} &= \frac{1}{\sum_k p_k} \sum_k p_k r[k], \\
\mu_0^{\text{yeni}} &= \frac{1}{\sum_k (1-p_k)} \sum_k (1-p_k) r[k], \\
(\beta_0^2)^{\text{yeni}} &= \frac{1}{\sum_k (1-p_k)} \sum_k (1-p_k)(r[k]-\mu_0^{\text{yeni}})^2, \\
(\beta_1^2)^{\text{yeni}} &= \frac{1}{\sum_k p_k} \sum_k p_k (r[k]-s^{\text{yeni}})^2. \quad (5)
\end{aligned}$$

Ardıl Dağılımın Hesabı: Beklenti adımını gerçekleştirmek için $p(\gamma_k|\boldsymbol{r},\theta^{\text{eski}})$ dağılımına ihtiyaç duyulmaktadır. Ardıl dağılım Şekil 3'te verilen bileşik olasılık yoğunluk fonksiyonu üzerinden $\alpha\beta$ yöntemi ile hesaplanabilir [6].

Bu yöntemde $\alpha(\gamma_k) = p(\gamma_k, r[0], \ldots, r[k])$, $\beta(\gamma_k) = p(r[k+1], \ldots, r[K-1]\,|\,\gamma_k)$ dağılımlarını göstermektedir. α fonksiyonu $\alpha(\gamma_0) = p(\gamma_0)p(r[0]\,|\,\gamma_0;\theta^{\text{eski}})$ ile ilk değeri belirlendikten sonra ve $k \geq 1$ için yinelemeli olarak

$$\alpha(\gamma_k) = p(r[k]\,|\,\gamma_k;\theta^{\text{eski}}) \sum_{t=0}^{1} p(\gamma_k|\gamma_{k-1}=t)\alpha(\gamma_{k-1})$$

ile hesaplanır. β-yinelemesi ise $\beta(\gamma_{K-1}) = 1$ ilk değeriyle, $2 \leq k \leq K-1$ aralığındaki azalan k değerleri için

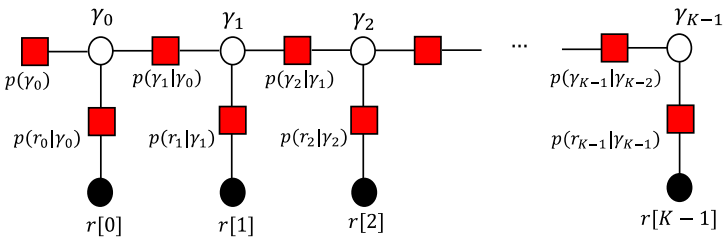
$$\beta(\gamma_{k-1}) = \sum_{t=0}^{1} p(r[k]\,|\,\gamma_k = t;\theta^{\text{eski}}) p(\gamma_k = t\,|\,\gamma_{k-1})\beta(\gamma_k = t)$$

ile hesaplanır. Ardıl dağılım ise

$$p(\gamma_k\,|\,\boldsymbol{r};\theta^{\text{eski}}) = \frac{\alpha(\gamma_k)\beta(\gamma_k)}{\alpha(\gamma_k=0)\beta(\gamma_k=0) + \alpha(\gamma_k=1)\beta(\gamma_k=1)}$$

olur. Gizli Markov yapılarının temelini oluşturan bu hesabın detayları için [6] numaralı makaleye bakabilirsiniz.

Beklenti-Enbüyütme Yöntemi için Başlangıç Noktası: Beklenti-Enbüyütme yönteminde yer alan ardıl dağılım hesabını gerçekleştirmek için θ^{eski} ile gösterilen parametre değerlerine ihtiyaç duyulmaktadır. Bu vektörün ilk değeri Beklenti-Enbüyütme yönteminin başarımı için çoğu problemde kritik



Şekil 3: Bileşik yoğunluk fonksiyonuna ait çarpan çizgesi

önemdedir. Aşağıda θ vektörünün ilk değerini kestirmek için bir yöntem verilmektedir.

İlk değeri oluşturmak için veriyi bölütleyerek işaret olan ve olmayan kısımları ayırmayı hedefleyen bir yöntem önereceğiz. Yöntem Şekil 1'deki örnek üzerinden anlatılacaktır. Şekil 1'de yer alan veriyi

$$r[k] = \begin{cases} N(r[k];\mu_1,\sigma_1^2), & c_0 = 0 \leq k < c_1 \\ N(r[k];\mu_2,\sigma_2^2), & c_1 \leq k < c_2 \\ N(r[k];\mu_3,\sigma_3^2), & c_2 \leq k < c_3 \\ N(r[k];\mu_4,\sigma_4^2), & c_3 \leq k \leq c_4 = K-1 \end{cases} \quad (6)$$

c_k ile gösterilen bölüt sınırları, (μ_k,σ_k^2) ile gösterilen parametreleri olan Gauss dağılımlı süreç örnekleri olarak düşünelim. Bu modeldeki parametreleri eldeki veriden enbüyük olabilirlik yöntemi ile kestirerek hem bilinmeyen $\{\mu_k,\sigma_k^2\}$ parametreleri hem de bölüt sınırları için kestirimler elde edebiliriz. Bilinmeyen $\{\mu_k,\sigma_k^2\}$ parametreleri için en büyük olabilirlik kestirimi yapılır ve olabilirlik fonksiyonuna kestirim değerleri yerleştirince, enyüksek olabilirlik değerli bölüt sınırları belirleme problemi $[c_1,c_2,c_3] = \text{argmin}_{c_1,c_2,c_3} J(c_1,c_2,c_3)$

$$J(c_1,c_2,c_3) = \sum_{l=1}^{4} (c_l - c_{l-1}) \log(\sigma^2(r(c_{l-1} : c_l - 1)))$$

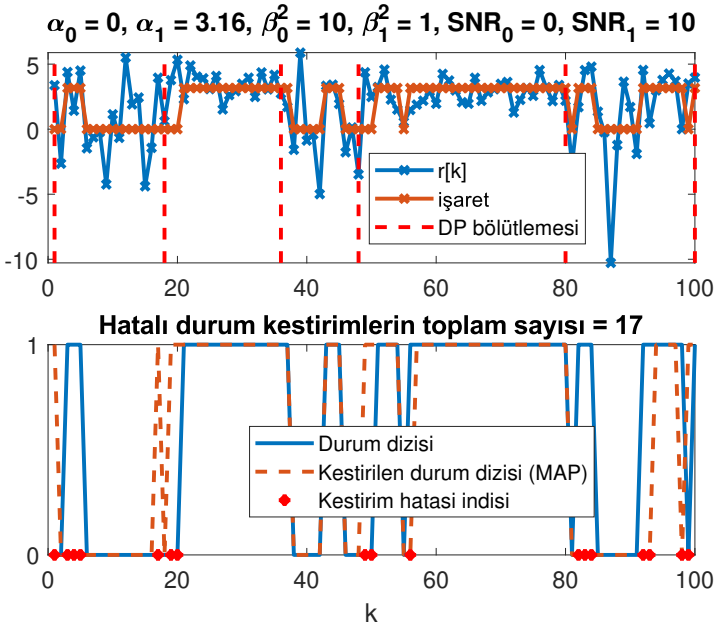
olur. Son ifadedeki $\sigma^2(r(c_{l-1} : c_l - 1))$ fonksiyonu l numaralı bölüte ait yanlı değişinti kestirimidir, $\sigma^2(r(c_{l-1} : c_l - 1)) = \frac{1}{c_l - c_{l-1}} \sum_{k=c_{l-1}}^{c_l-1} (r_k - \widehat{\mu}_k)^2$, $\widehat{\mu}_k = \frac{1}{c_l - c_{l-1}} \sum_{k=c_{l-1}}^{c_l-1} r_k$. $J(c_1,c_2,c_3)$ fonksiyonunun enküçültülmesi işlemini dinamik programlama ile verimli şekilde gerçeklemek mümkündür. Verimli gerçeklemenin detayları için [7]'ye bakabilirsiniz.

Yukarıda verilen bölütleme işlemini gerçekleştirmek için bölüt sayısına ihtiyaç duyulmaktadır. Bölüt sayısı kestirimi için model derecesi seçimi yöntemlerinden Bayesci bilgi kriteri (BIC - Bayesian information criterion) kullanılabilir [8]:

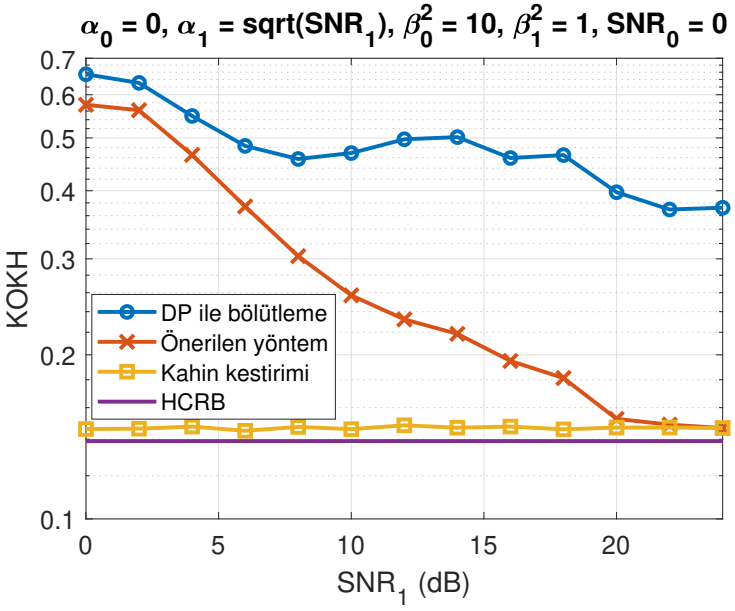
$$\text{BIC}(d) = -2\log\left(\max_{\widehat{\boldsymbol{\mu}}_n, \widehat{\boldsymbol{\sigma}}^2_n, \boldsymbol{c}_n} p(\boldsymbol{r};\widehat{\boldsymbol{\mu}}_n,\widehat{\boldsymbol{\sigma}}^2_n,\boldsymbol{c}_n) \right) + (3d-1)\log N.$$

Burada d bölüt sayısını göstermektedir. $\text{BIC}(d)$ değerini enküçülten bölüt sayısı bilgi kriteri seçim sonucudur. $\text{BIC}(d)$ ifadesinde yer alan enbüyük olabilirlik değeri daha önceden bahsedilen ve [7]'de detayları verilen dinamik programlama ile verimli şekilde hesaplanabilir. $\text{BIC}(d)$ ifadesinde yer alan $3d-1$ faktörü d adet bölüt için modeldeki toplam bilinmeyen sayısıdır [8].

Bölütleme işlemi tamamlandıktan sonra gözlem vektörünün bölütler içerisindeki ortalama değeri ve değişinti değerleri hesaplanır. Bulunan ortalama değerler aynı bölütleme işlemi



Şekil 4: Deney koşulları altında gözlemlenen bir koşum örneği



Şekil 5: Kestirim doğruluğu karşılaştırması

ile 2 gruba ayrılır. Yüksek ortalamalı bölütlerdeki örneklerin sağlıklı çalışma durumunda ($\gamma_k = 1$) toplandığı varsayılır ve bu örneklere $p_k = 1$, diğer örneklere $p_k = 0$ ataması yapılır ve denklem (5)'de yer alan ifadeler kullanılarak s, μ_0, β_0^2, β_1^2 parametrelerinin ilk değerleri elde edilir [7].

III. Benzetim Sonuçları

Algılayıcı sisteminin sağlıklı çalışma koşullarında $r[k] = s + w_1$, $w_1 \sim N(0,1)$ modeliyle, diğer durumunda ise $r[k] = w_0$, $w_0 \sim N(0,10)$ modeliyle veri topladığı varsayılsın. Bu durumda ilgilendiğimiz işaret s sistemin kötü çalışma durumunda gözlemleri etkilememektedir. (Bu senaryo (1)'de $\alpha_0 = 0$, $\beta_0 = \sqrt{10}$, $\beta_1 = 1$ seçimine denk gelmektedir.) Algılayıcı sistemin iyi/kötü çalışma durumları arasındaki geçiş olasılığı $\pi_{01} = \pi_{10} = 0.1$ olarak, sistemin ilk durumu (γ_0) ise eşit olasılıklı şekilde iyi/kötü durumlarından biri olarak seçilsin. Toplam $K = 100$ adet örnek toplandığı varsayılsın.

Şekil 4'de incelenen SNR$_1 = 10$ durumunda elde edilen bir koşum gösterilmektedir ($s = \sqrt{\text{SNR}_1}$). Şekil 4'ün üst kısmındaki grafikte mavi ve kahverengi çizgiler sırasıyla gözlemleri ve ilgilendiğimiz işareti göstermektedir. Bazı zaman dilimlerinde işaret gözlemlenememektedir (kesikli işaret durumu). Ayrıca işaretin gözlemlenemediği durumlarda gürültü değişintisi 10 kat artmaktadır (patlamalı gürültü). Üst grafikte verinin dinamik programlama ile bölütlenmesi sonucunda elde edilen bölüt sınırları gösterilmektedir. Dinamik programla ile hesaplanan bilinmeyen parametrelerin beklenti-enbüyütme yöntemi ile işlenmesinden sonra elde edilen algılayıcı sistem durum kestirimi Şekil 4'ün ikinci parçasında verilmektedir. Bu deneyde eldeki 100 örnekten 17 tanesine ait durumun yanlış kestirildiği görülmektedir. Hatalı kestirimler çoğunlukla işaretin kısa süreli olarak gözüktüğü zaman dilimlerine aittir.

Şekil 5'de sistemin sağlıklı çalışma durumuna ait farklı SNR$_1$ değerleri için önerilen yöntemin kök ortalama karesel hata (KOKH) değeri gösterilmiştir. Şekilde önerilen yöntemin ilk aşaması olarak düşünülebilecek olan dinamik programlama temelli bölütleme yönteminin kestirim başarısı, gizli değişkenlere ait durum vektörünü hatasız şekilde bilen kahin kestirimcisinin başarısı ve başarım alt sınırı olarak hibrit Cramer-Rao sınırı (HCRB - Hybrid Cramer Rao Bound) verilmektedir. Sonuçlar beklenti-enbüyütme yinelemelerinin parametre kestirim doğruluğunu önemli ölçüde artırdığı göstermektedir.

IV. Sonuç

Bu çalışmada klasik yaklaşıma göre daha karmaşık yapılı bir gürültü modeli altında parametre kestirim problemi incelenmektedir. Verilen yöntem gürültünün patlamalı, işaretin kesikli olduğu durumların tekil veya beraber olarak yaşandığı (benzetim sonuçları kısmındaki örnekte olduğu gibi) uygulamalarda kullanılabilir. Yöntem, işarete ait gözlemleri üreten algılayıcı sistemin kesikli olarak yaşanan girişim etkilerinden dolayı hatalı sonuçlar üretebildiği uygulamalar için geliştirilmiştir. Yöntemin başarımı bulut üzerinde bulunan, kullanıma hazır MATLAB kodları çalıştırılarak incelenebilir [7].

Kaynaklar

[1] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume 1: Estimation Theory* . Prentice Hall, 1993.

[2] N. Nahi, "Optimal recursive estimation with uncertain observation," *IEEE Trans. Inf. Theory*, vol. 15, no. 4, pp. 457–462, 1969.

[3] M. Hadidi and S. Schwartz, "Linear recursive state estimators under uncertain observations," *IEEE Trans. Automatic Control*, vol. 24, no. 6, pp. 944–948, 1979.

[4] B. Sinopoli, L. Schenato, M. Franceschetti, K. Poolla, M. I. Jordan, and S. S. Sastry, "Kalman filtering with intermittent observations," *IEEE Trans. Automatic Control*, vol. 49, no. 9, pp. 1453–1464, 2004.

[5] A. Logothetis and V. Krishnamurthy, "Expectation maximization algorithms for MAP estimation of jump Markov linear systems," *IEEE Trans. Signal Process.*, vol. 47, no. 8, pp. 2139–2156, 1999.

[6] D. Barber and A. T. Cemgil, "Graphical models for time-series," *IEEE Signal Process. Mag.*, vol. 27, no. 6, pp. 18–28, 2010.

[7] C. Candan. (2020) Parameter Estimation For Bursty-Intermittent Observations (MATLAB Code). [Online]. Available: https://codeocean.com/capsule/4933635/tree

[8] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, no. 4, pp. 36–47, 2004.

EK-3

Proper Definition and Handling of Dirac Delta Functions

Journal:	*IEEE Signal Processing Magazine*
Manuscript ID	SPM-Apr-2020-084.R1
Manuscript Type:	Columns and Forum
Date Submitted by the Author:	30-Jul-2020
Complete List of Authors:	Candan, Cagatay; METU, Electrical Engineering
Keywords:	generalized functions, dirac delta function, Fourier transform

Proper Definition and Handling of Dirac Delta Functions

Çağatay Candan
Department of Electrical and Electronics Engineering
Middle East Technical University (METU), Ankara, Turkey.
ccandan@metu.edu.tr

I. Scope

Dirac delta functions are introduced to students of signal processing in their sophomore year. Quite understandably, Dirac delta functions, which should be more aptly called as generalized functions or distributions, can not be comprehensively given to a young audience at the beginning of their engineering education. Instead, a simplified and abridged definition is presented and the implications of the definition in signal processing problems are illustrated through numerous examples following the footsteps of Oppenheim and Schafer [1, 2]. Students typically learn the properties by developing an affinity through their usage. As the mathematical knowledge of students matures, some students tend to notice inconsistencies related with the sugar-coated definitions and start questioning the mathematics behind them. Unfortunately, inquisitive questions of these students are rather difficult to answer convincingly due to the lack of sources on the generalized functions at the level of undergraduate/graduate engineering students. The goal of these notes is to scratch the sugar-coating a bit and provide the basics of generalized functions, generalized limits, generalized derivatives and their usage in signal processing problems. As an illustrative example, the Fourier transform of $f(t) = 1$, which is $F(\Omega) = 2\pi\delta(\Omega)$, is typically "proven" with the application of inverse Fourier transform on $F(\Omega) = 2\pi\delta(\Omega)$. Yet, according to the standard calculus results, the Fourier transform of $f(t) = 1$, which is $\mathcal{F}\{1\} = \int_{-\infty}^{\infty} \exp(-j\Omega t)dt$, ceases to exist for any Ω in the ordinary calculus sense. The plot further thickens when Fourier transform of unit step function, sign function and Hilbert transform discussions come into play. Generalized functions enable these calculations and they are indispensable tools of our field; yet, their proper understanding, true definition and why's & how's about their usage require an update on our classical calculus knowledge. Such an update, however it is incomplete, is the topic of these notes.

II. Relevance

Paul Dirac is one of the giants among the great physicists of early 20th century. It is a compliment to our profession that he received his first academic degree in electrical engineering (University of Bristol). He said that "I owe a lot to my engineering training because it [taught] me to tolerate approximations. Previously to that I thought one should just concentrate on exact equations all the time. Then I got the idea that in the actual world all our equations are only approximate. We must just tend to greater and greater accuracy. In spite of the equations being approximate, they can be beautiful." The function $\delta(t)$ introduced by him is now called Dirac delta function. Dirac delta function brings great computational and conceptual advantages in calculations involving diverging integrals such as the case for some Fourier integrals. In addition, the inclusion of Dirac delta function to the calculus of ordinary functions enables the differentiation of discontinuous (generalized) functions paving the way towards a consistent analysis of highly practical engineering problems, such as the circuit theory problems involving switches, unified treatment for the mixed random variables (random variables which are both discrete and continuous) and more. Despite the abundance of topics utilizing Dirac delta functions in signal processing, there are only a few sources explaining the true nature of the approximation involved to the signal processing audience, [3, Appendix I], [4]. This column is prepared to answer some of the questions on generalized functions, illustrate their properties and show their proper usage in some signal processing calculations.

Intended audience of these notes are instructors, researchers with an inclination towards theory and graduate students getting close to fulfilling their course requirements, say studying for Ph.D. qualification exams. For beginners to the topic, author suggests to follow the mainstream track and develop an affinity to the topic first by following the wisdom of Oppenheim and Schafer [1, 2]. The conventional treatment aims to develop a working knowledge of Dirac delta functions, which is a noteworthy goal on its own, and gives a good "first order approximation" to the topic. Science and engineering is built upon successively refined approximations as Paul Dirac has alluded to as a potential source of beauty. Especially in engineering, approximate models/explanations are important beyond their aesthetic value; because of the basic need for working tools and methods for the solution of practical problems. In a typical undergraduate course, the need for a working solution may easily shadow the need for a comprehensive theoretical treatment. As an example, the first course in physics studies the mechanics of inclined planes, stacked boxes with high/low friction surfaces etc. If we consider two stacked wood blocks on a flat surface, we may say that the weight of top block is balanced with the *normal* force so that the net force on the block is zero. This comment can be used to explain why two blocks does not coalesce into a single piece. Yet, if we think about the nature of normal force, the normal force is typically explained as a direct consequence of Newton's laws of motion (the law of action-reaction) and Newton's laws are brought upon students axiomatically in relation with Newton's empirical observations. Hence, the contents of Physics 101 correctly predict that two stacked wood blocks won't coalesce into a single piece without much saying about the mechanism behind the process! In spite of that Physics 101 students learn to use and appreciate the benefit of defining a normal force through a series exercises and problems; just like a beginner signal processing student working his/her way through a set of exercise problems on Dirac delta functions. Much later, the physicists with advanced degrees learn that the macroscopic normal force is due to the Pauli exclusion principle applied to the bulk matter [5]. Needless to say, such a comprehensive answer is of no help to Physics 101 students working on problems with inclined planes. The situation is almost analogous for signal processing students and Dirac delta functions. Hence, author believes that an exposure to Dirac delta functions beyond the conventional Oppenheim and Schafer level can be safely postponed to the graduate studies. Of course, the professionals of the field, lecturers and researchers, can refer quick learners with inquisitive questions to these notes disregarding the suggested timeline.

III. Prerequisites

Only prerequisites are working knowledge of freshman calculus, basic signal processing theory and a keen eye for detail.

IV. Problem Statement and Solution

A. Problem Statement

The main focus is on the handling of integrals, limits, derivatives which do not exist in the standard calculus sense. The Fourier transform of $u(t)$ (unit step function), $F(\Omega) = \int_{-\infty}^{\infty} u(t)\exp(-j\Omega t)dt$ is the prime illustrative example. This Fourier transform integral requires the evaluation of $\int_0^\infty \cos(\Omega t)dt$ and $\int_0^\infty \sin(\Omega t)dt$, which are known to diverge according to standard calculus results. Yet, signal processing textbooks express the result as $\mathcal{F}\{u(t)\} = 1/j\Omega + \pi\delta(\Omega)$, [1, Table 4.2]. The appearance of $\delta(\cdot)$ function hints the divergence of the Fourier integral to an experienced eye. However, this is not the case for all divergent integrals. The Fourier transform of $\mathrm{sgn}(t)$ (sign function) requires the evaluation of $\int_0^\infty \sin(\Omega t)dt$, which is a divergent integral. Yet, textbooks state that $\mathcal{F}\{\mathrm{sgn}(t)\} = 2/j\Omega$. The main problem is that the transform pair for both functions is not valid in the ordinary calculus sense; but valid in the generalized sense or in the sense of distributions. These notes study the definition of generalized functions and their use in signal processing problems.

B. Solution

We first present some basic definitions to better explain the upcoming definition on the Dirac delta and other generalized functions.

Function: Functions, as defined on the set of real numbers, map real numbers to real numbers. Functions are interpreted in a point-wise manner. For example, $\phi(t) = t^2$ maps t_0 in $(-\infty, \infty)$ to t_0^2 in $[0, \infty)$.

Linear Functional: A functional is a mapping from the space of functions to real numbers. For example, area functional defined as Area$\{\phi\} = \int_{-\infty}^{\infty} \phi(t)dt$ maps the function $\phi(t)$ to the numerical value of the total area under $\phi(t)$. A functional that satisfies the linearity conditions (homogeneity and additivity, [1, Section 1.6.6]) is called a linear functional. Our focus is entirely on linear functionals. Hence, the phrase functional should be interpreted as a linear functional in these notes.

It is easy to verify that the functional $T_f\{\cdot\}$

$$T_f\{\phi(t)\} \triangleq \langle f(t), \phi(t)\rangle = \int_{-\infty}^{\infty} f(t)\phi(t)dt \tag{1}$$

satisfies the conditions of linearity. We use notations of $T_f\{\phi(t)\}$ and $\langle f(t), \phi(t)\rangle$ interchangeably to denote functionals. $T_f\{\phi(t)\}$ explicitly shows that the "input" $\phi(t)$ is mapped an "output", i.e. a real number. The function $f(t)$ appearing in the subscript of $T_f\{\phi(t)\}$ characterizes the mapping. As an example, the area functional, previously given, can be realized by substituting $f(t)$ with 1 in (1). The second notation $\langle f(t), \phi(t)\rangle$ is handy in many calculations due to the symmetry between $f(t)$ and $\phi(t)$ in (1).

We refer the function $\phi(t)$ as the test function. Hence, $T_f\{\phi(t)\}$ is said to operate on test functions. Generalized functions or distributions, shown as $f(t)$, are built upon the "observed" action of functionals on the test functions, as described below.

Generalized equality: If functions $f(t)$ and $g(t)$ induce the same functional, that is $T_f\{\phi(t)\}$ and $T_g\{\phi(t)\}$ yield identical outputs for all test functions, functions $f(t)$ and $g(t)$ are said to be equal in the generalized sense. We show the generalized equality with the notation of $f(t) \overset{(g)}{=} g(t)$:

$$f(t) \overset{(g)}{=} g(t) \Leftrightarrow \langle f(t), \phi(t)\rangle = \langle g(t), \phi(t)\rangle \text{ for all } \phi(t). \tag{2}$$

To make the statements precise, we need to specify the function class for the test functions and also give a discussion of Lebesgue integration. We refer readers to [6, Ch.6] for a readable account of these topics. As readers can intuitively appreciate, the class for the test functions should be sufficiently "rich" and "refined" so that the generalized equality in (2) presents practically useful results. For example, if the test functions are limited constant functions, say $\phi(t) = c$ where c is a real number, the generalized equality in (2) only implies the equality of area under two functions, which is of rather limited value.

In this text, we assume that the test function class is infinitely differentiable functions in the form of Gaussian functions

$$\phi_{\mu,\sigma}(t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) \tag{3}$$

TABLE I
TABLE OF PROPERTIES FOR DIRAC DELTA FUNCTION AND ITS DERIVATIVES

Basic	Multiplication	$f(t)\delta(t-t_0) \overset{(g)}{=} f(t_0)\delta(t-t_0)$
	Scaling	$\delta(at) \overset{(g)}{=} \frac{1}{\lvert a\rvert}\delta(t)$
	Sifting	$\int_{-\infty}^{\infty} f(t)\delta(t-t_0)dt = f(t_0)$
	Convolution	$\delta(t) * f(t) \overset{(g)}{=} f(t)$
Advanced	Multiplication	$f(t)\delta^{(n)}(t-t_0) \overset{(g)}{=} \sum_{k=0}^{n}(-1)^k \binom{n}{k} f^{(k)}(t_0)\delta^{(n-k)}(t-t_0)$ where $\frac{d^n}{dt^n}\delta(t) = \delta^{(n)}(t)$ and $\frac{d^n}{dt^n}f(t) = f^{(n)}(t)$
	Scaling	$\delta(f(t)) \overset{(g)}{=} \sum_{k=1}^{K} \frac{1}{\lvert f'(t_k)\rvert}\delta(t-t_k)$ where t_k are zeros of $f(t)$, $f(t_k) = 0,\ k = \{1, 2, \ldots, K\}$
	Sifting	$\int_{-\infty}^{\infty} f(t)\delta^{(n)}(t-t_0)dt = (-1)^n f^{(n)}(t_0)$
	Convolution	$f(t) * \delta^{(n)}(t) \overset{(g)}{=} f^{(n)}(t)$

with arbitrary mean μ and spread σ. We take this class of test functions as sufficiently rich and refined so that the generalized equality $f(t) \overset{(g)}{=} g(t)$ in (2) becomes practically meaningful[1].

Dirac Delta Function: We consider a specific functional, called the evaluation functional, that maps the function $\phi(t)$ to $\phi(t_0)$, i.e. the evaluation functional maps $\phi(t)$ to the value of its sample at $t = t_0$. The evaluation functional is clearly linear, but it is not possible express the evaluation functional in the form of (1) with a regular $f(t)$ function. In spite of that, we substitute $f(t)$ with $\delta(t - t_0)$ in (1) and use the following as a formal definition of the evaluation functional:

$$\int_{-\infty}^{\infty} \delta(t-t_0)\phi(t)dt = \phi(t_0) \quad \text{for all } \phi(t). \tag{4}$$

We do not question the existence of $\delta(t)$ function at this point; but treat it as a regular function for now. Readers may interpret (4) as another notation for the evaluation functional from which some properties, such as the linearity of the functional, can be readily observed. Our goal is to derive some properties of $\delta(t)$, given in Table I, first and then answer existence questions.

Verification of Multiplication Property: Let's study the product of $f(t)$ and $\delta(t - t_0)$ which is $f(t_0)\delta(t - t_0)$ according to the multiplication property $f(t)\delta(t-t_0) \overset{(g)}{=} f(t_0)\delta(t-t_0)$ in Table I. To prove the generalized inequality, we need to show that $\langle f(t)\delta(t-t_0), \phi(t)\rangle = \langle f(t_0)\delta(t-t_0), \phi(t)\rangle$

[1]The class of infinitely differentiable test functions with rapid decay at infinity is called the Schwartz space [6, 7]. The Hermite functions, an orthonormal and complete set for L_2, are members of this class. Laurent Schwartz received the Fields Medal in 1950 for building the mathematical foundation (theory of distributions) to the framework of Dirac.

for all test functions. Let's focus on the term on left hand side $\langle f(t)\delta(t-t_0),\phi(t)\rangle$ first:

$$\begin{aligned}\langle f(t)\delta(t-t_0),\phi(t)\rangle = \int_{-\infty}^{\infty} f(t)\delta(t-t_0)\phi(t)dt &\overset{(a)}{=} \int_{-\infty}^{\infty}\delta(t-t_0)\widehat{\phi}(t)dt\big|_{\widehat{\phi}(t)=f(t)\phi(t)}\\ &\overset{(b)}{=} \widehat{\phi}(t_0)\\ &= f(t_0)\phi(t_0),\end{aligned} \tag{5}$$

In line-(a), $\widehat{\phi}(t)=f(t)\phi(t)$ is introduced and $\widehat{\phi}(t)$ is assumed to be a member of the test function class due to its "richness" and "fineness". Line-(b) is due to the definition of evaluation functional.

The right side of equality $f(t)\delta(t-t_0)\overset{(g)}{=}f(t_0)\delta(t-t_0)$ can be worked out as follows:

$$\begin{aligned}\langle f(t_0)\delta(t-t_0),\phi(t)\rangle = \int_{-\infty}^{\infty} f(t_0)\delta(t-t_0)\phi(t)dt &= f(t_0)\int_{-\infty}^{\infty}\delta(t-t_0)\phi(t)dt\\ &= f(t_0)\phi(t_0).\end{aligned} \tag{6}$$

Combining equations (5) and (6), we have

$$\langle f(t)\delta(t-t_0),\phi(t)\rangle = \langle f(t_0)\delta(t-t_0),\phi(t)\rangle, \quad \text{for all } \phi(t). \tag{7}$$

which concludes the proof of $f(t)\delta(t-t_0)\overset{(g)}{=}f(t_0)\delta(t-t_0)$.

An important take-away message from the proof of the first property is not the final result; but the proof procedure followed for the generalized equalities. The equality sign $\overset{(g)}{=}$ appearing in $f(t)\overset{(g)}{=}g(t)$ denotes the equality of the functionals for every member of test function class. It is indeed very different from the ordinary equality sign.

A rather silly, but a memorable, analogy given by one of my instructors can be repeated as follows: Assume that you are in a county fair and there is a contest going on about identifying an unknown animal. Contestants are allowed to ask only yes/no questions. After several rounds of questions, you learn that the animal is green, lives in a lake, capable of leaping significant distances and quacks. Given this information, can you say that the animal is a frog? If you have asked sufficiently large number of informative questions (richness and fineness of the question class), you can be pretty much sure that the animal is a frog! Yet, there is always a possibility that the animal is of another species which is capable of imitating a frog quite closely! But, if you are only interested in the actions of this animal, there is no harm that you call the animal, irrespective of its genus, as a frog or a generalized frog!

Analogous to the story, a generalized function $f(t)$ is characterized by its response to the probing test functions $\phi(t)$. Generalized functions are declared equal if they give the same response to all test functions.

The major mishap in the treatment of the impulse function or Dirac delta function in all signal processing texts is the usage of an ordinary equality sign $=$ instead of generalized equality sign $\overset{(g)}{=}$. This carries the potential of interpreting equations involving $\delta(t)$ in a pointwise manner which is prone to inconsistencies and calculation mishaps.

Verification of Scaling Property: Let's verify the scaling property $\delta(at)\overset{(g)}{=}\frac{1}{|a|}\delta(t)$ given in Table I. The left side of the equality can be written as:

$$\langle\delta(at),\phi(t)\rangle = \int_{-\infty}^{\infty}\delta(at)\phi(t)dt\,\big|_{u=at}\frac{1}{|a|}\int_{-\infty}^{\infty}\delta(u)\phi\left(\frac{u}{a}\right)du = \frac{\phi(0)}{|a|}. \tag{8}$$

Here $\phi\left(\frac{u}{a}\right)\delta(u)$ is assumed to be in the test function class, as in the proof of the first property and also treated $\delta(at)$ as a regular function and changed the integration variable from t to $u=at$ without any due diligence (more on this later).

The right side of the equality $\delta(at)\overset{(g)}{=}\frac{1}{|a|}\delta(t)$ can be written as:

$$\langle\frac{1}{|a|}\delta(t),\phi(t)\rangle = \frac{1}{|a|}\langle\delta(t),\phi(t)\rangle = \frac{\phi(0)}{|a|}. \tag{9}$$

Equations (8) and (9) imply the generalized equality of $\delta(at)\overset{(g)}{=}\frac{1}{|a|}\delta(t)$. Note that setting $a=-1$ in the scaling property gives $\delta(t)\overset{(g)}{=}\delta(-t)$ which is the evenness of function $\delta(t)$ in the generalized sense.

Generalized Limit: Up to this point, we have averted the existence questions on $\delta(t)$ function; but rather focused on its properties. Now, we present a limit argument for the construction of the Dirac delta function. The described limit operation is called as the generalized limit. In standard textbooks, Dirac delta function is introduced as the *point-wise* limit of ordinary functions, which is not the correct definition and the root cause of confusions in many discussions.

The generalized limit of ordinary functions $f_n(t)$ is said to be generalized function $f(t)$, if

$$\lim_{n\to\infty}\int_{-\infty}^{\infty}f_n(t)\phi(t)dt = \int_{-\infty}^{\infty}f(t)\phi(t)dt \tag{10}$$

is satisfied for all test functions $\phi(t)$. We denote the generalized limit as $f_n(t)\overset{(g)}{\to}f(t)$.

Dirac delta function can be given as the generalized limit of ordinary $f_n(t)$ functions defined as follows

$$f_n(t) = \begin{cases} n/\epsilon, & -\frac{\epsilon}{2n} < t < \frac{\epsilon}{2n} \\ 0, & \text{other}\end{cases}. \tag{11}$$

From Figure 1, it can be seen that $f_n(t)$ is a pulse of duration ϵ/n, centered around $t=0$. Area under $f_n(t)$ is unity for all n. With the running assumption that the test functions $\phi(t)$ are sufficiently smooth, we can expand the function into Taylor series around $t=0$:

$$\phi(t) = \phi(0)+\phi'(0)t+\phi^{(2)}(0)\frac{t^2}{2}+\text{h.o.t.} \tag{12}$$

Here h.o.t refers to the higher order terms of the Taylor series expansion. As $n\to\infty$, the support of function $f_n(t)$, as shown in Figure 1, approaches 0. Hence, the product $\phi(t)f_n(t)$ can be approximated

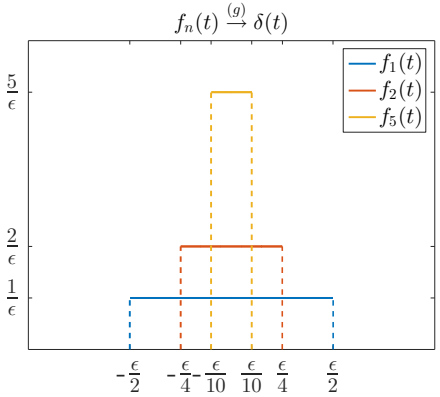


Fig. 1. An illustration showing the convergence of pulse sequences $f_n(t)$ to $\delta(t)$

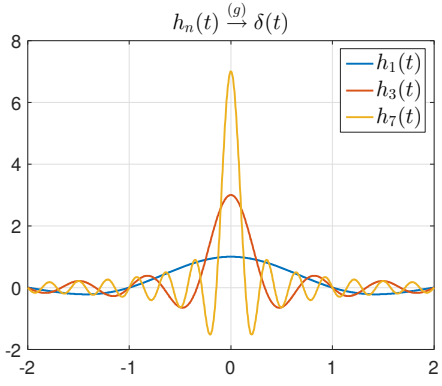


Fig. 2. An illustration showing the convergence of $h_n(t) = n\text{sinc}(nt)$ to $\delta(t)$. Convergence is not in the pointwise sense!

with the first term of the Taylor series expansion, which is $\phi(0)f_n(t)$, for large enough n. As a result, we have the equality of

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} f_n(t)\phi(t)dt = \phi(0) \tag{13}$$

in the usual calculus sense. Given the generalized limit definition, this concludes the proof of $f_n(t) \overset{(g)}{\to} \delta(t)$ as $n \to \infty$.

The definition of Dirac delta function as a generalized limit of ordinary functions is important in practice. Whenever in doubt, it is possible to replace $\delta(t)$ with $f_n(t)$ functions in (11), solve the problem of interest and then calculate the ordinary limit of the final result as $n \to \infty$. Readers are invited to do this calculation to have another verification of the scaling property in Table 1. Furthermore, the generalized limit definition establishes a connection with the "physical" interpretation of Dirac delta function as a very short duration pulse; but, readers of these notes should always keep in mind that the limit operation for getting shorter and shorter pulses is not an ordinary pointwise limit operation, as introduced in many undergraduate texts, but a generalized limit operation.

Dirac delta definition by a generalized limit argument is not specific to $f_n(t)$ given by (11). Readers can examine Problem 1.38 of [1] for some other ordinary functions whose generalized limit is $\delta(t)$. The basic requirement is the construction of a unit area function sequence with a diminishing support.

It can be verified that both

$$\begin{aligned} g_n(t) &= \sqrt{\frac{n}{2\pi}}\exp(-nt^2/2) \\ h_n(t) &= n\text{sinc(nt)} = \frac{\sin(\pi nt)}{\pi t} \end{aligned} \tag{14}$$

tend to $\delta(t)$ as $n \to \infty$ in the generalized sense.

Figure 2 shows the sketch of $h_n(t) = n\text{sinc(nt)}$ for different n values. The main lobe of the function $h_n(t)$ gets narrower and taller as n increases. Yet, however large n is, there exists two sidelobes, with a peak value about 1/5th of the maximum value on both sides of the mainlobe. Furthermore, by fixing t to a non-zero value, say t_0, and evaluating $\lim_{n\to\infty} h_n(t_0)$, we get

$$\lim_{n\to\infty} h_n(t_0) = \frac{1}{\pi t_0} \lim_{n\to\infty} \sin(\pi n t_0),$$

which does not exist in the usual sense. Hence, $h_n(t)$ does not approach to $\delta(t)$ in a manner that is as described in many undergraduate textbooks; but approaches in the generalized sense, or equivalently in the weak limit sense, [8].

Generalized Derivative of Dirac Delta Function: The derivative of Dirac delta function $d/dt\{\delta(t)\}$ is called doublet function, [1, Sec. 2.5.3]. No surprises, the differentiation operation in $d/dt\{\delta(t)\}$ is in generalized sense, that is according to introduced generalized limit definition. To understand this operation, let's examine the response of $d/dt\{\delta(t)\}$ to a test function:

$$\int_{-\infty}^{\infty} \frac{d}{dt}\{\delta(t)\}\phi(t)dt = \delta(t)\phi(t)\big|_{t=-\infty}^{t=\infty} - \int_{-\infty}^{\infty} \delta(t)\frac{d}{dt}\phi(t)dt = -\frac{d}{dt}\phi(t)\big|_{t=0}. \tag{15}$$

The calculation above is based on the application of integration-by-parts to the leftmost side of (15). Since the test function $\phi(t)$ a Gaussian function, the term $\delta(t)\phi(t)|_{t=-\infty}^{t=\infty}$ vanishes. The other term, the integral term of the integration-by-parts operation, can be expressed using the sifting property of Dirac delta function. Hence, we get the defining relation for the doublet function as

$$\int_{-\infty}^{\infty} \delta^{(1)}(t)\phi(t)dt = -\phi^{(1)}(0). \tag{16}$$

Here $\delta^{(n)}(t)$ and $\phi^{(n)}(t)$ refers to the n'th derivative of Dirac delta and test function $\phi(t)$ in the generalized and ordinary sense, respectively.

At this point, readers should be rightfully uncomfortable with the application of integration-by-parts with an integrand containing a Dirac delta function, as in (15). To the comfort of these readers (and also the ones still uneasy about the change of variables from t to $u = at$ in the scaling property discussion), we present an alternative proof path and suggest to replace $\delta(t)$ with an ordinary function $h_n(t)$ given in (14). The integration-by-parts operation with the substituted $h_n(t)$ function is now well defined and the final result becomes

$$\int_{-\infty}^{\infty} \frac{d}{dt}\{h_n(t)\}\phi(t)dt = -\int_{-\infty}^{\infty} h_n(t)\frac{d}{dt}\phi(t)dt. \tag{17}$$

By taking the limit of both sides in (17) as $n \to \infty$ and using the generalized limit definition in (10), we reach the conclusion that since $h_n(t) \overset{(g)}{\to} \delta(t)$, we have $\frac{d}{dt}\{h_n(t)\} \overset{(g)}{\to} \delta^{(1)}(t)$. The formal definition of $\delta^{(1)}(t)$ becomes the relation in (16).

Sifting and other properties for higher order derivatives of Dirac delta function are given in Table I. These results can be called advanced results; since they require more than a basic understanding of the generalized functions. Many signal processing textbooks avoid these properties, since even a partial justification of these results requires much more than a pictorial, or pointwise, justification of $\delta(t)$ function.

Derivative of Unit Step Function: By replacing $h_n(t)$ with an arbitrary regular function $f(t)$ in (17), we get

$$\int_{-\infty}^{\infty} \frac{d}{dt}\{f(t)\}\phi(t)dt = -\int_{-\infty}^{\infty} f(t)\frac{d}{dt}\phi(t)dt. \tag{18}$$

Substituting $f(t)$ in (18) with the unit step function $u(t)$ yields,

$$\begin{aligned}\int_{-\infty}^{\infty} \frac{d}{dt}\{u(t)\}\phi(t)dt &= -\int_{-\infty}^{\infty} u(t)\frac{d}{dt}\phi(t)dt \\ &= -\int_{0}^{\infty} \frac{d}{dt}\phi(t)dt \\ &\overset{(a)}{=} \phi(0) - \cancel{\phi(\infty)} \\ &= \langle \delta(t), \phi(t)\rangle \end{aligned} \tag{19}$$

where $\phi(\infty) = 0$ is used in line-(a) which is due to the test function class definition. Left-most and right-most sides of (19) imply that $\langle \frac{d}{dt}u(t), \phi(t)\rangle = \langle \delta(t), \phi(t)\rangle$ for all test functions. This statement is equivalent to $\frac{d}{dt}u(t) \overset{(g)}{=} \delta(t)$.

From this discussion, we reach the important conclusion that an ordinary function, such as $u(t)$, when interpreted as a generalized function has derivatives of all orders. In order words, function $u(t)$ is not differentiable function due to its discontinuity at $t = 0$, but it is differentiable of all orders in the generalized sense.

V. Application Examples

A number of examples are presented to illustrate the application of Dirac delta function. Our goal is to relate the applications to the generalized definitions on functions, limits, derivatives etc.

Example 1: Assume that a sequence $y[n]$ is formed by downsampling $x[n]$ by 2, $y[n] = x[2n]$. It is well known that the spectrum of $y[n]$, $Y(e^{j\omega})$, is related to the spectrum of $x[n]$, $X(e^{j\omega})$, according to the relation [2, Sec. 3.6.1]:

$$Y(e^{j\omega}) = \frac{1}{2}\left(X(e^{j\frac{\omega}{2}}) + X(e^{j(\frac{\omega}{2}+\pi)})\right). \tag{20}$$

In this example, we would like to illustrate the validity of this expression for $x[n] = \exp(j\omega_0 n)$. This exercise is quite trivial from the time-domain processing viewpoint. Since $y[n] = x[2n] = \exp(j2\omega_0 n)$, $y[n]$ is a complex exponential whose frequency is doubled after downsampling. The frequency domain representation of $x[n]$ and $y[n]$ are $X(e^{j\omega}) = 2\pi\delta(\omega-\omega_0)$, $Y(e^{j\omega}) = 2\pi\delta(\omega-2\omega_0)$, respectively. This example aims to verify this basic result directly from (20).

It should be remembered that the expressions for $X(e^{j\omega})$ and $Y(e^{j\omega})$ are periodic with 2π, as the notation implies. Let's check the validity of (20) for $X(e^{j\omega}) = 2\pi\delta(\omega - \omega_0)$:

$$\begin{aligned} Y(e^{j\omega}) = \frac{1}{2}\left(X(e^{j\frac{\omega}{2}}) + X(e^{j(\frac{\omega}{2}+\pi)})\right) &= \frac{2\pi}{2}\left[\delta\left(\frac{\omega}{2} - \omega_0\right) + \delta\left(\frac{\omega}{2} + \pi - \omega_0\right)\right] \\ &= \frac{2\pi}{2}\left[\delta\left(\frac{\omega - 2\omega_0}{2}\right) + \delta\left(\frac{\omega + 2\pi - 2\omega_0}{2}\right)\right] \\ &\overset{(a)}{=} \frac{2\pi}{2}\left[2\delta(\omega - 2\omega_0) + 2\delta(\omega + 2\pi - 2\omega_0)\right] \\ &\overset{(b)}{=} 2\pi\delta(\omega - 2\omega_0). \end{aligned}$$

In line-(a), we have used the scaling property of the Dirac delta function from Table I. In line-(b), the expression is rewritten to cover only a single 2π period, following the convention. As expected, the final result indeed matches the earlier result found from time-domain considerations.

Comment: Spectrum after downsampling operation is typically found with a frequency domain sketch that indicates the support of $X(e^{j\omega})$ and its translated versions, (see [2, Fig. 3.18]). Such a sketch is also useful to illustrate the aliasing concept. We see that when spectrum involves a Dirac

delta function, a sketch is not sufficient to explain the vanishing $1/2$ coefficient in (20). We need to bring the scaling property of the Dirac delta function into the play.

Example 2: Let X be a random variable with the probability density function (pdf) $f_X(x)$. The problem of interest is the pdf of the random variable $Z = X^2$.

This is a standard probability problem and we would like to illustrate the utility of Dirac delta function in this calculation:

$$\begin{aligned} f_Z(z) &\overset{(a)}{=} \int f_{X,Z}(x,z)dx \\ &\overset{(b)}{=} \int f_X(x) f_{Z|X}(z|X=x)dx \\ &\overset{(c)}{=} \int f_X(x)\delta(z-x^2)dx \\ &\overset{(d)}{=} \int f_X(x)\delta(x^2-z)dx \\ &\overset{(e)}{=} \int f_X(x)\left(\frac{\delta(x-\sqrt{z})}{2\sqrt{z}} + \frac{\delta(x+\sqrt{z})}{2\sqrt{z}}\right)dx \quad (\text{with } z \geq 0) \\ &\overset{(f)}{=} \frac{1}{2\sqrt{z}}\left(f_X(\sqrt{z}) + f_X(-\sqrt{z})\right). \end{aligned} \tag{21}$$

Line-(a) is the marginalization operation. Line-(b) includes a factorization for the joint density in terms of the conditional density. Line-(c) introduces $Z = X^2$ into the calculation. Line-(d) is due to evenness of Dirac delta function, $\delta(x) = \delta(-x)$. Line-(e) uses the scaling property of Table I (from the advanced section of the table). It is important to note that the integration variable in line-(d) is x. Hence, for the function $\delta(x^2 - z)$ appearing in the integrand, x is the variable and z is just a constant value. Therefore, the scaling property of Dirac delta function should be utilized by treating function $x^2 - z$ as a function of the variable x. Line-(f) is due to the sifting property.

Comment: We observe that inclusion of Dirac delta in the operational calculus results in significant shortening of algebra. Note that the calculation given above exactly mimics a similar calculation given for the discrete random variables (probability mass functions). More specifically, line-(c) of 21 can be interpreted as follows: Let's assume that $z = 100$ and consider the integral $\int_{-\infty}^{\infty} f_X(x)\delta(z-x^2)dx$. Since the function $\delta(z-x^2)$ is equal to 0 when $z \neq x^2$, this integral corresponds to checking all $x \in (-\infty, \infty)$ to find the ones satisfying the condition $x^2 = z = 100$ and "summing up" $f(x)$ values corresponding to these x values. The main difficulty of instructors is not this interpretation; but explaining the factor $1/2\sqrt{z}$, which is the Jacobian term arising during the functional mapping of random variables. The Jacobian term does not arise in discrete random variables and the "summing up" interpretation becomes exactly correct; that is, for the probability mass functions, the sum of the probability values for x that satisfy the condition $z = x^2$ gives the probability of z. With the inclusion of Dirac delta function in the calculus, $1/2\sqrt{z}$ term in line-(f) of (21) effortlessly comes out with the application of the scaling property.

Example 3: Let X and Y be two random variables with the joint pdf $f_{X,Y}(x,y)$. The problem is the derivation of the pdf for the random variable $Z = X + Y^2$.

$$\begin{aligned} f_Z(z) &\overset{(a)}{=} \int_x \int_y f_{X,Y}(x,y)\delta(z-x-y^2)dydx \\ &\overset{(b)}{=} \int_y \left(\int_x f_{X,Y}(x,y)\delta(z-x-y^2)dx\right)dy \\ &\overset{(c)}{=} \int_y \left(\int_x f_{X,Y}(x,y)\delta(x-z+y^2)dx\right)dy \\ &\overset{(d)}{=} \int_y f_{X,Y}(z-y^2,y)dy \\ &\overset{(e)}{=} \int_y f_Y(y) f_{X|Y}(z-y^2|Y=y)dy. \end{aligned} \tag{22}$$

Line-(a) is the "summing up" operation of $f_{X,Y}(x,y)$ values for which the condition $z = x + y^2$ is satisfied. In line-(b), the order of integration is exchanged, that is the inner integration is with respect to x after the exchange. Line-(c) is due to evenness of $\delta(x)$. Line-(d) is due to the sifting property. Line-(e) is the factorization of joint density in terms of conditional density of X given Y.

Comment: By changing the integration order in line-(c), the variable for the function $\delta(z-x-y^2)$ becomes x. After the order change, the variables z and y are treated as constants and we have the result in line-(d). If the inner integral in line-(c) were with respect to the variable y, that is if we do not change the order of integration; we need to use result given in Example 2 to evaluate the integral involving $\delta(z-x-y^2)$.

Example 4: Show that Fourier transform of $f(t) = 1$ is $F(\Omega) = 2\pi\delta(\Omega)$, where $F(\Omega) = \mathcal{F}\{f(t)\} = \int_{-\infty}^{\infty} f(t)\exp(-j\Omega t)dt$ is the Fourier transformation operation.

Freshman calculus results state that $\int_{-\infty}^{\infty} f(t)\exp(-j\Omega t)dt$ does not converge for any Ω for $f(t) = 1$. Hence, well-known Fourier transform pair of $1 \leftrightarrow 2\pi\delta(\Omega)$ should be interpreted in the generalized sense. To show $\mathcal{F}\{1\} \overset{(g)}{=} 2\pi\delta(\Omega)$, we need to examine the response of the function $F(\Omega) = \mathcal{F}\{1\}$ to a test function $\Phi(\Omega)$:

$$\begin{aligned} \langle \mathcal{F}\{1\}, \Phi(\Omega)\rangle &\overset{(a)}{=} \int_\Omega \mathcal{F}\{1\}\Phi(\Omega)d\Omega \\ &\overset{(b)}{=} \int_\Omega \left(\int_t 1e^{-j\Omega t}dt\right)\Phi(\Omega)d\Omega \\ &\overset{(c)}{=} \int_t \left(\int_\Omega \Phi(\Omega)e^{-j\Omega t}d\Omega\right)dt \\ &\overset{(d)}{=} \int_t 2\pi\phi(-t)dt \\ &\overset{(e)}{=} 2\pi\Phi(0) \\ &\overset{(f)}{=} 2\pi\int \delta(\Omega)\Phi(\Omega)d\Omega \\ &\overset{(g)}{=} \langle 2\pi\delta(\Omega), \Phi(\Omega)\rangle \end{aligned} \tag{23}$$

Line-(a) is due to the linear functional definition. Line-(b) results from the definition of Fourier transform. Line-(c) is the change of integration order. Line-(d) is due to the inverse Fourier transform relation for the ordinary, absolutely integrable functions, [1]. (With the assumed test function class (Gaussian functions), the Fourier integral, that is $F\{\phi(t)\} = \Phi(\Omega)$, is guaranteed to converge in the ordinary calculus sense.) Line-(e) is due to the fact that $\Phi(0) = \int_t \phi(t)dt$, that is the area of the time domain function is the value of its Fourier representation at $\Omega = 0$. Lines-(f) and (g) are different ways of writing line-(e).

Considering the left-most and right-most sides of (23) and remembering the generalized equality definition in (2), we can conclude the proof of $\mathcal{F}\{1\} \stackrel{(g)}{=} 2\pi\delta(\Omega)$.

Comment: In a first course, this relation is given by finding the inverse Fourier transform of $2\pi\delta(\Omega)$, i.e. $\mathcal{F}^{-1}\{2\pi\delta(\Omega)\}$, without mentioning the existence of Fourier integral for $f(t) = 1$. The Fourier integral for $f(t) = 1$ diverges in the usual sense; but only exists in the generalized sense or in the sense of distributions.

Example 5: Show that Fourier transform of $f(t) = \text{sgn}(t)$ is $F(\Omega) \stackrel{(g)}{=} 2/j\Omega$.

The Fourier transform of $\text{sgn}(t)$

$$\text{sgn}(t) = \begin{cases} 1 & t > 0 \\ -1 & t < 0 \end{cases}$$

can be written as the integral

$$\mathcal{F}\{\text{sgn}(t)\} = \frac{2}{j}\int_0^\infty \sin(\Omega t)dt, \tag{24}$$

which does not converge in the ordinary calculus sense. Hence, as suspected, $\mathcal{F}\{\text{sgn}(t)\}$ is equal to $2/j\Omega$ in the distribution sense. It is interesting to note that there is no Dirac delta function in the expression $\mathcal{F}\{\text{sgn}(t)\} \stackrel{(g)}{=} 2/j\Omega$ immediately giving away that the equality is in the generalized sense.

Let's define a regular function $g_T(\Omega)$ as $g_T(\Omega) = \int_0^T \sin(\Omega t)dt = \frac{1-\cos(\Omega T)}{\Omega}$. We would like to take the limit of $g_T(\Omega)$ as $T \to \infty$ with the goal of evaluating the transform in (24). To do that, we need to examine the response of $g_T(\Omega)$ to a test function $\Phi(\Omega)$, that is $\langle g_T(\Omega), \Phi(\Omega)\rangle$, and then evaluate the limit of the response as $T \to \infty$.

For a fixed T, $\langle g_T(\Omega), \Phi(\Omega)\rangle$ can be expressed as:

$$\begin{aligned} \langle g_T(\Omega), \Phi(\Omega)\rangle &= \langle \frac{1}{\Omega}, \Phi(\Omega)\rangle - \langle \frac{\cos(\Omega T)}{\Omega}, \Phi(\Omega)\rangle \\ &= \langle \frac{1}{\Omega}, \Phi(\Omega)\rangle - \langle \cos(\Omega T), \frac{\Phi(\Omega)}{\Omega}\rangle. \end{aligned} \tag{25}$$

As $T \to \infty$, the equality in (25) approaches

$$\lim_{T\to\infty} \langle g_T(\Omega), \Phi(\Omega)\rangle = \langle \frac{1}{\Omega}, \Phi(\Omega)\rangle - \lim_{T\to\infty} \langle \cos(\Omega T), \frac{\Phi(\Omega)}{\Omega}\rangle. \tag{26}$$

From (26) it is clear that we need to show $\lim_{T\to\infty}\langle \cos(\Omega T), \frac{\Phi(\Omega)}{\Omega}\rangle = 0$ to conclude the proof. Since the test function class is the class of Gaussian functions, the function $\Phi(\Omega)/\Omega$ is absolutely integrable in $\Omega \in (-\infty, \infty)$ in the Cauchy principle value sense, (Cauchy principle value integral is required due to the singularity of $\Phi(\Omega)/\Omega$ at $\Omega = 0$, [4, p.359]). We know from Dirichlet conditions that the Fourier transform of an absolutely integrable function exists in the regular sense, [1, p.290]. An important but less known fact by the signal processing audience is the Riemann-Lebesgue lemma stating that if $x(t)$ is absolutely summable, then $X(\Omega) \to 0$, as $\Omega \to \infty$, [3, p.278]. Armed with this knowledge, $\langle \cos(\Omega T), \frac{\Phi(\Omega)}{\Omega}\rangle$ can be interpreted as the real part of the $\mathcal{F}\{\frac{\Phi(\Omega)}{\Omega}\}$ with the transform domain variable T. Then due to the absolute integrability of $\Phi(\Omega)/\Omega$ and the Riemann-Lebesgue lemma, we have $\lim_{T\to\infty}\langle \cos(\Omega T), \frac{\Phi(\Omega)}{\Omega}\rangle = 0$.

By multiplying both sides of (26) by $2/j$ and replacing $g_T(\Omega)$ with $\int_0^T \sin(\Omega t)dt$, we reach

$$\lim_{T\to\infty} \int_{-\infty}^{\infty} \left(\frac{2}{j}\int_0^T \sin(\Omega t)dt\right)\Phi(\Omega)d\Omega = \int_{-\infty}^{\infty}\left(\frac{2}{j\Omega}\right)\Phi(\Omega)d\Omega \tag{27}$$

stating that $\mathcal{F}\{\text{sgn}(t)\} \stackrel{(g)}{=} 2/j\Omega$ via the generalized limit definition given in (10).

Comment: A first course in signal processing needs to sugar-coat some definitions and even some calculations due to pedagogical reasons. Among these, Fourier transformation of sign function and unit step function stands out. The sign function is clearly not absolutely or square summable; hence its Fourier transform can not be given in the usual sense. In spite of that, to show this result some instructors calculate the Fourier transform of a regular, absolutely summable function $\text{sgn}(t)e^{-\alpha|t|}$ and then evaluate the limit of the result as $\alpha \to 0$ and then present the limit as the Fourier transform of $\text{sgn}(t)$. The final result of this calculation matches the correct result; but the road taken to reach the final result, that is the intermediate steps, is dubious due to the existence problem of Fourier transform for $f(t) = \text{sgn}(t)$ and the the exchange of limit and Fourier transform operations in the last step. It should be clear that treatment of integrals diverging in the usual sense require much more effort to make sense/use of them in calculations. The generalized equality definition is an effort in this line.

As expected, the Fourier transform of $u(t)$ is also only valid in the generalized sense. By expressing $u(t)$ as $u(t) = (\text{sgn}(t) + 1)/2$ and applying the linearity of the Fourier transform, we can show $\mathcal{F}\{u(t)\} \stackrel{(g)}{=} 1/j\Omega + \pi\delta(\Omega)$.

Example 6: Find the inverse unilateral Laplace transform of $X(s) = s^2/(s+3)$.

This problem is typically solved by partial fraction expansion, that is

$$X(s) = \frac{s^2}{s+3} = s - 3 + \frac{9}{s+3}, \tag{28}$$

followed by inverse Laplace transformation via transform pair recognition. The final answer of this example is $x(t) = \delta^{(1)}(t) - 3\delta(t) + 9\exp(-3t)u(t)$. Our goal is to derive the same result via some alternative paths to illustrate the usage of generalized differentiation.

Let's first express $X(s)$ as $X(s) = s^2 X_p(s)$ where $X_p(s) = 1/(s+3)$. The inverse Laplace transform of $X_p(s)$ is $x_p(t) = 3\exp(-3t)u(t)$. Hence, the inverse Laplace transform $X(s) = s^2 X_p(s)$ becomes $x(t) = \frac{d^2}{dt^2}x_p(t)$. We can verify this result by remembering that the unilateral Laplace transform of $\frac{d}{dt}x(t)$ is $sX(s) - x(0^-)$. Note that in $x_p(t)$ and its derivatives are all zero at $t = 0^-$ due to the existence of $u(t)$ term in $x_p(t)$. Let's evaluate the first two derivatives of $x_p(t)$ and compare the result with the answer by partial fraction expansion:

$$\begin{aligned}
x_p^{(1)}(t) &= \frac{d}{dt}\{\exp(-3t)u(t)\} \\
&\overset{(a)}{=} \frac{d}{dt}\{\exp(-3t)\}u(t) + \exp(-3t)\frac{d}{dt}\{u(t)\} \\
&= -3\exp(-3t)u(t) + \exp(-3t)\delta(t) \\
&\overset{(b)}{=} -3\exp(-3t)u(t) + \delta(t), \\
x_p^{(2)}(t) &= \frac{d}{dt}x_p^{(1)}(t) = \frac{d}{dt}\left\{-3\exp(-3t)u(t) + \delta(t)\right\} \\
&\overset{(a)}{=} \frac{d}{dt}\{-3\exp(-3t)\}u(t) - 3\exp(-3t)\frac{d}{dt}\{u(t)\} + \frac{d}{dt}\{\delta(t)\} \\
&= 9\exp(-3t)u(t) - 3\exp(-3t)\delta(t) + \delta^{(1)}(t) \\
&\overset{(b)}{=} 9\exp(-3t)u(t) - 3\delta(t) + \delta^{(1)}(t). \qquad (29)
\end{aligned}$$

Line-(a) of both equations are due to the product rule for differentiation and the generalized equality of $\frac{d}{dt}u(t) \overset{(g)}{=} \delta(t)$. Line-(b) is due to the multiplication property of Dirac delta function from Table I. Note that equalities given in (29) are not ordinary equalities, but valid only in the generalized sense. The absence of $\overset{(g)}{=}$ symbol can be a source of inconsistencies and confusions; Yet, we go back to the conventional notation and symbols in this last example.

As a final exercise, let's redo the calculation by evaluating the second derivative of $x_p(t) = f(t)g(t)$ with $f(t) = \exp(-3t)$ and $g(t) = u(t)$ via the Leibniz's generalized product rule, $\frac{d^n}{dt^n}\{f(t)g(t)\} = \sum_{k=0}^{n}\binom{n}{k}f^{(n-k)}(t)g^{(k)}(t)$:

$$\begin{aligned}
x_p^{(2)}(t) &= f^{(2)}(t)g(t) + 2f^{(1)}(t)g^{(1)}(t) + f(t)g^{(2)}(t) \\
&= 9\exp(-3t)u(t) + 2(-3\exp(-3t))\delta(t) + \exp(-3t)\delta^{(1)}(t) \\
&\overset{(a)}{=} 9\exp(-3t)u(t) - 6\delta(t) + [\delta^{(1)}(t) + 3\delta(t)] \\
&= 9\exp(-3t)u(t) - 3\delta(t) + \delta^{(1)}(t). \qquad (30)
\end{aligned}$$

In line-(a), the basic and advanced version of the product rule in Table I is applied. The advanced product rule states that $f(t)\delta^{(1)}(t) = f(0)\delta^{(1)}(t) - f^{(1)}(0)\delta(t)$ and substituting $f(t) = \exp(-3t)$ into this relation gives the term in the square brackets of line-(a).

We see that the final result given either by (29) and (30) matches the one by the partial fraction expansion provided that we handle the differentiate $x_p(t)$ in the generalized sense and also respect the rules of Dirac delta function manipulation.

VI. What We Have Learned

We have studied generalized functions, generalized limit, generalized derivatives and their application in some signal processing problems. These notes aim to show that many familiar equalities are only valid in the generalized sense. Hence, the equality signs should be replaced with $\overset{(g)}{=}$ in many calculations involving Dirac delta functions, unit step functions etc. Interested readers can examine classical signal processing textbooks of Papoulis [3] and Bracewell [4] for a brief treatment of generalized functions. For more information, readers are invited to examine [7, 9, 10].

VII. Acknowledgments

Author would like to thank Prof. Bülent Sankur of Boğaziçi University, Istanbul, Turkey for his kind suggestions and comments.

VIII. Author

Çağatay Candan (ccandan@metu.edu.tr) is a professor at the Department of Electrical and Electronics Engineering in Middle East Technical University, Ankara, Turkey.

References

[1] A. V. Oppenheim, A. S. Willsky, and S. H. Nawab, *Signals & Systems (2nd Edition)*. Prentice Hall, 1999.

[2] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*. Prentice Hall, 1999.

[3] A. Papoulis, *The Fourier Integral and Its Applications*. McGraw-Hill, 1962.

[4] R. Bracewell, *The Fourier Transform & Its Applications (3rd Edition)*. McGraw-Hill, 1999.

[5] Wikipedia contributors, "Freeman Dyson — Wikipedia," 2020, [Online; accessed 29-July-2020]. [Online]. Available: https://en.wikipedia.org/wiki/Freeman_Dyson

[6] Y. Yamamoto, *From Vector Spaces to Function Spaces: Introduction to Functional Analysis with Applications*. Society for Industrial and Applied Mathematics, 2012.

[7] M. J. Lighthill, *An Introduction to Fourier Analysis and Generalised Functions*. Cambridge University Press, 1952, (online re-publication 2012).

[8] D. G. Luenberger, *Optimization by Vector Space Methods*. Wiley-Interscience, 1969.

[9] G. Temple, "Theories and applications of generalized functions," *Journal of the London Mathematical Society*, vol. s1-28, no. 2, pp. 134–148, 1953. [Online]. Available: https://londmathsoc.onlinelibrary.wiley.com/doi/abs/10.1112/jlms/s1-28.2.134

[10] ——, "The theory of generalized functions," *Proc. R. Soc. Lond. A*, no. 228, pp. 175–190, 1955. [Online]. Available: https://royalsocietypublishing.org/doi/10.1098/rspa.1955.0042

EK-4

Covariance Matrix Estimation of Compound-Gaussian Vectors With Texture Correlation

Çağatay Candan, Frédéric Pascal

Abstract

Covariance matrix estimation of compound-Gaussian vectors is studied for texture-correlated snapshots. The texture component of compound-Gaussian model is modeled with a Markov structure with states representing the texture value and transition probabilities establishing the correlation. A covariance matrix estimation solution is given for both noiseless and noisy snapshots. The suggested method can also be utilized in structured covariance matrix estimation with some simple modifications. The derivation for the persymmetric matrices is given. Numerical results indicate that the benefit of correlation information can be significant especially when the total number of snapshots is small. From applications viewpoint, the suggested model is well suited for the adaptive target detection application in sea-clutter environment where spatial correlation between range cells have been observed. The performance improvements with the suggested method especially for small number of snapshots (secondary data) make the method attractive in this application area. Some numerical results on the adaptive detection application are also given.

Index Terms

Covariance Matrix Estimation, Tyler's Estimator, Sample Covariance Matrix, Adaptive Radar Detectors

I. Introduction

The problem of accurate and reliable (outlier robust) covariance matrix estimation from a limited number of snapshots is critically important in many applications. The problem is typically treated for jointly Gaussian multivariates (Gaussian vectors); but in some applications, such as the clutter modeling, the total power of a snapshot vector can fluctuate significantly from snapshot-to-snapshot

Ç. Candan is with the Department of Electrical and Electronics Engineering, Middle East Technical University (METU), Ankara, Turkey. F. Pascal is with Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systémes, 91190, Gif-sur-Yvette, France. e-mail: {ccandan@metu.edu.tr, frederic.pascal@centralesupelec.fr}

leading to deviations from the standard model. For such applications the outlier robust estimators from statistics literature, such as Tyler's estimator [1], normalized sample covariance matrix [2] and Huber's M-estimator [3, 4], have been studied in depth to include some application specific needs in the design such as sample deficiency, low-rank constraint, known matrix structure, knowledge-aided operation [5–10]. The common assumption in these studies is the statistical independence of snapshot vectors. In this study, we examine the case of correlated snapshots for the covariance matrix estimation and present a versatile solution for noisy/noiseless snapshots. The case of noisy correlated snapshots is of practical importance for airborne/seaborne adaptive radar detectors which is the main motivation for this study.

Adaptive radar detectors aim to directly or indirectly estimate the interference (clutter plus noise) covariance matrix from a set of snapshots called secondary data set [11, 12]. More specifically, Kelly's test [13] and Adaptive Coherence Estimator (ACE) [14] indirectly estimate the unknown interference covariance matrix in the generalized likelihood ratio test (GLRT) framework. These two tests are similar but have important differences in their assumptions. Kelly's test assumes that the covariance matrix of the primary data, which is also called cell-under-test (CUT), is identical to the covariance matrix of the secondary data cells (homogeneous environment assumption). ACE assumes that covariances matrices of secondary data cells are identical; but differ from the CUT covariance matrix by an unknown scaling factor (partially homogeneous environment assumption), [11, Sec. 2.3]. Hence, ACE assumes that the primary cell can be at different average power level than secondary data cells. In literature, the compound-Gaussian models are suggested to take into account such power fluctuations [15]. For example, the clutter snapshot vector $\mathbf{x}_k$ can be modeled as a compound-Gaussian vector $\mathbf{x}_k = \sqrt{\tau_k}\mathbf{u}_k$ where τ_k (texture component) is a random variable independent of Gaussian vector $\mathbf{u}_k$ (speckle component). With the compound-Gaussian model, the texture parameter τ_k in Kelly's test becomes a global constant for both CUT and secondary data cells. ACE also assumes a constant texture parameter for the secondary data; but CUT texture parameter is independent of this constant. Note that both methods assume a constant texture value for secondary data cells. The constant texture parameter assumption is valid when the shape parameter ν of K-distributed clutter is arbitrarily large (as $\nu \to \infty$, the K-distribution approaches Rayleigh distribution with constant power [16]) or when the texture is fully correlated (spatial correlation) across the data cells. It is known that homogenous or partially homogenous environment assumption is not satisfied, or satisfied for only a limited number cells in spiky clutter environment [17, Fig.2 and Fig.3], [16, Sec. 4.3.2]. Hence, Kelly's test and ACE operate in mismatch with observed statistical description of sea-clutter. This mismatch is considered to be at least partially responsible for the observed deviation of the experimental results from theoretical predictions, [18, 19].

This study suggests a novel compound-Gaussian model whose texture-correlation is induced through

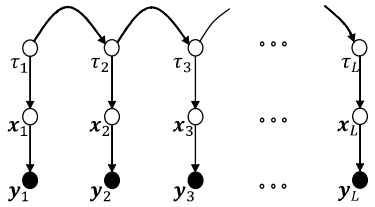


Fig. 1. Graphical model for texture-correlated compound-Gaussians

a Markov chain. From an adaptive detection perspective, the states of the chain can be considered to reflect the power ratio between CUT and a secondary data cell. This ratio is 1 and a deterministic unknown for Kelly's and ACE tests, respectively. Here, we allow this ratio to be a random variable and study the covariance matrix estimation for this set-up. The suggested Markov model is also in agreement with some empirical observations on sea-clutter where a sequence of high clutter power cells are observed to be followed by a sequence of low powered cells, as shown in Figure 2.21 from [16]. Also, it is possible to take into account the experimentally observed spatial texture correlation of sea clutter [17, Fig.2 and Fig.3] with the suggested model.

The main contribution of this study is the suggested hidden Markov model (HMM) based compound-Gaussian model for the texture-correlated compound-Gaussian snapshots and the development of the maximum likelihood covariance matrix estimator, via expectation maximization method. The suggested method is applicable for structured covariance matrix estimation and we also study the important case of persymmetric covariance matrix estimation in this study.

II. HMM Based Texture-Correlated Compound-Gaussian Model

Compound-Gaussian Signal Model: The snapshot vector $\mathbf{x}_k$ is formed by the product of a scalar variable $\sqrt{\tau_k}$ (texture amplitude) and $N \times 1$ dimensional vector $\mathbf{c}_k$ (speckle)

$$\mathbf{x}_k = \sqrt{\tau_k}\mathbf{c}_k, \quad k = \{1, \ldots, L\}. \tag{1}$$

The speckle component $\mathbf{c}_k$ is assumed to be independent and identically distributed (iid) with circularly symmetric complex Gaussian random vector with zero mean and covariance matrix $\mathbf{M}$, $\mathbf{c}_k \sim CN(\mathbf{0}, \mathbf{M})$. It is assumed that τ_k random variable is discrete valued (possibly formed by quantizing a continuous random variable) and generated according to the Markov structure:

$$p(\tau_1, \tau_2, \ldots, \tau_L) = p(\tau_1) \prod_{l=2}^{L} p(\tau_l | \tau_{l-1}). \tag{2}$$

The state-transition probabilities of the Markov chain is denoted as $\pi_{ij} = P(\tau_k = \kappa_i \,|\, \tau_{k-1} = \kappa_j)$ where $\{\kappa_1, \ldots, \kappa_S\}$ denotes the sample space of the random variable τ_k. It is easy to see from (1) that the snapshot vector $\mathbf{x}_k$ given τ_k has the distribution of $\mathbf{x}_k \mid \tau_k \sim CN(\mathbf{0}, \tau_k \mathbf{M})$.

One of the main goals of this study is the maximum-likelihood estimation of the covariance matrix $\mathbf{M}$ given the observations $\mathbf{x}_k$, $k = \{1, \ldots, L\}$. The joint-density of observations can be given as

$$f(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L) = \sum_{\tau_1, \tau_2, \ldots, \tau_L} \prod_{l=1}^{L} f(\mathbf{x}_l | \tau_l) p(\tau_1, \ldots, \tau_L) \tag{3}$$

where the summation is over the set of texture variables. Note that, the argument of the sum in (3) is the joint density $f(x_1, \ldots, x_L, \tau_1, \ldots, \tau_L)$ and (3) is the marginalization of the joint density over the random texture variables. We give an expectation-maximization based solution to the problem, by treating the texture variables as hidden variables in this study.

In many applications, the snapshot vector $\mathbf{x}_k$ is not observed, but its noisy version $\mathbf{y}_k$ is available:

$$\mathbf{y}_k = \mathbf{x}_k + \mathbf{n}_\mathbf{k}, \quad k = \{1, 1, \ldots, L\}. \tag{4}$$

Here the noise vector $\mathbf{n}_\mathbf{k}$ is assumed to be independent of all other random variables and is Gaussian distributed, $\mathbf{n}_k \sim CN(\mathbf{0}, \sigma_n^2 \mathbf{I})$. The graphical model for the random variables in this model is given in Figure 1. For this problem, the covariance matrix estimation reduces to the estimation of $\mathbf{M}$ given the noisy snapshots $\mathbf{y}_k$, $k = \{1, \ldots, L\}$.

In the context of adaptive detectors, the first cell, that is $\mathbf{y}_1$, can be considered to be CUT with the texture parameter τ_1. The other cells correspond to the secondary data cells with texture parameters τ_k. The texture variables are in correlated according to the density given in (2). In Section IV of this study, we suggest a method for the construction of the Markov structure given the clutter power fluctuations, that is the training data.

Identifiability Problem: Since the snapshot $\mathbf{x}_k = \sqrt{\tau_k}\mathbf{c}_k$ is the product of random variable τ_k and random vector $\mathbf{c}_k$, the factors of the product can be identified apart from a constant scaling factor. Stated differently, the substitution of τ_k with $\alpha\tau_k$, that is $\tau_k \leftarrow \alpha\tau_k$, and $\mathbf{c}_k \leftarrow \alpha^{-1}\mathbf{c}_k$ yields an identical product for $\tau_k \mathbf{c}_k$. Hence, the unknown covariance matrix $\mathbf{M}$ can only identified apart from a constant in this setting. To uniquely identify $\mathbf{M}$, a constraint on $\mathbf{M}$, such as $\mathrm{tr}\{\mathbf{M}\} = N$, is required. Without any loss of generality, we assume that the texture variable is scaled such that $E\{\tau_k\} = 1$.

III. Covariance Matrix Estimation for Texture Correlated Snapshots

We present an expectation-maximization based solution for the covariance matrix estimation for noise-free and noisy snapshots in this section.

A. Case 1: Noise-free Snapshots

We assume L snapshots $\mathbf{x}_k = \sqrt{\tau_k}\mathbf{c}_k, \quad k = \{1, 2, \ldots, L\}$ are given for the estimation of $\mathbf{M}$. The texture variable vector $\boldsymbol{\tau} = [\tau_1 \ \ldots \ \tau_L]$ is the hidden (unobserved) vector variable of the problem and $N \times L$ dimensional $\mathbf{X} = [\mathbf{x}_1 \ \ldots \ \mathbf{x}_L]$ is the matrix of observation vectors. For the expectation-maximization (EM) formulation, we specify the complete data-set as $\mathbf{X}_c = \{\mathbf{X}, \boldsymbol{\tau}\}$. The log-likelihood of complete data-set, $\Lambda(\mathbf{X}_c) = \log(f(\mathbf{X}_c)) = \log(f(x_1, \ldots, x_L, \tau_1, \ldots, \tau_L))$ can be expressed as

$$\Lambda(\mathbf{X}_c) \stackrel{c}{=} -L\log(|\mathbf{M}|) - \sum_{l=1}^{L} \frac{\mathbf{x}_l^H \mathbf{M}^{-1} \mathbf{x}_l}{\tau_l}. \tag{5}$$

Here $\stackrel{c}{=}$ shows to presence of additional terms on the right hand side of equation that do not affect the subsequent optimization steps.

The expectation-step (E-step) of EM-algorithm requires the expectation calculation of $\Lambda(\mathbf{X}_c)$ over the posterior distribution of hidden variables $\boldsymbol{\tau}$ given the observation $\mathbf{X}$. E-step can be explicitly written as

$$J(\mathbf{M}, \mathbf{M}^{(k)}) = E\{\Lambda(\mathbf{X}, \boldsymbol{\tau})|\mathbf{X}; \mathbf{M}^{(k)}\} \stackrel{c}{=} -L\log(|\mathbf{M}|) - \sum_{k=1}^{L} E_{\tau_k|\mathbf{X};\mathbf{M}^{(k)}} \left\{\frac{1}{\tau_k}\right\} \mathbf{x}_k^H \mathbf{M}^{-1} \mathbf{x}_k, \tag{6}$$

where $\mathbf{M}^{(k)}$ is an estimate for the unknown covariance matrix. The maximization-step (M-step) requires the optimization of $J(\mathbf{M}, \mathbf{M}^{(k)})$ over the deterministic variable $\mathbf{M}$. By differentiating (6), we can get the maximizer as:

$$\mathbf{M}^{(k+1)} = \frac{1}{L}\sum_{l=1}^{L} E_{\tau_l|\mathbf{X};\mathbf{M}^{(k)}} \left\{\frac{1}{\tau_l}\right\} \mathbf{x}_l \mathbf{x}_l^H = \frac{1}{L}\sum_{l=1}^{L} \left(\sum_{s=1}^{S} \frac{p(\tau_l = \kappa_s|\mathbf{X}; \mathbf{M}^{(k)})}{\kappa_s}\right) \mathbf{x}_l \mathbf{x}_l^H \tag{7}$$

where $\mathbf{M}^{(k+1)}$ is the updated covariance matrix estimate. EM-algorithm is run by inserting the updated variable $\mathbf{M}^{(k+1)}$ into E-step, repeating the process until convergence. EM-algorithm is an ascent algorithm that increases the likelihood of the parameter estimate at every update. The algorithm is guaranteed to convergence, possibly to a local maxima [20].

It should be noticed that the implementation of E-step requires the posterior calculation of hidden variables given the observations. For the problem of interest, this is the posterior of texture variables given the snapshots, that is $p(\tau_l = \kappa_s|\mathbf{X}; \mathbf{M}^{(k)})$. Below, we give $\alpha\beta$ recursion [21] for the marginal posterior density calculation.

Posterior probability calculation via $\alpha\beta$ recursion: The $\alpha\beta$ recursion is one of the methods for the marginal posterior probability calculation for HMM structures shown in Figure 1, [21]. The joint density of $\{\tau_l, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l\}$ is denoted as $\alpha_l(\kappa_s) = p(\tau_l = \kappa_s, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l)$, $l = \{1, \ldots, L\}$. The α-recursion is initialized with $\alpha_1(\kappa_s) = p(\tau_1 = \kappa_s) f(\mathbf{x}_1|\tau_1 = \kappa_s; \mathbf{M}^{(k)})$ and recursively evaluated for $l \geq 2$ via

$$\alpha_l(\kappa_s) \quad = \quad f(\mathbf{x}_l \,|\, \tau_l = \kappa_s; \mathbf{M}^{(k)}) \sum_{s'=1}^{S} \pi_{ss'} \alpha_{l-1}(\kappa_{s'}). \tag{8}$$

Algorithm 1: Proposed Method for Texture-Correlated Noise-free Snapshots

Input : $\mathbf{X} = [\mathbf{x}_1 \, \mathbf{x}_2 \, \ldots \, \mathbf{x}_L]$, maxiter: maximum number of iterations

Output: $\mathbf{M}_n$, ρ

1 Initial Conditions: $k = 0$, $\mathbf{M}^{(0)} = \mathbf{I}$

2 Marginal posterior density calculation:

$\alpha_1(\tau_s) = p(\tau_1 = \kappa_s) f(\mathbf{x}_1|\tau_1 = \kappa_s; \mathbf{M}^{(k)})$

for $l = 2 : L$, $\alpha_l(\kappa_s) = f(\mathbf{x}_l \,|\, \tau_l = \kappa_s; \mathbf{M}^{(k)}) \sum_{s'=0}^{S} \pi_{ss'} \alpha_{l-1}(\kappa_{s'})$; end

$\beta_L(\kappa_s) = 1$

for $l = L : -1 : 2$, $\beta_{l-1}(\kappa_s) = \sum_{s'=1}^{S} f(\mathbf{x}_l \,|\, \tau_l = \kappa_{s'}; \mathbf{M}^{(k)}) \pi_{s's} \beta_l(\kappa_{s'})$; end

$p(\tau_l = \kappa_s \,|\, \mathbf{X}; \mathbf{M}^{(k)}) \propto \alpha_l(\kappa_s)\beta_l(\kappa_s)$

3 Covariance matrix update:

$\mathbf{M}^{(k+1)} = \frac{1}{L}\sum_{l=1}^{L} \left(\sum_{s=1}^{S} p(\tau_l = \kappa_s|\mathbf{X}; \mathbf{M}^{(k)}) \frac{1}{\kappa_s}\right) \mathbf{x}_l \mathbf{x}_l^H$

4 $k \leftarrow k + 1$

5 **if** $k <$ maxiter, goto Step-2

6 **Return** $\mathbf{M}_n = \mathbf{M}^{(k)} \times N/\mathrm{tr}\{\mathbf{M}^{(k)}\}$, $\rho = \mathrm{tr}\{\mathbf{M}^{(k)}\}/N$

Here $\pi_{ss'}$ is the state-transition probability, $\pi_{ss'} = p(\tau_l = \kappa_s|\tau_{l-1} = \kappa_{s'})$ and $f(\mathbf{x}_l \,|\, \tau_l = \kappa_s; \mathbf{M}^{(k)})$ is the likelihood value of $\mathbf{x}_l|\tau_l = \kappa_s \sim CN(\mathbf{0}, \kappa_s \mathbf{M}^{(k)})$.

The function $\beta_l(\kappa_s) = f(\mathbf{x}_{l+1}, \ldots, \mathbf{x}_L \,|\, \tau_l = \kappa_s)$ denotes a likelihood value for the l'th texture variable. β-recursion is initialized with $\beta_L(\kappa_s) = 1$ and recursively evaluated, for decreasing l from L to 2:

$$\beta_{l-1}(\kappa_s) = \sum_{s'=1}^{S} f(\mathbf{x}_l \,|\, \tau_l = \kappa_{s'}; \mathbf{M}^{(k)}) \pi_{s's} \beta_l(\kappa_{s'}). \tag{9}$$

The posterior probability is normalized version of the product $\alpha\beta$ recursions, $p(\tau_l = \kappa_s \,|\, \mathbf{X}; \mathbf{M}^{(k)}) \propto \alpha_l(\kappa_s)\beta_l(\kappa_s)$ [21].

The suggested covariance matrix algorithm is summarized Algorithm Table 1. We would like to note that the algorithm suggested does not assume any structure on $\mathbf{M}$ matrix. Yet, any additional knowledge on the structure of $\mathbf{M}$ matrix can be easily taken into account in the M-step of the algorithm. As an example, when $\mathbf{M}$ is restricted to the class of persymmetric matrices, the M-step of the suggested scheme should be adapted to restrict the search space of the maximization operation to the persymmetric matrices. In the literature, there are several structured covariance matrix estimation methods for Gaussian vectors. The results of these studies can be used immediately in the M-step of the suggested method. For the persymmetric matrix constraint, the adaptation is simply the replacement

of M-step in (7) with

$$\mathbf{M}_{\text{persym}}^{(k+1)} = \frac{1}{2L}\sum_{l=1}^{L}\left(\sum_{s=1}^{S}\frac{p(\tau_l=\kappa_s|\mathbf{X};\mathbf{M}^{(k)})}{\kappa_s}\right)\left(\mathbf{x}_l\mathbf{x}_l^H + \mathbf{x}_l^R(\mathbf{x}_l^R)^H\right) \tag{10}$$

where $\mathbf{J}$ is the permutation matrix with all ones in its anti-diagonal, [22, 23] and $\mathbf{x}_k^R = \mathbf{J}\mathbf{x}_k^*$ to denote the reversed (flipped up-down) and conjugated vector $\mathbf{x}_k$ (see Appendix for more details).

As another example, the set of feasible covariance matrices can be restricted to $\mathcal{S} = \{\mathbf{R} : \mathbf{R} = \mathbf{M}+\sigma^2\mathbf{I}, \mathbf{M} \succeq 0, \text{rank}(\mathbf{M}) \leq r\}$ as in [24] for the goal of low-rank covariance matrix estimation. The algorithm in [24] can also be utilized in the set-up of this paper by replacing the sample covariance matrix calculation step (given as $\mathbf{X}\mathbf{X}^H$ in [24]) with (7) and running the low-rank covariance matrix estimation method in [24]) after this replacement in the M-step of the suggested scheme.

B. Case 2: Noisy Snapshots

In this section, we extend the algorithm to the noisy observation case. The compound-Gaussian snapshots $\mathbf{x}_k$ are assumed to be observed according to the following model:

$$\mathbf{y}_k = \mathbf{x}_k + \mathbf{n}_k = \sqrt{\gamma_k}\mathbf{c}_k + \mathbf{n}_k, \quad k = \{1, \ldots, L\}. \tag{11}$$

Here, the noise vector $\mathbf{n}_k$ is assumed to be zero-mean white Gaussian noise, $\mathbf{n}_k \sim CN(\mathbf{0}, \sigma_n^2\mathbf{I})$. The problem statement is the estimation of covariance matrix $\mathbf{M}$ given the observation matrix $\mathbf{Y} = [\mathbf{y}_1\, \mathbf{y}_2, \ldots, \mathbf{y}_L]$. To adapt earlier solution to this solution, we include the noise-free snapshots $\mathbf{x}_k$ as the hidden variables of the problem. Hence, the complete data-set for this problem is $\mathbf{Y}_c = \{\boldsymbol{\tau}, \mathbf{X}, \mathbf{Y}\}$ with $\boldsymbol{\tau} = [\tau_1, \ldots, \tau_L]$ and $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_L]$. The joint density function of the complete data-set can be factorized as follows:

$$f(\mathbf{Y}_c) = f(\tau_{1:L}, \mathbf{x}_{1:L}, \mathbf{y}_{1:L}; \mathbf{M}) = f(\tau_1)f(\mathbf{x}_1|\tau_1;\mathbf{M})f(\mathbf{y}_1|\mathbf{x}_1)\prod_{l=2}^{L}p(\tau_l|\tau_{l-1})f(\mathbf{x}_l|\tau_l;\mathbf{M})f(\mathbf{y}_l|\mathbf{x}_l). \tag{12}$$

Here $p(\tau_l = \kappa_i \,|\, \tau_{l-1} = \kappa_j) = \pi_{ij}$ is the state-transition probabilities of the Markov chain. Note that the conditional densities $f(\mathbf{x}_l|\tau_l;\mathbf{M})$ are the only factors depending on $\mathbf{M}$, where $\mathbf{x}_l \mid \tau_l \sim CN(\mathbf{0}, \tau_l\mathbf{M})$. Due to noisy observations, we have an additional conditional density for the observation vector which is $\mathbf{y}_l \mid \mathbf{x}_l \sim CN(\mathbf{x}_l, \sigma_n^2\mathbf{I})$.

Expectation Step: The conditional expectation of complete data log-likelihood $J(\mathbf{M}, \mathbf{M}^{(k)}) = E_{\mathbf{X},\boldsymbol{\tau}|\mathbf{Y},\theta^{\text{old}}}\{\log f(\mathbf{Y}_c)\}$ is

$$\begin{aligned} J(\mathbf{M}, \mathbf{M}^{(k)}) &\overset{c}{=} \sum_{l=1}^{L} E_{\mathbf{x}_l,\tau_l|\mathbf{Y}}\left\{\log f(\mathbf{x}_l|\tau_l;\mathbf{M})\right\} \\ &\overset{c}{=} -\sum_{l=1}^{L} E_{\mathbf{x}_l,\tau_l|\mathbf{Y}}\left\{\log|\tau_l\mathbf{M}| + \frac{\mathbf{x}_l^H\mathbf{M}^{-1}\mathbf{x}_l}{\tau_l}\right\} \\ &\overset{c}{=} -L\log|\mathbf{M}| - \sum_{l=1}^{L} E_{\tau_l|\mathbf{Y}}\left\{E_{\mathbf{x}_l|\tau_l,\mathbf{Y}}\left\{\frac{\mathbf{x}_l^H\mathbf{M}^{-1}\mathbf{x}_l}{\tau_l}\right\}\right\} \end{aligned} \tag{13}$$

The inner expectation in (13), $E_{\mathbf{x}_l|\tau_l,\mathbf{Y}}\left\{\mathbf{x}_l^H\mathbf{M}^{-1}\mathbf{x}_l/\tau_l\right\}$, can be evaluated by first rewriting the argument of the expectation as $\text{tr}(\mathbf{M}^{-1}\mathbf{x}_l\mathbf{x}_l^H)/\tau_l$ and interchanging trace - expectation operations and utilizing $E_{\mathbf{x}_l|\tau_l,\mathbf{Y}}\left\{\mathbf{x}_l\mathbf{x}_l^H\right\} = \overline{\mathbf{x}}_{l,\tau_l}\overline{\mathbf{x}}_{l,\tau_l}^H + \mathbf{K}_{\tau_l}$. This result follows from the conditional density of $\mathbf{x}_l$ given $\{\tau_l, \mathbf{Y}\}$. It can be easily verified from the graphical model in Figure 1 that $\mathbf{x}_l$ is independent of $\mathbf{x}_k$, $k \neq l$ conditioned on τ_l and $\mathbf{y}_l$. Since we have $\mathbf{y}_l = \mathbf{x}_l + \mathbf{n}_l$ with $\mathbf{x}_l \sim CN(0, \tau_l\mathbf{M}^{(k)})$ and $\mathbf{n}_l \sim CN(0, \sigma_n^2\mathbf{I})$; the conditional distribution is $\mathbf{x}_l \,|\, \{\mathbf{y}_l, \tau_l\} \sim CN(\overline{\mathbf{x}}_{l,\tau_l}, \mathbf{K}_{\tau_l})$ with

$$\begin{aligned} \overline{\mathbf{x}}_{l,\tau_l} &= \tau_l\mathbf{M}^{(k)}(\tau_l\mathbf{M}^{(k)} + \sigma_n^2\mathbf{I})^{-1}\mathbf{y}_l, \\ \mathbf{K}_{\tau_l} &= \left(\frac{1}{\sigma_n^2}\mathbf{I} + \frac{1}{\tau_l}(\mathbf{M}^{(k)})^{-1}\right)^{-1}. \end{aligned} \tag{14}$$

Once the inner expectation result is inserted in (13), we have

$$\begin{aligned} J(\mathbf{M}, \mathbf{M}^{(k)}) &\overset{c}{=} -L\log|\mathbf{M}| - \sum_{l=1}^{L} E_{\tau_l|\mathbf{Y}}\left\{\frac{\overline{\mathbf{x}}_{l,\tau_l}^H\mathbf{M}^{-1}\overline{\mathbf{x}}_{l,\tau_l} + \text{tr}\left(\mathbf{M}^{-1}\mathbf{K}_{\tau_l}\right)}{\tau_l}\right\} \\ &\overset{c}{=} -L\log|\mathbf{M}| - \sum_{l=1}^{L}\sum_{s=1}^{S}\frac{p(\tau_l=\kappa_s|\mathbf{Y})}{\kappa_s}\overline{\mathbf{x}}_{l,\kappa_s}^H\mathbf{M}^{-1}\overline{\mathbf{x}}_{l,\kappa_s} \\ &\quad -L\sum_{s=1}^{S}\frac{\overline{p}(\tau=\kappa_s|\mathbf{Y})}{\kappa_s}\text{tr}\left(\mathbf{M}^{-1}\mathbf{K}_{\kappa_s}\right) \end{aligned} \tag{15}$$

where

$$\overline{p}(\tau=\kappa_s|\mathbf{Y}) = \frac{1}{L}\sum_{l=1}^{L}p(\tau_l=\kappa_s|\mathbf{Y}). \tag{16}$$

denotes the average posterior of texture variable over all snapshots. The marginal posterior probability calculation is similar to the noiseless case. The main difference in marginal probability calculation is the replacement of earlier likelihood function $f(\mathbf{x}_l|\tau_l)$ with $f(\mathbf{y}_l|\tau_l)$. Details are given in Algorithm Table 2.

Maximization Step: The maximization step involves the maximization of $J(\mathbf{M}, \mathbf{M}^{(k)})$ given by (15) with respect to $\mathbf{M}$. By differentiation, we get the extrema as

$$\mathbf{M}^{(k+1)} = \frac{1}{L}\sum_{l=1}^{L}\sum_{s=1}^{S}\frac{p(\tau_l=\kappa_s|\mathbf{Y})}{\kappa_s}\overline{\mathbf{x}}_{l,\kappa_s}\overline{\mathbf{x}}_{l,\kappa_s}^H + \sum_{s=1}^{S}\frac{\overline{p}(\tau=\kappa_s|\mathbf{Y})}{\kappa_s}\mathbf{K}_{\kappa_s}. \tag{17}$$

Note that, the covariance matrix estimation step for the noiseless and noisy snapshots, given in (7) and (17) respectively, are quite similar. In the noisy case, the MSE-optimal estimate for the snapshot vector, that is $\overline{\mathbf{x}}_{l,\kappa_s}$ is utilized instead of $\mathbf{x}_l$. Also, the estimation error on $\mathbf{x}_l$ is taken into account with the inclusion of the error covariance matrix $\mathbf{K}_{\kappa_s}$ in (17).

The maximization step generating the covariance matrix estimate in (17) does not assume any matrix structure. As in noise-free case, the maximization step can be easily updated to include the structural constraints. For example, with the persymmetry contraint, the update equation becomes

$$\mathbf{M}_{\text{persym}}^{(k+1)} = \frac{1}{2L}\sum_{l=1}^{L}\sum_{s=1}^{S}\frac{p(\tau_l=\kappa_s|\mathbf{Y})}{\kappa_s}\left(\overline{\mathbf{x}}_{l,\kappa_s}\overline{\mathbf{x}}_{l,\kappa_s}^H + \overline{\mathbf{x}}_{l,\kappa_s}^R(\overline{\mathbf{x}}_{l,\kappa_s}^R)^H\right) + \sum_{s=1}^{S}\frac{\overline{p}(\tau=\kappa_s|\mathbf{Y})}{\kappa_s}\mathbf{K}_{\kappa_s}. \tag{18}$$

Algorithm 2: Proposed Method for Texture-Correlated Noisy Snapshots

Input : $\mathbf{Y} = [\mathbf{y}_1\, \mathbf{y}_2\, \ldots\, \mathbf{y}_L]$, maxiter: maximum number of iterations

Output: $\mathbf{M}_n$, ρ

1 Initial Conditions: $k = 0$, $\mathbf{M}^{(0)} = \mathbf{I}$

2 Marginal posterior density calculation:

$\alpha_1(\tau_s) = p(\tau_1 = \kappa_s) f(\mathbf{y}_1 | \tau_1 = \kappa_s; \mathbf{M}^{(k)})$

for $l = 2 : L$, $\alpha_l(\kappa_s) = f(\mathbf{y}_l \,|\, \tau_l = \kappa_s; \mathbf{M}^{(k)}) \sum_{s'=0}^{S} \pi_{ss'} \alpha_{l-1}(\kappa_{s'})$; end

$\beta_L(\kappa_s) = 1$

for $l = L : -1 : 2$, $\beta_{l-1}(\kappa_s) = \sum_{s'=1}^{S} f(\mathbf{y}_l \,|\, \tau_l = \kappa_{s'}; \mathbf{M}^{(k)}) \pi_{s's} \beta_l(\kappa_{s'})$; end

$p(\tau_l = \kappa_s \,|\, \mathbf{Y}; \mathbf{M}^{(k)}) \propto \alpha_l(\kappa_s) \beta_l(\kappa_s)$

Snapshots:

for $s = 1 : S$,

for $l = 1 : L$,

$\overline{\mathbf{x}}_{l,\kappa_s} = \kappa_s \mathbf{M}^{(k)} (\kappa_s \mathbf{M}^{(k)} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{y}_l$

end;

$\mathbf{K}_{\kappa_s} = \left(\frac{1}{\sigma_n^2} \mathbf{I} + \frac{1}{\kappa_s} (\mathbf{M}^{(k)})^{-1} \right)^{-1}$

end;

3 Covariance matrix update:

$\mathbf{M}^{(k+1)} = \frac{1}{L} \sum_{l=1}^{L} \sum_{s=1}^{S} \frac{p(\tau_l = \kappa_s | \mathbf{Y})}{\kappa_s} \overline{\mathbf{x}}_{l,\kappa_s} \overline{\mathbf{x}}_{l,\kappa_s}^H + \sum_{s=1}^{S} \frac{\bar{p}(\tau = \kappa_s | \mathbf{Y})}{\kappa_s} \mathbf{K}_{\kappa_s}$

4 $k \leftarrow k + 1$

5 **if** $k <$ maxiter, goto Step-2

6 **Return** $\mathbf{M}_n = \mathbf{M}^{(k)} \times N / \mathrm{tr}\{\mathbf{M}^{(k)}\}$, $\rho = \mathrm{tr}\{\mathbf{M}^{(k)}\}/N$

Note that $\mathbf{M}_{\text{persym}}^{(k+1)}$ in (18) is persymmetric provided that $\mathbf{K}_{\kappa_s}$ is persymmetric and suggested EM algorithm initialization $\mathbf{M}^{(0)} = \mathbf{I}$, as in Algorithm Table 2, guarantees that.

IV. Construction of Hidden Markov Model

In almost all applications, the texture parameter τ of the compound model is a continuous random variable. For example, the frequently used K-distributed sea-clutter model utilizes the gamma distribution as the texture distribution [16]. The gamma distribution has two parameters, namely shape and scale parameters. In sea clutter modeling, the shape parameter characterizes the spikiness of the clutter. A small shape parameter ν, say $\nu = 0.5$, corresponds to a highly spiky clutter and as ν increases, the spikiness is reduced. The scale parameter of the gamma distribution is related with its spread, that is when a gamma variate is multiplied/scaled by α, the resulting random variable is

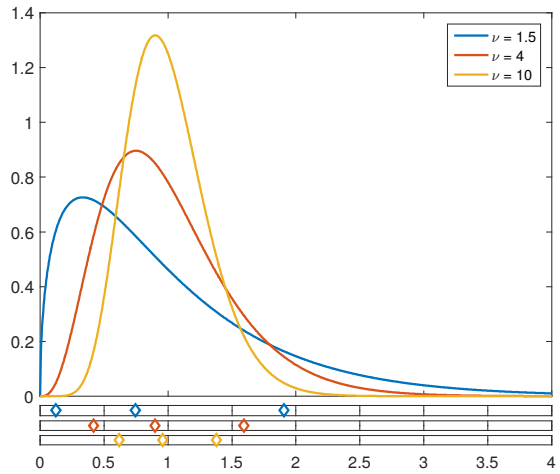


Fig. 2. Unit mean gamma distributions with shape parameters $\nu = \{1.5, 4, 10\}$ and 3 level quantization values for each distribution (shown with the matching color diamond shaped marker).

also a gamma variate with the same shape parameter, but its scale parameter is multiplied/scaled by α; hence, the name scale parameter. In this study, we consider texture variables having unit mean value. (When texture is gamma distributed, this corresponds to setting the scale parameter as $1/\nu$.) Working with a unit mean texture variable is not a limitation of the study as previously mentioned in the identifiability discussions in Section II.

The continuous valued texture random variable can be utilized in the suggested hidden Markov model (HMM) based covariance matrix estimation set-up provided that i. the texture random variable can be accurately quantized taking into account the needs of the covariance matrix estimation application; ii. the state transition probabilities of the chain can be properly assigned. The solution of these problems are studied under the topic of HMM training in the literature which can be further examined from the classic tutorial by Rabiner [25]. Here, we deviate from the standard HMM training practice for the first problem (quantization of continuous valued texture random variable) which is essentially due to the application specific needs of the covariance matrix estimation problem.

Quantization of Continuous-valued Texture Random Variable: Figure 2 shows the density of unit mean gamma distributions with different shape parameters. In the same figure, a set of 3-level quantization values for each distribution is also given. (The quantization values are indicated with

diamond markers whose color match the color of the continuous density in the figure.) From this illustrative figure, we note the choice for the number of quantization levels and their values as the first problem of HMM training [25]. For the specific application of interest, we suggest to utilize a modification of classical Lloyd-Max quantizer [26, 27].

The classical Lloyd-Max quantizer is the MSE-optimal quantizer that minimizes the quantization error, $E_x\{(x-Q(x))^2\}$ [26, 27]. Here $Q(\cdot)$ refers to the quantization operation. The main application of MSE-optimal quantizer, with the error definition specialized to the quantization error $x-Q(x)$, is the source coding. Different from the source coding problem, the continuous valued texture variable τ is not to be reconstructed in this application. In fact, τ is just a nuisance parameter of the problem whose value is not of any interest. An examination of Algorithm Table 1 should reveal that it is not the random variable τ, but its reciprocal $1/\tau$ affects the steps of estimation algorithm. More specifically, in the 3'rd step of Algorithm Table 1, the conditional mean of $1/\tau$, which is explicitly shown as $E_{\tau_k|\mathbf{X};\mathbf{M}^{(k)}}\{1/\tau_k\}$ in (6), is evaluated. Given this, we suggest to set the quantizer such that the reciprocal of τ is to be MSE-optimal represented. Hence, we suggest to optimize the following criteria $E_\tau\{(1/\tau - Q(\tau))^2\}$. An straightforward extension of the conventional Lloyd-Max quantizer to this problem leads to

$$\begin{aligned} \text{Step 1}: \quad & t_{q+1} = \frac{1}{2}(\kappa_q + \kappa_{q+1}) \quad q = \{0,\ldots,S-2\} \\ \text{Step 2}: \quad & \frac{1}{\kappa_q} = \frac{\int_{t_q}^{t_{q+1}} \frac{1}{\tau} f_\tau(\tau) d\tau}{\int_{t_q}^{t_{q+1}} f_\tau(\tau) d\tau} \quad q = \{0,\ldots,S-1\} \end{aligned} \tag{19}$$

where $t_0 \triangleq 0$ and $t_S \triangleq \infty$. In (19), κ_q, $q = \{0,\ldots,S-1\}$ indicate the quantization values and t_{q+1} indicates the separation boundary between κ_q and κ_{q+1}. Step 1 of (19) is identical to the classical Lloyd-Max quantizer, while the conditional mean of $1/\tau$, instead of τ, is calculated in Step 2. Similar to the conventional Lloyd-Max algorithm, these steps are run sequentially starting from an inital set of κ_q values, until convergence.

The quantization values shown in Figure 2 correspond to the application of modified Lloyd-Max quantizer for 3 quantization levels. The location of diamond markers on x-axis denotes the quantization values $\{\kappa_0, \kappa_1, \kappa_2\}$ for each distribution. It can be observed that the quantization values are more spread for small valued shape parameters and as shape parameter increases, the quantization levels concentrate around unity. Table I gives the numerical values for the quantization for different cases. The cases shown in 2 are marked in bold case. The numerical values reflect the skewness of the distribution towards 0 for small ν. As ν increases, the distribution becomes almost symmetric around 1; it is not exactly symmetric, since the distribution is a one-sided.

The suggested method in (19) for setting quantization values is applicable to any texture distribution with density $f_\tau(\tau)$. If the density is not available in close form, the conditional mean operation in

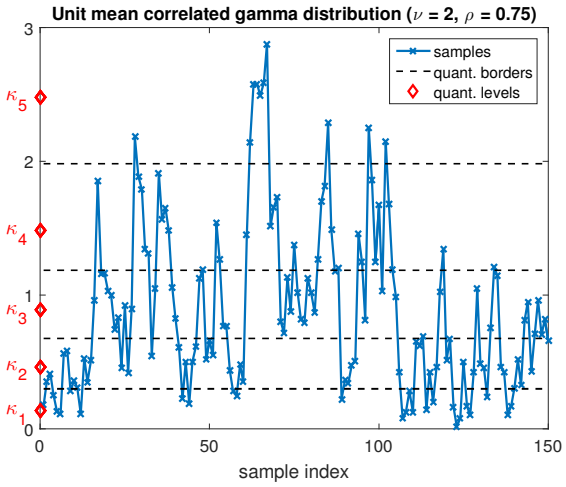


Fig. 3. Samples of unit mean correlated gamma distribution with parameters $\nu = 2$ $\rho = 0.75$ and 5-level quantization borders/values are shown.

Step 2 of (19) by evaluating the conditional mean of the empirical data. Hence, a distribution fitting to the empirical data is not required when setting the quantization values according to the suggested method.

Assignment of State Transition Probabilities: For the solution of second problem of HMM training, we follow the standard practice given in [25]. We assign the state transition probabilities according to the ratio of expected number of transitions from state-i to state-j and the expected number of times state-i is visited. Hence following the conventional method, the assignment of state-transition probabilities reduces to counting the number of transitions between states.

Comments: For illustration purposes, we generate a sequence of correlated unit mean gamma random variables with shape parameter $\nu = 2$ and scale parameter $1/2$ according to the linear transformation method described in [28]. The covariance function between the k'th and l'th element of the sequence is set as $\rho^{|k-l|}$ with $\rho = 0.75$. Figure 3 shows a sample run containing a total of 150 samples. The quantization values κ_q for 5-level quantization and the quantization borders are also shown in the figure. An examination of this figure reveals that pairs of consecutive quantized samples exhibit a tendency to stay in the same or neighboring state. This is perhaps most clearly seen for the samples quantized to κ_5 value. The set of next-states visited when the current-state is κ_5 is exclusively

TABLE I

SUGGESTED 3/5/7 LEVEL QUANTIZATION VALUES, IN DECIBELS, FOR UNIT MEAN GAMMA DISTRIBUTED TEXTURE PARAMETER WITH SHAPE PARAMETER ν

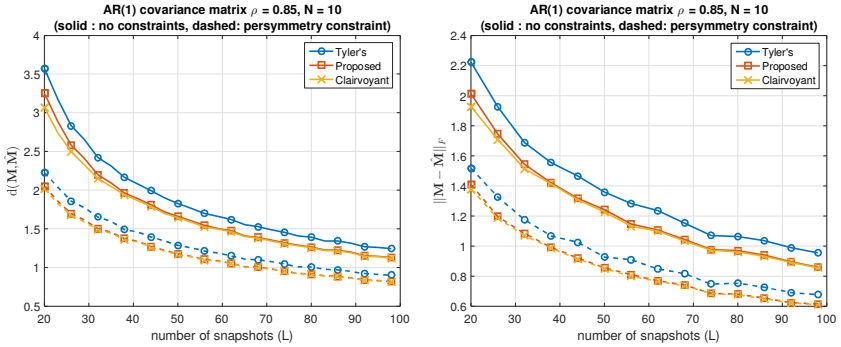
	3 Level			**5 Level**					**7 Level**						
	κ_{-1}	κ_0	κ_1	κ_{-2}	κ_{-1}	κ_0	κ_1	κ_2	κ_{-3}	κ_{-2}	κ_{-1}	κ_0	κ_1	κ_2	κ_3
$\nu = 1$	-17.05	-1.76	3.24	-19.8	-5.09	-0.8	2.26	5.15	-21.68	-7.24	-3.14	-0.41	1.81	3.87	6.11
$\nu = 1.25$	-11.77	-1.53	2.97	-14.37	-4.55	-0.76	2.03	4.71	-16.16	-6.5	-2.9	-0.43	1.6	3.5	5.61
$\nu = 1.50$	**-9.17**	**-1.27**	**2.81**	-11.51	-4.04	-0.66	1.9	4.4	-13.12	-5.81	-2.63	-0.39	1.48	3.25	5.24
$\nu = 1.75$	-7.7	-1.08	2.68	-9.83	-3.64	-0.58	1.8	4.15	-11.29	-5.27	-2.41	-0.36	1.38	3.05	4.94
$\nu = 2$	-6.74	-0.94	2.57	-8.69	-3.32	-0.51	1.71	3.95	-10.03	-4.84	-2.23	-0.32	1.31	2.89	4.69
$\nu = 4$	**-3.82**	**-0.45**	**2.02**	-5.1	-2.12	-0.26	1.32	3.01	-5.96	-3.15	-1.48	-0.18	0.99	2.17	3.56
$\nu = 6$	-2.88	-0.3	1.73	-3.89	-1.66	-0.18	1.12	2.54	-4.56	-2.48	-1.17	-0.12	0.84	1.82	3.01
$\nu = 8$	-2.39	-0.22	1.54	-3.25	-1.4	-0.13	0.99	2.25	-3.81	-2.1	-0.99	-0.09	0.74	1.61	2.67
$\nu = 10$	**-2.08**	**-0.18**	**1.4**	-2.83	-1.23	-0.11	0.9	2.05	-3.33	-1.85	-0.88	-0.07	0.67	1.46	2.42
$\nu = 100$	-0.57	-0.02	0.5	-0.79	-0.35	-0.01	0.32	0.71	-0.93	-0.54	-0.25	-0.01	0.23	0.5	0.84

limited κ_4 and κ_5 in this Monte Carlo run. In fact, a large number of Monte Carlo runs indicate that this is the case 90% of the time. The main of goal of this study is to make use of this knowledge, that is the observed persistence of states, in the covariance matrix estimation. Furthermore, the suggested HMM based model easily enables to empirical observations such as "the state κ_5 occurs rarely; but when it occurs, it persists for a number samples" into the covariance matrix estimation setting.

V. NUMERICAL EXPERIMENTS

This section gives the performance comparison results for the suggested method. We use two distance metrics in comparisons. The first metric is the Frobenius norm $||A||_F = \sqrt{\text{tr}(\mathbf{A}^T\mathbf{A})}$. The second metric is proposed by Förstner for the comparison of positive definite matrices and can be written as $d(\mathbf{M}, \widehat{\mathbf{M}}) = ||\ln(\mathbf{M}^{-1/2}\widehat{\mathbf{M}}\mathbf{M}^{-1/2})||_F$ or equivalently $d(\mathbf{M}, \widehat{\mathbf{M}}) = \sqrt{\sum_i \ln^2(\lambda_i)}$ where λ_i are the generalized eigenvalues of $\mathbf{M}$ and $\widehat{\mathbf{M}}$ [29]. Different from the Frobenius norm, Förstner's metric $d(\mathbf{M}, \widehat{\mathbf{M}})$ is affine-invariant. Hence, a change in the measurement units, say from meters to kilometers, has no effect on Förstner's metric.

We present comparisons of the proposed method with Tyler's estimator [1] and clairvoyant estimator. The clairvoyant estimator is assumed to have access to the true value of texture parameters



(a) Förstner's metric (b) Frobenius norm

Fig. 4. Covariance matrix estimation performance comparison with two metrics for noise-free snapshots

and can be expressed for matrices without any structure and with persymmetry structure as follows:

$$\widehat{\mathbf{M}}^{\text{clair}} = \frac{1}{L}\sum_{l=1}^{L}\frac{1}{\tau_l}\mathbf{y}_l\mathbf{y}_l^H,$$

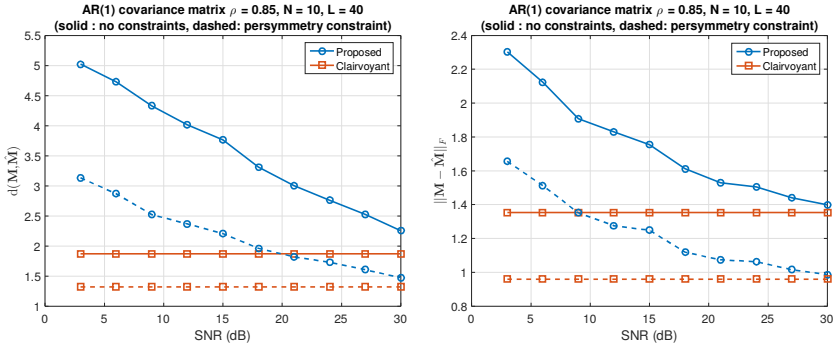
$$\widehat{\mathbf{M}}^{\text{clair}}_{\text{persym}} = \frac{1}{2L}\sum_{l=1}^{L}\frac{1}{r_l}\left(\mathbf{y}_l\mathbf{y}_l^H + \mathbf{y}_l^R(\mathbf{y}_l^R)^H\right). \quad (20)$$

The results for the clairvoyant estimator should be interpreted as a performance lower bound for other estimators. Tyler's estimator and its persymmetric version are described in the Appendix of this study.

A. Experiment 1: AR(1) Auto-correlation Matrix

In this case, the unknown covariance matrix $\mathbf{M}$ is taken as a Toeplitz matrix with the first row entries $[1\,\rho\,\rho^2\ldots\rho^{N-1}]$. This matrix corresponds the covariance matrix of a stationary AR(1) random process having the auto-correlation sequence $r[k] = \rho^{|k|}$. The texture parameter of this experiment is assumed to be quantized to 5-levels $\tau_k \in \{1/100, 1/10, 1, 10, 100\}$. The probability of preserving the same state is taken as 0.9, $\pi_{ii} = 0.9$. The transitions to other states are equally probable, that is $\pi_{ij} = 0.025$, $i \neq j$, in this experiment.

Figure 4 shows the results for $N = 10$ dimensional covariance matrix estimation. Since the unknown covariance matrix is Toeplitz, the matrix satisfies the persymmetry condition. Plots with dashed lines in all figures shows the results when persymmetry is taken into account in the estimation. Indeed, it can be observed from Figure 4 that the inclusion of persymmetry constraint brings significant gains. The suggested method can be noted to closely track the clairvoyant estimator at almost all L values.



(a) Förstner's metric (b) Frobenius norm

Fig. 5. Performance of the suggested method for noisy snapshots

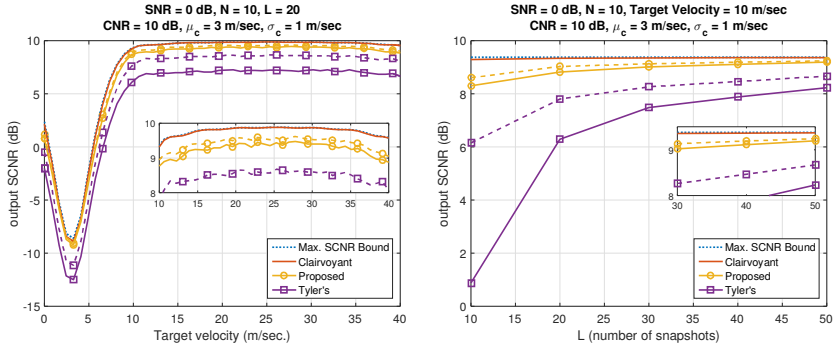
When compared with the Tyler's estimator, the suggested scheme achieves the performance of Tyler's estimator with 5 - 10 fewer number of snapshots. This result can be significant in applications where the number of snapshots is scarce.

We repeat the the same experiment in the presence of noise. The snapshots are assumed to be contaminated with white Gaussian noise with variance $1/\text{SNR}$. Note that in this experiment, the signal power is $r[k]|_{k=0} = 1$. Figure 5 shows that the performance of suggested method approaches to the clairvoyant estimator with increasing number of snapshots.

B. Experiment 2: Adaptive Radar Detection Application

We study the SNR loss due to imperfect covariance matrix estimation within the context of adaptive target detection. We consider an X-band (wavelength of $\lambda = 3$ cm) pulse Doppler radar system operating at a constant pulse repetition frequency (PRF) of 3 kHz, transmitting $N = 10$ pulses. The system operates under the influence of sea-clutter with mean velocity $\mu_c = 3$ m/sec and spread $\sigma_c = 1$ m/sec. The clutter-to-noise ratio (CNR) is 10 dB, noise variance is $\sigma_n^2 = 1$.

The clutter power spectral density is assumed to be in Gaussian form leading to the clutter auto-correlation of $r_c[k] = \text{CNR}\, e^{j2\pi f_d k}\rho^{k^2}$ with $f_d = 2\mu_c/(\lambda\,\text{PRF}) = 1/15$ and $\rho = \exp(-8\pi^2\sigma_c^2/(\lambda\,\text{PRF})^2) \approx 0.990$. The k'th row, l'th column entry of the clutter covariance matrix $\mathbf{M}_0$ is then $r_c[k-l]$. As in earlier experiment, the texture parameter is quantized to 5 levels, $\kappa_s \in \{-20, -10, 0, 10, 20\}$ dB; the state transition probabilities are set as $\pi_{ii} = 0.9$ and $\pi_{ij} = 0.025\ i \neq j$. These numerical values are also roughly inline with the measurements in [17, Fig.2 and Fig.3].



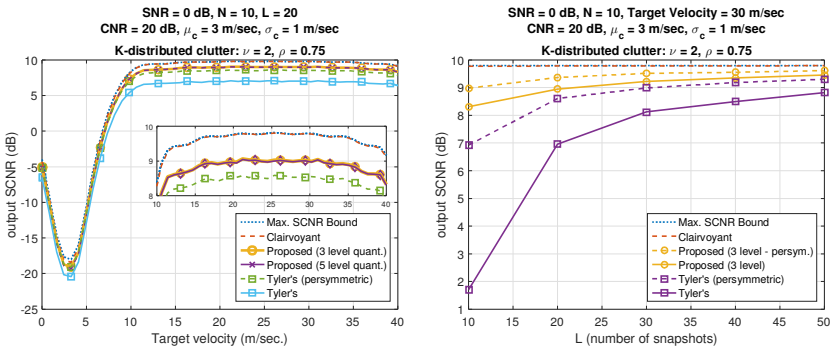
(a) SCNR vs. Target Velocity (b) SCNR vs. number of snapshots

Fig. 6. Average output SCNR comparison. Dashed and solid lines correspond to estimators with and without persymmetry constraint, respectively.

In this experiment, we present a comparison for the average output signal-to-clutter-and-noise ratio, $\text{SCNR}_{\text{out}}(\omega) = |\mathbf{w}^H\mathbf{s}_\omega|^2/\mathbf{w}^H(\mathbf{M} + \sigma_n^2\mathbf{I})\mathbf{w}$, for two-stage adaptive detector (AMF) with the linear combination weight $\mathbf{w} = (\widehat{\mathbf{M}} + \sigma_n^2\mathbf{I})^{-1}\mathbf{s}_\omega$ [11]. Here $\mathbf{s}_\omega = \sqrt{\text{SNR}}[1\ e^{j\omega}\ \ldots\ e^{j(N-1)\omega}]^T$ is the Doppler steering vector of the target with velocity v_t and $\omega = 4\pi v_t/(\lambda\,\text{PRF})$ is the phase progression due to the Doppler effect. In this comparison, we assume that $\widehat{\mathbf{M}}$ is generated from an observation of L snapshots via different methods and compare the achievable output SCNR for each method. We also present the maximum SCNR bound as a comparison metric. This bound is achieved when true $\mathbf{M}$ (instead of its estimate) is utilized in the weight vector calculation, $\text{SCNR}^{\max} = \mathbf{s}_\omega^H(\mathbf{M} + \sigma_n^2\mathbf{I})^{-1}\mathbf{s}_\omega$.

Figure 6a shows average SCNR attained with $L = 20$ snapshots at various target velocities. The proposed detector operates 0.3 dB (with persymmetry constraint) and 0.5 dB (without persymmetry constraint) below the performance bounds (clairvoyant detector and SCNR bound) at target velocity of 10 m/sec. The detector with Tyler's estimate has a loss around 1.5 dB (with persymmetry constraint) and 3.1 dB (without persymmetry constraint) and the SNR loss is not recoverable even at high target velocities. Note that, in this experiment Tyler's estimator does not estimate the clutter covariance matrix $\mathbf{M}$; but the interference matrix (clutter plus noise covariance matrix), that is $\mathbf{M}+\sigma_n^2\mathbf{I}$. Figure 6b presents the results when target velocity is fixed to 10 m/sec. and secondary data size (L) is varied. The SNR loss further increases for small number of snapshots.

K-distributed Clutter Experiment: This experiment examines the case of model mismatch which is the case in general for seaborne/airborne adaptive target detectors. The clutter is taken as K-distributed. That is, the texture parameter is a gamma distributed. The shape parameter of the



(a) SCNR vs. Target Velocity

(b) SCNR vs. number of snapshots

Fig. 7. Average output SCNR comparison. Dashed and solid lines correspond to estimators with and without persymmetry constraint, respectively.

distribution is set as $\nu = 2$. The texture correlation is $r_\tau[k] = \rho^{|k|}$ with $\rho = 0.75$. Figure 3 shows a realization of gamma variable sequence generated with these parameters.

We follow the procedure given in Section IV to construct the Markov model and use both 3 and 5 level quantization values given in Table I for $\nu = 2$ in this comparison. Figure 7 shows the results for $N = 10$ and $L = 20$. The other experiment parameters are given in the figure title.

We observe from Figure 7a that the performance of the proposed method for the quantization levels of 3 and 5 are almost identical. The performance plots for the proposed method in Figure 7a do not assume the persymmetry condition. The effect of persymmetry is examined in Figure 7b. From Figure 7a, we see that the SNCR loss for the velocity of 30 m/sec (high velocity region) when compared with SCNR bound or clairvoyant estimator is 0.8 dB for the proposed estimators (without persymmetry); while it is 1.2 dB and 2.7 dB for Tyler's estimator with and without persymmetry constraint, respectively.

Figure 7b gives the results for a fixed velocity of 30 m/sec for different number of snapshots. In this figure, only 3-level quantization results are shown for figure clarity; since 5-level quantization results are almost identical. In this figure, the results for the proposed method with and without persymmetry constraint are given. We see that for $L = 20$ case, the gap to the SCNR bound further reduces to 0.4 dB (with the persymmetry constraint) from 0.8 dB (without persymmetry). Hence the proposed method (with persymmetry constraint) brings an SNR improvement of 0.8 dB over the Tyler's method (with persymmetry constraint). Or stated different, Tyler's estimator requires $L = 50$ snapshots to operate at the output SCNR level of proposed method with $L = 20$ snapshots. Yet, we would like to

remind that in many adaptive radar detection applications, the number of snapshots can be limited by other factors such as clutter discretes, other targets etc. Hence, utilizing fewest possible snapshots can be mandated by other considerations. It can be seen from Figure 7b that for $L = 10$, the gap between proposed and Tyler's method is almost 2 dB.

VI. Conclusions

In this study, we examine the effect of texture correlation in the covariance matrix estimation problem. To the best of our knowledge, such a study is not given in the literature, in spite of observed texture correlation in some applications. Presented results show that the inclusion of this additional information can yield performance improvements especially for small number of snapshots. The presented covariance matrix estimation results can be especially in the adaptive radar detection applications where the number of snapshots is very limited in general.

Appendix

A. Tyler's Covariance Matrix Estimator for Persymmetric Matrices

A matrix is called persymmetric if it has symmetry across its anti-diagonal. Stated differently, matrices that satisfy $\mathbf{A} = \mathbf{J}\mathbf{A}^T\mathbf{J}$ with the exchange matrix $\mathbf{J}$ are called persymmetric. Exchange matrix is a permutation matrix which reverses the order of entries in a vector. ($\mathbf{J} = [\mathbf{e}_N\, \mathbf{e}_{N-1}\, \dots\, \mathbf{e}_1]$ and $\mathbf{e}_k$ is the kth column of $N \times N$ dimensional identity matrix.) An important example for persymmetric matrices are Toeplitz matrices. (Toeplitz matrices have constant values on the anti-diagonals, $\mathbf{T}_{k,k+i} = c_i$.)

Covariance matrices are by definition Hermitian-symmetric. In addition, some covariance matrices also satisfy the persymmetry condition. Among them, a particularly important covariance matrix class is the class of matrices associated with stationary random processes. These matrices are Toeplitz. Hence, they satisfy both Hermitian-symmetry $\mathbf{R} = \mathbf{R}^H$ and persymmetry $\mathbf{R} = \mathbf{J}\mathbf{R}^T\mathbf{J}$ conditions. Combining these two relations, we have $\mathbf{R}^H = \mathbf{J}\mathbf{R}^T\mathbf{J}$ which is the definition for centro-hermitian matrices. Centro-hermitian matrices are conjugate-symmetric across its center entry. As an example, a 3×3 centro-hermitian matrix $\mathbf{C}$ has elements which are hermitian symmetric about $\mathbf{C}_{2,2}$ entry, the central entry. Note that the centro-symmetry condition $\mathbf{R}^H = \mathbf{J}\mathbf{R}^T\mathbf{J}$ can also be written as $\mathbf{R} = \mathbf{J}\mathbf{R}^*\mathbf{J}$ where $\mathbf{R}^*$ is the complex-conjugate of the matrix $\mathbf{R}$.

We would like to give a derivation for Tyler's covariance matrix estimator when estimator is restricted by the persymmetry constraint. It is known that Tyler's estimator can be derived as the maximum-likelihood (ML) estimator for compound-Gaussian vectors with deterministic texture parameters. To extend the Tyler's estimator to the persymmetric matrices, we follow the same approach but impose the persymmetry constraint implicitly at an intermediate step of derivation.

The snapshots $\mathbf{x}_k = \sqrt{\tau}_k \mathbf{c}_k$ with deterministic texture parameters τ_k and iid Gaussian distributed speckle component $\mathbf{c}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{M})$ have the following joint distribution:

$$f(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L) = \prod_{k=1}^{L} \frac{1}{\pi^N |\tau_k \mathbf{M}|} \exp\left(-\frac{1}{\tau} \mathbf{x}_k^H \mathbf{M}^{-1} \mathbf{x}_k\right). \tag{21}$$

The log-likelihood function $\Lambda(\mathbf{M}, \tau_1, \ldots, \tau_L)$ can be written as

$$\Lambda(\mathbf{M}, \tau_1, \ldots, \tau_L) \stackrel{c}{=} -L \log(|M|) - \sum_{k=1}^{L} (N \log(\tau_k) + \frac{1}{\tau_k} \mathbf{x}_k^H \mathbf{M}^{-1} \mathbf{x}_k). \tag{22}$$

Here $\stackrel{c}{=}$ shows to presence of additional terms on the right hand side of the equation that do not affect the subsequent optimization steps. Differentiation with respect to τ_k yields the stationarity condition of $\tau_k^{(\text{opt})} = \mathbf{x}_k^H \mathbf{M}^{-1} \mathbf{x}_k / N$. Substitution of $\tau_k^{(\text{opt})}$ in (22) results in the compressed log-likelihood expression of

$$\Lambda(\mathbf{M}, \tau_1^{(\text{opt})}, \ldots, \tau_L^{(\text{opt})}) \stackrel{c}{=} -L \log(|M|) - N \sum_{k=1}^{L} \log\left(\mathbf{x}_k^H \mathbf{M}^{-1} \mathbf{x}_k\right) \tag{23}$$

Differentiating the compressed log-likelihood with respect to $\mathbf{M}$, we get the optimality condition as

$$-L\mathbf{M}^{-1} - N\mathbf{M}^{-1} \left(\sum_{k=1}^{L} \frac{\mathbf{x}_k \mathbf{x}_k^H}{\mathbf{x}_k^H \mathbf{M}^{-1} \mathbf{x}_k} \right) \mathbf{M}^{-1} = \mathbf{0}. \tag{24}$$

Multiplying the last expression from left and right $\mathbf{M}$, we get the fixed-point relation defining the Tyler's estimator:

$$\mathbf{M} = \frac{N}{L} \sum_{k=1}^{L} \frac{\mathbf{x}_k \mathbf{x}_k^H}{\mathbf{x}_k^H \mathbf{M}^{-1} \mathbf{x}_k}. \tag{25}$$

To impose the persymmetry constraint of $\mathbf{M} = \mathbf{J}\mathbf{M}^*\mathbf{J}$, we make the observation that it is possible to rewrite the quadratic product $\mathbf{x}_k^H \mathbf{M}^{-1} \mathbf{x}_k$ as

$$\mathbf{x}_k^H \mathbf{M}^{-1} \mathbf{x}_k = \mathbf{x}_k^H (\mathbf{J}\mathbf{M}^*\mathbf{J})^{-1} \mathbf{x}_k = (\mathbf{J}\mathbf{x}_k)^H (\mathbf{M}^*)^{-1} (\mathbf{J}\mathbf{x}_k) = \left((\mathbf{J}\mathbf{x}_k^*)^H \mathbf{M}^{-1} (\mathbf{J}\mathbf{x}_k^*)\right)^* = \left((\mathbf{x}_k^R)^H \mathbf{M}^{-1} \mathbf{x}_k^R\right)^*$$

where we introduce $\mathbf{x}_k^R = \mathbf{J}\mathbf{x}_k^*$ to denote the reversed (flipped up-down) and conjugated vector $\mathbf{x}_k$. Since the complex conjugate of a scalar is identical to its Hermitian, we can further simply the quadratic product $\mathbf{x}_k^H \mathbf{M}^{-1} \mathbf{x}_k$ to

$$\left((\mathbf{x}_k^R)^H \mathbf{M}^{-1} \mathbf{x}_k^R\right)^* = \left((\mathbf{x}_k^R)^H \mathbf{M}^{-1} \mathbf{x}_k^R\right)^H = (\mathbf{x}_k^R)^H \mathbf{M}^{-1} \mathbf{x}_k^R. \tag{26}$$

where $\mathbf{M} = \mathbf{M}^H$ is used one more time in the very last equality. Hence, for persymmetric $\mathbf{M}$ matrices, we have the identity $\mathbf{x}_k^H \mathbf{M}^{-1} \mathbf{x}_k. = (\mathbf{x}_k^R)^H \mathbf{M}^{-1} \mathbf{x}_k^R$. Using this identity, the log-likelihood expression in (22) for unstructured matrices can be written for persymmetric matrices as follows:

$$\Lambda(\mathbf{M}, \tau_1, \ldots, \tau_L) \stackrel{c}{=} -L \log(|M|) - \sum_{k=1}^{L} (N \log(\tau_k) + \frac{1}{2\tau_k} \left(\mathbf{x}_k^H \mathbf{M}^{-1} \mathbf{x}_k + (\mathbf{x}_k^R)^H \mathbf{M}^{-1} \mathbf{x}_k^R\right)). \tag{27}$$

The optimization for the texture variables yields $\tau_k^{(\text{opt})} = \frac{\mathbf{x}_k^H \mathbf{M}^{-1} \mathbf{x}_k + (\mathbf{x}_k^R)^H \mathbf{M}^{-1} \mathbf{x}_k^R}{2N}$. By inserting the optimized texture variable, into log-likelihood expression (27), we get

$$\Lambda(\mathbf{M}, \tau_1^{(\text{opt})}, \ldots, \tau_L^{(\text{opt})}) \stackrel{c}{=} -L \log(|M|) - N \sum_{k=1}^{L} \log\left(\mathbf{x}_k^H \mathbf{M}^{-1} \mathbf{x}_k + (\mathbf{x}_k^R)^H \mathbf{M}^{-1} \mathbf{x}_k^R\right). \tag{28}$$

Differentiating with respect to $\mathbf{M}$ and equating to the result to zero, we get the fixed-point relation defining the Tyler's estimator under the persymmetry constraint as:

$$\mathbf{M}_{\text{persym}} = \frac{N}{L} \sum_{k=1}^{L} \frac{\mathbf{x}_k \mathbf{x}_k^H + \mathbf{x}_k^R (\mathbf{x}_k^R)^H}{\mathbf{x}_k^H \mathbf{M}_{\text{persym}}^{-1} \mathbf{x}_k + (\mathbf{x}_k^R)^H \mathbf{M}_{\text{persym}}^{-1} \mathbf{x}_k^R} \tag{29}$$

where $\mathbf{x}_k^R = \mathbf{J}\mathbf{x}_k^*$. It is easy to verify that right hand side of (29) and therefore the fixed point of this relation is a persymmetric matrix. Hence, $\mathbf{M}_{\text{persym}}$ by (29) becomes the Tyler's estimate for the matrices constrained to persymmetric matrices. (This proof is an extension of the proof given for persymmetric covariance matrices for Gaussian vectors in [23] to the compound Gaussian vectors with deterministic texture parameters.)

REFERENCES

[1] D. E. Tyler, "A distribution-free m-estimator of multivariate scatter," *Ann. Statist.,*, vol. 15, no. 1, pp. 234–251, 1987. [Online]. Available: http://www.jstor.org/stable/2241079 (Cited on pages 2 and 13.)

[2] F. Gini and M. Greco, "Covariance matrix estimation for CFAR detection in correlated heavy tailed clutter," *Signal Processing*, vol. 82, no. 12, pp. 1847 – 1859, 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165168402003158 (Cited on page 2.)

[3] P. J. Huber, "Robust estimation of a location parameter," *Ann. Math. Statist.*, vol. 35, no. 1, pp. 73–101, 1964. [Online]. Available: http://www.jstor.org/stable/2238020 (Cited on page 2.)

[4] R. A. Maronna, "Robust m-estimators of multivariate location and scatter," *Ann. Statist.*, vol. 4, no. 1, pp. 51–67, 1976. [Online]. Available: http://www.jstor.org/stable/2957994 (Cited on page 2.)

[5] Y. I. Abramovich, B. A. Johnson, and N. K. Spencer, "Sample-deficient adaptive detection: Adaptive scalar thresholding versus CFAR detector performance," *IEEE Trans. Aerospace and Electronic Systems*, vol. 46, no. 1, pp. 32–46, Jan 2010. (Cited on page 2.)

[6] Y. Sun, A. Breloy, P. Babu, D. P. Palomar, F. Pascal, and G. Ginolhac, "Low-complexity algorithms for low rank clutter parameters estimation in radar systems," *IEEE Trans. Signal Processing*, vol. 64, no. 8, pp. 1986–1998, April 2016. (Cited on page 2.)

[7] A. Breloy, G. Ginolhac, F. Pascal, and P. Forster, "Robust covariance matrix estimation in heterogeneous low rank context," *IEEE Trans. Signal Processing*, vol. 64, no. 22, pp. 5794–5806, Nov 2016. (Cited on page 2.)

[8] I. Soloveychik and A. Wiesel, "Tyler's covariance matrix estimator in elliptical models with convex structure," *IEEE Trans. Signal Processing*, vol. 62, no. 20, pp. 5251–5259, Oct 2014. (Cited on page 2.)

[9] G. Pailloux, P. Forster, J. Ovarlez, and F. Pascal, "Persymmetric adaptive radar detectors," *IEEE Trans. Aerospace and Electronic Systems*, vol. 47, no. 4, pp. 2376–2390, OCTOBER 2011. (Cited on page 2.)

[10] F. Bandiera, O. Besson, and G. Ricci, "Knowledge-aided covariance matrix estimation and adaptive detection in compound-Gaussian noise," *IEEE Trans. Signal Processing*, vol. 58, no. 10, pp. 5391–5396, Oct 2010. (Cited on page 2.)

[11] F. Bandiera, D. Orlando, and G. Ricci, *Advanced Radar Detection Schemes Under Mismatched Signal Models*. Morgan & Claypool Publishers, 2009. (Cited on pages 2 and 16.)

[12] A. D. Maio and M. S. Greco, *Modern Radar Detection Theory*. Scitech Publishing, 2015. (Cited on page 2.)

[13] E. J. Kelly, "An adaptive detection algorithm," *IEEE Trans. Aerospace and Electronic Systems*, vol. AES-22, no. 2, pp. 115–127, Mar. 1986. (Cited on page 2.)

[14] E. Conte, M. Lops, and G. Ricci, "Asymptotically optimum radar detection in compound-Gaussian clutter," *IEEE Trans. Aerospace and Electronic Systems*, vol. 31, no. 2, pp. 617–625, April 1995. (Cited on page 2.)

[15] J. Ovarlez, F. Pascal, and P. Forster, "Covariance matrix estimation in SIRV and elliptical processes and their application in radar detection," in *Modern Radar Detection Theory*, A. D. Maio and M. S. Greco, Eds. Scitech Publishing, 2015, ch. 8, pp. 295–327. (Cited on page 2.)

[16] K. Ward, R. Tough, and S. Watts, *Sea Clutter: Scattering, the K distribution and radar performance*. The Institution of Engineering and Technology, 2013. (Cited on pages 2, 3, and 9.)

[17] S. Watts and K. D. Ward, "Spatial correlation in K-distributed sea clutter," *IEE Proc. F, Commun., Radar & Signal Process.,*, vol. 134, no. 6, pp. 526–532, October 1987. (Cited on pages 2, 3, and 15.)

[18] F. Gini, M. V. Greco, M. Diani, and L. Verrazzani, "Performance analysis of two adaptive radar detectors against non-Gaussian real sea clutter data," *IEEE Trans. Aerospace and Electronic Systems*, vol. 36, no. 4, pp. 1429–1439, Oct. 2000. (Cited on page 2.)

[19] A. D. Maio, G. Foglia, E. Conte, and A. Farina, "CFAR behavior of adaptive detectors: an experimental analysis," *IEEE Trans. Aerospace and Electronic Systems*, vol. 41, no. 1, pp. 233–251, Jan. 2005. (Cited on page 2.)

[20] B. C. Levy, *Principles of Signal Detection and Parameter Estimation*, 1st ed. Spring Street, NY, USA: Springer Science+Business Media, LLC, 2008. (Cited on page 5.)

[21] D. Barber and A. T. Cemgil, "Graphical models for time-series," *IEEE Signal Process. Mag.*, vol. 27, no. 6, pp. 18–28, 2010. (Cited on pages 5 and 6.)

[22] R. Nitzberg, "Application of maximum likelihood estimation of persymmetric covariance matrices to adaptive processing," *IEEE Trans. Aerospace and Electronic Systems*, vol. AES-16, no. 1, pp. 124–127, 1980. (Cited on page 7.)

[23] M. Jansson and P. Stoica, "Forward-only and forward-backward sample covariances – a comparative study," *Elsevier Signal Processing*, vol. 77, no. 3, pp. 235–245, 1999. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165168499000377 (Cited on pages 7 and 20.)

[24] A. P. Shikhaliev, L. C. Potter, and Y. Chi, "Low-rank structured covariance matrix estimation," *IEEE Signal Processing Lett.*, vol. 26, no. 5, pp. 700–704, May 2019. (Cited on page 7.)

[25] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989. (Cited on pages 10, 11, and 12.)

[26] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Information Theory*, vol. 28, no. 2, pp. 129–137, 1982. (Cited on page 11.)

[27] J. Max, "Quantizing for minimum distortion," *IRE Trans. on Information Theory*, vol. 6, no. 1, pp. 7–12, 1960. (Cited on page 11.)

[28] Y. Dong, L. Rosenberg, and G. Weinberg, "Generating correlated gamma sequences for sea-clutter simulation," Defence Science and Technology Organisation, Australia, Tech. Rep. DSTO–TR–2688, 2012. (Cited on page 12.)

[29] W. Förstner and B. Moonen, "A metric for covariance matrices," in *Festschrift for Erik W. Grafarend on the occasion of his 60th birthday. Also appeared in: Geodesy - The Challenge of the 3rd Millennium (2003, with editors Professor Dr. Erik W. Grafarend, Dr. Friedrich W. Krumm,Dr. Volker S. Schwarze, ISBN: 978-3-642-07733-3 (Print) 978-3-662-05296-9 (Online))*, F. Krumm and V. S. Schwarze, Eds., 1999, pp. 113–128. [Online]. Available: https://www.ipb.uni-bonn.de/pdfs/Forstner1999Metric.pdf (Cited on page 13.)

EK-5

Elsevier Editorial System(tm) for Signal Processing

Manuscript Draft

Manuscript Number: SIGPRO-D-20-00519

Title: Invariant Function Approach for Gridless and Non-iterative Maximum Likelihood Parameter Estimation and Its Application to Frequency Estimation of Real-Valued Sinusoids

Article Type: Research Paper

Keywords: Parameter Estimation; Maximum Likelihood Estimation; Frequency Estimation

Corresponding Author: Professor Cagatay Candan,

Corresponding Author's Institution: Middle East Technical Univ.

First Author: Cagatay Candan

Order of Authors: Cagatay Candan; Utku Celebi

Abstract: An invariant function approach for gridless, non-iterative maximum likelihood (ML) estimation of unknown parameters in the presence of additional nuisance parameters is described. The main attraction point of the approach is its potential to yield a ML-like performance at a significantly reduced computation load with respect to conventional ML estimator which requires repeated evaluation of an objective function or an application of numerical maximization routines. The suggested approach is illustrated on the problem of frequency estimation of real-valued sinusoids. The resulting estimator for this application closely tracks the Cramer-Rao bound and performs very close to the ML estimator. The approach with its extension to multiple parameter estimation is described in a constructive manner to enable the design of estimators for other parameter estimation problems.

Suggested Reviewers: Hing Cheung So
http://www.ee.cityu.edu.hk/~hcso/
hcso@ee.cityu.edu.hk
He is one of the leading researchers with several important publications in this specific area (frequency estimation).

Cover Letter

March 22, 2020

Dear Editors,

We are submitting a manuscript titled "*Invariant Function Approach for Gridless and Non-iterative Maximum Likelihood Parameter Estimation and Its Application to Frequency Estimation of Real-Valued Sinusoids*" for publication consideration at Elsevier Signal Processing.

The submitted manuscript describes a framework for parameter estimation which has the potential of yielding very good estimators, say closely tracking Cramer-Rao bound at high SNR, at a low computational cost. The details of the suggested approach is presented on the problem of frequency estimation for real-valued sinusoids.

Best regards,

Prof. Cagatay Candan
Department of Electrical-Electronics Engineering,
Middle East Technical University (METU),
Ankara, Turkey.

Invariant Function Approach for Gridless and Non-iterative Maximum Likelihood Parameter Estimation and Its Application to Frequency Estimation of Real-Valued Sinusoids

Çağatay Candan[a], Utku Çelebi[a]

[a]*Department of Electrical and Electronics Engineering, Middle East Technical University (METU), 06800, Ankara, Turkey.*

Abstract

An invariant function approach for gridless, non-iterative maximum likelihood (ML) estimation of unknown parameters in the presence of additional nuisance parameters is described. The main attraction point of the approach is its potential to yield a ML-like performance at a significantly reduced computation load with respect to conventional ML estimator which requires repeated evaluation of an objective function or an application of numerical maximization routines. The suggested approach is illustrated on the problem of frequency estimation of real-valued sinusoids. The resulting estimator for this application closely tracks the Cramer-Rao bound and performs very close to the ML estimator. The approach with its extension to multiple parameter estimation is described in a constructive manner to enable the design of estimators for other parameter estimation problems.

Keywords: Maximum Likelihood Estimation, Gridless Search, Frequency Estimation, Real-valued Sinusoids, Parameter Estimation.

1. Introduction

The maximum likelihood frequency estimation of a sinusoid is a classical problem of statistical signal processing with important applications in array signal processing, spectrum estimation, communications, time-series analysis and others [1, 2]. In this study, we present an alternative approach for the frequency estimation of real-valued sinusoids through invariant functions. The invariant function approach is quite general and easily applicable to other parameter estimation problems. The main advantage of the approach is its estimation accuracy in spite of its low computational complexity. Specifically for the frequency estimation problem, higher complexity methods utilize maximum likelihood search, eigen or subspace decompositions, [3–5]; while the suggested approach is based on transforming the input to Discrete Fourier Transform (DFT) domain and constructing a function of the Fourier spectrum samples which is *invariant* to the nuisance parameters of the problem. It should be underlined that the frequency estimation problem is chosen as a venue, which is of importance on its own, to illustrate the suggested approach.

The phrase frequency estimation can refer to the parameter estimation problem for both complex-valued ($Ae^{j(\omega n+\phi)}$) and real-valued ($A\cos(\omega n+\phi)$) sinusoids. The complex-valued sinusoids are in general the low-pass equivalent of a band-pass signal and utilized in spectrum modeling, direction of arrival estimation, communications problems. There are several methods for the parameter estimation of complex-valued sinusoids (also called complex exponentials), [2]. Among them, [6–15] overlap with the invariant function approach described herein where the goal is to develop a very low complexity estimator for this fundamental problem. For an illustrative example, the unknown frequency $\omega = 2\pi(k_p+\delta)/N$ is estimated in two-stages in [7, 8]. In the first stage N-point DFT of the input is calculated and the DFT index with the peak magnitude is declared as $\widehat{k}_p$, (also see Figure 1). In the second stage (fine frequency estimation stage), the remaining unknown δ (fine frequency part) is estimated through the relation

$$\widehat{\delta} = f^{-1}\left(\mathrm{Re}\left\{\frac{R[\widehat{k}_p-1]-R[\widehat{k}_p+1]}{2R[\widehat{k}_p]-R[\widehat{k}_p-1]-R[\widehat{k}_p+1]}\right\}\right) \quad (1)$$

where $f^{-1}(\cdot)$ refers to the inverse function of $f(\delta)=\tan(\pi\delta/N)/\tan(\pi/N)$ and $R[\cdot]$ is the N-point DFT samples, as illustrated in Figure 1. The complexity of the estimator given by (1) comprises of i. calculation of N-point DFT, ii. calculation of ratio appearing in the argument of $f^{-1}(\cdot)$, iii. evaluation of the inverse function. In spite of its extreme low complexity, the performance of the estimator is surprisingly good, [8].

Other low-cost frequency estimation methods have been proposed for complex-valued sinusoids, [9–16]. These methods mainly differ in the non-linear expression used in the fine-frequency estimation stage. In this study, we refer the functions $f(\cdot)$ involved in the fine-

Email addresses: ccandan@metu.edu.tr (Çağatay Candan), ucelebi@metu.edu.tr (Utku Çelebi)

frequency estimation, such as the one in (1), as the invariant function. As discussed later, these functions are designed to be invariant to the nuisance parameters of the problem (amplitude and phase for the frequency estimation problem) and solely depend on the parameter of interest.

The frequency estimation for real-valued sinusoids is a bit more complicated than its complex-valued counterpart due to the interaction of two complex-valued sinusoids forming real-valued sinusoid. In the literature, conventional linear prediction based approaches have been extended to the case of real-valued sinusoids in [4, 17, 18]. Sub-space based methods have been developed in [5] and recently, some low-complexity estimators have been proposed [19, 20]. Different from earlier efforts, the low complexity estimators suggested in the literature aim to cancel one of complex exponentials forming the cosine waveform and following the cancellation operation, available invariant function based approaches for the frequency estimation of complex-valued sinusoids are utilized. For example, the method of Djukanovìc aims to eliminate one of the complex exponentials forming $\cos(\omega n+\phi) = (e^{j\omega n+\phi} + e^{-j\omega n-\phi})/2$ by filtering, [19]. To do the elimination, a rough frequency estimate is generated and the complex exponential with the negative valued frequency is filtered. The method of Ye et al. [20, 21] is based on a similar principle and estimates not only frequency, but also the amplitude and phase to implement an interference cancellation procedure. Different from existing low complexity estimators, we describe a general invariant function approach for the parameter estimation problem and specifically develop an estimator for the real-valued sinusoid frequency estimation problem. The suggested invariant function based method is applicable to multiple unknowns and can be applied to a variety of parameter estimation problems.

2. Preliminaries

A sampled sinusoidal signal with an unknown amplitude A, phase ϕ and frequency ω is observed under zero mean additive white Gaussian noise (AWGN) $w[n]$ with variance σ_w^2,

$$r[n] = A\cos(\omega n+\phi) + w[n], \quad n = \{0,\dots,N-1\}. \tag{2}$$

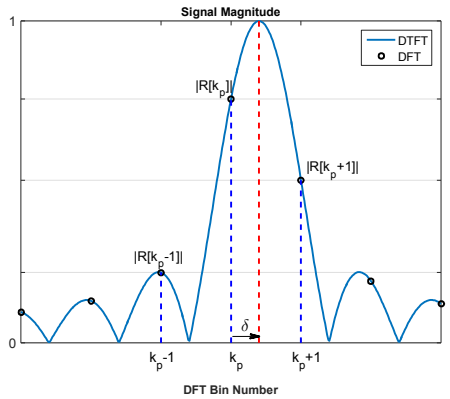


Figure 1: An illustration for the first stage of the proposed method

We consider the amplitude and the phase of the sinusoid as the nuisance parameters and treat the frequency ω as the sole parameter of interest. In fact, since $r[n]$ can be written as $r[n] = A_c\cos(\omega n) + A_s\sin(\omega n) + w[n]$ with $A_c = A\cos(\phi)$, $A_s = -A\sin(\phi)$; the maximum likelihood estimation of amplitude and phase reduces to a simpler problem with a linear observation model given the frequency estimate. The signal-to-noise-ratio (SNR) definition adopted in this study is $\mathrm{SNR} = A^2/(2\sigma_w^2)$.

Maximum-likelihood (ML) estimator: After some elementary manipulations outlined in Appendix A, the maximum-likelihood estimate of frequency is expressed as

$$\hat{\omega}_{\mathrm{ML}} = \arg\max_{\omega} \frac{|R(e^{j\omega})|^2 - \frac{\sin(\omega N)}{N\sin(\omega)}\mathrm{Re}\left\{R^2(e^{j\omega})e^{j\omega(N-1)}\right\}}{N - \frac{\sin^2(\omega N)}{N\sin^2(\omega)}}. \tag{3}$$

Here $R(e^{j\omega}) = \sum_{n=0}^{N-1} r[n]e^{-j\omega n}$ is the discrete-time Fourier transform (DTFT) of the input. The ML expression for the same problem also appears in [22, eq(7.65)] and [3]. Different from these expressions, (3) presents the relation in terms of DTFT of the input resulting in a simpler expression.

A straightforward implementation of the ML estimator in (3) is the application fast Fourier transform (FFT) for the calculation of DTFT samples and execution of a search for

the likelihood maxima. Interestingly, the peak location of the energy spectra, $\arg\max_\omega |R(e^{j\omega})|^2$, is not the maximum likelihood estimate for a finite N. Yet, as the number of observations N increases, the ML estimator converges to the peak localization in the magnitude DTFT spectra (periodogram). To reduce the computational complexity, a chirp-z transform based DTFT calculation over a portion of the spectrum can be also implemented. In this study, our goal is to present a much lower complexity estimator which is almost as good as the ML estimator given by (3).

Cramer-Rao Bound: For finite N, the Cramer-Rao bound (CRB) for the real-valued sinusoid depends on the nuisance parameter set (amplitude and phase) and on the parameter of interest (frequency). As $N \to \infty$, the term $\sum_{n=0}^{N-1} \sin(2\omega n + 2\phi)$ appearing in the Fisher information matrix (FIM) entries given in [22, p. 56] approaches 0 and CRB becomes independent of phase and frequency, [23]. The bound as $N \to \infty$ is called the asymptotic Cramer-Rao bound (ACRB):

$$E\{(\omega - \hat{\omega})^2\} \geq \text{ACRB} = \frac{12}{\text{SNR}(N^2-1)N}, \text{ as } N \to \infty. \tag{4}$$

We also note that the ACRB expression in (4) coincides with the hybrid Cramer-Rao bound (HCRB) expression for all finite N, provided that the signal phase is taken as a uniformly distributed random variable in $[0, 2\pi)$, [23]. HCRB is a Bayesian performance bound utilized in the presence of random nuisance parameters, [24]. In the numerical results section, we randomize the phase ϕ and utilize HCRB (i.e. the ACRB expression in (4)) as a performance benchmark. Since the value of phase can significantly affect the performance, its non-informative randomization results in a fair performance comparison for the frequency estimation problem. Lastly, we prefer to express the ACRB expression with the units of DFT bins. This unit convention is further explained in Section 4.

3. Proposed Method

The proposed method is composed of two stages. The first stage is the calculation of the $2N$-point DFT of the input $r[n] = A\cos\left(\frac{2\pi}{2N}(k_p + \delta)n + \phi\right) + w[n]$, $n = \{0, \ldots, N-1\}$. Note that we have switched the notation for the frequency variable from ω to $\frac{2\pi}{2N}(k_p + \delta)$. Here $k_p + \delta$ denotes the frequency in terms of $2N$-point DFT bins where k_p is an integer

between 0 and N and δ is a real number in $(-0.5, 0.5]$, (also see Figure 1). The first stage output is the DFT bin index with the maximum magnitude. This index is denoted by $\widehat{k}_p$, as shown in Algorithm 1 listing. In essence, the first stage implements a coarse search for the frequency via periodogram. Since the search is coarse, there is no need to utilize the exact ML expression (3) in this stage. (Interested readers can also examine earlier works on the frequency estimation of complex exponential signals utilizing the same model for more information, [6–15].)

The second stage, fine frequency estimation stage, utilizes three DFT outputs with the indices $\{\widehat{k}_p - 1, \widehat{k}_p, \widehat{k}_p + 1\}$ to estimate δ, as shown in Figure 1. To produce the fine frequency estimate $\widehat{\delta}$, a non-linear function of three variables is constructed and this function is evaluated at $R[\widehat{k}_p + l]$, $l = \{-1, 0, 1\}$. The operation is, in principle, similar to the one given by (1). The final estimate for the frequency becomes $\widehat{k}_p + \widehat{\delta}$ with the unit of DFT bins or $\widehat{\omega} = \frac{2\pi}{2N}(\widehat{k}_p + \widehat{\delta})$ radians per sample. Note that the overall complexity of the proposed method is $2N$ point DFT calculation and evaluation of a function of three complex-valued variables.

The performance of the proposed method critically depends on the non-linear function at the fine frequency estimation stage. By design, this function should be invariant to the nuisance parameters of problem in the absence of noise. To illustrate the invariant function, let's examine the complex exponential signal model with $r[n] = A\exp(j(\omega n + \phi)) + w[n]$, $n = \{0, \ldots, N-1\}$. In the absence of noise, DFT of $r[n]$ becomes $R[k] = A\exp(j\phi)\text{DFT}\{\exp(j\omega n)\}$. It can be noticed that the insertion of $R[k]$ in (1) results in the cancellation of $A\exp(j\phi)$ terms appearing on the numerator and denominator of ratio in the argument of the real part operator. Hence, this ratio is invariant to amplitude A and phase ϕ. Trivially, a function of this ratio is also an invariant function. It is possible to suggest different invariant functions for this problem [6–15]. The estimator performance depends on the properties of the invariant function, that is some invariant functions are more successful in the tracking of the Cramer-Rao bound such as the one proposed by Aboutanios and Mulgrew in [9].

Constructing Invariant Function for Real-valued Sinusoids: In the absence of

noise, $2N$-point DFT of the input,

$$R[k] = \sum_{n=0}^{N-1} r[n] e^{-j\frac{2\pi}{2N}nk} = A \sum_{n=0}^{N-1} \cos\left(\frac{2\pi}{2N}(k_p+\delta)n + \phi\right) e^{-j\frac{2\pi}{2N}nk} \tag{5}$$

can be expressed as

$$R[k] = \frac{A}{2} e^{-j\frac{\pi(N-1)}{2N}k} \left(\cos(\widetilde{\phi}) \left[\frac{\sin\left(\frac{\pi(k_p-k+\delta)}{2}\right)}{\sin\left(\frac{\pi(k_p-k+\delta)}{2N}\right)} + \frac{\sin\left(\frac{\pi(k_p+k+\delta)}{2}\right)}{\sin\left(\frac{\pi(k_p+k+\delta)}{2N}\right)} \right] + j\sin(\widetilde{\phi}) \left[\frac{\sin\left(\frac{\pi(k_p-k+\delta)}{2}\right)}{\sin\left(\frac{\pi(k_p-k+\delta)}{2N}\right)} - \frac{\sin\left(\frac{\pi(k_p+k+\delta)}{2}\right)}{\sin\left(\frac{\pi(k_p+k+\delta)}{2N}\right)} \right] \right) \tag{6}$$

where $\widetilde{\phi} = \phi + \frac{\pi}{2N}(k_p+\delta)(N-1)$, after some elementary manipulations.

In the absence of noise, the first stage output $\widehat{k}_p$ exactly matches the integer part of the unknown frequency, $\widehat{k}_p = k_p$. Hence, in the second stage, the DFT outputs with the indices $\{k_p-1, k_p, k_p+1\}$, i.e. $R[k_p+l]$ for $l = \{-1, 0, 1\}$, are used for fine frequency estimation. Here $l \triangleq k - k_p$, denotes the offset from the index k_p or more generally the offset from the index $\widehat{k}_p$, which is the first stage output.

To simplify the presentation, we define

$$\widetilde{R}[l] = R[k] e^{j\frac{\pi(N-1)}{2N}k} \,|_{k=l+k_p} = R[l+k_p] e^{j\frac{\pi(N-1)}{2N}(l+k_p)} \tag{7}$$

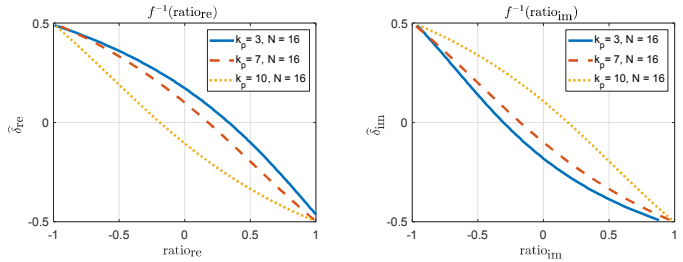
and denote the real and imaginary parts of $\widetilde{R}[l] = \widetilde{R}_{\rm re}[l] + j\widetilde{R}_{\rm im}[l]$ as $\widetilde{R}_{\rm re}[l]$ and $\widetilde{R}_{\rm im}[l]$, respectively. From (6), $\widetilde{R}_{\rm re}[l]$ and $\widetilde{R}_{\rm im}[l]$ can be expressed as

$$\begin{aligned} \widetilde{R}_{\rm re}[l] &\triangleq {\rm Re}\{\widetilde{R}[l]\} = M_{\rm re} \left[\frac{\sin\left(\frac{\pi(\delta-l)}{2}\right)}{\sin\left(\frac{\pi(\delta-l)}{2N}\right)} + \frac{\sin\left(\frac{\pi(2k_p+l+\delta)}{2}\right)}{\sin\left(\frac{\pi(2k_p+l+\delta)}{2N}\right)} \right], \\ \widetilde{R}_{\rm im}[l] &\triangleq {\rm Im}\{\widetilde{R}[l]\} = M_{\rm im} \left[\frac{\sin\left(\frac{\pi(\delta-l)}{2}\right)}{\sin\left(\frac{\pi(\delta-l)}{2N}\right)} - \frac{\sin\left(\frac{\pi(2k_p+l+\delta)}{2}\right)}{\sin\left(\frac{\pi(2k_p+l+\delta)}{2N}\right)} \right], \end{aligned} \tag{8}$$

where $M_{\rm re} = \frac{A}{2}\cos(\widetilde{\phi})$ and $M_{\rm im} = \frac{A}{2}\sin(\widetilde{\phi})$. Here $\widetilde{\phi} = \phi + \frac{\pi}{2N}(k_p+\delta)(N-1)$ is a constant depending on the unknown parameters, but independent of l.

Temporarily focusing on $\widetilde{R}_{\rm re}[l]$, it can be easily verified that the following expression involving $\widetilde{R}_{\rm re}[l]$, $l = \{-1, 0, 1\}$ is invariant to the nuisance parameters of the problem, namely the signal amplitude and phase,

$$\text{ratio}_{\rm re} \triangleq \frac{\widetilde{R}_{\rm re}[1] - \widetilde{R}_{\rm re}[-1]}{2\widetilde{R}_{\rm re}[0] - \widetilde{R}_{\rm re}[1] - \widetilde{R}_{\rm re}[-1]} \triangleq f_{\rm re}(\delta). \tag{9}$$



(a) $f_{\rm re}^{-1}(\text{ratio}_{\rm re})$ (b) $f_{\rm im}^{-1}(\text{ratio}_{\rm im})$

Figure 2: Inverse of the invariant functions $f_{\rm re}(\cdot)$ and $f_{\rm im}(\cdot)$ for a look-up table implementation.

Stated differently, the $\text{ratio}_{\rm re}$ given by (9) is solely a function of δ. We emphasize this fact with the notation of $\text{ratio}_{\rm re} = f_{\rm re}(\delta)$ in (9).

We also use the same invariant relation form for the imaginary part of DFT outputs given in (8) and define

$$\text{ratio}_{\rm im} \triangleq \frac{\widetilde{R}_{\rm im}[1] - \widetilde{R}_{\rm im}[-1]}{2\widetilde{R}_{\rm im}[0] - \widetilde{R}_{\rm im}[1] - \widetilde{R}_{\rm im}[-1]} \triangleq f_{\rm im}(\delta). \tag{10}$$

Note that $f_{\rm re}(\delta) \neq f_{\rm im}(\delta)$ due to the sign difference in the second component of the sum (8) in definition of $\widetilde{R}_{\rm re}[l]$ and $\widetilde{R}_{\rm im}[l]$.

Assuming that the inverse mapping for the invariant functions, $f_{\rm re}^{-1}(\cdot)$ or $f_{\rm im}^{-1}(\cdot)$, exists; an estimate for δ can be generated, in principle, via the application of the inverse function on the ratios, say $\widehat{\delta}_{\rm re} = f_{\rm re}^{-1}(\text{ratio}_{\rm re})$. Figure 2 shows the inverse mappings $f_{\rm re}^{-1}(\text{ratio}_{\rm re})$ and $f_{\rm im}^{-1}(\text{ratio}_{\rm im})$ for different parameter settings of N and k_p. Unfortunately, there is no simple analytical expression for the inverse function for the suggested invariant function. Hence, we need to utilize either a look-up table or numerical techniques for the inverse function mapping. Below, we describe a numerical procedure for the inverse function mapping when look-up table approach is not feasible or simply not preferred in order to achieve a better numerical accuracy. The suggested invariant function approach for the estimation of a single unknown δ is generalized to multiple unknowns in Appendix B.

Inversion of Invariant Function: In [8], a similar invariant function for the frequency estimation of *complex exponentials* is given as $f(\delta) = \tan(\pi\delta/N)/\tan(\pi/N)$ as

given in (1). For the complex exponential problem, it is possible to express the inverse function analytically in terms of elementary functions; but this is not the case for the one suggested in (9). Figure 2 shows the inverse of the invariant functions for different k_p and N values. Different from the complex exponential problem, the invariant function for the real-valued sinusoid depends on both k_p and N. Below, we describe a general, Taylor series based iterative method to establish the inverse mapping.

Focusing on the ratio_{re} given in (9), we consider $\widetilde{R}_{\text{re}}[l]$, $l = \{-1, 0, 1\}$ as a function of δ and expand the function around an arbitrary non-zero $\delta = \delta_0$ via the Taylor series and retain only the first two terms of the series. The approximation can be expressed as $\widetilde{R}_{\text{re}}[l] \approx K_l + \dot{K}_l(\delta - \delta_0)$. Here $K_l = \widetilde{R}_{\text{re}}[l]\,|_{\delta=\delta_0}$ and $\dot{K}_l = \frac{d}{d\delta}\widetilde{R}_{\text{re}}[l]\Big|_{\delta=\delta_0}$ can be explicitly given as

$$K_l = M_{\text{re}}\left[h\left(\frac{\pi(\delta_0 - l)}{2}\right) + h\left(\frac{\pi(2k_p + l + \delta_0)}{2}\right)\right] \tag{11}$$
$$\dot{K}_l = \frac{\pi}{2}M_{\text{re}}\left[h'\left(\frac{\pi(\delta_0 - l)}{2}\right) + h'\left(\frac{\pi(2k_p + l + \delta_0)}{2}\right)\right]$$

where $h(x) = \sin(x)/\sin(x/N)$ and $h'(x) = \frac{d}{dx}h(x) = \frac{N\cos(x)\sin(x/N) - \sin(x)\cos(x/N)}{N\sin^2(x/N)}$.

By substituting the Taylor series approximations for $\{\widetilde{R}_{\text{re}}[-1], \widetilde{R}_{\text{re}}[0], \widetilde{R}_{\text{re}}[1]\}$ in (9), we get

$$\text{ratio}_{\text{re}} \approx \frac{A_{\text{re}} + \dot{A}_{\text{re}}(\delta - \delta_0)}{B_{\text{re}} + \dot{B}_{\text{re}}(\delta - \delta_0)}, \tag{12}$$

where $A_{\text{re}} = K_1 - K_{-1}$, $\dot{A}_{\text{re}} = \dot{K}_1 - \dot{K}_{-1}$ and $B_{\text{re}} = 2K_0 - K_1 - K_{-1}$, $\dot{B}_{\text{re}} = 2\dot{K}_0 - \dot{K}_1 - \dot{K}_{-1}$.

To facilitate a simple approximation for $f_{\text{re}}^{-1}(\delta)$, valid around $\delta = \delta_0$; we multiply the numerator and denominator of the ratio on the right side of (12) by $B_{\text{re}} - \dot{B}_{\text{re}}(\delta - \delta_0)$ and ignore all second order terms to get,

$$\text{ratio}_{\text{re}} \approx \frac{A_{\text{re}}}{B_{\text{re}}} + \frac{\dot{A}_{\text{re}}B_{\text{re}} - A_{\text{re}}\dot{B}_{\text{re}}}{B_{\text{re}}^2}(\delta - \delta_0). \tag{13}$$

From (13), the unknown δ can be solved as

$$\widehat{\delta}_{\text{re}} = \frac{B_{\text{re}}^2}{\dot{A}_{\text{re}}B_{\text{re}} - A_{\text{re}}\dot{B}_{\text{re}}}\left(\text{ratio}_{\text{re}} - \frac{A_{\text{re}}}{B_{\text{re}}}\right) + \delta_0. \tag{14}$$

In practice, the inversion operation is applied iteratively; that is, we use the result of an earlier iteration as the expansion point (δ_0) of the next iteration. With the iterative

operation, the inverse mapping becomes a non-linear function of ratio_{re}. The accuracy of the suggested iterative inversion scheme is examined in the numerical results section. A detailed implementation is available in [25].

The same arguments can be repeated verbatim for the inversion of $f_{\text{im}}(\delta)$ function. Hence, a second estimate can be generated from the imaginary parts of $\widetilde{R}[l]$, $l = \{-1, 0, 1\}$ via the relation $\widehat{\delta}_{\text{im}} = \frac{B_{\text{im}}^2}{\dot{A}_{\text{im}}B_{\text{im}} - A_{\text{im}}\dot{B}_{\text{im}}}\left(\text{ratio}_{\text{im}} - \frac{A_{\text{im}}}{B_{\text{im}}}\right) + \delta_0$. where $A_{\text{im}}, \dot{A}_{\text{im}}$ and $B_{\text{im}}, \dot{B}_{\text{im}}$ are defined similarly.

Independence of $\widehat{\delta}_{\text{re}}$ and $\widehat{\delta}_{\text{im}}$: As also shown by Quinn in [26], the real and imaginary parts of $2N$-point DFT output calculated at step 1 of Algorithm 1 listing are correlated. Yet, as shown in Appendix C, once the step 3 of Algorithm 1 is executed, the resultant real and imaginary parts (step 4) becomes uncorrelated. Hence, the estimates, $\widehat{\delta}_{\text{re}}$ and $\widehat{\delta}_{\text{im}}$, which are derived solely from real and imaginary parts of $\widetilde{R}[l]$ (steps 5 to 15) are also independent random variables. This important result enables us to utilize a simple fusion rule for uncorrelated random variables as shown in the step 17 of Algorithm 1 Listing. We believe that the "demixing" operation at step 3, decorrelating real and imaginary parts, can be useful in other applications involving zero-padded DFTs.

Fusing Estimates: The final step of the suggested method is the fusion of the estimates produced from the real and imaginary parts of $\widetilde{R}[l]$. The estimates $\widehat{\delta}_{\text{re}}$ and $\widehat{\delta}_{\text{im}}$ are combined to reduce the estimation error. By construction, $\widetilde{R}_{\text{re}}[l]$ and $\widetilde{R}_{\text{im}}[l]$ are independent Gaussian random variables with variance $\sigma_w^2 N/2$, as shown in Appendix C. Other results of importance from Appendix C are as follows: The cross-correlation of $\widetilde{R}_{\text{re}}[l_1]$ and $\widetilde{R}_{\text{im}}[l_2]$ is zero for all (l_1, l_2) pairs. Yet $\widetilde{R}_{\text{re}}[l_1]$ is correlated with $\widetilde{R}_{\text{re}}[l_2]$ for odd valued $l_1 - l_2$. The same is also true for $\widetilde{R}_{\text{im}}[l_1]$. The most important fact for fusion purposes is that the estimates $\widehat{\delta}_{\text{re}}$ and $\widehat{\delta}_{\text{im}}$, derived from $\widetilde{R}_{\text{re}}[l]$ and $\widetilde{R}_{\text{im}}[l]$ respectively, are independent random variables. To combine them, we suggest to apply the best linear unbiased estimator (BLUE) rule which is applicable for uncorrelated random variables.

From (8), it can be noted that the input SNR for the estimates $\widehat{\delta}_{\text{re}}$ and $\widehat{\delta}_{\text{im}}$ are determined by the factors $M_{\text{re}} = \frac{A}{2}\cos(\widetilde{\phi})$ and $M_{\text{im}} = \frac{A}{2}\sin(\widetilde{\phi})$, respectively. Depending on the unknown parameter $\widetilde{\phi}$; SNR and therefore the accuracy of the estimates $\widehat{\delta}_{\text{re}}$ and $\widehat{\delta}_{\text{im}}$ can vary significantly. We suggest to use the following linear unbiased fusion rule to generate

the final estimate $\widehat{\delta}_{\rm F}$,

$$\widehat{\delta}_{\rm F} = \alpha\widehat{\delta}_{\rm re} + (1-\alpha)\widehat{\delta}_{\rm im}. \tag{15}$$

The fusion coefficient α should be ideally selected as $\alpha_{\rm ideal} = M_{\rm re}^2/(M_{\rm re}^2 + M_{\rm im}^2) = \cos^2(\widetilde{\phi})$. Yet, the signal phase is a nuisance parameter which is not estimated in the invariant function setting. Instead, we suggest to use the following approximation to the ideal fusion coefficient

$$\alpha \triangleq \frac{\widetilde{R}_{\rm re}^2[0]}{\widetilde{R}_{\rm re}^2[0] + \widetilde{R}_{\rm im}^2[0]} = \left(1 + \left(\frac{\widetilde{R}_{\rm im}[0]}{\widetilde{R}_{\rm re}[0]}\right)^2\right)^{-1}. \tag{16}$$

The approximation can be justified by noting from (8) that

$$\frac{\widetilde{R}_{\rm im}[0]}{\widetilde{R}_{\rm re}[0]} = \tan(\widetilde{\phi}) \underbrace{\left[\tan\left(\frac{\pi k_p}{2N}\right)\cot\left(\frac{\pi(k_p+\delta)}{2N}\right)\right]^p}_{\approx 1} \approx \tan(\widetilde{\phi}),$$

for $N \gg 1$ and $k_p \gg 1$. Here $p = (-1)^{\widehat{k}_p}$ is the parity of $\widehat{k}_p$ (first stage output), taking the value of 1 or -1 depending on $\widehat{k}_p$ is an odd or even integer, see Appendix D for details.

Finally, the suggested fusion coefficient, given by (16), can be expressed as $\alpha \approx \left(1+\tan^2(\widetilde{\phi})\right)^{-1} = \cos^2(\widetilde{\phi})$ for $N \gg 1$ and $k_p \gg 1$. This concludes the derivation of the estimator given in Algorithm 1.

Estimator MSE at High SNR: The MSE at high SNR can be written as $E\{(\delta - \widehat{\delta}_F)^2\} = \frac{f_{\rm re}+f_{\rm im}}{\rm SNR}$, where $f_{\rm re} = \frac{NB_{\rm re}^2 - 2A_{\rm re}B_{\rm re}\rho + \sigma_{\rm re}^2 A_{\rm re}^2}{(A_{\rm re}B_{\rm re} - A_{\rm re}\widetilde{B}_{\rm re})^2}$, $f_{\rm im} = \frac{NB_{\rm im}^2 + 2A_{\rm im}B_{\rm im}\rho + \sigma_{\rm im}^2 A_{\rm im}^2}{(A_{\rm im}B_{\rm im} - A_{\rm im}\widetilde{B}_{\rm im})^2}$,

$$\begin{aligned}
\sigma_{\rm re}^2 &= 3N - 4\gamma_1 - 2(-1)^{k_p}(\gamma_{2k_p+1} - \gamma_{2k_p-1}) \\
\sigma_{\rm im}^2 &= 3N - 4\gamma_1 + 2(-1)^{k_p}(\gamma_{2k_p+1} - \gamma_{2k_p-1}) \\
\rho &= (-1)^{k_p}(\gamma_{2k_p+1} + \gamma_{2k_p-1})
\end{aligned}$$

and $\gamma_k = \sin\left(\frac{\pi}{2N}k\right)$. The derivation of asymptotic MSE expression is given in Appendix E.

The analytical complexity of the MSE expression is due to the correlation of noise between neighboring DFT bins due to $2N$-point DFT operation. Retrospectively speaking, we have utilized the analytical MSE expression to optimize the parameter a of the invariant function $\frac{\widetilde{R}_{re}[1]-\widetilde{R}_{re}[-1]}{a\widetilde{R}_{\rm re}[0]-\widetilde{R}_{\rm re}[1]-\widetilde{R}_{\rm re}[-1]}$ and observed that $a = 2$ results in the smallest MSE value, [23].

4. Numerical Results

We compare the performance of the suggested estimator with the state-of-the art estimators in this section. The performance comparisons are conducted at challenging operational conditions of short data records ($N = 16$) and large frequency separation from DFT bins i.e. with an odd valued k_p and a rather small δ due to the definition of $\omega = \frac{2\pi}{2N}(k_p + \delta)$. The signal phase ϕ is independently sampled from uniform distribution in $[0, 2\pi)$ at each Monte Carlo trial. This enables the utilization of HCRB as a performance bound. The Cramer-Rao bound shown in the figures is calculated with the units of $2N$-DFT bins. Since $k_p + \delta = \omega\frac{2N}{2\pi}$, the product of $\sqrt{\rm ACRB}$ given by (4) and $\frac{2N}{2\pi}$ is the bound for the root mean square error (RMSE) with the unit of $2N$-point DFT-bins.

Accuracy of Iterative Inversion: Figure 3 shows the accuracy of the inverse function mapping method. The unknown frequency is set as $3+\delta$ bins where δ takes values in $[-0.5, 0.5]$. The Taylor series expansion point of the first iteration is taken as $\delta_0 = 0.25$. From Figure 3, it is seen that 10 iterations is sufficient to reach the numerical accuracy of the computing platform. In many cases, it suffices to have fewer iterations. For example, if the desired accuracy or the achievable accuracy at a given SNR is on the order of 1/100 of a DFT bin size, one can choose to terminate the scheme at the 6th iteration, given the information in Figure 3.

Estimator RMSE: Figure 4 shows RMSE of the estimators for a length $N = 16$ input with frequency $k_p + \delta$ DFT bins, where $k_p = 3$ and $\delta = -0.2$. It can seen from Figure 4 that the method of Djukanovi̇c [19] suffers from an error floor due to the estimator bias at high SNR. Suggested method, method of Ye et al. [20] and ML estimator closely track the CRB at a very small SNR gap. We have implemented two versions of the method of Ye et al. [20], denoted as YSA-N and YSA-2N. YSA-N (given in [20]) uses N-point DFT in the coarse localization stage, while YSA-2N (see Algorithm 2 listing) uses $2N$-point DFT[1]. For all practical purposes, estimators except the one of Djukanovi̇c act like ML

[1] We would like to record that YSA-N is the state-of-art estimation method published in the literature [20, 21]. YSA-2N is developed by authors of this study as an extension of the original YSA method. (Details of YSA-2N is in Algorithm 2.) After noticing that YSA-2N significantly outperforms YSA-N, we have included both methods in numerical comparisons for the sake of fairness.

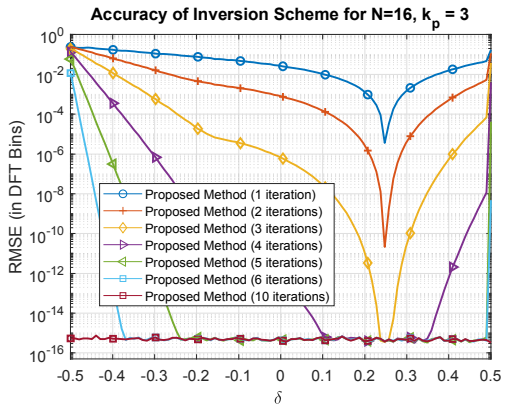


Figure 3: Accuracy of the iterative inversion scheme for different number of iterations.

estimator in the high SNR region. The suggested method, YSA-2N and ML estimator perform almost identically at all SNR values.

On Fusion Operation: Figure 5 studies the success of the fusion operation by comparing RMSE of $\widehat{\delta}_{\rm re}$, $\widehat{\delta}_{\rm im}$ and $\widehat{\delta}_{\rm F}$. The experiment parameters are $N = 16$, $k_p = 3$, $\delta = -0.2$, SNR $= 30$ dB. Figure 5 shows RMSE of the estimates as phase ϕ varies in $[0, 180]$ degrees. In this experiment, the signal phase is taken as a *non-random parameter* to observe its impact on $\widehat{\delta}_{\rm re}$ and $\widehat{\delta}_{\rm im}$.

As noted earlier, depending on the value of $\widetilde{\phi} = \phi + \pi\delta(1-\frac{1}{N}) - \frac{\pi}{N}k_p$, the input SNR of $\widetilde{R}_{\rm re}[l]$ and $\widetilde{R}_{\rm im}[l]$ can vary significantly, affecting the accuracy of the frequency estimates. It is seen that the suggested fusion rule, with the practical fusion coefficient given by (16), successfully combines both estimates so that the final error is almost independent of the signal phase. In Figure 5, CRB relation for finite N, given in [22, p. 56], is utilized. We note that in this experiment, all unknown parameters are non-random; hence, HCRB is not applicable and ACRB is not accurate for $N = 16$.

Computational Complexity Considerations: Numerical results indicate that the performance of suggested invariant function based method, method of Ye et al. [20] (with

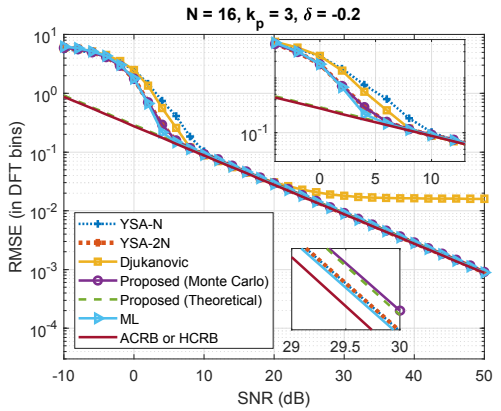


Figure 4: RMSE comparison of the proposed method with other methods.

2N-point DFT in the first stage) and the maximum-likelihood method are almost identical for a wide range of SNR values. If we compare the computational complexity of these three methods, the maximum likelihood by grid-search requires calculation of very high point DFT's to attain the Cramer-Rao bound especially at high SNR values. The method by Ye et al. requires only 2N-point DFT independent of the operational SNR. The computation complexity of this method is essentially determined by the calculation of X_p and $\widehat{a}$ given in the lines 6 and 9 of the Algorithm 2 listing. At every iteration, this method requires $2N$ complex multiplications. Typically, 2 to 5 iterations are required until convergence, [20, 21]. Hence, the complexity of the method by Ye et al. is 2N-point DFT calculation and an order of N complex multiplications (N is the observation vector length). The suggested method requires a 2N-point DFT; evaluation of ratio$_{\rm re}$ and ratio$_{\rm im}$ (line 5 of Algorithm 1 listing) and evaluation of inverse mapping on ratio$_{\rm re}$ and ratio$_{\rm im}$ (lines 6 - 15 of Algorithm 1 listing). In principle, the inverse function can be evaluated with a look-up table, hence its implementation cost can be free of any multiplications. Even if the iterative Taylor series based procedure (lines 9 - 14 of Algorithm 1 listing) is adopted, the cost is independent of number of observations N or SNR. Hence, the suggested method

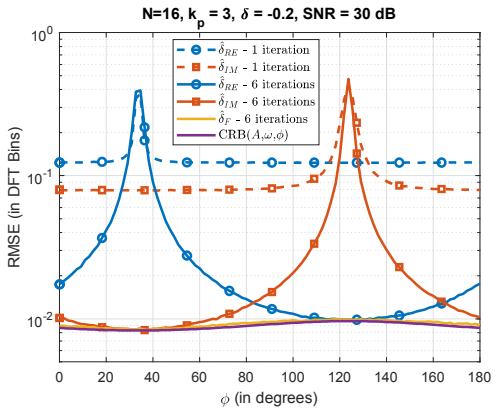


Figure 5: Estimator accuracy comparison before and after fusion operation.

has a complexity which is dominated by a single 2N-point DFT operation.

As an important note, we would like to state that the invariant function based estimators enjoy the extreme reduction of computational complexity not only in this specific application but in general. The main issue of concern is the estimation performance of these methods which changes from an invariant function to another.

Cautionary Remarks: RMSE results similar to the one given in Figure 4 can be obtained for different N, k_p and δ values except for very low ($k_p = \{0, 1\}$) and very high ($k_p = \{N-1, N\}$) frequencies, (see [23] for more simulations). This is essentially due to the special conditions on the DC and the maximum frequency bin. More simply put, the DC bin output is nothing but the summation of all input samples. Hence, DC bin output is purely real for the real-valued sinusoid problem. Therefore, DC bin value is of no value for the calculation of ratio_{im}. In short, very low and very high frequency cases in the real-valued sinusoid application are distinctly different from other cases and they should be individually treated.

5. Conclusions

We present an invariant function approach for parameter estimation and apply the approach on the problem of frequency estimation of real-valued sinusoids observed under AWGN noise. The suggested approach results in a very low complexity estimator performing as well as the maximum likelihood estimator in many scenarios. The suggested method can be either implemented via a look-up table resulting in a one-shot estimator or via suggested numerical method without any look-up table storage requirements. We invite readers to conduct additional Monte Carlo runs with the ready-to-use MATLAB implementation of the suggested method in [25].

We believe that the suggested invariant function approach can be utilized in other parameter estimation problems and can be a computationally efficient alternative to the grid-search based or numerical-search based ML estimators. An extension of the approach to multiple unknown parameters is given in Appendix B.

Appendix A. Maximum Likelihood Estimator

The maximum likelihood estimator for the frequency ω is given in [22, eq(7.65)] and [3]. We present a compact expression in terms of the discrete time Fourier transform (DTFT) of the input.

The maximum likelihood expression can be written as $\arg\max_\omega ||\mathbf{P}_\omega \mathbf{r}||^2$ where $\mathbf{P}_\omega$ is the projection matrix to sub-space spanned by $\boldsymbol{c} = [1 \;\; \cos(\omega) \ldots \cos(\omega(N-1))]^T$ and $\boldsymbol{s} = [0 \;\; \sin(\omega) \ldots \sin(\omega(N-1))]^T$, see [22, eq(7.65)] for more details. To get an analytical expression for the projection operation, we define $\widehat{\mathbf{s}} = \mathbf{s}_\text{u} + \mathbf{c}_\text{u}$ and $\widehat{\mathbf{c}} = \mathbf{s}_\text{u} - \mathbf{c}_\text{u}$ where $\mathbf{s}_\text{u} = \mathbf{s}/||\mathbf{s}||$, $\mathbf{c}_\text{u} = \mathbf{c}/||\mathbf{c}||$. The vectors $\widehat{\mathbf{s}}$ and $\widehat{\mathbf{c}}$ are orthogonal vectors in the span of $\boldsymbol{s}$ and $\boldsymbol{c}$. By normalizing $\widehat{\mathbf{s}}$ and $\widehat{\mathbf{c}}$ to unit norm; we get an orthonormal set of basis vectors for the span of $\boldsymbol{s}$ and $\boldsymbol{c}$. Hence, the maximum likelihood expression can be written as:

$$\begin{aligned}
\widehat{\omega} &= \arg\max_\omega ||\mathbf{P}_\omega \mathbf{r}||^2 \\
&= \arg\max_\omega \frac{(\widehat{\mathbf{s}}^T\mathbf{r})^2}{||\widehat{\mathbf{s}}||^2} + \frac{(\widehat{\mathbf{c}}^T\mathbf{r})^2}{||\widehat{\mathbf{c}}||^2} \qquad \text{(A.1)} \\
&= \arg\max_\omega \left[\frac{(\mathbf{s}_\text{u}^T\mathbf{r} + \mathbf{c}_\text{u}^T\mathbf{r})^2}{2(1+\mathbf{s}_\text{u}^T\mathbf{c}_\text{u})} + \frac{(\mathbf{s}_\text{u}^T\mathbf{r} - \mathbf{c}_\text{u}^T\mathbf{r})^2}{2(1-\mathbf{s}_\text{u}^T\mathbf{c}_\text{u})}\right], \qquad \text{(A.2)}
\end{aligned}$$

where $||\widehat{\mathbf{s}}||^2 = 2(1+\mathbf{s}_{\rm u}^T\mathbf{c}_{\rm u})$ and $||\widehat{\mathbf{c}}||^2 = 2(1-\mathbf{s}_{\rm u}^T\mathbf{c}_{\rm u})$. By adding two terms forming the argument of arg max in (A.2) and substituting $\mathbf{s}_{\rm u} = \mathbf{s}/||\mathbf{s}||$ and $\mathbf{c}_{\rm u} = \mathbf{c}/||\mathbf{c}||$; we get

$$\widehat{\omega} = \arg\max_{\omega} \frac{(\mathbf{s}^T\mathbf{r})^2||\mathbf{c}||^2+(\mathbf{c}^T\mathbf{r})^2||\mathbf{s}||^2-2(\mathbf{s}^T\mathbf{r})(\mathbf{c}^T\mathbf{r})(\mathbf{s}^T\mathbf{c})}{||\mathbf{s}||^2||\mathbf{c}||^2-(\mathbf{s}^T\mathbf{c})^2}. \tag{A.3}$$

Next, we derive the expressions for $||\mathbf{c}||^2, ||\mathbf{s}||^2$ and $\mathbf{s}^T\mathbf{c}$ by observing that

$$\begin{aligned} R+jX &= \sum_{n=0}^{N-1} e^{j2\omega n} = \sum_{n=0}^{N-1} \cos(2\omega n) + j\sin(2\omega n) \\ &= e^{j\omega(N-1)}\frac{\sin(\omega N)}{\sin(\omega)} \end{aligned} \tag{A.4}$$

and $R = \sum_{n=0}^{N-1}(\cos^2(\omega n) - \sin^2(\omega n)) = ||\mathbf{c}||^2 - ||\mathbf{s}||^2$ and $X = 2\sum_{n=0}^{N-1}\sin(\omega n)\cos(\omega n) = \mathbf{s}^T\mathbf{c}$. Also noting that $||\mathbf{c}||^2 + ||\mathbf{s}||^2 = N$, we get $||\mathbf{c}||^2 = (N+R)/2$, $||\mathbf{s}||^2 = (N-R)/2$ and $\mathbf{s}^T\mathbf{c} = X/2$. By substituting these expressions in (A.3), the ML expression becomes

$$\begin{aligned} \widehat{\omega} &= \arg\max_{\omega} \frac{2(\mathbf{s}^T\mathbf{r})^2(N+R)+2(\mathbf{c}^T\mathbf{r})^2(N-R)-4X(\mathbf{s}^T\mathbf{r})(\mathbf{c}^T\mathbf{r})}{N^2-R^2-X^2} \\ &= \arg\max_{\omega} \frac{2N((\mathbf{s}^T\mathbf{r})^2+(\mathbf{c}^T\mathbf{r})^2)+2R((\mathbf{s}^T\mathbf{r})^2-(\mathbf{c}^T\mathbf{r})^2)-4X(\mathbf{s}^T\mathbf{r})(\mathbf{c}^T\mathbf{r})}{N^2-(R^2+X^2)} \\ &= \arg\max_{\omega} \frac{2N|R(e^{j\omega})|^2+2\mathrm{Re}\{R^2(e^{j\omega})(R+jX)\}}{N^2-(R^2+X^2)} \end{aligned} \tag{A.5}$$

where $R(e^{j\omega})$ is the discrete-time Fourier transform (DTFT) of $r[n]$ which is $R(e^{j\omega}) = (\mathbf{s}^T\mathbf{r})+j(\mathbf{c}^T\mathbf{r})$. Substituting the expression for $R+jX$ and its magnitude square (R^2+X^2) from (A.4) into the last expression, we finalize the derivation of the ML estimator:

$$\widehat{\omega} = \arg\max_{\omega} \frac{|R(e^{j\omega})|^2 + \frac{\sin(\omega N)}{N\sin(\omega)}\mathrm{Re}\{R^2(e^{j\omega})e^{j\omega(N-1)}\}}{N - \frac{\sin^2(\omega N)}{N\sin^2(\omega)}}.$$

Appendix B. Invariant Function Approach for Multiple Unknowns

We generalize the invariant function approach given for a single unknown to multiple unknowns. The procedure is illustrated with two unknowns which are denoted as x_t and y_t. The observation model is given as:

$$r[n] = Ag_n(x_t, y_t) + w[n], \quad n = \{0, \ldots, N-1\}. \tag{B.1}$$

Here $g_n(\cdot,\cdot)$ is a known function of two variables, $w[n]$ is additive white Gaussian noise and A is the nuisance parameter of the problem.

In the coarse-search step, a sparsely populated grid for x and y unknowns is constructed and observations are processed with a bank of matched filters, each of which is matched to a grid-point:

$$R(k,l) = \sum_{n=0}^{N-1} g_n^*(x_k, y_l)r[n]. \tag{B.2}$$

Here $R(k,l)$ denotes the output of the matched filter which is matched to the grid point (x_k, y_l). The goal of matched filtering is to increase SNR so as to reliably locate the grid point which is closest to the true parameter pair (x_t, y_t). (A mismatched processing is also applicable as done in the application example given in Section 3.) The first stage output is the set of indices for the maxima of $|R(k,l)|^2$:

$$(\widehat{k}_p, \widehat{l}_p) = \arg\max_{k,l} |R(k,l)|^2. \tag{B.3}$$

When true parameter pair lies on the grid, first stage processing is the optimal processing which is equivalent to an M-ary composite hypothesis test for the detection of unknown parameters observed under AWGN noise. With high SNR assumption, we have $\widehat{k}_p = k_p$ and $\widehat{l}_p = l_p$ where (k_p, l_p) is the grid point which is closest to the true parameter pair. When true parameter pair is not guaranteed to be on the grid, which is the case in general; we proceed with the second processing stage.

The second stage estimates the off-grid component of the unknown parameters, that is δ_x and δ_y where $x_t = x_{k_p} + \delta_x$, $y_t = y_{k_l} + \delta_y$. Different from Section 3, an invariant function pair, say in the form

$$\begin{aligned} f_1(\delta_x, \delta_y) &= \frac{R(\widehat{k}_p+1, \widehat{l}_p) - R(\widehat{k}_p-1, \widehat{l}_p)}{2R(\widehat{k}_p, \widehat{l}_p) - R(\widehat{k}_p+1, \widehat{l}_p) - R(\widehat{k}_p-1, \widehat{l}_p)} \\ f_2(\delta_x, \delta_y) &= \frac{R(\widehat{k}_p, \widehat{l}_p+1) - R(\widehat{k}_p, \widehat{l}_p-1)}{2R(\widehat{k}_p, \widehat{l}_p) - R(\widehat{k}_p, \widehat{l}_p+1) - R(\widehat{k}_p, \widehat{l}_p-1)}, \end{aligned} \tag{B.4}$$

is required for the estimation of two unknowns. To simplify the presentation, we concatenate invariant functions and introduce a vector-valued $\boldsymbol{f}(\delta_x, \delta_y) = [f_1(\delta_x, \delta_y)\, f_2(\delta_x, \delta_y)]^T$ function. The second stage output is formed with the application of inverse function $\boldsymbol{f}^{-1}(r_1, r_2)$ on the ratios r_1 and r_2. Here r_1 and r_2 denote the ratios on the right side of (B.4). In the absence of noise, the mapping recovers δ_x and δ_y perfectly. When noise is present, the matched filter outputs, $R(\widehat{k}_p, \widehat{l}_p)$, are noisy and the accuracy achieved

by inverse function mapping depends on the manifold functions, first stage processing (matched/mismatched filtering), invariant functions and grid choice. As illustrated for the application in Section 3, a proper choice of invariant functions (as in (9) and (10)) and grid (grid utilized by $2N$-point DFT) can yield ML-like estimator performance at a very low computational complexity.

In Section 3, two invariant functions, $f_{\rm re}(\cdot)$ and $f_{\rm im}(\cdot)$, are suggested to estimate a single unknown δ. It should be remembered that unique to this application a successful fusion rule (based on the independence of estimates resulting from the application of inverse functions $f_{\rm re}^{-1}(\cdot)$ and $f_{\rm re}^{-1}(\cdot)$) is given to combine the estimates. Hence, it is possible to have more invariant functions than number of unknowns provided that the estimates can be fused successfully.

As a final comment, we would like to state that the iterative function inverse method suggested in Section 3 can also be extended to the multi-variable case via multi-variable Taylor series:

$$\boldsymbol{f}\left(\begin{bmatrix}\delta_x\\ \delta_y\end{bmatrix}\right) = \boldsymbol{f}\left(\begin{bmatrix}\delta_{x_0}\\ \delta_{y_0}\end{bmatrix}\right) + \boldsymbol{J}\begin{bmatrix}\delta_x-\delta_{x_0}\\ \delta_y-\delta_{y_0}\end{bmatrix} + \text{h.o.t.} \tag{B.5}$$

Here $(\delta_{x_0},\delta_{y_0})$ is the Taylor series expansion point and $\boldsymbol{J}$ is the standard Jacobian matrix whose entries are the first partial derivatives of invariant functions with respect to δ_x and δ_y. Once, function $\boldsymbol{f}(\cdot,\cdot)$ in (B.5) is approximated with constant and first order terms on the right hand of (B.5), that is by ignoring higher order terms (h.o.t.); the inverse function calculation reduces to a simple calculation involving $\boldsymbol{J}^{-1}$. As in Section 3, the initial Taylor series expansion point can be taken as $(0,0)$ (series expansion at one of the grid points) and approximate function inversion can be iteratively applied by using the estimate of an earlier iteration as the expansion point of the next iteration.

Appendix C. Independence of $\widehat{\delta}_{re}$ and $\widehat{\delta}_{im}$

The mean values of the $\widetilde{R}_{re}[l] = \text{Re}\{\widetilde{R}[l]\}$ and $\widetilde{R}_{im}[l] = \text{Im}\{\widetilde{R}[l]\}$, with $\widetilde{R}[l]$ definition given in (7), have no effect on the statistical properties of these random variables. We take the mean value (the signal component) of the random variables as zero to simplify

the derivation and define

$$S_l = \alpha^{k_p+l}\sum_{n=0}^{N-1} e^{-\frac{j2\pi(k_p-l)n}{2N}} w[n] \tag{C.1}$$

where $l = k-k_p$, $\alpha = e^{\frac{j\pi(N-1)}{2N}}$. S_l is identical to $\widetilde{R}[l]$, given in (7), in the absence of signal term. To determine the statistical properties of $\widetilde{R}_{re}[l]$ and $\widetilde{R}_{im}[l]$, we first examine the auto-correlation of the random variables:

$$\begin{aligned}
E\{S_{l_1}S_{l_2}^*\} &= \alpha^{l_1-l_2}E\left\{\left(\sum_{n_1=0}^{N-1} e^{-\frac{j2\pi(k_p+l_1)n_1}{2N}}w[n_1]\right)\left(\sum_{n_2=0}^{N-1} e^{\frac{j2\pi(k_p+l_2)n_2}{2N}}w^*[n_2]\right)\right\}\\
&\overset{(a)}{=} \sigma_w^2\alpha^{l_1-l_2}\sum_{n=0}^{N-1} e^{\frac{j2\pi(l_2-l_1)n}{2N}}\\
&= \sigma_w^2\alpha^{l_1-l_2}\frac{1-e^{\frac{j2\pi(l_2-l_1)}{2}}}{1-e^{\frac{j2\pi(l_2-l_1)}{2N}}}\\
&= \sigma_w^2\alpha^{l_1-l_2}e^{-j\frac{\pi}{2N}(l_2-l_1)(N-1)}\frac{\sin\left(\frac{\pi}{2}(l_2-l_1)\right)}{\sin\left(\frac{\pi}{2N}(l_2-l_1)\right)}\\
&= \sigma_w^2\frac{\sin\left(\frac{\pi}{2}(l_2-l_1)\right)}{\sin\left(\frac{\pi}{2N}(l_2-l_1)\right)} \triangleq \sigma_w^2 d(l_2-l_1)
\end{aligned} \tag{C.2}$$

where the function $d(x) = \sin\left(\frac{\pi}{2}x\right)/\sin\left(\frac{\pi}{2N}x\right)$ is introduced in the last line of (C.2).

By following the steps of (C.2) almost verbatim, it is also possible to show that $E\{S_{l_1}S_{l_2}\} = \sigma_w^2 d(2k_p+l_1+l_2)$. A rather surprising results is that after the phase multiplication in the step 3 of Algorithm 1 listing, both $E\{S_{l_1}S_{l_2}^*\}$ and $E\{S_{l_1}S_{l_2}\}$ becomes a real-valued function.

The covariance of $\widetilde{R}_{re}[l_1]$ and $\widetilde{R}_{re}[l_2]$ can be written as

$$\begin{aligned}
E\{\widetilde{R}_{re}[l_1]\widetilde{R}_{re}[l_2]\} &= E\{\text{Re}\{S_{l_1}\}\text{Re}\{S_{l_2}\}\}\\
&= \frac{1}{4}E\{(S_{l_1}+S_{l_1}^*)(S_{l_2}+S_{l_2}^*)\}\\
&= \frac{1}{2}\text{Re}\{E\{(S_{l_1}+S_{l_1}^*)S_{l_2}\}\}\\
&= \frac{\sigma_w^2}{2}[d(2k_p+l_1+l_2)+d(l_1-l_2)].
\end{aligned} \tag{C.3}$$

Note that when l_1-l_2 is a non-zero even number, $E\{\widetilde{R}_{re}[l_1]\widetilde{R}_{re}[l_2]\}$ reduces to zero as expected. Following similar lines of derivations, we can also get $E\{\widetilde{R}_{im}[l_1]\widetilde{R}_{im}[l_2]\} = \frac{\sigma_w^2}{2}[-d(2k_p+l_1+l_2)+d(l_1-l_2)]$.

The covariance of $\widetilde{R}_{re}[l_1]$ and $\widetilde{R}_{im}[l_2]$ can also be established as

$$\begin{aligned} E\{\widetilde{R}_{re}[l_1]\widetilde{R}_{im}[l_2]\} &= \frac{1}{4j}E\{(S_{l_1}+S_{l_1}^*)(S_{l_2}-S_{l_2}^*)\} \\ &= \frac{1}{2}\mathrm{Im}\{E\{(S_{l_1}+S_{l_1}^*)S_{l_2}\}\} = 0. \end{aligned}$$

Hence real and imaginary parts of $\widetilde{R}[l]$ are uncorrelated. Since these variables are jointly Gaussian distributed, uncorrelatedness implies independence. Once the independence of $\widetilde{R}_{re}[l]$ and $\widetilde{R}_{im}[l]$ are established, the fine frequency estimates ($\widehat{\delta}_{re}$ and $\widehat{\delta}_{im}$), which are derived from real and imaginary parts of $\widetilde{R}[l]$, are also independent random variables.

Appendix D. On Fusion Coefficient Calculation

The suggested fusion coefficient is given as $\alpha \triangleq \frac{\widetilde{R}_{\mathrm{re}}^2[0]}{\widetilde{R}_{\mathrm{re}}^2[0]+\widetilde{R}_{\mathrm{im}}^2[0]} = \left(1+\left(\frac{\widetilde{R}_{\mathrm{im}}[0]}{\widetilde{R}_{\mathrm{re}}[0]}\right)^2\right)^{-1}$. Here we provide the details on the ratio $\widetilde{R}_{\mathrm{im}}[0]/\widetilde{R}_{\mathrm{re}}[0]$ generating the fusion coefficient. Below, it is shown that

$$\frac{\widetilde{R}_{\mathrm{im}}[0]}{\widetilde{R}_{\mathrm{re}}[0]} = \tan(\widetilde{\phi})\underbrace{\left[\tan\left(\frac{\pi k_p}{2N}\right)\cot\left(\frac{\pi(k_p+\delta)}{2N}\right)\right]^p}_{\approx 1} \approx \tan(\widetilde{\phi}), \tag{D.1}$$

where $p=(-1)^{\widehat{k}_p}$ is the parity of $\widehat{k}_p$, taking the value of 1 or -1 depending on $\widehat{k}_p$ being an even or odd integer.

By substituting the definitions for $\widetilde{R}_{\mathrm{re}}[0]$ and $\widetilde{R}_{\mathrm{im}}[0]$ from (8) into $\frac{\widetilde{R}_{\mathrm{im}}[0]}{\widetilde{R}_{\mathrm{re}}[0]}$, we get

$$\frac{\widetilde{R}_{\mathrm{im}}[0]}{\widetilde{R}_{\mathrm{re}}[0]} = \tan(\widetilde{\phi})\frac{\sin(\frac{B}{N})\sin(A)-\sin(\frac{A}{N})\sin(B)}{\sin(\frac{B}{N})\sin(A)+\sin(\frac{A}{N})\sin(B)} \tag{D.2}$$

with $A=\frac{\pi}{2}\delta$ and $B=\frac{\pi}{2}(2k_p+\delta)$. We note that $\sin(B)=\sin(\pi k_p+\frac{\pi}{2}\delta))=(-1)^{k_p}\sin(A)$. Upon the substitution of $\sin(B)=(-1)^{k_p}\sin(A)$, into (D.2), we get:

$$\frac{\widetilde{R}_{\mathrm{im}}[0]}{\widetilde{R}_{\mathrm{re}}[0]}\cot(\widetilde{\phi}) = \frac{\sin(\frac{B}{N})-(-1)^{k_p}\sin(\frac{A}{N})}{\sin(\frac{B}{N})+(-1)^{k_p}\sin(\frac{A}{N})}. \tag{D.3}$$

Assuming, for now, k_p is an even number, (D.3) reduces to

$$\begin{aligned} \frac{\widetilde{R}_{\mathrm{im}}[0]}{\widetilde{R}_{\mathrm{re}}[0]}\cot(\widetilde{\phi}) &= \frac{\sin(\frac{B}{N})-\sin(\frac{A}{N})}{\sin(\frac{B}{N})+\sin(\frac{A}{N})} = \frac{\sin(\frac{B-A}{2N})\cos(\frac{B+A}{2N})}{\sin(\frac{B+A}{2N})\cos(\frac{B-A}{2N})} \\ &= \tan\left(\frac{B-A}{2N}\right)\cot\left(\frac{B+A}{2N}\right). \end{aligned} \tag{D.4}$$

Inserting $\frac{B-A}{2N}=\frac{\pi k_p}{2N}$ and $\frac{B+A}{2N}=\frac{\pi(k_p+\delta)}{2N}$ into (D.4), results in

$$\frac{\widetilde{R}_{\mathrm{im}}[0]}{\widetilde{R}_{\mathrm{re}}[0]} = \tan(\widetilde{\phi})\tan\left(\frac{\pi k_p}{2N}\right)\cot\left(\frac{\pi(k_p+\delta)}{2N}\right). \tag{D.5}$$

When k_p is an odd number, the numerator of the ratio on the right side of (D.3) is swapped with its the denominator and we have $\frac{\widetilde{R}_{\mathrm{im}}[0]}{\widetilde{R}_{\mathrm{re}}[0]} = \tan(\widetilde{\phi})\left[\tan\left(\frac{\pi k_p}{2N}\right)\cot\left(\frac{\pi(k_p+\delta)}{2N}\right)\right]^{-1}$. Both cases can be summarized as in (D.1).

Appendix E. Derivation of Asymptotic MSE Expression

To derive the asymptotic MSE expression, we assume that noise variance σ_w^2 is sufficiently low such that $\mathrm{ratio}_{\mathrm{re}}$ given in (9) can be approximated as $\mathrm{ratio}_{\mathrm{re}} \approx f_{\mathrm{re}}(\delta) +$ (equivalent-noise). Here the term (equivalent-noise) denotes the equivalent additive noise formed by ignoring noise-cross-noise terms, with the low noise variance (high SNR) assumption.

In the absence of noise, we have $\mathrm{ratio}_{\mathrm{re}}$ is $f_{\mathrm{re}}(\delta) \triangleq \mathrm{ratio}_{\mathrm{re}} = \frac{\widetilde{R}_{\mathrm{re}}[1]-\widetilde{R}_{\mathrm{re}}[-1]}{2\widetilde{R}_{\mathrm{re}}[0]-\widetilde{R}_{\mathrm{re}}[1]-\widetilde{R}_{\mathrm{re}}[-1]} = A_{\mathrm{re}}/B_{\mathrm{re}}$ where we have assumed $\delta-\delta_0 \approx 0$ in (12), i.e. the iterative inversion scheme is executed until convergence. In the presence of noise, $\mathrm{ratio}_{\mathrm{re}}$ becomes

$$\mathrm{ratio}_{\mathrm{re}} = \frac{\frac{A\cos(\widetilde{\phi})}{2}A_{\mathrm{re}}+w_{\mathrm{num}}^{\mathrm{re}}}{\frac{A\cos(\widetilde{\phi})}{2}B_{\mathrm{re}}+w_{\mathrm{denum}}^{\mathrm{re}}}, \tag{E.1}$$

where $w_{\mathrm{num}}^{\mathrm{re}} = \mathrm{Re}\{S_1\}-\mathrm{Re}\{S_{-1}\}$ and $w_{\mathrm{denum}}^{\mathrm{re}} = 2\mathrm{Re}\{S_0\}-\mathrm{Re}\{S_1\}-\mathrm{Re}\{S_{-1}\}$. Here, we use the random variable S_l defined in Appendix C. It is clear that $w_{\mathrm{num}}^{\mathrm{re}}$ and $w_{\mathrm{denum}}^{\mathrm{re}}$ are jointly Gaussian distributed zero-mean random variables. Using the results of Appendix C, we can express the marginal distributions as $w_{\mathrm{num}}^{\mathrm{re}} \sim \mathcal{N}(0,\, N\sigma_w^2)$ and

$$w_{\mathrm{denum}}^{\mathrm{re}} \sim \mathcal{N}\left(0,\, \sigma_w^2\left[3N-\frac{4}{\sin(\frac{\pi}{2N})}-2(-1)^{k_p}\left(\frac{1}{\sin(\frac{\pi(2k_p+1)}{2N})}-\frac{1}{\sin(\frac{\pi(2k_p-1)}{2N})}\right)\right]\right).$$

and the cross-correlation of $w_{\mathrm{num}}^{\mathrm{re}}$ and $w_{\mathrm{denum}}^{\mathrm{re}}$ as

$$E\{w_{\mathrm{num}}^{\mathrm{re}}w_{\mathrm{denum}}^{\mathrm{re}}\} = \sigma_w^2(-1)^{k_p}\left(\frac{1}{\sin(\frac{\pi(2k_p+1)}{2N})}+\frac{1}{\sin(\frac{\pi(2k_p-1)}{2N})}\right).$$

To get the equivalent noise term, we express ratio_{re} as

$$\begin{aligned}\text{ratio}_{\text{re}} &= \frac{\frac{A_{\text{re}}}{B_{\text{re}}} + \frac{w^{\text{re}}_{\text{num}}}{\frac{A\cos(\widetilde{\phi})}{2}B_{\text{re}}}}{1 + \frac{w^{\text{re}}_{\text{denum}}}{\frac{A\cos(\widetilde{\phi})}{2}B_{\text{re}}}} = \frac{\frac{A_{\text{re}}}{B_{\text{re}}} + \widetilde{w}^{\text{re}}_{\text{num}}}{1 + \widetilde{w}^{\text{re}}_{\text{denum}}}\frac{1 - \widetilde{w}^{\text{re}}_{\text{denum}}}{1 - \widetilde{w}^{\text{re}}_{\text{denum}}} \\ &\approx \frac{A_{\text{re}}}{B_{\text{re}}} + \widetilde{w}^{\text{re}}_{\text{num}} - \frac{A_{\text{re}}}{B_{\text{re}}}\widetilde{w}^{\text{re}}_{\text{denum}},\end{aligned}$$

where $\widetilde{w}^{\text{re}}_{\text{num}} = 2w^{\text{re}}_{\text{num}}/(A\cos(\widetilde{\phi})B_{\text{re}})$ and $\widetilde{w}^{\text{re}}_{\text{denum}} = 2w^{\text{re}}_{\text{denum}}/(A\cos(\widetilde{\phi})B_{\text{re}})$ and $w^{\text{re}}_{\text{num}} - \frac{A_{\text{re}}}{B_{\text{re}}}$ is the equivalent noise term which is formed by ignoring second powers of noise at the numerator and denominator of the ratio in (E.2).

The arguments given above when repeated almost verbatim for the ratio_{im} constructed from the imaginary part of $\widetilde{R}[l]$ (step 4 of Algorithm 1), we get the $\text{ratio}_{\text{im}} \approx \frac{A_{\text{im}}}{B_{\text{im}}} + \widetilde{w}^{\text{im}}_{\text{num}} - \frac{A_{\text{im}}}{B_{\text{im}}}\widetilde{w}^{\text{im}}_{\text{denum}}$ where $A_{\text{im}}, B_{\text{im}}$ and $\widetilde{w}^{\text{im}}_{\text{num}}, \widetilde{w}^{\text{im}}_{\text{denum}}$ are similarly defined.

Following (14), the asymtotic mean squared error expressions for $\widehat{\delta}_{\text{re}}$ and $\widehat{\delta}_{\text{im}}$ can be written as $c^2_{N_{\text{re}}}(\text{var}(\widetilde{w}^{\text{re}}_{\text{num}}) + \frac{A^2_{\text{re}}}{B^2_{\text{re}}}\text{var}(\widetilde{w}^{\text{re}}_{\text{denum}}) - 2\frac{A_{\text{re}}}{B_{\text{re}}}E\{\widetilde{w}^{\text{re}}_{\text{num}}\widetilde{w}^{\text{re}}_{\text{denum}}\})$ and $c^2_{N_{\text{im}}}(\text{var}(\widetilde{w}^{\text{im}}_{\text{num}}) + \frac{A^2_{\text{im}}}{B^2_{\text{im}}}\text{var}(\widetilde{w}^{\text{im}}_{\text{denum}}) - 2\frac{A_{\text{im}}}{B_{\text{im}}}E\{\widetilde{w}^{\text{im}}_{\text{num}}\widetilde{w}^{\text{im}}_{\text{denum}}\})$ respectively where $c_{N_{\text{re}}} = \frac{B^2_{\text{re}}}{A_{\text{re}}B_{\text{re}} - A_{\text{re}}B_{\text{re}}}$ and $c_{N_{\text{im}}} = \frac{B^2_{\text{im}}}{A_{\text{im}}B_{\text{im}} - A_{\text{im}}B_{\text{im}}}$.

With the fusion rule of $\widehat{\delta}_F = \cos^2(\widetilde{\phi})\widehat{\delta}_{\text{re}} + \sin^2(\widetilde{\phi})\widehat{\delta}_{\text{im}}$ (see Appendix D), the asymptotic MSE becomes

$$\begin{aligned}E[(\delta - \widehat{\delta}_F)^2] &= \frac{2[E\{\cos^2(\widetilde{\phi})\}f_1 + E\{\sin^2(\widetilde{\phi})\}f_2]}{\text{SNR}} \\ &= \frac{f_{re} + f_{im}}{\text{SNR}}\end{aligned}$$

where $f_{\text{re}} = \frac{NB^2_{\text{re}} - 2A_{\text{re}}B_{\text{re}}\rho + \sigma^2_{\text{re}}A^2_{\text{re}}}{(A_{\text{re}}B_{\text{re}} - A_{\text{re}}B_{\text{re}})^2}$, $f_{\text{im}} = \frac{NB^2_{\text{im}} + 2A_{\text{im}}B_{\text{im}}\rho + \sigma^2_{\text{im}}A^2_{\text{im}}}{(A_{\text{im}}B_{\text{im}} - A_{\text{im}}B_{\text{im}})^2}$,

$$\begin{aligned}\sigma^2_{\text{re}} &= 3N - 4\gamma_1 - 2(-1)^{k_p}(\gamma_{2k_p+1} - \gamma_{2k_p-1}), \\ \sigma^2_{\text{im}} &= 3N - 4\gamma_1 + 2(-1)^{k_p}(\gamma_{2k_p+1} - \gamma_{2k_p-1}), \\ \rho &= (-1)^{k_p}(\gamma_{2k_p+1} + \gamma_{2k_p-1}),\end{aligned}$$

and $\gamma_k = \sin\left(\frac{\pi}{2N}k\right)$. More detailed derivations are available at [23].

Algorithm 1: Proposed Method, (see [25] for a ready-to-use MATLAB implementation)

Input : $r[n]$: N samples of noisy real-valued sinusoid;

Output: $\widehat{\omega} = \frac{2\pi}{2N}(\widehat{k}_p + \widehat{\delta}_{\text{F}})$ rad./sample;

$R[k]$ =fft$(r[n], 2N)$ (2N-point FFT calculation);

$\widehat{k}_p = \underset{0 \le k \le N-1}{\arg\max}\ |R[k]|^2$ (locate maxima in spectrum);

Set $\widetilde{R}[l] = R[l + \widehat{k}_p]e^{j\frac{\pi(N-1)}{2N}(l+\widehat{k}_p)}, \quad l = \{-1, 0, 1\}$;

Set $\widetilde{R}_{\text{re}}[l] = \text{Re}\{\widetilde{R}[l]\}$ and $\widetilde{R}_{\text{im}}[l] = \text{Im}\{\widetilde{R}[l]\}$;

Evaluate ratio_{re} and ratio_{im}; use (9)

if lookup-table exists,

  $\widehat{\delta}_{\text{re}} = f^{-1}_{\text{re}}(\text{ratio}_{\text{re}}), \quad \widehat{\delta}_{\text{im}} = f^{-1}_{\text{im}}(\text{ratio}_{\text{im}})$

else

  Set $\delta^0_{\text{re}} = 0.25$, $\delta^0_{\text{im}} = 0.25$ and maxiter = 10;

  for iteration from 1 to maxiter,

    $\widehat{\delta}_{\text{re}} = f^{-1}_{\text{re}}(\text{ratio}_{\text{re}}, \delta^0_{\text{re}})$; use (14)

    $\widehat{\delta}_{\text{im}} = f^{-1}_{\text{im}}(\text{ratio}_{\text{im}}, \delta^0_{\text{im}})$;

    Set $\delta^0_{\text{re}} = \widehat{\delta}_{\text{re}}$ and $\delta^0_{\text{im}} = \widehat{\delta}_{\text{im}}$;

  end for

end

Evaluate the fusion coefficient α; use (16)

$\widehat{\delta}_{\text{F}} = \alpha\widehat{\delta}_{\text{re}} + (1 - \alpha)\widehat{\delta}_{\text{im}}$;

Return $\widehat{\omega} = \frac{2\pi}{2N}(\widehat{k}_p + \widehat{\delta}_{\text{F}})$.

Algorithm 2: YSA-2N Method, (Extension of [20, 21] to 2N points)

Input : $r[n]$: N samples of noisy real-valued sinusoid;

Output: $\widehat{\omega} = \frac{2\pi}{N}(\widehat{k} + \widehat{\delta})$ rad./sample;

$R[k]$ =fft$(r[n], 2N)$ (2N-point FFT calculation);

$\widehat{m} = \underset{0 \le k \le N-1}{\arg\max} \; |R[k]|^2$ (locate maxima in spectrum);

$\widehat{k} = \widehat{m}/2$;

Set $\widehat{\delta} = 0$ and $\widehat{a} = 0$;

  for iteration from 1 to Q,

    $X_p = \frac{1}{N} \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi}{N}(\widehat{k}+\widehat{\delta}+p)n}$, $p = \pm 0.5$;

    $\widehat{L}_p = \frac{\widehat{a}^*}{N} \frac{1+e^{-j4\pi\widehat{\delta}}}{1-e^{-j\frac{2\pi}{N}(2\widehat{k}+2\widehat{\delta}+p)}}$, and $\widehat{S}_p = X_p - \widehat{L}_p$;

    $\widehat{\delta} = \widehat{\delta} + \frac{1}{2}\mathrm{Re}\left(\frac{\widehat{S}_{0.5}+\widehat{S}_{-0.5}}{\widehat{S}_{0.5}-\widehat{S}_{-0.5}}\right)$;

    $\widehat{a} = \frac{1}{N}\left(\sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi}{N}(\widehat{k}+\widehat{\delta})n} - \widehat{a}^* \frac{1-e^{j4\pi\widehat{\delta}}}{1-e^{-j\frac{4\pi}{N}(\widehat{k}+\widehat{\delta})}}\right)$;

  end for

Return $\widehat{\omega} = \frac{2\pi}{N}(\widehat{k} + \widehat{\delta})$.

References

[1] P. Stoica, R. Moses, Spectral Analysis of Signals, Prentice Hall, 2005.

[2] P. Stoica, List of references on spectral line analysis, Signal Processing 31 (3) (1993) 329–340. doi:https://doi.org/10.1016/0165-1684(93)90090-W.
URL http://www.sciencedirect.com/science/article/pii/016516849390090W

[3] R. Kenefic, A. Nuttall, Maximum likelihood estimation of the parameters of a tone using real discrete data, IEEE J. Ocean. Eng. 12 (1) (1987) 279–280. doi:10.1109/JOE.1987.1145230.

[4] H. C. So, Kit Wing Chan, Y. T. Chan, K. C. Ho, Linear prediction approach for efficient frequency estimation of multiple real sinusoids: algorithms and analyses, IEEE Trans. Signal Process. 53 (7) (2005) 2290–2305. doi:10.1109/TSP.2005.849154.

[5] H. So, F. K. Chan, W. Sun, Efficient frequency estimation of a single real tone based on principal singular value decomposition, Digital Signal Processing 22 (6) (2012) 1005–1009. doi:https://doi.org/10.1016/j.dsp.2012.05.010.
URL http://www.sciencedirect.com/science/article/pii/S1051200412001352

[6] E. Jacobsen, P. Kootsookos, Fast, accurate frequency estimators, IEEE Signal Process. Mag. 24 (3) (2007) 123–125. doi:10.1109/MSP.2007.361611.

[7] C. Candan, A method for fine resolution frequency estimation from three DFT samples, IEEE Signal Process. Lett. 18 (6) (2011) 351–354. doi:10.1109/LSP.2011.2136378.

[8] C. Candan, Analysis and further improvement of fine resolution frequency estimation method from three DFT samples, IEEE Signal Process. Lett. 20 (9) (2013) 913–916. doi:10.1109/LSP.2013.2273616.

[9] E. Aboutanios, B. Mulgrew, Iterative frequency estimation by interpolation on Fourier coefficients, IEEE Trans. Signal Process. 53 (4) (2005) 1237–1242. doi:10.1109/TSP.2005.843719.

[10] L. Fan, G. Qi, Frequency estimator of sinusoid based on interpolation of three DFT spectral lines, Elsevier Signal Processing 144 (2018) 52–60. doi:https://doi.org/10.1016/j.sigpro.2017.09.028.
URL http://www.sciencedirect.com/science/article/pii/S016516841730350X

[11] J.-R. Liao, S. Lo, Analytical solutions for frequency estimators by interpolation of DFT coefficients, Elsevier Signal Processing 100 (2014) 93–100. doi:https://doi.org/10.1016/j.sigpro.2014.01.012.
URL http://www.sciencedirect.com/science/article/pii/S0165168414000309

[12] D. Belega, D. Petri, Frequency estimation by two- or three-point interpolated Fourier algorithms based on cosine windows, Elsevier Signal Processing 117 (2015) 115–125. doi:https://doi.org/10.1016/j.sigpro.2015.05.005.
URL http://www.sciencedirect.com/science/article/pii/S0165168415001747

[13] B. G. Quinn, Estimating frequency by interpolation using Fourier coefficients, IEEE Trans. Signal Process. 42 (5) (1994) 1264–1268.

[14] M. D. Macleod, Fast nearly ML estimation of the parameters of real or complex single tones or resolved multiple tones, IEEE Trans. Signal Process. 46 (1) (1998) 141–148.

[15] U. Orguner, C. Candan, A fine-resolution frequency estimator using an arbitrary number of DFT coefficients, Elsevier Signal Processing 105 (0) (2014) 17–21. doi:http://dx.doi.org/10.1016/j.sigpro.2014.05.013.
URL http://www.sciencedirect.com/science/article/pii/S016516841400231X

[16] Y. Chen, A. H. C. Ko, W. S. Tam, C. W. Kok, H. C. So, Non-iterative DOA estimation using discrete Fourier transform interpolation, IEEE Access 7 (2019) 55620–55630. doi:10.1109/ACCESS.2019.2913747.

[17] Y. Chan, J. Lavoie, J. Plant, A parameter estimation approach to estimation of frequencies of sinusoids, IEEE Trans. Acoust., Speech, and Signal Process. 29 (2) (1981) 214–219. doi:10.1109/TASSP.1981.1163543.